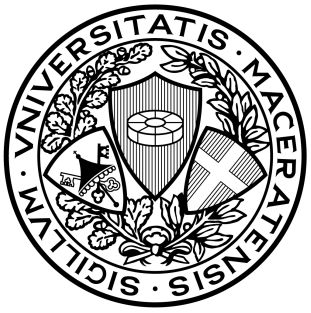


UNIVERSITÀ DEGLI STUDI DI MACERATA

CORSO DI DOTTORATO DI RICERCA IN
QUANTITATIVE METHODS FOR POLICY EVALUATION

CICLO XXXVII

OPENING THE BLACK BOX: NOWCASTING SINGAPORE'S GDP GROWTH AND ITS EXPLAINABILITY

SUPERVISORI DI TESI

Chiar.ma Prof.ssa Rosaria Romano
Chiar.mo Prof. Jamus Jerome Lim

DOTTORANDO

Dott. Luca Attolico

COORDINATORE
Chiar.ma Prof.ssa Margherita Scoppola

ANNO 2025

SDA
scuola di dottorato

Abstract

Effective economic policymaking requires timely assessments of current economic conditions rather than relying solely on past-quarter data, especially in a dynamic and open economic system like Singapore. Promptly identifying turning points becomes critical, considering how international fluctuations can swiftly influence domestic conditions. This real-time assessment, often called "nowcasting," is particularly challenging in rapidly changing economic conditions.

The econometric approach for GDP growth nowcasting has dominated the forecasting landscape for years. Dynamic Factor Models (DFMs) applications in macroeconomic studies have guided substantial research on reducing high-dimensional datasets into a few latent factors that evolve over time. Nonetheless, these factors may introduce noise when distilling key signals from a vast pool of predictors, and they may not consistently guarantee the most substantial forecasting capabilities—ultimately affecting model interpretability.

Traditional forecasting frameworks often encounter notable difficulties in handling structural breaks and crisis periods, emphasizing the need for flexible methods suited to volatile economic settings. Machine Learning (ML) models offer a highly effective path toward more accurate GDP growth nowcasting. Because these methods can approximate intricate nonlinear relationships, they frequently surpass older econometric techniques. In addition, recent advancements in explainable artificial intelligence (xAI) grant policymakers and researchers a clearer perspective on the internal mechanics of these models, enabling more targeted policy decisions. By integrating sub-period analyses, researchers can capture the heterogeneity of economic shocks, enhance real-time surveillance, and refine policy responses for smaller, highly exposed markets.

This research investigates how ML algorithms can improve the accuracy of nowcasting Singapore's GDP growth while clarifying the reasoning behind forecasts. Specifically, we introduce a structured pipeline designed to mitigate look-ahead bias, estimate model reliability via block bootstrap intervals, and highlight key features driving predictions. We also incorporate model combination techniques—such as simple averaging and weighted approaches—to boost forecast stability and monitor the dynamic weights assigned to individual models, thereby offering an interpretive lens for understanding which learners dominate at different points in time. These strategies are complemented by sub-period breakdowns that evaluate performance under varying market phases, revealing a broader perspective on robustness.

We analyze a set of approximately seventy variables, encompassing economic fundamentals (such as prices, production, trade, and investment), demographic factors (like population growth and age distribution), and societal indicators (e.g. employment rates). Data from the Singapore Department of Statistics and additional sources, including The Bank of Italy for external reference indices, span from 1990 Q1 to 2023 Q2. We implement penalized linear models (Lasso, Ridge, Elastic Net), dimensionality reduction approaches (Principal Component Regression, Partial Least Squares), ensemble learning methods (Random Forest, XGBoost), neural architectures (Multilayer Perceptron, Gated Recurrent Unit). Benchmarks such as a naïve Random Walk, an AR(3), and a Dynamic Factor Model are also included for comparison. A Bayesian optimization procedure is employed to select optimal hyperparameters within an expanding-window walk-forward validation framework,

accommodating the temporal dependencies of macroeconomic time series.

Our findings provide reassurance about the stability and reliability of ML-based models in economic forecasting. These models consistently outperform benchmark forecasts, including official projections by major financial institutions, and demonstrate resilience across crisis episodes. On average, penalized linear models, dimensionality reduction approaches, and neural networks consistently achieve substantial reductions in nowcast errors—typically ranging from 40% to 60% compared to established benchmarks such as Random Walk, AR(3), and Dynamic Factor Models. When these strong learners are combined, forecasts achieve further improvements in nowcast errors, as shown by performance metrics.

In addition to GDP growth nowcasting, we employ explainability approaches appropriate to each family model—such as coefficient evolution, Variable Importance in the Projection, Gini-based and permutation-based measures, and Integrated Gradients—to enhance interpretability. By combining these techniques with block bootstrap procedures, we pinpoint which features exert the most significant impact on GDP variations and gauge the inherent uncertainty surrounding them—thereby illustrating, for instance, how changes in foreign trade components or prices can trigger significant shifts in growth projections. Although xAI techniques are gaining traction in fields like image recognition or consumer analytics, they remain relatively unexplored in macroeconomic contexts. This approach enables policymakers to have a deeper understanding of how shocks traverse the economic landscape.

Our study contributes to the broader discourse on ML-based macroeconomic forecasting by showing how sub-period analysis, uncertainty quantification, combined models, and xAI can operate in tandem. Through this integration, we achieve more substantial predictive accuracy and foster greater confidence and clarity in the outputs—factors of particular value for small-scale, internationally interconnected economies. Future work could further refine our methodology by encompassing high-frequency data or trialing innovative ML models that reinforce our framework's versatility and reliability.

Acknowledgments

This thesis was made possible thanks to the institutional and personal support I received over the past years.

I am deeply grateful to the Università degli Studi di Macerata for funding and believing in this research project, and to ESSEC Business School, Asia-Pacific, for hosting me as a Visiting PhD Fellow and providing an intellectually stimulating environment in Singapore.

I wish to thank Dominique Lepore (Università Mercatorum) for immediately believing in my initial idea and for her invaluable advice in transforming it into a coherent research project. I am also grateful to the professors at Università Politecnica delle Marche—Maria Cristina Recchioni, Marco Gallegati (my master's thesis supervisor), Emanuele Frontoni, Donato Iacobucci, and Mauro Gallegati—for their support and for the references that made my doctoral application possible.

A special word of thanks goes to my year in Hong Kong in 2019: Andrea Croci and the late Eric Szeto were true mentors and role models, and they profoundly shaped my relationship with the Asia-Pacific region.

My heartfelt thanks go to my doctoral supervisors, Rosaria Romano (Università degli Studi di Napoli Federico II) and Jamus Jerome Lim (ESSEC Business School Asia-Pacific), for their guidance, patience, and constant support throughout this work. I am equally indebted to Pierre Alquier (ESSEC Business School Asia-Pacific) and Andrea Bucci (Università degli Studi di Macerata) for their mentorship and expert guidance on econometrics and machine learning.

This thesis was conceived and developed as an autonomous research endeavor, with a strong orientation toward real-world applicability: a robust approach to nowcasting, a focus on uncertainty quantification, and interpretable methods that can be transferred beyond academia. I am grateful to all those, in Italy and in Singapore, who have supported me over the years, even simply through moral encouragement; your help has been invaluable and will not be forgotten.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Bridging the Information Gap in Macroeconomics: The Importance of Nowcasting

Timely and accurate macroeconomic indicators are fundamental for effective policymaking and private-sector investment decisions. However, official economic statistics—including key measures such as Gross Domestic Product (GDP), inflation rates, and employment levels—typically exhibit considerable publication delays. These delays introduce an "information gap", imposing significant challenges to policymakers, central banks, and financial institutions that must operate under heightened uncertainty. Crucial economic information frequently becomes publicly accessible only weeks or even months after the respective reference period, thereby limiting the effectiveness, precision, and responsiveness of economic interventions (Bok et al., 2018; Giannone et al., 2008).

Economic researchers have progressively embraced the methodological paradigm known as "nowcasting" to address this gap. Originating from meteorological nowcasting and subsequently adapted to macroeconomic applications, nowcasting aims explicitly to generate contemporaneous assessments of economic conditions. Unlike traditional nowcasting—which predominantly targets future periods—nowcasting leverages timely and readily available high-frequency data to estimate current economic activity, effectively bridging the information lag before official releases (Bańbura & Rünstler, 2011; Baumeister et al., 2021). Consequently, nowcasting substantially reduces uncertainty, enhancing the accuracy and timeliness of economic decisions in both public and private sectors.

The global COVID-19 pandemic underscored the practical necessity and utility of robust nowcasting frameworks. During this period, traditional macroeconomic reporting frequencies proved insufficient, unable to capture rapid shifts in economic trajectories. As a result, prominent institutions such as the U.S. Federal Reserve heavily utilized advanced nowcasting methodologies, notably the GDPNow model from the Atlanta Fed (Higgins, 2014) and the New York Fed Staff Nowcast (Bok et al., 2018). These innovative frameworks successfully integrated unconventional real-time indicators—such as weekly unemployment insurance claims, mobility data, and financial market indices—to facilitate prompt, informed policy responses and clear communication with market stakeholders (Cajner et al., 2020; Foroni et al., 2020).

The benefits of nowcasting extend beyond public policy domains, providing substantial advantages for private investors, financial analysts, and corporate decision-makers. The ability to promptly assess economic conditions enables market participants to proactively manage risks, adjust investment portfolios, and refine strategic operations, thereby reducing informational asymmetries and improving overall market efficiency.

Nevertheless, despite their broad adoption and demonstrable benefits, nowcasting approaches face critical methodological limitations. Notably, these frameworks often lack sufficient interpretability and robust measures of predictive uncertainty. Policymakers and analysts increasingly emphasize the necessity for transparent and interpretable nowcasts,

demanding clear insights into the underlying economic drivers of predictive outcomes. Therefore, further methodological advancements that explicitly incorporate model interpretability and rigorous uncertainty quantification are urgently required to enhance nowcasting's practical utility (Petropoulos & Makridakis, 2020).

In response to these methodological challenges, our research investigates advanced machine learning methodologies tailored to Singapore. Given Singapore's economic openness, the exceptional quality of its macroeconomic data, and heightened sensitivity to global economic fluctuations, it represents a uniquely suitable empirical setting. This research aims to improve nowcasting accuracy, interpretability, and uncertainty management, providing robust tools for policymakers and financial analysts in contexts characterized by structural uncertainties and rapid economic developments. The rationale and potential of these machine learning-based techniques are detailed in the following sections.

1.2 Machine Learning as a Catalyst for Nowcasting

Machine learning (ML) techniques have rapidly gained prominence in macroeconomic nowcasting, driven by their distinctive ability to handle complex, high-dimensional data and effectively adapt to structural changes in economic dynamics. Classical econometric frameworks such as autoregressive (AR), vector autoregressive (VAR), and dynamic factor models (DFM) commonly encounter challenges related to multicollinearity, dimensionality, and non-linear dynamics, particularly under structural breaks and volatile economic conditions (Coulombe et al., 2022; Medeiros et al., 2021). In contrast, ML approaches effectively address these limitations by leveraging advanced computational frameworks and data-driven flexibility, significantly enhancing nowcasting accuracy and robustness in high-dimensional environments (Kim & Swanson, 2018).

A key methodological advantage of ML lies in its ability to efficiently manage high-dimensional, heterogeneous datasets through built-in regularization, adaptive model complexity, and non-linear approximations. Specifically, penalized regressions (LASSO, Ridge, Elastic Net), dimensionality reduction methods (PCR, PLSR), ensemble algorithms (Random Forest, eXtreme Gradient Boosting), and neural networks (MLPs, GRUs) inherently adapt to complex patterns, accommodating data sparsity and non-linear interactions typically overlooked by linear models (Breiman, 2001a; T. Chen & Guestrin, 2016; Hewamalage et al., 2021; Zou & Hastie, 2005).

Recent empirical evidence underscores ML's superior performance relative to traditional econometric benchmarks, particularly during periods of economic instability. For example, Medeiros et al. (2021) demonstrated enhanced nowcasting accuracy of ML techniques for inflation in volatile environments, while Coulombe et al. (2022) confirmed ML's advantages in adapting dynamically to structural economic shifts. The COVID-19 pandemic notably exemplified ML's responsiveness to rapid regime changes, effectively integrating unconventional, high-frequency data sources such as mobility metrics, financial indicators, and digital transactions to achieve timely and accurate nowcasts (Barbaglia et al., 2023; Foroni et al., 2020).

The methodological robustness of ML approaches principally originates from three core attributes:

- **Regularization and Variable Selection:** Penalized methods (e.g., LASSO, Elastic Net) automatically select relevant predictors, preventing overfitting and ensuring

parsimonious, stable nowcasting outcomes (Zou & Hastie, 2005).

- **Ensemble Learning:** Techniques like Random Forest and eXtreme Gradient Boosting aggregate multiple predictions, effectively reducing variance and capturing complex interactions inherent in macroeconomic data (Breiman, 2001a; T. Chen & Guestrin, 2016).

- **Temporal Dynamics Modeling:** Neural network architectures (MLP, GRU) explicitly model sequential dependencies, enhancing adaptability to structural breaks and evolving economic regimes (Cho et al., 2014; Hewamalage et al., 2021).

Nevertheless, despite their empirical successes, ML models frequently suffer from limited interpretability and challenges related to uncertainty quantification. Policymakers and financial analysts increasingly prioritize transparency and robust uncertainty estimates, where traditional econometric methods still hold clear advantages. Consequently, integrating explainable artificial intelligence (XAI) techniques, such as Integrated Gradients or permutation importance measures, becomes critical for enhancing the transparency and credibility of ML-based nowcasts (Lundberg et al., 2020; Sundararajan et al., 2017). Indeed, recent scholarship emphasizes interpretability as a fundamental prerequisite for deploying ML models effectively in high-stakes economic and policy environments (Rudin, 2019). Furthermore, the adoption of non-parametric uncertainty quantification methods, notably bootstrap procedures (e.g., stationary and block bootstrap), is essential to provide rigorous prediction intervals and quantify prediction uncertainty reliably (Li, 2021; Politis & Romano, 1994).

Addressing these interpretability and uncertainty challenges represents a key advancement in ML-based nowcasting research. The methodological framework proposed in this study explicitly tackles these critical dimensions, aiming to strengthen the robustness, transparency, and practical relevance of nowcasts in highly dynamic macroeconomic environments.

1.3 Why Singapore? An Ideal Testbed

Given its economic openness, sensitivity to global shocks, and exceptional data quality, Singapore provides an ideal empirical setting for methodological research on machine learning-based macroeconomic nowcasting.

Given its pronounced openness to trade and financial flows, Singapore rapidly transmits global shocks into domestic indicators like GDP and financial markets, highlighting the urgency for timely nowcasting methodologies.

Singapore's rigorous statistical framework ensures highly accurate and timely publication of key macroeconomic indicators (SingStat, 2023), significantly facilitating validation of advanced econometric and ML methodologies.

Despite these notable advantages, Singapore remains underrepresented in the nowcasting literature, especially regarding sophisticated ML frameworks. While advanced economies such as the United States or Eurozone have been extensively analyzed within empirical ML-based nowcasting studies, Singapore's case has scarcely been investigated, a gap implicitly confirmed by recent comprehensive reviews of nowcasting methodologies (e.g., Foroni et al., 2020). By explicitly addressing this gap, our research provides original

insights into ML performance within an economic context uniquely sensitive to global disruptions.

Existing nowcasting practices within Singapore, i.e., the GDP Nowcasting model regularly published by DBS Bank (DBS Group Research, 2025), predominantly rely on simpler year-over-year (YoY) methods and econometric frameworks lacking in sophistication. These models largely neglect advanced, interpretable ML approaches and do not adequately exploit the granular, quarterly-frequency data available. Introducing comprehensive ML methodologies—including penalized regressions, ensemble algorithms, and neural networks—to nowcast Singapore's quarterly GDP growth directly addresses this methodological shortcoming, promising substantial improvements in nowcasting accuracy, responsiveness, and transparency. Our research uniquely combines Singapore's economic sensitivity and exceptional data quality with advanced, interpretable ML methodologies—explicitly addressing critical gaps unaddressed by current practices and studies.

From a methodological perspective, Singapore's economic structure, combined with its high-quality data, offers an advantageous setting for rigorously evaluating ML models' predictive robustness and adaptability under varied economic conditions. The occurrence of recent economic shocks—such as the COVID-19 pandemic and geopolitical tensions affecting global supply chains—provides an ideal context to empirically assess how ML methodologies manage abrupt regime shifts and structural breaks. Such empirical evaluations yield vital methodological insights, extending applicability beyond the Singaporean context.

Its global financial prominence further enhances our research's practical and international relevance, benefiting policymakers, investors, and multinational firms through improved decision-making and risk management strategies. Indeed, addressing interpretability and uncertainty quantification is increasingly recognized as essential for the practical deployment of nowcasting models in policy-making contexts (Petropoulos & Makridakis, 2020).

By explicitly focusing on Singapore—an economically strategic yet understudied setting—this research bridges a crucial gap in existing nowcasting literature. It advances robust methodological frameworks with significant implications for global economic nowcasting practices.

1.4 Existing Approaches and Related Literature

The literature on macroeconomic nowcasting features a diverse range of methodologies, broadly categorized into traditional econometric models and emerging machine learning (ML) techniques. We critically synthesize these approaches, highlighting their strengths and limitations, and position our research clearly within this established literature.

Dynamic Factor Models (DFM), extensively developed by Bai and Ng (2008) and Stock and Watson (2002, 2011), represent one of the primary econometric tools for nowcasting. DFMs reduce data dimensionality by extracting common latent factors from large macroeconomic datasets. This approach enhances interpretability and nowcasting efficiency by summarizing numerous correlated variables into a few representative factors. Despite their widespread adoption, DFMs exhibit substantial methodological limitations. In particular, the selection of latent factors is inherently arbitrary and relies heavily on subjective econometric criteria, often leading to model misspecification or omitted variable bias. Bańbura et al. (2010) further underlines the complexities of factor selection, espe-

cially when dealing with large-scale datasets. Additionally, DFMs predominantly assume linear relationships among variables, failing to capture non-linear dynamics particularly pronounced during episodes of structural instability, such as the global financial crisis or the COVID-19 pandemic.

Other econometric approaches—autoregressive (AR) and vector autoregressive (VAR) models—are similarly prevalent in the nowcasting literature, particularly due to their straightforward implementation and clear interpretability. However, these methodologies suffer considerably from the "curse of dimensionality[1]", as the incorporation of numerous variables significantly diminishes model performance, resulting in parameter instability and nowcasting inefficiencies. Consequently, traditional econometric models often lack the robustness and adaptability to reflect rapid economic shifts adequately.

In response to these methodological limitations, recent literature increasingly explores ML-based nowcasting techniques. ML methodologies possess inherent advantages in managing complex, high-dimensional datasets characterized by intricate non-linear interactions and structural instabilities (Coulombe et al., 2022; Medeiros et al., 2021; Varian, 2014). This methodological shift has been systematically documented by Kim and Swanson (2018), who underscore ML's comparative advantages in predictive robustness. The principal ML methods in current nowcasting applications can be classified as follows:

- **Penalized Regression Models (e.g., LASSO, Ridge, Elastic Net)**: These methods, originating from statistical learning, incorporate regularization to address multicollinearity and dimensionality issues, providing systematic variable selection and robust predictive performance. Nonetheless, their linear structure limits their ability to capture complex non-linear relationships fully.

- **Dimensionality Reduction Techniques (PCR, PLSR)**: Principal Component Regression and Partial Least Squares Regression effectively compress large datasets into fewer components, mitigating dimensionality and multicollinearity. However, similar to DFMs, their interpretability remains limited, and their linear assumptions restrict performance during pronounced structural breaks.

- **Ensemble Algorithms (Random Forest, eXtreme Gradient Boosting)**: Popularized by Breiman (2001) and Chen and Guestrin (2016), ensemble methods employ multiple decision trees to reduce nowcasting variance and capture non-linear relationships. Although effective, their "black-box" nature constrains interpretability, presenting notable challenges for policy-relevant applications.

- **Neural Networks (MLP, GRU)**: Deep learning models like Multi-layer Perceptrons (MLP) and Gated Recurrent Units (GRU) explicitly accommodate sequential data, offering superior adaptability to structural changes and non-linear dynamics (Cho et al., 2014; Hewamalage et al., 2021). Despite these advantages, their complexity introduces substantial interpretability concerns and increased risk of overfitting.

Recent methodological discourse increasingly highlights the importance of interpretability and uncertainty quantification in nowcasting applications. Traditional econometric

[1]The "curse of dimensionality" refers to the statistical and computational challenges arising from high-dimensional datasets, where the inclusion of numerous variables can deteriorate model performance and nowcasting accuracy due to sparsity, multicollinearity, and parameter instability (Bellman, 1961).

models provide straightforward inferential frameworks (e.g., hypothesis tests, confidence intervals), enabling clear interpretability and robust uncertainty assessment. In contrast, ML methodologies often lack explicit statistical inference, necessitating alternative procedures—primarily non-parametric resampling methods such as stationary or block bootstrap—to deliver rigorous predictive uncertainty measures (Li, 2021; Politis & Romano, 1994).

Moreover, recent literature emphasizes integrating explainable artificial intelligence (XAI) techniques into ML-based nowcasting frameworks. Popular methodologies such as Integrated Gradients, SHAP values, or permutation feature importance have been applied extensively to elucidate ML predictions, thereby increasing transparency and credibility for policymakers and financial analysts (Lundberg et al., 2020; Rudin, 2019; Sundararajan et al., 2017).

Our research explicitly acknowledges and addresses these methodological gaps, proposing a robust ML-based nowcasting framework tailored to the Singaporean economic context. The study systematically implements a comprehensive range of ML methodologies—penalized regressions, dimensionality reduction techniques, ensemble learning algorithms, and neural network models—rigorously evaluated through transparent interpretability methods and robust bootstrap-based uncertainty quantification. This approach substantially improves existing Singaporean practices, currently dominated by traditional econometric and simplified nowcasting techniques, thereby contributing original empirical insights and methodological advancements to the broader nowcasting literature.

1.5 Limitations of Traditional and Emerging Approaches

Despite significant methodological advancements, traditional econometric and emerging machine learning (ML) approaches present noteworthy limitations in macroeconomic nowcasting. Critically examining these constraints provides valuable insights into their practical applicability and robustness, particularly under turbulent economic conditions.

Traditional econometric methods—dynamic factor models, autoregressive (AR), and vector autoregressive (VAR)—typically assume linear relationships and stationarity within data-generating processes. Such assumptions frequently become restrictive during pronounced non-linearities and abrupt structural shifts, limiting predictive accuracy and flexibility (Bańbura et al., 2010; Ng & Wright, 2013; Stock & Watson, 2011). The global financial crisis, the COVID-19 pandemic, and geopolitical disruptions exemplify periods when traditional models often experienced pronounced nowcast deterioration due to their limited ability to dynamically adapt to extreme circumstances (Bańbura et al., 2010; Stock & Watson, 2011). Additionally, the subjectivity in selecting latent factors and variables, particularly within dynamic factor models, introduces risks of omitted-variable bias and model misspecification, thus adversely impacting nowcast stability (Bai & Ng, 2008).

While conceptually superior in managing high-dimensional data and non-linear dynamics, ML methodologies exhibit their distinct limitations. Although theoretically adept at capturing complex economic interactions, empirical evidence on ML models' robustness remains mixed, especially during rapid regime shifts. Techniques such as Random Forest, eXtreme Gradient Boosting, and neural networks have demonstrated substantial predictive instability under abrupt economic changes (Coulombe et al., 2022; Medeiros et al., 2021), raising concerns about their consistency across diverse economic scenarios (Coulombe et al., 2022; Giannone et al., 2021; Medeiros et al., 2021).

Interpretability also remains a prominent challenge for ML approaches. Methods such

as ensemble learners and deep neural networks function predominantly as "black boxes," limiting transparency regarding internal decision processes. This opacity significantly diminishes their direct applicability in policy contexts, where the clarity of nowcasts and accountability are paramount. Recent advancements in explainable artificial intelligence (XAI), including Integrated Gradients, permutation importance, and SHAP values, partially mitigate these interpretability issues; however, achieving comprehensive transparency in ML-driven nowcasting remains an ongoing methodological challenge (Lundberg et al., 2020; Rudin, 2019).

Robust quantification of nowcast uncertainty presents another substantial obstacle within ML frameworks. Unlike traditional econometric methods, which offer well-established inferential techniques (confidence intervals, hypothesis testing), ML models primarily rely on algorithmic optimization, lacking direct statistical inference procedures. Thus, rigorous uncertainty estimation in ML necessitates alternative methods, such as non-parametric stationary or block bootstrap procedures. These bootstrap techniques, essential for valid uncertainty measures, must explicitly accommodate temporal dependencies and structural instabilities, complicating their methodological application and computational feasibility (Li, 2021; Politis & Romano, 1994).

Moreover, predictive reliability in ML methods significantly depends on meticulous hyperparameter tuning, extensive computational resources, and rigorous cross-validation. ML models are prone to overfitting, highlighting the necessity of systematic validation procedures and careful preprocessing of heterogeneous high-frequency data to prevent misleading inferences and spurious relationships (Hastie et al., 2015).

While existing literature underscores ML's potential benefits, empirical evidence consistently indicates substantial heterogeneity in predictive robustness across economic contexts and methodological implementations. Systematic empirical analyses under diverse macroeconomic conditions, especially during structural breaks, remain indispensable for thoroughly understanding the comparative advantages and limitations of ML and traditional econometric methodologies.

Recognizing these gaps, our research explicitly incorporates advanced bootstrap methods for robust uncertainty quantification and sophisticated XAI techniques for enhanced interpretability. This methodological integration directly addresses the limitations identified in traditional econometric and emerging ML approaches, thereby significantly advancing the robustness, transparency, and practical utility of macroeconomic nowcasts under complex economic dynamics.

1.6 Objectives and Contributions of the Research

This study advances the methodological frontier in macroeconomic nowcasting by addressing the following targeted research question: "*How can machine learning (ML) and advanced econometric methods significantly enhance the quality, interpretability, and robustness of quarterly GDP growth nowcasts for Singapore, particularly during periods of structural instability?*" Motivated by clearly identified gaps within both traditional econometric and emerging ML frameworks, we propose an original, rigorous multi-model nowcasting framework tailored explicitly to Singapore's highly sensitive and globally interconnected economic context.

Specifically, our contributions extend the existing nowcasting literature along five principal dimensions:

- *Comprehensive Multi-Model Evaluation*. We rigorously deploy and comparatively evaluate multiple ML methodologies—including penalized regressions (LASSO, Ridge, Elastic Net), dimensionality-reduction techniques (PCR, PLSR), ensemble algorithms (Random Forest, eXtreme Gradient Boosting), and neural networks (MLP, GRU)—to explicitly mitigate model uncertainty and enhance predictive robustness during structural shifts (e.g., the COVID-19 pandemic).

- *Rigorous Uncertainty Quantification via Block Bootstrap*. Addressing ML's limited inferential capabilities, we introduce block bootstrap techniques explicitly designed to handle temporal dependencies and significant structural breaks, including unprecedented economic shocks such as the COVID-19 pandemic. This methodological innovation significantly advances robust quantification of prediction intervals and feature-importance estimates, which are absent in existing ML-based nowcasting literature.

- *Enhanced Interpretability through Model-Specific and XAI Techniques*. Given the policy-critical need for transparent nowcasts, we systematically implement both model-specific interpretability methods—such as coefficient-based feature importance for penalized regressions, gain importance for XGBoost, and mean decrease in impurity (MDI) for Random Forest—as well as Integrated Gradients for our neural models, a XAI technique rarely utilized in macroeconomic nowcasting. Combining these complementary approaches significantly enhances transparency and the practical applicability of complex ML forecasts in policy and economic contexts.

- *Robust Model Selection via Model Confidence Set (MCS) and Nowcast Combinations*. We apply the Model Confidence Set methodology to systematically address model selection uncertainty, rigorously selecting statistically superior models. Furthermore, we implement advanced nowcast aggregation techniques (Weighted Average, Exponentially Weighted Average), substantially reducing individual model biases and enhancing overall nowcasting stability.

- *Novel Empirical Insights from the Singaporean Context*. Our research offers unique empirical evidence on ML performance and adaptability in small, highly open economies by applying advanced ML methodologies specifically within Singapore's understudied yet strategically key context. This geographic-methodological contribution fills an explicit gap identified in recent comprehensive reviews on macroeconomic nowcasting, simultaneously providing a robust methodological benchmark potentially transferable to similar small, open economies frequently exposed to global structural disruptions.

These original methodological and empirical advancements significantly extend existing knowledge, providing robust, transparent, and operationally relevant nowcasting solutions to policymakers, central banks, and financial stakeholders in conditions of pronounced economic volatility and structural uncertainty.

Chapter 2

Methodology and Pipeline

2.1 Tools and Pipeline Organization

2.1.1 Tools

We relied on a dual-language and multi-script workflow to execute our research, combining Python (version 3.x) and R (version 4.x) to exploit each environment's distinctive strengths. All scripts are managed within Spyder and RStudio integrated development environments. This section outlines the libraries employed, the families of nowcasting models under scrutiny, and the pipeline structure devised to accommodate predictive performance, prediction interval estimation, feature-importance analysis, and full-sample and sub-period evaluations.

Families of Models and Evaluations

We analyzed four major families of nowcasting algorithms:

- *Penalized Linear Models:* LASSO, Ridge, and Elastic Net (EN);
- *Dimension-Reduction Models:* Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR);
- *Ensemble Learning Models:* Random Forest (RF) and eXtreme Gradient Boosting (XGB);
- *Neural Models:* Multilayer Perceptron (MLP) and Gated Recurrent Unit (GRU) neural networks;
- *Aggregation/Combination Models:* Simple Average (SA), Weighted Average (WA), and Exponentially Weighted Average (EWA), each generating aggregated forecasts from a subset of best-performing models.

Programming Tools and Libraries

Common and General Libraries for Python-based Tools and Scripts

- `numpy` and `pandas`: supply the core data containers used in every pipeline stage.
  (i) `numpy` arrays underpin all numerical routines (vectorized algebra, rolling operations, linear algebra);
  (ii) `pandas` data-frames store the raw macro series, the recursively generated forecasts, and every intermediate diagnostic table; they also offer built-in indexing and data manipulation;

- `matplotlib` and `seaborn`: provide visualizations for
  (i) residual-diagnostic panels;
  (ii) forecast–vs.–actual plots;
  (iii) stacked area charts tracking the time-varying weights in forecast combinations.

- `statsmodels` and `scipy.stats`: supply econometric tests and distributional tools, such as
  (i) ADF and Ljung–Box for unit-root and serial-correlation checks;
  (ii) Shapiro–Wilk for normality assessment;
  (iii) OLS wrappers, further augmented by a custom Lumley–Heagerty covariance for Giacomini–White predictive-ability regressions.

- `scikit-learn` (`sklearn`): offers general machine-learning utilities. `StandardScaler` ensures homogeneous scaling of features before stationarity checks, whereas `IsotonicRegression` imposes monotonic constraints in the WEAVE (Lumley–Heagerty) covariance approach;

- `logging`, `tracemalloc`, `random`, and `os` handle reproducibility and runtime monitoring:
  (i) a single project-wide random seed ensures replicable bootstrapping steps;
  (ii) `logging` maintains timestamped records of execution phases;
  (iii) `tracemalloc` assists in memory usage tracking;
  (iv) `os` facilitates path management across different systems.

- *openpyxl* or `xlsxwriter`: enable standardized workbook exports. Each file may hold dedicated sheets for:

  - *(i)* point forecasts;
  - *(ii)* RMSFE and coverage measures per sub-period;
  - *(iii)* diagnostic test results, thus allowing external replication without rerunning the entire pipeline;
  - *(iv)* intermediate model outputs, such as MCS rankings, feature importances, and bootstrap resamples.

Specific Python Libraries for Scripted Model Families and Diagnostic Tests

- *Penalized Linear Models (Lasso, Ridge, Elastic Net):*
  `sklearn.linear_model` implements coordinate-descent estimators, while `statsmodels.api.OLS` may be used post-selection for inference. Hyper-parameters are explored through `optuna`, with parallelism via `joblib`. Confidence intervals for coefficients or forecasts are computed through a manual block-bootstrap routine summarized in `scipy.stats`;

- *Dimension-Reduction Models (Principal Component Regression, Partial Least Squares Regression):*
  `scikit-learn` libraries `sklearn.cross_decomposition.PLSRegression` and `sklearn.decomposition.PCA` extract latent factors, while `scipy.linalg.svd` underpins the numerical routines. Bayesian hyper-parameter search uses `optuna`.

Predictive intervals and coefficient bands follow from manual block-bootstrap sampling, summarized with `scipy.stats` for quantiles and confidence limits;

- *Ensemble Learning Models (Random Forest, XGB):*
  `sklearn.ensemble.RandomForestRegressor` and `xgboost.XGBRegressor` train the ensembles. Hyper-parameters are optimized via `optuna`. Feature importance is derived from built-in "feature importances" or permutations; predictive intervals and importance confidence ranges rely on in-house block-bootstrap logic coupled with `scipy.stats` percentile transforms;

- *Neural Networks (MLP, GRU):*
  `torch` is employed for network layers, training loops, and `torch.optim.Adam` for parameter updates. Hyper-parameter tuning uses `optuna`. Feature attribution is done via `captum`'s Integrated Gradients. Block-bootstrap procedures for intervals and attribution bounds are coded internally, with `scipy.stats` providing quantiles;

- *Confidence intervals and feature importance confidence ranges*:
  They rely on resampling techniques, including pair block-bootstrap implemented through
  `arch.bootstrap.MovingBlockBootstrap`, with summary statistics generated using `scipy.stats` for quantiles and confidence limits;

- *Forecast Aggregation Schemes (Simple Average, Weighted Average, Exponentially Weighted Average):*
  Weights are computed in a fully manual way (`numpy`-based) and logged over time. No hyper-parameter tuning is required, though each script saves the dynamic weight evolution and dominant-model frequencies using `matplotlib`;

- *Diagnostic and Stability Tests:*
  ADF, Shapiro–Wilk, and Ljung–Box are from `statsmodels.tsa.stattools` and `statsmodels.stats.diagnostic`. Giacomini–White regressions use the custom `lumley_heagerty.py` module, overriding default OLS covariances with the WEAVE correction for autocorrelated residuals.

Common and General Libraries for R-based Tools and Scripts

- `readxl` and `openxlsx`: simplify the process of importing and exporting data, ensuring the generation of loss matrices or summary tables can be read in R, and vice versa;

- `parallel`: enables multi-core support in the Model Confidence Set bootstraps, distributing the sampling tasks across available cores;

- `stats`: offers core time-series and sampling functions (e.g., `ar`, `filter`) plus descriptive statistics that are invoked throughout the R scripts;

- `base` usage: the Model Confidence Set (MCS) procedure is encapsulated in a custom S4 workflow[1], formalized in an auxiliary source file and instantiated through the class "Superior Set of Models" (SSM), executing the Hansen–Lunde–Nason algorithm[2].

Specialized R Routines Implemented in the Project

- *Model Confidence Set (MCS) Procedure*: a tailored S4 approach conducts the iterative removal of inferior models via block bootstrap replications of the test statistic. Specifically, it
  (i) reads the matrix of losses (squared errors, in our case) for all candidate models,
  (ii) generates B=10,000 contiguous-block resamples to approximate the distribution of the test, and
  (iii) prunes models until a final superior set emerges at a chosen confidence threshold (e.g., 10%). The results, including p-values, ranking, and MCS membership, are stored in the SSM object and exported to a structured data file for documentation.

All scripts implement partial logging, memory cleanup, and a seed fix (e.g., `np.random.seed(42)`, `os.environ["PYTHONHASHSEED"]="42"`, or `set.seed(42)`)[3] to guarantee consistency throughout the entire pipeline.

2.1.2 Research workflow

Our entire research pipeline comprises the following steps:

1. **Data ingestion and stationarity checks.** We import the quarterly dataset, apply Augmented Dickey-Fuller tests to filter out non-stationary features or transformations and incorporate shock dummy variables if needed;

2. **Individual model training.** Each script related to the model families (Penalized Linear Models, Dimension-Reduction Models, Ensemble Learning Models, Neural Networks) is executed, yielding point forecasts, sub-period diagnostics (Shapiro–Wilk, Ljung–Box), and block-bootstrap confidence intervals for forecast distributions or coefficient estimates;

3. **Model Confidence Set.** We merge all loss columns into a single matrix and invoke the MCS procedure in R to remove systematically inferior models at a 10% significance level, returning a final set of top-performing models;

[1] In R, S4 refers to a formal system for object-oriented programming. We employ it here to define the "Superior Set of Models" (SSM) class, which stores MCS outputs (i.e., rankings, p-values, bootstrap details) in a structured manner.

[2] Hansen, Lunde, and Nason (P. R. Hansen et al., 2011) developed an iterative algorithm which discards models found systematically inferior at a given significance level, thus retaining only those with statistically indistinguishable (superior) performance.

[3] The value "42" is often chosen as a "random seed" for reproducibility reasons as an Easter egg: in the humorous science fiction novel "The Hitchhiker's Guide to the Galaxy," British writer Douglas Adams identifies "42" as "the answer to life, the universe and everything" (Adams, 1979). Since then, the number has become a recurring cultural reference in programming languages and technical documentation.

4. **Combination of selected models.** The choice of best-performing individual models for each aggregated method is based on the MCS output, ensuring that only the top-performing models (those selected by the MCS procedure) are included in the ensemble combinations. This procedure iteratively discards inferior models based on bootstrap resampling, retaining only the top-performing models for the final aggregation. The three combined models we used are:

    - *Simple Average:* uniform weighting across MCS members;
    - *Weighted Average:* weighting inversely to partial RMSE up to each time step;
    - *Exponentially Weighted Average:* a two-level scheme that firstly tries multiple η parameters and then applies a second EWA across these "η-experts."

5. **Benchmarking.** We compare each final model and aggregator to three benchmarks—Random Walk (RW), AR(3), and a Dynamic Factor Model (DFM)—and collect RMSFE ratios over the overall sample (Overall) and sub-periods (Pre-COVID, COVID, Post-COVID, and Excluding-COVID)[4];

6. **Predictive-ability tests.** Finally, Giacomini–White regressions are run in Python to assess whether each model or aggregator systematically outperforms the benchmark. We rely on a tailored Python procedure to incorporate a monotonic correlation structure in the residuals, culminating in tabulated Wald tests and p-values stored in formatted data files.

The overall workflow, summarized in Figure 2.1.

This sequential approach ensures that each subtask—from forecast generation and intervals, through the MCS selection and final aggregation to the formal out-of-sample testing—remains consistent and reproducible across Python and R environments. The subsequent sections detail the core methodological elements, including data transformations, hyper-parameter optimization, sub-period evaluations of predictions and their reliability, feature-importance analyses, and interpretability of the aggregated results.

2.2 Data and Preprocessing

In this section, we detail the phases related to the construction of the dataset and the preprocessing techniques adopted to ensure consistency, uniformity, and reliability of the variables in the subsequent one-step-ahead nowcasting process. The central objective is twofold: on the one hand, to have a database adequately refined and free from distortions

[4]The "Overall" period encompasses the entire forecasting horizon analyzed, spanning from 2017 Q1 to 2023 Q2, and thus integrates all specified sub-periods, namely "Pre-COVID," "COVID," and "Post-COVID." This comprehensive timeframe captures the complete range of macroeconomic conditions, including stable economic phases and significant disruptions induced by the COVID-19 pandemic and subsequent recovery period.

The "Excluding COVID" period covers the entire forecasting horizon from 2017 Q1 to 2023 Q2, analogously to the "Overall" period, but explicitly omits the quarters defined as the "COVID" sub-period. This exclusion isolates structural relationships between predictors and economic outcomes in periods free of the extreme economic disruptions associated with the COVID-19 pandemic, providing a reference for model stability and feature importance under comparatively standard economic conditions.

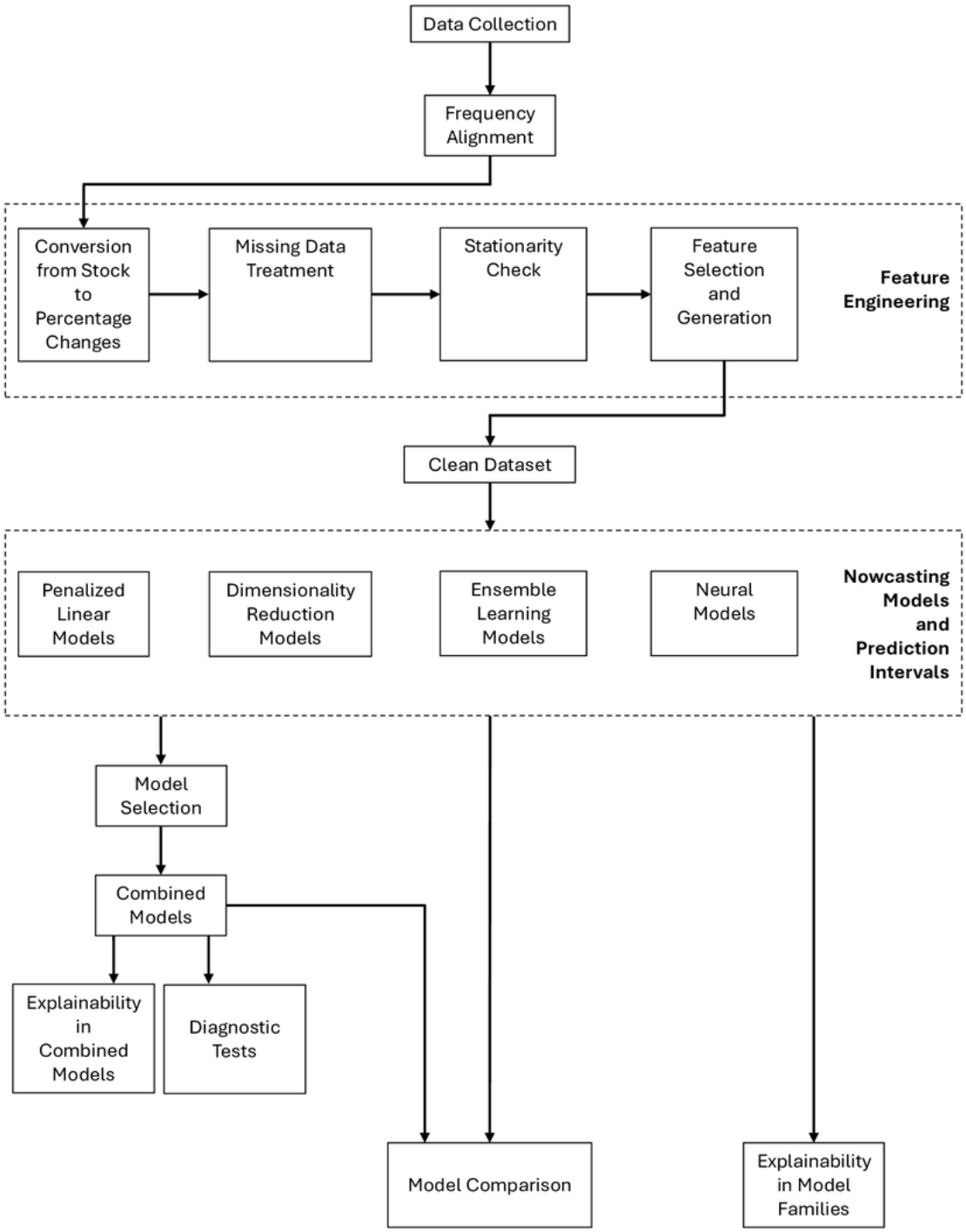


Figure 2.1: Research pipeline workflow

deriving from non-stationarity phenomena or non-uniform frequencies; on the other, to ensure that the transformation and standardization procedures strictly consider the temporal sequentiality, to avoid look-ahead bias (ex-post information contamination). At the end of the initial acquisition operations, the original dataset had 95 candidate variables (features), with coverage from the first quarter of 1990 (1990 Q1) to the second quarter of 2023 (2023 Q2), for a total of 134 quarterly observations. Following the procedures of filling, cleaning, feature engineering, removal of features excessively incomplete or not plausibly linked to economic activity, and exclusion of series that proved to be non-stationary according to the augmented Dickey–Fuller (ADF) test, a core of 58 features has been reached. Five dummy variables were then added to these ones—three for seasonality ("Seasonality Quarter 1", "Seasonality Quarter 2", "Seasonality Quarter 3") and two to identify positive or negative economic shocks ("Past Negative Shocks of Singapore's GDP Quarter over Quarter Growth", "Past Positive Shocks of Singapore's GDP Quarter over Quarter Growth")—bringing the total number of explanatory variables to 63. Together with the target variable (i.e., the deflated GDP growth rate), the final dataset has 64 columns, continuously covering over 134 quarters. This configuration, the result of the balance between maximizing information and minimizing methodological uncertainty, constitutes the solid basis for the nowcasting analyses in the following chapters.

2.2.1 Data Sources: Collection, Integration, and Indexing

To build the dataset, we collected data mainly from statistics produced by the Singapore Department of Statistics (SingStat, 2023). In particular, we used quarterly and monthly historical series relating to the main indicators of the Singapore economy: industrial production, foreign trade, price dynamics, real estate sector, household consumption, transport and airport movements, labor cost indices, balance of payments, exchange rates, business confidence indicators, as well as demographic variables. Each series required a preliminary acquisition, cleaning, and integration process into a single dataset, temporally indexed from the first quarter of 1990 to the most recent period covered by the official archives (second quarter of 2023, in our case). In addition to the SingStat sources, we had to integrate two specific exchange rate series (Euro/SGD and Renminbi/SGD) from the Bank of Italy (Banca d'Italia, 2023) since the corresponding observations in SingStat were incomplete for the period of interest. In particular, for the Euro/SGD, we used a series that also covered the years before the formal introduction of the euro, using data referring to the ECU (European Currency Unit) and the one-to-one correspondence between ECU and euro starting from 1 January 1999. Similarly, in the case of the Renminbi, the SingStat source had gaps up to June 1993, so the series provided by the Bank of Italy offered a more extensive and homogeneous coverage, allowing us to preserve the historical coherence necessary for our analysis horizon. Before proceeding with the transformations, we executed a preliminary check regarding the denomination, the frequency of release, and the correct association with the reference quarter of all the variables to avoid duplication or time misalignment errors, especially in the case of monthly series subsequently converted to quarterly frequency.

2.2.2 Transformations and Feature Engineering

Once the variables were consolidated, we performed transformation and processing operations to make the series comparable and manage the different frequencies.

Target Variable

Singapore's GDP growth at constant prices is our forecasting target in the nowcasting analysis. We obtained the variable from the quarterly GDP data at current prices and the corresponding GDP deflator (base 2015 = 100).

First, we deflated the nominal values of GDP to obtain a series expressed in real terms and adjusted for variations in price levels. Subsequently, the deflated series was transformed into quarterly growth rates (quarter-over-quarter), calculating the percentage difference between the current and previous quarters.

Frequency Alignment

The collected series had heterogeneous frequencies: some were available with monthly periodicity (e.g., price indicators, exchange rates, meteorological variables), while others were already processed in quarterly format (e.g., composite confidence indices or GDP data). Since the target variable is on a quarterly basis, we decided to standardize all monthly series to the same frequency to ensure consistency in the subsequent modeling process. The wide availability of data and the relative ease of converting them made it unnecessary to use specific methodologies for managing mixed frequencies (such as MIDAS models[5]) without compromising the robustness of the analysis. In detail, the choice of the aggregation method (sum, average, or end-of-period value) depends not only on the nature of the variable (e.g., stock vs. flow) but also on the original format of the SingStat publications. In the case of series already expressed as monthly cumulative (e.g., "Contracts Awarded" or "Duty-Paid Releases Of Tobacco"), the sum of the three corresponding months provides the quarterly value. The arithmetic mean over the quarter was calculated for variables released as a monthly average (such as exchange rates or specific price indices). For data cataloged at the end of each period (end-of-period), including "Currency In Circulation," the value recorded in the last month of each quarter was assumed. Although simplified, this approach adequately reflects the sources' structure and ensures a fully quarterly dataset, free from misalignments and gaps.

Conversion from Stock to Percentage Changes

Most of the initial variables reflect stocks or indicators on a nominal basis. We decided to convert these stocks into quarterly percentage changes to analyze the rates of change and the characteristics of macroeconomic nowcasting approaches. This practice not only grants comparability between variables with different reference scales but also allows for mitigating possible non-stationarity phenomena. The only exception concerns the availability of some series that already expressed quarterly differences in absolute terms (such as "Changes in Employment"). Since the original levels to calculate a percentage change were unavailable, we constructed an "Acceleration of Net Employment Flow"

[5]MIDAS (Mixed Data Sampling) models allow the combination of data observed at different sampling frequencies (e.g., monthly and quarterly) in one regression framework, without the need to convert all variables to the same frequency. For a detailed discussion, see (Ghysels et al., 2006)

measure to track the speed with which net job creation varies. Although it is a second difference, we may interpret this indicator as the intensity of employment expansion or contraction.

Missing Data and Variables Not Related to Economic Activity

Some variables began on dates after 1990 Q1. To limit the introduction of noise and preserve the length of the time series, we eliminated those features with ten or more missing observations in the initial part of the sample (e.g., "Landed Properties Supply Change" or "Foreign Trade Index Change"). We opted for the forward-filling technique for variables with a few missing initial values (less than ten), i.e., translating the last known value to the subsequent quarters in which it was missing. We may justify this imputation by the hypothesis that, in the initial phase of the analysis period, there were no international or internal shocks to the Singapore economy that could cause structural changes that would make this technique inadequate. Finally, we removed some variables that were poorly correlated with the performance of Singapore's real economy, such as elementary demographic data (change in live births, change in deaths) and meteorological variables (changes in air temperature and sunshine, relative humidity and rainfall) whose influence was neither plausible nor theoretically justifiable in a GDP growth forecasting context. This filtering allowed us to focus only on indicators intrinsically linked to macroeconomic developments.

Stationarity

Although the conversion to growth rates helps reduce the presence of trends or seasonality, we further verified the stationarity of each variable using the Augmented Dickey-Fuller (ADF) test, adopting a conventional significance level (5%). We eliminated those series that remained non-stationary even after the transformation (including changes in personal saving rate and electricity generation). This approach reflects the need to build reliable predictive models, avoiding the persistence of trends or unit-roots generating misleading results in the estimation phase.

Tracking Seasonality and Positive/Negative Economic Shocks

To present the non-artificial nature of the features and the target variable and allow the model to recognize seasonal patterns through the associated coefficients, we chose to handle seasonal effects by creating dummy variables (i.e., "Seasonality Quarter 1", "Seasonality Quarter 2", "Seasonality Quarter 3") rather than proceeding with an ex-ante seasonal adjustment.

Regarding macroeconomic shocks, we defined two additional dummies (i.e., "Past Negative Shocks of Singapore's GDP Quarter over Quarter Growth", "Past Positive Shocks of Singapore's GDP Quarter over Quarter Growth") to mark quarters characterized by strongly negative GDP changes or particularly pronounced rebounds. To this end, we adopted a threshold-based criterion (–2.5% to identify significant declines and +5% to recognize robust recoveries), in line with the literature that uses percentage thresholds to distinguish episodes of contraction or exceptional growth (Claessens et al., 2012; Hausmann et al., 2005; Kannan et al., 2009).

Integrating these values with the history of significant economic and financial events

for Singapore (e.g., the Asian financial crisis of the late 1990s, the Great Recession of 2007–2009, the COVID-19 pandemic) helps avoid ambiguous labeling between normal oscillations and shocks of significant magnitude, thereby more clearly distinguishing growth peaks from a context of statistical normality.[6]

Iterative Standardization of Non-Dummy Variables

The last phase of preprocessing is represented by the iterative standardization of continuous variables so that each forecast window reflects the actual information availability of the moment and avoid look-ahead bias (i.e., it does not benefit from future temporal knowledge). Specifically, at each nowcasting step, we separated the portion of data considered "training" up to that point and calculated the mean and standard deviation for each variable. These parameters were then applied to the validation and test subset to normalize the data consistently with the conditions present during training. The dummy variables were not involved in the standardization process, as they did not need to be scaled. We repeated this mechanism iteratively as the training time horizon expanded, ensuring a correct simulation of the actual practice of nowcasting (in which models are regularly updated as new observations become available) and preventing look-ahead bias as a source of overestimation of the predictive performance.

The phases described in this section ensure that our input dataset's economic variables are consistent, stationary, and comparable while preserving historical sequentiality. The choices made regarding exclusion, imputation, and transformation of the variables constitute a balanced compromise between maximizing the amount of information available and safeguarding the statistical robustness of the nowcasting experiments illustrated in the following chapters.

2.3 Nowcasting: Methodologies, Baseline Models, and Hyperparameter Optimization

2.3.1 Expanding+Rolling Window Methodology

The integrated implementation of the Expanding Window and Rolling Window methodologies forms the basis of our Singapore GDP quarterly nowcasting framework (Hyndman & Athanasopoulos, 2021; Inoue & Rossi, 2012; Tashman, 2000). This solution allows us to draw on the entire history of the data, maintaining a dynamic validation criterion that, at each iteration, provides updated feedback on the model's performance. At the same time, it is consistent with the need for a limited forecast horizon and accurate calibration of the hyperparameters, enhancing the robustness of the entire econometric analysis framework.

We adopt a walk-forward evaluation scheme with an expanding training window and a 12-quarter rolling validation window, followed by a one-step-ahead test quarter. This design preserves temporal ordering and avoids look-ahead bias while continuously assessing performance on the most recent data. The basic idea is to build a data window at each step that includes the training period and an updated validation set and then perform the estimate for a future quarter. This scheme preserves the flexibility needed to apply to

[6]For an in-depth analysis of threshold-based approaches to identifying recessions and expansions, see also (Barro & Ursúa, 2008) and (Bry & Boschan, 1971).

different families of algorithms (Tashman, 2000). A schematic of the Expanding+Rolling Window procedure is shown in Figure 2.2.

The choice of a progressively expanding training window (Expanding Window) is based on the principle of valorizing the entire information heritage from the first available quarter to the last, without discarding remote data that might contain relevant structural signals. In a macroeconomic context, this approach is sometimes preferable to a pure Rolling Window since it facilitates the identification of long-term patterns in the presence of external shocks or cyclical changes that characterize open economies (Clements & Hendry, 1998; Pesaran & Timmermann, 2007; Stock & Watson, 1996). In our implementation, we initially allocate approximately 80% of the available quarters to the training stage (ending at Q4 2016). [7] Subsequent iterations expand the training boundary by one quarter each time, preserving the entire economic history of Singapore (Giannone et al., 2008).

In parallel, we applied the Rolling Window concept to define a validation subset that "slides" forward by one quarter at each iteration. In this way, the model is fine-tuned on an expanded training set but always validated on the last 12 quarters (3 years) immediately preceding the forecast (Bańbura & Rünstler, 2011; Pesaran & Timmermann, 2007). Choosing a 12-quarter window reflects a balance between reactivity to recent changes and stability of the validation error, a common practice in short-horizon macroeconomic forecasting (Inoue & Rossi, 2012). [8]

The integration between the Expanding Window and the Moving Window thus represents a trade-off between maximizing historical information and evaluating the model on data consistent with the recent economic situation. Extending the training window allows us to capture long-term patterns, while the moving validation portion highlights any instability or regime changes (Clark & McCracken, 2009).

Once hyperparameters are selected based on the last 12 quarters of validation, the model is then re-fitted on the entire training-plus-validation window to incorporate all available data. This final fit forms the basis for generating the one-step-ahead forecast on the immediately following quarter, thus ensuring that the model parameters reflect the most up-to-date information collected prior to the test horizon.

At the end of each iteration, the forecast is made for the immediately following quarter (the Test Set Window). The result is a one-step-ahead forecasting mechanism that maximizes the use of historical information and isolates a subset of data to fine-tune the hyperparameters. Furthermore, this nowcasting procedure aligns with the operational practice of economic research centers and the mainstream literature on quarterly forecasts since it systematically updates the estimates for each newly available quarter (Bańbura & Rünstler, 2011; Giannone et al., 2008).

[7] An 80–20 division of the dataset is a common heuristic in time-series forecasting (Hyndman & Athanasopoulos, 2021), balancing the need for a sufficiently large training set with the requirement to retain enough observations for robust validation and testing.

[8] Shorter validation windows (e.g., 4–8 quarters) often yield higher variance in the validation metric, while significantly longer windows may dilute the impact of recent structural shifts. Three years is frequently adopted by institutions and empirical studies, as it provides enough recent data without overwhelming the validation with older observations (Clark & McCracken, 2009).

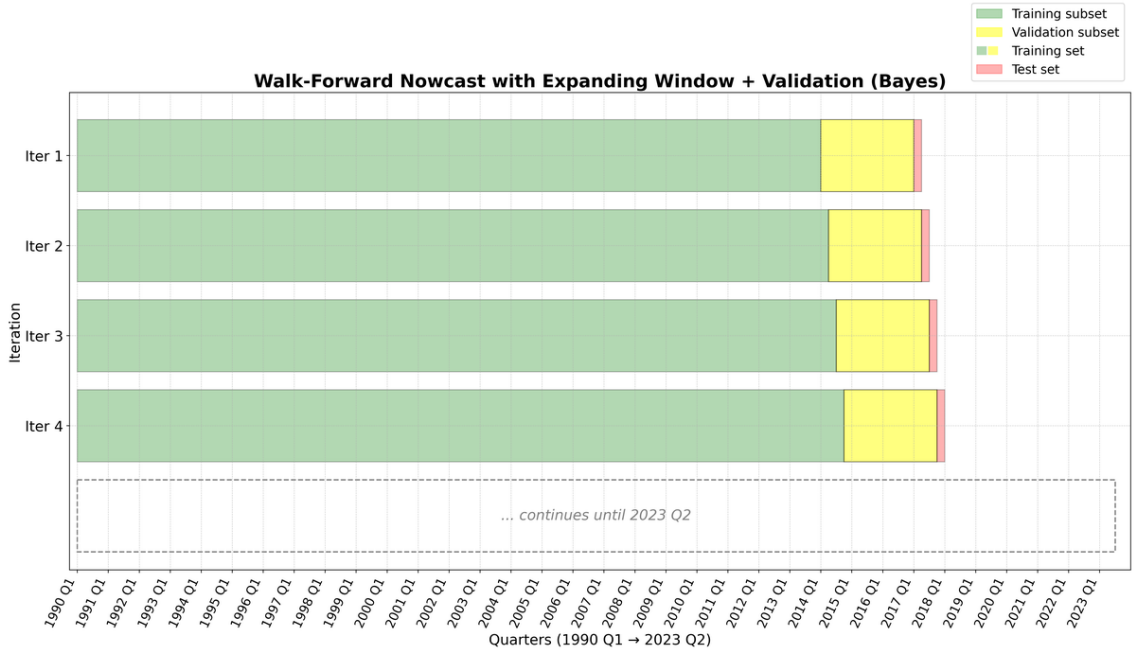


Figure 2.2: Walk-forward nowcast with expanding+rolling window procedure

Expanding Training Set Window

In the first iteration, we set 2016 Q4 as the final date of the initial raining Set Window (green+yellow area), covering about 80% of the initial observations. Each subsequent iteration extends the training subset boundary (green area) by an additional quarter, including progressively newer observations. This incremental approach leverages the entire economic history available for Singapore while ensuring that the model receives fresh data at each step (Giannone et al., 2008).

Rolling Validation Subset Window

We used the last 12 quarters (three years) of the Training Set Window to fine-tune the hyperparameters during each iteration. This Validation Subset Window (yellow area) shifts one quarter forward after each step, ensuring continuous performance examination of recent data (Inoue & Rossi, 2012; Pesaran & Timmermann, 2007). Twelve quarters are frequently cited in the macro-forecasting literature as a window length that balances the need to detect short-run regime shifts without discarding essential cyclical information (Bańbura & Rünstler, 2011; Clark & McCracken, 2009).

One-step-ahead Nowcasting and Test Set Window

Once the training-validation block is built, we use the immediately following quarter as the Test Set Window (pink area) for the out-of-sample evaluation. Adopting a one-step-ahead horizon is standard in nowcasting exercises: it reduces error propagation across multiple steps and focuses on the next immediate quarter, which is typically of greatest policy and practical interest (Marcellino et al., 2006; Tashman, 2000).

The combination of Expanding Window in training and Rolling Window in validation meets the requirements of a quarterly nowcasting system, where the progressive availability of new information consolidates the model's robustness (Giannone et al., 2008). The 12-quarter Validation Window offers a balance between continuity of estimates and attention to the most recent signals, while the single-step forecast horizon reflects the immediate macroeconomic outlook (Bańbura & Rünstler, 2011). It is worth recognizing that exceptional shocks or sudden regime shifts could attenuate the framework's ability to capture new conditions promptly (Pesaran & Timmermann, 2007). In such circumstances, the progressive accumulation of data may be insufficient to compensate for structural changes (Clark & McCracken, 2009). Nonetheless, the iterative mechanism illustrated here retains a high level of adaptability, as it incorporates the most recent observations at each round and, if necessary, allows for the reconsideration of variables or model configurations (Giraitis et al., 2013).

2.3.2 Hyperparameter Tuning in the Nowcasting Process

Validation and selection of hyperparameters are crucial steps in our nowcasting framework, as the optimal configuration of the prediction model's hyperparameters significantly affects forecast accuracy.

In our setup, we integrate hyperparameter optimization into the iterative Expanding+Rolling data scheme described in the previous subsection. This procedure unfolds in several steps, ranging from data loading and validation-split creation to the final choice of the hyperparameter set that minimizes the mean squared error (MSE) on the validation set.

Hyperparameter Optimization: Overview and Rationale for Bayesian Methods

Hyperparameter fine-tuning in a predictive model (e.g., the hidden dimension of a neural network, the regularization strength, or the learning rate) includes various strategies, for example:

- *Grid Search*: It exhaustively evaluates a grid of parameter combinations. Although conceptually straightforward, it becomes infeasible for large or continuous parameter spaces as the number of combinations rapidly grows.
- *Random Search*: This approach randomly samples parameter sets. It typically outperforms grid search in high-dimensional spaces but does not systematically leverage prior search outcomes.
- *Genetic Algorithms*: They "evolve" a population of candidate solutions (i.e., hyperparameter vectors) through mutation and crossover operators inspired by natural selection (Goldberg, 1989; Holland, 1975; Mitchell, 1998). While they can explore large search spaces, they often require numerous generations and careful tuning of evolutionary parameters (e.g., population size or mutation rate).

While each method has its merits, they can become computationally intensive in higher-dimensional scenarios. Grid Search suffers from combinatorial explosion, Random Search fails to incorporate past trial outcomes systematically, and Genetic Algorithms demand multiple generations and specialized parameter tuning. By contrast, Bayesian Optimization

updates its search distribution after every trial, directing the search toward the most promising hyperparameter regions and reducing the computational overhead in our nowcasting setting (Gustafsson et al., 2023; Snoek et al., 2012).

To thoroughly explore potentially optimal model configurations, we adopt broad hyperparameter ranges for all methods tested[9]. Theoretically, wide intervals help prevent overlooking unusual but beneficial hyperparameter values in macroeconomic contexts (Bergstra et al., 2011; Shahriari et al., 2016). Narrow ranges can bias or restrict the model to local optima, especially when relationships are nonlinear or the predictor set is large (Snoek et al., 2012). By contrast, allowing hyperparameters to span orders of magnitude (e.g., for regularization or hidden dimensions) gives the algorithm the flexibility to discover diverse solutions. Practically, macroeconomic nowcasting often involves heterogeneous data, from short, low-frequency series to many high-dimensional indicators (Bańbura & Rünstler, 2011; Varian, 2014). Broader ranges increase robustness, letting TPE and pruning mechanisms find "extreme" values that might be useful under regime shifts or strong nonlinearities (Cerqueira et al., 2020). Bayesian optimization also prunes less promising configurations early (Akiba et al., 2019; Snoek et al., 2012), focusing the search on promising regions without high computational costs. That is especially relevant for rolling or expanding-window procedures, where limiting hyperparameters could exacerbate misspecification risks (Probst et al., 2019; Tashman, 2000).

In the specific case of quarterly nowcasting for Singapore's deflated GDP, with a dataset of 63 explanatory variables and 134 total observations, we need a method that efficiently uses each evaluation and quickly prioritizes promising hyperparameter configurations, thus preventing excessive computational burden. Recent evidence from large Bayesian VARs highlights how careful hyperparameter selection (e.g., shrinkage priors) can markedly improve macroeconomic forecast accuracy[10] (Bańbura et al., 2010; Carriero et al., 2015a; Cimadomo et al., 2022; Giannone et al., 2015).

Bayesian Optimization, specifically the *Tree-structured Parzen Estimator* (TPE) method from the Python library Optuna, offers an attractive compromise since it updates a surrogate model at each new trial (where a trial, within each iteration, corresponds to a training and validation attempt using a given set of hyperparameters, here represented by the vector $\boldsymbol{\theta}$). This surrogate model steadily focuses on the most promising regions of the hyperparameter space.

Core Principles of Bayesian Optimization

Bayesian optimization aims to progressively locate those regions in the hyperparameter space that yield superior results without resorting to an exhaustive or purely random search. In general terms, consider defining an objective function:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathrm{MSE}_{\mathrm{val}}(\boldsymbol{\theta}) = \frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \Big(y_i^{(\mathrm{val})} - \hat{y}_i^{(\mathrm{val})}(\boldsymbol{\theta}) \Big)^2 \tag{2.1}$$

where $\boldsymbol{\theta} \in \mathcal{H}$ is a hyperparameter configuration (e.g., the hidden dimension of a

[9]Recall that these methods are: LASSO, Ridge, Elastic Net, Principal Component Regression, Partial Least Square Regression, Random Forest, eXtreme Gradient Boosting, Multilayer Perceptron, Gated Recurrent Unit.

[10]See also the foundational work on BVARs in (Litterman, 1986).

neural network, a regularization coefficient, or the learning rate), and $\mathcal{H}$ is the space of all admissible hyperparameter combinations. In $\hat{y}_i^{(\mathrm{val})}(\boldsymbol{\theta})$, the model configured by $\boldsymbol{\theta}$ yields the prediction for the i-th observation in the validation set. In contrast, n_{val} is the number of observations in that validation subset. Bayesian optimization iteratively constructs a surrogate model of $\mathcal{L}(\boldsymbol{\theta})$, updating it whenever we obtain a new MSE value for a particular set of hyperparameters.

We employ the Tree-structured Parzen Estimator algorithm in Optuna[11]. TPE splits trials into two classes: those with a "good" error (below a threshold γ) and those with a "less good" error (above or equal to γ). The threshold γ is a dynamic threshold used by TPE to classify trials (i.e., the different hyperparameter sets tested), distinguishing those with a validation error under γ from those with worse performance. Unlike a static approach, γ is often determined by selecting some fraction (e.g., 10%) of the lowest errors so far, and it updates as new results are generated. TPE then constructs two conditional probability distributions:

$$\ell(\boldsymbol{\theta}) = p\big(\boldsymbol{\theta} \mid (\mathrm{MSE} < \gamma)\big) \tag{2.2}$$

$$g(\boldsymbol{\theta}) = p\big(\boldsymbol{\theta} \mid (\mathrm{MSE} \geq \gamma)\big) \tag{2.3}$$

and selects the next hyperparameter configuration $\boldsymbol{\theta}$ by maximizing the ratio $\ell(\boldsymbol{\theta})/g(\boldsymbol{\theta})$. Consequently, the algorithm focuses on upcoming trials in regions of the hyperparameter space that, based on prior observations, appear most promising. This approach is especially effective with a limited dataset (134 quarters, in our case) since each evaluation provides valuable information immediately incorporated into the surrogate model. In practice, the algorithm works in this way:

- *Number of Trials* (n_{trials})*:* a maximum number of evaluations (for instance, 60) is set for each time-window iteration. If $n_{\mathrm{trials}} = 60$, then as many as 60 training+validation cycles—each using a different set of hyperparameters—can occur within that single iteration. The algorithm finally identifies the $\boldsymbol{\theta}^*$ combination that minimizes the average MSE across the entire validation set.

- *Validation Error Computation:* in each trial, the model is trained on the whole training set $(X_{\mathrm{train}}, y_{\mathrm{train}})$ up to a maximum number of iterations—or until the solver's convergence criterion is met—according to the model's specified configuration. It then generates 12 forecasts (one per quarter in the validation set) and computes the mean squared error (MSE) across these 12 points. Based on that MSE, TPE determines whether the trial is "good" $(\mathrm{MSE} < \gamma)$ or "less good."

Pruning Mechanism (Median Pruner) and Reducing Computational Overhead

The algorithm applies a pruning mechanism, i.e., early termination of trials whose partial performance is significantly worse than the median of completed trials at the same training epoch. Pruning avoids expending excessive computational resources on clearly suboptimal configurations. Specifically, the "Median Pruner" in Optuna compares the cumulative validation MSE in the early epochs of each trial to the median MSE recorded among

[11]The TPE algorithm was originated in the Hyperopt library and subsequently integrated into other frameworks, including Optuna.

completed trials. If the trial in progress lags well behind that median, Median Pruner truncates the trial prematurely. This strategy proves highly beneficial in our context, where the limited sample (134 quarters) and the large number of features (63) make it necessary to conserve computational resources. Additionally, a predefined number of "warm-up" trials run to completion, gathering baseline statistics for TPE and balancing exploration against potential exploitation of already-discovered good solutions.

Hence, there can be two levels of early termination in each trial:

- *Internal early stopping*, if the validation MSE fails to improve for 30 epochs,
- *Pruning*, if the trial's partial results are significantly above the median of previous trials.

Iterative Procedure under the Expanding+Rolling Framework

As explained in Section 2.3.1, at each time-window, iteration the data are partitioned as follows:

- *Training set* $(X_{\text{train}}, y_{\text{train}})$: covers the observations from the initial period up to a certain limit, expanded at each step;
- Validation set $(X_{\text{val}}, y_{\text{val}})$: includes the last 12 quarters of the training window, used for hyperparameter evaluation;
- *Test set* $(X_{\text{test}}, y_{\text{test}})$: the immediately following quarter, excluded from both hyperparameter choice and validation, thereby yielding an out-of-sample forecast.

In each iteration, TPE explores up to n_{trials} hyperparameter configurations internal to that same time window. For each trial:

1. The model is trained on $(X_{\text{train}}, y_{\text{train}})$ with the proposed hyperparameters $\boldsymbol{\theta}$.
2. The validation error is computed on $(X_{\text{val}}, y_{\text{val}})$.
3. The pruning procedure can halt those trials with notably poor performance.
4. After exhausting (or pruning) all n_{trials}, TPE selects the $\boldsymbol{\theta}^*$ that minimizes $\text{MSE}_{\text{val}}(\boldsymbol{\theta})$.

Afterward, the model is retrained on $\left(X_{\text{train}} \cup X_{\text{val}}, y_{\text{train}} \cup y_{\text{val}}\right)$ using $\boldsymbol{\theta}^*$, generally with the same training parameters (up to 200 epochs, early stopping at 30). This final model then produces an out-of-sample forecast on $(X_{\text{test}}, y_{\text{test}})$. Shifting the time window by one quarter completes one iteration; repeating this procedure yields a sequence of out-of-sample forecasts that aligns with real-world nowcasting practices.

Adopting TPE alongside a pruning mechanism (Median Pruner) effectively handles the search for hyperparameter configurations over a broad space, even with relatively few quarterly observations (Gustafsson et al., 2023). TPE rapidly directs trials toward promising regions, reducing the full training required compared to a fully random or exhaustive search; pruning halts configurations with high error, saving computational resources. Overall, this synergy iteratively improves the out-of-sample forecasts while maintaining a pipeline sufficiently agile to be updated every quarter.

2.3.3 Handling Shock Dummies to Prevent Look-Ahead Bias

Modeling macroeconomic shocks in a nowcasting context plays a crucial role in correctly interpreting economic evolution. In such a context, using dummy variables that signal disruptive events of a positive or negative nature—such as sudden financial crises, pandemics, or unexpected production recoveries—allows us to emphasize exceptional circumstances that usual cyclical trends cannot recognize (Carriero et al., 2022; Clements & Hendry, 1996; Salkever, 1976). However, such variables must be rigorously managed to ensure that the validation and forecasting process does not incorporate future information that is not yet available during the analysis (look-ahead bias) (Giannone et al., 2021; Lenza & Primiceri, 2022).

During the Expanding+Rolling iterative procedure and in order to prevent any temporal bias, we adopted the following strategy:

1. **Validation subset (12 quarters).** For the first eleven quarters of validation, the shock dummies maintain their effective value, reflecting the availability of complete historical data in retrospect; in the last validation quarter, the dummies are forced to zero, simulating the scenario in which forecasters did not know impending extraordinary events when they generate the validation estimates.

2. **Test set (1 quarter).** Likewise, during the out-of-sample prediction quarter, we neutralize the shock dummies. That ensures that the model's forecast does not inadvertently exploit knowledge of any future disturbance that has not yet appeared in the observable data.

By neutralizing shock dummies only in the last validation quarter and in the test quarter, the procedure reflects a genuine forecasting scenario in which new shocks are unknown until they manifest in real data. At the same time, the model can still learn from historically observed shocks, thereby retaining any insights gained from past anomalous events. This combination of selective neutralization and rolling estimation promotes robust and bias-free short-term projections, thus enhancing the credibility of the overall nowcasting framework.

2.3.4 Model Families and Benchmarks

The absence of an official nowcasting model for Singapore's deflated quarterly GDP growth motivates the adoption of three essential benchmarks, alongside which we apply several families of statistical-mathematical models. On the one hand, we establish a comparison with elementary or standard references (Random Walk, AR(3) selected from various orders, and a specially configured Dynamic Factor Model); on the other hand, we evaluate several types of models that, due to their structure and theoretical assumptions, present characteristics potentially suitable for a small, dynamic and highly integrated economy.

In this section, we describe the fundamental mathematical-statistical principles of benchmarks and nowcasting models, grouped according to their distinct estimation approaches and theoretical premises to provide a solid methodological basis essential to evaluate the effectiveness of each approach:

- **the benchmarks**, which act as fundamental baselines for evaluating the performance of our nowcasting models;

- **the families of nowcasting models**, grouped based on their respective estimation criteria and structural assumptions:
  - *Penalized Linear Models*, which exploit regularization (e.g., Lasso, Ridge, Elastic Net) as a way to handle high-dimensional data and mitigate overfitting;
  - *Dimensionality Reduction Models*, which employs techniques (e.g., Principal Component Regression, Partial Least Squares Regression) to project the original dataset onto a lower-dimensional subspace, preserving the core informational content;
  - *Ensemble Learning Models*, based on the combination of multiple elementary predictors (such as Random Forest or XGBoost) exploiting their internal diversity in order to improve stability and accuracy of predictions;
  - *Neural Models*, which adopt flexible hyperparametric architectures (Multilayer Perceptron, Gated Recurrent Unit) capable of approximating non-linear relationships and capturing temporal dependencies.

Benchmarks

Random Walk

The Random Walk (RW) or "naïve forecast" offers a basic yet essential strategy for predicting GDP growth without imposing structural constraints or including exogenous regressors (Makridakis et al., 1998; Muth, 1960; Tashman, 2000). Its core assumption is that the optimal predictor for period $t+1$ is simply the observed value at time t, thereby forgoing any additional hypothesis about underlying dynamics.

Formally, the model assumes a unit-root process:

$$y_{t+1} = y_t + \varepsilon_{t+1},$$

where:

- y_t represents the target variable (e.g., quarterly GDP growth),
- ε_{t+1} is a stochastic error term with zero mean and constant variance.

Consequently, the one-step-ahead forecast is given by:

$$\hat{y}_{t+1|t} = y_t.$$

This specification implies that any permanent shock is not reabsorbed and, in the absence of further information, the method cannot distinguish a long-term economic trend from a temporary fluctuation (Nelson & Plosser, 1982; Stock & Watson, 1996).

The forecast procedure can be summarized in three steps:

1. Define an initial historical window containing $y_1, y_2, \ldots, y_t$.
2. At time t, assign $\hat{y}_{t+1|t} = y_t$, positing that no supplementary information alters the most recent observed change.

3. Once the actual data y_{t+1} becomes available, compute the forecast error $\hat{y}_{t+1|t} - y_{t+1}$ to assess predictive performance.

The effectiveness of the Random Walk heavily depends on the data-generating process. If the time series predominantly follows a stochastic pattern, the naïve forecast can be highly competitive vis-à-vis more complex techniques (Makridakis et al., 1998; Tashman, 2000). However, in the presence of pronounced deterministic trends or clearly defined cyclical patterns, this approach may underestimate or overestimate structural changes, as it merely projects the latest observation forward without modeling any trend component. Because it requires no parameter or hyperparameter estimation, the Random Walk serves as a fundamental benchmark in forecasting exercises (Makridakis et al., 2022). Its minimalism highlights the gains in predictive accuracy offered by more sophisticated methods, making improvements more credible once compared with a technique that, by definition, integrates no additional information beyond the last observed data point.

Autoregressive Models

Autoregressive (AR) models rank among the most widespread approaches for analyzing economic time series, as they assume that the current value of a variable (e.g., quarterly deflated GDP growth) depends linearly on its past observations. The general form of an AR(p) model is:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

where:

- y_t denotes the endogenous variable, in this case, the deflated quarterly GDP growth at time t;
- α is the intercept that captures the systematic mean level of the series;
- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive coefficients, each measuring the influence exerted by the corresponding lag;
- ε_t is a random error term with zero mean and constant variance, representing shocks not accounted for by the model.

Such a structure enables the model to capture persistence over time in an economic process, especially when prior shocks continue to affect subsequent observations.

An AR(3) model expresses the current value of a time series (here, Singapore's deflated quarterly GDP growth) as a linear combination of its three most recent past values. Formally, this can be written as:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t,$$

where:

- y_t is the deflated GDP growth at time t;
- α denotes the long-run average component (intercept);

- ϕ_1, ϕ_2, ϕ_3 capture the influence of the first, second, and third lag, respectively;
- ε_t represents residual shocks or noise unaccounted for by the three lags (Hamilton, 1994; Marcellino et al., 2006).

This specification goes beyond the simplicity of, for instance, an *AR(1)* or a Random Walk (Nelson & Plosser, 1982), as it can capture short-run dynamics over multiple quarters yet remains more parsimonious than higher-order autoregressive models.

The estimation and forecast procedure for an AR(3) typically follows these steps:

1. **Defining the historical window.** An initial dataset $\{y_1, y_2, \ldots, y_T\}$ is isolated. If a rolling estimation approach is applied, the earliest points may be omitted to allow for progressive model updating (Banerjee & Marcellino, 2006).
2. **Model fitting.** At each iteration, one estimates the intercept α and the coefficients ϕ_1, ϕ_2, ϕ_3 via maximum likelihood or least squares, under standard assumptions of white noise errors (Hamilton, 1994).
3. **One-step-ahead forecasting.** Substituting $\hat{\alpha}, \hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$ and the last three observed values y_t, y_{t-1}, y_{t-2} into the AR(3) equation gives the forecast $\hat{y}_{t+1}$.
4. **Forecast error evaluation.** Once the actual y_{t+1} is released, $\hat{y}_{t+1} - y_{t+1}$ quantifies the prediction error for that quarter (Marcellino et al., 2006)[12].

In our context, AR(3) offers a middle-ground strategy that often outperforms simpler models, such as AR(1) or the Random Walk (Nelson & Plosser, 1982), while remaining more tractable than AR processes of higher order. Its practical effectiveness, however, ultimately depends on the stability of the underlying economic environment and on how frequently large, unanticipated shocks occur.

Dynamic Factor Models

Dynamic Factor Models (DFM) are a widely used tool in macroeconomic analysis. They allow the synthesizing of a large set of variables that describe the dynamic structure of an economic system in a small number of latent factors.

Starting from the studies of (Bai & Ng, 2002), which introduce optimal selection criteria for the number of factors, and of (Stock & Watson, 2002), which show how a limited set of principal components can explain most of the macroeconomic variability, DFMs have evolved by integrating different solutions for the management of incomplete data and heterogeneous frequencies. In particular, (Doz et al., 2011, 2012) employ a state-space

[12] In our analysis of Singapore's GDP nowcasting, we tested multiple AR orders, from AR(1) to AR(4), and compared them by the Root Mean Squared Forecast Error (RMSFE). Overall, AR(3) exhibits the lowest RMSFE, although in certain sub-periods (Pre-COVID, COVID, Post-COVID, Excluding-COVID) other orders may marginally outperform it. Hence, AR(3) is a balanced choice for the entire sample, as it captures up to three-lag dynamics without the additional parameters of higher-order models. Nonetheless, large structural shifts—like those witnessed during the pandemic—can alter which AR order best aligns with specific phases of the economy (Giannone et al., 2021; Lenza & Primiceri, 2022; Marcellino & Schumacher, 2010).

approach [13] and an Expectation–Maximization (EM) algorithm [14], combined with the Kalman filter, to address problems of ragged edge[15] (Bańbura & Rünstler, 2011) or time series with mixed frequency.

This type of approach provides maximum flexibility. However, in contexts where the data are aligned on a single frequency and are complete, these tools may exceed the actual operational needs and be computationally expensive without frequency discrepancies or significant missing data.

From a theoretical point of view, a DFM assumes that the vector
$x_t \in \mathbb{R}^N$,
the vector of the variables observed at time t, is composed of common factors and idiosyncratic components according to the equation:

$$x_t = \Lambda f_t + \varepsilon_t,$$

where:

- $f_t \in \mathbb{R}^r$ represents the latent factors (with $r \ll N$),
- Λ indicates the factor loading matrix and
- ε_t contains the portion of variability specific to each variable (not explained by the factors)[16].

In this work, we adopt a "light" DFM model with a more essential structure than the state-space one. In particular, we extracted the factors f_t through Principal Component Analysis. This choice reflects the perspective of (Stock & Watson, 2002), according to which a few principal factors can explain most of the macroeconomic behavior. To allow for dynamic modulation of the unrelated errors, we consider a structure

$$\varepsilon_t = \Phi(L)\, \varepsilon_{t-1} + e_t,$$

where:

- $\Phi(L)$ is a polynomial in the lag[17] of empirically selected order
- e_t is a white noise error term.

[13] In the state-space approach, we represent a system by a transition equation for the dynamics of latent variables (states) and an observation equation that links the states themselves to the empirical data. A Kalman filter continuously updates the states, even with missing data or non-homogeneous release timing.

[14] EM procedures alternate an Expectation phase, in which the distribution of unobserved variables is estimated based on the current parameters, with a Maximization phase, in which the parameters are updated to maximize the conditional likelihood. This iteration continues until convergence.

[15] With ragged edge, we indicate the situation in which some variables end before others or have different publication delays, generating an irregular structure in the last points of the sample.

[16] In many (Stock & Watson, 2011) implementations, ε_t is often treated as simple white noise or dynamically modeled.

[17] For example, with order p, we have $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$. Applying $\Phi(L)$ to ε_t allows us to introduce dependencies on past observations, making the idiosyncratic an autoregressive process.

Therefore, we preferred an AR for the idiosyncratic term[18], avoiding using a completely state-space model since the data were already aligned and did not present missing values that needed to be estimated iteratively. To automatically search for the optimal number of factors r and the order of $\Phi(L)$ for each quarter of nowcasting, we used the Bayesian-like iterative optimization process on a sliding validation window and pruning described in section 2.3.2. We omitted using Kalman filters and EM procedures since our series were already quarterly and free of missing data. This condition made unnecessary the data reconciliation or imputation mechanisms proposed in (Doz et al., 2011, 2012). During the analysis, we then directly integrated the dummy variables of seasonality and economic shocks into the regression component that links the factors to the target variable without adopting more sophisticated state-space smoothing techniques.

This approach is consistent with the reference literature. However, it simplifies the management of any temporal misalignments and missing data, exploiting a dataset in which the quarterly frequency is already uniform.

Penalized Linear Models

Penalized linear models offer a structured approach to controlling overfitting in contexts where numerous regressors may exhibit strong collinearity or marginal relevance. In the one-step-ahead quarterly nowcasting setting for a small open economy, introducing specific shrinkage terms in the objective function helps reconcile flexibility and parsimony. This mechanism prevents extraneous variables from overshadowing essential economic signals, enhancing predictive accuracy (Hoerl & Kennard, 1970; Tibshirani, 1996; Zou & Hastie, 2005).

At the core of this family are techniques that impose different regularization strategies. Lasso, for instance, incorporates an L1 penalty, which drives some coefficients to zero and facilitates automatic variable selection. Conversely, Ridge employs an L2 penalty, shrinking all coefficients continuously while retaining each predictor. Elastic Net combines L1 and L2 components, balancing feature selection with robust shrinkage. These frameworks adapt to evolving economic relationships and integrate new indicators without excessively complicating the model's structure. Their adaptability is particularly relevant when policy shifts or global disturbances arise, highlighting their utility in dynamic environments.

Several design principles distinguish penalized linear models from unpenalized regressions:

- *Regularization Control*: adjusting the penalty strength governs the trade-off between bias and variance, thus mitigating overfitting.
- *Sparse or Dense Solutions*: Lasso-induced sparsity excludes uninformative regressors, whereas Ridge retains them at reduced magnitudes.

Penalized linear models thus represent robust tools for macroeconomic forecasting, especially when data availability expands incrementally. They can handle variations in sample size and shifting data patterns without resorting to intricate architectures. While

[18]In this autoregressive model, the ε_t variables can retain persistences not captured by the factors. Since we have completed and without frequency mismatches in the quarterly series, we considered the adoption of iterative estimates of the latent states unnecessary.

these methods typically employ linear specifications and do not inherently capture highly nonlinear dynamics, they often balance predictive performance and simplicity. With careful and incremental hyperparameter tuning, they yield short-horizon forecasts that absorb shocks yet remain attuned to underlying trends. This robustness underscores their efficacy in capturing the dynamic nature of economic systems, frequently outperforming unregularized alternatives in volatile and high-dimensional contexts (Smeekes & Wijler, 2018; Uematsu & Tanaka, 2019).

Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression technique for forecasting models when dealing with several explanatory variables (Tibshirani, 1996; Zou & Hastie, 2005). It effectively manages many explanatory variables, including some that may not substantially influence the prediction. In this scenario, LASSO is theoretically helpful for selecting the most significant variables and avoiding overfitting[19] the model to the data, thus optimizing real-time forecasts.

The use of L1 regularization distinguishes LASSO (Hastie et al., 2015). In this context, the parameter λ imposes a penalty equal to the sum of the absolute values of the coefficients $\left(\sum_{j=1}^{p}|\beta_j|\right)$. This mechanism, often referred to as "shrinkage penalty," progressively reduces the coefficients of some variables until they are zero, thereby excluding the less relevant ones. By increasing λ, the extent of this reduction increases, producing a more accentuated "sparseness[20]" effect.

In the case of models with multiple regressors, where there is a high risk of overfitting, the LASSO allows maintaining a good balance between the estimates' precision and the specification's simplicity (Uematsu & Tanaka, 2019; Zou, 2006). LASSO differs from traditional linear regressions by automatically selecting variables based on coefficient values, thereby highlighting the most predictive variables (Mei & Shi, 2024).

From a theoretical point of view, LASSO[21] is based on minimizing a cost function that combines the estimation error with an L1 penalty, proportional to the sum of the absolute values of the coefficients. In summary form, the function to minimize is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \Big\{ \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\big(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j\, x_{ji}\big)^2}_{\text{error term (reduced MSE)}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{L1 regularization}} \Big\},$$

where:

- $\frac{1}{2n}\sum_{i=1}^{n}\big(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j\, x_{ji}\big)^2$ represents the measure of the estimation error (a "reduced form of MSE[22]");

[19]The term overfitting refers to excessive adaptation of the model to the training data, which compromises its generalization ability (Smeekes & Wijler, 2018).

[20]"Sparsity" is the property of L1 penalty to drive some coefficients exactly to zero, thus eliminating unimportant features. This principle was originally introduced by (Tibshirani, 1996), and later extended by (Zou & Hastie, 2005), demonstrating its practical advantages in high-dimensional regression scenarios.

[21]In this discussion, we follow the convention used by the `scikit-learn` library (Pedregosa et al., 2011; Scikit-learn Developers, 2025b), which inserts the term $\frac{1}{2n}$ to normalize the sum of squares and estimates an intercept parameter β_0 separately.

[22]The expression "reduced form of MSE" denotes that, instead of using the canonical formula

- $\lambda \sum_{j=1}^{p} |\beta_j|$ is the L1 penalty (regularization term), which controls how much the coefficients are shrunk (Hastie et al., 2015).

More specifically,

- y_i denotes the observed dependent variable;
- x_i represents the vector of independent variables (features) for observation i;
- β_j indicates the coefficients to be estimated;
- λ is the regularization parameter regulating the intensity of the L1 penalty (the higher λ, the more coefficients are reduced to zero).

Our approach to modeling with LASSO follows these fundamental steps, emphasizing that the penalty parameter λ is determined through an iterative process:

1. **Selection of the parameter λ by optimization and pruning.** The determination of λ, which controls the intensity of the L1 regularization, takes place at each iteration through a Bayesian optimization process based on the TPE and pruning techniques, implemented in Optuna (Gustafsson et al., 2023; Snoek et al., 2012). In each temporal block of data (train+validation), the algorithm searches the space of possible values of λ and selects the one minimizing the mean square error on the validation subset. This procedure is repeated quarter by quarter, thus adapting progressively to the latest economic conditions (Smeekes & Wijler, 2018);

2. **Model training and forecast generation.** Once λ is selected in the single optimization step, the LASSO model is trained on the combined training and validation data, then used to generate the forecast for the test quarter (or period). After each forecast, the training dataset is extended by including the most recent observations, and the process of hyperparameter optimization and training is repeated to account for the continuous inflow of available information. This sequential strategy — where λ is updated with each sample expansion — allows the model to keep track of evolving economic dynamics and reduces the risk of over-fitting in the long run.

Adopting LASSO in this context allows us to choose λ in a robust manner, minimizing the validation error and improving the model's predictive ability in future data (Hastie et al., 2015; Mei & Shi, 2024; Uematsu & Tanaka, 2019).

Ridge

Ridge Regression is a forecasting method used in regression analysis that involves several explanatory variables (Exterkate et al., 2016; Hoerl & Kennard, 1970; Kim & Swanson, 2014; Smeekes & Wijler, 2018). It efficiently handles cases where a large number of variables affect the outcome, even when none of them have a particularly strong effect on their own. In this context, using the L2 penalty allows for keeping all the regressors, even

$\frac{1}{n}\sum_{i=1}^{n}(\dots)^2$, the coefficient $\frac{1}{2n}$ is introduced in front of the squared residuals. This change neither alters the location of the minimum nor the interpretation of the mean squared error, but it simplifies the derivatives and numerical handling during optimization.

the less relevant ones, while ensuring rigorous control over overfitting by reducing the amplitude of the coefficients (Hastie et al., 2009).

The distinctive element of Ridge lies in the penalty that equals the square of the coefficients $\left(\sum_{j=1}^{p} \beta_j^2\right)$. This approach, known as the "L2 type shrinkage penalty," incrementally lowers the coefficients' values without eliminating them completely. By raising α, the degree of contraction experienced by the coefficients increases, thus helping to reduce the model's variance in situations that may be prone to multicollinearity.

In the presence of multiple regressors, the Ridge maintains a balance between the estimates' precision and the model's structural simplicity since each variable remains included but suffers a penalty proportional to its size. As a result, the influence of less significant predictors is reduced without implying their complete removal.

From a theoretical point of view, the Ridge[23] minimizes a cost function that combines the estimation error with an L2 penalty proportional to the sum of the squared coefficients. In summary form, the function to minimize is:

$$\hat{\beta} = \arg\min_{\beta} \Big\{ \underbrace{\tfrac{1}{2n} \sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ji}\big)^2}_{\text{estimation error (reduced form of MSE)}} + \underbrace{\alpha \sum_{j=1}^{p} \beta_j^2}_{\text{L2 regularization}} \Big\},$$

where:

- $\frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ji}\big)^2$ represents the measure of the estimation error (i.e., a "reduced" form of MSE);
- $\alpha \sum_{j=1}^{p} \beta_j^2$ is the L2 penalty, which establishes the extent of the coefficients' contraction.

More specifically,

- y_i indicates the observed dependent variable;
- x_i represents the vector of independent variables for observation i;
- β_j denotes the coefficients to be estimated;
- α is the regularization parameter, which regulates the intensity of the L2 penalty (the greater α, the stronger the contractions).

Our Ridge implementation follows these key steps, emphasizing that the parameter α is determined through an iterative process:

1. **Selection of the α parameter through optimization and pruning.** The determination of α, which governs the intensity of the L2 penalty, was conducted at each iteration through a Bayesian optimization process based on the TPE approach and pruning techniques, implemented in Optuna (Snoek et al., 2012). In each time block

[23] As with other penalties, in `scikit-learn` we follow the convention of introducing the factor $\frac{1}{2n}$ in front of the sum of the squared residuals and of estimating the intercept β_0 separately (Pedregosa et al., 2011; Scikit-learn Developers, 2025f).

of data (train+validation), the algorithm explored the space of possible values of α, choosing the one that minimized the mean squared error on the validation subset. This procedure was repeated each quarter, gradually adapting to the most recent economic conditions (Medeiros et al., 2021);

2. **Model training and forecast generation.** Once the value of α was determined for each optimization phase, the Ridge model was trained on the combined training and validation data, then used to generate the forecast for the test quarter (or period). After each forecast, the training dataset was extended to include the most recent observations, and the calibration and training process was repeated to handle the constant influx of new information. This sequential strategy — in which α is redefined at each sample expansion — makes the model responsive to evolving economic dynamics and counteracts any over-fitting phenomena in the long run.

The calibration and validation mechanism in our model, which uses Ridge regression as a "prediction engine," allows the configuration of the hyperparameter α while preserving the model's robustness for future data, thanks to effective shrinkage in reducing multicollinearity among the regressors (Exterkate et al., 2016; Kim & Swanson, 2014).

Elastic Net

Elastic Net (EN) is a regression methodology that integrates the L1 type penalty and the L2 type penalty in a single framework (Hastie et al., 2015; Zou & Hastie, 2005). This provides a flexible structure that combines the advantages of LASSO regression (partial selection of variables by zeroing some coefficients) with those of Ridge regression (variance containment thanks to the contraction of parameters). This double penalty is particularly effective in models with numerous regressors since it mitigates multicollinearity problems (Smeekes & Wijler, 2018; Uematsu & Tanaka, 2019) and gradually eliminates the less significant predictors.

From a theoretical point of view, Elastic Net[24] minimizes the following cost function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \Big\{ \underbrace{\frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ji} \Big)^2}_{\text{error term (reduced MSE)}} + \underbrace{\alpha \Big[(1-\gamma) \sum_{j=1}^{p} \beta_j^2 \; + \; \gamma \sum_{j=1}^{p} |\beta_j| \Big]}_{\text{mixed penalty L2 and L1}} \Big\},$$

where:

- $\frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ji} \Big)^2$ represents the estimate error component (reduced form of the MSE);
- $\alpha \Big[(1-\gamma) \sum_{j=1}^{p} \beta_j^2 \; + \; \gamma \sum_{j=1}^{p} |\beta_j| \Big]$ is the penalty term that mixes L2 and L1.

More specifically:

- $\alpha > 0$ governs the overall intensity of the penalty (i.e., the amount of regularization imposed on the coefficients);

[24] The `scikit-learn` library (Pedregosa et al., 2011; Scikit-learn Developers, 2025a) applies by default a $\frac{1}{2n}$ factor to the squared residuals and estimates the β_0 intercept separately.

- $0 \leq \gamma \leq 1$ determines the relative weight of L1 versus L2. If $\gamma = 1$, the algorithm becomes equivalent to pure LASSO; if $\gamma = 0$, it coincides with Ridge regression;
- since β_0 is excluded from the penalty, both the intercept and the coefficients are handled consistently within this mixed penalization framework.

Our Elastic Net implementation follows these key steps, emphasizing that the parameters α and γ are determined through an iterative process:

1. **Searching for the α and γ hyperparameters with optimization and pruning.** We calibrate the Elastic Net over multiple time steps using an iterative procedure for calibration and validation, regularly updating the α and γ hyperparameters in light of the most recent data. In each data block (union of a training subset and validation window), the model searches for the values of α and γ that minimize the mean squared error on the validation set, adopting a Bayesian optimization approach based on TPE (Tree-structured Parzen Estimator) (Akiba et al., 2019; Bergstra et al., 2011) and possible pruning techniques. This procedure ensures an accurate selection of the penalty levels (Kock et al., 2020) and a constant adaptability of the model to systematic changes in the regressors;
2. **Final training and forecast generation.** Once the optimal parameters are identified in a single block, the model is recalibrated on the entire training sample (*training+validation*) and used to generate the forecast. Subsequently, the *training* set is extended with the most recent observations, and the iteration is repeated, progressively aligning the Elastic Net regression to the dynamic characteristics of the analyzed phenomenon (Pesaran & Timmermann, 2007).

The sequential update strategy allows the Elastic Net to adjust its penalty parameters dynamically, progressively combining the L1 and L2 components. This approach improves the forecasting capacity of future data since it mitigates the risks of overfitting and maintains high performance in contexts characterized by continuous variations in the regressors.

Dimensionality Reduction Models

Dimensionality reduction techniques prioritize parsimony and robustness by mapping a potentially large set of interconnected variables onto fewer latent factors (Boivin & Ng, 2006; Jolliffe & Cadima, 2016; Stock & Watson, 2002). In quarterly nowcasting for a small open economy in Southeast Asia, such as Singapore, these methods reduce excessive noise and mitigate multicollinearity, thereby improving forecast stability. By transforming high-dimensional datasets into fewer orthogonal or covariance-driven components, dimensionality reduction models limit the risk of overfitting and accommodate unforeseen policy shifts or external shocks.

Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) exemplify this approach. PCR first identifies a subset of principal components that capture the bulk of variance in the predictor space (Jolliffe & Cadima, 2016), then fits a regression on these uncorrelated axes. This procedure attenuates issues arising from large predictor pools, helping the model focus on essential signals. Meanwhile, PLSR targets components that maximize the covariance between predictors and the target variable (Mehmood et al.,

2012; Wold, 1985), making it particularly appealing when the data contain many near-redundant series (Boivin & Ng, 2006). By directly emphasizing predictive relevance, PLSR can outperform variance-only methods in scenarios where traditional factor-based models might include uninformative directions.

Several design elements characterize these methods:

- *Orthogonality*: PCR relies on principal components that are mutually uncorrelated, enhancing coefficient stability.
- *Target-driven factors*: PLSR identifies latent directions aligned with the response variable, reinforcing predictive accuracy.
- *Shrinkage control*: in PCR, modern implementations often include penalties (e.g., Ridge penalization) to limit excessive parameter growth.
- *Adaptive calibration*: periodic re-estimation in an expanding window setup allows the system to integrate evolving patterns and structural changes.

Dimensionality reduction provides a more parsimonious representation, as superfluous features are filtered out. This streamlined perspective proves advantageous for nowcasting tasks in dynamic environments, where robust forecasts are often required. Moreover, these models do not impose the strict assumptions of purely linear frameworks, enabling flexible adaptations to evolving economic conditions.

Principal Component Regression

The Principal Component Regression (PCR) constitutes a hybrid methodological framework that combines the dimensionality reduction of a set of predictors with the estimation of a linear regression model. In particular, the PCR extracts a subset of principal components starting from a potentially large and high-dimensional, multicollinearity-prone (i.e., the presence of strong linear correlations among two or more predictors) set of regressors. These components are constructed as orthogonal linear combinations of the original predictors—meaning they are uncorrelated weighted sums of the input variables, each capturing a specific direction of maximal variance in the data. By concentrating most of the original information into a smaller set of uncorrelated components, PCA effectively reduces the problem's dimensionality and mitigates issues such as multicollinearity (Jolliffe & Cadima, 2016; Scikit-learn Developers, 2025d).

Once the original regressors are projected onto the new orthogonal axes defined by these components, linear regression is performed using the resulting transformed variables, known as principal components (or scores). The classical formulation of PCR uses Ordinary Least Squares (OLS)[25]—a method first proposed by (Massy, 1965)—which, however, may suffer from overfitting or high variance in the presence of noisy components. In traditional PCR, the components $\mathbf{Z}$ are used within a simple OLS regression on the dependent variable $\mathbf{y}$; such approaches have been successfully applied in macroeconomic forecasting (Stock & Watson, 2002). In our implementation, we replace the OLS estimator with a Ridge regression (Hoerl & Kennard, 1970; Scikit-learn Developers, 2025f), thereby introducing an L2 penalty term that regularizes the magnitude of the coefficients and extends the standard

[25]Ordinary regression (OLS) does not include any penalty term; it minimizes only the sum of the squared residuals.

PCR framework to enhance robustness and generalization. Recent studies have shown that Ridge regression can be a viable alternative to traditional PCR in high-dimensional settings (De Mol et al., 2008; Giannone et al., 2021).

The Principal Component Analysis (PCA) itself relies on the spectral decomposition of the regressor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ after appropriate centering or standardization. This process yields eigenvectors and eigenvalues[26] that characterize the data structure. Formally, one begins by computing the sample covariance (or correlation) matrix:

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^\top \mathbf{X},$$

and solving the eigenvalue problem:

$$\mathbf{S}\, \mathbf{v}_r = \lambda_r\, \mathbf{v}_r,$$

which yields a set of orthogonal eigenvectors $\mathbf{v}_r \in \mathbb{R}^p$, each associated with an eigenvalue λ_r. The eigenvectors define the directions of maximal variance in the data, while the eigenvalues indicate how much variance is explained in each direction.

We retain the first k components that capture most of the total variance by ordering the eigenvectors according to their associated eigenvalues in decreasing order. Letting $\mathbf{V}_k \in \mathbb{R}^{p \times k}$ denote the matrix whose columns are the top k eigenvectors, the matrix of scores—i.e., the transformed data in the reduced component space—is obtained as:

$$\mathbf{Z} = \mathbf{X}\, \mathbf{V}_k.$$

Each row of $\mathbf{Z}$ corresponds to an observation expressed in terms of the new orthogonal axes. The columns of $\mathbf{Z}$ are uncorrelated by construction, facilitating subsequent modeling steps and alleviating problems arising from multicollinearity in the original data.

In "traditional" PCR, the principal components $\mathbf{Z}$ are used within a simple OLS regression on the dependent variable $\mathbf{y}$. However, OLS may yield estimates with high variance when the data are noisy. Moreover, even if all variables carry some marginal signal, purely OLS-based estimates can become unstable if each predictor contributes only slightly but is still relevant; in such cases, leaving minor coefficients unshrunk can inflate variance.

Instead, our second stage adopts a Ridge regression in the principal component domain. Specifically, let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ be the vector of Ridge coefficients on the principal components $\mathbf{Z}$. Formally:

$$\hat{\boldsymbol{\gamma}}_{\text{Ridge}} \;=\; \arg\min_{\boldsymbol{\gamma}} \Big\{ \underbrace{\tfrac{1}{2n} \left\| \mathbf{y} - \mathbf{Z}\boldsymbol{\gamma} \right\|^2}_{\text{estimation error (MSE form)}} \;+\; \underbrace{\lambda \sum_{j=1}^{k} \gamma_j^2}_{\text{L2 penalty}} \Big\},$$

where λ controls the intensity of the regularization and $\mathbf{Z}$ is the matrix of the first k principal components of $\mathbf{X}$. Notably, the penalty $\lambda \sum_{j=1}^{k} \gamma_j^2$ acts directly on these coefficients in the (orthogonal) component space, rather than on the original predictors. The intercept γ_0 is estimated separately (and typically excluded from the penalty) if mean-centering is

[26] Eigenvectors are directions that remain unchanged when a linear transformation is applied. Eigenvalues are scalars that quantify the amount of stretching, shrinking, or mirroring along those directions.

performed. Once $\hat{\gamma}_{\text{Ridge}}$ is obtained, the corresponding coefficients in the original feature space arise via the loadings $\mathbf{V}_k$, namely:

$$\hat{\boldsymbol{\beta}}_{\text{PCR}} \;=\; \mathbf{V}_k \, \hat{\boldsymbol{\gamma}}_{\text{Ridge}},$$

which effectively re-expresses the penalized coefficients back to the domain of the initial predictors.

This variant of PCR (with Ridge regression in the second stage) can limit the estimates' variance while maintaining the benefits of dimensionality reduction. By imposing $\lambda \sum_{j=1}^{k} \gamma_j^2$ on the coefficients $\boldsymbol{\gamma}$, the Ridge regression shrinks large parameter estimates toward zero, thereby mitigating the risk of overfitting that could arise even after dimensionality reduction. Indeed, while the projection onto a reduced space alleviates multicollinearity by extracting orthogonal components, some of those components may still capture variance that is not strictly relevant for prediction. The L2 penalty thus serves as an additional safeguard by discouraging overly large weights and containing the impact of noise. This mechanism is especially beneficial in high-dimensional contexts, where the number of potential regressors—and hence principal components—may be considerable and the sample size is limited. Consequently, combining PCA with a Ridge penalty reduces the dimensional complexity and provides a second layer of shrinkage, promoting more stable and robust forecasts.

In our nowcasting context, PCR is adopted with an iterative approach on time blocks, where at each sample expansion, both the number k of principal components and the regularization parameter λ of the L2 regression are recalibrated. In each window (train+validation), the procedure follows two fundamental phases:

1. **Search for hyperparameters with Bayesian optimization and pruning.** The model explores the space of values for k (number of components) and λ employing a TPE (Tree-structured Parzen Estimator) approach (Bergstra et al., 2011). At each iteration, the algorithm finds the combination $(k,\ \lambda)$ that minimizes the mean squared error on the validation set. In the case of unpromising trials, a pruning mechanism halts the evaluation prematurely, thereby reducing computational overhead.

2. **Final training and prediction.** Once the optimal values of k and λ are determined, the PCA is performed on the entire train+validation subset, in order to produce the principal components $\mathbf{Z}$. A Ridge regression model is then applied to these components, and predictions are made for the test period.

At the end of each prediction, the training window is expanded to include the last tested observation, and the entire hyperparameter calibration and training cycle is repeated. This "expanding window" strategy allows constant updating of the principal component decomposition and the degree of L2 penalization to reflect any structural changes in the underlying economic relationships (Bańbura & Rünstler, 2011).

PCR emerges as a flexible forecasting engine that leverages dimensionality reduction and linear regression. The adoption of Ridge regression in our context (instead of the traditional OLS estimation) enhances the model's robustness, mitigates the risk of overfitting while maintaining a high predictive capacity, and is in line with recent findings in high-dimensional macroeconomic forecasting (Giannone et al., 2021). Furthermore, the iterative

hyperparameter calibration process, which employs Bayesian techniques, enables the model to adapt to changes in data (Chinn et al., 2023).

Partial Least Squares Regression

Partial Least Squares Regression (PLSR) originates from the work of (Wold, 1985) and serves as a regression strategy that helps reduce the dimensionality of the covariate system while identifying the latent directions that maximize the covariance between the regressors and the dependent variable (Mehmood et al., 2012). This procedure is advantageous in the presence of datasets with potential multicollinearity or a high number of variables, since it concentrates the relevant information in a small number of uncorrelated factors while preserving the relationship with the response.

From an applicative point of view, PLSR offers a compromise between the reduction of dimensionality (characteristic of component methods) and the focus on the relationships with the variable of interest. Compared to other linear techniques, PLSR does not only perform a decomposition of the regressor matrix $\mathbf{X}$[27], but also builds latent components that aim to explain both the variance of $\mathbf{X}$ and, simultaneously, the variability of $\mathbf{y}$.

This feature differentiates PLSR, for example, from Principal Component Regression. In PCR, the extraction of the components occurs by considering only the maximization of the variance of $\mathbf{X}$, without explicitly taking into account $\mathbf{y}$. If some aspects of $\mathbf{X}$ explain a large portion of variance but are poorly correlated with the target variable, they can still be included among the principal components, slowing or hindering the predictive capacity (Mehmood et al., 2012). On the contrary, PLSR integrates the variable of interest in the same decomposition procedure, selecting informative directions that correlate with $\mathbf{y}$. This approach tends to generate latent factors more relevant from the viewpoint of prediction, making PLSR an effective option in situations where the mere explanation of the variance of the regressors does not suffice to guarantee adequate model performance.

Furthermore, PLSR proves particularly robust when regressors exhibit strong collinearity or when the number of variables is high compared to the available observations. In such cases, focusing on components that maximize $\mathrm{cov}(\mathbf{X}, \mathbf{y})$ helps exclude directions with high variance but low predictive relevance, which could otherwise distort the estimate (Krämer & Sugiyama, 2011). Consequently, the set of latent factors tends to be better suited for capturing meaningful relationships between the regressors and the output variable, mitigating the risk of overfitting.

Thanks to this mechanism, PLSR can provide a competitive advantage in regression tasks by extracting dimensions that both explain data variance and contribute significantly to estimating the dependent variable. In nowcasting or macroeconomic forecasting contexts, where numerous potentially informative series are often available, PLSR supports a more direct focus on the components that truly matter for the phenomenon of interest (Groen & Kapetanios, 2016; Kelly & Pruitt, 2015; Stamer, 2024), thus allowing the model to concentrate on latent structures that meaningfully enhance predictive ability.

PLSR[28] searches for a sequence of latent factors $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a$ such that the projection

[27]The notation $\mathbf{X} \in \mathbb{R}^{n \times p}$ indicates n observations and p regressors; in PLSR, the dependent variable $\mathbf{y}$ is treated in parallel, identifying the components with maximum covariance between $\mathbf{X}$ and $\mathbf{y}$.

[28]In the implementation based on `scikit-learn` (Scikit-learn Developers, 2025c), the `PLSRegression` class jointly handles the regressors and the dependent variable, typically excluding the intercept from the penalization.

of $\mathbf{X}$ and $\mathbf{y}$ onto these factors is maximally correlated (Wold, 1985). Formally, let $\mathbf{w}_j$ be the weight vectors that define each factor $\mathbf{t}_j = \mathbf{X}\,\mathbf{w}_j$. Once a such factors are determined, the model can be written as:

$$y \;=\; \beta_0 \;+\; \sum_{j=1}^{a} t_j\, c_j \;+\; \varepsilon, \quad \mathbb{E}\big[\varepsilon \mid \mathbf{X}\big] \;=\; 0$$

where $\mathbf{c}_j$ denotes the coefficient associated with the j-th factor. The number of components a is a critical hyperparameter, as it dictates the level of dimensional reduction and the balance between explanatory capability and model complexity. The estimation error is typically evaluated via the Mean Squared Error (MSE):

$$\mathrm{MSE} \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \hat{y}_i\big)^2.$$

The search for latent factors proceeds iteratively, maximizing the covariance $\mathrm{cov}(\mathbf{t}_j, \mathbf{u}_j)$ between the projections $\mathbf{t}_j$ (from $\mathbf{X}$) and $\mathbf{u}_j$ (from $\mathbf{y}$) at each step (Mehmood et al., 2012).

Consistent with our other approaches, we employed PLSR as a "prediction engine" to generate one-step-ahead estimates through an expanding window technique. In each time block, the dataset was split into a training+validation subset and a test observation. The selection of the number of components a was carried out via Bayesian optimization based on TPE (Akiba et al., 2019; Bergstra et al., 2011; Snoek et al., 2012), which:

1. **Explored the space of possible configurations** for a (and other parameters, if relevant), evaluating their mean squared error on the validation portion;

2. **Identified the combination of hyperparameters** that minimized the validation error, halting in advance the least promising attempts (pruning).

After this procedure was completed, we recalibrated the model using the entire span (training+validation) with the optimal number of components, then generated the forecast for the test portion. Upon each sample expansion, the window was extended to include the newly available observation, and the calibration–validation cycle was repeated in full. These progressive hyperparameter updates yielded a flexible framework that adapted to the evolving nature of the data.

Adopting PLSR in this framework allows combining the selection of latent components and the optimization of their number in an iterative way. This methodology provides a model able to capture significant linear relationships even in the presence of potentially collinear regressors and updates the structure of the forecasting system with the arrival of new information.

Ensemble Learning Models

Ensemble learning methods combine multiple predictors into a single forecasting framework, leveraging the diversity among base learners to enhance predictive accuracy and stability (Breiman, 2001a; T. Chen & Guestrin, 2016). These approaches offer notable benefits in the context of quarterly one-step-ahead nowcasting for a small open economy

such as Singapore. By aggregating distinct trees or boosted sequences of estimators, ensemble models can more effectively accommodate sudden policy shifts, external shocks, and other structural changes than purely linear specifications (Chu & Qureshi, 2023; Yoon, 2021). Their adaptability is further heightened by their capacity to handle many potentially correlated variables without imposing rigid functional forms.

A key advantage of ensemble techniques lies in their ability to reduce the variance of individual learners through either bagging or boosting. Random Forest, for instance, grows multiple trees independently on bootstrap samples and averages their outputs, producing stable point estimates with limited sensitivity to outliers (Breiman, 2001a). Conversely, XGBoost trains trees sequentially, with each new iteration refining previous residuals through gradient-based error correction (T. Chen & Guestrin, 2016). Although bagging and boosting differ in managing errors and variance, both methods can address intricate, nonlinear patterns frequently encountered in macroeconomic series.

The methodological structure of these models is characterized by:

- *Randomization*: Bagging (e.g., Random Forest) applies sampling with replacement and random feature selection, enabling robust inferences even under high multicollinearity.
- *Sequential refinement*: Boosting (e.g., XGBoost) iteratively improves on residuals, reducing bias while incorporating regularization to avoid overfitting.
- *Hyperparameter control*: Tuning tree depth, learning rates, and sub-sampling balances model complexity and generalization capacity.
- *Nonlinearity*: Through tree-based splits, ensemble methods capture threshold effects and regime switches naturally.

Ensemble learning strategies adapt systematically to new observations when implemented in a rolling or expanding window setup. The model selection procedure involves identifying optimal hyperparameters by maximizing performance on a validation segment and then retraining on updated datasets. Such continuous recalibration proves particularly relevant for small and open economies, where shifts in global demand, exchange rates, or trade policies can alter traditional forecasting relationships. The combination of variance reduction, nonparametric flexibility, and straightforward interpretability of split structures makes ensemble learning a compelling paradigm for nowcasting tasks, complementing traditional factor models and more recent neural network approaches.

Random Forest

The Random Forest (RF) is an ensemble learning technique based on bagging (Biau & Scornet, 2016; Breiman, 2001a), where multiple decision trees are trained independently on bootstrap subsamples of the original dataset. At each tree node, a subset of regressors is randomly selected (Probst et al., 2019), thereby ensuring internal diversification among the various trees, reducing the variance of the single model, and mitigating the risks of overfitting. Furthermore, unlike some penalized regressions that require an explicit functional hypothesis, the Random Forest can capture potential nonlinear relationships between macroeconomic variables while maintaining robustness to correlations among predictors. Such methodological flexibility proves advantageous in nowcasting analyses,

where multiple potentially correlated series and heterogeneous indicators often render the specification of a strictly linear model problematic.

From a theoretical standpoint, the Random Forest is composed of B decision trees. Each tree T_b $\big(b = 1, \ldots, B\big)$ is built starting from a bootstrap sample of the training data (Breiman, 2001a), and a random subset of available features is considered at each node to find the optimal split. In terms of prediction, each of these B trees provides an estimate $\hat{y}_b(\mathbf{x}_i)$ for the observation $\mathbf{x}_i$. Aggregation then occurs by averaging the predictions of the individual trees, i.e.:

$$\hat{y}_i \;=\; \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}_i).$$

The training objective can be formalized as the maximization (or minimization with opposite sign) of a function related to the root mean square error (RMSE). More precisely, we define:

$$\mathrm{Obj} = \underbrace{-\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{y_i}_{\text{actual value}} - \underbrace{\frac{1}{B}\sum_{b=1}^{B} T_b\big(\mathbf{x}_i\big)}_{\text{ensemble-based forecast}}\Big)^2}}_{\text{negative RMSE (score function)}}.$$

where:

- y_i is the actual value of target variable;
- $\frac{1}{B}\sum_{b=1}^{B} T_b\big(\mathbf{x}_i\big)$ represents the Random Forest's aggregate prediction.

By combining the outputs of all B trees, the Random Forest reduces variance compared with a single tree and, through randomization both in the observations (bootstrap) and in the features (`max_features`), it preserves model stability even in scenarios characterized by high multicollinearity.[29]

The main hyperparameters of the Random Forest regulate each tree's complexity and define the sampling strategies (Probst et al., 2019). In particular, within our modeling framework[30]:

- B (`n_estimators`): higher values tend to reduce variance, albeit at the cost of increased computational demands;
- $d_{\max}$ (`max_depth`): limiting tree depth controls overfitting by preventing excessively complex splits;

[29] In high-dimensional macroeconomic contexts, this bagging framework and the random subspace selection (m_{try}) can effectively mitigate collinearity and overfitting (Biau & Scornet, 2016; Probst et al., 2019). Such diversity-promoting mechanisms bear a conceptual resemblance to the feature selection or shrinkage strategies adopted by penalized linear models.

[30] In brackets we have indicated the names of the hyperparameters used in our model in Python (Scikit-learn Developers, 2025e).

- n_{split} and n_{leaf} (`min_samples_split` and `min_samples_leaf`): set the minimum number of samples required to split a node or create a leaf, significantly influencing a single tree's variance;

- m_{try} (`max_features`): determines how many regressors are selected at each split, fostering internal ensemble diversity and counteracting correlations among predictors;

- boot (`bootstrap`): determines whether each tree is trained on subsets of observations drawn with or without replacement, promoting internal diversity within the ensemble model;

- κ (`criterion`): sets the splitting criterion (e.g., mean squared error or absolute error), thus affecting the model's sensitivity to outliers or unusual residual distributions.

Calibrating these parameters enables practitioners to balance bias and variance, flexibly adapt to the dataset's features, and ensure satisfactory predictive performance (Breiman, 2001a; Probst et al., 2019).

In our procedure, we implemented the Random Forest as a "prediction engine" within a sequential protocol. In each time block, we split the data into a train+validation segment (to calibrate the aforementioned hyperparameters) and a single test period corresponding to the subsequent observation. We relied on a Bayesian search strategy for hyperparameter optimization (Akiba et al., 2019; Shahriari et al., 2016; Snoek et al., 2012), enhanced by early "pruning" of less promising configurations. Once the optimal values were identified, we retrained the model on the same train+validation subset and generated the forecast for the test. The dataset was then expanded to incorporate the newly predicted observation, and the entire process was repeated at every fresh period. This "expandable window" approach enabled the Random Forest to adapt progressively to shifts in macroeconomic conditions, reducing the likelihood of prolonged misspecification and supporting more reliable estimates over time (Cerqueira et al., 2020).

Integrating the Random Forest with a Bayesian hyperparameter search permits exploring a vast configuration space, leveraging the trees' inherent capacity to detect nonlinear patterns. By iteratively updating these parameters at each dataset expansion, the model achieves greater stability in its estimates without relinquishing algorithmic flexibility, making the Random Forest a competitive option in nowcasting scenarios that demand dynamic and consistent forecasts.

Despite its advantages, our application uses an i.i.d. bootstrap for sampling, rather than a time-series block bootstrap (Künsch, 1989)[31]. The latter could better account for autocorrelation in quarterly macroeconomic data but would considerably increase computational times, already high due to ensemble methods and hyperparameter optimization, especially when timely nowcasts are needed on a quarter-by-quarter basis (Biau & Scornet, 2016; Probst et al., 2019). Implementing a fully time-series-friendly Random Forest, incorporating explicit lag features and a block bootstrap for each tree, further inflates data

[31] In mathematical terms, if we indicate with $\{(x_t, y_t)\}_{t=1}^{T}$ our series, a block bootstrap could consist in building a succession of blocks $\{B_j\}$ of length L, where $B_j = \{(x_{(j-1)L+1}, y_{(j-1)L+1}), \ldots, (x_{jL}, y_{jL})\}$. By sampling such blocks with replacement (possibly managing the remainder when T is not a multiple of L), a new artificial series $\tilde{\mathcal{D}}$ of length T is constructed, preserving part of the internal temporal correlation. Each Random Forest tree would then be trained on a subsample $\tilde{\mathcal{D}}_b$ generated via this block logic.

dimensionality and necessitates more elaborate data manipulations. Given the one-step-ahead nature of our nowcasting exercise, we opt for the simpler i.i.d. bootstrap, balancing real-time feasibility with robust predictive performance (Medeiros et al., 2021; Varian, 2014).

We acknowledge that this strategy does not fully preserve temporal dependence in the training process and may underreact to abrupt regime shifts (Carriero et al., 2015b). Nonetheless, adopting an expanding-window validation protocol and sequential hyperparameter tuning alleviates much of the risk of overfitting past regimes, as newly available data gradually refine the model (Hyndman & Athanasopoulos, 2021). Moreover, this streamlined approach has proven effective in multiple forecasting competitions, such as the M5 Competition, where computational overhead is a constraint and many participants successfully employed simpler i.i.d. or rolling holdout strategies (Makridakis et al., 2022). Future research could explore whether the gains from a block bootstrap or additional time-series-oriented structures outweigh the added complexity in contexts characterized by significant autocorrelation or frequent structural breaks.

Gradient Boosting with XGBoost (eXtreme Gradient Boosting)

The Gradient Boosting methodology is an ensemble learning technique based on a sequential boosting approach. This approach combines multiple "weak trees" to reduce the forecast error progressively. The key idea is to correct, at each iteration, the residual errors committed by the previous model, thus reducing the bias and strengthening the predictive capabilities.

In this context, eXtreme Gradient Boosting (XGBoost or XGB) (T. Chen & Guestrin, 2016; XGBoost Developers, 2023) integrates powerful regularization mechanisms, controlled rate iterative learning, and parallel computing facilities while preserving the ability to capture non-linear relationships between regressors and macroeconomic output. This flexibility presents advantages in datasets with multiple potentially correlated variables, where a purely linear model could prove limiting.

Formally, XGB builds an ensemble of M regression trees, each denoted as f_m. The set of such trees produces an aggregate prediction:

$$\hat{y}_i \;=\; \sum_{m=1}^{M} f_m\big(\mathbf{x}_i\big),$$

where $f_m\big(\mathbf{x}_i\big)$ represents the single prediction provided by the m-th tree for the observation $\mathbf{x}_i$. The optimization focuses on minimizing an error metric, such as RMSE (Root Mean Squared Error), which can be transformed into an objective function to be maximized by placing a negative sign in front of the square root. Concretely:

$$\text{Obj} \;=\; \underbrace{-\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{y_i}_{\text{actual value}} - \underbrace{\sum_{m=1}^{M} f_m\big(\mathbf{x}_i\big)}_{\text{sum of the predictions of the M trees}}\Big)^2}}_{\text{score function (RMSE with negative sign)}},$$

where:

- y_i indicates the actual value of the dependent variable;

- $\sum_{m=1}^{M} f_m(\mathbf{x}_i)$ synthesizes the sequential combination of the predictions provided by the individual trees.

In addition to the minimization of the loss function, XGB also incorporates a penalty term on the complexity of each tree, aiming to reduce overfitting. Specifically, the regularization includes an L2 penalty governed by λ (`reg_lambda`), which penalizes large magnitudes of the leaf weights:

$$\Omega(f_m) \;=\; \gamma\, T_m \;+\; \frac{\lambda}{2} \sum_{j=1}^{T_m} \big(w_j\big)^2,$$

where:

- T_m is the number of leaves of the m-th tree;
- γ (`gamma`) is a parameter that penalizes the addition of new leaves;
- w_j denotes the weight (i.e. the constant prediction) associated with each leaf j;
- λ (`reg_lambda`) penalizes large leaf weights w_j[32].

By incorporating $\Omega(f_m)$ into the global objective, XGB discourages both an excessive number of leaves and excessively large leaf weights, thereby mitigating variance and helping the model generalize better.

Each tree f_m operates as a step function that partitions the space of variables into distinct regions, assigning a constant estimate w_j to each region (leaf). The structure of these trees—splits, depth, and leaf values—is learned iteratively, correcting, at each iteration, the residual errors of its predecessors. However, the L2 penalty shrinks large leaf coefficients, thus reducing the risk of overfitting to high-frequency noise. As previously discussed, this shrinkage also helps preserve smaller yet potentially informative effects by proportionally reducing leaf weights rather than nullifying them altogether.

In our study, we implement XGB as a "prediction engine" within a sequential protocol analogous to that adopted for the Random Forest (Chu & Qureshi, 2023; Masini et al., 2023). In each time block, the dataset is divided into a train+validation subset (used to calibrate the hyperparameters) and a single test observation corresponding to the subsequent period.

For hyperparameter tuning, we rely on a Bayesian search strategy based on TPE(Akiba et al., 2019; Bergstra et al., 2011; Optuna Developers, 2023), augmented by a pruning mechanism that halts evaluations showing limited promise. Specifically, we explore the following main hyperparameters (with the names used in the script in brackets):

- M (`n_estimators`): total number of boosted trees;
- $d_{\max}$ (`max_depth`): maximum depth of each tree;
- η (`learning_rate`): rate at which new trees correct existing forecasts;
- γ (`gamma`): minimum gain threshold for splitting a node;

[32]This L2 penalty acts as a form of "weight decay," analogous to that seen in neural networks (Hastie et al., 2015; Krogh & Hertz, 1992), shrinking large coefficients to reduce overfitting.

- $w_{\min}$ (`min_child_weight`): minimum number of samples required for a split;
- $\text{col}_{\text{bytree}}$ (`colsample_bytree`): fraction of predictors to consider in each tree;
- sub_s (`subsample`): fraction of observations sampled before growing each tree;
- λ (`reg_lambda`): coefficient for the L2 penalty term $\Omega(f_m)$.

Once the best configuration of these hyperparameters is identified, we retrain the model on the combined train+validation subset and generate the forecast for the test observation. The dataset is then expanded to include the newly predicted quarter (or period), and the entire process is repeated at each fresh time step. This expanding window approach enables XGB to adapt gradually to structural shifts in the macroeconomic environment and to reduce the likelihood of long-lasting misspecification, thereby enhancing its predictive reliability.

This XGB approach, therefore, merges the predictive capacity of sequential trees with an accurate iterative error correction mechanism while maintaining control over the complexity thanks to targeted regularization parameters. Furthermore, the continuous updating of hyperparameters, made possible by Bayesian optimization, ensures a constant adaptation of the model to macroeconomic series changes without compromising the overall estimates' robustness.

Neural Models

Neural networks (Goodfellow et al., 2016; Hornik et al., 1989) offer a flexible framework for capturing nonlinear dynamics, complex interactions among variables, and evolving economic conditions. In the specific context of one-step-ahead nowcasting for a small and open economy such as Singapore, their capacity to adapt to rapidly changing environments proves advantageous. This section focuses on two architectures: the Multilayer Perceptron (MLP) and the Gated Recurrent Unit (GRU) (Cho et al., 2014), each equipped with regularization mechanisms to ensure robustness against overfitting.

Neural models differ from linear or factor-based approaches by incorporating activation functions, memory cells, or gating structures that introduce nonlinearity. Such features are crucial when macroeconomic indicators shift abruptly due to exogenous shocks, policy interventions, or other regime changes. Indeed, recent studies highlight the competitiveness of neural architectures in macroeconomic contexts characterized by volatility (Almosova & Andresen, 2023; Hewamalage et al., 2021). Neural networks respond to these shifts by updating their internal parameters in a rolling or expanding manner, thus preserving the ability to learn new data patterns as they arise.

In designing neural networks for quarterly nowcasting, there are several guiding considerations:

- *Nonlinear Relationships.* Traditional regression models rely on linearity assumptions or static factor structures, which may not fully capture the diverse and potentially intricate linkages that govern aggregate economic outcomes. Neural layers accommodate nonlinearities by design, thus more readily uncovering hidden patterns or threshold effects.

- *Regularization Strategies.* Given the limited sample size typical of quarterly data, shrinkage methods, and dropout help prevent the over-parametrization of the model. These techniques penalize large weights or stochastically exclude neurons, protecting the model from spurious correlations in historical data.

- *Temporal Evolution.* While MLPs treat each observation independently, GRU models incorporate memory states that track historical context, providing a natural handle on time-lagged effects. This property is especially relevant in small open economies, where external shocks, trade flows, and policy decisions can induce persistent fluctuations.

- *Hyperparameter Optimization.* Choosing the right network depth, hidden dimensionality, and learning rate is critical. We employ a Bayesian search strategy with pruning to explore these configurations efficiently. The model is then retrained on updated samples at each step, enhancing adaptability to unforeseen structural breaks.

Although neural approaches can be more opaque than penalized regressions, the iterative recalibration of weights and the inclusion of interpretability strategies (as already discussed in this study) facilitate a clearer understanding of which predictors drive the nowcast. Theoretically, the synergy between nonlinearity, sequence modeling, and automatic hyperparameter tuning renders neural models a competitive alternative in scenarios marked by high volatility or nonstationary data-generating processes.

Multilayer Perceptron

The Multilayer Perceptron (MLP) is a feedforward neural network architecture[33] (Hornik et al., 1989) designed to address regression problems that may exhibit nonlinear relationships[34] between the regressors and the variable of interest. Unlike pure linear models, the MLP introduces nonlinear activation functions[35] (Glorot et al., 2011) and multiple hidden layers[36] (hidden layers), allowing the representation of even complex links between economic inputs.

Adopting an MLP for nowcasting is particularly appropriate when the number of variables is high, and it is suspected that non-trivial interactions may influence the target variable (e.g., GDP growth). Furthermore, including regularization mechanisms (such as dropout[37] (Srivastava et al., 2014)) helps mitigate overfitting, favoring a robust adaptation in the face of unstable macroeconomic dynamics.

In such settings, it is common that some predictors become irrelevant or partially redundant, thus introducing noise and complicating the learning process. Regularization mechanisms such as dropout are advantageous in addressing extraneous or spurious regressors that overshadow the genuinely significant ones. These mechanisms foster a more parsimonious representation of the data.

[33] A feedforward neural network is a class of networks in which signals flow solely from inputs to outputs, with no feedback or internal loops.

[34] Nonlinear relationships are interactions between variables in which a slight change in a regressor can produce non-constant (or non-proportional) changes in the dependent variable.

[35] Nonlinear activation functions (e.g., Rectified Linear Unit (ReLU) or sigmoid) transform the output of a neuron after the linear combination of the inputs, allowing to capture more complex relationships.

[36] Hidden layers are the intermediate levels between the input and the output of the network, where internal representations of the data are elaborated.

[37] This technique randomly "turns off" some units during the training process, helping to prevent the model from passively storing the noise in the historical data.

Formally, the MLP expresses the relationship

$$y_i \;=\; f\big(\mathbf{x}_i;\, \boldsymbol{\theta}\big) \;+\; \varepsilon_i, \quad \mathbb{E}\big[\varepsilon_i \mid \mathbf{x}_i\big] \;=\; 0,$$

where:

- y_i is the target variable (e.g., quarterly GDP growth), expressed as a sum of a systematic component and a random noise term;
- $f(\mathbf{x}_i;\, \boldsymbol{\theta})$ is the function learned by the network, which maps regressors $\mathbf{x}_i$ to a predicted value of y_i;
- ε_i represents the unpredictable part (noise);
- $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$ assumes no systematic bias remains in the residuals once $\mathbf{x}_i$ is known;
- $\mathbf{x}_i \in \mathbb{R}^p$ is the vector of regressors;
- $\boldsymbol{\theta}$ collects all the weights and biases of the network.

Each neuron[38] linearly combines the inputs and applies an activation function (e.g., ReLU) to introduce nonlinearity. If a hidden layer has h neurons, each neuron j processes

$$z_j \;=\; \sigma\Big(\sum_{k=1}^{d} w_{jk}\, x_k \;+\; b_j\Big),$$

where:

- $\sigma(\cdot)$ is the activation function;
- w_{jk} and b_j are the parameters to optimize.

Connecting multiple layers in sequence creates a hierarchy of representations useful for capturing complex patterns.

In our empirical setup, we configure the MLP with a single hidden layer (or a limited number of layers). Although deeper architectures can approximate highly intricate relationships, in macroeconomic nowcasting with relatively short samples, additional hidden layers do not necessarily improve generalization and can exacerbate overfitting (Glorot et al., 2011; Hornik et al., 1989). A shallower network is often sufficient to capture the bulk of nonlinearities while retaining computational efficiency.

The estimate of $\boldsymbol{\theta}$ occurs by minimizing a cost function consisting of the mean squared error (MSE):

$$\min_{\boldsymbol{\theta}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i;\, \boldsymbol{\theta})\big)^2 \Big\}$$

[38] A neuron is the fundamental processing unit of a neural network, in which an activation function follows a linear combination of inputs.

The backpropagation process[39] (Rumelhart et al., 1986) calculates the gradients[40] of this MSE with respect to each weight and bias, allowing them to be updated iteratively through an optimizer (e.g., Adam) (Kingma & Ba, 2015).

In deep networks or datasets with a large number of variables, overfitting may emerge. Regularization strategies (i.e., dropout) are used to counteract overfitting, restricting the influence of unhelpful or purely noisy features. Hence, although the MLP is in principle flexible enough to incorporate multiple hidden layers, we typically keep the network shallow in nowcasting tasks to avoid excessive complexity and make training more stable in limited-sample scenarios.

Our empirical study implements the MLP as a "forecasting engine" according to a rolling protocol with an expandable window (Tashman, 2000). During each iteration, we explore the following main hyperparameters (with the names used in the script in brackets):

- α_{hd} (`hidden_dim`): the number of hidden units (neurons) per layer;
- α_{nl} (`num_hidden_layers`): the total number of hidden layers in the MLP;
- α_{dr} (`dropout_rate`): the percentage of neurons "turned off" at each epoch to prevent overfitting;
- η (`lr`): the learning rate of the optimizer, which determines the size of the weight update steps;
- β (`batch_size`): the size of the mini-batches used in training.

Also, in this case, the search for optimal parameters occurs during each iteration through the Bayesian optimization approach based on the TPE (Akiba et al., 2019; Bergstra et al., 2011; Snoek et al., 2012). The algorithm explores the space of configurations and, for each one, estimates the MSE on a validation subset. The search process is accelerated by "pruning" mechanisms that stop the least promising hyperparameter combinations early. At the end of the exploration, the combination of hyperparameters that minimizes the validation error is selected. To prevent excessive overfitting, we halt training when there is no improvement in validation error for a specified number of epochs. Once the optimal parameters are determined, the MLP is recalibrated on the entire training+validation block, then the out-of-sample forecast for the test quarter is generated. The last observation (just forecasted) is incorporated into the dataset, and the next block is moved on, repeating the entire process at the new time horizon. In this way, the model adjusts its weights and hyperparameters at each sample expansion, remaining sensitive to possible shocks or regime changes and reducing the probability of misspecification over time.

MLP, integrated with Bayesian optimization and regularization strategies, combines the ability to capture nonlinear relationships and adapt quickly to macroeconomic framework changes. Despite the lower transparency of network parameters compared to penalized linear models (such as LASSO or Ridge), the iterative selection of hyperparameters and the adoption of mechanisms such as dropout ensure a balance between flexibility and

[39] Backpropagation is the procedure by which the prediction error is propagated backward, allowing the partial derivatives of the error to be computed with respect to each weight and bias.

[40] The gradient is the vector of the partial derivatives of a function (in this case, the error) with respect to the variables to be optimized (weights and biases). It indicates the direction of maximum growth of the function itself and, by reversing its direction, the model iteratively reduces the error.

robustness, mitigating the impact of potentially irrelevant regressors. Frequent updating allows the system to respond promptly to structural changes, preserving good generalization to future data or unexpected economic scenarios.

Gated Recurrent Unit

Gated Recurrent Units (GRUs) are an architecture belonging to the class of recurrent networks[41] (Cho et al., 2014; Chung et al., 2014), designed to capture dynamic relationships between economic variables along the time axis. The introduction of the gating mechanism[42] allows to efficiently synthesize and transmit the memory between successive states, maintaining relatively low computational costs compared to other more complex recurrent networks (Jozefowicz et al., 2015).

A key motivation in using GRUs for regression is the need to contain the influence of any irrelevant variables, which could introduce noise in the estimation phases. In this context, it is appropriate to include forms of regularization, such as the L2 penalty (or weight decay) (Krogh & Hertz, 1992), to penalize the excessive growth of coefficients associated with non-informative predictors. In parallel, adopting dropout guarantees further control over overfitting, randomly switching off some neurons during the training phase (Gal & Ghahramani, 2016; Srivastava et al., 2014; Zaremba et al., 2014). Thanks to these measures, GRUs offer flexibility and simultaneously limit the risk of excessively including superfluous components, favoring a reduction in the complexity of the model (Merity et al., 2018).

Formally, the GRU model represents the relation

$$y_i \;=\; f\big(\mathbf{x}_i;\, \boldsymbol{\theta}\big) \;+\; \varepsilon_i, \quad \mathbb{E}\big[\varepsilon_i \mid \mathbf{x}_i\big] \;=\; 0,$$

where:

- y_i is the target variable (here, quarterly GDP growth), modeled as the sum of a systematic component and a random noise;
- $f(\mathbf{x}_i;\, \boldsymbol{\theta})$ is the function learned by the network, mapping the regressors $\mathbf{x}_i$ to a predicted value;
- ε_i is a noise term, capturing the unpredictable part of y_i;
- $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$ states that, on average, the noise has zero mean once $\mathbf{x}_i$ is given (no systematic bias remains in the residual);
- $\mathbf{x}_i \in \mathbb{R}^p$ denotes the vector of regressors;
- $\boldsymbol{\theta}$ collects the set of parameters (weights, biases, and gating constructs) to be optimized.

In GRUs, at each instant t, the *latent output* $\mathbf{h}_t$ selectively updates the information coming from the *previous state* $\mathbf{h}_{t-1}$ and from the current input $\mathbf{x}_t$.

[41] Recurrent networks (RNNs) are distinguished from feedforward models by their feedback loops that allow the store of information about the previous state, providing a dynamic structure to the forecast.

[42] In the GRU, the update gates (z_t) and the reset gates (r_t) regulate how the new information is combined with the previous state, mitigating phenomena such as the vanishing gradient typical of traditional RNNs.

- $\mathbf{h}_t$ (*latent output*) is the hidden state not directly observed but internally updated by the gating mechanism;
- $\mathbf{h}_{t-1}$ (*previous state*) is the hidden state carried over from the preceding time step;
- $\mathbf{z}_t$ (*update gate*) controls how much of the old state is retained vs. replaced by new input information;
- The *reset* part of $\mathbf{h}_{t-1}$ determines how much past information is forgotten before computing $\tilde{\mathbf{h}}_t$.

In particular, the update gate $\mathbf{z}_t$ balances the stored and new components:

$$\mathbf{h}_t \;=\; (1 - \mathbf{z}_t) \;\circ\; \mathbf{h}_{t-1} \;+\; \mathbf{z}_t \;\circ\; \tilde{\mathbf{h}}_t,$$

with $\tilde{\mathbf{h}}_t$ calculated by a linear combination of $\mathbf{x}_t$ and the "reset" part of the previous state. The operator $\circ$ indicates the element-by-element product. The number of hidden units and the number of layers filled determine the representativeness of the model.

In this study, we are employing a single hidden layer for the GRU.[43]. Although deeper recurrent architectures can be beneficial under extensive data availability, several empirical findings suggest that, for quarterly macroeconomic nowcasting, additional layers do not necessarily improve accuracy and may increase overfitting risk (Almosova & Andresen, 2023; Hewamalage et al., 2021). A single-layer GRU often suffices to capture short-run dynamics in limited-sample environments, preserving computational efficiency and converging more reliably (Zaremba et al., 2014).

The set of parameters $\boldsymbol{\theta}$ is estimated by minimizing a cost function consisting of the mean squared error (MSE) plus an L2 penalty

$$\min_{\boldsymbol{\theta}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i;\, \boldsymbol{\theta})\big)^2 \;+\; \lambda_{\mathrm{L2}} \, \|\boldsymbol{\theta}\|^2 \Big\}^{44}$$

using backpropagation through time (Werbos, 1990). Backpropagation is the algorithm that computes the gradient of the error with respect to all trainable parameters, systematically propagating the errors backward from the output to earlier layers (and across time steps in a recurrent network).

Backpropagation through time gradually propagates the error through the past recursive iterations of the network and allows weights to be updated using optimizers (e.g., Adam) with an L2 penalty. In addition to the dropout, these regularization elements help to contain the possible proliferation of coefficients relevant only to some data segments, preventing an uncontrolled growth of the effective dimensionality.

In the empirical study adopted here, the estimation of the GRU model followed a rolling iterative forecasting protocol with an expandable window: at each sample expansion, the best hyperparameters were automatically identified thanks to TPE (Akiba et al., 2019;

[43] A similar design choice applies to the MLP in the previous section, wherein a single hidden layer suffices to capture nonlinear relationships while controlling overfitting in limited-sample environments (Glorot et al., 2011; Hornik et al., 1989)

[44] Similarly to what we did in MLP's section, here we refer to the regularization parameter as λ_{L2} to highlight the neural-network weight decay perspective, but its function is analogous to λ in a linear Ridge model.

Bergstra et al., 2011) that evaluated candidate configurations based on the in-sample validation results. The search process was further accelerated by "pruning" mechanisms, which interrupted less promising trials early and thus reduced the computational overhead.

The search variables included are:

- α_{hd} (`hidden_dim`), i.e. the hidden dimensionality of each GRU layer;
- α_{nl} (`num_layers`), the number of overlapping recurrent layers;
- α_{do} (`dropout_rate`), the share of neurons stochastically excluded during training;
- λ_{L2} (`l2_reg`): the weight decay coefficient controlling the intensity of L2 regularization;
- η (`lr`), the learning rate of the optimizer;
- β (`batch_size`), the size of mini-batches in stochastic descent.

The search for hyperparameter configurations occurred at each estimation block, stopping training early (early stopping) if no tangible improvements in validation emerged. Once the optimal values were determined, the model was recalibrated on the entire training+validation set and produced the forecast for the out-of-sample horizon. Subsequently, it was advanced by one quarter, including the new historical information and the entire procedure was repeated. This iterative scheme guaranteed a constant updating of the network, allowing it to adapt to any structural changes and containing the risks of misspecification.

The GRU configured as a "predictive engine" is capable of exploiting temporal persistence, with a lighter internal structure than LSTMs, but equally suitable for modeling non-linear connections between macroeconomic inputs and the target variable. The dynamic optimization approach—through progressive sample expansion and Bayesian selection of hyperparameters—ensures its operational flexibility, maintaining a balance between learning ability and generalization robustness.

2.4 Prediction Intervals around nowcasts

Constructing "prediction intervals" in macroeconomic nowcasting allows to quantify the uncertainty associated with each point estimate (Chatfield, 1993; Tashman, 2000). In contexts where the residuals can be considered approximately linear and stationary[45], more traditional methods can be adopted, such as the residual-based Bootstrap (or, if appropriate, the sieve bootstrap in AR structures[46]). However, when the analysis is extended

[45]Residuals are defined as "linear" when they can be described by linear relationships with their lags or with the past values of explanatory variables (the so-called "state variables"). "Stationary" means those processes whose statistical properties (mean, variance, autocorrelation) remain constant over time.

[46]Bootstrap is a method for estimating the distribution of a parameter (or estimate) by resampling the original data with replacement.

"Time series bootstrap" methods consider the autocorrelation present in time series (i.e., subsequent values may depend on previous ones). Therefore, resampling procedures (e.g., "block bootstrap") preserve the order of the data or group contiguous intervals to reproduce the temporal structure more realistically.

In the "residual-based bootstrap," the estimated residuals from a model are sampled and fed back into the equations to generate new simulated paths.

The "sieve bootstrap" is a variant typically used in AR contexts, where the residuals are obtained from an autoregressive process approximating the dynamics of the series.

to multiple nonlinear models or structural breaks are suspected, these techniques risk not correctly capturing the data's complexity (Künsch, 1989). Thus, in the presence of potential nonlinearities and structural changes, the pair block bootstrap (Masarotto, 1990; Pan & Politis, 2016) offers the possibility of generating numerous alternative training trajectories while maintaining the temporal dependence in the data and being agnostic to the shape of the residuals, thus more suitable for a multi-model scenario and subject to possible regime shifts. Furthermore, it allows monitoring the evolution of the estimated uncertainty at each forecasting iteration, thus providing a measure of the predictive robustness even in the presence of endogenous variations and potential structural shocks (Grigoletto, 1998).

Consider any predictive model that, starting from regressors $\mathbf{x}_t \in \mathbb{R}^p$, produces an estimate $\widehat{y}_t$ of the macroeconomic variable y_t. Formally, we assume that

$$y_t \;=\; f\big(\mathbf{x}_t;\, \boldsymbol{\theta}\big) \;+\; \varepsilon_t, \quad \mathbb{E}\big[\varepsilon_t \mid \mathbf{x}_t\big] \;=\; 0,$$

where:

- y_t is the real value of the quantity of interest (in our case, the change in GDP);
- $f(\mathbf{x}_t;\, \boldsymbol{\theta})$ represents the functional mapping estimated by the model, subject to various parameters $\boldsymbol{\theta}$;
- ε_t indicates the noise component, free of conditioned bias and not explicitly modeled;
- $\mathbb{E}[\varepsilon_t \mid \mathbf{x}_t] = 0$, a condition that guarantees the absence of bias (zero mean value of the error) once $\mathbf{x}_t$ is known, i.e. ε_t reflects exclusively noise, without systematic tendencies to over- or underestimate the phenomenon.

The overall uncertainty around the forecast $\widehat{y}_t$ depends not only on the variability of the errors ε_t but also on possible limitations of the model (f), on the small size of the available sample and any breakpoints (e.g., crises, shocks, legislative changes) (Clark & Ravazzolo, 2015; Pesaran et al., 2006).

In a preliminary analysis, we applied a Bai-Perron test[47] using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) on the entire historical series of the quarterly GDP growth rate (from 1990 Q1 to 2023 Q2). According to these criteria, no significant structural breaks emerged since the model with zero breaks was preferred[48].

However, this outcome may reflect specific features of Singapore's economy—such as its relatively small size, high degree of openness, and history of rapid cyclical adjustments—which can limit the power of break tests in detecting short-lived or swiftly reversed shocks (Giraitis et al., 2013). Despite the Bai-Perron test preferring a model with zero breaks, many studies adopt ex-ante break dates (Clements & Hendry, 1996) whenever major global crises are known to have potentially altered economic regimes. Such an approach is economically motivated rather than purely statistical, justified in contexts where some shocks, though real, prove too transient or compensated to manifest as strong, persistent breaks in the data.

[47]The Bai-Perron test is a procedure to identify multiple breakpoints in a regression model. Using information criteria (AIC, BIC) evaluates whether the regression function should be split into segments with different parameters and determines the optimal number of breaks.

[48]Specifically: for $m = 0$ breaks we found RSS = 998.853 (AIC = 273.175, BIC = 278.971); for $m = 1$ break yielded RSS = 945.221 (AIC = 275.668, BIC = 281.412); increasing m to 2–5 further raised both AIC and BIC. Consequently, the no-break ($m = 0$) model remained preferred under both criteria.

Due to the well-known global economic shocks (Asian financial crisis of the late 1990s, the September 11 attacks in 2001, SARS, the 2008 crisis, the advent of COVID-19, and the Russian-Ukrainian conflict), we decided to impose certain structural breaks ex-ante corresponding to the quarters in which such events have likely affected global and Singaporean macroeconomic dynamics, regardless of the statistical test (Clements & Hendry, 1996). This choice is justified by the economic interest in assessing the possible impact of such shocks. It also reflects a model design approach intended to preserve relevant historical information, even if purely statistical criteria do not emphasize these breaks.

In our analysis and for each iteration, we divided the procedure into the following steps:

1. **Subdivision of the train+validation sample into sub-segments.** For each iteration, we subdivided the train+validation time interval (up to the last quarter considered in-sample) into multiple sections/sub-segments, separating them at the historical-economic breaks considered crucial (including 1997 Q3, 2001 Q1, 2003 Q1, 2008 Q3, 2020 Q1, 2022 Q1). Although the Bai-Perron test did not confirm these breaks statistically, we opted to maintain them ex-ante to reflect real-world events (such as global crises or shocks) that could affect the Singapore economy. This approach preserves economic coherence, allowing the distinction of pre- and post-crisis phases (Pesaran et al., 2006).

2. **Creation of contiguous blocks for each subsegment.** In each subsegment, the data $(\mathbf{x}_t, y_t)$ were grouped into contiguous blocks of fixed size (4 quarters) to safeguard any internal serial autocorrelation. Instead of randomly shuffling individual observations, we considered consecutive data groups, reflecting the temporal dynamics of an entire year of observations[49] (Masarotto, 1990).

   Instead of randomly shuffling individual observations, we considered consecutive data groups, reflecting the temporal dynamics of an entire year of observations. We set the block size to 4 quarters (one year) to preserve potential seasonality within each sampled block and to maintain key short-run dependencies that often manifest at an annual frequency in macroeconomic series (see (Masarotto, 1990; Pan & Politis, 2016)). This approach reduces the risk of breaking crucial within-year correlations that could otherwise bias the bootstrap re-sampling.

3. **Sampling with block replacement.** Several blocks of the same length (4 quarters) were randomly extracted from each subsegment, with replacement, until a quantity of data similar to the original one of the subsegment was reached. In this way, a new sample $\left(\mathbf{x}_{\text{boot}}, y_{\text{boot}}\right)$ was created that preserves the structure of intra-block dependencies but varies in terms of order or frequency of the blocks themselves. Unlike a simple row shuffling, this method does not assume that the errors are independent and identically distributed (i.i.d.) (Thombs & Schucany, 1990).

4. **Concatenation of the resampled subsegments.** The "bootstrapped" subsegments (i.e., reassembled into blocks) were then concatenated in sequence, forming an "alternative" train+validation dataset compared to the initial one, but with the same length and with a similar temporal subdivision, at least in the macro-phases. As

[49] Such a block size (one full year) preserves typical annual seasonality in macroeconomic series and reduces the risk of "breaking" within-year correlations, as recommended by (Makridakis et al., 1998).

a result, we obtained a complete and ready-to-be-used training set (Pan & Politis, 2016).

5. **Retraining the model and nowcast on the test quarter.** Once the dataset $\left(\mathbf{x}_{\text{boot}}, y_{\text{boot}}\right)$ was defined, the model used as a "predictive engine" (e.g., Elastic Net, PLSR, GRU) was recalibrated on the resampled data. Once training was completed, the estimate $\widehat{y}_{\text{B}}$ for the test quarter was generated. The operation was repeated $\text{B} = 1000$ times, producing a distribution of predictions $\{\widehat{y}^{(b)}\}_{b=1}^{B}$. In this way, multiple "predictive scenarios" are collected under different reorganizations to reflect the potential variability of the train+validation data (Hewamalage et al., 2021).

6. **Construction of the confidence interval.** The set of estimates $\{\widehat{y}^{(b)}\}_{b=1}^{B}$ was then ordered, and the quantiles associated with the desired levels (e.g., 2.5% and 97.5%), defined as:

   - $\underline{q}_{\alpha} \;=\; \text{quantile}\left(\{\widehat{y}^{(b)}\},\, \alpha\right)$ (lower quantile, i.e. 2.5%)
   - $\overline{q}_{\alpha} \;=\; \text{quantile}\left(\{\widehat{y}^{(b)}\},\, 1-\alpha\right)$ (upper quantile, i.e. 97.5%)

   which results in the Confidence Interval (Prediction Interval, PI) around the nowcast

$$\text{PI}_{1-2\alpha}(t) \;=\; \Big[\underline{q}_{\alpha},\, \overline{q}_{\alpha}\Big].$$

   This interval captures the sampling uncertainty generated by a potential undersampling of the starting data and the variability introduced by the possible presence (or imposition) of structural breaks (Chatfield, 1993).

These intervals are estimated within an Expanding+Rolling protocol with hyperparameter selection through TPE and pruning (and early stopping, in the case of neural networks). In particular:

- at each sample expansion, the best hyperparametric configuration is identified in the train+validation subset, training the model on updated historical data (Akiba et al., 2019; Bergstra et al., 2011; Shahriari et al., 2016);
- once the hyperparameters have been fixed, the nowcast for the following quarter (test quarter) is produced;
- finally, to evaluate the sensitivity of the forecast (and the corresponding confidence interval) to potential breaks, the pair block bootstrap was repeated on the same train+validation window, obtaining the distribution of predictions and the related extreme quantiles.

The process continues iteratively, moving forward one quarter until covering the entire range of out-of-sample forecasts, ensuring consistency with the Expanding+Rolling strategy (Bańbura et al., 2010).

Additionally, to evaluate how predictive uncertainty evolves under different macroeconomic regimes, we compute the average width of the prediction intervals across several sub-periods (Pre-COVID, COVID, Post-COVID, Excluding-COVID) and the entire forecast

horizon (Overall). This breakdown highlights how global disruptions (e.g., COVID-19) alter the scale of forecast uncertainty compared to more stable phases, offering a more granular assessment of each model's sensitivity to structural shocks.

The "prediction intervals around nowcasts" thus calculated preserve the series' internal correlation, integrate the effect of historical-economic breaks (even if not confirmed by the Bai-Perron statistical test), and provide empirical uncertainty measures valid on multiple models, reflecting the degree of risk associated with each nowcast. At the same time, these intervals serve as a crucial verification criterion for measuring forecast stability, contributing to a more robust and informed comparative analysis between different methodologies (Makridakis et al., 2022).

2.5 Feature Importance and Confidence Intervals

2.5.1 Explainability Methods per Model Family

Constructing a coherent interpretability framework for macroeconomic nowcasting requires reconciling each model's internal logic with the temporal dependencies inherent in time-series data. Our primary objective is twofold. First, we aim to align the feature importance measure with the specific structure of every model, whether linear or nonlinear, so that we retain the model's characteristic effects (e.g., shrinkage, gradient-based splits, or latent-factor extractions). Second, we seek to determine whether distinct estimation procedures converge on similar rankings of predictors, thus providing a robust cross-check of our nowcasting results.

We choose not to adopt computationally intensive methods, such as time-series variants of SHapley Additive exPlanations (SHAP)[50], for two main reasons:

- hardware constraints, since model-agnostic interpretability tools may require extensive resampling or repeated forward passes to compute contributions for each feature and each time point;
- our preference for model-specific techniques directly leveraging the estimation principles already embedded within each forecast.

Instead, we relied on measures compatible with rolling and expanding time windows, preserving seasonality and potential structural shifts while minimizing computational overhead. That maintained the temporal structure of our data, which is necessary for capturing real-world macroeconomic regimes.

We divided our methodology into a triple analysis to gain an overall view of the variables' importance:

[50] SHAP (Shapley Additive Explanations) is a "game-theoretic" approach to model interpretation in which the contribution of each explanatory variable is computed based on "Shapley values," i.e., the average marginal effect of including that variable across all possible coalitions, as derived from cooperative game theory. The so-called "time-series SHAP" methods extend this algorithm by introducing dynamic baselines—which replace a single static reference with time-varying or regime-specific points of comparison, thereby adapting the measure of each variable's impact to evolving macroeconomic conditions—and similar adaptations for capturing temporal dependencies and regime changes typical of macroeconomic forecasting. See Lundberg and Lee (Lundberg & Lee, 2017) for the theoretical foundations of SHAP, and Janizek et al. (Janizek et al., 2021) for applications to time-series predictions.

- *Global analysis*, where all the final values of feature importance, relative to the entire period, are accumulated, and their average (or other appropriate aggregate statistic) is calculated. That allows us to highlight which features are overall important beyond any structural variations in the economic system;

- *Analysis by sub-periods*, breaking down the observed horizon into macro-phases identified based on specific events or situations. For each sub-period, we calculate the average importance scores for the forecasts falling within that time interval. In this way, any differences in the role of specific features are highlighted: for example, some of them could prove critical in the contraction phases, while others could emerge in the recovery phase;

- *Quarterly analysis*, where we keep track of iteration by iteration, generating a real historical series of the feature importance (quarter-by-quarter evolution). That makes it possible to promptly identify gradual or sudden variations in the relevance of specific predictors, highlighting any moments of transition or macroeconomic shocks.

In each of these perspectives (global, sub-periods, and quarterly), we sorted the features by the absolute value of their importance scores. Finally, we identified the top 10 for immediate ranking and interpretability. This final step simplifies the comparison of relevant predictors, facilitating an at-a-glance overview of which variables dominate under various economic conditions or time frames.

This triple perspective extends well beyond a single global viewpoint, enabling in-depth exploration of how variables may fluctuate in importance when subjected to structural breaks or nonlinearities. By retaining the internal rationale of each model's fitting procedure, we capture a finer-grained picture of the drivers behind our forecasting. Consequently, our interpretability framework offers a scalable and consistent method across various model classes and provides valuable insights into the temporal progression of each feature's influence. This approach is especially pertinent in macroeconomic contexts, where economic indicators can shift rapidly due to policy changes, crises, or other exogenous shocks, calling for nuanced and computationally feasible interpretability solutions.

Coefficient-Based Feature Importance in Penalized Linear Models

In penalized linear models (e.g., LASSO (Scikit-learn Developers, 2025b; Tibshirani, 1996), Ridge (Hoerl & Kennard, 1970; Scikit-learn Developers, 2025f), and Elastic Net (Hastie et al., 2015; Scikit-learn Developers, 2025a; Zou & Hastie, 2005)), the estimated coefficients are direct indicators of the influence exerted by each explanatory variable on the forecasts. From this perspective, we can consider the Coefficient-Based Feature Importance (CBFI) as an interpretability measure for penalized linear models. It consists in interpreting the value (positive or negative) of each coefficient as a signal of the direction of the effect and its magnitude as a measure of the intensity of the impact on the target.

This approach hinges on the idea that, in a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.4}$$

the coefficient β_j associated with each predictor $\mathbf{X}_j$ reflects the sensitivity of the dependent variable to that predictor (Hamilton, 1994; Hastie et al., 2009). In the case of penalized

models, the estimation process introduces a constraint (of type L1, L2, or hybrid) that tends to reduce—or, in some cases, cancel—the coefficients associated with the less relevant features. Formally, for instance, a LASSO model solves

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \; + \; \lambda \sum_j |\beta_j| \right\}, \tag{2.5}$$

where λ controls the degree of shrinkage, pushing less important β_j toward zero (Tibshirani, 1996; Zou, 2006).

Consequently, those regressors characterized by coefficient values $\hat{\beta}_j$ significantly different from zero in terms of amplitude assume a central role in explaining the variability of the target quantity. In contrast, coefficients close to zero suggest a lower relevance. Furthermore, this approach maintains a model-specific connotation by relying entirely on the linear regression structure and the shrinkage effects of regularization (Smeekes & Wijler, 2018; Uematsu & Tanaka, 2019).

In our linear models, the procedure for calculating the importance of the variables via CBFI follows these steps:

1. **Iterated estimation of the penalized model.** After selecting the corresponding optimal hyperparameters, regression fitting (e.g., LASSO, Ridge, or Elastic Net) is performed during each iteration (Akiba et al., 2019; Bergstra et al., 2011).

2. **Extraction and storage of coefficients.** At the end of each iteration, the coefficients associated with only the non-dummy features are recorded. For each predictor, we note the numerical value of the coefficient, its sign (positive or negative), and the corresponding magnitude in absolute value. That allows us to trace its direction and intensity immediately (Müller & Guido, 2016).

3. **Aggregation over time.** The collected coefficients are subsequently managed according to three perspectives:

   - *Global*: all the iterations available on the entire time sample are considered, computing the average coefficient of each variable, e.g.

   $$\overline{\beta}_j^{(\text{global})} \; = \; \frac{1}{|\mathcal{T}_{\text{all}}|} \sum_{t \in \mathcal{T}_{\text{all}}} \hat{\beta}_j^{(t)}, \tag{2.6}$$

   where $\mathcal{T}_{\text{all}}$ indicates the set of all iterations in the sample (Makridakis et al., 1998; Salkever, 1976);

   - *By sub-periods*: the sample is divided into relevant economic phases (Pre-COVID, COVID, Post-COVID, Excluding-COVID), and the average of the coefficients is calculated only for the iterations whose test-quarters fall in each sub-period. Formally, if $\mathcal{T}_p$ is the subset of iterations referring to the p-th sub-period, then

   $$\overline{\beta}_j^{(p)} \; = \; \frac{1}{|\mathcal{T}_p|} \sum_{t \in \mathcal{T}_p} \hat{\beta}_j^{(t)}. \tag{2.7}$$

- *Quarterly*: the entire sequence of iteration-based coefficients is retained, forming a time series for each feature. Formally, for feature j we define

$$\mathcal{Q}_j = \left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \ldots, , \hat{\beta}_j^{(n)}\right), \tag{2.8}$$

where $\mathcal{Q}_j$ is the time sequence of the coefficients related to the variable estimated at each iteration and $\hat{\beta}_j^{(t)}$ is the coefficient obtained at iteration t. This quarter-by-quarter evolution facilitates the prompt detection of gradual or sudden changes in a variable's importance and potential shifts in the underlying macroeconomic regime (Clements & Hendry, 1999; Pesaran et al., 2006).

4. **Definition of the relevance ranking.** In each of the above perspectives (global, sub-periods, and quarterly), the features are ordered based on the absolute value of the average coefficient. Thus, the predictors that present values significantly different from zero in modulus (i.e., large $|\beta_j|$) emerge as the most influential (Belloni et al., 2014; Kapetanios & Zikes, 2018).

The choice of using a Coefficient-Based Feature Importance in linear contexts with the regularization logic adopted by penalized models proves advantageous and consistent because:

- it offers a feature-simple selection criterion based on a single summary value (the coefficient or its average over time) (Hastie et al., 2015; Hoerl & Kennard, 1970; Tibshirani, 1996);
- it does not require post hoc interpretative techniques since the coefficients are easily interpretable in economic or statistical terms (Smeekes & Wijler, 2018; Zou & Hastie, 2005);
- it directly reflects the penalizing dynamics imposed by the model. The L1 and/or L2 regularizers push the less important coefficients towards zero (zeroing or significant reduction), suggesting a negligible contribution in a given temporal context and facilitating the identification of a limited number of relevant predictors (Kim & Swanson, 2018; Zou, 2006);
- it is largely justified in our analysis framework, both for the need for a global overview of the main economic determinants and for the interest in isolating significant differences between different phases of macroeconomic evolution (Aliaj et al., 2023; Carriero et al., 2015a; Kapetanios & Zikes, 2018).

Coefficient-Based Feature Importance in Principal Component Regression

Also in PCR (Jolliffe, 1982; Jolliffe & Cadima, 2016; Massy, 1965), the estimated coefficients are direct indicators of each explanatory variable's influence on the forecasts, albeit through an intermediate transformation. From this perspective, we can also view the Coefficient-Based Feature Importance as an interpretability tool for PCR models. Specifically, each coefficient's sign signals the direction of the effect, and its magnitude measures the intensity of the impact on the target once those coefficients are re-expressed in the space of the original predictors.

This approach revolves around the idea that, in a standard linear relationship

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.9}$$

the coefficient β_j reflects how sensitive the dependent variable is to changes in predictor $\mathbf{X}_j$ (Hamilton, 1994; Hastie et al., 2009). However, under PCR, we first reduce $\mathbf{X}$ to $\mathbf{Z}$—a set of orthogonal principal components (Jolliffe & Cadima, 2016)—and then fit a regression (a penalized one using Ridge, in our case) of the form:

$$\min_{\boldsymbol{\gamma}} \Big\{ \tfrac{1}{2n} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}\|^2 \; + \; \lambda \sum_k \gamma_k^2 \Big\}, \tag{2.10}$$

where:

- $\boldsymbol{\gamma}$ are the coefficients in the principal component space;
- λ controls the degree of shrinkage.

Only after projecting these $\hat{\gamma}_k$ back to the original domain (via the loadings from the PCA) (Jolliffe, 1982), we obtain a set of $\hat{\beta}_j$ corresponding to each original predictor. As in penalized models (Smeekes & Wijler, 2018), large absolute values of $\hat{\beta}_j$ reveal the most relevant features, whereas small (or near-zero) coefficients indicate minor relevance. In this manner, PCR retains a model-specific interpretation anchored in a linear structure, though partially mediated by the dimension-reduction step (Massy, 1965).

In our PCR framework, the procedure for calculating the variables' importance via CBFI proceeds through the following steps:

1. **Iterated estimation of the PCR model.** We perform the PCR fitting at each iteration after selecting the optimal number of components and penalty hyperparameters (if used) (Akiba et al., 2019; Bergstra et al., 2011). That includes:
   - applying Principal Component Analysis to the current training set of $\mathbf{X}$;
   - estimating a penalized regression of $\mathbf{y}$ on the principal components $\mathbf{Z}$.
2. **Projection and storage of coefficients.** At the end of each iteration, we map the estimated coefficients $\hat{\gamma}_k$ in the Principal Component space back to the domain of the original predictors using the loadings matrix. Hence, for each non-dummy predictor, we obtain a numerical value $\hat{\beta}_j$, its sign, and its magnitude in absolute value, thereby identifying direction and intensity (Müller & Guido, 2016).
3. **Aggregation over time.** The resulting $\hat{\beta}_j$ values are then processed according to three distinct perspectives (similarly to what we did in penalized linear models):
   - *Global*: we compute the average coefficient for each variable over all iterations. Formally,
   $$\overline{\beta}_j^{(\text{global})} \; = \; \frac{1}{|\mathcal{T}_{\text{all}}|} \sum_{t \,\in\, \mathcal{T}_{\text{all}}} \hat{\beta}_j^{(t)}, \tag{2.11}$$
   where $\mathcal{T}_{\text{all}}$ indicates the set of available iterations (Makridakis et al., 1998; Salkever, 1976).

- *By sub-periods*: the sample is split into meaningful macroeconomic phases (Pre-COVID, COVID, Post-COVID, Excluding-COVID). Within each sub-period p, we average the coefficients obtained during the iterations that fall into that sub-period:

$$\overline{\beta}_j^{(p)} \;=\; \frac{1}{|\mathcal{T}_p|} \sum_{t \,\in\, \mathcal{T}_p} \hat{\beta}_j^{(t)}, \tag{2.12}$$

  enabling a comparison of variable importance across different economic regimes (Pesaran & Timmermann, 2007; Stock & Watson, 2007).

- *Quarterly*: for each feature, we consider the full-time series of iteration-based coefficients:

$$\mathcal{Q}_j \;=\; \Big(\hat{\beta}_j^{(1)},\; \hat{\beta}_j^{(2)},\; \ldots,\; \hat{\beta}_j^{(n)}\Big), \tag{2.13}$$

  which allows the detection of gradual shifts or abrupt changes in a variable's importance over the rolling sequence (Clements & Hendry, 1999).

4. **Definition of the relevance ranking.** The analysis ranks the features according to the absolute values of their average coefficients across different perspectives (global, sub-periods, quarterly). Predictors with large $|\beta_j|$ stand out as key drivers of the target, whereas those with coefficients near zero are deemed less influential (Belloni et al., 2014; Kapetanios & Zikes, 2018).

Implementing a Coefficient-Based Feature Importance in the context of Principal Component Regression offers a natural extension of the logic employed in penalized linear models with an added dimension-reduction step. In particular:

- it allows the adoption of a succinct selection criterion based on a single summary value (the back-projected coefficient or its temporal average) (Jolliffe & Cadima, 2016);

- it preserves interpretability by restoring each Principal Component coefficient to the space of the original explanatory variables, obviating the need for additional post hoc explanations (Smeekes & Wijler, 2018; Zou & Hastie, 2005);

- it aligns with the regularization framework of Ridge regression in the Principal Component domain, guiding coefficient shrinkage and focusing attention on the most relevant components and, ultimately, on the most relevant predictors (Hoerl & Kennard, 1970; Kim & Swanson, 2018);

- it is justified in scenarios requiring both a global perspective on dominant factors and a deeper scrutiny of distinct macroeconomic segments, enabling the identification of shifts in the role of explanatory variables over time (Carriero et al., 2015a; Pesaran & Timmermann, 2007; Stock & Watson, 2007).

This additional projection step does not alter the fundamental meaning of $\hat{\beta}_j$ regarding its direction and magnitude; instead, it embeds the penalty-induced shrinkage within the dimensionality-reduced space (Jolliffe, 1982). Therefore, PCR's Coefficient-Based Feature Importance approach remains consistent with its penalized linear counterparts while accommodating the PCA transformation.

Contribution of Components via Variable Importance in Projection in Partial Least Squares Regression

In Partial Least Squares Regression (PLSR) (Abdi, 2010; Mehmood et al., 2012; Wold, 1985), the extracted latent components are direct indicators of the influence exerted by each explanatory variable on the forecasts, albeit through an intermediate step tied to dimensionality reduction. From this perspective, the Variable Importance in Projection (VIP) (Akarachantachote et al., 2014; Chong & Jun, 2005) is an interpretability measure for PLSR. It consists of interpreting the VIP value of each feature as a signal of the intensity (magnitude) of its impact on the target. By definition, VIP values are always positive. Hence, they quantify only the magnitude of a predictor's contribution to the latent factors and do not convey any information about the sign of the relationship with the target.

This approach hinges on the idea that, in a PLS framework

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \boldsymbol{E} \quad \text{and} \quad \mathbf{y} = \mathbf{T}\mathbf{q} + \boldsymbol{f}, \tag{2.14}$$

the latent factors $\mathbf{T}$ reflect the directions in $\mathbf{X}$ that maximally covary with $\mathbf{y}$ (Krämer & Sugiyama, 2011). Consequently, those regressors characterized by large VIP scores assume a central role in explaining the variability of the target quantity. In contrast, VIPs close to or below unity suggest a lower relevance (Akarachantachote et al., 2014; Chong & Jun, 2005). Furthermore, this approach maintains a model-specific connotation by relying entirely on the PLS decomposition, which aligns $\mathbf{X}$ and $\mathbf{y}$ in a shared latent space.

As described in our PLSR forecasting section, this estimation routinely adopts an expanding or rolling window. Thus, we recalibrate the model with newly available observations in each iteration before measuring VIP (Fuentes et al., 2015). That consistency ensures that the interpretative measure aligns with our one-step-ahead forecast procedure.

In our PLSR setting, the procedure for calculating the importance of the variables via VIP follows these steps:

1. **Iterated estimation of the PLSR model.** After selecting the corresponding optimal hyperparameters (e.g., the number of PLS components), we perform regression fitting during each iteration. In this phase—consistently with the rolling or expanding window described earlier—the model extracts latent factors from $\mathbf{X}$ that are most predictive of $\mathbf{y}$.

2. **Extraction and storage of VIP scores.** At the end of each iteration, the VIP scores of only the non-dummy features are recorded. For each predictor, we note the numerical value of the VIP, which captures its relative contribution to the latent factors underlying the forecasts (Mehmood et al., 2012).

3. **Aggregation over time.** The collected VIP scores are subsequently managed according to three perspectives:

   - *Global*: we average each variable's VIP over all the iterations on the entire time sample, e.g.,

$$\overline{\mathrm{VIP}}_j^{(\mathrm{global})} = \frac{1}{|\mathcal{T}_{\mathrm{all}}|} \sum_{t \in \mathcal{T}_{\mathrm{all}}} \mathrm{VIP}_j^{(t)}, \tag{2.15}$$

where $\mathcal{T}_{\text{all}}$ indicates the set of all iterations (Makridakis et al., 1998; Salkever, 1976). This global average highlights the variable's overall relevance across the entire sample.

- *By sub-periods*: the sample is divided into relevant macroeconomic phases (Pre-COVID, COVID, Post-COVID, Excluding-COVID), and the average VIP is computed only for the iterations corresponding to each sub-period. Formally, if $\mathcal{T}_p$ is the subset of iterations referring to the p-th sub-period, then

$$\overline{\text{VIP}}_j^{(p)} = \frac{1}{|\mathcal{T}_p|} \sum_{t \in \mathcal{T}_p} \text{VIP}_j^{(t)}. \tag{2.16}$$

  That allows us to detect how a feature's influence changes across distinct economic contexts.

- *Quarterly*: the entire sequence of iteration-based VIP scores is retained, forming a time series for each feature. Concretely, for feature j we define

$$\mathcal{V}_j = \left(\text{VIP}_j^{(1)}, \text{VIP}_j^{(2)}, \ldots, \text{VIP}_j^{(t)}\right), \tag{2.17}$$

  where $\mathcal{V}_j$ represents the temporal progression of VIP estimates at each iteration. This quarter-by-quarter evolution facilitates the prompt recognition of gradual or sudden shifts in a variable's importance.

4. **Definition of the relevance ranking.** The features are ordered based on their average VIP in the above perspectives (global, sub-periods, and quarterly). Predictors that present high VIP_j values indicate a notable influence on the latent factors driving the forecasts, whereas variables with lower VIPs appear less relevant. This ranking helps isolate the few most influential regressors.

The choice of using VIP-based feature importance in a PLSR context proves advantageous and consistent because:

- it offers a straightforward selection criterion based on a single summary value (the VIP or its average) (Mehmood et al., 2012; Wold, 1985);
- it does not require post hoc interpretative techniques since the VIP is intrinsically connected to the latent factors discovered by PLS (Krämer & Sugiyama, 2011);
- it directly reflects the covariance-driven dynamics of PLS, prioritizing features that best align with the explanatory structure of $\mathbf{y}$ (Kim & Swanson, 2018);
- it is largely justified in our forecasting framework, both for obtaining a global overview of dominant economic determinants and for isolating significant differences across diverse macroeconomic regimes (Fuentes et al., 2015);
- it preserves a model-specific connotation by relying on PLS decomposition, ensuring that each predictor's importance is measured consistently with the structure used for predictive modeling.

The emphasis on maximizing covariance rather than solely capturing variance does not alter the fundamental role of each predictor with respect to the target; instead, it leverages the latent-factor structure of PLSR to highlight the most relevant explanatory directions. Therefore, the VIP-based assessment of feature importance remains fully coherent with the logic adopted in penalized linear models and PCR while incorporating an explicit alignment of $\mathbf{X}$ and $\mathbf{y}$ that enhances interpretability and predictive accuracy.

Impurity-Based Feature Importance in Random Forest

In Random Forest models (Breiman, 2001a; Breiman et al., 1984), the Impurity-Based Feature Importance (also known as Mean Decrease in Impurity, MDI) directly indicates the influence exerted by each explanatory variable on the forecasts. From this perspective, the Impurity-Based Feature Importance is a model-specific interpretability measure for Random Forests. It consists of interpreting the average reduction in node-level impurity[51] (for instance, mean squared error for regression trees) associated with each predictor, thus capturing the magnitude of its impact on the target (Mullainathan & Spiess, 2017).

This approach hinges on the idea that, in a Random Forest:

$$\mathrm{RF} \;=\; \Big\{\mathrm{Tree}_1,\; \mathrm{Tree}_2,\; \ldots,\; \mathrm{Tree}_B\Big\}, \tag{2.18}$$

each Tree_b is independently built on a bootstrap subsample, and the splitting at each node is chosen to maximize the decrease in impurity. In a regression setting, we can define a succinct formalism for the MDI of the variable j as:

$$\mathrm{MDI}_j \;=\; \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{\ell\,\in\,\mathcal{L}_b(j)} \Delta(\mathrm{Impurity})_\ell\right], \tag{2.19}$$

where:

- B is the total number of trees;
- $\mathcal{L}_b(j)$ denotes the set of nodes in the b-th tree that use feature j for splitting;
- $\Delta(\mathrm{Impurity})_\ell$ quantifies the local reduction in MSE at node ℓ.

In regression contexts, Random Forests may adopt either "squared error" or "absolute error" as the node-level impurity measure. In our study, the choice between these two criteria is determined dynamically via the TPE-based hyperparameter optimization (Borup et al., 2023; Shahriari et al., 2016). During each iteration, our model tries both options and selects the one minimizing the validation mean squared error:

[51] In a regression tree, "impurity" measures how heterogeneous the target values are within a node. Typically, for node ℓ containing t observations $\{y_1, \ldots, y_t\}$, impurity can be defined via the variance of the y_i (equivalent to the MSE). An alternative measure is the mean absolute deviation from the median (MAE), which may be more robust to outliers. In both cases, reducing impurity leads to more homogeneous nodes and enhanced predictive performance. For classification trees, criteria such as Gini or entropy are instead employed (Breiman et al., 1984).

- *Squared Error Criterion:* if the algorithm chooses "squared error," then the node-level impurity corresponds to the mean of squared residuals with respect to the sample mean. Specifically, if a node ℓ contains n observations $\{y_1, \ldots, y_n\}$ with average $\bar{y}$, its impurity is

$$\text{Impurity}(\ell) \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{2.20}$$

  When splitting ℓ on feature X_j into left and right child nodes, we define

$$\Delta(\text{Impurity})_\ell \;=\; \text{Impurity}(\ell) \;-\; \Big[\tfrac{n_L}{n}\,\text{Impurity}(\ell_L) \;+\; \tfrac{n_R}{n}\,\text{Impurity}(\ell_R)\Big], \tag{2.21}$$

  where n_L and n_R are the cardinalities of the two child nodes of the parent node n ($n = n_L + n_R$). The total importance for X_j accumulates these $\Delta(\text{Impurity})_\ell$ over all relevant splits in all trees of the ensemble.

- *Absolute Error Criterion:* if the algorithm chooses "absolute error," the impurity is based instead on the mean of absolute deviations from the median $\tilde{y}$. In that case,

$$\text{Impurity}(\ell) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big|\, y_i - \tilde{y}\big|, \tag{2.22}$$

  with the node-level decrease in impurity $\Delta(\text{Impurity})_\ell$ computed analogously as the difference between parent and children nodes' weighted absolute-error impurity.

Thus, Mean Decrease in Impurity (MDI) is conceptually similar in both cases: each split contributes its local decrease in impurity $\Delta(\text{Impurity})$ to the variable causing that partition, and we average these contributions across all trees to obtain

$$\text{MDI}_j \;=\; \frac{1}{B}\sum_{b=1}^{B}\sum_{\ell\,\in\,\mathcal{L}_b(j)}\Delta(\text{Impurity})_\ell.$$

Here, B is the total number of trees, and $\mathcal{L}_b(j)$ denotes the set of splits using X_j in the b-th tree. The only distinction is whether impurity is measured as variance-like (squared error) or deviation-like (absolute error), and the hyperparameter-tuning routine selects whichever yields a superior validation performance.

Regressors with high MDI values are central determinants of the overall predictive performance. On the other hand, regressors with near-zero MDI scores indicate low relevance. This method is closely tied to the model's specific features because it depends solely on how decisions are made through the tree-splitting process in the Random Forest (Breiman, 2001a).

In line with the principles illustrated in the previous sections (Coefficient-Based Feature Importance in penalized linear models, PCR, and PLSR), the procedure adopted to estimate the importance of the variables via Impurity-Based Feature Importance consists of four phases:

1. **Iterated estimation of the Random Forest model.** After selecting the optimal hyperparameters (for example, number of trees and maximum depth), we estimate the Random Forest at each iteration, associating each iteration with a specific time block (rolling or expanding). In each ensemble of trees, predictors that achieve splits with more significant error reduction at the nodes acquire higher importance (Borup et al., 2023).

2. **Extraction and storage of impurity-based importances.** At the end of each iteration, we obtain an MDI vector $\hat{I}_j$ for all the non-dummy features. Each value corresponds to the average reduction in impurity attributable to a given variable $\mathbf{X}_j$ in the set of trees in the forest. We record these values to monitor each predictor's contribution and trace its evolution. However, it is worth noting that highly correlated predictors can inflate the MDI (Nembrini et al., 2018; Strobl et al., 2008), similar to issues discussed with other methodologies in cases of multicollinearity. Alternative approaches, such as permutation-based importances, can mitigate correlation biases albeit at a higher computational cost.

3. **Aggregation over time.** Analogous to the approach illustrated for coefficients (CBFI) or VIP in PLSR, the MDI values collected at each iteration are then handled according to three perspectives:

   - *Global*: we consider all the iterations over the entire sample, computing the average MDI for each variable, for instance

   $$\overline{I}_j^{(\text{global})} \;=\; \frac{1}{|\mathcal{T}_{\text{all}}|} \sum_{t \,\in\, \mathcal{T}_{\text{all}}} \hat{I}_j^{(t)}, \tag{2.23}$$

   where $\mathcal{T}_{\text{all}}$ is the set of all available iterations (Mullainathan & Spiess, 2017). That offers a comprehensive view of the importance of $\mathbf{X}_j$ over the entire period.

   - *By sub-periods*: if the sample is split into macroeconomic phases (Pre-COVID, COVID, Post-COVID, Excluding-COVID), we compute the average $\hat{I}_j$ only for the iterations falling into each p-th sub-period, formally

   $$\overline{I}_j^{(p)} \;=\; \frac{1}{|\mathcal{T}_p|} \sum_{t \,\in\, \mathcal{T}_p} \hat{I}_j^{(t)}, \tag{2.24}$$

   thus identifying whether the contribution of each predictor is higher or lower in certain economic contexts.

   - *Quarterly*: for each variable $\mathbf{X}_j$, we retain the entire sequence of iteration-based MDI values, defining

   $$\mathcal{M}_j \;=\; \Big(\hat{I}_j^{(1)},\, \hat{I}_j^{(2)},\, \ldots,\, \hat{I}_j^{(t)}\Big), \tag{2.25}$$

   thereby highlighting gradual or abrupt shifts in $\mathbf{X}_j$'s importance on a quarterly basis (Borup et al., 2023).

4. **Definition of the relevance ranking.** The variables are ordered based on the mean MDI value in all three perspectives (global, sub-periods, quarterly). Predictors with substantially high $\overline{I}_j$ stand out as central determinants, whereas those with near-zero values indicate a marginal influence.

The choice of using Impurity-Based Feature Importance (Mean Decrease in Impurity) in a Random Forest context proves advantageous and consistent because:

- it provides a single-value criterion derived directly from the ensemble's internal splitting mechanism;
- it does not require post hoc interpretative tools since the MDI emerges from the calculation of the impurities on the nodes (Breiman et al., 1984);
- it accounts for nonlinear relationships and possible interactions, thanks to the hierarchical structure of regression trees (Medeiros et al., 2021);
- it is justified in our iterative framework, both for obtaining a global overview of primary economic determinants and for revealing significant differences across macroeconomic sub-periods (Borup et al., 2023);
- it remains a model-specific measure, fully reflecting how each predictor reduces the error in node partitions and thus integrating seamlessly into the rolling or expanding estimation procedure.

Gain-Based Feature Importance in XGBoost

In XGBoost models, the Gain-Based Feature Importance (sometimes referred to simply as "Gain Importance") (T. Chen & Guestrin, 2016; XGBoost Developers, 2023) directly indicates the influence exerted by each explanatory variable on the forecasts. From this perspective, the Gain-Based Feature Importance is a model-specific interpretability measure for boosting methods (Lundberg et al., 2020; Zhou & Hooker, 2021). It consists of interpreting the average improvement in the objective function (such as a regularized mean squared error) associated with each predictor at split nodes, thus capturing the magnitude of its impact on the target. In line with the interpretative rationale observed in linear and factor-based approaches (Hastie et al., 2009; Jolliffe, 1982; Wold, 1985), this measure offers a more adaptable perspective on variable significance, mainly when the data-generating process includes nonlinearities or intricate interactions. However, like other model-specific importance, it only pinpoints which features most effectively reduce the loss function without disclosing whether their relationship with the outcome is positive or negative (Nembrini et al., 2018; Strobl et al., 2008; Zhou & Hooker, 2021). Consequently, it complements methods such as Coefficient-Based Feature Importance, Principal Component Regression, and Partial Least Squares Regression, especially under conditions of strong nonlinearity.

This approach centers on the concept that, in an XGBoost model:

$$\text{XGB} \;=\; \sum_{b=1}^{B} \text{Tree}_b, \tag{2.26}$$

each Tree_b is trained sequentially to reduce a differentiable loss function $\mathcal{L}$ (often with L_2 regularization), and the splitting at each node is chosen to maximize the gain, which measures the drop in the overall loss once a node is partitioned into two child nodes. Formally, for a node ℓ split on feature j, the gain can be expressed as

$$\mathrm{Gain}_\ell \;=\; \mathrm{Score}(\ell) \;-\; \Big[\mathrm{Score}(\ell_L) \;+\; \mathrm{Score}(\ell_R)\Big] \;-\; \Gamma, \tag{2.27}$$

where:

- $\mathrm{Score}(\ell)$ is the node-specific value derived from first- and second-order gradient statistics[52];
- ℓ_L and ℓ_R denote the left and right child nodes, respectively;
- Γ is a regularization penalty that discourages overly complex splits[53].

Thus, the Gain-Based Feature Importance (GBI) of the variable j in regression settings with XGBoost can be summarized as:

$$\mathrm{GBI}_j \;=\; \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{\ell\,\in\,\mathcal{L}_b(j)} \mathrm{Gain}_\ell\right], \tag{2.28}$$

where:

- B is the total number of boosting rounds (or trees);
- $\mathcal{L}_b(j)$ denotes the set of nodes in the b-th tree where feature j is used to split;
- Gain_ℓ quantifies the local improvement in the objective function at node ℓ.

In regression contexts, XGBoost commonly employs a squared-error objective with L_2 regularization (T. Chen & Guestrin, 2016; XGBoost Developers, 2023), but it may also adopt alternative loss formulations. For example:

- *Squared Error Criterion:* if the algorithm chooses a squared-error loss, the node-level $\mathrm{Score}(\ell)$ stems from the sum of squared residuals and their second-order derivatives, with an additional L_2 regularization term applied. The gain Gain_ℓ thus measures how much splitting on X_j at node ℓ reduces the regularized MSE-based objective.

[52] In a typical regression setting with squared-error loss (and L_2 regularization), the node-level score can be approximated by

$$\mathrm{Score}(\ell) \;=\; -\,\frac{\left(\sum_{i\in\ell} g_i\right)^2}{\sum_{i\in\ell} h_i + \lambda},$$

where g_i and h_i denote the first- and second-order gradients of the loss function evaluated on each observation i in node ℓ, and λ is the L_2 regularization coefficient. For classification tasks (e.g., logistic loss), XGBoost employs the corresponding gradient statistics of the logistic objective, preserving the same notion that a higher (more negative) score indicates a better local fit (T. Chen & Guestrin, 2016).

[53] In some references, Γ is denoted by γ and represents the minimum loss reduction required to partition a leaf node further, thereby controlling overfitting. A larger Γ (or γ) penalizes additional splits that offer only marginal gains.

- *Absolute Error Criterion:* alternatively, an absolute-error-oriented loss (or a pseudo-Huber variant) can be employed. In that scenario, $\mathrm{Score}(\ell)$ relies on gradient statistics originating from an L_1-like loss, and the gain quantifies the improvement in minimizing the absolute deviations at the node level.

Hence, different objective functions alter the exact $\mathrm{Score}(\ell)$ calculation but preserve the overarching concept that each split's Gain indicates how much that split (and hence the feature used) advances the model's predictive accuracy. The hyperparameter-tuning routine (e.g., TPE-based) can systematically compare these objectives (Akiba et al., 2019; Bergstra et al., 2011; Shahriari et al., 2016) and retain the one yielding the best validation performance.

In our study, we typically adopt a squared-error loss with L_2 regularization, but alternative objectives (including absolute-error-based functions) may also be employed if indicated by the validation results. The choice of hyperparameters (e.g., number of boosted trees, maximum depth of each tree, learning rate, regularization terms) is similarly handled via TPE-based optimization (Akiba et al., 2019; Bergstra et al., 2011), mirroring the procedure described for Random Forest. The model determines the best combination of these hyperparameters during each iteration and re-estimates the XGBoost ensemble, yielding updated gain-based importance scores.

High GBI values indicate central determinants for the predictive performance. Conversely, regressors with near-zero GBI scores have a marginal influence. Like other model-specific importances, the Gain-Based measure depends solely on the gradient-based splitting decisions within the ensemble and does not reveal the sign of the underlying relationship.

Following the same principles illustrated in previous sections (Coefficient-Based Feature Importance in penalized linear models, PCR, and PLSR), the procedure adopted to estimate the importance of the variables via Gain-Based Feature Importance consists of four phases:

1. **Iterated estimation of the XGBoost model.** After selecting the optimal hyperparameters (e.g., number of boosting rounds, maximum depth, and learning rate), we fit the XGBoost model at each iteration, associating each iteration with a specific time block (rolling or expanding). Within each ensemble, predictors that yield splits with more significant loss reduction gain higher importance.

2. **Extraction and storage of gain-based importances.** At the end of each iteration, we obtain a GBI vector $\hat{I}_j$ for all non-dummy features. Each value corresponds to the average gain attributable to a given variable $\mathbf{X}_j$. We record these values to monitor each predictor's contribution. As in Random Forest, highly correlated predictors may inflate GBI, and permutation-based methods or additional regularization adjustments might address such correlation biases at higher computational cost (Nembrini et al., 2018; Strobl et al., 2008).

3. **Aggregation over time.** Analogous to the approach used for Coefficient-Based Feature Importance or VIP in PLSR, the GBI values from each iteration are examined under three perspectives:

   - *Global*: we compute the average GBI for each variable over all iterations,

$$\overline{I}_j^{(\mathrm{global})} = \frac{1}{|\mathcal{T}_{\mathrm{all}}|} \sum_{t \in \mathcal{T}_{\mathrm{all}}} \hat{I}_j^{(t)}, \tag{2.29}$$

offering a comprehensive assessment of how relevant $\mathbf{X}_j$ is throughout the entire sample.

- *By sub-periods*: if the sample is divided into macroeconomic phases (Pre-COVID, COVID, Post-COVID, Excluding-COVID), we calculate

$$\overline{I}_j^{(p)} \;=\; \frac{1}{|\mathcal{T}_p|} \sum_{t \,\in\, \mathcal{T}_p} \hat{I}_j^{(t)}, \tag{2.30}$$

thereby highlighting whether each predictor's importance increases or decreases in specific economic contexts.

- *Quarterly*: for each variable, we retain the entire sequence

$$\mathcal{M}_j \;=\; \Big(\widehat{I}_j^{(1)},\, \widehat{I}_j^{(2)},\, \ldots\, \widehat{I}_j^{(t)}\Big), \tag{2.31}$$

of iteration-based GBI scores, revealing gradual or abrupt shifts in $\mathbf{X}_j$'s influence.

4. **Definition of the relevance ranking.** The variables are ordered according to their mean GBI in all three perspectives (global, sub-periods, quarterly). Predictors with substantially high $\overline{I}_j$ consistently appear as key drivers in the XGBoost ensemble, whereas those with near-zero scores have negligible effects on the model.

Adopting a Gain-Based Feature Importance approach in XGBoost offers several advantages:

- it provides a single-value criterion that arises naturally from the gradient-boosted splitting mechanism;
- it encapsulates nonlinear relationships and potential interactions due to the sequential refinement of residuals;
- it aligns well with rolling or expanding frameworks, offering both an overall view of key economic determinants and insights on how these determinants evolve across different macroeconomic regimes;
- it remains strictly model-specific, reflecting how each predictor contributes to the boosted ensemble's objective function at the node level.

Impurity-Based Feature Importance (MDI) in Random Forest and Gain-Based Feature Importance (GBI) in XGBoost represent model-specific approaches to interpretability. They quantify how much each variable contributes to improving the model's performance by reducing errors (measured by chosen loss measure—an impurity measure or a gradient-based objective) at different decision points (split nodes). However, these methods do not indicate whether a variable has a positive or negative effect on the target outcome. While they do not show direct cause-and-effect relationships, they can effectively capture complex and nonlinear connections, especially in macroeconomic situations where simpler linear models might be inadequate.

Therefore, MDI and GBI offer an alternative and complementary insight to CBFI in penalized and Principal Component regressions, and VIP in Partial Least Squares Regression into which predictors matter the most under conditions of high complexity or time-varying structures while at the same time requiring careful handling of potential correlation biases (Nembrini et al., 2018; Strobl et al., 2008).

Integrated-Gradients-Based Feature Importance in MLP and GRU

In Multilayer Perceptron (MLP) and Gated Recurrent Unit (GRU) models, the Integrated-Gradients-Based Feature Importance (IG) quantifies the contribution of each explanatory variable to the neural network's forecast. From this perspective, the Integrated Gradients (IG) technique is a model-specific interpretability measure for differentiable models. It relies on computing the integral of the gradients of the network output concerning the input features, thus capturing how each predictor influences the final prediction (Sundararajan et al., 2017).

This approach hinges on the concept that, in a neural network F:

$$F(\mathbf{x}) \;=\; \widehat{y},$$

the partial derivatives $\frac{\partial F}{\partial x_j}$ indicate how small perturbations in the input feature x_j affect the output $\widehat{y}$. The Integrated Gradients formula extends this idea by evaluating and averaging those partial derivatives along a straight-line path from a reference baseline $\mathbf{b}$ to the actual input $\mathbf{x}$.

In practice, we define a linear trajectory

$$\mathbf{x}(\alpha) \;=\; \mathbf{b} \;+\; \alpha\,\big(\mathbf{x}-\mathbf{b}\big), \quad \alpha \;\in\; [0,1],$$

where $\mathbf{b}$ is often a zero vector or another neutral reference. The Integrated Gradient $\mathrm{IG}_j(\mathbf{x})$ of feature j is then given by

$$\mathrm{IG}_j(\mathbf{x}) \;=\; \big(x_j - b_j\big) \;\times\; \int_0^1 \frac{\partial F\big(\mathbf{x}(\alpha)\big)}{\partial x_j}\, d\alpha.$$

Hence, the total importance of predictor j is derived by multiplying the difference $\big(x_j - b_j\big)$ by the average gradient along that path. A numerical approximation typically divides α into discrete steps and sums the corresponding gradients.

In both MLP and GRU models, this technique quantifies the magnitude of each input's impact on the final output (Freeborough & van Zyl, 2022; Paudel et al., 2023), without indicating its sign (positive or negative effect). Consequently, it complements methods such as Coefficient-Based Feature Importance, Principal Component Regression, or Partial Least Squares by addressing highly nonlinear or interaction-heavy data-generating processes.

A key practical decision involves choosing the baseline $\mathbf{b}$. In many applications, $\mathbf{b}$ is set to the zero vector, but alternative references include the average of the training set or any meaningful neutral point in the feature space. Similarly, the discretization of α (i.e., the step size or number of steps) influences both the computational burden and the numerical precision of the integrated gradients.

In analogy to the Random Forest (MDI) and XGBoost (GBI) frameworks, the neural network model (MLP or GRU) is fitted at each iteration of a rolling or expanding window.

Formally, we define:

$$\mathrm{NN} \;=\; \Big\{\mathrm{NN}^{(1)},\ \mathrm{NN}^{(2)},\ \ldots,\ \mathrm{NN}^{(T)}\Big\}, \tag{2.32}$$

where each $\mathrm{NN}^{(t)}$ is either an MLP or a GRU trained on a specific time block. We assume the hyperparameters (e.g., number of layers, hidden dimension, learning rate, regularization) are selected via TPE-based optimization or a similar approach. Once the network parameters are fixed for iteration t, the calculation of IG-based importances proceeds through four main steps:

1. **Train the model**: estimate $\mathrm{NN}^{(t)}$ on the designated training (and possibly validation) set within the rolling or expanding window.

2. **Compute Integrated Gradients**: for each non-dummy predictor j, evaluate $\mathrm{IG}_j^{(t)}(\mathbf{x})$ over an appropriate subset of observations, using the chosen baseline $\mathbf{b}$ and step discretization in α.

3. **Aggregate into a single importance score**: optionally average the individual-observation attributions, obtaining a scalar $\widehat{I}_j^{(t)}$ for each variable j.

4. **Record and store**: collect the resulting IG-based importances into a vector $\big[\widehat{I}_1^{(t)}, \ldots, \widehat{I}_p^{(t)}\big]$ for subsequent analysis.

At the conclusion of iteration t, the vector $\widehat{I}_j^{(t)}$ is available for all non-dummy features. These values measure how much each variable $\mathbf{X}_j$ contributes, on average, to the neural network's output relative to $\mathbf{b}$. We preserve these IG values to:

- monitor the role of each predictor over time,
- detect potential instability of the importance measure in the presence of collinearity,
- compare integrated gradients across distinct macroeconomic contexts.

In neural networks, correlated inputs may lead to partial redundancies in IG scores, analogous to how collinearity affects MDI and GBI, albeit via a different allocation mechanism governed by backpropagation.

Consistent with the principles adopted for Random Forest and XGBoost, we aggregate the IG values across all iterations according to three perspectives:

- *Global*: we compute the mean of $\widehat{I}_j^{(t)}$ over every iteration, resulting in

$$\overline{I}_j^{(\mathrm{global})} \;=\; \frac{1}{|\mathcal{T}_{\mathrm{all}}|} \sum_{t \,\in\, \mathcal{T}_{\mathrm{all}}} \widehat{I}_j^{(t)}, \tag{2.33}$$

  which captures a holistic view of how $\mathbf{X}_j$ influences the forecast over the full sample.

- *By sub-periods*: if the sample is split into distinct macroeconomic phases (e.g., Pre-COVID, COVID, Post-COVID, Excluding-COVID), we similarly define

$$\overline{I}_j^{(p)} \;=\; \frac{1}{|\mathcal{T}_p|} \sum_{t \,\in\, \mathcal{T}_p} \widehat{I}_j^{(t)}, \tag{2.34}$$

  facilitating a comparison of each predictor's relevance in different economic environments.

- *Quarterly*: we store the entire sequence

$$\mathcal{M}_j \;=\; \Big(\widehat{I}_j^{(1)},\, \widehat{I}_j^{(2)},\, \ldots\, \widehat{I}_j^{(t)}\Big), \tag{2.35}$$

  thus enabling an examination of gradual or abrupt shifts in $\mathbf{X}_j$'s importance on a quarter-by-quarter basis.

Following the same logic adopted for MDI or GBI, we establish an importance ranking by ordering the features according to their average IG score in the perspectives described above (global, sub-periods, quarterly). Predictors with high $\overline{I}_j$ values consistently appear as principal contributors, whereas variables whose IG scores remain close to zero signify a limited effect on the neural network's output. Because IG measures how the output changes relative to the baseline, these values are strictly model-specific and do not provide information on whether increments in $\mathbf{X}_j$ associate with a positive or negative effect on the target.

Adopting an integrated-gradients-based feature importance in neural networks offers several benefits:

- it provides a single-value criterion derived from backpropagation gradients, thus cohering with the model's core optimization process;
- it accommodates nonlinear relationships and interactions, reflecting the flexible structure of MLPs and GRUs;
- it naturally aligns with rolling or expanding frameworks, yielding both a comprehensive perspective of dominant predictors and a window-by-window analysis of importance variation;
- it clarifies how each predictor alters the fitted function from a baseline state to the actual input, forming a conceptual parallel to "impurity reduction," yet grounded in gradients rather than splits.

Similarly to other model-specific metrics (MDI, GBI), IG-based importances do not inherently convey the sign of the underlying relationship, nor are they entirely immune to correlation-related distortions. Nevertheless, this methodology remains a powerful tool for illuminating predictors' intricate and evolving significance in deep or recurrent neural architectures (Rudin, 2019).

2.5.2 Confidence Intervals for Feature Importance via Pair Block Bootstrap

In this study, we calculated confidence intervals for feature importance across all models, including linear ones, instead of relying on p-values. This approach provides a consistent measure of uncertainty for feature importance, regardless of the type of model used. Specifically, p-values are not readily computable for models such as XGBoost or GRU, making it challenging to assess the stability of feature importance in those cases (Wasserstein & Lazar, 2016). By employing confidence intervals for all models, we ensure a common foundation for comparison, allowing us to evaluate the stability of feature importance across a diverse range of model types. This methodological choice strengthens the comparability of our results, offering a unified framework for interpreting feature importance across different predictive models (Molnar, 2022; Shmueli, 2010).

The construction of "confidence intervals for feature importance" represents an essential step in quantifying the uncertainty surrounding the contribution of each predictor in macroeconomic models. This process provides a more robust interpretation of the role of variables in the model's predictive capacity (Breiman, 2001b).

In contexts where residuals exhibit temporal dependencies or structural breaks, advanced methods such as the block bootstrap (Künsch, 1989; Lahiri, 2003; Masarotto, 1990; Pan & Politis, 2016; Politis & Romano, 1994) estimate these confidence intervals, offering a more reliable measure of feature importance. Specifically, we adopt a pair block bootstrap, meaning that each resampled block contains both the explanatory variables $\mathbf{x}_t$ and the target variable y_t, thus preserving the simultaneous temporal structure of input-output data. As previously described in Section 2.4, where a pair block bootstrap was used for prediction intervals, this technique ensures that the temporal structure of the series is respected, thereby providing a resampling procedure agnostic to the distribution of the residuals.

The block bootstrap procedure can be broken down into the following steps:

1. **Subdivision of the dataset into sub-segments.** In each iteration, the training and validation dataset (up to the last in-sample quarter) is divided into sub-segments corresponding to key historical economic breaks, such as global crises or other major economic shocks. This segmentation is based on ex-ante economic knowledge of the periods during which significant regime shifts occurred rather than solely relying on statistical tests (such as the Bai-Perron test). By adopting this approach, we ensure that periods affected by global disruptions (e.g., the 2008 financial crisis or the COVID-19 pandemic) are adequately considered in calculating feature importance, even if these breaks were not statistically detected. This step is crucial for capturing potential structural changes in the data that might influence the importance of the features.

2. **Creation of contiguous blocks for each subsegment.** Within each sub-segment, the data pairs $(\mathbf{x}_t, y_t)$ are grouped into contiguous blocks of fixed size (e.g., four quarters). The choice of a block size corresponding to one year (four quarters) is intended to preserve the seasonality and internal serial autocorrelation typical of macroeconomic time series. This block size ensures that intra-annual dependencies are maintained, which is especially important when modeling macroeconomic data, where seasonal patterns play a significant role (Makridakis et al., 1998).

3. **Sampling with block replacement.** Several blocks, each of the same length, are

randomly extracted from each subsegment, with replacement, until the size of the bootstrap sample matches that of the original subsegment. This sampling method guarantees that temporal dependencies within the blocks are preserved while allowing the order and frequency of the blocks to vary between iterations. The key feature of this step is that it ensures the temporal dependence across blocks is maintained, which is essential for accurately assessing feature importance over time, as temporal relationships are critical in macroeconomic forecasting (Lahiri, 2003; Politis & Romano, 1994).

4. **Concatenation of the resampled subsegments.** The resampled blocks are concatenated to form an alternative training and validation dataset. Although this new dataset preserves the overall length and temporal structure of the original dataset, it is rearranged in such a way as to reflect different configurations of the data. This step ensures that the variability in feature importance is captured through multiple resampled datasets, reflecting potential shifts in the underlying data-generating process (Pan & Politis, 2016).

5. **Retraining the model and importance calculation.** Once the bootstrap sample $(\mathbf{x}_{\text{boot}}, y_{\text{boot}})$ has been generated, the model is retrained on this new dataset. The feature importance values are recalculated for each bootstrap iteration. This process is repeated B times (typically $B = 1000$) to generate a distribution of feature importance scores. Each iteration yields a new estimate of feature importance, which accounts for potential changes in the data structure and provides insight into the uncertainty associated with each feature's contribution to the model (Altmann et al., 2010; Ishwaran & Lu, 2019; Mentch & Hooker, 2016; Wager et al., 2014; Williamson et al., 2023).

6. **Construction of the confidence interval for feature importance.** After obtaining the feature importance values from each bootstrap iteration, the values are ordered, and the quantiles corresponding to the desired confidence level (e.g., 2.5% and 97.5%) are calculated. The lower and upper quantiles denoted as $\underline{q}_\alpha$ and $\overline{q}_\alpha$, respectively, define the confidence interval for each feature's importance:

$$\text{CI}_{1-2\alpha}(\text{Feature}) = [\underline{q}_\alpha, \overline{q}_\alpha]$$

These confidence intervals reflect the uncertainty in the estimated feature importance, considering both the inherent variability due to sampling and the potential impact of structural shifts or regime changes. The construction of these intervals allows for a more nuanced evaluation of feature stability over time, capturing the effect of external shocks such as global economic crises or pandemics (Bergmeir et al., 2016; Li, 2021).

This procedure is also applied iteratively within an Expanding+Rolling framework. The best hyperparameter configuration is determined in each iteration, and the model is retrained on the updated historical data. After the hyperparameters are fixed, feature importance for the next test set is computed, and the block bootstrap is again employed to evaluate the sensitivity of the feature importance estimates to possible breaks or variations

in the training data. This iterative process ensures that feature importance measures are continuously refined and accurately reflect changes in the data structure over time.

Furthermore, the average width of the confidence intervals for feature importance is calculated across multiple sub-periods (Pre-COVID, COVID, Post-COVID, Excluding-COVID). That enables an assessment of how structural changes, such as economic crises or pandemics, influence the uncertainty associated with the feature importance measures. The analysis allows us to evaluate the robustness of feature importance estimates under varying macroeconomic conditions and to understand how global disruptions impact the model's sensitivity to predictors. By comparing these intervals across different phases, we gain valuable insights into the dynamic nature of feature importance over time.

The "confidence intervals for feature importance" calculated in this manner provide reliable, valid uncertainty measures across various models. These intervals offer a deeper understanding of the variability in feature importance, serving as an essential tool for assessing the stability of feature importance across different model configurations and historical periods. This process contributes to a more informed and robust analysis of the role of predictors in macroeconomic forecasting.

2.6 Model Selection and Construction of Combined Models

2.6.1 Identifying Superior Models for Combination via the Model Confidence Set

The selection of a coherent subset of forecasting models constitutes a critical prerequisite for effective ensemble methods. In particular, macroeconomic forecasting often encompasses multiple candidate specifications (linear and nonlinear, parametric and nonparametric), each potentially suited to different phases of the economic cycle. Consequently, identifying a "Superior Set of Models" (SSM)[54] in a statistically rigorous manner is essential for mitigating forecast uncertainty and avoiding the inclusion of systematically underperforming models. In this section, we adopt the "Model Confidence Set" (MCS) framework (P. R. Hansen et al., 2011) to fulfill this task. We first outline the theoretical underpinnings of the MCS approach and then detail its step-by-step implementation. Our exposition aims to clarify how the MCS can be applied effectively in a time-series context, ensuring that temporal dependence is duly respected.

A Model Confidence Set is a statistical procedure identifying the best subset of models based on predictive accuracy. It tests, in an iterative and statistically controlled manner, whether any model in a pool is inferior[55] to the others by a statistically significant margin.[56] In the presence of K candidate models $\{M_1, \ldots, M_K\}$, the procedure proceeds as follows:

[54]The term "Superior Set of Models" (SSM) refers to the subset of candidate models that, according to the "Model Confidence Set" (MCS), cannot be statistically rejected as inferior at a chosen significance level. In our application, we set this level to $\alpha = 0.10$. (P. R. Hansen et al., 2011; P. R. Hansen, 2005; White, 2000)

[55]"Inferior" here indicates that a model's predictive loss is significantly higher than that of at least one competing model. Specifically, we say that model M_j is "inferior" if the p-value computed from the MCS bootstrap procedure falls below our significance threshold $\alpha = 0.10$. In other words, "significantly higher" means that the estimated difference in losses (relative to at least one other model) is unlikely to have arisen by chance under the joint null hypothesis of equal predictive ability. (P. R. Hansen et al., 2011)

[56]A "statistically significant margin" is any observed difference in mean losses between two models (or between one model and the group) that, after block bootstrap resampling, yields a test statistic with an associated p-value $< \alpha$. In our empirical implementation, $\alpha = 0.10$, so significance implies that we reject the null hypothesis of equal performance at the 10% level.

1. **Computes pairwise loss differentials.** Let $\ell_t(M_j)$ denote the loss function for model M_j at time t, where $t = 1, \ldots, N$. We focus on squared errors in our application (though the approach accommodates other loss measures). For each pair (M_i, M_j), define the pairwise difference

$$d_t(M_i, M_j) \;=\; \ell_t(M_i) \;-\; \ell_t(M_j).$$

   Intuitively, $d_t(M_i, M_j)$ indicates whether model M_i is performing better (lower loss) or worse (higher loss) than model M_j at each time t. Aggregating these differences over $t = 1, \ldots, N$ provides a basis for measuring whether M_i systematically outperforms M_j.

2. **Performs a joint hypothesis test.** Using a block bootstrap scheme to account for potential autocorrelation in the series $\{d_t(M_i, M_j)\}_{t=1}^{N}$, the MCS procedure tests if at least one model exhibits significantly larger losses than all others. This approach addresses the multiple-testing issue (i.e., controlling the family-wise error rate) and prevents inflating Type I error across many comparisons (P. R. Hansen, 2005; White, 2000).

3. **Eliminates inferior models.** A user-chosen significance level α (e.g., 0.10) governs the threshold for deciding whether a model is significantly worse. If a model is found consistently inferior to at least one competitor across the bootstrap replications, it is removed from the set. The test is then iterated on the remaining models until no further rejections occur.

In adopting this iterative screening, the MCS obviates the need for ex-ante assumptions about model validity; instead, it relies on evidence from the block bootstrap to identify those models that do not show persistent inferiority relative to others (Aiolfi & Timmermann, 2006).

Below, we summarize the main conceptual steps followed in our methodological application:

1. **Definition of the loss matrix.** We gather and organize the out-of-sample forecasting errors (in our case, squared errors) into an $N \times K$ loss matrix, where each column corresponds to one of the K candidate models, and each N row represents a specific time point (e.g., a quarter).

2. **Block bootstrap for test statistics.** To check if a model M_j is distinctly inferior to others, we employ a block bootstrap with a chosen block size k. In our empirical study, we set $k = 4$ quarters to preserve any annual pattern or seasonal correlation. We replicate the bootstrap B times (with $B = 10{,}000$ in our final specification) to ensure a stable estimate of the distribution of the pairwise differentials. Statistics such as "Tmax" or "TR"[57] are computed in each replication to ascertain whether any

[57] Both Tmax and TR are test statistics frequently used in MCS frameworks. Tmax refers to the maximum absolute t-statistic of the loss differentials across all model pairs, thereby identifying whether any single model stands out as worse relative to the rest. TR instead focuses on the range (or span) of these t-statistics, effectively capturing whether the overall spread of model performance is large enough to conclude that one model is outperformed by others. In each bootstrap replication, these test statistics are recalculated, yielding an empirical distribution against which the observed statistics are tested.

model significantly exceeds a certain loss threshold.

3. **Iterative elimination and significance.** If a model repeatedly fails the test (i.e., it shows statistically higher average loss at level $\alpha = 0.10$), it is expunged from the current set of models. The algorithm continues until all remaining models cannot be rejected as inferior, producing the final SSM.

4. **Rankings and p-values.** For interpretability, the MCS returns a ranking of models based on their performance metrics (e.g., the average differential or a t-based ordering) and a global p-value indicating how strongly the data support the joint hypothesis that no surviving model is worse than the others. In practical terms, a model included in the SSM is considered not distinctly inferior to the top performers, thereby qualifying for subsequent combination in the nowcasting ensemble (Amendola et al., 2020; Samuels & Sekkel, 2017).

By structuring the procedure in these steps, we ensure that each model is subjected to a consistent, statistically rigorous evaluation. The block bootstrap is especially vital in macroeconomic contexts, where serial correlation in forecast errors is the rule rather than the exception.

The MCS focuses on comparative performance across multiple forecasts: it does not proclaim a single, absolute "best model," but rather identifies those models for which there is insufficient evidence of inferiority[58] relative to the rest of the group. This balanced perspective is aligned with common practices in forecast evaluation, especially in real-time macroeconomic settings characterized by model uncertainty.

Including only the surviving models from the final SSM in subsequent ensemble methods has several advantages:

- *Statistical reliability:* because the procedure controls for multiple comparisons in a bootstrap-based manner, the set of retained models is less likely to contain consistently poor performers (P. R. Hansen et al., 2011; White, 2000).

- *Efficiency and clarity:* dropping inferior models curtails the risk of "averaging down" when forecasts are combined, as a large pool of candidates can dilute improvements (Bates & Granger, 1969).

- *Flexibility for real-time updating:* as new data become available, the MCS can be re-estimated, and the SSM can change accordingly if certain models degrade or others improve in predictive accuracy.

While the MCS does not claim that any included model is definitively superior in all circumstances, it ensures that none of the retained models shows statistically validated inferiority at the chosen level (i.e., $\alpha = 0.10$). In other words, the methodology offers a robust filter, grounded in time-series bootstrap inference, before forming aggregated nowcasts. This systematic screening aligns with our broader principle of reducing model risk by leveraging iterative hypothesis testing and bootstrap-based diagnostics, and it is

[58] "Insufficient evidence of inferiority" means that, within the bootstrap-based hypothesis testing framework, the loss differentials do not exhibit a consistent or statistically robust indication that a given model is outperformed by any other model in the set at $\alpha = 0.10$. (P. R. Hansen et al., 2011)

computationally more demanding than standard pairwise tests but furnishes a more reliable subset of forecast candidates.

In the subsequent sections, we illustrate how these retained models (i.e., the final Superior Set of Models) are combined through various aggregation techniques, including unweighted and weighted averaging. Such a two-stage strategy—MCS selection followed by forecast aggregation—helps balance the competing goals of robust performance (through screening) and informational breadth (by blending multiple specifications).

2.6.2 Combining Superior Models

The search for robust predictive accuracy often transcends the reliance on a single specification, even when the model under scrutiny emerges from rigorous selection procedures such as the Model Confidence Set (MCS). Indeed, the practice of combining forecasts traces back to seminal studies arguing that a suitable aggregation of predictions from multiple parsimonious models can outperform (in terms of bias or variance reduction) any single best contender[59](Bates & Granger, 1969). Recent refinements reinforce this intuition, suggesting that pooled forecasts mitigate specification risk in dynamic and potentially unstable contexts (Timmermann, 2006).

In macroeconomic scenarios, structural shifts, regime changes, and incomplete information can undermine the stability of individual estimators. Aggregating the predictions of multiple "superior" models—each capturing distinct data features or leveraging different identification strategies—can help offset model-specific fragilities. Furthermore, exogenous shocks or evolutions in data collection processes may cause an abrupt deterioration in the performance of even well-tuned estimators. A pooled strategy, by design, spreads the risk of relying excessively on a single specification. Consequently, these combination models provide a pragmatic safeguard, especially where the sample size is modest, or the pattern of economic indicators is subject to frequent fluctuations (Aiolfi & Timmermann, 2006).

In the present work, we focus on three principal techniques for combining the forecasts of the MCS-retained models:

- *Simple Average (SA)*: it assigns uniform weights to each model, thus offering a baseline benchmark. Although it overlooks differential predictive accuracy across models, it remains widely adopted for its transparency and resistance to abrupt overfitting.
- *Weighted Average (WA)*: it constructs a set of weights that scale inversely with the historical error of each model. Hence, those estimators that consistently exhibit lower losses receive higher weights when generating new out-of-sample predictions.
- *Exponential Weighted Average (EWA)*: it extends this idea by applying a dynamic exponential decay to the cumulative errors, thereby emphasizing more recent performances. In certain variants, the procedure itself is nested into a "meta-layer" (Meta-EWA), where several possible decay parameters (η) compete through an additional weighting mechanism.

[59]In the literature, this approach is frequently referred to as "forecast combination methods," "aggregation models," or "ensemble models." In the present work, we adopt the term "combination models" to avoid confusion with standard machine-learning "ensemble methods," such as random forests or gradient boosting, which we already employ as base learners in certain specifications.

While differing in the explicit formulation of the weighting scheme, these three methods share a common architecture: they revisit and aggregate the predictions at each time step, recording residuals and partial metrics (e.g., MSFE, RMSFE, MAFE) in a rolling or iterative fashion. Such a parallel structure enables a coherent comparison of strengths, limitations, and overall forecast accuracy across the aggregation approaches.

The next subsections elaborate on each of the aforementioned combination strategies. For each, we proceed as follows:

1. **Theoretical Rationale and General Functioning.** We illustrate the conceptual foundation of each "combination model," focusing on its objectives and the high-level mechanisms by which it aims to consolidate multiple base forecasts;

2. **Combination and Weighting Scheme.** We describe the specific method by which the individual model outputs are aggregated, emphasizing how weights are assigned or updated over time in each approach;

3. **Implementation in Our Empirical Setting.** We summarize how the combination procedure is integrated into our overall forecasting framework, indicating if it is executed in a purely offline fashion[60] or through a rolling re-estimation protocol. We also outline any relevant hyperparameters, where applicable.

These approaches offer a nuanced perspective on balancing multiple models' relative merits and minimizing misspecification's impact on final nowcasts. By systematically updating (or recalibrating) the weighting structure, we aim to accommodate potential breaks in economic relationships while preserving the benefits of diversification. In the subsequent subsections, we detail each technique's design and present the respective empirical findings in a standardized format, thus ensuring a direct comparability of results. We also compare WA forecasts with both simpler and more advanced approaches in order to determine whether partial RMSE-based weighting yields measurable improvements in predictive stability under diverse economic conditions.

Simple Average

A Simple Average (SA) is the most straightforward strategy for aggregating multiple model forecasts. It assigns each base learner (model) the same weight, thereby avoiding discrimination based on historical performance or model complexity. The underlying rationale is to minimize the risk of over-reliance on a single specification by evenly distributing the contribution of each model in the final forecast (Bates & Granger, 1969; Clemen, 1989; Smith & Wallis, 2009). Consequently, the SA approach offers a transparent and stable

[60] We refer to a procedure as "offline" (batch) if it loads the entire dataset ex-post, simulating the forecasting process over all observations already available.

In an "online" (or streaming) approach, the model continuously updates its parameters or weights as soon as new data points arrive without reprocessing all past observations.

In this study, each combination method operates in a batch mode: the data for the entire sample (e.g., 2017 Q1–2023 Q2) is loaded at once, and the step-by-step evolution of weights is replayed retroactively. If new observations become available, the entire procedure must be rerun. Our choice aligns with typical macroeconomic data releases, which appear at discrete intervals rather than in a high-frequency streaming format.

benchmark, particularly valuable when the pool of candidate models exhibits comparable performance or when no clear ex-ante indication suggests favoring a specific learner. Additionally, in contexts characterized by abrupt regime shifts or data outliers, uniform weighting can provide a robust baseline that reduces the influence of sudden performance degradations from any one component model (Makridakis et al., 2020).

From a statistical standpoint, the Simple Average can reduce variance by blending forecasts from distinct estimation methods, data transformations, or structural assumptions (Armstrong, 2001; Stock & Watson, 2004; Timmermann, 2006). This property is especially desirable in situations marked by parameter instabilities. Moreover, the mechanism does not require frequent re-estimation of weights. Once the set of base models has been determined, each forecast is given an identical portion of the overall "voting power." Although the method disregards potential differences in predictive accuracy, it often yields robust outcomes in empirical settings where economic or financial indicators are subject to substantial noise. In principle, it would be possible to introduce a dynamic aspect to the weights, but this would deviate from the uniform philosophy at the core of SA. In our script implementation, partial error metrics (such as MSFE or RMSFE) are tracked at each time step purely for diagnostic or comparative purposes, without feeding back into the weights. As such, any deterioration in the performance of a specific model does not prompt a reallocation of weights (B. Chen & Hood, 2020; Hyndman & Athanasopoulos, 2021).

Let the MCS procedure select a subset of M forecasting models, indexed by $m \in \{1, 2, \ldots, M\}$. At each forecasting period t, each model m produces a predicted value $f_{m,t}$. Under the Simple Average aggregator, the final combination $\widehat{y}_t$ is defined as a uniform mean of all base forecasts:

$$\widehat{y}_t \;=\; \frac{1}{M} \sum_{m=1}^{M} f_{m,t}.$$

Hence, all models are assigned the weight $1/M$, regardless of their historical performance or estimation strategy. When the size of the model set M remains fixed over the evaluation window, the aggregation proceeds identically for all time steps (Genre et al., 2010; Zhemkov, 2021).

Partial metrics, such as MSFE or RMSFE, are often computed at each time step to assess accuracy patterns or to compare with more complex weighting schemes. However, these statistics do not feed back into the weights for SA, since the method preserves the same assignment $(1/M)$ throughout the sample. Thus, no dynamic mechanism adjusts any coefficient in response to recent forecast errors. This design choice allows a direct comparison with methods such as Weighted Average or Exponentially Weighted Average, which do update their weights as new data arrive (B. E. Hansen, 2008; Makridakis & Hibon, 2000; White, 2000).

This weighting scheme offers two principal advantages:

- it is computationally simple: once model outputs are obtained, the final prediction requires no further parameters or recursive updates;
- it provides a strong baseline against which more refined methods (e.g., Weighted Average or Exponentially Weighted Average) may be tested. Supposing a given model begins to deteriorate due to structural breaks or unforeseen economic shocks, its impact on the combined forecast remains limited, as its influence is capped at

$1/M$ of the total prediction. However, an important caveat of SA is its insensitivity to differences in individual accuracy: should one model exhibit systematically higher predictive power, the Simple Average offers no mechanism to grant that model a proportionally larger weight.

In this study, the SA method is applied offline (batch mode) to the models retained by the preceding MCS-based selection. Specifically, once the MCS step determines which learners are statistically superior under the chosen loss function, each of these M models contributes equally to generate the final aggregated nowcast. The procedure operates as follows:

1. **Initialization.** Identify the subset of base models deemed superior by the MCS. No additional hyperparameters or performance tracking are required for the Simple Average method.

2. **Uniform Aggregation.** At each time t, compute each model's forecast $f_{m,t}$; sum these predictions and divide by M to obtain the SA forecast, $\widehat{y}_t$.

3. **Batch Application.** Because the weighting remains fixed, there is no re-estimation of weights once the set of models is chosen. If new observations become available and the set of superior models is unchanged, the SA forecast is updated simply by averaging the newly generated model predictions.[61]

4. **Final Nowcast.** Record $\widehat{y}_t$ for all t in the evaluation sample to compare against realized outcomes. Although the average is constant in weighting, it can still incorporate underlying structural differences indirectly if the base models capture distinct aspects of the data. During this step, partial error metrics are computed at each t to facilitate subsequent comparison with other aggregation methods. Period-specific analyses (including sub-period evaluations) are also carried out, mirroring the approach used for Weighted Average or Exponentially Weighted Average, without any effect on the fixed uniform weights.

The Simple Average provides a robust and parsimonious benchmark by obviating the need for a dynamic adjustment mechanism. Its forecasts may, in certain instances, rival or surpass those derived from more elaborate procedures, especially in moderate samples or under conditions where no single approach maintains a lasting predictive edge. Later sections compare this uniform aggregator with adaptive schemes, investigating whether more sophisticated weighting improves accuracy and stability across diverse economic environments.

Weighted Average

The Weighted Average (WA) combination strategy adjusts each model's influence according to its observed performance over a historical window (Clemen, 1989; Timmermann, 2006). In contrast to a simple average, WA grants proportionally significant weight to models

[61] Should the sample period be extended, all data points are reprocessed ex-post in a batch fashion. This reprocessing does not alter the uniform nature of the combination. Additional logging and error metrics (e.g., RMSFE, MAFE) are stored for robustness but do not modify any weighting.

displaying lower cumulative forecasting errors (Bates & Granger, 1969; Stock & Watson, 2004). It relies on the idea that models with consistently minor deviations from actual outcomes are more likely to capture the prevailing economic dynamics.

The core principle of WA is that the accuracy of each model, when measured and tracked over multiple periods, furnishes a credible basis for guiding future weights (Armstrong, 2001; Atiya, 2020). In practice, one first defines a performance metric, such as the partial root mean square error (RMSE), which reflects how effectively a model has predicted the target variable from the initial period up to the current time. Models with lower partial RMSE are considered more reliable and receive higher weights.

Let there be M distinct models and $d_{m,t}$ the in-sample forecast error (in our case, the squared difference between observed and predicted) for model m at time t. A partial RMSE, spanning the interval $[1, \ldots, t]$, can be inverted to yield a factor $\alpha_{m,t}$ that increases with model quality[62]:

$$\alpha_{m,t} = \frac{1}{\mathrm{RMSE}_{m,t}} \quad \text{where} \quad \mathrm{RMSE}_{m,t} = \sqrt{\frac{1}{t}\sum_{\tau=1}^{t} d_{m,\tau}}.$$

If $\mathrm{RMSE}_{m,t}$ is especially small, $\alpha_{m,t}$ becomes correspondingly large, signifying robust past performance. After computing $\alpha_{m,t}$ for each model, one applies a normalization step to ensure the sum of the weights to unity:

$$w_{m,t} = \frac{\alpha_{m,t}}{\sum_{j=1}^{M} \alpha_{j,t}}.$$

The coefficient $w_{m,t}$ represents the proportion of the total ensemble weight allocated to model m at time t.

Once these weights are established, the combined WA forecast for time t is formed as a convex linear blend of the base models:

$$\widehat{y}_t = \sum_{m=1}^{M} w_{m,t}\, f_{m,t},$$

where $f_{m,t}$ is the prediction generated by model m. Whenever new data becomes available, each model's partial RMSE is updated to include the latest forecast error, and a fresh set of weights is derived. This routine encourages models with deteriorating accuracy to lose weight, while those maintaining a strong alignment with the observed values see their weight increase.

The main advantage of WA lies in its direct, performance-driven rationale: the better a model has performed recently, the more heavily it is relied upon for the ensemble forecast. Should a model's forecasts begin to diverge from actual realizations, its partial RMSE grows, automatically reducing its relative weight. Conversely, a model whose predictions remain close to observed values achieves a smaller partial RMSE, thereby gaining a larger share of the overall combination (Genre et al., 2013).

In this study, we apply WA to the base models retained by the Model Confidence Set (MCS) (P. R. Hansen et al., 2011). The key steps are:

[62] In this context, "increases with model quality" means that lower partial RMSE values correspond to higher $\alpha_{m,t}$. Hence, a model exhibiting a smaller RMSE up to time t is deemed to have performed better historically, thus warranting a proportionally greater weight in the combination.

1. **Performance Tracking**. At each quarterly time t, we calculate the partial RMSE for each model, comparing its forecasts with realized outcomes from the initial period up to t.

2. **Inverse RMSE**. For each model, we invert its partial RMSE so that lower RMSE values translate into higher $\alpha_{m,t}$.

3. **Normalization**. We divide each model's $\alpha_{m,t}$ by the sum of all $\alpha_{j,t}$ values, ensuring the weights sum to one.

4. **Aggregation**. The final WA forecast at time t is the weighted sum of the base models' predictions, with weights determined simultaneously.

We perform these steps in an offline (batch) fashion (Tashman, 2000), retrospectively simulating how the weights evolve each quarter. Each time new macroeconomic data become available, the entire sample is reprocessed, allowing the method to reflect changes in model performance over differing economic regimes.

The principal strength of WA lies in its simplicity and interpretability: the weights have a tangible meaning, namely that models performing better in the recent past receive proportionally higher emphasis. Moreover, the weights adjust automatically when a model's accuracy degrades, discouraging continued reliance on outdated performance.

Nonetheless, WA might fail to capture abrupt regime shifts if partial RMSE takes too long to incorporate recent, more severe errors. In highly volatile environments, a swifter reweighting mechanism may be necessary: supposing that multiple models exhibit nearly indistinguishable performance, their weights can converge to a near-uniform distribution, effectively approximating a simple average and thereby limiting the added value that a more selective weighting scheme would otherwise provide.

The Weighted Average (WA) combination adopted in this study adheres to these guidelines:

- *Initialization:* each model's partial RMSE is set to zero, and uniform weights are assigned for the first nowcast.

- *RMSE Update:* at each quarterly time t, compute the new forecast error for each model and incorporate it into the partial RMSE, thereby accumulating performance information.

- *Weighting and Normalization:* invert the partial RMSE to obtain a measure of each model's predictive accuracy, then normalize across all models so that the weights sum to unity.

- *Final Aggregation:* produce the combined nowcast as a weighted sum of the MCS-selected base models, with weights reflecting their respective performance records.

This procedure allows a quarter-by-quarter reconstruction of how partial RMSE and weights would have evolved in real-time. Whenever new macroeconomic data become available, the entire procedure is repeated so that the WA weighting mechanism remains responsive to recent patterns of model accuracy.

Exponentially Weighted Average

The Exponentially Weighted Average (EWA) constitutes a flexible mechanism for combining multiple forecasting models, each distinguished by its own pattern of predictive errors (Devaine et al., 2013; Littlestone & Warmuth, 1994). At a conceptual level, EWA assigns greater weight to those models that have demonstrated superior accuracy in the most recent past. Formally, at every time step, the method incorporates an exponential decay factor into each model's cumulative loss, progressively attenuating the impact of older forecast errors and allowing the ensemble to adjust promptly when new information arrives.

Macroeconomic series are prone to regime changes and structural breaks that can weaken the predictive capabilities of a single forecast combination with static weights (Clark & McCracken, 2010; Tian & Anderson, 2014). The EWA algorithm addresses this limitation by weighting models according to their latest performance, penalizing recent inaccuracies more heavily than older ones (Timmermann, 2006). It is grounded in the notion that, under evolving economic dynamics, an estimator's most reliable gauge of current accuracy is its error profile in the latest observations (Rossi, 2021).

From a formal perspective, let $m \in \{1, \ldots, M\}$ index the base models selected via the Model Confidence Set (MCS). At time t, each model m accumulates a loss $\mathcal{L}_{m,t}$, which could be measured as a sum of squared forecast errors up to period t. An exponential weighting function then transforms these cumulative losses into a set of normalized coefficients:

$$w_{m,t} \;=\; \frac{\exp\!\big[-\,\eta\,\mathcal{L}_{m,t}\big]}{\sum_{j=1}^{M}\exp\!\big[-\,\eta\,\mathcal{L}_{j,t}\big]},$$

where $\eta > 0$ denotes a *learning rate* (or decay parameter). It determines the speed at which the EWA scheme adapts to new forecast errors: larger η values amplify the penalty for recent mistakes, enabling quick revisions of weights; smaller values yield a more gradual adjustment. The coefficient $w_{m,t}$ represents the fraction of the total weight allocated to model m at time t. In all cases, these weights remain nonnegative ($0 \;\leq\; w_{m,t} \;\leq\; 1$) and sum to unity ($\sum_{m=1}^{M} w_{m,t} = 1$), thus yielding a valid convex combination. Because $w_{m,t}$ depends inversely on $\mathcal{L}_{m,t}$ through the exponential term, models that exhibit smaller cumulative losses (i.e., better historical performance) receive higher weights and those with larger losses see their weights diminished.

To generate a forecast at time t, EWA calculates:

$$\widehat{y}_t \;=\; \sum_{m=1}^{M} w_{m,t}\, f_{m,t},$$

where $f_{m,t}$ is the one-step-ahead forecast from model m. Once the actual y_t is observed, each model's loss (e.g., squared error, $(y_t - f_{m,t})^2$) is updated accordingly, and the new cumulative losses feedback into the weighting scheme. EWA proves particularly valuable in contexts where distinct models exhibit comparative advantages in different macroeconomic regimes. An approach that excels during expansions may temporarily lose favor if it underperforms during downturns, allowing more robust models in recessive phases to gain increased weight in the ensemble (de Rooij et al., 2014; Raftery et al., 2010). Hence, EWA systematically adjusts the relative influence of each model as the economic environment evolves without relying on prespecified thresholds or manual interventions.

In our analysis, we employ EWA to merge the forecasts of those models (also referred to as "experts" (Cesa-Bianchi & Lugosi, 2006) retained by the MCS procedure. At each quarterly step, we:

1. *Compute the per-model loss* based on squared deviations from the observed target;
2. *Accumulate the losses* to form $\mathcal{L}_{m,t}$ for each model m;
3. *Derive exponential weights* by applying $\exp\big[-\eta\,\mathcal{L}_{m,t}\big]$ to each cumulative loss;
4. *Normalize* these weights so that $\sum_{m=1}^{M} w_{m,t} = 1$;
5. *Construct the nowcast* as the weighted sum of the base forecasts.

While EWA can naturally operate in an online context—updating weights at each incoming observation—we implement it offline (batch mode) for consistency with our quarterly data releases (Tashman, 2000). Specifically, in each experiment, we retroactively "simulate" the time evolution of weights by sequentially introducing historical observations and recomputing the ensemble's forecast at each step.

We further enhance EWA by including a "meta-layer[63]" that accounts for multiple values of η. Each distinct learning rate η_j is treated as defining a separate EWA aggregator, wherein the base experts' weights are updated according to η_j. Concretely, for each $\eta_j \in \{\eta_1, \eta_2, \ldots, \eta_K\}$:

- We maintain a running cumulative loss for every base expert (as above) but apply η_j to obtain a specialized weighted forecast at each t.
- This procedure yields K parallel EWA-based predictions, each reflecting a different speed of adaptation.

Subsequently, we interpret these K aggregated forecasts as "meta-experts," each with its own incurred error (i.e., the difference between its aggregated forecast and the realized outcome). A second-tier exponential weighting framework then assigns mass across these η-based aggregators, using an analogous rule:

$$\omega_{j,t} \;=\; \frac{\exp\big[-\,\lambda\,\widetilde{\mathcal{L}}_{j,t}\big]}{\sum_{h=1}^{K}\exp\big[-\,\lambda\,\widetilde{\mathcal{L}}_{h,t}\big]},$$

where $\widetilde{\mathcal{L}}_{j,t}$ denotes the cumulative forecast loss for aggregator j up to time t, and λ is an additional learning rate controlling how aggressively we switch among different η_j-forecasters. By combining these η-aggregators via a second-level EWA, the "Meta-EWA" mechanism mitigates the risk of misspecification that arises from fixing a single decay parameter η. This layered scheme automatically identifies the η_j that best adapts to the prevailing data pattern, assigning a proportionally higher weight to its associated aggregator (Breiman, 1996; Montero-Manso et al., 2020; Wolpert, 1992).

[63]We refer to it as "Meta-EWA" because, after computing separate EWA forecasts for multiple values of η, we treat these parallel EWA aggregators themselves as "meta-experts" in a second-tier exponential scheme. This meta-layer automatically learns which decay parameter performs better at each stage, attenuating the risk of using a single η poorly suited to the entire sample.

Regarding uniform weighting, EWA provides a more dynamic mechanism that promptly favors models with lower recent errors. Consequently, EWA's adaptability may result in superior out-of-sample performance in volatile or rapidly shifting environments (Ballings et al., 2019). Because of these qualities, EWA remains widely adopted whenever the real-time responsiveness of weights is deemed essential. Its recursive structure ensures the ensemble accommodates new information without discarding the historical record entirely. As a result, it aligns well with macroeconomic nowcasting contexts, where the timely integration of evolving economic indicators is of primary concern.

The EWA combination in this study adheres to the following protocol:

- *Initialization*: set all models' losses to zero and assign uniform weights.
- *Cumulative update*: at time t, compute each model's new squared error and add it to the respective loss term.
- *Weighting and normalization*: convert the updated losses into exponential weights, then normalize.
- *Final aggregation*: generate the integrated forecast as a weighted sum of the MCS-selected base predictions.
- *Meta-EWA extension*: in some iterations, allow multiple η-based EWA forecasts to compete and apply an additional exponential weighting to choose among them adaptively.

These steps are repeated over the entire sample in an offline simulation, enabling a quarter-by-quarter reconstruction of how weights and forecasts would have evolved in real-time. Whenever new macroeconomic data become available, we rerun this offline procedure so that the EWA weights remain anchored to recent error patterns. In subsequent sections, we compare EWA outcomes with those from simpler weighting schemes, thereby assessing whether exponential adaptation enhances predictive accuracy under various macroeconomic conditions.

2.6.3 Dynamic Weighting Mechanisms and Explainability in Nowcast Combination Models

Weighted Average (WA) and Exponential Weighted Average (EWA) models offer robust frameworks for combining forecasts, capturing the dynamic contributions of individual models over time. These aggregation techniques are especially beneficial in time-series forecasting, where understanding each model's evolving relative importance enhances the accuracy and interpretability of predictions. By focusing on their shared architecture, we can highlight the conceptual transparency of these methods, which allows for a clear explanation of how each nowcasting model's relative importance changes as economic conditions evolve.

Both WA and EWA follow a similar process for weight assignment, which can be broadly outlined in the following steps (Aiolfi & Timmermann, 2006; Genre et al., 2013; B. E. Hansen, 2008):

1. *Update:* at each time step t, the procedure updates the forecast errors of the individual base models, quantifying how closely each model approximates the observed outcome. This step is the primary input to the subsequent weight adjustment (Diebold & Mariano, 1995).

2. *Normalize:* the performance indicators (such as root mean square errors, cumulative losses, or exponential losses) are then mapped to a common scale, ensuring that the resultant weights fall within the $[0, 1]$ interval and sum to unity. This normalization ensures that the weights are comparable across models, regardless of their scale or magnitude (Radchenko et al., 2023).

3. *Reweight:* the normalized weights are then used to combine the forecasts of all base models. This reweighting reflects how each model's recent performance influences its share in the final combined prediction, forming a dynamically updated forecast (Samuels & Sekkel, 2017).

In this shared framework, Bates and Granger (Bates & Granger, 1969) originally proposed calculating each model's relative importance by inverting its partial RMSE (Weighted Average, WA). A later work introduced exponential decay factors to reduce the influence of older forecast errors (Exponentially Weighted Average, EWA) (Devaine et al., 2013; Littlestone & Warmuth, 1994). Despite these mathematical differences, the interpretability of both methods is rooted in the same principle: if a model performs poorly over a specific period, its weight diminishes, and vice versa.

To explicitly explain how the weights are calculated and evolve at each time step t, we break down the process as follows:

1. *Forecast Errors Calculation (Update):* the forecast errors for each model m are computed based on the difference between the model's prediction $f_{m,t}$ and the observed target y_t. The forecast error is given by:

$$d_{m,t} = |f_{m,t} - y_t|$$

For WA, the error is simply the root mean square error (RMSE) calculated over a rolling window of observations. For EWA, the errors are accumulated into a cumulative loss function, which is then exponentially decayed as follows:

$$\mathcal{L}_{m,t} = \sum_{\tau=1}^{t} \exp\left(-\eta(t-\tau)\right) \cdot d_{m,\tau}$$

2. *Normalization of Performance Indicators (Normalize):* the forecast errors are then normalized to ensure the final weights are on a comparable scale. This step is critical for WA and EWA, ensuring that all weights are positive and sum to 1. For WA, this is done by computing the inverse of the RMSE for each model:

$$w_{m,t}^{WA} = \frac{1/\text{RMSE}_{m,t}}{\sum_{j=1}^{M} 1/\text{RMSE}_{j,t}}$$

For EWA, the cumulative losses are normalized using the exponential loss formula:

$$w_{m,t}^{EWA} = \frac{\exp\left(-\eta\mathcal{L}_{m,t}\right)}{\sum_{j=1}^{M} \exp\left(-\eta\mathcal{L}_{j,t}\right)}$$

3. *Reweighting Forecasts (Reweight):* once the normalized weights are calculated, they are applied to the individual forecasts of each model. This reweighting reflects the relative performance of each model, with better-performing models being assigned a higher weight. The combined forecast for the time step t is then given by:

$$\widehat{y}_t = \sum_{m=1}^{M} w_{m,t} f_{m,t}$$

This methodology ensures that the weights are updated in real-time (or on a rolling basis in a batch setting), reflecting the dynamic contribution of each model's forecast based on its recent accuracy.

Our analysis used "stacked area plots" to visually represent how model weights evolve and WA and EWA's interpretability. These plots allow a clear visual interpretation of the evolution of the weights assigned to each model at every time step as new data becomes available:

- each area in the plot represents the weight evolution of a single model in the total nowcast;
- as models gain or lose influence based on their performance, the size of their respective areas adjusts accordingly.

Using this visualization technique, we can track how each model adapts to changes in the economic environment. For example, a sudden increase in the weight of a specific model may suggest that it is better suited to the current economic regime. In contrast, a decrease in its weight may indicate underperformance. These visual tools enhance the interpretability of both WA and EWA models, allowing users to gain deeper insights into the underlying decision-making process.

Although both WA and EWA share the core logic of *Update* $\rightarrow$ *Normalize* $\rightarrow$ *Reweight*, they differ in how they penalize recent errors. WA relies on an inversion of errors, where the model's weight is inversely proportional to its RMSE (Genre et al., 2013). In contrast, EWA applies an exponential decay factor to cumulative forecast errors, gradually reducing the influence of older data (Devaine et al., 2013). This difference makes EWA more sensitive to abrupt changes in model accuracy, allowing it to respond more quickly to deteriorations in predictive performance. On the other hand, WA can appear more stable, requiring a more extended period of consistent underperformance before a significant weight adjustment occurs (Atiya, 2020).

In contrast to WA and EWA, the Simple Average (SA) aggregator assigns an equal weight to each model and does not modify these weights over time. As a result, SA lacks the dynamic dimension that enables the interpretability of WA and EWA. While SA is transparent in its equal allocation of weights across models, it does not provide insight into which models perform better at any given stage since the weights remain fixed (Clemen, 1989; Stock & Watson, 2004). Therefore, while SA promotes diversity by giving equal importance to all models, it does not allow for the interpretation of how individual model performance affects the aggregated forecast.

In our framework, WA and EWA models provide a comprehensible method for aggregation model interpretability by revealing a time-dependent set of weights that indicate which

models are deemed most reliable at each forecasting period (P. R. Hansen et al., 2011; Samuels & Sekkel, 2017). These methods enable an understanding of how the combined forecast is formed and offer a quantitative measure of the evolving trust placed in each model. These features foster transparency that is particularly valuable in economic and financial contexts, where understanding the sources of predictive power is as important as achieving high levels of forecast accuracy (Makridakis & Hibon, 2000; Stock & Watson, 2011; X. Wang et al., 2023).

2.7 Diagnostic Tests and Predictive Ability Tests

2.7.1 Diagnostic Tests for Combined Models

The robustness of macroeconomic forecasting models, particularly those based on combined approaches such as Simple Average (SA), Weighted Average (WA), and Exponentially Weighted Average (EWA)(Clemen, 1989; Genre et al., 2013; X. Wang et al., 2023), necessitates thorough diagnostic checks. These tests ensure that the models' residuals do not exhibit systematic patterns that would undermine the validity of the forecasts. In this section, we focus on two widely employed diagnostic tests: the Shapiro-Wilk test for normality(Shapiro & Wilk, 1965) and the Ljung-Box test for autocorrelation in the residuals(Ljung & Box, 1978). Both tests were implemented in the context of the model combination strategies discussed earlier, offering a systematic evaluation of their predictive performance.

The Shapiro-Wilk Test for Normality

The Shapiro-Wilk test(Barrow & Kourentzes, 2016; Shapiro & Wilk, 1965) is a statistical method used to assess whether a sample of data comes from a normally distributed population. This test is particularly relevant in time-series forecasting, where the assumption of normally distributed residuals often underpins various diagnostic checks and the construction of confidence intervals. In the context of combined models, the Shapiro-Wilk test examines the normality of residuals derived from the aggregate forecasts of the SA, WA, and EWA methods.

The Shapiro-Wilk test evaluates the null hypothesis that the model's residuals are normally distributed. It is based on the calculation of a test statistic W, which compares the observed distribution of the residuals with a theoretical normal distribution. A value of W close to 1 indicates that the residuals do not significantly deviate from normality, while a value considerably less than 1 suggests non-normality.

In our implementation, the residuals from each combination model—SA, WA, and EWA—are extracted after the predictions are generated. The Shapiro-Wilk test is then applied to these residuals, assessing whether they follow a normal distribution. This check is crucial because many forecasting models, particularly those relying on statistical techniques such as ordinary least squares or maximum likelihood estimation, assume normality in the residuals for valid inference.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where:

- x_i are the ordered sample values of the residuals;

- $\bar{x}$ is the sample mean;
- a_i are constants derived from the expected values of the order statistics of a normal distribution.

In time-series forecasting, especially in our context of nowcasting (from 2017 Q1 to 2023 Q2), the normality of residuals is often a desirable but not always attainable property. The presence of economic shocks (e.g., the COVID-19 pandemic)(Carriero et al., 2022) can lead to distributions with heavy tails or skewness, which deviates from normality. Despite these challenges, the Shapiro-Wilk test remains useful for preliminary diagnostics. Although it assumes independence of residuals (an assumption that is often violated in time-series data, where residuals can exhibit autocorrelation), its application can still highlight significant deviations from normality.

In this study, we justify using the Shapiro-Wilk test as a first diagnostic tool for evaluating residuals, even though it does not account for serial correlation. However, the test's application is more reliable, considering that the data used in this study were iteratively standardized to avoid look-ahead bias and that stationarity issues were fully addressed. Despite the economic shocks during the period, this preparation ensures the test is a valid preliminary check for normality in the residuals. It provides a simple and effective way to detect non-normality, which could suggest underlying issues such as model misspecification or inadequate handling of non-linearity in the data.

The Ljung-Box Test for Autocorrelation

The Ljung-Box test(Ljung & Box, 1978) is a diagnostic tool used to detect autocorrelation in the residuals of a time series model. Autocorrelation in residuals suggests that the model has failed to capture some of the time series structure, which could lead to biased or inconsistent forecasts. This test is crucial when dealing with time series data, where residual autocorrelation may indicate model misspecification or an inadequate model structure.

The Ljung-Box test evaluates the null hypothesis that no autocorrelation exists at any lagged values up to a specified maximum lag q. The test statistic Q is computed as:

$$Q = n(n+2)\sum_{k=1}^{q}\frac{\hat{\rho}_k^2}{n-k}$$

where:

- $\hat{\rho}_k$ is the sample autocorrelation at lag k;
- n is the number of residuals.

Under the null hypothesis of no autocorrelation, Q follows a chi-squared distribution with q degrees of freedom. A significant value of Q (i.e., larger than the critical value from the chi-squared distribution) indicates the presence of autocorrelation in the residuals, suggesting that the model may not have fully accounted for the time-series structure.

We applied the Ljung-Box test to the residuals of the combined forecasts produced by the SA, WA, and EWA models. The presence of autocorrelation in these residuals would indicate that the aggregation process might not be optimal and that further refinements in model selection or combination may be necessary.

For each combination model—SA, WA, and EWA—the residuals are first calculated as the difference between the actual observations and the model forecasts. These residuals are then subjected to the Shapiro-Wilk test for normality and the Ljung-Box test for autocorrelation. The implementation steps are as follows:

1. *Residual Calculation:* for each combination model, compute the residuals by subtracting the model's forecast from the actual observed values;

2. *Normality Test (Shapiro-Wilk):* apply the Shapiro-Wilk test to the residuals to check for normality. A non-significant result suggests that the residuals are normally distributed;

3. *Autocorrelation Test (Ljung-Box):* perform the Ljung-Box test on the residuals to detect any autocorrelation at different lags. A significant result indicates that the residuals exhibit autocorrelation and that model improvements are required.

The results of these diagnostic tests provide valuable insights into the reliability of the combined models' forecasts. In theory, a failure to meet the assumptions of normality or autocorrelation in the residuals would suggest potential issues in the model structure, necessitating further refinement or reconsidering of the combination approach(Tashman, 2000).

Applying the Shapiro-Wilk and Ljung-Box tests to the residuals of the combined models (SA, WA, EWA) ensures that the residuals are approximately normal and uncorrelated—conditions essential for the validity of many econometric techniques used in forecasting. The outcomes of these tests allow for the identification of potential issues within the aggregated models, guiding subsequent refinements, and enhancing the reliability and robustness of the forecasts. This process is crucial for providing accurate and dependable insights and informed macroeconomic decision-making.

2.7.2 Predictive Ability Tests for Model's Comparison

The evaluation of predictive performance is essential in macroeconomic forecasting, where consistently assessing models' relative accuracy and ability to outperform benchmark models is critical. This section discusses the application of the "Giacomini-White test"(Giacomini & White, 2006) to compare the predictive ability of the best individual models selected through the Model Confidence Set, and the combined models in comparison with our three reference models (Random Walk, AR(3), and Dynamic Factor Model). We aim to evaluate whether our individual and combined models (SA, WA, EWA) significantly improve predictive accuracy over these selected benchmarks.

The Giacomini-White test is a statistical method used to compare the predictive performance of a candidate model against a benchmark, testing whether there are significant differences in forecasting accuracy. Building on earlier work on forecast accuracy tests such as the Diebold-Mariano test(Diebold & Mariano, 1995) and the Reality Check(White, 2000), it focuses on conditional predictive ability rather than unconditional comparisons(Inoue & Rossi, 2012).

The null hypothesis of the Giacomini-White test posits that there is no systematic difference in predictive ability between the candidate and reference models. The alternative hypothesis suggests that at least one model outperforms the other statistically. To implement

the test, we compute the difference in forecast errors between the candidate and benchmark models and then regress these differences using ordinary least squares (OLS)(P. R. Hansen, 2005).

A key component of this test is using a robust covariance estimator to account for heteroskedasticity and autocorrelation in the residuals. Specifically, we employ the "Lumley-Heagerty covariance matrix" (also called WEAVE, Weighted Empirical Adaptive Variance Estimation)[64](Lumley & Heagerty, 1999), which adjusts for autocorrelation and heteroskedasticity, ensuring reliable statistical inference.

The Giacomini-White test performed in this study follows these steps:

1. **Loss Differential Calculation:** for each best individual and combined model, we first calculate the loss differential, which is the difference between the squared errors of the model and the reference model:

   $$d_t = \text{Error of model} - \text{Error of benchmark}$$

   where the errors are typically measured as squared forecast errors (e.g., $\epsilon_t^2 = (y_t - \hat{y}_t)^2$).

2. **Regress Loss Differentials:** the loss differential is then regressed using OLS, where the dependent variable is the loss differential d_t, and the independent variable is typically an intercept (although optional instruments can be used for additional refinements).

3. **Application of Lumley-Heagerty Covariance:** the covariance of the OLS residuals is adjusted using the Lumley-Heagerty covariance matrix. This adjustment accounts for potential autocorrelation and heteroskedasticity in the forecast errors, providing a more accurate covariance estimate.

4. **Wald Test:** the Wald test is conducted on the regression coefficients (including the intercept). A rejection of the null hypothesis (that all coefficients are zero) suggests a significant difference in predictive performance between the model and the benchmark.

5. **Significance Evaluation:** the results of the Wald test are used to determine whether the difference in predictive accuracy between the model and the reference model is statistically significant. A low p-value (typically below 0.05) indicates that the model significantly outperforms the reference model.

[64]The Lumley-Heagerty covariance matrix is a robust "sandwich estimator" designed to adjust the standard errors from an OLS regression for both heteroskedasticity and autocorrelation. It achieves that by applying adaptive weights—often derived through isotonic regression to enforce a monotonic decay—to the autocovariance of the residuals. These weights are then used to construct a more reliable covariance matrix estimator, thereby improving inference accuracy when model assumptions regarding constant variance and uncorrelated errors are violated. For the sake of clarity, a "sandwich estimator" is a succinct term for a robust variance estimator that typically appears in the product form bread $\times$ meat $\times$ bread, where "bread" represents the inverse Hessian of the regression model, and "meat" captures the empirical covariance of the residuals, thus accounting for possible misspecifications and error correlations.

The results of the Giacomini-White test for each best individual and aggregated model are summarized by the "Wald statistic" and the associated p-value[65]. These results provide valuable insights into the relative performance of the models:

- *Wald Statistic:* a large Wald statistic indicates a strong rejection of the null hypothesis, suggesting a significant difference in predictive performance between the model and the benchmark.
- *p-Value:* the p-value associated with the Wald test indicates the statistical significance of the result. A p-value below 0.05 suggests that the model provides a significantly better prediction than the reference model.
- *Intercept Estimate:* the intercept estimate in the regression represents the average difference in forecast errors between the model and the reference model. A negative intercept suggests the model consistently provides lower forecast errors than the benchmark, while a positive intercept suggests the opposite.

The outcomes of these tests are used to determine whether the best individual models and the combination models significantly outperform the reference models in terms of average predictive accuracy. A failure to reject the null might also indicate forecast breakdowns or model instability if combined with further analysis(Giacomini & Rossi, 2009). However, negative, significant intercepts in the Wald test results here indicate that the models generally outperform the benchmarks in forecasting accuracy, offering robust evidence of their predictive superiority.

In summary, the Giacomini-White test for model comparison, applied in conjunction with the Lumley-Heagerty covariance estimator, provides a robust framework for evaluating the predictive ability of different forecasting models. By comparing the selected best individual models and the combined models against benchmark models, we can confidently assess whether these models offer statistically significant improvements in forecasting accuracy. Using the Wald test and the Lumley-Heagerty covariance ensures that our results are robust to heteroskedasticity and autocorrelation, which are common in time-series data.

[65]The R-squared values are not reported, as the primary focus of the Giacomini-White test is on the Wald test and the significance of the intercept term. Since only the intercept is used in the regression model to capture the mean difference in forecast errors, the R-squared is expected to be low and is not critical for interpreting the test results.

Chapter 3

Empirical Results and Discussion

[opening remarks]

3.1 Empirical Results

3.1.1 Predictive Accuracy of Individual Models

In this subsection, we present the results related to the nowcast accuracy of individual models, evaluating both the Mean Square Forecast Error (MSFE), Root Mean Square Forecast Error (RMSFE), and Mean Absolute Forecast Error (MAFE) metrics in absolute value (Table 3.1, first part) and the values of RMSFE normalized to the three benchmarks considered (Random Walk, AR(3), Dynamic Factor Model), as illustrated in the second table (Table 3.2) in correspondence with different macroeconomic regimes.

The main objective is to highlight how each approach manages to contain the nowcast error (MSFE, RMSFE, MAFE) in different macroeconomic contexts and quantify the relative benefits (relative gain) compared to the benchmarks. Below, we provide a summary, isolating the most relevant trends in the overall nowcast period (Overall) and the sub-periods (Pre-COVID, COVID, Post-COVID, Excluding COVID).

The mechanisms underlying these differences and the methodological and operational implications will be explored in section 3.2.

Absolute results

The first table (*Forecast accuracy metrics across models and sub-periods*) reports the main error measures (MSFE, RMSFE, and MAFE) for each family of models. Although multiple metrics are provided, we focus on the RMSFE as it heavily penalizes large deviations and proves particularly relevant in highly volatile phases, typically observed during macroeconomic crises or turbulent periods. Below, we summarize each model family's behavior across the entire prediction horizon (Overall) and the four sub-periods (Pre-COVID, COVID, Post-COVID, and Excluding COVID), maintaining a qualitative perspective consistent with Table 3.1.

1. **Penalized linear models.**

   - *Average robustness and performance.* Ridge and Elastic Net exhibit moderate and stable error levels across the Overall prediction period and perform particularly well during low-volatility sub-periods (Excluding COVID). Conversely, LASSO achieves competitive performance exclusively during stable phases but its prediction errors during the COVID period markedly degrade its overall accuracy.
   - *Resistance to shocks.* During the COVID phase, Ridge and Elastic Net demonstrate relative robustness, experiencing only moderate error increases compared

Table 3.1: Forecasting performance metrics (MSFE, RMSFE, and MAFE) of models across sub-periods

Penalized linear models

LASSO

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	4.973	0.395	26.038	2.042	1.144
RMSFE	2.230	0.628	5.103	1.492	1.069
MAFE	1.164	0.538	3.194	1.104	0.795

Ridge

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	2.091	0.654	3.518	3.245	1.831
RMSFE	1.446	0.808	1.876	1.801	1.353
MAFE	0.978	0.635	1.184	1.306	0.940

Elastic net

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	1.669	0.599	3.582	2.189	1.322
RMSFE	1.292	0.774	1.893	1.480	1.150
MAFE	0.936	0.598	1.276	1.206	0.875

Dimensionality reduction-based models

Principal component regression

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	2.146	0.823	1.002	4.192	2.355
RMSFE	1.465	0.907	1.001	2.047	1.534
MAFE	1.164	0.825	0.902	1.675	1.211

Partial least squares regression

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	2.152	0.579	5.274	2.791	1.584
RMSFE	1.467	0.761	2.297	1.671	1.259
MAFE	1.038	0.653	1.433	1.342	0.966

Ensemble learning

Random forest

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	7.048	0.954	33.962	3.596	2.155
RMSFE	2.655	0.977	5.828	1.896	1.468
MAFE	1.750	0.841	4.991	1.545	1.161

eXtreme Gradient Boosting

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	6.655	0.739	35.912	2.051	1.335
RMSFE	2.580	0.860	5.993	1.432	1.156
MAFE	1.673	0.787	5.294	1.287	1.015

Neural networks

Multilayer perceptron

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	7.967	0.777	41.666	3.114	1.840
RMSFE	2.823	0.882	6.455	1.765	1.356
MAFE	1.537	0.691	5.059	1.144	0.897

Gated recurrent unit

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
MSFE	5.410	0.405	29.705	1.698	0.992
RMSFE	2.326	0.636	5.450	1.303	0.996
MAFE	1.305	0.557	4.159	1.061	0.786

to the more stable periods. In sharp contrast, LASSO shows pronounced vulnerability to this turbulent regime, dramatically escalating prediction errors.

2. **Dimensionality reduction-based models.**

   - *General trend.* Both models, based on factorial projections, exhibit error metrics comparable to or slightly higher than penalized models in stable sub-periods (Pre-COVID and Excluding COVID). However, PCR and PLSR display diverging behaviors across different phases, particularly when shocks occur.
   - *Crisis segment.* During the COVID period, PCR demonstrates remarkable robustness, effectively containing prediction errors, while PLSR experiences a substantial deterioration. Conversely, in the Post-COVID phase, PCR's performance worsens significantly, whereas PLSR exhibits a clear recovery, reducing its nowcast error notably.

3. **Ensemble learning models.**

   - *Overall performance.* Both models achieve relatively good accuracy during stable periods (Pre-COVID, Excluding COVID), though performing less favorably than penalized and Dimensionality reduction methods. However, substantial prediction errors during the COVID crisis significantly compromise their overall performance, making them among the worst-performing methods across the entire evaluation period.
   - *COVID and Post-COVID.* In the presence of crises, both models record significant error increases. In particular, Random Forest and XGB show a partial recovery in the time window following the COVID sub-period.

4. **Neural models.**

   - *Normal vs. crisis.* Neural networks perform well in stable periods but deteriorate substantially during the COVID phase. Notably, MLP registers one of the highest errors among all considered models. Conversely, despite a substantial increase in prediction errors during COVID, GRU shows greater robustness and faster error reduction in the subsequent Post-COVID period.
   - *Average Overall Results.* Neural networks' overall performances remain generally superior to Random Forest but remain behind penalized linear models and Dimensionality reduction methods. GRU consistently outperforms MLP, showing notably better overall stability and accuracy.

The absolute results, therefore, highlight that Ridge and Elastic Net consistently obtain the lowest and most stable error levels in most circumstances, particularly in phases characterized by volatility and shocks (COVID). PCR exhibits remarkable robustness, specifically during the COVID crisis, but deteriorates notably in the subsequent phase, while PLSR behaves oppositely, suffering significantly during COVID but recovering thereafter. Conversely, LASSO, ensemble methods (RF and XGB), and neural networks show significant nowcasting deterioration during turbulent periods. Among neural networks, GRU demonstrates relatively greater resilience compared to MLP.

Relative values to benchmarks

The second table (Prediction accuracy relative to benchmarks across sub-periods (RMSFE ratios)), Table 3.2, provides a comparison of the ratios between the RMSFE of each model and that of the three reference benchmarks (Random Walk, AR(3), Dynamic Factor Model). A ratio < 1 indicates an improvement to the benchmark, while a value > 1 corresponds to a relative deterioration.

1. **Compared to Random Walk.**

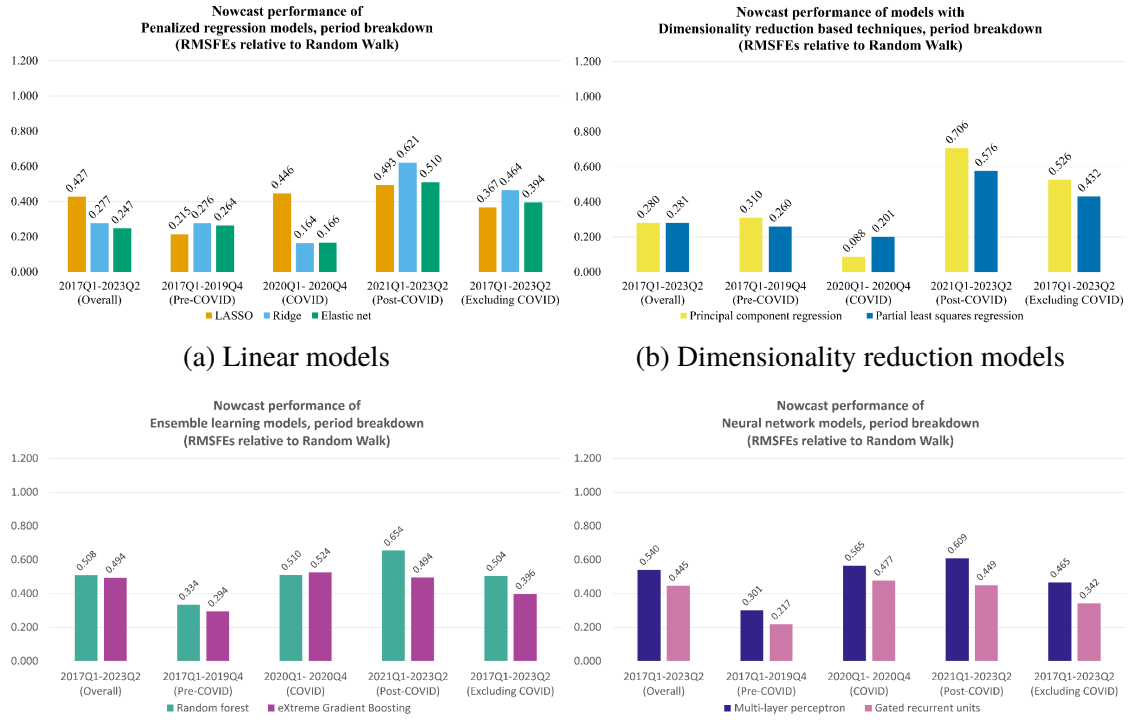
   - *Overall.* Penalized linear (Ridge, Elastic Net) and Dimensionality reduction-based models (PCR, PLSR) show, overall, ratios significantly lower than 1. Other approaches (ensemble learning and neural networks) remain below the unit threshold, although with slightly less pronounced margins.
   - *Pre-COVID.* In this stable window, most models—especially penalized ones and dimensionality reducers—realize a very marked advantage over random walk. Neural networks and ensembles confirm competitive performances, even if the gap from RW is, in some cases, less wide than that of penalized ones.
   - *COVID.* During the crisis, all methods worsened in absolute terms. However, PCR achieves the lowest relative error (ratio below 0.1), followed closely by Ridge and Elastic Net, highlighting exceptional robustness even in this turbulent phase. The other solutions (e.g., LASSO, XGB, GRU) suffer more significant increases in error but generally remain superior to the benchmark.
   - *Post-COVID.* In the periods following the most acute phase, some models (e.g., LASSO, XGB, and GRU) obtain more evident relative improvements, always offering better performance than RW. Ridge and particularly PCR remain below the unit threshold but demonstrate significant deterioration in their relative ratios, especially PCR.
   - *Excluding COVID.* If we eliminate the pandemic window, all models maintain a clear advantage over RW. The penalized ones remain particularly stable, while ensembles and neural networks can show slightly larger oscillations but remain on ratios < 1.

Table 3.2: Forecast performance (RMSFE) of models relative to benchmarks (Random Walk, AR(3), Dynamic Factor Model) to benchmarks across sub-periods (RMSFE ratios)

RMSFE, relative to Random walk	**Overall**	**Pre-COVID**	**COVID**	**Post-COVID**	**Excluding COVID**
Penalized linear models					
LASSO	0.427	0.215	0.446	0.493	0.367
Ridge	0.277	0.276	0.164	0.621	0.464
Elastic net	0.247	0.264	0.166	0.510	0.394
Dimensionality reduction-based models					
Principal component regression	0.280	0.310	0.088	0.706	0.526
Partial least squares regression	0.281	0.260	0.201	0.576	0.432
Ensemble learning models					
Random forest	0.508	0.334	0.510	0.654	0.504
eXtreme Gradient Boosting	0.494	0.294	0.524	0.494	0.396
Neural networks models					
Multi-layer perceptron	0.540	0.301	0.565	0.609	0.465
Gated recurrent units	0.445	0.217	0.477	0.449	0.342

RMSFE, relative to AR(3)	**Overall**	**Pre-COVID**	**COVID**	**Post-COVID**	**Excluding COVID**
Penalized linear models					
LASSO	0.636	0.334	0.660	0.728	0.557
Ridge	0.412	0.430	0.243	0.918	0.705
Elastic net	0.368	0.411	0.245	0.754	0.599
Dimensionality reduction-based models					
Principal component regression	0.418	0.482	0.129	1.044	0.800
Partial least squares regression	0.418	0.404	0.297	0.851	0.656
Ensemble learning models					
Random forest	0.757	0.519	0.754	0.967	0.765
eXtreme Gradient Boosting	0.735	0.457	0.775	0.730	0.602
Neural networks models					
Multi-layer perceptron	0.804	0.469	0.835	0.899	0.707
Gated recurrent units	0.663	0.338	0.705	0.664	0.519

RMSFE, relative to Dynamic factor model	**Overall**	**Pre-COVID**	**COVID**	**Post-COVID**	**Excluding COVID**
Penalized linear models					
LASSO	0.555	0.507	0.561	0.535	0.530
Ridge	0.360	0.653	0.206	0.674	0.670
Elastic net	0.321	0.625	0.208	0.554	0.569
Dimensionality reduction-based models					
Principal component regression	0.364	0.733	0.110	0.767	0.760
Partial least squares regression	0.365	0.614	0.253	0.626	0.623
Ensemble learning models					
Random forest	0.660	0.789	0.641	0.710	0.727
eXtreme Gradient Boosting	0.642	0.694	0.659	0.536	0.572
Neural networks models					
Multi-layer perceptron	0.702	0.712	0.710	0.661	0.672
Gated recurrent units	0.579	0.514	0.600	0.488	0.493



(a) Linear models

(b) Dimensionality reduction models

(c) Ensemble learning models

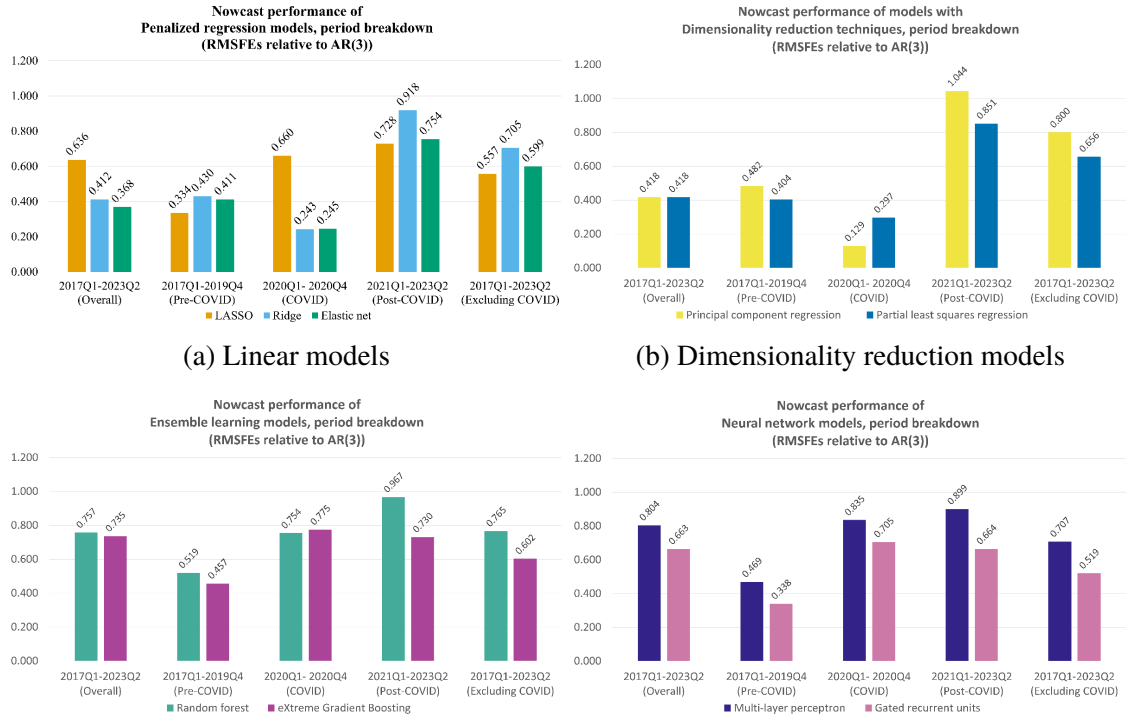
(d) Neural models

Figure 3.1: Nowcast accuracy (RMSFE ratios relative to Random walk) for Penalized linear, Dimensionality reduction, Ensemble learning, Neural models across different macroeconomic regimes.

2. **Compared to AR(3).**

   - *Overall.* The penalized ones (especially Ridge and Elastic Net) and the Dimensionality reduction-based models show very significant error reductions compared to AR(3). The ensemble learning and neural solutions remain below 1, although with more moderate gaps.
   - *Pre-COVID.* Before the crisis, penalized families and factorial models reached ratios below 1, indicating a significant gain in RMSFE compared to AR(3). Neural networks and ensembles also remain below 1, although with minor advantages.
   - *COVID.* In this phase of strong uncertainty, the performances of Ridge and Elastic Net continue to lead, with very low ratios. Neural networks and ensembles show greater oscillations, although GRU recovers faster than MLP. PCR improves significantly during this crisis phase, while PLSR's ratio rises moderately.
   - *Post-COVID.* At the end of the crisis, Ridge and Elastic Net underwent significant deteriorations compared to the previous phase. GRU and XGB recover compared to the previous time window while remaining advantageous compared to AR(3). PLSR and PCR significantly worsened compared to the previous subperiod, with PCR notably exceeding the unit threshold, thus underperforming compared to AR(3).

- *Excluding COVID.* Excluding the pandemic period, most models are solidly below 1, with penalized models and dimensional reducers almost always in the lead. Neural networks and ensemble learning show gaps that are, on average, smaller but still positive compared to AR(3).



(a) Linear models

(b) Dimensionality reduction models

(c) Ensemble learning models

(d) Neural models

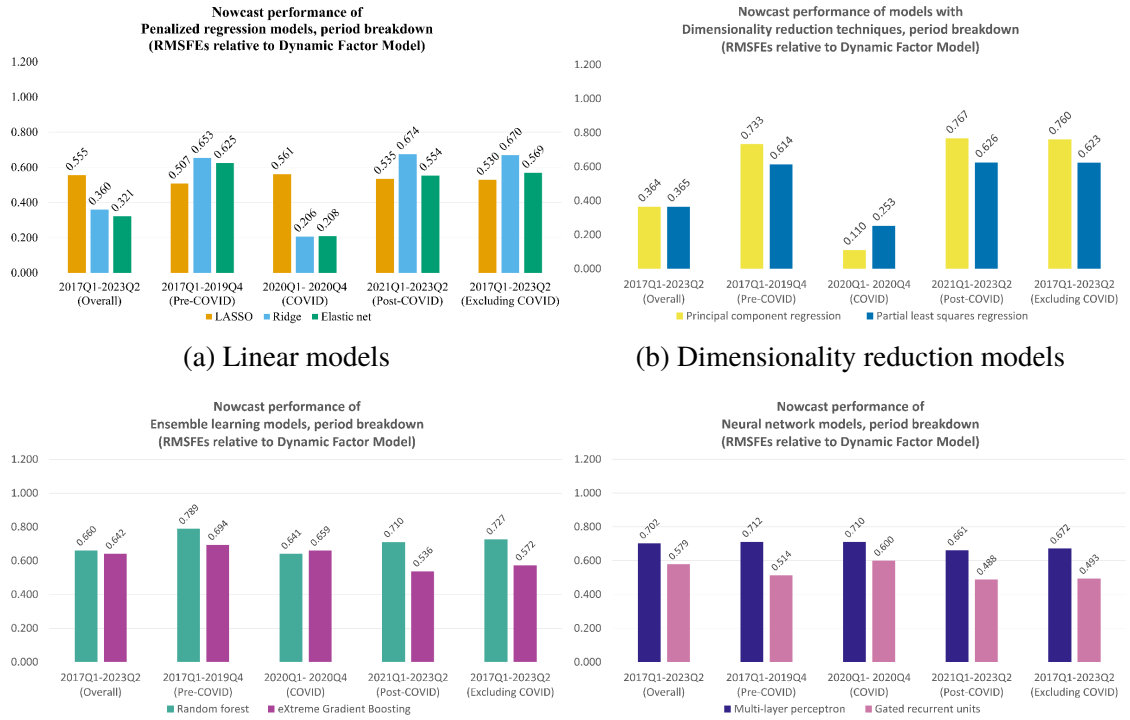
Figure 3.2: Nowcast accuracy (RMSFE ratios relative to AR(3)) for Penalized linear, Dimensionality reduction, Ensemble learning, Neural models across different macroeconomic regimes.

3. **Compared to Dynamic Factor Model.**

- *Overall.* The majority of techniques are competitive or significantly better than DFM: penalized and dimensionally reduced models consistently perform below 1, while ensemble learning and neural models also achieve significant advantages, albeit with generally larger variations across sub-periods.
- *Pre-COVID.* In low-volatility periods, nearly all models achieve substantial improvements compared to DFM, although PCR and Random Forest display less pronounced gains, remaining favorable.
- *COVID.* During the crisis, penalized models continue to show superior relative robustness. Some methods, such as MLP or Random Forest, experience more significant increases in the ratio but generally remain below 1, then perform better than the benchmark. PCR achieves particularly outstanding relative error reduction during COVID.
- *Post-COVID.* As uncertainty is gradually overcome, models such as GRU, MLP, and XGB recover quickly, always maintaining an advantage over DFM.

However, Ridge, PCR, and RF undergo considerable relative deteriorations compared to the COVID period, although still retaining performance below the unit threshold.

- *Excluding COVID.* By eliminating the most turbulent phase (COVID), penalized linear and factor models consistently are well below 1. However, this does not always improve performance since during COVID, the PCR records a greater reduction in the relative error metric, while ensemble learning and neural show less uniform but still positive improvements compared to the DFM.



(a) Linear models

(b) Dimensionality reduction models

(c) Ensemble learning models

(d) Neural models

Figure 3.3: Nowcast performance (RMSFE ratios relative to Dynamic Factor Model) for Penalized linear, Dimensionality reduction, Ensemble learning, Neural models across different macroeconomic regimes.

Penalized linear models (especially Ridge and Elastic Net) and Dimensionality reduction-based models generally achieve RMSFE ratios consistently below benchmarks across most periods. However, PCR shows significant variability: it displays exceptional robustness during COVID but deteriorates substantially in the subsequent Post-COVID phase. In contrast, ensemble learning and neural network models, though frequently outperforming benchmarks, demonstrate higher overall vulnerability to shocks, particularly pronounced during the COVID period. Among neural approaches, GRU systematically outperforms MLP, highlighting better stability and resilience to macroeconomic disturbances.

3.1.2 Prediction Uncertainty

The reliability of nowcasting methods is inherently linked to their capacity to accurately measure and represent prediction uncertainty, especially in contexts subject to structural changes or unexpected economic shocks. Prediction intervals offer valuable insights, quantifying how model predictions vary under different macroeconomic conditions. Table 3.3 presents the average widths of these intervals, computed across the whole prediction period as well as distinct sub-periods (Pre-COVID, COVID, Post-COVID, Excluding COVID), enabling a nuanced assessment of each model family's consistency and resilience.

Table 3.3: Average prediction intervals across models and sub-periods

	Overall	**Pre-COVID**	**COVID**	**Post-COVID**	**Excl. COVID**
Penalized linear models					
LASSO	2.117	1.278	5.460	1.787	1.510
Ridge	2.130	1.932	3.513	1.813	1.878
Elastic net	1.499	0.924	3.540	1.372	1.128
Dimensionality reduction-based models					
PCR	2.157	1.443	4.819	1.950	1.674
PLSR	2.125	1.124	4.918	2.209	1.617
Ensemble learning models					
RF	3.059	1.190	2.230	5.632	3.209
XGB	1.960	1.613	2.141	2.304	1.927
Neural network models					
MLP	2.785	1.584	6.296	2.822	2.147
GRU	2.526	1.991	3.261	2.873	2.392

Below, we outline the key patterns characterizing prediction uncertainty, highlighting differences and commonalities within and across model families over the analyzed sub-periods:

1. **Penalized linear models.**

    - *Overall trends.* Elastic Net consistently provides narrower prediction intervals than LASSO and Ridge across the entire nowcasting horizon. In particular, Elastic Net demonstrates reduced uncertainty during stable economic phases, while Ridge intervals remain relatively wider throughout.
    - *Period-specific variations.* During the COVID period, prediction intervals significantly expanded for all penalized models, reflecting heightened prediction uncertainty. However, Elastic Net intervals widened comparatively less, indicating better stability under economic volatility. In Post-COVID sub-period, prediction intervals notably contracted, with Elastic Net and LASSO returning closer to pre-crisis levels.

2. **Dimensionality reduction-based models.**

    - *General behavior.* PCR and PLSR show similar average interval widths In the overall prediction period, although exhibiting distinct intra-period dynamics. Specifically, intervals from PCR remained relatively broader than PLSR in most periods, except during the Post-COVID phase, when PLSR intervals notably exceeded PCR.

- *Economic shocks impact.* Both models experienced substantial interval widening during COVID, with PCR showing slightly narrower intervals, suggesting marginally better robustness under crisis conditions. Nevertheless, PLSR intervals significantly narrowed after the crisis, highlighting its quicker return to pre-crisis uncertainty levels.

3. **Ensemble learning models.**

   - *Aggregate results.* Random Forest presented the widest prediction intervals overall, signaling substantial uncertainty across periods. Conversely, XGBoost maintained narrower intervals, demonstrating enhanced stability and precision, especially in stable economic sub-periods.
   - *Sub-period dynamics.* Random Forest intervals notably expanded during the Post-COVID phase, surpassing considerably the interval widths observed during both Pre-COVID and COVID phases, where intervals were comparatively narrower. XGBoost exhibited less dramatic shifts, maintaining moderate interval widths even during turbulent economic phases.

4. **Neural models.**

   - *Interval widths and volatility.* MLP and GRU displayed elevated uncertainty compared to Penalized and Dimensionality reduction models, especially during COVID. MLP recorded the widest intervals of all models during the crisis, indicating pronounced sensitivity to extreme economic disturbances.
   - *Post-crisis behavior.* Post-COVID intervals for GRU slightly narrowed compared to the COVID period, while those of MLP narrowed markedly, indicating a faster normalization after the crisis. Considering the whole prediction period but excluding COVID, GRU consistently maintained broader intervals relative to other methods, highlighting ongoing caution in nowcasts even under normal conditions.

A detailed examination of the actual and predicted GDP trajectories across model families, visualized through prediction interval plots (see Figures 3.4 to 3.7), provides additional insights regarding prediction interval coverage and reliability:

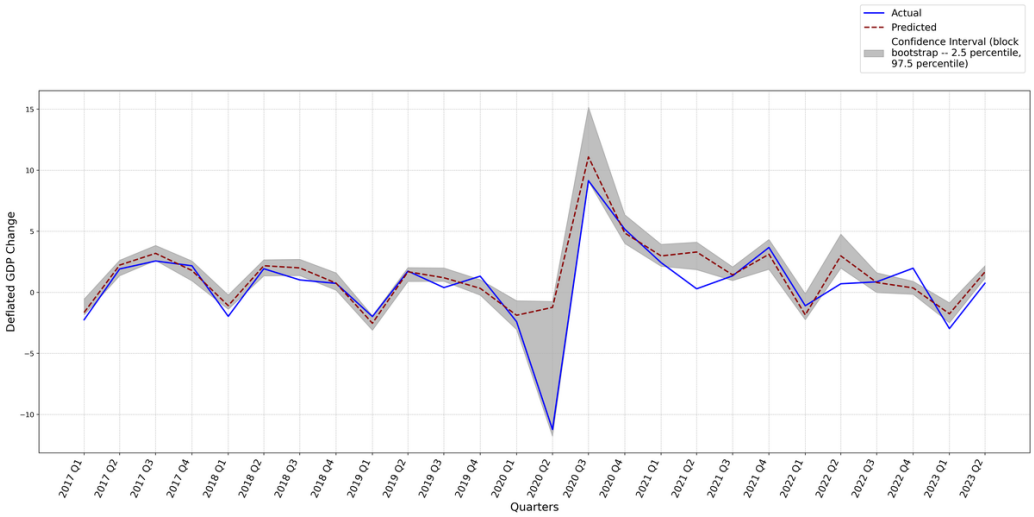
- Penalized linear models (Figure 3.4) present narrow prediction intervals closely following the predicted GDP path. Despite their precision, during periods of extreme volatility, actual GDP values frequently fall outside the prediction intervals, signaling potential limitations in capturing exceptional economic fluctuations.

- Dimensionality reduction-based models (Figure 3.5) generate prediction intervals that are slightly broader than penalized linear models, enhancing the coverage of actual GDP values during stable periods. Nonetheless, these intervals still demonstrate challenges in fully encompassing observed GDP dynamics, with actual values occasionally breaching the interval boundaries.

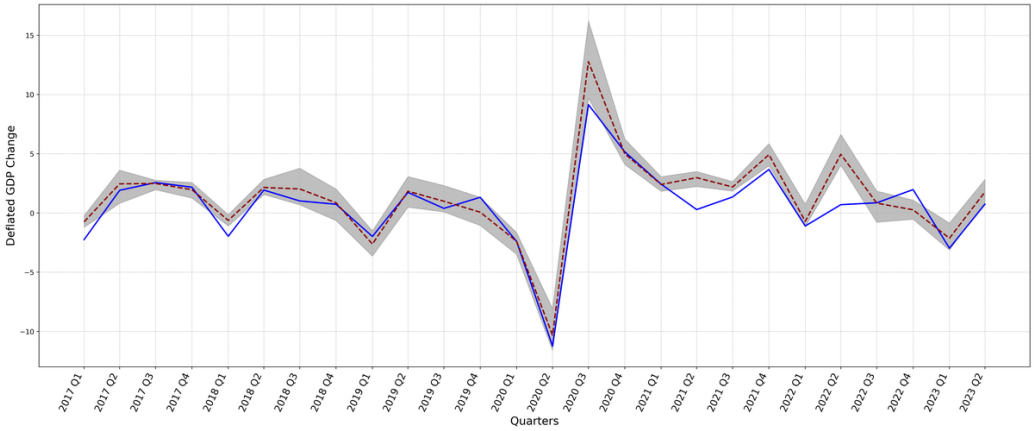
- Random Forest produces notably wider prediction intervals, especially in the Post-COVID period. Despite this wider uncertainty quantification, intervals sometimes

fail to encompass actual GDP data and the model's predictions, suggesting elevated uncertainty together with weaker predictive consistency. Conversely, XGBoost demonstrates moderately wide intervals that consistently envelop the predicted values and, to a greater extent, actual GDP figures across different economic phases (Figure 3.6).

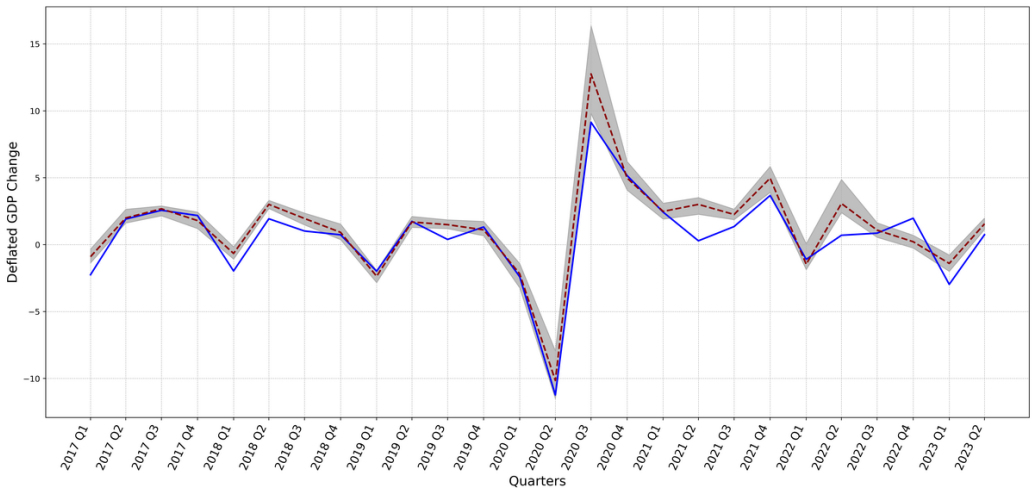
- GRU intervals exhibit moderate widths, capturing most of the observed GDP variations, even in volatile contexts. MLP, however, displays the broadest intervals during crisis periods, often encompassing extreme GDP values. This broader uncertainty quantification persists even in less turbulent sub-periods (Figure 3.7) .



(a) LASSO

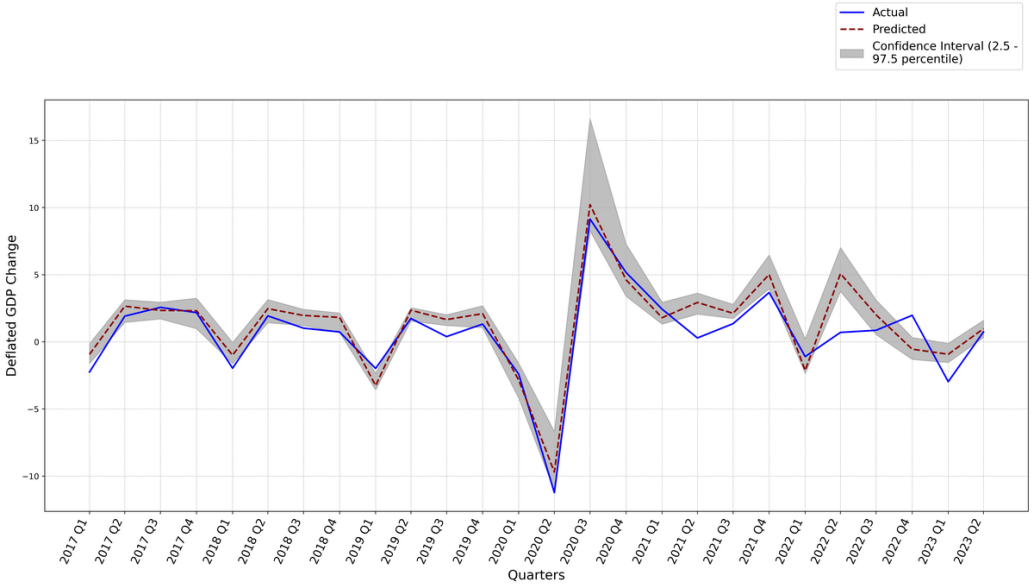


(b) Ridge

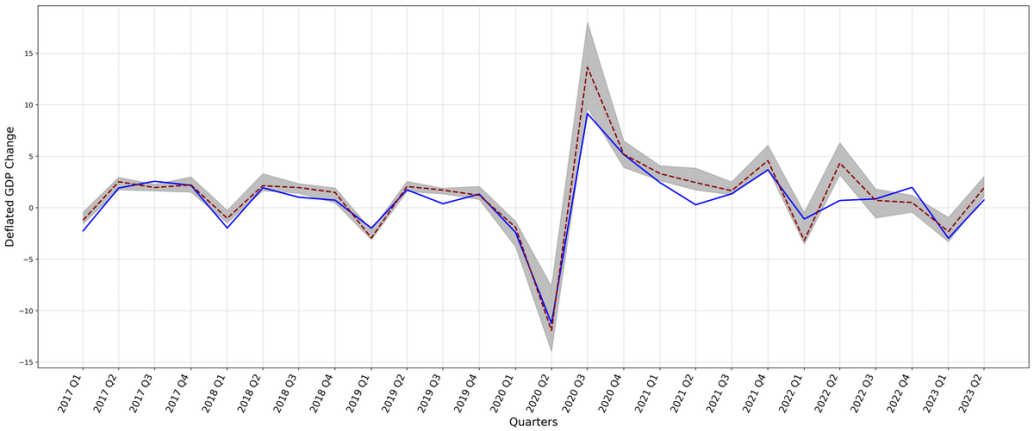


(c) Elastic Net

Figure 3.4: Actual and predicted deflated GDP growth rate for Penalized Linear models (LASSO, Ridge, and Elastic Net). Shaded areas represent the 95% prediction intervals obtained through the block bootstrap method.

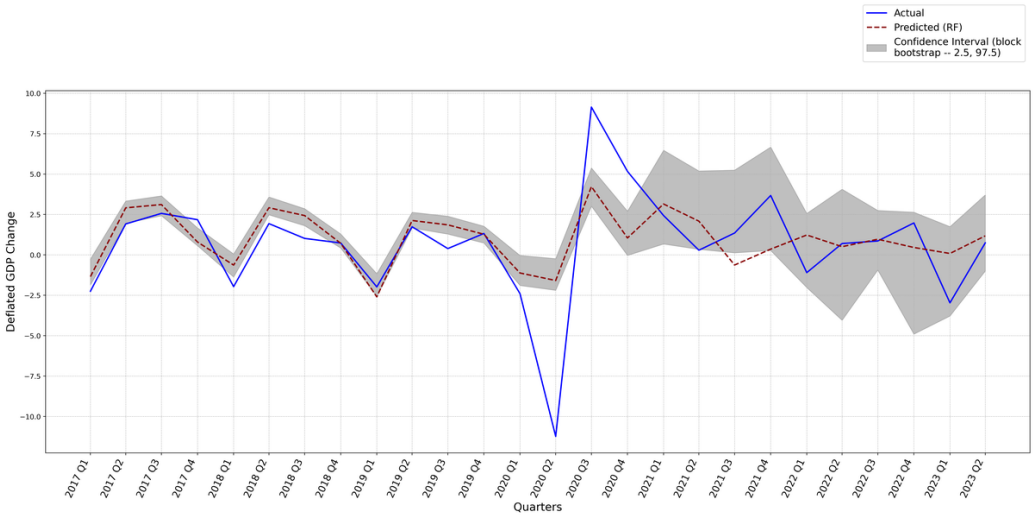


(a) Principal Component Regression

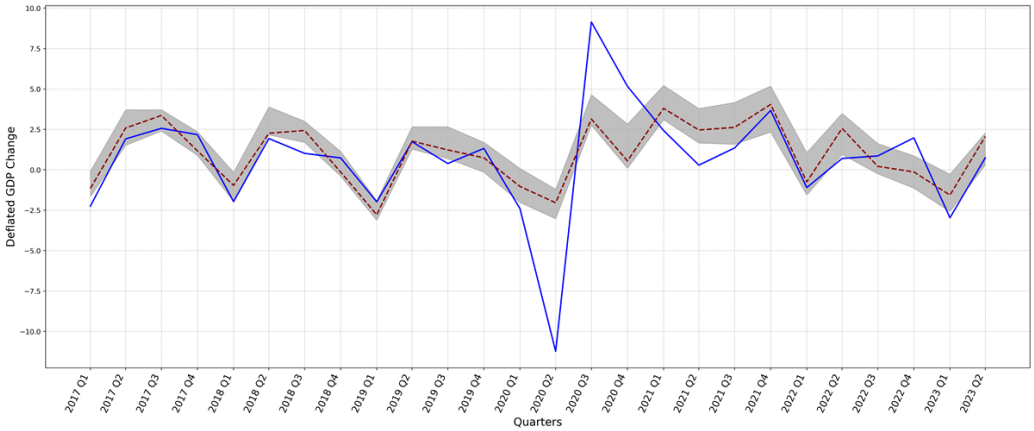


(b) Partial Least Squares Regression

Figure 3.5: Actual and predicted deflated GDP growth rate for Dimensionality Reduction-based models (Principal Component Regression and Partial Least Squares Regression). Shaded areas represent the 95% prediction intervals obtained through the block bootstrap method.

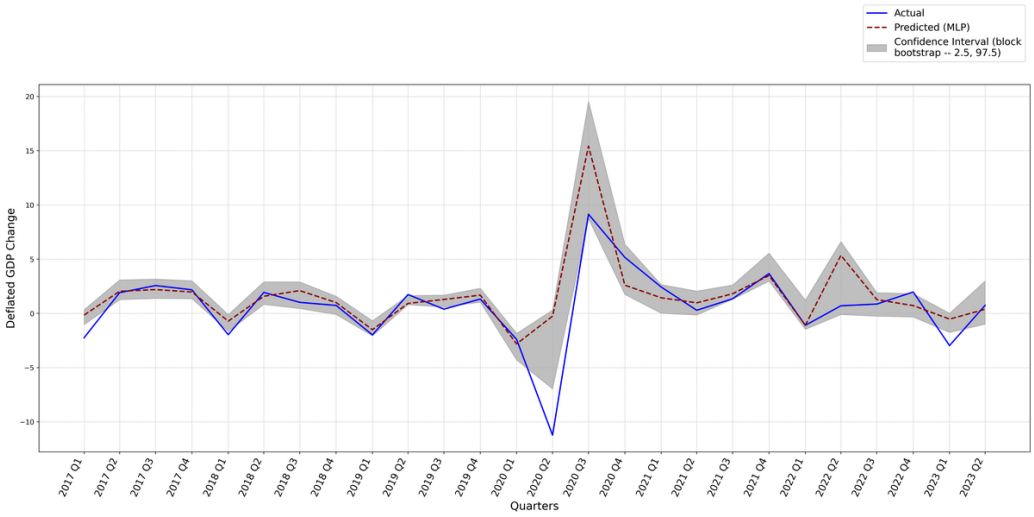


(a) Random Forest

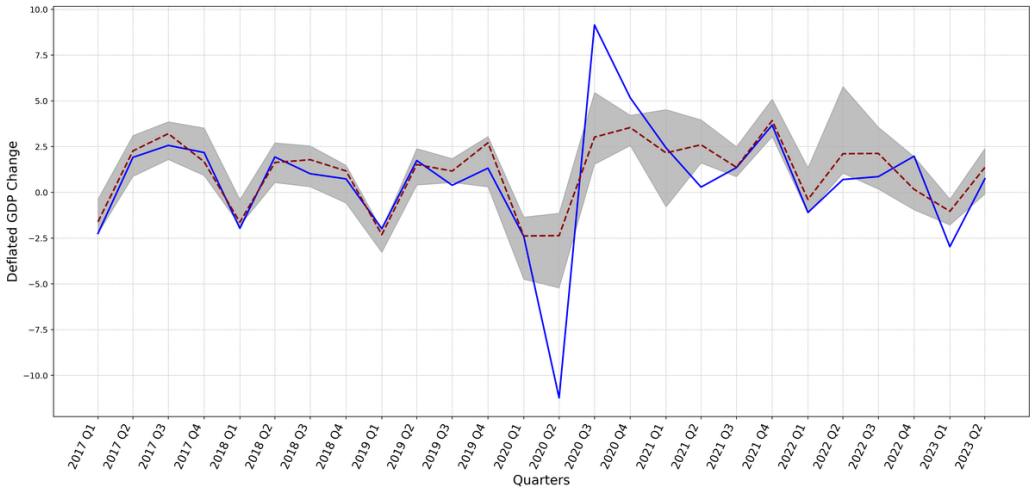


(b) eXtreme Gradient Boosting

Figure 3.6: Actual and predicted deflated GDP growth rate for Ensemble Learning models (Random Forest and eXtreme Gradient Boosting). Shaded areas represent the 95% prediction intervals obtained through the block bootstrap method.



(a) Multilayer Perceptron



(b) Gated Recurrent Unit

Figure 3.7: Actual and predicted deflated GDP growth rate for Neural models (Multilayer Perceptron and Gated Recurrent Unit). Shaded areas represent the 95% prediction intervals obtained through the block bootstrap method.

The analysis of prediction intervals reveals distinct uncertainty profiles across model families. Penalized linear models—particularly Elastic Net— consistently achieve narrower and more stable prediction intervals across most periods, indicating greater reliability in prediction uncertainty quantification, especially during phases of heightened volatility (e.g., COVID). Among Dimensionality reduction-based models, PCR exhibits considerable stability during the COVID crisis but performs less consistently in the subsequent Post-COVID phase. In contrast, PLSR intervals widen notably during the crisis but return to lower levels thereafter. Conversely, ensemble learning methods (particularly Random Forest) and neural network approaches (especially MLP) tend to produce substantially wider prediction intervals, signaling pronounced uncertainty and heightened sensitivity to macroeconomic shocks. Within the neural network family, GRU generally outperforms MLP, maintaining comparatively narrower intervals and exhibiting greater robustness

across the evaluated sub-periods.

3.1.3 Feature Importance and Explainability

Examining feature importance offers a complementary perspective to predictive accuracy and uncertainty, enhancing interpretability and transparency and providing deeper insights into the economic mechanisms underlying model predictions. To this end, we investigate how explanatory variables contribute to nowcasting outcomes across different model families, highlighting general trends and specific variations observed across macroeconomic regimes (Overall, Pre-COVID, COVID, Post-COVID, and Excluding COVID). These results, complementing previously presented analyses on predictive accuracy and uncertainty, offer additional insights on the relative relevance assigned to economic indicators by each modeling framework within the macroeconomic nowcasting context.

1. **Penalized linear models.**

    - *General trends.* Persistently ranked among crucial features are industrial production and gross operating surplus. Additionally, labor market indicators such as employee compensation and unit labor costs frequently appear among key predictors, particularly during turbulent economic periods.
    - *Period-specific variations.* During the COVID sub-period, penalized models significantly increased the importance of labor and production-related variables, effectively capturing volatility. In the Post-COVID interval, the dominance of industrial production notably intensified, clearly reflecting economic stabilization.
    - *Excluding COVID.* Excluding the COVID period, the hierarchy of relevant variables closely mirrors the overall pattern, confirming stability in predictor selection and importance, particularly regarding industrial production, gross operating surplus, and labor market variables.

2. **Dimensionality reduction-based models.**

    - *General trends.* PCR prominently features industrial production and fiscal indicators (taxes less subsidies), consistently maintaining these predictors' importance and relative ranking across all sub-periods. PLSR, on the other hand, notably emphasizes air transport variables such as air passenger arrivals and aircraft movements.
    - *Period-specific variations.* While PCR maintained relatively stable feature importance throughout all periods, PLSR highlighted air transport variables more intensely during and following the COVID period, capturing critical economic disruptions and subsequent adjustments.
    - *Excluding COVID.* In the Excluding COVID period, both PCR and PLSR reconfirmed the significance of industrial production and air transport variables. PCR maintained stability in attributing importance to fiscal indicators, while PLSR continued emphasizing air transport and international trade indicators.

Table 3.4: Linear models with penalization, top 10 predictors ranked by average feature importance and 95% confidence intervals, estimated via block bootstrap, across different sub-periods

	LASSO			Ridge			Elastic Net		
	Feature	**Feature (avg)**	**CI range**	**Feature**	**Feature (avg)**	**CI range**	**Feature**	**Feature (avg)**	**CI range**
Overall	employee compensation change	1.684	3.157	unit labor cost change	-1.355	4.234	industrial production index change	0.907	0.567
	unit labor cost change	-1.657	3.139	employee compensation change	1.276	4.161	gross operating surplus change	0.507	0.364
	industrial production index change	0.947	0.451	industrial production index change	0.732	0.549	air cargo loaded change	0.322	0.181
	gross operating surplus change	0.669	0.327	gross operating surplus change	0.435	0.361	fnb services index change	0.240	0.216
	fnb services index change	0.277	0.215	fnb services index change	0.237	0.276	taxes less subsidies change	0.222	0.160
	air cargo loaded change	0.272	0.253	air cargo loaded change	0.232	0.254	pawnshop pledges received change	0.200	0.144
	public progress payments change	0.262	0.098	taxes less subsidies change	0.229	0.083	public progress payments change	0.179	0.169
	taxes less subsidies change	0.227	0.055	pawnshop pledges received change	0.210	0.275	private properties available change	-0.160	0.130
	pawnshop pledges received change	0.226	0.118	composite leading index change	0.192	0.164	composite leading index change	0.158	0.102
	exchange rate hkd change	-0.173	0.163	exchange rate hkd change	-0.191	0.439	business expectations manufacturing	0.153	0.108
Pre-COVID	employee compensation change	3.100	0.153	unit labor cost change	-2.568	4.178	industrial production index change	0.945	0.286
	unit labor cost change	-3.075	0.153	employee compensation change	2.469	4.102	gross operating surplus change	0.538	0.193
	industrial production index change	0.763	0.076	industrial production index change	0.689	0.210	air cargo loaded change	0.328	0.112
	gross operating surplus change	0.571	0.061	gross operating surplus change	0.444	0.076	fnb services index change	0.240	0.136
	fnb services index change	0.240	0.072	exchange rate hkd change	-0.266	0.474	taxes less subsidies change	0.231	0.213
	public progress payments change	0.230	0.023	fnb services index change	0.239	0.145	pawnshop pledges received change	0.200	0.155
	taxes less subsidies change	0.216	0.015	composite leading index change	0.236	0.094	private properties available change	-0.186	0.127
	pawnshop pledges received change	0.204	0.011	taxes less subsidies change	0.235	0.059	public progress payments change	0.178	0.155
	air cargo loaded change	0.193	0.052	public progress payments change	0.182	0.065	composite leading index change	0.169	0.097
	composite leading index change	0.192	0.030	pawnshop pledges received change	0.175	0.056	business expectations manufacturing	0.153	0.144
COVID	employee compensation change	1.559	3.000	industrial production index change	0.755	0.208	industrial production index change	0.799	0.243
	unit labor cost change	-1.549	3.102	gross operating surplus change	0.398	0.106	gross operating surplus change	0.426	0.135
	industrial production index change	0.952	0.442	air cargo loaded change	0.293	0.027	air cargo loaded change	0.304	0.037
	gross operating surplus change	0.689	0.323	taxes less subsidies change	0.251	0.043	taxes less subsidies change	0.243	0.050
	air cargo loaded change	0.313	0.207	pawnshop pledges received change	0.191	0.054	pawnshop pledges received change	0.189	0.052
	public progress payments change	0.275	0.078	unit labor cost change	-0.185	0.067	private properties available change	-0.179	0.055
	taxes less subsidies change	0.237	0.046	private properties available change	-0.182	0.059	composite leading index change	0.175	0.018
	pawnshop pledges received change	0.206	0.048	composite leading index change	0.181	0.023	fnb services index change	0.174	0.052
	fnb services index change	0.205	0.064	business expectations manufacturing	0.174	0.054	unit labor cost change	-0.164	0.055
	domestic supply price change	-0.204	0.074	fnb services index change	0.171	0.052	business expectations manufacturing	0.162	0.046
Post-COVID	industrial production index change	1.167	0.027	industrial production index change	0.775	0.550	industrial production index change	0.904	0.562
	gross operating surplus change	0.780	0.108	gross operating surplus change	0.439	0.362	gross operating surplus change	0.503	0.361
	fnb services index change	0.351	0.208	unit labor cost change	-0.366	0.634	air cargo loaded change	0.320	0.167
	air cargo loaded change	0.351	0.137	employee compensation change	0.308	0.658	fnb services index change	0.266	0.216
	public progress payments change	0.295	0.046	air cargo loaded change	0.277	0.084	pawnshop pledges received change	0.204	0.111
	pawnshop pledges received change	0.261	0.112	fnb services index change	0.260	0.279	taxes less subsidies change	0.203	0.042
	taxes less subsidies change	0.236	0.056	pawnshop pledges received change	0.259	0.290	public progress payments change	0.188	0.156
	exchange rate hkd change	-0.222	0.084	taxes less subsidies change	0.211	0.050	business expectations manufacturing	0.148	0.048
	business expectations manufacturing	0.195	0.038	public progress payments change	0.200	0.240	composite leading index change	0.139	0.063
	private properties available change	-0.160	0.076	merchandise exports change	0.160	0.096	gross fixed capital formation current change	0.128	0.053
Excluding COVID	employee compensation change	1.707	3.161	unit labor cost change	-1.567	4.234	industrial production index change	0.926	0.576
	unit labor cost change	-1.677	3.141	employee compensation change	1.487	4.162	gross operating surplus change	0.522	0.365
	industrial production index change	0.947	0.451	industrial production index change	0.728	0.550	air cargo loaded change	0.325	0.184
	gross operating surplus change	0.666	0.312	gross operating surplus change	0.441	0.361	fnb services index change	0.252	0.216
	fnb services index change	0.290	0.200	fnb services index change	0.249	0.277	taxes less subsidies change	0.218	0.178
	air cargo loaded change	0.265	0.254	taxes less subsidies change	0.224	0.081	pawnshop pledges received change	0.202	0.156
	public progress payments change	0.260	0.097	air cargo loaded change	0.221	0.254	public progress payments change	0.182	0.176
	pawnshop pledges received change	0.230	0.110	pawnshop pledges received change	0.213	0.279	private properties available change	-0.156	0.135
	taxes less subsidies change	0.225	0.055	exchange rate hkd change	-0.209	0.452	composite leading index change	0.155	0.105
	exchange rate hkd change	-0.169	0.163	composite leading index change	0.194	0.164	business expectations manufacturing	0.151	0.119

Table 3.5: Dimensionality reduction based models with penalization, top 10 predictors ranked by average feature importance and 95% confidence intervals, estimated via block bootstrap, across different sub-periods

	Principal Component Regression			Partial Least Squares Regression		
	Feature	**Feature (avg)**	**CI range**	**Feature**	**Feature (avg)**	**CI range**
Overall	industrial production index change	0.595	0.340	industrial production index change	2.509	1.162
	taxes less subsidies change	0.326	0.365	air passenger arrivals change	2.081	1.778
	gross operating surplus change	0.315	0.109	air cargo loaded change	2.060	0.481
	merchandise reexports change	0.312	0.086	aircraft arrivals departures change	1.968	1.859
	air cargo loaded change	0.309	0.118	merchandise imports change	1.918	0.717
	merchandise imports change	0.268	0.096	gross operating surplus change	1.829	0.926
	merchandise exports change	0.244	0.077	merchandise reexports change	1.796	0.633
	composite leading index change	0.201	0.072	merchandise exports change	1.728	0.704
	liquor releases change	0.175	0.110	fnb services index change	1.583	0.954
	cpf due members change	-0.166	0.093	pawnshop pledges received change	1.327	0.632
Pre-COVID	industrial production index change	0.657	0.313	industrial production index change	2.821	0.032
	gross operating surplus change	0.349	0.042	air cargo loaded change	2.211	0.102
	taxes less subsidies change	0.330	0.387	merchandise imports change	2.202	0.096
	merchandise reexports change	0.321	0.074	gross operating surplus change	2.152	0.090
	air cargo loaded change	0.303	0.047	merchandise reexports change	2.060	0.091
	merchandise imports change	0.266	0.100	merchandise exports change	2.018	0.121
	merchandise exports change	0.241	0.078	air passenger arrivals change	1.455	0.028
	composite leading index change	0.199	0.067	domestic supply price change	1.412	0.104
	air passenger arrivals change	0.197	0.042	import price index change	1.395	0.135
	liquor releases change	0.187	0.104	aircraft arrivals departures change	1.320	0.057
COVID	industrial production index change	0.608	0.143	industrial production index change	2.545	0.609
	taxes less subsidies change	0.353	0.152	aircraft arrivals departures change	2.131	1.648
	gross operating surplus change	0.304	0.034	air cargo loaded change	2.042	0.385
	air cargo loaded change	0.291	0.030	merchandise imports change	1.903	0.542
	merchandise reexports change	0.288	0.038	air passenger arrivals change	1.872	0.984
	merchandise imports change	0.265	0.064	gross operating surplus change	1.852	0.578
	merchandise exports change	0.245	0.051	merchandise reexports change	1.756	0.522
	composite leading index change	0.192	0.040	merchandise exports change	1.696	0.496
	air passenger arrivals change	0.170	0.033	fnb services index change	1.581	0.582
	cpf due members change	-0.165	0.055	pawnshop pledges received change	1.379	0.266
Post-COVID	industrial production index change	0.516	0.163	air passenger arrivals change	2.917	0.680
	air cargo loaded change	0.324	0.125	aircraft arrivals departures change	2.680	0.800
	taxes less subsidies change	0.311	0.107	new vehicle registration change	2.451	0.691
	merchandise reexports change	0.310	0.068	industrial production index change	2.121	0.577
	gross operating surplus change	0.278	0.043	fnb services index change	1.920	0.465
	merchandise imports change	0.270	0.069	air cargo loaded change	1.886	0.283
	merchandise exports change	0.247	0.063	merchandise imports change	1.582	0.178
	composite leading index change	0.207	0.061	pawnshop pledges received change	1.507	0.346
	liquor releases change	0.174	0.107	merchandise reexports change	1.495	0.115
	gross fixed capital formation current change	0.168	0.076	gross operating surplus change	1.432	0.244
Excluding COVID	industrial production index change	0.593	0.345	industrial production index change	2.503	1.165
	taxes less subsidies change	0.322	0.371	air passenger arrivals change	2.120	1.779
	gross operating surplus change	0.317	0.110	air cargo loaded change	2.063	0.483
	merchandise reexports change	0.316	0.079	aircraft arrivals departures change	1.938	1.863
	air cargo loaded change	0.313	0.117	merchandise imports change	1.920	0.718
	merchandise imports change	0.268	0.098	gross operating surplus change	1.825	0.927
	merchandise exports change	0.244	0.078	merchandise reexports change	1.803	0.634
	composite leading index change	0.202	0.074	merchandise exports change	1.734	0.706
	liquor releases change	0.181	0.111	fnb services index change	1.584	0.947
	cpf due members change	-0.166	0.099	new vehicle registration change	1.326	2.485

3. **Ensemble learning models.**

    - *General trends.* Consistently appearing as dominant predictors in both RF and XGB models are industrial production index and air cargo loaded. XGB consistently assigns the highest importance to air cargo loaded across all periods, with narrower confidence intervals signaling stability. In contrast, RF significantly emphasizes this indicator only in the Post-COVID period.
    - *Period-specific variations.* Random Forest notably increased reliance on air cargo loaded in the Post-COVID period, clearly capturing structural shifts in economic dynamics. XGB's predictor importance remains stable across all sub-periods, underscoring its robustness to economic turbulence.
    - *Excluding COVID.* The Excluding COVID period shows strong convergence in ensemble models on international trade indicators (particularly air cargo) and industrial production, confirming robustness in variable selection outside pandemic conditions.

4. **Neural network models.**

    - *General trends.* MLP and GRU prominently feature industrial production, gross operating surplus, and unit labor cost. GRU additionally assigns considerable importance to employee compensation and pawnshop pledges, consistently showing less feature importance variability than MLP across all periods.
    - *Period-specific variations.* During COVID, neural networks emphasized labor and production indicators, demonstrating high sensitivity to macroeconomic disruptions. GRU exhibited stable predictor importance, whereas MLP's feature relevance displayed notably higher variability, particularly under volatile economic conditions, particularly under volatile conditions.
    - *Excluding COVID.* Without the COVID period, neural models emphasize the consistent relevance of industrial production, gross operating surplus, and unit labor cost. GRU confirms stability in feature selection, whereas MLP continues to reflect slightly greater variability.

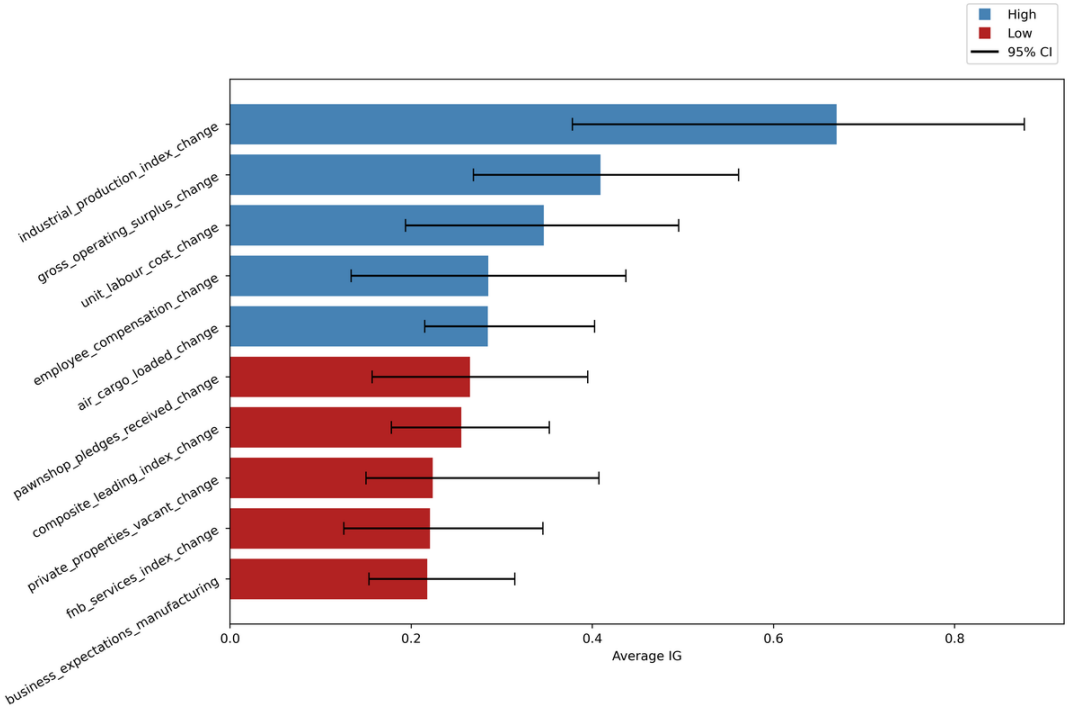
Figure 3.8 to Figure 3.10 illustrate the top 10 explanatory variables identified by GRU for each analyzed period, together with their 95% confidence intervals.

Table 3.6: Ensemble learning models with penalization, top 10 predictors ranked by average feature importance and 95% confidence intervals, estimated via block bootstrap, across different sub-periods

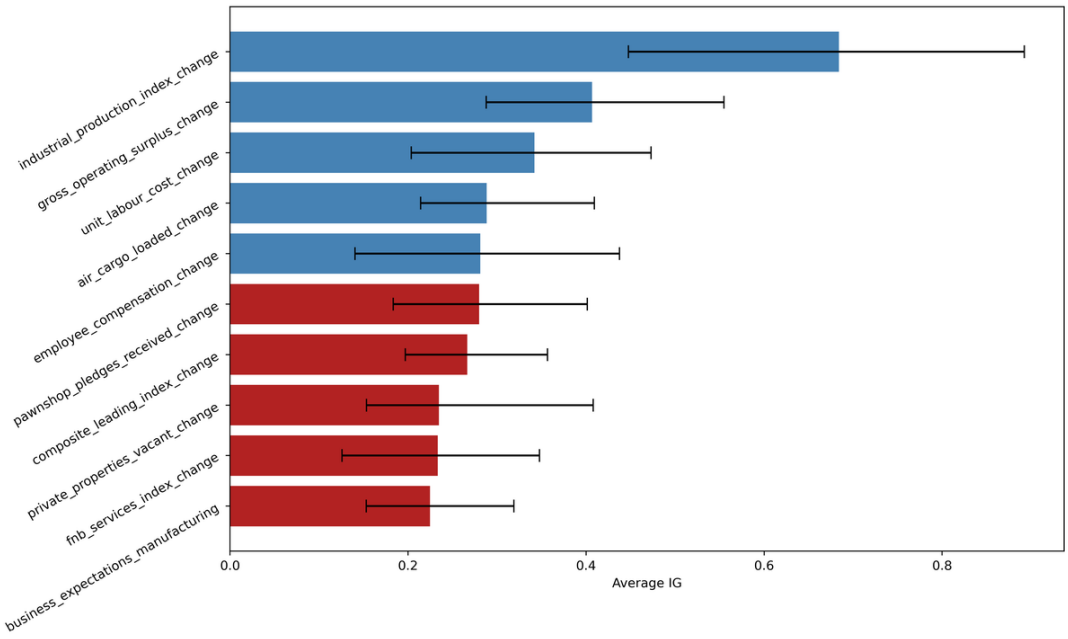
	Random Forest			eXtreme Gradient Boosting		
	Feature	**Feature (avg)**	**CI range**	**Feature**	**Feature (avg)**	**CI range**
Overall	industrial production index change	0.224	0.355	air cargo loaded change	0.097	0.098
	air cargo loaded change	0.217	0.431	merchandise imports change	0.066	0.116
	merchandise imports change	0.064	0.114	air passenger arrivals change	0.062	0.083
	air passenger arrivals change	0.046	0.091	merchandise reexports change	0.058	0.049
	gross operating surplus change	0.042	0.104	industrial production index change	0.054	0.066
	merchandise reexports change	0.036	0.071	gross operating surplus change	0.051	0.074
	business expectations manufacturing	0.030	0.057	aircraft arrivals departures change	0.050	0.080
	gross fixed capital formation current change	0.021	0.068	merchandise exports change	0.044	0.092
	merchandise exports change	0.021	0.045	business expectations manufacturing	0.027	0.024
	consumer price index change	0.020	0.047	fnb services index change	0.019	0.027
Pre-COVID	industrial production index change	0.262	0.383	air cargo loaded change	0.084	0.073
	air cargo loaded change	0.093	0.070	gross operating surplus change	0.065	0.066
	air passenger arrivals change	0.070	0.049	air passenger arrivals change	0.062	0.082
	gross operating surplus change	0.070	0.081	merchandise reexports change	0.061	0.048
	merchandise reexports change	0.055	0.034	aircraft arrivals departures change	0.057	0.069
	merchandise imports change	0.043	0.033	industrial production index change	0.055	0.056
	merchandise exports change	0.035	0.027	merchandise imports change	0.051	0.101
	business expectations manufacturing	0.024	0.026	merchandise exports change	0.046	0.094
	aircraft arrivals departures change	0.023	0.026	business expectations manufacturing	0.028	0.022
	pawnshop pledges received change	0.022	0.023	employee compensation change	0.021	0.035
COVID	industrial production index change	0.154	0.064	air cargo loaded change	0.095	0.063
	air cargo loaded change	0.117	0.098	merchandise imports change	0.066	0.052
	air passenger arrivals change	0.065	0.017	air passenger arrivals change	0.057	0.043
	merchandise reexports change	0.043	0.027	merchandise reexports change	0.052	0.013
	gross operating surplus change	0.042	0.041	aircraft arrivals departures change	0.050	0.059
	merchandise imports change	0.041	0.028	merchandise exports change	0.048	0.038
	business expectations manufacturing	0.031	0.012	gross operating surplus change	0.047	0.031
	aircraft arrivals departures change	0.027	0.023	industrial production index change	0.042	0.055
	merchandise exports change	0.025	0.034	employee compensation change	0.035	0.084
	pawnshop pledges received change	0.023	0.019	business expectations manufacturing	0.026	0.011
Post-COVID	air cargo loaded change	0.405	0.254	air cargo loaded change	0.112	0.085
	industrial production index change	0.208	0.109	merchandise imports change	0.085	0.106
	merchandise imports change	0.099	0.109	air passenger arrivals change	0.066	0.068
	consumer price index change	0.038	0.042	industrial production index change	0.058	0.063
	business expectations manufacturing	0.037	0.063	merchandise reexports change	0.056	0.044
	gross fixed capital formation current change	0.028	0.083	merchandise exports change	0.041	0.060
	pawnshop loans given change	0.018	0.046	aircraft arrivals departures change	0.040	0.062
	pawnshop pledges received change	0.017	0.034	gross operating surplus change	0.037	0.047
	merchandise reexports change	0.011	0.060	business expectations manufacturing	0.026	0.022
	air passenger arrivals change	0.011	0.031	private progress payments change	0.024	0.046
Excluding COVID	industrial production index change	0.237	0.348	air cargo loaded change	0.097	0.100
	air cargo loaded change	0.235	0.438	merchandise imports change	0.066	0.119
	merchandise imports change	0.068	0.117	air passenger arrivals change	0.063	0.084
	air passenger arrivals change	0.043	0.092	merchandise reexports change	0.059	0.050
	gross operating surplus change	0.042	0.109	industrial production index change	0.057	0.064
	merchandise reexports change	0.035	0.072	gross operating surplus change	0.052	0.076
	business expectations manufacturing	0.030	0.059	aircraft arrivals departures change	0.050	0.075
	gross fixed capital formation current change	0.022	0.071	merchandise exports change	0.044	0.093
	consumer price index change	0.021	0.047	business expectations manufacturing	0.028	0.024
	merchandise exports change	0.020	0.045	fnb services index change	0.019	0.021

Table 3.7: Deep learning models with penalization, top 10 predictors ranked by average feature importance and 95% confidence intervals, estimated via block bootstrap, across different sub-periods

	Multilayer Perceptron			Gated Recurrent Units		
	Feature	**Feature (avg)**	**CI range**	**Feature**	**Feature (avg)**	**CI range**
Overall	industrial production index change	0.395	0.615	industrial production index change	0.670	0.499
	gross operating surplus change	0.246	0.369	gross operating surplus change	0.409	0.293
	air cargo loaded change	0.180	0.231	unit labor cost change	0.346	0.302
	business expectations manufacturing	0.149	0.230	employee compensation change	0.285	0.303
	unit labor cost change	0.140	0.224	air cargo loaded change	0.285	0.187
	composite leading index change	0.124	0.152	pawnshop pledges received change	0.265	0.238
	merchandise imports change	0.118	0.149	composite leading index change	0.255	0.174
	fnb services index change	0.110	0.183	private properties vacant change	0.224	0.257
	merchandise reexports change	0.109	0.122	fnb services index change	0.221	0.220
	merchandise exports change	0.105	0.132	business expectations manufacturing	0.218	0.161
Pre-COVID	industrial production index change	0.486	0.528	industrial production index change	0.653	0.473
	gross operating surplus change	0.314	0.299	gross operating surplus change	0.440	0.270
	air cargo loaded change	0.211	0.174	unit labor cost change	0.370	0.209
	unit labor cost change	0.178	0.157	employee compensation change	0.279	0.268
	business expectations manufacturing	0.169	0.188	air cargo loaded change	0.275	0.146
	composite leading index change	0.143	0.155	composite leading index change	0.269	0.147
	merchandise imports change	0.135	0.118	private properties vacant change	0.248	0.217
	fnb services index change	0.115	0.164	pawnshop pledges received change	0.235	0.125
	employee compensation change	0.114	0.196	business expectations manufacturing	0.213	0.172
	merchandise reexports change	0.108	0.098	merchandise imports change	0.213	0.159
COVID	industrial production index change	0.429	0.356	industrial production index change	0.590	0.434
	gross operating surplus change	0.238	0.168	gross operating surplus change	0.422	0.276
	air cargo loaded change	0.187	0.156	unit labor cost change	0.370	0.273
	business expectations manufacturing	0.172	0.123	employee compensation change	0.308	0.278
	composite leading index change	0.146	0.102	air cargo loaded change	0.265	0.063
	unit labor cost change	0.142	0.127	composite leading index change	0.194	0.054
	merchandise imports change	0.138	0.091	pawnshop pledges received change	0.183	0.112
	merchandise exports change	0.127	0.043	business expectations manufacturing	0.179	0.019
	merchandise reexports change	0.126	0.039	private properties vacant change	0.163	0.030
	gross fixed capital formation current change	0.117	0.099	merchandise imports change	0.163	0.031
Post-COVID	industrial production index change	0.272	0.496	industrial production index change	0.722	0.200
	gross operating surplus change	0.168	0.285	gross operating surplus change	0.367	0.118
	air cargo loaded change	0.140	0.198	pawnshop pledges received change	0.334	0.131
	business expectations manufacturing	0.116	0.208	unit labor cost change	0.308	0.257
	merchandise reexports change	0.103	0.130	air cargo loaded change	0.304	0.185
	fnb services index change	0.101	0.178	fnb services index change	0.285	0.152
	merchandise exports change	0.095	0.133	employee compensation change	0.283	0.287
	unit labor cost change	0.093	0.197	composite leading index change	0.264	0.142
	composite leading index change	0.093	0.123	business expectations manufacturing	0.238	0.082
	merchandise imports change	0.091	0.125	private properties vacant change	0.219	0.246
Excluding COVID	industrial production index change	0.389	0.621	industrial production index change	0.684	0.445
	gross operating surplus change	0.248	0.374	gross operating surplus change	0.407	0.267
	air cargo loaded change	0.179	0.233	unit labor cost change	0.342	0.269
	business expectations manufacturing	0.145	0.233	air cargo loaded change	0.288	0.195
	unit labor cost change	0.140	0.225	employee compensation change	0.281	0.297
	composite leading index change	0.120	0.157	pawnshop pledges received change	0.280	0.218
	merchandise imports change	0.115	0.148	composite leading index change	0.266	0.160
	fnb services index change	0.109	0.186	private properties vacant change	0.235	0.255
	merchandise reexports change	0.106	0.125	fnb services index change	0.233	0.222
	merchandise exports change	0.101	0.133	business expectations manufacturing	0.225	0.166

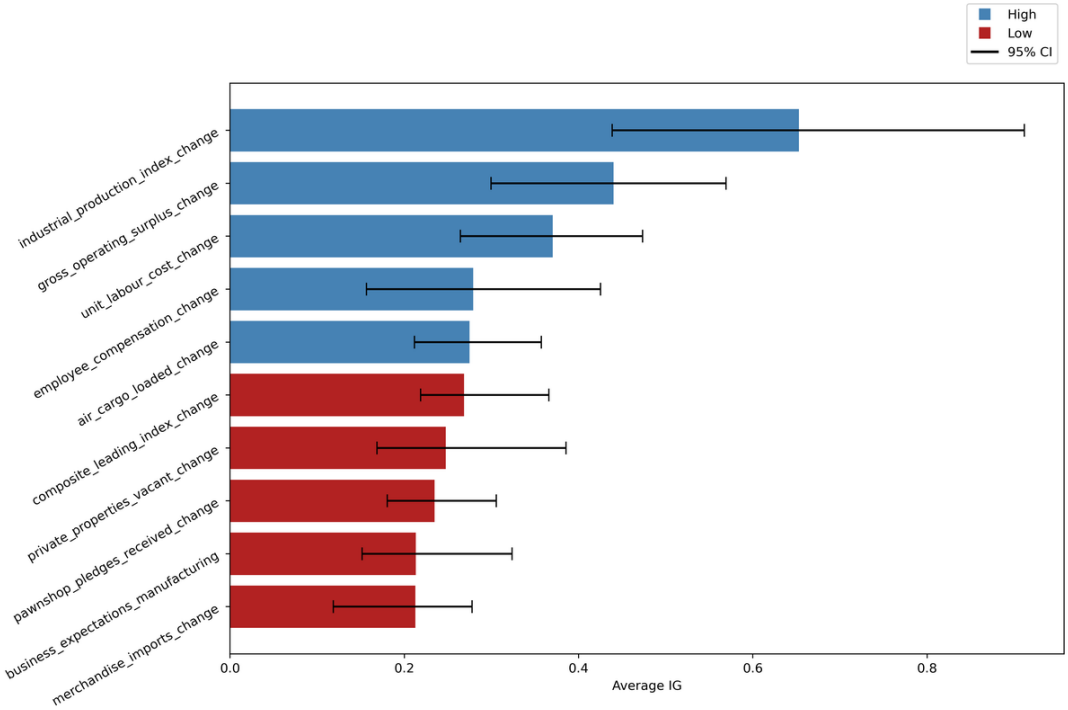


(a) Overall

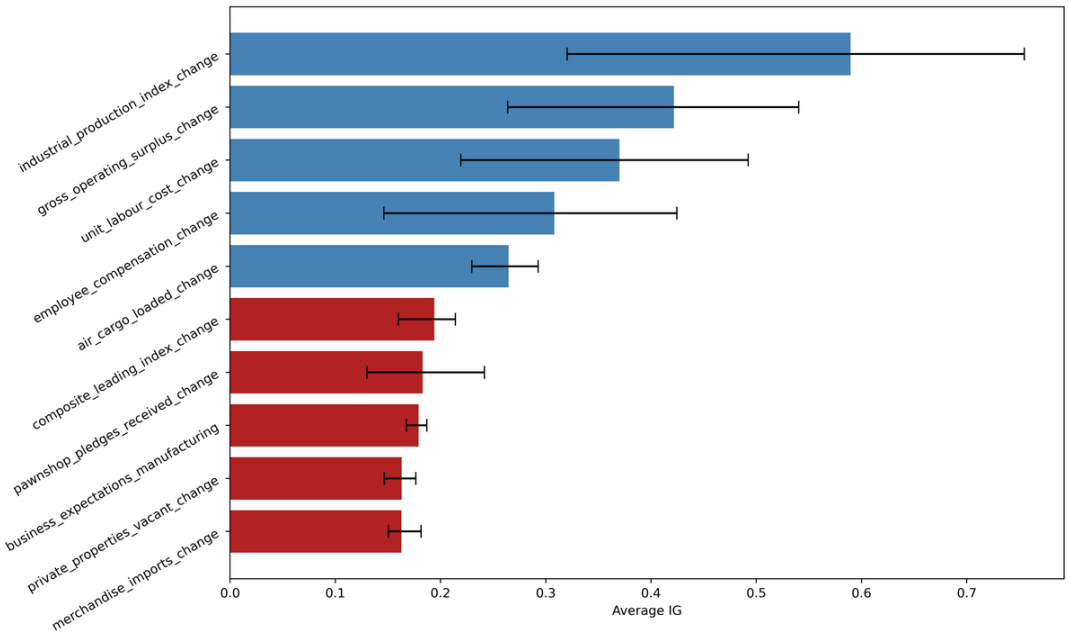


(b) Excluding COVID

Figure 3.8: GRU model, top 10 explanatory variables ranked by average Integrated Gradients in the Overall and Excluding COVID periods. Variables are color-coded based on importance magnitude (High, Low). Bars represent the average Integrated Gradients values; error bars denote the 95% confidence intervals obtained via block bootstrap.

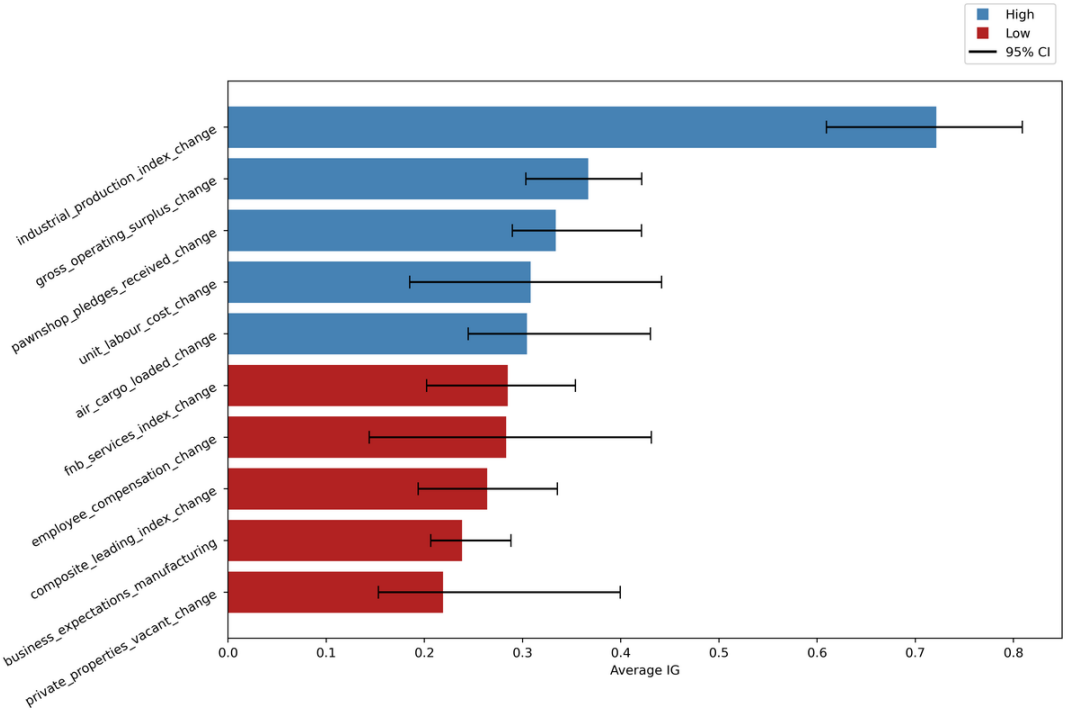


(a) Pre-COVID



(b) COVID

Figure 3.9: GRU model, top 10 explanatory variables ranked by average Integrated Gradients in Pre-COVID and COVID periods. Variables are color-coded based on importance magnitude (High, Low). Bars represent the average Integrated Gradients values; error bars denote the 95% confidence intervals obtained via block bootstrap.



(a) Post-COVID

Figure 3.10: GRU model, top 10 explanatory variables ranked by average Integrated Gradients in Post-COVID period. Variables are color-coded based on importance magnitude (High, Low). Bars represent the average Integrated Gradients values; error bars denote the 95% confidence intervals obtained via block bootstrap.

The graphical representation clearly shows stability and variability in selecting key predictors across macroeconomic regimes, consistently reflecting the descriptive results summarized in Table Y.

The reported patterns highlight differences in feature importance attribution among Penalized linear, Dimensionality reduction-based, Ensemble learning, and Neural network models across varying macroeconomic regimes.

Penalized linear and Dimensionality reduction-based approaches consistently identify industrial production, trade indicators, and labor market variables among their most influential predictors, maintaining stable relevance across various sub-periods.

Ensemble learning and neural network models display variability in feature importance, reflecting differences in how these methodologies select and emphasize predictors across economic regimes. Such variability is particularly evident during significant macroeconomic fluctuations (e.g., COVID).

Figure 3.11 provides an additional perspective to the static analyses presented. It demonstrates, for illustrative purposes, the quarterly evolution of the feature importance calculated via Integrated Gradients for the variable industrial production index change in the GRU model. This representation highlights the dynamic reactivity of the model in correspondence with relevant economic events, such as the onset and development of the COVID-19 pandemic.

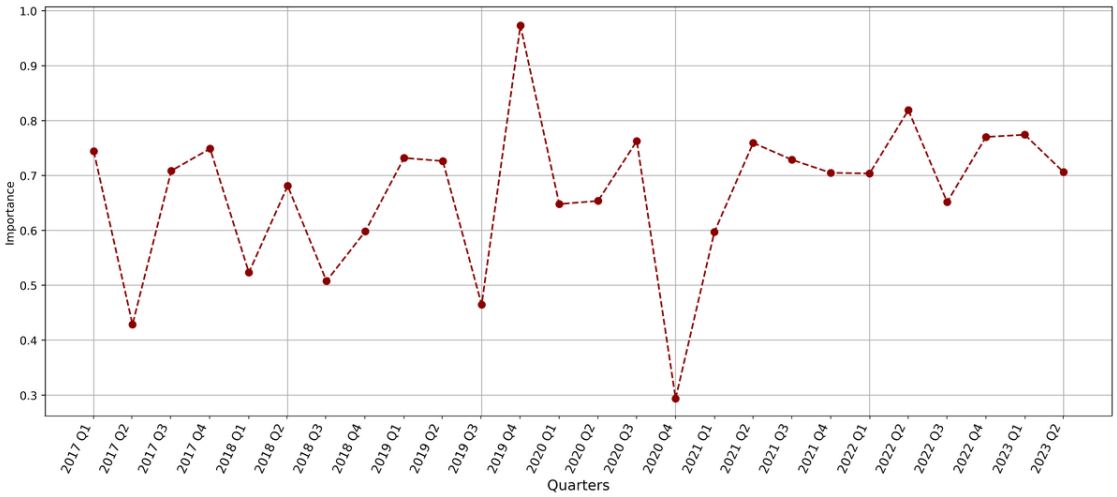


Figure 3.11: Quarterly evolution of the feature importance (Integrated Gradients) of the variable industrial production index change in the GRU model. Importance is estimated for each quarter between 2017 Q1 and 2023 Q2.

To further contextualize the relevance of the explanatory variables that emerged from the previous analyses, Figure 3.12 shows the Spearman correlation between each predictor variable and the target variable (quarterly Singapore's GDP growth). This representation highlights that some of the variables consistently selected by the models, such as industrial production, air cargo loaded and pawnshop pledges received, show statistically significant correlations with GDP, empirically confirming the robustness and coherence of the selections made by the analyzed models.

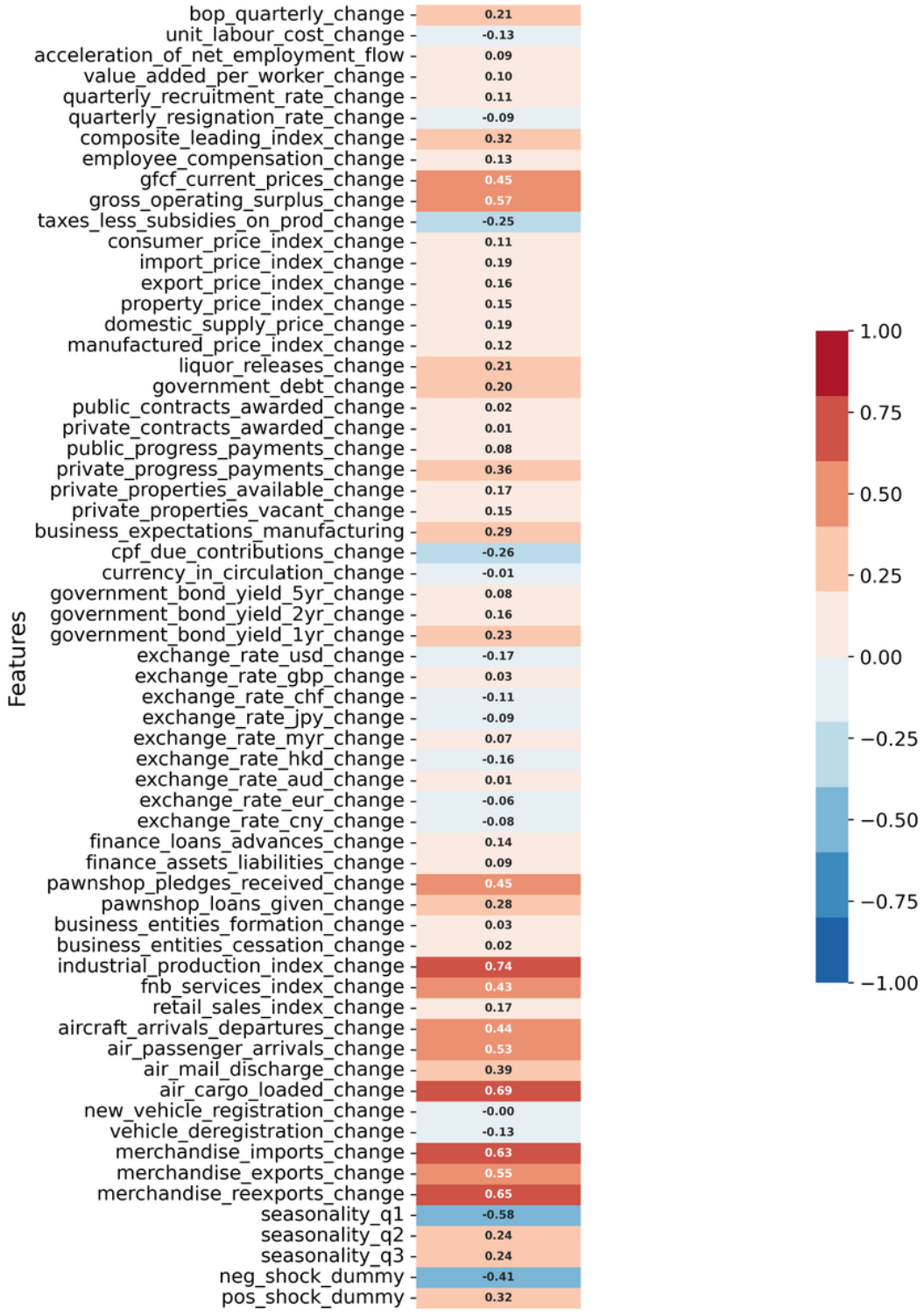


Figure 3.12: Spearman correlations between the predictors and the target variable. Predictors are sorted in the same order as they appear in the dataset. Positive values (red) indicate positive correlations with GDP, while negative values (blue) indicate inverse correlations.

3.1.4 Model Selection and Aggregated Predictions

Nowcast combination methods enhance prediction accuracy and robustness, mitigating model-specific errors and uncertainties (Bates & Granger, 1969; Makridakis et al., 2020; Timmermann, 2006). In our approach, we combined predictions from models selected through the Model Confidence Set (MCS) procedure (P. R. Hansen et al., 2011) into three aggregation strategies: Simple Average (SA), Weighted Average (WA), and Exponentially Weighted Average (EWA). WA and EWA explicitly track dynamic model weights, substantially enhancing interpretability.

The SA approach assigns equal weights to all selected models, disregarding differences in individual model performance. While simple, this method provides a robust benchmark, effectively mitigating risks associated with over-reliance on any nowcasting approach. In contrast, WA assigns static weights to individual models based on their relative predictive performance over the entire historical period. Differently, EWA dynamically updates model weights, assigning greater importance to recent nowcast performances through exponential smoothing. The dynamic weighting scheme of EWA explicitly captures structural breaks and evolving economic conditions, significantly enhancing explainability by transparently illustrating shifts in each model's importance over time.

Initially, the MCS procedure identified a subset of models demonstrating statistically indistinguishable predictive ability. The MCS selected penalized linear regression models, dimensionality reduction techniques, and GRU among neural networks (excluding ensemble learning models and MLP), leading to the subsequent application of these selected models for aggregation.

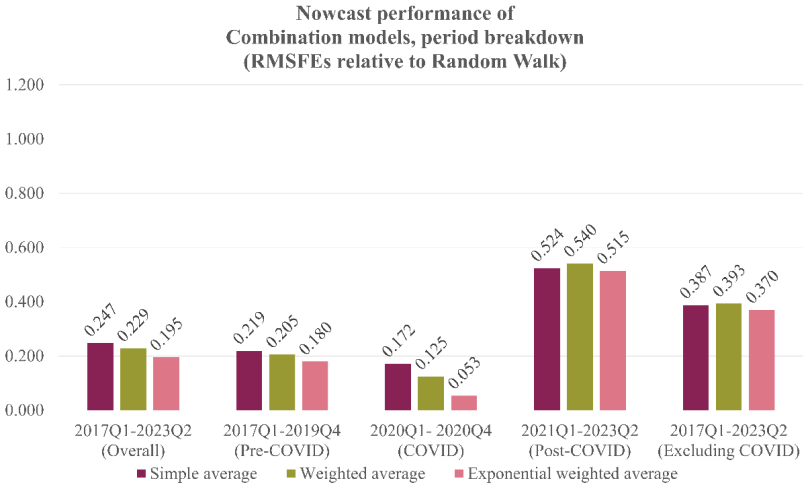
Empirical findings, summarized visually in Figure 3.13 and detailed quantitatively in Tables 3.8 and 3.9, highlighted that:

- All combined methods (SA, WA, and EWA) improved prediction accuracy relative to individual baseline benchmarks, i.e., the Random Walk, AR(3), and the Dynamic Factor Model, in overall prediction windows and sub-periods.

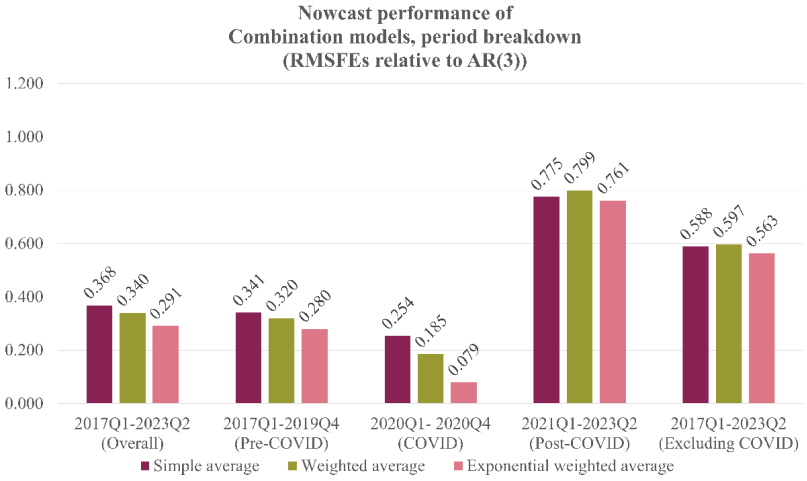
- Among the aggregation strategies, EWA consistently provided lower prediction errors compared to WA and SA across various sub-periods, especially during the COVID-19 crisis (2020 Q1-2020 Q4). This improved performance is directly attributable to its dynamically adaptive weighting scheme. It also serves as a valuable interpretative tool by reflecting each model's changing relevance during economic shifts

- SA offered stable improvements, mitigating misspecification risks through equal weighting. Its static weights, however, led to comparatively higher errors during pronounced instability, underscoring EWA's adaptive advantage. Still, SA occasionally outperformed WA post-COVID, reaffirming its robustness under moderate uncertainty.

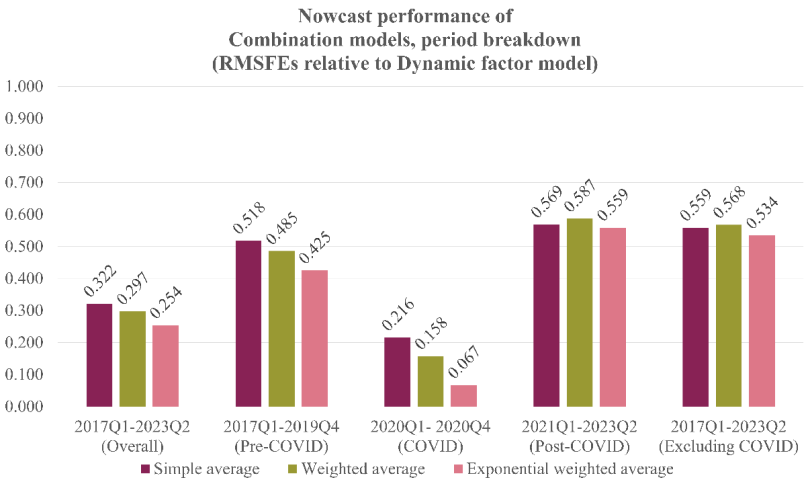
- RMSFE results highlighted notably lower prediction errors for EWA during the COVID-19 crisis than all benchmark models.



(a) Benchmark: Random Walk



(b) Benchmark: AR(3)



(c) Benchmark: Dynamic Factor Model

Figure 3.13: Nowcast performance (RMSFE ratios relative to benchmarks) for Simple Average, Weighted Average, and Exponential Weighted Average aggregation strategies across different macroeconomic regimes.

Table 3.8: Forecasting performance metrics (MSFE, RMSFE, and MAFE) of combined models (Simple Average, Weighted Average, and Exponentially Weighted Average) across sub-periods

Combinations models

Simple average

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excl. COVID
MSFE	1.671	0.412	3.852	2.309	1.274
RMSFE	1.293	0.641	1.963	1.520	1.129
MAFE	0.918	0.527	1.417	1.188	0.827

Weighted average

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excl. COVID
MSFE	1.427	0.362	2.052	2.456	1.314
RMSFE	1.195	0.601	1.433	1.567	1.146
MAFE	0.877	0.497	1.132	1.231	0.831

Exponentially weighted average

Metric	Overall	Pre-COVID	COVID	Post-COVID	Excl. COVID
MSFE	1.043	0.278	0.372	2.229	1.165
RMSFE	1.021	0.527	0.610	1.493	1.079
MAFE	0.730	0.433	0.462	1.194	0.779

Figure 3.14 visually captures the evolution of model weights in WA and EWA frameworks, providing immediate insight into each model's shifting importance. Complementary detailed quarterly breakdowns in Table 3.10 (WA) and Table 3.11 (EWA) further clarify how individual models predominantly influenced aggregated predictions across different economic regimes, explicitly addressing the explainability objective:

- Considering the whole prediction period, the GRU model was initially assigned the highest weights in WA and EWA strategies until late 2019. WA maintained essentially stable and constant weights over the entire period, demonstrating minimal fluctuations in model importance. Conversely, EWA dynamically adjusted model weights based on recent nowcasting performance, highlighting substantial shifts in model relevance, particularly during episodes of economic instability such as the COVID-19 period. This adaptive mechanism improved prediction accuracy and interpretability, especially during significant economic disruptions such as COVID-19.

- During the COVID-19 disruption, dimensionality reduction-based models, particularly PCR, experienced substantial weight increases within the WA and EWA methods. This dynamic adjustment, especially in the EWA plot, reflects PCR's superior predictive reliability under heightened economic volatility. Conversely, SA maintained equal weighting across models without adjustments, implicitly limiting its responsiveness during rapid economic shifts.

- In the Post-COVID sub-period, Elastic Net consistently received the highest weights within the WA and EWA frameworks. Again, the adaptability of EWA is distinctly evident, showing how quickly and effectively Elastic Net gained prominence as economic conditions stabilized but remained uncertain. WA confirms Elastic Net's

Table 3.9: Forecasting performance (RMSFE ratios) of selected individual models (from MCS) and combination models relative to benchmarks (Random Walk, AR(3), Dynamic Factor Model) across distinct sub-periods.

RMSFE, relative to Random walk

	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
Best models (MCS)					
LASSO	0.427	0.215	0.446	0.493	0.367
Ridge	0.277	0.276	0.164	0.621	0.464
Elastic net	0.247	0.264	0.166	0.510	0.394
Principal component regression	0.280	0.310	0.088	0.706	0.526
Partial least squares regression	0.281	0.260	0.201	0.576	0.432
Gated recurrent unit	0.445	0.217	0.477	0.449	0.342
Combination models					
Simple average	0.247	0.219	0.172	0.524	0.387
Weighted average	0.229	0.205	0.125	0.540	0.393
Exponential weighted average	0.195	0.180	0.053	0.515	0.370

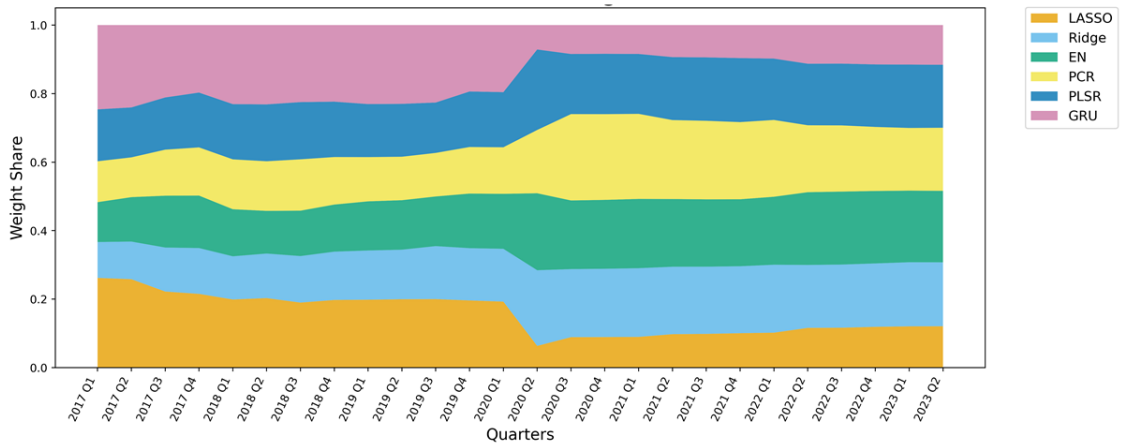
RMSFE, relative to AR(3)

	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
Best models (MCS)					
LASSO	0.636	0.334	0.660	0.728	0.557
Ridge	0.412	0.430	0.243	0.918	0.705
Elastic net	0.368	0.411	0.245	0.754	0.599
Principal component regression	0.418	0.482	0.129	1.044	0.800
Partial least squares regression	0.418	0.404	0.297	0.851	0.656
Gated recurrent unit	0.663	0.338	0.705	0.664	0.519
Combination models					
Simple average	0.368	0.341	0.254	0.775	0.588
Weighted average	0.340	0.320	0.185	0.799	0.597
Exponential weighted average	0.291	0.280	0.079	0.761	0.563

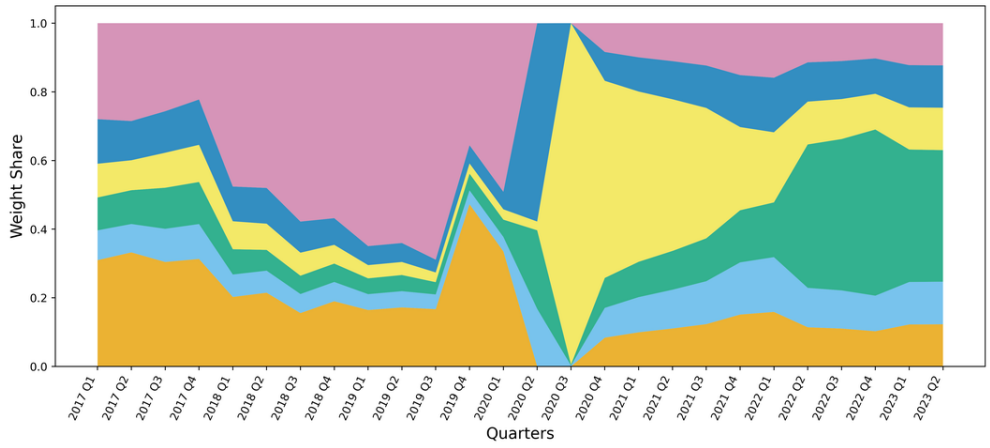
RMSFE, relative to Dynamic factor model

	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
Best models (MCS)					
LASSO	0.555	0.507	0.561	0.535	0.530
Ridge	0.360	0.653	0.206	0.674	0.670
Elastic net	0.321	0.625	0.208	0.554	0.569
Principal component regression	0.364	0.733	0.110	0.767	0.760
Partial least squares regression	0.365	0.614	0.253	0.626	0.623
Gated recurrent unit	0.579	0.514	0.600	0.488	0.493
Combination models					
Simple average	0.322	0.518	0.216	0.569	0.559
Weighted average	0.297	0.485	0.158	0.587	0.568
Exponential weighted average	0.254	0.425	0.067	0.559	0.534

high historical average reliability, maintaining constant but elevated weights. With its equal-weight structure, SA continued to allocate identical importance across all models, thus not reflecting these nuanced shifts.



(a) Weighted Average



(b) Exponential Weighted Average

Figure 3.14: Stacked area chart illustrating the evolution of individual model weights within the Weighted Average and Exponential Weighted Average aggregation frameworks, from 2017 Q1 to 2023 Q2.

A detailed quantitative breakdown of dominant models across each sub-period and aggregation framework (SA, WA, EWA) is provided in Table 3.12. This analysis confirmed the frequent selection of PCR and Elastic Net after the structural break induced by the pandemic, particularly under the WA and EWA aggregation frameworks, while highlighting significant contributions from GRU and PLSR in earlier and transitional periods.

A final diagnostic check was performed on the residuals from each aggregated nowcasting methodology to ensure robustness and statistical reliability. Table 3.13 summarizes results from the Shapiro-Wilk test for normality and the Ljung-Box test for residual autocorrelation. The tests suggest that residuals from SA, WA, and EWA approaches largely exhibit

Table 3.10: Evolution of individual model weights within the Weighted Average (WA) aggregation framework, and corresponding dominant model, from 2017 Q1 to 2023 Q2

Quarters	LASSO weight	Ridge weight	EN weight	PCR weight	PLSR weight	GRU weight	Dominant model
2017 Q1	0.262	0.105	0.116	0.119	0.152	0.246	LASSO
2017 Q2	0.258	0.110	0.130	0.116	0.146	0.240	LASSO
2017 Q3	0.222	0.128	0.151	0.134	0.152	0.211	LASSO
2017 Q4	0.215	0.134	0.153	0.141	0.160	0.197	LASSO
2018 Q1	0.199	0.126	0.137	0.146	0.161	0.231	GRU
2018 Q2	0.203	0.130	0.125	0.145	0.166	0.232	GRU
2018 Q3	0.190	0.136	0.133	0.150	0.167	0.225	GRU
2018 Q4	0.198	0.141	0.137	0.139	0.162	0.224	GRU
2019 Q1	0.198	0.144	0.143	0.129	0.155	0.231	GRU
2019 Q2	0.200	0.145	0.145	0.127	0.154	0.230	GRU
2019 Q3	0.200	0.155	0.145	0.127	0.147	0.226	GRU
2019 Q4	0.196	0.153	0.159	0.136	0.162	0.194	LASSO
2020 Q1	0.193	0.154	0.160	0.136	0.161	0.196	GRU
2020 Q2	0.064	0.221	0.225	0.185	0.235	0.071	PLSR
2020 Q3	0.089	0.199	0.200	0.252	0.175	0.084	PCR
2020 Q4	0.089	0.199	0.201	0.250	0.176	0.084	PCR
2021 Q1	0.090	0.200	0.202	0.248	0.175	0.084	PCR
2021 Q2	0.098	0.197	0.198	0.230	0.184	0.094	PCR
2021 Q3	0.099	0.196	0.197	0.229	0.185	0.094	PCR
2021 Q4	0.101	0.195	0.195	0.225	0.187	0.096	PCR
2022 Q1	0.102	0.198	0.198	0.224	0.179	0.098	PCR
2022 Q2	0.116	0.184	0.212	0.195	0.180	0.113	EN
2022 Q3	0.117	0.184	0.213	0.193	0.180	0.113	EN
2022 Q4	0.119	0.185	0.211	0.187	0.182	0.115	EN
2023 Q1	0.121	0.187	0.209	0.183	0.185	0.115	EN
2023 Q2	0.121	0.187	0.209	0.184	0.184	0.116	EN

Table 3.11: Evolution of individual model weights within the Exponential Weighted Average (EWA) aggregation framework, and corresponding dominant model, from 2017 Q1 to 2023 Q2

Quarters	LASSO weight	Ridge weight	EN weight	PCR weight	PLSR weight	GRU weight	Dominant model
2017 Q1	0.310	0.087	0.096	0.098	0.130	0.280	LASSO
2017 Q2	0.332	0.083	0.099	0.087	0.114	0.285	LASSO
2017 Q3	0.304	0.097	0.120	0.102	0.121	0.257	LASSO
2017 Q4	0.313	0.102	0.123	0.108	0.132	0.222	LASSO
2018 Q1	0.203	0.065	0.074	0.081	0.101	0.476	GRU
2018 Q2	0.215	0.064	0.061	0.076	0.104	0.480	GRU
2018 Q3	0.156	0.055	0.053	0.067	0.091	0.578	GRU
2018 Q4	0.190	0.056	0.054	0.055	0.078	0.568	GRU
2019 Q1	0.164	0.046	0.046	0.038	0.055	0.649	GRU
2019 Q2	0.172	0.047	0.047	0.038	0.055	0.640	GRU
2019 Q3	0.167	0.043	0.036	0.028	0.037	0.689	GRU
2019 Q4	0.472	0.040	0.049	0.030	0.053	0.356	LASSO
2020 Q1	0.335	0.042	0.050	0.030	0.051	0.491	GRU
2020 Q2	0.000	0.168	0.229	0.025	0.578	0.000	PLSR
2020 Q3	0.000	0.003	0.003	0.994	0.000	0.000	PCR
2020 Q4	0.084	0.087	0.087	0.574	0.084	0.084	PCR
2021 Q1	0.100	0.102	0.103	0.496	0.100	0.100	PCR
2021 Q2	0.111	0.113	0.113	0.442	0.111	0.111	PCR
2021 Q3	0.123	0.125	0.125	0.380	0.124	0.123	PCR
2021 Q4	0.151	0.152	0.152	0.242	0.151	0.151	PCR
2022 Q1	0.159	0.159	0.159	0.204	0.159	0.159	PCR
2022 Q2	0.114	0.115	0.418	0.125	0.114	0.114	EN
2022 Q3	0.110	0.111	0.441	0.116	0.111	0.110	EN
2022 Q4	0.103	0.103	0.484	0.104	0.103	0.103	EN
2023 Q1	0.122	0.124	0.386	0.123	0.123	0.122	EN
2023 Q2	0.123	0.124	0.383	0.123	0.123	0.123	EN

Table 3.12: Frequency of dominant models selected within each aggregation framework (Simple Average, Weighted Average, Exponentially Weighted Average) by sub-period

Simple Average

	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
Penalized linear models					
LASSO	2	2	0	0	2
Ridge	4	2	0	2	4
Elastic net	3	2	0	1	3
Dimensionality reduction-based models					
Principal component regression	2	0	1	1	1
Partial least squares regression	8	3	2	3	6
Neural networks					
Gated recurrent units	7	3	1	3	6

Weighted Average

	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
Penalized linear models					
LASSO	5	5	0	0	5
Ridge	0	0	0	0	0
Elastic net	5	0	0	5	5
Dimensionality reduction-based models					
Principal component regression	7	0	2	5	5
Partial least squares regression	1	0	1	0	0
Neural networks					
Gated recurrent units	8	7	1	0	7

Exponential Weighted Average

	Overall	Pre-COVID	COVID	Post-COVID	Excluding COVID
Penalized linear models					
LASSO	5	5	0	0	5
Ridge	0	0	0	0	0
Elastic net	5	0	0	5	5
Dimensionality reduction-based models					
Principal component regression	7	0	2	5	5
Partial least squares regression	1	0	1	0	0
Neural networks					
Gated recurrent units	8	7	1	0	7

acceptable statistical properties and support the statistical validity of the combined predictions across different aggregation strategies. In particular, the Shapiro-Wilk tests do not strongly reject the null hypothesis of residual normality for any of the aggregation methods, indicating that residual distributions from all combinations remain approximately normal. The Ljung-Box test reveals that residual autocorrelation issues are modest, with SA and WA residuals showing marginal evidence of autocorrelation at conventional significance levels, while EWA residuals appear more consistently free of significant autocorrelation.

Aggregating predictions using MCS-selected models systematically enhanced predictive accuracy and robustness compared to baseline benchmarks (Random Walk, AR(3), DFM). The dynamic weighting of EWA proved particularly beneficial during periods of heightened volatility, such as the COVID-19 pandemic, significantly improving both predictive accuracy and interpretability. Meanwhile, the simpler SA approach outperformed WA across various

Table 3.13: Residual diagnostics (Shapiro-Wilk and Ljung-Box tests) for aggregated models (Simple Average, Weighted Average, Exponentially Weighted Average

	SA		WA		EWA	
	Statistic	**p-value**	**Statistic**	**p-value**	**Statistic**	**p-value**
Shapiro-Wilk	0.928	0.071	0.959	0.365	0.928	0.070
Ljung-Box (lag=4)	10.136	0.038	9.649	0.047	7.952	0.093

regimes, confirming its reliability as a robust benchmark.

3.1.5 Formal Predictive-Ability Tests

Evaluating the predictive performance of nowcasting models necessitates a rigorous statistical approach to determine whether observed improvements in prediction accuracy are statistically meaningful rather than due to random variation. This section presents empirical findings from the Giacomini-White (GW) predictive ability test (Giacomini & White, 2006). Unlike classic tests, such as the Diebold-Mariano test (Diebold & Mariano, 1995), the GW test explicitly assesses conditional predictive accuracy, essential when nowcast models adapt dynamically to evolving economic information.[1] We assess whether our models selected by the Model Confidence Set and our nowcast aggregation methods demonstrate statistically significant predictive superiority over our three established benchmarks.

The Giacomini-White test was systematically employed across these comparative scenarios, explicitly accommodating conditional predictive abilities and relying on robust covariance estimation through the Lumley-Heagerty (WEAVE) estimator. This estimator ensures robust covariance estimation, effectively adjusting standard errors for heteroskedasticity and autocorrelation in prediction errors. As summarized in Table 3.14, when assessing predictive performance relative to the Random Walk benchmark, all MCS-selected individual models and combined nowcasts consistently demonstrated statistically significant improvements.

Specifically:

- The Wald test statistics were uniformly high (ranging approximately from 21 to 29), firmly rejecting the null hypothesis of equal predictive accuracy.
- All tested models showed statistically significant improvements (all p-values < 0.001), with EN and EWA consistently outperforming across benchmarks.
- Among individual models, Elastic Net (EN) yielded the most pronounced reduction in prediction errors (intercept estimate: -25.63), closely followed by PLSR and Ridge, underscoring their predictive reliability.
- Among combined nowcasts, EWA showed the most significant improvement, indicated by the lowest intercept estimate (-26.26) and strongest statistical significance (Wald p-value = 0.000016), highlighting its capacity to dynamically adapt to periods of higher uncertainty, such as the COVID-19 pandemic.

[1] In this context, predictive ability refers to a model's capacity to consistently deliver more accurate nowcasts compared to alternative models, considering all information available at each prediction step.

Table 3.14: Giacomini-White test results comparing predictive accuracy of MCS-selected individual and aggregated models (SA, WA, EWA) against benchmark models (Random Walk, AR(3), Dynamic Factor Model). Negative intercepts indicate superior nowcast performance relative to benchmarks.

Random Walk

Model	Intercept (Estimate)	p-value (Intercept)	Wald statistic	Wald p-value
LASSO	-22.324630389	0.000111374	20.954358337	0.000111374
Ridge	-25.207332539	0.000019777	27.514105600	0.000019777
EN	-25.628654847	0.000015209	28.594954554	0.000015209
PCR	-25.151611279	0.000030197	25.820526531	0.000030197
PLSR	-25.146212250	0.000015814	28.433066364	0.000015814
GRU	-21.888333365	0.000046176	24.179216204	0.000046176
SA	-25.627322138	0.000020503	27.367642584	0.000020503
WA	-25.245151148	0.000021503	27.174652336	0.000021503
EWA	-26.255301547	0.000016205	28.331950232	0.000016205

AR(3)

Model	Intercept (Estimate)	p-value (Intercept)	Wald statistic	Wald p-value
LASSO	-7.338380433	0.000000494	45.035271911	0.000000494
Ridge	-10.221082583	0.000041759	24.562593668	0.000041759
EN	-10.642404890	0.000021016	27.267321898	0.000021016
PCR	-10.165361322	0.000060456	23.167166436	0.000060456
PLSR	-10.159962293	0.000036978	25.030507111	0.000036978
GRU	-6.902083408	0.000001242	40.156413956	0.000001242
SA	-10.641072182	0.000012711	29.346717758	0.000012711
WA	-10.258901191	0.000013859	28.983078148	0.000013859
EWA	-11.269051591	0.000015030	28.644198103	0.000015030

Dynamic Factor Model (DFM)

Model	Intercept (Estimate)	p-value (Intercept)	Wald statistic	Wald p-value
LASSO	-11.186248419	0.000008183	31.239860475	0.000008183
Ridge	-14.068950569	0.000037424	24.984211817	0.000037424
EN	-14.490272876	0.000022913	26.918502887	0.000022913
PCR	-14.013229308	0.000061349	23.112753861	0.000061349
PLSR	-14.007830280	0.000029409	25.924740344	0.000029409
GRU	-10.749951394	0.000005111	33.339347101	0.000005111
SA	-14.488940168	0.000021251	27.222283274	0.000021251
WA	-14.106769177	0.000021861	27.107988472	0.000021861
EWA	-15.116919577	0.000021619	27.152988717	0.000021619

Note: The intercept estimates represent the difference in prediction errors between the tested models and benchmarks. Negative intercepts indicate that tested models outperform benchmarks.

These outcomes indicate that all evaluated models significantly enhance predictive accuracy over the naive Random Walk, with dynamic weighting schemes such as EWA proving especially valuable during volatile economic conditions.

Comparisons against the AR(3) model further reinforced the robustness of our selected nowcasting methodologies. All individual and aggregated models demonstrated statistically significant superiority, according to the GW test:

- Wald statistics remained notably large, exceeding 23 across all scenarios, indicative of robust model superiority.
- P-values consistently fell far below the standard significance level (all below 0.0001), indicating enhanced nowcasting performance compared to the AR(3) benchmark.
- Again, EN emerged as the top-performing individual method, with an intercept estimate of -10.64, whereas among aggregations, EWA maintained its leadership in nowcast improvement, reflected by the lowest intercept estimate (-11.27).

This consistent improvement against an established econometric benchmark underlines the efficacy of penalized regression approaches and adaptive weighting mechanisms in capturing dynamic macroeconomic relationships and structural shifts.

The DFM comparison, frequently adopted in macroeconomic nowcasting literature due to its capacity for dimensionality reduction and common factor extraction, yielded similarly strong support for our modeling strategies. Specifically:

- Wald test statistics uniformly exceeded 23, strongly rejecting the null hypothesis of equivalent predictive accuracy between tested models and DFM.
- Low p-values (below 0.0001 in all cases) provided definitive evidence of nowcasting improvement.
- The best individual models again included EN, Ridge, and PLSR, each showing intercept estimates close to -14. Among combination strategies, EWA exhibited the highest predictive enhancement (intercept -15.12), reiterating its adaptive strengths across diverse economic regimes.

These empirical findings substantiate that both individual MCS-selected models and dynamic aggregations significantly surpass the predictive accuracy of standard factor-based methods, reinforcing their potential for robust macroeconomic nowcasting.

The empirical evidence from the Giacomini-White tests highlights several important implications for macroeconomic nowcasting:

- Models selected through rigorous MCS procedures consistently deliver statistically significant nowcasting improvements across benchmarks, validating their empirical relevance.
- Adaptive weighting methodologies, particularly EWA, substantially enhance predictive performance in turbulent or regime-switching periods, ensuring robustness and responsiveness.
- Simple aggregation methods (SA), despite their straightforward nature, still provide stable and significant improvements over benchmarks and remain valuable as reliable nowcasting baselines.

These findings underscore the critical importance of conditional predictive ability testing and adaptive aggregation strategies for effective macroeconomic nowcasting.

These empirical outcomes highlight how conditional predictive ability testing and adaptive aggregation techniques systematically differentiate nowcasting models regarding predictive accuracy. The Giacomini-White test results consistently indicate statistically significant improvements in the predictive accuracy of individual and aggregated models compared to traditional benchmark models typically employed in macroeconomic nowcasting. These findings reinforce the practical value of dynamic aggregation methods (especially EWA) for policymakers, particularly under uncertain economic conditions where adaptability in nowcasting methods is crucial.

3.2 Discussion and Implications

This section aims to contextualize the empirical evidence discussed in Section 3.1 and prepares the ground for the subsequent methodological considerations and policy insights. Specifically, we investigate multiple sub-periods, including Pre-COVID, COVID, and Post-COVID, applying a block bootstrap procedure to enhance interval coverage even under sudden macroeconomic shifts. Our objective here is twofold: first, to distill the central quantitative findings arising from assessments of predictive accuracy, interval-based uncertainty, feature relevance, and nowcast-combination strategies; second, to underscore the significance of these observations for subsequent model evaluations and policy-oriented applications.

From a methodological viewpoint, our empirical results highlight several points of interest. Penalized linear models (particularly Ridge and Elastic Net) exhibit relatively low and stable prediction errors across diverse macroeconomic phases. That suggests regularization can effectively harness key explanatory signals without overfitting, even under heightened volatility. Dimensionality-reduction methods (notably Principal Component Regression and Partial Least Squares Regression) display more pronounced fluctuations around crises such as COVID-19, albeit with phases of commendable robustness. Meanwhile, ensemble learning and neural architectures prove competitive during tranquil periods yet appear more susceptible to abrupt structural shifts.

We further validate these findings through the Giacomini-White test, which confirms the statistical significance of the observed accuracy enhancements over classic benchmarks. The test outcome thus reinforces the reliability of our models under varying macroeconomic scenarios.

Parallel to these performance metrics, the analysis of prediction intervals sheds light on each model's approach to managing uncertainty. While some of our techniques produce tighter intervals in stable phases yet encounter challenges with extreme shocks, others adopt a more conservative stance with broader intervals, achieving more robust coverage in turbulent settings. Moreover, exploring feature importance reveals several consistently influential indicators (e.g., industrial production, trade-related variables, and labor market proxies) whose relative weighting can shape the scale of prediction intervals during volatility spikes.

The investigation of combination strategies, including Simple Average, Weighted Average, and Exponentially Weighted Average (EWA), complements these empirical results. EWA, in particular, demonstrates heightened adaptability in volatile environments by favoring models with strong current performance, thereby achieving either competitive or superior accuracy. Moreover, this dynamic weighting framework enhances transparency by

clarifying the individual model's impact on the aggregated prediction.

In the following subsections, we build on these findings to analyze:

- how the proposed techniques compare against established benchmarks under distinct macroeconomic regimes,
- the reliability and interpretability of the estimated prediction intervals,
- the efficacy of static and dynamically weighted nowcasting methods, specifically focusing on how EWA can offer additional insights into model contributions.

These perspectives inform broader debates on macroeconomic nowcasting and policy-making, highlighting the potential of flexible model combinations to mitigate the limitations of single-model approaches and reinforce predictive resilience during economic disruptions.

3.2.1 Comparative Nowcasting Performance across Macroeconomic Regimes: Evaluating Penalized, Dimensionality Reduction, Ensemble, and Neural Models

In light of the descriptive evidence presented in Section 3.1.1, this subsection provides a comparative evaluation of nowcasting model families (penalized linear, dimensionality reduction-based, ensemble learning, and neural network models) across different macroeconomic regimes, offering valuable insights into the relative robustness, stability, and adaptability of various methodological frameworks.

The primary evaluation metric remains the Root Mean Square Forecast Error (RMSFE), which is considered both in absolute terms and relative to traditional benchmarks (Random Walk, AR(3), and Dynamic Factor Model).

Penalized linear approaches, specifically Ridge and Elastic Net (Hastie et al., 2009; Hoerl & Kennard, 1970; Kim & Swanson, 2018; Smeekes & Wijler, 2018; Uematsu & Tanaka, 2019; Zou & Hastie, 2005), consistently demonstrated superior predictive performance relative to traditional benchmarks. The regularization mechanisms inherent to these methods effectively control model complexity, enhancing stability during periods of high volatility, such as the COVID-19 crisis. Conversely, LASSO (Tibshirani, 1996) exhibited noticeable vulnerability due to its stringent variable selection criterion, which proved detrimental when structural economic shifts occurred. Key insights from penalized linear models include:

- *Stability and robustness*: Ridge and Elastic Net consistently maintained low and stable prediction errors across regimes, highlighting their resilience to macroeconomic volatility.
- *Relative superiority*: These methods systematically outperformed traditional benchmarks, reinforcing their value for real-time macroeconomic nowcasting applications.
- *LASSO vulnerability*: Despite widespread use in variable selection, LASSO significantly deteriorated under structural shocks, suggesting limited applicability in volatile economic contexts.

Dimensionality reduction approaches (Principal Component Regression and Partial Least Squares Regression) (Fuentes et al., 2015; Groen & Kapetanios, 2016; Jolliffe,

1982; Jolliffe & Cadima, 2016; Krämer & Sugiyama, 2011; Massy, 1965; Mehmood et al., 2012; Wold, 1985) provided valuable yet regime-dependent nowcasting capabilities. PCR showed remarkable resilience during extreme volatility periods, effectively capturing critical macroeconomic information while filtering out noise. However, its performance deteriorated significantly during the Post-COVID regime transitions. Conversely, PLSR experienced significant deterioration during the COVID crisis but demonstrated a notable recovery in prediction accuracy during the Post-COVID normalization phase, although it is still not reaching pre-crisis accuracy levels. Crucial observations include:

- *Adaptability to economic regimes*: PCR's strong performance in crisis periods contrasted sharply with its reduced adaptability during recovery phases.
- *Complementary dynamics*: The opposing behavior of PCR and PLSR during and after crises suggests strategic complementarities between these two methodologies.
- *Overall competitiveness*: Both methods offered substantial improvements over benchmarks, albeit PCR exhibited greater volatility across regimes, suggesting the necessity of regime-specific model selection.

Ensemble learning models (Random Forest and eXtreme Gradient Boosting) (Biau & Scornet, 2016; Borup et al., 2023; Breiman, 2001a; T. Chen & Guestrin, 2016; Coulombe et al., 2022; Goulet Coulombe, 2024; Probst et al., 2019; Soybilgen & Yazgan, 2021) displayed mixed prediction effectiveness. While theoretically well-suited to capturing complex nonlinear relationships, both models showed pronounced vulnerability during abrupt economic disruptions. XGB, however, demonstrated better recovery capabilities post-crisis than RF, reflecting relative strengths in managing transition dynamics. Notable insights include:

- *Nonlinearity handling limitations*: Despite inherent flexibility, ensemble methods underperformed in highly volatile periods, highlighting their sensitivity to rapid structural economic changes.
- *Complexity trade-offs*: High error rates observed during crises underscored the need for dynamic calibration and continuous retraining to maximize their predictive potential. Nevertheless, despite partial recovery, ensemble models remained generally inferior in overall prediction performance compared to penalized linear and dimensionality-reduction-based methods.

Neural network models, particularly Gated Recurrent Units (GRU), outperformed traditional Multilayer Perceptrons (MLP), reflecting the advantage of explicitly modeling temporal dependencies (Almosova & Andresen, 2023; Cho et al., 2014; Chung et al., 2014; Coulombe et al., 2022; Hewamalage et al., 2021; Hornik et al., 1989; Jozefowicz et al., 2015). Nonetheless, both neural architectures experienced significant prediction degradation during the COVID crisis, underscoring challenges related to data availability, quality, and model calibration under conditions of economic stress. Strategic observations include:

- *Architectural differences*: GRU's systematic superiority emphasizes the necessity of recurrent structures to capture temporal dynamics effectively.

- *Economic shock sensitivity*: Both neural models demonstrated substantial vulnerability during severe macroeconomic disturbances, highlighting their role primarily as complementary rather than stand-alone prediction tools, especially in volatile environments. Moreover, although neural models consistently outperformed Random Forest, they generally underperformed compared to penalized linear and dimensionality reduction-based models.

The collective evidence highlights a nuanced trade-off between model adaptability (the ability to rapidly and flexibly adjust predictions to unexpected and lasting structural changes in the economy) and stability (the capacity to consistently deliver accurate predictions without being overly influenced by short-term volatility or transient noise in economic data) (Clark & McCracken, 2010; Elliott & Timmermann, 2016; Giraitis et al., 2013; Pesaran et al., 2013; Rossi, 2021; Timmermann, 2006). Highly flexible models (ensemble and neural networks) exhibited sensitivity to extreme shocks, whereas penalized linear approaches offered greater robustness with potential trade-offs in short-term responsiveness to structural changes. Strategic recommendations derived from this comparative analysis include:

- *Adaptive stability trade-offs*: Optimal model selection necessitates careful balancing between stability (penalized linear) and adaptability (ensemble, neural networks).
- *Regime-informed model selection*: Differential performance across economic regimes underlined the importance of dynamically adjusting model choice and calibration based on prevailing macroeconomic conditions.
- *Benchmarking as strategic guidance*: Consistent outperformance of advanced econometric models relative to traditional benchmarks (RW, AR(3), DFM) underscored the empirical value of these methodological innovations.

From this overview, regularization-based models—particularly Ridge and Elastic Net—demonstrated solid and robust performance for quarterly nowcasting of Singapore's GDP growth across different macroeconomic regimes. Dimensionality reduction models (PCR and PLSR) generally achieved good accuracy but exhibited higher variability across subperiods: PCR notably performed exceptionally well during the COVID crisis yet experienced marked deterioration in the subsequent Post-COVID period, whereas PLSR recovered steadily after initial setbacks. Conversely, ensemble learning and neural network models showed less consistent advantages: while competitive in stable periods, their prediction accuracy significantly deteriorated during turbulent phases (COVID). Among neural networks, GRU consistently outperformed MLP, providing relatively greater stability and resilience during periods of macroeconomic uncertainty.

3.2.2 Analysis and Discussion of Prediction Uncertainty

The prediction intervals presented in Section 3.1.2 shed light on each model's ability to quantify uncertainty under evolving economic regimes. In this subsection, we examine how interval width and coverage interact with model structure and macroeconomic volatility, thereby highlighting strengths and limitations of different approaches.

Models producing narrower intervals may occasionally fail to encompass actual values during extreme disruptions, potentially underestimating latent macroeconomic risks. By contrast, methods that generate broader ranges may be more successful in covering tail events, yet they often incur a loss of precision. The pivotal objective is thus to calibrate prediction intervals to reflect prediction uncertainty while avoiding excessive spread.

1. **Penalized linear models.** Our results suggest that penalized linear techniques, especially Elastic Net, tend to produce relatively tight intervals, indicative of an efficient regularization strategy that mitigates overfitting. Nevertheless, during acute volatility episodes (e.g., the COVID-19 crisis), Elastic Net may understate uncertainty if the structural break is abrupt and substantial. During COVID-19 specifically, Ridge exhibited notably narrower intervals compared to LASSO, underscoring distinct behaviors within penalized linear methods under heightened volatility conditions. Elastic Net's dual penalty scheme (combining L1 and L2 regularization) appears to support more stable intervals than LASSO or Ridge, balancing variable selection against the risk of overreacting to transient shocks. During abrupt regime changes, Elastic Net intervals expand sufficiently to account for higher uncertainty but not to the point of rendering the nowcast signal uninformative. This feature is valuable for operational nowcasting, where interval interpretability is critical. Penalized linear models show a marked yet relatively moderate expansion of intervals in crisis contexts.

2. **Dimensionality reduction-based models.** Although PCR and PLSR yield, on average, comparable width measures, their interval patterns diverge under shifting conditions. PCR presents narrower confidence bands during and after the COVID-19 shock, indicating greater resilience in uncertainty recovery compared to PLSR, whose intervals remain comparatively wider in the Post-COVID period. Such contrasting behaviors suggest that the choice between these two techniques should reflect the risk tolerance for crisis versus recovery phases. Dimensionality reduction approaches, such as Principal Component Regression and Partial Least Squares Regression, reveal larger jumps in uncertainty during crises, although they display distinct recovery trajectories.

3. **Ensemble learning models.** Ensemble methods demonstrate notable interval fluctuations. Random Forest yields the widest intervals, particularly in the Post-COVID phase, without consistently capturing realized values. Random Forest's broad intervals may reflect a heightened reaction to anomalies, occasionally overshooting the true volatility range. Conversely, XGBoost exhibits a more controlled expansion, aided by its intrinsic form of shrinkage, thus preserving steadier prediction bands even under macroeconomic shocks. Random Forest's broader intervals might seem "safer" at first glance, but occasionally fail to align with realized outcomes, hinting at potential oversensitivity to transient or extreme observations. XGBoost, on the other hand, retains tighter intervals through an additional layer of internal shrinkage. This balancing act between flexibility and constraint can be advantageous in volatile settings where moderate expansions of intervals suffice to capture most disturbances without an unwarranted loss of precision.

4. **Neural network models.** Neural networks exhibit divergent behaviors while capturing nonlinear patterns. MLP and GRU networks expand their prediction intervals markedly during turbulent periods, though the MLP's band growth is often more pronounced. Extensive intervals may hinder practical decision-making, while heightened caution is preferable to missing significant shocks entirely. GRU's less extreme interval widening and its recurrent structure may provide a more controlled uncertainty assessment in real-time contexts. GRU intervals remain marginally broader in the Post-COVID phase than MLP's ones, indicating a continued cautious approach during the recovery period.

An important observation is the lack of perfect alignment between point nowcast errors (e.g., Root Mean Squared Forecast Error) and interval widths (Christoffersen, 1998). Some techniques achieve respectable accuracy on average, yet produce intervals that can be disproportionately large. Others exhibit narrower intervals but occasionally exclude realized values, suggesting an underestimation of volatility. As a result, nowcast users should evaluate not only the magnitude of errors but also the alignment of predicted uncertainty with observed economic variations.

In a structurally evolving economy, models may underestimate uncertainty if they rely on historical patterns that fail to anticipate sudden shocks (B. E. Hansen, 2006). LASSO, for instance, can be slow to adjust interval scales during the early phase of an unexpected crisis if key covariates are penalized too aggressively. Analogously, MLP networks may show inertia in re-scaling intervals due to stable training scenarios. Detecting early signs of regime shifts is thus pivotal to ensuring accurate uncertainty quantification.

We derived intervals via a block bootstrap procedure that respects temporal dependencies (see Politis & Romano, 1994, for the original stationary bootstrap framework), yet each model reacts distinctively to resampled data. More rigid regularization methods, including Ridge, tend to exhibit minor expansions. By contrast, more flexible approaches, such as MLP or Random Forest, can respond sharply to outlying resampled blocks, yielding abrupt shifts in interval width. This heterogeneity underscores why similar resampling techniques may produce dissimilar uncertainty profiles across models.

This analysis underscores distinct uncertainty profiles across model families, revealing their different sensitivity and adaptability to macroeconomic volatility. Penalized linear methods, particularly Elastic Net, and ensemble methods such as XGBoost consistently achieve balanced prediction intervals, effectively responding to regime shifts without excessive uncertainty. Dimensionality reduction models (PCR, PLSR) and neural networks (MLP, GRU) exhibit contrasting behaviors, emphasizing that interval width alone does not always indicate coverage reliability. Ultimately, these insights stress the critical need to carefully calibrate interval estimation methods according to the specific macroeconomic conditions and forecasting objectives.

3.2.3 Feature Importance Dynamics across Macroeconomic Regimes

The interpretability analyses previously described in Section 3.1.3 provide critical insights into the economic dynamics captured by different modeling frameworks. They highlight the robustness and variability of predictor importance across distinct macroeconomic regimes. Below, we systematically discuss key interpretive observations derived from these analyses.

All modeling approaches—penalized linear models, dimensionality-reduction methods,

ensemble learning, and neural networks—consistently emphasize a common subset of macroeconomic indicators:

- Industrial production index,
- Gross operating surplus, and
- Key labor market variables, notably employee compensation and unit labor cost.

This convergence underscores the structural significance of these variables, suggesting their persistent informational value in tracking Singapore's short-term GDP dynamics, irrespective of economic stability or turmoil.

Although a core set of predictors remains influential across periods, notable shifts in variable importance are observed in response to macroeconomic disturbances, particularly the COVID-19 pandemic. Specifically:

- *Penalized linear models* increased emphasis on labor market and production indicators (e.g., employee compensation and unit labor cost), reflecting immediate labor-market disruptions (Cajner et al., 2020).
- *Dimensionality reduction methods*, especially Partial Least Squares Regression (PLSR), elevated the importance of air transportation metrics (Hakim & Merkert, 2016), capturing the abrupt contraction and subsequent recovery of global passenger and cargo traffic.

These findings indicate the capacity of models to internally recalibrate their predictor selection in real-time, effectively capturing structural shifts driven by sudden macroeconomic events.

Modeling methodologies exhibit distinct behaviors concerning feature selection stability and adaptability:

- *Penalized linear models* (LASSO, Ridge, Elastic Net) show relatively stable predictor hierarchies, with controlled fluctuations attributable to their regularization mechanisms.
- In contrast, *ensemble methods* (Random Forest (Breiman, 2001a), XGBoost) and *neural networks* (Rudin, 2019) demonstrate substantial variability in their feature importance rankings, significantly altering their internal structures in response to short-term fluctuations. This adaptability enhances responsiveness to regime changes but diminishes the immediate interpretability of economic drivers.
- *Principal Component Regression* (PCR) consistently emphasizes production and fiscal indicators, whereas *Partial Least Squares Regression* (PLSR) prioritizes variables related to trade and logistics. Such methodological nuances underline the differential focus inherent in dimensionality-reduction strategies.

Our analyses highlight labor-market variables as critical conduits for economic shocks, particularly during crises. Neural network models and penalized regressions notably increased the relevance of employee compensation and unit labor cost during COVID-19

(Cajner et al., 2020), reinforcing the labor market's centrality in macroeconomic adjustments. Concurrently, air transport variables (air cargo loaded and air passenger arrivals) emerged as robust leading indicators. Their consistent prominence, especially in dimensionality-reduction and ensemble methods, underscores Singapore's sensitivity to global trade dynamics and international mobility (Hakim & Merkert, 2016), which are vital to its economic performance.

The Spearman correlation analysis empirically reinforces the robustness of predictor selection. Variables frequently identified by various models correlate strongly and consistently with GDP. The industrial production index, in particular, consistently displays high and stable correlations, substantiating its role as a primary macroeconomic indicator.

Our results highlight a critical interpretative trade-off:

- *Stability in feature selection*, typical of linear and penalized methods, facilitates constructing coherent long-term economic narratives.
- *Greater adaptability*, characteristic of ensemble and neural network approaches, captures nonlinearities and sudden economic disruptions, albeit at the cost of interpretative clarity.

Thus, the optimal methodological choice depends on the prevailing macroeconomic context and the specific analytical or operational objectives, whether focused on immediate policy response or longer-term strategic forecasting.

The quarterly evolution analysis of the industrial production index (illustrated through the GRU model's Integrated Gradients (Sundararajan et al., 2017)) vividly demonstrates its heightened importance during the COVID-19 crisis. This dynamic interpretability underscores industrial production as a reliable barometer of economic resilience and subsequent recovery.

Across all evaluated models, the industrial production index change consistently emerges as the most influential predictor, underscoring its fundamental role in GDP nowcasting. Other consistently important variables include air cargo loaded change, gross operating surplus change, and air passenger arrivals change, reflecting their sensitivity to both cyclical economic fluctuations and structural shifts.

These results emphasizes how different modeling frameworks dynamically recalibrate feature importance in response to evolving macroeconomic conditions. These empirical observations illustrate the inherent methodological trade-off between stability in predictor selection—favoring interpretability—and flexibility in adapting to abrupt economic shifts. A deeper exploration of the econometric and practical implications of these findings, particularly concerning model selection for forecasting and policy purposes, will be provided in the next subsection.

3.2.4 Forecast Combination: Robustness, Adaptability, and Implications

The findings presented in Section 3.1.4 illustrate that combining forecasts from multiple models—via aggregation methods such as Simple Average (SA), Weighted Average (WA), and Exponential Weighted Average (EWA)—can substantially enhance accuracy and resilience in nowcasting tasks (Bates & Granger, 1969; Claeskens et al., 2016; Smith & Wallis, 2009). Notably, the initial selection of models via the Model Confidence Set (MCS)

procedure ensures that only methods with robust statistical predictive power contribute to these aggregations, strengthening the reliability of combined forecasts. This result holds particular significance in the presence of structural breaks or sudden economic disruptions, corroborating established best practices in the forecasting literature (Diebold & Lopez, 1996; Hendry & Clements, 2004).

A key rationale for forecast combination lies in the attenuation of biases and uncertainties exhibited by individual approaches. Notably, the performance of single-model strategies deteriorated sharply during the COVID-19 crisis, characterized by unprecedented volatility (Foroni et al., 2020), whereas aggregated forecasts retained robust predictive power. This result underscores how methodologically diverse ensemble strategies offer greater protection against structural shocks (Makridakis et al., 2020). By leveraging each model's unique strengths, combined methods reduce the risk of failure when any single model proves vulnerable to unexpected shifts in macroeconomic dynamics.

EWA emerges as a particularly effective solution among the three examined techniques in volatile contexts. Its dynamic adjustment of weights, based on recent performance (see Figure 3.14), enables swift adaptation to regime changes and breaks in the data-generating process (Koop & Korobilis, 2012). This adaptability was quantitatively evident during the COVID-19 crisis, with EWA achieving notably lower RMSFE (0.610) compared to SA (1.963) and WA (1.433), as shown in Table 3.8. Additionally, EWA transparently illustrates each model's contribution in real-time, enabling policymakers to identify dominant modeling approaches during rapidly evolving economic conditions (Raftery et al., 2010).

The three aggregation methods exhibit distinct trade-offs between simplicity and reactivity:

- *Simple Average* assigns equal weights without active reweighting, serving as a robust and stable benchmark. It occasionally surpasses WA only in the post-COVID period (RMSFE: 1.520 vs. 1.567, Table 3.8), thus avoiding over-reliance on potentially outdated historical performance metrics (Smith & Wallis, 2009).

- *Weighted Average* employs static weights derived from long-run predictive performance. This approach stabilizes weighting but faces challenges if economic conditions shift abruptly, as historical performance may quickly become obsolete.

- *Exponentially Weighted Average* dynamically recalibrates weights, excelling in rapidly changing environments. Nevertheless, it requires vigilant parameter tuning and close monitoring to guard against excessive weight volatility, thus demanding greater implementation effort (Timmermann, 2006).

Detailed analyses of WA and EWA weight trajectories (Tables 3.10, 3.11, Figure 3.14) reveal that distinct models temporarily dominate under varying macroeconomic regimes. Principal Component Regression (PCR) notably excelled during peak COVID-19 volatility; Elastic Net (EN) gained prominence during post-crisis stabilization; and the Gated Recurrent Unit (GRU) demonstrated robust performance in pre-crisis and dynamic recovery phases. These complementary strengths suggest that no single approach uniformly outperforms across all scenarios, highlighting the strategic advantage of forecast combinations (Batchelor & Dua, 1995).

Diagnostic evaluations (Shapiro–Wilk and Ljung–Box tests, Table 3.13) confirm that aggregated forecasts do not introduce problematic biases or autocorrelation structures.

Indeed, EWA's dynamic approach yields residuals with notably fewer autocorrelation issues (Ljung-Box p-value: 0.093) compared to SA (0.038) and WA (0.047), substantiating its statistical soundness. These characteristics, alongside lower RMSFE values, underscore the empirical robustness of dynamic aggregation methods (Diebold & Lopez, 1996), particularly under abrupt economic changes.

From an econometric standpoint, results align with the forecasting literature consensus advocating forecast combinations as essential tools in macroeconomic nowcasting (Genre et al., 2013; Koop & Korobilis, 2012). Predictive ability tests (e.g., Giacomini-White) reinforce that aggregated methods consistently outperform traditional benchmarks—Random Walk, AR(3), and Dynamic Factor Models—across overall performance and distinct subperiod analyses (Tables 3.9). This evidence further supports their utility in informed policymaking and economic decision-making under heightened uncertainty (Claessens et al., 2012; Diebold & Mariano, 1995).

The choice of aggregation method depends on context-specific balances between simplicity and responsiveness:

- *Exponentially Weighted Average* is recommended in high-volatility contexts, offering rapid adaptation and interpretability (Koop & Korobilis, 2012).
- *Simple Average* remains appropriate for stable environments or when operational simplicity and lower computational complexity are prioritized.
- *Weighted Average* provides a balanced intermediate solution when historical model performance maintains predictive relevance.

Each strategy significantly enhances resilience against shocks while providing interpretative insights into model effectiveness.

These findings substantiate that careful implementation of forecast combination methods—validated through rigorous empirical evaluation and robust statistical diagnostics—significantly enhances predictive accuracy, resilience, and operational transparency. By leveraging complementarities among diverse modeling frameworks, combined forecasting approaches prove highly effective for real-time decision-making, particularly under frequent structural changes or high uncertainty (Hendry & Clements, 2004; Timmermann, 2006).

3.2.5 Methodological and policymaking implications for macroeconomic nowcasting

The empirical findings in this study offer meaningful insights into methodological practices and policy-oriented applications within macroeconomic nowcasting. This section synthesizes key implications from our analysis, addressing methodological robustness, model selection strategies, interpretability, uncertainty quantification, and strategic policymaking.

Methodological Implications

- *Superiority over traditional benchmarks.* Across all macroeconomic regimes analyzed, advanced methods (penalized regression, dimensionality reduction, ensemble, and neural networks) consistently outperformed traditional benchmarks, including Random Walk, AR(3), and Dynamic Factor Models (DFM). Furthermore, formal predictive-ability tests (e.g., Diebold & Mariano, 1995; Giacomini & White, 2006)

indicate that these models deliver statistically significant improvements over standard references even under regime shifts, thereby reinforcing the reliability of the recommended modeling frameworks. This superiority validates investments in more sophisticated econometric techniques for enhancing nowcast accuracy (Makridakis & Hibon, 2000; Makridakis et al., 2020).

- *Regularization under volatility.* Penalized regression models, notably Ridge and Elastic Net (De Mol et al., 2008; Hastie et al., 2009; Hoerl & Kennard, 1970; Kim & Swanson, 2018; Smeekes & Wijler, 2018; Tibshirani, 1996; Uematsu & Tanaka, 2019; Zou & Hastie, 2005), demonstrated consistently robust performance during severe economic disruptions, such as the COVID-19 shock. These regularized approaches proved essential in maintaining prediction accuracy and reliability by effectively controlling overfitting and managing extensive macroeconomic datasets. Nevertheless, Elastic Net, characterized by narrower prediction intervals, occasionally risked insufficient coverage of extreme outcomes, highlighting a critical trade-off between precision and comprehensive coverage.

- *Dimensionality reduction and structural adaptability.* Principal Component Regression (PCR) revealed significant resilience during crisis periods, despite performance deterioration post-crisis. Conversely, Partial Least Squares Regression (PLSR) experienced immediate accuracy reductions during shocks, though it recovered more swiftly post-crisis. This differential performance underscores the necessity of aligning dimensionality-reduction methods with specific nowcasting objectives, such as initial shock resistance versus accelerated post-crisis normalization (Fuentes et al., 2015; Jolliffe, 1982; Krämer & Sugiyama, 2011; Wold, 1985).

- *Flexibility–interpretability trade-off.* Neural network models, particularly recurrent architectures such as Gated Recurrent Unit (Almosova & Andresen, 2023; Cho et al., 2014; Hewamalage et al., 2021), exhibited superior adaptability to nonlinearities and complex temporal structures compared to Multilayer Perceptrons and linear or ensemble methods (XGBoost, Random Forest). Due to its gating mechanism, Gated Recurrent Unit can store and update relevant historical signals more effectively (Chung et al., 2014), which proves advantageous when abrupt economic shocks occur. However, this enhanced flexibility comes at the cost of reduced interpretability. Therefore, in such circumstances it is necessary to adopt techniques that allow greater transparency and comprehensibility of the results obtained, for instance Explainable Artificial Intelligence (XAI) techniques (e.g., Integrated gradients, time-series variants of SHapley Additive exPlanations).

- *Prediction intervals and uncertainty management.* Penalized regression methods generated comparatively narrower prediction intervals than neural or ensemble approaches, which produced broader and sometimes overly conservative intervals during periods of economic turbulence. Nevertheless, while these methods excel at capturing complex nonlinearities, an overly cautious approach to interval width in neural and ensemble frameworks may occasionally compromise the practical utility of nowcasts, highlighting the importance of carefully balancing coverage and precision in operational settings. In addition, adopting a block bootstrap or related resampling approaches attuned to temporal dependence (Li, 2021; Politis & Romano,

1994) can enhance the reliability of these intervals, particularly in the presence of structural breaks. Future model selection should explicitly consider policymakers' risk tolerance and contextual requirements, balancing interval precision and adequate coverage. Furthermore, incorporating scoring rules that penalize interval under- and over-coverage (Gneiting & Katzfuss, 2014; Gneiting & Raftery, 2007) may refine interval evaluation and increase their practical utility.

- *Feature importance and interpretability.* Key economic indicators—industrial production, gross operating surplus, labor market variables, and air transport metrics—consistently emerged as central predictors across models, emphasizing their structural relevance. Employing XAI techniques, such as Integrated Gradients (Molnar, 2022; Sundararajan et al., 2017), maintained interpretability even within complex models, facilitating a clearer understanding of underlying economic drivers.

- *Forecast combination and model stability.* Our results reinforce established evidence that forecast aggregation—through Simple Average, Weighted Average, and Exponentially Weighted Average—effectively mitigates individual model biases (Armstrong, 2001; Atiya, 2020; Bates & Granger, 1969; Clemen, 1989; Smith & Wallis, 2009; Timmermann, 2006). The preliminary use of the Model Confidence Set procedure further improved predictive stability, particularly under volatile macroeconomic conditions (Aiolfi & Timmermann, 2006; Genre et al., 2013; P. R. Hansen et al., 2011).

Policy-making Implications

- *Enhanced decision-making in crisis contexts.* Forecast combination approaches, particularly EWA methods, proved crucial in delivering robust and timely nowcasts during high-volatility phases. Policymakers can rely on these forecasts to formulate rapid, targeted counter-cyclical monetary or fiscal interventions (Clemen, 1989; Hendry & Clements, 2004).

- *Prioritizing key economic indicators.* Given Singapore's status as a small, open, and highly industrialized economy (International Monetary Fund, 2024), consistent evidence underscores the predictive centrality of core macroeconomic indicators—industrial production, labor costs, and air cargo volumes. Policymakers should prioritize frequent and granular monitoring of these variables, as timely and accurate data are essential for proactive and effective policy responses (Bańbura & Rünstler, 2011; Hakim & Merkert, 2016).

- *Modular forecasting and policy dashboards.* Given the variability in model performance across macroeconomic regimes, policymakers should adopt integrated forecasting dashboards that aggregate predictions from diverse methodological frameworks (Bok et al., 2018). Such dashboards enable clear differentiation between temporary sector-specific disruptions and broader systemic crises, optimizing policy targeting and responsiveness through frequent recalibration (Genre et al., 2013; Koop & Korobilis, 2012).

- *Transparency and stakeholder communication.* Employing interpretable methodologies and communicating feature importance insights can significantly enhance public and market confidence in policy decisions. Transparent explanation of forecast drivers strengthens institutional credibility, reduces market uncertainty (Blinder et al., 2008), and facilitates coordinated economic actions among stakeholders.

- *Preparedness for future economic shocks.* The COVID-19 experience highlights the necessity for forecasting frameworks capable of rapid adaptation to structural changes. Strategically employing dynamic weighting mechanisms such as EWA allows rapid recalibration, ensuring forecasting models remain resilient and actionable in future unexpected economic disturbances. Robust and timely forecasts facilitate more efficient allocation of resources and mitigate potential impacts of recessions or financial instability (Barbaglia et al., 2023). These findings provide robust methodological guidance and actionable insights for policymakers and practitioners involved in macroeconomic nowcasting, promoting enhanced stability, transparency, and adaptability in economic forecasting and policymaking processes.

Chapter 4

Conclusions

4.1 Summary of Contributions and Main Findings

This dissertation systematically evaluates and advances macroeconomic nowcasting methodologies by integrating state-of-the-art econometric techniques with machine and deep learning approaches. It offers policymakers enhanced predictive tools for navigating economic uncertainty.

Unlike existing studies, which primarily focus on single-model families, our research comparatively assesses penalized regression (LASSO, Ridge, Elastic Net), dimensionality-reduction techniques (PCR, PLSR), ensemble algorithms (Random Forest, XGBoost), and neural network architectures (MLP, GRU).

An essential innovation lies in developing a robust nowcasting pipeline explicitly designed to mitigate look-ahead bias, a frequent source of overly optimistic performance estimates in macroeconomic nowcasting. This methodological framework combines iterative expanding and rolling windows with Bayesian hyperparameter optimization and pruning, effectively balancing dynamic adaptability and predictive stability.

Additionally, we introduce systematic quantification of predictive uncertainty through block bootstrap intervals, responding directly to recent calls for robust prediction methodologies in the presence of structural instabilities and economic shocks. This advancement provides critical insights into the reliability and confidence of macroeconomic nowcasts.

Another significant contribution involves enhancing interpretability in complex macroeconomic ML models using Explainable AI techniques, specifically Integrated Gradients. This innovation fills critical interpretability gaps in recent literature, enabling policymakers to derive actionable insights from traditionally opaque models.

Empirically, our study leverages Singapore's characteristics—small, open, advanced, and highly diversified—as an informative testbed. By analyzing model performances across distinct macroeconomic regimes (Pre-COVID, COVID, Post-COVID), we empirically validate the superior adaptability of advanced econometric and ML models relative to traditional benchmarks, directly addressing recent methodological concerns on robustness and adaptability. Our findings indicate substantial predictive improvements—typically ranging between 40% and 60% in terms of RMSFE reductions—over widely established benchmarks such as Random Walk, AR(3), and Dynamic Factor Models. These results underline the empirical validity and practical advantage of our integrated methodological framework.

Furthermore, we introduce a nowcasting combination approach integrating Model Confidence Set procedures with dynamic averaging strategies (Exponentially Weighted Average), significantly enhancing aggregate nowcast accuracy. This method provides critical insights regarding dynamic model performances under shifting economic conditions.

Finally, extensive diagnostic checks (Shapiro-Wilk, Ljung-Box) and formal predictive-ability tests (Diebold-Mariano, Giacomini-White) underscore the adopted approach's methodological rigor and empirical robustness, fully aligning with the rigorous analytical

standards required by leading nowcasting literature.

Our work contributes substantially to macroeconomic nowcasting by introducing methodological innovations, robust interpretability frameworks, rigorous uncertainty quantification, extensive empirical validation, and practical nowcasting solutions tailored for effective policy formulation.

4.2 Methodological and Policy-making Implications

Leveraging the empirical insights detailed in Chapter 3, we synthesize key implications for methodological practices and policy-oriented applications in macroeconomic nowcasting, explicitly addressing issues of methodological robustness, model selection strategies, interpretability, uncertainty quantification, and strategic policymaking.

4.2.1 Methodological Implications

- *Superiority over traditional benchmarks:* Across all macroeconomic regimes analyzed, advanced methods (penalized regression, dimensionality reduction, ensemble, and neural networks) consistently outperformed traditional benchmarks, including Random Walk, AR(3), and Dynamic Factor Models (DFM). Furthermore, formal predictive-ability tests (e.g., Diebold & Mariano, 1995; Giacomini & White, 2006) indicate that these models deliver statistically significant improvements over standard references even under regime shifts, thereby reinforcing the reliability of the recommended modeling frameworks. This superiority validates investments in more sophisticated econometric techniques for enhancing nowcast accuracy (Makridakis & Hibon, 2000; Makridakis et al., 2020).

- *Regularization under volatility:* Penalized regression models, notably Ridge and Elastic Net (De Mol et al., 2008; Hastie et al., 2009; Hoerl & Kennard, 1970; Kim & Swanson, 2018; Smeekes & Wijler, 2018; Tibshirani, 1996; Uematsu & Tanaka, 2019; Zou & Hastie, 2005), demonstrated consistently robust performance during severe economic disruptions, such as the COVID-19 shock. These regularized approaches proved essential in maintaining prediction accuracy and reliability by effectively controlling overfitting and managing extensive macroeconomic datasets. Nevertheless, Elastic Net, characterized by narrower prediction intervals, occasionally risked insufficient coverage of extreme outcomes, highlighting a critical trade-off between precision and comprehensive coverage.

- *Dimensionality reduction and structural adaptability:* Principal Component Regression (PCR) revealed significant resilience during crisis periods, despite performance deterioration post-crisis. Conversely, Partial Least Squares Regression (PLSR) experienced immediate accuracy reductions during shocks, though it recovered more swiftly post-crisis. This differential performance underscores the necessity of aligning dimensionality-reduction methods with specific nowcasting objectives, such as initial shock resistance versus accelerated post-crisis normalization (Fuentes et al., 2015; Jolliffe, 1982; Krämer & Sugiyama, 2011; Wold, 1985).

- *Flexibility–interpretability trade-off:* Neural network models, particularly recurrent architectures such as Gated Recurrent Unit (Almosova & Andresen, 2023; Cho et al.,

2014; Hewamalage et al., 2021), exhibited superior adaptability to nonlinearities and complex temporal structures compared to Multilayer Perceptrons and linear or ensemble methods (XGBoost, Random Forest). Due to its gating mechanism, Gated Recurrent Unit can store and update relevant historical signals more effectively (Chung et al., 2014), which proves advantageous when abrupt economic shocks occur. However, this enhanced flexibility comes at the cost of reduced interpretability. Therefore, in such circumstances it is necessary to adopt techniques that allow greater transparency and comprehensibility of the results obtained, for instance Explainable Artificial Intelligence (XAI) techniques (e.g., Integrated gradients, time-series variants of SHapley Additive exPlanations).

- *Prediction intervals and uncertainty management:* Penalized regression methods generated comparatively narrower prediction intervals than neural or ensemble approaches, which produced broader and sometimes overly conservative intervals during periods of economic turbulence. Nevertheless, while these methods excel at capturing complex nonlinearities, an overly cautious approach to interval width in neural and ensemble frameworks may occasionally compromise the practical utility of nowcasts, highlighting the importance of carefully balancing coverage and precision in operational settings. In addition, adopting a block bootstrap or related resampling approaches attuned to temporal dependence (Li, 2021; Politis & Romano, 1994) can enhance the reliability of these intervals, particularly in the presence of structural breaks. Future model selection should explicitly consider policymakers' risk tolerance and contextual requirements, balancing interval precision and adequate coverage. Furthermore, incorporating scoring rules that penalize interval under- and over-coverage (Gneiting & Katzfuss, 2014; Gneiting & Raftery, 2007) may refine interval evaluation and increase their practical utility.

- *Feature importance and interpretability:* Key economic indicators—industrial production, gross operating surplus, labor market variables, and air transport metrics—consistently emerged as central predictors across models, emphasizing their structural relevance. Employing XAI techniques, such as Integrated Gradients (Molnar, 2022; Sundararajan et al., 2017), maintained interpretability even within complex models, facilitating a clearer understanding of underlying economic drivers.

- *Forecast combination and model stability:* Our results reinforce established evidence that nowcast aggregation—through Simple Average, Weighted Average, and Exponentially Weighted Average—effectively mitigates individual model biases (Armstrong, 2001; Atiya, 2020; Bates & Granger, 1969; Clemen, 1989; Smith & Wallis, 2009; Timmermann, 2006). The preliminary use of the Model Confidence Set procedure further improved predictive stability, particularly under volatile macroeconomic conditions (Aiolfi & Timmermann, 2006; Genre et al., 2013; P. R. Hansen et al., 2011).

4.2.2 Policy-making Implications

- *Enhanced decision-making in crisis contexts:* Nowcast combination approaches, particularly EWA methods, proved crucial in delivering robust and timely nowcasts

during high-volatility phases. Policymakers can rely on these nowcasts to formulate rapid, targeted counter-cyclical monetary or fiscal interventions (Clemen, 1989; Hendry & Clements, 2004).

- *Prioritizing key economic indicators:* Given Singapore's status as a small, open, and highly industrialized economy (International Monetary Fund, 2024), consistent evidence underscores the predictive centrality of core macroeconomic indicators—industrial production, labor costs, and air cargo volumes. Policymakers should prioritize frequent and granular monitoring of these variables, as timely and accurate data are essential for proactive and effective policy responses (Bańbura & Rünstler, 2011; Hakim & Merkert, 2016).

- *Modular nowcasting and policy dashboards:* Given the variability in model performance across macroeconomic regimes, policymakers should adopt integrated nowcasting dashboards that aggregate predictions from diverse methodological frameworks (Bok et al., 2018). Such dashboards enable clear differentiation between temporary sector-specific disruptions and broader systemic crises, optimizing policy targeting and responsiveness through frequent recalibration (Genre et al., 2013; Koop & Korobilis, 2012).

- *Transparency and stakeholder communication:* Employing interpretable methodologies and communicating feature importance insights can significantly enhance public and market confidence in policy decisions. Transparent explanation of nowcast drivers strengthens institutional credibility, reduces market uncertainty (Blinder et al., 2008), and facilitates coordinated economic actions among stakeholders.

- *Preparedness for future economic shocks:* The COVID-19 experience highlights the necessity for nowcasting frameworks capable of rapid adaptation to structural changes. Strategically employing dynamic weighting mechanisms such as EWA allows rapid recalibration, ensuring nowcasting models remain resilient and actionable in future unexpected economic disturbances. Robust and timely nowcasts facilitate more efficient allocation of resources and mitigate potential impacts of recessions or financial instability (Barbaglia et al., 2023). These findings provide robust methodological guidance and actionable insights for policymakers and practitioners involved in macroeconomic nowcasting, promoting enhanced stability, transparency, and adaptability in economic nowcasting and policymaking processes.

4.3 Limitations and Future Developments

The methodologies presented significantly enhance macroeconomic nowcasting by integrating advanced econometric and machine learning approaches within a comprehensive nowcasting pipeline. Nevertheless, we explicitly acknowledge several critical limitations, clearly defining the scope of our findings and highlighting areas for further methodological refinements and empirical expansion.

4.3.1 Main Limitations

We identify the following principal limitations:

- *Data frequency and granularity constraints:* Our empirical analysis primarily relied on monthly and quarterly data, limiting the systematic exploitation of higher-frequency indicators such as daily or weekly data. Furthermore, we did not explicitly incorporate Mixed Data Sampling (MIDAS) methods, potentially constraining predictive timeliness and responsiveness to rapidly evolving economic conditions. Prior research shows mixed-frequency (MIDAS) regressions can yield sizable accuracy gains when high-frequency information is available (Andreou et al., 2013; Ghysels & Marcellino, 2016).

- *Limited temporal coverage:* Despite covering significant macroeconomic shocks —including the COVID-19 pandemic—our evaluation period (2017 Q1-2023 Q2) remained relatively short. Consequently, this restricted the comprehensive assessment of model robustness and adaptability over extended economic cycles, raising caution concerning broader long-term generalizability. Longer evaluation windows are usually required to detect forecast breakdowns and gauge stability across structural regimes (Giacomini & Rossi, 2009; Rossi, 2021).

- *Geographical generalizability constraints:* Our empirical application exclusively focused on Singapore, which is characterized by its small, open, and advanced economic structure. Therefore, generalizing our proposed methodologies to structurally and institutionally diverse economies should be approached cautiously. Cross-country nowcasting evidence indicates that model performance can deteriorate when applied to economies with different data vintages or transmission mechanisms (Bok et al., 2018; Giannone et al., 2008).

- *Operational complexity and computational feasibility:* The nowcasting pipeline integrated multiple sophisticated approaches—including penalized regression, dimensionality reduction techniques, ensemble learning methods, neural network models, sequential Bayesian optimization, extensive block-bootstrap procedures, and advanced interpretability techniques. Such methodological comprehensiveness entails significant computational complexity, posing practical challenges for real-time implementation, particularly within resource-constrained policy environments (Bok et al., 2018; Shahriari et al., 2016).

- *Sensitivity to initial variable selection:* Although we employed rigorous variable selection methods, alternative predictor sets could significantly affect nowcasting stability and predictive performance, highlighting methodological sensitivity as a critical consideration. That mirrors well-documented evidence on the importance of predictor screening in high-dimensional macroeconomic forecasting (Ng, 2013; Z. Wang et al., 2023).

- *Reliability of predictive intervals under extreme events:* Despite robust uncertainty quantification via block-bootstrap methods, uncertainties persist regarding predictive interval reliability during unprecedented economic shocks. Additionally, our study did not explicitly account for uncertainties stemming from measurement errors,

subsequent data revisions, or the general quality of input data. Recent work shows that conventional predictive intervals often under-cover during crises such as COVID-19 (B. E. Hansen, 2006; Li, 2021).

4.3.2 Future Methodological and Empirical Developments

To systematically address these limitations and further advance macroeconomic nowcasting methodologies, we propose several future research avenues:

- *Leveraging alternative and granular data sources:* Integrating nontraditional datasets —such as social media sentiment analysis, financial news-based uncertainty indices, granular transactional data, satellite imagery, high-frequency mobility indicators (e.g., Google Mobility), and web-scraped economic indicators—could markedly improve nowcast accuracy and responsiveness, thus providing a richer informational basis for timely policy interventions (Bok et al., 2018; Coulombe et al., 2022).

- *Explicit integration of macro-financial risk indicators:* Future methodological advancements should explicitly integrate macro-financial stress indicators and systemic risk measures within the nowcasting framework. Such integration would enhance predictive capacity by proactively capturing and managing potential spillovers from financial markets to real economic activities.

- *Sensitivity analyses and feature interaction assessments:* Detailed sensitivity analyses focusing on the interactive effects among predictor variables could provide valuable insights into underlying macroeconomic dynamics, enhancing predictive robustness and interpretive clarity.

- *Extension to multi-step nowcasting horizons:* Future research should systematically evaluate model performance over extended nowcasting horizons (e.g., two-step-ahead nowcasts). Such assessments would strengthen operational applicability and test our methodological pipeline's robustness and practical utility for medium-term economic nowcasting scenarios.

- *Regional comparative analysis and generalizability assessments:* Systematically extending our proposed methodology to Southeast Asian economies lacking sophisticated GDP nowcasting models—such as Brunei Darussalam, Laos, Myanmar, and Timor-Leste—would substantially enhance geographical generalizability. At the same time, several regional economies have recently developed GDP nowcasting frameworks (e.g., Indonesia (Alifatussaadah et al., 2019), Malaysia (Jasni et al., 2022; Shabli et al., 2023), Hong Kong, Thailand, the Philippines, and Japan); however, these studies typically differ in scope, complexity, and forecasting objectives compared to our proposed pipeline. DBS Bank's nowcasting model for Singapore, China, India, and Indonesia focuses primarily on year-over-year (YoY) forecasts using an autoregressive framework with high-frequency proxies, rather than providing quarter-over-quarter (QoQ) nowcasts and comprehensive uncertainty quantification. Explicit comparative analyses between our integrated forecasting pipeline and existing regional models would yield meaningful insights, particularly regarding relative predictive performance, robustness, uncertainty management, and interpretability.

- *Explicit integration of high-frequency data via MIDAS models:* Future studies should explicitly incorporate Mixed Data Sampling (MIDAS) methods, enabling systematic integration of daily or weekly indicators with quarterly GDP nowcasts. That would significantly enhance predictive responsiveness, timeliness, and accuracy (e.g., see Andreou et al., 2013; Ghysels & Marcellino, 2016).

- *Exploratory hybrid econometric-deep learning frameworks:* While recent research indicates limited utility of large language models (LLMs) and Transformer architectures for direct time-series forecasting tasks (Tan et al., 2024), carefully designed hybrid models that combine structural econometric approaches with selected deep-learning components (potentially attention-based architectures) remain a promising—but still exploratory—methodological avenue. Future research should rigorously investigate under which specific conditions, contexts, or model configurations such hybridizations might provide tangible advantages, ensuring that meaningful gains in forecasting performance and interpretability justify any added complexity.

- *Comparative assessment between frequentist and Bayesian machine-learning approaches:* Explicit comparative analyses of predictive performance and uncertainty quantification between frequentist and Bayesian machine-learning frameworks would yield valuable methodological insights, informing optimal methodological choices for macroeconomic nowcasting.

- *Automated real-time nowcasting with dynamic interpretability:* Designing fully automated nowcasting dashboards capable of continuous real-time updates and enhanced dynamic interpretability—via advanced Explainable AI (XAI) techniques such as dynamic SHAP values and adaptive Integrated Gradients—would greatly enhance transparency, interpretability, and practical usability for policymakers (Molnar, 2022; Sundararajan et al., 2017).

- *Advanced robustness and structural adaptability assessments:* Comprehensive Monte Carlo simulations and extensive scenario analyses would further validate nowcasting stability, adaptability, and predictive reliability across diverse macroeconomic conditions, including extreme and unprecedented economic shocks.

- *Investigating determinants of predictive-interval characteristics:* Further research could systematically examine factors affecting predictive intervals, including interval-width variations, peak behavior, and plateau formations. Such analyses would substantially improve uncertainty-quantification reliability, robustness, and interpretability.

- *Expansion of the nowcasting framework to additional macroeconomic indicators as target variables:* Future methodological applications could systematically extend our nowcasting pipeline to other critical macroeconomic indicators—including inflation rates and industrial production indices—significantly broadening our research's methodological relevance and practical policy impact.

4.4 Concluding Remarks

Addressing the explicitly identified limitations and rigorously pursuing the proposed future developments could enhance macroeconomic nowcasting methodologies' robustness, transparency, interpretability, and practical relevance. In particular, integrating high-frequency data, alternative data sources, and advanced hybrid modeling techniques will allow policymakers and practitioners to navigate economic uncertainties more precisely and confidently. Moreover, expanding methodological applications to additional economic indicators and conducting comparative analyses across diverse economic contexts could significantly broaden nowcasting frameworks' generalizability and empirical validity. Ultimately, such comprehensive methodological advancements promise to deliver timely, reliable, and actionable economic insights, thereby improving the quality and effectiveness of macroeconomic policy and decision-making processes in rapidly evolving and increasingly complex global economic environments.

Appendices

Appendix A. Variables and Descriptive Statistics Used in the GDP Nowcasting Models

Table A.1: Table A.1: Variables included in the nowcasting models for Singapore's quarterly GDP growth (categorization by economic indicator type)

Variable name	Original name	Category
business entities formation change	Formation Of All Business Entities By Industry, Quarterly (Number)	Feature - Leading Indicator
business entities cessation change	Cessation Of All Business Entities By Industry, Quarterly (Number)	Feature - Leading Indicator
composite leading index change	Composite Leading Index (2015=100), Quarterly (Index)	Feature - Leading Indicator
public contracts awarded change	Contracts Awarded By Sector And Development Type, Quartely - Total Public Sector (Million Dollars)	Feature - Leading Indicator
private contracts awarded change	Contracts Awarded By Sector And Development Type, Quartely - Total Private Sector (Million Dollars)	Feature - Leading Indicator
landed properties supply change	Supply Of Private Residential Properties In The Pipeline By Development Status (End Of Period), Quarter - Total Landed Properties (Number Of Units)	Feature - Leading Indicator
non landed properties supply change	Supply Of Private Residential Properties In The Pipeline By Development Status (End Of Period), Quarter - Total Non-Landed Properties (Number Of Units)	Feature - Leading Indicator
business expectations manufacturing	Business Expectations Of The Manufacturing Sector - Forecast For Next 6 Months, Quarterly (In Percentage Terms)	Feature - Leading Indicator
business outlook services change	Business Expectations For The Services Sector - General Business Outlook For The Next 6 Months, Net Weighted Balance, Quarterly (In Percentage Terms)	Feature - Leading Indicator
services employment forecast change	Business Expectations For The Services Sector - Employment Forecast For The Next Quarter, Net Weighted Balance, Quarterly (In Percentage Terms)	Feature - Leading Indicator
services revenue forecast change	Business Expectations For The Services Sector - Operating Revenue Forecast For The Next Quarter, Net Weighted Balance, Quarterly (In Percentage Terms)	Feature - Leading Indicator
services outlook weighted change	Business Expectations For The Services Sector - General Business Outlook For Next 6 Months, Weighted % Of Up, Same, Down, Quarterly (In Percentage Terms)	Feature - Leading Indicator

Continued on the next page

Variable name	Original name	Category
services revenue weighted change	Business Expectations For The Services Sector - Operating Revenue Forecast Next Quarter, Weighted % Of Up, Same, Down, Quarterly (In Percentage Terms)	Feature - Leading Indicator
services employment weighted change	Business Expectations For The Services Sector - Employment Forecast Next Quarter, Weighted Percentages Of Up, Same, Down, Quarterly (In Percentage Terms)	Feature - Leading Indicator
bop quarterly change	Singapore's Balance Of Payments, (BPM6 Format), Quarterly (Million dollars)	Feature - Coincident Indicator
household net worth change	Household Sector Balance Sheet (End Of Period), Quarterly - Household Net Worth (Million Dollars)	Feature - Coincident Indicator
unit labour cost change	Unit Labour Cost Index (2015 = 100), Quarterly (Index)	Feature - Coincident Indicator
acceleration of net employment flow	Acceleration of Net Employment Flow	Feature - Coincident Indicator
value added per worker change	Changes In Value Added Per Worker In Chained (2015) Dollars, By Industry (SSIC 2020), Quarterly (Per Cent)	Feature - Coincident Indicator
quartely recruitment rate change	Average Quartely Recruitment Rate, Quarterly (Per Cent)	Feature - Coincident Indicator
quartely resignation rate change	Average Quartely Resignation Rate, Quarterly (Per Cent)	Feature - Coincident Indicator
employee compensation change	Compensation Of Employees By Industry At Current Prices, Quarterly (Million Dollars)	Feature - Coincident Indicator
gfcf current prices change	Gross Fixed Capital Formation At Current Prices, Quarterly (Million Dollars)	Feature - Coincident Indicator
gross operating surplus change	Gross Operating Surplus By Industry At Current Prices, Quarterly (Million Dollars)	Feature - Coincident Indicator
taxes less subsidies on prod change	Other Taxes Less Subsidies On Production By Industry At Current Prices, Quarterly (Million Dollars)	Feature - Coincident Indicator
personal saving rate change	Personal Disposable Income At Current Prices, Quarterly - Personal Saving Rate (Per Cent)	Feature - Coincident Indicator
consumer price index change	Consumer Price Index (CPI), 2019 As Base Year, Quarterly (Index)	Feature - Coincident Indicator
hdb resale price index change	Housing And Development Board Resale Price Index (1Q2009 = 100), Quarterly (Index)	Feature - Coincident Indicator
property price index change	Property Price Index By Type (4th Quarter 1998 = 100), Quarterly (Index)	Feature - Coincident Indicator

Continued on the next page

Variable name	Original name	Category
public progress payments change	Progress Payments Certified By Sector And Development Type, Quartely - Total Public (Million Dollars)	Feature - Coincident Indicator
private progress payments change	Progress Payments Certified By Sector And Development Type, Quartely - Total Private Sector (Million Dollars)	Feature - Coincident Indicator
private properties available change	Available And Vacant Private Residential Properties (End Of Period), Quarterly - All Types Private Residential Properties Available (Number Of Units)	Feature - Coincident Indicator
private properties vacant change	Available And Vacant Private Residential Properties (End Of Period), Quarterly - All Types Private Residential Properties Vacant (Number Of Units)	Feature - Coincident Indicator
electricity generation change	Electricity Generation, Quartely (Gigawatt Hours)	Feature - Coincident Indicator
piped gas sales change	Piped Gas Sales, Quarterly (Million Kilowatt Hours)	Feature - Coincident Indicator
finance loans advances change	Loans And Advances Of Finance Companies (End Of Period), Quartely (Million Dollars)	Feature - Coincident Indicator
industrial production index change	Index Of Industrial Production (2019 = 100), Quarterly (Index)	Feature - Coincident Indicator
fnb services index change	Food & Beverage Services Index, (2017 = 100), In Chained Volume Terms, Quarterly (Index)	Feature - Coincident Indicator
retail sales index change	Retail Sales Index, (2017 = 100), In Chained Volume Terms, Quarterly (Index)	Feature - Coincident Indicator
domestic trade index change	Domestic Wholesale Trade Index, (2017 = 100), In Chained Volume Terms, Quarterly (Index)	Feature - Coincident Indicator
foreign trade index change	Foreign Wholesale Trade Index, (2017 = 100), In Chained Volume Terms, Quarterly (Index)	Feature - Coincident Indicator
aircraft arrivals departures change	Civil Aircraft Arrivals And Departures, Passengers, Air Cargo Tonnage, Direct And Transhipment Tonnage And Mail, Changi Airport, Quartely - Total Aircraft Arrivals And Departures (Number)	Feature - Coincident Indicator
air passenger arrivals change	Air Passenger Arrivals By Region-Country Of Embarkation, Quartely - Total Passengers (Number)	Feature - Coincident Indicator
air mail discharge change	Air Cargo Discharged By Region-Country Of Origin, Quartely - Total Mail (Tonne)	Feature - Coincident Indicator
air cargo discharge change	Air Cargo Discharged By Region-Country Of Origin, Quartely (Tonne)	Feature - Coincident Indicator

Continued on the next page

Variable name	Original name	Category
air cargo loaded change	Air Cargo Loaded By Region-Country Of Destination, Quartely (Tonne)	Feature - Coincident Indicator
motor vehicle population change	Motor Vehicle Population By Type Of Vehicle (End Of Period), Quartely (Number)	Feature - Coincident Indicator
new vehicle registration change	New Registration Of Motor Vehicles Under Vehicle Quota System, Quartely - Total New Motor Vehicles Registered (Number)	Feature - Coincident Indicator
vehicle deregistration change	Motor Vehicles De-Registered Under Vehicle Quota System, Quartely (Number)	Feature - Coincident Indicator
vehicle population quota change	Motor Vehicle Population Under Vehicle Quota System, (As At End Of Period), Quartely (Number)	Feature - Coincident Indicator
vessel arrivals change	Sea Cargo And Shipping Statistics, Quartely - Vessel Arrivals (Number)	Feature - Coincident Indicator
vessel tonnage change	Sea Cargo And Shipping Statistics, Quartely - Vessel Arrivals - Shipping Tonnage (Thousand Gross Tonnes)	Feature - Coincident Indicator
sea total cargo change	Sea Cargo And Shipping Statistics, Quartely - Total Cargo (Thousand Tonnes)	Feature - Coincident Indicator
sea container throughput change	Sea Cargo And Shipping Statistics, Quartely - Total Container Throughput (Thousand Twenty-Foot Equivalent Units)	Feature - Coincident Indicator
bunker sales change	Sea Cargo And Shipping Statistics, Quartely - Bunker Sales (Thousand Tonnes)	Feature - Coincident Indicator
merchandise imports change	Merchandise Trade By Commodity Section, (At Current Prices), Quartely - Total Merchandise Imports, At Current Prices (Thousand Dollars)	Feature - Coincident Indicator
merchandise exports change	Merchandise Trade By Commodity Section, (At Current Prices), Quartely - Total Domestic Exports At Current Prices (Thousand Dollars)	Feature - Coincident Indicator
merchandise reexports change	Merchandise Trade By Commodity Section, (At Current Prices), Quartely - Total Re-Exports At Current Prices (Thousand Dollars)	Feature - Coincident Indicator
import price index change	Import Price Index, Base Year 2018 = 100, Quartely - Overall Items (Index)	Feature - Exogenous Variable
export price index change	Export Price Index, Base Year 2018 = 100, Quartely - Overall Items (Index)	Feature - Exogenous Variable
domestic supply price change	Domestic Supply Price Index, Base Year 2018 = 100, Quartely - Domestic Supply Price Index - Overall Items (Index)	Feature - Exogenous Variable

Continued on the next page

Variable name	Original name	Category
manufactured price index change	Singapore Manufactured Products Price Index, Base Year 2018 = 100, Quartely - Singapore Manufactured Products Price Index - Overall Items (Index)	Feature - Exogenous Variable
government debt change	Government Debt, (End Of Period), Quarterly (Million Dollars)	Feature - Exogenous Variable
government bond yield 5yr change	Current Banks Interest Rates (End Of Period), Quartely - Government Securities - 5-Year Bond Yield (Per Cent Per Annum)	Feature - Exogenous Variable
government bond yield 2yr change	Current Banks Interest Rates (End Of Period), Quartely - Government Securities - 2-Year Bond Yield (Per Cent Per Annum)	Feature - Exogenous Variable
government bond yield 1yr change	Current Banks Interest Rates (End Of Period), Quartely -Government Securities - 1-Year Bond Yield (Per Cent Per Annum)	Feature - Exogenous Variable
exchange rate usd change	Exchange Rates, (Average For Period), Quartely - US Dollar (Singapore Dollar Per US Dollar)	Feature - Exogenous Variable
exchange rate gbp change	Exchange Rates, (Average For Period), Quartely - Sterling Pound (Singapore Dollar Per Pound Sterling)	Feature - Exogenous Variable
exchange rate chf change	Exchange Rates, (Average For Period), Quartely - Swiss Franc (Singapore Dollar Per Swiss Franc)	Feature - Exogenous Variable
exchange rate jpy change	Exchange Rates, (Average For Period), Quartely - Japanese Yen (Singapore Dollar Per 100 Japanese Yen)	Feature - Exogenous Variable
exchange rate myr change	Exchange Rates, (Average For Period), Quartely - Malaysian Ringgit (Singapore Dollar Per Malaysian Ringgit)	Feature - Exogenous Variable
exchange rate hkd change	Exchange Rates, (Average For Period), Quartely - Hong Kong Dollar (Singapore Dollar Per Hong Kong Dollar)	Feature - Exogenous Variable
exchange rate aud change	Exchange Rates, (Average For Period), Quartely - Australian Dollar (Singapore Dollar Per Australian Dollar)	Feature - Exogenous Variable
exchange rate eur change	Exchange Rates, (Average For Period), Quartely - Euro (Singapore Dollar Per Euro)	Feature - Exogenous Variable
exchange rate cny change	Exchange Rates, (Average For Period), Quartely - Renminbi (Singapore Dollar Per Renminbi)	Feature - Exogenous Variable
foreign reserves change	Official Foreign Reserves (End Of Period), Quartely (Million Dollars)	Feature - Exogenous Variable
temp max change	Air Temperature And Sunshine, Relative Humidity And Rainfall - Air Temperature Means Daily Maximum (Degree Celsius)	Feature - Exogenous Variable
temp min change	Air Temperature And Sunshine, Relative Humidity And Rainfall - Air Temperature Means Daily Minimum (Degree Celsius)	Feature - Exogenous Variable

Continued on the next page

Variable name	Original name	Category
total rainfall change	Air Temperature And Sunshine, Relative Humidity And Rainfall - Total Rainfall (Millimetre)	Feature - Exogenous Variable
daily sunshine change	Air Temperature And Sunshine, Relative Humidity And Rainfall - Bright Sunshine Daily Mean (Hour)	Feature - Exogenous Variable
mean humidity change	Air Temperature And Sunshine, Relative Humidity And Rainfall - 24 Hours Mean Relative Humidity (Per Cent)	Feature - Exogenous Variable
liquor releases change	Duty-Paid Releases Of Liquors, Quartely (Litre)	Feature - Other Variable
tobacco releases change	Duty-Paid Releases Of Tobacco, Petroleum, Compressed Natural Gas And Motor Vehicles, Quartely - Tobacco (Kilograms)	Feature - Other Variable
cpf due contributions change	Central Provident Fund Contributions, Withdrawals And Amount Due To Members, Quartely - Amount Due To Members (Million Dollars)	Feature - Other Variable
currency in circulation change	Currency In Circulation (End Of Period), Quartely - Gross Circulation (Million Dollars)	Feature - Other Variable
finance assets liabilities change	Assets And Liabilities Of Finance Companies (End Of Period), Quartely - Assets(=Liabilities) Of Finance Companies (Million Dollars)	Feature - Other Variable
pawnshop pledges received change	Pledges At Pawnshops, Quartely - Number Of Pledges Received (Number)	Feature - Other Variable
pawnshop pledges redeemed change	Pledges At Pawnshops, Quartely - Number Of Pledges Redeemed (Number)	Feature - Other Variable
pawnshop loans given change	Pledges At Pawnshops, Quartely - Amount Of Loans Given Out (Million Dollars)	Feature - Other Variable
pawnshop loans redeemed change	Pledges At Pawnshops, Quartely - Amount Of Loans Redeemed Including Interest (Million Dollars)	Feature - Other Variable
registry ships count change	Sea Cargo And Shipping Statistics, Quartely - Singapore Registry Of Ships, End Of Period (Number)	Feature - Other Variable
registry ships tonnage change	Sea Cargo And Shipping Statistics, Quartely - Singapore Registry Of Ships (End Of Period) ('000 GT - Thousand Gross Tonnes)	Feature - Other Variable
live births change	Live-Births By Birth Order, Quarterly	Feature - Other Variable
deaths by ethnicity change	Deaths By Ethnic Group And Sex, Quartely	Feature - Other Variable
hospital admissions change	Admissions To Public Sector Hospitals, Quartely	Feature - Other Variable
lag1 deflated gdp change	Lag of Order 1 of Singapore's GDP growth	Lagged Variable - Feature

Continued on the next page

Variable name	Original name	Category
lag2 deflated gdp change	Lag of Order 2 of Singapore's GDP growth	Lagged Variable - Feature
lag3 deflated gdp change	Lag of Order 3 of Singapore's GDP growth	Lagged Variable - Feature
lag4 deflated gdp change	Lag of Order 4 of Singapore's GDP growth	Lagged Variable - Feature
seasonality q1	Dummy variable tracking of seasonality effect	Dummy Variable - Feature
seasonality q2	Dummy variable tracking of seasonality effect	Dummy Variable - Feature
seasonality q3	Dummy variable tracking of seasonality effect	Dummy Variable - Feature
neg shock dummy	Dummy variable tracking past negative shocks of Singapore's GDP quarter over quarter growth	Dummy Variable - Feature
pos shock dummy	Dummy variable tracking past positive shocks of Singapore's GDP quarter over quarter growth	Dummy Variable - Feature
deflated gdp change	Deflated Singapore's GDP quarter over quarter growth	Target Variable

Table A.2: Table A.2: Descriptive statistics of variables employed in GDP growth nowcasting for Singapore (1990 Q1 –2023 Q2)

Variable	Mean	SD	Min	Q1	Q2	Q3	Max	Skewness	Kurtosis
bop_quarterly_change	−387.854 675 2	4261.953 956	−49 145	−83.912 199 3	−27.632 830 4	70.260 384 85	2361.275 755	−11.425 003 55	131.661 715 2
unit_labour_cost_change	0.613 493 083	8.874 130 893	−27.727 272 73	−6.568 417 664	1.013 525 082	6.631 683 725	17.355 371 9	−0.126 063 224	−0.381 197 982
acceleration_of_net_employment_flow	9.017 399 902	236.694 468 3	−750	−50.241 935 49	−4.613 942 248	35.749 738 77	1700	3.488 161 268	25.558 363 31
value_added_per_worker_change	2.270 895 522	4.824 433 811	−13	0.125	1.8	4.7	19.4	−0.106 163 627	1.878 226 406
monthly_recruitment_rate_change	0.313 098 592	12.583 575 5	−36.842 105 26	−10	0	9.303 977 273	33.333 333 33	−0.005 333 43	−0.170 388 208
monthly_resignation_rate_change	−0.055 259 44	12.746 598 99	−23.529 411 76	−11.111 111 11	0	10.526 315 79	28.571 428 57	0.109 561 892	−0.976 900 118
composite_leading_index_change	0.586 963 4	2.350 141 427	−9.448 818 898	−0.671 230 441	0.475 580 344	1.745 476 974	7.304 785 894	−0.070 099 024	2.461 584 586
employee_compensation_change	1.993 777 583	9.200 423 498	−12.231 633 35	−6.655 604 016	2.533 805 821	7.527 427 754	22.200 820 44	0.313 762 723	−0.808 424 075
gfcf_current_prices_change	1.623 082 86	7.016 882 361	−27.799 064 02	−2.722 006 996	1.893 342 122	5.717 518 446	24.896 994 83	−0.112 840 434	2.326 867 898
gross_operating_surplus_change	2.117 017 751	7.512 638 937	−13.979 507 09	−3.684 206 11	1.361 217 2	6.749 015 413	22.603 491 57	0.322 347 537	−0.547 974 335
taxes_less_subsidies_on_prod_change	−2.422 435 72	133.670 710 3	−677.970 236 6	−21.366 789 2	3.883 421 718	14.392 838 6	1148.826 816	3.137 651 231	48.301 000 71
consumer_price_index_change	0.457 796 233	0.639 915 586	−1.209 926 877	0.005 658 907	0.438 015 708	0.807 985 187	2.276 703 058	0.326 830 706	0.637 289 49
import_price_index_change	−0.033 310 02	2.742 091 273	−11.598 269 55	−1.305 820 179	−0.014 834 849	1.216 235 501	10.412 397 07	−0.348 318 822	3.656 777 723
export_price_index_change	−0.350 189 543	2.576 767 469	−9.634 308 084	−1.633 381 852	−0.268 828 137	0.799 845 362	7.824 129 334	0.094 362 826	1.567 142 525
property_price_index_change	0.328 715 501	4.696 657 532	−12.489 991 99	−2.669 550 926	0.380 466 877	2.359 447 245	15.770 925 11	0.250 005 774	1.031 299 029
domestic_supply_price_change	0.104 788 373	3.795 639 587	−20.162 440 89	−1.612 827 847	0.265 450 935	2.180 952 486	11.750 796 78	−1.060 854 289	5.948 642 557
manufactured_price_index_change	−0.249 255 158	2.973 742 6	−16.759 512 91	−1.938 462 725	−0.140 074 251	1.359 217 122	6.377 791 344	−1.056 888 203	6.264 425 634
liquor_releases_change	1.441 173 092	13.195 27	−35.013 933 56	−7.245 633 031	2.249 524 877	9.128 936 175	46.168 864 59	0.021 213 954	0.892 901 959
government_debt_change	2.426 707 258	2.105 709 554	−3.449 687 278	1.444 290 385	2.377 933 072	3.099 616 064	13.051 552 06	1.596 237 894	8.090 371 641
public_contracts_awarded_change	12.072 832 49	59.291 694 41	−66.698 393 42	−21.961 888 35	3.498 837 324	29.836 393 49	400.225 279 6	2.810 441 113	13.848 285 64
private_contracts_awarded_change	8.312 892 052	41.720 956 81	−69.523 742 39	−21.493 201 02	−1.524 659 906	24.868 829 85	160.084 076 7	1.199 261 554	1.627 754 459
public_progress_payments_change	2.396 484 479	13.859 595 91	−64.954 239 54	−4.331 322 464	3.035 643 917	8.719 685 296	79.647 394 35	0.229 991 578	10.243 769 76
private_progress_payments_change	2.164 398 553	11.336 980 85	−45.505 419 23	−4.468 883 482	1.555 718 442	8.935 350 625	65.326 218 51	0.752 289 991	8.067 802 379
private_properties_available_change	1.592 232 759	5.141 993 402	−0.072 419 053	0.540 828 485	1.068 069 247	1.557 887 44	59.383 171 82	10.877 340 17	122.522 878
private_properties_vacant_change	1.950 766 729	8.680 900 58	−19.922 430 51	−3.454 811 258	0.818 251 722	7.648 704 6	33.217 930 03	0.397 595 81	0.570 662 064
business_expectations_manufacturing	6.417 910 448	17.211 771 85	−57	−2	7	20	39	−0.851 916 101	1.385 690 736
cpf_due_contributions_change	2.069 676 288	1.131 382 139	−5.871 021 7	1.794 646 518	2.196 864 189	2.577 191 059	4.176 547 73	−2.760 962 835	17.531 099 51
currency_in_circulation_change	1.696 761 705	3.201 686 981	−12.876 255 36	0	1.080 103 595	2.895 860 174	19.688 839 85	1.176 338 861	9.986 565 81
government_bond_yield_5yr_change	2.193 732 724	24.768 561 16	−46.464 646 46	−9.700 887 393	−0.131 578 948	9.216 573 419	116.666 666 7	1.778 810 036	5.552 334 038
government_bond_yield_2yr_change	4.452 275 76	32.704 941 3	−67.441 860 47	−13.950 693 2	−1.475 347 934	17.712 096 34	150	1.389 850 095	3.886 011 299
government_bond_yield_1yr_change	3.918 579 326	33.062 792 68	−100	−13.378 059 66	0	18.571 428 57	126.470 588 2	0.611 371 902	1.952 264 035
exchange_rate_usd_change	−0.252 820 86	2.287 330 847	−4.923 088 785	−1.877 587 899	−0.443 597 568	1.112 332 5	7.258 227 111	0.655 167 564	0.609 630 642
exchange_rate_gbp_change	−0.398 657 986	3.316 769 985	−16.025 131 48	−1.840 728 469	−0.241 572 683	1.273 826 184	9.557 800 163	−0.960 220 479	4.263 060 965
exchange_rate_chf_change	0.206 826 73	3.266 219 471	−9.729 270 305	−1.616 819 844	0.080 831 797	2.250 144 67	11.444 377 79	0.160 460 911	1.697 082 89
exchange_rate_jpy_change	−0.165 501 635	4.098 385 683	−10.997 103 12	−2.411 426 896	−0.363 566 208	1.873 748 341	19.262 053 22	0.817 347 93	3.468 499 96
exchange_rate_myr_change	−0.634 455 549	2.304 754 818	−14.531 027 21	−1.588 053 517	−0.256 233 462	0.632 058 242	3.618 730 549	−2.058 295 658	9.930 552 77
exchange_rate_hkd_change	−0.255 019 922	2.312 941 655	−4.943 988 051	−1.959 493 118	−0.409 841 468	1.030 760 044	7.337 380 746	0.711 587 493	0.759 767 785
exchange_rate_aud_change	−0.326 687 086	3.645 351 56	−19.360 649 57	−2.004 432 502	−0.239 567 885	1.257 286 717	11.261 903 97	−0.665 253 55	5.301 718 722
exchange_rate_eur_change	−0.255 572 65	3.127 392 53	−10.172 955 97	−2.074 079 591	−0.160 268 324	1.633 082 652	10.850 627 99	0.069 086 819	1.442 327 403
exchange_rate_cny_change	−0.624 201 739	4.101 480 839	−33.202 971 9	−1.764 933 223	−0.284 254 551	1.029 660 184	7.682 505 717	−4.654 516 916	33.966 917 24
finance_loans_advances_change	0.734 981 564	4.361 525 136	−25.631 172 08	−1.051 710 244	0.980 724 703	3.480 693 722	9.710 162 545	−2.669 590 965	13.766 099 04

Continued on the next page

Variable	Mean	SD	Min	Q1	Q2	Q3	Max	Skewness	Kurtosis
finance_assets_liabilities_change	0.487 954 517	4.356 243 665	−28.946 611 28	−0.916 887 409	0.535 627 788	2.694 332 103	8.964 255 397	−2.972 455 314	17.450 657 7
pawnshop_pledges_received_change	0.548 312 652	6.528 806 046	−33.538 875 78	−2.639 146 593	0.547 172 897	2.452 719 963	39.693 463 86	0.799 030 402	14.417 229 85
pawnshop_loans_given_change	2.011 362 284	7.639 137 422	−28.952 470 08	−0.687 254 86	1.063 312 109	3.181 072 657	40.421 528 04	1.803 450 55	10.731 124 06
business_entities_formation_change	1.236 486 053	9.090 412 555	−28.471 104 61	−4.015 830 028	0.164 896 204	6.257 581 669	39.756 376 09	0.494 614 477	2.957 971 339
business_entities_cessation_change	2.409 815 525	18.256 496 86	−46.430 593 67	−6.866 435 642	−0.074 536 072	7.253 464 435	81.270 010 67	1.619 108 741	5.253 456 83
industrial_production_index_change	1.511 398 708	6.325 005 299	−14.221 046 32	−2.149 571 894	0.553 769 163	5.294 913 637	21.428 873 75	0.254 491 126	0.336 251 33
fnb_services_index_change	0.305 874 458	7.730 516 959	−43.362 682 97	−2.479 774 508	0.373 705 034	2.793 642 595	49.389 048 42	0.526 878 792	18.846 454 64
retail_sales_index_change	0.981 481 919	8.668 689 555	−37.598 018 05	−3.785 178 186	0.413 191 147	5.899 290 757	53.780 732 11	0.898 765 123	12.230 224 01
aircraft_arrivals_departures_change	1.881 354 371	10.040 737 11	−82.716 862 79	−0.115 661 425	1.721 195 246	3.388 619 738	39.628 166 52	−4.179 355 805	39.803 735 53
air_passenger_arrivals_change	6.003 984 419	28.425 202 66	−99.113 448 6	−1.108 301 444	2.430 834 018	5.118 626 344	185.674 196 1	3.724 287 545	22.604 590 36
air_mail_discharge_change	1.568 283 302	15.376 474 53	−48.059 946 62	−8.023 173 44	0.575 225 589	9.725 477 178	51.764 705 88	0.399 675 597	0.905 491 696
air_cargo_loaded_change	1.019 301 91	7.765 828 576	−30.751 853 17	−4.720 149 486	2.750 747 164	5.941 706 297	20.902 750 15	−0.759 665 672	1.160 536 795
new_vehicle_registration_change	3.744 418 608	36.370 585 98	−79.787 283 24	−6.945 841 404	−1.215 238 046	6.572 880 136	383.870 967 7	8.641 513 111	91.082 608 32
vehicle_deregistration_change	2.897 811 955	20.340 684 56	−71.261 540 94	−5.615 022 377	1.375 985 576	10.404 474 78	130.361 289 1	1.912 901 227	14.849 796 86
merchandise_imports_change	1.486 421 583	6.716 990 794	−20.825 078 67	−2.536 578 328	2.057 601 303	6.118 227 995	14.620 510 05	−0.571 279 65	0.785 019 906
merchandise_exports_change	1.438 532 504	7.519 382 594	−25.072 598 68	−3.611 593 927	1.553 458 455	5.701 334 767	24.199 844 29	−0.306 972 915	1.000 999 499
merchandise_reexports_change	1.928 216	6.819 794 891	−16.585 209 98	−2.748 901 021	1.369 840 661	6.601 902 315	25.756 026 55	0.105 666 303	0.771 065 973
seasonality_q1	0.253 731 343	0.436 778 486	0	0	0	0.75	1	1.144 745 027	−0.700 235 743
seasonality_q2	0.253 731 343	0.436 778 486	0	0	0	0.75	1	1.144 745 027	−0.700 235 743
seasonality_q3	0.246 268 657	0.432 453 521	0	0	0	0	1	1.191 229 786	−0.590 003 067
neg_shock_dummy	0.059 701 493	0.237 822 011	0	0	0	1	1	3.758 858 954	12.312 586 75
pos_shock_dummy	0.037 313 433	0.190 239 913	0	0	0	0	1	4.937 943 499	22.722 237 46
deflated_gdp_change	1.336 099 115	2.740 469 915	−11.230 866 35	−0.398 795 114	1.384 522 168	3.060 714 521	9.143 570 053	−0.640 084 835	2.781 360 371

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106. https://doi.org/10.1002/wics.51 (cit. on p. 62).

Adams, D. (1979). *The hitchhiker's guide to the galaxy*. Pan Books. (Cit. on p. 12).

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1–2), 31–53. https://doi.org/10.1016/j.jeconom.2005.07.017 (cit. on pp. 77, 79, 87, 142, 146).

Akarachantachote, N., Chadcham, S., & Saithanu, K. (2014). Cutoff threshold of variable importance in projection for variable selection in PLSR. *Chemometrics and Intelligent Laboratory Systems*, *138*, 42–49. https://doi.org/10.1016/j.chemolab.2014.08.008 (cit. on p. 62).

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701 (cit. on pp. 22, 35, 40, 43, 45, 49, 51, 55, 58, 60, 69).

Aliaj, T., Ciganović, M., & Tancioni, M. (2023). Nowcasting inflation with lasso-regularized vector autoregressions and mixed frequency data. *Journal of Forecasting*, *42*(3), 507–526. https://doi.org/10.1002/for.2987 (cit. on p. 59).

Alifatussaadah, A., Primariesty, A. D., Soleh, A. M., & Andriansyah. (2019). Nowcasting indonesia's gdp growth using dynamic factor model: Are fiscal data useful? *Proceedings of the 1st International Conference on Statistics and Analytics (ICSA 2019)*. https://doi.org/10.4108/eai.2-8-2019.2290478 (cit. on p. 149).

Almosova, A., & Andresen, N. (2023). Nonlinear inflation forecasting with recurrent neural networks. *Journal of Forecasting*, *42*(2), 240–259. https://doi.org/10.1002/for.2892 (cit. on pp. 46, 51, 133, 141, 145).

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134 (cit. on p. 75).

Amendola, A., Braione, M., Candila, V., & Storti, G. (2020). A model confidence set approach to the combination of multivariate volatility forecasts. *International Journal of Forecasting*, *36*(3), 873–891. https://doi.org/10.1016/j.ijforecast.2019.12.004 (cit. on p. 78).

Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business and Economic Statistics*, *31*(2), 240–251. https://doi.org/10.1080/07350015.2013.767199 (cit. on pp. 148, 150).

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Springer. https://doi.org/10.1007/978-0-306-47630-3_19 (cit. on pp. 81, 83, 142, 146).

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*(1), 197–200. https://doi.org/10.1016/j.ijforecast.2019.06.020 (cit. on pp. 83, 89, 142, 146).

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221. https://doi.org/10.1111/1468-0262.00273 (cit. on p. 28).

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317. https://doi.org/10.1016/j.jeconom.2008.08.010 (cit. on pp. 4, 6).

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2019). Evaluating multiple forecast combination techniques with an application to credit scoring. *European Journal of Operational Research*, *283*(2), 601–618. https://doi.org/10.1016/j.ejor.2019.01.052 (cit. on p. 87).

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large bayesian vector autoregressions. *Journal of Applied Econometrics*, *25*(1), 71–92. https://doi.org/10.1002/jae.1137 (cit. on pp. 4, 6, 22, 55).

Bańbura, M., & Rünstler, G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting gdp. *International Journal of Forecasting*, *27*(2), 333–346. https://doi.org/10.1016/j.ijforecast.2010.01.011 (cit. on pp. 1, 19–22, 29, 38, 142, 147).

Banca d'Italia. (2023). Exchange rate time series [Accessed: 2023-11-16]. https://tassidicambio.bancaditalia.it/terzevalute-wf-ui-web/timeSeries (cit. on p. 15).

Banerjee, A., & Marcellino, M. (2006). Are there any reliable leading indicators for us inflation and gdp growth? *International Journal of Forecasting*, *22*(1), 137–151. https://doi.org/10.1016/j.ijforecast.2005.03.005 (cit. on p. 28).

Barbaglia, L., Frattarolo, L., Onorante, L., Pericoli, F. M., Ratto, M., & Tiozzo Pezzoli, L. (2023). Testing big data in a big crisis: Nowcasting under covid-19. *International Journal of Forecasting*, *39*(4), 1548–1563. https://doi.org/10.1016/j.ijforecast.2022.10.005 (cit. on pp. 2, 143, 147).

Barro, R. J., & Ursúa, J. F. (2008). Macroeconomic crises since 1870. *Brookings Papers on Economic Activity*, *Spring 2008*, 255–335 (cit. on p. 18).

Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. https://doi.org/10.1016/j.ijpe.2016.04.013 (cit. on p. 90).

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, *41*(1), 68–75. https://doi.org/10.1287/mnsc.41.1.68 (cit. on p. 139).

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. https://doi.org/10.1057/jors.1969.103 (cit. on pp. 78–80, 83, 88, 121, 138, 142, 146).

Baumeister, C., Korobilis, D., & Lee, T. K. (2021). Energy markets and global economic conditions. *International Journal of Forecasting*, *37*(3), 1056–1076. https://doi.org/10.1016/j.ijforecast.2020.09.016 (cit. on p. 1).

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press. (Cit. on p. 5).

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50. https://doi.org/10.1257/jep.28.2.29 (cit. on pp. 59, 61).

Bergmeir, C., Hyndman, R. J., & Benítez, J. (2016). Bagging exponential smoothing methods using stl decomposition and box–cox transformation. *International Journal of Forecasting*, *32*(2), 303–312. https://doi.org/10.1016/j.ijforecast.2015.07.002 (cit. on p. 75).

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems (NIPS)*, *24*, 2546–2554 (cit. on pp. 22, 35, 38, 40, 45, 49, 51, 55, 58, 60, 69).

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7 (cit. on pp. 41–43, 133).

Blinder, A. S., Ehrmann, M., Fratzscher, M., de Haan, J., & Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, *46*(4), 910–945. https://doi.org/10.1257/jel.46.4.910 (cit. on pp. 143, 147).

Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, *132*(1), 169–194. https://doi.org/10.1016/j.jeconom.2005.01.027 (cit. on pp. 35, 36).

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, *10*, 615–643. https://doi.org/10.1146/annurev-economics-080217-053255 (cit. on pp. 1, 142, 147–149).

Borup, D. D., Christensen, B. J., Mühlbach, N. S., & Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, *39*(2), 841–868. https://doi.org/10.1016/j.ijforecast.2022.10.001 (cit. on pp. 64, 66, 67, 133).

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/BF00117832 (cit. on p. 86).

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324 (cit. on pp. 2, 3, 40–43, 64, 65, 133, 137).

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726 (cit. on p. 74).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth. (Cit. on pp. 64, 67).

Bry, G., & Boschan, C. (1971). *Cyclical analysis of time series: Selected procedures and computer programs*. National Bureau of Economic Research. (Cit. on p. 18).

Cajner, T., Crane, L. D., Decker, R. A., Hamins-Puertolas, A., & Kurz, C. (2020). Tracking labor market developments during the covid-19 pandemic: A preliminary assessment. *Federal Reserve Board Finance and Economics Discussion Series*, (2020-030). https://doi.org/10.17016/FEDS.2020.030 (cit. on pp. 1, 137, 138).

Carriero, A., Clark, T. E., & Marcellino, M. (2015a). Bayesian vars: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, *30*(1), 46–73. https://doi.org/10.1002/jae.2322 (cit. on pp. 22, 59, 61).

Carriero, A., Clark, T. E., & Marcellino, M. (2015b). Realtime nowcasting with a bayesian mixed frequency model with stochastic volatility. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(4), 837–862. https://doi.org/10.1111/rssa.12103 (cit. on p. 44).

Carriero, A., Clark, T. E., Marcellino, M., & Mertens, E. (2022). Addressing covid-19 outliers in bvars with stochastic volatility. *Review of Economics and Statistics*, *104*(4), 1–38. https://doi.org/10.1162/rest_a_01010 (cit. on pp. 25, 91).

Cerqueira, V., Torgo, L., Soares, C., & Machado, L. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, *109*(11), 1997–2028. https://doi.org/10.1007/s10994-020-05927-9 (cit. on pp. 22, 43).

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press. (Cit. on p. 86).

Chatfield, C. (1993). Calculating interval forecasts (with discussion). *Journal of Business & Economic Statistics*, *11*(2), 121–143 (cit. on pp. 52, 55).

Chen, B., & Hood, K. (2020). *Nowcasting of advance estimates of personal consumption of services in the U.S. national accounts: Individual versus forecast combination approach* (tech. rep. No. BEA Working Paper 2021-3). Bureau of Economic Analysis. Washington, DC. (Cit. on p. 81).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785 (cit. on pp. 2, 3, 40, 41, 44, 67, 68, 133).

Chinn, M., Meunier, A., & Stumpner, S. (2023). *Nowcasting world trade with machine learning: A three-step approach* [Working Paper 31419], National Bureau of Economic Research. https://doi.org/10.3386/w31419 (cit. on p. 39).

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 (cit. on pp. 3, 5, 46, 50, 133, 141, 145).

Chong, I. .-., & Jun, C. .-. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, *78*(1–2), 103–112. https://doi.org/10.1016/j.chemolab.2004.12.011 (cit. on p. 62).

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862. https://doi.org/10.2307/2527341 (cit. on p. 136).

Chu, B., & Qureshi, S. (2023). Comparing out-of-sample performance of machine learning methods to forecast U.S. gdp growth. *Computational Economics*, *62*(4), 1567–1609. https://doi.org/10.1007/s10614-022-10328-y (cit. on pp. 41, 45).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. https://arxiv.org/abs/1412.3555 (cit. on pp. 50, 133, 141, 146).

Cimadomo, J., D'Agostino, A., & Monarch, R. (2022). Nowcasting with large bayesian vector autoregressions: Real-time evidence from a global perspective. *Journal of Econometrics*, *231*(2), 500–519. https://doi.org/10.1016/j.jeconom.2022.03.002 (cit. on p. 22).

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762. https://doi.org/10.1016/j.ijforecast.2015.09.005 (cit. on p. 138).

Claessens, S., Kose, M. A., & Terrones, M. E. (2012). How do business and financial cycles interact? *Journal of International Economics*, *87*(1), 178–190. https://doi.org/10.1016/j.jinteco.2011.11.008 (cit. on pp. 17, 140).

Clark, T. E., & McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, *50*(2), 363–395. https://doi.org/10.1111/j.1468-2354.2009.00533.x (cit. on pp. 19–21).

Clark, T. E., & McCracken, M. W. (2010). Averaging forecasts from vars with uncertain instabilities. *Journal of Applied Econometrics*, *25*(1), 5–29. https://doi.org/10.1002/jae.1052 (cit. on pp. 85, 134).

Clark, T. E., & Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, *30*(4), 551–575. https://doi.org/10.1002/jae.2388 (cit. on p. 53).

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5 (cit. on pp. 80, 82, 89, 90, 142, 146, 147).

Clements, M. P., & Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics*, *11*(5), 475–494. https://doi.org/10.1002/(SICI)1099-1255(199608)11:5<475::AID-JAE2>3.0.CO;2-N (cit. on pp. 25, 53, 54).

Clements, M. P., & Hendry, D. F. (1998). *Forecasting economic time series* [Book, multiple chapters. For reference, see also subsequent Journal of Econometrics publications.]. Cambridge University Press. (Cit. on p. 19).

Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. MIT Press. (Cit. on pp. 59, 61).

Coulombe, P. G., Leroux, M., Stevanovič, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920–964. https://doi.org/10.1002/jae.2910 (cit. on pp. 2, 5, 6, 133, 149).

DBS Group Research. (2025). Dbs gdp nowcasting for asia. https://www.dbs.com.sg/corporate/research-and-insights/insights/macro-analytics/GDP/gdp.html (cit. on p. 4).

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*(2), 318–328. https://doi.org/10.1016/j.jeconom.2008.04.004 (cit. on pp. 37, 141, 145).

de Rooij, S., van Erven, T., Grünwald, P., & Koolen, W. (2014). Follow the leader if you can, hedge if you must [AdaHedge algorithm discussion]. *Journal of Machine Learning Research*, *15*(1), 1281–1316 (cit. on p. 85).

Devaine, M., Gaillard, P., Goude, Y., & Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts: A review of the sequential aggregation of experts for time series. *Machine Learning*, *90*(2), 231–260. https://doi.org/10.1007/s10994-012-5317-8 (cit. on pp. 85, 88, 89).

Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination. In G. S. Maddala & C. R. Rao (Eds.), *Handbook of statistics (vol. 14)* (pp. 241–268). Elsevier. https://doi.org/10.1016/S0169-7161(96)14010-X (cit. on pp. 139, 140).

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599 (cit. on pp. 88, 92, 128, 140, 145).

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, *164*(1), 188–205. https://doi.org/10.1016/j.jeconom.2011.02.012 (cit. on pp. 28, 30).

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*(4), 1014–1024. https://doi.org/10.1162/REST_a_00227 (cit. on pp. 28, 30).

Elliott, G., & Timmermann, A. (2016). Forecast combinations. In *Economic forecasting* (pp. 310–344). Princeton University Press. (Cit. on p. 134).

Exterkate, P., Groenen, P. J. F., Heij, C., & van Dijk, D. (2016). Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, *32*(3), 736–753 (cit. on pp. 32, 34).

Foroni, C., Marcellino, M., & Stevanović, D. (2020). *Forecasting the covid-19 recession and recovery: Lessons from the financial crisis* (tech. rep. No. ECB Working Paper No. 2468). European Central Bank. (Cit. on pp. 1–3, 139).

Freeborough, W., & van Zyl, T. (2022). Investigating explainability methods in recurrent neural network architectures for financial time series data. *Applied Sciences*, *12*(3), 1427. https://doi.org/10.3390/app12031427 (cit. on p. 71).

Fuentes, J., Poncela, P., & Rodríguez, J. (2015). Sparse partial least squares in time series for macroeconomic forecasting. *Journal of Applied Econometrics*, *30*(4), 576–595. https://doi.org/10.1002/jae.2384 (cit. on pp. 62, 63, 132, 141, 145).

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. *Advances in Neural Information Processing Systems 29 (NIPS)*, 1019–1027 (cit. on p. 50).

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2010). *Combining the forecasts in the ECB SPF: Can anything beat the simple average?* (Tech. rep. No. Working Paper No. 1277). European Central Bank. (Cit. on p. 81).

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121. https://doi.org/10.1016/j.ijforecast.2012.06.004 (cit. on pp. 83, 87, 89, 90, 140, 142, 146, 147).

Ghysels, E., & Marcellino, M. (2016). The econometric analysis of mixed-frequency data sampling. *Journal of Econometrics*, *193*(2), 377–395. https://doi.org/10.1016/j.jeconom.2016.01.002 (cit. on pp. 148, 150).

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: How to get the most out of returns data sampled at different frequencies. *Journal of Econometrics*, *131*(1–2), 59–95. https://doi.org/10.1016/j.jeconom.2005.01.004 (cit. on p. 16).

Giacomini, R., & Rossi, B. (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies*, *76*(2), 669–705 (cit. on pp. 94, 148).

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x (cit. on pp. 92, 128, 140, 145).

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451. https://doi.org/10.1162/REST_a_00483 (cit. on p. 22).

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437. https://doi.org/10.3982/ECTA19524 (cit. on pp. 6, 25, 28, 37, 38).

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676. https://doi.org/10.1016/j.jmoneco.2008.05.010 (cit. on pp. 1, 19–21, 148).

Giraitis, L., Kapetanios, G., & Price, S. (2013). Adaptive forecasting in the presence of recent and ongoing structural change. *Journal of Econometrics*, *177*(2), 153–170. https://doi.org/10.1016/j.jeconom.2013.04.003 (cit. on pp. 21, 53, 134).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *15*, 315–323 (cit. on pp. 47, 48, 51).

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831 (cit. on pp. 142, 146).

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437 (cit. on pp. 142, 146).

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley. (Cit. on p. 21).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press. (Cit. on p. 46).

Goulet Coulombe, P. (2024). The macroeconomy as a random forest. *Journal of Applied Econometrics*, *39*(3), 401–421. https://doi.org/10.1002/jae.3030 (cit. on p. 133).

Grigoletto, M. (1998). Bootstrap prediction intervals for autoregressions: Some alternatives. *International Journal of Forecasting*, *14*(4), 447–460 (cit. on p. 53).

Groen, J. J. J., & Kapetanios, G. (2016). Revisiting useful approaches to data-rich macroeconomic forecasting. *Computational Statistics & Data Analysis*, *100*, 221–239. https://doi.org/10.1016/j.csda.2015.06.002 (cit. on pp. 39, 132).

Gustafsson, P., Herbst, E., Keller, M., & Pfarrhofer, M. (2023). Bayesian optimization of hyperparameters from noisy marginal likelihood estimates. *Journal of Applied Econometrics*, *38*(4), 577–595. https://doi.org/10.1002/jae.3012 (cit. on pp. 22, 24, 32).

Hakim, M. M., & Merkert, R. (2016). The causal relationship between air transport and economic growth: Empirical evidence from south asia. *Journal of Transport Geography*, *56*, 120–127. https://doi.org/10.1016/j.jtrangeo.2016.09.006 (cit. on pp. 137, 138, 142, 147).

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. (Cit. on pp. 28, 57, 60).

Hansen, B. E. (2006). Interval forecasts and parameter uncertainty. *Journal of Econometrics*, *135*(1–2), 377–398. https://doi.org/10.1016/j.jeconom.2005.07.008 (cit. on pp. 136, 149).

Hansen, B. E. (2008). Least squares model averaging. *Econometrica*, *76*(2), 365–379. https://doi.org/10.1111/j.1468-0262.2008.00832.x (cit. on pp. 81, 87).

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. https://doi.org/10.3982/ECTA5771 (cit. on pp. 12, 76, 78, 83, 90, 121, 142, 146).

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365–380. https://doi.org/10.1198/073500104000000563 (cit. on pp. 76, 77, 93).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer. (Cit. on pp. 33, 57, 60, 67, 132, 141, 145).

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Chapman & Hall/CRC Press. (Cit. on pp. 7, 31, 32, 34, 45, 57, 59).

Hausmann, R., Pritchett, L., & Rodrik, D. (2005). Growth accelerations. *Journal of Economic Growth*, *10*(4), 303–329 (cit. on p. 17).

Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. In D. F. Hendry & M. P. Clements (Eds.), *Forecasting in the presence of structural breaks and model uncertainty* (pp. 31–50). Oxford University Press. (Cit. on pp. 139, 140, 142, 147).

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.001 (cit. on pp. 2, 3, 5, 46, 51, 55, 133, 141, 146).

Higgins, P. (2014). Gdpnow: A model for gdp "nowcasting". *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2580350 (cit. on p. 1).

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634 (cit. on pp. 30, 32, 36, 57, 59, 61, 132, 141, 145).

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press. (Cit. on p. 21).

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8 (cit. on pp. 46–48, 51, 133).

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd) [31 May 2021 Edition]. OTexts. https://OTexts.com/fpp3/ (cit. on pp. 18, 19, 44, 81).

Inoue, A., & Rossi, B. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, *30*(3), 432–453. https://doi.org/10.1080/07350015.2012.707592 (cit. on pp. 18–20, 92).

International Monetary Fund. (2024). *Singapore: 2024 article iv consultation—press release; staff report; and statement by the executive director for singapore* (IMF Country Report No. 24/255) (Accessed: 2025-05-10). International Monetary Fund. Washington, D.C. https://www.imf.org/en/Publications/CR/Issues/2024/07/30/Singapore-2024-Article-IV-Consultation-Press-Release-Staff-Report-and-Statement-by-the-552788 (cit. on pp. 142, 147).

Ishwaran, H., & Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, *38*(4), 558–582. https://doi.org/10.1002/sim.8033 (cit. on p. 75).

Janizek, F., Tsunoda, S., & Lundberg, S. M. (2021). Explaining Time Series Predictions with Dynamic Masks. https://arxiv.org/abs/2106.05303 (cit. on p. 56).

Jasni, S. K., Mohd Aris, F. E., & Jamilat, V. S. (2022). Nowcasting malaysia's gdp with machine learning [Conference paper, 5 October 2022]. *Proceedings of the 9th Malaysia Statistics Conference (MyStats 2022): Dealing with Uncertainties: Unearthing Measures for Recovery* (cit. on p. 149).

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *31*(3), 300–303. https://doi.org/10.2307/2348005 (cit. on pp. 59–61, 67, 132, 141, 145).

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202 (cit. on pp. 35, 36, 59–61, 133).

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2342–2350 (cit. on pp. 50, 133).

Kannan, P., Scott, A., & Terrones, M. E. (2009). From recession to recovery: How soon and how strong? In *World economic outlook, april 2009* (pp. 103–138). International Monetary Fund. (Cit. on p. 17).

Kapetanios, G., & Zikes, F. (2018). Time-varying lasso. *Economics Letters*, *169*, 1–6. https://doi.org/10.1016/j.econlet.2018.05.021 (cit. on pp. 59, 61).

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, *186*(2), 294–316. https://doi.org/10.1016/j.jeconom.2015.02.043 (cit. on p. 39).

Kim, H. H., & Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, *178*, 352–367 (cit. on pp. 32, 34).

Kim, H. H., & Swanson, N. R. (2018). Mining big data for forecasting using factor models, forecast combinations, and shrinkage. *International Journal of Forecasting*, *34*(2), 302–325. https://doi.org/10.1016/j.ijforecast.2018.01.002 (cit. on pp. 2, 5, 59, 61, 63, 132, 141, 145).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization [arXiv:1412.6980]. *3rd International Conference on Learning Representations (ICLR)* (cit. on p. 49).

Kock, A., Medeiros, M. C., & Vasconcelos, G. F. (2020). Penalized time series regression. In P. Fuleky (Ed.), *Macroeconomic forecasting in the era of big data* (pp. 193–228, Vol. 52). Springer. (Cit. on p. 35).

Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*(3), 867–886. https://doi.org/10.1111/j.1468-2354.2012.00709.x (cit. on pp. 139, 140, 142, 147).

Krämer, N., & Sugiyama, M. (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, *106*(494), 697–705. https://doi.org/10.1198/jasa.2011.tm10171 (cit. on pp. 39, 62, 63, 133, 141, 145).

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems 4*, 950–957 (cit. on pp. 45, 50).

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241. https://doi.org/10.1214/aos/1176347265 (cit. on pp. 43, 53, 74).

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer. https://doi.org/https://doi.org/10.1007/978-1-4757-3803-2 (cit. on pp. 74, 75).

Lenza, M., & Primiceri, G. E. (2022). How to estimate a var after march 2020. *Journal of Applied Econometrics*, *37*(4), 688–699. https://doi.org/10.1002/jae.2927 (cit. on pp. 25, 28).

Li, Q. (2021). Block bootstrap prediction intervals for parsimonious var. *Journal of Forecasting*, *40*(3), 512–527. https://doi.org/10.1002/for.2772 (cit. on pp. 3, 6, 7, 75, 141, 146, 149).

Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, *4*(1), 25–38 (cit. on p. 22).

Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation*, *108*(2), 212–261. https://doi.org/10.1006/inco.1994.1009 (cit. on pp. 85, 88).

Ljung, G. M., & Box, G. E. P. (1978). On a measure of a lack of fit in time series models. *Biometrika*, *65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297 (cit. on pp. 90, 91).

Lumley, T., & Heagerty, P. J. (1999). Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 459–477 (cit. on p. 93).

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9 (cit. on pp. 3, 6, 7, 67).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://doi.org/10.5555/3122009.3122047 (cit. on p. 56).

Makridakis, S., & Hibon, M. (2000). The m3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1 (cit. on pp. 81, 90, 141, 145).

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014 (cit. on pp. 81, 121, 139, 141, 145).

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*(3), 897–906. https://doi.org/10.1016/j.ijforecast.2021.07.007 (cit. on pp. 27, 44, 56).

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd). John Wiley & Sons. (Cit. on pp. 26, 27, 54, 58, 60, 63, 74).

Marcellino, M., & Schumacher, C. (2010). Factor-midas for nowcasting and forecasting with ragged-edge data: A model comparison for german gdp. *Oxford Bulletin of Economics and Statistics*, *72*(4), 518–550. https://doi.org/10.1111/j.1468-0084.2010.00678.x (cit. on p. 28).

Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, *135*(1–2), 499–526 (cit. on pp. 20, 28).

Masarotto, G. (1990). Bootstrap prediction intervals for autoregressions. *International Journal of Forecasting*, *6*(2), 229–239 (cit. on pp. 53, 54, 74).

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111. https://doi.org/10.1111/joes.12519 (cit. on p. 45).

Massy, F. J. (1965). Principal component regression in exploratory statistical research. *Journal of the American Statistical Association*, *60*(309), 234–250 (cit. on pp. 36, 59, 60, 133).

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119 (cit. on pp. 2, 5, 6, 34, 44, 67).

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010 (cit. on pp. 35, 39, 40, 62, 63, 133).

Mei, Z., & Shi, Z. (2024). On lasso for high dimensional predictive regression. *Journal of Econometrics*, *242*(2), Article 105809 (cit. on pp. 31, 32).

Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, *17*, 1–41 (cit. on p. 75).

Merity, S., Keskar, N. S., & Socher, R. (2018). Regularizing and optimizing lstm language models [arXiv:1708.02182]. *6th International Conference on Learning Representations (ICLR)* (cit. on p. 50).

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT Press. (Cit. on p. 21).

Molnar, C. (2022). *Interpretable machine learning* (2nd) [https://christophm.github.io/interpretable-ml-book/]. Leanpub. (Cit. on pp. 74, 142, 146, 150).

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). Fforma: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. https://doi.org/10.1016/j.ijforecast.2019.03.010 (cit. on p. 86).

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87 (cit. on pp. 64, 66).

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media. (Cit. on pp. 58, 60).

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, *55*(290), 299–306. https://doi.org/10.1080/01621459.1960.10482064 (cit. on p. 26).

Nelson, C. R., & Plosser, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics*, *10*(2), 139–169. https://doi.org/10.1016/0304-3932(82)90012-5 (cit. on pp. 26, 28).

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the gini importance? *Bioinformatics*, *34*(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373 (cit. on pp. 66, 67, 69, 71).

Ng, S. (2013). Variable selection in predictive regressions. *Journal of Empirical Finance*, *21*, 129–144. https://doi.org/10.1016/j.jempfin.2012.12.002 (cit. on p. 148).

Ng, S., & Wright, J. H. (2013). Facts and challenges from the great recession for forecasting and macroeconomic modeling. *Journal of Economic Literature*, *51*(4), 1120–1156. https://doi.org/10.1257/jel.51.4.1120 (cit. on p. 6).

Optuna Developers. (2023). Optuna v3 Documentation [Accessed: 2025-01-15]. https://optuna.readthedocs.io/ (cit. on p. 45).

Pan, L., & Politis, D. N. (2016). Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions (with discussion). *Journal of Statistical Planning and Inference*, *168*, 1–27 (cit. on pp. 53–55, 74, 75).

Paudel, D. R., de Wit, A., Boogaard, H., Marcos, D., Osinga, S., & Athanasiadis, I. N. (2023). Interpretability of deep learning models for crop yield forecasting. *Computers and Electronics in Agriculture*, *206*, 107663. https://doi.org/10.1016/j.compag.2023.107663 (cit. on p. 71).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python [Accessed: 2025-01-15]. *Journal of Machine Learning Research*, *12*, 2825–2830. https://scikit-learn.org/ (cit. on pp. 31, 33, 34).

Pesaran, M. H., Pettenuzzo, D., & Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, *73*(4), 1057–1084 (cit. on pp. 53, 54, 59).

Pesaran, M. H., Pick, A., & Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, *177*(2), 134–152. https://doi.org/10.1016/j.jeconom.2013.04.002 (cit. on p. 134).

Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, *137*(1), 134–161. https://doi.org/10.1016/j.jeconom.2006.03.010 (cit. on pp. 19–21, 35, 61).

Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PLoS ONE*, *15*(3), e0231236. https://doi.org/10.1371/journal.pone.0231236 (cit. on pp. 2, 4).

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870 (cit. on pp. 3, 6, 7, 74, 75, 136, 141, 146).

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301 (cit. on pp. 22, 41–43, 133).

Radchenko, P., Vasnev, A. L., & Wang, W. (2023). Too similar to combine? on negative weights in forecast combination. *International Journal of Forecasting*, *39*(1), 18–38. https://doi.org/10.1016/j.ijforecast.2021.08.002 (cit. on p. 88).

Raftery, A. E., Kárny, M., & Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Journal of the*

American Statistical Association, *105*(489), 1794–1803. https://doi.org/10.1198/jasa.2010.tm09113 (cit. on pp. 85, 139).

Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature*, *59*(4), 1135–1190. https://doi.org/10.1257/jel.20201363 (cit. on pp. 85, 134, 148).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x (cit. on pp. 3, 6, 7, 73, 137).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0 (cit. on p. 49).

Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors, and confidence intervals. *Journal of Econometrics*, *4*(4), 393–397. https://doi.org/10.1016/0304-4076(76)90036-3 (cit. on pp. 25, 58, 60, 63).

Samuels, J. D., & Sekkel, R. M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, *33*(1), 48–60. https://doi.org/10.1016/j.ijforecast.2016.09.010 (cit. on pp. 78, 88, 90).

Scikit-learn Developers. (2025a). ElasticNet — scikit-learn 1.6.1 documentation [Accessed: 2025-01-15]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html (cit. on pp. 34, 57).

Scikit-learn Developers. (2025b). Lasso — scikit-learn 1.6.1 documentation [Accessed: 2025-01-15]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html (cit. on pp. 31, 57).

Scikit-learn Developers. (2025c). PLSRegression documentation (version 1.6.1) [Accessed: 2025-01-15]. https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html (cit. on p. 39).

Scikit-learn Developers. (2025d). Principal Component Regression vs Partial Least Squares Regression — scikit-learn 1.6.1 documentation [Accessed: 2025-01-15]. https://scikit-learn.org/stable/auto_examples/cross_decomposition/plot_pcr_vs_pls.html (cit. on p. 36).

Scikit-learn Developers. (2025e). RandomForestRegressor — scikit-learn 1.6.1 documentation [Accessed: 2025-01-15]. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (cit. on p. 42).

Scikit-learn Developers. (2025f). Ridge — scikit-learn 1.6.1 documentation [Accessed: 2025-01-15]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html (cit. on pp. 33, 36, 57).

Shabli, R., Tajul Arus, F. R., Mohd Aris, F. E., & Jamilat, V. S. (2023). Leveraging textual data in nowcasting malaysia's gross domestic product. *64th ISI World Statistics Congress*. https://www.isi-next.org/abstracts/submission/754/view/ (cit. on p. 149).

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218 (cit. on pp. 22, 43, 55, 64, 69, 148).

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591 (cit. on p. 90).

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330 (cit. on p. 74).

SingStat. (2023). Singstat data services [Accessed: 2023-11-16]. https://www.singstat.gov.sg/find-data/search-by-a-z (cit. on pp. 3, 15).

Smeekes, S., & Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, *34*(3), 408–430 (cit. on pp. 31, 32, 34, 58–61, 132, 141, 145).

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355. https://doi.org/10.1111/j.1468-0084.2008.00541.x (cit. on pp. 80, 138, 139, 142, 146).

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*, 2951–2959 (cit. on pp. 22, 32, 33, 40, 43, 49).

Soybilgen, B., & Yazgan, E. (2021). Nowcasting us gdp using tree-based ensemble models and dynamic factors. *Computational Economics*, *57*(2), 387–417. https://doi.org/10.1007/s10614-020-10024-5 (cit. on p. 133).

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958 (cit. on pp. 47, 50).

Stamer, V. (2024). Thinking outside the container: A sparse partial least squares approach to forecasting trade flows. *International Journal of Forecasting*, *40*(4), 1336–1358. https://doi.org/10.1016/j.ijforecast.2023.06.008 (cit. on p. 39).

Stock, J. H., & Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, *14*(1), 11–30. https://doi.org/10.1080/07350015.1996.10524626 (cit. on pp. 19, 26).

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179. https://doi.org/10.1198/016214502388618960 (cit. on pp. 4, 28, 29, 35, 36).

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430. https://doi.org/10.1002/for.928 (cit. on pp. 81, 83, 89).

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(S1), 3–33. https://doi.org/10.1111/j.1538-4616.2007.00014.x (cit. on p. 61).

Stock, J. H., & Watson, M. W. (2011). Dynamic factor models. In M. P. Clements & D. F. Hendry (Eds.), *The oxford handbook of economic forecasting* (pp. 35–59). Oxford University Press. (Cit. on pp. 4, 6, 29, 90).

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307. https://doi.org/10.1186/1471-2105-9-307 (cit. on pp. 66, 67, 69, 71).

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3319–3328 (cit. on pp. 3, 6, 71, 138, 142, 146, 150).

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., & Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? [arXiv preprint arXiv:2406.16964].

Advances in Neural Information Processing Systems (NeurIPS 2024). https://arxiv.org/abs/2406.16964 (cit. on p. 150).

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0 (cit. on pp. 18–20, 22, 26, 27, 49, 52, 84, 86, 92).

Thombs, L. A., & Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, *85*(410), 486–492 (cit. on p. 54).

Tian, J., & Anderson, H. M. (2014). Forecast combinations under structural break uncertainty. *International Journal of Forecasting*, *30*(1), 161–175. https://doi.org/10.1016/j.ijforecast.2013.02.014 (cit. on p. 85).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288 (cit. on pp. 30, 31, 57–59, 132, 141, 145).

Timmermann, A. G. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. G. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Elsevier. (Cit. on pp. 79, 81, 82, 85, 121, 134, 139, 140, 142, 146).

Uematsu, Y., & Tanaka, S. (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. *The Econometrics Journal*, *22*(1), 34–56 (cit. on pp. 31, 32, 34, 58, 132, 141, 145).

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3 (cit. on pp. 5, 22, 44).

Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, *15*, 1625–1651 (cit. on p. 75).

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518–1547. https://doi.org/10.1016/j.ijforecast.2022.11.005 (cit. on p. 90).

Wang, Z., Zhu, Z., & Yu, C. (2023). Variable selection in macroeconomic forecasting with many predictors. *Econometrics and Statistics*, *36*(6), 72–89. https://doi.org/10.1016/j.ecosta.2023.01.003 (cit. on p. 148).

Wasserstein, R. L., & Lazar, N. A. (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108 (cit. on p. 74).

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, *78*(10), 1550–1560. https://doi.org/10.1109/5.58337 (cit. on p. 51).

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126. https://doi.org/10.1111/1468-0262.00152 (cit. on pp. 76–78, 81, 92).

Williamson, B. D., Gilbert, P. B., Simon, N., & Carone, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, *118*(543), 1645–1658. https://doi.org/10.1080/01621459.2022.2087701 (cit. on p. 75).

Wold, H. (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 581–591, Vol. 6). Wiley. (Cit. on pp. 36, 39, 40, 62, 63, 67, 133, 141, 145).

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1 (cit. on p. 86).

XGBoost Developers. (2023). XGBoost Documentation [Accessed: 2025-01-15]. https://xgboost.readthedocs.io/ (cit. on pp. 44, 67, 68).

Yoon, J. (2021). Forecasting of real gdp growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), 247–265. https://doi.org/10.1007/s10614-020-10010-x (cit. on p. 41).

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. https://arxiv.org/abs/1409.2329 (cit. on pp. 50, 51).

Zhemkov, M. (2021). Nowcasting Russian gdp using forecast combination approach. *International Economics*, *168*, 10–24. https://doi.org/10.1016/j.inteco.2021.03.002 (cit. on p. 81).

Zhou, Z., & Hooker, G. (2021). Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, *15*(2), 1–21. https://doi.org/10.1145/3447542 (cit. on p. 67).

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429 (cit. on pp. 31, 58, 59).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320 (cit. on pp. 2, 3, 30, 31, 34, 57, 59, 61, 132, 141, 145).