Regularization method in the variable selection for logistic regression on BRFSS data

Jinbo Niu

University of Pennsylvania, School of Social Policy & Practice

jn1014@upenn.edu

Abstract

Stroke remains a leading cause of death and disability worldwide, yet effective prediction of stroke risk using large-scale population data remains challenging due to data imbalance and high-dimensional features. In this study, we develop and evaluate regularized logistic regression models for stroke prediction using data from the 2022 Behavioral Risk Factor Surveillance System (BRFSS), comprising 445,132 U.S. adult respondents and 328 health-related variables. To address data imbalance, we apply several resampling techniques including oversampling, undersampling, class weighting, and the Synthetic Minority Oversampling Technique (SMOTE). We further employ Lasso, Elastic Net, and Group Lasso regularization methods to perform feature selection and dimensionality reduction. Model performance is assessed using ROC-AUC, sensitivity, and specificity metrics. Among all methods, the Lasso-based model achieved the highest predictive performance (AUC = 0.761), while the Group Lasso method identified a compact set of key predictors—Age, Heart Disease, Physical Health, and Dental Health. These findings demonstrate the potential of regularized regression techniques for interpretable and efficient prediction of stroke risk from large-scale behavioral health data.

Introduction

Stroke remains a major global health concern, ranking as the fifth leading cause of death in the United States and a significant contributor to severe long-term disability worldwide[1][2] . Despite advances in medical care, the prompt identification of stroke symptoms in acute settings continues to pose challenges, often resulting in missed opportunities for timely intervention [3][4]. As a result, stroke persists as a threat to public health, which requires improved strategies for prevention, detection, and treatment.

The complex etiology of stroke involves multiple risk factors, including hypertension, smoking, diabetes, and obesity [5][6][7]. Understanding these risk factors and their interplay is crucial for developing effective prevention and management strategies[8]. Additionally, socioeconomic factors and access to healthcare have been shown to influence stroke outcomes, highlighting the need for a comprehensive approach to stroke research and care[9].

To address these challenges, large-scale epidemiological studies and surveillance systems play a vital role in monitoring stroke trends and risk factors across populations[10] One such initiative is the Behavioral Risk Factor Surveillance System (BRFSS), an annual telephone survey conducted in the United States[11]. The BRFSS aims to identify risk factors and track emerging health trends among adults, encompassing a wide range of topics including diet, physical activity, HIV/AIDS status, tobacco use, immunizations, existing health conditions, access to healthcare, and alcohol consumption[11].

The data applied in this research were collected from all 50 states, the District of Columbia, Guam, Puerto Rico, and the US Virgin Islands by conducting landline telephone surveys and cell phone-based

surveys in 2022. Disproportionate stratified sampling (DSS) has been used for the landline sample, and cell phone respondents are randomly selected, with each respondent having an equal probability of selection. The dataset we are working with contains 328 variables and a total of 445,132 observations in 2022. Missing values are represented by ”NA”. The data was retrieved from the CDC official website and it is originally in the SAS transport format.

The data consists of survey responses from 445,132 U.S. adults aged 18 and over. It is based on a large stratified random sample. Potential biases including non-response, incomplete interviews, missing values, and convenience bias. To alleviate the bias, some responses with blank, refused to answer, and not sure will be removed.

From the 328 variables included in the dataset, 20 are considered in the analysis. The variables are renamed and also recoded as below for further analysis:

Age: Age of the patient

Smoke: Smoke at least 100 cigarettes (0: No, 1: Yes)

Drinking: Average alcoholic beverages per day over the past 30 days

BMI: Computed body mass index (BMI)

Mental: Number of days mental health not good

Physical: Number of days physical health not good

COPD: Ever diagnosed with COPD (0: No, 1: Yes)

Education: 1: Elementary, 2: Some high school, 3: High school graduate, 4: Some college, 5: College or more

Veteran: Veteran status (0: No, 1: Yes)

Dental: Teeth removed (0: 0 teeth, 1: 1–5 teeth, 2: 6+ teeth)

Sleep: Hours of sleep per day (1–24)

Insurance: Have any health insurance (0: No, 1: Yes)

Male: Sex (0: Female, 1: Male)

Stroke: Ever diagnosed with stroke (0: No, 1: Yes)

Heart: Ever had a heart attack (0: No, 1: Yes)

Visual: Blind or visual impairment (0: No, 1: Yes)

Hearing: Serious difficulty hearing or deaf (0: No, 1: Yes)

Drug: Injected any non-prescribed drug (0: No, 1: Yes)

Type I: Type I diabetes (0: No, 1: Yes)

Type II: Type II diabetes (0: No, 1: Yes)

Material & Methods

The 2022 Behavioral Risk Factor Surveillance System (BRFSS) data will serve as the foundation for our data analysis, with R chosen as the programming language for both data wrangling and modeling tasks. Initially, the data will be transformed from the SAS format to an R-compatible file format, enabling seamless integration into our analytical workflow. Subsequently, the modeling and analysis phases will be conducted exclusively within the R environment.

Our predictive model will include a diverse set of variables, including gender, education level, smoking habits, alcohol consumption, sleep patterns, diabetes status, drug usage, BMI, age, history of heart attack, visual and auditory impairments, veteran status, physical and mental health indicators, presence of depression disorder, oral health status, and chronic obstructive pulmonary disease (COPD). Prior to model construction, we will address the challenge of imbalanced data, a common issue in stroke prediction tasks, where the majority of patients have not experienced a stroke. To mitigate this imbalance, we will employ various techniques such as class weights adjustment, synthetic sample generation, over-sampling, and under-sampling.

The performance of each imbalanced data handling method will be assessed using Receiver Operating Characteristic (ROC) analysis, enabling us to identify the most effective approach for further model development. Logistic regression, elastic net regularization, and lasso methods, including both standard lasso and group lasso techniques, will be employed as our learning algorithms.

Initially, a logistic regression model will be built, incorporating all variables with a probability cutoff of 0.5. Then regularization techniques such as lasso, elastic net, and group lasso logistic regression will be applied, with the optimal regularization parameter (lambda) determined through 10-fold cross-validation.

Following regression analysis, feature selection will be performed to identify the most informative variable sets. New logistic regression models will be constructed based on these selected features, allowing for a comparative evaluation of model accuracies between the initial and refined models. Variables demonstrating consistent predictive power across these models will be deemed as the most critical risk factors.

To avoid the issue of overfitting, 10-fold cross-validation will be employed throughout our modeling process, ensuring the robustness and generalizability of our predictive models.

Results

Descriptive statistics results

The Outcome variable

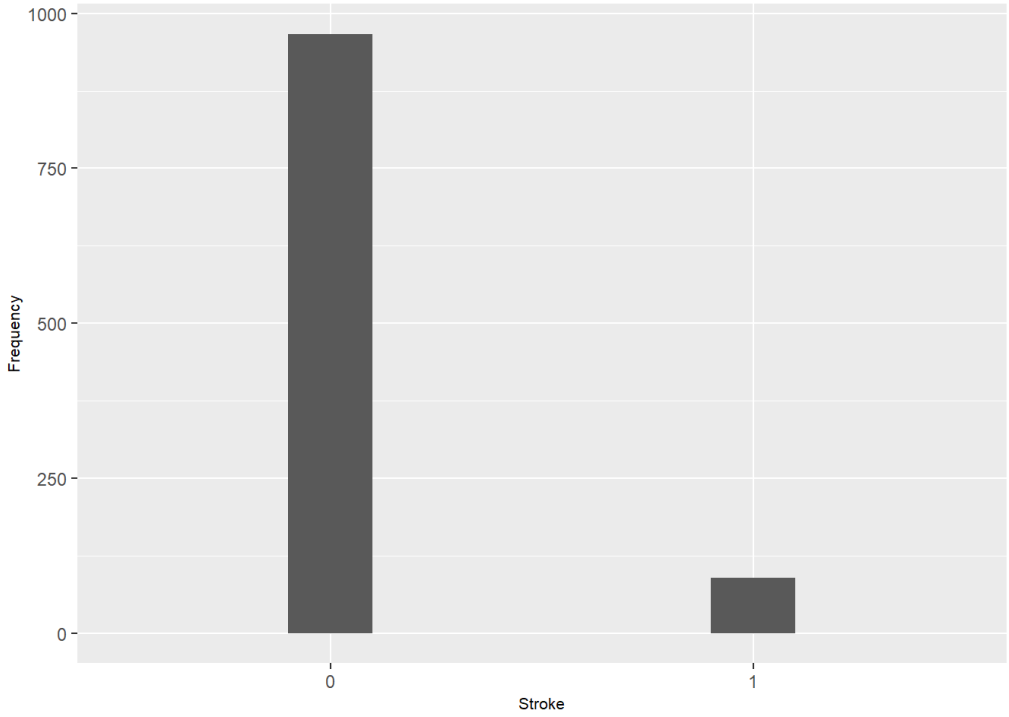

Figure 1: Distribution of the outcome variable

According to the bar chart representing the distribution of the outcome variable, it becomes evident that the dataset exhibits a notable imbalance, with a relatively small proportion of individuals experiencing stroke. This disparity in distribution is understandable, given the diverse age range of our study population and stroke occurrence may be less prevalent among certain age cohorts.

Recognizing the importance of addressing this class imbalance to ensure the robustness of our predictive model, we will employ various strategies such as oversampling, under-sampling, and the Synthetic Minority Over-sampling Technique (SMOTE)[12][13]. These techniques aim to rectify the imbalance by either augmenting the minority class instances, reducing the dominance of the majority class, or generating synthetic samples to balance the dataset [14]. By implementing these actions, we aim to enhance the model's ability to accurately predict stroke occurrences across all demographic groups, thereby improving its overall performance and reliability[8].

Categorical variables

The series of bar plots are the distributions of categorical variables. Generally most variables have even distributions. For some variables like diabetes and heart attack, the distribution could be as uneven as the outcome variable but the distributions are still reasonable as most people are not likely to have those diseases.

Also, we noticed that the patients not having type I diabetes are all having type II diabetes. So in this case, we will only take the Type I variable into the model.

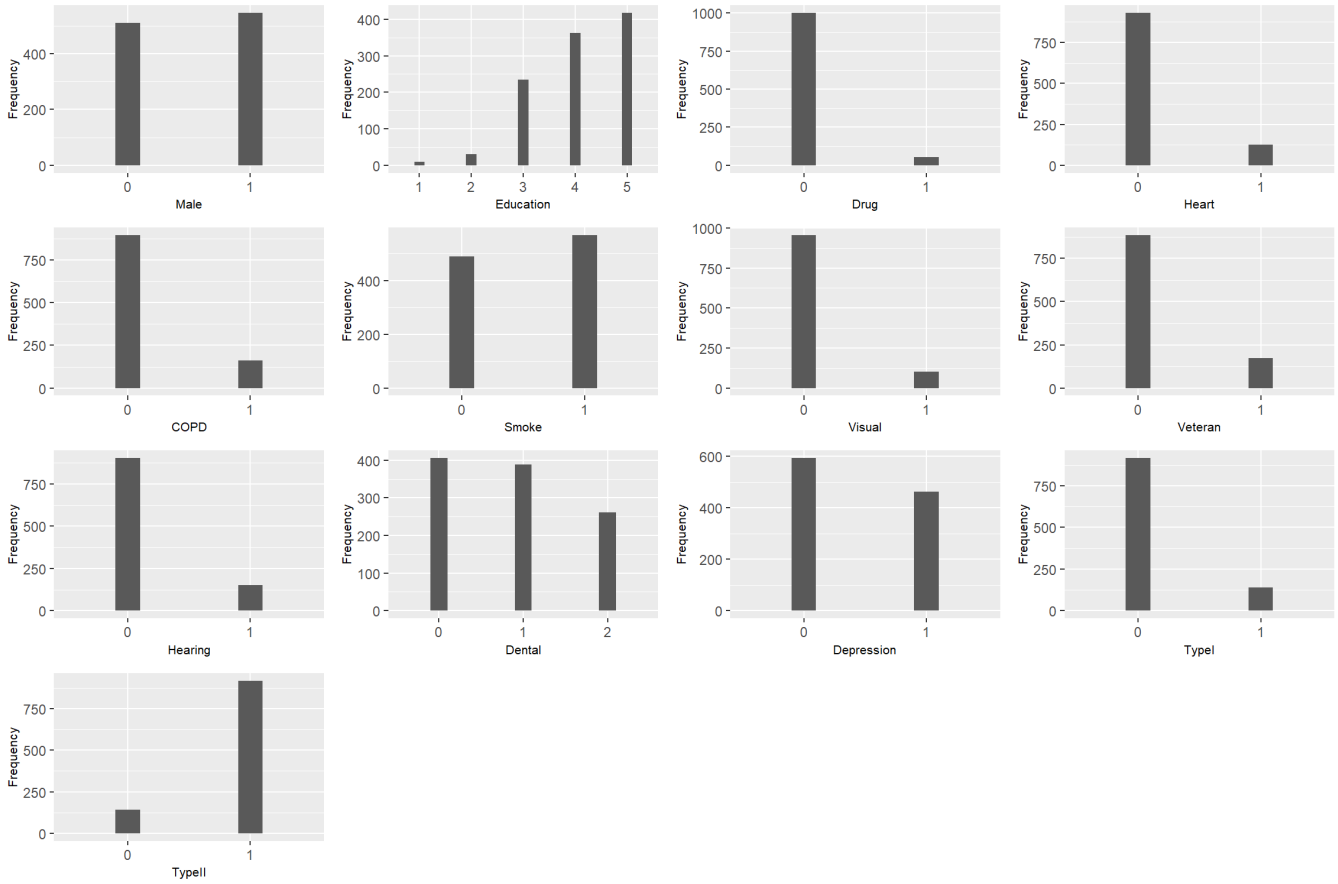

Figure 2: Distributions of categorical variables

Continuous variables

Box plots are utilized to visually represent the distribution of continuous variables. To ensure uniformity across variables, a log10 transformation is implemented. It is notable that the majority of variables exhibit outliers; however, given their behavioral nature, these outliers are deemed reasonable and thus are retained rather than removed.

Provided below is a comprehensive table detailing the descriptive statistics of the continuous variables:

name	n	mean	sd	median	max	min	skew	se
Sleep	1057	6.80	1.95	7	23	1	1.42	0.06
BMI	1057	3264.94	736.80	3165	7586	1414	0.85	22.66
Age	1057	58.41	13.46	60	80	18	-0.50	0.41
Drinking	1057	2.30	3.42	2	72	1	11.38	0.11
Mental	1057	11.83	10.38	8	30	1	0.75	0.32
Physical	1057	9.11	11.30	3	30	0	0.98	0.35

Table 1: Summary statistics of continuous variables

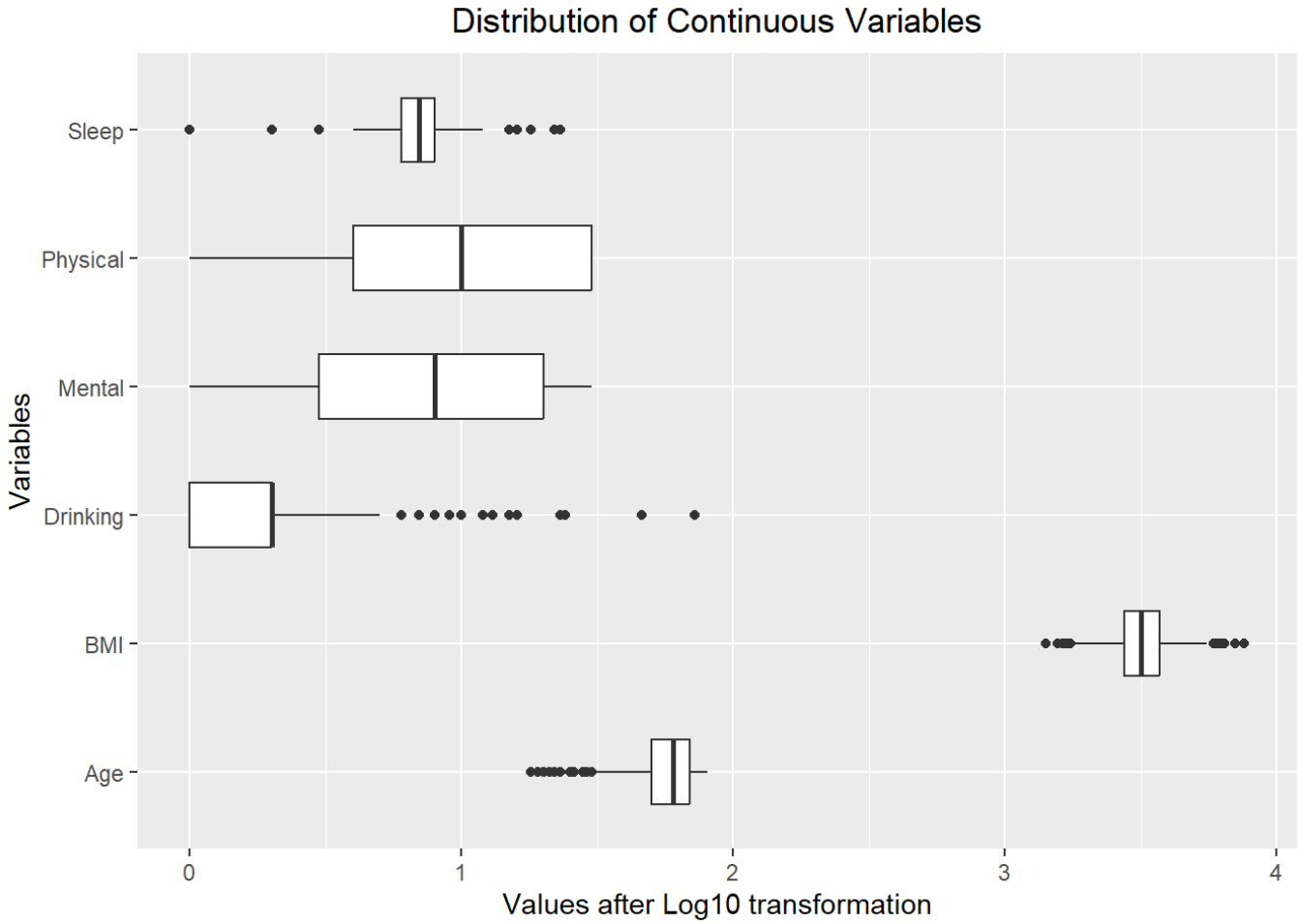

Figure 3: Distribution of continuous variables

Data Sampling results

To address the challenge of class imbalance, four methods—class weights, oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE)—were employed. Analysis of the results presented in the table reveals that the SMOTE method yields the most favorable performance, achieving the highest AUC among the evaluated methods. Although the Class Weight method demonstrates a competitive AUC value of 0.7396, its low sensitivity of approximately 0.067 raises concerns regarding its effectiveness in correctly identifying positive cases. As such, despite its comparable AUC performance, the Class Weight method is deemed less suitable for addressing imbalanced data due to its poor sensitivity. Consequently, the SMOTE method emerges as the preferred approach for handling class imbalance in subsequent analyses.

Name	AUC	Sens	Spec
SMOTE	0.739851	0.6	0.725773
Over sampling	0.728938	0.633333	0.699173
Under sampling	0.717773	0.6	0.677298
Class weight	0.7396	0.066667	0.992751

Table 2: Data Sampling results

Modeling results

grouping results

Five grouping policies were employed in this study. The first policy involves grouping variables based on their data type, resulting in two distinct groups: continuous variables and categorical variables. The second policy categorizes variables into three broad categories: Demographic, Behavior, and Health condition. The third policy further subdivides variables into four categories: Demographic, Behavior, Physical, and Mental Health. The fourth policy expands upon the third, adding Dental as a separate category. Lastly, the fifth policy treats each variable as an individual group. The outcomes of these grouping policies are detailed in Table 3.

Table 3: Grouped variables

Variables	group 1	group 2	group 3	group 4	group 5
Male	1	1	1	1	1
Education	2	1	1	1	2
Sleep	2	2	2	2	3
Drug	1	2	2	2	4
BMI	2	1	1	1	5
Age	2	1	1	1	6
Heart	1	3	3	3	7
Drinking	2	2	2	2	8
Mental	2	3	4	4	9
COPD	1	3	3	3	10
Smoke	1	2	2	2	11
Visual	1	2	2	2	12
Veteran	1	3	3	3	13
Hearing	1	1	1	1	14
Physical	2	3	3	3	15
Dental	2	3	3	5	16
Depression	1	3	4	4	17
TypeI	1	3	3	3	18

variable selection results

Utilizing various regularization techniques such as lasso, elastic net, and group lasso with distinct groupings, certain variables have had their coefficients effectively shrunk to zero. Consequently, variables associated with zero coefficients are deemed redundant and will be eliminated from further analysis, while those retaining non-zero coefficients will be retained for model refinement. Implementation of group lasso methods was facilitated through the gglasso package; however, it is noteworthy that this package has the propensity to shrink all variables to zero. To mitigate this potential limitation, alternative groupings may be explored, prioritizing groups with the least number of non-zero coefficients to ensure the retention of informative variables.

Table 4: Variables removed by non-group lasso methods

Model	**Variables removed**
Stepwise	Education, BMI, Hearing
Lasso	Sleep, BMI, Drinking, Smoke, Hearing, TypeI
Elasticnet	BMI, Drinking, Hearing

Table 5: Variable Selection Results of gorup lasso models

Variables	group1	group2	group3	group4(1)	group4(2)	group5(1)	group5(2)	group5(3)	group5(4)
Male									
Education									*
Sleep									
Drug									
BMI									
Age		*	*		*		*	*	*
Heart		*	*		*			*	*
Drinking		*	*		*				
Mental		*	*		*				*
COPD		*	*		*				
Smoke	*		*						
Visual	*		*						
Veteran	*		*						
Hearing	*		*						
Physical	*		*			*	*	*	*
Dental	*		*	*	*		*	*	*
Depression	*			*	*				
TypeI	*								

Logistic Models with different variable selections

In this section, the selected variables will be incorporated into the logistic regression model and compared with the original logistic regression model comprising all variables. The performance of each model will be assessed based on key metrics such as the Area Under the Curve (AUC), Sensitivity, and Specificity, as presented in Table 6. Upon examination, it becomes evident that the logistic model with the variables selected from the Lasso model exhibits the most favorable performance, boasting the highest AUC among all models tested. Following closely behind is the logistic model with variables from group 5(3), which achieves an AUC of 0.7488 and a sensitivity of 0.6111, showcasing minimal discrepancies compared to the Lasso model. Notably, the group 5(3) derived from group lasso achieves dimension reduction from 18 variables to 4, surpassing the Lasso model's reduction from 18 to 12 variables. Consequently, the variables from group 5(3) derived from group lasso emerges as the preferred choice for the final model selection.

Therefore, four variables—**Age, Heart, Physical, and Dental**— are selected and are considered as significant predictors for future research endeavors. These variables exhibit substantial predictive power and warrant further investigation in subsequent studies.

Logistic model with selected variables	ROC	Sens	Spec
Full Logistic Regression	0.7311	0.5778	0.7220
variables from Stepwise	0.7464	0.6111	0.7280
variables from Lasso	0.7613	0.6667	0.7259
variables from Elastic Net	0.7503	0.6333	0.7116
variables from Group 1	0.6502	0.5556	0.6371
variables from Group 2	0.7124	0.5333	0.7672
variables from Group 3	0.7341	0.5889	0.7270
variables from Group 4(1)	0.6483	0.6222	0.5957
variables from Group 4(2)	0.7289	0.6222	0.7352
variables from Group5(1)	0.6465	0.4889	0.7115
variables from Group5(2)	0.6991	0.6333	0.6412
variables from Group5(3)	0.7488	0.6111	0.7311
variables from Group5(4)	0.7450	0.6000	0.7363

Table 6: Model performance based on different variable selections

Discussion

Modeling results and advantages

In this study, we employed regularization methods with a particular focus on the group lasso technique, to effectively reduce dimensions and filter variables most pertinent to stroke. Regularization methods offer a powerful means to address the challenges of high-dimensional data by imposing penalties on model coefficients, thereby encouraging sparsity and feature selection[15]. The group lasso, in particular, extends the benefits of traditional lasso regularization by incorporating group structures within the data, allowing for the simultaneous selection of entire groups of correlated variables[16]. This approach not only helps identify key predictors associated with Stroke but also facilitates the interpretation of model results.

Compared to Principal Component Analysis (PCA), which operates solely on continuous variables and may obscure the interpretability of the underlying data, regularization methods offer a more flexible framework that can accommodate both continuous and categorical predictors[17][18]. Moreover, while PCA may struggle with non-linear relationships and interactions among variables[19], regularization methods, such as the group lasso, are better equipped to capture complex patterns in the data, leading to more robust and interpretable models[20].

Despite the reduction in the number of variables retained by the group lasso method, our logistic models constructed with these selected variables consistently demonstrated strong performance in predicting stroke outcomes. This highlights the efficiency of the group lasso in identifying a compact subset of informative predictors while maintaining predictive accuracy[21]. By prioritizing relevant variables and discarding noise, the group lasso effectively balances model complexity and performance, offering a practical solution for dimension reduction in predictive modeling tasks[22].

Limitations

In terms of methodology, the group lasso method offers the advantage of preserving the correlation of features within groups and handling both continuous and categorical variables. However, defining these groups can be subjective, potentially impacting model performance and interpretability[23]. Moreover, the sensitivity of Group Lasso to outliers, particularly those within group clusters, can further affect model performance[24]. Given the presence of outliers that cannot be removed, the efficacy of the group lasso model may be compromised.

Furthermore, when dealing with health status surveillance data, relying solely on self-reported prevalence rates may introduce bias, as respondents may lack awareness of their true risk status [25]. Thus, to obtain more precise estimates, it may be necessary to supplement self-reported data with laboratory tests[26].

Additionally, the gglasso package may exhibit a tendency to shrink all variables to zero, necessitating a thorough examination of the coefficient matrix and individual group selections with less optimal lambda values[27]. In such cases, alternative tools may be explored to complement the group lasso approach.

Conclusion

This study demonstrates the effectiveness of regularization-based logistic regression models for stroke risk prediction using large-scale behavioral health survey data. By systematically evaluating Lasso, Elastic Net, and Group Lasso methods under various data balancing strategies, we show that regularization can achieve competitive predictive performance while maintaining model interpretability. The integration of SMOTE with Lasso-based models yielded the best trade-off between sensitivity and specificity, highlighting the importance of addressing data imbalance in health-related machine learning tasks.

From a methodological perspective, our findings emphasize that group-structured penalization methods can serve as a powerful and interpretable alternative to black-box models in population-level health prediction. The selected key predictors—Age, Heart Disease, Physical Health, and Dental Health—reflect consistent behavioral and physiological relevance, demonstrating that interpretable models can still provide high performance on large-scale heterogeneous data.

Future work will extend these models to nonlinear and ensemble approaches such as gradient boosting and neural networks, and explore domain adaptation for cross-year BRFSS data. Additionally, integrating explainability techniques (e.g., SHAP or LIME) will further enhance transparency and clinical trust in machine learning–driven public health modeling.

References

[1] Muhammad Ahmed, Shaheer Bin Shafiq, Junaid Razzak, Khubaib Tariq Mansoor, Muhammad Abdullah Naveed, Ahila Ali, Muhammad Shaheer Bin Faheem, Sumaya Samadi, Himaja Dutt Chigurupati, and Sivaram Neppala. Trends and disparities in stroke mortality among adults with hyperlipidemia in the united states, 1999–2023. *Journal of Epidemiology and Global Health*, 15(1):116, 2025.

[2] M. Katan and A. Luft. Global burden of stroke. *Seminars in Neurology*, 38(2):208–211, 2018.

[3] Amie W. Hsia, Amanda Castle, Jeffrey J. Wing, Dorothy F. Edwards, Nina C. Brown, Tara M. Higgins, Jasmine L. Wallace, Sara S. Koslosky, M. Chris Gibbons, Brisa N. Sánchez, Ali Fokar, Nawar Shara, Lewis B. Morgenstern, and Chelsea S. Kidwell. Understanding reasons for delay in seeking acute stroke care in an underserved urban population. *Stroke*, 42(6):1697–1701, 2011.

[4] A. K. Boehme, C. Esenwa, and M. S. Elkind. Stroke risk factors, genetics, and prevention. *Circulation Research*, 120(3):472–495, 2017.

[5] Y. Kono, Y. Terasawa, K. Sakai, Y. Iguchi, Y. Nishiyama, C. Nito, S. Suda, K. Kimura, Y. Murakami, T. Kanzawa, K. Yamashiro, R. Tanaka, and S. Okubo. Association between living conditions and the risk factors, etiology, and outcome of ischemic stroke in young adults. *Internal medicine (Tokyo, Japan)*, 62(19):2813–2820, 2023.

[6] Jiao Shang, Yanmei Wu, Lixin Zhang, Xueting Jiang, and Ruiping Zhang. Joint effect of modifiable risk factors and genetic susceptibility on ischaemic stroke. *Journal of Stroke and Cerebrovascular Diseases*, 34(6):108313, 2025.

[7] R. Ng, R. Sutradhar, Z. Yao, W. P. Wodchis, and L. C. Rosella. Smoking, drinking, diet, and physical activity—modifiable lifestyle risk factors and their associations with age to first chronic disease. *International Journal of Epidemiology*, 49(1):113–130, 2020.

[8] Cheryl Bushnell, Walter N. Kernan, Anjail Z. Sharrief, Seemant Chaturvedi, John W. Cole, William K. Cornwell, Christine Cosby-Gaither, Sarah Doyle, Larry B. Goldstein, Olive Lennon, Deborah A. Levine, Mary Love, Eliza Miller, Mai Nguyen-Huynh, Jennifer Rasmussen-Winkler, Kathryn M. Rexrode, Nicole Rosendale, Satyam Sarma, Daichi Shimbo, Alexis N. Simpkins, Erica S. Spatz, Lisa R. Sun, Vin Tangpricha, Dawn Turnage, Gabriela Velazquez, and Paul K. Whelton. 2024 guideline for the primary prevention of stroke: A guideline from the american heart association/american stroke association. *Stroke*, 55(12):e344–e424, 2024.

[9] Camila Pantoja-Ruiz, Rufus Akinyemi, Diego I. Lucumi-Cuesta, Daniel Youkee, Eva Emmett, Marina Soley-Bori, Wasana Kalansooriya, Charles Wolfe, and Iain J. Marshall. Socioeconomic status and stroke: A review of the latest evidence on inequalities and their drivers. *Stroke*, 56(3):794–805, 2025.

[10] David C. Goff, Lawrence Brass, Lynne T. Braun, Janet B. Croft, Judd D. Flesch, Francis G.R. Fowkes, Yuling Hong, Virginia Howard, Sara Huston, Stephen F. Jencks, Russell Luepker, Teri Manolio, Christopher O'Donnell, Rose Marie Robertson, Wayne Rosamond, John Rumsfeld, Stephen Sidney, and Zhi Jie Zheng. Essential features of a surveillance system to support the prevention and management of heart disease and stroke. *Circulation*, 115(1):127–155, 2007.

[11] Centers for Disease Control and Prevention. Behavioral risk factor surveillance system. Web page, 2024.

[12] Y. Wu and Y. Fang. Stroke prediction with machine learning methods among older chinese. *International Journal of Environmental Research and Public Health*, 17(6):1828, 2020.

[13] Javad Hassannataj Joloudari, Abdolreza Marefat, Mohammad Ali Nematollahi, Solomon Sunday Oyelere, and Sadiq Hussain. Effective class-imbalance learning based on smote and convolutional neural networks. *Applied Sciences*, 13(6), 2023.

[14] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, et al. Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *Knowledge and Information Systems*, 65:31–57, 2023.

[15] Clara Yokochi, Regina Bispo, Fernando Ricardo, and Ricardo Calado. Regularization methods for high-dimensional data as a tool for seafood traceability. *Journal of Statistical Theory and Practice*, 17(3):44, 2023.

[16] J. Klosa, N. Simon, P. O. Westermark, et al. Seagull: Lasso, group lasso and sparse-group lasso regularization for linear regression models via proximal gradient descent. *BMC Bioinformatics*, 21:407, 2020.

[17] Gerhard Tutz and Jan Gertheiss. Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200, 2016.

[18] Szymon Nowakowski, Piotr Pokarowski, Wojciech Rejchel, and Agnieszka Sołtys. Improving group lasso for high-dimensional categorical data. In Jiří Mikyška, Clélia de Mulatier, Maciej Paszynski, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M.A. Sloot, editors, *Computational Science – ICCS 2023*, pages 455–470, Cham, 2023. Springer Nature Switzerland.

[19] A. A. Portnova-Fahreeva, F. Rizzoglio, I. Nisky, M. Casadio, F. A. Mussa-Ivaldi, and E. Rombokas. Linear and non-linear dimensionality-reduction techniques on full hand kinematics. *Frontiers in Bioengineering and Biotechnology*, 8:429, 2020. Published May 5, 2020.

[20] S. Friedrich, A. Groll, K. Ickstadt, et al. Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research*, 32(2):425–440, 2023.

[21] Xiaoli Liu, Jianzhong Wang, Fulong Ren, and Jun Kong. Group guided fused laplacian sparse group lasso for modeling alzheimer's disease progression. *Computational and Mathematical Methods in Medicine*, 2020(1):4036560, 2020.

[22] R Muthukrishnan and R Rohini. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pages 18–20, 2016.

[23] Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286, 2008.

[24] Y. Jiang, Y. Wang, J. Zhang, B. Xie, J. Liao, and W. Liao. Outlier detection and robust variable selection via the penalized weighted lad-lasso method. *Journal of Applied Statistics*, 48(2):234–246, 2020.

[25] R. Rosenman, V. Tennekoon, and L. G. Hill. Measuring bias in self-reported data. *International Journal of Behavioural and Healthcare Research*, 2(4):320–332, 2011.

[26] P. St Clair, É. Gaudette, H. Zhao, B. Tysinger, R. Seyedin, and D. P. Goldman. Using self-reports or claims to assess disease prevalence: It's complicated. *Medical Care*, 55(8):782–788, 2017.

[27] Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalized learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.