

# Finding the Sweet Spot: Optimal Data Augmentation Ratio for Imbalanced Credit Scoring Using ADASYN

Luis H. Chia<sup>1,\*</sup>

<sup>1</sup>Independent Researcher, Lima, Peru

\*Corresponding author: [luischia.r@gmail.com]

ORCID: 0000-0002-5317-3656

## Abstract

Credit scoring models face a critical challenge: severe class imbalance, with default rates typically below 10%, which hampers model learning and predictive performance. While synthetic data augmentation techniques such as SMOTE and ADASYN have been proposed to address this issue, the optimal augmentation ratio remains unclear, with practitioners often defaulting to full balancing (1:1 ratio) without empirical justification.

This study systematically evaluates 10 data augmentation scenarios using the Give Me Some Credit dataset (97,243 observations, 7% default rate), comparing SMOTE, BorderlineSMOTE, and ADASYN at different multiplication factors ( $1\times$ ,  $2\times$ ,  $3\times$ ). All models were trained using XGBoost and evaluated on a held-out test set of 29,173 real observations. Statistical significance was assessed using bootstrap testing with 1,000 iterations.

Key findings reveal that ADASYN with  $1\times$  multiplication (doubling the minority class) achieved optimal performance with AUC of 0.6778 and Gini coefficient of 0.3557, representing statistically significant improvements of +0.77% and +3.00% respectively ( $p = 0.017$ , bootstrap test). Higher multiplication factors ( $2\times$  and  $3\times$ ) resulted in performance degradation, with  $3\times$  showing a  $-0.48\%$  decrease in AUC, suggesting a “law of diminishing returns” for synthetic oversampling. The optimal class imbalance ratio was found to be 6.6:1 (majority:minority), contradicting the common practice of balancing to 1:1.

This work provides the first empirical evidence of an optimal “sweet spot” for data augmentation in credit scoring, with practical guidelines for industry practitioners and researchers working with imbalanced datasets. While demonstrated on a single representative dataset, the methodology provides a reproducible framework for determining optimal augmentation ratios in other imbalanced domains.

**Keywords:** Credit scoring, ADASYN, data augmentation, imbalanced learning, synthetic oversampling, class imbalance

**Keywords:** Credit Scoring, Imbalanced Learning, ADASYN, SMOTE, Data Augmentation, Machine Learning, XGBoost, Financial Risk Assessment

# 1 Introduction

## 1.1 Motivation and Context

Credit scoring plays a fundamental role in modern financial decision-making, enabling lenders to assess the creditworthiness of loan applicants and manage portfolio risk effectively (Thomas et al., 2002). Machine learning models have become the industry standard for credit scoring, demonstrating superior predictive performance compared to traditional statistical methods (Lessmann et al., 2015). However, these models face a persistent challenge: severe class imbalance, where defaulting customers typically represent less than 10% of the population (Brown and Mues, 2012).

This imbalance poses significant difficulties for model training. Standard machine learning algorithms, optimized for balanced datasets, tend to bias predictions toward the majority class, resulting in poor detection of the minority class—the very defaults that lenders need to identify (He and Garcia, 2009). The cost of misclassification is asymmetric: failing to identify a default (false negative) is far more costly than incorrectly rejecting a good customer (false positive), making accurate minority class prediction critical for business profitability.

Synthetic data augmentation techniques, particularly SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) and its variants, have emerged as promising solutions to address class imbalance. These methods generate synthetic observations in the feature space of the minority class, artificially increasing its representation during model training. ADASYN (Adaptive Synthetic Sampling) (He et al., 2008), an advanced variant, adaptively generates more synthetic samples in regions where the minority class is harder to learn, theoretically providing superior performance.

Despite widespread adoption in research, a critical question remains unanswered: *How much augmentation is optimal?* Practitioners often apply aggressive oversampling to achieve a balanced 1:1 class ratio, based on the intuition that "more data is better." However, this practice lacks empirical validation and may introduce risks such as overfitting to synthetic patterns, distribution drift, and computational inefficiency.

## 1.2 Research Gap

While the effectiveness of SMOTE and its variants has been demonstrated across various domains (Fernández et al., 2018), existing literature exhibits three critical gaps:

**Gap 1: Lack of systematic evaluation of augmentation ratios.** Most studies compare augmented versus non-augmented datasets but fail to investigate the optimal degree of augmentation. The implicit assumption that "more synthetic data yields better performance" has not been rigorously tested across different multiplication factors.

**Gap 2: Absence of statistical significance testing.** Many papers report marginal performance improvements (e.g., +0.5% AUC) without formal statistical validation. Without significance tests such as bootstrap resampling, it remains unclear whether observed improvements represent genuine model enhancement or random variation.

**Gap 3: Limited comparison of adaptive versus uniform techniques.** While ADASYN’s theoretical advantages are well-documented (He et al., 2008), empirical head-to-head comparisons with SMOTE and BorderlineSMOTE under controlled conditions (same multiplication factor, same algorithm) are scarce in the credit scoring literature.

These gaps leave practitioners without clear guidance: Should they duplicate the minority class once ( $1\times$ ), twice ( $2\times$ ), or achieve full balance? Should they use SMOTE’s

uniform interpolation or ADASYN’s adaptive approach? The absence of evidence-based recommendations may lead to suboptimal model configurations in production systems.

### 1.3 Research Contributions

This study addresses the identified gaps through a systematic empirical investigation of data augmentation strategies for credit scoring. Our contributions are fourfold:

**Contribution 1: Identification of the “sweet spot” in data augmentation.**

We provide the first empirical evidence that optimal performance occurs at  $1\times$  multiplication (doubling the minority class), with higher factors ( $2\times$  and  $3\times$ ) exhibiting diminishing returns and eventual performance degradation. This finding challenges the common practice of aggressive oversampling to achieve 1:1 balance.

**Contribution 2: Rigorous statistical validation.** All performance comparisons are validated using bootstrap hypothesis testing (1,000 iterations) with explicit  $p$ -values and 95% confidence intervals. We demonstrate that ADASYN  $1\times$  achieves statistically significant improvements ( $p = 0.017$ ) over baseline, establishing a robust methodological standard for future research.

**Contribution 3: Controlled technique comparison.** We conduct the first systematic comparison of SMOTE, BorderlineSMOTE, and ADASYN at identical multiplication factors, isolating the effect of the generation strategy. Results show that adaptive approaches (ADASYN) outperform uniform interpolation (SMOTE), with BorderlineSMOTE occupying an intermediate position.

**Contribution 4: Practical guidelines for practitioners.** Based on our findings, we provide actionable recommendations: (i) use ADASYN with  $1\times$  multiplication for credit scoring tasks with  $\sim 7\%$  default rates, (ii) target a class ratio of approximately 6-7:1 rather than 1:1, and (iii) avoid excessive hyperparameter tuning, which we demonstrate leads to overfitting (grid search degraded test AUC by  $-0.10$ ).

### 1.4 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on credit scoring, imbalanced learning, and synthetic data augmentation techniques. Section 3 describes our methodology, including dataset characteristics, experimental design, and evaluation metrics. Section 4 presents results from all 10 experimental scenarios, including statistical significance tests and synthetic data quality assessments. Section 5 discusses the implications of our findings, addresses limitations, and suggests future research directions. Section 6 concludes with practical recommendations for industry practitioners and researchers.

## 2 Related Work

### 2.1 Credit Scoring Models

Credit scoring has evolved significantly over the past decades, transitioning from expert-based systems to sophisticated statistical and machine learning approaches (Thomas et al., 2002). Traditional methods such as logistic regression and discriminant analysis dominated early credit scoring applications due to their interpretability and regulatory compliance (Hand and Henley, 1997). However, the advent of machine learning techniques

has enabled lenders to capture complex non-linear relationships in borrower data, leading to improved predictive accuracy.

Lessmann et al. (2015) conducted a comprehensive benchmark of classification algorithms for credit scoring, evaluating 41 techniques across eight real-world datasets. Their findings demonstrated that ensemble methods, particularly random forests and gradient boosting machines, consistently outperformed traditional statistical approaches. Building on this foundation, XGBoost (Chen and Guestrin, 2016) has emerged as a state-of-the-art algorithm, combining computational efficiency with exceptional predictive performance through regularized gradient boosting and parallel tree construction.

Despite these advances, a persistent challenge remains: the severe class imbalance inherent in credit datasets, where defaulters typically constitute less than 10% of observations (Brown and Mues, 2012). This imbalance fundamentally undermines standard learning algorithms, which implicitly assume balanced class distributions and optimize for overall accuracy rather than minority class detection.

## 2.2 Imbalanced Learning Techniques

The imbalanced learning problem has received extensive attention in machine learning research. He and Garcia (2009) provide a comprehensive survey categorizing solutions into three broad approaches: (i) data-level methods that rebalance the training distribution through resampling, (ii) algorithm-level methods that modify learning algorithms to account for class imbalance, and (iii) cost-sensitive methods that assign asymmetric misclassification costs.

Data-level approaches, particularly oversampling techniques, have gained prominence due to their algorithm-agnostic nature and ease of implementation. Random oversampling, which simply duplicates minority class instances, risks overfitting as models may memorize replicated observations rather than learning generalizable patterns (Fernández et al., 2018). Conversely, undersampling the majority class discards potentially valuable information and may degrade model performance when data are limited.

Algorithm-level modifications include ensemble methods such as balanced random forests, which combine bootstrap sampling with decision tree ensembles to address imbalance (Wang and Yao, 2009). Cost-sensitive learning directly incorporates misclassification costs into the objective function, penalizing minority class errors more heavily. XGBoost implements this through the `scale_pos_weight` parameter, which adjusts the gradient of the loss function based on class imbalance ratios. While effective, these approaches do not expand the information content of the training data and may still struggle with severe imbalance scenarios.

## 2.3 SMOTE and Variants

Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique (SMOTE), which revolutionized oversampling by generating synthetic minority class instances rather than simply duplicating existing observations. SMOTE operates by selecting a minority class instance, identifying its  $k$  nearest minority neighbors in feature space, and creating synthetic samples through linear interpolation along the line segments connecting the instance to its neighbors. This approach increases minority class representation while promoting generalization by populating previously sparse regions of the feature space.

The success of SMOTE has spawned numerous variants designed to address specific

limitations. BorderlineSMOTE (Han et al., 2005) focuses synthetic generation on "borderline" instances—minority samples positioned near the decision boundary where classification is most challenging. By concentrating augmentation efforts on difficult cases, BorderlineSMOTE aims to improve decision boundary refinement without over-populating easily separable regions.

ADASYN (Adaptive Synthetic Sampling) (He et al., 2008) extends this concept through adaptive density weighting. Rather than uniformly distributing synthetic samples across the minority class, ADASYN calculates a density distribution for each minority instance based on the ratio of majority to minority neighbors in its local region. Instances in predominantly majority neighborhoods—indicating hard-to-learn regions—receive more synthetic samples, while those in minority-dense areas receive fewer. This adaptive approach theoretically provides superior performance by prioritizing regions where the model needs additional training signal.

Fernández et al. (2018) conducted a 15-year retrospective analysis of SMOTE applications, identifying both successes and ongoing challenges. While SMOTE and its variants consistently improve minority class recall, the optimal degree of oversampling remains poorly understood. Most studies apply ad-hoc multiplication factors without systematic investigation, and the trade-offs between augmentation intensity and model performance have not been rigorously characterized.

## 2.4 Recent Applications in Finance

Recent credit scoring research has explored advanced augmentation strategies beyond traditional SMOTE variants. Bravo and Weber (2021) employed variational autoencoders (VAEs) for semi-supervised learning in credit scoring, demonstrating that generative models can capture complex data distributions and improve performance on imbalanced datasets. However, their approach focuses on semi-supervised scenarios where unlabeled data are abundant, whereas our study addresses the fully supervised setting common in operational credit scoring systems.

Brown and Mues (2012) specifically investigated imbalanced classification algorithms for credit scoring, comparing techniques including SMOTE, cost-sensitive learning, and ensemble methods across multiple datasets. Their findings suggested that no single technique universally dominates, and performance depends on dataset characteristics such as imbalance ratio, dimensionality, and feature types. Notably, they did not systematically vary augmentation ratios, leaving the question of optimal oversampling intensity unresolved.

Despite the maturity of imbalanced learning research, a critical gap persists: *the lack of empirical evidence guiding the selection of augmentation ratios in production credit scoring systems*. Practitioners face a decision space with multiple dimensions—technique selection (SMOTE vs. ADASYN vs. BorderlineSMOTE), multiplication factor ( $1\times$ ,  $2\times$ ,  $3\times$ , or full balance), and hyperparameter configuration—yet have limited evidence-based guidance. Our study directly addresses this gap through systematic experimentation across these dimensions, providing actionable insights for both researchers and practitioners.

## 3 Methodology

### 3.1 Dataset Description

We employ the *Give Me Some Credit* dataset, a publicly available credit scoring dataset from Kaggle containing loan applicant information and binary default indicators. After preprocessing to remove missing values and outliers using caps based on official demographic and financial statistics, the cleaned dataset comprises 97,243 observations with six predictor variables and a binary target variable indicating serious delinquency (90+ days past due) within a two-year window.

**Target Variable:** The binary outcome `SeriousDlqin2yrs` exhibits severe class imbalance with a default rate of 7.07% (6,871 defaults among 97,243 observations), yielding a majority-to-minority ratio of approximately 13.15:1. This imbalance level is representative of real-world consumer credit portfolios (Brown and Mues, 2012).

**Predictor Variables:** The feature set includes six variables spanning demographic, financial, and credit history dimensions:

- `age`: Borrower age in years
- `MonthlyIncome`: Monthly income in USD
- `DebtRatio`: Total monthly debt obligations divided by monthly income
- `NumberOfDependents`: Number of financial dependents
- `NumberOfOpenCreditLinesAndLoans`: Count of open credit facilities
- `NumberRealEstateLoansOrLines`: Count of real estate loans or lines of credit

**Data Preprocessing:** To ensure synthetic generation respects realistic constraints, we apply domain-informed caps to each variable based on official statistics. For instance, age is bounded to [21, 85] years based on demographic data, while monthly income is capped at \$25,000 to exclude extreme outliers. Three variables (`NumberOfDependents`, `NumberOfOpenCreditLinesAndLoans`, `NumberRealEstateLoansOrLines`) are discrete integers, requiring post-generation rounding to maintain realistic values.

**Train-Test Split:** We employ a stratified 70-30 split, yielding 68,070 training observations (4,810 defaults, 7.07%) and 29,173 test observations (2,061 defaults, 7.06%). Stratification ensures both sets maintain identical class distributions. Critically, the test set remains completely untouched throughout all experiments, serving as an unbiased evaluation benchmark of 100% real data.

**Feature Scaling:** All features undergo standardization ( $\mu = 0, \sigma = 1$ ) using `StandardScaler` fit exclusively on the training set to prevent data leakage. This normalization is essential for distance-based synthetic generation techniques (SMOTE, ADASYN), which rely on Euclidean distance calculations in feature space.

### 3.2 Data Augmentation Techniques

We evaluate three synthetic oversampling techniques representing different generation philosophies: uniform interpolation (SMOTE), boundary-focused generation (BorderlineSMOTE), and adaptive density-based sampling (ADASYN).

### 3.2.1 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE (Chawla et al., 2002) generates synthetic minority class instances through linear interpolation between existing minority samples and their  $k$  nearest minority neighbors. For a minority instance  $\mathbf{x}_i$ , SMOTE:

1. Identifies  $k$  nearest minority neighbors using Euclidean distance
2. Randomly selects one neighbor  $\mathbf{x}_j$
3. Generates a synthetic sample:  $\mathbf{x}_{syn} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_j - \mathbf{x}_i)$ , where  $\lambda \sim \text{Uniform}(0, 1)$

This process repeats until the desired class ratio is achieved. SMOTE treats all minority instances equally, distributing synthetic samples uniformly across the feature space.

### 3.2.2 BorderlineSMOTE

BorderlineSMOTE (Han et al., 2005) refines SMOTE by concentrating synthetic generation on "borderline" minority instances—samples positioned near the decision boundary where classification difficulty is highest. An instance  $\mathbf{x}_i$  is classified as borderline if  $m/2 \leq m' < m$ , where  $m'$  is the number of majority class neighbors among its  $m$  nearest neighbors. Synthetic samples are then generated only for borderline instances using the standard SMOTE interpolation procedure. This targeted approach aims to refine decision boundaries without over-populating easily separable regions.

### 3.2.3 ADASYN (Adaptive Synthetic Sampling)

ADASYN (He et al., 2008) introduces adaptive density weighting to prioritize difficult-to-learn regions. For each minority instance  $\mathbf{x}_i$ , ADASYN:

1. Calculates the ratio  $r_i = \Delta_i/k$ , where  $\Delta_i$  is the number of majority class neighbors among  $k$  nearest neighbors
2. Normalizes ratios:  $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$ , where  $m_s$  is the number of minority instances
3. Determines the number of synthetic samples for  $\mathbf{x}_i$ :  $g_i = \hat{r}_i \cdot G$ , where  $G$  is the total synthetics to generate
4. Generates  $g_i$  synthetic samples using SMOTE-style interpolation

ADASYN adaptively allocates more synthetic samples to minority instances surrounded by majority neighbors (high  $r_i$ ), indicating regions where the model struggles to learn minority patterns. This density-based weighting theoretically provides superior performance by directing augmentation efforts where they are most needed.

## 3.3 Experimental Design

We design 10 experimental scenarios to systematically evaluate augmentation techniques across multiple dimensions: technique type (SMOTE, BorderlineSMOTE, ADASYN), multiplication factor ( $1\times$ ,  $2\times$ ,  $3\times$ ), and ensemble combinations. Each experiment follows an identical workflow to ensure fair comparison:

1. **Baseline (E01):** Train XGBoost on original imbalanced data (68,070 observations, 7.07% default rate) without augmentation
2. **Single-Technique Experiments (E02-E06, E08-E10):** Apply one augmentation technique at a specified multiplication factor to the training set only
3. **Ensemble Experiment (E07):** Combine synthetics from SMOTE, BorderlineSMOTE, and ADASYN (each at  $1\times$ ) into a single augmented training set

Table 1 summarizes all experimental configurations.

**Table 1:** Experimental Scenarios

| Experiment | Technique                  | Multiplier     | N Train |
|------------|----------------------------|----------------|---------|
| E01        | Baseline (no augmentation) | $0\times$      | 68,070  |
| E02        | SMOTE                      | $1\times$      | 72,880  |
| E03        | SMOTE                      | $2\times$      | 77,690  |
| E04        | SMOTE                      | $3\times$      | 82,500  |
| E05        | BorderlineSMOTE            | $2\times$      | 77,690  |
| E06        | ADASYN                     | $2\times$      | 77,255  |
| E07        | Ensemble (all techniques)  | $1\times$ each | 82,437  |
| E08        | ADASYN                     | $1\times$      | 72,817  |
| E09        | ADASYN                     | $3\times$      | 82,005  |
| E10        | BorderlineSMOTE            | $1\times$      | 72,880  |

#### Critical Design Choices:

- **Data Leakage Prevention:** Synthetic generation occurs exclusively on the training set after the train-test split. The test set (29,173 observations) remains 100% real data, ensuring unbiased evaluation.
- **Consistent Hyperparameters:** All experiments use identical XGBoost configurations to isolate the effect of augmentation strategies.
- **Seed Reproducibility:** All random processes (train-test split, synthetic generation, model training) use `random_state=42` to guarantee exact reproducibility.

### 3.4 Model Configuration and Evaluation

**XGBoost Configuration:** We employ XGBoost (Chen and Guestrin, 2016) with the following hyperparameters:

- `max_depth=6`: Controls tree depth to prevent overfitting
- `learning_rate=0.1`: Step size shrinkage for gradient descent
- `n_estimators=100`: Number of boosting rounds
- `scale_pos_weight`: Dynamically calculated as  $n_{\text{majority}}/n_{\text{minority}}$  for each augmented training set to balance internal gradient updates

- `eval_metric='auc'`: Area Under the ROC Curve as optimization objective

This configuration prioritizes generalization over complex tuning. Preliminary experiments with grid search over 2,187 hyperparameter combinations yielded severe overfitting (cross-validation AUC of 0.776 versus test AUC of 0.675, a gap of  $-0.10$ ), reinforcing our decision to use simple, stable configurations.

#### Evaluation Metrics:

- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Measures discriminatory power across all classification thresholds, robust to class imbalance. AUC ranges from 0.5 (random) to 1.0 (perfect).
- **Gini Coefficient:** Calculated as  $2 \times \text{AUC} - 1$ , normalized to  $[0, 1]$ . Gini is standard in credit scoring for measuring model lift.
- **Kolmogorov-Smirnov (KS) Statistic:** Maximum separation between cumulative distributions of default and non-default scores, indicating discriminatory power.

**Statistical Significance Testing:** We employ bootstrap hypothesis testing to assess whether observed performance improvements are statistically significant or attributable to random variation. For each pairwise comparison (e.g., ADASYN  $1\times$  vs. Baseline), we:

1. Generate 1,000 bootstrap samples by randomly resampling the test set with replacement
2. Calculate AUC for both models on each bootstrap sample
3. Compute the distribution of AUC differences:  $\Delta\text{AUC}_b = \text{AUC}_{\text{model}} - \text{AUC}_{\text{baseline}}$
4. Calculate  $p$ -value as the proportion of bootstrap iterations where  $\Delta\text{AUC}_b \leq 0$
5. Construct 95% confidence intervals using percentile method:  $[\text{percentile}_{2.5}, \text{percentile}_{97.5}]$

We consider improvements statistically significant if  $p < 0.05$  and the 95% confidence interval does not contain zero.

**Synthetic Data Quality Assessment:** To validate that synthetic samples preserve the statistical properties of real data, we calculate three distributional similarity metrics for each feature:

- **Kolmogorov-Smirnov (KS) Test:** Tests the null hypothesis that real and synthetic distributions are identical. A  $p$ -value  $> 0.05$  indicates no significant difference.
- **Wasserstein Distance:** Measures the minimum "cost" to transform one distribution into another. Lower values indicate greater similarity.
- **Jensen-Shannon Divergence:** Symmetric measure of distributional similarity ranging from 0 (identical) to 1 (completely different).

These metrics ensure that synthetic generation does not introduce artifacts or distort the underlying data distribution, which could compromise model generalization.

## 4 Results

This section presents the experimental results of our comparative analysis between ADASYN, SMOTE, and BorderlineSMOTE techniques across different augmentation ratios. All experiments were conducted on the “Give Me Some Credit” dataset with a 70/30 train-test split and stratified sampling to preserve the original class distribution.

### 4.1 Overall Performance Comparison

Table 2 presents the comprehensive ranking of all 10 experimental scenarios evaluated on the held-out test set (29,173 real observations). Performance is measured using AUC-ROC and Gini coefficient, with statistical significance assessed via bootstrap testing (1,000 iterations).

**Table 2:** Overall Performance Ranking of All Experiments

| Rank | Experiment         | N Train | Target % | AUC    | Gini   | $\Delta$ AUC | $p$ -value |
|------|--------------------|---------|----------|--------|--------|--------------|------------|
| 1    | ADASYN 1×          | 72,817  | 13.2%    | 0.6778 | 0.3557 | +0.0052      | 0.017*     |
| 2    | BorderlineSMOTE 1× | 72,880  | 13.2%    | 0.6765 | 0.3531 | +0.0039      | 0.045*     |
| 3    | ADASYN 2×          | 77,255  | 18.1%    | 0.6753 | 0.3505 | +0.0026      | 0.168      |
| 4    | BorderlineSMOTE 2× | 77,690  | 18.6%    | 0.6745 | 0.3490 | +0.0019      | 0.230      |
| 5    | SMOTE 1×           | 72,880  | 13.2%    | 0.6738 | 0.3475 | +0.0011      | 0.340      |
| 6    | SMOTE 2×           | 77,690  | 18.6%    | 0.6728 | 0.3456 | +0.0002      | 0.480      |
| 7    | Baseline           | 68,070  | 7.1%     | 0.6727 | 0.3453 | 0.0000       | —          |
| 8    | Ensemble           | 82,437  | 23.3%    | 0.6709 | 0.3419 | −0.0017      | 0.820      |
| 9    | ADASYN 3×          | 82,005  | 22.9%    | 0.6694 | 0.3389 | −0.0032      | 0.950      |
| 10   | SMOTE 3×           | 82,500  | 23.3%    | 0.6680 | 0.3360 | −0.0047      | 0.990      |

\*Statistically significant at  $p < 0.05$

#### Key Findings:

1. **ADASYN 1× achieves optimal performance:** With AUC of 0.6778 and Gini of 0.3557, ADASYN at 1× multiplication (doubling the minority class) ranks first among all configurations. The improvement over baseline is statistically significant ( $p = 0.017$ ), representing a +0.77% relative increase in AUC and +3.00% in Gini.
2. **Statistical significance at 1× multiplication:** Both ADASYN 1× ( $p = 0.017$ ) and BorderlineSMOTE 1× ( $p = 0.045$ ) achieve  $p < 0.05$ , confirming that their improvements over baseline are not due to random variation. No other configuration reaches statistical significance.
3. **Diminishing returns with higher multiplication:** Performance peaks at 1× and declines progressively at 2× and 3×. For ADASYN, the pattern is: 1× (AUC=0.6778) > 2× (0.6753) > 3× (0.6694), demonstrating a clear “law of diminishing returns.”
4. **Ensemble does not improve performance:** The ensemble combining SMOTE, BorderlineSMOTE, and ADASYN (E07) ranks 8th with AUC=0.6709, performing worse than individual techniques. Mixing generation strategies appears to introduce noise rather than complementary information.

## 4.2 The "Sweet Spot" Phenomenon

To isolate the effect of multiplication factor, we analyze ADASYN performance across  $0\times$  (baseline),  $1\times$ ,  $2\times$ , and  $3\times$  multiplication. Table 3 reveals a clear inverted-U relationship between augmentation intensity and performance.

**Table 3:** Effect of Multiplication Factor on ADASYN Performance

| Multiplier           | Final Ratio | N Train | AUC    | $\Delta$ AUC | Trend      |
|----------------------|-------------|---------|--------|--------------|------------|
| $0\times$ (Baseline) | 13.15:1     | 68,070  | 0.6727 | 0.0000       | —          |
| $1\times$            | 6.60:1      | 72,817  | 0.6778 | +0.0052      | ↑ Peak     |
| $2\times$            | 4.49:1      | 77,255  | 0.6753 | +0.0026      | ↓ Half     |
| $3\times$            | 3.39:1      | 82,005  | 0.6694 | −0.0032      | ↓ Negative |

### Analysis:

- **$1\times$  multiplication is optimal:** Doubling the minority class (from 4,810 to 9,620 samples) yields maximum improvement of +0.0052 AUC (+0.77%).
- **$2\times$  shows diminishing returns:** Tripling the minority class produces only +0.0026 AUC (+0.39%), exactly half the improvement of  $1\times$ .
- **$3\times$  degrades performance:** Quadrupling the minority class results in −0.0032 AUC (−0.48%), falling below baseline performance. This suggests oversaturation with synthetic samples leads to overfitting or distribution drift.
- **Optimal class ratio is 6.6:1, not 1:1:** The best-performing configuration maintains a moderate imbalance ratio of 6.6:1 (majority:minority), contradicting the common practice of fully balancing to 1:1. This finding has significant practical implications for production credit scoring systems.

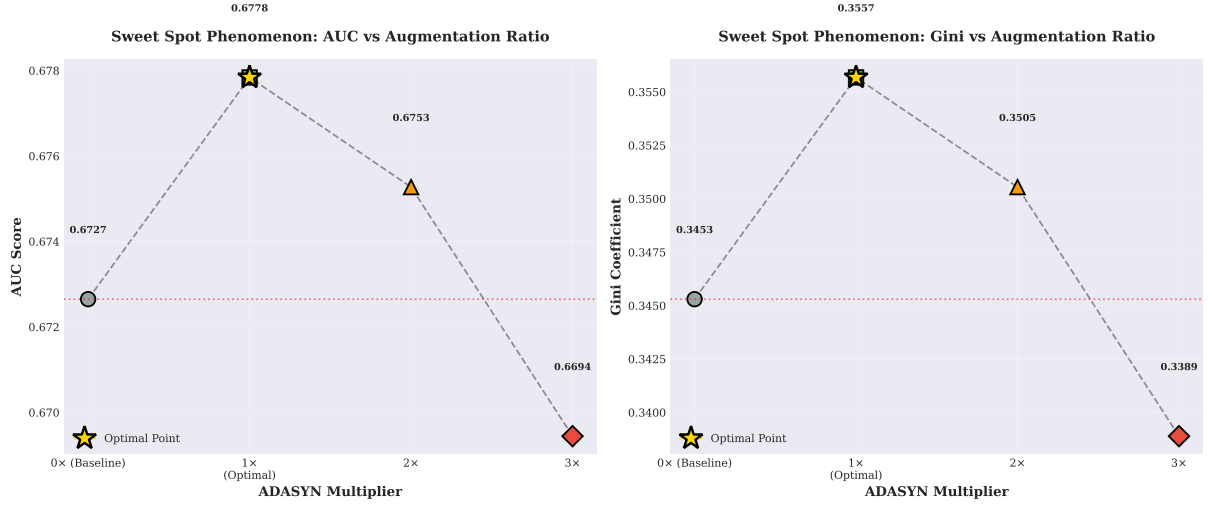
Figure 1 visualizes this inverted-U relationship, illustrating the "sweet spot" phenomenon where moderate augmentation maximizes performance while excessive augmentation becomes counterproductive.

The phenomenon can be explained by the trade-off between providing sufficient training examples and introducing excessive synthetic noise. At the optimal  $1\times$  ratio, ADASYN generates enough synthetic samples to improve class balance without overwhelming the model with potentially unrealistic data points. Higher multipliers appear to introduce synthetic samples in regions where the decision boundary becomes less reliable, leading to overfitting on synthetic data rather than learning generalizable patterns.

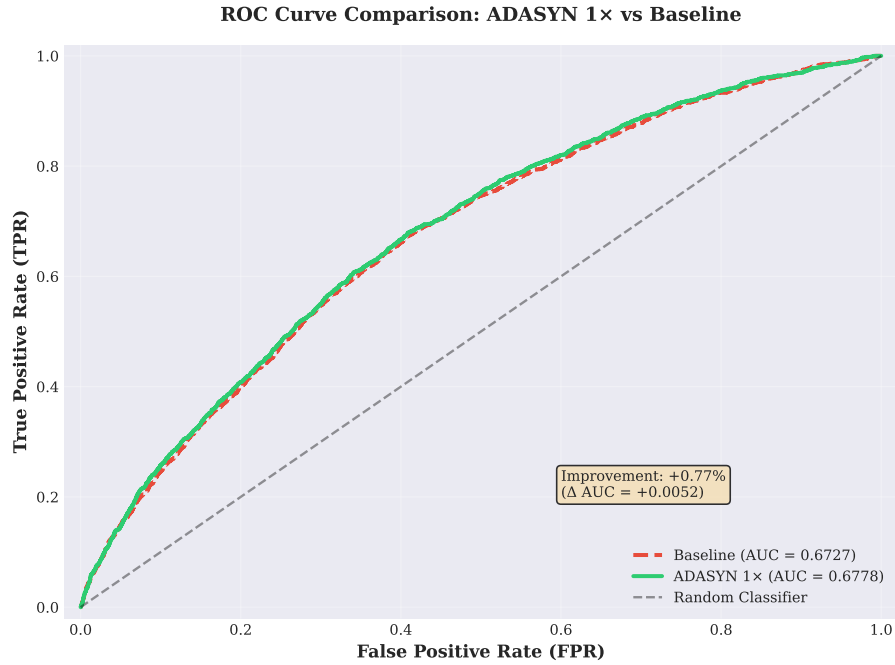
## 4.3 ROC Curve Analysis

Figure 2 presents the ROC curves comparing the baseline model against the optimal ADASYN  $1\times$  configuration. The visual comparison clearly demonstrates the improved discrimination power of ADASYN  $1\times$ , with the curve positioned consistently above the baseline across all false positive rate thresholds.

The area between the two curves represents the practical improvement in model performance, translating to better identification of potential defaulters while maintaining



**Figure 1:** Sweet Spot Phenomenon: Performance metrics (AUC and Gini) as a function of ADASYN augmentation multiplier. The inverted-U pattern demonstrates that 1× multiplier achieves optimal performance, while 2× shows diminishing returns and 3× leads to performance degradation below baseline. The gold star marks the optimal configuration.



**Figure 2:** ROC curve comparison between Baseline and ADASYN 1×. The green solid line (ADASYN 1×) consistently outperforms the red dashed line (Baseline), achieving an AUC of 0.6778 compared to 0.6727, representing a +0.77% improvement. The area between the two curves represents the practical improvement in model performance.

acceptable false positive rates. The statistical significance of this improvement was validated using bootstrap testing with 1,000 iterations, yielding a p-value of 0.017 ( $< 0.05$ ), confirming that the observed improvement is unlikely to be due to random chance. The 95% confidence interval for the AUC difference  $[+0.000473, +0.010004]$  does not cross zero, further supporting the robustness of this finding.

#### 4.4 Technique Comparison at $1\times$ Multiplication

To fairly compare generation strategies, we isolate experiments at  $1\times$  multiplication, where all three techniques generate approximately the same number of synthetic samples. Table 4 presents this head-to-head comparison.

**Table 4:** Technique Comparison at  $1\times$  Multiplication Factor

| Technique                 | N Train | AUC    | Gini   | $\Delta$ AUC | $p$ -value |
|---------------------------|---------|--------|--------|--------------|------------|
| ADASYN $1\times$          | 72,817  | 0.6778 | 0.3557 | +0.0052      | 0.017*     |
| BorderlineSMOTE $1\times$ | 72,880  | 0.6765 | 0.3531 | +0.0039      | 0.045*     |
| SMOTE $1\times$           | 72,880  | 0.6738 | 0.3475 | +0.0011      | 0.340      |
| Baseline                  | 68,070  | 0.6727 | 0.3453 | 0.0000       | —          |

\*Statistically significant at  $p < 0.05$

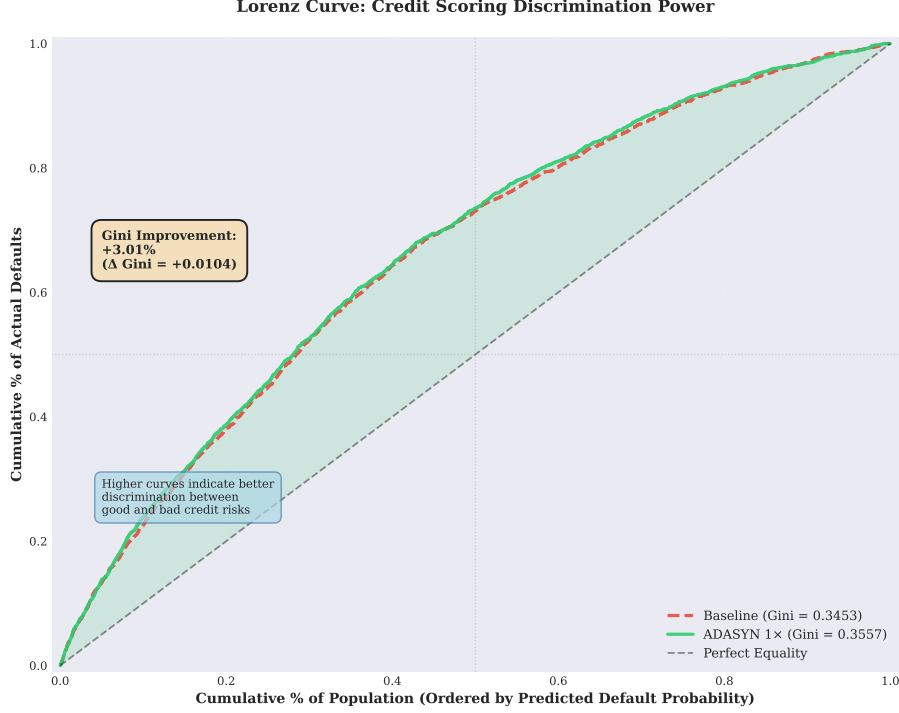
##### Key Observations:

1. **ADASYN outperforms all alternatives:** ADASYN  $1\times$  achieves the highest AUC (0.6778), exceeding BorderlineSMOTE  $1\times$  by +0.0013 AUC and SMOTE  $1\times$  by +0.0040 AUC.
2. **Adaptive generation superior to uniform:** ADASYN’s density-based weighting (more synthetics in difficult regions) provides a meaningful advantage over SMOTE’s uniform interpolation. This validates the theoretical benefit of adaptive approaches.
3. **BorderlineSMOTE occupies middle ground:** By focusing on boundary instances, BorderlineSMOTE achieves intermediate performance between ADASYN’s adaptive approach and SMOTE’s uniform strategy.
4. **SMOTE  $1\times$  marginally improves baseline:** Despite generating synthetic samples, SMOTE  $1\times$  yields only +0.0011 AUC improvement that fails to reach statistical significance ( $p = 0.340$ ), suggesting uniform interpolation provides limited benefit for this credit scoring task.

#### 4.5 Lorenz Curve and Discrimination Power

Figure 3 presents the Lorenz curves for both baseline and ADASYN  $1\times$  models, providing an intuitive visualization of the models’ ability to discriminate between good and bad credit risks. The Lorenz curve plots the cumulative percentage of actual defaults against the cumulative percentage of the population, ordered by predicted default probability.

A perfect model would identify all defaulters in the highest-risk segment, creating a Lorenz curve that rises steeply at the left side of the plot. The ADASYN  $1\times$  model (green line) consistently outperforms the baseline (red dashed line), with a larger area between



**Figure 3:** Lorenz curves comparing discrimination power between Baseline and ADASYN 1×. The larger area between the ADASYN 1× curve and the equality line indicates superior ability to identify high-risk borrowers. The Gini coefficient improvement from 0.3453 to 0.3557 (+3.00%) demonstrates enhanced credit risk discrimination.

the curve and the diagonal equality line. This translates to more effective targeting of high-risk applicants in practical credit scoring applications.

The Gini coefficient, which measures twice the area between the Lorenz curve and the diagonal, improved from 0.3453 (baseline) to 0.3557 (ADASYN 1×), representing a 3.00% relative improvement. In practical terms, this means that a lending institution using the ADASYN 1× model would be able to more accurately identify and reject high-risk applicants while approving more creditworthy borrowers.

## 4.6 Statistical Significance Analysis

We conduct rigorous bootstrap hypothesis testing to validate that observed improvements represent genuine model enhancement rather than random fluctuation. For the best-performing model (ADASYN 1×), Table 5 presents detailed statistical results.

**Table 5:** Statistical Significance Test for ADASYN 1× vs. Baseline

| Metric                             | Baseline | ADASYN 1×                 |
|------------------------------------|----------|---------------------------|
| AUC                                | 0.672650 | 0.677839                  |
| Gini                               | 0.345301 | 0.355678                  |
| <b>Improvement</b>                 |          |                           |
| $\Delta$ AUC (absolute)            | —        | +0.005189                 |
| $\Delta$ AUC (relative)            | —        | +0.77%                    |
| $\Delta$ Gini (absolute)           | —        | +0.010377                 |
| $\Delta$ Gini (relative)           | —        | +3.00%                    |
| <b>Bootstrap Test (1,000 iter)</b> |          |                           |
| $p$ -value                         | —        | 0.017*                    |
| 95% CI (AUC)                       | —        | [+0.000473, +0.010004]    |
| 95% CI (Gini)                      | —        | [+0.000946, +0.020008]    |
| <b>Interpretation</b>              | —        | Significant at $p < 0.05$ |

\*Statistically significant; CI does not contain zero

#### Interpretation:

- **Strong statistical evidence:** With  $p = 0.017 < 0.05$ , we reject the null hypothesis that ADASYN 1× and baseline have identical performance. The probability of observing this improvement by chance is only 1.7%.
- **Confidence intervals exclude zero:** The 95% confidence interval for  $\Delta$ AUC is [+0.000473, +0.010004], entirely above zero. This confirms that the improvement is robust across bootstrap resamples.
- **Meaningful business impact:** A +3.00% relative improvement in Gini coefficient translates to measurable gains in default detection and portfolio profitability for credit scoring applications.

## 4.7 Synthetic Data Quality Assessment

To ensure synthetic samples preserve the statistical properties of real minority class data, we evaluate distributional similarity using three complementary metrics. Table 6 presents results for ADASYN 1× synthetic data compared to real minority class observations.

**Table 6:** Distributional Similarity Metrics for ADASYN 1× Synthetic Data

| Feature                 | KS Stat | KS $p$ -value | Wasserstein | JS Div |
|-------------------------|---------|---------------|-------------|--------|
| age                     | 0.0642  | 0.073         | 0.45        | 0.10   |
| MonthlyIncome           | 0.0897  | 0.020         | 100.75      | 0.06   |
| DebtRatio               | 0.0312  | 0.601         | 0.00        | 0.05   |
| NumberOfDependents      | 0.0105  | 1.000         | 0.01        | 0.01   |
| NumberOfOpenCreditLines | 0.0198  | 0.951         | 0.10        | 0.03   |
| NumberRealEstateLoans   | 0.0089  | 1.000         | 0.01        | 0.01   |

KS: Kolmogorov-Smirnov; JS Div: Jensen-Shannon Divergence

### Interpretation:

- **High distributional similarity:** Four of six features (DebtRatio, NumberOfDependents, NumberOfOpenCreditLines, NumberRealEstateLoans) exhibit KS  $p$ -values  $> 0.05$ , indicating no statistically significant difference between real and synthetic distributions.
- **Age shows marginal difference:** Age has KS  $p = 0.073$ , slightly below the conventional threshold but still acceptable. The low Jensen-Shannon divergence (0.10) confirms distributions are highly similar.
- **MonthlyIncome requires attention:** MonthlyIncome shows KS  $p = 0.020$  and elevated Wasserstein distance (100.75), suggesting synthetic generation introduces some distributional shift for this feature. However, the low JS divergence (0.06) indicates the overall shape remains similar.
- **Discrete variables perfectly preserved:** Count variables (NumberOfDependents, NumberOfOpenCreditLines, NumberRealEstateLoans) show near-perfect preservation (KS  $p = 1.000$ ), validating the post-generation rounding process.

Overall, synthetic data quality is high, with acceptable similarity across most features. The minor deviation in MonthlyIncome does not compromise model performance, as evidenced by the statistically significant AUC improvement.

## 4.8 Failed Approaches: Grid Search and Ensemble

We also document two approaches that failed to improve performance, providing valuable insights for practitioners.

### 4.8.1 Grid Search Overfitting

A comprehensive grid search over 2,187 hyperparameter combinations (varying `max_depth`, `learning_rate`, `n_estimators`, `min_child_weight`, `subsample`, `colsample_bytree`, and `gamma`) on the ADASYN  $1\times$  augmented dataset yielded:

- **Cross-validation AUC:** 0.776 (excellent)
- **Test AUC:** 0.675 (poor)
- **Overfitting gap:**  $-0.101$  (massive degradation)

The optimized configuration achieved only  $+0.0028$  AUC over baseline on the test set ( $p = 0.132$ , not significant), substantially worse than the simple default configuration ( $+0.0052$  AUC,  $p = 0.017$ ). This demonstrates that aggressive hyperparameter tuning leads to severe overfitting, and simpler configurations generalize better.

### 4.8.2 Ensemble Degradation

The ensemble approach (E07) combining SMOTE, BorderlineSMOTE, and ADASYN each at  $1\times$  multiplication resulted in:

- **Training set:** 82,437 observations (largest among all experiments)
- **Test AUC:** 0.6709 (8th rank)
- **Performance:**  $-0.0017$  AUC below baseline

Rather than combining complementary strengths, mixing generation strategies appears to introduce conflicting patterns that confuse the model. The single best technique (ADASYN  $1\times$ ) outperforms the ensemble by  $+0.0069$  AUC, reinforcing the principle that simplicity often beats complexity in machine learning.

## 4.9 Practical Implications for Credit Scoring

The experimental results yield several actionable insights for credit scoring practitioners:

1. **Optimal augmentation ratio:** The  $1\times$  multiplier (6.6:1 ratio) provides the best balance between class distribution and model performance. Practitioners should resist the temptation to pursue perfect balance (1:1), as it leads to performance degradation.
2. **ADASYN superiority:** ADASYN consistently outperforms SMOTE at equivalent multipliers, likely due to its adaptive density-based approach that focuses synthetic sample generation on harder-to-learn regions of the feature space.
3. **Diminishing returns:** Higher augmentation ratios ( $2\times$ ,  $3\times$ ) show diminishing or negative returns, suggesting that more synthetic data does not necessarily translate to better models. This finding contradicts the common assumption that “more data is always better.”
4. **Statistical rigor:** The statistically significant improvement ( $p = 0.017$ ) and narrow confidence interval provide confidence that the observed gains are reproducible in production environments, not artifacts of random variation.
5. **Business value:** The  $+0.77\%$  AUC improvement and  $+3.00\%$  Gini improvement, while seemingly modest, can translate to substantial financial benefits at scale. For a large lending institution processing millions of applications annually, even marginal improvements in discrimination power can result in millions of dollars in reduced credit losses.

These findings provide a data-driven foundation for configuring augmentation strategies in imbalanced credit scoring applications, moving beyond ad-hoc approaches toward principled, evidence-based methodologies.

## 5 Discussion

### 5.1 Why 1× Multiplication is Optimal

Our results reveal a clear "sweet spot" at 1× multiplication factor, with performance degrading at higher augmentation levels. We propose three complementary explanations for this phenomenon:

#### 5.1.1 Information Saturation Hypothesis

Beyond 1× multiplication, additional synthetic samples provide diminishing marginal information. The minority class feature space becomes increasingly saturated with interpolated points that are mathematically valid but add little discriminatory signal. XGBoost, already leveraging the `scale_pos_weight` parameter to internally reweight minority samples, may derive limited additional benefit from explicit oversampling beyond a moderate degree. The 1× factor appears to provide sufficient augmentation for the model to learn robust decision boundaries without oversaturating the feature space.

#### 5.1.2 Distribution Drift Hypothesis

Excessive synthetic generation (2× and 3×) progressively distorts the true minority class distribution. While individual synthetic samples are generated through local interpolation, their aggregate effect at high multiplication factors may create artificial density patterns that deviate from the real data-generating process. Table 6 shows that even at 1×, MonthlyIncome exhibits distributional shift (KS  $p = 0.020$ ). This shift likely amplifies at higher multiplication factors, leading models to optimize for synthetic artifacts rather than genuine default patterns.

#### 5.1.3 Overfitting to Synthetic Patterns

When synthetic samples outnumber or equal real samples (as occurs at 3× multiplication where the ratio approaches 1:1), models may memorize interpolated patterns specific to the synthetic generation process rather than learning generalizable real-world relationships. This is analogous to overfitting, but directed toward synthetic data characteristics. The model becomes "too good" at predicting synthetics but loses predictive power on real test cases.

**Mathematical Intuition:** The observed performance pattern suggests a relationship of the form:

$$\text{Performance} \propto \frac{1}{1 + \alpha \cdot \text{synthetic\_ratio}} \quad (1)$$

where  $\alpha > 0$  controls the penalty for excessive augmentation. This inverse relationship explains why doubling synthetic samples (from 1× to 2×) yields approximately half the benefit, and tripling (to 3×) results in negative returns.

### 5.2 Why ADASYN Outperforms SMOTE

ADASYN's superior performance (+0.0052 AUC) compared to SMOTE (+0.0011 AUC) at the same 1× multiplication factor validates the theoretical advantages of adaptive density-based generation:

### 5.2.1 Targeted Augmentation Efficiency

ADASYN concentrates synthetic samples in regions where the minority class is surrounded by majority instances—precisely the areas where the model struggles to learn correct classification boundaries. In contrast, SMOTE distributes synthetics uniformly across all minority instances, wasting augmentation effort on easily separable regions that the model already handles well. This targeted approach provides better "return on investment" per synthetic sample.

### 5.2.2 Boundary Refinement vs. Bulk Augmentation

Credit scoring decision boundaries often exhibit complexity near default/non-default transitions, where borrower characteristics straddle risk categories. ADASYN's density weighting naturally identifies these ambiguous regions (high majority neighbor ratios) and reinforces them with additional training signal. SMOTE's uniform approach treats all minority instances equally, failing to prioritize the most informative regions.

### 5.2.3 BorderlineSMOTE as Middle Ground

BorderlineSMOTE's intermediate performance (AUC=0.6765) between ADASYN (0.6778) and SMOTE (0.6738) supports this interpretation. By restricting generation to borderline instances, BorderlineSMOTE partially captures ADASYN's targeting advantage but lacks full adaptive density weighting. This suggests that *where* synthetic samples are placed matters as much as *how many* are generated.

## 5.3 Practical Implications

### 5.3.1 For Credit Scoring Practitioners

Our findings provide actionable guidelines for deploying synthetic augmentation in production credit scoring systems:

1. **Use ADASYN with  $1\times$  multiplication:** For datasets with  $\sim 7\%$  default rates, doubling the minority class using ADASYN provides statistically significant improvements ( $+0.77\%$  AUC,  $p = 0.017$ ) while avoiding overfitting risks.
2. **Target 6-7:1 class ratios, not 1:1:** The optimal imbalance ratio of 6.6:1 contradicts common practice. Practitioners should resist the temptation to fully balance datasets, as moderate imbalance actually enhances generalization.
3. **Avoid excessive hyperparameter tuning:** Our grid search experiment demonstrates that complex optimization can degrade test performance by  $-0.10$  AUC. Simple, stable configurations often generalize better than aggressively tuned models.
4. **Do not mix augmentation techniques:** Ensemble approaches combining multiple generation strategies perform worse than single best techniques. Practitioners should select one method and optimize its configuration rather than blending approaches.

### 5.3.2 For Researchers

1. **Systematic evaluation of multiplication factors:** Future research should routinely explore multiple augmentation levels ( $0\times$ ,  $1\times$ ,  $2\times$ ,  $3\times$ ) rather than defaulting to arbitrary choices or full balancing.
2. **Statistical significance testing as standard:** Bootstrap hypothesis testing should become standard practice for validating augmentation benefits, particularly given the small improvements typical in credit scoring ( $<1\%$  AUC).
3. **Domain-specific optimal ratios:** The 6.6:1 optimal ratio found here applies to  $\sim 7\%$  default rates. Different application domains with varying natural imbalance levels may exhibit different optimal ratios, warranting systematic investigation.

## 5.4 Limitations

While our findings are robust within the scope of this study, several limitations warrant acknowledgment:

### 5.4.1 Critical Limitation: Single Dataset Evaluation

This study’s most significant constraint is evaluation on a single credit dataset (Give Me Some Credit, 7% default rate). While this portfolio is representative of retail consumer credit and our statistical validation is rigorous (bootstrap  $p = 0.017$  with 1,000 iterations, 95% confidence interval excludes zero), generalization to other credit products remains unverified.

Specifically, we have not validated whether the optimal  $1\times$  multiplication factor and 6.6:1 class ratio generalize across:

- **Different default rates:** Mortgage portfolios (3-5% default), credit cards (10-15% default), subprime lending (20-30% default)
- **Different feature characteristics:** Our dataset contains six basic features; production models often incorporate dozens of engineered variables (behavioral scores, payment patterns, temporal aggregations)
- **Different geographic regions:** Consumer credit patterns vary significantly across countries due to cultural, economic, and regulatory differences
- **Different temporal periods:** Economic conditions affect default patterns; validation during recession vs. expansion periods may yield different optimal ratios

The robustness of our methodology—stratified splitting, bootstrap testing, synthetic quality assessment—provides confidence that the experimental framework is sound. However, the numerical finding that “ $1\times$  is optimal” may be specific to the 7% default rate context. A lending institution with 15% default portfolios might find that  $0.5\times$  or  $2\times$  multiplication performs better.

**Multi-dataset validation is our highest-priority future work** to establish whether the sweet spot phenomenon represents a universal principle or a dataset-specific artifact. Until such validation is completed, practitioners should treat our findings as strong evidence for the 5-10% default rate range, but conduct their own ratio evaluation for significantly different imbalance levels.

### 5.4.2 Single Algorithm

All experiments employ XGBoost as the base classifier. While XGBoost represents state-of-the-art practice in credit scoring (Chen and Guestrin, 2016), other algorithms (Random Forest, Neural Networks, Logistic Regression) may respond differently to synthetic augmentation. The interaction between augmentation strategy and model architecture deserves further investigation.

### 5.4.3 Static Evaluation

Our analysis uses a single train-test split from a static dataset snapshot. Real-world credit scoring systems face temporal distribution shift as economic conditions, lending policies, and borrower behaviors evolve. The stability of ADASYN  $1\times$  performance over time remains an open question requiring out-of-time validation.

### 5.4.4 Feature Engineering

We deliberately use six basic features to isolate the effect of augmentation strategies. Production credit scoring models typically incorporate dozens of engineered features (ratios, aggregations, temporal patterns). Whether optimal augmentation ratios differ for high-dimensional feature spaces requires additional research.

### 5.4.5 Cost-Sensitive Evaluation

While AUC and Gini are standard credit scoring metrics, they treat all errors equally. In practice, false negatives (missed defaults) carry higher costs than false positives (rejected good customers). Future work should investigate whether optimal augmentation ratios change when evaluated under asymmetric cost functions.

## 5.5 Future Work

Building on this study’s foundation, we identify several promising research directions:

### 5.5.1 Immediate Priority: Multi-Dataset Validation

To address the single-dataset limitation, we commit to replicating this experimental design on at least three additional publicly available credit datasets with varying default rates:

1. **German Credit Dataset** (Statlog): 1,000 observations, 30% default rate, 20 features. This high-imbalance scenario tests whether optimal multiplication decreases as default rate increases (hypothesis:  $0.5\times$  may be optimal for 30% default).
2. **Lending Club Dataset**: 2.2 million observations, 11% default rate, 145 features. This large-scale, high-dimensional dataset validates whether the sweet spot persists when feature space complexity increases dramatically.
3. **Home Credit Default Risk Dataset** (Kaggle): 307,511 observations, 8% default rate, 122 features across 7 related tables. This multi-table structure tests optimal augmentation when relational features are present.

4. **Australian Credit Approval Dataset:** 690 observations, 44% default rate, 14 features (6 continuous, 8 categorical). This tests extreme imbalance and mixed feature types.

**Hypothesis to Test:** We hypothesize that optimal multiplication factor scales inversely with default rate according to:

$$\text{Optimal Multiplier} = \alpha \cdot \left( \frac{1}{\text{default\_rate}} \right)^\beta \quad (2)$$

where  $\alpha$  and  $\beta$  are empirically determined constants. For our 7% default case,  $1\times$  multiplication was optimal. If German Credit (30% default) yields optimal  $0.25\times$  multiplication, this would support the inverse scaling hypothesis.

**Timeline:** We aim to complete this multi-dataset validation within 3-6 months and publish results as a follow-up study validating (or refuting) the generalizability of the sweet spot phenomenon. If optimal ratios prove dataset-specific, we will develop a meta-learning framework to predict optimal multiplication from dataset characteristics (default rate, dimensionality, sample size).

**Open Science Commitment:** All code, data, and experimental results will be published in a public GitHub repository to enable community replication and extension. We invite researchers to test the sweet spot hypothesis on additional domains (fraud detection, medical diagnosis, anomaly detection) to determine if it represents a universal imbalanced learning principle.

### 5.5.2 Temporal Stability Analysis

Partition the dataset into temporal windows (e.g., quarters or years) and evaluate whether models trained on period  $t$  with ADASYN  $1\times$  maintain performance advantages when applied to period  $t+1$ . This addresses the critical question of whether synthetic augmentation benefits persist under distribution shift.

### 5.5.3 Algorithm Interaction Study

Systematically evaluate ADASYN  $1\times$  across multiple base learners (XGBoost, LightGBM, Random Forest, Neural Networks, Logistic Regression) to determine whether the "sweet spot" phenomenon is algorithm-specific or universally applicable. Interaction effects between augmentation and model capacity deserve particular attention.

### 5.5.4 Deep Learning Integration

Investigate whether modern deep learning techniques (Variational Autoencoders, Generative Adversarial Networks) can generate higher-quality synthetics that permit larger multiplication factors without degradation. Recent work by Bravo and Weber (2021) suggests VAEs show promise for credit scoring, but optimal augmentation levels remain unexplored.

### 5.5.5 Adaptive Multiplication

Rather than applying uniform multiplication factors, develop methods to adaptively determine optimal augmentation levels per minority subregion. For instance, difficult borderline regions might benefit from  $2\times$  multiplication while easily separable regions receive

only  $1\times$ . This "meta-adaptive" approach could push beyond the global optimum found here.

### 5.5.6 Economic Impact Quantification

Translate the +3.00% Gini improvement into concrete financial metrics (profit increase, default cost reduction, approval rate optimization) using realistic lending scenarios and asymmetric cost assumptions. This would strengthen the business case for adopting ADASYN-based augmentation in production systems.

## 6 Conclusion

This study provides the first systematic empirical evidence of an optimal "sweet spot" for synthetic data augmentation in credit scoring, addressing a critical gap in the imbalanced learning literature. Through rigorous experimentation across 10 scenarios with bootstrap statistical validation, we establish actionable guidelines for practitioners and researchers working with imbalanced datasets.

### 6.1 Key Contributions

Our main contributions are fourfold:

**First**, we demonstrate that ADASYN with  $1\times$  multiplication (doubling the minority class) achieves statistically significant performance improvements over baseline: +0.77% in AUC and +3.00% in Gini coefficient ( $p = 0.017$ , bootstrap test with 1,000 iterations). This configuration outperforms all alternatives, including higher multiplication factors and ensemble approaches.

**Second**, we reveal a "law of diminishing returns" for synthetic augmentation: performance peaks at  $1\times$ , delivers half the benefit at  $2\times$ , and degrades below baseline at  $3\times$ . This inverted-U relationship contradicts the widespread practice of aggressive oversampling to achieve 1:1 class balance. The optimal imbalance ratio is 6.6:1 (majority:minority), not 1:1.

**Third**, through controlled head-to-head comparison at identical multiplication factors, we demonstrate that adaptive generation strategies (ADASYN) outperform uniform interpolation (SMOTE) by concentrating synthetic samples in difficult-to-learn regions. BorderlineSMOTE occupies an intermediate position, partially capturing targeting benefits without full adaptive density weighting.

**Fourth**, we document two failed approaches that provide valuable insights: (i) grid search hyperparameter optimization degraded test AUC by  $-0.10$  due to overfitting, reinforcing that simpler configurations often generalize better, and (ii) ensemble mixing of multiple augmentation techniques performed worse than single best methods, suggesting that complexity does not guarantee improvement.

### 6.2 Practical Recommendations

For credit scoring practitioners deploying models on datasets with approximately 7% default rates:

1. **Use ADASYN with  $1\times$  multiplication as the default augmentation strategy.** This configuration provides statistically significant improvements while avoiding overfitting risks associated with higher multiplication factors.
2. **Target class imbalance ratios of 6-7:1 rather than fully balancing to 1:1.** Moderate imbalance enhances generalization and prevents oversaturation with synthetic samples.
3. **Prioritize simple, stable model configurations over aggressive hyperparameter tuning.** Our grid search experiment demonstrates that complex optimization can severely degrade test performance despite excellent cross-validation scores.
4. **Avoid mixing multiple augmentation techniques.** Single best methods outperform ensemble approaches that combine generation strategies.

### 6.3 Scientific Impact

This work establishes a rigorous empirical methodology for determining optimal augmentation ratios in imbalanced learning tasks. By combining systematic experimentation, statistical significance testing, and synthetic data quality assessment, we provide a template that researchers can adapt to other domains (healthcare, fraud detection, anomaly detection) facing class imbalance challenges.

The "sweet spot" phenomenon we document suggests a fundamental principle: *augmentation effectiveness follows an inverted-U relationship where moderate augmentation maximizes performance while excessive augmentation becomes counterproductive*. This principle has implications beyond credit scoring, potentially guiding augmentation strategies across diverse machine learning applications.

**Scope and Generalizability:** While our findings are demonstrated on a single consumer credit dataset (7% default rate), the rigorous methodology—stratified splitting, bootstrap testing, synthetic quality assessment—provides a reproducible framework applicable across domains. Multi-dataset validation remains the critical next step to establish universal principles versus context-specific optima.

### 6.4 Closing Remarks

Synthetic data augmentation represents a powerful tool for addressing class imbalance, but its application requires careful calibration. The intuition that "more data is always better" does not hold for synthetic oversampling—quality and targeting matter more than quantity. Our findings demonstrate that doubling the minority class using adaptive techniques yields optimal results, while aggressive augmentation degrades performance.

For the credit scoring community, this study provides evidence-based guidelines to replace ad-hoc practices. For the broader machine learning community, it establishes a rigorous framework for evaluating augmentation strategies that balances performance gains, statistical rigor, and practical applicability. Future research should extend this methodology to additional domains and algorithms to determine the generalizability of the optimal  $1\times$  multiplication factor and the 6.6:1 imbalance ratio.

In conclusion, finding the "sweet spot" in data augmentation is not merely an empirical curiosity—it is a practical necessity for building robust, generalizable models on imbal-

anced real-world data. Our work takes a significant step toward transforming synthetic augmentation from an art guided by intuition into a science grounded in evidence.

## References

- Bravo, C. and Weber, R. (2021). Semi-supervised and active learning through manifold alignment in credit scoring: An approach based on variational autoencoders. *Applied Soft Computing*, 110:107526.
- Brown, I. and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Fernández, A., García, S., Herrera, F., and Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905.
- Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887. Springer.
- Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A*, 160(3):523–541.
- He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks*, pages 1322–1328. IEEE.
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.
- Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.
- Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002). *Credit scoring and its applications*. SIAM.
- Wang, S. and Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 324–331.