# A Comparative Study on Loan Default Classification with Imbalanced Data Processing

Ziang Wang

Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario, L8S 4K1, Canada
wangz834@mcmaster.ca

**Abstract.** Credit risk default classification is a cornerstone of modern financial risk management, enabling institutions to optimize lending, allocate capital efficiently, and mitigate losses, with accurate predictions directly impacting financial system stability amid economic volatility. A critical challenge is data imbalance: default samples typically make up only 5–15% of datasets, biasing models toward the majority class and harming recall, the key metric for minimizing losses. This study compares four models (Logistic Regression, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Random Forest) combined with four imbalance-handling methods, using Accuracy, Recall, and F1 score as metrics. Results show tree-based models outperform Logistic Regression across metrics. For Logistic Regression, class weighting effectively improves recall; for tree-based models, class weighting boosts recall but slightly reduces F1, while Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) enhance F1 but risk noise. These findings highlight optimal strategies, with future work needed on ensemble methods and interpretability to refine credit risk assessment.

**Keywords:** Weight processing, Logistic Regression, Tree-based Models

## 1 Introduction

Credit risk default classification is a cornerstone of modern financial risk management, as accurate predictions of loan defaults enable financial institutions to optimize lending decisions, allocate capital efficiently, and mitigate potential losses [1]. In an era of increasing economic volatility, the ability to identify high-risk borrowers—those likely to default—directly impacts the stability of financial systems and the sustainability of credit markets.

Existing research has underscored the critical challenge of data imbalance in this domain: default samples typically constitute only 5–15% of real-world datasets, while non-default cases dominate [2]. This imbalance biases traditional machine learning models toward the majority class, leading to poor recall of defaults—the metric most critical for minimizing financial losses [3]. To address this, scholars have proposed various strategies: Chawla et al. introduced SMOTE, a synthetic oversampling technique that interpolates minority-class samples to balance datasets, though it risks


A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_33

overfitting [4]. To improve minority-class detection in decision trees, Elkan et al. defined cost-sensitive techniques that increase the misclassification costs for minority classes; and recent advancements like ADASYN, an extension of SMOTE, prioritize "hard-to-classify" minority samples to enhance model robustness [5, 6]. However, most studies focus on isolated methods rather than comparing hybrid strategies, and many rely on accuracy as the primary metric, overlooking the primacy of recall in credit risk assessment [7].

# 2 Methodology

## 2.1 Overview of the Process

This paper focuses on the task of loan default classification. After completing the feature engineering (including missing value filling, outlier processing, numerical feature standardization, and category feature encoding), a classification system is constructed by using multiple machine learning models combined with different data imbalance processing methods. By comparing the combined effects of different models and imbalanced samples processing methods, the optimal solution is explored.

## 2.2 Dataset Description

The dataset used is from Kaggle, named “credit-risk-dataset”, which provides a simulated world scenario for credit risk analysis [8]. It contains 32,564 loan records with 12 feature variables.

The key challenge lies in the imbalanced data, where only 26.3% of the samples are default cases. Traditional machine-learning models may be skewed towards the majority class (non-default) as a result of this imbalance, making it difficult to anticipate the minority class (default), which is crucial for evaluating credit risk.

Key features are: Age, Annual income, Home ownership, Loan intent, Loan grade, Loan amount, and Interest rate. The target feature is Loan status (0 is non-default, 1 is default).

## 2.3 Models

**Logistic Regression (LR)**. A model for linear classification. The likelihood that the sample is in the positive class (default) is determined by applying the Sigmoid function transformation to the linear regression findings. The probability is transformed into classification results by establishing a probability threshold, typically 0.5.

**XGBoost (XGB).** XGB is a tree model integration method based on the gradient boosting framework. It adopts the regularization boosting technique and introduces the regularization term in the loss function to control the model complexity [9].

**LightGBM (LGB)**. LGB also belongs to the gradient boosting tree family and adopts the histogram algorithm and the leaf-wise growth strategy to optimize the training efficiency. The core is to construct a histogram by discretizing the features to reduce the amount of calculation [10]. The objective function is similar to XGB, but it has better computing and storage efficiency and is suitable for large-scale data scenarios [11].

### 2.4 Weight Processing

**Baseline**. It is directly trained using the original data as a benchmark control to reflect the model's performance under the natural data distribution, without any data imbalance manipulation.

**Class Weighting**. Higher weights are given to minority classes and lower weights to majority classes to address sample imbalance.

**Synthetic Minority Over-sampling Technique (SMOTE).** SMOTE is an oversampling method that generates new samples by interpolating among samples of a few classes. The specific steps are as follows:

Firstly, for each minority class sample, calculate its k-nearest neighbors in the feature space (usually k=5). Then, randomly select samples from the nearest neighbor and generate new samples according to the following formula:

$$x_{new} = x_i + \lambda \cdot (x_j - x_i) \tag{1}$$

Here, $x_i$ are current minority class samples; $x_j$ are neighbor samples; $\lambda \in [0,1]$ is a random interpolation coefficient. Balance the data distribution by expanding the number of minority class samples.

**Adaptive Synthetic Sampling (ADASYN)**. An enhanced SMOTE, ADASYN, adaptively creates minority class samples according to sample difficulty. First, calculate the "difficulty level" of the samples (usually based on the proportion of samples of the majority class in the k-nearest neighbors). Then allocate the number of generated samples according to the difficulty level. The formula is:

$$g_i = \frac{\gamma_i}{Z} \times (n_{majority} - n_{minority}) \tag{2}$$

$i$ is class index; $\gamma_i$ is class imbalance ratio for class $i$,

$$\gamma_i = \frac{n_i}{k} \tag{3}$$

$n_i$ is the number of samples of the majority classes in the k-nearest neighbor. $Z$ is normalized coefficient.

$$Z = \sum_{i=1}^{k} \gamma_i \tag{4}$$

Smaller $\gamma_i$ will generate more new samples and enhance the model's recognition ability for difficult-to-classify minority classes. Then perform the same $x_{new}$ formula as SMOTE $g_i$ times for each minority class sample $x_i$.

### 2.5 Evaluation Metrics

Recall. The model's capacity to recognize minority classes is gauged by the recall rate. The recall effect of the model on default samples is better when the value is nearer 1.

F1 Score. The model's overall classification performance is reflected in the F1 score, which is the harmonic average of precision and recall. The model performs better in terms of balancing accuracy and Recall rate when the F1 value is higher.

## 3 Results

### 3.1 The Performance of SMOTE and ADASYN on Data

The influence of SMOTE and ADASYN on the category distribution of the loan default dataset (after feature engineering) can be visually observed through the bar chart.

Before sampling: As the left graph shows, the number of non-default samples (17,770) was much higher than that of default samples (5,024), and the category imbalance ratio was approximately 3.5:1, which was in line with the realistic characteristic of " German Credit Data" in some abnormal financial scenarios [3].

After SMOTE: As Fig. 1(b) shows, the defaulted samples were expanded to 17,770 through synthetic interpolation, which was completely balanced with the number of non-defaulted samples. This is consistent with the design goal of the SMOTE algorithm of "balancing categories through interpolation in the feature space" [12].

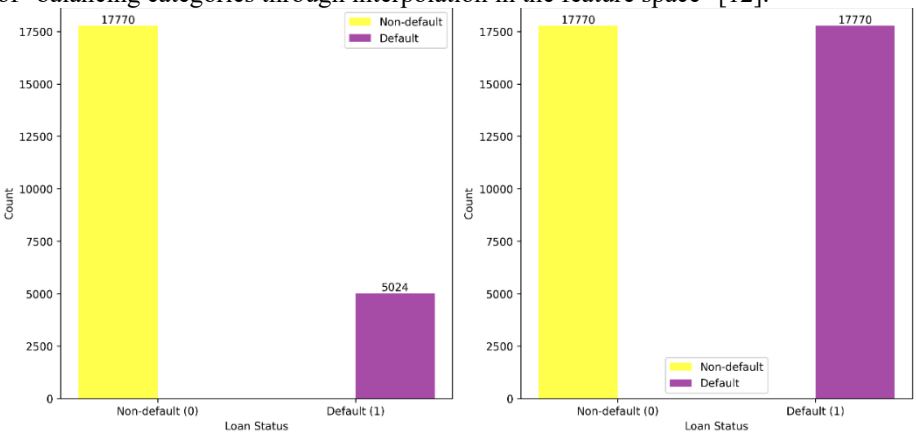

**Fig. 1.** Number of samples before and after SMOTE; (a) Loan Status Distribution Before Oversampling; (b) Loan Status Distribution After SMOTE (Picture credit: Original)

After ADASYN: As Fig. 2(b) shows, the number of defaulted samples was enhanced to 17,809 (slightly more than that of non-defaulted samples), reflecting the characteristic of ADASYN of "adaptive sampling for 'difficult-to-classify' minority samples", i.e., more synthetic data will be generated on the samples which are surrounded by the majority [13].

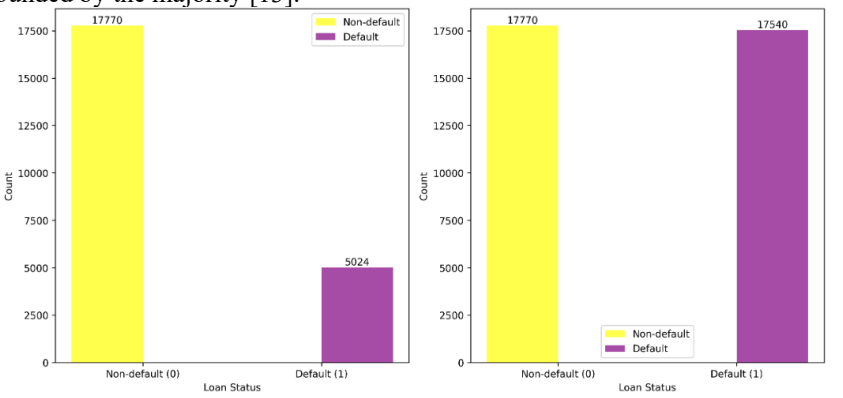


**Fig. 2.** Number of samples before and after ADASYN, (a) Loan Status Distribution Before Oversampling; (b) Loan Status Distribution After ADASYN (Picture credit: Original)

### 3.2 Key Observations Across Metrics

Figs. 3-5 show the results of model scoring. The y-axis on each graph represents accuracy, recall, and F1, while the x-axis shows the methods of handling imbalanced samples. The blue line represents LR, the green line represents XGB, the red line represents LGB, and the purple line represents RF.

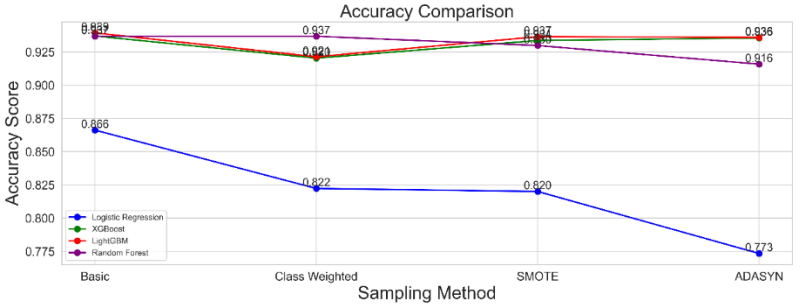


**Fig. 3.** Accuracy Comparison Graph (Picture credit: Original)


**Fig. 4.** Recall Comparison Graph (Picture credit: Original)

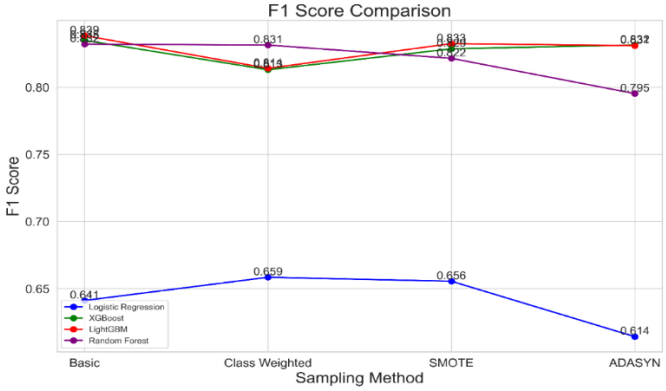


**Fig. 5.** F1 Score Comparison Graph (Picture credit: Original)

**Model Performance Overview**. From Figs. 3-5, tree-based models (XGB, LGB, RF) outperform LR across all metrics. Boosting and RF achieve the best Accuracy and F1 score, aligning with the literature that tree-based models handle non-linear relationships in financial data better [3]. RF shows strong Recall performance but less improvement with over-sampling methods on Accuracy and F1. This is consistent with Barandela et al., who noted ensemble methods are less sensitive to imbalanced data [2].

LR struggles with sample imbalance: F1 scores remain 0.6~0.65, and AUC = 0.87 (lowest among models). This reflects its reliance on linear assumptions, which fail to capture complex risk patterns [14]. However, the performance in terms of Recall is encouraging, with an obvious growth curve.

**Effect of Imbalanced Sample Processing Methods**. For XGB/LGB, Class Weighting boosts Recall (XGB Weighted Recall = 0.812; LGB Weighted Recall = 0.806) but slightly harms F1 (XGB Weighted F1 = 0.813; LGB Weighted F1 = 0.813). This tradeoff between Recall and Precision is consistent with cost-sensitive learning theory [5].

For LR, Class Weighting marginally improves Recall (from 0.554 to 0.775). This indicates that the Class Weighting method is more effective when applied to LR and is more suitable for loan scenarios that particularly require attention to the Recall rate.

SMOTE improves F1 for most models (XGB SMOTE F1 = 0.832; LGB SMOTE F1 = 0.831) but slightly reduces AUC for XGB/LGB (from 0.95 to 0.94). This matches Chawla et al., who warned SMOTE may introduce noisy synthetic samples [4].

ADASYN performs similarly to SMOTE for tree-based models, but has the most on LR (Precision drops around 0.05). This aligns with He et al.'s discovery, which noted ADASYN's focus on "hard" samples can amplify noise in weak models [6].

## 4 Conclusion

The problem of class imbalance in loan default classification is the main topic of this study. By comparing the performance of four models (Logistic Regression, XGBoost, LightGBM, and Random Forest) with four imbalanced sample processing methods (basic method, Class Weighting, SMOTE, and ADASYN). Key findings reveal that the performance of the four models varies after being processed with different imbalanced methods. Notably, the performance of Logistic Regression shows the most obvious changes, with its Recall increasing alongside the application of imbalanced sample processing methods.

It is worth emphasizing that the Class Weighting method stands out as the only effective approach for LR, proving highly useful when applied to this model. In loan scenarios, people may tend to pay excessive attention to over-sampling methods while overlooking this simple yet effective technique. A conjectured reason for this is that, unlike over-sampling methods, which modify the original data, Class Weighting adjusts the loss function. This adjustment avoids introducing "fake" samples and thus preserves the integrity of risk patterns inherent in real data.

When considering the tradeoffs of sampling methods, it is found that SMOTE and ADASYN can boost the F1 score of tree-based models but may introduce noise, as indicated by a decrease in Recall. Interestingly, their effect on LR is opposite.

This study places greater emphasis on addressing sample or quantity imbalance, while paying less attention to feature-level issues—such as the sparsity of risk-related features in high-dimensional data. Additionally, the interpretability of the models employed remains insufficient.

Moving forward, one key area for exploration involves further testing the Class Weighting method on loan default classification data with a more extensive set of features, while also conducting comparisons with over-sampling methods. In addition, integrating business rules into over-sampling methods could serve to ensure that the synthetic samples generated are more realistic and aligned with real-world scenarios. Moreover, leveraging tools like SHAP or LIME to interpret "latent default" decisions holds promise, as this could provide valuable insights to aid in the design of more effective risk-control rules.

## References


1. Siddiqi, N.: Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Wiley, Hoboken (2006)
2. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36(4), 849–851 (2003)
3. Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 297(1), 1–22 (2022)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
5. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), pp. 973–978 (2001)
6. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE Transactions on Neural Networks and Learning Systems 31(3), 1178–1191 (2020)
7. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring datasets. Expert Systems with Applications 39(3), 3446–3453 (2012)
8. Kaggle: Credit Risk Dataset. https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data, last accessed 2024/08/04
9. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, New York (2020)
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154 (2017)
11. Biau, G., Scornet, E.: Random forests and their variants. Annual Review of Statistics and Its Application 8, 457–486 (2021)
12. Douzas, G., Bacao, F.: Geometric SMOTE for imbalanced datasets: A robust oversampling method. Expert Systems with Applications 176, 114798 (2021)
13. Liu, Y., Li, X., Zhang, Z.: A hybrid sampling algorithm combining M-SMOTE and ENN for imbalanced data. Expert Systems with Applications 185, 115633 (2021)
14. Zhang, J., Li, X.: Limitations of linear models in financial risk prediction: An empirical study. Journal of Risk Model Validation 17(1), 1–18 (2023)