A Comparative Analysis of Machine Learning Models for Predicting Loan Defaults under Imbalanced Data Conditions

Yi Zhou

Department of Mathematics, Imperial College London, London, UK
mz2023@ic.ac.uk

Abstract. Accurate loan default prediction is crucial for credit risk management. This study used Kaggle's Loan Default Prediction dataset, applying preprocessing (cleaning, encoding, scaling) and testing six models: Logistic Regression, Decision Trees, Random Forest, eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and a neural network. Performance was evaluated using precision, recall, F1-score, and Receiver Operating Characteristic (AUC-ROC), with emphasis on handling class imbalance. Gradient boosting (especially LightGBM, AUC ~0.76) outperformed linear and tree-based models. Adjusting XGBoost's decision threshold improved the minority class F1-score from 0.17 to 0.36 without losing Area Under the Curve (AUC). Ensemble methods like Voting Classifiers balanced recall and precision effectively. Key takeaways include model selection, threshold tuning, and ensemble strategies for imbalanced data. Future work could explore richer features, imbalance-aware loss functions, and explainability tools for transparency. This study provides a reproducible benchmark for default prediction, laying the groundwork for more robust, interpretable, and fair credit scoring systems in real-world financial applications.

Keywords: LightGBM, XGBoost, Predicting Loan Defaults

1 Introduction

Many people apply for bank loans as a result of the banking industry's recent improvements and the rising trend of loan taking. The rising rate of loan defaults, however, is one of the main issues facing the banking industry in this dynamic economy. It is becoming increasingly challenging for banking regulators to accurately evaluate loan applications and address the risks of loan default [1, 2].

Recent studies have highlighted the strong potential of machine learning in predicting loan defaults, especially under imbalanced data conditions. Techniques such as Random Forest, XGBoost, and LightGBM, when combined with SMOTE or threshold tuning, have shown significant improvements in identifying minority (default) cases without sacrificing overall accuracy [3-6]. Neural networks are increasingly adopted for their ability to model complex borrower behavior, though they often require careful

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_25

tuning and remain less interpretable [7, 8]. Despite the popularity of these methods, few works systematically compare classical, ensemble, and deep models under a unified preprocessing and evaluation framework. This motivates the work: to benchmark a variety of classifiers under consistent conditions and explore how preprocessing, model architecture, and imbalance strategies interact to impact predictive performance. To illustrate, it finds that adjusting the decision threshold of XGBoost improved the minority class (default) F1-score from 0.17 to 0.36, while maintaining AUC at 0.75. Such results highlight how proper tuning of imbalance-handling strategies can yield significant practical gains.

In this study, a supervised learning pipeline is developed to predict loan defaults using a publicly available dataset containing borrower-level information. The research evaluates multiple classification models, including logistic regression, decision trees, XGBoost(eXtreme Gradient Boosting), and LightGBM(Light Gradient Boosting Machine), while addressing data imbalance challenges through techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and decision threshold adjustment. The objective is to establish a reliable predictive baseline and examine how model choice and imbalance handling affect performance.

2 Methodology

2.1 Data Preprocessing

The dataset used in this study is the publicly available Loan Default prediction Dataset, obtained from Kaggle [2]. It contains borrower-level information, including numerical and categorical variables, and a binary target variable indicating loan default.

A comprehensive preprocessing pipeline was applied to ensure data quality before model training. Missing values were imputed with appropriate statistics, median for numerical attributes and mode for categorical ones, while outliers and inconsistent data types were mitigated through winsorisation and type - casting. All categorical features, including binary indicators, were transformed via label encoding, and continuous variables were standardised (z-score) to stabilise model convergence. After cleaning, the dataset was stratified into training and test subsets using an 80:20 split, enabling unbiased supervised learning.

2.2 Evaluation Metrics

To evaluate the performance of the models in predicting loan default, the research uses several standard classification metrics. Accuracy measures the proportion of correct predictions among all predictions but can be misleading in imbalanced datasets, where the majority class dominates. To address this, the research also considers precision, which indicates the proportion of correctly predicted defaults among all predicted defaults, and recall, which reflects the proportion of actual defaults that were correctly identified. These two metrics are combined in the F1-score, the harmonic mean of precision and recall, offering a balanced measure when false positives and false negatives

are both costly. In addition, the Area Under the ROC Curve (AUC-ROC), which assesses the model's ability to distinguish between the default and non-default classes across all classification thresholds, is reported. A higher AUC indicates better discrimination capability and is particularly useful when the class distribution is highly skewed. These metrics provide a comprehensive view of model performance under both balanced and imbalanced evaluation criteria.

2.3 Different Models

Logistic Regression. Logistic Regression is a widely used baseline model for binary classification problems and is particularly suitable for credit risk tasks such as predicting loan default. In this context, the goal is to estimate the probability that a borrower will default on a loan given their individual financial and demographic characteristics (e.g., income, loan amount, education level).

After calculating a linear combination of the input features, the model maps the result into a probability between 0 and 1 using the sigmoid function.

The model is trained by minimizing the binary cross-entropy loss function, which penalizes the difference between predicted probabilities and actual class labels.

Logistic regression is well-suited to financial applications because of its interpretability. Each coefficient indicates the direction and strength of the influence of a feature on the log-odds of default. However, the model assumes linearity in the log-odds, which may limit its capacity to capture complex feature interactions compared to non-linear models such as decision trees or neural networks.

Decision Tree and Random Forest. To model the likelihood of loan default using borrower-level attributes, `it` employs Decision Tree and Random Forest classifiers. The input features include numerical variables such as Income, Loan Amount, Credit Score, Months Employed, Interest Rate, and categorical variables such as Education, Employment Type, Marital Status, and Has CoSigner, all of which may influence a borrower's ability to repay. The target variable is the binary indicator Default, where 1 represents default.

The Decision Tree classifier constructs a hierarchical model by iteratively splitting the dataset based on feature values that minimize a selected impurity measure—commonly the Gini index, calculated as $\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2$, where p_{i} is the proportion of samples belonging to class i in a node. This allows the tree to isolate homogeneous groups with respect to the default status.

To improve predictive performance and reduce overfitting, the Random Forest model is also implemented, which builds an ensemble of decision trees using bootstrap samples and random feature subsets. The final prediction is obtained through majority voting across all T trees $\hat{y} = \text{mode}\big(h_1(x), h_2(x), \dots, h_T(x)\big)$ where $h_{\text{t}}(x)$ denotes the prediction of the $t - th$ tree on input x. Random Forests offer improved generalization by reducing variance and capturing complex interactions among features.

Both models are evaluated using stratified train-test splits and assessed on classification metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC, providing a robust comparison in modeling the probability of loan default.

Gradient Boosting (XGBoost and LightGBM). This study implements two widely used gradient boosting algorithms, XGBoost and LightGBM, to predict the likelihood of loan default. Both methods are ensemble learning techniques that construct an additive model by sequentially combining multiple shallow decision trees.

XGBoost aims to minimize a regularized objective function, which balances prediction accuracy and model complexity.

LightGBM improves upon traditional boosting methods by adopting a leaf-wise tree growth strategy with depth constraints, which allows it to focus on minimizing loss more aggressively. It also employs histogram-based split finding and supports native categorical feature processing, which enhances efficiency, especially for large datasets.

Both models are trained using borrower-level features, such as Income, Loan Amount, Credit Score, Employment Type, and Has CoSigner. Categorical variables are label-encoded (in XGBoost) or directly handled by LightGBM. Hyperparameters, including learning rate, number of estimators, maximum depth, and subsample ratio, are tuned through cross-validation. These models are especially effective in imbalanced classification settings due to their ability to integrate class weighting and threshold tuning.

2.4 Handling Imbalance Data

The dataset used for predicting loan defaults exhibits a major class imbalance, where the majority of borrowers are labeled as non-defaulters, with only a small fraction defaulting. This skew creates a challenge for supervised learning models, which tend to focus on the dominant class and may struggle to correctly identify the minority default cases. To address this, several strategies are employed to balance the dataset. First, the Synthetic Minority Oversampling Technique (SMOTE) is applied, which creates new synthetic samples for the default class by interpolating between existing default instances. This approach helps balance the class distribution without simply duplicating data [9]. Next, for models like Logistic Regression, Decision Trees, XGBoost, and LightGBM, it incorporates class weights into the loss function to assign higher importance to misclassified default cases. This adjustment encourages models to pay more attention to the minority class during training. Finally, the classification threshold is optimized by testing various probability cutoffs to find the one that maximizes the F1-score or appropriately balances recall and precision. These methods are consistently applied across all models, enabling fair comparisons and greatly enhancing the models' ability to identify high-risk borrowers in an imbalanced setup.

3 Result

Table 1 and 2 report the test-set performance of all 12 model variants on the loan default

prediction task, including class–specific precision, recall and F_1-scores, overall accuracy, and AUC.

Table 1. Test set performance of all model variants (part 1)

Model	Precision(0)	Precision(1)	Recall(0)	Recall(1)
Logistic Regression	0.94	0.22	0.67	0.7
Decision Tree	0.9	0.2	0.88	0.23
Random Forest	0.94	0.22	0.69	0.68
LightGBM	0.94	0.23	0.69	0.69
XGBoost (SMOTE)	0.89	0.51	0.99	0.1
XGBoost (Adj Thresh=0.3241)	0.93	0.29	0.84	0.49
MLP (sklearn)	0.89	0.6	0.99	0.07
MLP (SMOTE)	0.94	0.18	0.56	0.74
MLP (RandomOverSample)	0.95	0.17	0.48	0.82
Voting Classifier	0.92	0.32	0.88	0.42
Voting (SMOTE)	0.91	0.35	0.93	0.31
Neural Net (TF)	0.93	0.27	0.81	0.53

Table 2. Test set performance of all model variants (part 2)

Model	Precision(0)	Precision(1)	Recall(0)	Recall(1)
Logistic Regression	0.94	0.22	0.67	0.7
Decision Tree	0.9	0.2	0.88	0.23
Random Forest	0.94	0.22	0.69	0.68
LightGBM	0.94	0.23	0.69	0.69
XGBoost (SMOTE)	0.89	0.51	0.99	0.1
XGBoost (Adj Thresh=0.3241)	0.93	0.29	0.84	0.49
MLP (sklearn)	0.89	0.6	0.99	0.07
MLP (SMOTE)	0.94	0.18	0.56	0.74
MLP (RandomOverSample)	0.95	0.17	0.48	0.82
Voting Classifier	0.92	0.32	0.88	0.42
Voting (SMOTE)	0.91	0.35	0.93	0.31
Neural Net (TF)	0.93	0.27	0.81	0.53

Table 1 and Table 2. Test-set performance of various classifiers. Prec(0)/Rec(0)/F_1(0) refer to the non-default class; Prec(1)/Rec(1)/F_1(1) to the default class.

Overall, the three gradient-boosting models—XGBoost, LightGBM and Random Forest, offer the best ability to separate defaulters from non-defaulters. Each achieves an AUC of about 0.75 and default recall near 0.69, confirming that boosting methods often beat linear models on unbalanced credit data. Logistic Regression also performs well, with AUC = 0.75 and overall accuracy of 0.67. The single Decision Tree has the highest recall for non-defaulters (0.88) but it rarely identifies defaulters (recall = 0.23, AUC = 0.55), a sign of its tendency to overfit.

The imbalance-handling experiments reveal clear trade-offs. Simple oversampling can introduce noise, as demonstrated by the fact that SMOTE oversampling raises non-default recall for both XGBoost and the MLP to 0.99 while default recall drops to 0.10 and 0.07. In contrast, changing the decision threshold of XGBoost maintains the AUC at 0.75 while increasing the default recall to 0.49 and the F_1 score to 0.36. With default

accuracy = 0.32, default recall = 0.42, F_1 = 0.37, and AUC = 0.76, a voting ensemble outperforms defaulters, demonstrating the value of combining models.

The TensorFlow neural network captures nonlinear patterns as well, achieving AUC = 0.74 and default recall = 0.53 (F_1 = 0.35), but it requires careful tuning and regularization. In summary, boosting models provide the strongest baseline. To improve detection of the minority class, it is more effective to tune thresholds or use ensembles than to rely on oversampling alone, which can harm overall discrimination.

4 Discussion

4.1 Model Strength and Performance

The experiments demonstrate that ensemble models—particularly LightGBM and XGBoost—outperform traditional classifiers such as logistic regression and decision trees in both AUC and default recall. This is consistent with findings by Nnaji and Imoudu and Yang et al., who highlight the effectiveness of gradient boosting on credit datasets with imbalanced classes [3, 4]. Neural networks also showed competitive performance, especially in default recall (e.g., 0.53 for our TensorFlow model), in line with results from Alghamdi et al. [7]. Recent studies also highlight the value of combining explainability with ensemble methods in credit risk modeling [10].

4.2 Suggestions for Optimization

While boosting models performed well, integrating explainability tools like SHAP can enhance transparency, as shown in Monje et al. [11]. Further improvements may include combining SMOTE with under-sampling to reduce oversampling noise, or using dynamic thresholding methods to balance precision and recall more effectively. Chen et al. demonstrated that hybrid sampling with adaptive thresholds significantly boosts classification performance in skewed credit risk tasks [5].

4.3 Limitations of this Study

The study is limited by its reliance on a single static dataset. It lacks dynamic borrower behavior data, such as payment timelines or transaction sequences. Moreover, the neural network architecture used was relatively simple and did not incorporate advanced designs like attention mechanisms or time-series modeling. The work also focused on offline batch modeling without testing for model drift in real-time environments.

4.4 Future Research Direction

Future work could expand the feature set with behavioral or alternative credit indicators, aligning with suggestions from Zhang and Li who observed performance gains from richer datasets [8]. There is also potential to explore more complex neural architectures, such as recurrent neural networks or transformer-based models. Additionally,

developing cost-sensitive or focal loss functions tailored to loan defaults, especially under extreme imbalance, could further improve model robustness. Finally, systematic comparisons under varying regulatory or business constraints would enhance the applicability of ML models in real-world lending.

5 Conclusion

This study presented a structured benchmarking framework for predicting loan defaults using various machine learning models under imbalanced data conditions. By implementing a unified preprocessing pipeline and systematically applying imbalance-handling strategies such as SMOTE, class-weighting, and threshold tuning, it ensured a fair and reproducible evaluation of six classification models.

Rather than focusing solely on accuracy, the analysis emphasized metrics sensitive to minority class detection—such as recall, F1-score, and AUC—providing a more realistic assessment of model suitability in credit risk tasks. Gradient boosting models, particularly LightGBM and XGBoost, proved to be consistently strong performers. Additionally, threshold optimization and ensemble voting further enhanced minority class detection without sacrificing overall discrimination.

Importantly, the work goes beyond isolated model comparisons by establishing a consistent evaluation environment that integrates both classical and modern architectures. This not only clarifies the trade-offs between interpretability and performance but also offers a practical reference for researchers and practitioners dealing with imbalanced classification problems in finance.

Future research can build on this foundation by introducing real-time features, advanced neural designs, and interpretable modeling components to support transparent, high-stakes decision-making in lending systems.

References

1. Madaan, M., Kumar, A., Keshri, C., Jain, R., Nagrath, P.: Loan default prediction using decision trees and random forest: A comparative study. IOP Conference Series: Materials Science and Engineering **1022**(1), 012042 (2021)
2. Kaggle: Loan Default Dataset. https://www.kaggle.com/datasets/nikhil1e9/loan-default, last accessed 2025/08/05
3. Nnaji, G., Imoudu, E.: Machine learning models for credit risk assessment. Journal of Financial Data Science (2023)
4. Yang, Z., Li, M., Xu, Y.: An end-to-end machine learning pipeline for p2p loan default prediction. In: AI Finance Review, Springer, Heidelberg (2024)
5. Chen, W., Liu, J., Yu, K.: Hybrid sampling and ensemble learning for credit scoring. electronics (MDPI) (2023)
6. Chen, T., Zhang, H.: Enhanced SMOTE in highly imbalanced credit risk data. ACM Transactions on Knowledge Discovery from Data (TKDD) (2022)
7. Alghamdi, F., Yousif, A., Iqbal, M.: Deep neural networks for credit scoring: performance and limitations. IEEE Access (2023)

8. Zhang, J., Li, R.: Comparing XGBoost and deep learning in imbalanced credit risk prediction. In: Proceedings of the AAAI Finance Workshop 2022, pp. 1–10. AAAI Press, Palo Alto (2022)
9. Liu, X., Jiang, Y., Zhao, J.: A SMOTE-based hybrid sampling method for imbalanced credit scoring. Expert Systems with Applications **198**, 116840 (2022)
10. Wang, J., Chen, L., Xu, Y.: Explainable ensemble learning for credit default prediction. IEEE Transactions on Computational Social Systems **10**(2), 478–489 (2023)
11. Monje, A., Liu, X., Han, R.: Interpretable boosting models in financial credit risk. In: AI Finance Review, Springer, Heidelberg (2025)