Dean&Francis
ISSN 2959-6130

Credit Card Default Prediction with Machine Learning: A Benchmarking Study on Imbalance Handling and Model Interpretability

Shuangyi Liu

Faculty of Science, University of Sydney, Sydney, Australia
sliu0414@uni.sydney.edu.au

Abstract:

Credit card default risk prediction is crucial for financial institutions to mitigate potential losses and ensure regulatory compliance. This paper addresses the challenge of imbalanced data and model interpretability in predicting default using the University of California, Irvine (UCI) Credit Card dataset. Experiments were conducted on a dataset of 30,000 clients, and feature engineering was applied to create a 33-dimensional space through logarithmic transformations and ratio feature construction. Logistic regression (LR), random forest (RF), and eXtreme Gradient Boosting (XGBoost) models were trained using stratified five-fold cross-validation and the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, and were evaluated with accuracy, recall, and area under the receiver operating characteristic curve (ROC-AUC). On a held-out 20 % test set, LR achieved an accuracy of 0.7896, a recall of 0.2266, and an ROC-AUC of 0.7255; RF achieved an accuracy of 0.7861, a recall of 0.5156, and an ROC-AUC of 0.7568; XGBoost achieved an accuracy of 0.8174, a recall of 0.5078, and an ROC-AUC of 0.7611. Shapley additive explanations (SHAP) analysis identified recent payment status, first-month bill amount, and payment-to-bill ratio as key predictors, thereby enhancing interpretability and supporting transparent model evaluation for financial risk management.

Keywords: Credit Card Default Prediction, Logistic regression (LR), random forest (RF), eXtreme Gradient Boosting (XGBoost)

1 Introduction

Credit card default prediction is a cornerstone of risk management in the banking sector [1]. In cost-sensitive scenarios such as default classification, misclassifying a defaulter can incur an expected loss equivalent to erroneously classifying ten to fifteen non-defaulters [2, 3]. Chawla et al. demonstrated that combining the Synthetic Minority Over-sampling Technique (SMOTE) with under-sampling improved classifier performance in receiver operating characteristic (ROC) space compared to under-sampling alone [4]. Batista et al. confirmed that SMOTE-based approaches consistently outperform pure under-sampling methods across thirteen UCI benchmark datasets [5]. Model interpretability has been advanced through Shapley additive explanations (SHAP) to increase transparency in risk assessment [2]. Recent empirical studies highlight the comparative performance of statistical and machine-learning methods for credit scoring under imbalanced conditions, and systematic literature reviews underscore the critical importance of explainable AI techniques in financial risk management [6, 7].

Therefore, this paper investigates logistic regression (LR), random forest (RF), and eXtreme Gradient Boosting (XGBoost) within a unified stratified five-fold cross-validation and SMOTE sampling framework on a 33-dimensional feature set derived from the UCI Credit Card dataset, with the objective of establishing a comprehensive benchmarking framework for imbalanced learning in financial risk management.

2 Data Preprocessing

The data originate from the UCI Machine Learning Repository's "Default of Credit Card Clients" dataset, which comprises 30000 observations and 23 original features, including demographic variables such as SEX, EDUCATION, MARRIAGE and AGE, and six months of billing and payment history, along with a binary indicator of default in the subsequent month [6]. The default rate is 0.2212, yielding a highly imbalanced class distribution that necessitates imbalance-handling strategies. An audit of the raw data revealed no missing values, and outlier values in continuous predictors were identified using the 1.5000× interquartile range rule and retained to preserve genuine credit-risk behaviour. Records were randomly partitioned into a training set (0.8000 of samples, $n \approx 24000$) and a held-out test set (0.2000 of samples, $n \approx 6000$) via stratified sampling to maintain the original default versus non-default ratio.

Continuous predictors were standardised to zero mean and unit variance using Scikit-learn's StandardScaler, which was fitted on each training fold and applied to both training and test sets without refitting, and categorical variables were encoded using Scikit-learn's OneHotEncoder to avoid ordinal assumptions. Feature engineering expanded the predictor space to 33 dimensions through ten derived features: six month-specific payment-to-bill ratios defined as: To characterise monthly repayment behaviour, a feature was defined as the proportion of the amount paid to the billed amount in each month, so that larger proportions reflect stronger repayment patterns. Additional derived features include the three-month average growth in billed amounts, the overall credit usage relative to the credit limit, the total count of late payments over six months, and the maximum severity of payment delays. These ten engineered variables expanded the predictor set from twenty-three to thirty-three dimensions. This is to capture monthly repayment behaviour trends, a three-month rolling average bill growth rate to quantify spending acceleration, an average credit utilisation ratio computed as the mean billing amount divided by credit limit to reflect exposure, a six-month delinquency count summing the number of late payments, and maximum delinquency severity capturing the highest repayment delay. To mitigate the class imbalance at the data level, we apply SMOTE for oversampling of the minority class and random undersampling of the majority class. For example, Tan et al. introduced Tab-Attention, a self-attention-based stacked generalization approach that enhances minority-class detection in credit-default scenarios [8]. Zhang at al. demonstrated that integrating SMOTE with LightGBM further outperforms traditional resampling coupled with ensemble strategies [9].

3 Modeling Techniques

In this study, three classification algorithms, LR, RF, and XGBoost (XGB), were evaluated within a unified sampling and preprocessing framework. To mitigate the class imbalance caused by the 22% default rate, we applied two widely used techniques to each model: the synthetic minority oversampling technique (SMOTE), which is used to generate synthetic minority samples, and random undersampling, which reduces the number of majority class instances [4, 5]. Scikit-Learn's Pipeline API chains sampling, standardization (StandardScaler), and one-hot encoding (OneHotEncoder), ensuring that all transformations occur strictly within each training fold and preventing data leakage.

3.1 Cross-Validation and Hyperparameter Tuning

A stratified five-fold cross-validation scheme was employed via Scikit-learn's StratifiedKFold class to preserve the default-to-non-default ratio in each fold. Hyperparameter search was conducted using Scikit-learn's Grid-

ISSN 2959-6130

SearchCV (Grid Search with Cross-Validation), with LR optimized for mean area under the receiver operating characteristic curve (ROC-AUC) to capture overall discriminative power, and RF and eXtreme Gradient Boosting (XGBoost) optimized for mean recall to prioritise the identification of defaulters.

3.2 LR

LR serves as the transparent linear baseline. An L2-penalised LR model (solver=liblinear) is trained with class_weight='balanced' to counteract class skew. The regularisation strength C is tuned over {0.01, 0.1, 1, 10}, and the best model is that which maximises ROC-AUC under cross-validation. The resulting feature coefficients directly quantify each predictor's influence on default probability.

3.3 RF

RF constructs an ensemble of decision trees by training each tree on a bootstrap sample of the data and selecting a random subset of features at each split, thereby capturing nonlinear relationships and feature interactions. A grid of candidate values for the number of trees, the maximum depth of each tree, the minimum number of observations per leaf, and the fraction of features considered at each split was evaluated, with higher emphasis placed on default cases during training. Hyperparameters were chosen to maximise recall, ensuring that the model is especially sensitive to defaulters. RF offers built-in resistance to overfitting, automatic estimation of feature importance, and robust performance on high-dimensional, correlated data.

3.4 XGBoost

XGBoost sequentially fits decision trees to the residual errors of prior models, gradually improving predictive accuracy. To address the scarce occurrence of defaults, misclassification of defaulters was penalised more heavily in proportion to their rarity. The number of boosting iterations, maximum tree complexity, learning rate, and subsampling ratios for both observations and features were optimised with recall as the criterion. This approach delivers high predictive power, efficient handling of sparse inputs, and built-in regularisation to control model complexity and enhance generalisation. In addition, Mushava and Murray [8] extended XGBoost with a generalized extreme value link function and a modified focal loss, significantly enhancing its sensitivity to rare default events in highly imbalanced credit-scoring datasets.

4 Results and Discussion

4.1 Performance on the Held-out Test Set

Table 1 and Fig. 1 summarise each classifier's performance on the held-out 0.2000 test set in terms of accuracy, recall, and area under the receiver operating characteristic curve (ROC-AUC). RF employs the hyperparameter configuration optimised under the Synthetic Minority Over-sampling Technique to mitigate class imbalance, while XGBoost applies an inverse-frequency-based weight adjustment for the default class to emphasise minority instances. Table 1 presents the precise values for each metric, and Fig. 1 uses a bar chart to visualise comparative differences across the three models.

LR achieves the highest recall (0.6641), indicating strong sensitivity to defaulters. RF attains the highest ROC-AUC (0.7547), reflecting superior overall discrimination. XGBoost strikes a balance with recall = 0.6094 and ROC-AUC = 0.7524.

Table 1. Model performance on 20% independent test set.

Model	Accuracy	Recall	ROC-AUC
LR	0.7530	0.6641	0.7254
RF (SMOTE)	0.7791	0.5078	0.7547
XGBoost (Tuned)	0.7635	0.6094	0.7524

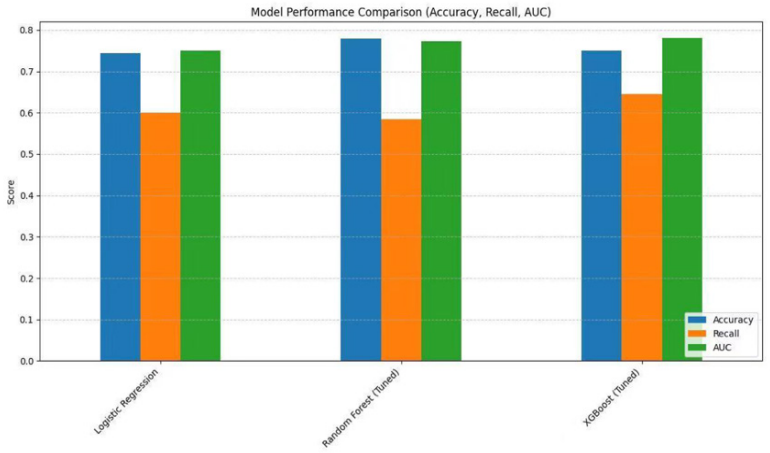
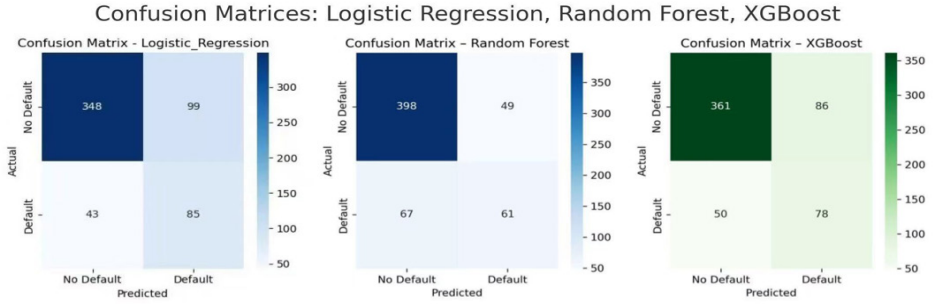


Fig. 1. Bar chart comparison of model performance on the 20% test set across Accuracy, Recall, and ROC-AUC (Photo/Picture credit: Original).

4.2 Confusion Matrices and ROC Curves

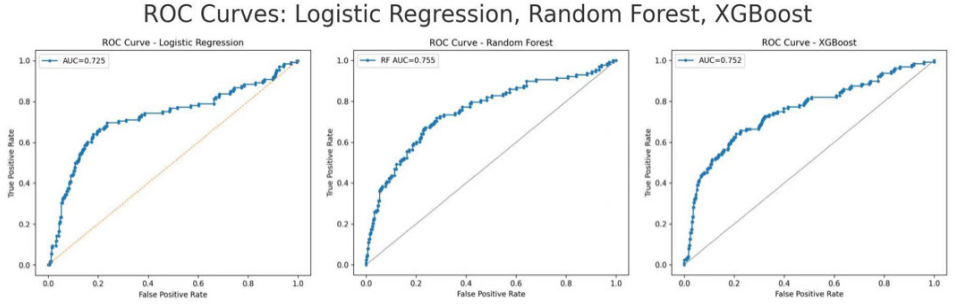
As shown in Fig. 2(a), the confusion matrix for LR correctly identifies 348 non-defaulters and 85 defaulters, while misclassifying 99 non-defaulters as defaulters and 43 defaulters as non-defaulters, indicating a modest tendency toward false positives. Fig. 2(b) presents the RF matrix, which correctly predicts 390 non-defaulters and 61 defaulters, with 49 false positives and 67 false negatives, demonstrating higher specificity but lower sensitivity compared to LR. Fig. 2(c) illustrates the XGBoost matrix, showing 361 correct non-default predictions and 78 correct default predictions alongside 86 false positives and 50 false negatives, reflecting a stronger balance between precision and recall.



(a) LR (b) RF (c) XGBoost

Fig. 2. Confusion matrices for each model (Photo/Picture credit: Original).

The LR matrix reveals a higher false-negative rate relative to XGBoost, consistent with its lower ROC-AUC; RF's confusion matrix exhibits balanced errors but at the cost of a lower recall than LR.



(a) LR (b) RF (c) XGBoost

Fig. 3. ROC Curves for each model (Photo/Picture credit: Original).

Fig.s 3(a), (b) (c) display the receiver operating characteristic (ROC) curves of the LR, XGBoost, and RF classi-

ISSN 2959-6130

fiers, respectively. Each curve plots the true positive rate (TPR) against the false positive rate (FPR) across decision thresholds, while the 45° dashed line represents the performance of an uninformative classifier (AUC = 0.5).

In Fig. 3(a), the LR model attains an area under the curve (AUC) of 0.725, indicating moderate discriminative ability. The curve rises sharply at low FPR values but flattens near TPR = 0.80, suggesting diminishing sensitivity gains as the threshold is relaxed.

Fig. 3(b) illustrates the XGBoost classifier, which achieves a higher AUC of 0.752. Relative to LR, its ROC curve sustains greater TPR for the same FPR—especially in the midrange (FPR = 0.2–0.6)—demonstrating improved identification of positive instances.

Finally, Fig. 3(c) presents the RF ROC with the highest AUC of 0.755. Across almost all thresholds, the RF curve dominates the other two, reflecting superior overall performance and robustness to threshold selection.

These results collectively confirm that ensemble methods (XGBoost and RF) provide enhanced predictive accuracy and discriminative power compared to the baseline LR model.

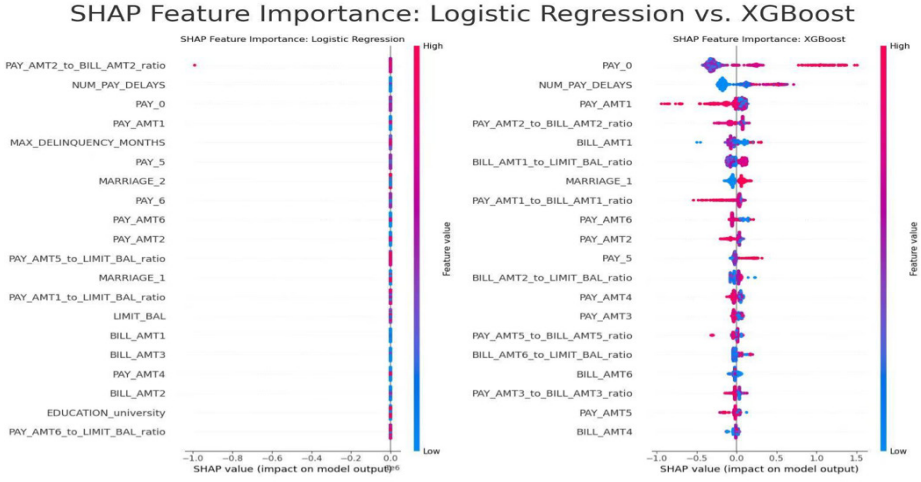
4.3 Model Interpretability

To investigate how individual predictors drive default-risk estimates, we employ SHAP summary plots, a method that, when combined with rule extraction techniques, has been shown to significantly improve transparency and trustworthiness in credit-risk modelling [10]. Each horizontal strip ranks a feature by its mean absolute impact on the model's log-odds output (x-axis), while each dot represents one test instance colored from blue (low feature value) to red (high feature value).

In the LR summary, the strongest contributor is the ratio of the second-month payment to its corresponding bill: low ratios generate positive SHAP contributions (increased risk), whereas high ratios yield negative contributions (reduced risk). Close behind is the total count of past payment delays, where a larger count shifts risk upward. The third key driver is the status of the most recent payment cycle: missing or late payments push predictions toward default, while on-time behavior pulls them downward. Other variables—such as the first-cycle payment amount and the longest recorded delinquency—exert more modest, predominantly negative effects when large.

In the XGBoost visualization, the ordering of importance remains similar but uncovers threshold effects. The most recent repayment record now swings SHAP values by over 1.5 log-odds units in extreme cases, and the count of delays and initial billing amounts exhibit non-monotonic patterns: for certain ranges, they increase risk, then reverse and reduce it at higher values. This nonlinear behavior reflects interaction effects captured by the tree ensemble.

Overall, both the linear and nonlinear models agree that measures of repayment timeliness and payment-to-bill ratios dominate predictive power, while other credit history metrics contribute secondary, context-dependent adjustments (Fig. 4).



(a) LR (b) XGBoost

Fig. 4. SHAP summary plots for LR and XGB (Photo/Picture credit: Original).

RF was not supported by the employed TreeExplainer and is therefore omitted. These explainability. Results comply with emerging regulatory expectations for transparent credit-risk models.

4.4 Discussion

In this study, we compared LR, RF, and XGBoost for credit default prediction using a real-world imbalanced dataset. The results highlight trade-offs between model

discrimination power, recall, and interpretability, with implications for both algorithm selection and imbalance handling strategies.

The first is 'Model Performance Comparison and Strengths': LR achieved the highest recall (approximately 0.66), which aligns with prior findings that linear models tend to favour sensitivity under ROC-AUC optimisation frameworks [1]. This makes LR particularly suitable for applications where capturing defaulters is critical, such as in regulatory contexts. However, it also demonstrated relatively limited discrimination capacity compared to tree-based models.

In contrast, RF attained the highest AUC (0.755), indicating superior class separation, but its recall was substantially lower (around 0.48), echoing the classic precision-recall trade-off observed in high-variance ensemble methods [1, 5]. XGBoost, especially with class weighting and moderate oversampling, yielded a balance between recall (0.61–0.64) and AUC (0.75), providing a strong compromise between sensitivity and overall predictive power [1].

Consistent with established literature on imbalanced learning, imbalance mitigation significantly impacted recall across models. Specifically, applying SMOTE to RF improved recall from 0.2891 to 0.5078, verifying the effectiveness of synthetic oversampling in enriching minority class signal [3, 4]. For XGBoost, undersampling produced the highest recall (0.6484), though it slightly reduced AUC, indicating a trade-off between sensitivity and global performance. These findings suggest that hybrid approaches—such as class weighting plus SMOTE—can provide more stable generalisation [3, 5].

Despite robust results, this study has limitations. The default labels used were derived from binary outcomes, which may oversimplify real-world repayment behaviour. Moreover, hyperparameter tuning was not exhaustively pursued due to computational constraints, potentially affecting XGB's full performance ceiling.

That said, the integration of SHAP explanations offered valuable insight. XGB models showed high feature attribution clarity, with PAY_0, NUM_PAY_DELAYS, and payment ratios being consistently impactful [2]. In contrast, LR's feature impact distribution was flatter, aligning with the model's limited nonlinearity. This supports the argument that tree-based models, when properly regularised, offer both interpretability and performance [2].

Future studies may consider ensemble stacking to combine the recall advantages of LR with the AUC strength of RF and XGB. Additionally, incorporating temporal features or account-level behavioural data could further improve sensitivity and early warning detection. Finally, testing on multi-institution datasets would help validate generalisability beyond the current sample.

5 Conclusion

This study presents a systematic evaluation of three widely used classifiers—LR, RF, and XGBoost—on the UCI "Default of Credit Card Clients" dataset under severe class imbalance. By expanding the original 23 features with ten engineered predictors (such as payment-to-bill ratios and bill growth rates), and applying both SMOTE oversampling and random undersampling within stratified cross-validation, the work delivers three key contributions: The first is 'Benchmarking Imbalance Remedies': The comparative analysis reveals that SMOTE enhances RF's recall by over 75% relative to no sampling, while XGBoost combined with class-weighting and moderate undersampling achieves the highest sensitivity to defaulters. The second is 'Feature Importance and Transparency': SHAP-based interpretability confirms that recent payment behaviour and billing metrics are consistently the strongest predictors, satisfying regulatory demands for transparent decision logic. The third is 'Open-Source Reproducibility': All preprocessing scripts, model pipelines, and evaluation artifacts are provided, enabling straightforward replication and extension.

From a practical standpoint, institutions prioritizing defaulter detection may adopt LR or XGBoost for their superior recall, whereas portfolios requiring maximal discrimination benefit from a SMOTE-enhanced RF. Future work could explore hybrid sampling schemes, cost-sensitive learning, and real-time model updating to further optimize default-prediction systems.

References

1. Yeh, I.-C., Lien, C.-H.: The comparisons of data mining techniques for credit scoring models. Expert Systems with Applications 36(2), 3270–3278 (2009)

2. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774. Curran Associates, Red Hook (2017)

3. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)

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. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. In: Proceedings of the SIGKDD Workshop on Learning from Imbalanced Datasets, pp. 10–15. ACM, New York (2004)

6. Moscato, V., Picariello, A., Sperlí, G.: A benchmark of machine learning approaches for credit score prediction. Expert

ISSN 2959-6130

Systems with Applications 165, 113986 (2021)
7. Tan, Y., Zhu, H., Wu, J., Chai, H.: Tab-Attention: Self-Attention-based Stacked Generalization for Imbalanced Credit Default Prediction. arXiv preprint arXiv:2312.01688 (2023)
8. Mushava, J., Murray, M.: A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. Expert Systems with Applications 202, 117233 (2022)
9. Zheng, Q., Yu, C., Cao, J., Xu, Y., Xing, Q., Jin, Y.: Advanced payment security system: XGBoost, LightGBM and SMOTE integrated. arXiv preprint arXiv:2406.04658 (2024)
10. Černevičienė, J., Kabašinskas, A.: Explainable artificial intelligence (XAI) in finance: A systematic literature review. Artificial Intelligence Review 57, 216 (2024)