Credit Card Default Prediction Based on Machine Learning

Shujun Yao

Eberly College of Science, The Pennsylvania State University, State College PA 16802
Sxy5411@psu.edu

Abstract. In recent years, credit cards have become deeply integrated into personal financial activities. While they provide ease and flexibility, they also introduce new challenges for managing financial risk. As the volume of credit card usage grows, concerns over potential defaults have drawn growing interest from the banking sector and related financial entities. Conventional approaches to evaluating credit risk often depend on rigid assumptions, making it difficult to account for the nuanced and dynamic nature of consumer behavior. This study investigates how machine learning techniques can improve default prediction by utilizing a real-world dataset. Three ensemble models—AdaBoost, Gradient Boosted Decision Tree (GBDT), and Random Forest—are implemented and assessed for their effectiveness in recognizing high-risk defaulters. Model performance is evaluated based on commonly used indicators such as accuracy, precision, recall, F1 score, and Area Under the Curve (AUC). Among the models, Random Forest demonstrates the strongest overall performance, especially in terms of balanced classification results and high AUC values. To further assess practical utility, the models are tested on two synthetic customer scenarios. All three models produce consistent outcomes, reinforcing their applicability to real-world cases. This research underscores the value of machine learning in refining credit risk analytics and contributes actionable insights for enhancing early warning frameworks in the finance sector.

Keywords: AdaBoost, GBDT, Random Forest, Credit Card Default Prediction

1 Introduction

The rising popularity of credit cards has significantly influenced personal financial behavior, becoming a cornerstone of modern consumer economies. With advantages such as convenience, short-term financing, and deferred payment options, credit cards have been widely adopted by individuals across different income levels and regions. In many developing and developed markets, this trend is further supported by digitization in the banking sector and government-led financial inclusion initiatives. However, the convenience of credit card use is closely accompanied by increasing concerns about credit risk—particularly, the risk of borrower default. Credit card default not only disrupts

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_37

personal creditworthiness but also directly impacts the financial stability of issuing institutions. According to Cheng et.al, credit risk is among the core risks banks must control through sound and forward-looking assessment systems [1]. As the scale of credit card lending grows, so does the urgency of building accurate and adaptive prediction models that help lenders screen high-risk applicants before credit is issued. Traditional risk assessment methods, such as the 5C model and rule-based scoring, often fall short due to their limited ability to process large, nonlinear, and imbalanced financial datasets. These limitations call for more intelligent, data-driven approaches to default prediction.

In response to this challenge, researchers have developed and tested a wide range of analytical models to improve the accuracy of credit risk classification. Logistic regression has long served as a foundational model in credit scoring due to its simplicity and interpretability. Recent studies, such as Levy and Baha, have reaffirmed its effectiveness, demonstrating that logistic regression consistently outperforms linear discriminant analysis across various credit risk datasets, particularly in small business lending contexts [2]. To overcome the limitations of linear modeling, Odeh et al. proposed a genetic algorithm-based approach, which better captured interactions among variables in multiclass loan default prediction [3]. Agbemava et al. examined default risk among microfinance clients using logistic regression and identified several important predictors, including marital status, type of loan, and repayment duration [4]. These studies contributed valuable insights but were largely based on relatively small datasets or focused on a narrow range of features. Many traditional models lack the flexibility to handle real-world challenges like data imbalance and nonlinear relationships—issues that frequently arise in credit risk modeling. In a recent review, Dastile et al. emphasized that although logistic regression continues to be widely used, newer approaches such as ensemble algorithms and support vector machines are gaining traction due to their superior adaptability and prediction capabilities in complex environments [5]. Separately, recent research has also explored the use of alternative data sources, such as borrower narratives, to improve prediction accuracy and address the limitations of traditional feature sets [6]. In recent years, ensemble machine learning models have emerged as promising alternatives. By integrating multiple weak learners, these algorithms can improve prediction robustness, reduce variance, and better capture hidden patterns in large-scale, high-dimensional data. Methods such as AdaBoost, Gradient Boosted Decision Tree (GBDT), and Random Forest have been successfully applied in various financial domains, but there remains a lack of comparative research that evaluates their performance side by side using consistent datasets and evaluation metrics. Additionally, practical testing using realistic case scenarios is often omitted, limiting our understanding of their real-world applicability.

This study builds and compares three ensemble learning models—AdaBoost, GBDT, and Random Forest—using the University of California, Irvine (UCI) credit card default dataset. Their prediction performance is evaluated using five common metrics, and their decision-making consistency is tested through hypothetical borrower profiles.

2 Methodology

2.1 Data Processing

Dataset Used. The dataset employed in this study, titled "Default of Credit Card Clients," originates from the UCI Machine Learning Repository managed by the University of California, Irvine. It was initially contributed by Yeh Yicheng and Lian Zhehui in 2009 [7]. Comprising 30,000 entries, the dataset includes 23 attributes such as credit limit, demographic details (e.g., gender, age, education, marital status), payment history over the past six months, bill amounts, and repayment records. The binary target variable, "default payment next month," is labeled as 0 for non-default and 1 for default.

Data Balancing Process. Before initiating model construction, the distribution of the target variable is examined. As shown in Fig. 1, the variable includes two categories: default and no default. Among the total users, 6,636 have defaulted, representing 22.12%, while the remaining 23,364 users, or 77.88%, have not defaulted. The sample ratio of the two categories is about 3.5: 1. The dataset exhibits class imbalance, with more positive samples than negative ones, which may introduce bias into the modeling process.

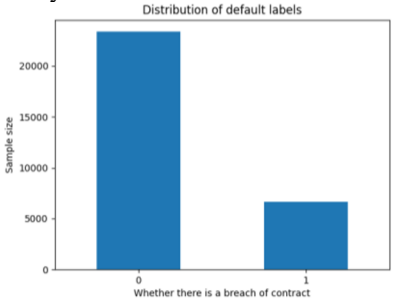

Fig. 1. Distribution of target variables in raw data (Picture credit: Original)

To meet the course requirements of between 3000 and 10000 entries and to balance the data, 5000 were randomly selected from defaults and non-defaults to form a new dataset with a total of 10000 data sets. The processed data is shown in Fig. 2 below.

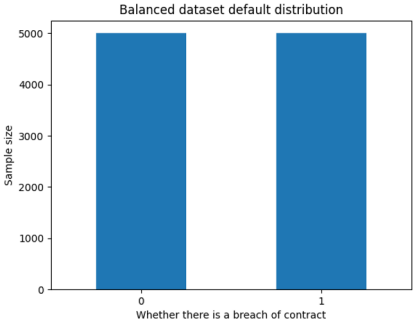

Fig. 2. Distribution of data after balanced processing (Picture credit: Original)

Data Segmentation. To build and validate the models, the dataset was partitioned into training and test subsets using a 4:1 ratio. Specifically, the training set includes 7,499 samples, comprising 3,774 default cases and 3,725 non-defaults. The remaining 2,500 samples form the test set, with 1,226 defaults and 1,274 non-defaults.

2.2 Introduction to Credit Card Default Prediction Modeling Process

The flow of the scheme studied in this paper mainly consists of five steps: data import, data preprocessing, model training, comparison and evaluation, and summarization, as shown in Fig. 3.

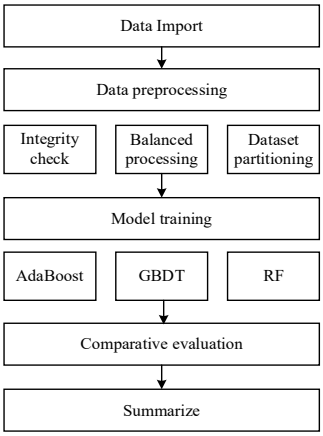

Fig. 3. Modeling Flowchart (Photo/Picture credit: Original)

2.3 Indicators for Model Evaluation

In binary classification tasks, model effectiveness is typically assessed through a confusion matrix, which yields several key metrics. Accuracy reflects the share of correctly classified instances, while precision and recall evaluate how well the model identifies relevant positive cases. The F1 score offers a harmonic balance between precision and recall. The ROC curve illustrates the trade-off between correctly identifying positives and incorrectly classifying negatives, whereas AUC provides a single-value summary of the model's discriminative ability—higher values indicate stronger performance.

3 Results

3.1 Credit Card Default Prediction Model based on AdaBoost

AdaBoost improves the overall prediction performance by gradually correcting the samples misclassified in the previous round by a weighted combination of a series of weak classifiers (a CART decision tree with max_depth=2 is used as the base learner in this study). In this paper max_depth=2, iteration number=200, learning rate=0.05, random state=42. The evaluation index scores are shown in Table 1.

Table 1. AdaBoost model metrics

	AUC	Accuracy	Precision	Recall	F1-score
Score	0.7569	0.7032	0.76	0.58	0.66

The ROC curve is shown in Fig. 4.

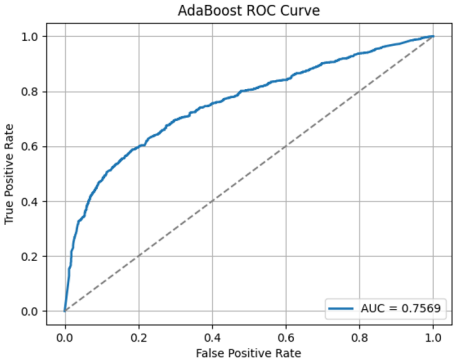

Fig. 4. AdaBoost ROC curve (Picture credit: Original)

The results suggest that the model performs reasonably well overall, though it shows clear limitations in identifying actual default cases. The AUC score of 0.7569 suggests a fair level

of discriminative power in identifying defaulted versus non-defaulted customers. An accuracy of 0.7032 indicates solid overall predictive capability, while a high precision of 0.76 suggests that the model is effective in correctly identifying defaults when it makes a positive prediction. However, the recall rate stands at 0.58, revealing that a substantial portion of true defaults remains undetected. With an F1 score of 0.66, the model demonstrates a reasonable balance between precision and recall, although the lower recall rate weakens its overall effectiveness. Additionally, the ROC curve confirms that the model performs notably better than random classification.

3.2 Credit Card Default Prediction Model Based on GBDT

GBDT enhances overall predictive accuracy by iteratively minimizing the residual errors of the previous model. At each iteration, the algorithm focuses on samples that were harder to classify correctly, gradually improving performance while incorporating a learning rate to mitigate the risk of overfitting. In this study, the model is configured with an exponential loss function, 100 boosting iterations, a learning rate of 0.1, a maximum tree depth of 4, and a fixed random seed of 42 to ensure reproducibility. The resulting evaluation metrics are summarized in Table 2.

Table 2. GBDT model metrics

	AUC	Accuracy	Precision	Recall	F1-score
Score	0.7602	0.7024	0.73	0.63	0.67

The ROC curve is shown in Fig. 5.

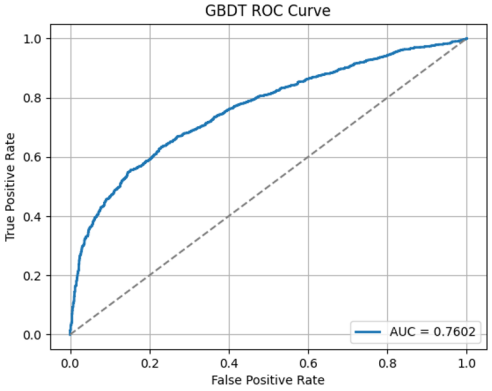

Fig. 5. GBDT ROC curve (Picture credit: Original)

The GBDT model exhibits stable performance on the test set. Its AUC score of 0.7602 suggests a moderate ability to distinguish between defaulters and non-defaulters, though the

overall classification capability remains at a general level. The model achieves an accuracy of 0.7024, indicating solid overall prediction performance. A precision score of 0.73 reflects the model's effectiveness in correctly identifying default cases among predicted positives. However, the recall rate is relatively low at 0.63, showing that the model still struggles to capture all actual default instances. The F1 score stands at 0.67, suggesting a fair balance between precision and recall, but also highlighting potential for further optimization. As shown in the ROC curve, the model's curve lies above the random baseline, confirming better-than-chance predictive ability.

3.3 Credit Card Default Prediction Model Based on Random Forest

Random Forest enhances model stability and generalization by introducing dual randomness in both data and feature selection. It creates several bootstrap samples drawn with replacement from the initial training data and builds an individual decision tree on each one. To minimize overfitting, a random selection of features is used at each split during the training process. The model's final output is then obtained through majority voting from all the constructed trees. In this study, the model is configured with a total of 200 trees and a fixed random seed of 1234. The performance outcomes are summarized in Table 3.

Table 3. RF model metrics

	AUC	Accuracy	Precision	Recall	F1-score
Score	0.7639	0.7020	0.72	0.64	0.68

The ROC curve is shown in Fig. 6.

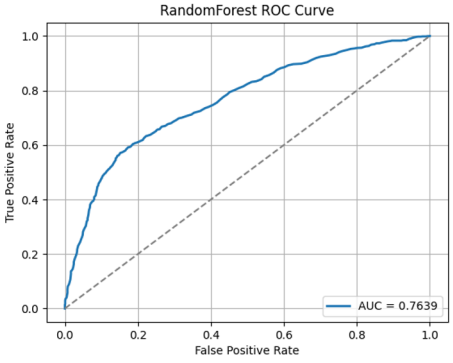

Fig. 6. RF ROC curve (Picture credit: Original)

With an AUC score of 0.7639, the Random Forest model shows a stronger classification capability than random guessing and marginally exceeds the performance of GBDT. Its accuracy of 0.7020 reflects stable overall classification performance. With a precision of

0.72, the model effectively identifies true defaulters among predicted positives. However, the recall rate is relatively low at 0.64, suggesting that some actual default cases remain undetected. The F1 score of 0.68 demonstrates a reasonable balance between precision and recall. Furthermore, the ROC curve of the Random Forest model consistently lies above the diagonal reference line, reinforcing its robustness and predictive stability.

3.4 Summary of Experimental Results

By comparing the performance of the three ensemble models, namely AdaBoost, GBDT, and Random Forest, in predicting credit card defaults, this study finds that although their overall predictive performance is similar, each model excels in different aspects. Random Forest achieves the highest AUC value (0.7639), slightly higher than that of GBDT (0.7602) and AdaBoost (0.7569), suggesting greater effectiveness in identifying and classifying default outcomes versus non-default ones. In terms of accuracy, all three models perform at a comparable level, with values around 70 percent. However, GBDT and Random Forest demonstrate higher recall rates, suggesting they are more effective at identifying actual defaulters. In contrast, AdaBoost achieves the highest precision, implying it is better suited for applications that emphasize the accurate identification of high-risk customers. Therefore, the selection of models in practice should be guided by specific objectives, balancing precision and recall to achieve more effective prediction results.

4 Model Performance Discussion

The comparative results of AdaBoost, GBDT, and Random Forest models indicate that while all three algorithms exhibit similar levels of overall accuracy (approximately 70%), they differ in their performance across specific metrics. Notably, Random Forest achieves the highest AUC value (0.7639), suggesting its superior capacity to differentiate between defaulters and non-defaulters. GBDT performs comparably, particularly in terms of recall, which reflects a stronger capability in identifying actual defaulters. AdaBoost, by contrast, demonstrates the highest precision, indicating its strength in minimizing false positives and thereby reducing misclassification of non-defaulting customers.

These findings are consistent with observations in prior literature. For example, Madaan et al. compared decision trees and Random Forest models for loan default prediction and concluded that Random Forest outperforms other models in classification accuracy and generalization ability, especially on tabular financial data [8]. Similarly, Bahnsen et al. highlighted the importance of feature interactions and ensemble methods in improving fraud detection and credit risk classification tasks [9]. Their work supports the argument that Random Forest and similar ensemble methods are more robust in capturing complex, nonlinear relationships within financial datasets. However, when minimizing the cost of

false positives is critical—such as when targeting high-risk customers for more stringent review—AdaBoost's higher precision may be more desirable.

Therefore, the choice of model should be guided by the specific application scenario. In cases where missing a real defaulter is more damaging (e.g., issuing large unsecured credit), models with higher recall, like GBDT or Random Forest, may be preferred. Conversely, in contexts requiring accurate targeting of high-risk individuals with limited resources, AdaBoost's high precision may offer operational advantages. These findings reinforce the idea proposed by Shaheen and Elfakharany, who emphasized that no single model dominates across all tasks, and integrated learning algorithms must be evaluated relative to the business goal and data characteristics [10]. Recent advances, such as the work by Song et al., further demonstrate that customizing ensemble learning schemes to credit rating levels and optimizing across multiple objectives can significantly enhance default prediction performance, especially under conditions of class imbalance and heterogeneous risk profiles [11].

5 Conclusion

This paper evaluates the predictive performance of three ensemble learning algorithms—AdaBoost, GBDT, and Random Forest—on credit card default using real-world data. Although the overall accuracy of the three models is similar, each shows unique strengths. Random Forest performs best in terms of AUC (0.7639) and recall, making it more effective in identifying actual defaulters. GBDT follows closely, offering a good balance between recall and precision. AdaBoost, while slightly behind in recall, achieves the highest precision, indicating its ability to reduce false positives. These differences suggest that model selection should depend on the specific risk control priorities of financial institutions. To assess real-world applicability, two hypothetical customer profiles were tested, and all three models produced consistent and logical prediction outcomes. Overall, the findings confirm that ensemble learning techniques can offer reliable support in credit risk assessment, especially when tailored to different operational goals. There are some limitations in this study: the overall recall of the models is not high, and the ability to identify real default users' needs to be improved; the data only comes from the publicly available UCI dataset, and richer and more varied data sources in real scenarios are not taken into account, which may affect the generalization ability of the models. More credit-related features and external data can be introduced in the future to improve the model's predictive ability. More sophisticated approaches, such as deep learning, may further enhance predictive accuracy. Other imbalance data processing methods (e.g., SMOTE) could also be explored to improve the ability to capture defaulting users.

References

1. Cheng, Y.: Research on credit risk early warning model of commercial banks based on neural network algorithm. arXiv (2024)
2. Levy, R., Baha, H.: Credit risk assessment: a comparison of the performances of logistic regression and linear discriminant analysis. International Journal of Entrepreneurship and Small Business 42(1–2), 169–186 (2021)
3. Odeh, O., Koduru, P., Featherstone, A.: A multi-objective approach for the prediction of loan defaults. Expert Systems with Applications 38(7), 8850–8857 (2011)
4. Agbemava, E., Nyarko, I.K., Adade, T.C.: Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal 12(1) (2016)
5. Dastile, X.N., Celik, T., Potsane, M.: Statistical and machine learning models for credit scoring: A review of recent literature. Heliyon 6(2), e03463 (2020)
6. Xia, Y., He, L., Li, Y.: Predicting loan default in peer-to-peer lending using narrative data. Journal of Forecasting 39(2), 260–280 (2020)
7. Yeh, I.-C., Lien, C.-H.: Default of credit card clients dataset. UCI Machine Learning Repository (2009). https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients, last accessed 2025/08/05
8. Madaan, M., Kumar, A., Keshri, C.: Loan default prediction using decision trees and random forest: a comparative study. In: IOP Conference Series: Materials Science and Engineering, vol. 1022(1), 012042. IOP Publishing, Bristol (2021)
9. Bahnsen, A.C., Aouada, D., Stojanovic, A.: Feature engineering strategies for credit card fraud detection. Expert Systems with Applications 51, 134–142 (2016)
10. Shaheen, S.K., Elfakharany, E.: Predictive analytics for loan default in the banking sector using machine learning techniques. In: Proceedings of the 28th International Conference on Computer Theory and Applications (ICCTA), pp. 66–71. IEEE, New York (2018)
11. Song, Y., Wang, Y., Ye, X.: Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. Information Sciences 629, 599–617 (2023)