Proceedings of ICEMGD 2025 Symposium: Innovating in Management and Economic Development
DOI: 10.54254/2754-1169/2025.LH23945

Research on Credit Card Default Prediction Based on Random Forest and Logistic Regression Models

Silong Lin

Business School, Hong Kong Metropolitan University, Hong Kong, China
s1336620@live.hkmu.edu.hk

Abstract: In this era of advanced payment methods, credit cards have become an indispensable financial instrument for individuals and business alike. The rapid development of credit card business has led to an escalation in credit card default issues, resulting in significant economic losses and risks for financial institutions. This study utilizes the Kaggle credit card default dataset to conduct credit card default prediction using the Random Forest model, with comparative analysis against the Logistic Regression model. The research findings demonstrate that the Random Forest model outperforms the Logistic Regression model across various evaluation metrics, including accuracy, recall, F1 score, ROC curve, and AUC value, particularly excelling in handling nonlinear relationships and high-dimensional data. Through feature selection, the study identifies key characteristics influencing credit card default, such as repayment status, credit limit, and bill amount. The research indicates that the Random Forest model can effectively identify potential default customers, assisting financial institutions in reducing default risks and enhancing risk management capabilities.

Keywords: Credit Card Default Prediction, Random Forest Model, Logistic Regression Model, Feature Selection, Risk Management.

1. Introduction

With the rapid advancement of technology and the evolution of consumer behavior, credit cards have become an indispensable payment instrument in modern life. However, the swift expansion of credit card services has been accompanied by a growing incidence of credit card defaults. Such defaults not only inflict direct financial losses on financial institutions but also pose potential threats to the stability of the broader financial market. Consequently, the effective prediction and mitigation of credit card defaults have emerged as critical objectives in financial risk management. The development of a credit card default prediction model presents a viable solution. For financial institutions, this model facilitates the precise assessment of customer credit risk, enabling the formulation of differentiated lending strategies tailored to varying credit profiles, thereby reducing potential losses. Furthermore, the model aids in the identification of at-risk customers, allowing for preemptive measures such as interest rate adjustments or credit limit reductions to mitigate risk. For individual users, the credit card default prediction model offers credit rating evaluations, empowering them to gain a clearer understanding of their financial standing and manage their finances more effectively.

DOI: 10.54254/2754-1169/2025.LH23945

With the widespread application of machine learning in the financial sector, credit card default prediction has emerged as a prominent research focus. Early studies predominantly employed traditional statistical methods, such as Logistic Regression, which has been extensively utilized in financial risk management due to its interpretability and implementation simplicity [1,2].

However, Logistic Regression struggles to handle complex nonlinear relationships, prompting an increasing number of scholars to adopt more advanced machine learning methodologies. For instance, Zhou et al. compared models including Support Vector Machines, Decision Trees, and Random Forests, demonstrating that Random Forests exhibit superior performance in handling high-dimensional data, feature selection, and nonlinear modeling [3].

Liu et al. further empirically demonstrated that Random Forests maintain stable performance in credit scoring scenarios, particularly when applied to datasets with complex inter-variable interactions [4]. Zhang and Wang integrated Logistic Regression with Random Forests, validating that ensemble learning enhances model robustness, especially in imbalanced default scenarios where it outperforms individual models [5].

Bahnsen et al. emphasized the cost implications of different types of misclassification in actual credit operations through cost-sensitive Logistic Regression models, highlighting the importance of considering business feasibility in model evaluation [6]. Lessmann et al. conducted a comprehensive comparison of dozens of machine learning models, concluding that Random Forests and ensemble models demonstrate the strongest overall performance in credit scoring [7].

In addition to traditional models, deep learning methodologies have been increasingly introduced into such problems in recent years. The XGBoost boosting tree model proposed by Chen and Guestrin has demonstrated superior performance across various financial scenarios due to its efficient parallelism and accuracy, although its model complexity and interpretability have constrained its large-scale implementation [8].

Tsai and Chen explored the construction of hybrid models by combining multiple algorithms, proposing that the complementary nature of logistic regression and neural networks could enhance performance [9]. Baesens et al. also validated the efficacy of neural networks in personal loan risk analysis, highlighting their advantages in predicting long-term default trends [10].

In summary, recent research indicates that while deep learning and ensemble models generally achieve higher accuracy, random forests remain widely adopted in practical financial operations due to their superior interpretability, robustness, and modeling efficiency. Consequently, this paper selects random forests and logistic regression as two representative models for comparison, offering both theoretical value and significant practical guidance.

This study aims to compare the predictive performance of Random Forest and Logistic Regression models in forecasting credit card defaults, while also identifying the most influential features affecting default prediction. Additionally, it seeks to offer actionable risk management recommendations for financial institutions to enhance their decision-making processes.

2. Method and data

2.1. Data source

This study utilizes the Kaggle Credit Card Default dataset, which include information on credit cards default payments, demographic factors, credit rating data, payment history, and billing status of Taiwanese credit card customers from early April 2005 to late September 2005. The dataset comprises 30,000 samples and 25 features, as shown in Table 1.

DOI: 10.54254/2754-1169/2025.LH23945

2.2. Method

This study initiates with an aggressive preprocessing of the data, including data cleaning, handling missing values, and addressing outliers. Subsequently, feature selection is conducted, ranking features based on their importance, and the top ten most significant features are selected. Finally, the performance of two models, Random Forest and Logistic Regression, is compared after their respective implementations, as shown in Table 1.

(1) RF Model introduction: The Random Forest model is an ensemble learning method capable of handling nonlinear relationships and high-dimensional data. By aggregating multiple decision trees, Random Forest exhibits robust resistance to overfitting and remains insensitive to noisy data.

(2) LR Model introduction: The Logistic Regression model is a linear model employed for classification tasks, particularly suited for binary classification problems. It maps the output of linear regression to probabilities via the logistic function. While it offers rapid training and prediction speeds, Logistic Regression is inherently limited in directly addressing nonlinear relationships and is susceptible to significant performance degradation in the presence of outliers.

Table 1: Feature abbreviations

Client ID	ID
Granted Credit Limit	LIMIT_BAL
Gender	SEX
Educational Attainment	EDUCATION
Marital Status	MARRIAGE
Age (Years)	AGE
Repayment Status as of September 2005	PAY_0
Client's Repayment Status from August 2005 to April 2005	PAY_2~6
Client's Billing Amount from September 2005 to April 2005	BILL_AMT1~6
Client's Previous Repayment Amount from September 2005 to April 2005	PAY_AMT1~6
Default Status	Default.payment.next.month

3. Result

3.1. Analysis of key features

The impact level of features was evaluated to identify the top 10 most influential factors determining whether credit card customers will default in the following month, as shown in Figure 1.

Feature Explanation:

(1) PAY-0: As the most critical feature, it reflects the customer's most recent monthly repayment behavior. The latest repayment status directly indicates both repayment capacity and willingness, with customers who have recently delayed payments being more likely to default.

(2) ID: The customer's ID is correlated with their credit history. Customers with poor historical credit records are more prone to default.

(3) AGE: Age is closely related to a customer's earning capacity and financial responsibility. Younger customers may have a higher likelihood of credit default due to lower repayment capacity or inadequate personal financial planning.

DOI: 10.54254/2754-1169/2025.LH23945

(4) BILL-AMT1: The September bill amount reflects the user's recent consumption level and debt status. A higher bill amount may indicate significant debt repayment pressure, potentially increasing the risk of default.

(5) LIMIT-BAL: The granted credit limit reflects the customer's credit rating and the financial institution's trust in them. A lower credit limit suggests the customer may have a lower repayment capacity, likely increasing the risk of default.

(6) BILL-AMT2: The previous month's bill amount also reflects the customer's recent consumption level and debt status. Consistently high bill amounts indicate potential financial stress, suggesting an increased risk of default.

(7) PAY-AMT1: The September repayment amount reflects the customer's recent repayment capacity. A lower repayment amount may indicate financial stress, potentially increasing the risk of default.

(8), (9), (10). BILL-AMT (3/4/5): The bill amounts for May, June, and July reflect the customer's earlier long-term consumption level and debt capacity. While the impact may not be as significant as the top-ranked factors, they provide valuable reference for financial institutions to identify customers under long-term financial stress.

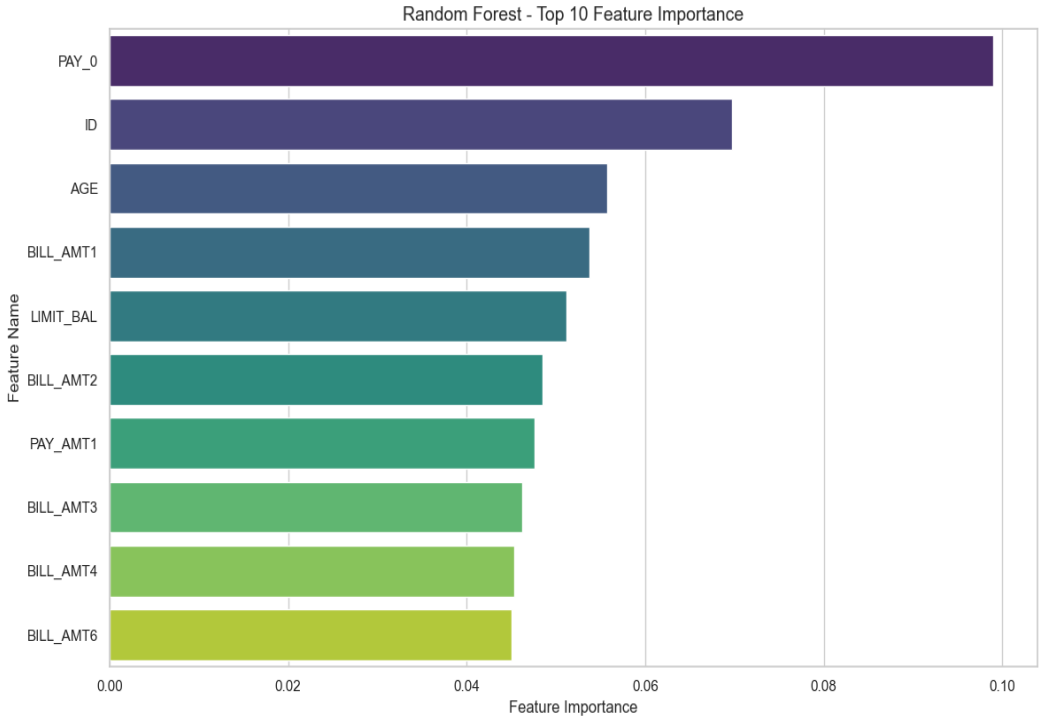

Figure 1: Top 10 features influencing credit card default risk among customers

3.2. Model performance comparison

To further demonstrate the superiority of the Random Forest model in this research, a comparative analysis was conducted against the Logistic Regression model.

The confusion matrices of both models reveal that the Random Forest model exhibits superior overall prediction accuracy and better alignment with actual outcomes. Notably, in predicting default customers, the Random Forest model significantly outperforms the Logistic Regression model.

Specifically, the Random Forest model correctly identified 395 default customers, whereas the Logistic Regression model only identified 215. This substantial difference indicates that the Random

DOI: 10.54254/2754-1169/2025.LH23945

Forest model has greater potential to mitigate financial losses for financial institutions, as shown in Figure 2, Figure 3, Figure 4, Figure 5.

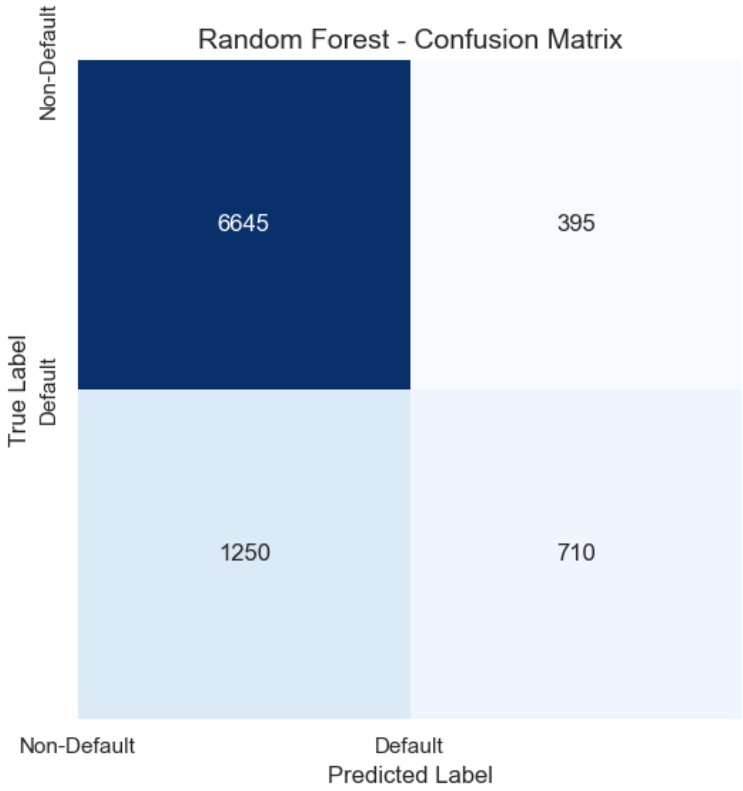

Figure 2: Confusion matrix of the random forest model

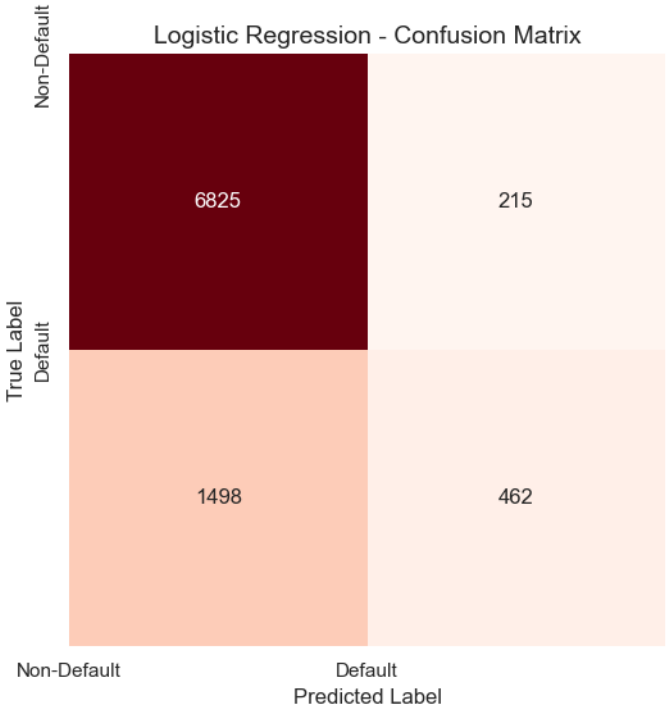

Figure 3: Confusion matrix of the logistic regression model

DOI: 10.54254/2754-1169/2025.LH23945

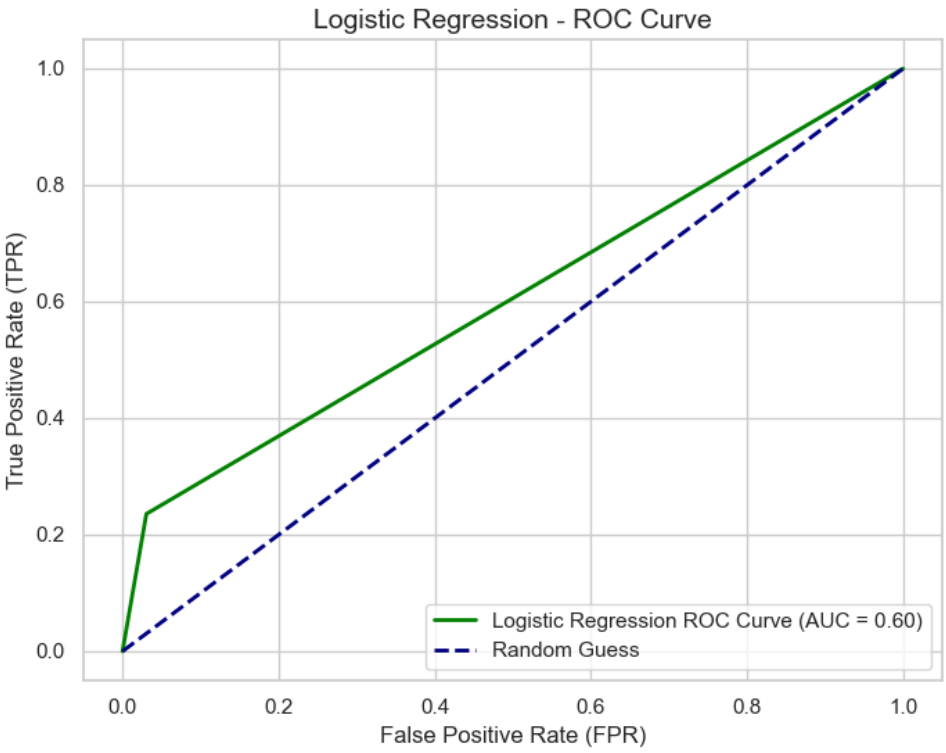

Figure 4: ROC curve of the logistic regression model

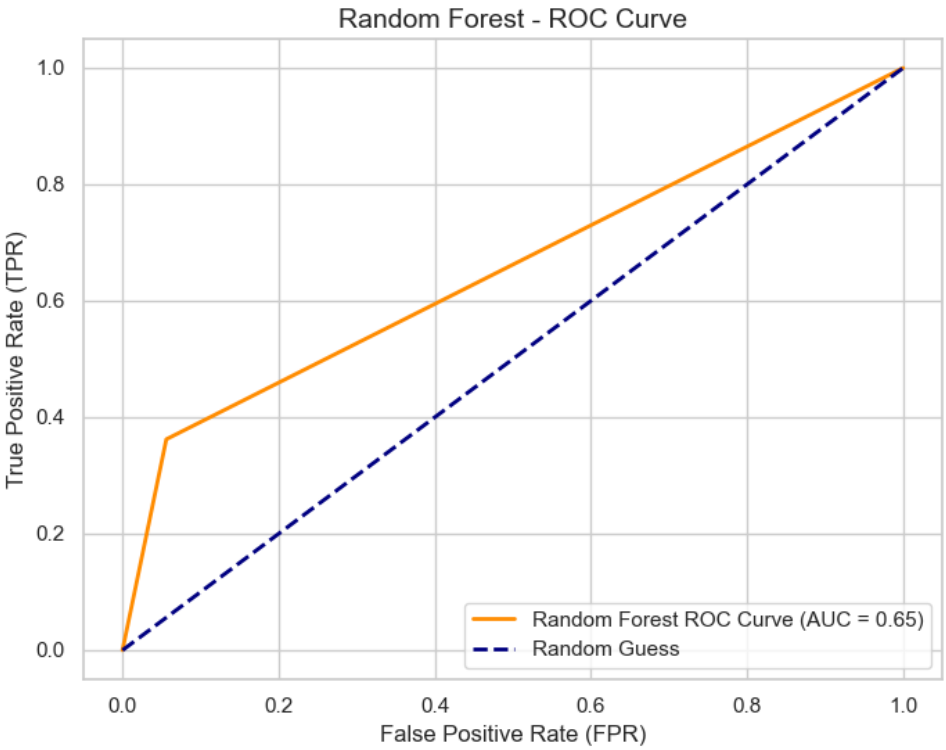

Figure 5: ROC curve of the random forest model

The ROC curve serves as a visual tool for assessing the classification performance of models. The horizontal axis represents the false positive rate (the proportion of instances that are actually negative

DOI: 10.54254/2754-1169/2025.LH23945

but incorrectly predicted as positive), while the vertical axis denotes the true positive rate (the proportion of instances that are actually positive and correctly predicted as positive). The diagonal line indicates the performance of random guessing (AUC=0.5).

AUC Value: The AUC value represents the area beneath the ROC curve and serves as critical metric for evaluating the classification performance of a model on given dataset. A higher AUC value indicates superior model performance, whereas a lower AUC value reflects diminished model performance. The logistic regression model achieves an AUC value of 0.60, whereas the random forest model attains an AUC value of 0.65. The random forest model demonstrates enhanced efficacy in distinguishing between normal and defaulting customers, as shown in Figure 6.

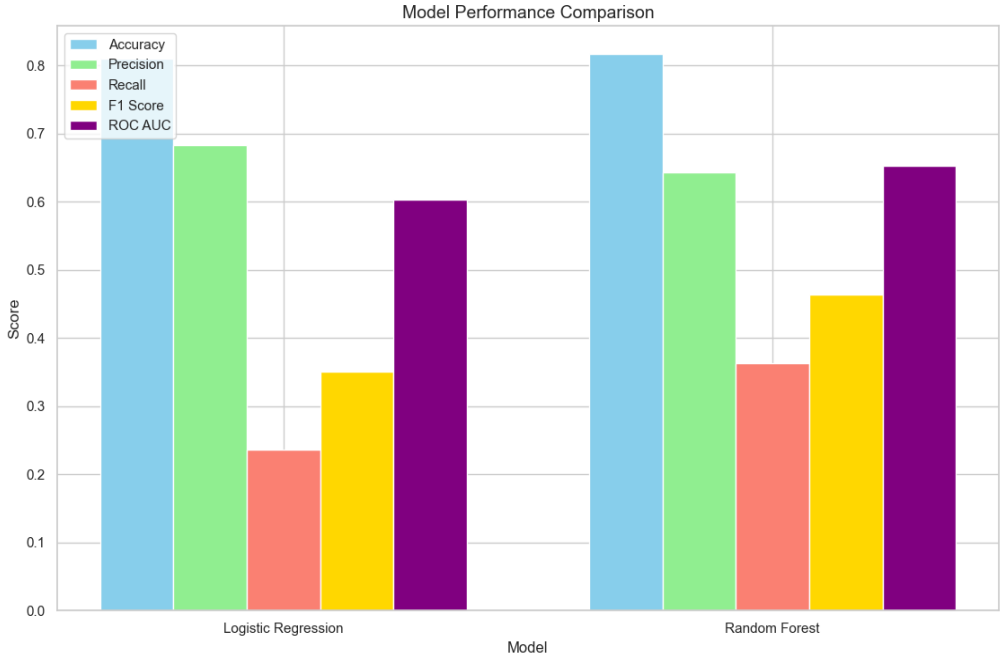

Figure 6: Model performance metrics chart

Model Evaluation Metrics Comparison: Model performance is assessed through following metrics (Table 2).

Table 2: Model performance

	Random Forest Model	Logistic Regression Model
Precision (Accuracy)	0.7012	0.6726
Recall (Sensitivity)	0.3813	0.2558
F1 Score	0.4932	0.3706
ROC AUC	0.65	0.60

Precision (Accuracy) indicates that the random forest model is more accurate in predicting normal customers.

Recall (Sensitivity) suggests that the random forest model is capable of identifying a greater number of defaulting customers.

F1 Score demonstrates that the random forest model achieves a better balance between precision and recall.

DOI: 10.54254/2754-1169/2025.LH23945

In summary, across all evaluation metrics, the random forest model outperforms the logistic regression model, particularly excelling in the ROC curve and AUC value.

4. Discussion

4.1. Reasons for the superiority of the random forest model

The random forest model demonstrates exceptional capability in handling nonlinear relationships and high-dimensional data, which is particularly relevant in the context of credit card default prediction where complex interactions among multiple features are predominantly nonlinear. By constructing multiple decision trees, the random forest model effectively captures these intricate nonlinear relationships.

Through the ensemble approach of aggregating predictions from multiple decision trees via voting or averaging mechanisms, the risk of overfitting of the random forest model is significantly lower than individual decision trees. This ensemble methodology substantially enhances the model's generalization capability.

Furthermore, the random forest model's implementation of random sampling for both features and instances renders it remarkably robust against noise and outliers in individual customer characteristics. This feature enables effective exclusion of exceptional cases, thereby further strengthening the model's generalization performance.

4.2. Practical recommendations and future research directions

Practical Recommendations: Financial institutions should prioritize key features such as recent repayment status and customer age, leveraging the random forest model to optimize risk management strategies, including credit limit adjustments and interest rate modifications.

Future Research: Potential research avenues include the incorporation of more sophisticated models (e.g., neural networks) or the integration of external data sources (e.g., consumer behavior data) to enhance predictive performance. Additionally, investigating model interpretability and generalization capabilities across different regions and time periods represents valuable research directions.

5. Conclusion

This study investigates the application of Random Forest and Logistic Regression models in credit card default prediction based on the Kaggle credit card default dataset. Through data preprocessing, feature selection, and comparative model analysis, the research aims to identify key features influencing credit card defaults and evaluate the predictive performance of different models. The research focuses on target variable distribution analysis, feature importance ranking, and model performance metrics such as accuracy, recall, F1 score, and AUC value.

The Random Forest model outperforms the Logistic Regression model in credit card default prediction, particularly demonstrating superior performance in handling non-linear relationships and high-dimensional data. Key features such as recent payment status (PAY 0), customer age (AGE), credit limit (LIMIT BAL), and bill amounts (BILL AMT1~6) significantly impact default prediction. The Random Forest model achieves an AUC value of 0.65, surpassing the Logistic Regression model's 0.60, and exhibits better performance in recall and F1 score, enabling more effective identification of potential default customers.

Future research on credit card default prediction could explore the use of ensemble models such as XGBoost and neural networks for modeling. Additionally, incorporating external data such as

DOI: 10.54254/2754-1169/2025.LH23945

credit social networks and employment status could enhance the accuracy of predicting potential default customers.

References

[1] *Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. "Expert Systems with Applications, 36" (2), 2473–2480.*

[2] *Hand, D. J., & 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.*

[3] *Zhou, L., Wang, H., & Zhang, X. (2018). Credit scoring using machine learning: A comparative study. "International Journal of Information Technology & Decision Making, 17" (4), 1093–1118.*

[4] *Liu, Y., Wang, Y., & Zhang, J. (2021). A new machine learning approach to credit scoring: Evidence from Chinese P2P lending markets. "Journal of Risk Model Validation, 15" (2), 1–22.*

[5] *Zhang, W., & Wang, T. (2020). Ensemble learning for credit scoring: A hybrid approach combining logistic regression and random forest. "Journal of Finance and Data Science, 6" (1), 1–15.*

[6] *Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Cost-sensitive logistic regression for credit scoring. "Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM)" (pp. 145–153). IEEE.*

[7] *Lessmann, S., Baesens, B., Seow, H.-V., & 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.*

[8] *Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" (pp. 785–794). ACM.*

[9] *Tsai, C.-F., & Chen, M.-L. (2010). Credit scoring by hybrid neural networks. "Expert Systems with Applications, 37" (4), 2660–2667.*

[10] *Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2005). Benchmarking state-of-the-art classification algorithms for credit scoring. "Journal of the Operational Research Society, 54" (6), 627–635.*