Interpretable Machine Learning Comparison for Credit Card Default Prediction

Jiaxin Guo

City University of Macau, Macau, 999078, China
b24090109085@cityu.edu.mo

Abstract. Credit card default has become an increasingly urgent issue in the financial field, causing huge economic losses and systemic risks. Traditional statistical methods, are no longer sufficient to handle the complexity and scale of modern financial data, especially their ability to manage nonlinear relationships and class imbalances is also very limited. The research aims to enhance credit card default prediction through interpretable machine learning. Three models - logistic regression (LR), random forest (RF), and eXtreme Gradient Boosting (XGBoost) - were evaluated using real-world credit card datasets from Taiwan. Optimize the model using grid search and validate the model through cross-validation. Performance is evaluated using Area Under Curve(AUC), precision, recall rate, f1 score and accuracy. SHapley Additive exPlanation (SHAP) is used to explain feature contributions and model decisions. The results show that the ensemble methods (RF and XGBoost) are significantly superior to LR, especially in dealing with imbalanced data. Repayment Status in the Most Recent Month (PAY_0), Credit Limit (LIMIT_BAL) and Repayment Status 2 Months Before the Most Recent Month (PAY_2) are the most influential predictors. On this basis, a dynamic analysis framework is proposed to help financial institutions identify high-risk customers and take preemptive measures. The research highlights the potential of explainable machine learning in credit risk analysis and provides actionable insights for financial decision-making.

Keywords: Credit Risk, Machine Learning, Credit Card Default, Model Interpretability, SHAP Analysis

1 Introduction

Credit card default has become a significant challenge in the financial sector today. When individuals fail to repay their credit card debts on time, it not only damages their personal credit scores but also results in the formation of non-performing assets for banks, thereby increasing the risk exposure of financial institutions. Accumulated bad debts can lead to substantial economic losses. In the long run, if credit card defaults become widespread, they may threaten the stability of the entire financial system and negatively affect the broader socio-economic order. According to a recent report released by the Federal Reserve Bank of New York, the total amount of credit card debt

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_31

held by American consumers has reached an unprecedented $1.14 trillion. Moreover, an increasing proportion of individuals are falling into credit card delinquency, indicating a growing financial strain among households [1]. The problem of credit card default is becoming increasingly serious. If this continues, it may also trigger a series of financial risks or economic fluctuations. Therefore, to deal with such problems, financial institutions need to predict the defaulting customer groups more accurately, not tend to customers who are very likely to cause bad debts and give reminders and warnings when customers are very likely to default and cut off funds in a timely manner. However, traditional statistical methods are clearly insufficient to complete these predictions; they can only rely on static indicators.

LR has traditionally been one of the most applied techniques in credit default prediction, primarily owing to its straightforward implementation and high level of interpretability [2]. However, it struggles to capture complex nonlinear patterns in high-dimensional data, limiting its predictive power [3]. As a result, its applicability in modern credit risk modeling has been increasingly questioned.

To address data imbalance issues in credit datasets, Naik employed the Synthetic Minority Oversampling Technique (SMOTE) as an initial step in their modeling process [4]. This approach significantly improved model performance by balancing the training samples.

The field has undergone substantial transformation in recent years, primarily propelled by rapid progress in big data technologies and advanced data analytics. Modern machine learning techniques, including RF, SVM, and neural networks, have been widely utilized in credit scoring applications due to their strong predictive capabilities and adaptability to complex data patterns. Modern methods are well-suited for handling high-dimensional and large-scale datasets and have consistently exhibited superior performance relative to traditional approaches. For example, Yeh and Lien compared various data mining techniques in predicting credit card default and found that these methods substantially outperform LR in terms of accuracy [5].

This study conducts a comparative analysis of three machine learning algorithms—logistic regression (LR), random forest (RF), and XGBoost—in the context of credit card default prediction. The study includes extensive data preprocessing, model development, and SHAP-based interpretability analysis to determine the most influential features contributing to default risk. Based on the experimental results, this paper proposes a dynamic customer profiling framework designed to improve the early identification of high-risk customers.

2 Methodology

2.1 Data Processing

The dataset adopted in this study is the Default of Credit Card Clients dataset, which was originally derived from the UCI machine learning library and has been publicly released on the Kaggle platform. This dataset covers the relevant information of credit card holders in Taiwan from April 2005 to September 2005, containing a total of 30,000 samples, each of which is composed of 25 characteristic

variables [6]. The dataset has no missing values; thus, the step of handling missing values is omitted. However, the ID column cannot be a factor affecting the default value of the credit card, so the ID column has been deleted. Convert categorical variables such as SEX, EDUCATION and MARRIAGE into numerical variables. And the data was divided. The data shows that 22% of people will default, which means that one out of every five people on the streets of Taiwan defaults. It can be imagined how large the number of credit card defaults is.

2.2 Model Selection

In selecting machine learning algorithms, this paper employs three models: LR, RF, and XGBoost. To enhance the overall performance of each model, hyperparameter tuning is conducted through grid search combined with cross-validation, ensuring robust and reliable optimization results. During the model training phase, multiple repetitions were conducted on the training and validation sets to ensure model stability and generalization capability. In a large-scale benchmark test study, Lessmann et al. systematically evaluated the performance of over 40 classification algorithms in credit scoring. The research results show that ensemble learning methods, especially RF and Gradient Boosting, consistently outperform traditional methods such as LR on multiple datasets, demonstrating stronger predictive capabilities. This finding further supports our decision to include XGBoost and RF in this study [7].

LR. LR is a commonly used supervised machine learning method, especially suitable for dealing with binary classification problems where the target variable is categorical [8]. As a generalized linear regression model, it is widely applied in fields such as data mining, automated disease diagnosis, and economic forecasting. The LR algorithm is simple and fast, and due to its probability output—which can be converted into binary probabilities—it also offers strong interpretability.

RF. RF is a classifier composed of multiple decision trees, and its final classification result is determined by counting the mode of the predicted categories of each decision tree. The earliest random forest model was proposed by Tin Kam Ho in 1995, and he utilized this method to implement the "stochastic discrimination" classification strategy proposed by Eugene Kleinberg [9]. RF offer several advantages: they are resistant to overfitting, capable of handling high-dimensional data, and generally achieve high accuracy.

XGBoost. XGBoost is an advanced machine learning technique that builds upon the Gradient Boosting Machine (GBM) framework. It is widely used in classification, regression, and ranking tasks. XGBoost builds a robust predictive model by combining multiple decision trees, with each tree iteratively correcting the errors

of its predecessors [10]. Thanks to its high flexibility, accuracy, and ability to perform feature importance analysis, XGBoost has become a popular choice in many data science competitions.

2.3 Evaluation Metrics

To evaluate the performance of credit card default prediction models, this paper adopts five indicators: accuracy rate, precision rate, recall rate, F1 value and AUC, comprehensively measuring the overall predictive ability of the model, the effect of identifying real defaulters, and the balance between precision rate and recall rate in the case of category imbalance.

3 Result

3.1 Model Performance Comparison

Table 1 comparing three models—LR, RF, and XGBoost—using five metrics: AUC, Accuracy, Precision, Recall, and F1 Score. RF shows highest F1 and Recall, XGBoost leads in AUC.

Table 1. Model performance comparison

Model	AUC	Accuracy	Precision	Recall	F1
XGBoost	0.759755	0.808667	0.618294	0.353089	0.449488
RF	0.756979	0.811889	0.630017	0.362632	0.460312
LR	0.685486	0.804111	0.653639	0.243596	0.354921

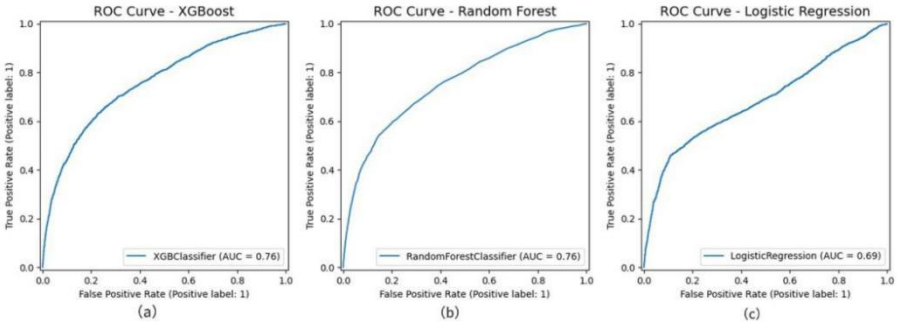

Fig. 1.ROC Curves for Three Models. (a) LR; (b) RF; (c) XGBoost (Photo/Picture credit: Original).

As shown in Fig. 1, both XGBoost and RF exhibit strong classification performance, with an AUC of 0.76. In contrast, the AUC of LR was relatively low at 0.69, indicating that its discriminative ability was relatively weak. Although the overall AUC values of XGBoost and RF are similar, Compared to XGBoost, Random Forest demonstrates a slight advantage in both recall and F1 score, as shown in Table 1. This advantage is

particularly valuable in cases of classification imbalance, as accurately identifying defaulters is of vital importance in such situations. The superior performance of the integrated model further highlights the limitations of traditional linear classifiers when dealing with complex nonlinear patterns in financial data.

3.2 Model Interpretability with SHAP

To understand the decision-making logic of the models, we applied SHAP to evaluate feature contributions. SHAP, proposed by Lundberg and Lee, is a unified approach based on cooperative game theory that attributes prediction scores to individual features in a consistent and interpretable manner [11].

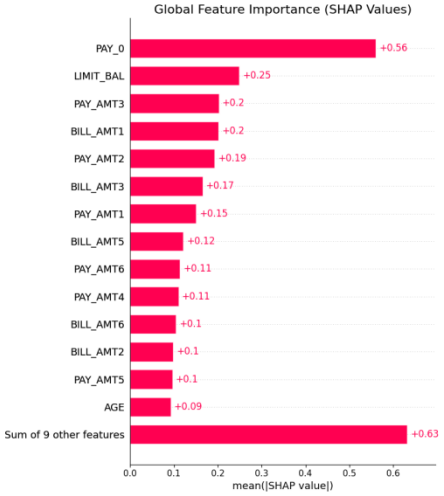

Fig. 2. Global Feature Importance based on SHAP Values (Photo/Picture credit: Original).

As shown in Fig. 2, this table shows the global importance ranking of each feature in the XGBoost model. PAY_0 is the most influential variable, with an average SHAP value of +0.56, followed by LIMIT_BAL, PAY_2 and other related variables.

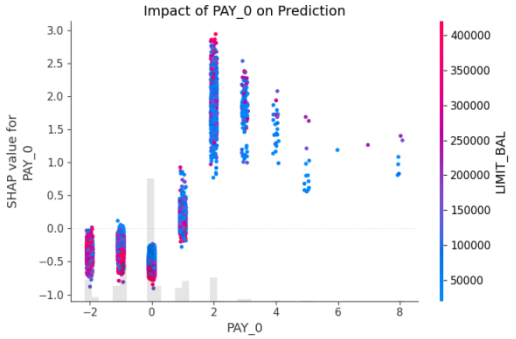

Fig. 3. SHAP Value Distribution for PAY_0 (Photo/Picture credit: Original).

As shown in Fig. 3, the chart shows the dependency between the PAY_0 variable and its SHAP value. The higher the value of PAY_0, the greater the severity of the overdue period, the larger the corresponding SHAP value, and the greater the possibility of being predicted as "default". Furthermore, customers with lower credit limits (blue dots) are more likely to be judged as having a high risk of default under the same overdue status, indicating that the model not only focuses on the default status but also comprehensively analyzes the liability capacity.

3.3 Analysis of Key Features

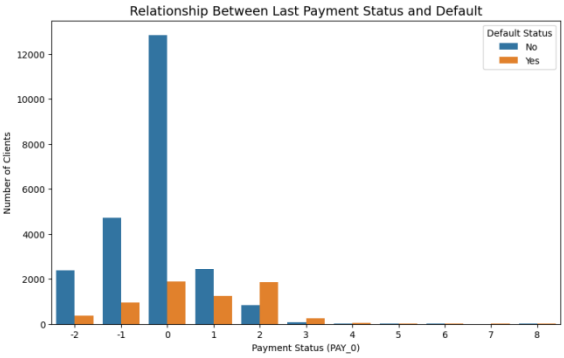

Fig. 4. PAY_0 vs Default Distribution (Photo/Picture credit: Original).

As shown in Fig. 4, the chart shows the distribution of PAY_0 between defaulting and non-defaulting customers. When PAY_0 is 0, the number of defaulting customers is much smaller than that of non-defaulting customers. However, when PAY_0 is larger, especially when PAY_0 = 2, the corresponding number of defaulting customers increases significantly. This trend indicates that whether a customer has defaulted recently is an important indicator for judging their future default risk.

4 Discussion

The comparison results of the three models provide important implications for the prediction of credit card defaults. The integrated models - RF and XGBoost - outperform LR in terms of AUC, recall rate and F1 score, especially in dealing with class imbalance problems. This finding is consistent with the comprehensive benchmark study by Lessmann et al., which evaluated over 40 credit scoring algorithms and concluded that ensemble models such as RFs generally outperform traditional methods like LR on multiple datasets [7].

To comprehensively evaluate the model's performance, the research prioritizes recall rate, AUC and F1 score rather than merely using accuracy. Although accuracy is a commonly used metric, it can be misleading in category-imbalanced data because it

focuses on overall prediction rather than identifying a few categories (defaulters). Relevant research indicates that TPR (Recall rate), precision rate and F1 score are more suitable for credit card default detection because they can better reflect the model's ability to identify key risk groups [12].

In terms of model interpretability, SHAP analysis provides us with a deeper understanding. PAY_0 is the most important variable affecting the prediction result. This finding aligns with previous research, indicating that both repayment behavior and individual characteristics are strongly associated with the risk of default. Customers with poor repayment status and low credit limits are more likely to be identified as defaulters by the model, indicating a significant superimposed effect of repayment behavior and credit exposure [13].

5 Conclusion

The research predicted credit card defaults by comparing the performance of three machine learning models: LR, RF, and XGBoost. The research results indicate that the ensemble learning methods, such as XGBoost and RF, outperform traditional LR, particularly in addressing class imbalance. The research uses a real dataset of credit card customers in Taiwan, optimizes the model through grid search, and conducts cross-validation. The model performance is comprehensively evaluated using AUC, accuracy, recall rate and F1 score. Although XGBoost performed slightly better in AUC, RF achieved the highest recall rate and F1 score, making it more suitable for identifying potential defaulters.

To improve interpretability, SHapley Additive explanation was adopted in this study. The results show that the most recent repayment status (PAY_0), credit limit (LIMIT_BAL), and the last most recent repayment status (PAY_2) are the biggest variables affecting default prediction. These findings highlight the significance of recent repayment behavior and credit exposure in risk assessment, reflecting the model's ability to capture the complex interactions among financial variables.

On this basis, the research has proposed a dynamic customer analysis framework to support financial institutions in the early identification and intervention of high-risk customers, thereby reducing potential losses and systemic risks. Future research may incorporate more behavioral data, such as transaction records, consumption frequency and geographic information, to further enhance the accuracy of predictions. In addition, the integration of deep learning technology and interpretable tools is expected to develop more intelligent and transparent credit risk management systems.

References

1. Federal Reserve Bank of New York: Household debt and credit report. https://www.newyorkfed.org/microeconomics/hhdc.html, last accessed 2024/08/04
2. Sperandio, S.: Understanding LR analysis. Biochemia Medica **24**(1), 12–18 (2014)
3. Zheng, A., Casari, A.: Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Sebastopol (2018)

4. Naik, K.S.: Predicting credit risk for unsecured lending: A machine learning approach (arXiv:2102.09511). https://arxiv.org/abs/2102.09511, last accessed 2024/08/04
5. Yeh, I.C., Lien, C.H.: 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 (2009)
6. Kaggle: Default of credit card clients dataset. https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset, last accessed 2024/08/04
7. 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 **247**(1), 124–136 (2015)
8. Couronne, R., Probst, P., Boulesteix, A.-L.: RF versus LR: a large-scale benchmark experiment. BMC Bioinformatics **19**(1) (2018)
9. Breton, M.L., Michelangeli, A., Peluso, E.: A stochastic dominance approach to the measurement of discrimination. Journal of Economic Theory **147**(4) (2012)
10. Hassan, M.A.M., RG, T., Mansur, U.M., Jha, R., Fahim, M.F.H., MT, R.: Interpretable machine learning models for credit risk assessment. In: Proceedings of the International Conference on Computing for Sustainable Global Development (2024), pp. 361–365. Publisher, Location (2024)
11. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. **30** (2017)
12. Baecova, A., Babice, F.: Predictive analytics for default of credit card clients. In: Proceedings of the 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 329–334 (2021)
13. Kalid, S.N., Khor, K.-C., Ng, K.-H., Tong, G.-K.: Detecting frauds and payment defaults on credit card data inherited with imbalanced class distribution and overlapping class problems: A systematic review. IEEE Access (2024)