Explainable Machine Learning for Personal Credit Default Prediction

Zixuan Qin

Hainan International College, Minzu University of China, Hainan, 572423, China
24160077@muc.edu.cn

Abstract. Predicting personal credit default is a vital part of managing financial risk for lenders. This paper proposes an explainable machine learning approach for predicting whether a credit card customer will default in the next month. The experiment employed the complete dataset of credit card clients from the UCI Machine Learning Repository, which includes 30,000 customer records. Three models are compared: Logistic Regression, Random Forest, and XGBoost. In order to make the study more meaningful than a simple model comparison, financial behavior features are constructed, including maximum repayment delay, average repayment status, credit utilization, and payment aggregates. The analysis further examines classification-threshold sensitivity and evaluates the stability of top-ranked explanatory features across five random seeds. The study revealed that XGBoost achieved the highest AUC score of 0.7822 and the highest recall score of 0.6436 after incorporating engineered features. Two of the most important and stable explanatory factors are the maximum repayment delay and the recent repayment status. The study argues that credit default models should be assessed not just for their predictive accuracy, but also for their decision thresholds and explanation quality.

Keywords: credit default prediction; explainable machine learning; financial behavior features; threshold sensitivity; SHAP stability

1 Introduction

Predicting a borrower's credit default is a key challenge in financial risk assessment. Banks and credit institutions must evaluate the risk of default before extending credit, setting credit limits, or creating risk-management policies. Therefore, credit default prediction is a decision-support problem with serious practical implications. In addition, credit decisions can impact customers' access to financial services. The prediction framework should therefore offer both reasonable predictive accuracy and explanations that analysts and decision-makers can easily review.

Previous studies have laid an important basis for credit default prediction. Yeh and Lien used a dataset of 30,000 credit card clients, compared multiple data mining methods like logistic regression, decision tree, neural networks, and discriminant analysis. In this work, data mining methods showed the ability to calculate the default probability, to help banks with credit risk management, not just to get simple ‘default’

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_154

or 'non-default' results [1]. Lessmann et al. benchmarked a large group of classification methods for credit scoring datasets and found that the more flexible machine learning models often have stronger predictive performance than traditional statistical models, but the best model depends on the dataset and evaluation standard [2]. Li and Wu applied an explainable machine learning approach to loan default prediction and adopted SHAP values to discover important risk factors. Their results showed that repayment-related variables and borrower-related variables were important predictors of default risk, and this supports the use of interpretation tools in credit risk analysis [3]. Chen et al. examined imbalanced credit scoring datasets and found that explanation methods such as SHAP can become less stable if the class distribution is hard to handle, and the interpretability results should also be checked carefully [4]. Song et al. designed a credit-rating-specific and a multi-objective ensemble learning method to predict the default probability of the loan, and their results also showed that ensemble designs can improve the prediction in more challenging credit risk settings [5]. Han et al. proposed an explainable oversampling method for credit scoring and found that sampling strategies can have an impact on the performance and explanation qualities of the model [6]. Overall, these studies suggest that the modern credit scoring research should be concerned not only with prediction accuracy but also with model comparison, threshold selection, feature interpretation, and explanation validity.

This paper develops a transparent machine learning model for personal credit default prediction. It leverages the full UCI credit card default dataset, creates financial behavior features, runs comparisons between Logistic Regression, Random Forest, and XGBoost, explores thresholds ranging from 0.30 to 0.70, and verifies SHAP feature stability. In the experiment, XGBoost achieves an AUC of 0.7822 and a recall of 0.6436, while the most stable factors are maximum repayment delay and recent repayment status.

2 Method

2.1 Dataset and Financial Behavior Feature Construction

The dataset used in this paper is the Default of Credit Card Clients dataset from the UCI Machine Learning Repository [7]. The dataset includes 30,000 observations and 23 descriptive variables encompassing credit limit, demographic information, repayment status, bill amounts, and payment amounts. The target variable is the default payment in the next month. In this paper, it is renamed as default status, where 1 means that the customer defaults in the following month and 0 means that the customer does not default. The ID variable is removed because it is only an index number and does not describe customer behavior.

The dataset is suitable for this study for three reasons. First, the issue is directly linked to a personal credit default, matching the research topic. Second, it combines both financial behavior variables and customer profile variables, which makes it possible to study prediction and interpretation at the same time. Third, it is a widely used public dataset, so the experiment can be easily verified and repeated by other

researchers. The data are moderately imbalanced. There are 23,364 non-default cases and 6,636 default cases, so the default rate is 22.12%.

To extend the previous basic model comparison, this paper constructs several financial behavior features (FBFs). The maximum repayment delay (MRD) records the worst repayment delay in the six-month period. Average repayment status (ARS) summarizes the general delay level. Recent utilization and average utilization measure how much of the credit limit is used. Average payment amount and average bill amount summarize repayment ability and bill burden. Bill trend captures whether the bill amount is increasing or decreasing. These variables are not new information from outside the dataset. They are reorganized from the original monthly variables, but they make the financial meaning clearer.

2.2 Models and Explainability Tools

Three predictive models are evaluated. Logistic Regression (LR) is a benchmark. The model is relatively simple, and the direction of the coefficients is relatively easy to read [8]. Random Forest (RF) is employed here because it combines multiple decision trees, which can capture nonlinear relationships between variables [9]. XGBoost is included because boosting methods often perform well on tabular classification data and can handle complex interactions between repayment, bill, and payment variables [10].

For a more detailed explanation, this paper uses SHAP. SHAP is a method that assigns a contribution score to each feature in a model prediction [11]. The study primarily used SHAP with XGBoost because that combination delivered the best AUC in the main experiment. The goal here is not to cover every mathematical aspect of SHAP, but to illustrate how it can be used as a practical tool to identify customer behaviors that steer the prediction towards the default. In order to avoid relying on one explanation result only, SHAP feature rankings are also compared across five random seeds.

2.3 Experimental Design

The dataset is separated into an 80/20 stratified train-test split, resulting in 24,000 observations for training and 6,000 for testing. Stratification keeps the default rate close to 22.12% in both sets. Categorical variables are cleaned and coded as needed, and numerical variables are standardized for LR.

The experiment is made up of three parts. First, the three models are compared after adding engineered FBFs. The metrics are accuracy, precision, recall, F1-score, and AUC. Second, the impact of these engineered features is assessed by comparing the predictions made by XGBoost using only raw variables with those made by XGBoost using both raw and engineered variables. Third, the study looks at the model's performance under different thresholds of 0.30 to 0.70. It's important to note that a default threshold of 0.50 might not be ideal in credit risk. Finally, the stability of SHAP explanations is assessed by recording the top features from five random runs and calculating their ranking consistency.

3 Results

3.1 Overall Model Performance

Table 1 shows the key predictive outcomes after incorporating engineered FBFs. XGBoost achieved the highest AUC of 0.7822 and the highest recall of 0.6436. RF achieves a slightly higher accuracy of 0.7560 and an almost identical F1-score. LR provides a clear baseline, although it's less effective than the ensemble models in terms of AUC.

The study reveals that model selection is influenced by the user's goals. If the primary objective is to rank customers by risk, AUC is critical, and XGBoost tends to perform slightly better. For a well-balanced classification result, RF is also a good choice. If the user prioritizes easy understanding and a more traditional statistical approach, LR is still a reasonable starting point. In this sense, the best model is not only a technical answer. The decision also depends on the type of credit the institution wants to offer.

Table 1. Model performance with engineered financial behavior features.

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.7327	0.4288	0.6285	0.5098	0.7305
Random Forest	0.7560	0.4618	0.6240	0.5308	0.7785
XGBoost	0.7485	0.4519	0.6436	0.5309	0.7822

3.2 Effect of Financial Behavior Feature Construction

Table 2 shows the impact that feature construction has on the XGBoost model. After adding these features, AUC increases from 0.7801 to 0.7822, and recall increases from 0.6292 to 0.6436. At the same time, accuracy and F1-score decrease slightly. The addition of these features, therefore, enhances sensitivity to default cases without improving all metrics at once.

For risk analysis, this trade-off is meaningful. In many financial situations, missing a real default customer can be more expensive than wrongly flagging a low-risk one. So the improved recall has practical value, even if it's not seen across all metrics. The engineered features also make the model more approachable and easier to discuss. For instance, it's easier to explain the MRD than to outline six different repayment statuses.

Table 2. XGBoost performance before and after adding engineered financial features.

Feature set	Accuracy	Precision	Recall	F1-score	AUC
Raw features only	0.7612	0.4702	0.6292	0.5382	0.7801
Raw + engineered features	0.7485	0.4519	0.6436	0.5309	0.7822

3.3 Threshold-Sensitive Analysis

Table 3 reveals how the XGBoost prediction varies across different thresholds. At 0.30, recall increases to 0.8877, but false positives rise significantly. At 0.50, the model's middle point is more balanced.

At 0.70, precision rises to 0.6254, and false positives decline to 330, but recall drops to 0.4152, and false negatives increase to 776. These findings reveal that a fixed threshold is not neutral; it reflects a policy decision about whether to emphasize default detection or to avoid unnecessary warnings.

Table 3. XGBoost threshold-sensitive evaluation.

Threshold	Accuracy	Precision	Recall	F1-score	False negatives	False positives
0.3000	0.5215	0.3021	0.8877	0.4507	149.0000	2722.0000
0.4000	0.6647	0.3721	0.7506	0.4975	331.0000	1681.0000
0.5000	0.7485	0.4519	0.6436	0.5309	473.0000	1036.0000
0.6000	0.7983	0.5449	0.5350	0.5399	617.0000	593.0000
0.7000	0.8157	0.6254	0.4152	0.4991	776.0000	330.0000

3.4 SHAP Interpretation and Stability

Table 4 lists the ten SHAP variables that are most influential when using XGBoost with engineered features. MRD takes first place, followed by most recent repayment status (MRRS), most recent credit utilization ratio, and ARS.

Table 5 shows how stable these five random seed explanations are. The frequency of MRD and MRRS in the top ten rankings is consistent across all runs, with a pairwise Spearman rank correlation of 0.9583. The explanation ranking was relatively stable across the different random splits in this experiment.

Table 4. Top SHAP features for the XGBoost model.

Rank	Feature	Mean absolute SHAP value
1	maximum repayment delay	0.4460
2	most recent repayment status	0.3091
3	most recent credit utilization ratio	0.1159
4	average repayment status	0.1086
5	most recent bill statement amount	0.0978
6	credit limit	0.0919
7	average payment amount	0.0743
8	average credit utilization ratio	0.0738
9	bill amount trend	0.0667
10	most recent payment amount	0.0665

Table 5. SHAP top-feature stability across five random seeds.

Feature	Appearances in the top 10	Best rank	Average rank
maximum repayment delay	5	1	1.0000
most recent repayment status	5	2	2.0000
most recent credit utilization ratio	5	3	3.8000
average repayment status	5	3	4.2000
credit limit	5	3	4.2000
most recent bill statement amount	5	6	6.6000
average payment amount	5	6	7.2000
average credit utilization ratio	5	5	7.2000
previous payment amount	4	9	9.2500
average bill statement amount	3	8	9.0000

4 Discussion

Model evaluation should consider recall, AUC, engineered behavior features, and threshold choices, because accuracy alone cannot reflect missed defaults and false-alarm trade-offs in credit risk.

The SHAP stability result is a useful addition to the explanation part. The study looked at whether the key features were the same across multiple runs. The stable ranking of MRD and MRRS supports the use of these features as major warning signals. Meanwhile, recent studies remind researchers that explanations might be unstable, particularly in unfair or cost-sensitive settings [4, 12, 13]. This study only looks at stability in a simple way, but it is a step toward more careful use of explainable machine learning.

There are still some issues that need to be addressed. The data is sourced from a single credit card market and an earlier time period, so the findings may not be broadly applicable to all countries or the current online lending market. The engineered features use only the variables available in the dataset. Income, employment, macroeconomic conditions, and real lending costs are not included. The paper also emphasizes prediction and explanation, rather than causal inference. Future research can use newer datasets, incorporate cost-sensitive learning, test fairness across different customer groups, and investigate whether local explanations for individual customers are stable enough for real credit decisions.

5 Conclusion

This paper presents an explainable machine learning framework for predicting personal credit default, based on the full UCI credit card default dataset. The study

compares LR, RF, and XGBoost, develops interpretable FBFs, tests several classification thresholds, and investigates the stability of SHAP rankings. The results reveal that XGBoost's performance is best when engineered features are incorporated. RF maintains strong accuracy and F1-score. An analysis of thresholds shows that default detection and false alarm rates can vary significantly depending on how the cutoff is set. SHAP analysis indicates that MRD and MRRS have the largest and most stable predictive contributions, suggesting that repayment delay and recent repayment status are key indicators of default risk. The study is constrained by the historical and market-specific nature of the dataset and by its focus on prediction rather than causal explanation. Future work can explore how to make the system fair, how to calibrate it, how stable the local explanations are, and how the decision rules can account for the economic costs of credit-risk errors.

References

1. Yeh, I.C., Lien, C.H.: The comparison 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)
2. 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)
3. Li, H., Wu, W.: Loan default predictability with explainable machine learning. Finance Research Letters 60, 104867 (2024)
4. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research 312(1), 357–372 (2024)
5. Song, Y., Wang, Y., Ye, X., Zaretzki, R., Liu, C.: Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. Information Sciences 629, 599–617 (2023)
6. Han, S., Jung, H., Yoo, P.D., Provetti, A., Cali, A.: NOTE: non-parametric oversampling technique for explainable credit scoring. Scientific Reports 14(1), 26070 (2024)
7. UCI Machine Learning Repository: Default of Credit Card Clients Dataset, https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients, last accessed 2026/06/16
8. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. 3rd edn. Wiley, Hoboken (2013)
9. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
10. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
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. Ballegeer, M., Bogaert, M., Benoit, D.F.: Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. European Journal of Operational Research 326(3), 630–640 (2025)
13. Lin, L., Wang, Y.: Shap stability in credit risk management: A case study in credit card default model. Risks 13(12), 238 (2025)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

