# Comparative Analysis of Bank Loan Risk Forecasting Based on Machine Learning Models Versus SHAP Interpretation

Yuning Wang

School of Management and Economics, Tianjin University, Tianjin, 300072, China
18704168455@163.com

**Abstract.** With the rapid development of the social economy, loan scales have expanded quickly. While banks and other financial institutions can generate considerable profits, this also means a significant increase in the risk of loan defaults. Given the vast amount of high-dimensional customer data, machine learning technology needs to be introduced. Based on 148,670 loan records from the Kaggle platform, this study systematically compared the performance of eight models—logistic regression, decision trees, random forests, gradient boosting, XGBoost, LightGBM, and soft voting and stacked integration—in default prediction, and model stability is evaluated through 5-fold cross-validation, the Shapley Additive Explanations method (SHAP) is used to interpret global feature importance, and robustness is verified from two dimensions: grouping tests based on loan amount and noise resistance tests on standardized interest rates. The results show that tree-based ensemble models are significantly superior to logistic regression models; Random Forest achieved the highest Area Under Curve (AUC) but low recall, while stacked integration achieved a better balance between accuracy and recall; Interest rates, credit types, negative amortization, debt income, and loan-to-valueratios are key risk characteristics; Interest ratecharacteristics are highly sensitive to noise, while the model performs stably across different loan amount groups.

**Keywords:** Loan Default Prediction, Machine Learning, Ensemble Learning, SHAP Interpretation, Model Robustness

## 1 Introduction

With the development of machine learning, scholars at home and abroad have conducted more extensive research on credit default prediction. First, for a single model, such as Qu et al., a logistic regression model was used to predict corporate credit risk and establish a credit strategy optimization model [1, 2]. With the development of internet finance, traditional single machine learning algorithms can no longer meet the needs of high-dimensional features and large-scale data processing. Therefore, researchers have adopted ensemble models for comparison. For example, Du selected traditional models, KNN, logistic regression, random forest, and stacked ensemble models to analyze personal loan defaults, finding that the accuracy of single models

© 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_157

was lower than that of integrated models [3]. Cao and others used a voting integration model to predict farmers' loan defaults [4].

According to current academic literature, the main gaps in the field of loan default prediction are as follows. First, the model comparison is not comprehensive enough; many studies only compare 2-4 models, lacking a comprehensive horizontal comparison between multiple single models and multiple integrated models. Second, much research focuses solely on improving prediction accuracy but neglects the explanation of the model's decision logic, affecting business personnel's understanding. Third, there is a lack of feature sensitivity and robustness, with little exploration of the impact of a core feature on model predictions.

This study employs eight machine learning models, including two single models and six integrated models. Through cross-validation, SHAP interpretation, group testing, and noise disturbance, the system evaluates and compares the performance of each model and analyzes feature importance and robustness, aiming to provide interpretable and stable decision support for loan default prediction.

# 2 Method

## 2.1 Datasets and Data Preprocessing

The loan default dataset used in this study contains 148,670 loan application records, each containing 34 raw features and a binary target variable [5]. Original characteristics include loan amount, interest rate, loan term, applicant's income, property value, type of credit record, and demographic information. All data were collected in 2019 and do not include time series information.

Before modeling, this paper performs the following data preprocessing steps.

Constructing advanced features, converting age segments into median values; Calculate the ratio of loan amount to income, loan amount to property value, and income to property value; The three classification variables—early approval mark, commercial loan mark, and open credit mark—were converted into a 0/1 mark; Construct the interaction between loan term and loan amount; If the original loan-to-value ratio (LTV) ismissing, the ratio of the loan amount to the property value is used to fill the gap.

Next, remove the leaked features to ensure the model's authenticity and generalization capability. Leakage features refer to information that cannot be obtained in real life before lending or variables directly related to whether a default occurs. Therefore, based on business experience, interest rate spreads, interest rate spread interaction items, and prepaid fees have been removed.

Then the missing values are handled, first deleting columns with a missing rate exceeding 50% to avoid excessive insertion of padding values that would reduce model performance; For numerical features, the median is used to fill the missing value to eliminate the influence of extreme values on the fill value; Since category features have no numerical intermediate value, modes are used to fill in missing values.

Numerical features and category features are standardized and coded separately. After standardization, the mean is 0, and the variance is 1; Assign sequence numbers to category features, with unknown category codes set to -1.

## 2.2 Selection and Application of related Models

This study selected eight machine learning models, including foundational models and integrated models, for loan default prediction.

Logistic Regression, as a linear reference model, uses L2 regularization with a regularization strength of C=0.1 and a maximum of 1000 iterations. It can output default probability and has good interpretability, but struggles to capture nonlinear interactions between features. The Decision Tree is set to a maximum depth of 5 and a minimum number of split samples of 20 to prevent overfitting. It can automatically handle nonlinear relationships and feature interactions, and the generated rules can be directly understood by business personnel. The Random Forest contains 100 decision trees, each with a maximum depth of 10 and a minimum splitting sample size of 20. It reduces variance through bootstrap sampling and random feature subspaces, enhancing generalization capability. Gradient Boosting uses 100 regression trees, with a learning rate of 0.1 and a maximum depth of 3, allowing gradual optimization and prediction through iterative fitting of residuals. XGBoost is set with 100 trees, a maximum depth of 3, a learning rate of 0.1, logarithmic loss as the evaluation metric, and introduces second-order derivatives and regularization terms, achieving fast training speed and excellent performance. LightGBM parameters are basically the same as XGBoost, but it uses a histogram-based decision tree algorithm, which uses less memory and is faster to train.

Additionally, Voting Classifier integrates Random Forests, Gradient Boosts, and XGBoost as base learners, integrating soft voting (i.e., average prediction probability). Stacking Classifier uses base learners for soft voting, with metalearners using logistic regression (C=0.1), and base learners' predicted probabilities as meta feature input [6]. All models were randomly seeded at 42 to ensure reproducibility.

## 2.3 Design of the Evaluation Process for Loan Default Experiments

**Model Evaluation Metrics and Cross-Validation**. First, accuracy, precision, recall, F1 score, and AUC were selected as evaluation indicators and applied to the eight selected models [7].

To more stably evaluate the model's generalization ability and stability, 5-fold cross-validation was performed on all models on the training set, and finally, the mean and standard deviation of the five AUC values were calculated. The mean reflects the model's average discrimination ability, while the standard deviation measures the degree of performance fluctuation when dealing with different datasets. The smaller the standard deviation, the less sensitive the model is to sampling changes in training data, and the more stable the generalization performance.

Since cross-validation uses all training data for multiple training and validation, the results better reflect the model's adaptability to different data subsets compared to

previous single-test divisions. Therefore, cross-validation of the previous eight models further reflects the overall performance of these models in this study.

**Relevant Explanations of the Model**. To facilitate reviewers and regulatory agencies in understanding the relevant models and to enhance the practical value of the research, this study uses the SHAP method to analyze the importance of all features of all models, quantifying the specific contribution of each feature to the model's prediction.

By summarizing the SHAP values of all samples on the test set, a global feature importance map is obtained, from which features that have the greatest impact on the overall probability of default can be identified. At the same time, the SHAP value has direction, indicating whether each feature increases or reduces default risk, helping to better understand the impact of each feature on prediction results [8].

**Robustness Verification of the Model**. Finally, to test the model's performance stability under interference from different customer groups and feature noise, this study continues robustness validation of the model from two perspectives.

First, considering that loan amounts may affect the model's discriminating ability, the test set was divided into low, medium, and high groups based on the loan amount quartiles, and AUC was calculated for each sample. If the AUC differences among the three groups are too large, it indicates performance bias in loan amount changes, and this verification demonstrates model fairness.

Next, for the most important features found in the previous SHAP analysis, errors during data collection or entry are simulated by artificially adding Gaussian noise of varying intensities. Specifically, zero-mean Gaussian noise with standard deviations of 0, 0.05, 0.1, 0.2, and 0.5 is sequentially added to the interest rate features of the test set, and the AUC on the disturbed test set is recalculated. By observing the trend of AUC as noise standard deviation increases, the model's sensitivity to the quality of key feature data can be assessed. If AUC drops sharply with increasing noise, it indicates that the model is highly dependent on the accuracy of this feature, and its quality needs to be specially monitored in practical applications; Conversely, if the AUC decreases slowly, it indicates the model has strong anti-interference capability.

# 3 Result

## 3.1 Preliminary Comparison of All Model Performance Predictions

The performance of all models on the test set is shown in Table 1 (sorted in descending order by AUC metric). From the AUC indicator, random forests performed best, followed by stacked integration and voting integration, with very little difference among the three. Gradient Boost, XGBoost, and LightGBM also achieve very high differentiation capabilities. The decision tree's AUC is 0.97276, significantly lower than the aforementioned tree ensemble model. Logistic regression performed the worst, with an AUC value of only 0.74467, indicating a strong nonlinear relationship between

features and default risk, and tree and ensemble models can effectively capture these findings.

In terms of accuracy and recall, random forests have the highest accuracy but relatively low recall, indicating a tendency to conservatively predict defaults, i.e., only when there is high certainty that a default is judged. In contrast, XGBoost and LightGBM have recall rates exceeding 0.86960 while maintaining an accuracy of about 0.84500, striking a good balance between false positives and missed positives. Gradient Boosting has both accuracy and recall rates between 0.85822 and 0.85974, with the highest F1 score, making it the model that best balances misjudgments and missed positives. The F1 score of the integrated model is about 0.85159~0.85708, with stable overall performance. The stacked ensemble model performs slightly better than the voting ensemble model in all aspects under this scenario.

**Table 1.** Performance metrics of each model on the test set

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| Random Forest | 0.91794 | 0.97127 | 0.68736 | 0.80502 | 0.98465 |
| Stacking (RF+GB+XGB+LR) | 0.93129 | 0.87929 | 0.83597 | 0.85708 | 0.98452 |
| Voting RF+GB+XGB) | 0.93018 | 0.89429 | 0.81277 | 0.85159 | 0.98445 |
| Gradient oosting | 0.93055 | 0.85974 | 0.85822 | 0.85898 | 0.98422 |
| XGBoost | 0.92863 | 0.84523 | 0.86968 | 0.85728 | 0.98358 |
| LightGBM | 0.92833 | 0.84358 | 0.87063 | 0.85689 | 0.98350 |
| Decision ree | 0.90751 | 0.80158 | 0.83024 | 0.81566 | 0.97276 |
| Logistic egression | 0.78614 | 0.68918 | 0.24086 | 0.35696 | 0.74467 |

### 3.2 Partial Representation of the Cross-Validation Results of the Model

Due to space limitations, after cross-validating all models, the results of cross-validation for four models are listed here. It chose to show these four models because they all had high AUC values in the preliminary test and represented tree and composite model structures, respectively, covering a wide range.

Cross-validation of the four representative models by 50% is shown, with the mean and standard deviation of AUC shown in Table 2 (sorted in descending order by the average AUC indicator). Stack integration achieved the highest average AUC, indicating that the stacked integration model has the strongest average discrimination ability, followed by Random Forest, XGBoost, and LightGBM. All models have standard deviations less than 0.00070, indicating that the models perform very stably across different training subsets, have strong generalization capabilities, and can effectively predict the test set and future new loan applications.

**Table 2.** 5-fold cross-validation AUC results

| Model | CV AUV Average | CV AUC Standard Deviation |
|---|---|---|
| Stacking (RF+GB+XGB + LR) | 0.98404 | 0.00058 |
| Random Forest | 0.98394 | 0.00061 |
| XGBoost | 0.98321 | 0.00057 |
| LightGBM | 0.98310 | 0.00050 |

### 3.3 Model Interpretation and Robustness

After SHAP interpretation of all models, due to space limitations, only the XGBoost model is selected here for the results presentation, as it performs well overall and represents the most widely used tree model in the study.

Fig. 1 shows that for the XGBoost model, the main characteristics affecting default probability are the standardized interest rate, credit type, negative amortization type, debt-to-income ratio, and loan-to-value ratio.

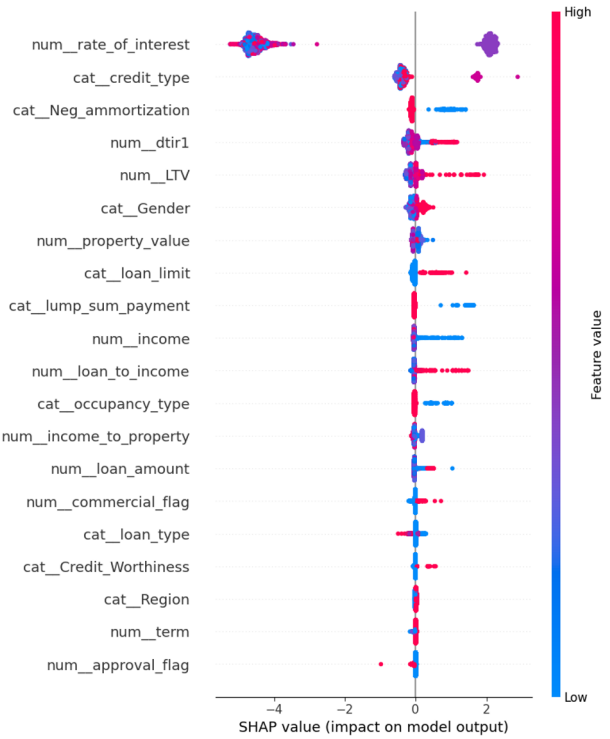



**Fig. 1.** Importance of global features of the XGBoost model in SHAP (Photo/Picture credit: Original).

### 3.4 Grouping Robustness and Noise Resistance Results of the XGBoost Model

To ensure research continuity and readability, the results of grouping robustness and noise resistance of the XGBoost model discussed earlier are continued here.

Based on the loan amount, the test set was divided into low, medium, and high groups. The XGBoost model's AUCs in each group were 0.97920, 0.98230, and 0.98860, respectively. As shown in Fig. 2, the XGBoost model performed well across different loan size groups and had stronger differentiation ability for applicants with high loan amounts.

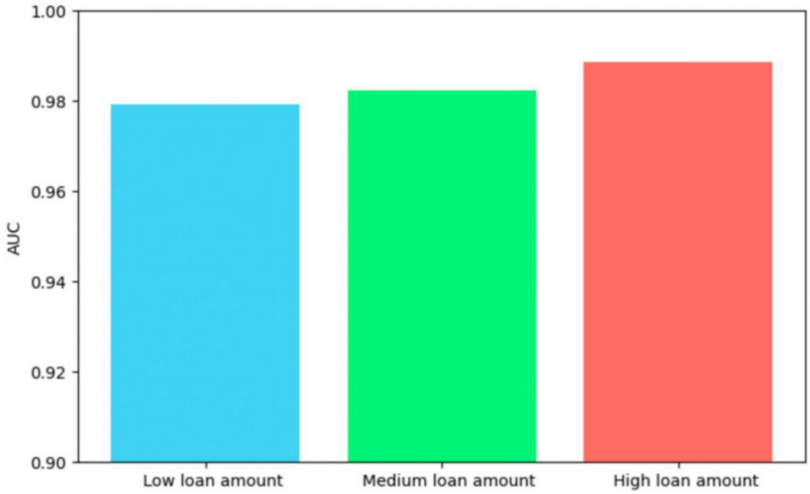



**Fig. 2.** Model AUC performance for different loan amount groupings (Photo/Picture credit: Original).

Afterwards, Gaussian noise is added to the most critical feature—the standardized interest rate feature. As the noise standard deviation increases from 0 to 0.05, as shown in Fig. 3, the AUC sharply drops from 0.98360 to 0.86180; as the noise continues to increase to 0.5, the AUC slowly decreases to 0.85220. This indicates that the model heavily relies on interest rate features, and once these features are contaminated, such as data entry errors, model performance will deteriorate significantly.

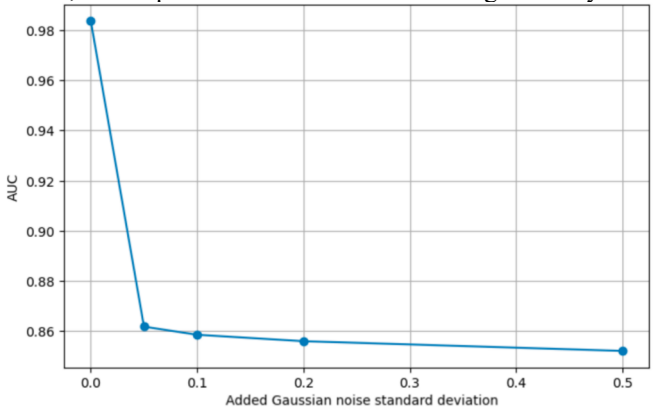



**Fig. 3.** Rate Feature Changes in AUC After Gaussian Noise Addition (Photo/Picture credit: Original).

## 4 Discussion

This study mainly uses 2019 cross-sectional data for modeling analysis, systematically compares the performance of eight machine learning models on loan default prediction tasks, and uses the SHAP method to analyze the model's interpretability, while group testing and noise disturbance are used to verify the model's robustness. The following discussion will be conducted from three aspects: its relationship with existing research, its limitations, and directions for improvement.

The results of this study show that tree-based ensemble models significantly outperform logistic regression in loan default prediction, which is consistent with the mainstream conclusions in the academic community. For example, Duan compared decision trees and logistic regression based on datasets from state-owned commercial banks, finding that the classification effect of decision tree methods under large samples was far superior to logistic regression and discriminant analysis [9]. Meanwhile, for the non-equilibrium data used in the study, the stacked ensemble model performed slightly better than the voting ensemble model, which is consistent with Chen Yihui's research on personal loan default prediction based on the Stacking fusion model under non-equilibrium data [8].

Since the data year used in this study is 2019, the generalization ability of the model in the face of a future environment or different years is unknown. It should try to collect multi-year loan data for verification and evaluate the stability of the model in different economic cycles. At the same time, the data in this study are unbalanced, that is, the number of customers with normal repayment is much higher than the number of customers with default. In this case, although the AUC index will not be affected too much, the research of Zheng and others shows that imbalance will lead to the classification model being biased towards the majority of classes, which will degrade the performance when predicting a few classes, to solve this problem [10]. It can adjust the distribution of data at the data level in the future, and solve the problem of unbalanced distribution of data sets through data oversampling, undersampling, or data grouping, such as using the classic oversampling SMOTE algorithm to generate new samples to balance data [11]. In addition, due to space constraints, the 50% cross validation results of all models and the results interpreted by Shap are not listed, and the grouping robustness and anti-noise tests only deal with the standardized interest rate characteristics of the XGBoost model; the processing results of all models can be listed in subsequent studies. The research of Li Qing and others showed that in the process of credit data collection and transmission, multiple features may be affected by transmission noise at the same time and change the prediction results [12]. So the simultaneous disturbance of multiple features can be added in the future anti-noise test.

## 5 Conclusion

By comparing a variety of machine learning models, this study found that the tree-based integration methods (Random Forest, Gradient Boosting, XGBoost, LightGBM) were far better than logistic regression in the task of loan default

prediction, which was consistent with the existing research conclusions. Random Forest achieved the highest AUC, but the recall rate was low, indicating that its failure rate in the default sample was high. The voting integration model and the stacking integration model have achieved a better balance between recall rate and accuracy rate. The overall performance of the voting integration model is better than that of the single model, and the stacking integration model is slightly better than the voting integration model, but the improvement range is limited. This may be because the basic learning devices, such as Random Forest, Gradient Boosting, XGBoost, etc., are already very powerful, and the marginal benefit of further integration is small.

It is worth noting that the model used in the study shows high sensitivity to the characteristics of interest rates - adding small Gaussian noise will lead to a significant decline in AUC. This finding proves that the accuracy of interest rate data must be guaranteed in practical applications. At the same time, multiple checks or robust regression can be considered for interest rates. In addition, the performance of the model in different loan amount groups has little difference, which indicates that it has good fairness and will not significantly favor a certain loan amount group.

This study also has limitations, and future research can start from these limitations. If the year in the data is only 2019, multi-year data can be further collected to verify the model; At the same time, the problem of category imbalance has not been specially dealt with, and SMOTE oversampling method can be used in data processing in subsequent research; In addition, due to content limitations, some processing results only show part of the model, which can be supplemented later; The noise test is only for a single feature, and the simultaneous disturbance of multiple features can be considered in the subsequent research.

The loan default prediction method constructed in this paper can be widely used in the scenarios of credit approval of commercial banks, anti-fraud of financial companies, and risk control audit of Internet Financial platforms. In the future, the trained model can be encapsulated as an API service and integrated into the bank's credit approval system to significantly reduce the approval time of customers' loan applications, to achieve rapid and accurate risk assessment, carry out dynamic risk monitoring, and assist financial institutions in making better and more scientific lending decisions.

## References

1. Su, X., Guo, H., Liang, L.: Multi-model prediction of personal loan default risk based on machine learning. Journal of Changsha University, 1–11 (2026)
2. Qu, Z., Wu, H., Li, F.: Credit risk assessment and credit strategy optimization modeling of SMEs based on logistic regression. Journal of Shandong University (Natural Science) (8), 94–102 (2024)
3. Du, S.: Research on loan default prediction model based on stacking ensemble learning. Master’s thesis, Dongbei University of Finance and Economics (2022)
4. Cao, W., Zhang, J.: Prediction of farmers' loan default risk based on unbalanced data processing and weighted soft voting heterogeneous ensemble. Computer Applications and Software 42(8), 71–79 (2025)

5. Yasser, M.H.: Loan default dataset, https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data, last accessed 2026/06/16
6. Liang, Z., Liang, H., Huang, Y.: Application of machine learning in financial loan default prediction. Modern Computer 30(24), 103–107, 113 (2024)
7. He, D., Mu, Y.: Credit card default risk assessment and prediction based on stacking model fusion. Mathematical Modeling and Its Applications 14(4), 49–56 (2025)
8. Chen, Y.: Personal loan default prediction and SHAP explanation based on stacking fusion model under imbalanced data. Master's thesis, Jiangxi University of Finance and Economics (2025)
9. Duan, Y., Zhang, J., Cao, L.: The comparative analysis of the models in default warning of the credit clients in commercial banks. The Journal of American Science 3(2), 30–33 (2007)
10. Zheng, L., Liu, X.: Credit default risk prediction model for unbalanced data. Journal of Hubei University of Technology 41(2), 139–144 (2026)
11. Sun, W.: Research on credit default prediction based on distributed machine learning. Master's thesis, Beijing University of Technology (2023)
12. Li, Q., Peng, F., Zhang, X.: Research on loan default probability prediction based on supervised machine learning models. Financial Development Review (6), 1–11 (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.

