Available Online: 24 June 2026 DOI: 10.54254/2977-5701/2026.34691

Research on loan risk assessment based on random forest

Shuo Wang

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China

15688702999@163.com

Abstract. Against the backdrop of the current macroeconomic environment, financial institutions face mounting pressure to strengthen credit risk management throughout the lending process. This study introduces machine learning algorithms to establish a loan risk assessment system based on the Random Forest model. Based on the '*Credit Risk Dataset*' published on the Kaggle platform, this study employed the Bootstrapping method to generate a subsample, and applied an ensemble of 100 trained decision trees to vote on the test dataset. Experimental results show that the model achieves an overall accuracy of 93.38% and an AUC as high as 0.9345, significantly outperforming logistic regression and single decision tree models. Furthermore, its high F1-score demonstrates that it has found a relatively ideal balance between "accurately identifying default applications" and "maximizing the interception of potential risks" which helps banks and other lending platforms maximize their profits while ensuring their own capital security.

Keywords: random forest, loan risk assessment, financial leverage, bootstrapping, F1 score

1. Introduction

The evolution of the market economy has catalyzed the diversification and expansion of consumer finance. To gain a competitive advantage in an increasingly intense market environment, banks and various financial institutions have continuously increased the marketing efforts of credit platforms, driving rapid growth in overall credit volume. However, behind the prosperity of the credit market lies huge risks. In recent years, non-performing loan behavior has increased significantly, and the concealment and complexity of default risk have become increasingly stronger. Therefore, it is of great significance to build an accurate and efficient loan risk assessment system to protect the asset security of banks and various financial institutions and maintain the effectiveness of financial supervision. Idczak pointed out that the essence of credit scoring is to use statistical models to transform relevant data into quantitative indicators to guide credit decisions [1]. In the current big data environment, how to extract effective features from massive, multidimensional borrower data and build a high-precision assessment model has become a focus point of academic attention.

In recent years, breakthroughs in data resources and algorithm technologies have prompted the research paradigm of bank loan risk assessment to shift from traditional discriminant analysis and logistic regression to the fields of machine learning and deep learning [2, 3]. Luo conducted an early study on the risk assessment of personal housing loans using the Fuzzy Analytic Hierarchy Process (AHP) [4]. By constructing five criterion levels and 19 specific features, and introducing a fuzzy consensus matrix for pairwise comparisons, the study determined the relative importance weights of each criterion level and feature indicator. Concurrently, a

personal credit scoring table was established, realizing the transformation of traditional qualitative subjective judgment into quantitative risk assessment standards [4]. Fan and Zhou, on the other hand, introduced the Logit model based on a combination of the Analytic Hierarchy Process and the Delphi method, comprehensively assessing the specific probability of default for each loan customer [5]. Lu et al. constructed a decision tree for risk assessment of small and micro enterprises lacking credit loan records. Empirical results showed that the accuracy of the pruned decision tree on the validation set could be improved to 96% [6]. Bu compared the performance of logistic regression, XGBoost, LightGBM, and CatBoost models in personal loan risk assessment based on the Lending Club dataset, and concluded that ensemble learning algorithms have excellent performance in processing nonlinear financial data [7]. Lu used the Multi-Layer Perceptron **(**MLP) neural network model to assess the risk of personal loans of small and medium-sized banks, and concluded that the MLP model performed well in identifying high-risk default customers with an accuracy of over 95% [8]. Wang Yifei and Song used Python data visualization technology and introduced the K-Means clustering algorithm to accurately segment borrowers. This not only verified the core role of features such as the dti ratio in loan risk assessment, but also laid a solid and easy-to-operate technical foundation for the subsequent introduction of more complex prediction algorithms through standardized data processing procedures [9]. Although Lu used random forests to assess the risk of microloans and verified its effectiveness [10], her selection of indicators was mainly limited to the attributes of the loan contract itself (such as the number of loan terms, the amount of bad debt, etc.), and she focused on using post-loan data for risk warning, paying less attention to the borrower's own occupational stability and financial leverage. Secondly, her study can only be used to assess customers with credit loan records. For customers without records, her model will fail due to the lack of characteristic indicator values. In contrast, this study aims to construct an assessment indicator system for loan customers themselves, focusing on using random forests to make predictions based on the basic economic characteristics of borrowers. It is applicable to a wider range of individuals, and the assessment of loan default risk is more forward-looking, providing a more scientific basis for decision-making for banks and other lending platforms.

2. Method

2.1. Data source

The dataset used in this study comes from the '*Credit Risk Dataset*' released on the Kaggle platform, which contains 32,581 real personal loan records from a financial institution. This study uses 11 borrower characteristics as independent variables for training the decision trees in a random forest: person_age (borrower's age), person_income (borrower's annual income), person_home_ownership (borrower's home ownership), person_emp_length (borrower's employment years), loan_intent (loan intention) , loan_grade (borrower's credit rating), loan_amnt (loan amount), loan_int_rate (loan interest rate), loan_percent_income (borrower's income-to-debt ratio), cb_person_default_on_file (borrower's default record), and cb_person_cred_hist_length (borrower's credit history duration). Loan_status (loan transaction status) is used as the dependent variable for training each decision tree (1 represents default, 0 represents on-time repayment).

2.2. Data preprocessing

2.2.1. Filling in missing values

During the decision tree training process, missing values were observed in two predictor variables: 'person_emp_length' and 'loan_int_rate'. To preserve distributional integrity while avoiding bias from arbitrary

imputation, missing entries for each variable were replaced with its respective median—computed from the non-missing observations ordered ascendingly.

2.2.2. Outlier removal

In processing the data, this study removed loan records corresponding to outliers (e.g., borrower's employment years: 123). The formal criteria applied to identify and exclude outliers are specified in Table 1.

Table 1. Outlier removal criteria

Feature Filtering	Description
Feature Filtering 1	The borrower's age (numerical value) must not exceed 100.
Feature Filtering 2	The borrower's years of employment (numerical value) must not exceed 60.

2.2.3. Encoding assignment

For the non-quantitative categorical features in the dataset—namely, person_home_ownership, loan_intent, loan_grade, and cb_person_default_on_file—this study applies one-hot encoding. As a standard preprocessing technique in machine learning, one-hot encoding transforms nominal categorical variables into a set of binary indicator variables (dummy variables), thereby enabling compatibility with numerical algorithms such as random forest, which require strictly numeric input representations.

In addition, converting categorical variables into binary form improves the performance and reliability of the random forest model. Because Random Forest builds decision trees based on feature splits, binary dummy variables enable the algorithm to evaluate each category independently and capture complex, nonlinear associations between borrower characteristics and default risk. To further maintain numerical stability and avoid perfect multicollinearity among the generated dummy variables, one category from each feature is intentionally omitted. This "dropfirst" treatment eliminates redundant information and prevents the dummy variable trap during model estimation. The detailed coding of this study is shown in Tables 2, 3, 4, and 5.

Table 2. One-hot encoding for the feature person_home_ownership

person_home_ownership	coding
MORTGAGE	(benchmark)
OWN	[0, 1, 0]
RENT	[0, 0, 1]
OTHER	[1, 0, 0]

Table 3. One-hot encoding for the feature loan_intent

loan_intent	coding
DEBTCONSOLIDATION	[0, 0, 0, 0, 0] (benchmark)
EDUCATION	[1, 0, 0, 0, 0]
HOMEIMPROVEMENT	[0, 1, 0, 0, 0]
MEDICAL	[0, 0, 1, 0, 0]
PERSONAL	[0, 0, 0, 1, 0]
VENTURE	[0, 0, 0, 0, 1]

Table 4. One-hot encoding for the feature loan_grade

loan_grade	coding
A	[0, 0, 0, 0, 0, 0] (benchmark)
B	[1, 0, 0, 0, 0, 0]
C	[0, 1, 0, 0, 0, 0]
D	[0, 0, 1, 0, 0, 0]
E	[0, 0, 0, 1, 0, 0]
F	[0, 0, 0, 0, 1, 0]
G	[0, 0, 0, 0, 0, 1]

Table 5. One-hot encoding for the feature cb_person_default_on_file

cb_person_default_on_file	coding
N	[0]
Y	[1]

2.3. Sample set partitioning

The original dataset selected for this study yielded 32,574 valid samples. Eighty per cent of these valid samples were randomly selected to form the training set for building the decision tree, whilst the remaining 20 per cent were used as the test set. The specific allocation of the sample size is shown in Table 6:

Table 6. Training and testing set sample size allocation

No.	Sample	Size
1	Original sample size	32,581
2	Analytic sample size	32,574
3	Training set size	26,059
4	Testing set size	6,515

2.4. Model construction

2.4.1. Bootstrappingself-sampling

The random forest model employed in this study comprises 100 decision trees. To ensure both ensemble diversity and consistent training set size, bootstrap sampling was applied during the construction of each tree: specifically, 26,059 observations were drawn with replacement from the full set of 26,059 valid training samples. This procedure guarantees that each tree is trained on a distinct, stochastically varied subset

2.4.2. Decision tree integration

After data preprocessing, a total of $n = 22$ features were obtained for modeling, as shown in Table 7. During the splitting process at each node of each decision tree, 5 features were randomly selected ($\sqrt{22} \approx 4.69$) (The square root of the total number of features is commonly chosen as the number of features for each decision tree) to constitute a candidate subset, and the optimal segmentation feature and its corresponding threshold are determined by comparing the Gini coefficient.

Table 7. Independent variable and their encoded dimensions

Original independent variable	Variable type	Number of variables after One-Hot encoding
person_age	Numerical	1
person_income	Numerical	1
person_home_ownership	Category	3
person_emp_length	Numerical	1
loan_intent	Category	5
loan_grade	Category	6
loan_amnt	Numerical	1
loan_int_rate	Numerical	1
loan_percent_income	Numerical	1
cb_person_default_on_file	Category	1
cb_person_cred_hist_length	Numerical	1
total		21

2.4.3. Random forest voting

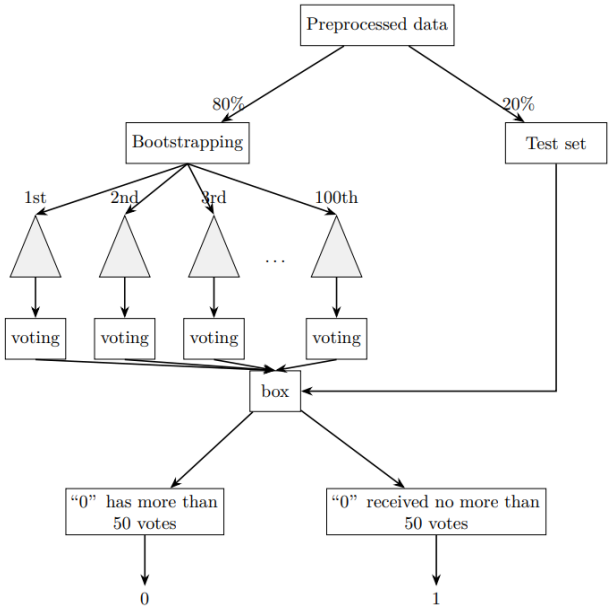

Figure 1. Flowchart of the random forest model

Figure 1 illustrates the detailed workflow of the random forest algorithm adopted in this study. The 100 generated decision trees are used to vote on the test set. For each test sample, if more than 50 decision trees in the random forest predict that loan_status is 0, the output is 0 (normal); otherwise, the output is 1 (default).

3. Result

3.1. Overall model evaluation

3.1.1. Confusion matrix of the test set

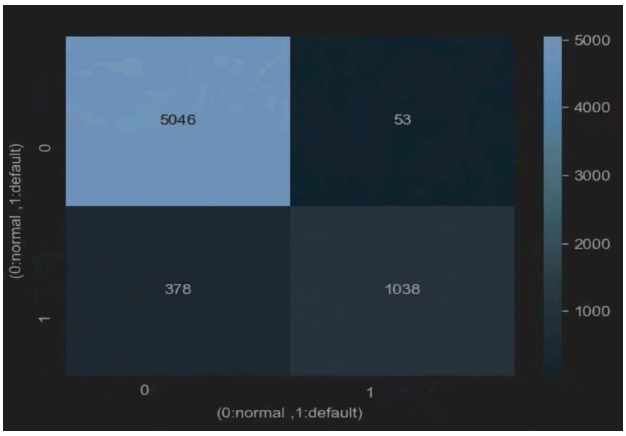

Figure 2. Heatmap results of the test set

Figure 2 shows a heatmap of the test sample classification results, with the color intensity visually reflecting the model's judgment on 6,515 samples. The results demonstrate that Random Forest exhibits excellent robustness in identifying normal samples. In the 6,515 test samples, the model accurately identified 5,046 creditworthy loan customers, with a false positive rate of only 1.0394%. This indicates that the model has a significant advantage in protecting the bank's existing high-quality customers and preventing business loss. Simultaneously, the model performs well in capturing potential defaulting loans, achieving a recall rate of 73%, providing a reliable decision-making basis for banks and other lending platforms to identify high-risk loan applications.

3.1.2. Model accuracy

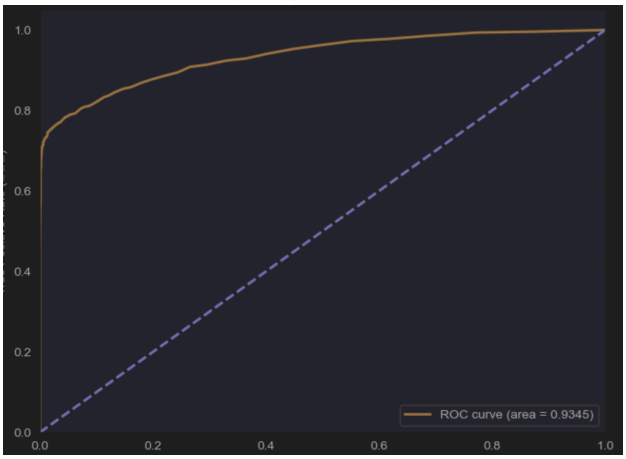

Figure 3. Receiver Operating Characteristic (ROC) curve of the Random Forest model

Experimental results show that, in the field of loan risk assessment, Random Forest possesses exceptional risk stratification capabilities, with an Area Under Curve (AUC) as high as 0.9375 presented in Figure 3.

3.2. Analysis of the contribution ratio of characteristic indicators to decision-making

To further explore the underlying logic of the model's decision-making, this study measured the contribution of 22 feature variables involved in the construction of the decision tree by calculating the average decrease in Gini impurity at each node split. Figure 4 shows the top 10 feature variables ranked by contribution.

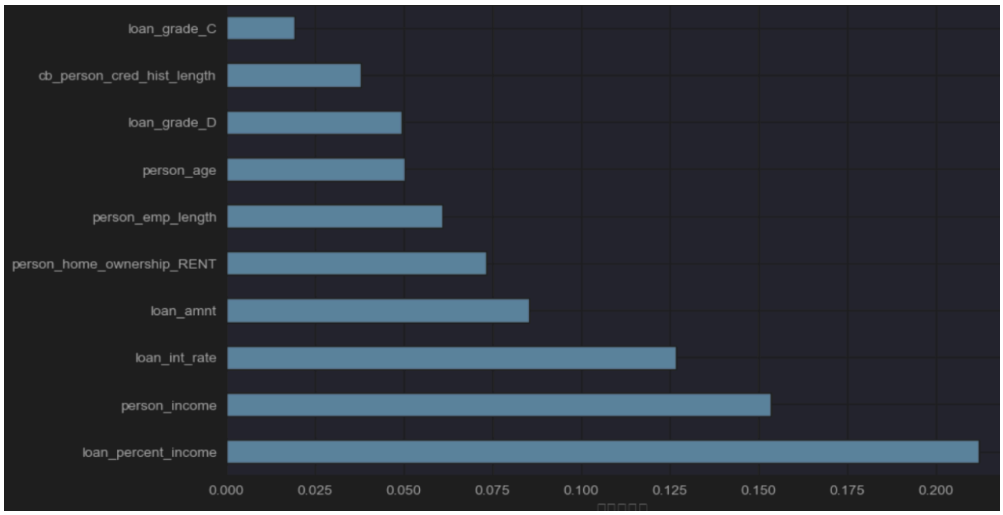

Figure 4. Model feature variable importance ranking evaluation

As can be seen from Figure 4, loan_percent_income (borrower's debt-to-income ratio), person_income (borrower's annual income) and loan_int_rate (loan interest rate) are the core risk drivers. In contrast, person_age (borrower's age) and cb_person_cred_hist_length (length of the borrower's credit history) have limited value as references for credit risk assessment. For the categorical feature variable 'loan_grade' (borrower's credit rating), 'loan_grade_D' and 'loan_grade_C' rank highest in importance. This indicates that the Random Forest model is capable of accurately capturing the risk characteristics within specific credit rating ranges, providing empirical evidence for banks and other lending platforms to implement precise risk control measures.

4. Discussion

The data obtained from the Kaggle platform were fed into logistic regression, decision tree, and LightGBM models, and the results of defaulting customer identification are shown in Table 8.

Table 8. Performance comparison of different classification model

Model	accuracy	Accuracy	Recall rate	F1 score	AUC value
Logistic Regression	85.03%	75.37%	46.26%	0.5733	0.8626
Single decision tree	89.47%	74.80%	77.75%	0.7625	0.8524
LightGBM	93.66%	97.00%	73.09%	0.8337	0.9501
Random Forest	93.38%	95.14%	73.31%	0.8281	0.9345

4.1. Overall performance evaluation of the model

As shown in Table 8, the overall accuracy (93.66%) and AUC (0.9345) of the random forest model are significantly better than those of the logistic regression model and the single decision tree model, demonstrating its obvious advantages in handling high-dimensional and non-linear data such as loan data.

4.2. Model for identifying defaulting customers

Experimental results show that the Random Forest model achieved a recall rate of 73.31%, ranking first among all compared models. This demonstrates its strong ability to intercept potential credit risks, ensuring that the bank can effectively avoid the vast majority of bad debt risks and, to a certain extent, safeguard capital security. In terms of accuracy, the Random Forest model improved by approximately 25% compared to logistic regression and the single decision tree, significantly reducing the 'false rejection rate' and minimizing profit losses resulting from missed opportunities for high-quality loan transactions.

4.3. F1-score overall recognition performance

The F1-score, as the harmonic mean of precision and recall, is widely used for evaluating imbalanced samples [11]. It is an authoritative indicator for measuring the model's comprehensive ability to identify default samples. Random Forest has an F1-score of 0.8281, which is much higher than logistic regression and decision tree models. It is on the same level as the LightGBM model, which helps banks and other lending platforms find a relatively good balance between accurately identifying defaulted loan applications and maximizing the interception of potential risks.

5. Conclusion

This study utilized the Random Forest model to validate its outstanding performance in loan risk assessment. Through empirical analysis of 32,581 loan data obtained from the Kaggle platform, the model's AUC is as high as 93.45%, and the overall accuracy exceeds 93%. The recall rate for customers with normal repayments reaches 99%, which indicates that the model can help banks and other lending platforms greatly reduce the loss of high-quality customers and potential interest income losses caused by "false rejections". For defaulting customers, the model achieves a recall rate of 73% and an overall accuracy of 95%. This balanced performance enables the construction of a high-confidence risk interception framework—one that effectively identifies the majority of high-risk applicants while preserving lending volume, thereby aligning with banks' prudential risk appetite. The empirical findings further underscore the critical influence of three key determinants in the risk assessment system: borrower financial leverage, annual income, and loan interest rate. Nevertheless, the model exhibits several limitations that warrant acknowledgment. Wang et al. pointed out that when dealing with imbalanced data, standard algorithms often produce problems such as learning bias and distortion of evaluation indicators [12]. Consequently, this study is use of a default classification threshold of 0.5 during the decision tree voting stage, without fully considering the specific characteristics of the samples in the credit data, has certain limitations. Future research on Random Forests could optimize the model's voting mechanism by incorporating methods such as Bayesian decision-making, thereby providing banks and other lending platforms with a more precise and comprehensive evaluation system.

References

[1] Idczak, P. A. (2019). Remarks on statistical measures for assessing quality of scoring models. *Acta Universitatis Lodziensis. Folia Oeconomica, 4*(343), 21–38. https: //doi.org/10.18778/0208-6018.343.02

[2] Guzmán-Castillo, S., Garizabalo-Davila, C., Luis, A. G., & Rodríguez, J. (2023). Credit risk scoring model based on the discriminant analysis technique. *Procedia Computer Science, 220*, 928–933. https: //doi.org/10.1016/j.procs.2023.03.127

[3] Barasa, N. I., Wanyonyi, W. S., & Kololi, M. (2025). Application of logistic regression in enhancing digital credit risk management in commercial banks. *Asian Journal of Probability and Statistics, 27*(2), 13–26. https: //doi.org/10.9734/AJPAS/2025/V27I2710

[4] Luo, Z. (2013). On assessment of individual housing loans risk based on FAHP method. *Journal of Hunan University of Technology (Social Science Edition), 18*(3), 30–34.

[5] Fan, S. G., & Zhou, S. Y. (2019). Construction of risk assessment model for rural microloans based on Delphi and AHP methods. *China Market Marketing*, (36), 47–50. https: //doi.org/10.13939/j.cnki.zgsc.2019.36.047

[6] Lu, L. J., Lu, C., & Dong, H. Y. (2022). Data analysis of credit risk for small, medium and micro enterprises based on decision tree. *Popular Science & Technology, 24*(9), 5–9.

[7] Bu, H. H. (2025). Research on personal loan credit risk assessment model based on machine learning. *Computer Knowledge and Technology, 21*(10), 25–28. https: //doi.org/10.14004/j.cnki.ckt.2025.0504

[8] Lu, H. L. (2025). Research on credit risk assessment of personal loans in small and medium-sized banks based on MLP neural network. *Computer Knowledge and Technology, 21*(21), 19–21. https: //doi.org/10.14004/j.cnki.ckt.2025.1040

[9] Wang, Y. F., & Song, Y. R. (2023). Visual analysis of financial data based on data mining and K-Means model. *Computer Era*, (7), 96–99, 104. https: //doi.org/10.16644/j.cnki.cn33-1094/tp.2023.07.022

[10] Lu, H. L. (2025). Research on microloan credit risk assessment empowered by random forest. *Fujian Computer, 41*(11), 26–31. https: //doi.org/10.16707/j.cnki.fjpc.2025.11.005

[11] Chen, Z. L., Charles, E., & Balakrishnan, N. (2014). Optimal thresholding of classifiers to maximize F1 measure. In *Machine learning and knowledge discovery in databases: European Conference, ECML PKDD 2014, Proceedings*(Part III, pp. 225–239). Springer. https: //doi.org/10.1007/978-3-662-44851-9_15

[12] Wang, D., Wu, T., Yu, Z. H., Li, G. C., & Ma, Z. Q. (2025). Cost-sensitive convolutional neural network classification method based on federated learning. *Journal of Xi'an University of Science and Technology, 45*(3), 591–606. https: //doi.org/10.13800/j.cnki.xakjdxxb.2025.0316