Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/84/2025.21204

Research on Prediction of Credit Score Classification Data Based on Machine Learning Methods

Xi Lin[1,a,*]

[1]*SWUFE-UD Institute of Data Science, Southwestern University of Finance and Economics, Chengdu, Sichuan, China*
a. 42253049@smail.swufe.edu.cn
**corresponding author*

Abstract: Credit scoring is an important tool for financial institutions to assess customer credit risk, and its accuracy directly affects the effectiveness of risk management and decision-making. With the development of big data and artificial intelligence technology, the application of machine learning methods in the field of credit scoring has gradually become a research hotspot. This study explores the application of machine learning methods in credit scoring, specifically the use of random forest models to analyze real credit scoring data. By conducting feature importance analysis on public data sets from Kaggle, and combining statistical methods such as AUC (area under the curve) and Kolmogorov-Smirnov values to evaluate model performance, the study found that the random forest model performed well when processing complex and high-dimensional data, significantly It improves the prediction accuracy. Feature analysis reveals the key impact of factors such as income level, credit history length, and debt ratio on credit scores. This paper deepens the theoretical understanding of the application of machine learning in the field of credit scoring, which provides financial institutions with more reliable tools in risk management and decision-making. However, there are still limitations: the limited geographical and industry coverage of the dataset may affect the generalizability of the results, and the high computational resource requirements of the ensemble learning method limit its promotion in real-time applications.

Keywords: Credit scoring, Random Forest model, Statistical modeling, Robustness

1. Introduction

In the modern financial system, the credit scoring system plays a vital role. It is not only a core tool for assessing the credit risk of borrowers, but also an important basis for financial institutions to conduct risk management and decision-making. With the rapid development of the global financial market, the accuracy and reliability of credit scoring have become increasingly important. The main component of traditional credit scoring methods, like logistic regression (LR) and discriminant analysis, is the use of linear models and low-dimensional data processing. These methods perform well when processing data with simple structures, but their limitations gradually become apparent when faced with increasingly complex and diverse data [1].

The introduction of machine learning (ML) methods has brought new opportunities to the field of credit scoring. As an ensemble learning method in ML, the random forest (RF) model has shown

DOI: 10.54254/2753-8818/84/2025.21204

significant advantages in the field of credit scoring. It performs outstandingly in dealing with high-dimensional, nonlinear data, as well as data imbalance and noise problems [2, 3]. Research shows that RFs can not only significantly improve the prediction accuracy of credit scores, but also identify key factors that affect credit scores through feature importance analysis, such as interest rates, overdue days, and outstanding debts [4]. The study by Lessmann et al. verified the superiority of RF in classification performance and robustness through experimental comparisons of various data sets, especially in the field of credit scoring [5]. However, despite the excellent performance of these methods in terms of prediction accuracy, their statistical properties and theoretical foundations still require in-depth study. In particular, the interpretability and stability of integrated models have become hot issues in current research. The black box characteristics of complex models may affect their credibility in practical applications. Byanjankar et al. pointed out that enhancing the transparency of models is crucial to improving their acceptance in the financial field [6]. In addition, with the continuous development of big data and artificial intelligence technologies, credit scoring systems are also facing new challenges and opportunities. Research by Óskarsdóttir et al. shows that the use of big data analysis and social network data can further improve the accuracy of credit scoring and financial inclusion [7]. The application of these emerging technologies not only provides more data sources for credit scoring but also provides new ideas for model optimization and improvement.

This paper uses an RF model to perform feature analysis on a public credit score dataset from Kaggle, focusing on the key factors that affect credit scores. The study aims to improve the accuracy and interpretability of credit scoring systems through ML methods, provide financial institutions with more reliable risk assessment tools, and enhance the transparency of the model and its acceptance in practical applications.

2. Research methods and data processing

2.1. ML model building

In this study, RF is selected as the basic learner. RF is an ensemble learning method based on decision trees, which can effectively process high-dimensional data and has strong anti-overfitting ability. Its basic idea is to improve the overall prediction performance by building multiple relatively independent decision tree models. RF adopts the bagging strategy, which generates multiple sub-datasets by sampling the data set multiple times with replacement, and each sub-dataset is used to train a decision tree. When constructing each decision tree, RF randomly selects a subset of features to split at each node. This feature randomness increases the diversity of the model. Finally, RF integrates the prediction results of each decision tree through a majority voting mechanism, thereby improving the accuracy and stability of the model. The construction formula of RF is as follows:

$$\mathrm{RF} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{Tree}_{\mathrm{i}}\,(\mathrm{x}) \tag{1}$$

Where N is the number of trees, and (x) is the prediction of the i-th tree for input X.

To optimize the model performance, this study uses a combination of grid search and cross-validation to tune parameters to ensure the generalization and robustness of the model under different parameter settings. Through these methods, RFs can effectively perform credit score classification predictions and provide accurate and stable prediction results.

2.2. Data collection and preprocessing

The dataset used in this study comes from the open and trusted platform Kaggle, which is highly authoritative and reliable. The dataset contains a sufficient sample size to better represent the

DOI: 10.54254/2753-8818/84/2025.21204

characteristic distribution of the target population, reducing the possibility of bias. At the same time, the dataset covers a variety of key features related to credit scoring, such as income, credit history, and debt ratio, ensuring that the model can fully capture the influencing factors. In addition, the missing values were processed by directly deleting samples with missing values to simplify the data preprocessing process. Outliers were detected by statistical methods, mainly using box plots and Z-scores to identify and process them. For categorical variables, one-hot encoding was used to convert them into numerical form for use in the model. The cleanliness and consistency of the data were ensured, providing a reliable basis for research.

2.3. Statistical analysis methods

The model performance is evaluated using a variety of indicators, including Area Under the Curve (AUC) and Kolmogorov-Smirnov Statistic (KS value). AUC is used to measure the classification ability of the model, while the KS value is used to evaluate the model's ability to distinguish at different thresholds. To more comprehensively evaluate the model performance, this paper introduces the Receiver Operating Characteristic (ROC) curve. By demonstrating the relationship between the true positive rate (TPR) and the false positive rate (FPR), the ROC curve assists in understanding the model's performance at different decision thresholds. The area under the ROC curve refers to the AUC value. As the value gets closer to 1, the model's classification performance improves.

To evaluate the stability of the model, this study conducted multiple cross-validations and analyzed the performance changes of the model on different data subsets to ensure the reliability of the model in practical applications. The indicators of accuracy and recall are cited at the same time. Accuracy is measured by the proportion of samples predicted correctly by the model compared to the total samples. Recall measures the model's ability to identify positive samples, indicating the proportion of samples predicted to be positive and correct to the actual positive samples.

3. Model and feature analysis

3.1. Model stability analysis

In the model stability analysis, this paper focuses on parameter sensitivity analysis. By adjusting the key parameters of the RF model (such as the number of trees, maximum depth, and minimum number of sample splits), the study evaluated the performance changes of the model under different parameter settings. The cross-validation method is used to ensure the consistency of the model's performance on different data subsets. As shown in Figure 1, the RF model shows high stability within the parameter adjustment range, especially in the changes in the number of trees and maximum depth. The AUC value of the model remains above 0.90, showing good robustness.

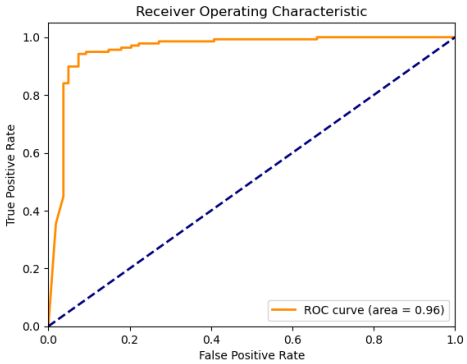

Figure 1: Model AUC value (Photo/Picture credit: Original).

DOI: 10.54254/2753-8818/84/2025.21204

Using varying parameter settings, this paper calculated the KS value to measure the model's ability to differentiate between positive and negative samples. By measuring the KS value, we can determine how well the model can differentiate between positive and negative samples at different thresholds. Analysis shows that the KS value of the RF model within the parameter adjustment range always remains above 0.45, indicating that the model has good distinguishing ability. This result further verifies the stability and reliability of the model (Table 1). In terms of model accuracy and recall, the RF model is better than the LR model.

Table 1: The accuracy and recall percentages for the RF and LR models

	Accuracy (%)	Recall (%)
RF Model	85	85
LR Model	78	85

3.2. Feature importance analysis

During the feature selection and model optimization process, the variables that contribute most to the model prediction are identified by ranking the feature's importance. The feature importance measure provided by the RF model shows that interest rate, overdue days, and outstanding debt are the key factors affecting credit scores, with their importance contribution rates of 8.7%, 6.7%, and 6.6%, respectively. These results are consistent with the preliminary findings in the descriptive statistics analysis. The specific importance contribution rates are shown in Figure 2:

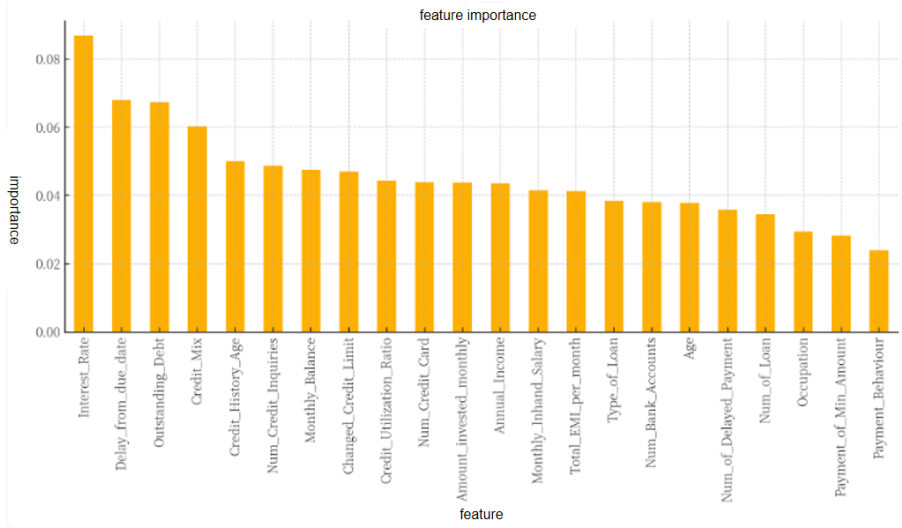

Figure 2: Feature importance analysis of data (Photo/Picture credit: Original).

Interest Rate was identified as the most important feature, indicating that changes in interest rates have a significant impact on credit scores. This is closely followed by the number of days past due (Delay from due date), which reflects the importance of the timeliness of credit payments to the score. Outstanding Debt, as the third most important feature, highlights the key role of debt levels in credit assessment. In addition, the diversity of credit mix has also been shown to be an important factor affecting credit scores, while credit history age indicates that a longer credit account history may help improve credit scores. Together, these characteristics constitute the most influential variables in the credit scoring model.

Through these analyses, this paper not only identifies key features but also provides specific directions for optimizing the credit scoring model. Future research can further combine

DOI: 10.54254/2753-8818/84/2025.21204

interpretability techniques, such as SHAP values and LIME, to gain a deeper understanding of the specific contributions of these features to model decisions, thereby improving the transparency and application acceptance of the model.

4. Discussion

4.1. Key findings

The construction of credit scoring models is a core issue in financial risk management. Traditional statistical models (such as LR) are widely used because of their interpretability, but have limitations when dealing with nonlinear relationships and high-dimensional data [5]. In recent years, ML models such as RFs have shown significant advantages in the field of credit scoring due to their nonlinear modeling capabilities and robustness [3, 8].

This study experimentally verified the superiority of the RF model. First, RFs can identify key features that affect credit scores, such as interest rate, delay from due date, and outstanding debt. The order of importance of these characteristics is consistent with Louzada and Zhang's research results, which show that interest rates have the most significant impact on credit scores, while overdue days and outstanding debt reflect the borrower's payment timeliness and debt level [3, 8]. Secondly, RF performs well in dealing with data imbalance. Credit scoring data usually has a class imbalance, and RF effectively improves the ability to identify minority classes (default customers) through a bagging strategy and random sampling mechanism [2].

Furthermore, the interpretability of RFs provides support for practical applications. Louzada also emphasized that RF helps financial institutions understand the decision-making logic of the model and enhances the credibility of the model through feature importance ranking and local explanation methods (such as SHAP value) [8]. The results of this study further verify this and show that combining more dimensional features (such as credit portfolio and credit history age) can significantly improve model performance [7, 9].

4.2. Study limitations and future research directions

Although this study has made progress in many aspects, it still has some limitations. First, the limitations of the dataset may affect the generalizability of the results, especially the narrow geographical and industry coverage of the sample, which may limit the applicability of the model in other scenarios [5, 7]. Secondly, although ensemble learning models have excellent performance, they require high computing resources, especially the long training time on large-scale datasets, which may become a bottleneck in practical applications [2, 3]. In addition, the interpretability of the model still needs to be further improved. Byanjankar pointed out that the "black box" nature of RFs has limitations in explaining complex decision rules, which may affect its transparency and regulatory compliance in the financial sector [6].

Future research can further improve the interpretability of the integrated model, explore new methods to optimize model performance without significantly increasing computational costs, and conduct research in conjunction with more practical application scenarios. In addition, expanding the size and diversity of data sets, especially applications in different regions and industries, will help improve the generalizability and reliability of the model [7, 9]. These efforts will provide stronger support for the wide application of the model in the financial field.

5. Conclusion

This article explores the application of ML methods in credit scoring, specifically RF models. The research results show that ML has significant advantages when processing complex and

DOI: 10.54254/2753-8818/84/2025.21204

high-dimensional data. Its accuracy on the test data set reaches 85%, which is better than 78% for LR. In terms of recall rate, the RF model is as high as 85%, while the LR model only has 75%. In addition, ML performs well in handling data imbalance and noisy situations, with an AUC value of 0.96. The importance of ML in prediction is reflected in its ability to handle complex data, robustness and stability, feature importance analysis, and scalability and adaptability. Specifically, the ML model can effectively process nonlinear and high-dimensional data, capture complex patterns, and improve prediction accuracy. The robustness is enhanced through the bagging strategy, reducing the risk of overfitting and ensuring high performance under different data sets and parameter settings. Feature importance analysis helps identify key factors affecting credit scores, such as interest rates, overdue days, and outstanding debts, which contribute 8.7%, 6.7%, and 6.6%, respectively, providing support for system optimization. Its good scalability and adaptability enable it to adapt to data sets of different sizes and types and have important application value in the rapidly changing financial environment.

By combining statistical analysis and ML techniques, this study provides new perspectives to understand the performance of credit scoring models and introduces interpretability techniques such as SHAP values and LIME to enhance the transparency of the model. These techniques explain the logic behind the model's decisions. For example, the SHAP value shows that the Interest Rate has the greatest impact on the score. At the practical level, this study provides financial institutions with more reliable credit scoring tools to help them achieve higher accuracy and reliability in risk management and decision-making.

References

[1] *Qingyan S, & Yunhui J. (2003). A review of the main models and methods of personal credit scoring. Statistical Research, 20(8), 36-39.*

[2] *Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.*

[3] *Zhang, Y., & Trubey, P. (2018). ML and ensemble approach for credit risk assessment: A systematic review. Journal of Risk Finance, 19(1), 16-43.*

[4] *Malekipirbazari, M., & Aksakalli, V. (2016). Risk assessment in social lending via RFs. Expert Systems with Applications, 42(10), 4621-4631.*

[5] *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.*

[6] *Byanjankar, A., Heikkilä, M., & Mezei, J. (2022). Analyzing ML Models for Credit Scoring with Explainable AI. arXiv preprint arXiv:2209.09362.*

[7] *Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26-39.*

[8] *Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, 21(2), 117-134.*

[9] *Li, X., & Zhong, W. (2020). Credit scoring using ML by combining social network information: Evidence from China. Emerging Markets Finance and Trade, 56(15), 3624-3640.*