Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/97/20241348

Credit Default Prediction Based on Random Forest

Siyu Xiang[1,3,*], Haowen Yan[2]

[1]Chengdu Experimental Foreign Languages School, Sichuan, 611134, China
[2]Taiyuan Experimental Middle School, Taiyuan, 030031, China

[3]xiangsiyu@ldy.edu.rs
*corresponding author

Abstract. To help the lender to make reasonable prediction judgments on the lender in advance, to get the most appropriate way of lending amount, as well as facing the risk of a reasonable response program and the ability to cope with the need to make predictions in advance before lending for personal credit risk. This paper aims to predict credit default based on random forest, logistic regression and decision tree algorithms, by comparing and analyzing the advantages and disadvantages of these algorithms, this paper finally chooses the random forest algorithm. This paper concludes that in predicting the risk of credit default, the three characteristics of Credit amount, Duration and Job have the most significant influence in predicting the risk value of the borrower, Credit amount is the most important factor that affects the risk value, Duration is also a key factor, a more relaxed repayment period will reduce the pressure of repayment, and thus reduce the risk of default. Job, different occupations, different incomes and different income stability will lead to different repayment abilities of each borrower, according to the repayment ability and the loan amount of the comparison, you can initially arrive at the corresponding risk value and the development of the corresponding risk control strategy.

Keywords: Credit default, prediction, machine learning.

1. Introduction

Credit default is a situation in which the borrower is unable to make principal or interest payments to the lender by the time specified in the contract, or is unable to fulfil other contractual commitments of the loan. This usually occurs when the borrower fails to repay the loan on time, violates the terms of the contract, or has other circumstances that seriously affect his ability to repay the debt. Nowadays, more and more people will be burdened with car loans or home loans, for such a form of public acceptance has been very high [1,2]. However, under such a trend, the instances of loan defaults have also become more and more frequent. This may lead to an insufficient chain of funds in banks and lower interest rates on deposits. If there are a large number of credit defaults, it may lead to bank failures or the need for government funding, and the government may tighten the corresponding policies as a result, leading to difficulties in lending in the future. And the government will increase the fiscal pressure in funding banks or actively intervening in the financial market [3]. Therefore, predicting credit defaults and analyzing various aspects of the borrower, can help lenders evaluate the risk level of the borrower when lending money, and reduce the risk of lending appropriately. This also gives banks a better ability to face risks, to foresee them in advance and to block them.

DOI: 10.54254/2755-2721/97/20241348

Zhang predicts credit default risk based on adaptive feature interaction learning, which solves the problems of some existing credit default prediction models that are unable to analyze the potential relationship of data and adaptive generalization of reoccurring combinations of features, and at the same time improves the performance of credit default prediction and provides a convenient and fast interaction method [4]. Zhang et al. based on a Transformer encoder and residual network for the credit default prediction model, solved the problem of lack of effective treatment of high-dimensional sparse category features in the traditional credit default prediction model [5]. Wang Yuan et al. studied the credit risk and management of local commercial banks under the background of big data and solved the problem of how to improve the management level of credit risk and the comprehensive evaluation of credit risk on sample data using information technology [6].

This paper is dedicated to the study of advance prediction of credit default risk based on different information about the borrower such as: whether the borrower owns a personal home, the borrower's occupation, the borrower's savings, and so on. It also analyzes the loan information itself, such as the loan amount and repayment period. After analyzing the borrower's information and the loan information, the bank can make adjustments to the loan amount or the repayment period by analyzing the risk value of the loan and can use the risk value data as a reference to make a good risk control strategy in advance. To achieve the goal of risk prediction and to be as accurate as possible, this paper cites a publicly available dataset on credit risk in Germany and analyzes this data set using five different algorithms. This paper intends to compare the five algorithms and come up with the one that has the best overall performance to help in the final VaR prediction.

2. Data and methodology

The data for this paper is a publicly available data set on credit risk in Germany. The dataset consists of 1000 rows of data, each row containing 10 different features. (https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk/data_)

There are five methods chosen in this paper, which are the LR logistic regression algorithm, DT decision tree algorithm, Bagging algorithm and Ada Boost algorithm [7]. LR logistic regression algorithm is a commonly used machine learning algorithm for solving classification problems. DT decision tree algorithm, which recursively performs the feature selection to categorize the data from the training set.

Bagging is an integrated learning algorithm, usually used to reduce the variance in a noisy dataset, using substitution to select random data samples in the training set. Ada Boost is an iterative algorithm, whose core idea is to train different classifiers (weak classifiers) for the same training set, and then aggregate these weak classifiers to form a stronger final classifier (strong classifiers). RF Random Forest algorithm, is in the decision tree construction, first from the training set have put back randomly selected samples, to get a new training set (samples can be selected multiple times as well as not selected). Secondly, a subset of features is selected from the feature set and the optimal features from the subset are selected for further splitting. Finally, the decision tree is constructed using the selected samples as well as the feature set until a condition is met (e.g., the number of samples is less than a specific threshold) [8,9].

Each decision tree makes a prediction for the input samples, and the forest mean is taken to get the final prediction structure.

Four basic concepts commonly used in the evaluation metrics used in this paper include true positive examples (TP), true negative examples (TN), false positive examples (FP), and false negative examples (FN) [10]. Specifically, TP refers to the number of positive class samples correctly detected by the model; TN refers to the number of negative class samples correctly detected by the model; FP stands for the number of negative class samples incorrectly determined to be positive; and FN refers to the number of positive class samples incorrectly determined to be negative.

Precision and Accuracy are two important metrics when evaluating model performance, but they are not the same. Accuracy is the number of samples that the model predicts correctly as a proportion of the total number of samples. Accuracy, on the other hand, is the proportion of samples that are actually in

DOI: 10.54254/2755-2721/97/20241348

the positive category out of those predicted to be in the positive category by the model. These two metrics reflect the performance of the model from different perspectives, with the accuracy rate focusing on the overall correctness of the predictions, while the precision rate emphasizes the reliability of the results predicted as positive classes.

3. Analysis of results

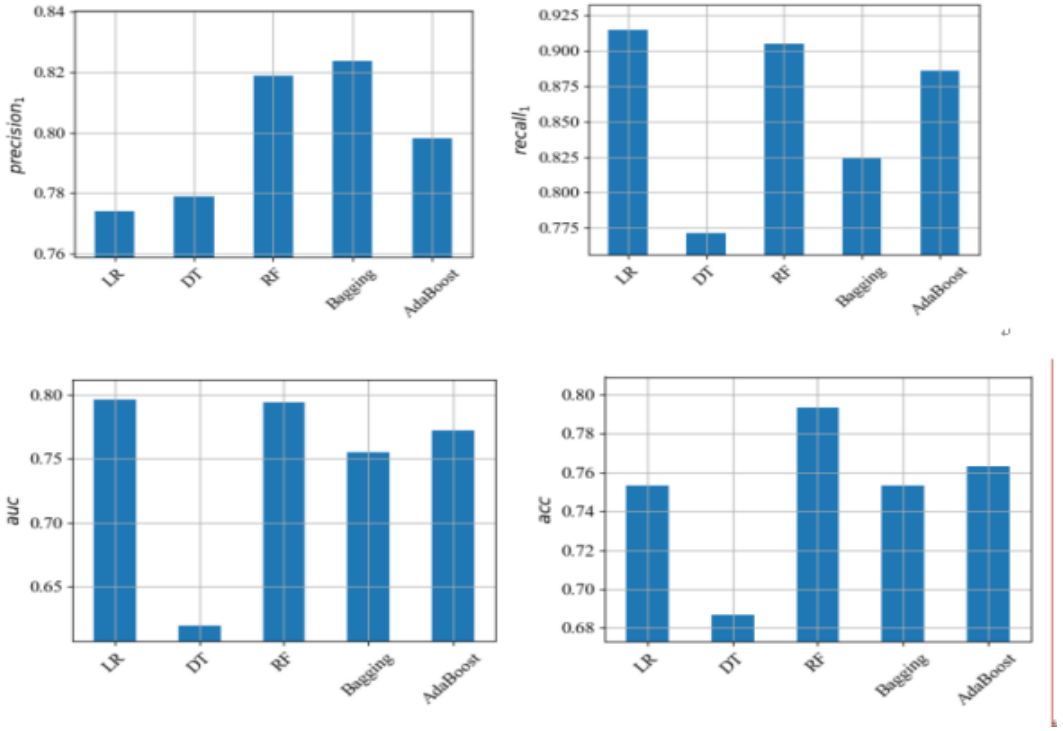

Figure 1. Model performance with different metrics. Left corner: Precision comparison analysis of each model; upper right corner: Recall comparison analysis by model; Bottom left, Area Under the Curve (AUC); Bottom right, acc comparison by model

In Figure 1 (left corner), it is seen that Bagging has a better performance, while LR has a worse performance. In the chart of recall, LR has a better performance, DT has the weakest performance, and all others are at a high level. In the AUC chart, we see that LR is the best performer, RF is almost as good as LR, and DT is the worst. In the acc graph, we see that RF is the best performer and DT is the worst performer.

Random Forest (RF) has better performance in Accuracy (ACC) and AUC compared to other base models.RF has a significant advantage in Precision and performs in the middle to high level in Recall.

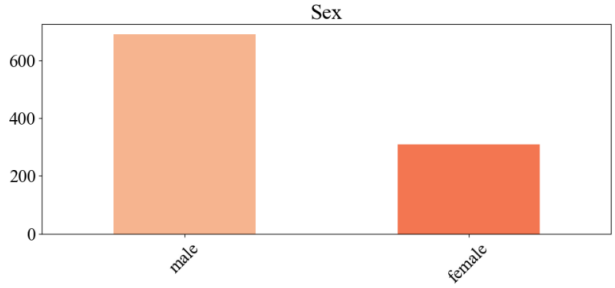

Figure 2. Visualization of the gender ratio

DOI: 10.54254/2755-2721/97/20241348

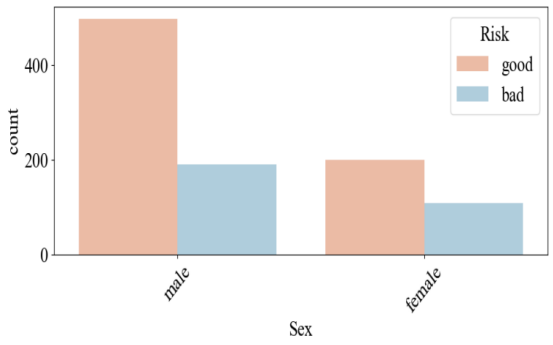

Figure 3. Risky quantity statistics

As can be seen in Figure 2 and Figure 3, the number of males far exceeds the number of females in the entire dataset. The number of males in Figure 1 is as high as 680, while the number of females is at around 300, with the number of males being more than double the number of females. Adding to this, we can see from Figure 2 that the high borrowing rate of males is accompanied by an increase in their high borrowing risk rate. This is why the risk of borrowing by gender needs to be discussed separately in the subsequent discussion, and the reasons for borrowing differ for males and females.

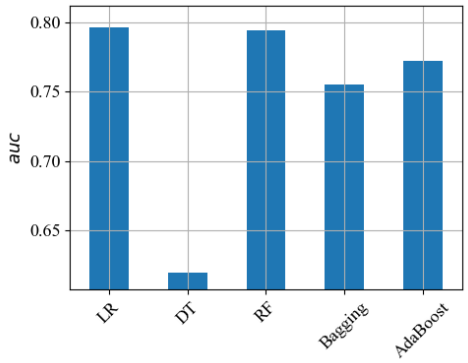

Figure 4. Number of algorithms

From Figure 4, the LR model is the best run, its AUC is close to 0.8, and the RF model looks good too, but it's still a bit worse than LR. DT, on the other hand, is so bad that the AUC looks only about 0.61. So in the final choice, we will choose either the LR or RF model, which will have better classification performance than the other models

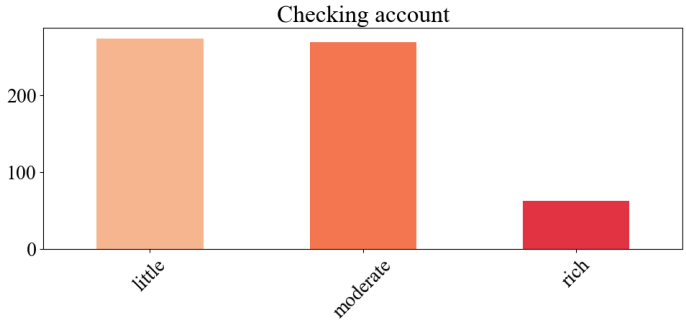

Figure 5. Comparison of existing account balances

DOI: 10.54254/2755-2721/97/20241348

Figure 5 shows that only a small percentage of people with checking account balances are wealthy and eighty percent of people just have low or moderate balances, so most users do not have high checking account balances, which can lead to high credit risk.

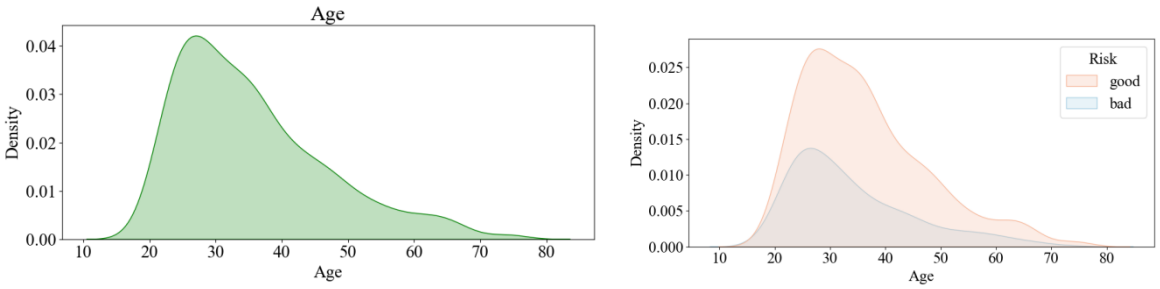

Figure 6. Age analysis

As can be seen from Figure 6, most of the borrowers are young and middle-aged around 25-45 years old. This group of people is in the working period and has a great possibility to borrow money to buy a house, a car, and to maintain their life, etc. At the same time, their risk of borrowing money increases proportionally. At the same time, their risk of borrowing also increases proportionally. On the other hand, the risk of borrowing for the younger and older age groups is extremely low. The younger age group is not in a position to borrow, while the older age group has very little need to do so.

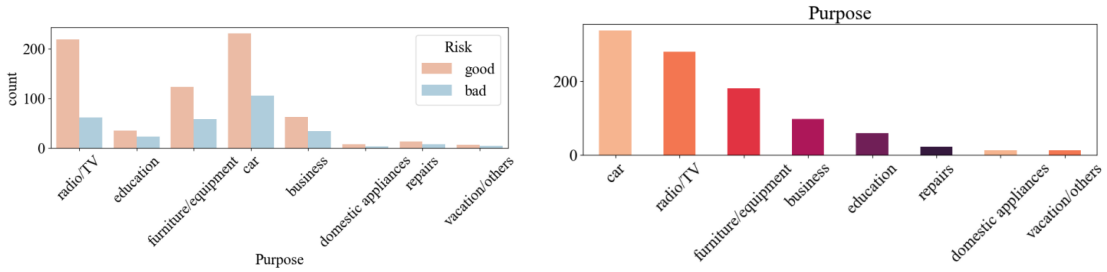

Figure 7. Impact analysis of borrowing purposes

Figure 7 shows some of the mainstream directions of popular borrowing, the chart shows that the highest borrowing rate is for auto repair, which the chart shows as high as about 320, and the lowest is for leisure and entertainment, which the chart shows as low as about 10. The items with higher borrowing rates are basically the necessities of people's lives. The increase or decrease of these borrowing risks changes as individuals have different amounts of needs for the above areas, and different standards of living may also be associated with different levels of borrowing risks.

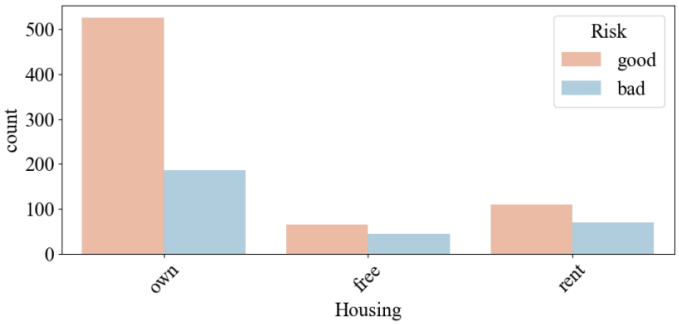

Figure 8. Housing situation

DOI: 10.54254/2755-2721/97/20241348

The data in Figure 8 relies on the housing situation of different households to analyze different levels of borrowing risk. From the figure, we can conclude that individuals who own their properties have a low risk of borrowing with high credibility. We can guess that this group of individuals have some stable jobs and incomes. On the contrary, the population of individuals with free housing has a low borrowing rate but relatively high risk.

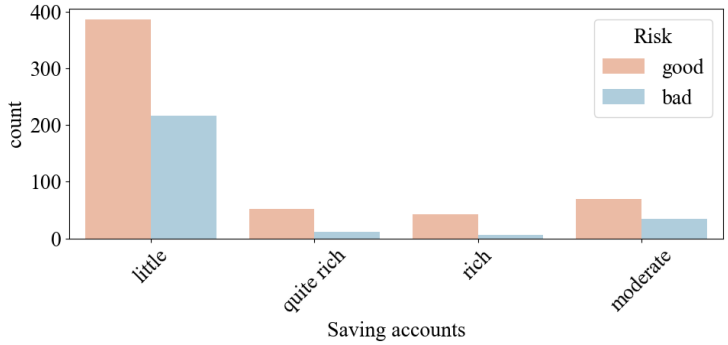

Figure 9. Comparison of account balances

Figure 9 analyzes the borrowing risk associated with personal savings. In this respect, people with very low private savings have a high risk of borrowing and are not immune to usury, with a coefficient of risk of around 380, and we suspect that most of the reasons for borrowing in this group are subsistence.

4. Conclusion

In this paper, for the prediction of credit default risk, after comparing different algorithms, we concluded that the Random Forest Model is the best-performing algorithm. We analyzed the ACC, AUC, recall and precision of each algorithm and found that the RF model has the advantage of 78.66% ACC, 79.71% AUC, 70.79% recall and 75.44% precision (the last two are the average of two calculations). After the mean value). Eventually, with the help of Random Forest and by analyzing the dataset, we determined that three features, loan amount, repayment term and borrower occupation, have high importance in predicting the value at risk. The relationship between these three feature quantities and the value at risk helps in assessing the default risk of the lender more accurately and formulating the risk control strategies accordingly. Ultimately this paper hopes that the results obtained can be used as a reference by banks or lenders when lending money and that this VaR data can be used as a reference as well as to reduce the percentage of credit defaults in the future for the better operation of banks as well as the financial market. It is also hoped that every borrower can get the amount they need after the risk value is derived. However, the research in this paper is still insufficient, for example, if a lender lends through different channels, it is impossible to know the final loan amount, which affects the final result. Therefore, this paper expects that in the future, we can have more information to analyze through other ways, for example, whether the same user has taken out loans in different banks, and how much the loan amount is respectively, to help the bank to make better judgments.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Pan H 2024 Research and application of personal credit default prediction method based on integrated learning (Xi'an: Xi'an University of Technology) PhD Thesis.

[2] Li Y, Peng Y, Xu M, et al. 2024 Loan default prediction model selection based on integrated learning algorithm enhancement method China Management Informatization 27(09) 141–144.

DOI: 10.54254/2755-2721/97/20241348

[3] Shao X 2023 Research on Investment Decision Making Based on Default Risk Prediction in Online Lending (Shandong: Shandong Institute of Commerce and Industry) PhD Thesis.
[4] Lai J 2023 Research on personal loan default prediction based on random forest model (Shihezi: Shihezi University) PhD Thesis.
[5] Zhang Z 2024 Research on the application of gradient enhancement method in credit risk assessment Automation Application 65(13) 26–31 DOI:10.19769/j.zdhy.2024.13.008.
[6] Zhang A 2024 Credit default prediction based on adaptive feature interaction learning (Yinchuan: Northern Nationalities University) PhD Thesis.
[7] She K-W, Shen Y, Zhao S 2024 The impact of big data on bank credit behavior: evidence from a digital social credit platform Economic Research 59(03) 147–165.
[8] Zhang Y, Zhuo P, Liu Z, et al. 2024 Credit default prediction model based on Transformer encoder and residual network Computer Applications 44(S1) 324–329.
[9] Han X, Jie P, He T 2024 Influence mechanism and empirical test of digital inclusive finance to alleviate financing constraints of small and medium-sized enterprises from the perspective of commercial banks West Finance (02) 12–19.
[10] Wang Y, Zhai G 2024 Research on credit risk management of local commercial banks under the background of big data Journal of Huaibei Institute of Vocational Technology 23(04) 101–106.