ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

Prediction method of financial market risk behavior based on big data mining algorithm

Haoyu Bi[1,*], Shihao Fang[2,a], Xinyi Bi[3,b]

[1]Central South University, Changsha, Hunan, China
[2]Beijing Institute of Technology, Zhuhai, Guangdong, China
[3]Hong Kong Baptist University, Hong Kong, China
[a]1668586530@qq.com, [b]b508192042@163.com
**Corresponding author:1981455279@qq.com*
These authors contributed equally to this work.

Abstract: *This paper predicts the default behavior of the financial market, reduces the bad debt rate in bank loans and securities investment, and discovers potential risks in time. The currently used technologies mainly rely on models with static weights, such as simple linear models. The advantage of these algorithms is that they are fast. But in a large number of samples, these algorithms also face inaccurate problems, requiring the use of machine learning modeling methods to train models. This paper proposes a modeling framework for financial data mining algorithms based on random forests, which can accurately predict microscopic behaviors and reduce financial risks. The experimental results show that the method proposed in this paper has certain application value, the prediction accuracy (precision) reaches 85%, and the recall rate (recall) reaches 90%.*

Keywords: *Decision tree, Random forest, Logistic regression, Risk prediction*

1. Introduction

In recent years, with the continuous development of economy, the threshold to enter the financial market has been constantly lowered, and investment methods have become increasingly rich. More and more people allocate assets through the financial market to realize the appreciation or preservation of assets, but this has also led to frequent defaults, and many investors and enterprises have suffered heavy losses.[1][2]Therefore, we should predict the default behavior in the financial market, reduce the bad debt rate of bank loan securities investment and debt securities investment, improve the yield, minimize the risk of default behavior in advance, and try to avoid the financial crisis caused by corporate losses affecting more enterprises. At the same time, it also strengthens the supervision of market default behaviors to minimize the risks of financial behaviors and maximize the benefits.

Under the background that big data methods are widely used in financial risk research[3]-[5], the current techniques for predicting default behavior in financial markets rely primarily on models with static weights, such as simple linear models. This method is to transfer the trained weights to another place to run. Generally speaking, the weight information and weight distribution will basically remain unchanged. Therefore, although the training and prediction speed of this type of algorithm is very fast, it also faces the problem of inaccuracy in a large number of samples. In addition, a simple linear models cannot solve nonlinear problems, but there are many problems in the financial market that cannot be divided linearly, which also leads to inability to predict default behaviors more accurately.

Based on the limitations of current forecasting methods, we propose to use data mining and machine learning modeling methods to train models, that is, to use machine learning-based financial data mining algorithm modeling frameworks to accurately predict microscopic behaviors, thereby reducing financial risks[6]. Machine learning is a technique that enables computers to learn from supplied data[7]. It aims to model the relationships of input data and reconstruct knowledge scenarios. With the increase of computing power of computer servers, the performance of machine learning has been significantly improved, and classification and prediction based on known data can achieve high accuracy and reliability. That is, it can learn and update itself after each new data is obtained to achieve a more accurate prediction result. For example, banks use machine learning and big data technology to calculate possible risks and fraudsters. Such machine learning-driven fraud detection systems can actively learn and calibrate new real or potential threats, detect and report abnormal behaviors, in order

ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

to reduce the bad debt rate and losses.

After experimental verification, our machine learning-based financial data mining algorithm and model have certain application value. In this model, the prediction precision rate of machine learning samples has reached more than 85%, that is, the correct samples investigated account for 85% of the total samples, and the accuracy rate is high; and the recall rate is also as high as 90% under this model, that is, the investigated The correct samples account for 90% of the total correct samples, indicating that the model has strong machine learning capabilities, can respond to dynamics in real time, and more comprehensively analyze and predict default behaviors. To sum up, our machine learning-based financial data mining algorithm and model can accurately, efficiently, and comprehensively predict financial market defaults, reduce the bad debt rate in bank loans, trust investment and other fields, improve yields, and timely. Identify potential risks.

2. Method

2.1. Model principle

Decision tree is a common machine learning method in the field of data mining, named because of its tree-like logical structure. The algorithm makes judgment and decision based on the tree structure, which is similar to the decision processing mechanism of the human brain. Decision tree belongs to supervised machine learning. The model uses information entropy as a measurement basis to construct a tree with the fastest entropy drop. When the node entropy of a leaf drops to zero, the sample data in the leaf belongs to the same class. The most significant disadvantage of the decision tree model is that when the dataset or noise is too large, it is easy to fall into overfitting. In order to further improve this shortcoming, Breiman proposed a random forest algorithm. The random forest algorithm is a combination of decision tree and Bagging Ensemble Learning methods. The steps are: first, random sampling with replacement is performed on all training samples to form multiple training sets containing 63.2% of the original samples, and then use these training sets to construct decision trees respectively. When constructing each decision tree, n features are randomly selected, and the features are screened according to the criterion with the smallest Gini coefficient, so as to select the optimal feature generation node, so that each tree can fully grow without pruning. Finally, a simple vote is performed, and the classification with the highest number of votes is output. Since random forest randomly selects the training set and features from the original data and has the characteristics of random training set and random attribute, so it has good anti-noise ability, which is suitable for the financial market with a lot of changes and noise. In order to compare with the random forest model and highlight the superiority of the model, this paper also introduces the logistic regression method for comparative research.

2.2. Model implementation

If the dependent variable Y has n observations, there are k independent variables related to it. When building a classification tree, the random forest will randomly reselect n observations in the original data. Some observations are selected multiple times and some are not selected. This is called Bootstrap's method of resampling. At the same time, random forest randomly selects some variables from k independent variables to determine the nodes of the classification tree. In this way, the classification tree constructed each time may be different. In general, random forests randomly generate hundreds to thousands of classification trees, and then select the tree with the highest degree of repetition as the final result. When using the random forest model, first use the self-help sampling method to randomly select n samples from the sample data set to form a sample set Y_i （i=1,2,3,...,k).Then randomly select m independent variables from all independent variables, form the feature input vector X from these m independent variables, and establish a non-pruning decision tree classifier $f(Y_i, X_i)$（i=1,2,3,...,k), after that using these k base learners to form a combined classification decision system, and use the simple voting method to make the final prediction.

2.3. Parameter selection

The characteristic of random forest is that it has two layers of randomness, namely sampling randomness and node randomness. Due to its double randomness, the two most important parameters in building a random forest model are the number of trees and the number of node features. Before building the model, some parameters need to be determined in advance. The most important parameter

ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

is nestimators, which represents the number of basic learners. In this article, it refers to the number of decision trees in the random forest model. We take n_stimators = 100. In addition, the basic learner in the random forest does not traverse all the feature variables when dividing the nodes, but randomly selects some of the feature variables as a set of feature variables. Next, selecting the optimal feature variable from the candidate variables. Therefore, the number of feature variables in the candidate feature variable set is very important, which is controlled by the max_features parameter. The smaller the parameter, the smaller the feature variables in the candidate set, and the more different the formed base classifiers. Here, we use the GridSearchCV module function in the sklearn library, adopt the method of cross-validation. Finally, we determine max_features=2.

3. Results

3.1. Description of data statistics

There are 150000 data in the easy. The variables we want to predict are: probability PR of behavior risk, that is, overdue arrears exceeding 90 days or worse. We set it as y variable and other variables as independent variable x. The independent variables X are: V1 (borrower's age at that time), V2 (debt ratio), V3 (number of open-end loans and loans, number of open-end loans (installment payments such as auto loans or mortgages) and credits (such as credit cards), V4 number of real estate loans or lines: Mortgage loans and real estate loans, including home equity credit lines) Number of time 30-59 days past due not worth (35-59 days overdue but not bad), monthly income (monthly income), number of dependents (number of family members: number of family members excluding myself), a total of 7 variables.

We describe different x variables separately, which is divided into seven parts.

3.1.1. Age (the age of the borrower at that time)

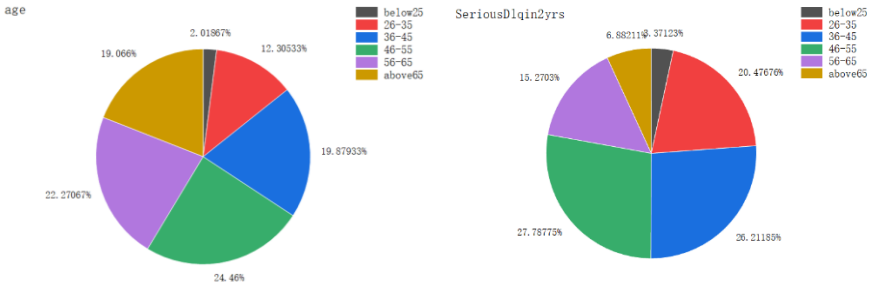

Figure 1: The age of the borrower

There are 150000 data in this variable, of which 2.0187% are under the age of 25, and the default rate among people under the age of 25 is 11.1625%. The number of people aged 26-35 is 18458, accounting for 12.3053%. The default rate accounts for 11.1226%, and the number of people aged 36-45 accounts for 19.8793%. The number of people in default is 2628, accounting for 8.8131% of the population aged 36-45. The population aged 56-65 accounted for 22.2707%, and the number of defaulters in this population was 1531, accounting for 4.5830%. The number of people over 65 years old is 28599, and the number of people in breach of contract accounts for 2.4127%. From these data, it can be seen that the high incidence of default is among people under the age of 25, followed by those aged 26-35. We can learn that defaults are high among young people who are not sensitive to credit rating.

3.1.2. Debt Ratio

There are 150000 data in this variable. The number of people with a debt ratio less than 0.5 accounts for about 62.4720%, and the number of people with a debt ratio greater than 0.5 accounts for 37.5280%. From this data, we can see that the majority of people have a debt ratio below 0.5, which accounts for 3.77% of the total population, and the remaining number of defaulters accounts for 2.914% of the total population.

ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

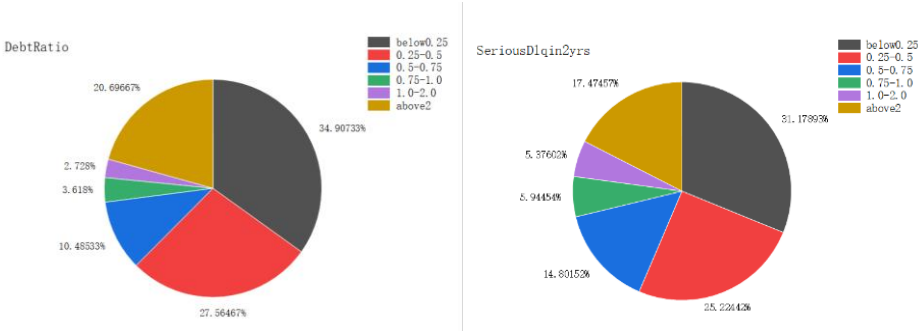

Figure 2: Debt ratio

3.1.3. Number of open credit lines and loans (number of open-end loans and loans, number of open-end loans (installment payments such as auto loans or mortgages) and credits (such as credit cards))

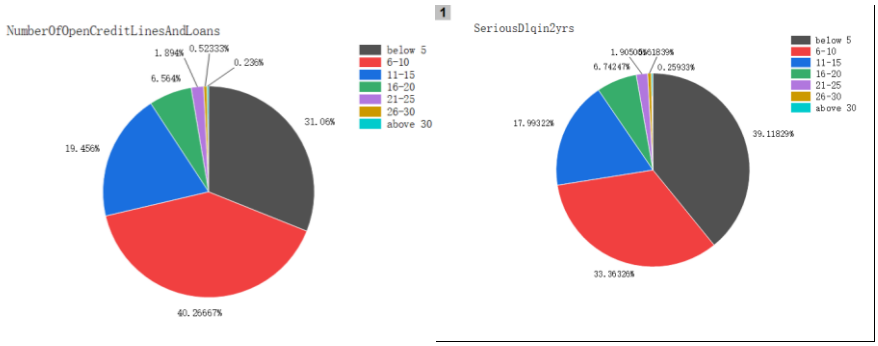

Figure 3: Number of open-end loans

There are 150000 data in this group. There are data gaps in the number of open-end loans and loans at 15 places. The total number of open-end loans and loans is 136174, accounting for 90.7827%, but the number of defaults only accounts for 60.4733% of the total number of defaults. It can be seen that when the number of open-end loans and loans, open-end loans and loans is too high (higher than 15), the default risk will increase, And the default rate of such credit increased by a cliff in 15 places.

3.1.4. Number of real estate loans or lines (mortgage loans and real estate loans, including home equity credit lines)

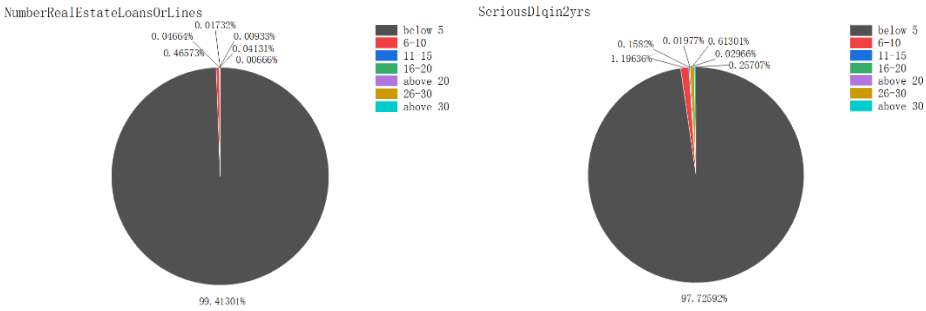

Figure 4: Number of real estate loans or lines

There are 150000 data in this group. The number of people with less than 5 real estate accounts for 99.4713% of the total number, but the default rate is only 6.6243%, while the default rate of the rest is quite high. For people with 6-10 real estates, there are about 120 defaults for every 700 people, accounting for about 17.3104% of this population. The default rates of people with 11-15, 16-20 and more than 20 real estate accounts for 22.8571%, 21.4286% and 20% respectively. It shows that the number of real estate is roughly proportional to the default rate.

ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

3.1.5. Number of time 30-59 days past due not worse (35-59 days overdue but not bad)

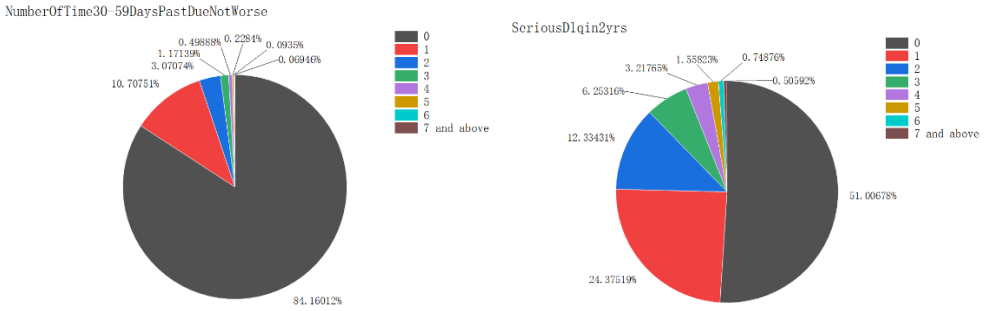

Figure 5: Number of time 30-59 days past due

There are 150000 data in this group. We can see that about 84% of people are not overdue for 35-59 days, and the default rate of such people is only 4%. However, about 2% of people are still overdue but not bad three or more times. In addition, the number of defaulters in this group accounts for 30% - 50% of their respective groups, and the default rate is quite high. It can be seen that the number of overdue times is roughly positively correlated with the situation of breach of contract.

3.1.6. Monthly Income (Monthly income)

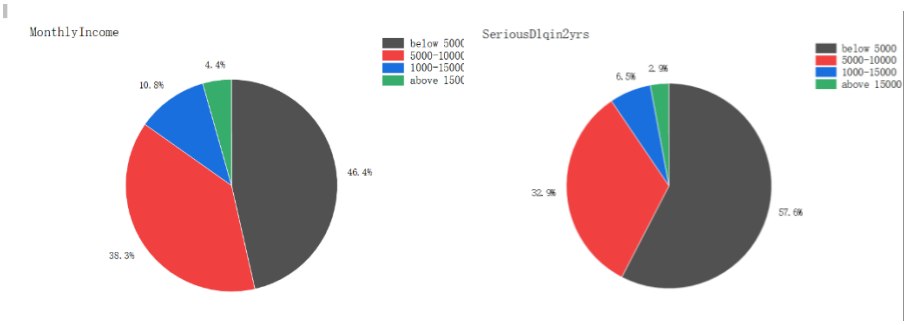

Figure 6: Monthly income

There are 150000 data in this group. From this group, people with monthly income of less than 5000 account for 57.6% of the defaulters, followed by people with monthly income of 5000-10000, accounting for about 33% of the total defaulters. On the whole, the higher the monthly income, the smaller the probability of default. People with lower income may be unable to make ends meet. The default rate is much higher than that of high-income people.

3.1.7. Number of dependents (number of family members: number of family members excluding myself)

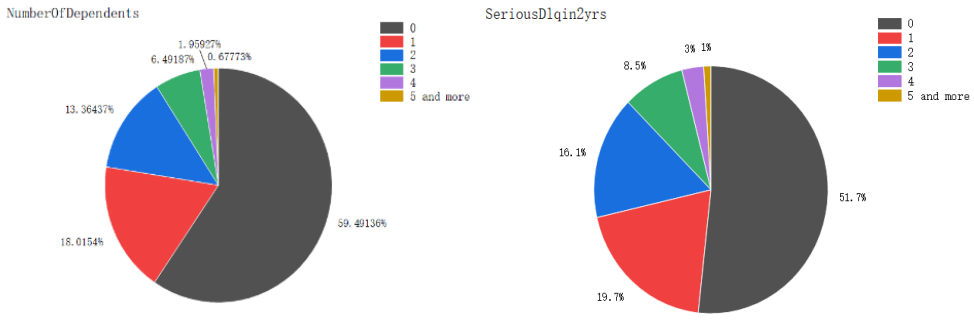

Figure 7: Number of family members

There are 150000 data in this group. The population with 0 family members accounts for more than half of the total population. The proportion of defaulters in each group is basically the same. The default rate of the population with 0 family members is the lowest, only 5.8629% of this group. The rest are between 7% - 10%.

ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

(Tips: the pie chart in the analysis of each group of variables, the former is the population distribution of the variable, and the total amount of the latter is the total number of defaulters under this variable, expressing the different proportion of each type of defaulters under this variable.)

3.2. Sample index analysis

Accuracy is the correct number of investigations divided by the total number. We can see from the analysis and results of two data that the prediction accuracy of machine learning training samples and verification samples under this model can reach more than 85%. We need to correctly understand the difference between accuracy rate and recall rate. Recall rate reflects the proportion of positive cases correctly judged in the total positive cases. With the data support of this paper, we can get that the recall rate of our data under the financial data mining algorithm and model has reached 90%. The recall rate can reflect the results of machine learning. The higher the recall rate, the stronger the "recall" ability of machine retrieval. The financial data mining algorithm and model based on this kind of machine learning are very efficient and accurate, which reflects the great advantages of our algorithm. We can predict the default behavior of the financial market through this model algorithm. To reduce the bad debt rate in bank loans and securities investment and find the potential risks in time. We can see from the expression ($F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$) of F1 that when the accuracy rate and recall rate are very high, the F1 value naturally increases.

3.3. Analysis of test results

First of all, we should know that AUC is defined as the area surrounded by the curve ROC and the x-axis of the coordinate axis, and this area is always less than 1. At the same time, since the ROC curve is generally located above the primary function of y = x, we can guide the value of AUC to be between 0.5-1. When the AUC is closer to 1, it can be proved that the authenticity of our detection method is higher. On the contrary, if the value of AUC is equal to 0.5, the lower its authenticity is. If the authenticity is too low, the detection method has no application value. We can see from the sample test results:

```
the best parameter: {'max_features': 2, 'min_samples_leaf': 50}
the best score: 0.862773390611
auc= 0.907379483293
auc= 0.864284730976
```

Figure 8: Test results

The results of sample test data are above 0.8, while the best test data is 0.8628, which is close to 1, indicating that our detection method has high authenticity, which can further verify that the detection method under our model has high application valu

4. Related research

Data mining is a technology that combines the results of multiple disciplines. Traditional methods relying on static weights such as simple regression analysis methods may face inaccurate problems in dealing with nonlinear problems. On this basis, machine learning methods such as decision trees, random forests, neural networks, logistic regression, regression Decision trees, etc., solve the nonlinear boundary problem well.

Among the current mainstream machine learning methods. Neural networks can be applied to classification and regression, and still have good performance in solving multi-parameter problems. Decision trees and random forests are mainly used in classification problems. Logistic regression can handle the sub-linear boundary problem in the binary classification problem well, and it is also a kind of nonlinear regression.

5. Conclusion

This paper proposes a new method, which is used in the fields of bank loan, bond investment, securities investment, trust investment, etc. It provides a new technical method to reduce the bad debt rate, reduce investment risk and discover risks for these fields. Using machine learning and big data to

ISSN 2616-5775 Vol. 5, Issue 3: 78-84, DOI: 10.25236/AJCIS.2022.050311

predict risks in the financial market has certain reference significance, which is reflected in several aspects. The results show that the method in this paper has the advantages of accurate, comprehensive and real-time dynamic feedback.

References

[1] LIU Na. Research on financial engineering in the field of financial market risk management [J]. Chinese Industry & Economy, 2021(17):136-137.
[2] WANG Peng ,TANG Zhen Yuan.The integration of big data application and management science and its impact on financial risk control [J].Chinese Agricultural Accounting,2021(11):65-67. DOI:10.13575/j.cnki.319.2021.11.022.
[3] LI Fei.Analysis of the application of big data method in the study of systemic financial risk [J].China CIO News,2021(09):87-89.
[4] WU Da Sheng.Talking about the research of big data technology in the field of financial risk control [J].China New Telecommunications,2021,23(18):99-100.
[5] HE Qing Quan,XU Jie,WANG Hui.Analysis of the impact of financial technology on commercial banks [J].Modern Business,2021(25):87-89.DOI:10.14097/j.cnki.5392/2021.25.028.
[6] TIAN Xiao Li,DING Jing Bo,BAI han.Causes and solutions of digital financial credit risk in the context of big data [J].Investment And Entrepreneurship,2021,32(15):1-3.
[7] ZHANG Shi Jie.Application of Machine Learning in Stock Market Prediction with Big Data[J].Journal of Guiyang University(Social Sciences),2021,16(04):43-48.DOI:10.16856/j.cnki. 52-1141/c.2021.04.007.