Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI: 10.61784/tsshr3009

FINANCIAL CREDIT RISK ASSESSMENT BASED ON MACHINE LEARNING

MingYue Gao
School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: 925483956@qq.com

Abstract: In the era of big data, the financial industry is facing new challenges and opportunities. Through big data and artificial intelligence technology, we can more accurately assess and manage various types of financial risks, including credit risk, market risk, fraud risk, etc. In this paper, the decision tree model is used to model and analyze the credit risk in financial risk, and the financial risk prevention and control system is established. For credit risk, according to the bank customer information data, after data processing, the decision tree classification method is used to judge whether the customer may default in the future through the customer's basic information, and finally the main discriminant basis is age, expected income, balance and number of credit cards. Then, from the five dimensions of establishing a sound risk assessment and early warning mechanism, improving citizens' financial risk awareness, promoting the construction of financial stability guarantee fund, strengthening industry supervision and self-discipline, and improving the legal and regulatory system, feasible suggestions are put forward for financial risks.

Keywords: Financial credit risk; Machine learning; Data mining; Decision tree model

1 INTRODUCTION

In the context of globalization and economic integration, financial risks have become an important factor affecting national economic security and social stability. Especially under the impetus of scientific and technological progress, the financial industry is undergoing unprecedented changes, which not only brings development opportunities, but also brings unprecedented challenges.

In recent years, China's financial industry has undergone tremendous changes driven by science and technology. With its unique advantages and broad application prospects, financial technology has become an important force to promote financial innovation and development. However, the development of financial technology has also brought a series of new financial risks. For example, the rise of Internet finance has made the boundaries of financial services more blurred, and the participation of non-traditional financial institutions has increased the complexity and uncertainty of the market. At the same time, the application of big data, artificial intelligence and other technologies in the financial field has also brought new risks such as data leakage and algorithm discrimination. In addition to technological factors, macroeconomic environment, policy changes, market volatility and other factors also have an important impact on financial risks. In the context of globalization, the unstable factors of the international financial market may be transmitted to the domestic financial market through channels such as trade and capital flows, posing a threat to China's financial stability. Therefore, we need to pay close attention to the changes in the economic and financial situation at home and abroad, and strengthen the construction of risk monitoring and early warning mechanisms.

In order to cope with these challenges, it is particularly important to promote the construction of financial stability guarantee fund. As a special financial system arrangement, the financial stability guarantee fund aims to enhance the ability of the financial system to cope with sudden risks by raising funds in advance and effectively dealing with risks afterwards. The sources of funds such as the payment of financial institutions, the return of disposal funding and the injection of financial funds provide a solid material basis for the operation of the financial stability guarantee fund. However, in the face of systemic and cross-industry financial risks, the existing guarantee fund may be difficult to respond effectively. Therefore, it is of practical significance and urgency to establish a financial stability guarantee fund. By improving the operation mechanism of the financial stability guarantee fund, it can better play its important role in preventing and defusing financial risks.

In order to monitor and respond to financial risks in a timely manner, the construction of early warning systems is particularly important. By integrating various types of risk information and establishing a scientific risk assessment model, the early warning system can issue an alarm before the risk breaks out, providing sufficient time and space for decision makers to respond. In the process of dealing with financial risks, enterprises need to constantly innovate. Through technological innovation and business transformation, we can improve our competitiveness and ability to resist risks. By improving the company's governance structure and strengthening internal control, we can reduce risks and improve the company's market reputation and brand value.

2 RESEARCH STATUS AND METHODS

2.1 Research Status

Credit score is the core evaluation index in customer default prediction. Fisher first proposed to provide a quantitative risk judgment basis for lenders by comprehensively evaluating the borrower's credit history, financial status, repayment ability and other factors[1]. Through credit scoring, lenders can more accurately assess the default risk of borrowers, so as to formulate more reasonable loan policies and risk management strategies. In addition to credit scoring, a variety of classification prediction models are widely used abroad to predict customer defaults. These models include but are not limited to logistic regression decision trees, random forests, neural networks, etc.

Wiginton used Logistic regression model to explore the credit score in his study[2]. He emphasized that there is no need to impose special restrictions on the distribution of explanatory variables when making judgment analysis, which increases the flexibility and universality of the model in practical applications. Subsequently, Makowski introduced the decision tree model into personal credit evaluation for the first time[3]. This innovation has brought new perspectives and methods to the field of credit evaluation. The credit scoring model constructed by Coats and Fant is based on the neural network method, and widely collects the actual loan data of various countries for empirical research, thus verifying the validity and practicability of the model. These studies not only enrich the theoretical framework in the field of credit scoring, but also provide lenders with more diversified and more accurate assessment tools[4].

In recent years, with the progress of science and technology and the continuous innovation of data analysis methods, the research in the field of customer default prediction has also made significant progress. More and more studies have confirmed that machine learning algorithms show better performance than traditional methods in predicting customer default risk. Obare et al. fully proved this point. They deeply explored the application of machine learning in customer default prediction and achieved remarkable results. Machine learning algorithms can process a large amount of data and extract complex patterns and rules that are difficult to identify manually[5]. Chopra and Bhilare conducted in-depth research on bank loan data sets and found that the gradient boosting model showed significant advantages over the decision tree in prediction performance[6]. Ampountolas et al. conducted empirical research using actual data to compare the performance of various machine learning models in classification prediction[7]. The results show that the random forest model has certain advantages in classification prediction, which provides a new reference for loan institutions in selecting and applying risk prediction models.

The above research not only promotes the application of machine learning in the field of loan risk prediction, but also provides a more accurate and efficient tool for the risk management of financial institutions. By training and optimizing these algorithms, the default risk of customers can be predicted more accurately, so as to provide more reliable decision support for loan institutions. Compared with traditional credit scoring methods and classification prediction models, machine learning algorithms have stronger flexibility and adaptability. They can automatically adjust parameters to adapt to the characteristics of different data sets and maintain high prediction accuracy in the face of complex and changeable market environment and customer behavior.

In addition, with the continuous development of big data and artificial intelligence technology, the customer default prediction system is also constantly upgrading and improving. By using more abundant data sources and more advanced algorithm models, lenders can achieve a more comprehensive and in-depth analysis of the borrower's credit status, and further improve the accuracy and reliability of loan default prediction.

2.2 Decision Tree Algorithm

The decision tree is a tree structure in the form of a flow chart. It is used to calculate the probability that the expected value of the net present value is greater than or equal to zero by constructing a decision tree based on the known probability of occurrence of various situations, so as to evaluate the project risk and judge its feasibility. This decision branch is drawn into a graph like the branch of a tree, so it is named decision tree. In machine learning, decision tree is a prediction model, which represents a mapping relationship between object attributes and object values.

The decision tree contains three types of nodes: decision nodes (usually represented by rectangular boxes), opportunity nodes (usually represented by circles), and endpoints (usually represented by triangles). Each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category or prediction result.

The decision tree classification in this paper is based on the Gini value in the decision tree, that is, the Gini coefficient, which is an important indicator for evaluating the uniformity of data distribution or the purity of nodes. In the decision tree algorithm, the Gini coefficient is used for the decision-making process of feature selection and node division. The smaller the value is, the more uniform the data distribution is. The larger the value, the more uneven the data distribution.

The classification calculation formula of Gini coefficient is:

$$\mathrm{gini(T)} = \frac{\mathrm{S_1}}{\mathrm{S_1+S_2}}\mathrm{gini(T_1)} + \frac{\mathrm{S_2}}{\mathrm{S_1+S_2}}\mathrm{gini(T_1)} \tag{1}$$

Among them, S_1 and S_2 are the respective sample sizes of the two types after division.

In the process of building a decision tree, we first need to draw a decision tree according to the actual situation, which includes predicting various events that may occur in the future and expressing these situations in a tree diagram. Then, the expected value of each node is calculated, and the scheme is optimized by comparing the expected values of different schemes. Finally, pruning may be required to further simplify the decision tree and improve the prediction accuracy. Decision tree is divided into classification tree and regression tree. The classification tree is used to deal with discrete variables, while the regression tree is used to deal with continuous variables.

In machine learning and data mining, decision tree is a very common technology, which can be used for classification, regression, feature selection and other tasks. It has been widely used in the financial field because of its intuitive, easy to implement and strong explanatory. The decision tree divides the data into different subsets by constructing a tree structure, so as to realize the classification and prediction of the data. In the risk management of customer default, the decision tree can help financial institutions identify the key factors affecting customer default, establish a customer default prediction model, and then provide decision support for risk management. In general, decision tree is an intuitive and powerful decision analysis method, which can help people make wise decisions in complex situations. By constructing decision trees, people can better understand various possible results and probabilities, so as to make more reasonable choices.

3 ESTABLISHMENT OF DECISION TREE MODEL

This study aims to use decision tree technology to judge and predict the default risk of bank customers, which has important theoretical and practical significance. From a theoretical point of view, this study can further improve and enrich the theoretical system of risk management. By applying decision tree technology to customer default risk management, we can deeply explore the inherent laws and characteristics of customer default risk, and reveal the key factors affecting customer default and its mechanism of action. This will help to promote the innovation and development of risk management theory and provide financial institutions with more scientific and effective risk management methods and tools. From a practical point of view, this study has important application value. First of all, by constructing a customer default prediction model, financial institutions can achieve accurate identification and evaluation of customer default risks, so as to formulate more reasonable credit policies and risk management strategies. Secondly, the interpretability of the decision tree model is strong, so that risk managers can intuitively understand the working principle and prediction results of the model, which is convenient for application and adjustment in practice. In addition, with the development of big data technology, financial institutions can obtain more abundant customer data, which provides a broader space and possibility for the application of decision tree model.

3.1 Data Source

In foreign countries, especially in Western countries, the long-standing credit culture and consumption concept have prompted people to be more inclined to achieve their life goals through borrowing. In this environment, borrowing is regarded as a normal economic behavior, which helps to improve the quality of life. In China, due to the different historical, cultural and economic development stages, people pay more attention to savings and stability, and hold a more cautious attitude towards borrowing. In addition, the policy and social environment have also affected people’s consumption and borrowing habits to a certain extent. In recent years, China has also promoted the development of consumer finance to encourage reasonable consumer credit to meet people’s growing consumer demand.

In this paper, the data are selected from 10000 customer data of banks in three regions abroad, and ten relevant information about customers are selected from them, namely: credit score, geographical location, gender, age, tenure, balance, number of products, whether to have a credit card, whether to be an active user, and estimated income.

3.2 Descriptive Statistics

Figure 1 shows that boys account for 45.43 % and girls account for 54.57 %, indicating that the gender ratio of the subjects surveyed is relatively balanced. At the same time, from the geographical point of view, France has a larger number, Germany and Spain have the same number. Secondly, from the perspective of age distribution, most people are between 20 and 80 years old, which meets the age limit requirements of bank customers. Finally, in terms of tenure, the number of people in office for 1 to 9 years is relatively uniform, indicating that most people work relatively stable.

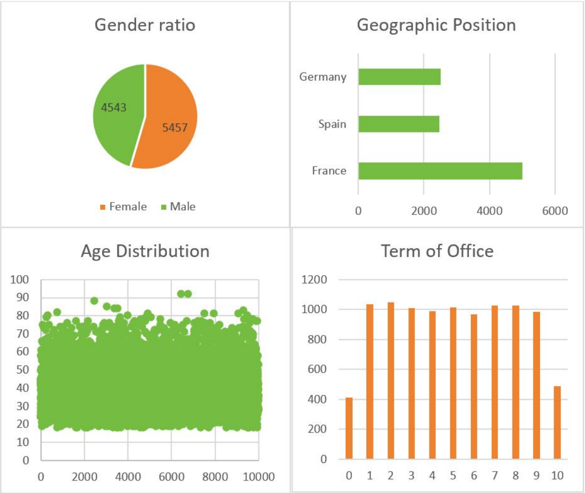

Figure 1 Descriptive Statistic

3.3 Data Preprocessing

Through observation, it is found that the data does not have missing values and outliers, so it does not need to be processed. However, there are two non-numeric characteristic variables. For the subsequent modeling analysis, the two characteristic variables of country and agenda are converted into values through the relevant code in python. At the same time, in order to facilitate the examination of whether the customer defaults, the credit score variable is divided into two categories according to the average value. The data below the average value is recorded as 0, representing default, and the data above the average value is recorded as 1, representing trustworthiness. In order to construct a decision tree model to predict bank customer defaults, the processed data is divided into a training set and a test set according to the ratio of 8:2, as shown in the Table 1.

Table 1 Dataset Classification

	Number of trustworthy people	Number of defaults	Grand total
Training sets	4109	3891	8000
Testing set	1027	973	2000
Grand total	5136	4864	10000

3.4 Basic Evaluation

ROC curve, also known as receiver operating characteristic curve or receiver operating characteristic curve, is to draw the relationship curve between True Positive Rate and False Positive Rate under different thresholds by changing the decision threshold of the two classifiers. On the ROC curve, the point closest to the upper left of the coordinate map is the critical value with high sensitivity and specificity. When the ROC curve of the model is closer to the critical value, the classification result of the method is more effective. The area under the curve is a quantitative index used to evaluate the overall performance of the diagnostic method. The value range is 0 to 1, and the larger the value is, the better the classification effect is.

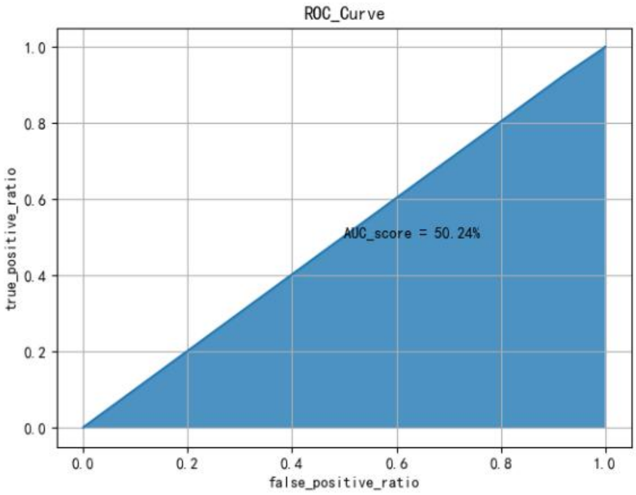

Figure 2 ROC Curve

As shown in the Figure 2, the AUC value is 50.24 %, indicating that the gap between the two samples of the data is obvious, and the classification result is better using the decision tree.

3.5 The Results of Customer Default Prediction Based on Decision Tree Model

First, the initial decision tree model will be trained using the training set of the divided bank customer credit score data. By inputting the training data into the model, the model will learn how to predict whether customers default based on their credit score characteristics. After the training, we get an initial decision tree model with predictive ability.

After obtaining the trained decision tree model, our next step is to import the test set of bank customer credit score data into this model. The role of the test set is to evaluate the predictive performance of the model, rather than participating in the training process of the model. We will use the model to predict the data in the test set to determine whether the customer defaults. In order to evaluate the prediction performance of the model, we need to calculate some evaluation indicators. These indicators will help us understand the accuracy, recall rate and other key information of the model in predicting whether customers default. By comparing these indicators, we can have a comprehensive understanding of the performance of the model.

Finally, in order to improve the prediction effect of the model on customer default, we need to optimize the parameters

of the decision tree model. This can be achieved by modifying the parameter values of the model, such as adjusting the maximum depth of the tree, the minimum number of samples required for leaf nodes, etc. In the process of adjusting the parameters, we evaluate the influence of different parameter settings on the performance of the model according to the change of the evaluation index, so as to find the optimal parameter combination. Through this process, a more accurate and reliable decision tree model can be obtained to predict the default risk of bank customers.
According to the above process, the final decision tree is obtained in the Figure 3.

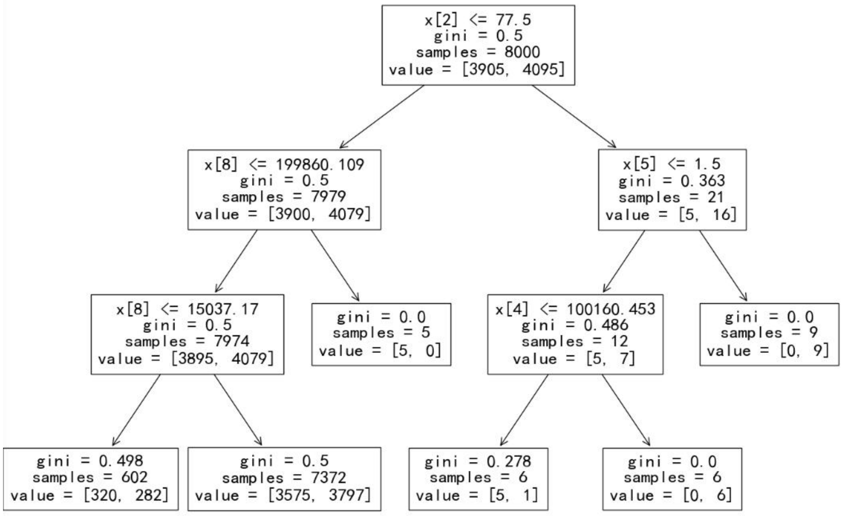

Figure 3 Decision Tree

According to the results, judging whether a bank customer defaults has nothing to do with geographical location and gender. The overall decision tree is first divided into two categories based on age. For those younger than 77.5 years old (left half branch), the estimated income is used as the classification point, and then divided downward. For those who are older than 77.5 years old (the right half branch), first take whether they have a credit card as the classification point, and then examine the balance status of the customers who have a credit card, and finally judge whether they default according to their term of office.

4 CONCLUSIONS AND RECOMMENDATIONS

4.1 Research Conclusion

Based on the bank customer data, the decision tree is established to determine whether the customer defaults. Finally, the customer will be judged to be in default in two cases. One is when the customer is younger than 77.5 years old and the expected income is lower than 15037.17, and the other is older than 77.5 years old. At the same time, the number of credit cards is small and the balance is less than 100160.453. This result is in line with expectations, and by analyzing the various attributes and historical data of customers, it can help banks assess customer credit risk, help banks understand customer repayment ability more accurately, and thus manage loan risk more effectively.
At the same time, banks can understand the differences between different customer groups, and adjust the customer service strategy accordingly, which can provide decision support for banks and help them formulate more intelligent loan approval strategies. For example, stricter risk control measures can be adopted for high-risk customers, while more favorable loan conditions can be provided for low-risk customers, thereby improving customer satisfaction and loyalty.
Finally, banks can conduct customer default risk assessment based on decision trees, and banks can reduce the losses caused by loan defaults. Early detection of customers who may default, you can take appropriate measures, such as raising interest rates, reducing the amount of loans or requiring guarantees to reduce the risk.
In summary, the customer default prediction model based on decision tree is of great significance to banks, which can help banks better manage risks, optimize customer service, and improve the accuracy and efficiency of loan approval.

4.2 Suggestions

With the continuous development of the global economy, financial risks have increasingly become an important issue that we cannot ignore. Especially in the financial industry reform driven by scientific and technological progress, the complexity and diversity of financial risks are becoming more and more prominent. According to the conclusion of this paper and the current financial risk situation, the following suggestions are put forward:

4.2.1 Establish a sound risk assessment and early warning mechanism

Financial institutions should establish a sound risk assessment system to regularly assess and monitor various financial risks. Through big data analysis, machine learning and other technical means, a risk early warning model is constructed to detect and warn potential risks in time. At the same time, we should strengthen the comprehensive monitoring of market, credit, liquidity and other risks to ensure that the risks are measurable, controllable and bearable.

4.2.2 Improve citizens' awareness of financial risks

Through media publicity and educational activities, citizens' awareness and awareness of financial risks should be improved. Enable citizens to rationally view financial market fluctuations and avoid blind investment and excessive borrowing. At the same time, we should strengthen the protection of financial consumers' rights and interests and maintain the fairness, transparency and stability of financial markets.

4.2.3 Promote the construction of financial stability guarantee fund

Drawing on international experience and combining with China's actual situation, we will accelerate the construction of financial stability guarantee funds. The source of funds should be diversified, including financial institutions to pay, financial capital injection and so on. Through the establishment of a financial stability guarantee fund, it can provide financial support for the disposal of systemic financial risks and reduce the risk exposure of financial institutions and the entire financial system.

4.2.4 Strengthen industry regulation and self-discipline

The regulatory authorities should strengthen the supervision of financial institutions to ensure that their business operations are compliant and robust. At the same time, promote financial institutions to strengthen self-discipline, establish and improve the internal risk control system, improve risk management and disposal capacity. In addition, information sharing and collaboration between industries should be strengthened to jointly cope with cross-industry financial risks.

4.2.5 Improve the legal system

In view of the new problems and challenges in the financial field, the relevant laws and regulations system should be improved in time. Through legislative means to clarify the power and responsibility relationship between financial institutions and regulatory authorities, and regulate the order of financial markets. At the same time, increase the punishment of illegal acts, to form an effective deterrent and restraint.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Fisher R A. The use of multiple measurements in taxonomic problems. Annals of eugenics, 1936, 7(2): 179-188.

[2] Wiginton J C. A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 1980, 15(3): 757-770.

[3] Makowski P. Credit scoring branches out. Credit World, 1985, 75(1): 30-37.

[4] Coats P K, Fant L F. Recognizing financial distress patterns using a neural network tool. Financial management, 1993: 142-155.

[5] Obare D, Murary M. Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. Am. J. Appl. Math. Stat, 2018, 6(6): 266-271.

[6] Chopra A, Bhilare P. Application of ensemble models in credit scoring models. Business Perspectives and Research, 2018, 6(2): 129-141.

[7] Ampountolas A, Nyarko Nde T, Date P, et al. A machine learning approach for micro-credit scoring. Risks, 2021, 9(3): 50.