Research on risk assessment of clients before loan based on decision tree algorithm

Yongkai Wang1,*, Ditao Duan2

1Shandong Management University, Jinan City, China 250357
2School of Finance, Zhongnan University of Economics and Law, Wuhan, 430073, China

Corresponding author and e-mail: Yongkai Wang, 14438120190531@sdmu.edu.cn

Abstract. How to transform credit funds into the endogenous development capacity of the market and promote the sustainable development of the economy, the first consideration is how to conduct risk management. How to do a good job in risk assessment has gradually become an important part of current financial risk management. This paper uses the decision tree algorithm to conduct a research on the risk assessment of customers before loan. After the model verification of the cross-validation method and the random validation method, the results show that the risk estimation accuracy of the decision tree algorithm reaches 81.2% and 83.6%, which proves that this model can be an effective reference for pre-loan risk assessment.

1. Introduction

Finance is the core of modern economy and the core competitiveness of the country. The orderly development of finance can effectively promote the development of the real economy. Many financial institutions are currently formulating their own future development strategies in accordance with my country's economic development situation in order to coordinate with the pace of economic construction and economic development. During the development and expansion of the financial industry, due to the lag in the development of risk management methods, a large number of financial risk problems have arisen. Under this situation, financial institutions across the industry are required to actively adjust their financial risk management. For this reason, many experts and scholars have conducted corresponding risk management research.

In 2020, Tingting Xie and Yi Li build a comprehensive evaluation index system from 8 dimensions including macroeconomic development risks, asset price bubble risks and government regulation risks. The entropy method is used to assign weights and measure the evolution trend of Xinjiang's financial risks from 2013 to 2018 [1]. Yuanyuan Huo et al. established a credit risk assessment method for manufacturing enterprises based on the Probit model. Through index selection and testing, after measuring the credit risk of listed manufacturing enterprises in China, the accuracy of identifying the occurrence of corporate default events reached 92.97%, and it was able to Early warning of credit risk crises for companies 1 to 8 quarters in advance can better predict the probability of corporate default events and provide decision-making support for the prevention of credit risks for manufacturing companies [2]. Tingting Wu and others used extreme quantile regression methods to measure the contribution of China’s 33 listed financial institutions to systemic financial risks. The results showed that the return on assets of Chinese financial institutions has obvious non-normal distribution
characteristics, and extreme quantile regression. The method can more accurately measure the linkage of tail risk [3]. Based on the financial risk detection data of the 5 administrative counties and districts under the jurisdiction of S city, Gang He obtained the comprehensive evaluation value of the financial risk of the three counties and two districts within the jurisdiction of S city according to the quantitative equation and entropy weight method. The results were 48.98, 61.58, 18.46, 54.66, 133.94, found that the county-level regional risk is much higher than the municipal-level region, and got some risk control enlightenment [4].

However, the above-mentioned research on financial risk management only considers systemic risks within a region, and it is difficult to solve risk problems in practical applications. This article will establish a customer pre-loan risk assessment model based on a decision tree algorithm to accurately control financial risks to individual customers. By making reasonable pre-judgments, it will provide pre-loan references for bank loans, credit card installments, online loans, etc., and provide credit risk Prevent and provide decision support.

2. Algorithm principle
The decision tree algorithm is a method of approximating the value of a discrete function. It is a typical classification method. It first processes the data, uses induction algorithms to generate readable rules and decision trees, and then uses decisions to analyze new data. In essence, a decision tree is a process of classifying data through a series of rules.

2.1. The structure of the decision tree
The decision tree is logically represented in the form of a tree, including nodes and edges. In general, a decision tree contains a root node, several internal nodes and several leaf nodes. The root node includes the complete set of samples. The path from the root node to each leaf node corresponds to a discriminative test sequence. Internal nodes represent a feature and attribute. Each internal node is a judgment condition, and includes a data set, a set of data that meets all conditions from the root node to the node. According to the result of the attribute test of the internal node, the set of data corresponding to the internal node is divided into two or more child nodes. The leaf node represents a class, corresponding to the decision result. The leaf node is the final category. If the data is contained in the leaf node, it belongs to this category. As shown in Figure 1, the circles and boxes represent internal nodes and leaf nodes, respectively.

![Figure 1. The structure of the decision tree.](image)

2.2. Node selection based on Gini coefficient as the criterion
After establishing the model structure of the decision tree, this article will introduce the Gini coefficient as a reference basis for selecting decision tree nodes. The Gini coefficient was first used in economics and was mainly used to measure the fairness of income distribution. In the decision tree algorithm, the Gini index is used to measure the impurity or uncertainty of the data, and the Gini coefficient is used to determine the optimal dichotomous value of the categorical variables.

In the problem, suppose there are \( k \) classes and the probability that the sample point belongs to the \( k \)-th class is \( P_k \), then the Gini coefficient of the probability distribution is defined as:

\[
Gini(P) = \sum_{k=1}^{k} P_k (1 - P_k) = 1 - \sum_{k=1}^{k} P_k^2
\]

(1)

If the sample set \( D \) is divided into two parts \( D_1 \) and \( D_2 \) according to a certain feature \( A \), then under the condition of feature \( A \), the Gini coefficient of set \( D \) is defined as:

\[
Gini(D, A) = \frac{D_1}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2)
\]

(2)

Gini(D,A) represents the uncertainty of the data set \( D \) in different groups of feature \( A \). The greater the Gini coefficient value, the greater the uncertainty of the sample set, which is similar to the probability of entropy. This paper uses the gini coefficient to determine the optimal segmentation point of the feature, while ensuring the minimum value of the segmentation point coefficient to construct the optimal decision tree.

2.3. Algorithm flow

The decision tree algorithm constructs a decision tree to discover the classification rules contained in the data. Constructing a high-precision, small-scale decision tree is the core content of the decision tree algorithm. The decision tree construction can be done in two steps. The first step is to generate a decision tree: the process of generating a decision tree from the training sample set. In general, the training sample data set is a data set with a history and a certain degree of comprehensiveness according to actual needs, and is used for data analysis and processing. The second step, the pruning of the decision tree: The pruning of the decision tree is the process of checking, correcting and modifying the decision tree generated in the previous stage, mainly using the data in the new sample data set (which becomes the test data set) Verify the preliminary rules generated during the decision tree generation process, and prune those branches that affect the accuracy of the pre-balance. The algorithm flow of generating the decision tree is shown in Figure 2.
3. Experimental design

3.1. Sample and feature selection
The pre-loan survey is mainly divided into two types: personal loans and loans from enterprises and institutions. Compared with corporate loans, personal loans have a more intuitive pre-loan survey, focusing on checking the customer’s previous credit records, followed by the customer’s education, experience, management ability, performance, social relations, hobbies, living habits, and character, Age, health, etc. This article will select the personal information of 3422 customers for the construction and verification of the decision tree model, and extract the annual income as the output criterion of whether to lend. When the annual income is greater than 50k, the loan is agreed, and when the annual income is less than 50k, the loan is not granted. The independent and dependent variables designed are: age X1, gender X2, job category X3, education time X4, average weekly working hours X5, and annual income Y. Table 1 shows the design variables and their assignment descriptions.

Table 1. Variable assignment.

| Factor         | Variable parameter | Assignment description |
|----------------|--------------------|------------------------|
| Age            | X1                 | “>35”=1;“<=35”=0       |
| Sex            | X2                 | Male=1;Female=0        |
| Work class     | X3                 | Private=1;Civil servant=2;Freelancer=3 |
| Education-num  | X4                 | “>14yr”=1;“<=14yr”=0   |
| Hours-per-week | X5                 | “>40h”=1;“<=40h”=0     |
| Income         | Y                  | “>50K”=1;“<=50K”=0     |
3.2. Result analysis

The SPSS 26 software is used for statistical analysis of the sample data, and the CRT algorithm is selected as the growth method of the decision tree. CRT is an algorithm that divides decision tree nodes based on Gini coefficient. This paper chooses two model verification methods to calculate the risk of the decision tree model, namely the cross-validation method and the random allocation method. The random allocation method selects 75% of the samples as the training set and 25% of the samples as the test set. The output results obtained are shown in Table 2 and Table 3.

Table 2. Risk calculated by cross-validation.

| Method          | Estimate | Standard error |
|-----------------|----------|----------------|
| Resubstitution  | 0.188    | 0.007          |
| Cross-validation| 0.188    | 0.007          |

The results of the cross-validation method show that the estimated value of the decision tree has 18.8% sample decision bias when re-substituting, and the estimated value of the decision tree has 18.8% sample decision bias during cross-validation. It can be found that the decision tree has high accuracy in making choices under a large number of data samples and has a high reference value.

Table 3. Risk calculated by random allocation.

| Sample   | Estimate | Standard error |
|----------|----------|----------------|
| Training set | 0.187    | 0.008          |
| Test     | 0.193    | 0.014          |

The results of the random verification method show that the estimated value of the decision tree of the training set samples has a sample decision bias of 18.7%, and the estimated value of the decision tree of the test set samples has a sample decision bias of 19.3%. It can also be found that the decision tree has high accuracy in making choices under a large number of data samples and has a high reference value.

According to the Gini value of each factor, the degree of direct influence of each factor on the evaluation result can be obtained, in order of importance from low to high: job category, gender, age, weekly working hours, marital status, and education time, as shown in the figure 3 shown.

Figure 3. Importance of independent variables.
As shown in Figure 3, it can be seen that the time of education is the most important, which is consistent with the reality and objective performance. The longer the education time, the higher the academic qualifications obtained. The academic qualifications are indispensable as the first condition for high-paying jobs. The second is marital status. When people’s income reaches a certain amount, they will consider getting married while maintaining the basic life of a family. Working hours can also explain the income problem to a certain extent. Under the same type of work, the longer the working hours, the longer they will be able to get more income, but different jobs will have greater differences, so the importance of working hours is relatively low. From the point of view of importance, gender, age, and job type have a low impact on income, and will be removed in the pruning process of the decision tree construction to avoid overfitting of the model.

4. Conclusions
This paper uses the decision tree algorithm to conduct a research on the risk assessment of customers before loan. After the model verification of the cross-validation method and the random validation method, the results show that the risk estimation accuracy of the decision tree algorithm reaches 81.2% and 83.6%. This model can become Effective reference method for pre-lending risk assessment. With the further development of the modern economy, both individuals and enterprises have begun to transform and adjust, but the ultimate goal is to maximize the benefits. The emergence of financial credit has brought convenience to the financial investment of individuals and enterprises, and it also brings certain risks. Some risks exist and are unavoidable, and some risks are potential and can be avoided. Perform corresponding risk assessment. It is of great significance to the stable development of the economy.

References
[1] Tingting Xie, Yi Li. Study on Regional Financial Risk Measurement in Xinjiang Based on Entropy Method [J]. Journal of Shihezi University (Philosophy and Social Sciences Edition), 2020, (No. 5).
[2] Yuanyuan Huo, Tianyi Yao, Jiang Li. Research on Credit Risk Measurement of Chinese Manufacturing Enterprises Based on Probit Model[J]. Forecasting, 2019, (No. 4).
[3] Tingting Wu, Fei Hua, Shiyin Jiang. Quantitative Research on the Contribution of Systemic Financial Risks of Chinese Financial Institutions—CoVaR Model Based on Extreme Quantile Regression [J]. Jiangxi Social Sciences, 2020, (9th).
[4] Gang He. Research on the Quantitative Evaluation and Distribution of Regional Financial Risks—Taking S City as an Example [J]. Finance, 2015, (No. 9).
[5] Xinya Wang. Research on big data risk control model based on GBDT algorithm [J]. Journal of Zhengzhou University of Aeronautics, 2020, (No. 5).