ISSN 2616-5902 Vol. 4, Issue 16: 99-102, DOI: 10.25236/AJBM.2022.041617

Small and Medium Enterprise Credit Risk Assessment

Zunhui Li

College of Science, Shihezi University, Shihezi, Xinjiang, 832003, China

Abstract: *In recent years, the state encourages the development of small and medium-sized enterprises and promotes economic growth, but small and medium-sized enterprises lack real assets and cash, so they need to borrow from banks. Most SMEs are small and have unstable supply chains, making them prone to bankruptcy, and leaving banks unable to recover their loans. In this paper, the random forest model is used to determine the credit risk strategy, and the SVM support vector machine is used for prediction and verification. The results show that the credit risk strategy has high accuracy and can help banks avoid credit risk.*

Keywords: *small and medium enterprises, random forest, SVM*

1. Introduction

Since the beginning of the century, small and medium-sized enterprises in our country have sprung up like mushrooms after a rain and become the main force in my country's economic development. Policy Support. However, due to the relatively small scale of small and medium-sized enterprises and the lack of real assets for mortgage loans, banks should comprehensively consider the invoice transaction information of such enterprises, whether the supply and demand relationship is stable, credit history, etc., to determine whether to issue loans.

At present, in the aspect of bank credit risk prediction, models such as random forest and SVM vector machine are the main research methods. Wang Zhuangzhi [1] and others introduced the SVM model to predict the credit risk of small and medium-sized enterprises and solved the practical problems such as small sample size, non-linearity, and many dimensions in the credit risk measurement of small and medium-sized enterprises. Yuan Yisheng [2] used multiple decision trees as meta-classifiers based on the random forest algorithm of corporate tax and other characteristics, which can improve the stability and accuracy of corporate evaluation and classification. Pu Zhao [3] analyzed the importance of the indicators through the random forest model and then used the support vector machine to establish the model to determine the summation function, and obtained the most suitable evaluation method for the credit risk of listed companies. Based on the previous research, this paper uses a random forest decision tree and an SVM support vector machine to determine the credit strategy for small and medium-sized enterprises.

2. Data description

2.1 Data source

The data in this study are extracted from the 2020 “Higher Education Society Cup” National Mathematical Contest in Modeling for Undergraduates, which includes 123 small and medium-sized enterprises with credit rating records and 302 unrated small and medium-sized enterprises. It contains all the transaction records of each company from 2017 to 2019, with a total of more than 100,000 transaction records.

In the fact, the main factors that affect the stability of enterprises include annual profit, bad debt amount and bad debt growth rate, and upstream and downstream supply chains. Whether the supply is stable is mainly reflected in the number of purchases and shipments.

ISSN 2616-5902 Vol. 4, Issue 16: 99-102, DOI: 10.25236/AJBM.2022.041617

2.2 Data Description

Table 1: Metric Description

Indicator name	symbol
Total input value tax	Hj
Total sales tax	Hx
profit growth rate	Lz
Input void invoice amount	J_i
Sales void invoice amount	X_i
The growth rate of the input voided invoice amount	L_{J_i}
The growth rate of sales voided invoices	L_{X_i}

Among them, $i = 2017$、2018、2019represents the year

$$Lz = \frac{Hx-Hj}{Hx} \tag{1}$$

$$Lj_i = \frac{J_i-J_{i-1}}{J_{i-1}} \tag{2}$$

$$Lx_i = \frac{X_i-X_{i-1}}{X_{i-1}} \tag{3}$$

2.3 Data preprocessing

Before solving the problem and establishing the model, data preprocessing should be carried out first. The main purpose is to unify or eliminate the unit dimension to prevent errors in the results when solving the model; the second is to sort out the chaotic data and remove Some meaningless data bars, and some missing values are rounded off to get the indicators and data needed to solve the problem , as shown in Table 1.

3. Determine the main indicators that affect corporate credit

There are many methods for evaluating the importance of indicators, such as the AHP analysis hierarchy process, principal component analysis, etc., but relatively speaking, there is no random forest decision tree widely used.

3.1 Establish a credit strategy prediction index system

Random Forest (RF) is a "combination classification algorithm based on decision tree algorithm. It resamples in the training set to obtain a considerable number of samples, and then randomly selects some indicators from the original indicators in each sample to form a decision tree, which can be obtained. Different classification results, and finally each decision tree votes on different results to select the optimal solution." Can effectively avoid overfitting.

First, the data of 123 enterprises were screened, the enterprises with incomplete information were eliminated, and the data of the remaining 63 enterprises were processed. The original data were processed to obtain preset index data, which was used as a training set sample, and 13 indicators were determined from the CCP. They are: the growth rate of the net profit of each company, the total amount of purchases, the total amount of shipments, the number of invalid invoices for incoming and outgoing items in each year in 2018 and 2019, and the growth rate of the number of invalid invoices, in addition to 2017 Input voided invoice amount and output voided invoice amount. The importance weight of the indicator is calculated by the out-of-bag error rate (OOB). The principle is to take the data of our 63 companies as a sample and randomly divide it into two parts: the training set and the verification set. Suppose the name of an indicator is T, and the T indicator is in the first The formula for calculating the importance weight of IMT in k trees is:

$$IMT = \frac{\sum_{i=1}^{n_k} I(B_i=B_i')}{m_k} - \frac{\sum_{i=1}^{n_k} I(B_i=B_i'')}{m_k} \tag{4}$$

IMT represents the importance weight of the T index, n_k represents the number of observations

ISSN 2616-5902 Vol. 4, Issue 16: 99-102, DOI: 10.25236/AJBM.2022.041617

outside the kth tree bag, B_i represents the actual value of the ith observation, B_i' is the predicted value before replacement, and B_i'' is the predicted value after replacement. Takes 1 when the values in the function parentheses of (4) are equal, otherwise takes 0. In this paper, 123 companies with credit ratings are divided into 0-1 events according to the existing ABCD grades. A, B, and C grades can issue loans as 1, and D grades cannot issue loans as 0.

3.2 Weight Ranking

During the establishment of the random forest model, 1000 decision trees were selected, and the ratio of the training set and test set was 2:8 for processing. Determine the level of importance of metrics that affect lending strategy. The results are shown in Figure 1 below.

Get the final index importance ranking:

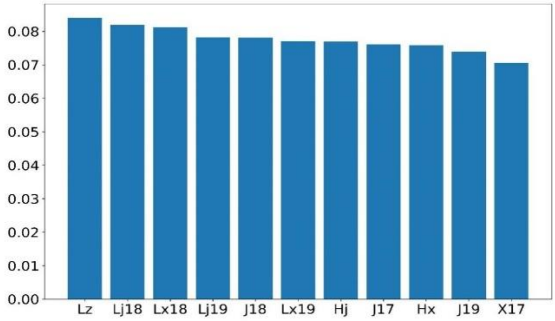

Figure 1: Indicator importance

4. SVM Model Prediction Model

4.1 Establish SVM Model

SVM Support Vector Machine is a machine learning method mainly used for regression and classification [4].

Depending on the selected data, vector machines can be roughly divided into linear and nonlinear support vector machines. Different types of data use different optimization methods. The kernel function is a key step of the support vector machine. An appropriate kernel function can simplify the operation. Otherwise, it may increase the amount of calculation. Therefore, it is necessary to select an appropriate kernel function.

Five common kernel functions are linear kernel function, Gaussian kernel function, polynomial kernel function, Laplace kernel function, and Sigmoid kernel function.

Among them, the Gaussian kernel function is the most widely used, and the expression:

$$K(x_i, x_j) = \exp\left(\frac{-|x_i - x_j|^2}{2\sigma^2}\right) \tag{5}$$

In this paper, the Gaussian kernel function is selected as the kernel function of the model, because the parameters of the Gaussian kernel function are relatively small, which is convenient for machine calculation [5]. And choose the FM index of the confusion matrix for evaluation, and select the top six indicators of the importance of the random forest model as the evaluation index of the SVM model.

After continuously adjusting the model parameters, it is found that when selected $c = 53, \gamma = 0.0013$, the model accuracy is better, and the cross-validation accuracy reaches 70%.

5. Summary

Small and medium-sized enterprises have a low starting point, and lending to banks is a necessary means. However, in recent years, there are many small and medium-sized enterprises due to unstable

ISSN 2616-5902 Vol. 4, Issue 16: 99-102, DOI: 10.25236/AJBM.2022.041617

supply chains or the inability to produce normally due to the impact of emergencies, resulting in the bankruptcy of the enterprises and the inability of banks to recover the loans issued. Therefore, the establishment of a loan risk assessment model for small and medium-sized enterprises is very important. The important indicators of risk assessment of small and medium-sized enterprises are screened through random forest decision trees, and the risk of small and medium-sized enterprises is assessed based on the SVM support vector machine model. Through actual comparison, the accuracy rate can reach 70%. This shows that the credit risk assessment model for small and medium-sized enterprises established in this paper has strong adaptability and reality. In addition, due to the limitations of the data itself, it is not enough to be applied to most enterprises. In the future, a larger amount of data should be collected to further optimize the model.

References

[1] Zeng Jianghong, Wang Zhuangzhi, Cui Xiaoyun. Research on individual credit risk measurement of collective bond financing of small and medium-sized enterprises based on SVM [J]. Journal of Central South University (Social Science Edition), 2013, 19(02):8-11+ 19.

[2] Yuan Yisheng. Credit risk assessment of small and medium-sized enterprises based on random forest algorithm [D]. Shandong University, 2021. DOI: 10.27272/d.cnki.gshdu.2021.003222.

[3] Pu Zhao. Based on RF and integrated SVM Research on Green Credit Risk Assessment Model of Listed Companies [D]. Shanghai Normal University, 2019.

[4] Zhang Xiaoli, Wang Pin, Xiong Chao, Ge Jianjun. Operation Prediction of Domestic Listed Companies Based on Support Vector Machines [J]. Green Technology, 2022, 24(09):251-254. DOI: 10.16663/j.cnki.lskj.2022.09.064.

[5] Hu Fangfeng. Construction and Research of Risk Control Model Based on Credit Card Transaction Data [D]. East China Normal University, 2021. DOI: 10.27149/d.cnki.ghdsu.2021.002150.