DOI: 10.23977/ferm.2022.050405
ISSN 2523-2576 Vol. 5 Num. 4

Analysis of Corporate Credit Risk Based on Random Forest and TOPSIS Models

Wenshu Zhao*, Jie Hou, Qili Ran

College of Mathematics and Physics, Beijing University of Chemical Technology, Beijing 100029, China

Keywords: credit strategy, credit risk, random forest, TOPSIS

Abstract: Based on the assumption that the input invoices and output invoices can fully reflect the overall income and profitability of enterprises. This paper firstly mines and enriches the data information in the annex, and then establishes the risk evaluation index system under the problem scenario by combining the risk evaluation indexes commonly used within banks; then uses the random forest model to derive the valuation of enterprise default probability and the weight coefficients of various indexes, and on this basis achieves the quantitative analysis of enterprise credit risk by TOPSIS method.

1. Introduction

At this stage, the rapid development and expansion of the domestic financial market, coupled with the transformation of China's economic structure and industrial restructuring [1], have increased the credit risk of each bank to a large extent. It directly affects the competitiveness of the financial sector and the efficiency of economic development, which further affects social stability [2].Therefore, it is crucial to strengthen banks' credit risk analysis and credit strategy formulation. For MSMEs, banks make an assessment of credit risk based on the strength and reputation of the enterprise, and then determine the credit strategy based on credit risk and other factors [3]. Therefore, this paper will quantify the credit risk of 123 firms and give the bank's credit strategy for these firms when the total annual credit is fixed.

2. Model construction

2.1 Model analysis

Because credit risk refers to the possibility that the borrower will not be able to return the principal and interest on time and the bank will suffer losses, there is a strong correlation between credit risk and the probability of default, and it is necessary to value the probability of default before quantitatively analyzing credit risk. We combine the risk evaluation indicators with the random forest model to analyze the credit default of 123 enterprises, and solve for the default probability of 123 enterprises and the weight coefficients of each evaluation indicator. The quantitative analysis results of each enterprise's credit risk, i.e., the enterprise credit score, are then obtained by the TOPSIS integrated evaluation method with weights. Finally, we classify and quantify the size of enterprises according to their size, establish quantitative indicators of loan amount and quantitative indicators of interest rate respectively, and further analyze to get bank lending strategies.

2.2 Calculation of default probability and default factor weights

2.2.1 Overview of random forest methods

Random forest is an algorithm for optimizing decisions by multiple decision trees, in which the decision tree CART pruning method is performed as follows.

Step 1: set up $k = 0$, $T = T_0$, $\alpha = +\infty$.

Step 2: Bottom-up computation of C (T_t) for each internal node t, $|T_t|$ and

$$g(t) = \frac{C(t) - C(T_t)}{|T_t| - 1} \tag{1}$$

$$\alpha = \min(\alpha, g(t)) \tag{2}$$

Here, T_t denotes the subtree with t as the root node, C (T_t) is the prediction error on the training data, and $|T_t|$ is the number of leaf nodes of T_t.

Step 3: Visit the internal node t top-down, prune it if g(t) = α, and decide the class of the leaf node t by majority voting to obtain the tree T.

Step 4: Let $k = k + 1$, $\alpha_k = \alpha$, $T_k = T$, if T is not a tree composed of root nodes alone, then go back to Step 3.

Step 5: The optimal subtree T_α is selected in the subtree sequence using cross-validation method.

The random forest implementation process is shown in Figure 1.

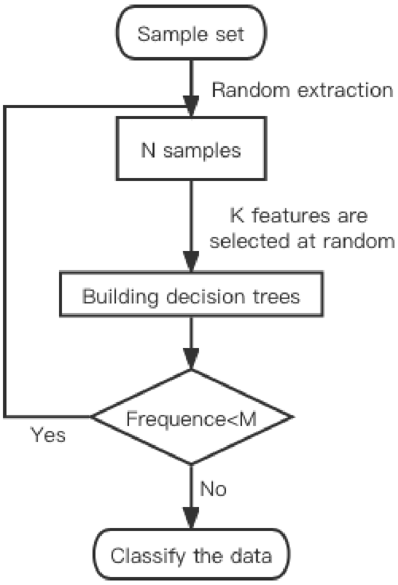

Figure 1 Random Forest Flow Chart

2.2.2 Random forest creation

Based on the above pre-processing of data and the selection of default analysis indexes, the sample set is divided into two major parts: the model construction sample set and the model application sample set, each sample set contains 123 samples and 8 feature indexes, in which the model construction sample set is divided into the training set and the test set according to the ratio of 7:3, with no default as 0 and default as 1.After preliminary screening, there are 96 positive samples (no default) and 27 negative samples, which is clearly a balanced sample. In this paper, the sklearn library included in Python is used to construct the corresponding code to train the random forest model, so as to obtain the default probability, default situation (default probability >0.5 is considered as default) and the importance of 8 characteristics of the 123 enterprises[4]. In this model, the rfc interface score is calculated to obtain a model accuracy of 0.973 and a ROC score of 0.991, which is a high model accuracy. the ROC curve is an important index used to measure the model effectiveness, and the curves are plotted below with FPR (false positive class rate) as the horizontal coordinate and TPR (true class rate) as the vertical coordinate.

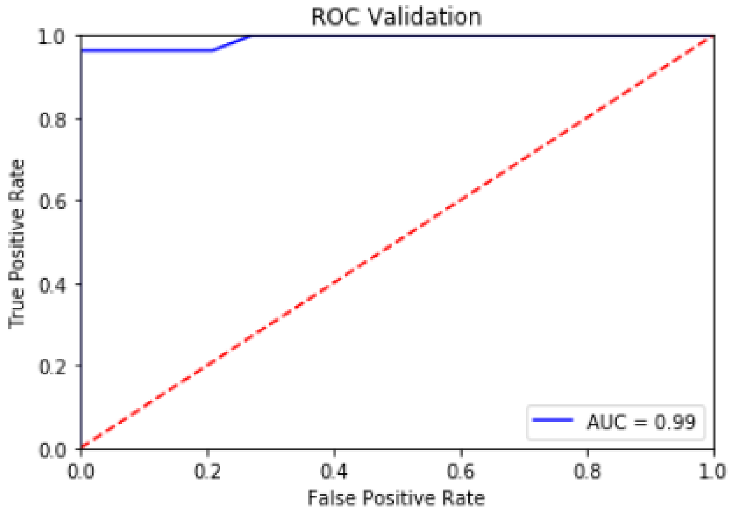

Figure 2 ROC graph

From the Figure 2, it can be seen that the value of the vertical coordinate is larger when the horizontal coordinate is fixed, i.e., the model is correct. Second, the area under the ROC curve is noted as AUC, and the closer the AUC is to 1, the better the model works. The random forest model gives a better AUC of 0.99.

2.2.3 Cross-validation and superparameter tuning for random forests

Cross-validation is a method used to prevent over-fitting caused by overly complex models. In this paper, we call the cross score function in Python to cross-validate the model, and the validation score is 0.976, which is a good fit for the model.

To perform hyperparameter tuning, this paper selects the optimal hyperparameters by grid search Grid Search CV. By further search, the maximum number of hyperparameter decision trees estimators is 63, the maximum depth of decision trees max-depth is 10, the type weight parameter weight is balanced, and the criterion of feature evaluation is information gain entropy. The ROC score of the hyperparameter-tuned classifier is 0.994, which is an improvement over the default parameter score of 0.991. Therefore, the new classifier is used to re-analyze the probability of default, default situation (default probability >0.5 is considered as default), and the importance of the 8 feature indicators for 123 firms, As shown in the following Figure 3 and Table 1.

Table 1 Characteristic index weights

Number	Characteristic indicators	Weights
0	Income annual average	0.109185
1	Sales annual average	0.261278
2	Average annual growth rate of profit	0.077159
3	Average annual growth rate of revenue	0.051584
4	Invalid transaction rate	0.037794
5	Pinnacle fluctuations (standard deviation)	0.061926
6	Income fluctuation (standard deviation)	0.065011
7	Credit Rating	0.336063

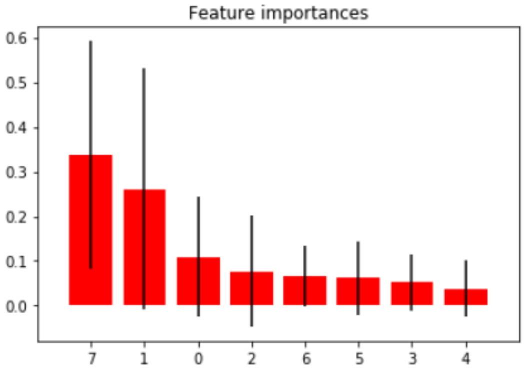

Figure 3 Comparison of characteristic index weights

2.3 Quantitative analysis of credit risk

Given that the number of analyzed enterprises is 123 and the number of evaluation indicators is 8, if the hierarchical analysis method is used, too many decision levels of evaluation may lead to too much difference between the judgment matrix and the consistency matrix, and at the same time to ensure the objectivity of the model and clear quantitative analysis results, after comprehensive consideration we use the TOPSIS method with weights[5]. Firstly, the original data matrix is forwarded to very large indicators, and we use max - x to convert very small indicators such as the standard deviation of input price and tax, the standard deviation of output price and tax, and the default rate to very large indicators. Afterwards, the normalization matrix is normalized to eliminate the influence of the magnitudes between different indicators, and for each element in the normalization matrix Z there are:

$$z_{ij} = x_{ij} / \sqrt{\sum_{i=1}^{n} x_{ij}^2} \tag{3}$$

Later in the normalized matrix

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nn} \end{bmatrix} \tag{4}$$

Define maximum value

$Z^{+}=(Z_1^{+}\ ,Z_2^{+},\cdots Z_m^{+})=(\max\{z_{11},z_{21},\ldots,z_{n1}\},\max\{z_{12},z_{22},\ldots,z_{n2}\},\ldots,\max\{z_{1m},z_{2m},\ldots,z_{nm}\})$

Define minimum value

$Z\text{-}=(Z_1^{-},Z_2^{-},\cdots Z_m^{-})=(\min\{z_{11},z_{21},\ldots,z_{n1}\},\min\{z_{12},z_{22},\ldots,z_{n2}\},\ldots,\min\{z_{1m},z_{2m},\ldots,z_{nm}\})$

Define the distance of the (i=1,2,...,n) evaluation object from the maximum value

$$D_i^{+} = \sqrt{\sum_{j=1}^{m} \omega_j \left(Z_j^{+} - z_{ij}\right)^2} \qquad (5)$$

Distance of the (i = 1, 2, … , n) evaluation object from the minimum value

$$D_i^{-} = \sqrt{\sum_{j=1}^{m} \omega_j \left(Z_j^{-} - z_{ij}\right)^2} \qquad (6)$$

By calculate the overall evaluation score of each enterprise, finally, normalization is performed. The normalized evaluation score is used as the result of quantitative analysis of credit risk. Since the evaluation is done with very large indicators, the closer the normalized evaluation score is to 1, the lower the credit risk to the company or the higher the credit rating of the company. Therefore, we refer to the quantitative analysis results as corporate credit score.

3. Credit strategy development

3.1 Credit line strategy development

In general, the loan amount is mainly related to the size of the business that reflects the annual revenue, annual profit, annual consumption of funds and the risk of default of the business. For example, if a company is small and needs very little capital, but has a low risk rating and a high credit rating, the traditional strategy may result in a large loan, resulting in waste. To prevent this from happening, we plan to introduce a quantification of firm size. Therefore, we further quantify the size as well as the corporate credit of 123 firms. According to the above, the quantification of enterprise credit directly uses enterprise credit scores. We classify 123 enterprises by industry according to the *Statistical Classification of Large, Small, Medium, and Micro Enterprises (2017)*, and on this basis, we classify enterprises into medium, small, and micro categories according to different industry standards. Finally, the size T of small, medium, and micro enterprises was quantified as 10, 5, and 1, respectively, based on the proportion of annual revenue that these three categories of size enterprises have, as shown in the Figure 4 below.

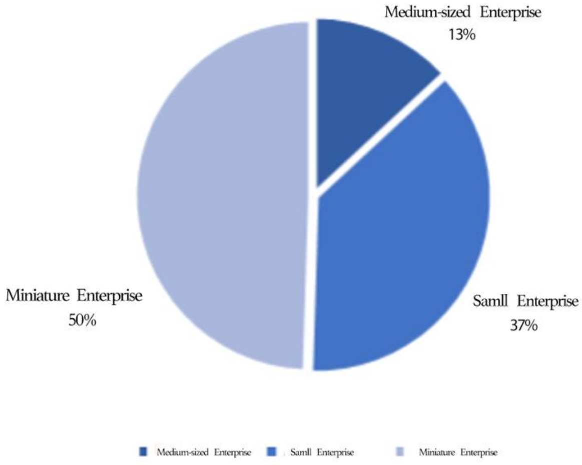

Figure 4 Percentage of small, medium and micro enterprises

The larger the size of an enterprise, the larger the loan amount it needs; the higher the credit of an enterprise, the larger the loan amount the bank can provide. By combining the enterprise size T and the enterprise credit score R, we establish the quantitative index of the enterprise loan amount La as follows

After filtering out the companies with credit rating of D, the La distribution of 123 companies' credit score is shown in the Figure 5 below.

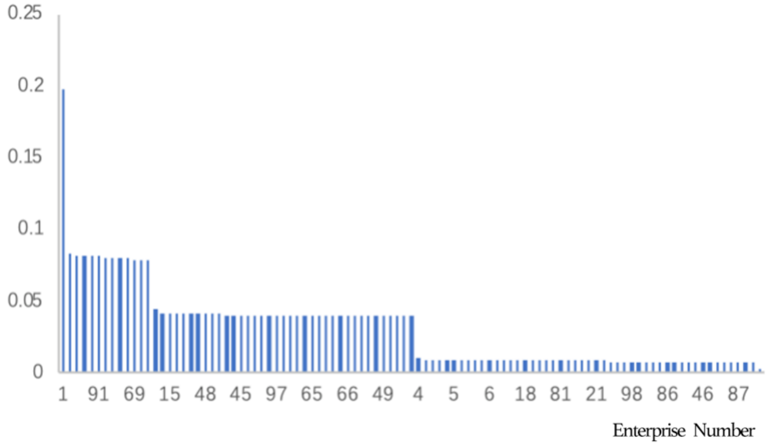

Figure 5 Quota score La distribution chart

To prevent these outliers from influencing the calculation of the loan amount, we assign a line of credit of $1 million to E1 with a very high line of credit score and a line of credit of $100,000 to E96 with a very low line of credit score. Then, the highest remaining score is marked as 1 million yuan and the lowest remaining score is marked as 100,000 yuan, and the linear function is derived from these two endpoints, as shown in the Figure 6 below.

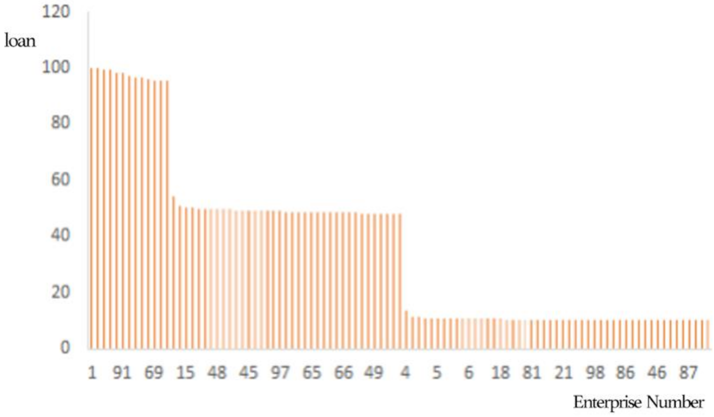

Figure 6 Loan Distribution Map

3.2 Credit rate strategy development

By analyzing the study, we found that firms with different credit ratings have different churn rates for different loan APRs. For banks, the higher the APR, the higher the interest earned, but the higher the churn rate, the lower the interest earned. Therefore, we maximize the loan interest rate Y and the customer churn rate L, and establish the quantitative indicators of corporate loan interest rate It as follows

$$It=Y(1-L) \tag{7}$$

The relationship between the quantitative interest rate index It and the annual interest rate Y is also plotted.

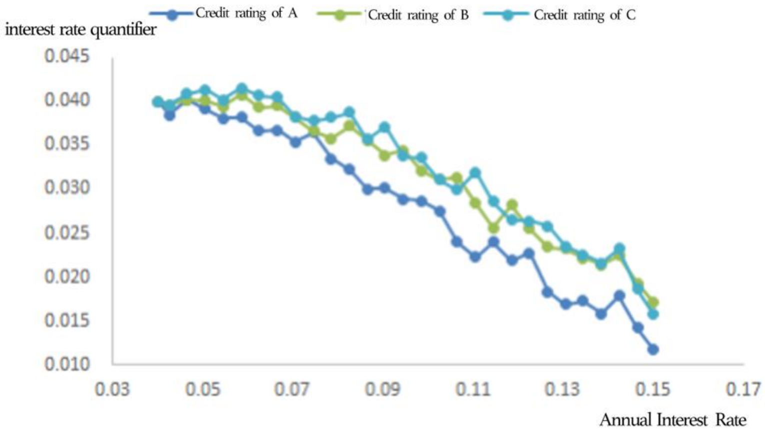

Figure 7 The relationship between the interest rate quantifier It and the annual interest rate Y

From the Figure 7, it can be seen that companies with a credit rating of A have the highest interest rate earned by the bank for a loan with an annual interest rate of 0.0465, while companies with credit ratings of B and C have the highest interest rate earned by the bank for a loan with an annual interest rate of 0.585. Therefore, the annual interest rates for loans to companies with credit ratings of A, B, and C are taken to be 0.0465, 0.0585, and 0.0585, respectively.

4. Conclusion

4.1 Advantages of the model

(1) When constructing the risk evaluation system, we fully combined the risk evaluation indexes commonly used by banks and the extended data information, and the accuracy of the model in the subsequent default probability valuation reached over 97%, reflecting the scientificity of the index selection and the rationality of the evaluation system.

(2) For the quantitative analysis of credit risk, we adopt the TOPSIS method with weights. The advantage of this method is that it can make full use of data information and accurately reflect the gaps among evaluation objects, while introducing indicator weights generated by random forest, which avoids the subjectivity of weights and improves the objective validity of the model.

4.2 Disadvantages of the model

(1) For the quantitative analysis of enterprise size, although reference is made to each *Statistical Classification of Large, Small, Medium and Micro Enterprises (2017)*, the small differences in the division of enterprise size among various industries are ignored in order to reduce the complexity of the model, making the quantitative model of size somewhat crude and subjective.

(2) Only the relationship between the annual interest rate and the customer churn rate is considered when setting the bank's annual interest rate on credit, making only one annual interest rate for each credit rating, without considering the differences in the amount borrowed by different companies, which may lead to differences between the bank's interest rate and the theoretical optimal value.

References

[1] Cheng, Hao. Study on Credit Risk Management of Small and Micro Enterprises in XT Bank [D]. Hebei University, 2020.GB 38755-2019, Guidelines for security and stability of power systems [S].
[2] Ulantuya. Study on the improvement program of risk management of micro and small enterprise credit business of Baoshang Bank [D]. Lanzhou University, 2019.
[3] Ren Zanbin, Liu Bingjie. Credit risks and management strategies of commercial banks [J]. Times Finance, 2017(33):54+66.
[4] Xia Yuchi. Support vector machine based credit evaluation model for small and medium enterprises and application research [D]. Central South University, 2013.
[5] Qin Fayan. An empirical study on the credit rating of small enterprises by commercial banks in China: the case of Yichang City, Hubei Province [J].Time Finance, 2009(01):52-54.