2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI) | 978-1-7281-8115-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICETCI55101.2022.9832083

Credit Risk Intelligent Assessment Model Based on Machine Learning

Haofeng Shi[1, *, †], Zhihua Wang[1, †], Xinming Wang[2, †]
[1]School of Economics, Northwest Minzu University, Lanzhou, Gansu, 730124, China
[2]College of Educational Science and Technology, Northwest Minzu University, Lanzhou, Gansu, 730124, China
*Corresponding author: shihaofeng0106@163.com
†These authors contributed equally.

***Abstract*—In today's Chinese economic system, market system, and development goals, MSMEs play an essential role. The availability of capital is a necessary condition for the development of MSMEs. However, in China, MSMEs still cannot effectively eliminate the development dilemma that financing is difficult and expensive. The deep-seated reason lies in the imperfect credit evaluation system of the banking industry for MSMEs. The purpose of this paper is to establish a mathematical model for credit risk evaluation of MSMEs and to establish a new bank credit evaluation strategy to support and back up MSMEs of different sizes accurately.**

Keywords—Credit risk assessment, XGBOOST, Analytic hierarchy process, Machine learning model, Mpai

I. INTRODUCTION

As economic globalization develops, Chinese MSMEs have seen rapid development. However, with the further development of MSMEs, the shortage of capital often becomes a significant factor limiting the further development of MSMEs. Due to the limitation of the size of MSMEs, their risk tolerance is weak, leading to further increase of enterprise credit risk and the difficulty of credit assessment by commercial banks. The purpose of this paper is to establish a mathematical model for assessing the credit risk of MSMEs and to establish new bank credit strategies to support and back up MSMEs of different sizes accurately. In this paper, an evaluation model is established by combining the machine learning model XGBoost algorithm with AHP, and the combined model is used to construct and analyze the model to quantify further and evaluate the credit risks of MSMEs. While ensuring the interests of commercial banks, more MSMEs are provided with loans to increase the amount of loanable capital of the whole society. Promote the further development of MSMEs and promote social progress.

II. MODEL DATA DESCRIPTION

The data sources in this paper are authentic. The model in this paper is based on the 2020 China National Mathematical Contest in Modeling for College Students, and all the data and assumptions are from this contest.

III. MODEL ESTABLISHMENT AND SOLUTION

A. Data preprocessing and essential model establishment

First of all, the data in the attachment shall be converted to the data available in this paper, and a single company shall process the data required in this paper, and the outlier shall be cleared. The outlier test method is 3sigma outlier identification, and the data in the attachment shall be defined in Table I below:

TABLE I DATA DEFINITIONS

Define the name	The original name
Total of input bills	Total of all input bills of a single enterprise(Excluding negative invoices and invalid invoices)
Total of written bills	Total of all cancellation bills of a single enterprise(Excluding negative invoices and invalid invoices)
Number of valid invoices	The total invoice - Number of invalid invoices - Number of negative invoice
The company together	Amount of the cancellation invoice - Amount of input invoice(Excluding negative invoices and invalid invoices)
Bad single rate	(Number of invalid invoices + Number of negative invoice)/ The total invoice
Bad single amount	Negative invoice amount + Invalid invoice amount
Credit risk ranking	Company strength*Weight vector + The stability of supply and demand*Weight vector + Customer turnover rate*Weight vector
Note: All the data above are based on individual enterprises	

After thoroughly preparing the data, this paper classifies the eigenvalues similar to machine learning, screens the auxiliary data, establishes a mathematical model, and substitutes the effective indexes into the machine learning model xgboost regression for training. By adjusting the number of decision trees to 100, and substitute the above processed quantitative data as the primary reference index into mpai software for training. The xgboost objective function is defined as:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

After rewriting a series of formulas, the optimal formula of W and the objective function is obtained by using the vertex formula:

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (2)$$

According to the final data processing results, this paper classifies the eigenvalues similar to machine learning and substitutes the effective indicators into the machine learning model xgboost regression for training. Adjust the number of

decision trees to 100, take the above processed quantitative data as the main reference index, train in mpai software, and process the data successfully.

B. Modeling steps

(1) Establish analytic hierarchy process structure model

This paper selects three factors affecting credit risk: company strength (difference between output bills and input bills), credit value, stable supply and demand (proportion of obsolete bills in all bills), and customer churn rate is divided into three layers, with the target layer at the top, the criterion layer or index layer in the middle, and the scheme layer at the bottom, as shown in Figure 1:

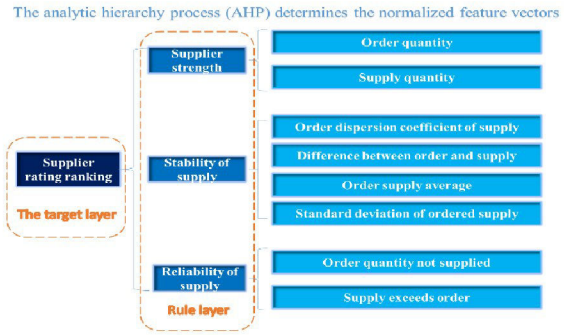

Fig. 1 Hierarchy analysis structure model

(2) Construct pairwise comparison matrix

Decision-making problem: compare the influence of N factors $x_1, x_2, x_3 \ldots\ldots x_n$ on goal Z. In this paper, we want to determine their proportion in Z, that is, the relative importance of these n factors to target Z. The importance of each factor is quantified pairwise comparison. Take two factors x_i and x_j each time, and use a positive number a_{ij} to represent the importance ratio of x_i to x_j. The matrix $A=(a_{ij})_{n\times n}$ obtained from all comparison results is called the pairwise comparison matrix, which obviously has λ=3.0055.

This paper establishes the scale degree table, and the processed data indicators are compared in proportion. Finally, the weight is calculated according to the comparison results, and the degree table is established.

In this paper, the processed data is imported into SPSS software for analysis, and the degree distribution is carried out in proportion. The distribution results are shown in Table II:

TABLE II DESCRIPTIVE STATISTICS

	Total input price tax	Total sales tax on fake prices	Bad entry rate	Bad debt ratio	The difference between sales and income
The case number	121	123	123	123	123
Minimum value	280 . 0000000	36742.00000	0.0000000000	0.0126315789	-2363128053
Maximum value	7212019834	4848891781	0.2564102564	0.6896551724	189595472
Average value	90744714.37	127477186.8	0.342775709	0.1423575276	38207996.29
Standard deviation	655596469.4	490420825.2	0.0347741335	0.1113168320	301113923.4

The scale method is used to compare the indicators respectively, and finally, the weight is calculated according to the comparison results, and the weight judgment matrix is established, as shown in Table III:

TABLE III WEIGHT JUDGMENT MOMENT

	Company strength	The stability of supply and demand	Customer churn
Company strength	1	2	5
The stability of supply and demand	1/2	1	2
Customer churn	1/5	1/2	1

The credit risk indicator matrix is as follows:

$$A=\begin{bmatrix} 1 & 2 & 5 \\ \frac{1}{2} & 1 & 2 \\ \frac{1}{5} & \frac{1}{2} & 1 \end{bmatrix} \tag{3}$$

In this paper, MATLAB software is used to calculate the weight vector of the matrix and the maximum eigenvalue of the criterion layer is λ=3.0055. MATLAB is used to test the consistency of the analytic hierarchy process structure. After normalization, the weight of the eigenvector is: $W=\{0.5954,0.2764,0.1283\}$

$$CI=\frac{\lambda_{\max}-n}{n-1}=0.0028 \tag{4}$$

$$C_R=\frac{C_I}{R_I}=0.0053 \tag{5}$$

Through a series of calculations above, the purpose is to obtain the credit risk factor coefficient, and the normalized eigenvector weight is:

$$W=\{0.5954,0.2764,0.1283\}$$

The bank's evaluation data on "credit rating" and "default" of enterprises are transformed into quantitative data of credit risk, and the feature engineering for credit rating classification is established based on "default." The feature vectors included in the feature engineering are shown in Table IV below.

TABLE IV EIGENVECTOR CLASSIFICATION

Company strength
Total input price tax
Total output price tax
The total difference between sales and inputs
Number of valid invoices

Because too much data is given in establishing Feature Engineering, there will be errors in the process of data processing. Therefore, it is necessary to process the abnormal value of the data. The abnormal value inspection method is 3sigma abnormal value identification, and the abnormal filling method is 3 times the standard deviation. After the comparison data is average, the frequency analysis shall be conducted regarding the "credit rating" to prove the representativeness and credibility of the analysis of relevant problems, as shown in Figure 2

variable	frequency	percentage	cumulative percentage
3	22	18.182	18.182%
0	27	22.314	40.496%
1	34	28.099	68.595%
2	38	31.405	100.000%

Fig. 2 Frequency analysis

The above variables 0, 1, 2, and 3 correspond to grade A, B, C, and D in the credit rating, respectively, resulting from data label conversion. It can be seen that the data is not too balanced. In order to balance the data, this paper will use the smote oversampling to supplement the additional samples generated after interpolation to make their frequencies similar.

In this paper, the eigenvalues are classified and screened similar to machine learning, and the effective indexes are substituted into the machine learning model xgboost regression for training. By adjusting the number of decision trees parameter to 100, it takes [credit rating] as the main reference index and substitutes it into mpai software for training. The evaluation results of test data are shown in Figure 3.

The evaluation index	The evaluation results
MSE	64.2167
RMSE	8.0135
MAE	41.0061
R^2	0.4064
MAPE	Inf

Fig. 3 Test data evaluation results

Finally, the sensitivity test of the model is carried out: it can be seen that = 0.4064 meets the test conditions. That is, the test results are valid.

C. Model results

Using the processed data in 3.1, this paper classifies and screens the eigenvalues similar to machine learning, substitutes the effective indicators into the machine learning model xgboost regression for training, substitutes the normalized weight vector in the analytic hierarchy process using xgboost machine learning, uses mpai to calculate the final enterprise credit risk ranking, and gives the credit strategy scientifically and reasonably according to the ranking. According to the nested xgboost regression model of analytic hierarchy process and simple data processing of Excel, the ranking after quantitative analysis of data is published in Table V below:

TABLE V LIST OUT

Enterprise	Score
E17	0.585450093
E19	0.580834447
E6	0.567753341
E8	0.559237371
E18	0.558185069
E11	0.549882138
E2	0.548528098
E1	0.547067641
E27	0.515122021
E22	0.485928991

Therefore, the higher the score, the stronger the enterprise's ability to deal with risks. According to this classification, the bank can reduce the interest and make more loans to the enterprises with higher scores. The loan amount can be classified according to the number of enterprises. The enterprises with higher scores can make large loans, and those with lower scores can only make small loans. The loan amount is distributed in the same amount of 100000 ~ 1 million respectively, and the interest rate increases by 4%~15%.

IV. MODEL EVALUATION AND CONCLUSIONS

This paper establishes a reasonable mathematical model and formulates the bank's credit strategy by evaluating and ranking the credit risk. We first use the analytic hierarchy process to determine the factors affecting the credit risk and then use the comparative method to quantify the importance of each factor. After the consistency test is passed, an acceptable pairwise comparison matrix is constructed. MATLAB software calculates the proportion of N factors in this layer in target Z. The proportion in the criterion layer is written into a vector and normalized to obtain the weight vector.

Then, this paper uses SPSS software to quantify the stability of supply and demand in the criterion layer and then quantify the relationship between customer churn rate and reputation level given in the data to calculate the average value of the three levels of ABC. After data preparation, this paper classifies the eigenvalues for similar machine learning. We use Mpai software to substitute effective indicators into the machine learning model XGBOOST regression for training. In this paper, xgboost machine learning is used to substitute the normalized weight vector in the analytic hierarchy process to obtain the quantitative regression value of the company's strength. Finally, the credit risk rating evaluation formula is constructed using the criteria layer weight and the lower level quantitative data. The final enterprise credit risk ranking is obtained using Mpai

software, and the credit strategy is given scientifically and reasonably according to the ranking.

REFERENCES

[1] Huang Yuan Research on credit risk based on logistic regression algorithm [J] Scientific consulting (Science and technology Management), 2019 (04): 71

[2] Yang Jun, Xia Chenqi Credit risk assessment of small enterprises based on gradient boosting algorithm [J] Zhejiang finance, 2017 (09): 44-50

[3] Shang Miao Credit risk assessment and management of small and medium-sized enterprises based on logit model [J] Times finance, 2018 (12): 210 + 220

[4] Liu Dongying Credit risk and management of small and micro enterprises in commercial banks [J] Times finance, 2019 (33): 26-27

[5] Wang Xing, Li Yaqiong Research on financial default prediction of Listed Enterprises Based on xgboost [J] Economic mathematics, 2020,37 (03): 195-201

[6] Li Wei Objective oriented analytic hierarchy process and its application [D] Jilin University, 2019

[7] Chai Rui, Luo Jiajia Application of analytic hierarchy process and fuzzy mathematics in financial risk evaluation [J] Modern business, 2020 (23): 169-172