Advances in Economics, Business and Management Research, volume 166
Proceedings of the 6th International Conference on Financial Innovation and Economic Development (ICFIED 2021)

Research on the Default Risk and Credit Strategy of Small and Medium-Sized Enterprise

Siqi Guo[1], Wanqing Peng[1], Haoxuan Li[1,*]

[1] *College of Mathematics, Sichuan University, Chengdu, Sichuan 610064*
[*] *Corresponding author. Email: hxli.scu@gmail.com*

ABSTRACT
With the rapid development of China's multilevel capital markets, small and medium-sized enterprises gradually occupy an important part of the market economy. Due to the characteristics of small scale, simple technology, low financing and capital utilization efficiency of SMEs, commercial banks will have concerns when facing loans, therefore a more effective credit evaluation criteria should be designed. This article studies the default risk and credit strategy of SMEs from the perspective of banks. For the default risk assessment system, the default risk prediction model is constructed based on the Logistic regression model, and the corporate reputation rating model is constructed based on the CART decision tree. For bank credit strategy decisions, risk reduction and profit gain should be balanced reasonably. The establishment of the dual-objective optimization model can better deal with the trade-off between risk and profit, thus determine the bank's loan strategy. The Pareto efficiency of the model is obtained by NSGA-II optimization technology.

Keywords: *Small and medium-sized enterprise, Default risk, Credit strategy, CART decision tree, NSGA-II.*

1. INTRODUCTION

In the development of China's national economy, Small and Medium-Sized Enterprises (SME) have an irreplaceable position, not only as a catalyst for economic growth, but also as a driving force for stimulating new vitality of the national economy [1]. Due to the characteristics of small scale, simple technology, low financing and capital utilization efficiency of SMEs, the state has issued a number of policies to support the development of SMEs in recent years, including credit policies [2, 3]. Therefore, credit loans have become an important way for SMEs to flourish, and it is also the economic lifeline of commercial banks. Bank credit is an economic activity in which banks temporarily lend part of their deposits to enterprises and institutions, then collect them within an agreed time and charge a certain amount of interest. In general, SMEs are susceptible to factors such as policies and markets, and the default risk is higher than that of the large enterprises. Therefore, the realization of risk avoidance and profit maximization of bank loans to SMEs has become an important concern of the contemporary era.

In this article, the corporate default risk assessment model is constructed based on the statistical methods such as Logistic regression and decision trees to conduct risk assessments on companies with and without credit records, respectively [4-6]. All data is extracted from the invoices of 123 companies with credit records and 302 companies without credit records. For numerical variables, the invoice data of a specific company is merged, so that the final indicator data is presented quarterly. For text variables, based on the sequence relationship and specific meaning, the text content of the variable is mapped to a numerical representation.

2. INDEX EXTRACTION

In order to measure the credit risk of SMEs more comprehensively, this article uses non-financial indicators and qualitative indicators to construct the credit risk assessment model, and obtains 5 primary indices and 17 secondary indices, as shown in Table 1.

The correlation analysis of the 17 secondary indices selected above indicates the existence of multicollinearity, which would reduce the robustness of the credit evaluation model if the indices are used directly, and the prediction results will be inaccurate as a result. In order to reduce the impact of the correlation between indices and retain the effective information in the indices as much as possible, stepwise regression analysis is used to select indices. The idea is to introduce indices into the model individually: after each index is introduced, every index in the model need to be tested individually.

The stepwise regression results of the 17 secondary indices are shown in Table 2. The stepwise regression results of 17 secondary indicators are shown in Table 2. The analysis shows that the p-value of the F test is less than 0.005, and the model is valid through the significance test. The VIF value of each index is much less than 10, and there is no multicollinearity between the indices.

Table 1. Primary and secondary index of the enterprises

Primary index	Secondary index	
Enterprise scale	Tax payment	X_1
	The total number of invoice	X_2
	The number of income invoice	X_3
	The number of output invoice	X_4
Profitability	Tax payment growth rate	X_5
	Operating revenue	X_6
	Net profit rate on sales	X_7
Development ability	High-tech enterprise	X_8
	Operating revenue growth rate	X_9
	Net profit rate growth rate	X_{10}
Supply and demand stability	The number of buyers	X_{11}
	The number of sellers	X_{12}
	Negative invoice rate	X_{13}
	Trade stability	X_{14}
Corporate reputation	Income invoice invalid rate	X_{15}
	Output invoice invalid rate	X_{16}
	Enterprise credit rating	X_{17}

Table 2. Stepwise regression results of the indices

	B	RMSE	t-value	p-value	VIF
Intercept	1.056	0.115	9.195	< 0.001	N/A
X_{17}	(0.012)	0.001	(9.139)	< 0.001	1.251
X_{16}	0.730	0.256	2.850	0.005	1.226
X_7	(0.047)	0.018	(2.564)	0.012	1.140
X_{14}	(0.012)	0.006	(2.172)	0.032	1.112

After stepwise regression, 4 effective indices are finally extracted: *net profit rate on sales* X_7, *trade stability* X_{14}, *output invoice invalid rate* X_{16} and *enterprise credit rating* X_{17}.

3. DEFAULT RISK ASSESSMENT SYSTEM

From the perspective of bank interests, the risk assessment of a company is a necessary preparation before operating a loan. In order to quantify the credit risk more accurately, this chapter takes the four indices extracted above, combined with the credit rating of each company, and builds a reasonable credit risk evaluation system through the Logistic binary model to effectively reduce the credit risk faced by banks.

3.1. Default Risk Prediction Model

3.1.1. Logistic Regression Model

Logistic regression is a non-linear model used to predict the probability of events affected by multiple factors. The default risk prediction model is established based on Logistics regression as follows:

$$\ln(Y) = \ln\frac{P}{1-P} = \beta_0 + \beta_1 X_7 + \beta_2 X_{14} + \beta_3 X_{16} + \beta_4 X_{17}$$

Where Y is the enterprise default label, and P is the enterprise default probability.

Take 0.5 as the critical value, and provide loans to enterprises whose regression default probability does not exceed 0.5. The estimation of the model parameters is obtained through ten-fold cross-validation, and the regression results are as follows:

$$\ln\frac{P}{1-P} = 13.95 - 0.41X_7 - 0.07X_{14} - 0.87X_{16} - 0.26X_{17}$$

3.1.2. Model Testing and Evaluation

Comparing the enterprise default probability based on Logistic regression prediction with the actual default record, and combining the results of ten-fold cross-validation, the confusion matrix of the default risk prediction model is shown in Table 3.

Table 3. Confusion matrix of the regression model

		Predicting value		Accuracy	Error rate
		0	1		
Actual value	0	96	0	100.00%	0.00%
	1	3	24	88.89%	11.11%
Total value				97.56%	2.44%

From Table 3, the model's prediction accuracy is 97.56%. The p-value of the likelihood ratio chi-square test is less than 0.001, the AUC is 0.973, the recall is 0.976, and the F_1-value is 0.975, which is the harmonic mean of the precision and the recall. The AUC, accuracy, precision, recall and the F_1-value are all close to 1, which

reflects the effectiveness of the default risk prediction model.

3.2. Corporate Reputation Rating Model

Corporate reputation is an important indicator for banks to assess the risk of default, divided into four levels: A, B, C, and D. Banks will not grant loans to companies with a corporate reputation of D [7, 8].

For companies with credit records, take corporate reputation as a known independent variable and establish a Logistic regression model according to section 3.1 to quantify the corporate credit risk. However, for companies without credit records, their corporate reputation cannot be directly obtained, then the credit risk assessment model lacks the indicator data. Therefore, banks first need to indirectly evaluate the corporate reputation level through factors such as corporate scale and corporate operating revenue, then bring it into the Logistic regression model to estimate the probability of default. The following provides a method to evaluate the reputation of companies without credit records by constructing a decision tree.

The CART decision tree algorithm consists of feature selection, tree generation and pruning [9]. A bipartite recursive segmentation technique divides the current sample into two sub-sample sets, so that each non-leaf node generated has two branches. The CART decision tree algorithm for classification takes the Gini index instead of the information gain ratio [10]. The Gini index represents the impurity of the model, and a lower Gini index represents a higher model effectiveness.

The factor data and corporate reputation of 123 companies with credit records are used to train the CART decision tree as shown in Figure 1. Then the corporate reputation estimates of 302 companies with no credit records are obtained from their factor data through CART decision trees. Finally, through the Logistic regression model established in Section 3.1, the company's default probability P_i is estimated to determine whether to lend.

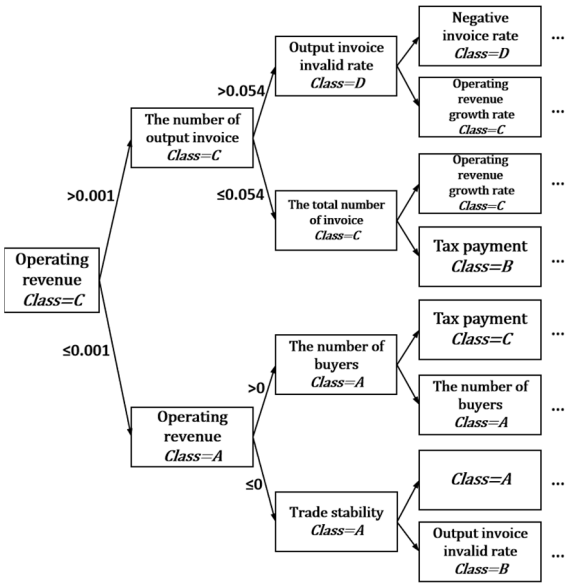

Figure 1 CART corporate reputation decision tree.

4. CREDIT STRATEGY DECISION

After evaluating the default risks of enterprises, banks should maximize their own interests while avoiding risks. Therefore, the credit strategy design of commercial banks needs to optimize both the tolerable risk and profit.

In this chapter, in order to effectively balance default risk and loan interest, a dual-objective optimization model is constructed to propose reasonable and beneficial credit strategies for banks, and the optimal solution of the dual-objective optimization model is determined by genetic algorithm [11-14].

4.1. Dual-Objective Optimization Model

Banks will not grant loans to companies with a corporate reputation of D. For companies with corporate reputations of A, B, and C, the amount of loans given to them by banks is also different due to various default risks. Assuming that the amount of bank loans to each enterprise is between 100,000 and 1 million, the annual interest rate for loans to each enterprise is between 4% and 15%.

From the perspective of banks, it is necessary to simultaneously consider obtaining as high profit as possible while reducing the risk of corporate default and minimizing the churn rate of potential customers, so as to avoid the gradual decrease of bank customers in the future. Therefore, assuming that the relationship between the annual interest rate of loans of different corporate reputation and the customer churn rate is known, this section constructs a dual-objective optimization model to balance the relationship between profits and risks.

Suppose the number of companies that the bank agrees to lend is n, the amount of the bank lends to the i-th company is $a_i (i = 1, 2, \dots, n)$, the annual interest rate

is $b_i(i = 1, 2, \dots, n)$, and f_1 represents the total income obtained by the bank through loans in one year. Then introduce the reputation level parameter $\zeta(i) \in \{A, B, C\}$ of the i-th company, let $\mu(b_i)$ denote the customer churn rate when the bank sets the annual interest rate of the i-th company as b_i, and suppose the customer churn rate of A, B, and C levels are $\mu_1(b_i), \mu_2(b_i)$ and $\mu_3(b_i)$, thus:

$$\mu(b_i) = \begin{cases} \mu_1(b_i), & \zeta(i) = A \\ \mu_2(b_i), & \zeta(i) = B \\ \mu_3(b_i), & \zeta(i) = C \end{cases}$$

$$f_1 = \sum_{i:P_i \le 0.5} a_i \cdot b_i \cdot [1 - \mu(b_i)]$$

When making loans for enterprises, it is necessary to consider the losses suffered by the bank in default of the enterprise. Generally, the larger the loan amount, the higher the risk of default by the enterprise, and the greater the profit that the bank may lose. Suppose the bank's loss amount is f_2, thus:

$$f_2 = \sum_{i:P_i \le 0.5} \mu(b_i)/n$$

Assuming that the bank's total annual credit is 100 million, the two-objective optimization model is established as follows:

$$\begin{cases} f_1 = max \sum_{i:P_i \le 0.5} a_i \cdot b_i \cdot [1 - \mu(b_i)] \\ f_2 = min \sum_{i:P_i \le 0.5} \mu(b_i)/n \end{cases} \text{ s.t. } \begin{cases} 10 \le a_i \le 100 \\ 0.04 \le b_i \le 0.15 \\ \sum_{i:P_i \le 0.5} a_i \le 10000 \end{cases}$$

4.2. Model Solving by Genetic Algorithm

First, randomly generate an initial population with a size of N. After non-dominated sorting, the first generation progeny population is obtained through the three basic operations of genetic algorithm selection, crossover and mutation. Secondly, starting from the second generation, the parent population and the offspring population are merged to perform fast non-dominated sorting, and at the same time, the crowding degree of the individuals in each non-dominated layer is calculated, and the appropriate individuals are selected according to the non-dominated relationship and the crowdedness of the individuals to form a new parent population. Finally, a new progeny population is generated through the basic operations of genetic algorithm, and iterates until the conditions for the end of the program are met.

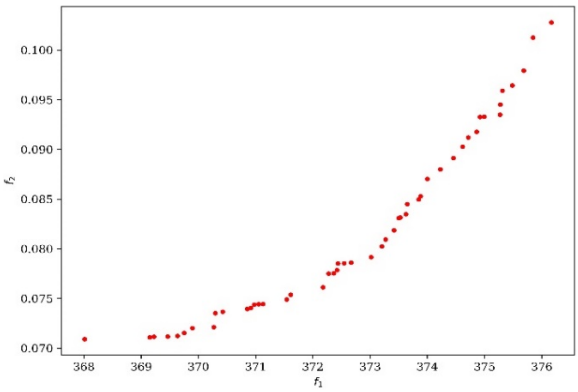

Figure 2 Pareto optimal solutions of the model.

Set the number of iterations to 250, the training set is the data of 123 companies with credit records, and all Pareto optimal solutions are obtained as shown in Figure 2. Each scattered point in the figure represents a specific lending strategy, that is, banks can issue different loan amounts to 123 companies in accordance with their own different preferences for customer churn rate and loan returns, where f_1 represents the total profit of the bank, and f_2 represents the average customer churn rate. The annual interest rate set by the bank when issuing loans directly determines the bank's total loan profit, and Figure 2 shows that f_1 and f_2 are positively correlated. When the bank increases the annual interest rate and expects to obtain high profits, it will also cause an increase in customer churn rate. Banks need to balance profit and customer churn rate. Therefore, the solution that balances the total profit and the risk should be selected as the final solution of the model.

5. CONCLUSION

The development of small and medium enterprises is in a bottleneck period. Based on the data set of 123 companies with credit records and the data set of 302 companies without credit records, this article establishes a default risk assessment system through Logistic regression and CART decision tree, and balances bank profits and risks and design the credit strategy through dual-objective optimization model.

In the social environment of the rise of big data, banks can establish big data risk assessment models to quantify credit risks. Combining the background of the times and market demand, banks can help small and medium-sized enterprises while effectively avoiding risks and promoting the vigorous development of the Chinese economy.

REFERENCES

[1] Zhiyun L. Banking Structure and the Small-Medium-Sized Enterprise Financing[J]. Economic Research Journal, 2002.

[2] Yida N, Hehua L. Research on credit risk management of small and medium-sized enterprise in the industrial and commercial bank of China[J]. The International Journal of Electrical Engineering & Education, 0020720920983509.

[3] Cook P, Nixson F. Finance and small and medium-sized enterprise development. Manchester: Institute for Development Policy and Management, University of Manchester, 2000.

[4] Menard S. Applied logistic regression analysis[M]. Sage, 2002.

[5] Bolton C. Logistic regression and its application in credit scoring[D]. University of Pretoria, 2010.

[6] Zhu Y, Xie C, Sun B, et al. Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models[J]. Sustainability, 2016, 8(5): 433.

[7] Barnett M L, Jermier J M, Lafferty B A. Corporate reputation: The definitional landscape[J]. Corporate reputation review, 2006.

[8] Davies G, Chun R, Da Silva R V, et al. Corporate reputation and competitiveness[M]. Psychology Press, 2003.

[9] Lewis R J. An introduction to classification and regression tree (CART) analysis[C]. Annual meeting of the society for academic emergency medicine in San Francisco, California. 2000, 14.

[10] Priyam A, Abhijeeta G R, Rathee A, et al. Comparative analysis of decision tree classification algorithms[J]. International Journal of current engineering and technology, 2013, 3(2): 334-337.

[11] Zain N A M . Overview of NSGA-II for Optimizing Machining Process Parameters[J]. Procedia Engineering, 2011.

[12] rey Horn J, Nafpliotis N, Goldberg D E. Multiobjective optimization using the niched pareto genetic algorithm[J]. IlliGAL report, 1993, 93005.

[13] Horn J, Nafpliotis N, Goldberg D E. A niched Pareto genetic algorithm for multiobjective optimization[C]. Proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence. Ieee, 1994: 82-87.

[14] Vlennet R, Fonteix C, Marc I. Multicriteria optimization using a genetic algorithm for determining a Pareto set[J]. International Journal of Systems Science, 1996, 27(2): 255-260.

[15] Deb K, Agrawal S, Pratap A, et al. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II[C]. International conference on parallel problem solving from nature. Springer, Berlin, Heidelberg, 2000: 849-858.

[16] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE transactions on evolutionary computation, 2002, 6(2): 182-197.

[17] Hamdani T M, Won J M, Alimi A M, et al. Multi-objective feature selection with NSGA II[C]. International conference on adaptive and natural computing algorithms. Springer, Berlin, Heidelberg, 2007: 240-247.