Probabilistic Calibration and Genetic Algorithm-based Bank Credit Strategies for MSMEs and Enlightenment to Tobacco Enterprise Management

Liu Haixu, Undergraduate
Zhang Yong, Professor
Li Hui,Undergraduate
Mao Tianjun,Undergraduate
ZhengWenhui,Undergraduate
LiJiao,Undergraduate

Liu Haixu, Bachelor in Mathematical Modeling and Machine Learning, Faculty of Bigdata and Artificial intelligence, Chizhouuniversity, Chizhou, Anhui 247000, China. Zhang Yong, Professor in Nonlinear Analysis, Faculty of Bigdata and Artificial intelligence, Chizhou University, Chizhou, Anhui 247000, China.Li Hui, Bachelor in Mathematics and Applied Mathematics, Faculty of Bigdata and Artificial intelligence, Chizhou University, Chizhou, Anhui 247000, China. Mao Tianjun, Bachelor in Multi-modal Learning and Information System Management, Dalian University of technology, Dalian, Liaoning, China.Zheng Wenhui, Bachelor in Statistics, Faculty of Bigdata and Artificial intelligence, Chizhou University, Chizhou, Anhui 247000, China.Li Jiao, Bachelor in Big Data and Cloud Computing, Faculty of Bigdata and Artificial intelligence,Chizhou University, Chizhou, Anhui 247000, ChinaCorrespondence author: Zhang Yong;yongzhang@czu.edu.cn

Objectives:To further strengthen the role of Micro, Small and Medium Enterprises (MSMEs) in maintaining the vitality of national economy, governments around the world introduced many special policies. They kept guiding the banking industry to increase the support for MSMEs and reduce their financing difficulties in banks. Basing on the analysis of the bank's credit strategy for small and medium-sized enterprises of similar size, this paper gives the management strategy for small and medium-sized enterprises in tobacco industry to obtain bank credit when they cannot expand their turnover. In this paper, we proposed a binary classification model-based probabilistic calibration algorithm to calculate the default probability of enterprises in the formation of risk measurement model, and found the optimal solution of credit strategy using an improved genetic algorithm. Firstly, we discovered the enterprise's information and invoice data of 123 micro and medium-sized enterprises with existing credit ratings. We extracted several features from multiple perspectives, such as size, relationship in supply chain, profitability, performance ability, and level of development, and removed the correlations among the indicators using principal component analysis (PCA). Secondly, the retained principal components were used as covariates, and we determined the credit ratings of the firms and the probability of default using discrete variables such as the credit ratings of the firms and whether they defaulted. Finally, we substituted the probability of default into the credit risk model to calculate the loss expectation and profit

expectation of the credit portfolio, and used the profit expectation of the credit portfolio as the objective function of the 0-1 programming equation to derive the credit strategy with the lowest risk exposure and the highest return basing on the genetic algorithm.

Key words:PCA-Logistic regression model, Credit risk model, 0-1 planning, Genetic algorithm

Tob Regul Sci.™ 2021;7(6): 5726-5740
DOI: doi.org/10.18001/TRS.7.6.56

INTRODUCTION

In 2005, theUnited Nations introduced theconcept of "financial inclusion", which refers to the provision of effective financial services at an affordable cost to all social classes and groups in need of financial services. One of the key service targets is micro, small and medium enterprises (MSMEs), which are considered to be the main driving force of national economic development. However, MSMEs are characterized by relatively small scale, lack of collateral assets, opaque information, weak risk resistance, short life cycle, etc. Thus, they are not favored by bank capital compared with large enterprises listed companies. [1] On the other hand, most of the current methods for calculating corporate credit risk measures in the banking industry are derived from the risk measure models designed by European and American financial institutions for the credit business of listed companies in the last century, such as the KMV model and the risk metric technique model.[2] Therefore, it leads to the fact that the data available from MSMEs in some developing countries can hardly help banks to quantify credit risk, and ultimately financial institutions have difficulties in completing effective risk control and credit granting.

In summary, the current credit rationing is not conducive to the further development of SMEs. From the perspective of bank development, homogeneous competition in the banking industry is becoming increasingly fierce, and the current corporate credit business reflects pain points such as rigid lending system, lengthy approval process, high credit cost, and difficulty in avoiding manual risks. Only by effectively improving the existing corporate credit business system and process can we improve the risk control capability and operation level of corporate credit business to cope with the increasingly fierce competition in the market. In this context, it is of great significance to analyze the use of big data and statistical learning technologies to solve the financing difficulties of micro and small enterprises and alleviate their difficulties.

Enterprises related to the tobacco industry are involved in a series of economic activities, such as tobacco planting, processing, cigarette manufacturing, packaging, auxiliary materials manufacturing, cigarette production, sales, tobacco international trade and so on.It contains a long industrial value chain, covers a wide range of fields, and its supply chain relations and group management are also very complex.A few oligopolies in the core cigarette manufacturing industry monopolize the worldwide market,[3]but in fact most of the enterprises related to the tobacco industry are typically local small, medium and micro enterprises.[4]Tobacco enterprises are often levied high income tax in the world, so as an important part of the national economy, they play a significant role in economic growth.It is worth mentioning that China National Tobacco Corporation, one of the world's four tobacco giants, is also defined as a small and micro enterprise due to its unique management structure.[5]In recent years, the number of smokers worldwide continues to decline, and the related tax increases continue to make tobacco industry growth be sluggish.The continued rise in tobacco prices has barely kept the industry growing at a modest pace.The tobacco industry and the banks that have financed itin the long term are coming to realize the unsustainability of relying on higher prices to sustain profit growth.Meanwhile, enterprises in the tobacco industry cannot apply for pledge loans because of the characteristics of their own products.Therefore, the loan conditions for enterprises related to the tobacco industry are getting worse and worse, which is not only

conducive to the development and transformation of tobacco enterprises, but also makes the banking industry lose potential profits.

LITERATURE REVIEW

In recent years, many scholars have studied risk measurement models as well as credit strategy models. The research group from the People's Bank of China Business Management Department[6] investigated a variety of big data financial financing methods. Wang[7] et al. studied the early warning mechanism of financial crisis of listed companies in China from the perspective of cash flow by combining the principal component analysis method with logistic regression. Guo[8] et al. established a logistic regression model and LDA model to empirically measure the credit risk of SMEs in China, and determined the optimal model by comparing the model prediction accuracy. Ma[9] et al. extracted features from the historical loan data of a lending institution and compared the pros and cons of common classification models. They finally used the Light GBM algorithm to establish a prediction model for credit overdue behavior. Based on the sample data of 1173 loans, Lv[10] et al. calculated the default rate variable in Credit Risk+ model using logistic regression model for analyzing the factors influencing credit risk of farmland management rights mortgage loans, measuring credit risk of farmland management rights mortgage loans, and predicting default probability. Based on Guo[8] et al.'s study, Xie et al. [11]used a BP neural network model to distinguish and eliminate enterprises with high default risk, and used PCA to construct an evaluation model to calculate and rank the creditworthiness scores of the remaining enterprises, award the final credit rating and use the scores as input to build a planning model that gives the solution with the greatest difference between the return and the risk of loss. In addition, several researchers have used data mining techniques, logistic regression algorithms, as well as entropy weighting TOPSIS method, to accurate classification models as well as evaluation models, e.g., Zhou et al..[12] In fact, the evaluation model of[11] is more subjective, while the assumptions used in the linear programming model established are difficult to hold in reality and do not have a high practical value. Zhou[12], Wang[13] and other scholars used the classical logistic model to predict only the default rate without rating the firm's creditworthiness. This result is not in line with the working specification of banks to determine the credit rate and whether to grant credit based on the firm's creditworthiness rating, and also does not clearly give a feasible credit strategy model.

In this paper, based on the basic assumption that there is no correlation between default events, we first extracted the note information of the dataset using a data mining algorithm to obtain 30 indicators reflecting the business status of enterprises; then combined with principal component analysis to construct new principal components to remove the repetition among the original indicators, and used the new principal components to train a multivariate logistic regression model for the rating for forecasting. The new principal components are then used to train a multivariate logistic regression model to forecast the enterprise's rating. We also applied the Sigmoid function in the logistic regression model to calculate the default rate of all firms and retained firms with credit ratings higher than D. The maximum credit limit for these enterprises is calculated based on the real banking industry's risk limit calculation criteria for small and medium-sized enterprises. We substitute the enterprise default probability and risk limit into the simplified Credit Risk+ credit risk model based on the calculation of credit asset portfolio default loss expectation and profit expectation, and list the 0-1 linear programming credit decision model with the expected profit maximum of credit asset portfolio as the objective function. Finally, the 0-1 linear programming credit decision model is solved by integer programming algorithm and genetic algorithm.

DATA PREPROCESSING

We collected invoice data from 425 micro and medium-sized enterprises. The invoices were dated from October 2016 to February 2020, all input and output invoices were desensitized with information, and the types of invoices included valid invoices, invalid invoices, and negative invoices. Among them, 123 firms had credit records as well as credit ratings, and 302 firms had no credit records as well as credit ratings. We

construct indicators that reflect the creditworthiness of enterprises from five perspectives: enterprise size, supply relationship, profitability, performance capability, and development, combine the input and output invoices of each enterprise, and distinguish them by invoice status. After that, the invoices and their status and years were counted according to total price, quantity, ratio, and average value, and the results were further explored. For example, the number of valid invoices was obtained by subtracting twice the number of negative invoices from the number of valid invoices. To reflect whether the supply chain of the enterprise is stable or not, we count the number of supply times of the selling unit as well as the purchasing unit to the enterprise, and design the threshold value to determine whether the unit is a stable supplier or distributor of the enterprise, and count the stable suppliers and stabledistributors of each enterprise. Finally, we use the replace function to fill in the missing values, and use the merge function to establish the enterprise information table and invoice information table. The constructed 28 indicators are shown in Table 1.

Table 1
Metrics Obtained through Data Mining Metric

Symbol	Description	Symbol	Description
X_1	Number of failed invoices	X_{16}	Number of suppliers
X_2	Invoice invalidation rate	X_{17}	Stable suppliers
X_3	Profits	X_{18}	Number of dealers
X_4	Actual tax payment	X_{19}	Number of stable dealers
X_5	Turnover	X_{20}	Total input price tax in 2018
X_6	Total input tax	X_{21}	Total outbound tax in 2018
X_7	Profit before tax	X_{22}	Profit level in 2018
X_8	Number of Negative Bills	X_{23}	Profit amount in 2018
X_9	Valid invoices	X_{24}	Total input price tax in 2019
X_{10}	De-negative valid invoices	X_{25}	Total outbound tax in 2019
X_{11}	Incoming expenses	X_{26}	Profit level in 2019
X_{12}	Sales revenue	X_{27}	Profit amount in 2019

X_{13}	Profit margin	X_{28}	Annual sales growth rate
X_{14}	Income Tax	X_{29}	Default record
X_{15}	Taxes on sales	X_{30}	Credit rating

METHODOLOGY

PCA-based Logistic Regression Model

Since there are many types of classification models, we use $X_1 \Lambda\ X_{28}$ as features to select the classification models with excellent performance by Classification Learner toolbox in Matlab. The selection results are shown in Table 2.

Table 2
Classifier Selection Results

Symbol	Description	Symbol	Description
X_1	Number of failed invoices	X_{16}	Number of suppliers
X_2	Invoice invalidation rate	X_{17}	Stable suppliers
X_3	Profits	X_{18}	Number of dealers
X_4	Actual tax payment	X_{19}	Number of stable dealers
X_5	Turnover	X_{20}	Total input price tax in 2018
X_6	Total input tax	X_{21}	Total outbound tax in 2018
X_7	Profit before tax	X_{22}	Profit level in 2018
X_8	Number of Negative Bills	X_{23}	Profit amount in 2018
X_9	Valid invoices	X_{24}	Total input price tax in 2019
X_{10}	De-negative valid invoices	X_{25}	Total outbound tax in 2019
X_{11}	Incoming expenses	X_{26}	Profit level in 2019
X_{12}	Sales revenue	X_{27}	Profit amount in 2019

X_{13}	Profit margin	X_{28}	Annual sales growth rate
X_{14}	Income Tax	X_{29}	Default record
X_{15}	Taxes on sales	X_{30}	Credit rating

We compressed features using principal component analysis, and significantly improve the accuracy of the logistic regression classifier up to 10.6%. Then we adjusted the model hyperparameters and the classification accuracy increased to 89.4%. Another advantage of logistic regression is that it calculates the probability of classification, so there is no need to calibrate the results with probability again. Therefore, we chose PCA-based logistic regression to build a prediction model of corporate default rate under the condition of unknown credit rating.

Since we have considered a large number of features and these features are correlated, if we directly use the logistic regression, there will be a serious redundancy problem. Therefore, we use PCA to reduce the dimension and remove the correlation between indicators. The steps of PCA are as follows.

Firstly, we normalize the data using the Z-score method, as shown in Eq. (1), where the parameters are calculated based on Eq. (2).

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} (\mathrm{i}, j = 1,2,..., p) \quad (1)$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} (\mathrm{j} = 1,2,..., p) \quad (2)$$

Then we compute the correlation coefficient matrix,

$$R = \begin{bmatrix} r_{11} & r_{12} & ... & r_{1p} \\ r_{21} & r_{22} & ... & r_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ r_{p1} & r_{p2} & ... & r_{pp} \end{bmatrix} \quad (3)$$

where $r_{ij}(i, j = 1,2,...p)$ is the correlation coefficient between the x_i and x_j, which is calculated as Eq. (4).

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}} \quad (4)$$

After that, we find the characteristic roots λ_i of the correlation coefficient matrix and the corresponding eigenvectors e_i, and calculate the principal component contribution and the cumulative contribution based on the following equation.

$$\mathrm{b}_j = \frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k} (5)$$

In general, we select the principal components corresponding to the eigenroots with more than 85% of the cumulative eigenroots. We calculated the principal component loadings based on Eq. (6), and normalized the principal component loadings as follows.

$$a_{ij} = \sqrt{\lambda_i} e_{ij} \qquad a_{ik}^* = \frac{a_{ik}}{\sqrt{\sum_{k=1} a_{ik}^2}} (6)$$

Finally, we can derive the expressions of each principal component as follows.

$$X_i = a_{1i}x_1 + a_{2i}x_2 + a_{3i}x_3\Lambda + a_{ni}x_i \quad (7)$$

Logistic regression model is a generalized linear regression model, which is mainly used for binary classification with discrete variables as the dependent variable in a small sample training set, or calculating the probability of a certain event. Using the logistic regression model, we calculate the probability of default events of 123 firms with credit records as the training set, and the steps are as follows.

Firstly, we use the Sigmoid function to map the predicted values to probability ranging from 0 to 1.

$$P_i = \frac{1}{1+e^{-Z}} \quad (8)$$

where the power of the exponential function is the linear decision boundary function, where β_0 is the constant term and $\beta_1, \beta_2 \cdots \beta_{28}$ is the regression coefficient.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n \quad (9)$$

We take the principal components constructed by PCA as input, and compute the distance between the regression coefficients and the intercept using machine learning, and substitute the linear decision boundary function into the Sigmond function to obtain Eq. (10), and finally find the probability of default for each firm:

$$P_i = \frac{1}{1+\exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)]} \quad (10)$$

To identify the enterprises with credit rating D, i.e., those cannot be granted credit, we consider using a multivariate logistic regression model to classify the credit rating of these 123 enterprises.

Multivariate logistic regression can be considered as multiple dichotomous logistic regressions. We assume that the dependent variable is a categorical variable, the number of classes is 4 and there is no order between the classes. We assume that the values of Y are a, b, c, d , and they are chosen as the common reference group, then we have the following model:

$$\begin{aligned} \ln\left[\frac{P(Y=b)}{P(Y=a)}\right] &= \alpha_b + \beta_{11} x_1 + \cdots + \beta_{1p} x_p \\ \ln\left[\frac{P(Y=d)}{P(Y=a)}\right] &= \alpha_d + \beta_{31} x_1 + \cdots + \beta_{3p} x_p \\ \ln\left[\frac{P(Y=c)}{P(Y=a)}\right] &= \alpha_c + \beta_{21} x_1 + \cdots + \beta_{2p} x_p \end{aligned} \quad (11)$$

The probability of each class can be defined as follows.

$$P(Y=a \mid x, \omega) = \frac{1}{(1+\sum_{i=1} e^{\omega_i \cdot x})} \quad (12)$$

$$P(Y=i \mid x, \omega) = \frac{e^{\omega_i \cdot x}}{(1+\sum_{i=1} e^{\omega_i \cdot x})} \quad (13)$$

Finally, we use the classification result with the highest probability (larger than 50%) as the prediction result.

Credit Risk+ Based Risk Evaluation Model

The Credit Risk+ model only considers two states, default and non-default. The traditional Credit Risk+ model assumes that the default probability is fixed during the observation period and the number of defaults follows a Poisson distribution, and the probability of a loan default in the asset portfolio is given by Eq. (14).

$$P(\text{nde}faults) = \frac{\mu^n e^{-\mu}}{n!} \quad (14)$$

where μ is the average number of defaults per year, $\mu = \sum P_M$, P_M is the probability of default for a single debtor M, and the number of defaults $\text{n} = 0,1,2\Lambda$. When we calculate the loss distribution, Credit Risk+ divides the exposures into many groups, each of which is considered as a separate portfolio of assets. Since the Credit Risk+ model assumes that the number of defaults obeys the Poisson distribution, the probability generating function of the asset portfolio can be computed using Eq. (15):

$$G(\text{z}) = \prod_i^m e^{-\mu i + \mu i z^{v_i}} \quad (15)$$

where μ_{i} denotes the number of expected defaults in the i-th group and v_i denotes the i-th group in L. The portfolio default loss distribution is Eq. (16).

$$P(loss = nL) = \frac{1}{n!} \frac{d^n G(z)}{dz^n}\Big|_{z=0} (1,2,3,\Lambda\ ,\text{m}) \quad (16)$$

To measure credit risk using the Credit Risk+ model, the required parameters include: default loss rate, default probability, standard deviation of default rate, and exposure. In this study, we keep the traditional Credit Risk+ idea, but since the probability of default and its standard deviation can be predicted by the logistic model, so the Poisson distribution cannot describe the

probability of default of n loans in the asset portfolio.

Furthermore, in the scenario of this paper, the credit limits are different for different enterprises, and the calculation is based on the common risk limit calculation formula for the banking industry, Risk Limit = Solvency Base × Limit Multiplier. The solvency base varies by industry, but generally consists of a 50% weighting of sales revenue and a 50% weighting of net assets, although the weighting may change depending on the industry to which the enterprise belongs. In cases where net worth cannot be estimated, banks often limit the valuation of net worth by the maximum ratio of sales revenue to net worth. For MSMEs, the limit multiplier is limited to 55%, and the risk limit of banks' credit to MSMEs will not exceed 40% of their sales revenue in the previous year (2019), which can be approximated as a limit multiplier k based on business revenue equal to 0.4.

Unlike mortgage loans, credit loans are not collateralized, so the risk exposure is 100% of the risk limit. In the short term, the default loss rate is close to or even exceeds 100% if the enterprise defaults. Different banks have significantly different default loss rates, so we consider the default loss rate as 100% in our subsequent modeling.

Based on the above analysis, once the default rates are determined, we can calculate the expected loss of the debt portfolio with the risk limit of each enterprise. There are in total 2^n possible default scenarios in the portfolio containing n credits. The probability of each possible occurrence is shown in Eq. (17).

$$P(Ndefault) = \prod_{i \in N} P_i \prod_{j \notin N} (1 - P_j) \quad (17)$$

where N is the set of defaulted credit firms in the asset portfolio and has the following relationship.

$$P_{sum} = \sum_{i=1}^{2^n} P_i(Ndefault) = 1 \quad (18)$$

In this case, if the firm defaults, it loses all risk limits. If the firm performs, it gains all the profit, i.e., the annual interest rate multiplied by the risk limit. Therefore, the expected return in this case can be calculated as in Eq. (17), where ir_Y denotes the annual interest rate of the credit facility when the firm's interest rate credit rating is Y. The APR for credit facilities is:

$$R(Ndefault) = \sum_{j \notin N} (kx_j \cdot ir_Y) - \sum_{i \in N} (kx_i \cdot LGD) \quad (19)$$

So we can obtain Eq. (20), which represents the expectation of return for this credit strategy.

$$E(Ndefault) = \sum_{i=1}^{2^n} P(Ndefault) R(Ndefault) \quad (20)$$

So we can obtain Eq. (20), which represents the expectation of return for this credit strategy.

Credit Strategy Model based on 0-1 Planning

The objective function isto maximize the credit strategy revenue expectation, when the number of enterprises with credit demand is set to n, we need to maximize the revenue expectation by deciding which enterprises can be lent and which enterprises should be rejected. Then we need to determine the most reasonable portfolio of credit assets N. In addition, to make the model more practically useful, we need to add some constraints to the 0-1 planning model. This is because the total amount of credit granted cannot exceed a fixed value M, which is determined by the bank based on the cost budget of the MSME credit business. If an enterprise is not selected in $R(Ndefault) = \sum_{j \notin N} (kx_j \cdot ir_Y) - \sum_{i \in N} (kx_i \cdot LGD)$, all the gains or losses associated with it are cancelled out. In the formula for computing the probability $P(Ndefault) = \prod_{i \in N} P_i \prod_{j \notin N} (1 - P_j)$, no changes are needed. Since when a firm is not selected, no matter what the default rate of the firm is, it cannot have any effect on the corresponding profit, and it is easy to find that in the case where the number of firms n is determined, if m firms are excluded, then theoretically there exists 2^{n-m} possibilities of different defaults, and from the original point of view, there are 2^m cases that are combined into one case. Then the basic form of Eq. (21) of the 0-1 planning model for this problem is as follows.

$$\min Z = \sum_{i=1}^{2^n} \prod_{i\in N} P_i \prod_{j\notin N} (1-P_i) \cdot [\sum_{j\notin N} (kx_j \cdot ir_Y \cdot X_j) - \sum_{i\in N} (kx_i \cdot LGD \cdot X_i)]$$
$$s.t. \begin{cases} 0 \le \sum_{k=1}^{n} kx_k \cdot X_k \le M \quad k = 1,2\Lambda\ ,n \\ X_{i,j} = 0,1 \quad i,j = 1,2,\Lambda\ ,n \end{cases} \tag{21}$$

With the above model, we can quickly give credit strategies for multiple MSMEs' credit applications and submitted invoice data by determining the default loss rate and the risk exposure of credit operations or pledged loan operations with a fixed cost budget for the bank's credit operations. In addition to 0-1 planning, this model can be changed to a nonlinear planning model with constrained variables. In a general sense, there is no fixed solution to nonlinear integer programming, which can then be solved with the help of simulation optimization algorithms. The traditional 0-1 planning algorithm is mainly solved by dynamic planning algorithm or branch-and-top-bound method, whose time complexity increases dramatically with the increase of dimensionality, when we consider using 0-1 planning algorithm and genetic algorithm to simultaneously find the optimal solution for the objective function.

Improved Genetic Algorithm based on Solving Credit Strategy Model

Genetic algorithm (GA) is a stochastic global search and optimization method developed based on the principle of natural selection and natural genetic mechanism, which imitates the biological evolution mechanism in nature. The essence of the GA is to obtain the optimal solution or quasi-optimal solution through the population search technique, which evolves generation by generation according to the principle of survival of the fittest. It has the following steps.

(1) Chromosome encoding and decoding

The encoding maps the expression space to the genotype space, while decoding is the inverse process. Since we are dealing with a 0-1 planning problem, so we adopt binary coding. If there are n companies eligible for loans, we can generate 2^n different codes of length n, which are feasible solutions. If the i-th number of the code is 0, it means that no loan will be made to the i-th company; on the contrary, if it is 1, it means a loan will be made to this company.

$x_1 = 00000\Lambda\ 000$

$x_2 = 00000\Lambda\ 001$

$x_3 = 00000\Lambda\ 010$

Λ

$x_{2^n} = 11111\Lambda\ 111$

(2) Individual fitness function

The fitness function is transformed from the objective function, for the objective function of the maximization problem, $Fitness(f(x)) = f(x)$, where $f(x)$ is the objective function, but since the objective function is not determined, it will not meet the non-negative requirements for selection probability of operator roulette, so we need to construct the boundary.

For the objective function maximization problem, the individual (chromosome) fitness function is as follows,

$$Fitness = \begin{cases} f(x) - C_{min}, f(x) > C_{min} \\ 0 \end{cases} \tag{22}$$

where $C_{\min}$ is the estimated minimum value of the objective function.

(3) Selection process

The selection rule is that the higher the fitness of a chromosome, the greater the chance of getting replicated. Assuming that there are N chromosomes x_i in the population, the fitness function of the chromosome is $f(x_i)$, the probability of the chromosome being selected is

$$p(x_i) = \frac{f(x_i)}{\sum_{j-1}^{n}(x_j)} \tag{23}$$

This is the famous roulette wheel selection.

(4) Crossover operation

The crossover operation occurs between two chromosomes, which is the most important genetic operation. It randomly pairs individuals in the

population, generating a random crossover for each pair and exchanging some of their genes (codes) after the crossover with a predetermined probability.

(5) Mutation operation

The mutation operation is a modification of the genetic algorithm, in which an individual (chromosome) is randomly selected in the population. A mutation operation is performed with a certain probability to randomly change the value of some genes in the binary coded chromosome, e.g. from 0 to 1 or from 1 to 0. The mutation operation may create new individuals.

(6) Elimination of individuals

Once the crossover and mutation operations are completed, we rank the fitness of individuals (chromosomes) by the fitness function. According to the idea of selection of the fittest, less adapted chromosomes are eliminated at a certain rate.

After completing the above steps, we derive the offspring population that will enter the next generation, and return to step 1 to complete the next iteration until we find the global optimal solution or quasi-optimal solution.

RESULTS AND DISCUSSION

Experimental Results

Firstly, we conducted correlation analysis on the indicators. By observing the correlation matrix and the heat map, the lighter color of the heat map indicates the stronger positive correlation, we found that the correlation between the indicators constructed by data mining is strong, and we used PCA to retain the eight principal components whose cumulative contribution exceeds 95%.

The heat map (Figure1) shows that the correlations among the retained principal components are completely eliminated after the dimensionality reduction by PCA.

Table 3
Confusion Matrix for Accuracy of Multivariate Logistic Regression Model

	Prediction				
Ground truth	1	2	3	4	Accuracy
1	19	6	2	0	70.4%
2	3	29	6	0	76.3%
3	2	12	20	0	58.8%
4	0	0	0	24	100.0%
Percentage	19.5%	38.2%	22.8%	19.5%	74.8%

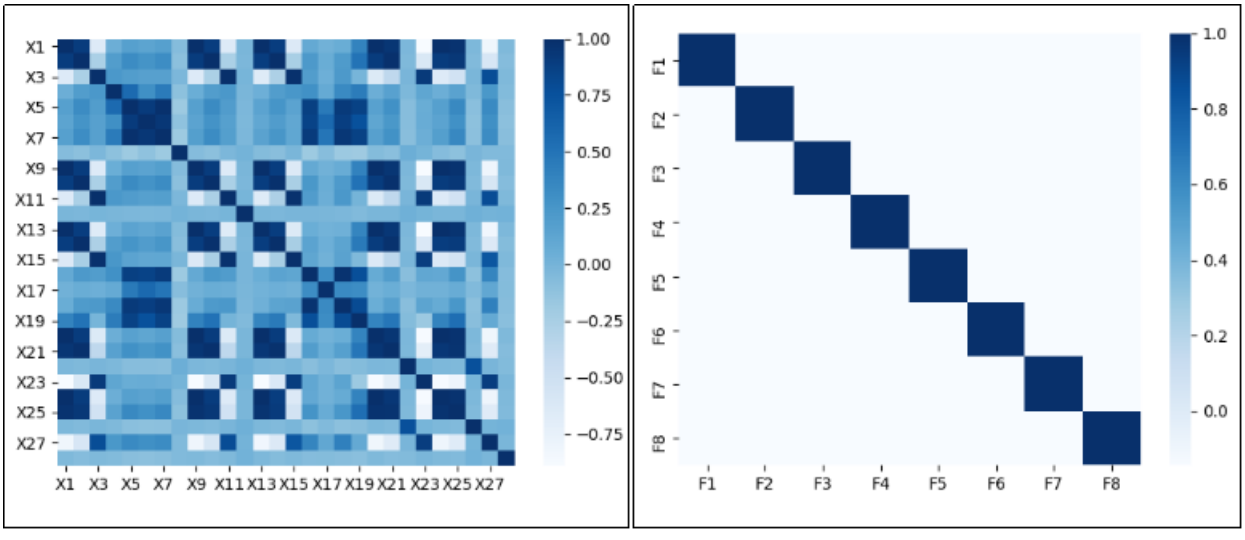

Figure 1 Heat Map of Correlation between Data Mining Metrics Compared with the Heat Map of Correlation between Principal Components after Dimensionality Reduction

Then we compute the scores of each new principal component for each sample, and the scores of each principal component for 123 companies were used as training input to the multivariate logistic regression model. The confusion matrix is presented in the Table 3.

From the above table, we can see that the multivariate logistic regression classification model achieves an accuracy rate of 75%, and also good at classifying the enterprises with a credit rating of D (enterprises that will not be lent). The model performs poor at the enterprises with a credit rating of B and C. The difference in interest rates between B and C enterprises at the time of bank credit is not significant, so the effect is not significant.

After that, we used a logistic regression model to predict the default situation of enterprises, and derived the weights of each feature and calculated the default rate. The results of the weights of each feature are as follows.(Table 4)

Table 4
Parameter Estimation Table

Feature	B	B(Z-Score)	Feature	B	B(Z-Score)
Number of invoices voided	0.23	-13745.347	**Sales tax**	0[b]	0
Invoice invalidation rate	-.024	-1152.189	**Number of suppliers**	-.035	-16.391
Profit	0b	0	**Stable suppliers**	.008	6.31
Actual taxes paid	-.060	-27.791	**Number of distributors**	-.155	-7.489
Turnover	-.006	-37.607	**Number of stable dealers**	.011	0.346
Total input tax	.039	11.226	**Total input tax for 2018**	-.012	-372.413
Profit before tax	.008	45.561	**Total outgoing price tax for 2018**	.013	261.011
Number of negative invoices	-12.342	-0.947	**Profitability in 2018**	.001	0.7
Valid invoices	.269	13963.075	**Amount of earnings for 2018**	0[b]	0
Valid tickets without negative	.034	1392.893	**Total input price tax for 2019**	-.003	-64.577
Incoming expenses	0b	0	**Total outgoing price taxes for 2019**	-.016	-244.51
Sales revenue	2.496	1.107	**Profitability for 2019**	-.002	-1.251
Profit margin	0b	0	**Profit for 2019**	0[b]	0
Income taxes	0b	0	**Annual sales growth**	.422	1.327

Intercept	-1.964	67.726

We found that the flow of business operating funds (incoming expenses, outgoing income and other related indicators) has a more significant impact on the default rate.

Finally, the classification results are as follows.(Table 5)

Table 5
Logistic Regression Default Rate Prediction Confusion Matrix

	Observation		
Ground truth	0	1	Accuracy
0	90	6	93.8%
1	7	20	74.1%
Percentage			89.4%

As seen from the table, the dichotomous logistic regression model works better, so the probability given by the maximum likelihood function of the logistic regression model can be used as the basis for calculating the default rate of enterprises.

Then we input the prediction set into the model. We use the multivariate logistic regression classification model to classify the enterprises in the prediction set into four different credit ratings, then 24 enterprises with D rating are not granted credit, and the remaining 236 enterprises are input into the dichotomous logistic regression model to calculate the default probability. The risk limit is obtained based on the business income of the remaining enterprises in the previous year, and is solved by substituting the improved risk measurement model based on Credit Risk+ with a 0-1 nonlinear programming decision model. The following figure is an iterative convergence graph plotted by GA method of Optimization toolbox in Matlab.

As can be seen from the above figure 2, the algorithm has fully converged at the 41st iteration and achieved a quasi-optimal solution of 8329.8194. The exact value obtained using Matlab's intlinprog command is 8349.09. These two solutions are quite close. If we consider sacrificing more time complexity and expanding the number of individuals in the population, then it is possible to obtain the exact solution.

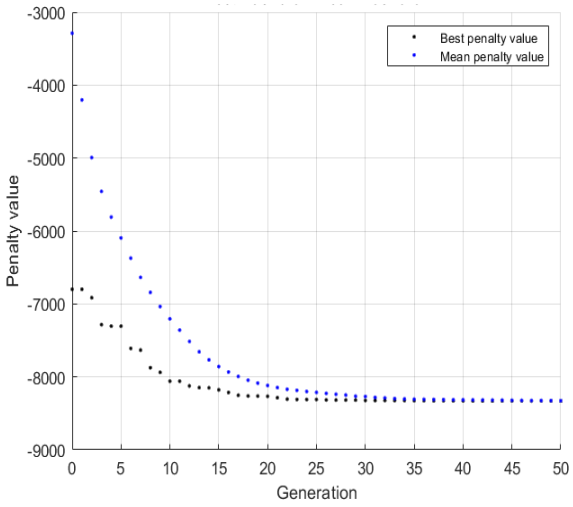

Figure 2 The Convergence Plot

Results Analysis

We analyzed the correlation between the 28 indicators and the bank's credit rating, and found that the correlation between the number of stable dealers and the credit rating was 0.4, which was moderately positive. The correlation between the number of dealers, the number of negative invoices and the credit rating was more than 0.3, which was weakly positively correlated. In addition, the correlation between invoice invalidation rate and credit rating is -0.36, which is a weak negative correlation. From the perspective of the company, it is clear that if they want to improve their reputation rating, they should start by increasing the number of transactions, strengthening the downstream relationship in the supply chain, and controlling the invoice invalidation rate. This also reflects the fact that our credit rating model focuses on the downstream quality of a company's supply chain. From the bank's perspective of examining the default rate of the enterprise, the correlation between the invoice invalidation rate and the default rate is 0.362 with a weak positive correlation, and other indicators with weak correlation are also correlated with the number of suppliers and the number of invoices, which shows that the default rate of the enterprise does not directly depend on its operation scale or profitability, but is closely related to the quality of its supply chain.

Three enterprises with different ratings in the same industry were selected from the data set and put into the model for analysis: E81 with credit rating A, E57 with credit rating B, and E77 with credit rating C. With a fixed annual interest rate of 6%, the model calculates their possible default rates, as well as the risk limits and expected returns of a single loan are shown in Table 6.

Table 6
Partial Results of the Risk Quantification Model Calculations are Shown

Enterprise rating	Probability of default	Risk limit	Expected return
E81(A)	0.09772	30.92408	-1.34777
E57(B)	0.01311	945.75585	43.60256
E77(C)	0.00225	10.92124	0.62923

Based on the results, it is not the case that the higher the credit rating of the enterprise, the greater the expected return. Instead, the credit rating of the enterprise can only objectively reflect the business situation and the quality of the enterprise's supply chain, as shown in the table above, if the credit rating is granted according to the credit rating, it may cause losses. On the contrary, if the credit enterprise is selected according to the expected return, then the overall credit asset portfolio will have a higher return. Taking 123 enterprises in the training pool as an example, if credit is granted according to credit rating when there is sufficient capital, the expected return is calculated to be 66,016,000 RMB, while the expected return is 83,490,900 RMB after solving the 0-1 planning equation by our optimization algorithm, which is an increase of 17,480,300 RMB, or 26.5%, while the exposure to risk capital decreased from 146,865,800 to 141,039,800.

CONCLUSION AND FUTURE WORK

Basing on the analysis of the above results, we suggest that the tobacco companies for business met with resistance to provide perfect water bill, can optimize the supply chain structure, increase the stability of the supplier and the distributor, should strictly control the quality to reduce invalid orders to improve the credit rating, in order to enhance the possibility of access to bank loans, and lower loan interest rates.

In this paper, we designed a planning algorithm consists of feature engineering, classification model, regression model, and optimization model. We modified the Credit Risk+ model and use a PCA-based logistic regression model to predict the credit default rate of MSMEs. We replaced the number of defaulted loans that obey Poisson distribution in the Credit Risk+ model. Not only does it make it more practical and accurate in credit default rate prediction for MSMEs, but also the improved risk measure model can be solved as

the objective function of a 0-1 linear credit decision model by a simulation optimization algorithm such as improved genetic algorithm, thus helping banks to make credit decisions that can maximize expected returns with limited available investment capital.

Our research successfully shows the possibility of using probabilistic calibration of classification results from highly accurate machine learning classification models to calculate the probability of default in risk metric models. With the rapid development deep learning techniques, the dichotomous and even multi-classification models become more and more accurate. It makes the classification probabilities calculated from their classification results after probabilistic calibration increasingly accurate. Under this condition, enterprises do not have to provide a wide range of supporting information, and banks do not have to conduct due diligence on enterprises. The credit rating, default probability and expected revenue of MSMEs can be calculated by simply inputting certain verifiable information, such as invoice flow, into the model for data mining and machine learning, and supporting the solution of the bank's credit strategy through simulation optimization algorithms. This provides a method to calculate the default rate of MSMEs without credit history and credit rating at a certain time in the future. It also eliminates the need for enterprises to provide numerous supporting documents and banks to conduct due diligence on these enterprises, which intuitively reduces credit cost and manual risk, simplifies the credit approval process, and optimizes the credit strategy. Our study helps the banking industry to reduce credit risks, thus enhancing the confidence of the banking industry in the credit business of MSMEs, as well as effectively improving the systems and processes of the banks' credit business. It enhances the banks' risk control ability and operation level in the enterprise credit business, and enables the banks to gain advantages in the increasingly competitive market. At the same time, we also put forward suggestions and measures to solve the financing difficulties of MSMEs based on the model analysis results from the perspective of banks' credit preferences, and reduce the difficulty of credit for MSMEs. It brings a win-win situation for both MSMEs and the banking industry, and has high promotion value in the policy context of developing inclusive finance.

Finally, we propose some directions for future work. In fact, most of the enterprises that lend to banks have credit ratings, we substitute credit ratings into the logistic regression equation and find that the success rate of predicting the default rate of enterprises reaches 100%. The enterprises that are judged to have no default risk at this time have a default probability lower than 0.000001, which is exactly in line with the traditional Credit Risk+ model has a very small probability of default for companies that are not correlated with each other. Therefore, when dealing with credit-rated MSMEs, we can directly calculate their maximum credit limits and then use the traditional Credit Risk+ model to calculate the portfolio default probability and loss expectation of credit assets. We also need to study a more realistic situation, that is, how to calculate the portfolio default probability and loss expectation of credit assets when both non-credit-rated enterprises and credit-rated enterprises lend to banks. In addition, to minimize the risk and maximizes profit, the model of 0-1 planning is slightly rigid. We propose to use linear planning instead of 0-1 planning in the future, and solve it by simplex method and simulated annealing method, to formulate the credit limit of each enterprise more precisely and ensure that the financial inclusion policy can benefit more MSMEs. Moreover, our model only takes into account the internal situation of enterprises, but we need to combine it with the macro situation of the external economic environment to make more accurate predictions on the probability of default rate or credit rating changes. We plan to integrate the variables into the macro-simulation model developed by McKinsey, and calculate the credit rating state transfer matrix with more data. This enables us to design a more practical model that can measure default risk and credit spread risk using macro-simulation model and risk measurement technique model.

Conflict of Interest

We all declare that we have no conflict of interest in this paper.

Acknowledgements

I would like to thank Associate Professor Zhang Yong and Professor Bao Xiaobing for their encouragement and guidance, Dean Lu Kezhong and Associate Professor Wang Haibin for their support of my research work, and Ms. Li Hui who is willing to accompany me to do my favorite mathematical modeling work.Finally, I want to thank my mother for everything she has done for me.

Author Declaration

This research is not funded by any organization related to tobaccoproduction.

References

1. Zou W, Ling JH. Financial Inclusion and Financing Constraints of MSMEs-Empirical Evidence from Chinese MSMEs. *Finance and Economics Series*. 2018;(06):34-45.
2. Li W. Comparative analysis of modern credit risk management models. *Financial Economics*.2016;(04):164-165.
3. Levy DT. Chaloupka F. Lindblom EN. Sweanor DT. O'Connor, RJ. Shang C. Borland R.The US Cigarette Industry: An Economic and Marketing Perspective. *Tobacco Regulatory Science*.2019;Mar:156-168.
4. .D'Angelo H. AyalaGX. Gittelsohn J. Laska MN, Sindberg LS. Horton L. Kharmats AY. Ribisl KM. An Analysis of Small Retailers' Relationships with Tobacco Companies in 4 US Cities.*Tobacco Regulatory Science*.2020;JAN:3-14.
5. Jiang QY.Tobacco enterprise financial risk analysis and prevention strategy discussion.*China International Business*.2017;(19):216.
6. Research Group of the Business Administration Department of the People's Bank of China. The main modes, problems and policy suggestions of big data finance supporting small and micro enterprises financing. *Beijing Financial Review*. 2019;(04):143-153.
7. Wang JY. Financial Crisis Warning Research of Listed Company in China Based on the Perspective of Cash Flow. Tianjin University. 2013.
8. Guo Y,Zhang LG, Liu J. Study on credit risk measurement model for small and medium-sized enterprises--an empirical analysis based on Shandong Province. *Dongyue Series*. 2013;34(07):58-61.
9. Ma W. Prediction of credit overdue behavior based on data mining algorithm. Shanxi University. 2020.
10. Lv DH, Zhang WK. Research on credit risk influencing factors of agricultural land management rights mortgage and its measurement - estimation based on CreditRisk+ model. *Journal of Huazhong Agricultural University (Social Science Edition)*. 2018;136(04):143-153+179.
11. Xie R, Wang HM, Wang YX, Zhu JM. Evaluation of credit risk and credit strategy planning for micro and small enterprises based on factor analysis. *Journal of Natural Sciences*. Harbin Normal University. 2020;36(06):53-63.
12. Zhou ZL, Li C. Credit Risk Evaluation and Credit Strategy Planning of Small and Micro Enterprises Based on Factor Analysis [10]. *Modern Trade Industry*. 2021;42(06):112-113.
13. Wang C, Han Z, Xu ZQ, Wang L. A study on credit strategy of small and medium-sized enterprises based on logistic regression. *Business News*. 2021;(02):91-92.