ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

Integrated Algorithm-based Credit Risk Assessment and Credit Decision Guidance

Yating Peng[1, *], Wenjun Li[2]

[1]School of Information and Computer Engineering, Northeast Forestry University, Harbin, 150006, China
[2]School of Computer Science and Engineering, Changchun University of Technology, Changchun, 130102, China
**Corresponding author: pengyating1129@163.com.*

Abstract: *This paper investigates helping banks to assess the strength and creditworthiness of firms based on limited data on MSMEs, which leads to credit risk assessment and lending decisions. In the data of enterprise invoice with complete information, four indicators of annual profit, annual profit rate, annual profit growth rate and reputation rating were selected as independent variables (the last one symbolizes enterprise reputation, the other three symbolize enterprise strength), and whether to default or not as the dichotomous dependent variable, six single classifiers such as LDA, KNeighborsClassifier, NB, SVM, LR and MLP classifiers were trained with one The integrated classifier was finally selected through performance evaluation. For the data with missing reputation ratings, we use three indicators such as enterprise capital turnover, order completion rate, and effective invoicing rate from the data with complete information as features and reputation ratings as labels, and fit a classification prediction model by fisher linear discriminant, and then make predictions for the data with missing information. After excluding firms with a credit rating of D and a high risk factor, we allocated the loan amount to the total amount of the three types of firms according to their size. After fitting the equation for the relationship between APR and customer churn rate, a multi-objective nonlinear programming model was developed to solve the credit strategy with minimum customer churn rate and maximum profit as the objectives and loan amount and APR as the decision variables.*

Keywords: Cr*edit risk assessment, credit strategy, integrated classifier, nonlinear programming, linear discriminant analysis*

1. Introduction

In the lending business of banks, since MSMEs are relatively small and also lack collateral assets, banks usually provide loans to strong enterprises with stable supply and demand relationships based on credit policies, information on their transaction notes and the influence of upstream and downstream enterprises, and can offer preferential interest rates to enterprises with high creditworthiness and low credit risks.

It is a pressing problem to extract information about the strength and creditworthiness of MSMEs based on the limited information about their ticket bureau, and then assess the credit risk and determine whether to lend and credit strategies such as loan amount, interest rate and maturity. Previously, Yanqiu Xu[2] has proposed SVM and Yuan-Yuan Huo[3] et al. have proposed models such as Probit for credit risk measurement. And Logit has been used to predict credit risk in the previous research work of Jinggui Zhang[4].

This paper analyzes the solvency and credit risk of the enterprises based on their invoices and loan records and helps banks to develop the best credit strategy.

2. Data processing

The data in Annex 1 (invoice data of 123 enterprises with complete information) are split into sales invoice data part1-out.csv, input invoice data part2-in.csv, credit rating part1_crerate.csv, and default record table_R.csv. The data in Annex 2 (invoice data of 302 enterprises with missing credit rating and default record) data are organized accordingly as part1-out.csv, part2-in.csv, etc.

ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

As we subsequently study invoice amount information by year (12 months), about 32% of the sample firms in the available data are missing data for 2016 or 2020, and we exclude data for these two years to facilitate the study. A small number of firms are missing individual monthly data for 2017 to 2019, which we add by reference to the proportion of the current month's amount in the annual total for the adjacent year; if data for the same month is missing for 3 years, we exclude that sample; and finally collate the data for the year.

The data were processed using mysql software.

3. Risk factor prediction model

3.1. Selection of indicators

Referring to the credit risk prediction studies by Guo Yan [5] and Xiao Beiming [6], we organize the credit risk assessment indicators and select the following variables.

The creditworthiness rating, G, of an enterprise intuitively reflects the credibility of its repayment over a certain period of time. The four grades A, B, C and D in the data are numerically assigned as 1, 0.67, 0.33 and 0 respectively.

The firm's average annual profit, A, represents the number of profits the firm makes in a given number of years.

$$A_i = \frac{1}{3}\sum_{n=2017}^{2019}(Annual\ input\ amount_i^n - Annual\ output\ amount_i^n) \quad (1)$$

(1) The firm's average annual profit margin, B, represents the profit earned by the firm per unit of cost of goods sold in a given year.

$$B_i = \frac{1}{3}\sum_{n=2017}^{2019}\frac{Annual\ input\ amount_i^n - Annual\ output\ amount_i^n}{Annual\ input\ amount_i^n} \quad (2)$$

(2) The average annual profit growth rate of an enterprise, C, reflects the change in the operating profit of an enterprise in a given year.

$$C_i = \frac{1}{2}\sum_{n=2017}^{2018}\frac{B_i^{n+1} - B_i^n}{B_i^n} \quad (3)$$

(3) The risk factor R, the default rate of the firm after the loan, the dependent variable of the model, is calculated in the model as the probability of being classified as a default, defined in the domain [0, 1].

3.2. Performance measurement

We decided to use confusion matrices to observe model performance due to the heavily skewed dataset and the small number of both positive and negative samples; in addition, the prediction accuracy (ACC) was similarly employed and they are defined below.

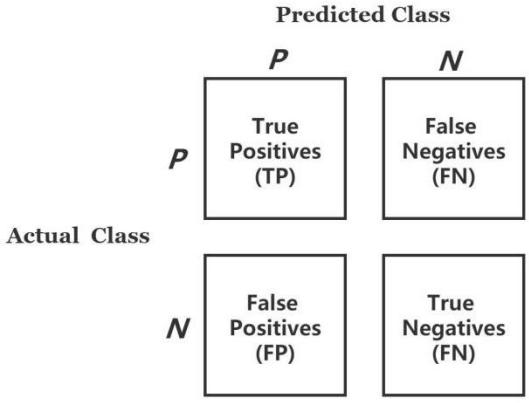

Figure 1: Confusion matrix

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

3.3. Single classifier

We compared six classification methods:linear discriminant classifier, nearest neighbor classifier, plain Bayes, SVM, logistic regression classifier, and MLP classifier, all of which are implemented by python.

Table 1: Predicting performance of LDA, KNN, NB, SVM, LR, and MLP on the test set about the risk of corporate credit-default

method	ACC	confusion matrix
LDA	0.9032	[26 2] [1 2]
KNeighborsClassifier	0.8709	[27 1] [3 0]
SVC	0.9032	[28 0] [3 0]
MLP	0.9032	[28 0] [3 0]
LR	0.8387	[25 3] [2 1]
NB	0.4516	[11 17] [0 3]

LDA, SVC, and MLP all perform relatively well, but the small sample size of the dataset may have overfitting issues.

3.4. Bagging integrated classifier

Bagging is a typical integrated learning approach that can reduce overfitting by reducing the variance of the results while improving its accuracy and stability. **Grid search is** an exhaustive search method for specifying parameter values, and the optimal learning algorithm is obtained by optimizing the parameters of the estimation function by cross-validation methods.

Han Xu[7] proposed a model based on Gaussian mixture Model (GMM) and SMOTE (Synthetic Minority Oversampling) Technique combined with GSRA to carry out Under sampling for most samples, in order to improve the prediction performance of classification algorithm in credit imbalance data.Bagging is used to integrate multiple decision trees and obtain the optimal number of decision trees (n_estimators) and the maximum number of samples per random sampling (max_samples) by grid search (5-fold cross-validation). The above process is implemented by python.

Table 2: Predicting performance of ensemble classifier on the test set about the risk of corporate credit-default

method	ACC	confusion matrix
bagging	0.9355	[27 0] [2 2]

The model performs well and improves accuracy to some extent.

4. Credit rating prediction model

With reference to the existing online creditworthiness evaluation index system of enterprises[8] and the study by Yueh-Ping Wang et al[9], the credit risk assessment indexes were organized and the following variables were selected.

(1) The probability that a business has an average annual order fulfillment rate of D, no customer returns for refunds, cancellations, etc., and a positive invoice denomination.

$$D_i = \frac{1}{3} \sum_{n=2017}^{2019} \frac{\text{Annual positive invoice number}_i^{n+1}}{\text{Annual effective invoice number}_i^{n}} \quad (5)$$

(2) The average annual effective invoicing rate of a business E, the probability that a business is invoicing legally and effectively.

ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

$$E_i = \frac{1}{3} \sum_{n=2017}^{2019} \frac{\text{Annual positive invoice number}_i^{n+1}}{\text{Annual invoice number}_i^{n+1}} \quad (6)$$

(3) The average annual economic size of an enterprise, F, expressed as the total annual turnover of the enterprise, reflects the total economic volume of the enterprise.

$$F_i = \frac{1}{3} \sum_{n=2017}^{2019} (Annual\ input\ amount_i^n + Annual\ output\ amount_i^n) \quad (7)$$

(4) Credit ratings, as dependent variables, in the classification model A, B, C, D are numerically assigned the values 4, 3, 2, 1 of the nominal variable GI.

Linear discriminant analysis was performed using SPSS software, and the data from Annex 1 were fitted and predicted for the data from Annex 2 using the fisher four classifier.

Table 3: Unnormalized discriminant function coefficients

	function		
	1	**2**	**3**
D	**35.172**	**-9.674**	**-20.098**
E	**11.957**	**-.004**	**16.125**
F	**2.435**	**9.526**	**-2.227**
(constant)	**-45.136**	**9.192**	**4.724**

Table 4: Structure matrix

	function		
	1	**2**	**3**
D	.759*	-.326	-.563
E	.188	.972*	-.139
F	.509	.070	.858*

Table 5: Fisher classifier fit results for Annex 1 data

	GI	Forecasting team members				aggregate
		1	**2**	**3**	**4**	
reckoning	1	6	5	1	0	12
	2	5	12	10	1	28
	3	1	5	26	3	35
	4	1	0	4	21	26

a. 64.4 per cent of the original grouped cases were correctly classified.

The quad classifier achieved an accuracy of 64.4%. Due to the lack of a sufficient number of feature variables and the fact that in the cases of misclassification 88.9% of the samples were misclassified into adjacent categories, the accuracy was acceptable.

5. Credit Decision Planning Model

5.1. Proportion of loan amount allocated

First neither firms with too low a credit rating nor those with too high a credit risk are granted loans, so data for firms with a credit rating of D and a credit risk factor above a threshold of 0.80 are excluded from the planning model.

With reference to the study of Liu, Zhenhai [10] et al,classified enterprises into the following three categories according to their economic size F: medium, small, and micro, and classified the loan amount to total ratio according to the total ratio of economic size of the three categories.

Table 6: Proportion of loan amounts according to the economic size of the enterprise

scope	Economic size of the enterprise F	reckoning	Percentage of allocation
in	F >= 10^8	8	0.7
small	10^8 >= F >= 10^7	21	0.3
micro-	F < 10^7	60	0.1

ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

5.2. Fitting APR and customer churn rate

The variability of the three curves is small, i.e., the relationship between customer churn rate α and annual interest rate β is very similar for firms with different ratings, so the mean value of customer churn rate for the three data sets is taken and refitted.

$$\alpha = 0.659 * \log(\beta) + 2.188 \quad (8)$$

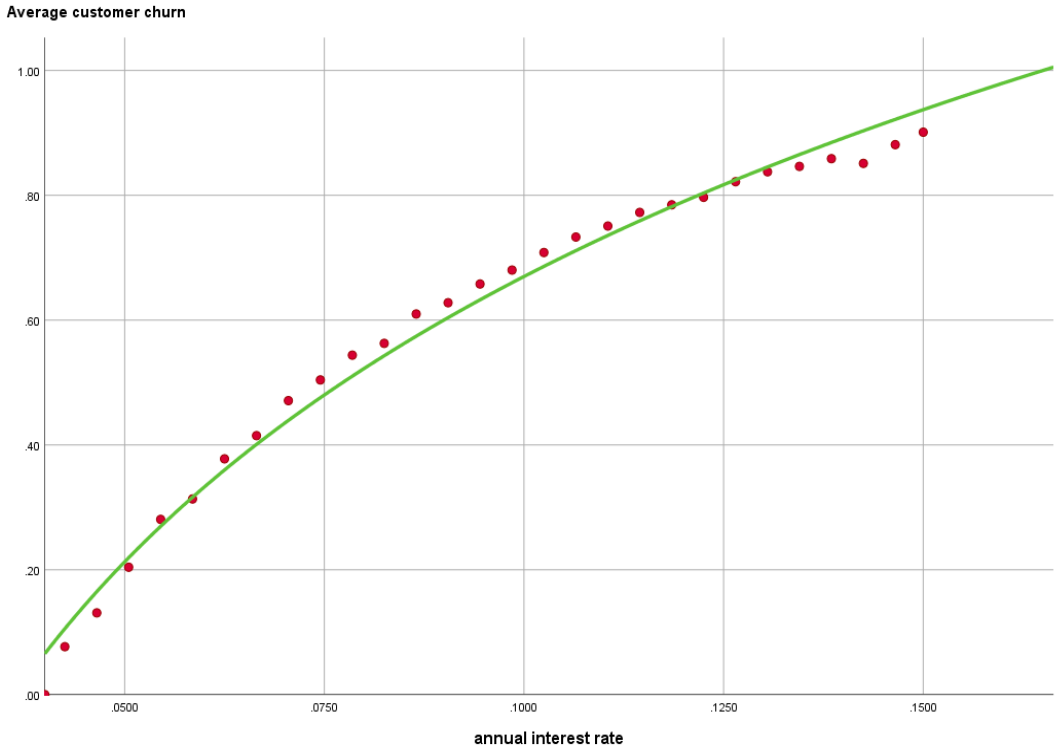

Figure 2: APR and average customer churn fit

The fitted $R^2 = 0.988$, which is a very good fit.

5.3. Multi-objective nonlinear programming models

A planning model was established in the research work of Zhao Lingling et al. [11] after AHP was used to obtain the influence weights of credit risks and possible unexpected factors on various enterprises. Establishing a multi-objective nonlinear programming model to solve the credit strategy by taking the minimum customer churn rate and maximum profit as the objective and the loan amount and annual interest rate as the decision variables.

Let there be a total of K enterprises participating in the planning model, and the enterprises are listed in descendingorder of their size, and the number of medium, small and micro enterprises are S_a, S_b, S_c, the proportion of the allocated quota to the total amount is T_a, T_b, T_c;

Total annual loans amounted to M;

The loan amount and the annual interest rate are used as decision variables,where the annual interest rate α is given by X_1, X_2, ...X_K denoted by , and the loan amount m is given by X_{K+1}, X_{K+2}, ...X_{2K} Denote;

(a) To simplify the model, assume that the revenue P and the customer churn rate D are equally important and weighted equally.

Due to the large difference in magnitudes, dividing P and D by their respective maximum values P_max (P_max=$\alpha_{max} * m_{max}$) with D_max (D_max value is 1) and summed as the objective function [12].

$$min\ f(x) = \frac{-1}{P\_max}\sum_{i=1}^{K} X_i * X_{i+K} * (1-\alpha_i) * (1-R_i) + \frac{1}{D\_max}\sum_{i=K+1}^{2K} X_i \quad (9)$$

ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

$$s.t. \quad \begin{cases} \sum_{i=1}^{S_a} X_i <= T_a * M \\ \sum_{i=S_a+1}^{S_a+S_b} X_i <= T_b * M \\ \sum_{i=S_a+S_b+1}^{K} X_i <= T_c * M \\ X_i >= 0.04, \quad i = 1,2,\ldots,K \\ X_i <= 0.15, \quad i = 1,2,\ldots,K \\ X_i >= 100000, \quad i = K+1, K+2,\ldots,2K \\ X_i <= 1000000, i = K+1, K+2,\ldots,2K \end{cases}$$

To ensure the robustness of the results, the interior-point, sqp, active-set three algorithms were used to solve the problem and interior-point gave the best results. The above procedures are implemented by matlab.

6. Conclusion

In this paper, building the risk factor prediction model, compare the performance of different single classifiers with the integrated classifier, and choose the integrated classifier as the final model. For the problem of imbalanced data samples, using confusion matrix along with ACC to evaluate the model performance; for the problem of overfitting caused by small data samples, using cross-validation (included in grid search) and bagging integration approach to reduce overfitting; finally, using grid search to select the best parameters. For the prediction of reputation ratings, the fisher classifier was used, and because the direct boundaries of the 4 ratings are not obvious, the samples misclassified into adjacent categories account for a large proportion of the total misclassified samples, and the model within a certain error range was accepted. With a few more valuable feature variables, the model accuracy should be significantly improved. Before constructing the planning model for the credit decision first that neither firms with too low credit rating nor firms with too high credit risk were granted loans, so data for firms with a credit rating of D and a credit risk factor above a threshold of 0.80 were excluded from the planning model. These firms were classified into the following 3 categories of medium, small and micro based on their economic size F and classify the loan amount as a percentage of the total according to the total ratio of the 3 categories of economic size. After fitting the equation for the relationship between annual interest rate and customer churn rate, a multi-objective nonlinear planning model was developed with minimum customer churn rate and maximum profit as the objectives and loan amount and annual interest rate as the decision variables to finally give the credit strategy for 247 enterprises in Annex 2.

The production, operation and economic performance of enterprises may be affected by some unexpected factors, and the unexpected factors often have different impacts on different industries and categories of enterprises. Taking the new crown as an example, the enterprise's repayment ability decreases and the credit rating receives an impact, the average GDP growth rate of each industry from the fourth quarter of 2019 to the second quarter of 2020 compared with the growth rate of the same period in previous years can be considered to obtain the disturbance coefficient H of the epidemic on the economic returns of different industries, which is added to the credit rating model after determining the weights based on expert scoring.

Biographical notes

Yating Peng is an undergraduate student of Northeast Forestry University, China. Her research interest focuses on Big Data.

Wenjun Li is an undergraduate student of Changchun University of Technology, China. His research interest focuses on Software Engineering.

References

[1] Han Liang. Research on credit risk identification and prevention strategies of small and micro

ISSN 2616-5902 Vol. 3, Issue 11: 103-109, DOI: 10.25236/AJBM.2021.031120

enterprises of INDUSTRIAL and Commercial Bank of M City [D]. Southwest university of science and technology, 2019. DOI: 10.27415 /, dc nki. GXNGC. 2019.000071.
[2] Yanqiu Xu. Research on the model of enterprise credit loan risk assessment system based on support vector machine [D]. Shanghai International Studies University, 2017.
[3] Yuanyuan Huo, Tianyi Yao, Jiang Li. Research on Credit Risk Measurement of Chinese Manufacturing Enterprises based on Probit Model [J]. Prediction, 2019, 38 (4): 76-82.
[4] Jingui Zhang, Hou Yu. Empirical analysis of credit risk of small and medium-sized enterprises based on Logit Model [J]. Friends of Accounting, 2014(30): 40-45.
[5] Yan Guo, Liguang Zhang, Jia Liu. Research on credit risk measurement model of small and medium enterprises--an empirical analysis based on Shandong Province [J]. Dongyue Series, 2013, 34(07): 58-61. DOI:10.15981/j.cnki.dongyueluncong.2013.07.001.
[6] Beiming Xiao. A study on credit risk prediction model combining macro and micro analysis [J]. Financial Forum, 2004(10): 57-61+63. DOI:10.16529/j.cnki.11-4613/f.2004.10.010.
[7] Han X. Classification and feature selection of imbalanced data for credit default risk [D]. Tianjin University, 2019. DOI: 10.27356 /, dc nki. Gtjdu. 2019.001573.
[8] GB/T 39887-2021, online reputation evaluation index system of enterprises [S].
[9] Yueping Wang. Analysis of evaluation indicators of corporate reputation [J]. Science and Technology Entrepreneurship Monthly, 2006(02): 146-148.
[10] ZH Liu, JZ Yang. Firm size based credit decision support: an experiment using external credit ratings [J]. Financial Research, 2008(08): 177-185.
[11] Lingling Zhao, Jin Chen, Xiaoying Li, Jiaming Zhu. Credit Strategy analysis of Small, medium and micro enterprises based on K-means Clustering Analysis [J]. Science Journal of Normal University, 201, 41(09): 14-20.