Computing and Informatics, Vol. 40, 2021, 145–168, doi: 10.31577/cai_2021_1_145

CREDIT RISK ASSESSMENT OF BANKS' LOAN ENTERPRISE CUSTOMER BASED ON STATE-CONSTRAINT

Renjing Liu, Xuming Yang, Xinyu Dong, Boyang Sun

School of Management, Xi'an Jiaotong University
No. 28, Xianning West Road
Xi'an, China
e-mail: 18223205757@163.com

Abstract. Commercial banks are facing increasingly complex enterprise loan customers and businesses. It is important for banks' enterprise loan business to efficiently assess credit risks. Our study builds an enterprise credit risk assessment model based on the state and constraint of bank and customer, and get empirical researches with RF, SVM and DT algorithms. The results show that our model has excellent performance with accuracy 99 % and great characteristic importance in the evaluation of enterprise credit risk. The study can provide important decision-making reference for bank loan business and enrich the theoretical system of bank credit risk research.

Keywords: Credit risk assessment, state and constraint, enterprise loan, machine learning

Mathematics Subject Classification 2010: 68U35

1 INTRODUCTION

As it is universally acknowledged that commercial loan is the core source of bank profit and takes the most important position in banking business. However, in recent years, the non-performing loan ratio of commercial banks has been constantly on the rise. Taking China as an example, according to the announcement of China Banking Regulatory Commission in August 2020, the non-performing loan ratio

of China's commercial banks has risen to 1.94 %, with an increase of 7.2 % over last year. As one of the core subjects of commercial loans, enterprises have complex market economy network, huge loan amount, increasingly difficult asset quality management, and high credit risk. Enterprise's loans are known as the saying "one loan losses can lose nine loan profits". At the same time, in the current financial environment with the integration of Internet and digital technology, the bank's traditional business model that used to rely on the expansion of corporate credit scale to increase profits has squeezed the profit space of banks and accumulated a large number of customer risks [1]. With the continuous innovation and development of financial technology and the accumulation of massive customer data information in the banking industry, it has become one of the most urgent and important tasks for commercial banks to effectively control risks by means of digital technology based on the digital transformation of the banking industry driven through big data [2]. It is of important value for commercial bank credit business development to combine financial technology and big data of banks to get scientific and efficient evaluation of enterprise credit risk, which can help bank timely and objective assessment of enterprise customers and credit conditions, also provide more powerful decision support for loan business.

The integration of big data analysis and other new information technologies into the financial field has triggered a new round of digital reform in the financial industry that attracted a lot of scholars' interest in the research of corporate loan credit evaluation. Using the real-time data of bank customers to assess the credit risk of enterprises through big data analysis [3, 4] and machine learning [5, 6] methods are also increasingly intensified. At present, relevant researches are mainly carried out from two aspects. Firstly, based on the in-depth study of enterprise credit risk assessment model, the evaluation model was constantly improved and optimized by adding or introducing characteristic factors which affect enterprise credit, so as to improve the performance effect of assessment. For example, Minnis and Sutherland added tax and other indicators into the model construction to improve the effectiveness of the model for predicting corporate default risk [7]. However, too many characteristic indicators can cause the model to suffer from "dimensional disaster" [8]. In this regard, Tong's improvements put forward the LSOMAP-RVM credit model to evaluate the credit risk of domestic listed companies and solve the high-dimensional problem through algorithm and index fusion [9], but its application scenario was extremely limited. Only focusing on extraction of feature elements and optimization of credit model will, to a certain extent, result in all "preferences" of certain aspects such as enterprise state [10], managerial ability [11] overall industry characteristics [12] in the evaluation results, and fail to comprehensively evaluate enterprise credit risk.

Secondly, based on the research on the improvement of enterprise credit risk assessment methods, some machine learning algorithms were improved to overcome data and algorithm problems in enterprise credit assessment, such as data imbalance [13] and algorithm stability [14]. For example, Tian et al. built a new fuzzy set and the most advanced credit risk assessment algorithm model using the kernel-free

QSSVM basic model to deal with information label error [15]. Bu et al. created a new mixed information method, using mixed integrated information to predict enterprise credit risk, and demonstrated its advantages in short-term credit assessment [16]. Huang et al. improved the robustness of the probabilistic neural network model by determining the dimensions in advance [17]. The research based on the improvement of evaluation method can solve the data processing problems in the practice of algorithm evaluation, but the internal relationship mining of enterprise credit characteristic indexes are not enough.

Both the model optimization and the algorithm improvement in the existing research on enterprise credit risk evaluation focus more on the perspective of financial market overall credit risk and enterprise financial operation direction of the microcosmic perspective, but pay few attention to the enterprise and the bank individual state and inherent relationship, which can lead to some specific "bias" and "distortion" for credit evaluation results. Meanwhile it will also result in data structure problems such as unbalanced data. However, the generation of enterprise credit risk is not only closely related to the operation state of enterprises, but also correlated with the operation state and risk control ability of banks [10, 18]. Different from the previous single enterprise credit evaluation model of customer perspective, this paper analyzes the internal connection and restriction relationship between banks and their enterprise customers, and the model of enterprise credit risk evaluation is constructed based on the respective state and constrains of both bank and enterprise, and the SMOTE sampling method is used to overcome the imbalance data. Then we use the random forests and support vector machine (SVM) algorithms to evaluate the customer's credit state of the enterprise. The results show that the credit evaluation model based on the perspective is of high rationality and credibility. Compared with other scholar's researches, our relevant parameters all have 10 % to 20 % improvement, and data imbalance problem is solved well. The conclusion of the study can provide a certain bank loan management decision-making reference, at the same time enrich research perspectives and research model of enterprise credit assessment.

The rest of our article is structured as follows. We firstly introduce the algorithm basis used in our study in Section 2, including three machine learning algorithms and SMOTE algorithm. Next, based on the bank constraint theory, we construct a credit risk assessment model from both sides of the bank and the enterprise in Section 3. Then we get an empirical analysis of our enterprise credit risk assessment model with three machine learning algorithms and SMOTE algorithm with the data of a commercial bank in Section 4. Finally, in Section 5, we offer some concluding remarks.

2 ALGORITHMS BASIS

With the development of bank digital transformation and the generation of massive data, it has obvious advantages for machine learning algorithms in complex rela-

tional data analysis [19]. Machine learning can be divided into supervised learning and unsupervised learning. In this paper, the classification algorithm with supervised learning is selected. According to research of Choi et al. about all kinds of machine learning algorithms [6], meanwhile thinking about the bank and customer data missing, imbalance and associated features and so on, we finally choose random forest (RF) and support vector machine (SVM) algorithm, to evaluate the credit risk of bank corporate loans, and select a decision tree algorithm for supplementary research. The following is a brief introduction of these algorithms.

2.1 Random Forest Algorithm

2.1.1 The Decision Tree

Decision tree, known as classification tree, is the underlying tree structure applied to random forest algorithm. There are many kinds of decision trees, and the typical binary CART decision tree is widely used in classification problems. CART decision trees utilize Gini minimization criteria for feature selection and recursive modeling. For the training data set D with K categories, C_K represents the sample subset of class K, $|C_K|$ and $|D|$ are respectively the size of and D, then the Gini coefficient of set D is

$$Gini(D) = 1 - \mathrm{K}\sum_{K=1}^{K}\left(\frac{|C_K|}{|D|}\right)^2. \tag{1}$$

Suppose the discrete feature A is used to segment the data, then D is divided into D_1 and D_2 according to the value of A

$$\begin{aligned} D_1 &= \{D \mid A = a\}, \\ D_2 &= \{D \mid A \neq a\}. \end{aligned} \tag{2}$$

Then, under the condition of discrete feature A, Gini index of set D is:

$$Gini(D) = \frac{|D_1|}{D} Gini\left(\mathrm{D}_1\right) + \frac{|D_2|}{D} Gini\left(D_2\right). \tag{3}$$

Gini coefficient represents sample impurity degree, so the attribute with small Gini index is preferred during tree construction. While the type of attribute is greater than two, the Gini coefficient will be calculated for each combination of two categories of classification, and the classification combination that minimizes Gini index will be automatically selected. When the Gini coefficient of sample sets in nodes is less than the reference threshold, the number of samples is less than the reference threshold, or there are no more features, the algorithm converges.

2.1.2 Random Forest

Stochastic forests are essentially integrated learning derived from decision trees, proposed by Tin Kam Ho of Bell Laboratories in 1955. It works by generating

multiple single classification trees that can be learned and predicted independently. The steps to establish each classification tree are as follows:

1. Assuming that the sample set size is N, bootstrap sampling is adopted. N training samples are random and put back from the sample set to be the training set without a classification tree.
2. Assuming that the feature dimension of each sample is M, m features are randomly selected from M features as feature subset (m is far less than M), and the tree is divided from these m features each time, so as to calculate the optimal splitting mode.
3. Random sampling can ensure that there is no overfitting, so each decision tree is allowed to grow completely without pruning until it meets the predetermined requirement.

Random forest is mainly based on bagging thoughts [23], as shown in Figure 1. Every time there are replacement samples from the population N, about 2/3 of the samples forming the training set. The remaining one-third of the sample is called out of bag (OOB), and the OOB is excluded from each tree. Then the OOB error estimation model is used to estimate the accuracy and internal error of the prediction. Since OOB error rate is an unbiased estimate of random forest generalization error, its junction effect is approximately equal to k-fold cross validation. Therefore, the random forest does not need to be cross-verified, and at the same time, the model can be well generalized [24].

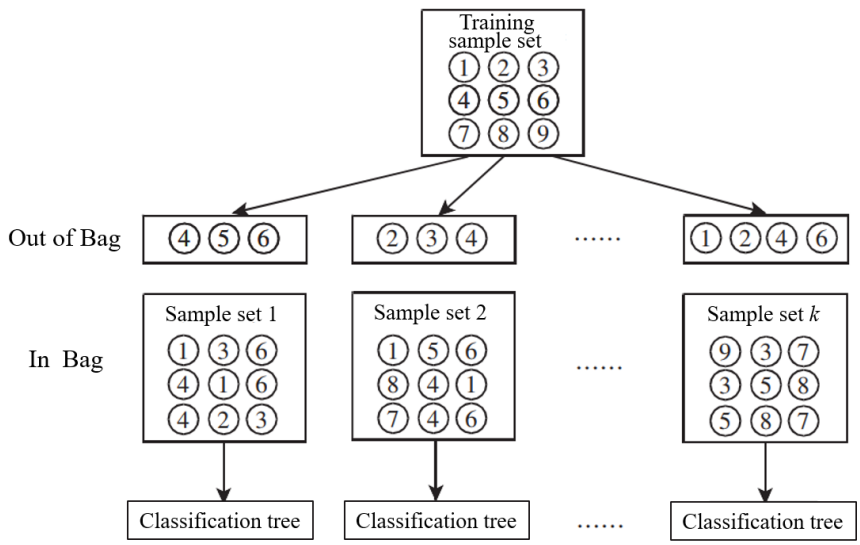


Figure 1. Schematic diagram of the Bagging process

2.2 Support Vector Machine

Support vector machine (SVM) is a machine learning algorithm based on the principle of structural risk minimization and statistical theory [25]. It is widely used in statistical regression and classification problems. For the credit classification problem of bank corporate customers, SVM has good classification performance and learning ability, meanwhile it has strong nonlinear approximation ability and can better overcome dimensional disasters, which can help process bank customer data very efficiently.

2.2.1 Linear SVN Model

First, choosing the hyperplane of the classifier in the sample space:

$$w^T x + b = 0. \tag{4}$$

The distance from any point in the sample space to the hyperplane is:

$$\mathrm{r} = \frac{\left|w^T x + b\right|}{||w||}. \tag{5}$$

Using hyperplane to classify samples:

$$\begin{cases} w^T x + b \geq +1, y_i = +1, \\ w^T x + b \leq -1, y_i = -1. \end{cases} \tag{6}$$

As shown in Figure 2, the samples are closest to the hyperplane so that the above equation holds are called support vectors. The sum of the distances from the support vectors to the hyperplane is the "interval":

$$r = \frac{2}{||w||}. \tag{7}$$

To make the maximum interval of the partition hyperplane of the classifier

$$\begin{cases} \min_{w,b} \frac{1}{2}||w||^2, \\ y_i \left(w^T x + b\right) \geq 1, i = 1, 2, \ldots, m. \end{cases} \tag{8}$$

2.2.2 Nonlinear SVM Model

However, many sample spaces are not linearly separable in reality, so a nonlinear transformation method is needed to transform the problem into a linear separable problem in the feature space of a certain dimension, so as to train linear support vector machines in the feature space of a higher dimension.

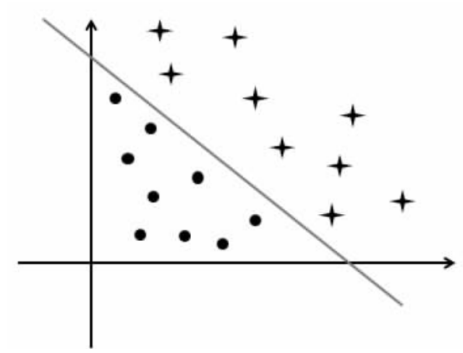


Figure 2. Schematic diagram of linear SVM model

The common transformation method is to use kernel function to transform. As shown in Figure 3, the given data set is shown in the figure on the left, and the hyperplane is divided into ellipses, which cannot be linearly divided. At this point, input vectors can be mapped into the high-dimensional feature space by introducing kernel functions, and be converted into the linear form of the figure on the right, so as to be converted into the form of linear SVM. Common nonlinear kernel functions are as follows:

1. Polynomial kernel function

$$\mathrm{k}\left(x_i, x_j\right) = \left(x_i^T x_j\right)^d. \tag{9}$$

2. Radial basis kernel function

$$\mathrm{k}\left(x_i, x_j\right) = \exp\left(\frac{||x_i - x_j||^2}{2\sigma^2}\right). \tag{10}$$

3. Sigmoid and functions

$$\mathrm{k}\left(x_i, x_j\right) = \tanh\left(\beta x_i^T x_j + \theta\right). \tag{11}$$

2.3 SMOTE Algorithm

Synthetic minority oversampling technique (SMOTE) is an improved scheme based on random oversampling algorithm. Its basic idea is to analyze minority samples and artificially synthesize new samples based on minority samples to add to the data set, which can well solve the problem of over-fitting and low model generalization resulted from simple random oversampling. The details are shown in Figure 4, and the algorithm flow is as follows.

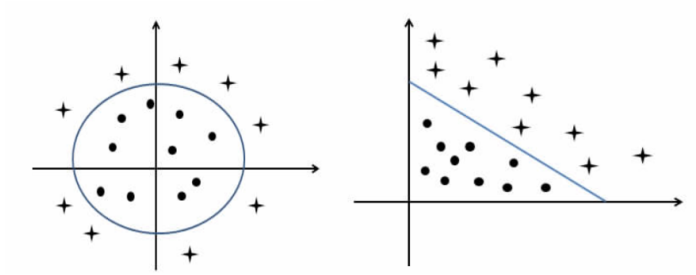


Figure 3. Schematic diagram of non-sexual SVM model

1. For each minority sample P, K nearest neighbor is obtained from the minority samples around it.
2. A minority class sample P_{bour} is selected among K nearest neighbors randomly.
3. The composite sample P_{new} is obtained by interpolation between P and P_{bour}, as shown in Formula (12).

$$P_{new} = P + rand(0,1) \times (P_{bour} - P) \tag{12}$$

where rand$(0,1)$ is the random number between $[0,1]$.

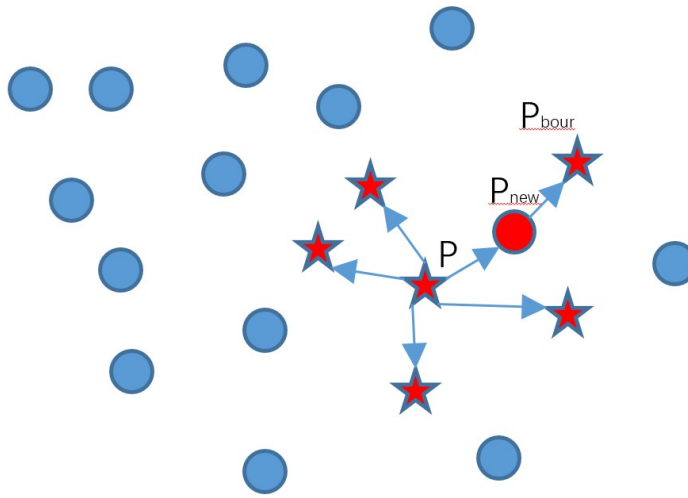


Figure 4. Schematic diagram of SMOTE algorithm sampling

3 THE CONSTRUCTION OF CREDIT RISK ASSESSMENT MODEL BASED ON STATE-CONSTRAINT

In general, the research on enterprise credit risk assessment focuses more on the assessment of enterprises as on an independent entity [3, 4, 15, 9], therefore, the influence of credit risk bearers will be ignored to a certain extent. However, in fact, the bearing party of credit risk will also affect the credit risk value of relevant enterprises. For example, when enterprises are faced with loans from banks and other financial institutions, their risk measurement is inconsistent because banks have a more complete credit system and solvency [26, 27]. Therefore, the credit risk of an enterprise is closely related to the enterprise itself, the loan bank and the whole market industry. Based on the bank constraint theory, this section constructs a credit risk assessment model from both sides – from the bank and from the enterprise side.

The theory of constraints (TOC) [28] was first proposed by Dr. Goldratt in his optimization production technique (OPT), which requires firms to establish management system to identify and eliminate constraints in the process of achieving goals. TOC emphasizes the importance of treating the enterprise as a system, considering and dealing with problems from the perspective of overall benefits. The enterprise credit risk involves not only the enterprise itself, but also the bearers of credit risk (generally referring to banks or investment companies) and the whole industry, etc. These factors will jointly act on the formation of enterprise credit risk, forming the whole integration of enterprise credit risk [27, 29].

The essence of the credit risk management of the bank's enterprise loan customers is to set a series of constraints for the enterprise, so as to reduce the probability of the loan enterprise to break the promise. For an enterprise, the higher the constraint force is, the smaller the risk of dishonesty will be. As for enterprise constraints, there are generally two aspects: self-constraints and external constraints [30]. From the subjective point of view, under the guidance of banks, enterprises will form a self-restraint system to restrict the occurrence of dishonest behaviors, including the loss of credibility at the present stage and the influence of dishonesty at the future stage, etc. These constraints can restrict enterprises' willingness of dishonest behaviors to a certain extent [31]. From an objective point of view, corporate dishonesty is subject to the relevant constraints of external banks, such as floating interest rate, overdue penalty, etc.

In addition, establishing relevant constraints and the guarantee of constraints need to depend on the operational state of both parties, and only a good operational state can ensure the operation of the constraint mechanism [32]. However when an enterprise is in a poor state of operation, even if it has no intention to break its promise, it has to break its promise because of its bad enterprise assets state. Of course, if the bank has a good risk control system, the warning is made in advance and this may help the enterprise to solve the problem, and the enterprise may still solve the loan repayment after the recovery.

To sum up, according to the research fully based on a domestic commercial bank loan business and the state-constraint theory, we have integrated the enterprise itself, risk takers, the overall industry and the environment generated by credit risk into a complete system. From the perspective of the states and constraints of both the enterprise and the bank, there are 15 characteristic indicators in 4 aspects selected from the bank's corporate loan database to construct a corporate loan credit risk assessment model, which is shown in Figure 5. The description and interpretation of the characteristics indicators is in Table 1. The following will introduce the characteristic indicators of the model.

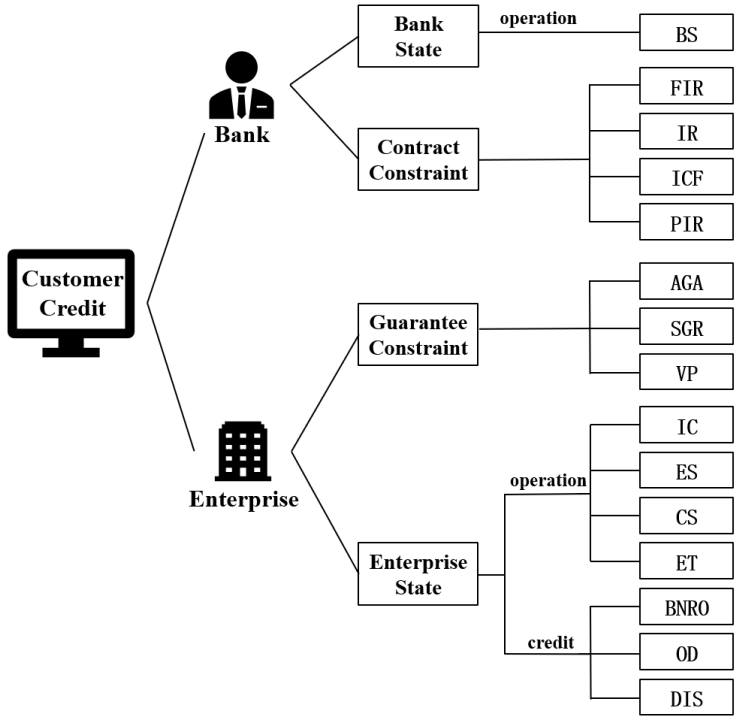


Figure 5. Credit risk assessment model based on state-constraint

3.1 Characteristic Indicators of Enterprise

3.1.1 Enterprise State

The enterprise state is an index reflecting the viability and operating state of an enterprise. The more ideal the state of an enterprise is, the less its credit risk will be, because "no enterprise wants to deliberately fight against the bank under nor-

Feature Dimension	**Characteristics**	**Characteristics Abbreviation**	**Characteristics Description**
Label	Customer Credit	CC	Dichotomous variable $CC \in 0, 1$
Bank State	Bank State	BS	Multiple categorical variables $BS \in [0, 100]$
Contract Constraint	Flexible Interest Rate	FIR	Continuous variable $FIR \in [0, +\infty)$
	Interest Rate	IR	Continuous variable $IR \in [0, +\infty)$
	Interest Calculation Frequency	ICF	Multiple categorical variables $ICF \in 0, 1, 2, 3$
	Penalty Interest Rate	PIR	Continuous variable $PIR \in [0, +\infty)$
Guarantee Constraint	Account of Guarantee Amount	AGA	Continuous variable $AGA \in [0, +\infty)$
	Security Guarantee Reliability	SGR	Continuous variable $SGR \in [0, +\infty)$
	Value of Pledges	VP	Continuous variable $VP \in [0, +\infty)$
Enterprise State	Industry Categories	ICF	Multiple categorical variables $CC \in 0, 1, 2 \cdots 17,$
	Enterprise Scale	ES	Multiple categorical variables $ES \in 0, 1, 2, 3$
	Customer State	CS	Multiple categorical variables $CS \in 0, 1, 2, 3$
Enterprise credit State	Extension Times	ET	Continuous variable $ET \in [0, +\infty)$
	Borrow New to Returnthe Old Times	BNRO	Continuous variable $BNRO \in [0, +\infty)$
	Overdue Days	OD	Continuous variable $OD \in [0, +\infty)$
	Debit Interest State	DIS	Dichotomous variable $DIS \in 0, 1$

Table 1. Description the characteristics indicators of the state-constrained enterprise credit evaluation model

mal circumstances". For credit risk assessment, the enterprise state mainly includes operation state and credit state.

The enterprise state is a comprehensive reflection index to measure the business state and financial state of the enterprise. In this paper, it mainly includes three categories of industries: enterprise scale (ES), customer state (CS) and industry category (IC). IC means the industry category in which the enterprise is located, which can reflect the overall market state of the current industry. It covers 18 main industry categories. ES represents the size of the enterprise and re-

flects the development state and overall size of the enterprise at the present stage. It is divided into four grades: large, medium, small and micro. The CS indicates the bank's assessment of the business state of the enterprise at the present stage. There are four levels of development, consolidation, adjustment and elimination.

Enterprise credit state is a comprehensive response index to measure the past dishonest behavior and credit state of enterprises, mainly including extension times (ET), Borrow New to Return the Old Times (BNRO), Overdue Days (OD) and Debit Interest State (DIS). ET is the total number of times in the past that the loan of the enterprise could not be repaid upon maturity and was approved to extend the repayment time, which can reflect the dishonest behavior and dishonest intention of the enterprise in the past. BNRO is the total number of times it fails to repay the loan on time and applies for a new loan again to repay part or all of the original loan after the loan is due (including the maturity after the extension). The larger BNRO is, the higher the credit risk of the company. OD means the total number of overdue days in the past when the enterprise loan is due and cannot be repaid on time, which can very cleanly reflect the state and degree of corporate credit default. DIS means that the company has defaulted or not on loan interest in the bank in the past, which can reflect the credit risk state of the company.

3.1.2 Guarantee Constraint

For the loan enterprise, its main constraint is the guarantee constraint in the loan. Guarantee constraint is the funds and assets paid by enterprises for guarantee when they draw loans, and it is one of the most important constraints for banks to restrict the credit loss of enterprises. Loan guarantee generally has the value equal to the enterprise loan fund, once the enterprise breaks the promise, the bank can collect the loan guarantee to repay part or the whole loan. Therefore, guarantee constraint plays an important role in alleviating enterprise credit risk. The guarantee constraint mainly includes three indexes – the guarantee amount (AGA), security guarantee reliability (SGR) and value of pledges (VP).

AGA is the total amount of the relevant account connected by the loan account, which can reflect the financial stability and reliability of the loan account. The higher the connected amount is, the more reliable the financial constraint of the loan account will be. SGR is the product of the enterprise security guarantee coefficient and the total amount of the security guarantee, which can reflect the security of the guarantee fund and the reliability of the guarantor, and high reliability can effectively restrict the occurrence of misconduct. VP means the total amount of assets mortgaged to the bank when the enterprise draws a loan, which can reflect the minimum guarantee of the loan. The existence of collateral can well restrict the enterprise to break the promise and reduce the risk of the bank's dead loan.

3.2 Characteristic Indicators of Bank

3.2.1 Bank State

Bank state (BS) in our study is the comprehensive evaluation score made by the head office of bank for each branch in the enterprise loan business, which reflects the credit evaluation ability, risk control ability and loan recovery ability of the bank in the enterprise loan business. Banks with a higher state/status of the bank have better risk control ability and means, and the cost and impact of their dishonesty is higher when facing such banks. Therefore, the state of the bank can indirectly affect the credit of the enterprise from the external environmental conditions.

3.2.2 Contract Constraint

For the bank of enterprise loan business, the main constraint it can carry out on the enterprise is the contract constraint of loan. Contract constraint is the related constraint of the contract signed by the bank and the enterprise when it grants loans to enterprises. Contract constraint has explicit legal effect and is of great importance to the dishonesty of enterprises. Contract constraints mainly include four indicators, such as flexible interest rate (FIR), interest rate (IR), interest calculation frequency (ICF) and penalty interest rate (PIR). FIR is the maximum floating interest rate provided by the bank according to the credit evaluation of the loan enterprise, which can encourage the enterprise to repay the loan on time and restrain the enterprise's dishonest behavior to some extent. IR is the loan interest rate set by the bank when it lends money to the enterprise. IR with different credit levels is different, so it can restrain the subsequent influence of the enterprise due to its bad records. ICF is the frequency of interest calculation agreed by both parties when the bank makes a loan, and faster frequency is with relative higher interest and faster reimbursement frequency. PIR means the overdue penalty interest rate that the loan enterprise does not repay the loan in accordance with the contract. PIR is generally higher than the contract interest rate, so it can restrain the enterprise from breaking the promise and generate excess penalty interest.

4 EMPIRICAL ANALYSIS OF ENTERPRISE CREDIT RISK ASSESSMENT

4.1 Data Source and Data Preprocessing

The research data in our study are from the enterprise-loan database of a commercial bank in China. After sorting and screening, 1 467 enterprise loan data are obtained, among which 119 are in default and 1 348 are in normal credit. Then based on the enterprise credit risk assessment model, the integrated screening was carried out, and each data obtained contained a total of 16 indicators, and the data set was

further preprocessed including missing items processing, numerical processing and standardized processing.

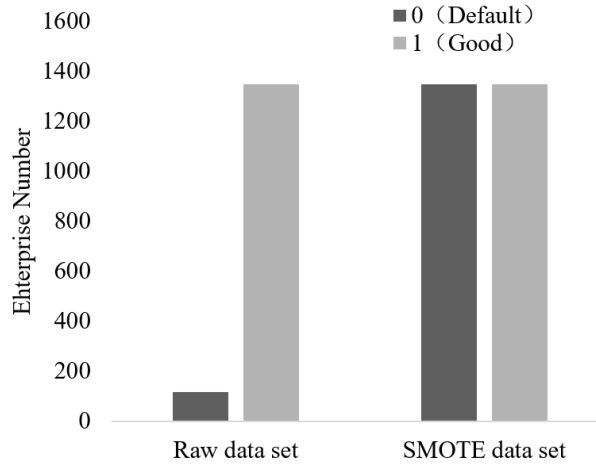


Figure 6. Credit distribution of enterprise customers

As shown in Figure 6 (left), the gap between the data set of default enterprises and normal enterprises is very large, in this case the unbalanced data problems in machine learning will affect the learning performance. Comparing with methods of under-sampling and over-sampling [33], we selected a SMOTE over-sampling algorithm dealing with unbalanced data sets, to obtain the new data set distribution as shown in Figure 6 on the right. Since the balance coefficient of the data set is lower than 15, the machine learning performance of the original data set is still reliable, so the subsequent analysis in this paper adopts two data sets for the comparative analysis.

4.2 Indicators Correlation Analysis

In this section, we used correlation analysis to test the rationality and correlation of the selection of indicators for the construction of the enterprise credit risk assessment model based on the state-constraint theory. Because the distribution of some indicators was unknown and does not show a normal distribution, meanwhile model included classification indicators, the Spearman rank correlation coefficient was finally used for the analysis.

As shown in Table 2, in the 15 indicators of the model, all indicators except the PIR were significantly related to customer credit (CC) at the 95 % confidence level, and 13 indicators were significantly related to CC at the 99 % confidence level. Significant correlation indicates that there is a significant correlation between the 15 indicators selected in the model and the dependent variable CC, which means

that the model construction is reasonable.

Index Name	**Case Number**	**Correlation Coefficient**	**Significance (P)**
CC	1 467	1.000**	0
BS	1 467	0.393**	0
FIR	1 467	0.075**	0.004
IR	1 467	0.201**	0
ICF	1 467	0.053*	0.042
PIR	1 467	0.006	0.808
AGA	1 467	0.117**	0
SGR	1 467	0.122**	0
VP	1 467	0.094**	0
ICF	1 467	0.194**	0
ES	1 467	0.137**	0
CS	1 467	0.245**	0
ET	1 467	0.230**	0
BNRO	1 467	0.279**	0
OD	1 467	0.652**	0
DIS	1 467	0.643**	0

Table 2. Spearman correlation coefficient between each indicator and CC

4.3 Selection of Model Evaluation Indexes

In machine learning, the commonly used indexes include accuracy, precision, recall, specificity, F1-value, AUC, etc. In our study, the emphasis of enterprise credit risk assessment is put to accurately identify enterprises with high credit risk, so as to provide decision-making support for bank enterprise credit. Therefore, we comprehensively selected four model evaluation indexes, including accuracy, recall rate, precision and F1-value.

Real Situation	Predicted Results	
	Case	Not the Case
Case	TP	FN
Not the Case	FP	TN

Table 3. Results of dichotomous problems refer to table

Accuracy. The accuracy rate represents the proportion of all correctly classified enterprises in all enterprise sampIs caseles, and the value interval is $[0, 1]$. The closer to 1 it is, the higher the identification accuracy of enterprise credit risk assessment is. Its calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \tag{13}$$

Recall rate. Recall rate means the proportion of enterprises that will be identified as high credit risk enterprises in the truly high credit risk enterprises, and the value interval is [0,1]. The closer to 1 it is, the stronger the ability to identify enterprises with high credit risk is. Its calculation formula is as follows:

$$Recall = \frac{TP}{TP + FN}. \tag{14}$$

Precision. Precision means the true proportion of all enterprises with high credit risk identified by the model as high credit risk, and the value interval is [0,1]. The closer to 1 it is, the higher the credibility of the identification result is. Its calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP}. \tag{15}$$

F1-value. F1-value is the harmonic average of recall rate and precision, which can comprehensively reflect the overall effect of recall rate and precision, so it is often used as the comprehensive evaluation parameter of machine learning. Its value interval is $[0, 1]$, the closer to 1 it is, the overall performance of the model is better. Its calculation formula is as follows:

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}. \tag{16}$$

4.4 Empirical Research Results of Credit Evaluation

In this section, three algorithms including random forest (RF), support vector machine (SVM) and decision tree (DT) on R language platform are respectively used to empirically analyze the enterprise credit risk data of banks. Each experiment contained the original data set and the data set processed by SOMTE. Due to the sufficient data sample size, in order to improve the test credibility of the model, the ratio of analysis training set and sample set in RF and SVM is 6:4. In the DT algorithm, the ten fold cross validation method was used to train and test the model. The overall experimental results and comparative analysis are shown in Table 4.

4.4.1 Results of RF-Based Enterprise Credit Assessment Analysis

Regarding the parameter setting of the RF algorithm, the number of decision trees contained in the random forest was finally set to ntree = 600, and the number of variables used in the binary tree in the node mtry = 3.

The results are shown in Table 4. In the experiment with the original set, the overall sample corporate credit classification accuracy rate and F1-value was 99 % and 97 %, indicating that the overall recognition performance of the model was very good, and the high-risk enterprise samples credit classification recall and precision

Research	Characteristics	Algorithm	Accuracy	Recall	Precision	F1
Our study	State-constraint characteristics	RF	**0.99**	**0.96**	**0.99**	**0.97**
		SVM	**0.99**	**0.94**	**0.99**	**0.96**
		DT	0.95	0.64	0.89	0.74
		SMOTE +RF	0.94	0.87	0.90	0.88
		SMOTE +SVM	**0.99**	**0.99**	**0.99**	**0.99**
		SMOTE +DT	**0.99**	**0.99**	**0.99**	**0.99**
Qiu W et al. (2019) [35]	Credit history and company information characteristics	RF	0.84	0.75	0.50	0.60
		GBDT	0.87	0.76	0.59	0.66
		XGBOOST	0.82	0.85	0.47	0.61
Jain et al. (2020) [37]	Trade and loan characteristics	DT	**0.99**	0.78	0.81	0.79
		RF	**0.99**	0.78	**0.97**	0.86
		XGBOOST	**0.99**	0.83	0.95	0.89
Wang F et al. (2020) [34]	Online supply chain characteristics	LS-SVM	0.97	**0.96**	**0.97**	**0.96**
Jingming L et al. (2020) [36]	Enterprise competence characteristics	GSO-ELM	0.91	/	0.91	/

The top 5 values of each evaluation index are bolded;
GBDT- Gradient Boosting Decision Tree;
LS-SVM- Least Squares SVM;
GSO-ELM- Group Search Optimizer- Extreme Learning Machine.

Table 4. Empirical analysis results of enterprise credit evaluation

was 96 % and 99 %, indicating that the model's ability to identify key risks was outstanding as well. The result is significantly better than similar studies of Qiu [35] and Jain [37] with the same RF algorithm but different model characteristics, as well as better than research results of Qiu [35] and Li [36] with their advanced methods but different model characteristics, which shows that our credit risk assessment model on state-constraint of both bank and enterprise is reliable with pretty performance. In the experiment of SMOTE set, the overall classification accuracy rate and F1-value dropped to 94 % and 88 %, and the recall and precision dropped 87 % and 90 %, and the overall risk identification ability of the model drops significantly. After inspection and analysis, we believe that the reason for the decline in the performance of the SMOTE data set model may be that the imbalance of the total sample set is relatively small, and the RF algorithm itself has greater adaptability and tolerance for data imbalance, and high-risk enterprise sample groups have high similarities. Using the SMOTE algorithm will oversample a small number of samples with high risk to form new samples with a smaller gap from the original samples, resulting in similar "overfitting" problems, leading to learner model performance

worse. At the same time, we found that this problem does not exist in other two algorithms.

4.4.2 Results of SVM-Based Enterprise Credit Assessment Analysis

Regarding the parameter setting of SVM, by comparing the classification accuracy of high-risk enterprise samples in the SVM algorithm, the radial basis kernel function (RBF) was finally selected as the kernel function of the SVM model. There are two important parameters in RBF: gamma and cost. Gamma is a parameter of the kernel function that controls the shape of the segmented hyperplane. The larger the gamma, the more support vectors and the wider the range of training samples. Cost represents the cost parameter of the model's error cost. The greater the cost, the greater the model's penalty for errors, the more complex the generated classification boundary, and the smaller the error in the corresponding training set, but it is also possible that the too small cost can lead to overfitting problems. Considering the balance of learning performance and efficiency, we finally let the gamma parameter range to $(10^{-6}, 10)$ and the cost parameter range to $(10^{-6}, 10^{10})$.

In the SVM algorithm experiment, the overall classification accuracy and F1-value of the original data set was 99 % and 96 %, meanwhile the recall and precision was 94 % and 99 %, indicating that the learner's recognition ability for the samples of low-risk enterprises was higher than that of high-risk enterprises. But the comprehensive effect of our model is still pretty fine, which is a little senior than research of Wang et al. with LS-SVM in 2020 [34]. In the SMOTE set, overall classification accuracy and precison of the dataset was still 99 %, but recall rate increased from 94 % to 99 %, and F1-value reached to 99 %. This indicated that SMOTE had a certain improving effect on the SVM model, and our final results were also better than other researches.

4.4.3 Results of DT-Based Enterprise Credit Assessment Analysis

Since the SMOTE algorithm had different effects in the RF and SVM models, we added the experiment of the decision tree algorithm to obtain more reliable results. In the DT experiment, because there were too many categorical feature variables, the CHAID decision tree suitable for processing multivariate and categorical variables was selected. The parent node and child node of the minimum number of cases were 100 and 50, respectively, and the maximum number of classes was 3.

As can be seen from the results in Table 4, the overall classification accuracy of the original data set was 95 %, but the F1-value was only 74 %, and the recall and precision was only 64 % and 89 %, indicating that the decision tree model has poor recognition ability for high-risk groups, even though this result was better than result of Qiu et al. with GBDT [35]. In SMOTE set, four evaluation indexes all increased to 99 %, especially the recall rate suggested a 33 % increase over the original set. This result confirms our analysis in the stochastic forest algorithm, and indicates that SMOTE has an obvious effect in treating the imbalance data set,

which can resolve the data imbalance and improve the performance of the learner to some extent.

4.5 Importance of Risk Assessment Characteristic Indicators

The stronger the ability of the risk assessment characteristic indicators to distinguish between enterprise customer credit (default and normal), the higher its importance, that is, the characteristic indicators have obvious individual effects on customer credit assessment, and play a significant role in risk credit risk assessment. Figure 7 shows the accuracy reduction characteristics importance based on the RF algorithm and the characteristics importance of the Gini coefficient. Although the importance of a few indicators is inconsistent, the overall results were regionally consistent. In our study, the 15 characteristic indicators based on the state-constraint model all had a certain level of credibility. Among them, the top 5 indicators such as OD, DIS, CS, IC and IR were ranked among the top 5 with the importance of accuracy contribution over 0.015 and the importance of Gini coefficient over 8, which showed a strong ability to distinguish enterprise customer credit. PIR and BNRO were both at the bottom, which showed poor ability to distinguish customer credit. The importance of the Gini coefficient for the other eight indicators, all exceeded 4, which was at an intermediate level, and played a certain role in promoting customer credit differentiation.

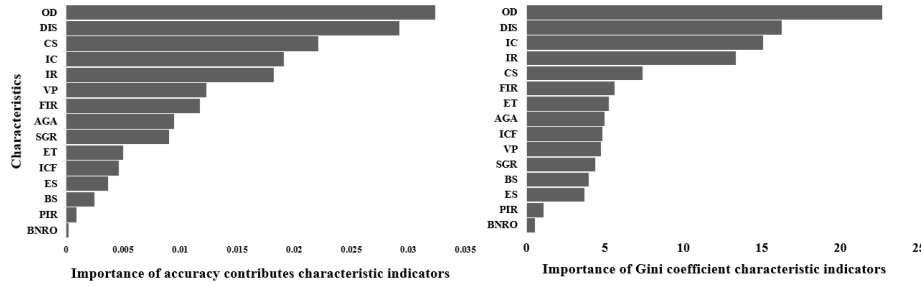


Figure 7. Importance of characteristic indicators of a risk assessment model

5 CONCLUSION AND DISCUSSION

The digital transformation of the financial industry is constantly developing. Commercial banks and other financial institutions will inevitably face increasingly complex enterprise loan customers and businesses. The scientific and digital assessment of enterprise loan customer credit risk is crucial. Based on the theory of constraint, our study extracts characteristic indicators from the state and constraint of both the bank and enterprise and builds an enterprise credit risk assessment model. In

our empirical research, the comprehensive evaluation recognition rate of the overall sample and high-risk credit enterprises can both reach up to 99 %. The results based on RF and SVM in our study are better than others' researches with same methods but different model characteristics, as well as better than others' researches with both different methods and model characteristics, which shows that our model has excellent performance and reliability for the evaluation of enterprise credit risk. At the same time, it shows that the unbalanced data processing based on the SMOTE algorithm does not significantly improve the performance of the RF algorithm, but it has a significant improvement in the performance of SVM and DT algorithms, which can well overcome the problem of data imbalance. Considering both the performance and robustness of the state-constrain model, SVM seems to be a better choice for the credit risk assessment.

In addition, about 90 % of the characteristic indicators in our research model have a significant correlation with customer credit at a 99 % confidence level. At the same time, the importance of the Gini coefficient is great enough, indicating that the characteristic indicators are highly distinguishable among customer credits. The ability to accurately assess credit risk of our model is strong, and the model construction in this article is very reasonable. This study's credit risk assessment model and the importance of its characteristic indicators can help banks to better understand the internal relationship between banks and enterprise customers, so that banks can better review and control enterprise credit risks in the corresponding feature dimensions. Thereby our study can provide important decision-making references for banks and enrich theoretical system of the credit risk research.

There are some limitations in our study. Firstly, the data used in our study comes from a commercial bank in China, and the data information is limited, although we have done a lot of exploration and attempted to find the optimal algorithm for the model. Due to the nature of machine learning and the limited data it is hard to figure out the internal relationship between model data and machine learning algorithms and give a fixed optimal algorithm. At the same time, the applicability of the conclusion in different countries and different regulatory policies needs to be further explored. In addition, we only study the credit risk assessment of commercial bank customers, but do not expand the study to other non-bank financial institutions. In the future, we will try to focus on customer credit risk of banks and financial institutions in different countries, continuously improve the research model, and explore the universality of credit risk assessment of our study.

REFERENCES

[1] Katre, S. M.: Analysis of Problems Faced by Public Sector Banks and Cooperative Banks and Strategies to Overcome w.s.r. to Ahmednagar City. IBMRD's Journal of Management and Research, Vol. 1, 2012, No. 1, pp. 84–87.

[2] Laurenceson, J.—Chai, J. C. H.: State Banks and Economic Development in China. Journal of International Development, Vol. 13, 2001, No. 2, pp. 211–225, doi: 10.1002/jid.727.

[3] Luvizan, S. S.—Nascimento, P. T.—Yu, A.: Big Data for Innovation: The Case of Credit Evaluation Using Mobile Data Analyzed by Innovation Ecosystem Lens. 2016 Portland International Conference on Management of Engineering and Technology (PICMET), IEEE, 2016, pp. 925–936, doi: 10.1109/picmet.2016.7806738.

[4] Hurley, M.—Adebayo, J.: Credit Scoring in the Era of Big Data. The Yale Journal of Law and Technology, Vol. 18, 2016, pp. 148–216.

[5] Golbayani, P.—Wang, D.—Florescu, I.: Application of Deep Neural Networks to Assess Corporate Credit Rating. 2020, arXiv: 2003.02334v1.

[6] Choi, J.—Suh, Y.—Jung, N.: Predicting Corporate Credit Rating Based on Qualitative Information of MD & A Transformed Using Document Vectorization Techniques. Data Technologies and Applications, Vol. 54, 2020, No. 2, pp. 151–168, doi: 10.1108/dta-08-2019-0127.

[7] Minnis, M.—Sutherland, A.: Financial Statements as Monitoring Mechanisms: Evidence from Small Commercial Loans. Journal of Accounting Research, Vol. 55, 2017, No. 1, pp. 197–233, doi: 10.1111/1475-679x.12127.

[8] Oseledets, I. V.—Tyrtyshnikov, E. E.: Breaking the Curse of Dimensionality, Or How to Use SVD in Many Dimensions. SIAM Journal on Scientific Computing, Vol. 31, 2009, No. 5, pp. 3744–3759, doi: 10.1137/090748330.

[9] Tong, G.—Li, S.: Construction and Application Research of Isomap-RVM Credit Assessment Model. Mathematical Problems in Engineering, Vol. 2015, 2015, Art. No. 197258, doi: 10.1155/2015/197258.

[10] Yin, W.—Liu, X.: Bank Versus Nonbank Financial Institution Lending Behaviour: Indictors of Firm Size, Risk or Ownership. Applied Economics Letters, Vol. 24, 2017, No. 18, pp. 1285–1288, doi: 10.1080/13504851.2016.1273473.

[11] Bonsall, S. B.—Holzman, E. R.—Miller, B. P.: Managerial Ability and Credit Risk Assessment. Management Science, Vol. 63, 2016, No. 5, pp. 1425–1449, doi: 10.1287/mnsc.2015.2403.

[12] Bourgain, A.—Pieretti, P.—Zanaj, S.: Financial Openness, Disclosure and Bank Risk-Taking in MENA Countries. Emerging Markets Review, Vol. 13, 2012, No. 3, pp. 283–300, doi: 10.1016/j.ememar.2012.01.002.

[13] Huang, Y. M.—Hung, C. M.—Jiau, H. C.: Evaluation of Neural Networks and Data Mining Methods on a Credit Assessment Task for Class Imbalance Problem. Nonlinear Analysis: Real World Applications, Vol. 7, 2006, No. 4, pp. 720–747, doi: 10.1016/j.nonrwa.2005.04.006.

[14] Lou, Y.: The Research on Corporate Credit Risk Evaluation Model Based on Fuzzy Neural Network. Journal of Central South University, Vol. 19, 2013, No. 5 (in Chinese).

[15] Tian, Y.—Sun, M.—Deng, Z.—Luo, J.—Li, Y.: A New Fuzzy Set and Nonkernel SVM Approach for Mislabeled Binary Classification with Applications. IEEE Transactions on Fuzzy Systems, Vol. 25, 2017, No. 6, pp. 1536–1545, doi: 10.1109/tfuzz.2017.2752138.

[16] Bu, D.—Kelly, S.—Liao, Y.—Zhou, Q.: A Hybrid Information Approach to Predict Corporate Credit Risk. The Journal of Futures Markets, Vol. 38, 2018, No. 9, pp. 1062–1078, doi: 10.1002/fut.21930.

[17] Huang, X.—Liu, X.—Ren, Y.: Enterprise Credit Risk Evaluation Based on Neural Network Algorithm. Cognitive Systems Research, Vol. 52, 2018, pp. 317–324, doi: 10.1016/j.cogsys.2018.07.023.

[18] Kim, J. B.—Song, B. Y.—Stratopoulos, T. C.: Does Information Technology Reputation Affect Bank Loan Terms? The Accounting Review, Vol. 93, 2018, No. 3, pp. 185–211, doi: 10.2308/accr-51927.

[19] Manogaran, G.—Chilamkurti, N.—Hsu, C. H.: Special Issue on Advancements in Artificial Intelligence and Machine Learning Algorithms for Internet of Things, Cloud Computing and Big Data. International Journal of Software Innovation, Vol. 7, 2019, No. 2.

[20] Janitza, S.—Strobl, C.—Boulesteix, A.-L.: An AUC-Based Permutation Variable Importance Measure for Random Forests. BMC Bioinformatics, Vol. 14, 2013, No. 1, Art. No. 119, doi: 10.1186/1471-2105-14-119.

[21] Gislason, P. O.—Benediktsson, J. A.—Sveinsson, J. R.: Random Forests for Land Cover Classification. Pattern Recognition Letters, Vol. 27, 2006, No. 4, pp. 294–300, doi: 10.1016/j.patrec.2005.08.011.

[22] Rodriguez-Galiano, V. F.—Ghimire, B.—Rogan, J.—Chica-Olmo, M.—Rigol-Sanchez, J. P.: An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 67, 2012, pp. 93–104, doi: 10.1016/j.isprsjprs.2011.11.002.

[23] Breiman, L.: Random Forest. Machine Learning, Vol. 45, 2001, No. 1, pp. 5–32, doi: 10.1023/A:1010933404324.

[24] Qin, X.—Li, Q.—Dong, X.—Lv, S.: The Fault Diagnosis of Rolling Bearing Based on Ensemble Empirical Mode Decomposition and Random Forest. Shock and Vibration, Vol. 2017, 2017, Art. No. 2623081, doi: 10.1155/2017/2623081.

[25] Chapelle, O.: Training a Support Vector Machine in the Primal. Neural Computation, Vol. 19, 2007, No. 5, pp. 1155–1178, doi: 10.1162/neco.2007.19.5.1155.

[26] Lookman, A. A.: Bank Borrowing and Corporate Risk Management. Journal of Financial Intermediation, Vol. 18, 2009, No. 4, pp. 632–649, doi: 10.1016/j.jfi.2008.08.006.

[27] Laeven, L.—Levine, R.: Bank Governance, Regulation and Risk Taking. Journal of Financial Economics, Vol. 93, 2009, No. 2, pp. 259–275, doi: 10.1016/j.jfineco.2008.09.003.

[28] Stoi, R.—Kühnle, B. A.: Theory of Constraints. Controlling – Zeitschrift für Erfolgsorientierte Unternehmenssteuerung, Vol. 14, 2002, No. 1, pp. 55–56, doi: 10.15358/0935-0381-2002-1-55.

[29] Lee, J.—Naranjo, A.—Sirmans, S.: Exodus from Sovereign Risk: Global Asset and Information Networks in the Pricing of Corporate Credit Risk. The Journal of Finance, Vol. 71, 2016, No. 4, pp. 1813–1856, doi: 10.1111/jofi.12412.

[30] Wagner, J.: Credit Constraints and Exports: Evidence for German Manufacturing Enterprises. Applied Economics, Vol. 46, 2014, No. 3, pp. 294–302, doi: 10.1080/00036846.2013.839866.

[31] Hall, K.: The Psychology of Corporate Dishonesty. Australian Journal of Corporate Law, Vol. 19, 2006, p. 268–286.

[32] Sassi, S.—Gasmi, A.: The Effect of Enterprise and Household Credit on Economic Growth: New Evidence from European Union Countries. Journal of Macroeconomics, Vol. 39, 2014, Part A, pp. 226–231, doi: 10.1016/j.jmacro.2013.12.001.

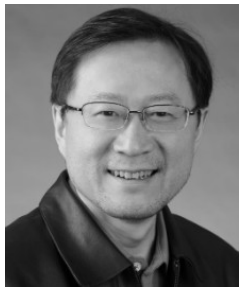
[33] Yap, B. W.—Khatijahhusna, A. R.—Rahman, H. A. A.—Fong, S.—Khairudin, Z.—Abdullah, N. N.: An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In: Herawan, T., Deris, M., Abawajy, J. (Eds.): Proceedings of the First International Conference on Data Engineering (DaEng-2013). Springer, Singapore, Lecture Notes in Electrical Engineering, Vol. 285, 2014, pp. 13–22, doi: 10.1007/978-981-4585-18-7_2.

[34] Wang, F.—Ding, L.—Yu, H.—Zhao, Y.: Big Data Analytics on Enterprise Credit Risk Evaluation of e-Business Platform. Information Systems and e-Business Management, Vol. 18, 2020, pp. 311–350, doi: 10.1007/s10257-019-00414-x.

[35] Qiu, W.—Li, S.—Cao, Y.—Li, H.: Credit Evaluation Ensemble Model with Self-Contained Shunt. 2019 5th International Conference on Big Data and Information Analytics (BigDIA), 2019, pp. 59–65, doi: 10.1109/bigdia.2019.8802679.

[36] Li, J.—Li, X.—Dai, D.—Ruan, S.—Zhu, X.: Research on Credit Risk Measurement of Small and Micro Enterprises Based on the Integrated Algorithm of Improved GSO and ELM. Mathematical Problems in Engineering, Vol. 2020, 2020, Art. No. 3490536, doi: 10.1155/2020/3490536.

[37] Jain, V.—Agrawal, M.—Kumar, A.: Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 86–88, doi: 10.1109/icrito48877.2020.9197762.



Renjing Liu received his graduation degree from the Xinjiang University of Mathematics, Xinjiang, China, in 1987. He is Professor in the Xi'an Jiaotong University, Xi'an China. His current research interests include artificial intelligence, complex system management, multi project management, risk management, business intelligence.

Xuming Yang received his graduation degree from the Chongqing University of Industrial Engineering, Chongqing, in 2019. He is a graduate student of Xi'an Jiaotong University, Xi'an, China. His current research interests include information management and business intelligence.

Xinyu Dong is a graduate student of the Dietrich School of Arts and Sciences of University of Pittsburgh. Her current research interests include big data analysis and artificial intelligence.

Boyang Sun is a graduate student of the School of Management of Xi'an Jiaotong University, Xi'an, China. His current research interests include information management and data mining.