ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20220623113041
Original scientific paper

A Text Mining and Ensemble Learning Based Approach for Credit Risk Prediction

Yang MAO, Shifeng LIU, Daqing GONG*

Abstract: Traditional credit risk prediction models mainly rely on financial data. However, technological innovation is the main driving force for the development of enterprises in strategic emerging industries, which is closely related to enterprise credit risk. In this paper, a novel prediction framework utilizing technological innovation text mining data and ensemble learning is proposed. The empirical data from China listed enterprises in strategic emerging industries were applied to construct prediction models using the classification and regression tree model, the random forest model and extreme gradient boosting model. The results show that the model uses the technological innovation text mining data proven to have significant predict ability, and top management team's attention to innovation variables offer the best prediction capacities. This work improves the application value of enterprise credit risk prediction models in strategic emerging industries by embedding the mining of technological innovation text information.

Keywords: credit risks; ensemble learning; strategic emerging industries; text mining

1 INTRODUCTION

Technological innovation is the main driving force of sustainable economic and social development. China's government has announced that it will shift its economy from traditional industries to a more advanced and green technology-driven model and proposes to vigorously cultivate and develop strategic emerging industries to gain technological advantages in the international competition [1]. The strategic emerging industries suggested by the China government are a group of seven industries-energy conservation and environmental protection, new energy, new energy automobiles, biotechnology and medicine, new materials, new generation IT, high-end equipment manufacturing, and corresponding sub-industries [2]. Strategic emerging industries play a key role in promoting ecological civilization construction, green China, and sustainable economic and social development [3]. Strategic emerging industry enterprises (SEIEs) constitute the backbone of the strategic emerging industry. However, it is difficult for SEIEs to raise funds from commercial banks since the uncertainty of the R&D prospect of SEIEs makes their credit risk higher than that of traditional enterprises.

On one hand, SEIEs contribute considerably to the sustainable development of the economy and technology; on the other hand, lending to SEIEs may come with greater risks than lending to traditional enterprises. This dilemma has attracted considerable research interest among both academics and practitioners. Since the groundbreaking work of Altman [4], many credit risk prediction models focusing on enterprises and utilizing various traditional financial indicators have been proposed [5-10]. However, there are few works of literature on credit risk prediction of SEIEs. Due to the ability of traditional financial data to predict the credit risk caused by the uncertainty of future scientific and technological competitiveness is limited.

Another form of data, innovation information, can solve this problem; such data include R&D investment, intangible assets, and the top management team's attention to innovation (TMTAI) that can be constructed through text mining technology.

Therefore, the purpose of this study is to prove that using traditional financial ratios and adding innovative information of text mining can significantly predict the credit risk of SEIEs. To prove the innovation information feasibility and applicability in the major mainstream models, we selected the two most popular ensemble learning models - random forest (RF) and extreme gradient boosting (XGBoost) and their base model-classification and regression tree (CART) - to test the result. The collected data were then subjected to CART, RF, and XGboost model tests to demonstrate their reliability using various classification methods. In total, external data of 70 listed SEIEs from 2017 to 2019 were selected as the research sample in this study.

The main contributions of this study are as follows. Firstly, we build an ensemble learning model of credit risk assessment embedded with text mining technology. Secondly, this study compares the credit risk prediction ability of strategic emerging industry enterprises with various types of innovative information. Thirdly, this work expands the literature on credit risk prediction models for enterprises in strategic emerging industries.

The rest of the paper is organized as follows: In section 2, we review the literature. In section 3, we put forward a research framework as the basis of the research problem, and explain the research data and variables. We also introduce the text mining method and the ensemble learning model. The model is verified by the listed companies in China's strategic emerging industries. In section 4, we discuss the results of the empirical analysis. Finally, the conclusion is given in the last section.

2 LITERATURE REVIEW

2.1 Literature Review of Credit Risk Prediction Methodologies

There are two main types of enterprise credit risk forecast methods: one is the traditional statistical analysis method, such as discriminant analysis [4] and logistic regression analysis [5]. Discriminant analysis model is the earliest statistical model used in the prediction of enterprise credit risk, the other is machine learning methods, such as ensemble learning approach [6, 7], artificial neural network (ANN) approach [8, 9], support vector machine (SVM) approach [10] etc. Parisa Golbayani et al. [11] applied all machine learning methods to the same datasets from 2009 to 2018 for enterprises in energy and healthcare industries, all their results show that ensemble learning methods have better prediction performance than SVM and ANN. Meanwhile, the present research concentrates on the ensemble learning approach which is an effective method and possesses excellent credit risk forecast performance [12, 13].

As CART is widely used in the base model of ensemble learning, the RF and XGBoost are frequently utilized ensemble learning methods for credit risk prediction. All the three methods were adopted in this study.

2.2 Literature Review of Variables Used in Credit Risk Prediction

In terms of credit risk identification indicators, in the field of enterprise credit risk research, most researchers used financial ratio indicators to build risk identification models [4, 14]. Some scholars believe that most of the previous research on enterprise credit risk prediction is based on quantitative data and lacks the mining of qualitative text information. Therefore, many scholars began to introduce text analysis into the research of credit risk prediction, quantify the qualitative text information such as news reports, network public opinion and legal decisions, and then add it to the credit risk prediction model [15, 16]. However, because the above research information comes from outside the company, its value is relatively limited, and the information conveyed by the management as an internal person is more valuable. Some scholars have verified that the management tone of the text content transmission has indeed improved the effectiveness of the enterprise credit risk early warning model by analysing the text characteristics of the press conference and annual report of listed companies [17, 18].

In recent years, the innovation attention of senior executives has become an important dimension to measure the innovation ability of enterprises. Scholars' research shows that CEOs' attention to new technologies has affected the number of patent applications for high-tech industrial enterprises [18, 19]. The attention of the top management team to technological innovation is the degree of attention paid by the top management team to technological innovation. The more attention the senior management team pays to technological innovation, the more attention the senior management team pays to issues related to technological innovation. To sum up, the attention of the senior management team to technological innovation represents the attention of the senior management team to all issues related to technological innovation. The higher the attention, the stronger the enterprise's technological innovation ability in general.

In terms of research on the relationship between enterprise innovation capability and credit risk, Wojan et al.'s research results show that the vast majority of long-term surviving manufacturing enterprises will shift from non-innovation strategies to incremental or broader innovation directions [20]; Lee et al. analysed the technological innovation characteristics that affect the survival period of SMEs in the Korean manufacturing industry, and found that the innovation characteristics related to technology have a positive effect on the survival of SMEs [21]. The research results of Zhang et al. show that innovation measured by patents, innovation efficiency and enterprise import and export activities can improve the survival rate of Chinese high-tech enterprises [22]. Li et al. reported that the R&D input of firms can significantly decrease firm death rates [23]. Fernandes and Paunov find that new products can increase the likelihood of firm survival under certain conditions [24]. Helmers and Rogers suggest that firm survival duration is positively and significantly associated with innovation [25].

It can be seen from the above literature that in the current research on enterprise credit risk identification methods, the evaluation indicators mostly use quantitative indicators such as financial indicators and text information such as management tone, legal judgment, media, and investor sentiment. For SEIEs, these pieces of information ignore the important factor of technological innovation ability that leads to the credit risk of SEIEs, through text analysis of the text technological innovation information of the TMTAI transmitted by the annual reports of SEIEs; this study establishes a credit risk identification method for SEIEs based on the combination of technological innovation information reflecting the technological innovation ability of enterprises and financial data reflecting the financial status of enterprises in combination with common financial indicators and indicators measuring the technological innovation ability of enterprises. The selection of indicators cannot only objectively and comprehensively reflect the current operating conditions of SEIEs, but also reflect the potential impact of SEIEs' technological innovation level on future credit risk from the three aspects of TMTAI, R&D and technological achievements. It has important academic value in updating the credit risk identification methods of SEIEs.

3 RESEARCH METHODS

3.1 Research Framework

This paper establishes a research method framework for credit risk assessment of SEIEs based on text analysis and ensemble learning, as shown in Fig. 1. Firstly, the unstructured text data related to technological innovation in the annual reports of listed SEIEs are reprocessed in natural language, including Chinese word segmentation, the stop words, etc. in combination with reading the annual reports of listed SEIEs and screening the technological innovation keyword set based on relevant literature, the TMTAI is constructed through TF-IDF algorithm. Then, through characteristic importance analysis, the impact of a single indicator of financial and technological innovation on enterprise credit risk is compared and analysed. Finally, through the ensemble learning model, the impact of financial indicators and technological innovation combination indicators on enterprise credit risk is compared and analysed, to realize the construction of the research framework of credit risk assessment method system for SEIEs.

3.2 Data and Variable

As the previous work has repeatedly shown, it is not reliable to use only the financial data for SEIEs credit risk prediction. Therefore, this study uses financial data and innovative information based on text mining to train the proposed prediction model. We collected the annual data of 70 listed companies in strategic emerging industries in China from 2017 to 2019. There are 15 high credit risk listed enterprises with *st in the data set, and 55 non *st low credit risk listed enterprises. The *st listed enterprises refer to special listed enterprises that suffer delisting risk warning due to abnormal financial conditions. We use $t-2$ year data of listed companies in strategic emerging industries to model and predict whether the listed companies will have credit risk in t year. From 2019 to

2021, 15 listed companies in strategic emerging industries by *st were taken as research samples, and the corresponding 55 normal listed companies in strategic emerging industries were taken as matched samples. As the number of high credit risk enterprises is less than that of low-risk enterprises, and the data is unbalanced, we use SMOTE algorithm to balance the samples, and finally obtain 55 samples of high credit risk enterprises and 55 samples of low credit risk enterprises, a total of 110 samples. The data in this paper comes from the Choice financial terminal and the annual reports of listed companies. The collected text data mainly includes the discussion and analysis of the operation part of the annual reports - "overview" and "prospects for future development".

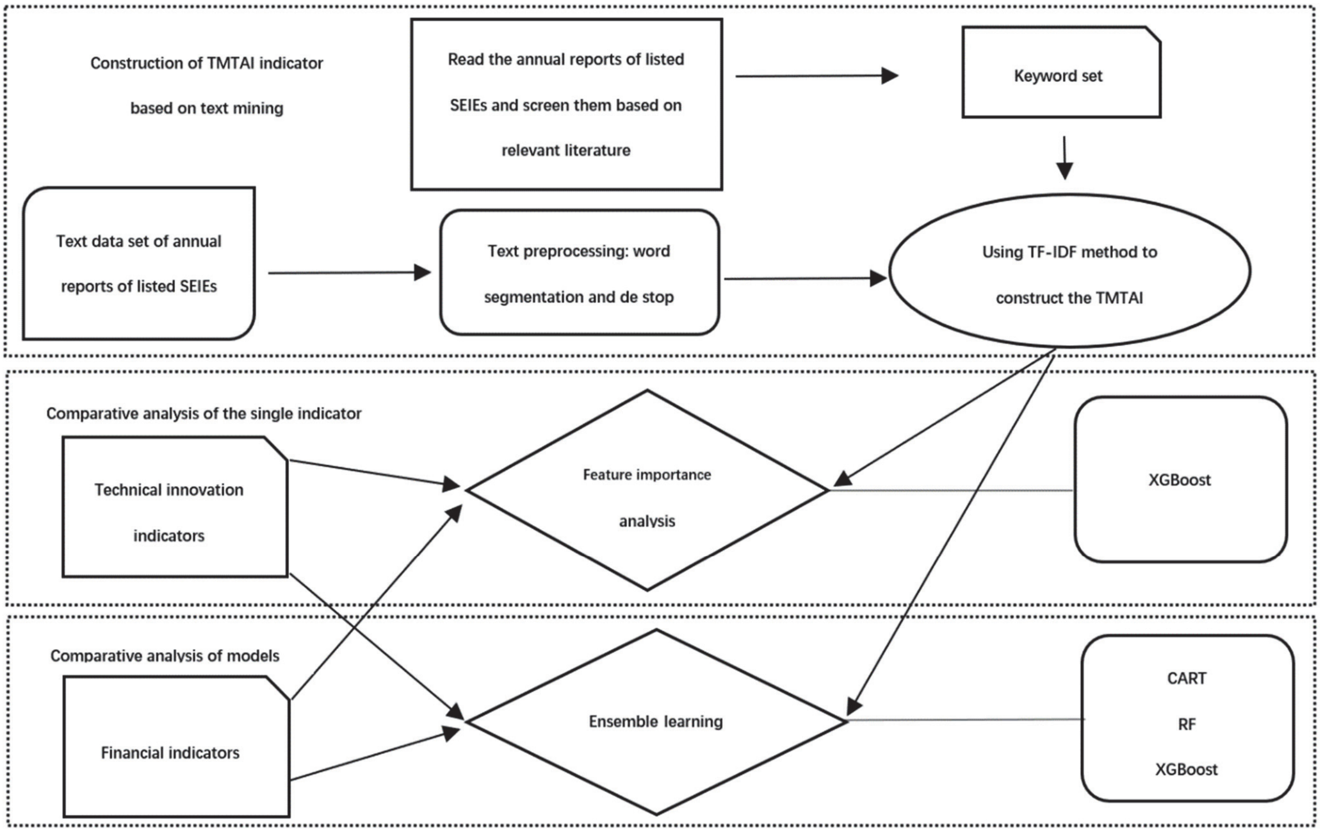
Construction of TMTAI indicator based on text mining
Read the annual reports of listed SEIEs and screen them based on relevant literature
Keyword set
Text data set of annual reports of listed SEIEs
Text preprocessing: word segmentation and de stop
Using TF-IDF method to construct the TMTAI
Comparative analysis of the single indicator
Technical innovation indicators
Feature importance analysis
XGBoost
Comparative analysis of models
Financial indicators
Ensemble learning
CART
RF
XGBoost

Figure 1 A framework of the research method of credit risk assessment of SEIEs based on text analysis and ensemble learning

There are two main indicators to evaluate the credit risk of SEIEs: financial indicators and innovation indicators. Financial indicators are divided into 11 evaluation indices based on the recommendations of Zhang et al. [26] and innovation indicators are divided into 3 evaluation indices, which are based on suggestions of Angilella et al. [27] and Chen et al [18]. These 14 indicators were used as the source independent variables of the models. These independent variables are defined in Tab. 1.

The main financial indicators to evaluate the credit risk of SEIEs are: liquidity index, leverage index, profitability index, and growth index. The innovation information was categorized into three types: management innovation attention information - top managers team attention to innovation (TMTAI), innovation capital investment information- proportion of R&D investment in sales revenue, and intellectual property information, which can be represented by the proportion of intangible assets in fixed assets.

Table 1 List of independent variables

Serial number	Variable name	Variable categories
1	Current Ratio	Liquidity
2	Quick Ratio	Liquidity
3	Asset-liability Ratio	Liquidity
4	Inventory Turnover	Leverage
5	Accounts Receivable Turnover	Leverage
6	Total Asset Turnover	Leverage
7	Fixed Asset Turnover	Leverage
8	ROA	Profitability
9	Net Profit Margin	Profitability
10	ROE	Profitability
11	Capital Increment Rate	Growth
12	R&D/Sales	Innovation
13	Intangible Assets/Fixed Assets	Innovation
14	TMTAI	Innovation

In this paper, the construction method of the top management team's attention to innovation is based on the text data of the annual reports of listed companies in strategic emerging industries, and the word segmentation module of "Jieba" in Python programming language is used to segment the words of "overview" and "outlook for the future development of the company" in the chapter of "discussion and analysis of operation and management" of the annual reports. Based on the research of Chen et al. [18], we use the file based method. By counting the number of words reflecting technological innovation in the annual reports of listed companies in strategic emerging industries, reconstructing the TMTAI indicators in the sample, 9 keywords related to technological innovation constitute the vocabulary search dictionary for the content analysis of technological innovation of SEIEs (Tab. 2), including intellectual property, innovation, patent, technology, R&D, experiment, talent, science and technology, and research. With reference to the measurement method of Chen et al. [18] on innovation attention, the sum of TF-IDF of key words related to technological innovation of each enterprise is counted as the TMTAI indicator.

Table 2 Glossary of technological innovation

Chinese	chuangxin	jishu	yanfa
English	innovation	technology	research and development
Chinese	rencai	zhuanli	zhishichanquan
English	talent	patent	intellectual property
Chinese	shiyan	yanjiu	keji
English	experiment	research	science and technology

Using text analysis, we searched for the above key words in SEIEs' annual reports. We calculate the TMTAI of SEIEs as follow, where n_j is the number of occurrences of innovation keyword in annual report j, and the denominator is the sum of the occurrences of all the words in annual report j. $|D|$ is the total number of files contained in the corpus. The denominator is the number of files containing the innovative keyword t.

$$TMTAI = TF - IDF = \frac{n_j}{\sum_k n_{kj}} \times \log \frac{|D|}{1+\left|\left\{j : t \in d_j\right\}\right|} \tag{1}$$

3.3 Model

This research adopts ensemble learning method as the main research method. Ensemble learning method is a method to improve the accuracy of machine learning algorithm by constructing and combining the integration of base machine learning classifiers. Each base classifier only needs to have medium accuracy on the training set. Decision tree is usually used as the base classifier, and the two most popular ensemble learning methods are bagging ensemble and boosting ensemble.

3.3.1 Base Classifier Model-Classification and Regression Tree (CART)

It is a decision tree algorithm proposed by Breiman et al. [28]. Classified regression tree is a typical binary decision tree, which can be used for classification or regression. If it is used to predict discrete data, a classification tree is generated. If it is used to predict continuous data, a regression tree is generated. In this study, the classification decision tree algorithm is used as the base classifier of the random forest algorithm, and the regression decision tree is used as the base classifier of the extreme gradient boosting algorithm, to classify high credit risk enterprises and low credit risk enterprises.

The algorithm of the classification decision tree is as follows:

Input: the data set is $D = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i)\}$;

(1) The Gini index of the existing feature pair training data set D is calculated, for each feature A, for all possible values of a, the Gini index of the training data set D for each feature a is divided into two parts: D_1 and D_2 according to $a = A$ and $a! = a$, the formula for calculating the Gini index of $A = a$ is as follows:

$$\mathrm{Gini}\left(D, A\right) = \frac{|D_1|}{D}\mathrm{Gini}(D_1) + \frac{|D_2|}{D}\mathrm{Gini}(D_2) \tag{2}$$

(2) After traversing all features A, the Gini index of all possible values of A is calculated. The feature and cut-off point corresponding to the minimum Gini index of D are selected as the optimal partition, and the data is divided into two subsets.

(3) Continue to call steps (1) (2) on both subareas until the stop condition is met.

(4) The CART is generated.

Output: CART

The algorithm of regression tree is shown as the following:

Input: the data set is $D = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i)\}$;

(1) Traverse variable j, scan the splitting point s for the fixed partition variable j:

$$\hat{c}_1 = \mathrm{ave}\left(y_i \mid x_i \in R_1\left(j, s\right)\right) \tag{3}$$

$$\hat{c}_2 = ave(y_i \mid x_i \in R_2\left(j, s\right)) \tag{4}$$

Select the optimal segmentation variable j and the slitting point s, which makes the equation reach the minimum value.

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} \left(y_i - c_1\right)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} \left(y_i - c_2\right)^2\right] \tag{5}$$

(2) The region is divided by the selected (j, s) and the corresponding output value is determined:

$$R_1\left(j, s\right) = \{x \mid x^{(j)} \le s\}, R_2\left(j, s\right) = \{x \mid x^{(j)} > s\} \tag{6}$$

$$\hat{c}_m = \frac{1}{N_m}\sum_{x_i \in R_m(j,s)} y_i,\ x \in R_m,\ m = 1, 2 \tag{7}$$

(3) Continue to call steps (1) (2) on both subareas until the stop condition is met.

(4) The input space is divided into m regions R_1, R_2, ..., R_m, and the decision tree is generated:

$$f\left(x\right)=\sum_{m=1}^{M}\hat{c}_m I\left(\mathrm{x}\in R_m\right) \tag{8}$$

Output: $f(x)$

3.3.2 Bagging Ensemble - Random Forest Model (RF)

The principle of bagging ensemble is to improve the classification effect by combining the classification results of multiple training sets, that is, creating and combining multiple classifiers. Its representative algorithm is random forest algorithm. The random forest (RF) algorithm was proposed by Breiman [29]. It is an integrated algorithm based on classification tree. The algorithm is a classifier containing multiple decision trees. By selecting several groups of independent decision trees, the final classifier is formed, and then the final classification is obtained by averaging the output of each decision tree. In addition to randomly selecting data subsets, random forest also randomly selects subsets on each node, and calculates the best partition on that node only within a given subset. This structure provides uncorrelated or weakly correlated predictions. Random forest has the advantages of processing unbalanced data sets, effectively reducing over fitting probability, and fast training speed. According to the research of Breiman [30], the principle of random forest algorithm is as follows:

K samples with the same sample size as the original training set are extracted from the original training set. These samples are used to establish K decision trees respectively, and the final classification is obtained by voting according to the K classification results. Random forest uses training sets with different structures to improve the differences between classification models to improve the generalization prediction ability of the integrated model. After K rounds of training, the classification model sequence $\{h_1(X), h_2(X), \ldots, h_k(X)\}$ is formed. They form a classification model that uses simple majority voting to determine the result.

Final classification decision:

$$H\left(x\right)=\arg\max_{Y}\sum_{i=1}^{k}I\left(h_i\left(x\right)=Y\right)\begin{cases}I\left(\alpha\right)=1 \text{ if } \alpha \text{ is true,}\\ I\left(\alpha\right)=0 \text{ otherwise.}\end{cases} \tag{9}$$

$H(x)$ represents the combined classification model, $\mathrm{h_i}$ represents the single classification tree model, Y represents the target variable, and Eq. (9) represents that the final classification result is determined by voting.

3.3.3 Gradient Boosting Ensemble - Extreme Gradient Boosting Model (XGBoost)

Gradient boosting ensemble constructs a composite classifier by training the classifiers in turn and increases the weight of error classification observations through iteration. Compared with the example of correct prediction, the observation results of previous classifiers' wrong prediction are selected more frequently. Bosting ensemble combines the prediction of classifier set with weighted majority voting and gives more weight to more accurate prediction. Gradient boosting decision tree is a widely used boosting ensemble algorithm, which is an iterative decision tree algorithm. The algorithm is composed of multiple regression trees, and the results of all regression trees jointly determine the results.

Extreme gradient boosting (XGBoost) algorithm proposed by Chen and Guestrin [30] is a supervised ensemble tree algorithm derived from the gradient boosting tree algorithm. In the iterative optimization process, the extreme gradient boosting algorithm performs a second-order Taylor expansion on its loss function, so it can estimate the loss function more accurately. In addition, the extreme gradient boosting algorithm can also effectively deal with missing values. The regularization term is added to its objective function, which can effectively avoid over fitting. XGBoost solves the classification problem by using residuals to iteratively fit the final value for many times, which is the key point of the algorithm different from random forest. The algorithm principle is as follows:

Input: the data set is $D = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i)\}$; The prediction result of the i-th tree can be expressed as:

$$\widetilde{y_l}=\sum_{k=1}^{k}f_k\left(x_i\right) \tag{10}$$

$f_k\left(x_i\right)$ represents the prediction result of the k-th tree, and:

$$loss=\sum_i l(\widetilde{y_l},y_i)+\sum_k\Omega\left(f_k\right) \tag{11}$$

$$\Omega\left(f_k\right)=T+\frac{1}{2}\lambda\left\|w\right\|^2 \tag{12}$$

where: $\widetilde{y_l}$ represents the predicted value of the model; y_i represents the actual value of the sample; K represents the number of trees; f_k represents the k-th tree model; T represents the number of leaf nodes of the tree; W represents the score at each leaf node; λ represents a super parameter. The training process of XGBoost model is as follows:

$$\widetilde{y_l}^{(0)}=0 \tag{13}$$

$$\widetilde{y_l}^{(1)}=f_1\left(x_i\right)=\widetilde{y_l}^{(0)}+f_1\left(x_i\right) \tag{14}$$

$$\widetilde{y_l}^{(t)}=\sum\nolimits_{k=1}^{t}f_k\left(x_i\right)=\widetilde{y_l}^{(t-1)}+f_t\left(x_i\right) \tag{15}$$

where, $\widetilde{y_l}^{(t)}$ represents the prediction result of round t, and a new function is added to the prediction result of round $t-1$. Therefore, the objective function of round t is:

$$loss^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widetilde{y_l}^{(t-1)} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) \tag{16}$$

After Taylor expansion, keep the first three terms and remove the minimum term, the objective function can be transformed into:

$$loss^{(t)} = \sum_{i=1}^{n}\left[l\left(y_i, \widetilde{y_i}^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \\ + \sum_{k=1}^{k} \Omega\left(f_t\right) \tag{17}$$

$$g_i = \partial_{\widetilde{y_i}^{(t-1)}} l\left(y_i, \widetilde{y_i}^{(t-1)}\right) \tag{18}$$

$$h_i = \partial^2_{\widetilde{y_i}^{(t-1)}} l\left(y_i, \widetilde{y_i}^{(t-1)}\right) \tag{19}$$

Substitute the optimal value of leaf node into Eq. (17), and the objective function can be expressed as:

$$loss^{(t)}\left(q\right) = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) W_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j{}^2\right] + \gamma T \tag{20}$$

3.4 Empirical Procedure

This paper mainly uses "sklearn" library to realize the construction of ensemble learning model. First, we import the "Decision Tree Classifier" module, "Random Forest Classifier" module and "XGB Classifier" module respectively, and instantiate the CART model, RF model and XGBoost model. Second, in order to obtain reliable and stable prediction results, we use the "10x cross validation" method to verify the effectiveness of the model, and randomly divide the data set into 10 equal sized parts *d*1, *d*2, …, *d*10. Secondly, train *d*2, *d*3, …, *d*10 and verify d1, then train *d*1, *d*3, *d*4, …, *d*10 and verify *d*2, and repeat this process ten times until each group is verified once. At last, the average value of 10 results is used as the estimation of algorithm accuracy. Using CART model, RF model and GBoost model, credit risk identification models of SEIEs are established respectively.

3.5 Experimental Performance Measure

3.5.1 Model Performance Indicators

The standard measurement of credit risk prediction of SEIEs is established by using performance indicators. The indicators include "*precision rate*", "*recall rate*" and "*F*1", which are defined as:

$$Precision\ Rate = TP/(TP + FP) \tag{21}$$

$$Recall\ Rate = TP/(TP + FN) \tag{22}$$

$$F1 = 2PR/(P + R) \tag{23}$$

Among them, *TP*, *FP*, *TN* and *FN* denote "true positive", "false positive", "true negative" and "false negative"; negative means "no risk" and "positive" indicates "risk"; *P* and *R* respectively represent "*precision rate*" and "*recall rate*". *Precision rate* is the ratio of accurate “true positive” cases to "predicted positive" cases. *Recall rate* is the ratio of correct "predicted positive" cases to "true positive" cases. The high value of *P* and *R* mean excellent performance. *F*1 is the arithmetic mean of precision rate and recall rate. As shown in Eq. (23), the value of *F*1 is positively correlated with the value of *P* and *R*. The larger the *F*1 value, the better the prediction performance of the classifier.

3.5.2 Characteristic Importance

The influence of financial variables and technological innovation information variables on the credit risk of SEIEs may not be equal. Therefore, we use the relative importance method to study their significance to credit risk. This method estimates the sum of the improvements made when the target variables are used to segment the internal nodes of the decision tree. The improved measurement method is to merge the whole subtree into the terminal node, and divide the given node and the proposed variable into the difference between the square error of the given node. According to the research of Luo et al. [31], the specific principles are as follows:

Given a set of variables $V = \{(v_1, v_2, ..., v_i\}$, is a decision tree, *T* is an intermediate node, $t \in T$, $\hat{i}_t^2\left(v_l\right)$ represents the improvement of the splitting solution point t by the variable *V*, and the characteristic importance of variable $v_l \in V$ in tree *T* can be defined as:

$$I_l\left(T\right) = \sum_{t\in T} \hat{i}_t^2\left(v_l\right) \tag{24}$$

$I_l\left(T\right)$ is the variable v_l relative feature importance in the decision tree *T*. On all internal nodes, the sum of the right end divided by the variable *V* is the maximum estimated improved square error due to the division. In general, the relative importance of *V* is the sum of all these square improvements on all internal nodes, where *V* is selected as the splitting variable. After determining the relative importance of variable *V* in a single tree, the overall degree of variable importance can be constructed. For the decision tree set, the characteristic importance of variable *V* can be summarized by the mean value. Given a set of decision trees T_1, …, T_M, variables v_l the importance in this group of decision trees is:

$$I_l\left(T_1, \ldots T_M\right) = \frac{1}{M}\sum_{m=1}^{M} I_l\left(T_M\right) \tag{25}$$

4 RESULTS AND DISCUSSION

4.1 Feature Importance Analysis

Through the application of XGBoost model, the characteristic importance of explanatory variables can be determined. Feature importance is measured as a score, which indicates the dominant position of each feature in the model. The higher the score, the greater the predictive ability of the variable.

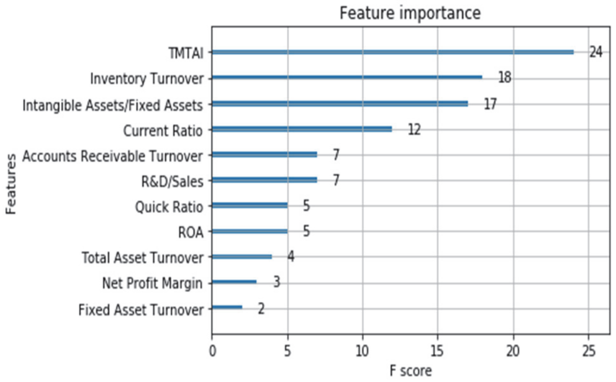

Figure 2 Feature importance evaluation graph based on XGBoost

The results shown in Fig. 2 show that the top management team's attention to technological innovation shows the strongest impact, followed by inventory turnover, intangible assets/fixed assets, current ratio, accounts receivable turnover, R&D investment/sales, quick ratio, ROA, total asset turnover, net profit ratio, and fixed asset turnover. Compared with the financial data, the TMTAI indicator based on text mining proposed in this paper has a more important impact on the credit risk of SEIEs.

4.2 Model Result Analysis

The CART model, RF model, and XGBoost model are applied to the credit risk prediction of SEIEs based on technological innovation information and financial data. The three technological innovation information indicators are combined with traditional financial indicators separately and in combination with the three models. The results show that the combined data set of technological innovation information and traditional financial data significantly improves the recognition performance of the model's credit risk, among all the identification models combined with financial data, the use of three innovative information combined with traditional financial data contributes the most to the improvement of the identification ability of the three models. In addition, we found that the RF model has the best performance, with $F1$ reaching 0.96 at the highest, and the XGBoost model has $F1$ reaching 0.95 at the highest, which is only slightly lower than the RF model.

The construction of a scientific and effective enterprise credit risk evaluation index system needs to meet several important principles, including the principle that the index system should be objective and comprehensive. In order to test the scientific effectiveness of the credit risk evaluation index system of SEIEs constructed in this paper by using text mining and ensemble learning, the traditional index system based on financial quantitative data is compared with the new index with the indicator system combined with quantitative indicators of technological innovation including TMTAI, which comprehensively compares the credit risk identification ability of SEIEs, and makes in-depth analysis from the aspects of management attention text mining, indicator comparison and dimension combination of technological innovation indicators.

Table 3 Performance results with different methods and feature combinations

	CART	RF	XGBoost
Financial indicator	0.86	0.92	0.93
Financial indicator TMTAI	0.87	0.95	0.93
Financial indicator R&D investment	0.91	0.90	0.93
Financial indicator Intangible assets	0.87	0.91	0.94
Financial indicator TMTAI R&D investment	0.87	0.94	0.93
Financial indicator Intangible assets R&D investment	0.90	0.89	0.92
Financial indicator TMTAI Intangible assets	0.91	0.95	0.93
Financial indicator TMTAI Intangible assets R&D investment	0.88	0.96	0.95

From the perspective of management attention text mining, this study finds for the first time that from the perspective of management attention, through the text analysis of the annual reports of listed companies in strategic emerging industries, the TMTAI to identify credit risks can be mined. Attention is the limited cognitive resource of enterprise strategists, and should be effectively managed to stimulate enterprise technological innovation. The more attention resources point to technological innovation, the more likely the company will be more competitive, thus reducing the credit risk of the enterprise. If the senior management team does not pay enough attention to technological innovation, the strategic judgments related to technological innovation will not be developed in a timely and appropriate manner, which will lead to challenges to the development of technological innovation capability and high credit risk for technological innovation enterprises in strategic emerging industries. From the characteristic importance analysis, the characteristic importance value of the TMTAI indicator constructed by this paper based on the text data of the annual report through the text analysis technology is as high as 25, which is significantly higher than the commonly used technical innovation indicators such as R&D investment, intangible assets, and traditional financial indicators. As for the two mainstream ensemble learning models of credit risk assessment, RF model and XGBoost model, whether compared with the traditional manufacturing credit risk assessment system with pure financial indicators, or the credit risk assessment system with common technological innovation indicators such as R&D investment and intangible assets, the new credit risk assessment model embedded in the text analysis of TMTAI has more effective identification ability.

To sum up, based on the text mining of the annual reports of listed companies in strategic emerging industries, the attention of the senior management team on technological innovation is represented by the index of the attention of the management of Listed Companies in strategic emerging industries to technological innovation can significantly identify the credit risk of Listed Companies in strategic emerging industries, Moreover, the relevant qualitative text content of the annual report of listed enterprises that the management pays attention to technological innovation is an effective supplement to the quantitative financial data for the credit risk identification of listed enterprises in strategic emerging industries. This conclusion has important reference value for the credit risk identification of SEIEs.

From the comparison of indicators, it can be seen from the analysis of characteristic importance that financial indicators such as inventory turnover, current ratio, and accounts receivable turnover still have a significant impact on credit risk identification. The characteristic importance of XGBoost model is 18, 12, and 7 respectively. From the empirical results of ensemble learning model, the $F1$ indicators of CART, RF and XGBoost models using only traditional financial data have reached more than 0.85, The traditional financial indicators still have a high predictive ability for SEIEs. In terms of the impact of a single indicator of technological innovation on the credit risk of SEIEs, from the characteristic importance analysis of the three indicators of TMTAI, R&D investment and intangible assets, it can be found that the three technological innovation indicators are more significant than most traditional financial indicators. Compared with traditional financial indicators, technological innovation indicators have a more significant impact on the credit risk of SEIEs. R&D investment is the first step in the R&D innovation process of SEIEs. The intensity of R&D investment is positively related to the output of technological innovation. The more R&D investment, the higher the probability of producing high-value and leading technologies, and the greater the possibility of obtaining and maintaining the industry competitive advantage and then obtaining excess profits. Therefore, the more SEIEs invest in R&D, the more likely they are to develop new products and technologies and enhance product competitiveness. If SEIEs lack R&D investment, their technological innovation ability will inevitably be affected. Intangible assets such as patents and intellectual property rights represent the current technological innovation capability of SEIEs and reflect the core competitiveness of enterprises in technological innovation. The larger the proportion of intangible assets such as patents and intellectual property rights in assets, the stronger the technological innovation capability of SEIEs. In the demonstration of various indicator combination models of financial indicators and technological innovation, the recognition ability of the combination forecasting model combining financial data with technological innovation information indicators such as TMTAI, R&D investment and intangible assets have been significantly improved. The $F1$ indicators of CART, RF and XGBoost models combining financial data and three kinds of technological innovation information are 0.88, 0.96 and 0.95 respectively. It is significantly higher than 0.86, 0.92 and 0.93 of the traditional model using financial data alone. The combination of technological innovation indicators obviously provides the incremental information of credit risk of SEIEs that the traditional financial indicators do not have. The stronger the technological innovation ability of enterprises, the lower the credit risk. The new credit risk assessment system has significantly improved the risk identification ability of the credit risk assessment model of SEIEs. After adding technological innovation indicators, the empirical results of the significant advantages over the traditional manufacturing credit risk assessment system have obvious reference value for the credit risk identification of SEIEs.

From the perspective of the dimensional combination of technological innovation indicators, compared with the traditional enterprise credit risk assessment system, the new credit risk assessment system for SEIEs proposed in this paper not only adds the annual report text analysis technology as a credit risk information supplement to the traditional quantitative data, but also based on the financial indicators commonly used in the credit risk identification of traditional manufacturing enterprises. From the perspective of management attention, R&D investment and technological achievements that affect the technological innovation of enterprises, we add three important indicators of technological innovation, namely, the TMTAI, R&D investment and intangible assets, to identify the credit risk of SEIEs. From the empirical comparison of the impact of financial indicators and technological innovation portfolio indicators on the credit risk of SEIEs, it can be found that the combination of three indicators of technological innovation and traditional indicators can significantly improve the credit risk identification ability of SEIEs. This may be because technological innovation capability is the internal performance of the core competitiveness of SEIEs. The senior management team of SEIEs with healthy and sustainable development can pay more attention to technological innovation, increase R&D investment, so as to obtain more intangible assets such as patents and intellectual property rights, improve the technological innovation ability and market competitiveness of enterprises, and reduce credit risk. However, poorly managed SEIEs may fall into a continuous shortage of funds. The senior management team has no time to consider technological innovation, thus reducing investment in technology research and development, thus losing the core competitiveness of technology represented by intangible assets such as technology patents and intellectual property rights, and eventually leading to breach of contract. Therefore, from this point of view, the reference value of the research conclusion of this paper to the credit risk identification of SEIEs lies in that the technological innovation ability can be used as a symbol of the operation status and even the credit risk status of SEIEs. Whether healthy, sustainable, chaotic or short-sighted, technological innovation capability can be effectively measured from three dimensions: management attention, R&D investment and technological achievements.

5 CONCLUSION

This paper establishes a credit risk assessment framework for SEIEs based on text mining and ensemble learning. In this study, the TMTAI indicator is constructed to describe the top management team's attention to innovation, and the three ensemble learning models are used to evaluate the credit risk of SEIEs. At the same time, combining with the listed SEIEs in China, empirical research is carried out, and management suggestions are put forward according to the analysis results. Due to the limited credibility of traditional financial data and credit risk prediction ability, this paper focuses on the impact of technical innovation ability indicators including the TMTAI based on text mining on the credit risk of SEIEs. Based on the traditional financial index system, through text mining of the annual reports of SEIEs, the innovative attention text information index of the top management team and the common technological innovation index are combined with the ensemble learning model to construct a credit risk assessment system for SEIEs, which combines the three technological innovation dimensions of management attention, R&D investment, and technological achievements. In light of the existing literature, the current research mainly focuses on the traditional manufacturing industry and financial indicators. However, there is a scant investigation into the ensemble learning model for credit risk assessment of SEIEs based on technological innovation information text mining. Therefore, this paper tries to fill this literature gap.

In summary, the main contributions of this paper are as follows.

(1) Theoretically: from the current literature, the research on credit evaluation mostly focuses on the traditional manufacturing industry and financial indicators, and the research on scientific and technological innovation information of SEIEs is less. This study establishes a credit risk assessment system for SEIEs based on scientific and technological innovation information and traditional financial indicators. The evaluation system is more comprehensive and objective. On the other hand, the variables proposed in this paper add the TMTAI based on text mining, which is different from previous studies. The research also verifies the key factor affecting the credit risk of SEIEs: scientific and technological innovation information. Through the importance of characteristics analysis of a single indicator and the comparative analysis of the model of the combined index system, this paper demonstrates that compared with the traditional credit risk assessment system, the new credit risk assessment system based on the three dimensions of technological innovation of technological strength has the advantage of identifying the credit risk of SEIEs. Finally, through the contribution of the TMTAI index constructed by this text analysis to the improvement of the credit risk identification ability of the model, it can be concluded that the relevant qualitative text content of the listed SEIEs annual report is an effective supplement to the quantitative financial data for the credit risk identification of SEIEs.

(2) Method: In the method, this research uses a combination of ensemble learning model and text mining technology. We also analyze the importance of variable characteristics in combination with XGBoost algorithm to make the evaluation system more perfect. In order to prevent the credit risk of enterprises in strategic emerging industries, it is necessary to build a credit risk assessment ensemble learning model embedded in the text analysis of management's attention to technological innovation.

(3) Practically: This paper proposes a management dimension technological innovation index based on the text analysis of the annual reports of listed companies in strategic emerging industries' technological innovation: the TMTAI index, combined with R&D investment indicators in the dimension of research and development investment, intangible assets indicators of existing technical strength and traditional financial indicators, an objective and comprehensive credit risk evaluation index system for SEIEs is constructed. The index system can be applied to banks and other financial institutions for enterprise credit evaluation. The evaluation system cannot only evaluate the credit risk of strategic emerging industry enterprises for financial institutions but also have certain guiding significance for SEIEs to optimize their own internal management.

However, this paper has some limitations. On the existing basis, more variables need to be added to make the model results more accurate. At the same time, this paper focuses on the research of SEIEs, and pays more attention to the technological innovation information of enterprises. Therefore, the influence of human capital and external capital factors is ignored in the process of variable selection. We will solve these problems in future research.

Acknowledgements

This work was supported by Beijing Logistics Informatics Research Base. We appreciate their support very much. The work presented in this study remains the sole responsibility of the authors.

6 REFERENCES

[1] Luo, Q., Miao, C., Sun, L., Meng, X., & Duan, M. (2019). Efficiency evaluation of green technology innovation of China's strategic emerging industries: An empirical analysis based on Malmquist-data envelopment analysis index. *Journal of Cleaner Production*, *238*, 117782. https://doi.org/10.1016/j.jclepro.2019.117782

[2] Prud'hommea, D. (2016). Dynamics of China's provincial-level specialization in strategic emerging industries. *Research Policy*, *45*, 1586-1603. https://doi.org/10.1016/j.respol.2016.03.022

[3] Sun, L-Y., Miao, C-L., & Yang, L. (2018). Ecological environmental early-warning model for strategic emerging industries in China based on logistic regression. *Ecological Indicators*, *84*, 748-752. https://doi.org/10.1016/j.ecolind.2017.09.036

[4] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*, 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[5] Fernandes, G. B. & Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European journal of operational research*, *2*, 517-524. https://doi.org/10.1016/j.ejor.2015.07.013

[6] Zhu, Y., Xie, C., Wang, G.-J., & Yan, X.-G. (2016). Predicting China's SME credit risk in supply chain finance based on machine learning methods. *Entropy*, *18*, 195,2016. https://doi.org/10.3390/e18050195

[7] Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science*, *174*, 150-160. https://doi.org/10.1016/j.procs.2020.06.070

[8] Zhu, Y., Xie, C., Sun, B., Wang, G.-J., & Yan, X.-G. (2016). Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models. *Sustainability*, *8*, 433. https://doi.org/10.3390/su8050433

[9] Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. *Cognitive Systems Research*, *34*(52), 317-324. https://doi.org/10.1016/j.cogsys.2018.07.023

[10] Wang, G. & Ma, J. (2012). A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. *Expert Systems with Applications*, *39*, 5325-5331. https://doi.org/10.1016/j.eswa.2011.11.003

[11] Golbayani, P., Florescu, I., & Chatterjee, R. (2020). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *North American Journal of Economics and Finance*, *54*, 101251. https://doi.org/10.1016/j.najef.2020.101251

[12] Wang, G. & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications*, *38*, 13871-13878. https://doi.org/10.1016/j.eswa.2011.04.191

[13] Zhu, Y., Xie, C., Wang, G.-J., & Yan, X.-G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing*, *28*, 41-45. https://doi.org/10.1007/s00521-016-2304-x

[14] Chen, X., Wang, X., & Wu, D. D. (2010). Credit risk measurement and early warning of SMEs: An empirical study of listed SMEs in China. *Decision Support Systems*, *49*, 301-310. https://doi.org/10.1016/j.dss.2010.03.005

[15] Lu, Y.-C., Shen, C.-H., & Wei, Y.-C. (2013). Revisiting early warning signals of corporate credit default using linguistic analysis. *Pacific-Basin Finance Journal*, *24*, 1-21. https://doi.org/10.1016/j.pacfin.2013.02.002

[16] Yin, C., Jiang, C., Jain, H. K., & Wang, Z. (2020). Evaluating the credit risk of SMEs using legal judgments. *Decision Support Systems*, *136*, 113364. https://doi.org/10.1016/j.dss.2020.113364

[17] Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*, 164-175. https://doi.org/10.1016/j.dss.2010.07.012

[18] Chen, S., Bu, M., Wu, S., & Liang, X. (2015). How does TMT attention to innovation of Chinese firms influence firm innovation activities? A study on the moderating role of corporate governance. *Journal of Business Research*, *68*, 1127-1135. https://doi.org/10.1016/j.jbusres.2014.11.002

[19] Kaplan, S. (2008). Cognition, capabilities, and incentives: Assessing firm response to the fiber-optic revolution. *Academy of Management Journal*, *51*(4), 672-695. https://doi.org/10.5465/AMJ.2008.33665141

[20] Wojan, T. R., Crown, D., & Rupasingha, A. (2018). Varieties of innovation and business survival: Does pursuit of incremental orb far-ranging innovation make manufacturing establishments more resilient? *Research Policy*, *47*, 1801-1810. https://doi.org/10.1016/j.respol.2018.06.011

[21] Lee, J.-W. (2021). Analysis of technology-related innovation characteristics affecting the survival period of SMEs: Focused on the manufacturing industry of Korea. *Technology in Society*, *67*, 101742. https://doi.org/10.1016/j.techsoc.2021.101742

[22] Zhang, D., Zheng, W., & Ning, L. (2018). Does innovation facilitate firm survival? Evidence from Chinese high-tech firms. *Economic Modelling*, *75*, 458-468. https://doi.org/10.1016/j.econmod.2018.07.030

[23] Li, S., Shang, J., & Slaughter, S. A. (2010). Why do Software Firms Fail? Capabilities, Competitive Actions, and Firm Survival in the Software Industry from 1995 to 2007. *Information Systems Research*, *21*(3), 631-654. https://doi.org/10.1287/isre.1100.0281

[24] Fernandes, A. M. & Paunov, C. (2015). The risks of innovation: are innovating firms less likely to die? *Review of Economics and Statistics*, *97*(3), 638-653. https://doi.org/10.1162/REST_a_00446

[25] Helmers, C. & Rogers, M. (2010). Innovation and the survival of new firms in the UK. *Review of Industrial Organization*, *36*(3), 227-248. https://doi.org/10.1007/s11151-010-9247-7

[26] Zhang, M., He, Y., & Zhou, Z.-F. (2013). Study on the influence factors of high-tech enterprise credit risk: empirical evidence from China's list companies. *Procedia Computer Science*, *17*, 901-910. https://doi.org/10.1016/j.procs.2013.05.115

[27] Angilella, S. & Mazzù, S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. *European Journal of Operational Research*, *244*, 540-554. https://doi.org/10.1016/j.ejor.2015.01.033

[28] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *Biometrics*, 173.

[29] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

[30] Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

[31] Zhichao, L., Pingyu, H., & Ni, X. (2020). SME Default Prediction Framework with the Effective Use of External Public Credit Data. *Sustainability*, *12*, 7575. https://doi.org/10.3390/su12187575

Contact information:

Yang MAO
School of Economics and Management,
Beijing Jiaotong University, Haidian, 100044, China
E-mail: 18113060@bjtu.edu.cn

Shifeng LIU, Professor
School of Economics and Management,
Beijing Jiaotong University, Haidian, 100044, China
E-mail: shfliu@bjtu.edu.cn

Daqing GONG
(Corresponding author)
School of Economics and Management,
Beijing Jiaotong University, Haidian, 100044, China
E-mail: dqgong@bjtu.edu.cn