ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20230316000447
Original scientific paper

A Hybrid Technological Innovation Text Mining, Ensemble Learning and Risk Scorecard Approach for Enterprise Credit Risk Assessment

Yang MAO, Shifeng LIU, Daqing GONG*

Abstract: Enterprise credit risk assessment models typically use financial-based information as a predictor variable, relying on backward-looking historical information rather than forward-looking information for risk assessment. We propose a novel hybrid assessment of credit risk that uses technological innovation information as a predictor variable. Text mining techniques are used to extract this information for each enterprise. A combination of random forest and extreme gradient boosting are used for indicator screening, and finally, risk scorecard based on logistic regression is used for credit risk scoring. Our results show that technological innovation indicators obtained through text mining provide valuable information for credit risk assessment, and that the combination of ensemble learning from random forest and extreme gradient boosting combinations with logistic regression models outperforms other traditional methods. The best results achieved 0.9129 area under receiver operating characteristic. In addition, our approach provides meaningful scoring rules for credit risk assessment of technology innovation enterprises.

Keywords: ensemble learning; risk assessment; risk scorecard; technological innovation; text mining

1 INTRODUCTION

With the sustained growth of the global economy and the increasing scale of enterprises, enterprise credit risk assessment has become an increasingly important area of research. Enterprise credit risk assessment is to evaluate the credit quality and solvency of an enterprise so that financial institutions, investors and regulators can better understand the enterprise's risk level and make better risk management and investment decisions.As technology innovative enterprises are characterized by asset-light and high R&D risks, the backward-looking financial information used in historical literature to assess traditional enterprises can hardly reflect their risk characteristics; in addition, the selection of credit evaluation indexes is not objective, and it is difficult to balance the transparency and accuracy of the assessment results. The goal of this study is to establish a set of credit risk evaluation system that combines objectivity, accuracy and transparency without additional expertise compared with the traditional evaluation methods for the characteristics of technology innovation enterprises.

The theoretical framework of enterprise credit risk assessment can be divided into three parts: the analysis of enterprise credit risk factors, access to enterprise credit risk assessment indicators, and the selection of assessment model, and the following literature is sorted out and discussed according to the framework.

Firstly, most of the above methods are based on financial variables, the most obvious drawback being that financial information is inherently backward looking and based on historical information rather than an assessment of the future [1]. For a more accurate credit assessment analysis it is necessary to use information that can look forward. With regard to the factors affecting the credit risk of technology innovators, in recent yearsmany scholars have found that technological innovation can significantly affect enterprise survival. Mao et al. discovered that adding innovative key indicators of executive teams from the annual reports of listed companies significantly improves the credit risk prediction ability of strategic emerging industry companies [2]. Lee et al. conducted an analysis of the technological in-novation characteristics that influence the survival period of small and medium-sized enterprises (SMEs) in the Korean manufacturing industry. They found that innovation features related to technology have a positive impact on the survival of SMEs [3]. Fernandes and Paunov found that new products can increase the likelihood of firm survival under specific conditions [4]. Zhang et al. showed that innovation, as measured by patents, innovation efficiency, and enterprise import and export activities, can improve the survival rate of high-tech enterprises in China [5], and Wojan et al. demonstrated that the vast majority of manufacturing companies that survive in the long-term shift from non-innovative strategies to incremental or broader innovation orientations [6].

Secondly, traditional methods of obtaining risk assessment indicators are subjective and difficult to standardize, and mining textual information can avoid the influence of individual subjective preferences. Text mining technology can obtain textual information from relevant texts to predict corporate credit risk. Previously, scholars have applied text mining technology to mine public opinion text information and news text information [7], legal text information [8], listed company related text executive tone [9] or attention information [2] to assess corporate credit risk. Unfortunately, text mining techniques are less frequently applied to the construction of technological innovation index systems for predicting corporate credit risk.

Finally, with regard to the selection of corporate credit risk assessment models, many researchers have attempted to use statistical analysis and machine learning techniques to build automated classification systems to solve the credit prediction problem. The former includes discriminant analysis [10] and logistic regression [11], while the latter includes ensemble learning [12], neural networks [13] and support vector machines [14]. However, while these are well-established techniques for credit assessment and prediction, a number of problems have arisen. Logistic regression (LR) is widely used in credit risk assessment because of its effectiveness and robustness [15]. Many studies have focused on using the LR method to score the credit risk of enterprises based on the probability of default. For example, Pederzoli and Thomas [16] used LR method to analyze the default probability of enterprises. However,

although logistic regression models do not require assumptions about the probability distribution of feature variables, nor do they require equal covariance, with strong model interpretability and better model robustness, and are able to derive risk scores based on classification probabilities, only a small number of variables can be selected, making them unable to be directly applied to risk assessment in a big data environment, and thus have limited prediction accuracy. Machine learning models do not need to be modelled under strict assumptions and do not require any data distribution, and have high prediction accuracy. However, the approach is poorly interpretable, suffers from algorithmic 'black box' problems, is not robust to changes in risk data, and is difficult to grasp the logic within the machine learning model. Some scholars have therefore combined the two, selecting important predictive variables based on the importance of the variables obtained from machine learning, and using different independent variables as input to logistic regression or other risk assessment models that can identify classification logic. To find the relative importance of potential predictor variables, Yeh et al. used random forest (RF) for screening of indicators and used the screened indicators to analyze the credit rating of firms using rough sets [17], while others combined extreme gradient boosting (XGBoost) and logistic regression to build a logistic regression model for uncontrolled screening of new coronary pneumonia using extreme gradient boosting of screened indicators [18], but existing studies lack a comparative analysis and combined application of these two different ensemble learning screening metrics methods, especially for enterprise credit risk assessment studies.

In this study, we use information based on technological innovation as a predictor variable. In the proposed approach, text mining techniques are used as a tool to extract technological innovation information for each technological innovation enterprise. To validate the proposed approach, we use a hybrid model that combines RF and XGBoost, two ensemble learning algorithms with different principles, and risk scorecard based on LR to improve the accuracy of credit risk assessment. Firstly, we used financial and technological innovation variables as potential predictor variables for credit risk scoring. Secondly, RF and XGBoost were used for variable selection because of their reliability in obtaining the relative importance of predictor variables. Next, we used the significant predictor variables obtained from RF and XGBoost as inputs to the LR model. We can then compare the results obtained to see if the models, including various machine learning models, provide better classifi-cation accuracy. Finally, we produce results in the form of scoring rules that are transparent and easy to understand for decision makers.

The rest of the paper is organized as follows. Section 2 describes the methods used in this paper: TF-IDF, Word2Vec, RF, XGBoost, risk scorecard based on LR, and the hybrid text mining, combined ensemble learning with risk scorecard strategy and experimental framework used in this study, respectively. Section 3 presents the experimental results of the proposed methods. Section 4 provides an analysis and discussion of the experimental results, and finally Section 5 concludes.

2 RESEARCH METHODS

2.1 Text Mining

2.1.1 *TF-IDF* Statistical Method

Term Frequency-inverse Document Frequency (TF-IDF) is a statistical method for evaluating the importance of words to a document or a document collection. The importance of a word increases proportionally with the number of times it appears in a document, but decreases inversely with the frequency with which it appears in the corpus. If a word or phrase occurs more frequently in one category and less frequently in other categories, it is considered to have good category differentiation ability and is suitable for classification. The Inverse Document Frequency (IDF) of a given word can be obtained by dividing the total number of documents by the number of documents containing the word, given the total number of documents in the corpus.

Using text analysis, we searched for top *TF-IDF* key words in the word set. We calculate the *TF-IDF* as follow, where n_j is the number of occurrences of innovation keyword in technical innovation keyword set j, and the denominator is the sum of the occurrences of all the words in technical innovation keyword set j. $|D|$ is the total number of files contained in the corpus. The denominator is the number of files containing the innovative keyword t.

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

2.1.2 Word2Vec Method

Word vectors, or Word2Vec, is a natural language processing-based word processing technique, researched and developed by Mikolov T. et al [19], which essentially uses shallow neural networks to autonomously learn the frequency and position of utterances or words in a corpus and embed them in a space of moderate dimensionality for vectorization purposes. Word2Vec provides a new research tool in the field of natural language processing by enabling the rapid and efficient representation of words as vectors based on a given corpus, using optimal training patterns.

The vector of words consists of two training models: the continuous bag-of-words model (CBOW) and the skip-gram. -gram predicts the current word from the context, while skip-gram predicts each of the j words in the context from the current word. Both training models are shallow neural networks with an input layer, an implicit layer and an output layer.

In the skip-gram model, the weights between the input and implicit layers are represented as a $V \times N$ matrix $\boldsymbol{W}$, where V represents the number of feature words in the technological innovation lexicon formed by text mining and N represents the number of neurons in the implicit layer. Each row in $\boldsymbol{W}$ is an $\boldsymbol{N}$ - dimensional vector associated with the corresponding feature word in the input layer, and the ith feature word w_i. The corresponding row vector in W is denoted as v_{w_i} . Assuming that the input layer has a unique thermal encoding vector $x \in R^V$ for the feature word w_i, where only $x_i = 1$ and the rest are 0, the

corresponding implicit layer vector h for x can be expressed as.

$$h = x^T \cdot \boldsymbol{W} = v_{W_i} \tag{2}$$

The weights from the implicit layer to the output layer are represented by an $N \times C$ dimensional matrix $\boldsymbol{W}'$, where C is the number of feature word w_i window context vectors. Let the corresponding vector of the jth contextual word in $\boldsymbol{W}'$ be u_j, then the feature word w_i is related to the jth contextual word as the vector of words consists of two training models: the continuous bag-of-words model (CBOW) and the skip-gram. -gram predicts the current word from the context, while skip-gram predicts each of the j words in the context from the current word. Both training models are shallow neural networks with an input layer, an implicit layer and an output layer.

$$u_{i,j} = v_{W_i} \cdot u_j \tag{3}$$

By training the model and adjusting the weight matrices $\boldsymbol{W}$ and $\boldsymbol{W}'$ to maximize the probability of the feature words generating context words, the word vectors for all words are found. The formula for the cosine similarity between word vectors is as follows:

$$\cos\theta = \frac{\sum(v_1 \times v_2)}{\sqrt{\sum(v_1)^2} + \sqrt{\sum(v_2)^2}} \tag{4}$$

where v_1 and v_2 represent the two-word vectors for which the cosine similarity is to be calculated. The resulting cosine similarity between the corresponding word vectors is used as an important objective basis for the subsequent determination of the content of the variable system.

2.2 Ensemble Learning

2.2.1 Bagging Ensemble -Random Forest Model (RF)

The bagging ensemble principle aims to enhance classification performance by combining multiple classifiers that generate classification results from distinct training sets. The Random Forest (RF) algorithm, proposed by Breiman [20], is a representative algorithm based on bagging ensemble. This algorithm comprises multiple decision trees within a classifier. The final classifier is formed by selecting independent decision trees from multiple groups, and the ultimate classification is obtained by averaging the output of each decision tree. The RF algorithm incorporates random selection of data subsets and subsets at each node, where the best partition is calculated solely within a given subset. This structure provides uncorrelated or weakly correlated predictions. RF offers the advantages of processing unbalanced data sets, reducing the probability of overfitting, and fast training speed, the principle of RF algorithm can be summarized as follows.

K samples, each with the same size as the original training set, are drawn from the original training set to establish K decision trees, and the final classification result is determined by voting according to the K classification results. Random Forest utilizes training sets with diverse structures to enhance the diversity between classification models, which in turn improves the generalization prediction ability of the integrated model. After K rounds $\{h_1(x), h_2(x), ..., h_k(x)\}$of training, a sequence of classification models is formed. The models are then integrated into a single classifier using the simple majority voting method to determine the final result.

Final classification decision:

$$H(x) = arg\max \sum_{i=1}^{k} I(h_i(x)Y) = \\ = \begin{cases} I(\alpha) = 1 \text{ if } \alpha \text{ is true,} \\ I(\alpha) = 0 \text{ other wise.} \end{cases} \tag{5}$$

$H(x)$ represents the combined classification model, h_i represents the single classification tree model, Y represents the target variable, and Eq. (5) represents that the final classification result is determined by voting.

2.2.2 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 [21] 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 ith tree can be expressed as:

$$\tilde{y}_1 = \sum_{k=1}^{k} f_k(x_i) \tag{6}$$

$f_k(x_i)$ represents the prediction result of the kth tree, and:

$$loss = \sum_i l(\tilde{y}_1, y_i) + \sum_k \Omega(f_k) \tag{7}$$

$$\Omega\left(f_k\right)=T+\frac{1}{2}\lambda\left\|w\right\|^2 \tag{8}$$

Where: $\tilde{y}_1$ 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 kth 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.

2.3 Scoring Model Based on Logistic Regression

Logistic regression, also known as a logit model, is a widely used credit risk forecasting tool. It uses a non-linear maximum log-likelihood technique to estimate the conditional probability of default of a firm based on external variables. The logit model is exported in the following form:

$$P_i=E\left(Y=1\middle|X_i\right)=\frac{1}{1+e^z}= \\ =\frac{1}{1+e^{-\left(\alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_nX_{in}\right)}} \tag{9}$$

where P_i is a vector of attribute variables X_{ij} given firm i, the probability of failure of firm i, and β_j is the parameter to be estimated. According to most of the existing literature, we refer to the model as a model constructed using a dummy variable Y, which has a value of 1 to indicate high risk and 0 to indicate low risk.

Alternatively, Eq. (9) can be deformed as follows:

$$\log\left(odds\right)=\log\left(\frac{P_i}{1-P_i}\right)=\alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_nX_{in} \tag{10}$$

What is obtained by Eq. (10) is an estimate of the firm's probability of default, but in practice it usually needs to be converted into a simple and intuitive score, which is to normalize the probability value to a score value. The calibration of the scores is the process of establishing a certain correspondence between the final model scores and the good/bad ratio (*odds*) through a linear transformation. odds is the ratio of the good sample probability P to the bad sample probability $1-P$ and is expressed as $odds = P/(1-P)$.

The scale of scores set for risk scoring can be defined by expressing the scores as a current expression for the logarithm of the ratio as follows [22].

$$score=A-B\cdot\log\left(odds\right)= \\ =A-B\left(\alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_nX_{in}\right) \tag{11}$$

The score for a particular point with a ratio of x (i.e. odds) is set to P_0, and PDO is the score that needs to be raised for odds to be twice as high. The score for a point with a ratio of $2x$ is $P_0 + PDO$. Taking Eq. (11) into account gives the values of both P_0 and PDO are known constants, the calculated A and B values are brought into Eq. (11) to obtain the risk score scores for different x [23]. It is worth noting that only the logistic regression model can calculate the score for each variable giving the form of a scorecard, other models have no way of calculating variable scores as they cannot give information on weights or parameters, hence the logistic regression model was chosen to construct risk scorecard for this study.

2.4 Experimental Results Metrics

2.4.1 Feature Importance

RF and XGBoost algorithms utilize a variable importance index to rank variables based on their significance to classification accuracy, while considering interactions between variables. The importance of a variable is estimated by examining the increase in prediction error when data not in the bootstrap sample of the variable are replaced while all other data remain constant. In the context of credit risk for technology innovation companies, the influence of financial variables and technological innovation information variables may not be equal. Thus, the relative importance method is utilized to examine their significance to credit risk. This method estimates the sum of improvements made by using target variables to segment the internal nodes of a decision tree. Specifically, the method involves merging the subtree into a terminal node and calculating the difference in squared error between the given node and the proposed variable. According to the research of Luo et al. [23], the principle is as follows:

Given a set of variables $V=\{(v_1, v_2, ..., v_i\}$, is a decision tree, T is an intermediate node, $t\in T$, $\hat{l}_t^2\left(v_l\right)$ 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:

$$I_l\left(T\right)=\sum_{t\in T}\hat{l}_t^2\left(v_l\right) \tag{12}$$

In the decision tree T, the relative importance of variable v_l is denoted by $I_l(T)$. Specifically, on each internal node, the ratio of the improvement in the splitting solution attributed to variable V to the total improvement in the same solution is calculated. Typically, the relative importance of variable V is the sum of all these improvement ratios across all internal nodes where V is chosen as the splitting variable. Once the relative importance of variable V is determined for a single tree, an overall measure of variable importance can be constructed. For a set of decision trees, the importance of variable V can be summarized using the mean value. Specifically, given a set of decision trees $T_1, ..., T_M$ the importance of variable v_l in this set of trees is determined as follows:

$$I_l\left(T_1,\ \ldots,T_M\right)=\frac{1}{M}\sum_{m=1}^{M}I_l\left(T_M\right) \tag{13}$$

2.4.2 Model Performance Metrics

We use the area under curve (AUC) metric to measure the performance of the model and quantify the performance of the classification model by calculating the area under the

receiver operating characteristic curve (ROC) curve. Using the false positive rate (*FPR*) as the horizontal axis and the true positive rate (*TPR*) as the vertical axis, the ROC curve can be derived. Where *FPR* and *TPR* are calculated as follows.

$$FPR = FP/(FP + TN) \tag{14}$$

$$TPR = TP/(TP + FN) \tag{15}$$

where *FP* stands for false positive, which is the number of actual low-risk firms in the sample that the model incorrectly identifies as high-risk. *TP* stands for true positive, which is the number of actual high-risk firms in the sample that the model correctly identifies as high-risk. *FN* stands for false negative, which is the number of low-risk enterprises identified by model errors in all actual high-risk enterprises in the sample. *TN* stands for true negative, which is the number of firms in the sample that were correctly identified as low risk by the model, out of all the actual low risk firms.

2.5 Proposed Methodology

This study proposes a hybrid credit risk assessment method based on text mining, combined ensemble learning and risk scoring card, in which text mining is used to mine the technical innovation indicator system, and combines with the traditional financial indicator system to form the credit risk assessment indicator system, and then combines the RF and XGBoost Ensemblelearning algorithms to screen indicators, finally, the credit risk is scored using the scorecard model based on logical regression. The detailed process is as follows.

Firstly, we preprocess the text of the technical innovation part in the inquiry letter of the listed audit experts of the science and technology innovation board, and form the technical innovation keyword set by combining the dictionary established by reading the documents screened by the co-citation analysis, then calculated the *TF-IDF* indicators of these technological innovation keywords and ranked them to obtain the primary indicators, and then used Word2Vec to calculate the similarity of word vectors to obtain the secondary indicators to form the technological innovation indicator system, and then combined the financial variables and the technological innovation variables obtained from text mining were then used as potential predictor variables.

Secondly, two ensemble learning methods, RF and XGBoost, were used for variable selection because of their reliability in obtaining the relevant importance of the predictor variables. Next, the significant predictor variables obtained from RF and XGBoost as input to the LR model, and the dataset was divided into two sub-datasets, with 60% of the collected data serving as the training set and the remaining 40% as the test set. The divided dataset was also used to balance the data using the Synthetic Minority Oversampling Technique (SMOTE) [24] .

Finally, the credit risk scorecard based on logical regression was used to credit risk assessment,where the performance of the method proposed in this study can be assessed by AUC of logistic regression.

The methodological framework diagram shown in Fig. 1. The integration of different technologies (technology innovation, text mining, ensemble learning, risk scorecard) is based on the following principles: firstly, text mining extracts technology innovation related information from the audit questionnaire to refine the technology innovation indexes, and then the extracted technology innovation indexes and the traditional financial indexes are sorted and screened for feature importance by two different ensemble learning methods, namely "bagging" and "boosting". The screening results of the two methods are combined as the credit risk assessment index system, and finally the index system is input into the risk scorecard model to get the credit risk score of the enterprise.

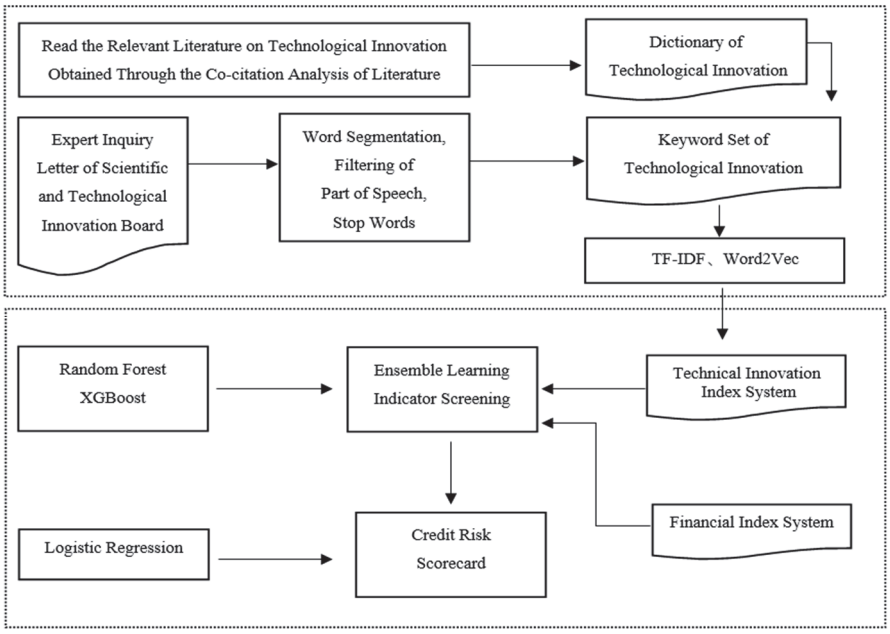


Figure 1 Framework of hybrid approach to credit risk assessment for technology innovation enterprises

3 RESULTS

3.1 Data

A listing audit enquiry letter is a letter with a series of questions and requests sent to a company by the securities regulator when reviewing its application for listing. The Science and Technology Innovation Board is a board set up to provide a financing and development platform for science and technology innovation enterprises, and the companies on the Science and Technology Innovation Board are usually involved in the technology field, so the questions in the audit enquiry letters often involve technological innovation, core technology and other aspects, with more questions on technological innovation. In this study, we downloaded and collated all the texts of technological innovation in the expert audit enquiry letters of 298 companies listed on the Science and Technology Innovation Board of the Shanghai Stock Exchange, and ob-tained approximately 160000 bytes of initial technological innovation text data. The data cleaning process was carried out by removing deactivated words, separating words from text, filtering lexicons and setting dictionary of technological in-novation.

In order to make a dictionary of technological innovation, we use the ISI Web of Science core journal citation index database for retrieval, and the retrieval content is related to technological innovation. Therefore, the search field is "TI = (technological innovation)", and the papers are retrieved by title. From 2012 to 2022, 937 related papers were retrieved. Then, CiteSpace software was used to analyze the co-citation of these papers. A total of 11 papers with citation frequency greater than 5 were selected, as shown in Fig. 2. Read in the order of the number of citations, and collect the terms used in each paper to record the mode, behavior and influencing factors of the enterprise's technological innovation activities, and finally translate them into Chinese to form a technological innovation dictionary.

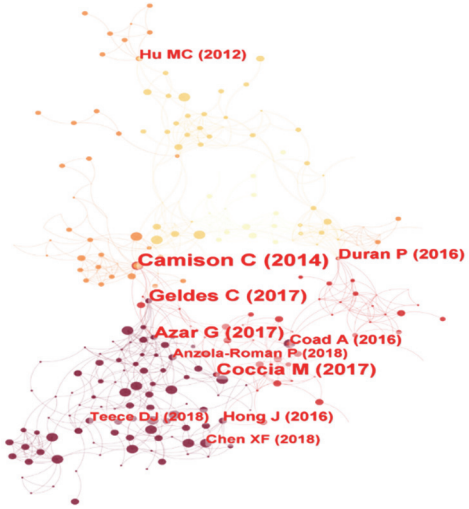


Figure 2 Literature co-citation analysis

After completing the data cleaning process, a large lexicon of approximately 130000 words was obtained that was particularly relevant to the subject term 'technological innovation'. This can be achieved by using the jieba in the PyCharm software based on Python, combined with the corresponding programming techniques. This provides research ideas and technical support for the construction of the first level indicators of the technology innovation index system below.

The model sample includes 460 listed technology innovation enterprises in China. The industries of technology innovation enterprises include the seven sectors of strategic emerging industries proposed by the Chinese government: energy conservation and environmental protection, new energy, new energy vehicles, biotechnology and medicine, new materials, new generation IT, high-end equipment manufacturing [25]. There are 392 low-risk enterprises and 68 high-risk enterprises. High-risk enterprises are defined as ST or *ST listed enterprises, with ST and *ST indicating special listed enterprises that have received delisting risk warnings due to abnormal financial conditions, and low-risk enterprises are defined as non-ST and *ST normally listed enterprises that have not received delisting risk warnings. We validate the validity of the proposed risk assessment methodology by using data from 2021 to predict whether a delisting risk warning will be received in 2023. 60% of the collected data was used for training and the remaining 40% was used for testing. Given the imbalance between the high-risk and low-risk samples, we used the SMOTEto balance the split sample data separately, resulting in a total of 784 data for 392 high- risk samples and 392 low-risk samples. The financial data and technical innovation data except for technical disputes in this study were sourced from Choice Financial Terminal, and the technical dispute data were sourced from the China Judgment Document Network.

3.2 Potential Predictor Variables

By referring to the existing representative literature on technological innovation and by ranking the high *TF-IDF* words in the text database as shown in Tab. 1, the main content of the technological innovation indicator system was divided into five aspects according to the characteristics of these keywords, namely: industry, R&D investment, R&D achievement, core technology products, and the existence of technology licensing disputes. The five first level variables of the index system were thus determined, namely industry, R&D investment, R&D achievement, core technology products, and technology disputes, and the keywords of R&D investment and R&D achievement were used as the first level variables for the similarity search of the second level variables under them.

When the angle between two vectors is assumed to be between 0 and 90°, the value of the cosine can indicate the magnitude of the angle value. As the vectors in this paper are all generated by the word vector method, their angle range is generally limited to 0 to 90°, and the value of the cosine of the angle is (1, 0). The larger the value of the cosine of the angle, the more adjacent the two vectors are, and the stronger the correlation between the words; the smaller the value of the cosine of the angle, the more distant the two vectors are, and the weaker the correlation between the two words. Therefore, in the use of the word vector method, the magnitude of the cosine value can indicate the magnitude of the two angles, and the similarity or semantic relevance of the two language vectors in the corpus can also be obtained.

Table 1 Key words *TF-IDF* ranking table

Serial number	Keyword nam (Chinese)	Keyword name (English)	*TF-IDF*
1	hexinjishu	Core Technology	0.2497
2	jishu	Technology	0.1593
3	chanpin	Product	0.1466
4	zhuanli	Patent	0.1436
5	yanfa	R&D	0.1112
6	yanfafeiyong	R&D expense	0.0811
7	jishurenyuan	Technical personnel	0.0677
8	yanfarenyuan	R&D personnel	0.0573
9	famingzhuanli	Invention patent	0.0519
10	hangye	Industry	0.0492
11	zhishichanquan	Intellectual property	0.0481
12	jiufen	Dispute	0.0401
13	yanfaxiangmu	R&D Project	0.0380
14	cunzaijiufen	Existing dispute	0.0332
15	yanfatouru	R&D investment	0.0330
16	jingying	Operating	0.0318
17	yewu	Operation	0.0303
18	jishushuiping	Technology level	0.0300
19	shouquan	Licensing	0.0292
20	chengguo	Achievement	0.0290

In this paper, we use first level variables previously obtained through *TF-IDF* to expand the second level variables using word vectors. First, input the previously established keyword set of technological innovation into the word2vec library of python and establish a word vector model. Then, the words "R&D investment" and "R&D achievements" make the program output words with high cosine similarity, and combine the rationality of the output result data to further search for words with high cosine similarity, as shown in Tab. 2. Finally, proportion of R&D expenses in operating income and proportion of R&D personnel as the second level variables of R&D investment, and intellectual property, invention patent and utility models patent as the second level variables of R&D achievement.

The financial index system of technology innovation enterprises refers to the study of Luo and Zhang [26] while eliminating indicators with many missing values. The financial index system is divided into 4 first level variables, namely profitability, leverage, growth and operational capability, as well as 9 second level variables, namely ROA, ROE, current ratio, quick ratio, asset liability ratio, net asset growth rate, net profit growth rate, current assets turnover ratio, and accounts receivable turnover rate.

Based on the definition of the concept of technological innovation, this paper focuses on representative industry sectors, and takes into account the availability and comparability of data, and constructs a technological innovation indicator system with 5 first level variables and 8 second level variables, as well as a financial indicator system with 4 first level variables and 9 second level variables, as shown in Tab. 3.

Table 2 Word2 Vec word vector similarity correlation table

Original keyword （Chinese）	Original keyword (English)	Similar keyword (Chinese)	Similar keyword (English)	Degree of similarity
yanfatouru	R&D investment	yanfarenyuan	R&D personnel	0.9977
yanfarenyuan	R&D personnel	yanfafeiyong	R&D expenses	0.9948
yanfachengguo	R&D achievement	zhishichanquan	Intellectual property	0.9943
zhishichanquan	Intellectual property	datadata	Invention	0.9972
zhishichanquan	Intellectual property	zhuanli	Patent	0.9954
zhuanli	Patent	shiyongxinxing	Utility models	0.9951

Table 3 Technological innovation information and financial information indicator system table

Variable Category	First Level Variables	Second Level Variables	Indexes	Variables Description
Technological Innovation Information	Industry	Industry	X_I	Industry the enterprise belonged to
	R&D investment	Proportion of R&D expenses in operating income	X_{PREO}	Proportion of R&D expenses in operating income
		Proportion of R&D personnel	X_{PRP}	Proportion of R&D personnel
	R&D achievement	Intellectual property	X_{INTP}	Number of invention patent
		Invention patent	X_{INVP}	Number of intellectual property
		Utility model patent	X_{UMP}	Number of utility model patent
	Core technology products	Proportion of core technology revenue	X_{PCTR}	Proportion of core technology revenue
	Technical disputes	Technical disputes	X_{TD}	Whether there is a technical dispute lawsuit
Financial Information	Profitability	ROA	X_{ROA}	Net profit/total assets
		ROE	X_{ROE}	Net profit/average net assets
	Leverage capacity	Current ratio	X_{CR}	Current assets/ current liabilities
		Quick ratio	X_{QR}	Quick assets/current liabilities
		Asset liability ratio	X_{ALR}	Total liabilities/total assets
	Growth capacity	Net asset growth rate	X_{NAGR}	(Closing net assets - opening net assets)/opening net assets
		Net profit growth rate	X_{NPGR}	(Closing net profit opening net profit)/opening net profit
	Operating capacity	Current assets turnover ratio	X_{CATR}	Net main business income/ average total current assets
		Accounts receivable turnover ratio	X_{ARTR}	Net credit sales income/average of accounts receivable

3.3 Variable Importance Ranking

Variable importance is measured as a score indicating the degree of dominance of each variable in the model. The higher the score, the greater the predictive power of the variable. In order to find the relative importance of potential predictor variables, we used the RandomForestClassifier module and the XGBoost module from the sklearn library in python to calculate the importance values of the variables.

We determine the significance of the characteristics of the explanatory variables by applying RF. The results shown in Fig. 3 indicate that invention patent show the strongest ability, followed by net asset growth rate, current ratio, asset liability ratio, intellectual property, ROE, ROA, proportion of R&D expenses in operating income, net profit growth rate, quick ratio, proportion of R&D personnel, utility model patent, accounts receivable turnover ratio, current asset turnover rate, proportion of core technology revenue, industry, and technology disputes. The five indicators that are significantly lower are accounts receivable turnover ratio, current assets turnover rate, proportion of core technology revenue, industry, and technology disputes. When we use RF to filter variables, we can delete them.

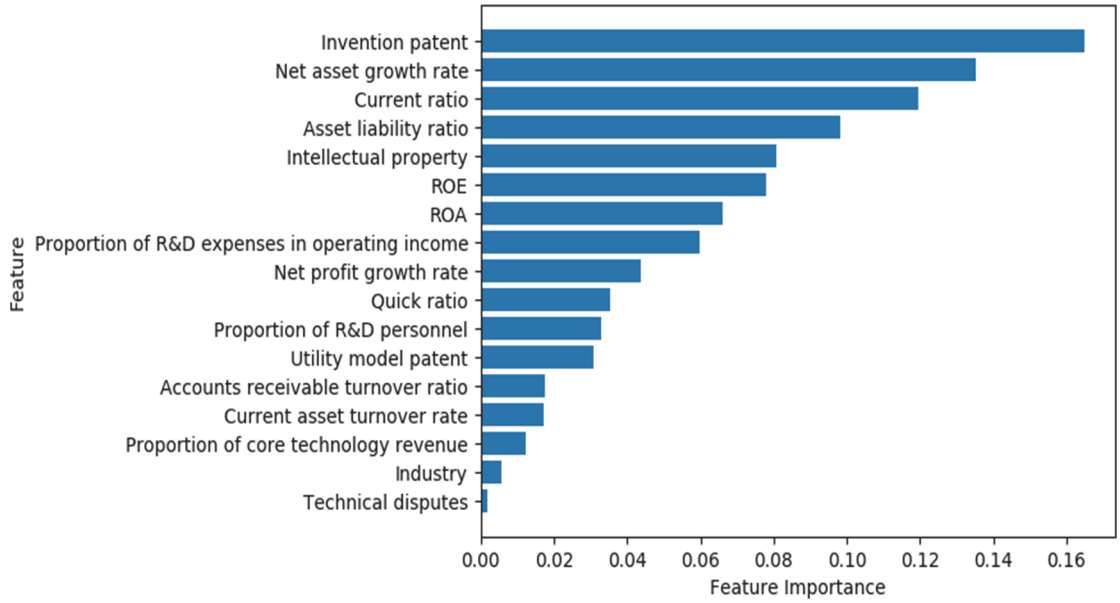


Figure 3 RF feature importance ranking chart

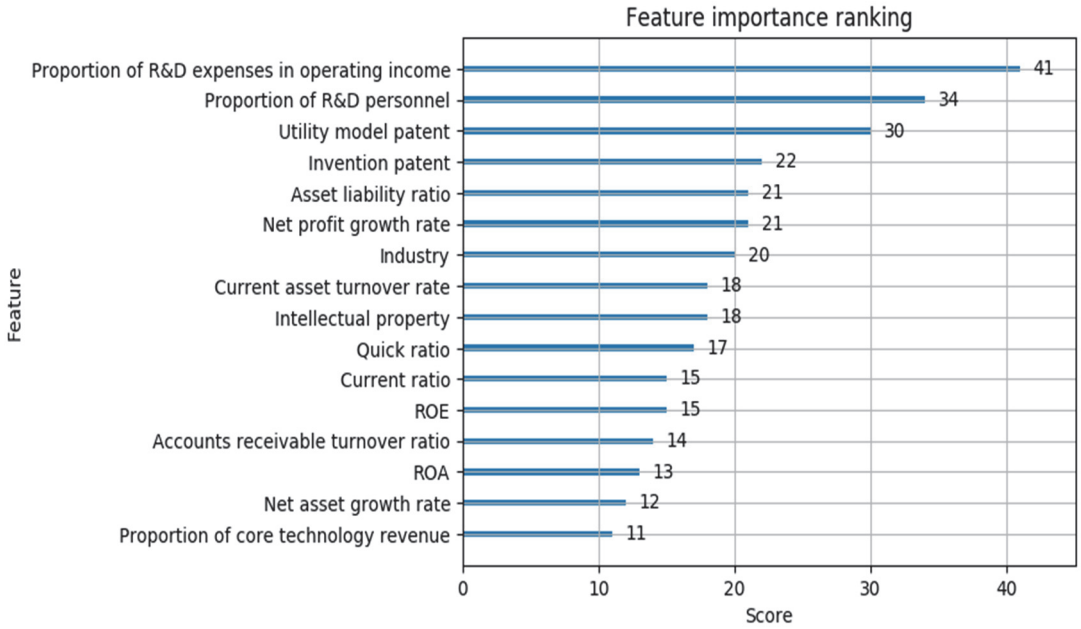


Figure 4 XG Boost feature importance ranking chart

We can also determine the significance of the characteristics of the explanatory variables by applying XGBoost. The results shown in Fig. 4 indicate that proportion of R&D expenses in operating income shows the strongest power, followed by proportion of R&D personnel, utility model patent, invention patent, asset liability ratio, net profit growth rate, industry, current asset turnover ratio, intellectual property, quick ratio, current ratio, ROE, accounts receivable turnover rate, ROA, net asset growth rate and proportion of core technology

revenue. When we use XGBoost to filter variables, the top twelve indicators are retained: proportion of R&D expenses in operating income, proportion of R&D personnel, utility model patent, invention patent, asset liability ratio, net profit growth rate, industry, current asset turnover ratio, intellectual property, quick ratio, current ratio and ROE

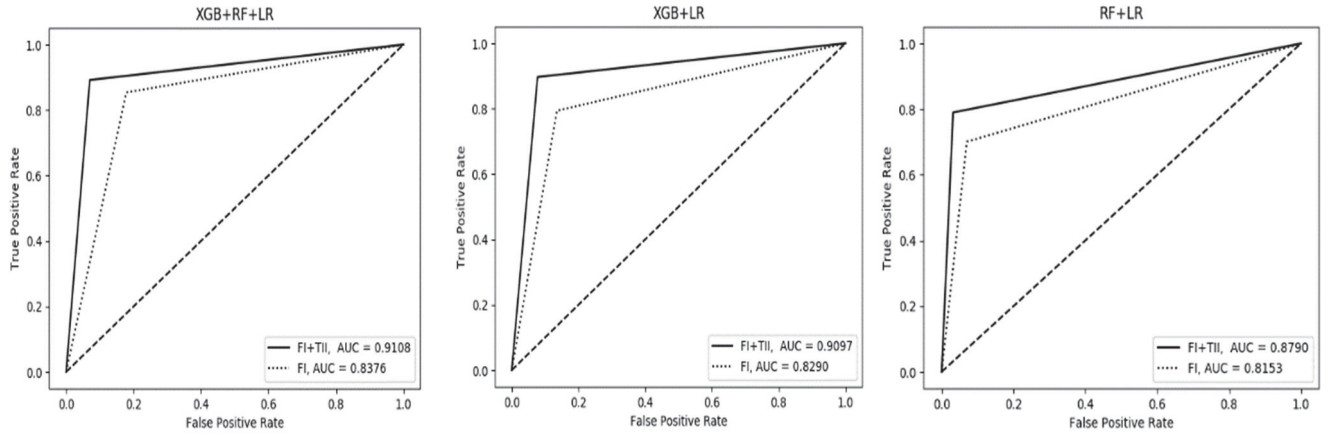
The combined ensemble learning variable screening method proposed in this study comprehensively considers the importance ranking results of features from Random Forest and XGBoost, the ten indicators of proportion of R&D expenses in operating income, invention patent, proportion of R&D personnel, asset liability ratio, intellectual property, quick ratio, net profit growth rate, utility model patent, current ratio and ROE, which were relatively high in both rankings, were retained.

3.4 Comparison of Different Models and Indicator Systems

After selecting the important independent variables, the LR classifier was implemented. For the performance evaluation of the proposed hybrid model, a hold-out approach was used to verify the validity of the proposed model, where 60% of the collected data was used for training and the remaining 40% for testing. LR was used to construct our default prediction model, which contains all the independent variables filtered by the integrated algorithm. Tab. 4 and Fig. 5 show the difference in AUC results between these different metrics and models, with the best AUC result being the proposed hybrid XGBoost-RF-LR model based on technological innovation indicators (*TII*) and financial indicators (*FI*) (0.9129). The RF-LR, XGBoost-LR, and XGBoost-RF-LR models using only financial indicators were evaluated at 0.8153, 0.8290, and 0.8387, respectively, while the AUC results of the single integrated models RF-LR and XGBoost-LR using technological innovation indicators and financial indicators were 0.8790, 0.9097, respectively and 0.9097, it can be seen that the addition of text mining technology innovation information and hybrid integrated learning screening methods have improved the accuracy of the risk assessment model, and the hybrid credit risk assessment system proposed in this study is superior to the traditional single financial indicator assessment system and single integrated learning indicator screening methods.

Table 4 AUC performance of different hybridmodels

Title 1	*FI*	*FI* + *TII*
RF-LR	0.8153	0.8790
XGBoost-LR	0.8290	0.9097
XGBoost-RF-LR	0.8387	0.9129



(a) (b) (c)

Figure 5 AUC performance of different hybrid models

3.5 Technology Innovation Enterprise Credit Risk Scorecard

Transforming the logistic regression model into a risk scorecard can evaluate the probability of enterprise default. The objective is to establish a scoring rule that regulators or investors can use to help make the best risk management decisions, firstly, the sample of XGBoost-RF screened indicators are equifrequency binned and weight of evidence (WOE) [23] coded to merge similar terms, followed by LR model for training to obtain the parameters.

Based on the score transformation method described in Section 2.3, the equations for solving constants A and B can be obtained as follows.

$$\begin{cases} P_0 = A - B \cdot \log(odds) \\ P_0 + PDO = A - B \cdot \log(2odds) \end{cases} \tag{16}$$

Assuming that the expected score $P_0 = 600$, when the odds is equal to {1:60} (default: normal) and $PDO = 20$ (different P_0 and PDO can be set as references in different studies [30]). Substituting the parameters into Eq. (16), $A = 481.86$ and $B = 28.85$ can be calculated. Substitute the B, the WOE corresponding to the variable, and the parameters of the trained LR model into Eq. (11),the variable score of the corresponding variable can be calculated, then we calculate the base score ,the basic score is not affected by the characteristics in the scorecard,base score = $A - B\alpha = 481.86 - 28.85 \times 0.13 = 478.11$. When using the credit risk scorecard, each variable of each technological innovation enterprise can only correspond to one score. After calculating the score of all variables, the final credit risk score can be obtained by increasing or decreasing the variable score on the basis of the basic score. The higher the final score, the lower the credit risk of technological innovation enterprises. Therefore, the established credit risk scorecard for technological innovation enterprises is shown in Tab. 5.

Table 5 Technology innovation enterprise credit risk scorecard

Variable Name	Variable Possible Values	Score
Base score		478.11
Proportion of R&D personnel	$X_{PRP} < 12.57$	−44.89
	$12.57 < X_{PRP} < -30.3$	4.83
	$X_{PRP} > 30.3$	31.78
Proportion of R&D expenses in operating income	$X_{PREO} < 4.69$	−20.28
	$4.69 < X_{PREO} < 7.05$	−2.44
	$7.05 < X_{PREO} < 12.26$	3.67
	$X_{PREO} > 12.26$	18.74
Invention patent	$X_{INVP} < 2$	−89.75
	$2 < X_{INVP} < 18$	−4.83
	$X_{INVP} > 18$	16.49
Utility model patent	$X_{UMP} = 0$	−28.91
	$1 < X_{UMP} < 19$	0.85
	$20 < X_{UMP} < 75$	5.10
	$X_{UMP} > 75$	11.39
Intellectual property	$X_{INPT} < 18$	−29.75
	$18 < X_{INPT} < 175$	5.75
	$X_{INPT} > 175$	8.67
ROE	$X_{ROE} < -18.7$	−76.51
	$-18.7 < X_{ROE} < 3.32$	−28.64
	$X_{ROE} > 3.32$	44.45
Current ratio	$X_{CR} < 1.06$	−16.82
	$1.06 < X_{CR} < 2.14$	−1.04
	$2.14 < X_{CR} < 4.06$	3.23
	$X_{CR} > 4.06$	8.47
Quick ratio	$X_{QR} < 0.82$	21.82
	$0.82 < X_{QR} < 1.8$	4.35
	$1.8 < X_{QR} < 3.56$	−5.81
	$X_{QR} > 3.56$	−18.31
Asset liability ratio	$X_{ALR} < 21.15$	49.48
	$21.15 < X_{ALR} < 38.38$	20.07
	$38.38 < X_{ALR} < 71.55$	−10.15
	$X_{ALR} > 71.55$	−89.33
Net profit growth rate	$X_{NPGR} < -61.35$	−18.86
	$-61.35 < X_{NPGR} < 11.55$	9.44
	$X_{NPGR} > 11.55$	10.27

4 DISCUSSION

Based on the above experiments, the results and implications of the analysis are as follows and are presented below.

(1) Based on Tab. 4 and Fig. 5, the experimental results provide evidence that the technological innovation variable system constructed through text mining has shown to be effective for credit risk assessment of technological innovation enterprises in this study, broadening the research area of the traditional literature on methods for constructing credit risk assessment variable systems for technological innovation enterprises.

(2) This study also shows that a model that includes both financial and technological innovation variables provide better classification results than a model that uses only financial variables. The inclusion of 'forward-looking' technological innovation variables helped to improve the classification performance of the predictive models more than the traditional 'backward-looking' financial variables only. Overall, this suggests that technological innovation variables can compensate for the shortcomings of traditional financial variables in the credit risk assessment of technological innovation enterprises.

(3) There are many variables that can be considered in the assessment of enterprise credit risk; therefore, it is critical to find the most important variables as it will affect the accuracy of the model developed. Instead of selecting variables based on domain knowledge, we automatically selected variables based on the importance calculated by two different types of ensemble learning methods, XGBoost and RF. As evidenced in Tab. 4 and Fig. 5, the RF and XGBoost methods were effective in improving accuracy regardless of the metric system used, and the best predictive performance was achieved by the metric model screened by the combined RF and XGBoost methods. The above analysis shows that the hybrid approach proposed in this study can help select variables to improve the performance of the model in evaluating enterprises credit risk without requiring special domain knowledge.

(4) In the further analysis results in Fig. 3 and Fig. 4, there are some differences between RF-based and XGBoost-based variable importance ranking. The top four variables screened by RF in terms of importance are invention patent, net asset growth rate, current ratio, asset liability ratio, with obvious financial variables accounting for a larger proportion, while the top four variables screened by XGBoost in terms of importance are proportion of R&D expenses in operating income, proportion of R&D personnel, utility model patentand invention patent, all of which are technological innovation variables, showing the difference in the effectiveness of the two differentensemble learning methods in screening indicators. In addition, from the combination of the two algorithms, invention patent, asset liability ratio, proportion of R&D expenses in operating income, intellectual property and net profit growth rate have considerable influence on the risk of technology innovation enterprises in both algorithms.

(5) Based on the above screened predictor variables, scoring the credit risk of technology innovation enterprises through risk scorecard based on LR is an understandable

and meaningful set of scoring rules that can be easily applied in a credit risk scoring system. The proposed hybrid model can provide a new and intelligent way to evaluate the credit risk of technology innovation enterprises.

5 CONCLUSION AND FUTRUE WORK

Enterprise credit risk assessment forecasting plays an important role in financial regulation and investment decisions. Among the current forecasting models, many studies have focused on financial information based on historical information rather than assessment of the future. To establish a set of credit risk evaluation system that combines objectivity, accuracy and transparency without additional expertise compared with the traditional evaluation methods for the characteristics of technology innovation enterprises, in this study, we use information based on technological innovation as a predictive variable and propose a hybrid model which combines RF, XGBoost and LR to improve the accuracy of credit risk assessment. In the proposed approach, text mining techniques are used as tools to extract technology innovation-based information for each company. To prove the proposal, we also used XGBoost + LR and RF + LR from the previous literature as benchmarks and applied them to technological innovation companies in China. Based on the experimental results, the findings of this study can be summarized as follows.

First, information based on technological innovation is a useful ex ante determinant for credit risk assessment. The new predictor variable, i.e. technology innovation-based information, improves the classification effect of credit risk assessment for technology innovation firms. We do this by highlighting the link between technology innovation-based information and credit risk assessment. In particular, we have shown that information on technological innovation is an important ex ante indicator of credit risk for technology firms. Second, the experimental results show that there are differences between the two different integrated learning methods of RF and XGBoost for screening indicators, and that the hybrid method that combines both integrated learning methods for indicator screening outperforms previous methods in terms of higher accuracy and fewer variables. Finally, the understandable scoring rules derived from the new hybrid approach make it easier to read the categorized credit risk scores and determine which high-tech companies are worth investing in or not, and the results of this work are useful and feasible for decision makers in investment institutions and regulators.In summary, the mainfindings and contributions of this paper are as follows.

(1) Theory: From the current literature, the research on credit risk assessment mainly focuses on traditional manufacturing and financial indicators, and the research on innovation information of technological innovation enterprises is less. This study has established a credit risk assessment system for technological innovation enterprises based on the combination of technological innovation information and traditional financial information. Through the comparative analysis of the combined indicator system model, the new credit risk assessment system with technological innovation information has the advantage of identifying the credit risk of technological innovation enterprises. It can be concluded that technological innovation information is an effective supplement to the credit risk identification of technological innovation enterprises compared with traditional financial information, this broadens the research field of credit risk of technological innovation enterprises.

(2) Methods: This paper proposes a new mixed evaluation method of credit risk based on text mining, integrated learning and logical regression. In this method, the technical innovation index system is built based on text mining technology, the index screening is based on integrated learning model, and the credit risk scoring is based on logistic model to form a complete credit risk evaluation system. The technical innovation indicator system proposed in this paper is based on the technical innovation information mined from the audit inquiry letter of listed companies through text mining technology, which is more scientific and objective than the subjective technical innovation indicator system established in previous studies, In addition, the ensemble learning model for indicator screening combines the XGBoost algorithm based on Gradient Boosting ensemble and the random forest algorithm based on Bagging ensemble, which are two different principles of the ensemble algorithm, and then constructs a method for screening indicators through the importance of variable characteristics, so that the evaluation system is more perfect than the previous single integrated learning indicator screening method.

(3) Practicality: This paper proposes a credit risk assessment method with strong applicability. Compared with the traditional logistic method, this method improves the prediction accuracy of the model by adding technical innovation information and embedding the optimized ensemble learning index screening method. At the same time, the evaluation principle of this method is more transparent than the single machine learning methods such as support vector machine, neural network and integrated learning. The specific source of risk can be analyzed through credit risk scoring rules of risk scorecard, the easy-to-understand credit risk scoring rules are also more practical for listed company regulators, banks and other financing institutions as well as investors. At the same time, this method also has certain guiding significance for technological innovation enterprises to optimize their internal management.

In practical applications, it is recommended that regulatory authorities can use this risk assessment framework to assess the risks of listed companies on the Technology Innovation Board for targeted auxiliary supervision. Banks can use this risk assessment method to determine whether to lend to enterprises. Enterprise managers can use this method to understand the risk situation of enterprises and assist in management decision-making. Investors can use this method to understand enterprise risks and assist in investment decision-making. However, there are several issues that warrant further research. First, we use information based on technological innovation as a predictor variable for credit risk assessment predictions. Future research on listed companies could include market information on listed companies as part of the credit risk assessment. Secondly, we can extend to other rule-based classifiers to test the validity of technological innovation information. Finally, we should

continue to compare our proposed approach with the most recently researched classification models.

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] He, X. & Gong, P. (2008). Research on internal credit ratings for listed companies, *Kybernetes*, *37*, 1339-1348. https://doi.org/10.1108/03684920810907634

[2] Yang, M., Shifeng, L., & Daqing, G. (2023). A text mining and ensemble learning based approach for credit risk prediction. *Technical Gazette*, *30*(1), 138-147. https://doi.org/10.17559/TV-20220623113041

[3] Jun-won L. (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

[4] 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

[5] Dongyang, Z., Wenping, Z., & Lutao, N. (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

[6] Timothy, R. W., Daniel, C., & Anil, R. (2018). Varieties of innovation and business survival: Does pursuit of incremental orfar-ranging innovation make manufacturing establishments more resilient? *Research Policy*, *47*, 1801-1810. https://doi.org/10.1016/j.respol.2018.06.011

[7] Yang-Cheng, L., Chung-Hua, S., & Yu-Chen, W. (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

[8] Chang, Y., Cuiqing, J., Hemant, K. J., & Zhaoy, W. (2020). Evaluating the credit risk of SMEs using legal judgments. *Decision Support Systems*, *136*, 113364. https://doi.org/10.1016/j.dss.2020.113364

[9] Mark, C., Haldun, A., Gary, J. K., & Praveen, 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

[10] 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

[11] Fernandes, G. B. & Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European journal of operational research*, *249*, 517-524 https://doi.org/10.1016/j.ejor.2015.07.013

[12] You, Z., Chi, X., Gang-Jin, W., & Xin-Guo, Y. (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

[13] Xiaobing, H., Xiaolian, L., & Yuanqian, R. (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

[14] Parisa, G., Ionuţ, F., & Rupak, C. (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

[15] Harrell, F. E. & Lee, K. L. (1985). *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*. NewYork: North-Holland.

[16] Pederzoli, C., Thoma, G., & Torricelli. C. (2013). Modelling credit risk for innovative firms: The role of innovation measures. *Journal of Financial Services Research*, *44*, 111-129. https://doi.org/10.1007/s10693-012-0152-0

[17] Ching-Chiang, Y., Fengyi, L., & Chih-Yu, H. (2012). A hybrid KMV model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, *33*, 166-172. https://doi.org/10.1016/j.knosys.2012.04.004

[18] Chunjiao, D., Yixian, Q., Chunheng, S. Xiwen, L., Xiaoning, Y., Qin, C., Yuxuan, L., Jianan, Z., Yunfeng, W., Yahong, C., Qinggang, G., & Yurong, B.(2022). Non-contact screening system based for COVID-19 on XGBoost and logistic regression. *Computers in Biology and Medicine*, *141*, 105003. https://doi.org/10.1016/j.compbiomed.2021.105003

[19] Mikolov, T., Sutskever, I., Chen, K. et al. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111-3119.

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

[21] 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

[22] Lingxiao, F. & Yong, D. Q. (2022). Research on risk scorecard of sick building syndrome based on machine learning. *Building and Environment*, *211*, 108710. https://doi.org/10.1016/j.buildenv.2021.108710

[23] Luo, Z., Hsu, P., & Xu, N. (2020). SME default prediction framework with the effective use of external public credit data. *Sustainability*, *12*, 7575. https://doi.org/10.3390/su12187575

[24] Anahita, N., Mohammad, S., Fethi, R., & Mohsen, N. (2018). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*, *11*, 925-935. https://doi.org/10.2991/ijcis.11.1.70

[25] 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

[26] Qi, L. & Mu, Z. (2022). Research on credit risk assessment of listed companies in science and technology sector by introducing industry research report information. *Procedia Computer Science*, *214*, 1317-1324. https://doi.org/10.1016/j.procs.2022.11.311

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