International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 4213-4226

https://doi.org/10.70517/ijhsa463360

Research on Construction and Application of Enterprise Financial Risk Prediction Model Based on XGBoost Algorithm

Ke Sun[1] and Yupeng Li[2,*]
[1] School of Accountancy, Guangzhou College of Technology and Business, Guangzhou, Guangdong, 510850, China
[2] School of Accountancy, Anyang Institute of Technology, Anyang, Henan, 455000, China
Corresponding authors: (e-mail: kelly_sun2024@163.com).

Abstract The risk problems of listed companies will not only bring great losses to the enterprise, how to find out the risk information from the many data of the enterprise and take risk prevention in advance is crucial for the healthy development of the enterprise. Combining relevant research results with network crawling technology to obtain research data, and pre-processing the disturbing information in the data. After setting the enterprise financial risk prediction index system and model evaluation indexes, the KPCA algorithm is used to downsize the prediction index system, followed by the construction of the enterprise financial risk prediction model based on the XGBoost algorithm, and the application of the model is analyzed. The prediction accuracy of this paper's model for ST enterprise financial risk in 2022 is more than 0.9, indicating that the prediction model constructed in this paper has excellent application value and promotes the development of intelligent detection of enterprise financial risk.

Index Terms XGBoost algorithm, KPCA algorithm, prediction model, enterprise financial risk

I. Introduction

In the new period, the development environment faced by Chinese enterprises is becoming more and more complex, and the challenges faced by enterprises are becoming more and more severe [1]. Strengthening the enterprise's financial risk prediction is an important foundation for guaranteeing the healthy development of enterprises in the new period, through the financial risk prediction can reduce the incidence of enterprise financial risk, so that the enterprise can adapt to the environment to enhance the ability to improve the internal control level of the enterprise [2]-[4]. This can be in the business expansion at the same time to avoid the enterprise due to the lack of internal control level and encounter the impact of risk, and then to protect the healthy and stable development of the enterprise [5]. However, in the traditional enterprise management, there is no scientific and effective management method to improve the level of internal control, which can not help enterprises to better avoid risks [6], [7]. In the big data environment, massive data and information can provide an effective basis for the prediction and assessment of enterprise risk, realize the evidence-based enterprise risk management, and further improve the quality of enterprise risk management [8], [9]. Especially after the application of relevant data analysis technology for prediction, the problems existing in the enterprise financial data can be accurately excavated, to avoid some potential risks are not found in time and cause serious economic losses to the enterprise [10]-[12].

For financial risk prediction has been a lot of workers have made a study, usually the financial statements of listed companies as the object of study. Literature [13] studied the application of artificial intelligence technology in enterprise financial statement heritage prediction, through the introduction of deep learning algorithms to build financial data analysis model, can effectively analyze the historical financial data, and predict the future financial risk, for enterprise managers to provide good risk management and investment decision-making suggestions. Literature [14] combines the evidence theory-random forest (DS-RF) model to construct the enterprise financial risk early warning framework, based on the credibility of the evidence to integrate and analyze the enterprise financial data information, so as to improve the reliability of the enterprise financial risk early warning. Literature [15] proposes a financial risk prediction method that contains a fusion deep learning model, which shows a better risk prediction effect in the face of nonlinear multivariate corporate financial data with complex structure. Literature [16] used the evaluation model based on the support vector machine model to analyze the enterprise financial data, by extracting the financial features in the financial historical data and combining the relevant optimization algorithms, so as to accurately assess the enterprise financial risk. Literature [17] explored effective big data mining methods applicable to enterprise financial risk management, showing that the enterprise financial risk management model based on information fusion is more effective in financial risk management and classification, and at the same time provides accurate risk prediction results for enterprises. Literature [18] shows that traditional enterprise financial risk

management methods lack awareness of complex relationships between enterprises, for which a tribal style map oriented hierarchical graph neural network (TH-GNN) is proposed to fully consider information dissemination between enterprises and shareholders so as to achieve effective and efficient financial risk assessment. However, the above studies targeting financial statements are mostly based on the process of analyzing structured data, which is incomplete in terms of predicting results. Therefore, the research can use the panel data as the input condition of the machine learning model, which can reflect the real financial risk status of the enterprise in a more comprehensive and integrated way.

This paper prioritizes the identification of sample enterprises and time windows for this research, adopts network crawling technology to obtain the text data of enterprise annual reports, and its corresponding preprocessing. After completing the above preparatory work, the enterprise financial risk prediction index system and model evaluation index are constructed. In order to reduce the difficulty and time of the algorithm, the kernel principal component analysis algorithm is used to reduce the dimensionality of the prediction indicators, and on this level, the XGBoost algorithm is used to design the enterprise financial risk prediction model, and the prediction model designed in this paper is verified and analyzed in application by combining the corresponding research data and evaluation indexes, respectively.

II. Enterprise financial risk prediction model

II. A. Sample selection

II. A. 1) Criteria for selection of sample enterprises

At present, there are various opinions on the definition of “financial risk”. Foreign scholars define it mostly in terms of a company's inability to pay its debts on time, a company's suffering significant losses, or even a company's collapse. However, most of my definitions of this issue are based on the acceptance of disposals. Therefore, this paper, similarly, divides the sample into two categories based on disposal. In addition to this, the paper obtains a detailed list of bond defaults of real estate companies, and in addition to companies receiving disposals being categorized as financially risky firms, cases such as bond defaults and financial trusts failing to pay are also identified as financially risky firms.

II. A. 2) Time window selection

The purpose of this paper is to forecast the financial risk of a company in advance, and to help investors, consumers and other stakeholders of the relevant companies to know the outbreak of the company's financial risk in advance. Generally speaking, the closer the data year used for modeling is to the forecasting year, the better the predictive effect of the model is. However, since the previous year's financial report of listed companies is usually published in April or May, if the company's financial report in 2021 is known in April or May 2022, in order to know whether the company has a significant financial risk in 2022, it is too lack of timeliness. So it is not meaningful to use the relevant data of 2021 to predict the results of 2022. In addition, if the data of 2019 is used to construct the early warning model, its prediction effect is too poor. Therefore, under comprehensive consideration, this paper selects the data of 2020 to construct the enterprise financial crisis prediction model, and warns the company two years before, which can leave the company with enough time to adjust its business behavior. To sum up, this paper chooses the real estate listed enterprises from 2020 to 2023 as the research samples, and obtains 1,953 samples, of which 92 are financial crisis samples.

II. B. Data acquisition

II. B. 1) Quantitative data acquisition

The predictive indicators in this paper cover both financial and non-financial indicators. In practice, there are many reasons for the financial risk of listed companies in the manufacturing industry, and the selection of prediction indexes often varies due to the imperfect unified theoretical foundation and the different research focuses. Based on the principles of reliable accuracy, relevance, feasibility and comprehensiveness in the selection of financial risk prediction indexes, this paper selects the corresponding financial indexes based on the enterprise financial diagnosis theory in the previous article, starting from the causes of financial risk. In this paper, from the six aspects of solvency, risk level, profitability, cash flow, development ability, operating ability, 18 financial indicators that meet the theoretical basis and are used frequently are selected as the financial risk prediction indicators in this paper.

II. B. 2) Access to the text of the annual report

According to the theory of information asymmetry and signaling in the previous section, text-based information such as company announcements can reflect the company's production and operation and other situations, and only use financial indicators and other numerical indicators to analyze the financial risk of enterprises can not reflect the business management status of the enterprise in a timely and comprehensive manner, so this paper introduces the

text of the company's annual report on the basis of quantitative indicators. First of all, the annual reports of 92 companies were obtained from Juchao.com and Oriental Fortune.com using web crawlers, the same as the quantitative financial and non-financial data, and if a listed company is subject to special treatment by the stock exchange in 2022, the annual report of the listed company in 2019 was downloaded through web crawlers. The annual report of a company consists of a large number of chapters, such as basic company profile, discussion and analysis of operations and corporate governance, which makes the information of listed companies' annual reports redundant and long, and the predictive ability of the model may be reduced due to too much noise if the whole input is input.

II. C.Cleaning and processing of data

II. C. 1) Quantitative data processing

(1) Data pre-processing

Before screening the indicators, data preprocessing should be carried out. Firstly, the missing values in the data should be filled in, and this paper adopts the regression method, i.e., the feature T with missing values is treated as a label, and the other remaining features and the original label form a new feature matrix, and the missing values of the feature T are predicted by training the new feature matrix. Secondly, for the subtyped variables and discrete value indicators, the coding process and the binning process for the distribution of indicators are carried out respectively. Finally, due to the large number of quantitative financial indicators selected in this paper, in order to ensure the comparability of the data, but also in order to avoid the bias caused by different units, the indicators are normalized, so that all the data are dimensionless and are in the same order of magnitude, and the formula for the standard normalization treatment is as follows:

$$x_{normalization} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

(2) Indicator Screening

For the screening of the selected financial and non-financial quantitative indicators, this paper refers to the information and uses the significance test, a common indicator screening method in financial risk prediction, to carry out the corresponding test on whether each continuous indicator has a significant difference between listed manufacturing companies with financial risk and financial normality, so as to screen out the indicators that have a strong identification ability and contribute to the enhancement of the model's performance improvement. Specifically, the normality of the indicator variables should be tested, and all 18 evaluation indicators pass the test, and these 18 indicators are subsequently used as inputs to the prediction model.

II. C. 2) Annual report text processing

(1) Text pre-processing

As structured data (quantitative financial and non-financial table data), stop words, invalid numbers, English and so on often appear in the chapter of “Discussion and Analysis of Business Situation”, this paper does the following four aspects of processing on the text of annual reports of listed companies before word separation:

a) Delete punctuation marks, tabs, line breaks and other symbols that have no practical significance.

b) Delete auxiliary words, prepositions and other deactivated words.

c) Remove deactivated words that contain numbers (dates and amounts) to avoid neural networks overfitting number formats to specific formats

d) Constructing vocabulary lists and ignoring words with less than 10 occurrences.

(2) Text representation

Before feeding text into a deep learning model, it usually needs to be mapped to a completely new space and converted to a multi-dimensional real number vector. Existing methods for text representation mainly include statistical-based methods such as the one-hot method, and two methods based on language models. Due to the shortcomings of matrix sparsity, dimensional catastrophe, and semantic missing using one-hot coding, this paper utilizes the language-based model Word2Vec in order to obtain the word vector matrix of the text of the annual reports of listed companies. After processing, the text inputs of the 92 selected listed companies are turned into 300×300 word vector matrices, which are used as inputs for the subsequent financial risk prediction model.

II. D. Construction of Financial Risk Prediction Indicator System

II. D. 1) Indicator system

Through the processing of quantitative data and financial text mentioned above, 18 financial and non-financial indicators as well as the 300×300 word vector matrix of each listed company are obtained as the input of the model, and the enterprise financial risk evaluation index system is shown in Table 1.

Table 1: Enterprise financial risk evaluation index system

Theme	First index	Symbol	Secondary index	Symbol
Financial risk	Debt-paying ability	X1	Current ratio	X11
			Cash ratio	X12
			Interest coverage ratio	X13
	Business ability	X2	Accounts receivable turnover rate	X21
			Business cycle	X22
			Working capital turnover ratio	X23
	Profitability	X3	Return on assets	X31
			Net profit margin on total assets	X32
			Return on net assets	X33
	Cash flow	X4	Net cash content of operating income	X41
			Cash meets the investment ratio	X42
			Cash suitability ratio	X43
	Develop ability	X5	Capital preservation and appreciation rate	X51
			Capital accumulation rate	X52
			Growth rate of total assets	X53
	Governance Structure	X6	Board size	X61
			The number of senior executives	X62
			Management shareholding ratio	X63

II. D. 2) Model evaluation indicators

The model evaluation metrics in this section are constructed using the constructed model by predicting the sample companies and comparing the final prediction results with the real labels of these sample companies. Some of the most common evaluation metrics are Accuracy, Precision, Recall, and F1-score, which will be subsequently combined with the model evaluation metrics to predict the financial risks of the companies.

II. E. Predictive modeling

II. E. 1) KPCA-based dimensionality reduction of predictors

In the process of forecasting the risk of financial crisis in SMEs, because there are some correlations and overlaps between each financial index variable, and the dimensionality of the financial data is relatively high, which will lead to a significant increase in the calculation of the subsequent prediction model, which in turn affects the efficiency and accuracy of the prediction. In order to solve this problem, it is particularly important to introduce the principal component analysis (KPCA) method to reduce the dimensionality of the data [19]. Specifically, the dimensionality reduction process of KPCA can be divided into the following steps: first, a suitable kernel function (e.g., Gaussian kernel function or polynomial kernel function) is selected, based on which, each sample point of the original dataset is corresponded to a higher dimensional feature space, so as to obtain the covariance matrix of a high-dimensional sample:

$$j(X)j(X)^T \omega_j = l_j \omega_j, j = 1,2,\cdots,L,d \tag{2}$$

In Eq. $j(X)$ denotes the vector of high-dimensional mapping samples of dimension $D*l(D\square\ d)$. l_j denotes the eigenvalues. ω_j denotes the vector of dimensions corresponding to the eigenvalues. Next, the kernel matrix, which is the inner product representation of each sample point in the new space, is computed. Then, the kernel matrix is centered to remove the mean shift in the data. Then, the eigenvectors and eigenvalues are obtained by eigenvalue decomposition. On this basis, the largest set of each eigenvector is used as the eigenspace after dimensionality reduction. Finally, on this basis, the original sampling points are mapped to obtain the dimensionality reduced data:

$$x^{new} = \omega_i^T j(x) \tag{3}$$

In the formula, x^{new} denotes the data after dimensionality reduction. Through the dimensionality reduction process of KPCA, the redundant information and noise in financial data can be effectively removed and the most representative features can be retained. At the same time, the dimensionality reduced data is easier to understand and visualize, which helps to better analyze and explain the patterns and laws in the financial data.

II. E. 2) XGBoost Algorithm

XGBoost model is a kind of integrated learning, integrated learning by training multiple weak learners to solve the same problem, and combining multiple weak learners to achieve a better integrated learner [20], [21].Bagging and Boosting are both integrated learning methods that combine learners to generate a strong learner, which makes the newly generated learner have better learning effect [22], [23].The XGBoost model is also an algorithm based on Boosting, which uses numerical optimization more effectively than GBDT by making the loss function, i.e., the error between the predicted value and the true value, more complex by adding a regularization to the loss function. The objective function is still the sum of the predicted values of all trees equal to the final predicted value, and its objective function is as in equation (4). That is:

$$Obj = \sum_{i}^{n} L(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{4}$$

The first of these terms, Training Loss, also known as Empirical Risk, is used to describe how well the classification model fits the sample data and measures the training error for n samples. The second term is Regularization, the regularization term, also known as structural risk, is used to indicate the complexity of the model. We want to minimize the objective function so that the model fits the data better, minimizes the bias, and yet keeps the complexity of the model down. The objective function of XGBoost is to be optimized in three steps:

Step 1: The loss function is subjected to a second-order Taylor expansion to remove the constant term and optimize it.

Step 2: The regular term is expanded to remove the constant term.

Step 3: Combine the expanded primary and secondary terms to get the final optimized objective function.

Since XGBoost adopts a forward distribution algorithm, the further t machine learner can be unfolded to get the relationship with the previous round of iterations, which is organized into equation (5). Namely:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

Therefore, in the t round, what we need to determine is the function $f_t(x)$, so that after organizing it, we can get the objective function of the t round as in equation (6). Namely:

$$\begin{aligned} Obj^{(t)} &= \sum_{i}^{n} L(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} L(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + cons\tan t \end{aligned} \tag{6}$$

In round t of the GBDT algorithm, we use the negative gradient corresponding to the objective function, i.e., the loss function, to the predictor function as the sample labels to fit the decision tree for this round. While in XGBoost, we use the first-order bias, i.e., the gradient and the second-order bias to fit the decision tree in this round. Since the second order partial derivative has better approximation than the first order partial derivative, it can be utilized to the second order Taylor's formula as shown in equation (7). Namely:

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \tag{7}$$

In the above equation x corresponds to $\hat{y}^{(t-1)}$ and Δx corresponds to $f_t(x_i)$, so the objective function for the t th round can be organized as Eq. (8). Namely:

$$Obj^{(t)} = \sum_{i=1}^{n}\left[L(y_i, \hat{y}_i^{t-1} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + constant \tag{8}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} L\left(y_i, \hat{y}^{(t-1)}\right)$, $h_i = \partial^2_{\hat{y}^{(t-1)}} L\left(y_i, \hat{y}^{(t-1)}\right)$. g_i represents the first-order partial derivative of the error term, and h_i represents the second-order partial derivative of the error term. This item can be ignored since Eq. $L\left(y_i, \hat{y}^{(t-1)}\right)$ is generated for the first $t-1$ round of iterations does not affect the magnitude of the objective function of the first t round. After removing irrelevant terms and retaining only the terms related to the objective function of the first t round, the objective function of the first t round is transformed into equation (9). i.e:

$$Obj = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{9}$$

In a decision tree, $f_t(x_i)$ is a simple mapping function that passes samples into the mapping function and out the predicted classification result. The specific expression of $f_t(x_i)$ is shown in equation (10). Namely:

$$f_t(x) = \omega_{q(x)}, \omega \in \square^T, q: \square^d \to \{1, 2, \cdots, T\} \tag{10}$$

In the above equation, ω is the estimation value corresponding to each leaf node, and q is a simple relation which maps each sample to the corresponding leaf node. For the second canonical term of the objective function, instead of using the conventional L1 canonical term and L2 canonical term, a new model complexity is constructed for the special structure of the decision tree in the form of Eq. (11). That is:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{11}$$

where T represents the number of leaf nodes and ω_j represents the predicted value corresponding to each leaf node. When the number of leaf nodes is more, the tree is deeper, and it is easier to produce overfitting problems. When ω_j is larger, i.e., the larger the predicted value of the leaf node is, then the larger the proportion of values in all the regression trees is, and the whole prediction is easy to be dominated by this tree. Thus the first term of the regular term penalizes the number of leaf nodes and the second term penalizes the node values. At this point, the objective function is composed of samples as a unit, and further transforming the objective function to be composed of individual leaf nodes, we can obtain Eqs. (12) and (13). Namely:

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_T(x_i) + \frac{1}{2} h_i f_T^2(x_i) \right] + \Omega(f_t) \tag{12}$$

$$Obj^{(t)} = \sum_{i=1}^{n}\left[g_i \omega_q(x_i) + \frac{1}{2} h_i \omega_q^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{13}$$

The above equation is actually the sum of T quadratic functions associated with ω. It is further organized as Eq. (14) as well as Eq. (15). To wit:

$$Obj^{(t)} = \sum_{j=1}^{T}\left[\sum_{i \in I_j} g_i \cdot \omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \tag{14}$$

$$Obj^{(t)} = \sum_{j=1}^{T}\left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda \right) \omega_j^2 \right] + \gamma T \tag{15}$$

where $G_j = \sum_{i \in I_j} g_i$ is the sum of first-order derivatives. $H_j = \sum_{i \in I_j} h_i$ is the sum of second order derivatives. According to Eq. (16) as well as Eq. (17), H_j is the second-order derivative of the loss function, which is a convex

function, and thus the second-order derivative of the loss function is greater than 0. ω_j is a quadratic function with an opening upward that can be taken as a minimum. To wit:

$$\arg\min_x Gx + \frac{1}{2}Hx^2 = -\frac{G}{H}, H > 0 \tag{16}$$

$$\min_x Gx + \frac{1}{2}Hx^2 = -\frac{G^2}{H} \tag{17}$$

Therefore, when the structure of the tree is given, the predicted values corresponding to each leaf node can be obtained from Eq. (18) as well as Eq. (19) based on the above solution process. Namely:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{18}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{19}$$

From this, when given which leaf node each sample belongs to, the optimal predicted value of the leaf node and the minimum value of the objective function can be found.

II. E. 3) General Model Architecture

In constructing a financial crisis prediction model, the XGBoost algorithm shows its strong prediction ability and efficiency. In the XGBoost algorithm, a regularized loss function is introduced, multiple weak classifiers are constructed, and then each weak classifier is fused together to form a high-precision strong classifier. The loss function can be expressed as:

$$L(y, F(x)) = |y - F(x)| \tag{20}$$

In the formula, $L(y, F(x))$ denotes the loss function. $F(x)$ denotes the approximation function. (x) denotes the training sample input vector. (y) denotes the training sample output variable. I.e:

$$\hat{y} = \sum_{i=1}^{K} f_k(x_i), f_k \in F \tag{21}$$

In the formula, $\dot{y}$ denotes the prediction model. k denotes the number of trees. f denotes a specific CART tree. $f_k(x_i)$ denotes the predicted value of the sample x_i on the k th tree. The model is composed of k CART trees.

Using the above prediction model, the following steps were followed to realize the prediction of the risk of financial crisis of SMEs:

Step 1: Input financial data variables containing 36 dimensions that comprehensively reflect the business conditions of SMEs.

Step 2: The method divides the whole sample into two parts: a testing part and a training part.

Step 3: The XGBoost model constructs new feature vectors by learning the structure of the tree. The new feature vectors are either 0 or 1, and they all correspond one-to-one to the leaf nodes of the tree in the model. In this process, the sampling point passes through a tree and reaches a leaf node, then the element of the eigenvector corresponding to this leaf node becomes 1, while the other leaf nodes become 0.

Step 4: Finally, a new eigenvector is obtained whose length is equal to the sum of the leaf nodes in each tree in the XGBoost model. Then, the Logistic regression model is trained using the new features. On this basis, this project proposes a new fusion algorithm and realizes the prediction of this algorithm through the application of this algorithm in practical applications.

III. Empirical analysis of corporate financial risk prediction models

III. A. Model validation analysis

III. A. 1) Model hyperparameter settings

The configuration of hyperparameters in a model can have an impact on its prediction accuracy and generalization performance. Therefore, we should adjust these parameters according to the characteristics of various models to determine the best combination of parameters for optimal prediction. For the logistic regression model, the parameters set in this paper are the regularization parameter, the loss function optimizer, and the inverse of the regularization strength. In the decision tree model, the hyperparameters set in this paper include the maximum number of leaf nodes and the maximum depth of the tree. In the study of XGBoost, the following hyperparameters are set in this paper: maximum depth, minimum weight of samples in child nodes, feature sampling ratio, percentage of sub-samples in the whole sample set, and minimum loss reduction required when performing segmentation. The alternative parameters for each model are shown in Table 2. For logistic regression and decision tree models, this paper uses grid search to regulate all parameters simultaneously. For XGBoost, on the other hand, in order to improve the speed of model tuning, the first three parameters are adjusted first to determine the optimal values, and then the last two parameters are adjusted to determine the optimal parameters of the model in turn.

Table 2: Alternative parameters for each model

Model	Name of hyperparameter	Hyperparameter	Hyperparameter Settings
Logistic regression	Regularization parameter	C	[0.001,0.01,0.1,1,10,100,1000]
	Loss function optimizer	Penalty	['l1','l2']
	The reciprocal of the regularization intensity	Solver	['liblinear','saga']
Decision tree	Maximum depth	max_depth	[1,2,3,4,5,6]
	The maximum number of leaf nodes	max_leaf_nodes	[3,4,5,6,7,8]
	Maximum depth	max_depth	[i for i in range(3,6,1)]
XGBoost	The sum of the minimum sample weights in the child nodes	min_child_weight	[i for i in range(1,6,2)]
	The proportion of feature sampling	Colsample_bytree	[i/100 for i in range(80,100,5)]
	The proportion of sub-samples in the entire sample set	Subsample	[i/10 for i in range(6,10)]
	The minimum reduction in loss required for segmentation	Gamma	[i/10 for i in range(0,5)]

III. A. 2) Data analysis

This section begins with a fit using 2022 data. Prior to constructing the model, the data were standardized. Standardized processing refers to transforming the indicators into values between [0,1] to eliminate the influence of unit and scale on the fitting results. Then the four evaluation indicators were fitted to three models, logistic regression, decision tree and XGBoost, respectively, for a total of 18 sets of models. Because of the large randomness of dividing the test set and training set according to a certain ratio, which is commonly used in modeling, this paper uses the method of five-fold cross-validation to compare the accuracy of each model. Table 3 gives the summed confusion matrices of the above 18 groups of models for the five-fold cross-validation, as well as the corresponding correctness, accuracy, recall, and F1 scores. As a whole, the fit of the XGBoost model for the combination of indicators for all the metrics performs optimally, with all values being optimal, much better than the other two groups of control algorithms (logistic regression, decision tree).

Table 3: Fitting performance of 2022 year data

Indicator category	Model	No.	Accuracy	Precision	Recall	F1-score
Financial data	Logistic regression	1	0.7815	0.7935	0.7725	0.7831
	Decision tree	2	0.7815	0.7945	0.7725	0.7835
	XGBoost	3	0.7845	0.8045	0.7744	0.7915
Financial data+Text data	Logistic regression	4	0.7715	0.8005	0.7544	0.7745
	Decision tree	5	0.7715	0.805	0.7544	0.7745
	XGBoost	6	0.7942	0.8405	0.7721	0.8033
All indicators	Logistic regression	7	0.9234	0.9403	0.9115	0.9605
	Decision tree	8	0.9234	0.9403	0.9112	0.9605
	XGBoost	9	0.9544	0.9845	0.9234	0.9722

The same method was used to fit the 2021 and 2020 data, respectively. The summed confusion matrices and the corresponding correctness, accuracy, recall, and F1 scores for the five-fold cross-validation of the 18 sets of models constructed for the 2021 and 2020 data are given in Tables 4 and 5, respectively. Whether it is the 2021 corporate

financial risk data or the 2022 corporate financial risk data, the values of the four predictive performance indexes of the XGBoost algorithm are 0.9545 (0.9722), 0.9922 (0.9927), 0.9245 (0.9435), and 0.9732 (0.9728), which are much better than those of the logistic regression and decision tree, indicating that the XGBoost algorithm can be used to predict corporate financial risk in 2021 and 2020, indicating the superiority of XGBoost algorithm in corporate financial risk prediction.

Table 4: Fitting performance of 2021 year data

Indicator category	Model	No.	Accuracy	Precision	Recall	F1-score
Financial data	Logistic regression	1	0.7545	0.7835	0.7425	0.7631
	Decision tree	2	0.7545	0.7841	0.7425	0.7625
	XGBoost	3	0.7623	0.7845	0.7433	0.7635
Financial data+Text data	Logistic regression	4	0.7715	0.8111	0.7532	0.7815
	Decision tree	5	0.7715	0.8111	0.7532	0.7815
	XGBoost	6	0.7715	0.8111	0.7532	0.7815
All indicators	Logistic regression	7	0.9525	0.9705	0.9225	0.9535
	Decision tree	8	0.9533	0.9705	0.9233	0.9541
	XGBoost	9	0.9545	0.9922	0.9245	0.9732

Table 5: Fitting performance of 2020 year data

Indicator category	Model	No.	Accuracy	Precision	Recall	F1-score
Financial data	Logistic regression	1	0.7635	0.7711	0.7633	0.7641
	Decision tree	2	0.7642	0.7715	0.7633	0.7645
	XGBoost	3	0.7905	0.8144	0.7729	0.7935
Financial data+Text data	Logistic regression	4	0.7415	0.8144	0.7121	0.7604
	Decision tree	5	0.7415	0.8144	0.7121	0.7604
	XGBoost	6	0.7405	0.8167	0.7531	0.7745
All indicators	Logistic regression	7	0.9615	0.9523	0.9325	0.9631
	Decision tree	8	0.9618	0.9515	0.9331	0.9617
	XGBoost	9	0.9722	0.9927	0.9435	0.9728

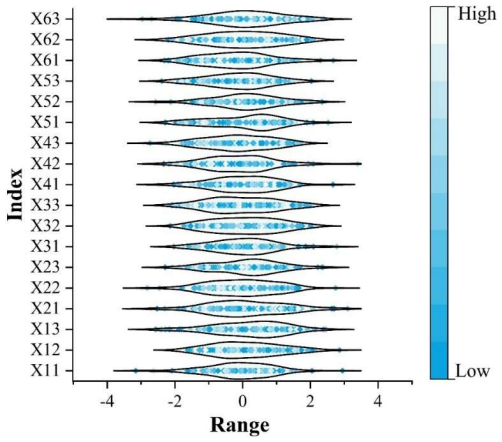

Figure 1: The Shap influence of Model features

Although the XGBoost algorithm performs well in prediction, like many machine learning techniques, it faces the challenge of insufficient interpretation, which is like a “black box” that cannot accurately assess the actual effect of each indicator. Therefore, in this paper, we use the Shap model to calculate the Shap value of each financial risk impact factor to improve the explanatory power of the model, Shap is an a posteriori inference tool for parsing more complex machine learning models. In most cases, the machine learning model is regarded as an unknown box, and the expected results can be obtained directly after the model is trained by simply inputting the financial risk predictors in the front end of the model. Figure 1 shows us the Shap impact of each feature by sorting the features by the sum

of the Shap values of all the samples, with each row representing a feature and each dot representing a sample. Taking X11 as an example, a high Shap value of X11 has a negative impact on the prediction and a low Shap value of X11 has a positive impact on the prediction. Figure 2 shows the mean of the absolute value of the Shap value of each indicator as a reflection of the importance of each indicator. Based on the data in the table, we can find that X21 has the most important contribution to predicting financial risk in the XGBoost model, with a mean value of 0.97 for the absolute value of Shap, which is much higher than the contribution of other indicators. This in turn helps us to better understand how the model prediction results are formed, thus improving the accuracy and robustness of the model prediction.

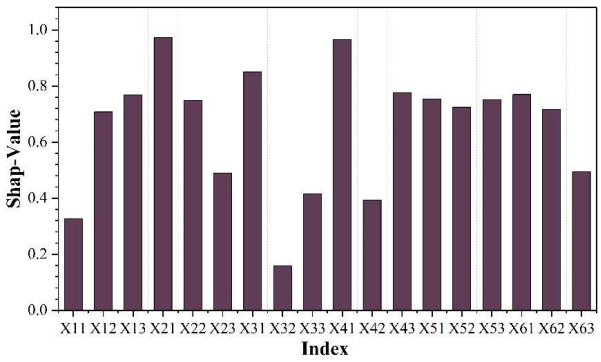

Figure 2: The global mean Shap of the features of Model

III. B. Analysis of model applications

III. B. 1) Business profile

This paper selects ST enterprise as the case enterprise analyzed in this paper, which was founded in 1992, completed the shareholding system reform in 1994, became a joint-stock limited company, and in September 2000, it was listed on the stock exchange. The enterprise is under the jurisdiction of the Sixth Division of the Production and Construction Corps, and is also a national agricultural leading enterprise supported by the Corps, and in June 2016, it was renamed as “ST Health Industry Co. Relying on the unique geographical advantages of Xinjiang and the intensive management of the Corps, the company is committed to the development of tomato industry, with 10 branches and 3 holding subsidiaries as well as a red industrial science and technology park and a biological products company under its jurisdiction, 16 participating and holding enterprises, 16 modernized tomato product processing factories, 29 world-class production lines, a comprehensive production capacity of 604,500 tons, and a large-scale industrial scale, with a production capacity of 604,500 tons. With a production capacity of 604,500 tons, the industry scale ranks the second in China and the third in the world, the products are exported to more than 70 countries and regions in the world, and it is a long-term and stable supplier of global famous food enterprises.

On the evening of August 7, 2022, the enterprise announced that the use of shell companies and not constitute a related party transit company, fictitious purchase and sale of business for many years, a total of 6 years of inflated profits of 530 million, this matter was investigated by the Securities and Exchange Commission (SEC), the company has received the SEC's “Decision on Administrative Penalties”, the SEC warned the company and imposed a fine of 200,000 yuan, on the seven executives imposed a fine ranging from 30,000 yuan to 300,000 yuan. Fines ranging from 30,000 to 300,000 yuan were imposed on seven executives.

III. B. 2) Reasons for selection

In order to save its listed company's “shell”, ST enterprises do not hesitate to carry out massive counterfeiting, forging company transactions, inflating company revenues and profits, the starting point and the use of means are very typical of China Foundation. The starting point and the means used by CKI are very typical. The financial analysis and model warning of typical companies can better show that the results of the model's financial risk prediction are more persuasive and the conclusions revealed are more common.

III. B. 3) Statement of the underlying financial position of the enterprise in 2022

In accordance with the previous analysis, the financial data of ST Enterprises for the year 2022 is selected here to analyze the basic financial status of the company, while the financial data for the year 2023 is selected for comparative analysis. Table 6 shows the balance sheet of ST enterprise in 2022 and 2023. It can be found that in the asset side of ST enterprise in 2023, the total value of assets still shows an increase, the percentage of increase

in 48.83%, mainly reflected in the money funds increased significantly by 164.59%, ST enterprise reserves a large amount of money funds. Inventory increased by 39.47%, this may be mainly due to the downstream stagnant sales leading to the increase in inventory, but also does not exclude the possibility of the expansion of adult production, inventory turnover rate is low. Construction in progress increased by 72.24%, fueling a significant increase in assets in 2023. In addition current liabilities increased by 42.79% and non-current liabilities increased by 49.81%, driving an increase in liabilities and owner's equity in 2023.

Table 6: Balance sheets for 2022 and 2023

Name	Balance at the end of 2023	Balance at the end of 2022
Monetary funds	33449.67	12642.02
Accounts receivable	15125.06	11042.36
Advance payment	1533.81	1619.06
Other receivables	9826.26	18838.33
Inventory	58370.28	41851.04
Total current assets	128201.04	86141.81
Available-for-sale financial assets	1915.21	1921.74
Long-term receivables	505.19	542.13
Long-term equity investment	15524	15524
Investment real estate	8506.46	8787.61
Fixed assets	96458.92	82031.46
Construction in progress	911.42	529.16
Total non-current assets	134290.18	110459.16
Total assets	262491.73	193550.87
Total current liabilities	153289.48	107355.4
Total non-current liabilities	6333.49	12617.26
Total liabilities	159622.97	119972.66

Compare the income statement status of ST enterprises in 2022 and 2023, Table 7 shows the income statement of ST enterprises in 2022 and 2023, the operating profit of ST enterprises in 2022 was already negative, but due to debt restructuring of subsidiaries in 2022, a total of 336 million debt settlement obligation exemption was obtained, and the debt restructuring profit amounted to 339 million, which made the deductible net profit of that year from negative to positive. Turn positive. And through this way to increase profits is obviously unsustainable, and sure enough in 2023 the ST enterprise was a substantial loss, net profit fell -147.34% year-on-year.

Table 7: ST profits in 2022 and 2023

Name	Balance at the end of 2023	Balance at the end of 2022
Total operating Revenue	68144.83	41338.61
Total operating cost	68144.83	41338.61
Taxes and surcharges	85456.41	65134.49
Sales expenses	61938.36	36258.01
Administrative expenses	214.34	257.19
Financial expenses	9312.53	5641.36
Asset impairment loss	5527.61	3841.4
Operating profit	1605.23	3844.71
Non-operating income	7228.32	15317.33
Gains from the disposal of non-current assets	-17311.08	-23328.37
Non-operating expenses	11733.12	34419.11
Loss from disposal of non-current assets	69.65	8.14
Total profit	53.12	305.09
Income tax expense	6.27	61.89
Net profit	-5908.02	10808.26

Table 8 shows the cash flow statement of ST enterprise in 2022 and 2023, ST enterprise in 2022 and 2023, its operating side of the cash flow are positive outflow, the enterprise operating side began to be unable to make ends

meet. The investment side is also a positive outflow, as the enterprise has outflowed a large amount of money in the acquisition of assets. The financing side of the inflow, reflecting the enterprise through the financing into a large amount of funds, the financing side of the cash inflow in 2023 compared with 2022 increased by 49.23 times. There are generally two possibilities for this situation, one is that the enterprise is in the start-up period, the enterprise has not yet or just opened the product market, the operating side can not generate cash inflows, the enterprise still needs a lot of investment, it is necessary to invest a large amount of funds, the formation of production capacity, to develop the market, at this time the enterprise can not generate a certain amount of internal fund balance, its source of funds is only debt, financing and other fund-raising activities. The second is the poor operating conditions of enterprises, resulting in the enterprise operating cash is not enough to make ends meet, at the same time, the financial situation is also very poor, frequent and imprudent investment and mergers and acquisitions require a large amount of external funds, can only rely on fund-raising continue to continue to life. In view of the fact that ST Zhongji has been listed for more than ten years, this paper judges that it is in the second situation.

Table 8: Cash flow statements of ST Enterprise for 2022 and 2023

Name	Balance at the end of 2023	Balance at the end of 2022
Cash flows generated from operating activities	67077.36	40065.36
Cash received from the sale of goods and the provision of services	7007.06	5923.24
Tax and fee refunds received	23283.26	7435.1
Receive other cash related to business operations	97367.47	53264.04
Small amount of cash inflow from operating activities	66142.17	43148.92
Cash paid for purchasing goods and accepting services	7636.86	5226.3
Cash paid to and on behalf of employees	26391.33	19063.4
Pay other cash related to business operations	101290.38	69068.22
Small amount of cash outflow from operating activities	-3822.41	-15303.18
Cash flows generated from investment activities	88.08	
Small amount of cash inflow from investment activities	30219.44	3644.2
Cash paid for the purchase and construction of fixed assets, intangible assets and other long-term assets	77.73	
Obtain the net cash paid by subsidiaries and other business units	30158.67	3644.2
A small amount of cash outflow from investing activities	-30110.2	-3644.2
Cash flows generated from financing activities	46465	31435
Obtain the cash received from the loan	58418.08	25000
Receive other cash related to fundraising activities	105183.53	56435
Small amount of cash inflows from financing activities	47117.14	52014.93
Cash paid for repaying debts	3323.31	3743.51
Cash for distributing dividends, profits or paying interest	50728.4	55409
A small amount of cash outflow from financing activities	54305.13	1049
Net cash flows generated from financing activities	67127.22	40061.26

Combined with the above analysis, the financial position of ST enterprises in 2022 is already worse, only due to a debt restructuring of the inflow of profits, so that the net profit of CKI can barely turn from loss to profit, and 2024 CKI's operations and finances deteriorated in all aspects, the risk is exposed.

III. B. 4) ST Enterprise Data Calculation and Forecasting Analysis

The XGBoost model constructed above is used to predict the financial risk data of ST enterprises in order to determine the possibility of their facing financial risks in the subsequent years, and the results of the prediction analysis are shown in Table 9. At the same time, according to the financial indicators selected by XGBoost, the financial data of ST enterprises in 2022 are analyzed, and the results of the financial indicators are used to further expose the risk situation encountered by the enterprises. From the XGBoost model prediction accuracy are kept above 0.9, which is consistent with the fact, so the XGBoost financial risk prediction model constructed in this paper on the company's prediction results are correct, fully verified this paper's model prediction application effectiveness.

Table 9: Predictive analysis results

Name	Actual value	Predictive value	Accuracy
Main business ratio	-2.1722	-2.1601	0.9944
The ratio of retained earnings to total assets	-0.3521	-0.3402	0.9662
The ratio of earnings before interest, taxes, depreciation and amortization to total operating income	0.5646	0.5511	0.9761
Return on net assets (Weighted)	0.0676	0.0621	0.9186
The ratio of cash received from the sale of goods and provision of services to operating income	0.9623	0.9502	0.9874
Long-term asset suitability ratio	0.8318	0.8254	0.9923
Return on net assets (deducted/average)	-0.493	-0.4815	0.9767
The ratio of working capital to total assets	-0.1086	-0.1044	0.9613
Capital fixation ratio	1.3407	1.3201	0.9846
The ratio of net cash flows generated from operating activities to non-current liabilities	-1.2326	-1.2163	0.9868
Gross profit margin on sales	0.1229	0.1114	0.9064
Tangible net asset debt ratio	1.4746	1.4373	0.9747

IV. Conclusion

This paper first determines the sample enterprise and time window, with the help of relevant literature and crawling software, respectively obtains the quantitative data and annual report text data of this research. In order to ensure the validity of the results of this study, the collected data are preprocessed, and then the enterprise financial risk prediction index system and model evaluation index are set. Finally, on the basis of indicator dimensionality reduction of KCPA algorithm, XGBoost algorithm is used to construct enterprise financial risk prediction model, and the model is applied and analyzed. Through the comparative analysis of model prediction performance, it is found that the four prediction performance index values of this paper's model are 0.9545 (0.9722), 0.9922 (0.9927), 0.9245 (0.9435), 0.9732 (0.9728), which are better than the other two prediction models, and the prediction performance of the XGBoost algorithm is verified. In addition, the prediction accuracy of this paper's prediction model for ST enterprise's financial risk reaches more than 0.9, which is consistent with the actual situation of the enterprise, which not only shows that this paper's model has excellent prediction application effectiveness, but also provides a reference for the intelligent management of enterprise's finance.

References

[1] Martynenko, O. V., Makarova, O. N., & Makarova, Y. N. (2019, January). Role of Financial Statements of the Production Enterprise at Assessment of Risks and Threats of Foreign Economic Activity. In International Scientific Conference" Far East Con"(ISCFEC 2018) (pp. 311-314). Atlantis Press.

[2] Wang, J., Liu, G., Xu, X., & Xing, X. (2024). Credit risk prediction for small and medium enterprises utilizing adjacent enterprise data and a relational graph attention network. Journal of Management Science and Engineering, 9(2), 177-192.

[3] Yi, J. (2022). Research on enterprise financial economics early warning based on machine learning method. Journal of Computational Methods in Sciences and Engineering, 22(2), 529-539.

[4] Sun, X., & Lei, Y. (2021). Research on financial early warning of mining listed companies based on BP neural network model. Resources Policy, 73, 102223.

[5] Wang, J., Tian, B., & Gao, H. (2024, November). Mathematical Model of Enterprise Financial Risk Early Warning Based on SVM Method. In Proceedings of the 2024 4th International Conference on Big Data, Artificial Intelligence and Risk Management (pp. 854-858).

[6] Zhang, T., Chen, Q., & Zhu, X. (2024). Analysis of Enterprise Financial Risk Early Warning Model Based on the Evidence Theory and Whitening Weight Function. Scientific Programming, 2024(1), 9207782.

[7] Ding, Q. (2021). Risk early warning management and intelligent real-time system of financial enterprises based on fuzzy theory. Journal of Intelligent & Fuzzy Systems, 40(4), 6017-6027.

[8] Xinxian, C., & Jianhui, C. (2022). Digital transformation and financial risk prediction of listed companies. Computational intelligence and neuroscience, 2022(1), 7211033.

[9] Qi, Q. (2021). Study on financial risk prediction of enterprises based on logistic regression. Journal of Computational Methods in Science and Engineering, 21(5), 1255-1261.

[10] Jiang, H. (2022). Risk prediction model of enterprise financial data based upon sensor signal fusion. Mobile Information Systems, 2022(1), 7819224.

[11] Chen, H., Guo, R., & Jin, Y. (2025). Risk Early Warning of Enterprise Financial Management Risk Based on CNN and BiLSTM under Accounting Informatization. J. COMBIN. MATH. COMBIN. COMPUT, 127, 7425-7440.

[12] Zhang, H., & Luo, Y. (2022). Enterprise financial risk early warning using BP neural network under internet of things and rough set theory. Journal of interconnection networks, 22(03), 2145019.

[13] Dong, X., Dang, B., Zang, H., Li, S., & Ma, D. (2024). The prediction trend of enterprise financial risk based on machine learning arima model. Journal of Theory and Practice of Engineering Science, 4(01), 65-71.

[14] Zhu, W., Zhang, T., Wu, Y., Li, S., & Li, Z. (2022). Research on optimization of an enterprise financial risk early warning method based on the DS-RF model. International review of financial analysis, 81, 102140.

[15] Chen, X., & Long, Z. (2023). E-commerce enterprises financial risk prediction based on FA-PSO-LSTM neural network deep learning model. Sustainability, 15(7), 5882.
[16] Xin, Q. (2024, July). Construction of a machine-learning-based risk management evaluation model for enterprise financial reporting. In International Conference on Communication, Information, and Digital Technologies (CIDT2024) (Vol. 13185, pp. 19-26). SPIE.
[17] Yue, H., Liao, H., Li, D., & Chen, L. (2021). Enterprise financial risk management using information fusion technology and big data mining. Wireless Communications and Mobile Computing, 2021(1), 3835652.
[18] Bi, W., Xu, B., Sun, X., Wang, Z., Shen, H., & Cheng, X. (2022, August). Company-as-tribe: Company financial risk assessment on tribe-style graph with hierarchical graph neural networks. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 2712-2720).
[19] Qingdong Zhu,Jian Wang,Zhaoliang Gu,Xuelei Wang,Qihui Cui,Longlong Li... & Wenbing Zhu. (2025). Research on partial discharge pattern recognition of transformer based on KPCA and JS-SVM algorithm. Journal of Physics: Conference Series,2963(1),012013-012013.
[20] Lei Xu,Shaomu Wen,Hongfa Huang,Yongfan Tang,Yunfu Wang & Chunfeng Pan. (2025). Corrosion failure prediction in natural gas pipelines using an interpretable XGBoost model: Insights and applications. Energy,325,136157-136157.
[21] Xuan Zhao,Pei Fu Zhang,Daxu Zhang,Qi Zhao & Yiliyaer Tuerxunmaimaiti. (2025). Prediction of interlaminar shear strength retention of FRP bars in marine concrete environments using XGBoost model. Journal of Building Engineering,105,112466-112466.
[22] YuanyuanGuo,FeihongLi & WumaierailiAimaitikali. (2025). Exploring the Effects of Built Environment on Traffic Microcirculation Performance Using XGBoost Model. Journal of Advanced Transportation,2025(1),8821071-8821071.
[23] Zisheng Zeng,Bin Song,Shaocheng Wu,Yongwen Li,Deyu Nie & Linong Wang. (2025). Prediction of Breakdown Voltage of Long Air Gaps Under Switching Impulse Voltage Based on the ISSA-XGBoost Model. Energies,18(7),1800-1800.