Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Enhancing Corporate Financial Transparency and Performance Assessment through Big Data and Machine Learning

Xiaoyang Meng[1,✉], Ying Jin[2]

[1] *School of Accounting, Jiaozuo University, Jiaozuo, Henan, 454100, China*
[2] *College of Continuing Education, Jiaozuo University, Jiaozuo, Henan, 454100, China*

ABSTRACT

The higher the corporate financial transparency, the more it can reduce the information asymmetry, which can enhance the market trust and improve the corporate performance. In order to improve corporate financial transparency, the study constructs a financial fraud identification model by improving the machine learning model based on XGBoost algorithm from the financial fraud factors. Based on the XGBoost algorithm, the model integrates the decision rules through the weighted fusion method to generate a new decision tree to determine the financial fraud. In order to improve the ability of enterprise performance assessment, the baryon support vector machine method is used to classify the performance of enterprise employees, and the nonlinear baryon support vector machine is used to establish the enterprise performance assessment model. In the process of verifying the effect of the two models, text indicators are extracted using big data technology to provide a rich feature set for the financial fraud identification model. The data from ERP, CRM and other systems are integrated to provide a comprehensive and high-quality data set for the enterprise performance assessment model. After empirical analysis, the combination of big data and machine learning can improve the effect of financial fraud identification, and then effectively improve the transparency of corporate finance. The enterprise performance evaluation model provides a scientific and efficient quantitative evaluation tool for enterprise managers, and effectively improves the enterprise performance evaluation capability.

Keywords: Big Data, Machine Learning, XGBoost Algorithm, Twin Support Vector Machines, Financial Transparency

✉ Corresponding author.
E-mail address: hnjzmxy2023@163.com (X. Meng).
Received 03 February 2024; Revised 12 April 2024; Accepted 10 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-383

1. Introduction

Big data has brought new development opportunities and changes for enterprise financial management. The traditional financial management model has been difficult to meet the requirements of modern enterprises for data processing speed, analysis depth and decision-making accuracy [1-2]. Big data analysis technology, by virtue of its advantages in mass data processing, multi-dimensional analysis and real-time monitoring, is gradually becoming an important tool to improve the level of enterprise financial management [3-4].

The development overview and application of big data analytics in enterprise financial management is undergoing profound changes. Leading international enterprises have taken the lead in applying big data analytics to financial management, forming a technical system centered on predictive analysis, real-time monitoring and intelligent decision-making [5-6]. More than 75% of the world's large enterprises have incorporated big data analytics into their financial management strategic planning, with an average annual investment of 15% to 20% of the IT budget. Chinese enterprises have gradually constructed a financial analysis model based on big data in the process of digital transformation of financial management [7-9]. Typical applications include the application of machine learning algorithms for accounts receivable forecasting, the use of deep learning technology to establish a financial risk early warning system, and the use of blockchain technology to optimize the management of financial records [10-11]. Leading companies such as Alibaba and Huawei have established financial big data platforms to realize multi-dimensional analysis and real-time monitoring of financial data. At the technical level, enterprise financial big data analysis mainly adopts distributed computing frameworks such as Hadoop, Spark, etc., combined with data analysis tools such as Python, R, etc., to build a complete technical chain of financial data collection, storage, analysis and visualization [12-13]. The wide application of cloud computing technology reduces the cost threshold for enterprises to build financial big data platforms, and promotes the popularization of related technologies in small and medium-sized enterprises.

Meanwhile, digital transformation has become a key initiative for enterprises to enhance competitiveness and achieve sustainable development, especially the wide application of artificial intelligence technology, which provides powerful technical support for the transformation of enterprise financial accounting [14-15]. Through intelligent data processing, automated financial analysis, accurate risk prediction and other technical means, artificial intelligence can effectively improve the efficiency and precision of financial management, reduce the occurrence of human error, and provide timely and accurate decision support for senior management [16-18]. Therefore, an in-depth study of how artificial intelligence can promote the digital transformation of enterprise financial accounting is of great practical significance for promoting the innovation of enterprise management mode [19-20].

With the rapid development of information technology, artificial intelligence has gradually become the core driving force to promote the digital transformation of various industries, especially in the field of enterprise financial accounting. Combining the information asymmetry theory, principal-agent theory and risk management theory, Ren, S elucidated that big data empowered enterprise financial decision-making, which helps to lead enterprise business expansion well, and at the same time promotes the level of internal management and competitiveness of enterprises [21]. Gao, J talked about the introduction of big data technology into the work of enterprise financial management, effectively reducing the perceived operational errors and improving the efficiency and quality of financial management [22]. Shang, H et al. based on the empirical research and analysis methodology, specified that the application of big data technology in the field of enterprise financial management significantly improves the efficiency of financial data processing, and at the same time, also facilitates the realization of the goal of accurate prediction of financial risks [23].

Yi, J compared and analyzed the performance of Support Vector Machine (SVM) and Back

Propagation Neural Network (BPNN) algorithms in machine learning algorithms for corporate financial risk prediction, and found that SVM fused with BP algorithm performs better in terms of model convergence speed and prediction accuracy [24]. Peng, C and Tian, G, in order to improve the efficiency and accuracy of the design work of corporate finance, envisioned a intelligent auditing strategy for corporate finance, and the feasibility was verified by numerical simulation tests [25]. All of the above studies confirm that big data technology and machine intelligence learning algorithms play a positive role in promoting the efficiency and precision of enterprise financial management and financial auditing work, but there are fewer studies on how big data technology and machine intelligence learning algorithms work on enterprise financial management transparency and financial performance.

The study constructs a financial fraud identification model and an enterprise performance assessment model based on machine learning to improve the financial transparency and performance assessment ability of enterprises, respectively. The financial fraud identification model is an interpretable machine learning model based on the XGBoost framework, which uses a weighted cross-entropy loss function to deal with unbalanced samples, and reduces the weighted model to a single decision tree to explain financial fraud in a clear tree structure. The enterprise performance assessment model is to use the twin support vector machine method to classify all the important factors affecting human resource management with effective arithmetic and complete the employee performance assessment of the enterprise. Text indicators and data sets are constructed based on big data technology during the use of the two models respectively to improve the identification accuracy and application efficiency of the models.

2. Methods of enhancing corporate financial transparency

Financial transparency refers to the extent to which an enterprise's financial information and all information that could lead to changes in its financial position is publicly available or known. Untimely and incomplete disclosure of information and financial fraud are the key factors affecting the financial transparency of enterprises. Accordingly, this chapter improves the machine learning model based on XGBoost algorithm and constructs a financial fraud identification model in order to improve the financial transparency of enterprises.

2.1. Optimizing the XGBoost objective function

Early financial fraud identification relied heavily on statistical analysis and discriminant analysis models, such as classic models like F-Score, M-Score, and Z-Score. With the advancement of technology, researchers started to use machine learning to build more efficient models.

XGBoost is a machine learning algorithm based on the gradient boosting framework, which builds strong learners by iteratively adding weak learners (e.g., decision trees) [26]. Compared to GBDT, XGBoost uses a unique objective function consisting of a training loss and a regularization term. The training loss is used to measure the difference between the predicted and actual values of the model, and the regular term is used to control the model complexity to prevent overfitting. The efficiency and powerful adaptability of XGBoost make it a suitable data processing tool for a wide range of tasks. Modified XGBoost reduces the effect of different errors in the classification task by introducing a weighted cross-entropy loss function, specifically the optimized loss function is:

$$Loss = y \times a_{FP} \times \ln p + (1-y) \times a_{FN} \times \ln(1-p) \tag{1}$$

where: y is the actual label, indicating the positive and negative categories of the sample. p is the predicted probability of the model. a_{FP} & a_{FN} are the coefficients used to adjust the error weights.

Unlike determining the optimal weights through grid search [27], which focuses on reducing false-negative errors in financial distress prediction, this paper uses Bayesian optimization to determine the

best parameters. After the optimization is completed, the XGBoost model is retrained and a_{FP} & a_{FN} are replaced with Average False Positive Rate [Average False Positive Rate, $a_{FP} = FP/(FP+TN)$] and Average False Negative Rate [Average False Negative Rate, $a_{FN} = FN/(FN+TP)$] to better reflect the impact of different types of errors in financial distress detection.

2.2. Decision rules for weighted fusion XGBoost

Since the essence of the modified XGBoost algorithm is to approximate the weighted XGBoost by an interpretable decision tree, in order to better evaluate the model's generalization ability and performance variation at different time points, the decision rule merging mechanism is used: by comparing the decision paths of neighboring decision trees, rules with the same features but different values are identified, adjusted accordingly to form a cross-rule set, and calculating its accuracy. The following is the core formula for updating the probability of new decision paths when merging decision paths:

$$p_{t,i} \leftarrow \frac{1}{2}(p_{t,i} + p_{t-1,j}) \tag{2}$$

where $p_{t,i}$ denotes the probability of the i rd path in the set of decision paths of the t nd decision tree and $p_{t-1,j}$ denotes the probability of the i th path in the set of decision paths of the $t-1$ th decision tree. A smoother tree structure is achieved by weighted fusion of the decision paths of two neighboring decision trees t and $t-1$ to form a new decision path. The accuracy of the cross rule set is calculated as:

$$Rule\_Acc = \frac{Rule_t + Rule_{t-1}}{2} \tag{3}$$

where, $Rule_t$ denotes the decision path accuracy of the t nd decision tree and $Rule_{t\text{-}1}$ denotes the decision path accuracy of the $t-1$ th decision tree, and $Rule\_Acc$ is obtained by calculating the mean of the two.

2.3. Generating a new decision tree using fusion rules

After cyclically merging all the decision tree rules output from XGBoost, the decision path intersection RT is obtained.To control the complexity of the new decision tree, the modified XGBoost algorithm selects the top L decision paths to form a rule set based on the ranking of the predicted probabilities of the decision paths, and at the same time, the rules in the rule set are extracted by splitting features and stored in a hash table. Subsequently, the decision tree is constructed layer by layer from top to bottom by prioritizing the use of the best rules up to a set depth of the decision tree, aiming to approximate the decision logic of the original weighted XGBoost.

In constructing a new decision tree, the modified XGBoost algorithm selects the best splitting features by maximizing the information gain (IG) and recursively constructs the tree structure. The information gain is based on the loss function formula (1), which reduces the overall prediction error by minimizing the loss of each node during the splitting process of each decision node. The information gain is calculated as follows:

$$IG = \frac{|Rule_L| \times E(Rule_L) + |Rule_R| \times E(Rule_R)}{|Rule_L| + |Rule_R|} - E(Rule) \tag{4}$$

where $|Rule_L|$ and $|Rule_R|$ denote the number of rules in the left and right subsets, respectively, $E(Rule_L)$ and $E(Rule_R)$ denote the loss in the left and right subsets, respectively, and $E(Rule)$ denotes the loss in selecting all paths.

3. Enterprise performance assessment methodology

Effective arithmetic categorization of all important factors affecting human resource management through machine learning methods can further enhance the efficiency of enterprise performance evaluation.

3.1. Bipartite Support Vector Machines

As an improved version of traditional machine learning, bosonic support vector machines have more excellent classification capabilities and are well suited for solving the problem of classifying samples of approximate types [28]. In addition, compared with the traditional support vector machine (SVM), the bosonic support vector machine searches for a pair of non-parallel hyperplanes, and thus is more computationally efficient.

In practical application cases, most data samples are not simple binary categorization, for example, the enterprise performance data samples studied in this paper are not simple two-dimensional attribute data, and simple linear straight line division can no longer distinguish all samples. Therefore, the kernel function solution is introduced to solve the problem when linearly indivisible.

Then the method of seeking hyperplane is:

$$K(x^T, C^T)u_+ + b_+ = 0, K(x^T, C^T)u_- + b_- = 0(1) \tag{5}$$

where $C = [1; B] \in R^{l \times n}$, K are the kernel functions.

Similarly, the planes dividing the positive and negative classes, respectively, satisfy the following conditions:

$$\begin{aligned} &\min_{u_+ b_+ \xi_-} \frac{1}{2} \| K(A, C^T)u_+ + e_+ b_+ \|^2 + c_1 e_-^T \xi_- \\ &-(K(B, C^T)u_+ + e_- b_+) + \xi_- \geq e_-, \xi_- \geq 0 \end{aligned} \tag{6}$$

$$\begin{aligned} &\min_{u_- b_- \xi_+} \frac{1}{2} \| K(B, C^T)u_- + e_- b_- \|^2 \\ &+ c_2 e_+^T \xi_+ (K(A, C^T)u_- + e_+ b_-) \\ &+ \xi_+ \geq e_+, \xi_+ \geq 0 \end{aligned} \tag{7}$$

To further simplify Eqs. (6) and (7), they are pairwise transformed:

$$\begin{aligned} &\max_\alpha e_-^T \alpha - \frac{1}{2} \alpha^T R (S^T S)^{-1} R^T \alpha \\ &s.t.\ 0 \leq \alpha \leq c_1 e_- \end{aligned} \tag{8}$$

$$\begin{aligned} &\max_\gamma e_+^T \gamma - \frac{1}{2} \gamma^T S (R^T R)^{-1} S^T \gamma \\ &s.t. 0 \leq \gamma \leq c_2 e_+ \end{aligned} \tag{9}$$

where $R = [\bar{K}(B, C^T)e] S = [\bar{K}(A, C^T)e_+]$. Solving Eqs. (8) and (9) yields:

$$(u_+^T, b_+)^T = -(S^T S)^{-1} R^T \alpha \tag{10}$$

$$(u_-^T, b_-)^T = -(R^T R)^{-1} S^T \gamma \tag{11}$$

According to equations (10) and (11), u_+, u_-, b_+, b_- yields the hyperplane of the classification as:

$$classlabel = \arg \min_{k=+,-} | K(x^T, C^T)u_k + b_k | \tag{12}$$

3.2. Nonlinear Baryonic Support Vector Machines for Firm Performance Evaluation

According to the personnel data samples, combined with the twin support vector machine method, the enterprise performance is categorized. There are more elements of personnel composition, and the core elements of enterprise personnel are selected for indicator extraction and quantification, and the finalized performance assessment indicators are shown in Table 1. Through the extraction of key elements of personnel composition and the optimization of similar elements, a total of 8 elements are selected as the basic components of the personnel involved in the classification sample.

Table 1. Personnel data table

Index	Variable	Value
Personnel major	X1	Integer
Working hours	X2	Integer
Post grade	X3	The unit corresponds to 1-10
Technical results	X4	The number of results identified by the company
Professional skill grade	X5	Grade
Training situation	X6	The number of training
rewards	X7	laureates
Project experience	X8	The number of projects hosted

Then the baryon support vector machine model is established, and the process of enterprise performance evaluation by nonlinear baryon support vector machine is as follows:

1) Personnel ability index element selection.
2) Training sample and test sample selection.
3) Sample initialization and vectorization.
4) Kernel function selection.
5) Relaxation and control variables setting.
6) The first binary support vector machine operation to categorize the poorer samples as “medium” and “poor”.
7) The second binary support vector machine operation classifies the worse samples as “medium” and “poor”.
8) The third binary support vector machine operation classifies the better samples as “excellent” and “good”.

After three binary support vector machine operations, the personnel performance classification results are obtained.

4. Empirical analysis

In order to verify the effect of the financial fraud identification model constructed in the previous section, the study uses big data technology to combine the textual information of management discussion and analysis (MD&A) in the annual reports of listed companies to extract textual indicators and provide a rich feature set for the financial fraud identification model. In the enterprise performance evaluation model, big data technology can also be used to integrate data from ERP, CRM and other systems to provide a comprehensive and high-quality data set for the machine learning model.

4.1. Analysis of the effect of improving the financial transparency of enterprises

The research selects BP neural network (BPNN), SVM, and random forest (RF) to construct financial fraud recognition model respectively. Using big data technology to construct text information,

respectively, financial and non-financial indicators and samples added to the text indicators are put into the above three models as well as the financial fraud identification model constructed in this paper for identification.

4.1.1. Sample Data Selection and Indicator System Construction:

1) Sample Selection. The sample of this paper is selected from A-share listed companies from 2019-2024, and the fraud samples come from penalized companies announced by SSE and SZSE. The specific types of penalties include: material omissions, fictitious profits, false records, and false listing of assets. At the same time, non-financial companies in the same industry with similar total asset size and not penalized are selected as non-fraud company samples according to the 1:1 matching principle. The screening criteria for the matched companies are as follows:

(1) The sample of matched companies is selected according to the 1:1 matching principle.

(2) Matching is done according to the 2012 industry categorization criteria and firms in the financial sector are excluded.

(3) The total assets of non-fraud matched companies are comparable or similar to the sample of financial fraud companies.

By screening the company samples according to the above screening conditions, a total of 390 companies were selected as the data set for fraud identification, of which 195 fraud company samples, while 195 non-fraud company samples were matched 1:1.

2) Fraud identification index system construction:

(1) Financial and non-financial index system construction. This paper takes financial fraud related research as the theoretical basis, fully considers the accessibility of data, the relevance of indicators, and constructs the indicator system based on financial and non-financial indicators as shown in Table 2.

Table 2. Based on financial and non-financial indicators

	Index type	Index name
Financial index	Solvency	Mobility ratio
		Speed ratio
		Asset ratio
	Operational capacity	Receivable turnover
		Inventory turnover
	Profitability	Total asset profit rate
		Return on equity
	Development ability	Total asset growth rate
		Rate of operating profit
	Risk level	Integrated lever
Non-financial indicators	Contract conflict of executive pay	Executive shareholding ratio
	Internal control	Internal control is defective
	Equity structure	Domestic share ratio
		The largest shareholder shareholding ratio
	Governance structure	The chairman and the general manager are concurrently
		Independent directors account for the number of directors

(2) Measurement index system construction based on MD&A management tone. Management Discussion and Analysis (MD&A) reflects the management's analysis of the past company's operating conditions and prediction of future development trends, and is an important textual information in the annual report. Research has shown that obtaining information about the company's performance and future development prospects from the management's perspective through MD&A is effective in solving the fraud identification problem.

In this paper, the bag-of-words approach is used to extract management tone indicators based on sentiment lexicon from the MD&A section of annual financial reports issued by various companies. Sentiment lexicon resources used in this paper include “Chinese Sentiment Analysis Word Collection” and “Chinese Sentiment Vocabulary Ontology” in “Chinese Sentiment Analysis Word Collection” published by Zhi.com. “The Chinese Sentiment Analysis Vocabulary is categorized into negative sentiment, positive sentiment, negative evaluation, positive evaluation, degree level words, and assertion words. The “Ontology of Chinese Emotion Words” divides emotion into seven aspects: happiness, goodness, anger, sadness, fear, evil, and surprise. Positive emotion words are regarded as positive tone and negative emotion words are regarded as negative tone, and the positive and negative tones constitute the net tone of the text.

Using python to extract management tone. Firstly, we use “jieba” library in python, which is one of the most commonly used and matured Chinese lexical tools, and in this paper, we use “jieba” to lexically process the text of management discussion and analysis. Then, in order to delete those words which are not meaningful to the analysis, the list of deactivated words will be introduced to remove the deactivated words in the text, and the frequency of various kinds of emotion words in the MD&A will be counted according to the “Chinese Sentiment Analysis Words Collection” and the “Chinese Sentiment Vocabulary Ontology Library”, and finally divided by the frequency of occurrence of these words. Based on the “Chinese words for sentiment analysis” and “Chinese sentiment vocabulary ontology”, we counted the frequency of each type of sentiment words in the MD&A, and finally divided it by the total number of MD&A words to get the relative length of the variable, which was used as the management tone based on MD&A.

The MD&A management tone measure is combined with financial and non-financial indicators to investigate the impact of the application of big data technology on the fraud identification effect.

(3) Data normalization. Due to the large differences in the nature of the indicators and indicators, if the original data are directly analyzed, the larger-scaled indicators have a greater impact on the smaller-scaled indicators, weakening the role of the smaller-scaled indicator data on the results. In order to enhance the comparability of the data and reduce the fluctuation range of the data, this paper normalizes the data in the fraud dataset and converges the data to between [0,1], as shown in equation (13):

$$x^{'} = \frac{x - min(x)}{max(x) - min(x)} \tag{13}$$

where x denotes the raw data, $x^{'}$ denotes the normalized result, $\min(x)$ denotes the minimum value of the same type of indicator, and $\max(x)$ denotes the maximum value of the same type of indicator.

4.1.2. Financial fraud identification model evaluation indicators. In order to evaluate how the model recognition performance is, as well as to better compare the performance of each model, this paper selects the first type of error (false alarms), the second type of error (missed alarms) and the AUC value as auxiliary evaluation indexes in addition to the most important recognition accuracy.

1) The accuracy rate, i.e., the overall correct rate of model recognition, is calculated as shown in equation (14), and the closer the value is to 1, the better the model recognition effect is:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{14}$$

2) The first type of error, also known as false alarm, is the probability that the model will classify a non-fraud sample as fraud. As shown in equation (15) below, the closer the value is to 0, the better the model recognition performance:

$$\text{First type of error} = \frac{FP}{FP + TN} \tag{15}$$

3) The second type of error, also known as underreporting, is the probability that the model will classify a fraud sample as non-fraud. As shown in equation (16) below, the closer the value is to 0, the better the model recognition performance:

$$\text{Second type of error} = \frac{FN}{TP + FN} \tag{16}$$

4) AUC value, which is the area enclosed by the ROC curve formed by taking FPR (False Positive Rate) as the horizontal axis and TPR (True Positive Rate) as the vertical axis. The range of this value is [0.5, 1], the closer to 1, indicating that the model can better identify financial fraud.

4.1.3. Analysis of model identification results:

1) BPNN. The statistical results of the accuracy, Type I error, Type II error and AUC value of the financial fraud recognition model constructed based on BP neural network are shown in Figure 1. When only financial and non-financial indicators are considered (denoted by Financial), the model's recognition accuracy, first-type error, second-type error, and AUC value are 78.34%, 29.61%, 20.41%, and 83.22%, respectively. After the introduction of text metrics (with Financial+Text), the recognition accuracy and AUC value improved by 1.2% and 4.81%, respectively. Type I and Type II errors decreased by 9.41% and 0.89%, respectively.

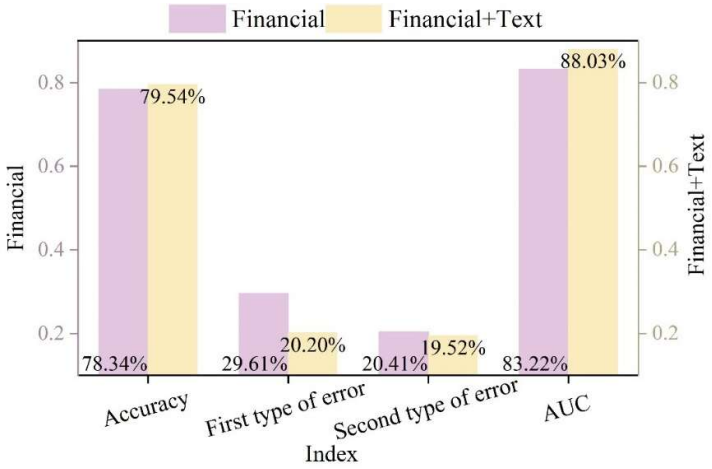

Fig. 1. Evaluation results of the fraud identification model of BPNN

2) SVM. The statistical results of the accuracy rate, the first type of error, the second type of error, and the AUC value of the financial fraud recognition model constructed based on SVM are shown in Figure 2. When only financial and non-financial indicators are considered, the model's recognition accuracy, first type of error, second type of error, and AUC value are 76.6%, 29.46%, 18.36%, and 77.39%, respectively. After the introduction of textual metrics, the recognition accuracy and AUC value were improved by 1.07% and 8.64%, respectively. Type I and Type II errors decreased by 1.62% and 0.98%, respectively.

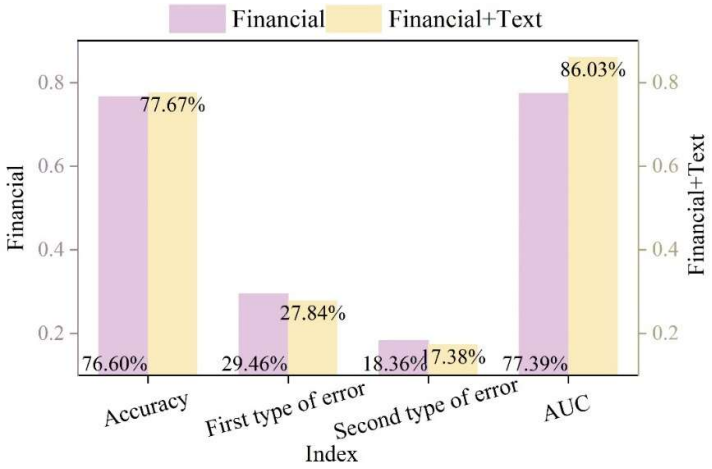

Fig. 2. Evaluation results of the fraud identification model of SVM

3) RF. The statistical results of the accuracy, type I error, type II error, and AUC value of the financial fraud identification model constructed based on random forest are shown in Figure 3. When only financial and non-financial indicators are considered, the model's recognition accuracy, first type of error, second type of error, and AUC value are 80.15%, 19.64%, 20.49%, and 84.12%, respectively. After the introduction of textual metrics, the recognition accuracy and AUC values were improved by 1.72% and 2.3%, respectively. Type I and Type II errors decreased by 2.66% and 0.87%, respectively.

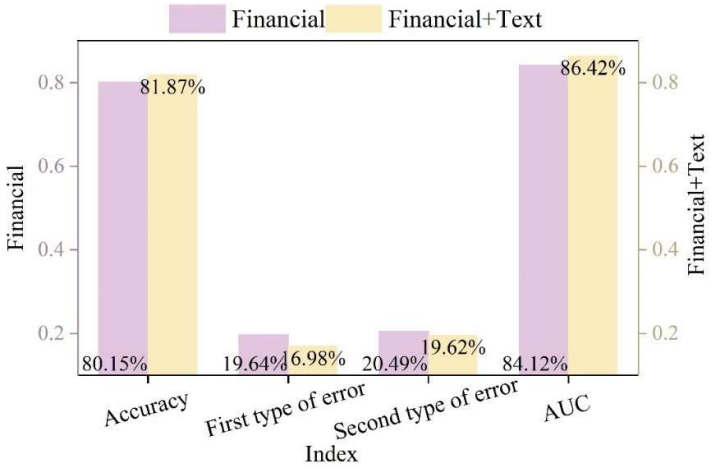

Fig. 3. Evaluation results of the fraud identification model of RF

4) Modified XGBoost. The statistical results of the accuracy, the first type of error, the second type of error and the AUC value of the financial fraud recognition model based on modified XGBoos constructed in this paper are shown in Figure 4. When only financial and non-financial indicators are considered, the model's recognition accuracy, type I error, type II error, and AUC value are 80.31%, 18.71%, 18.02%, and 85.12%, respectively. After the introduction of textual metrics, the recognition accuracy and AUC value were improved by 2.09% and 4.04%, respectively. Type I and Type II errors decreased by 2.9% and 0.97%, respectively.

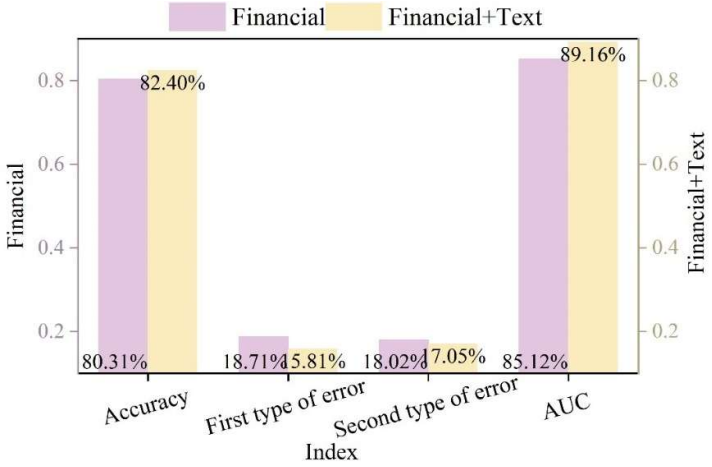

Fig. 4. Evaluation results of the fraud identification model of improved XGBoost

In summary, when textual indicators collected based on big data technology are added to the model, the accuracy of model identification can be improved to a certain extent and the probability of model misreporting can be effectively reduced, i.e., the probability of the model determining non-fraud samples as fraud is reduced. The textual indicators collected based on big data provide incremental information to improve the accuracy of model identification.

In order to see the recognition effect of each model more clearly, the following is a comparative analysis of the recognition effect of these four models from the situation of financial and non-financial indicators and the introduction of text indicators, respectively. The model comparison results are shown in Fig. 5, (a) and (b) represent the cases of financial and non-financial indicators and the introduction of text indicators, respectively. Regardless of which indicator scenario, the classification effect of the model constructed based on modified XGBoost is overall better than that of the model constructed based on BPNN, SVM and RF. From Fig. 5(b), after combining the big data technology and

XGBoost algorithm, the recognition accuracy and AUC value of the financial fraud recognition model are 82.4% and 98.16%, respectively, which is a big improvement compared with the models of other methods. The first type of error and the second type of error are 15.81% and 17.05%, respectively, which are decreased to a larger extent compared to other method models.

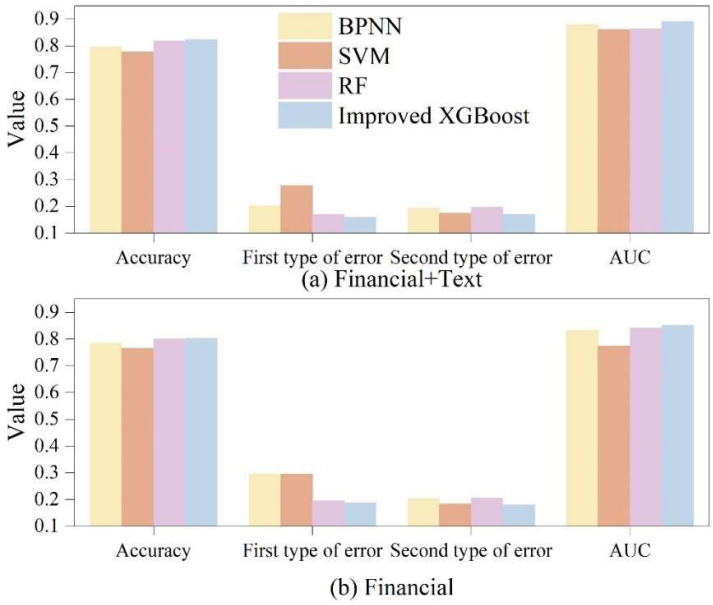

Fig. 5. Model comparison results

4.1.4. Application of financial fraud identification models. In order to verify the accuracy and reliability of the method model of this paper in practical application. From the typical cases of violations released by the Securities and Exchange Commission (SEC), eight listed companies in the manufacturing industry with financial fraud were selected, and the method of this paper was used to predict the financial fraud behavior of these listed announcements and compare them with the time of investigation and punishment by the SEC, and the results of the comparison are shown in Table 3. In the beginning year of the financial fraud behavior of listed companies, the model predicts the fraud time is earlier than the time of investigation and punishment by the SEC. That is, if the model constructed in this paper is used to predict the financial fraud of listed companies, the financial fraud of listed companies can be detected in time. In addition, most of the earliest time of fraud predicted by the model for listed companies is consistent with the start time of fraud for listed companies, and the earliest start time of fraud predicted for a small number of listed companies is earlier than the start time of fraud identified by the SEC. For the prediction of financial fraud earlier than the time identified by the SEC does not affect the validity of the model identification. To a certain extent, the above results show that the combination of big data and improved XGBoost can improve the effect of financial fraud identification, and then effectively improve the transparency of corporate finance.

Table 3. Forecast the time of the fraud is compared with the CSRC

Company code	Age of fraud	CSRC check time	Model prediction time
688311	2021	2024	2021
300903	2018	2024	2018
002587	2019	2023	2019
688001	2019	2024	2016
688559	2018	2023	2018
603809	2019	2023	2018
603758	2019	2024	2017
688569	2019	2024	2018

4.2. Analysis of the effect of the improvement of enterprise performance evaluation capacity

4.2.1. Analysis of model classification effects. Taking the performance of enterprise A in 2024 as an example, using big data technology to integrate the data of ERP, CRM and other systems, 10 employees are selected from them, and their data about the performance evaluation indexes involved in Table 1 are counted and classified using the enterprise performance evaluation model constructed above.

In order to verify the classification performance of the model, this paper compares the classification results of the nonlinear bosonic support vector machine with the classification results of the radial basis function neural network (RBF) classification method that is more commonly used at present, as well as the results of the enterprise's original manual evaluation of the 10 employees, in order to test the classification effect of the enterprise performance evaluation model.

Comparison of RBF neural network, SVM and manual classification results are shown in Table 4 and Figure 6. Setting the manual classification results as the real value, the method model and RBF neural network model in this paper correctly classify 9 and 5 respectively, with classification accuracies of 90% and 50% respectively. It shows that employee performance evaluation using nonlinear bosonic support vector machine is closest to manual evaluation.

Table 4. Comparison of classification results

Staff	SVM	RBF	Artificial
1	Bad	Bad	Bad
2	Medium	Medium	Bad
3	Medium	Medium	Medium
4	Medium	Bad	Medium
5	Medium	Bad	Medium
6	Good	Excellent	Good
7	Good	Good	Good
8	Excellent	Excellent	Excellent
9	Good	Medium	Good
10	Medium	Medium	Medium

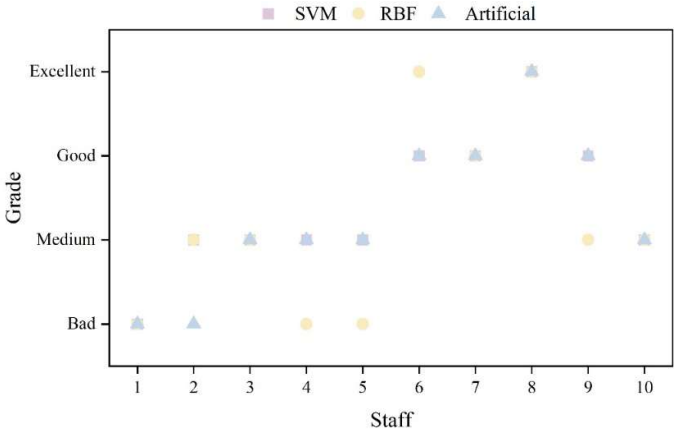

Fig. 6. Comparison of classification results

4.2.2. Application of performance assessment models. The comprehensive performance evaluation values of 100 listed agricultural companies are collected as shown in Figure 7. Use the model of this paper for performance classification. The results of enterprise performance classification are shown in Figure 8. There are 3, 15, 67 and 15 enterprises with "excellent, good, medium and poor" performance ratings respectively, indicating that the enterprise performance evaluation model constructed in this paper can effectively categorize the evaluation ratings of companies according to the comprehensive performance evaluation values.

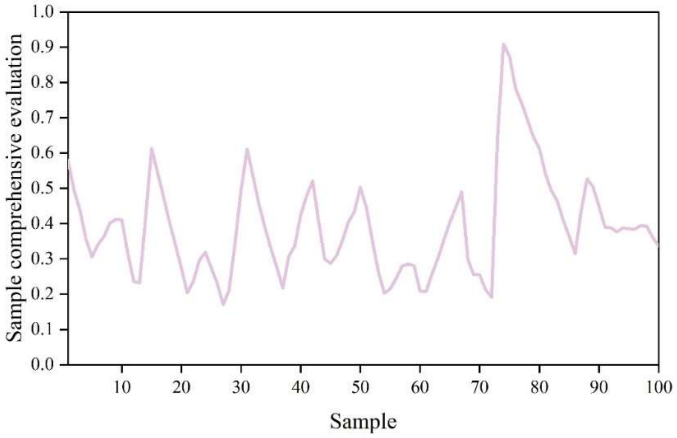

Fig. 7. Comprehensive performance evaluation

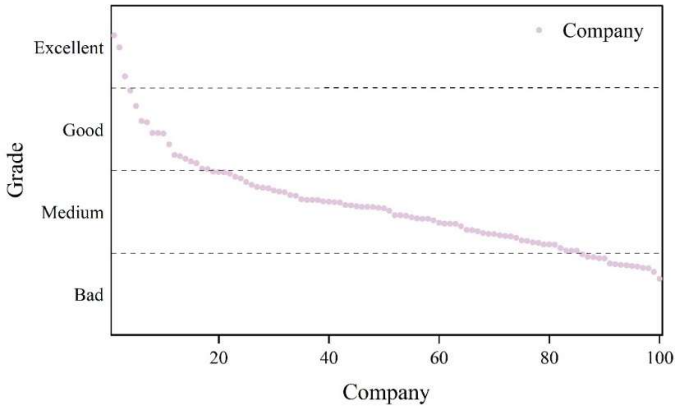

Fig. 8. Corporate performance classification results

5. Conclusion

In order to improve the enterprise financial transparency and performance assessment ability, this paper constructs the financial fraud identification model and enterprise performance assessment model based on machine learning, respectively. Combined with big data technology for empirical analysis respectively, the following conclusions are drawn:

1) After the introduction of textual metrics constructed based on big data technology, the recognition accuracies of the financial fraud recognition models based on BPNN, SVM, RF, and modified XGBoost improved by 1.2%, 1.07%, 1.72%, and 2.09%, respectively. And the financial fraud recognition model based on modified XGBoost has the largest accuracy and AUC value, and the smallest type I error and type II error compared with other methods. This shows that big data technology combined with the modified XGBoost algorithm can improve the accuracy of financial fraud identification, and thus promote corporate financial transparency.

2) The classification accuracy of the enterprise performance evaluation model combined with the use of big data technology and nonlinear baryonic support vector machine reaches 90%, which is extremely close to manual evaluation. The method can effectively categorize enterprise performance in practical application.

References

[1] Sari, M., Irfan, I., Jufrizen, J., & Deli, L. (2020). The Testing Model of Financial Management Ability of Small and Medium Enterprises (SMEs). Jurnal Reviu Akuntansi dan Keuangan, 10(3), 584-601.

[2] Liu, Y. (2022, August). Research on the Application of Big Data Technology in Enterprise Financial Decision Making. In Proceedings of the 6th International Conference on Big Data Research (pp. 55-59).

[3] Rauf, M. A., Shorna, S. A., Joy, Z. H., & Rahman, M. M. (2024). DATA-DRIVEN TRANSFORMATION: OPTIMIZING ENTERPRISE FINANCIAL MANAGEMENT AND DECISION-MAKING WITH BIG DATA. Academic Journal on Business Administration, Innovation & Sustainability, 4(2), 94-106.

[4] Rushchyshyn, N., Nikonenko, U., & Kostak, Z. (2017). Formation of financial security of the enterprise based on strategic planning. Baltic Journal of Economic Studies, 3(4), 231-237.

[5] Wei, R., & Yao, S. (2021). Enterprise financial risk identification and information security management and control in big data environment. Mobile Information Systems, 2021(1), 7188327.

[6] Xiong, L. (2021, April). Enterprise financial shared service platform based on big Data. In Journal of Physics: Conference Series (Vol. 1852, No. 3, p. 032004). IOP Publishing.

[7] Gao, Y., Wang, Z., Wang, K., Zhang, R., & Lu, Y. (2023). Effect of big data on enterprise financialization: Evidence from China's SMEs. Technology in Society, 75, 102351.

[8] Liu, L., & Li, L. (2020, November). Innovation of Enterprise Financial Management Model Under the Background of Big Data. In 2020 International Conference on Social Sciences and Big Data Application (ICSSBDA 2020) (pp. 307-312). Atlantis Press.

[9] Liu, H., & Sun, G. (2021). Research on the application of big data technology in the integration of enterprise business and finance. Journal on Big Data, 3(4), 175.

[10] Karadag, H. (2017). The impact of industry, firm age and education level on financial management performance in small and medium-sized enterprises (SMEs) evidence from Turkey. Journal of Entrepreneurship in emerging economies, 9(3), 300-314.

[11] Feng, Q., Chen, H., & Jiang, R. (2021). Analysis of early warning of corporate financial risk via deep learning artificial neural network. Microprocessors and Microsystems, 87, 104387.

[12] Yang, W., Zhou, Y., Xu, W., & Tang, K. (2022). Evaluate the sustainable reuse strategy of the corporate financial management based on the big data model. Journal of Enterprise Information Management, 35(4/5), 1185-1201.

[13] 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.

[14] Zheng, M. (2022). Advanced artificial intelligence model for financial accounting transformation based on machine learning and enterprise unstructured text data. Mobile Information Systems, 2022(1), 5708652.

[15] Zhang, Z., Wang, Z., & Cai, L. (2025). Predicting financial fraud in Chinese listed companies: An enterprise portrait and machine learning approach. Pacific-Basin Finance Journal, 90, 102665.

[16] Huang, B., Wei, J., Tang, Y., & Liu, C. (2021). Enterprise risk assessment based on machine learning. Computational Intelligence and Neuroscience, 2021(1), 6049195.

[17] Ismail, M. M., & Haq, M. A. (2024). Enhancing Enterprise Financial Fraud Detection using Machine Learning. Engineering, Technology & Applied Science Research, 14(4), 14854-14861.

[18] Cheng, J. (2023, June). Enterprise Accounting Decision Support System Based on Deep Learning Algorithm. In International Conference on Artificial Intelligence and Communication Technology (pp. 37-46). Singapore: Springer Nature Singapore.

[19] Wang, X., Luo, X., & Hu, Y. (2021, September). Enterprise accounting and financial risk analysis system based on decision tree and SVM. In 2021 4th International Conference on Information Systems and Computer Aided Education (pp. 2015-2018).

[20] Dong, J. (2018). Application research of artificial intelligence technology in enterprise financial management. In Proceedings of the 2018 International Conference on Economics, Finance and Business Development (ICEFBD).

[21] Ren, S. (2022). Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data. Journal of Mathematics, 2022(1), 1708506.

[22] Gao, J. (2022). Analysis of enterprise financial accounting information management from the perspective of big data. International Journal of Science and Research (IJSR), 11(5), 1272-1276.

[23] Shang, H., Lu, D., & Zhou, Q. (2021). Early warning of enterprise finance risk of big data mining in internet of things based on fuzzy association rules. Neural Computing and Applications, 33(9), 3901-3909.

[24] 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.

[25] Peng, C., & Tian, G. (2023). Intelligent auditing techniques for enterprise finance. Journal of Intelligent Systems, 32(1), 20230011.

[26] Jieqi Liu. (2024). Construction of a geological identification model for small faults in coal fields based on Bayes-XGBoost. Journal of Biotech Research196-208.

[27] Yang JinRong,Chen Qiang,Wang Hao,Hu XuYang,Guo YaMin & Chen JianZhong. (2022). Reliable CA-(Q)SAR generation based on entropy weight optimized by grid search and correction factors. Computers in biology and medicine105573-105573.

[28] Thakur Ujwala,Vidyarthi Ankit & Prajapati Amarjeet. (2024). A Bilateral Assessment of Human Activities Using PSO-Based Feature Optimization and Non-linear Multi-task Least Squares Twin Support Vector Machine. SN Computer Science(3).