Financial Fraud Prediction in Chinese Growth Enterprise Board Listed Companies

--Based on the Machine Learning Experience of GWO+XGBoost

Xi Chen

Tianjin Foreign Studies University, Tianjin, 300011, China

18812373931@163.com

Abstract. In recent years, there have been frequent incidents of financial fraud in China's A-share GEM-listed companies. In order to effectively detect instances of financial fraud, this study focuses on 980 listed companies in China's Growth Enterprise Market (GEM) and utilizes the GWO+XGBoost algorithm to develop a predictive model for identifying such fraudulent activities. This study incorporates both financial and non-financial information from the companies. The empirical studies reveal that machine learning-based models such as SVM and XGBoost exhibit superior predictive performance compared to traditional statistical methods, including Naive Bayes and Logistic regression.; The GWO+XGBoost model outperforms other machine learning models in terms of Precision, Recall, F_1 and AUC; The incorporation of non-financial information indicators, such as corporate governance and audit information, significantly enhances the predictive accuracy of the model, underscoring the efficacy of non-financial information in providing valuable incremental information content for financial fraud prediction.; The study also employs Shapley's value method to examine the contribution of characteristic variables in predicting financial fraud. This analysis provides valuable decision-making guidance for auditors, investors, and regulators, helping to reduce information asymmetry in the capital market and enhance resource allocation efficiency.

Keywords: Financial Fraud; GWO+XGBoost ; Machine Learning; Financial Characteristics Variables; Non-financial Characteristic Variables.

1 Introduction

Financial fraud refers to the conduct of a company during the process of disclos-ing its financial information to external parties, resulting in materially misleading financial reports due to subjective and objective factors. This behavior signifi-cantly impacts the decision-making judgment of third parties[1]. In China, there has been an increasing trend in both the frequency and amount of financial fraud cases, which has raised concerns among investors, auditors, and regulators [2-4]. Since China's A-share market GEM listing began just over ten years ago with a relatively short development time and an

Z. Wang et al. (eds.), *Proceedings of the 4th International Conference on Economic Development and Business Culture (ICEDBC 2024)*, Advances in Economics, Business and Management Research 299,
https://doi.org/10.2991/978-94-6463-538-6_23

immature market system, major in-stances of financial fraud have occurred in companies such as Kerong Environ-ment Technology Co, Firstar Panel Technology Co, LeEco, etc. Financial fraud not only hampers a company's own development but also causes losses for third parties while severely disrupting the functioning order of securities markets.

The report of the 20th CPC National Congress emphasizes the necessity to en-hance and refine modern financial supervision, subject all types of financial ac-tivities to legal supervision. The Ministry of Finance has issued Cai Kuai Docu-ment No. 28 of 2022, which mandates heightened attention towards areas that have been prone to and highly susceptible to financial fraud in recent years, as well as effective response measures to potential fraud risks. The importance of constructing an effective fraud prediction model to prevent and respond to po-tential fraud risks and better serve the development of the real economy has be-come more and more prominent in the current environment.

Currently, research on financial fraud by domestic and international scholars primarily relies on case studies and theoretical analyses.There is a lack of litera-ture focusing on the utilization of quantitative models for proactive prediction of financial fraud. Furthermore, most quantitative analyses only consider company financial data and employ traditional statistical models for analysis, neglecting the incorporation and discussion of non-financial information.Therefore, there is still room for improvement in enhancing the accuracy of existing financial fraud prediction models.

In the 1960s, Beaver used a single financial variable to predict the financial distress of a company[5], while Altman used multiple linear discriminant analysis to construct a Z-Score model with multiple financial variables, which analysed the position of the decision point to determine whether the company was in fi-nancial distress, changing the single variable past[6]. After that, many scholars constructed M-score model[7] and F_1 model[8] based on the historical data of finan-cial indicators. Domestic scholars have also used linear discriminant analysis model and logistic model to carry out researches on the prediction of financial distress in Chinese companies [9-11], The aforementioned models demonstrate the existence of a certain level of explanatory power in financial data for detecting financial malpractice. However, the significance of non-financial indicators, such as corporate governance and textual information, can easily be overlooked when identifying instances of financial malpractice. Moreover, most studies on these variables primarily employ linear analyses, which typically yield low performance in terms of identification accuracy. Additionally, the evaluation index used to measure model performance is often limited to accuracy rate, failing to comprehensively assess issues related to imbalances between positive and nega-tive samples in detecting financial fraud.

In recent years, scholars have employed machine learning techniques to con-duct research on financial and fiscal forecasting, thereby affirming the evident advantages of machine learning over traditional models for financial identifica-tion. Firstly, it has been demonstrated that machine learning has demonstrated its proficiency in extracting non-linear features from data, leading to an enhanced accuracy rate[12]. Many scholars use a single machine learning for financial fraud identification, and Zhou Weihua[13] and Zhang Qinglong[3] have selected non-financial data related to the degree of govern-

ance of listed companies as well as financial data including company profitability, solvency, operating ca-pacity, growth capacity, and cash-generating capacity for financial risk identifi-cation, all of which improve the accuracy of risk prediction. Among many main-stream machine learning models, The XGBoost model proposed by Chen and Guestrin in 2016, has gained widespread adoption in the field of risk prediction due to its superior performance compared to traditional models in classification and prediction tasks[14]. Wu Shinong et al. employed a comprehensive characteris-tic variables from financial information, internal and external corporate govern-ance, and macro and micro to analyze the early warning of corporate financial fraud, and the empirical results showed that the combined model of XGBoost model is superior to other models[15]. GWO algorithm is an efficient new swarm intelligence optimization algorithm with simple structure, generalizability and high accuracy[16-17]. The study of Xiao Yanli and Xiang Youtao (2021) revealed that the GWO-XGBoost model exhibits superior performance in terms of predic-tion accuracy, stability and statistical significance compared with Support Vector Machine(SVM), K-Nearest Neighbours(KNN), Decision Tree, LDA, and RF, and it has important applications in data prediction and decision making[18]. There-fore, Previous studies have proposed research ideas for this paper, suggesting that the utilization of intelligent optimization algorithms can be considered to facilitate model selection and tuning, thereby enhancing the accuracy of predic-tion results during predictive tasks.

Furthermore, machine learning algorithms offer a proficient solution to ad-dress the issue of data imbalance. For instance, the SMOTE method proposed by Chawla et al. in 2002 generates multiple sub-datasets through oversampling, trains distinct models in separate sessions, and ultimately evaluates the model performance on real data [19]. The proposed method effectively leverages all available data and eliminates subjective selection of matching samples, thereby minimizing the discrepancy between the training samples and the actual sample distribution. Consequently, it significantly mitigates the risk of model overfitting [20]. In comparison to conventional models, the SMOTE method enables a more objective matching of approximate samples, thereby preventing detachment of the model from real-world data samples and enhancing the predictive accuracy of the model.

Additionally , The application of machine learning in assessing the risk of fi-nancial fraud necessitates attention not only to enhancing prediction accuracy but also to ensuring the interpretability of model results, thereby addressing the challenge posed by the “black box” problem. In recent years, Shapley Additive exPlanations (SHAP) values, derived from the concept of Shapley values, have b have gained significant traction in elucidating intricate models[21]. SHAP values can be utilized to quantify the contribution of each attribute towards classifica-tion predictions, thereby elucidating the prediction outcomes of intricate models. By visually representing the SHAP values and their rankings for diverse attrib-utes, it can establish a theoretical foundation for further investigating the causal relationships between different variables and financial fraud.

This study focuses on 980 listed companies in China's Growth Enterprise Market (GEM) and uses the GWO+XGBoost algorithm to construct a financial fraud prediction model by utilizing both disclosed financial and non-financial information as feature

variables. Conventional statistical models (Logistic and Naive Bayes) as well as machine learning techniques such as Support Vector Machines, Extreme Gradient Boosting Model, and the optimized GWO+XGBoost prediction model are constructed respectively. The impact of incorporating non-financial characteristics on predicting corporate fraud is compared, and the predictive accuracy of traditional statistical models and machine learning models is assessed to optimize the financial fraud prediction model. Lastly, the SHAP value method is used to visualize and explain feature variables.

Based on the literature reviewed, this study contributes significantly in three aspects: (1) It enriches existing research on financial fraud prediction of GEM-listed companies by demonstrating the effectiveness of the GWO-XGBoost combined model. (2) By comparing it with various other approaches, this paper confirms the outstanding performance of the GWO-XGBoost model in handling imbalanced samples and predicting corporate financial fraud risk. It also empha-sizes the importance of non-financial information in enhancing prediction accu-racy. (3) Additionally, this paper employs the SHAP value method to analyze and discuss feature variables within the GWO-XGBoost model, deepening our under-standing of key factors influencing corporate financial fraud in China through visual representation.

2 The Design of the Empirical Study

2.1 The Selection of Characteristic Variables

Theoretically, machine learning models in a big data environment can have an abundance of feature variables; however, an excessive number of features may lead to a "dimensional catastrophe" and increase the risk of overfitting. Review-ing existing literature shows that selecting appropriate features requires domain knowledge and statistical properties In this study, based on previous works [3,13], the selection of characteristic variables primarily focuses on two aspects: non-financial information and financial feature types. Financial feature types are chosen from six dimensions encompassing solvency, cash flow ability, profita-bility, development ability, operating ability and asset-liability structure within the company's fundamental framework consisting of thirteen indicators. Non-financial feature types include corporate governance environment totaling six variables. Detailed information is presented in the table 1.

Table 1. Indicators of Early Warning Characteristics of Corporate Financial Fraud Risk

Feature Type	Feature Category	Variable Category	Variable Meaning
Non-financial Characteristic Variables	Corporate Governance	duarbc	The chairman and general manager are the same person.
		indpr	The ratio of independent directors
		jsgm	The number of supervisory board members in the annual report
		tetn	The total number of senior executives in the annual report
		exctn	The total remuneration of senior management
		audf	The company's fees for audit services

Financial Characteristics Variables	Solvency	cr	(Cash and Cash equivalents) / Total current liabilities
		em	Total assets/Shareholders' equity
	Cash Flow Ability	eeoi	Increase in value of operating indices over the last year
		fcf	Operating cash flow - Capital expenditure
		fcfet	Cash flows that the company provides to equity investors
	Profitability	roe	Net Income/Shareholders' Equity
		roa	Net Income/Total Assets
	Develop-ment Ability	tagr	Ending total assets−Beginning total assets /Beginning total assets
		orgro	Operating profit growth rate over last year's growth value
		droe	Growth rate of return on net assets for the previous year
		iar	Value of intangible assets/Total assets
	Operating Ability	tat	Net sales/Average total assets
	Asset-Lia-bility Struc-ture	wcr	Current assets/Current liabilities

2.2 The Selection and Process of Data

The data in this paper is sourced from the China Stock Market & Accounting Research (CSMAR) database, which focuses on China's A-share market GEM companies as the research subject. The observation period for the sample of company financial fraud spans from 2009 to 2022, and a total of 980 GEM companies were collected. The CSMAR database enables multi-table join queries through cross-table unions, specifically selecting 'fictitious profit' and 'fictitious assets' from the 'irregularities' and 'financial indicators' libraries, as well as 'false assets' and 'false records' from the 'violations' and 'financial indicators' libraries to identify instances of fraud.

(1)Remove Financial companies and vacant values.(2) The data is backdated to reflect the most recent year when penalized companies committed financial fraud, ensuring uniqueness despite potential multiple violations within a single year due to delayed penalties. (4)Through the steps above, a total of 4,745 samples were collected - comprising 241 fraud samples and 4,501 non-fraud samples ,resulting in an extremely imbalanced positive-to-negative sample ratio of 1:20.

2.3 The Design of Research Models

2.3.1 The XGBoost Model.

The XGBoost (Extreme Gradient Boosting) algorithm is a powerful ensemble learning technique that combines the Gradient Boosting algorithm with a decision tree model. It is extensively utilized in regression and classification problems to enhance overall performance by iteratively constructing an ensemble of weak classifiers and integrating their predictions.

The study of financial fraud involves two potential scenarios for listed company i: the occurrence of financial fraud and non-financial fraud. The sample dataset of GEM listed companies comprises multiple data points, each divided into x_i and y_i. Here, x_i represents various characteristic variables of compan y_i, while y_i denotes the output

(0 or 1) from the XGBoost model. Specifically, y_i is assigned a value of 1 when the company has engaged in financial fraud and 0 when it has not.

Let the data sample be $S = (x_i, y_i), i = 1,2,3,\cdots,n, x_i \in R^m$, $y_i \in R$, where m is the dimension of the data sample, and n is the number of samples. Assuming that there are h decision trees (h=1,2,3...t), the loss function is defined as follows:

$$\text{Objective}^{(t)} = \sum_{i=1}^{n} l\left(y_i, y_i^{(t)}\right) + \sum_{h=1}^{t} \Omega(f_h) \tag{1}$$

$$\text{Objective}^{(t)} \approx \sum_{i=1}^{n} \left[\frac{1}{2} w_i f_t^2(x_i) + g_j f_t(x_i)\right] + \Omega(f_t) + C \tag{2}$$

The g_i and w_i in the Taylor expansion are defined as $\alpha_{\hat{y}(t-1)} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $\alpha^2_{\hat{y}(t-1)} l\left(y_i, \hat{y}^{(t-1)}\right)$, respectively. Upon solving the model, we obtain an XGBoost-based early warning model for financial fraud risk of listed companies. Given that there are multiple parameters in the XGBoost model, selecting appropriate parameter values becomes crucial as it significantly impacts the prediction results.

2.3.2 The GWO Optimization Algorithm.

The parameter selection of the XGBoost model significantly influences the prediction results. The grey wolf algorithm serves as a heuristic optimization algorithm that achieves optimal solutions. This algorithm offers advantages such as rapid search speed, ease in obtaining global optimal solutions, and high stability. In this study, we employ the grey wolf optimization algorithm (GWO) to optimize the parameter settings for learning_rate, number of weak classifiers (n_estimators), and maximum depth (max_depth) within the XGBoost model. Subsequently, we apply the GWO-XGBoost model to predict financial fraud risk for a company. The GWO algorithm is outlined as delineating intra-pack wolf hierarchies, finding target prey and attacking target prey.The grey wolf algorithm mimics the hierarchy within a grey wolf pack, a wolf pack can be divided into 4 hierarchies, namely the first status grey wolf α, the second status grey β, the third status grey wolf δ, and the fourth status grey wolf ω. In the search optimization process, the grey wolf $\omega$is responsible for finding the path, and the 3 types of grey wolves, namely, α, β, and δ, are responsible for optimizing and updating the optimal search path, and then obtaining the general solution δbased on the hierarchical assignments, suboptimal solution β, optimal solution $\alpha$and other solutions ω.

The main implementation steps of the grey wolf algorithm:

Step 1: Encircle the prey and calculate the relative distance between the grey wolf and the target prey D. The wolf pack will first encircle the prey in the early stage of predation, and the relative distance D between an individual wolf pack and the target prey is denoted as:

$$D = |C \cdot X_P(t) - X(t)| \tag{3}$$

$$X(t+1) = X_P(t) - A \cdot D \tag{4}$$

Where t is the current iteration number; D is the distance between the grey wolf and the prey; $X(t)$ is the current grey wolf position; $X_p(t)$ is the current prey position; $X(t+1)$ is the position of the grey wolf after the next iteration; and A and C are coefficient vectors, with A representing the convergence impact factor and C being the impact factor, which are updated as follows:

$$A = 2a^* r_2 - a, a = 2 - 2t^* \frac{1}{G_{\max}} \tag{5}$$

$$C = 2 * r_1 \tag{6}$$

Where r_2 and r_1 arethe uniformly distributed random numbers between 0 and 1, a is the convergence factor, the initial value of 2, the convergence factor will be with the number of iterations from 2 linearly decreasing to 0, G_{max} is the maximum number of iterations.

Step 2: The grey wolf chases the prey and calculates the fitness value of each grey wolf individual. The top three grey wolves in terms of fitness value are denoted as α, β and δ. The grey wolves α, $\beta$and δ, which are relatively shorter from the target prey, lead the three grey wolves to drive the remaining grey wolves ω to approach the target prey when rounding up the prey, and at the same time, update the position movement of the remaining grey wolf according to the remaining grey wolf's position information, and capture the target prey in accordance with $X_p(t+1)$.

$$D_\alpha = |C_1 \cdot X_\alpha(t) - X(t)|, X_1 = X_\alpha - A_1 * D_\alpha \tag{7}$$

$$D_\beta = |C_2 \cdot X_\beta(t) - X(t)|, X_2 = X_\beta - A_2 * D_\beta \tag{8}$$

$$D_\delta = |C_3 \cdot X_\delta(t) - X(t)|, X_3 = X_\delta - A_3 * D_\delta \tag{9}$$

$$X_p(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{10}$$

Step 3: The wolves attack the prey and capture the prey, and calculate the optimal solution. Judge whether the result meets the requirement, if it meets the set value, then the algorithm runs to the end and outputs the optimal solution; otherwise, return to Step 2 to continue to search for the optimal calculation.

2.3.3 The SMOTE Sampling Model.

The SMOTE technique, an oversampling method, enhances the classification performance of a learning algorithm by augmenting the number of samples in the minority class through the generation of synthetic instances interpolated between existing minority class samples. Instead of merely duplicating minority class samples, this approach analyzes the feature space occupied by these samples and subsequently generates novel synthetic instances within the linear subspace connecting them. This enables the model to capture additional information pertaining to the often overlooked minority classes, thereby aiming to improve classification accuracy on imbalanced datasets and foster more robust models for predicting minority class outcomes.

3 The Analysis of the Empirical Results

3.1 The Indicators of Evaluation

Since the corporate fraud prediction problem is essentially a binary classification problem, this paper uses Precision, Recall, F_1, and AUC values, which are commonly used in unbalanced binary classification problems, as the evaluation metrics of the model. For each test sample, the model has four possible prediction results, as shown in the table 2:

Table 2. Mixed matrices of financial fraud discriminatory results at the disaggregated level

Actual results	Projected results		
	Projected fraud	Projected non-fraud	Total
Actual fraud	TP	FN	TP+FN
Actual non-fraud	FP	TN	FP+TN
Total	TP+FP	FN+TN	TP+FP+FN+TN

The names and formulas of the evaluation metrics are presented in Table 3. In the context of this study's financial fraud prediction problem, real-world interest lies more in accurately identifying non-fraudulent firms among the samples predicted as non-fraud, rather than simply predicting a high number of non-fraud samples correctly. Similarly, it is important to identify truly fraudulent firms among the samples predicted as fraudulent. This aspect is captured by recall, which emphasizes correctness in predicted outcomes. Precision, on the other hand, tends to vary inversely under similar conditions compared to recall. To strike a balance between these two indicators, F_1 serves as an average measure that comprehensively evaluates precision and recall together. It particularly suits cases with category imbalance and provides a comprehensive assessment of their balance. Another performance indicator for model evaluation is the AUC value - representing the area enclosed by the ROC curve and axes - where ROC curve ranks predictions based on financial fraud identification model probabilities in descending order while cumulatively calculating false positive rate and true rate of the model to obtain an upward sloping ROC curve. A higher AUC value suggests better classification effectiveness of the model overall. Therefore, when both F_1 value and AUC value are higher for a given model, it signifies its enhanced capability in recognizing diverse samples along with superior overall performance.

Table 3. Model performance evaluation metrics

Name of evaluation indicator	Meaning of evaluation indicator
Recall	$Recall = \frac{TP}{TP + FN}$
Precision	$Precision = \frac{TP}{TP + FP}$
Composite indicator 1:F_1	$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
Composite indicator 2	AUC is the area enclosed by the ROC and the coordinate axes

3.2 The Analysis of Model Results

3.2.1 The Introduction of SMOTE Sampling Algorithm.

After applying the SMOTE sampling algorithm, the initial imbalanced distribution of positive and negative data samples was rectified to achieve a balanced classification, with an equal ratio of financial fraud samples to non-fraud samples at 1:1. Specifically, both non-fraud and fraud samples were adjusted to 3622 each. By employing the SMOTE algorithm for data rebalancing, the model's performance remains uncompromised while effectively enhancing its capability in identifying rare instances (i.e., fraud cases) and mitigating bias. Consequently, the model exhibits improved accuracy in distinguishing between positive and negative cases, thereby increasing its practical utility.

3.2.2 The Analysis of the Performance of Models.

This paper compares and analyses two classical statistical models and three machine learning models, which include Logistic, Naive Bayes and SVM, XGBoost, and GWO-XGBoost. It is evident from the ROC curves that regardless of whether based on financial characteristic variables or the inclusion of corporate governance characteristic side variables, the traditional statistical approach in Logistic regression outperforms in terms of financial fraud prediction performance. However, it falls short compared to machine learning-based models. Therefore, machine learning-based models leverage data more effectively for better fitting and prediction accuracy while enhancing model identification performance.

Table 4. Financial fraud modelling predictions

	Models	Precision	Recall	F_1	AUC
Financial Variables	Bayes	0.050	0.556	0.092	0.540
	Logistic	0.062	0.741	0.114	0.670
	SVM	0.066	0.611	0.119	0.660
	XGBoost	0.131	0.500	0.208	0.740
	GWO+XGBoost	0.368	0.519	0.431	0.790
All Financial Variables	Bayes	0.070	0.712	0.127	0.660
	Logistic	0.072	0.797	0.132	0.700
	SVM	0.095	0.678	0.167	0.720
	XGBoost	0.333	0.407	0.366	0.810
	GWO+XGBoost	0.690	0.592	0.574	0.860

In conjunction with Table 4 and Figure 1, it is evident that the XGBoost model outperforms the other three models in terms of both financial features and all features. The GWO algorithm optimizes the accuracy of the XGBoost model by determining the optimal number of weak classifiers, learning rate, and maximum depth. Following GWO algorithm optimization, significant enhancements are observed in Precision, Recall, F_1 score, and AUC value for the XGBoost model based on financial features. Moreover, when considering all feature variables (i.e., both financial and non-financial), the optimized GWO-XGBoost model exhibits improved performance across all four metrics compared to default alternative models. For this comprehensive set of feature variables, the optimal parameter configuration for the GWO-XGBoost model is determined as n_estimators=2, learning_rate=1, and max_depth=0.

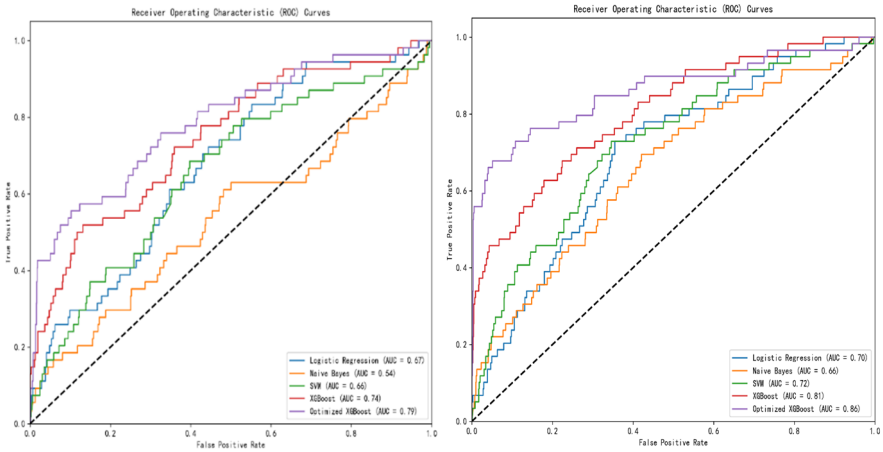

Fig. 1. ROC curves of financial fraud prediction models for financial variables (left) and all variables (right)

The GWO algorithm optimizes the XGBoost model by prioritizing precision rate as the optimization goal, despite dealing with highly unbalanced data where the number of financial fraud samples is small. Consequently, an increase in precision rate may lead to misclassification of a few fraud samples, resulting in a smaller or slightly reduced increase in recall. However, it is important to note that the optimized model exhibits higher values for F_1 and AUC. This can be attributed to the fact that F_1 combines comprehensive metrics provided by Precision and Recall, enabling an objective assessment of the model's performance. Additionally, AUC metric considers both positive and negative sample classification abilities simultaneously, making it suitable for evaluating classifiers with imbalanced samples. Thus, when a model demonstrates higher values for F_1 and AUC, it signifies stronger ability to identify different samples and overall improved performance. Based on this analysis using data from this paper, optimizing the XGBoost model through GWO algorithm enhances its performance.

3.2.3 The Analysis of the Importance of Features.

The Shapley Value method is employed to allocate gains among participants in a cooperative game, quantifying the individual contributions towards overall cooperation. In the context of machine learning, features can be regarded as participants and prediction outcomes as gains derived from this cooperative game. SHAP values, inspired by the concept of Shapley Value, enable the assessment of each feature's contribution to a specific prediction outcome. As depicted in Figure 2 below, the top five influential features according to SHAP value ranking are: the number of supervisors(jsgm), the total audit fee （audf）, the net profit margin of total assets(roa), the number of senior managers(tetn), and growth rate of total assets(tagr). Notably, more than half of these non-financial feature variables highlight their significant impact on financial fraud predictions. Among these characteristics mentioned above, two stand out as particularly crucial: number of supervisors and total audit fees.

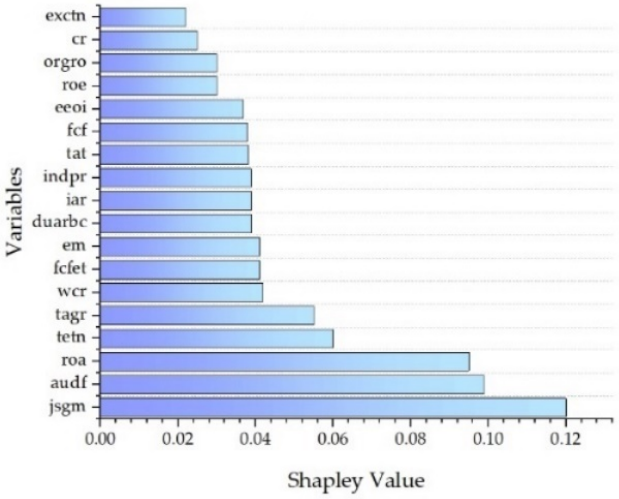

Fig. 2. Ranking of characteristic indicator SHAP values

The number of supervisors ranked first in terms of contribution among the indicators of important characteristics of financial fraud in GEM-listed companies, i.e., the number of supervisors can seriously affect the occurrence of financial fraud in a company. Supervisory committee is an important organ of Chinese companies under the Company Law, and its main function is to supervise whether the board of directors and the management are diligent in their duties. The number of supervisors in this paper is derived from the total number of supervisory board members at the end of the year as disclosed in the annual report. Due to the limited human resources of most GEM listed companies, shareholders are often concurrently directors or executives of the company, and do not pay attention to the positions and functions of the supervisory board, thus resulting in a relative weakening of the function of the supervisory board to supervise the board of directors and the managerial layer, and the number of supervisors reflects the efficiency of the decision-making process and the effectiveness of the supervision of management. The small number of supervisors and weakened functions of the supervisory board may result in internal control overriding the supervision of the supervisory board of GEM companies, reducing the effectiveness of supervision and triggering financial fraud. Therefore, the smaller the number of supervisors and the weaker the function of the supervisory board, the higher the possibility of financial fraud.

For GEM-listed companies, audit fees constitute the second most significant determinant in predicting financial fraud. In auditing practice, auditors need to evaluate the risk of financial fraud associated with a business based on the concept of risk-based auditing. When an audit client presents a high risk of financial fraud, due to information asymmetry issues, auditors tend to increase their proposed project fees as compensation for assuming such risks. Consequently, companies opt to reduce other expenditures including internal control and monitoring mechanisms while allocating more resources

towards the audit process in order to cooperate with auditors. This inadvertently elevates the risk of financial fraud and delays its disclosure. Therefore, it can be concluded that higher audit fees are indicative of a greater likelihood of financial fraud within a company.

4 Conclusions of the Empirical Study

This study focuses on a sample of 980 listed companies in China's Growth Enterprise Market (GEM) from 2009 to 2022, selecting financial characteristics (solvency, cash flow ability, profitability, development ability, operating ability, asset-liability structure) and non-financial characteristics (corporate governance, auditing information) to construct a GWO-XGBoost-based model for predicting financial fraud and conducting an analysis.

The empirical results demonstrate that the GWO-XGBoost financial fraud prediction model employed in this study surpasses traditional statistical models such as plain Bayes and logistic regression models, exhibiting superior generalization ability, higher prediction accuracy, and stronger stability, it effectively anticipates corporate financial fraud.

Additionally, the incorporation of non-financial characteristic variables in this research introduces incremental information to corporate financial fraud detection, thereby enhancing predictive capabilities.

Furthermore, the utilization of the SHAP method quantitatively illustrates the impact of corporate fraud risk while augmenting interpretability of machine learning models. This approach mitigates concerns associated with “black box” issues during default risk prediction. This study focuses on the significant impact of the top two SHAP indicators, namely audit fees and the number of supervisory board members, on financial fraud based on the results obtained from SHAP ranking.

In conclusion, this study demonstrates the significant role of introducing machine learning GWO-XGBoost in the research field of corporate financial fraud prediction. By comparing it with traditional statistical models and three other machine learning methods, our findings provide valuable insights for auditors, investors, regulators, and other stakeholders to authenticate financial performance, reduce information asymmetry in capital markets, and enhance resource allocation efficiency. Moreover, this study serves as a crucial decision-making reference and positively contributes to improving the accuracy of financial fraud prediction.

References

1. Zhang, X. Y., Shi, Y. (2023) Financial fraud recognition model based on meta-learning. J. Journal of Management Sciences, 26(10):95-113. DOI: 10.19920/j.cnki.jmsc.2023.10.006.
2. Zhang, Y., Liu, T. X., Huang, Y. J. (2023) Algorithm for Quantitative Assessment of Predictors of Financial Fraud in Listed Companies . J. Friends of Accounting, 10:117-123.

3. Zhang, Q. L., Xin, C. Y., Zhang, Y. B., et al. (2023) Research on the Analysis and Prediction of Financial Irregularities of Listed Companies ——Empirical evidence based on methods of enterprise portraits and machine learning. J. Auditing Research, 02:73-87.
4. Luo, D. L., Huang, Y. X., He, J. M. (2022) Governance of financial fraud in listed companies: empirical and theoretical analyses. J. Finance and Accounting Monthly, 22: 29-37. DOI : 10.19641/j.cnki.42-1290/f.2022.22.004.
5. Beaver, W. H. Empirical Research in Accounting: Selected Studies 1966--Discussion of Financial Ratios As Predictors of Failure. J. Journal of Accounting Research, 4:71-111. DOI : 10.2307/2490172
6. Altman, E. I. (1968) Financial Ratios, Discriminant Analysis and The Prediction of Corporate Bankruptcy. J. Journal of Finance, 23(4):589-609. https://doi.org/10.2307/2978933.
7. Beneish, M. D. (1999) The Detection of Earnings Manipulation. J. Financial Analysts Journal, 55(5): 24-36. https://doi.org/10.2469/faj.v55.n5.2296.
8. Bell, T. B. Carcello, J. V. (2000) A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. J. Auditing A Journal of Practice & Theory, 19(1):169-184. DOI: 10.2308/aud.2000.19.1.169.
9. Hong, W. Z.. Wang, X. X., Feng, H. Q. (2014) Research on the Financial Report Fraud Detection of Listed Companies based on Logistic Regression Model. J. Chinese Journal of Management Science, 22 (S1): 351-356. DOI: 10.16381/j.cnki.issn1003-207x.2014.s1.002.
10. Ye, Fan., Ye, Q. H., Huang, S. Z. (2021) Identifying and Responding to Currency Fund Fraud - A Case Study Based on Yuganite. J. Finance & Accounting, 11:37-42.
11. Zhang, Z. L., Gao, Y. (2017) Construction and Empirical Study of Financial Fraud Identification Model. J. Statistics & Decision, 09: 172-175. DOI: 10.13546/j.cnki.tjyjc.2017.09.044.
12. Hastie, T., Tibshirani, R., Friedman, J. H., et al. (2009) The Elements of Statistical Learning: Data Mining, Inference and Prediction. By. J. The Mathematical Intelligencer, 27(2): 83-85. DOI: 10.1007/BF02985802.
13. Zhou, W. H., Zhai, X. F., Tan, H. W. (2022) Research on Financial Frauds Prediction odelof Chinese Public Companie s with XGBoost. J. The Journal of Quantitative & Technical Economics, 39(07): 176-196. DOI: 10.13653/j.cnki.jqte.2022.07.009.
14. CHEN, T., GUESTRIN, 2016. Xgboost：a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. pp. 785- 794. https://doi.org/10.1145/2939672.2939785.
15. Wu, S. N., Chen, Z. Y. (2023) Bond Default Early Warning Models Based on Financial and Non－financial Information: Empirical Evidence from Machine Learning Methods. J. Journal of Xiamen University(A Quarterly for Studies in Arts & Social Sciences).
16. Mirjalili, S. (2015) How Effective is the Grey Wolf Optimizer in Training Multi-Layer Perceptrons. J. Applied Intelligence, 43(1):150-161. DOI: 10.1007/s10489-014-0645-7.
17. Rezaei, H., Chu, X. (2017) Grey wolf optimization (GWO) algorithm In: Bozorg-Haddad, O.(Eds.), Advanced Optimization by Nature -Inspired Algorithms. Springer, Singapore, pp. 81-91. https://link.springer.com/book/10.1007/978-981-10-5221-7
18. Xiao, Y. L., Xiang, Y. T. (2021) Research on Financial Frauds Prediction Model of Chinese Public Companies with XGBoost. J. Shanghai Finance, 10: 44-54. DOI:10.13910/j.cnki.shjr.2021.10.005.
19. Chawla, V. N., Bowyer, W. K., Hall, O. L., et al.(2002) SMOTE: Synthetic Minority Oversampling Technique. J. The Journal of Artificial Intelligence Research, 16(0):321-357. DOI: https://doi.org/10.1613/jair.953.

20. Wu, Y. H., Liu, X. X. (2024) Chen Yun Yan .Optimisation and Enhancement of Bond Default Early Warning Models -- A SMOTETomek-GWO-XGBoost Based Approach. J. Friends of Accounting, 06:73-81.
21. Lundberg, S., Lee, S. I. (2017) A Unified Approach to Interpreting Model Predictions. J. Advances in Neural Information Processing Systems, 30. DOI: 10.48550/arXiv.1705.07874.