Int J Syst Assur Eng Manag (December 2023) 14(6):2354–2364
https://doi.org/10.1007/s13198-023-02083-z

ORIGINAL ARTICLE

# Design of XGBoost prediction model for financial operation fraud of listed companies

**Yi Liu**[1]

Received: 6 March 2023 / Revised: 24 June 2023 / Accepted: 27 July 2023 / Published online: 16 August 2023

**Abstract** To resolve the issue of untimely discovery of fraud phenomena in the supervision process of listed companies' financial fraud, this research studies the establishment of a financial operation fraud prediction model for companies based on XGBoost and introduces regularization items and column sampling to enhance the robustness. Meanwhile, the data processing strategy of the model is designed according to the characteristics of the fraud data of listed companies, and the gray samples in the data samples are eliminated. Finally, the research uses the financial fraud prediction test method to test the model. The results indicated that the AUC of the XGBoost model designed in the study was 0.91, which was the maximum value of AUC in the comparison model. And its prediction effect reached the greatest level. The XGBoost prediction model designed by the research had a KS value of 0.65 in the prediction, which was the best value among the comparison models. This value was within the range of the KS value of the ideal model and can well distinguish positive and negative samples. The XGBoost listed company financial operation fraud prediction model designed by the research can effectively and dynamically predict financial fraud, laying a foundation for the establishment of a comprehensive capital market supervision system.

**Keywords** Financial fraud · XGBoost · Financial supervision · Chi-square test

Yi Liu
liuyi2101@yandex.com

[1] School of Economics and Management, Changsha Normal University, Changsha 410100, China

## 1 Introduction

Financial fraud is a form of corporate fraud that deliberately provides misleading financial statements in essence. The most common means of fraudulent financial statements are not to write off outdated inventory and recognize revenue before sales. With the rapid development of the domestic capital market, the market regulatory system has not been perfected with the development of the market, which has led to the formation of a gap between market behavior and market regulation. This gap has led to listed companies having more space and opportunities to engage in financial fraud. Financial fraud is repeatedly prohibited in the modern capital market, bringing serious economic and financial impacts to investors and the market. As an important venue for resource allocation in the capital market, its efficiency and smoothness are crucial for the efficiency of resource allocation (Zhang et al. 2020; Kong et al. 2019; Naumovska et al. 2020). However, financial fraud has long plagued the capital market, and financial fraud cases continue to emerge, directly threatening market confidence and healthy operation. The backwardness of regulatory systems makes it difficult to detect fraudulent activities of listed companies in a timely manner, and often fraudulent activities have already developed to a certain stage at the beginning of regulatory activities, causing certain losses (Niu et al. 2019; Ghafoor et al. 2022; Saluja and Sugiat 2023). In this environment, it is necessary to predict the financial fraud of listed companies. Listed companies not only bear market responsibility but also bear certain social responsibility. Timely investigation and supervision of fraudulent behavior is conducive to correcting the negative development of the capital market while protecting the collective interests of the public in the market (Solomon and Soltes 2021; Pradesyah et al. 2021; Suh et al. 2019). In order to prevent financial fraud and maintain the stability and healthy development of the capital market, a more robust, comprehensive, and accurate financial fraud prediction model is constructed using the XGBoost algorithm based on previous research results. This model evaluates the financial fraud risk of a certain company for a certain year and quickly determines the possibility of financial fraud in the enterprise. The research has innovatively explored research methods and paradigms, providing new ideas and references for financial fraud research.

## 2 Related work

In recent years, the research on financial fraud has gradually diversified and deepened. Wang's team designed a semi-supervised attention map neural network model, which can mark multiple views and provide a new method for financial operation fraud detection. This method was reflected in third-party user fraud prediction. It has a higher detection accuracy and can provide a safer and faster financial fraud and fraud detection method in practical applications (Wang et al. 2019). Liao's team explored the influence of corporate social duty on the issue of financial fraud. The research results showed that the social responsibility of the enterprise itself was negatively correlated with financial fraud, which indicated that the more socially responsible the enterprise was, the less likely it was to have financial fraud, and the disclosure of social duty was more helpful to avoid corporate financial fraud. At the same time, the study also found that continuous participation in socially responsible behaviors can make companies with insufficient internal control less financial fraud (Liao et al. 2019). Yiu et al. focused on the perspective of the corporate governance system in transition and designed a triangular alternative agency supervision mechanism based on the dual system logic. This supervision mechanism can effectively alleviate corporate financial fraud. The research results showed that alternative agency supervision can effectively establish a regulatory logic among the three elements of relationship governance, administrative governance, and foreign governance, and form an effective deterrent effect on corporate financial fraud (Yiu et al. 2019). Xu's team analyzed the financial supervision system that was not perfected in time under the rapid development of domestic capital. The research cited many financial fraud and fraud incidents and analyzed auditors, corporate financial management, and market supervision from the perspective of game theory and regulatory game behavior among departments. The study put forward practical suggestions for market financial regulation by establishing a payment matrix (Xu 2022). The Li team conducted a correlation analysis on the social responsibility activities of financial fraud companies and financial fraud and used the strategy of the propensity score to conduct a matching analysis of the two factors. The research results showed that the social responsibility performance scores of financial fraud companies in the fraud period were significantly promoted compared with the non-fraud period, and were higher than those of the same type of non-fraud companies during the same period (Li et al. 2021).

On the other hand, the XGBoost model has been generally applied in various fields as a computer intelligence tool. The Bi team combined the XGBoost model with different types of new sequence coding strategies and developed a machine learning-based bioinformatics discrimination technology, which can effectively distinguish chemically modified gene expression types and identify different types of outstanding features (Bi et al. 2020). Ma B et al. studied a survival forecasting model based on the improved XGBoost model, which can analyze high-dimensional genetic data. The research optimized the model by constructing a new loss function and used cross-validation to optimize the model. performance was verified. The research results showed that the model showed superior performance on different data sets (Ma et al. 2022). The Ryu team applied the XGBoost model to the dementia risk prediction model, by extracting hyperparameters and derived variables from digital medical data about dementia, to achieve the effect of using data characteristics for disease data prediction (Ma et al. 2022). The Osman team applied the XGBoost model to predict groundwater levels, which predicted groundwater levels from local rainfall, evaporation, and seasonality. The research carried out model testing based on rainfall data. The research results showed that the XGBoost model designed in the research was relatively ineffective in the first two cases when the rainfall data was delayed by 1 day, 2 days, and 3 days. The model effect of the third case was better (Osman et al. 2021). Zhou et al. combined the XGBoost model with the Bayesian algorithm and utilized the model to the prediction of the propulsion speed of the hard rock tunnel excavator. At the same time, they studied and collected 1286 groups of hard rock tunnel data, and used the data for the actual model prediction test. Among them, the cross-validation method was mainly used in the experiment to prove the function. From the results, the prediction model designed by the research can form a better prediction effect than the traditional model (Zhou et al. 2021). The XGBoost model was mostly used in the prediction of various occasions, financial cheating was a means of corporate cheating, and the prediction of corporate fraud in the supervision process can strengthen the effectiveness and timeliness of market supervision. Therefore, the study applies the XGBoost model to corporate financial supervision and forecasting to provide an effective path for market supervision of listed companies.

Firstly, previous research on financial fraud was mostly based on traditional statistical methods, lacking the means to analyze and predict using advanced computer technology and intelligent algorithms. The XGBoost model introduced in this study is a computer intelligence tool that can effectively conduct data mining and prediction in multiple fields. By combining this model, research can more accurately determine whether a company has financial fraud behavior, reduce the occurrence of false negatives and false positives, and provide an effective reference for regulatory authorities. Secondly, many studies predict the possibility of financial fraud based on the characteristics of enterprises, such as company size, debt ratio, operating cash flow, etc. However, these factors may not be sufficient to capture anomalies in the financial statements. This study introduces the XGBoost model and combines it with other advanced data mining technologies to analyze more financial data and reveal potential financial fraud risks. Meanwhile, previous research has often focused on a specific industry or market, lacking examples for validation on larger datasets. This study introduced a large amount of financial data, enhancing the representativeness and applicability of the research results. Finally, research on financial fraud often focuses on analysis and explanation, while there is less research on prediction. This study compensates for this by using the XGBoost model to focus on predicting the likelihood of financial fraud. It constructs a practical and effective prediction model, which fills the gaps in previous research on financial fraud prediction, data mining technology, and empirical research. It provides stronger support for preventing and monitoring financial fraud behavior and opens up new directions for research in related fields.

## 3 Design of XGBoost prediction model for financial operation fraud of listed companies

### 3.1 Design and construction of XGBoost fraud prediction model

The XGBoost algorithm is an integrated learning means based on the forward assignment algorithm to realize the additive model. The core idea of the ensemble model is to build a powerful model by building a series of weak basic models. The XGBoost model is an improvement to the GBRT (Gradient Boosted Regression Tree) model, and the serial iterative process of the GBRT model is expressed as Eq. (1)

$$\begin{cases} \hat{y}^{(0)} = 0 \\ \hat{y}^{(1)} = vf_1(x;\theta_1) = \hat{y}^{(0)} + vf_1(x;\theta_1) \\ \cdots\cdots \\ \hat{y}^{(T)} = v\sum_{j=1}^{T} f_j\left(x;\theta_j\right) = \hat{y}^{(T-1)} + vf_T(x;\theta_T) \end{cases} \tag{1}$$

In Eq. (1), $T$ represents the number of basic regression trees used for ensemble modeling; $\theta$ represents the structure of the regression tree, $\theta_j$ represents the structure of the first $j$ regression tree; $v$ represents the scaling weight coefficient, that is, the learning rate, which is used to scale a single regression tree to the ensemble; $\hat{y}^{(j)}$ indicates the forecasting result of the $j$ th regression tree; $f_j(x;\theta_j)$ indicates the outcome of the $j$ th regression tree when the scaling weight coefficient is not considered. With the continuous superposition of $j$, the residual will decrease in gradient, and the model effect will increase in gradient. The gradient boosting procedure of the GBRT model is shown in Fig. 1.

The study proposes an extended gradient boosting algorithm XGBoost, which follows the basic principles of regularized learning and integrates regularized terms into the traditional loss function of the GBRT algorithm. The basic idea of the meta-model of the commonly used XGBoost model is to expand the negative gradient of the approximate model through the second-order Taylor equation of the loss function and to give higher accuracy to the samples with lower accuracy in the previous round of regression tree training. By learning the weights, one can improve precision and enable serial iteration of multiple

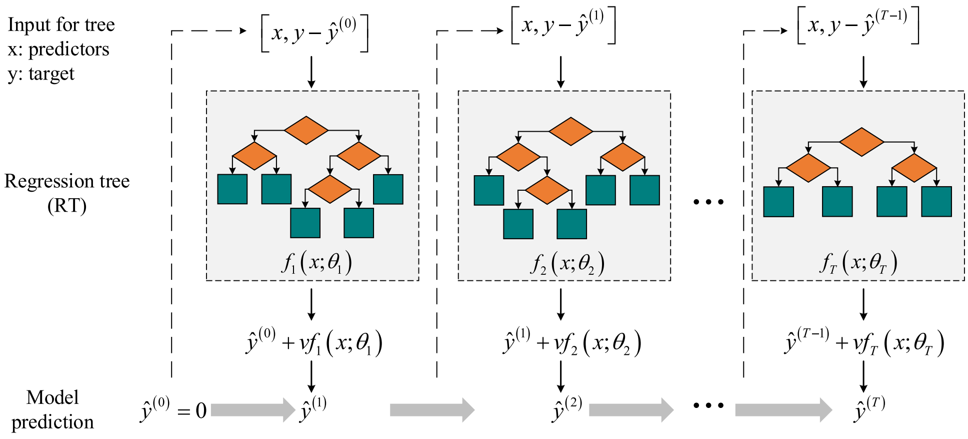


**Fig. 1** Gradient lifting process of the GBRT model

models. This allows for gradual correction of deviation until the loss satisfies the convergence condition. The core task of the gradient boosting tree model $f_j\left(x;\theta_j\right)$ is to find each optimal single regression tree $\theta_j$ and establish a decision function in the first step by minimizing the objective function $j$, and its objective function is shown in the Eq. (2).

$$\hat{\theta}_1 = \arg\min_{\theta_j}\left\{\sum_{i=1}^{N}\left[\hat{y}_i^{(j-1)} - y_i + vf_j\left(x_i;\theta_j\right)\right]^2 + \omega\left(\theta_j\right)\right\} \tag{2}$$

In Eq. (2), $N$ represents the sample size; $\omega\left(\theta_j\right)$ represents the regular term $j$ of the first regression tree, and its principle is shown in Eq. (3).

$$\omega\left(\theta_j\right) = \alpha M_j + \frac{1}{2}\lambda\left\|w_k\right\| = \alpha M_j + \frac{1}{2}\lambda\sum_{k=1}^{M_j}\left(w_k^{(j)}\right)^2 \tag{3}$$

In Eq. (3), $w_k^{(j)}$ represents the leaf score of the $k$ th leaf node of the $j$ th regression tree; $M_j$ represents the number of leaf nodes of the $j$ th regression tree; $\alpha$ represents the minimum loss of each leaf node branch added; $\lambda$ = L2 represents the regularization of the leaf score of the regression tree item. The 2nd order Taylor extension of the loss function in the XGBoost algorithm makes it more efficient to find the optimal solution. The rule items included in the objective function penalize the sophistication of every regression tree, and the model can avoid overfitting more effectively. Introducing the 2nd order Taylor extension equation of the loss function, the target function can be expressed as Eq. (4).

$$\hat{\theta}_j \approx \arg\min_{\theta_j}\left\{\begin{array}{r}\sum_{i=1}^{N}\left[L\left(y_i,\hat{y}_j^{(j-1)}\right) + vg_i^{(j)}f_j\left(x_i;\theta_j\right) + \frac{1}{2}v^2h_i^{(j)}f_j^2\left(x_i;\theta_j\right)\right] \\ + \alpha M_j + \frac{1}{2}\lambda\sum_{k=1}^{M_j}\left(w_k^{(j)}\right)^2\end{array}\right\} \tag{4}$$

In Eq. (4), $L(y,\hat{y}) = (\hat{y} - y)^2$; $g_i^{(j)}$ is expressed as Eq. (5).

$$g_i^{(j)} = \frac{\partial L\left(y_i,\hat{y}_j^{(j-1)}\right)}{\partial\hat{y}_j^{(j-1)}} = 2\left(\hat{y}_j^{(j-1)} - y_i\right) \tag{5}$$

In Eq. (4), $h_i^{(j)}$ is expressed as Eq. (6).

$$h_i^{(j)} = \frac{\partial L\left(y_i,y_j^{(j-1)}\right)}{\left(\partial\hat{y}_j^{(j-1)}\right)^2} = 2 \tag{6}$$

It can be seen from the equation that the greater the quantity of leave $M_j$ s, the greater the penalty term $\alpha$, and it is more necessary to obtain a tree with a simple structure. Since the sample set $\hat{y}^{(j-1)}\ y_i$ is fixed and determined in the first $j-1$ step, it can be regarded as offsetting the constant term $L\left(y_i,\hat{y}_j^{(j-1)}\right)$. After each optimal single regression tree is determined $\theta_j(j = 1, 2, \cdots, T)$, the XGBoost model completes the training. The XGBoost model improves the robustness of the model through the introduction of regular items and column sampling; It adopts a parallelization strategy when selecting the split point of each tree, which significantly improves the speed of the model; The tree model can be used as a meta-model through its serial results. The cross-combination between features is further realized, which has learning advantages compared to artificial feature engineering. In the financial fraud prediction scenario, the modeling process is actually to abstract a binary classification decision function from the feature data distribution, input the features and fraud labels into the XGBoost algorithm, and finally output the mapping result of the high-dimensional feature space. Model parameters need to be specified during model training, and the research realizes model automatic parameter search according to self-defined rules. It is expected that the gap between the KS value of the training set of the model and the KS value of the out-of-time sample is small to ensure the stability of the model across time, and the KS value of the out-of-time sample set needs to be large enough to ensure the high accuracy. The distribution weight of the KS value $\mu$ will be modified depending on the real condition. In situations where stability fails to meet the desired standards, adjusting the weight can help reduce the disparity between the two KS values. The study adopts a greedy search method for the target KS value, which only considers a single parameter for forward and backward search each time, and continues to search after increasing the target value KS value, otherwise stops the search in this direction.

The selection of financial fraud characteristic variables should have both domain knowledge and statistical characteristics. The derived characteristic variables can be constructed by feature combination and feature aggregation. The characteristic variables of the study are mainly selected from four aspects: corporate governance, accounting supervision, financial indicators, and business operations. The main emphasis of corporate governance lies in the equity attributes of the company, as well as the governance environment of the board of directors, board of supervisors, and management. Financial and accounting supervision focuses on internal control and external audit indicators, whereas financial indicators prioritize a company's solvency, profitability, operating ability, cash flow ability, and development potential. In terms of business operations, analysis is carried out from four different angles: sales collection, purchase payment, capital flow, and related transactions to determine whether the financial data aligns with expectations. Through comparison, trends build indicators that reflect abnormalities in business operations. The financial fraud risk index system is shown in Fig. 2.

### 3.2 XGBoost fraud prediction model data selection and processing

The financial fraud samples come from the CSMAR China listed company financial annual report database and the violation information summary table of the violation event database.

After determining the initial sample, the study processed the data set. The data were discretized using WOE encoding to avoid the influence of extreme values in the model data. WOE coding is a binning process for discrete variables, which represents the variation between the ratio of fraudulent samples in the present group to all fraudulent samples $p_{y_i}$ in the sample and the ratio of non-fraudulent samples in the present group to all non-fraudulent samples $p_{n_i}$ in the sample. The ratio of the two is expressed, and the greater the WOE, the greater the difference. The equation is as in Eq. (7).

$$woe_i = \ln\left(\frac{p_{y_i}}{p_{n_i}}\right) \tag{7}$$

Data quality and feature engineering affect the forecasting capability of the model to some degree. The study screened features from five aspects: feature missing rate, feature information, the correlation between features, feature stability PSI and recursive feature deletion. The IV value of feature information refers to the weighted summation of feature WOE codes. The IV of a single feature is equal to the accumulation corresponding to each value of the feature $iv_i$. The definition of sum refers to Eq. (7), and the IV equation is shown in Eq. (8).

$$\begin{cases} iv_i = \left(p_{y_i} - p_{n_i}\right) \times woe_i \\ IV = \sum iv_i \end{cases} \tag{8}$$

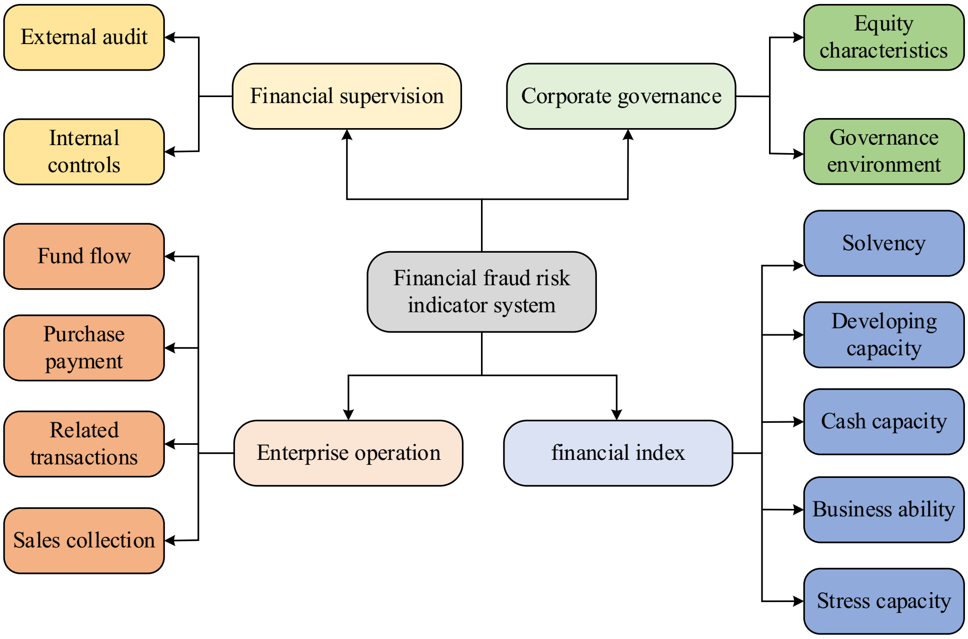
IV can reflect the contribution of a single feature to label discrimination to a certain extent. Features with IV values less than 0.02 are usually not informative and should be removed from the dataset. The study used the Spearman correlation coefficient to measure the correlation between features. The Spearman coefficient is solved according to the sorting position of the original data, as shown in Eq. (9).

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{9}$$

In Eq. (9), $n$ represents the quantity of data; $d_i$ represents the difference between two features $X$ and $Y$ data order, Features that have a relevance larger than 0.7 are deleted from the dataset by the computational study. The feature stability PSI value is to calculate the distribution difference of the same index on two different data sets, and then measure the stability of features and models. $p_t^i$ and $p_b^i$ are used to represent the proportion of the samples $i$ in the target set and the base distribution in the first bin of the population, respectively. The definition of the PSI equation is shown in Eq. (10).

$$PSI = \sum_i \left(p_t^i - p_b^i\right) * \ln\left(\frac{p_t^i}{p_b^i}\right) \tag{10}$$

**Fig. 2** Financial fraud risk indicator system



According to the equal-frequency binning strategy, the basic distribution is equally divided into 10 parts, the target distribution is binned using the same threshold, and the sum of the second bins is calculated respectively, and the obtained values are replaced with Eq. (10) for the calculation to get the stability of the two distributions The features with a PSI value less than 0.02 were removed from the data set. To strike a trade-off between the sophistication and the capability to characterize the dataset, the OLS linear regression model was used in stepwise regression, and AIC and BIC were used as the standard to weigh the estimated model complexity and fit.

$q$ is defined as the spreading probability of $A$ and $B$ of the law, it is the spread of $A$ and $B$ of the financial data of each company's balance sheet, income statement, and cash flow statement. If it is close to 0, it means that the possibility of fraudulent financial data of companies is higher. The calculation of the correlation coefficient is as Eq. (11).

$$q(A,B) = \frac{Cov(A,B)}{\sqrt{V(A)\cdot V(B)}} \tag{11}$$

In Eq. (11), $Cov(A,B)$ represents the covariance of $A$ and $B$; $V(A)$ and $V(B)$ represents the variance of $A$ and $B$. The chi-square test is used to verify the level of agreement between the assignment of sample values and the theoretical distribution of Benford's law. Taking the financial data of the balance sheet, income statement, and cash flow statement of listed companies as observation samples, the statistics $\chi^2$ are calculated and compared with the standard value. If the sample group statistics $\chi^2$ are greater than the standard value, the original hypothesis is not valid, and the possibility of financial fraud is high. The calculation of statistics $\chi^2$ is as Eq. (12).

$$\chi_n^2 = \sum\nolimits_{i=0}^{9} \frac{\left(F_n(i) - F_0(i)\right)^2}{F_0(i)} \tag{12}$$

In Eq. (12), $F_n(i)$ is the actual observed value, that is, the frequency of the sample with the digit $n$ of 0 in the whole sample $i$; $F_0(i)$ represents the Benford theoretical distribution value. The definition of the Benford risk factor includes three items. The associated factor of the first number with the law is lower than 0.9; the probability allocation of the second number "0" exceeds 0.18; the chi-square test result is "rejected".

According to the definition of the Benford risk factor, the samples that do not meet the definition in the sample are eliminated. LOF is a density-based anomaly detection method, which introduces a local reachable density method that can describe the data density, and uses this method to measure the degree of abnormality of the sample as shown in Eq. (13).

$$l_k(a) = \frac{1}{\frac{\sum_{b\in N_{k(a)}} dis_k(a,b)}{\left|N_k(p)\right|}} \tag{13}$$

In Eq. (13), $a$ and $b$ are two sample points respectively, and $dis_k(a,b)$ is the reachable distance between them. The k-nearest neighbor $N_k(a)$ of the sample point $a$ indicates that the distance between the sample point and the sample point $a$ is not larger than the distance between the first nearest point $k$ and the sample point $a$. The relative density of the sample point $a$ and its k-nearest neighbors are calculated, and then the notion of the local anomaly factor is obtained as in Eq. (14).

$$LOF_k(a) = \frac{\sum_{b\in N_{k(a)}} \frac{l(b)}{l(a)}}{\left|N_k(a)\right|} = \frac{\frac{\sum_{b\in N_{k(a)}} l(b)}{\left|N_k(a)\right|}}{l(a)} \tag{14}$$

It can be seen that the smaller the local reachability density of a sample point $a$ is compared with the mean local accessibility intensity of its k-nearest neighbors, the better the degree of anomaly. The study removes samples with an anomaly score greater than 0.7 percentile from the data sample. The IF algorithm is an integrated algorithm based on the idea of random partitioning of space. Taking two-dimensional space as an example, each coordinate axis represents a feature, as shown in Fig. 3.

Outliers $z_i$ are shown in Fig. 3, $z_0$ representing normal samples. A straight line is randomly selected to divide the space along the two coordinates, $z_0$ is the probability of being isolated $z_i$ before $z_0$ is much greater than the probability of $z_i$ being isolated before. The IF algorithm finally outputs the IF exception as Eq. (15).

$$S\left(z_i\right) = 2^{\frac{E\left(t\left(z_i\right)\right)}{C(m)}} \tag{15}$$

In Eq. (15), $E\left(t\left(z_i\right)\right)$ represents the average path length of $z_i$ on all isolated trees, $m$ is the number of training samples on one particle number, and $C(m)$ is the mean path length of the binary tree trained with samples. The value range of the IF abnormal score is [0,1]. The nearer the value is to 1, the more abnormal the data is. In this study, samples with an abnormal score greater than 0.7 percentile were removed from the data samples, and the overall samples obtained after data processing were reduced to 3842.

## 4 Analysis of prediction effect of XGBoost financial fraud prediction model

The study selected violations of "fictitious profits", "false assets" and "false records" in the processing documents issued by the regulatory agency The data of the company in the year of violation was used as the sample of financial fraud in this research. The study selected samples from 2000 to 2020 as the time observation window, split the data set into training samples and test samples based on the ratio of 7:3, and the last section of the time slice in the entire modeling sample was the out-of-time verification sample. The training samples and test samples were selected from the fraud incidents up to 2018, the samples in the financial industry were eliminated, and the initial sample distribution was finally determined as presented in Table 1.

The final model parameters are presented in Table 2.

The study used the highly operable Benford's law, Local Outlier Factor (LOF), and unsupervised learning isolation forest algorithm (Isolation Forest, IF) to eliminate gray samples in non-fraud samples. Benford's law refers to the distribution of the first digit in a large number of natural data sets. It is a common method for fraud identification. The frequency of the theoretical spread of the first and second numbers of this law was presented in Table 3.

The research compared the function of the constructed XGBoost fraud forecasting model with the commonly used Fscore model and Cscore model. To test the generalization ability and the stability of the model, the study selected out-of-time samples to verify the accuracy rate, PR curve, ROC curve, KS value, fluctuation point, and PSI.

In the binary classification problem, the overall accuracy of positive and negative sample classification is usually used to evaluate the classification effect of the classifier. TP is the number of real fraud samples that are properly forecasted as fraud samples, TP and FN are the numbers of fraud samples that are incorrectly forecasted and correctly predicted, and TN and FP are the correctly identified samples and incorrectly identified in non-fraud samples quantity. The accuracy is calculated through the confusion matrix, and the positive

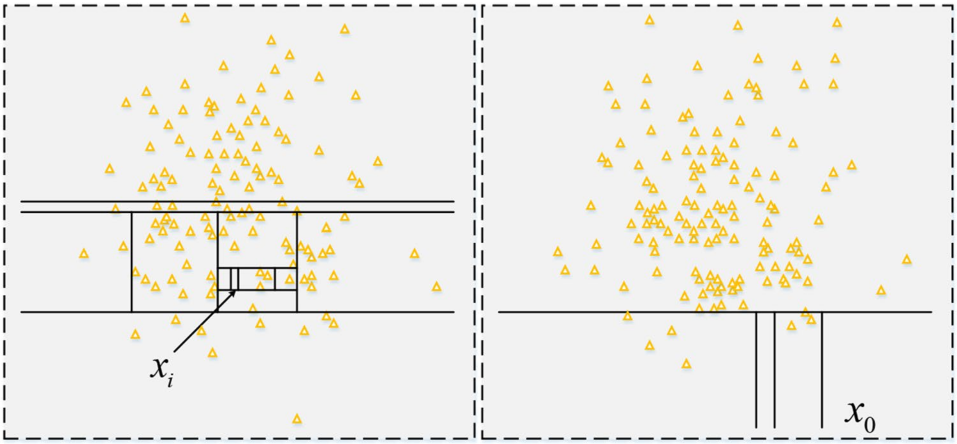


**Fig. 3** IF algorithm two-dimensional space single isolated tree

**Table 1** Initial sample distribution

| Data set | Number of samples | Fraud sample | Nonfraud sample | Proportion of fraud samples (%) |
|---|---|---|---|---|
| Training Set | 10,124 | 1841 | 8283 | 18.18 |
| Test Set | 4256 | 752 | 3504 | 17.67 |
| Out-of-time samples | 4210 | 485 | 3725 | 11.52 |
| Total | 18,590 | 3078 | 15,512 | 16.56 |

**Table 2** Model parameter

| Parameter | Parameter name | Parameter value |
|---|---|---|
| learning_rate | Learning rate | 0.05 |
| n_esti | Maximum Iterations | 200 |
| max_depth | Maximum depth of a subtree | 3 |
| min_child_w | Child node weight threshold | 1 |
| sub_sam | Sampling proportion of training samples | 0.7 |
| reg_lambda | L2 regularization coefficient | 300 |
| scale_pos_w | Adjust the weight of positive and negative samples | 1 |

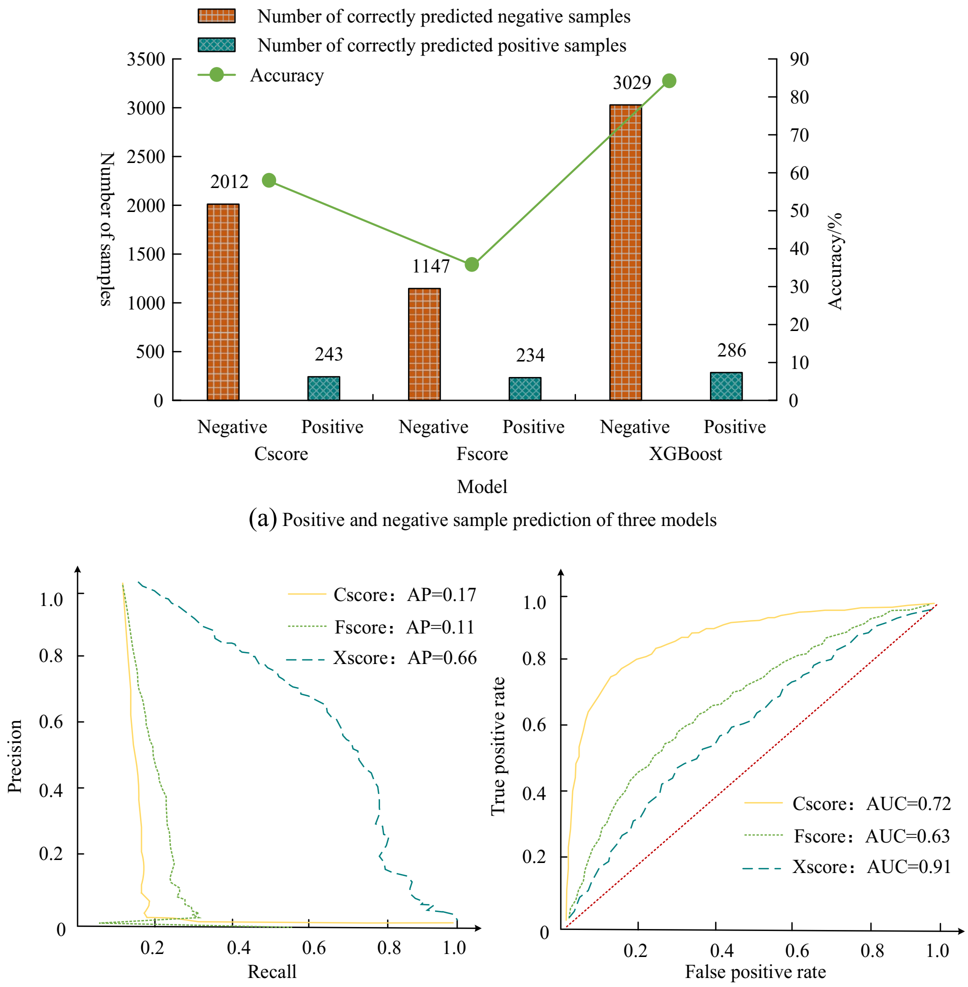
**Table 3** Theoretical frequency of distribution of first digit and second digit of Benford's law

| Digit position | First digit | Second digit | Digit position | First digit | Second digit |
|---|---|---|---|---|---|
| 0 | 0 | 0.120 | 5 | 0.079 | 0.097 |
| 1 | 0.301 | 0.114 | 6 | 0.067 | 0.093 |
| 2 | 0.176 | 0.109 | 7 | 0.058 | 0.090 |
| 3 | 0.125 | 0.104 | 8 | 0.051 | 0.088 |
| 4 | 0.097 | 0.100 | 9 | 0.046 | 0.085 |

and negative sample predictions of the three models are shown in Fig. 4.

From Fig. 4, the number of non-fraud samples correctly predicted by the Cscore model was 2012, the quantity of cheat samples properly predicted was 243, and the accuracy rate of model prediction was 58%. The quantity of non-cheat samples properly forecasted by the Fscore model was 1147, the number of fraud samples correctly predicted was 234, and the precision rate of model forecasting was 36%. The XGBoost model correctly predicted the quantity of non-cheat samples was 3029, the quantity of cheat samples was correctly predicted at 286, and the accuracy rate of model prediction was as high as 86%. Comparing the three models, the model predicted more positive and negative samples, had higher prediction accuracy, and had better performance of the model. From the above indicators, in terms of predicting non-fraudulent samples, the XGboost model had a higher recall rate, accuracy, and F1 value than other models; In terms of fraud sample prediction, the XGboost model had a higher recall rate, F1 value, and higher accuracy than the other two models. Overall, the XGBoost prediction model performed better than other models in predicting both fraudulent and non-fraudulent samples. Therefore, when conducting credit risk assessment, the XGBoost prediction model can more accurately distinguish between fraudulent and non-fraudulent samples than other models, providing more reference results for business decision-making.

**Fig. 4** Positive and negative sample prediction and ROC



(a) Positive and negative sample prediction of three models

(b) PR Curve

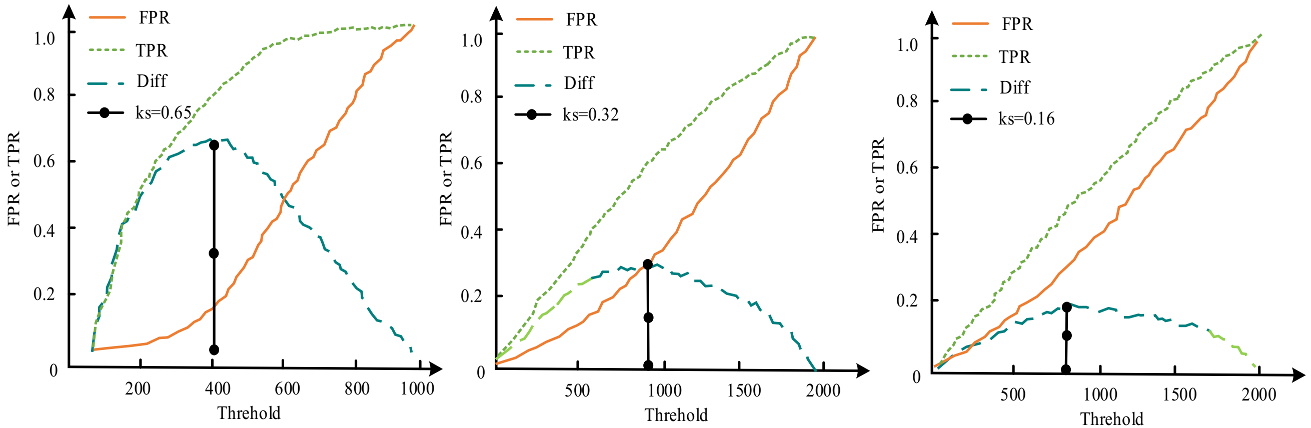
(c) ROC Curve

The PR curve can intuitively reflect the trade-off relationship between the precision rate and recall rate. Each point on the curve corresponds to the precision and recall under a classification threshold. The threshold increases, and the recall decreases. The PR curve was evaluated by computing the area AP surrounded between the PR curve and the coordinate axis. The region below the curve (AUC) was calculated to evaluate the ROC curve. The AUC of the random classifier was 0.5, and the closer the AUC was to 1, the better the classifier was. The PR curves and ROC curves of the three prediction models were shown in subgraphs (b) and (c). From subgraphs (b), it can be seen that the AP value of the Cscore model was 0.17, the AP value of the Fscore model was 0.11, and the AP value of the XGBoost prediction model proposed in the study was 0.66. By comparing the AP values of the three models, it can be concluded that the XGBoost prediction model proposed in the study had the highest AP value, indicating that the model had the best performance among the three models. In subgraphs (c), the AUC of the Cscore model was 0.72, the AUC of the Fscore model was 0.63, and the AUC of the XGBoost model proposed in the study was 0.91. By comparing the three, it can be seen that the model proposed in the study had the highest AUC and the best predictive performance. In the data of China, the ROC curve of the Fscore model showed a relatively low amplitude of change, and the AUC value was also very low, indicating that the model cannot distinguish positive and negative samples well; The ROC curve of the Cscore model had a certain upward curvature and a higher AUC value, indicating that the model performs better in data from China. In the XGBoost prediction model, the upward bending of the ROC curve was more pronounced, with an AUC value above 0.8, indicating that the model had very good performance. Overall, the XGBoost prediction model performed better than the other two models in the data of China, with an excellent ability to distinguish positive and negative samples. At the same time, the ROC curve showed an upward bending trend, and the AUC value was high, reflecting the excellent performance of the model.

The abscissa of the KS curve was the descending order of thresholds from 1 to 0, the ordinate was the variation between TPR and FPR below various thresholds, and the KS value was the maximum value of the difference. The model with a KS value between 0.3 and 0.75 was ideal, and the greater the KS value within this range, the stronger the identification ability of the model. When the KS value exceeded 0.75, the model training may have the problem of overfitting. The KS curves of the three models were shown in Fig. 5.

Analyzing Fig. 5, the Fscore model performed the worst in the data in China, with a KS value of only 0.16, indicating that the model cannot distinguish positive and negative samples; the Cscore model performed well, with a KS value of 0.32, which had a certain degree of positive and negative sample discrimination ability; The XGBoost model performed the best among the three models, with a KS value of 0.65, which was within the range of the KS value of the ideal model and had a very good ability to distinguish positive and negative samples. The fluctuation point reflects the ability of the model to sort positive and negative samples. The grouping KS value and the proportion of negative samples of the three models were shown in Fig. 6.

From Fig. 6, starting from the 7th box sample, the fraud ratio of the XGBoost fraud prediction model after grouping exceeded the fraud ratio of the 6th box sample, so the ranking fluctuation point of the model was the 7th box sample. However, the maximum group KS value of the model was taken from the sixth box sample, so it can be seen that the KS value cannot directly reflect the ranking ability of the model, and it was necessary to use the ranking fluctuation points to assist in the evaluation of the model. The more



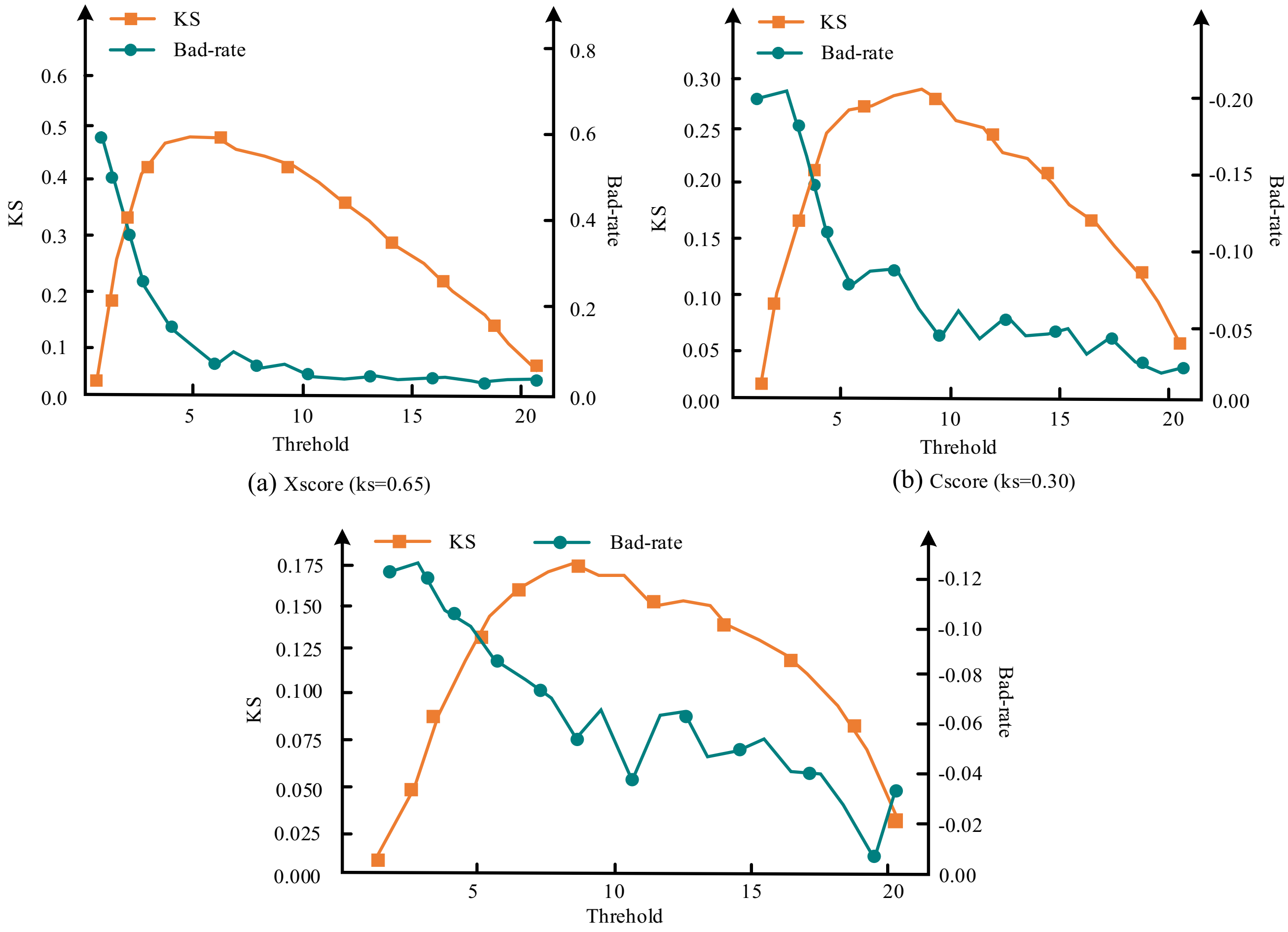
(a) KS Curve (ks=0.65) (b) KS Curve (ks=0.32) (c) KS Curve (ks=0.16)

**Fig. 5** KS curves of three models

lagging the fluctuation point of the model, the stronger the sorting capability. Compared to the Cscore model and the Fscore model, the XGBoost model proposed by the study had smaller fluctuations, a gentler overall trend, and a stronger ranking ability. The model stability index PSI can measure the distribution difference between test samples and model training samples. It was commonly considered that if the PSI was lower than 0.1, the consistency was high, if the value of the PSI was between 0.1 and 0.2, the stability of the model was average, and if the value of the PSI was larger than 0.2, the consistency was poor. After comparison, the PSI of the XGBoost model was only 0.0824, indicating that the model was extremely stable and had superior performance, while the PSI of the Cscore model and the Fscore model were both greater than 0.1, indicating that the stability of the two was not high. Compared with the XGBoost model, the consistency was worse, which indicates this model had a strong generalization ability.

## 5 Conclusion

The study applies the XGBoost model to the financial operation fraud prediction of listed companies to resolve the issue of untimely inspection of fraud. In designing the model, the regularization term and column sampling were introduced to enhance the robustness of the XGBoost prediction model. At the same time, a data processing strategy in line with the characteristics of financial fraud data was designed to process the model data, while avoiding the model from being affected by extreme data values and improving the data processing efficiency. The results indicated that the AUC value of the XGBoost financial fraud prediction model designed by the research in the prediction test was 0.91, and the KS value was 0.65, which were the maximum values among the same type of comparison models. The XGBoost financial fraud prediction model designed by the research had a relatively high performance. At the same time, the XGBoost fraud prediction model showed better stability than other models from the 7th box sample. It can be seen that the



(a) Xscore (ks=0.65)

(b) Cscore (ks=0.30)

(c) Fscore (ks=0.18)

**Fig. 6** Trend chart of grouping KS value and negative sample proportion of three models

model designed in the research had a stronger generalization ability in the actual risk fraud prediction, and can lay a technical foundation for the prediction and supervision of financial cheating of listed companies.

Firstly, research needs to focus on the predictive accuracy of the XGBoost financial fraud prediction model designed for research. Its predictive performance is better than that of similar models, which can more accurately evaluate the financial operational risks of enterprises. Research can use this model to assist in investment decision-making, thereby better avoiding potential risks and protecting research investments. Secondly, research needs to understand the financial implications of the survey results. If the survey results indicate issues with the use of funds, it will affect the research's confidence in the company's operations. Research needs to understand whether the company can manage funds correctly to ensure its financial robustness. If a company's funding management is inappropriate, research needs to be cautious about its investment potential. From the two main perspectives of financial forecasting accuracy and company funding management, it is possible to more accurately assess the potential risks of the enterprise and make more informed decisions.

**Fundings** The research is supported by: The Hunan Philosophy and Social Science Fund, Research on residents' sense of taxation burden in the background of individual income tax reform in China (No. 19YBQ008).

**Declarations**

**Conflict of interest** The authors report there is no competing interests to declare.

## References

Bi Y, Xiang D, Ge Z et al (2020) An interpretable prediction model for identifying N7-methylguanosine sites based on XGBoost and SHAP. Mol Ther-Nucleic Acids 22:362–372

Ghafoor A, Zainudin R, Mahdzan NS (2022) Factors eliciting corporate fraud in emerging markets: the case of firms subject to enforcement actions in Malaysia. Business and the Ethical Implications of Technology. Springer, Cham, pp 281–302.

Kong D, Xiang J, Zhang J et al (2019) Politically connected independent directors and corporate fraud in China. Acc Finance 58(5):1347–1383

Liao L, Chen G, Zheng D (2019) Corporate social responsibility and financial fraud: evidence from China. Account Fin 59(5):3133–3169

Li X, Kim JB, Wu H et al (2021) Corporate social responsibility and financial fraud: The moderating effects of governance and religiosity. J Bus Ethics 170(3):557–576

Ma B, Yan G, Chai B et al (2022) XGBLC: an improved survival prediction model based on XGBoost. Bioinformatics 38(2):410–418

Naumovska I, Wernicke G, Zajac EJ (2020) Last to come and last to go? The complex role of gender and ethnicity in the reputational penalties for directors linked to corporate fraud. Acad Manag J 63(3):881–902

Niu G, Yu L, Fan GZ et al (2019) Corporate fraud, risk avoidance, and housing investment in China. Emerg Mark Rev 39:18–33

Osman AIA, Ahmed AN, Chow MF et al (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J 12(2):1545–1556

Pradesyah R, Yuslem N, Batubara C (2021) Fraud In Financial Institutions. J Int Conf Proc (JICP) 4(2):341–348

Saluja S, Sugiat M (2023) Corporate fraud during COVID-19: Evidence from India and Indonesia. In: Acceleration of Digital Innovation and Technology towards Society 5.0. Routledge, pp 400–406

Solomon DH, Soltes E (2021) Is, "not guilty" the same as "innocent"? Evidence from SEC financial fraud investigations. J Empir Leg Stud 18(2):287–327

Suh JB, Nicolaides R, Trafford R (2019) The effects of reducing opportunity and fraud risk factors on the occurrence of occupational fraud in financial institutions. Int J Law Crime and Justice 56:79–88

Wang D, Lin J, Cui P, et al. (2019) A semi-supervised graph attentive network for financial fraud detection.In: 2019 IEEE International conference on data mining (ICDM). IEEE, pp 598–607

Xu H (2022) CPA audit and corporate financial fraud: an analysis based on game theory model. In: 2022 International conference on artificial intelligence, internet and digital economy (ICAID 2022). Atlantis Press, pp 996–1002.

Yiu DW, Wan WP, Xu Y (2019) Alternative governance and corporate financial fraud in transition economies: evidence from China. J Manag 45(7):2685–2720

Zhang J, Wang J, Kong D (2020) Employee treatment and corporate fraud. Econ Model 85:325–334

Zhou J, Qiu Y, Zhu S et al (2021) Estimation of the TBM advance rate under hard rock conditions using XGBoost and Bayesian optimization. Undergr Space 6(5):506–515