Innovative Design and Intelligent Manufacturing
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA241178

Evaluating the Effectiveness of Machine Learning Algorithms for Financial Fraud Detection

Tianbao LI[a] and Jingbang ZHOU[b,1]
[a]*Fuyang Normal University, Fuyang 236037, China*
[b]*Hefei University of Economics, Hefei 230012, China*

Abstract. In order to study the problem of financial statement fraud identification, the evaluation of the application effect of machine learning algorithms in financial fraud detection is proposed. Taking the financial statements of Shenzhen and Shanghai A-share listed companies from 2011 to 2020 as sample data, the information value was introduced to build an indicator screening model, and 17 financial variables and 4 non-financial variables were extracted. After cleaning and normalizing the sample data, we used XGBoost algorithm classifies sample data. The experimental results show that the financial statement fraud identification model built based on the XGBoost algorithm has the best prediction effect, with an accuracy of 86.96% and a precision of 88.57%. Conclusion: The financial statement fraud identification model based on the XGBoost algorithm is better than the logistic regression, support vector machine and random forest algorithms in machine learning algorithms in all performance indicators.

Keywords. listed companies; financial statement fraud; machine learning; XGBoost

1. Introduction

In the 1980s, the stock exchange market was born in China. After several rounds of bulls and bears in the stock market, the economy burst into a new vitality in 2000, and people are paying more and more attention to the stock market, and the enthusiasm of researching and predicting the stock market has never subsided [1]. When investing in the stock market, we should not only consider the macro factors, but also pay more attention to the value of the company itself, especially the profitability of the enterprise, which is fully reflected in the financial statements. Financial statements summarize and reflect the overall operating results and financial position of the enterprise in the past period of time, and provide information for investors to make decisions [2]. According to the provisions of accounting standards, the notes to the financial statements should disclose news that may have a negative impact on the company's estimated value, and news that has a positive impact on the company's valuation should be disclosed cautiously so as not to mislead investors [3]. However, in this case, honesty and trustworthiness may lead to damage to the company's reputation, a significant drop in share price, loss of wealth of the company's senior management and employee unemployment and other problems [4]. In this case, in order to maintain their reputation,

[1] Corresponding Author: Jingbang ZHOU E-mail: 13236599797@163.com

to ensure the stability of their own stock price does not fall or in order to obtain cash through financing, the listed company is the most effective means of financial statement data modification and falsification. This kind of breach of trust will make the users of financial statements obtain false information, make wrong judgments, and ultimately suffer great losses [5]. This is the most effective means to modify and falsify financial statement data

In April 2020, Ruixing Coffee announced that it had admitted to financial fraud, fictitious transactions of more than 2 billion yuan, and eventually suffered the fate of "18 months" delisting, triggering a number of meltdown mechanisms and leading to the suspension of trading, which brought heavy losses to investors. Subsequently, it was fined a huge amount of money by the relevant regulatory agencies in China and the U.S.[6] In 2020, Ruixing Coffee was officially delisted from NASDAQ, and the details of the fictitious transactions were subsequently exposed, which made people think very carefully. In order to expel the interests, a year-long systematic counterfeiting project was opened, and as many as forty enterprises were associated with this financial counterfeiting case [7].

2. Literature review

In foreign financial fraud research, sample selection mainly relies on four major databases: the Government Accountability Office, the Audit Analysis Database Restatement Announcement (AA), the securities class action database of the Stanford Securities Class Action Clearinghouse (SCAC), and the U.S. Securities and Exchange Commission Accounting and Auditing Executive Reports (AAERs) of the Exchange Commission. Each of these databases has its own strengths and limitations, such as differences in coverage, date of first identification of fraud, and omitted or effectively omitted events, so no one database can dominate and the choice of an appropriate database depends on the researcher's specific research question. This has a significant impact on the empirical test results.

Among them, the GAO, AA and CFRM databases partially omit the events they are trying to capture. When these missing cases are associated with the researchers' variables of interest, they may cause bias. For example, research using AAERs sample data found that the frequency of discovered financial fraud in companies is usually less than 1% of all companies per year [8]. Researchers estimate that only about half of serious financial reporting violation cases are discovered by the U.S. Securities and Exchange Commission (SEC). This problem may be more prominent in less developed countries with weaker institutional environments [9], indicating that the number of financial fraud cases that have been discovered The scarcity continues to bring challenges to the identification of financial fraud in listed companies.

In terms of structured data, researchers divided company samples into financial manipulation categories and aggressive accruals categories, and used the Probit model to find that high lagged accruals can help identify earnings manipulation companies. The existence of high lagged accruals indicates that management has adopted the ultimate reasonable earnings management strategy [10]. Another study constructed 8 explanatory variables based on financial statement data and used the weighted exogenous sampling maximum likelihood (WESML) model to identify earnings manipulation. Seven of the financial statement ratios represent indicators. The higher the index value, the more likely it is that earnings are overstated. The bigger [11].

This study builds a financial statement fraud identification model based on the XGBoost algorithm in machine learning to improve financial statement users' awareness of potential fraud, identify financial statement fraud, reduce losses caused by financial statement fraud, and maintain the sustainable development of the capital market. .

3. The research methodology

3.1. Data acquisition

The data used in this study come from the annual financial statements of Shenzhen and Shanghai A-share listed companies from 2011 to 2020 in the China Stock Market and Accounting Research (CSMAR) database, in which 283 fraudulent financial statements were selected, involving a total of 126 listed companies. In order to control the external environment and industry factors, this study refers to two criteria when selecting non-fraud samples: first, the listed companies involved in the fraud sample data and the non-fraud sample data belong to the same industry; second, the fraud sample data and the non-fraud sample data belong to the same industry. from the same year. According to these two criteria, a total of 566 non-fraudulent financial statements of 252 listed companies were selected with a matching ratio of 1:2. Finally, this study selected 849 financial statements as detection samples for the financial statement fraud identification model, involving a total of 378 listed companies. The summary of sample industry types and the distribution of sample years are shown in Table 1.

Table 1. Summary of sample industry types

type of industry	non-fraud sample size	The number of fraud samples	Total	As a percentage
Agriculture, forestry, animal husbandry and fisheries	20	10	30	3.53%
Mining	12	6	18	2.12%
Manufacturing	370	185	555	65.37%
Construction	18	9	27	3.18%
Wholesale and retail trade	32	16	48	5.65%
Transportation, storage and postal services	8	4	12	1.41%
information transmission, software and information technology	46	23	69	8.13%
Services	20	10	30	3.53%
Real Estate	22	11	33	3.89%
Leasing and business services	4	2	6	0.71%
Water, Environment and Utilities Management	4	2	6	0.71%
health and social work	10	5	15	1.77%
Consolidated Total	566	283	849	100%

As can be seen from Table 1, listed companies in the manufacturing industry are most involved in financial statement fraud, accounting for more than 50% of the total, and the frequency of financial statement fraud occurred in the period of 2015 to 2017 is on the high side.

3.2. Variable selection

3.2.1 Initial selection of variables

In order to improve the accuracy of model prediction, it is crucial to select appropriate indicators for financial fraud identification. Therefore, on the basis of existing research, based on five dimensions, namely, solvency, operating ability, profitability, development ability and governance structure, this study initially selects 26 indicators for measuring financial statement fraud, which are composed of 22 financial variables and 4 non-financial variables.

3.2.2 Variable Screening Models

Information value (IV) can be used to evaluate the impact of variables on the target, that is, to measure the predictive ability of variables. The calculation of information value is based on weight of evidence (WOE), a coding form that processes raw variables by grouping [12]. For the i-th group, the weight of evidence is calculated as follows (1):

$$\mathrm{WOE_i} = \ln \frac{f(x_i \mid X_n)}{f(y_i \mid Y_n)} \tag{1}$$

where f(Xi/Xn) is the ratio of the number of financial report fraud samples in this group to the total number of financial report fraud samples after grouping; f(yi/yn) is the ratio of the number of non-fraud financial report samples in this group to the total number of non-fraud financial report samples after grouping [13]. Therefore, the greater the weight of evidence, the greater the number of samples of financial reporting fraud. The information value is calculated by the weighted sum of the evidence weights, which is calculated as follows (2):

$$IV = \sum\left(f(x_i \mid X_n) - f(y_i \mid Y_n)\right)\ln \frac{f(x_i \mid X_n)}{f(y_i \mid Y_n)} \tag{2}$$

As can be seen from equation (2), the information value is non-negative. The larger the information value of a variable, the stronger the predictive ability of the variable for the target classification. Therefore, this study introduces the information value to construct the financial fraud indicator screening model.

An indicator with an information value greater than 0.03 is an indicator with predictive ability. Therefore, this study finally selected 21 indicators: current ratio (X01), quick ratio (X02), inventory turnover rate (X05), accounts payable turnover rate (X06), accounts receivable turnover rate (X07), Accounts receivable to income ratio (X08), total asset turnover rate (X09), inventory to income ratio (X10), shareholders' equity turnover rate (X11), return on assets (X12), return on invested capital (X13), Net profit rate on total assets (X15), long-term capital return rate (X17), growth rate of total assets (X18), growth rate of total operating income (X20), growth rate of total operating costs (X21), growth rate of net assets per share (X22), proportion of independent directors (X23), shareholding proportion of the board of directors (X24), shareholding proportion of the board of supervisors (X25), shareholding proportion of the top ten shareholders (X26)[14].

3.3. XGBoost algorithm

The XGBoost algorithm is based on the gradient boosting tree algorithm and adds a regularization term to the objective function, which can reduce the complexity of the model and avoid overfitting. Its objective function is as shown in formula (3) and formula (4):

$$\text{Obj}\ (\Phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{3}$$

$$\text{where}\ \Omega(f) = \gamma T + \frac{1}{2}\lambda\omega^2 \tag{4}$$

Among them$\hat{y}_i$is the predicted value, yi is the true value,$\Omega(f_k)$is the regular term, fk is the decision tree, T represents the number of leaf nodes, w represents the proportion of leaf nodes,γ Controlling the number of leaf nodes, theλ Control leaf node proportions.

The XGBoost algorithm performs iterative operations and second-order Taylor expansion during the solution process of the objective function, as shown in formula (5), which improves the solution speed and model training speed.

$$Obj^t = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

Among them. $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the first-order and second-order derivatives of the loss function respectively. The XGBoost algorithm sorts the eigenvalues in advance and then saves them as block structures, so it can maximize the determination of the criteria for segmentation points. In addition, in order to meet the situation where the eigenvalues after data processing are sparse, the XGBoost algorithm sets a certain diversion for missing values. This greatly improves the efficiency of the algorithm.

3.4. Model construction

This study sampled a total of 849 data samples, involving 378 listed companies, and determined 21 indicators through the indicator screening model, including 17 financial variables and 4 non-financial variables [15]. After data normalization, the sample data was divided into a training set and a test set using the five-fold cross-validation method, and the XGBoost algorithm was used as a classifier to build a financial statement fraud identification model.

4. Analysis of results

4.1. Model parameterization

The XGBoost setting parameters using grid search are shown in Table 2.

Table 2. XGBoost parameter settings

parameters	The meaning of the parameter	parameter values
learning_rate	The learning rate, which controls the step size at each iteration when updating the weights	0.008
n_estimators	The total number of iterations, i.e., the number of decision trees	2300
max_depth	the maximum depth of the decision tree	9
colsample_bytree	The proportion of all features used in training each tree	0.8
subsample	The proportion of data used to train each tree out of the total training set	0.8
reg_alpha	regularization factor	0.0001

4.2. Experimental results

The model generates sample memory during the training process, and if the training set is used for testing it will lead to high test results and affect the performance of the model[16]. Therefore, this study adopts the model validation method of five-fold cross-validation to improve the generalization ability of the model.

This study uses three machine learning algorithms: logistic regression, support vector machine, and random forest to compare with the XGBoost algorithm as a financial statement fraud identification classifier. The classification results of each machine learning algorithm are shown in Table 2.

Table 3. Comparison of evaluation metrics for classification results of various machine learning algorithms

Model	Accuracy	Accuracy	Recall values	F1 value
logistic regression	70.40%	69.11%	65.93%	65.02%
support vector machines	71.13%	65.38%	62.74%	67.88%
Random Forests	80.32%	81.36%	77.22%	79.24%
XGBoost	86.96%	88.57%	83.61%	81.98%

Taking various evaluation indicators into consideration, it can be seen that the financial statement fraud identification model based on the XGBoost algorithm has the best prediction effect, with an accuracy of 86.96% and a precision of 88.57%.

4.3. Analysis of experimental results

Ensemble learning combines the variances and biases of multiple individual learners and is a more comprehensive strong supervised learning algorithm that can achieve better performance. Therefore, the performance of the financial statement recognition model based on the random forest and XGBoost algorithm in the ensemble learning algorithm is significantly higher than that of models based on individual learners such as logistic regression and support vector machines [17]. Each decision tree of the random forest randomly selects a feature subset, while the XGBoost algorithm uses a greedy algorithm to determine the optimal feature subset, and serially generates a series of individual learners, and then uses the difference between the predicted value and the true value as the objective function. Optimizing parameters, the final predicted value is the sum of the predicted values of individual learners. Therefore, for imbalanced data sets, the prediction model built based on the XGBoost algorithm has better classification results.

5. Conclusion

With the rapid development of computer technology, various fields have entered the era of big data and artificial intelligence. Machine learning has been widely used because it can process large amounts of data quickly and effectively. Building a financial statement fraud identification model based on machine learning algorithms can improve the shortcomings of traditional financial statement fraud identification methods that rely too much on manpower. Therefore, this article proposes to build a financial statement fraud identification model based on the XGBoost algorithm in machine learning. This article draws the following conclusions: ① Comparing the prediction models constructed by multiple machine learning algorithms, experiments have proven that the financial statement fraud identification model based on the integrated learning algorithm is better than Individual learner. ② Comparing the random forest algorithm and the XGBoost algorithm, which are both ensemble learning algorithms, experiments have proven that the financial report fraud identification model based on the XGBoost algorithm has better prediction ability.

Funding

This research was funded by Anhui Provincial Department of Education College Humanities and Social Sciences Key Project,project number 2022AH040202.

References

[1] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, 10, 39700-39715.

[2] Hashemi, S. K., Mirtaheri, S. L., & Greco, S. (2022). Fraud detection in banking data by machine learning techniques. IEEE Access, 11, 3034-3043.

[3] Wang, H., Wang, W., Liu, Y., & Alidaee, B. (2022). Integrating machine learning algorithms with quantum annealing solvers for online fraud detection. IEEE Access, 10, 75908-75917.

[4] Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. Human-Centric Intelligent Systems, 2(1), 55-68.

[5] Hamal, S., & Senvar, Ö. (2021). Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. Int. J. Comput. Intell. Syst., 14(1), 769-782.

[6] Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1), 111-138.

[7] Riskiyadi, M. (2024). Detecting future financial statement fraud using a machine learning model in Indonesia: a comparative study. Asian Review of Accounting, 32(3), 394-422.

[8] Li, R., Liu, Z., Ma, Y., Yang, D., & Sun, S. (2022). Internet financial fraud detection based on graph learning. IEEE Transactions on Computational Social Systems, 10(3), 1394-1401.

[9] Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology, 41(1), 33-40.

[10] Wyrobek, J. (2020). Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. Procedia Computer Science, 176, 3037-3046.

[11] Khamainy, A. H., Ali, M., & Setiawan, M. A. (2022). Detecting financial statement fraud through new fraud diamond model: the case of Indonesia. Journal of Financial Crime, 29(3), 925-941.

[12] Khamainy, A. H., Amalia, M. M., Cakranegara, P. A., & Indrawati, A. (2022). Financial statement fraud: The predictive relevance of fraud Hexagon theory. JASF, 5(1), 110-133.

[13] Shonhadji, N., & Maulidi, A. (2021). The roles of whistleblowing system and fraud awareness as financial statement fraud deterrent. International Journal of Ethics and Systems, 37(3), 370-389.
[14] Aviantara, R. (2023). Scoring the financial distress and the financial statement fraud of Garuda Indonesia with «DDCC» as the financial solutions. Journal of Modelling in Management, 18(1), 1-16.
[15] Seifzadeh, M., Rajaeei, R., & Allahbakhsh, A. (2022). The relationship between management entrenchment and financial statement fraud. Journal of Facilities Management, 20(1), 102-119.
[16] Fitriana, F., Saepudin, D., & Santoso, R. A. (2021). Fraud Diamond Theory Detect Financial Statement Fraud in Manufac-turing Companies on The Indonesia Stock Exchange. International Business and Accounting Research Journal, 5(2), 93-105.
[17] Widnyana, I. W., & Widyawati, S. R. (2022). Role of forensic accounting in the diamond model relationship to detect the financial statement fraud. International Journal of Research in Business and Social Science (2147-4478), 11(6), 402-409.