Hindawi
Journal of Mathematics
Volume 2021, Article ID 9241338, 9 pages
https://doi.org/10.1155/2021/9241338

Research Article

Empirical Analysis of Financial Statement Fraud of Listed Companies Based on Logistic Regression and Random Forest Algorithm

Xinchun Liu

School of Economics and Management, Panzhihua University, Panzhihua 617000, Sichuan Province, China

Correspondence should be addressed to Xinchun Liu; liuxinchun@pzhu.edu.cn

Received 23 October 2021; Accepted 1 December 2021; Published 17 December 2021

Academic Editor: Miaochao Chen

Financial supervision plays an important role in the construction of market economy, but financial data has the characteristics of being nonstationary and nonlinear and low signal-to-noise ratio, so an effective financial detection method is needed. In this paper, two machine learning algorithms, decision tree and random forest, are used to detect the company's financial data. Firstly, based on the financial data of 100 sample listed companies, this paper makes an empirical study on the fraud of financial statements of listed companies by using machine learning technology. Through the empirical analysis of logistic regression, gradient lifting decision tree, and random forest model, the preliminary results are obtained, and then the random forest model is used for secondary judgment. This paper constructs an efficient, accurate, and simple comprehensive application model of machine learning. The empirical results show that the comprehensive application model constructed in this paper has an accuracy of 96.58% in judging the abnormal financial data of listed companies. The paper puts forward an accurate and practical method for capital market participants to identify the fraud of financial statements of listed companies and has certain practical significance for investors and securities research institutions to deal with the fraud of financial statements.

1. Introduction

Data is considered to be the source of knowledge. Massive data often contains a lot of valuable information [1], just as Walmart found that, after buying baby paper diapers, men usually buy beer to reward themselves, so they bundle the two products. With the increase of sales, it can be seen that reasonable data analysis will obtain great value information from the data. With the massive increase of data, the dimension of data has also been greatly improved [2]. Massive and high-dimensional data has far exceeded the ability of human beings to process and analyze data. In order to more accurately mine all kinds of valuable knowledge and information contained in massive data, the question is how to complete these tasks with the help of the power of machines [3]. It has become one of the most important tasks for scientific researchers. Compared with human beings, computer has powerful computing power [4]. With the help of computer's ability to process data, data mining can be quickly active, and it has become one of the most cutting-edge research fields in the 21st century [5]. The task of data mining is to extract implicit and potentially valuable information and knowledge from a large amount of actual data through data analysis [6]. The common directions of data mining include data classification, regression, and clustering. As one of the most important methods in the field of data mining, data classification is applied to many fields, including customer relationship classification, spam recognition, and facial expression recognition. Data mining often needs the help of data statistics and machine learning algorithms [7]. The algorithms used include nearest neighbor algorithm, decision tree, naive Bayes, logistic regression, and artificial neural network [8]. Among many machine learning algorithms, random forest is a combined classifier, which was proposed by Leo Breiman in 2001 and registered with random forest as a trademark [9]. Subsequently, many scientists have made a

lot of contributions to the optimization and development of random forest. Random forest contains multiple decision tree classifiers [10]. Each decision tree classifier generated is a weak classifier generated by random sampling and training model from all training data sets and feature sets [11]. The results of random forest data classification are determined by the comprehensive voting of all or part of the decision trees. As a combined classifier, random forest can carry out distributed parallel operation and has many advantages that other classification algorithms cannot compare, but there are still some shortcomings in algorithm design. For example, there is not a very reasonable method to specify the size of the random forest when the decision tree is generated [12]. Too large or too small scale of the model will affect the final decision result of the model. In addition, because the data sets and feature sets used in the generation of each decision tree are selected from the total training data set and feature set by random put back method [13], the quality of each decision tree cannot be guaranteed, and the final voting result is determined by the decision results of all or part of the decision tree, which brings great uncertainty to the accuracy of model classification [14]. Poor decision tree not only reduces the generalization ability and prediction accuracy of the model but also increases the time of data prediction [15].

Financial data is abnormal. In short, it means that the statistical data we get are inconsistent with the actual data [16]. The occurrence of data anomalies is affected by subjective and objective factors and many other aspects. The emergence of objective factors mainly comes from technical problems, such as backward statistical methods, unscientific investigation methods, and difficult to obtain certain data [17]. The subjective reason involves the interests of all aspects. In order to seek private interests, some companies and individuals subjectively falsify statistical data through illegal means such as concealment, false reporting, and tampering, which has caused great obstacles to the smooth development of statistics, review, and other works. In short, the subjective reason for the falsification of financial data of listed companies is that enterprises take risks and operate in violation of regulations for their own economic interests. How to judge the operating level and ability of listed companies and ensure the compliance of public financial statements of listed companies have attracted the keen attention of investors, which also directly affects the vital interests of minority shareholders. In the process of enterprise operation, both systemic risks and risks caused by their own operation may bring many problems. In order to maintain their own stock price, listed companies have a certain motivation to hide risks, which needs to be studied and found through the supervision of the CSRC [18]. As an important research direction of artificial intelligence, machine learning has naturally become a very important prediction means and candidate scheme in the prediction of financial violations [19]. Machine learning and deep learning include many learning models, such as supervised learning, unsupervised learning, and semisupervised learning. These different research categories have different labeling degrees of data samples. Supervised learning needs to make early judgment on the samples (such as whether they have good academic performance). Unsupervised learning automatically judges through the classification and processing of data, while semisupervised learning is between the two and needs a certain degree of data labels. This paper uses the decision tree model and random forest (RF) model in supervised learning to detect the company's financial violations. The reason is that the decision tree and random forest algorithm have certain advantages in the processing of financial data. For the decision tree, the preparation of data is often simple or unnecessary, and it can deal with data and conventional attributes at the same time and can make feasible and effective results for large data sources in a relatively short time. The advantage of random forest algorithm is that it can produce unbiased estimation of the generalized error internally when building the forest. It calculates the closeness in each case, which is very useful for data mining, detecting outliers, and visualizing data.

2. Related Work

As a combined classifier algorithm, random forest contains many tree classifiers. Each tree classifier is a weak classifier (decision tree), and the training and decision of each decision tree do not affect each other, which greatly improves the speed of training and decision-making of the overall model and is very convenient for parallel operation and multicore operation [20]. Random forest algorithm can play an important role in financial data. For example, we can establish multiple decision trees by using random forest algorithm, so that each decision tree has its own emphasis, and pay attention to the analysis and judgment of different financial data, so that each has its own advantages. Finally, the output results of all individual modules are summarized by using the method of projection, and a total judgment result is obtained. Çömert et al. [21] used the random forest model to predict the prediction ability of RF on large data sets. It is found that the prediction ability of the random forest model based on bagging algorithm is better; Jun [22] et al. studied the prediction ability of SVM, logistic model, and random forest model. The conclusion is that the prediction ability of random forest model is always better than that of SVM model, while the prediction ability of SVM model with parameter selector is better than that of logistic model [23]. Due to its high prediction accuracy and difficult overfitting characteristics, random forest theory and application have developed rapidly [24]. Random forest model is used in prediction algorithms in all walks of life. Sahin et al. [25] used the random forest model to predict the high-dimensional protein domain structure, with an accuracy of 79.78%, which is higher than the best accuracy predicted by other models; Quiroz [26] used the random forest model to predict gene sequences and used DLDA (direct linear discriminant analysis), KNN (proximity algorithm), and SVM model to make the same prediction. It is found that the random forest prediction effect has obvious advantages, and it is said that the random forest also plays an important role in gene prediction. Demestichas [27] predicted the remote sensing location by using the random forest model and other commonly used prediction algorithm models and

found that the prediction accuracy of the random forest model is the highest. Kurniawan [28] simplified the financial indicators based on the correlation and importance of the indicators, constructed the financial early warning model, and established the variable precision weighted average roughness decision tree, which significantly improved the noise prevention ability and classification accuracy of the early warning model. Yan [29] proposed that, in previous studies, for the comparison of model effects of financial early warning models, more consideration was given to the accuracy of comparison model results, and the probability of false error of the model was not discussed. Tosiri [30] believed that the stochastic forest model has better interpretability. The contribution stochastic forest model is introduced to study the credit risk of corporate bonds, and the importance of each index for the determination of corporate default rate is obtained through the index contribution rate. Chan [31] believed that the machine learning algorithm integrating multiple classifiers has higher prediction accuracy and more stability than a single classification model. According to the data of listed enterprises for two consecutive years, the prediction results are compared with the logistic model, which proves that the prediction effect of random forest is better. Sakiyani [32] constructed the credit risk assessment model with the random forest method through the combination of financial indicators and nonfinancial indicators and compared the assessment model with the prediction accuracy and model performance, respectively, which confirmed that the prediction ability of the random forest was better than that of the decision tree. Onan et al. [33] used three neural language models, two unsupervised term weight functions, and eight supervised term weight functions for irony recognition task. The classification accuracy of the model is 95.30%, and the results are satisfactory. Korukoğlu et al. [34] found that consistent clustering and elite Pareto based multiobjective evolutionary algorithm can be effectively applied to integrated pruning. Through the experimental analysis of the traditional integration method and pruning algorithm, the effectiveness and efficiency of the scheme are proved. Toaar et al. [35, 36] proposed an integrated feature selection method, which aggregates multiple individual feature lists obtained by different feature selection methods, so as to obtain a more robust and efficient feature subset.

3. Decision Tree Algorithm and Random Forest Model Construction

3.1. Decision Tree Algorithm. Decision tree is a tree classifier, which belongs to a supervised learning method. It is composed of root node, internal node, and leaf node. Each decision tree has only one root node, and the data prediction and training are extended level by level from the root node. The data on each nonleaf node will be divided into two or more sub-data-sets according to the characteristic attributes of the current node and handed over to the node at the next level for processing. After the data reaches the leaf node, it is no longer necessary to continue the division. The leaf node where the data is located is the result of this classification prediction. The process of decision tree prediction data is essentially a top-down process. Each level of nonleaf node performs data analysis and completes data classification according to its own characteristic attribute classification rules. The decision tree can be roughly divided into two categories according to the different criteria for dividing data sets by nonleaf nodes: one is the decision tree based on information entropy, such as ID3 decision tree algorithm, and the other is a decision tree based on Gini index, such as CART decision tree algorithm.

There are many algorithms to build decision tree, but they basically use top-down greedy algorithm to form a tree model, including multiple child nodes. There are two types of child nodes: nonleaf nodes and leaf nodes. Each nonleaf node selects the feature element with the best classification effect from the current feature set as the feature attribute of the current node to divide the data set. The data set is divided into two or more sub-data-sets. According to this method, the sub-data-sets are iterated repeatedly until the stop condition of decision tree growth is reached; then the leaf node is the result of classification. The data set is no longer divided at the leaf node, and the data classification is completed at the leaf node. Figure 1 shows a binary decision tree algorithm model.

In the above algorithm model diagram, a binary decision tree is shown. The decision tree extends downward from the root node. Each nonleaf node will have a different feature attribute t, and different binary classification conditions are formulated according to different feature attributes to classify the input data. The decision tree finally divides the data to be predicted into a leaf node through data attribute division. The leaf node has a classification attribute label 1, which represents the classification results of the current data. According to different node attribute division standards, decision tree algorithms mainly include ID3 decision tree algorithm, C4.5 decision tree algorithm, and CART decision tree algorithm.

3.1.1. ID3 Decision Tree Algorithm. In ID3 algorithm, the criterion to measure the classification ability of feature attributes on data sets is information gain. From information theory, we know that the greater the information gain is after data set division, the more the entropy of data set decreases, and the higher the purity of data set is. ID3 algorithm uses information gain as the standard for the classification of nonleaf node feature attributes. In the current feature set, select the feature attribute with the largest information gain after data set division as the feature attribute of nonleaf node to create nonleaf node. According to this method, recursive operation is carried out to construct child nodes, meet the conditions for creating leaf nodes, stop the continuous growth of decision tree, and complete the construction of a decision tree model.

$$\text{Entrop}(E) = \frac{\lambda \ln t}{2} + \prod_{i=1}^{n} t. \tag{1}$$

Information gain is defined as the difference in the entropy of the data set before and after the data set is divided according to a certain characteristic attribute T, which is called information gain. Now set t as the feature element in the feature set t, and divide the sample data set D into m subsample sets of $D1, D2, \ldots, Dm$ according to feature t. Then,

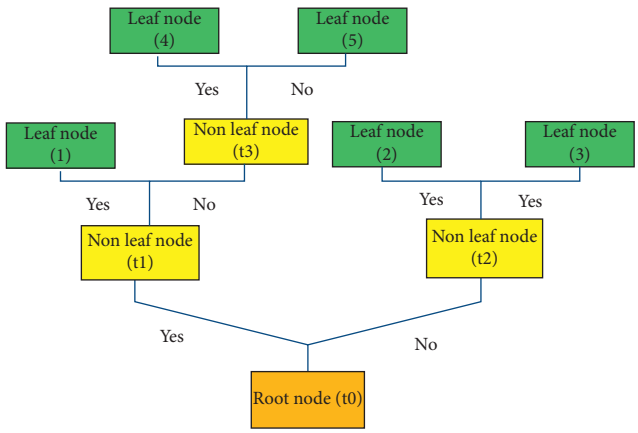

Figure 1: Two-classification decision tree algorithm model diagram.

the information gain of data set d after being divided according to feature t can be expressed as

$$\text{Gain}(x) = \text{Entrop}(E) + \frac{\int_{i=1}^{n} (L - M)}{\text{Entrop}(E)}, \tag{2}$$

where Dj is the number of samples in the j-th subsample data set and D is the total number of samples in the data set. Information gain describes the difference in the entropy of the total sample before and after the data set is divided according to a certain characteristic attribute and the sum of the weighted entropy of each subsample data set obtained after the division (the expected entropy of all subsamples). ID3 decision tree algorithm takes the information gain as the standard of feature attribute selection, recursively calculates the information gain after data set division according to the current feature set elements for each sample data set, selects the feature attribute with the largest information gain for node creation, and completes the construction of an ID3 decision tree until the condition that the decision tree stops growing is reached.

3.1.2. C4.5 Decision Tree Algorithm. Different from ID3 algorithm, C4.5 algorithm uses the information gain ratio as the judgment standard for the selection of feature attributes of nonleaf nodes. The information gain ratio used in C4.5 introduces split information. Data set D is divided according to feature t to obtain m sub-data-sets. The split information after data set division is defined as

$$\text{split\_info}(x) = \lim_{n \longrightarrow \infty} \prod_{i=1}^{n} \left(\frac{L}{M}\right)^2 - \ln \int_{i=1}^{n} (L - M). \tag{3}$$

D_i is the data volume of the i-th sub-data-set after division, D is the total data volume of the data set before data set division, and the split information can be regarded as the weighted sum of the possible entropy of each subsample data set. With the definitions of information gain and split information, the information gain ratio after data set division according to feature t can be obtained:

$$\text{GainRation}(x) = \frac{\text{Gain}(x) - \text{split\_info}(x)}{1 - x}. \tag{4}$$

The construction process of decision tree in C4.5 decision tree algorithm is the same as that in ID3, except that the judgment criterion for selecting the optimal feature attribute from the current feature set every time is the information gain ratio. In the process of each iteration, the feature with the largest information gain ratio is selected from the current feature set as the attribute for node creation until the condition that the decision tree stops growing is reached, and the construction of the whole decision tree is completed.

3.1.3. CART Decision Tree Algorithm. CART decision tree algorithm is classified regression tree. CART decision tree can be used to predict both the classification problem of discrete values and the regression problem of continuous data. Different from ID3 algorithm and C4.5 algorithm, CART algorithm always divides the data set into two parts when dividing the data set on each nonleaf node, no matter how many types of current feature attributes are in the data set, so the decision tree constructed by CART algorithm is a binary tree model. In addition, the standard for CART algorithm to select the optimal feature attribute is Gini index. For sample data set D, it is divided according to a feature attribute t. The divided subsample data set contains samples of k categories; then the Gini index of subsample data set is expressed as follows:

$$\text{Gini}(x) = \lambda + \bigcup_{i=1}^{n} |E|. \tag{5}$$

The data set D is divided according to the characteristic attribute t to obtain m sub-data-sets. The Gini index obtained after classification is

$$\text{Gini}(x) = \bigcup_{i=1}^{n} \lambda x \text{Gini}(x). \tag{6}$$

3.2. Random Forest Algorithm. When building a single decision tree model, it is easy to be affected by the training data set. In order to avoid overfitting, it is necessary to prune the decision tree, and excessive pruning will reduce the prediction ability of the decision tree. In addition, because the local greedy method is used in the iterative feature attribute selection during the construction of the decision tree, each time the optimal feature is selected from the current feature set to create a new node, the model will not be backtracked, which is easy to cause the local optimal solution, so that the final prediction model converges to the local optimal solution. Figure 2 shows the flow chart of a random forest model.

Random forest (RF) is a combined classifier. The metaclassifier used is CART decision tree. CART decision tree algorithm belongs to unstable learning algorithm. Therefore, Bagging algorithm is introduced to conduct multiple random independent sampling on the training sample set to train a single decision tree classifier, so as to build a combined model. This can effectively improve the overall

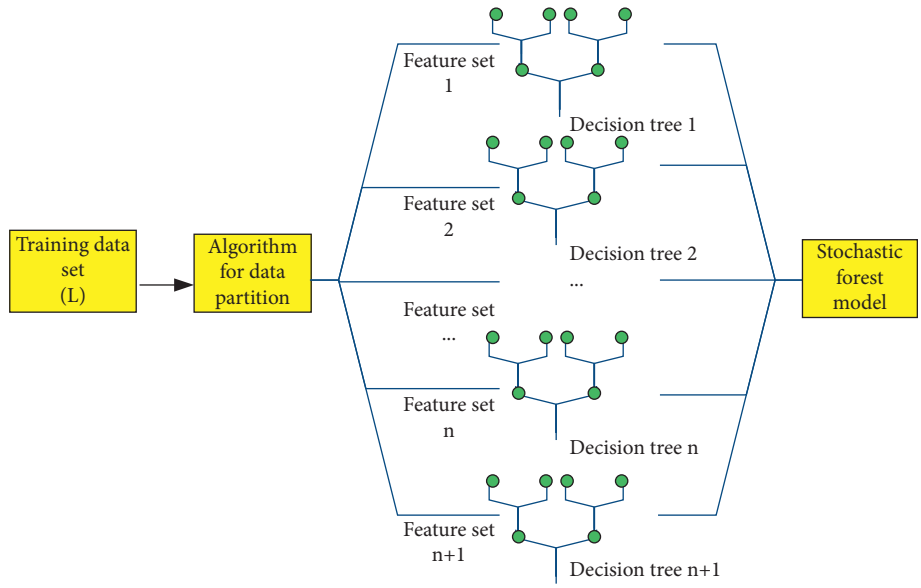

Figure 2: Flow chart of random forest model construction.

generalization ability of the model and increase the accuracy of model prediction. The description of random forest algorithm is that random forest is a set containing multiple tree classifiers, such as

$$\left\{\int f(t,p)dt\ ,\quad t=1,2,3,\cdots\right\}, \tag{7}$$

where $H(x, e)$ is the metaclassifier constituting the model. Here, the classification regression tree constructed by CART algorithm without pruning operation is used. X represents the training data set constructed by random forest, which is a multidimensional vector set. In general, for the regression problem, the random forest calculates the prediction results of each decision tree and then obtains the average value of the prediction results of all decision trees by averaging. Output this average value as the final prediction result. For the prediction of classification problems, the mode voting method is the most used voting method in random forest, that is, counting the votes of decision tree received by each classification label and outputting the classification label with the largest number of votes as the final prediction result. For example, the total number of tags predicted in facial expression recognition is 6. Unknown facial expression data is input. Each decision tree model in the random forest makes an independent prediction for the expression data and gives the prediction results. The facial expression tag with the largest number of votes is output as the final prediction result. While giving the final prediction result, the random forest can also give the probability that the prediction result is a certain expression by counting the voting of each expression.

There are n prediction classification labels of random forest model $\{h-(x), h{:}\,(x), \ldots, \mathrm{hm}\,(x)\}$, so the algorithm flow of data prediction of the random forest model is as follows (Figure 3). First, classify different data and add classification labels, then calculate the corresponding mode and average according to the classification problem and regression problem, and finally output the results.

The random forest model can be described as a combined classifier:

$$\{f_i(t),\quad i=1,\cdots n\}. \tag{8}$$

It is a classifier set composed of k $(k>1)$ subclassifiers. The prediction output result obtained by inputting the prediction vector X is Y. For the sample data set (x, y), the boundary function is defined as

$$margin(m,n)=k\sum(f(m)-n)+\frac{k}{\lambda}\iint f(m)+mn. \tag{9}$$

The boundary function calculates the combined classifier to predict a sample vector, predict the correct average number of votes and the maximum number of votes in the case of wrong prediction, and calculate the difference between the two indexes. Obviously, the larger the value of the boundary function, the stronger the prediction ability of the classifier set and the higher the confidence. The generalization error of combined classifier is defined as

$$\mathrm{PE}^{*}=Q_{m,n}(margin(m,n)\geq -10). \tag{10}$$

We hope that the larger the value of the boundary function of the combined classifier, the better. This shows that the classifier set has strong generalization ability and high prediction ability. However, there will also be a case where the number of metaclassifiers with wrong prediction in the combined classifier is greater than the number of metaclassifiers with correct prediction, that is, margin $(x, y) < 0$. In this case, the combined classifier will give wrong prediction results. The generalization error represents the probability that the margin $(x, y) < 0$ obtained by the combined classifier for data prediction, that is, the error rate of the combined classifier for data prediction.

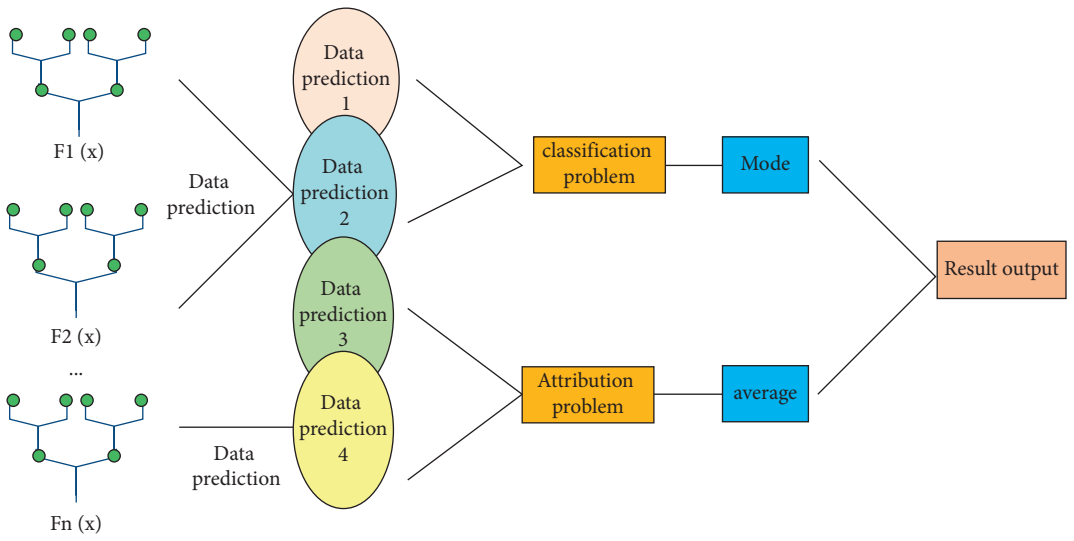

Figure 3: Flow chart of random forest model data prediction.

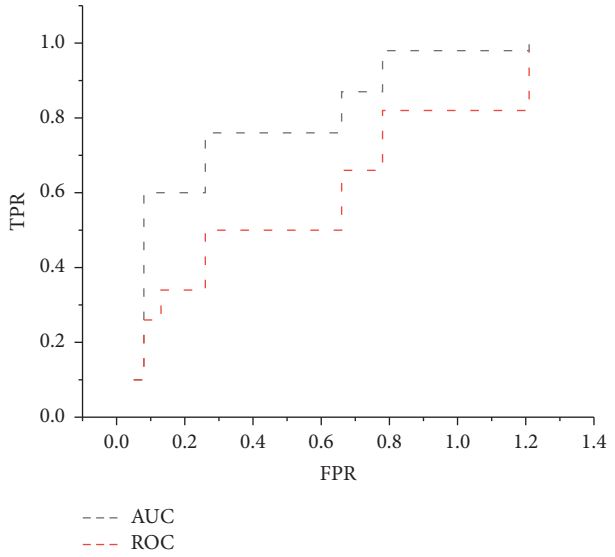

Figure 4: ROC curve and AUC value of gradient lifting decision tree algorithm.

4. Financial Data Detection Based on Decision Tree and Random Forest Algorithm

4.1. Analysis Results of Single Machine Learning Model. Firstly, this paper uses the gradient boosting decision tree model to analyze and judge the financial index data of 100 sample companies. The whole sample is divided into training set and test set according to the ratio of 75 : 25. The idea of gradient lifting decision tree algorithm makes it have natural advantages. It can find a variety of distinguishing financial index features and feature combinations, so as to surpass logical regression in judging the fraud of financial statements. Specifically, the gradient lifting decision tree algorithm has an accuracy of 86.28%, a precision of 84.66%, a recall of 53.28%, and an $F1$ score of 0.603 on the test set. AUC is 0.738, as shown in Figure 4.

Then, this paper uses random forest to study the financial index data of sample companies. On the test set, the accuracy of random forest algorithm is 74.14%, the precision is 100%, the recall rate is 26.18%, $F1$ score is 0.4186, and AUC = 0.865. Although the accuracy of random forest has decreased, its precision is higher, as shown in Figure 5.

4.2. Empirical Analysis Results of Comprehensive Application Model. The purpose of this paper is to comprehensively use appropriate machine learning algorithms to identify the abnormal financial data of listed companies simply, accurately, and generally. In this paper, the decision tree and random forest algorithms are used to judge the results independently, and the final results are obtained. The comprehensive application model also performs well in the test set. In particular, the accuracy of the method in the test set shows that the correct prediction results account for 96.58% of the test set samples, the precision is 100%, the recall rate is 91.91%, $F1$ score is 0.9414, and

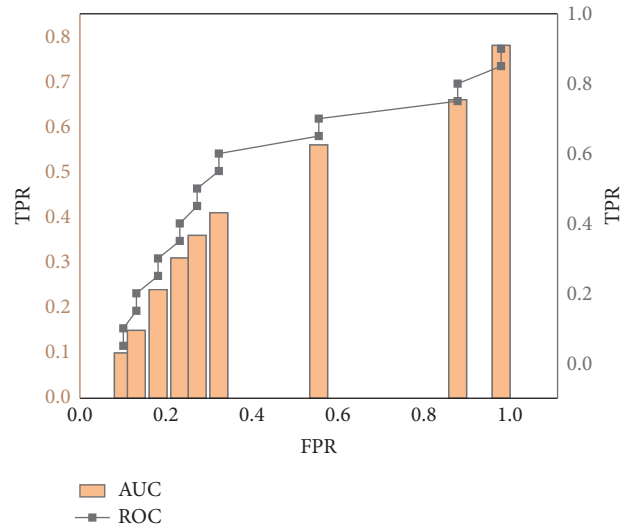

Figure 5: ROC curve and AUC value of random forest algorithm.

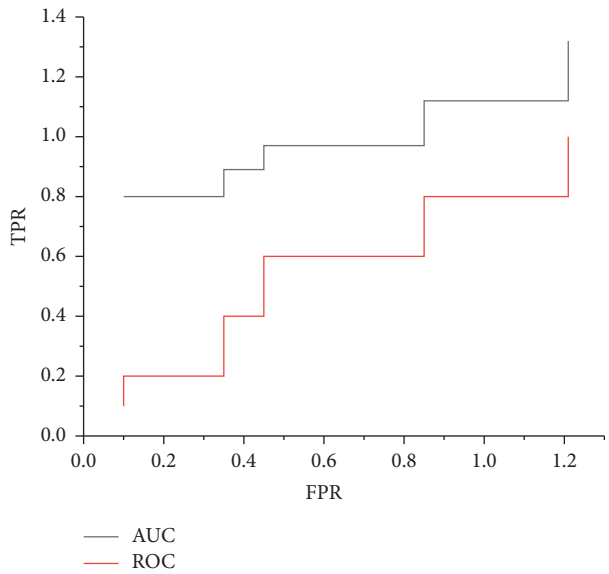

Figure 6: ROC curve and AUC value of comprehensive application model.

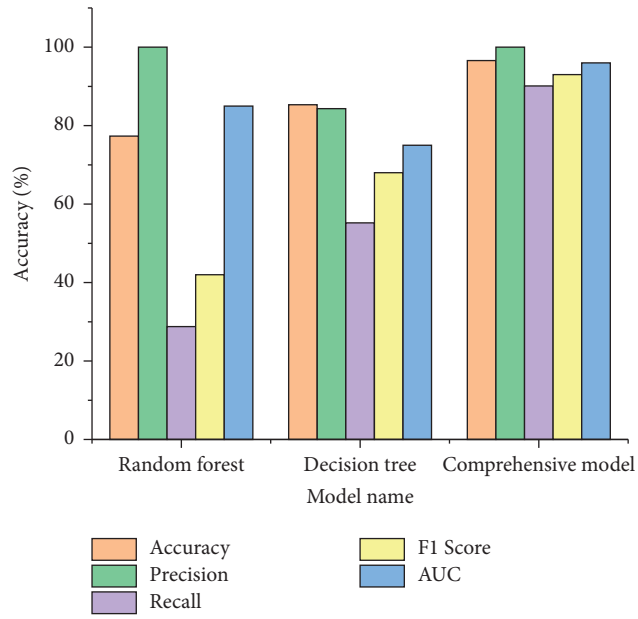

Figure 7: Comparison of experimental results of each machine learning model.

AUC = 0.96. This result completely eliminates the concern of overfitting in the comprehensive application model, as shown in Figure 6.

4.3. Comparative Analysis of Machine Learning Model Results. According to the empirical test results of each machine learning model, this paper summarizes and analyzes the analysis results of each machine learning model as follows in order to get valuable suggestions (Figure 7).

This paper evaluates the advantages and disadvantages of a model from three aspects: simplicity, accuracy, and universality. The comprehensive application model of machine learning constructed in this paper far exceeds the single machine learning technology in accuracy, precision, recall, and *F*1 score. In single machine learning technology, the recall rate of random forest is much lower than that of decision tree, but the precision of random forest is higher than that of decision tree.

5. Conclusion

This paper reviews and summarizes the common financial data anomaly detection methods and identification models. By using the random forest algorithm to judge the independent results of decision tree and random forest algorithm again to get the final result, this paper constructs a machine learning model which can identify the fraud of financial statements of listed companies efficiently, accurately, and simply. It provides a new tool for investors in the capital market to deal with the fraud of financial statements of listed

companies. The comprehensive application model ranks first in four machine learning technologies in terms of accuracy, precision, recall, *F*1 score, and AUC. This is because the model is based on three single machine learning technologies to judge the financial indicators of sample companies and then the random forest model is used to judge the results of the three algorithms, so as to improve the accuracy of model judgment. Due to many reasons, such as the amount of sample data, time constraints, and the limited application of machine learning technology, the comprehensive application model constructed in this paper has little gap with other models in judging nonfinancial statement fraud companies. The main difference is reflected in the accuracy of judging companies with fake financial statements. The accuracy of the comprehensive application model in the test set is 96.58%, which is far higher than those of the other three machine learning models. The comprehensive application model can minimize the type I error on the basis of ensuring that the type II error is maintained at a certain level.

The tree-like machine learning model of decision tree can be widely used in image recognition, word processing, and so on. Through this experiment, we confirm that it can also play a role in the financial field. Through the treatment of financial analysis, we can better supervise the behavior and operation of listed companies and better protect the interests of minority shareholders. In terms of use, the decision tree model can fit higher-dimensional parameters and has a considerable degree of fitting effect.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Veny Amilia, R. Rachmadita, and M. Azani, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Computer Science*, vol. 161, pp. 765–772, 2019.

[2] H. Esmaily, M. Tayefi, H. Doosti, M Ghayour-Mobarhan, H Nezami, and A Amirabadizadeh, "A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes," *Journal of Research in Health Sciences*, vol. 18, no. 2, Article ID e00412, 2018.

[3] S. Buschjäger and K. Morik, "Decision tree and random forest implementations for fast filtering of sensor data," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 1, pp. 209–222, 2017.

[4] K. Zhang, X. Wu, and R. Niu, "The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area," *Environmental Earth Sciences*, vol. 76, no. 11, pp. 1–20, 2017.

[5] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," *Baltic Journal of Modern Computing*, vol. 5, no. 2, p. 221, 2017.

[6] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019.

[7] D. Denisko and M. M. Hoffman, "Classification and interaction in random forests," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1690–1692, 2018.

[8] X. Qiu, L. Zhang, P. Nagaratnam Suganthan, and G. A. J. Amaratunga, "Oblique random forest ensemble via Least Square Estimation for time series forecasting," *Information Sciences*, vol. 420, pp. 249–262, 2017.

[9] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *Journal of Petroleum Science and Engineering*, vol. 174, pp. 776–789, 2019.

[10] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 2, pp. 01–09, 2018.

[11] T. Berhane, C. Lane, Q. Wu et al., "Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory," *Remote Sensing*, vol. 10, no. 4, p. 580, 2018.

[12] J. Abellán, C. Mantas, and G. Castellano, "A random forest approach using imprecise probabilities," *Knowledge-Based Systems*, vol. 134, pp. 72–84, 2017.

[13] Y. Kim, R. Hardisty, and E. Torres, "Seismic facies classification using random forest algorithm," Society of Exploration Geophysicists, Oklahama, OK, USA, 2018, pp. 2161–2165, SEG Technical Program Expanded Abstracts.

[14] X. Tan, S. Su, Z. Huang et al., "Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm," *Sensors*, vol. 19, no. 1, p. 203, 2019.

[15] Y. Li, C. Yan, W. Liu, and M. Li, "A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification," *Applied Soft Computing*, vol. 70, pp. 1000–1009, 2018.

[16] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017.

[17] H. Aydadenta and A. Adiwijaya, "A clustering approach for feature selection in microarray data classification using random forest," *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1167–1175, 2018.

[18] X. Ye, L.-A. Dong, and D. Ma, "Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score," *Electronic Commerce Research and Applications*, vol. 32, pp. 23–36, 2018.

[19] S. Sarkar, R. Raj, S. Vinay, J. Maiti, and D. K. Pratihar, "An optimization-based decision tree approach for predicting slip-trip-fall accidents at work," *Safety Science*, vol. 118, pp. 57–69, 2019.

[20] Z. Chen, F. Han, L. Wu et al., "Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents," *Energy Conversion and Management*, vol. 178, pp. 250–264, 2018.

[21] K. Çömert and U. Avdan, "Object based burned area mapping with random forest algorithm," *International Journal of Electronic Governance*, vol. 4, no. 2, pp. 78–87, 2019.
[22] S. Jun, S. Lee, and H. Chun, "Learning dispatching rules using random forest in flexible job shop scheduling problems," *International Journal of Production Research*, vol. 57, no. 10, pp. 3290–3310, 2019.
[23] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees," *Journal of Cleaner Production*, vol. 203, pp. 810–821, 2018.
[24] S. Balli, A. Sağbaş, and M. Peker, "Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm," *Measurement and Control*, vol. 52, no. 1-2, pp. 37–45, 2019.
[25] E. K. Sahin, I. Colkesen, and T. Kavzoglu, "A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping," *Geocarto International*, vol. 35, no. 4, pp. 341–363, 2020.
[26] J. C. Quiroz, N. Mariun, M. R. Mehrjou, M. Izadi, N. Misron, and M. A. Mohd Radzi, "Fault detection of broken rotor bar in LS-PMSM using random forests," *Measurement*, vol. 116, pp. 273–280, 2018.
[27] K. Demestichas, N. Peppes, and T. Alexakis, "An advanced abnormal behavior detection engine embedding autoencoders for the investigation of financial transactions," *Information*, vol. 12, no. 1, p. 34, 2021.
[28] G. Kurniawan, "Differences abnormal return and cumulative abnormal return financial sector issuers for the previous period and time of the covid-19 pandemic," *Financial Management Studies*, vol. 1, no. 2, pp. 1–11, 2021.
[29] X. Yan, H. Liu, G. Xin, H. Huang, Y. Jiang, and Z. Guo, "Research on real-time elimination of ultra-wideband radar ranging abnormal value data," *Geoscientific Instrumentation, Methods and Data Systems*, vol. 10, no. 2, pp. 153–160, 2021.
[30] S. Tosiri, T. Sethjinda, and N. Tangjitprom, "Abnormal return on stock split-revisiting the evidence of Thailand during 2009-2018," *AU-GSB e-Journal*, vol. 13, no. 2, pp. 24–37, 2020.
[31] H. Chan, K. Hee, and J. Wang, "Abnormal audit fees and stock price crash risk," *International Journal of Economics and Business Research*, vol. 22, no. 1, pp. 54–74, 2021.
[32] A. Sakiyani and N. Salehi, "Effects of financial constraints on crash risk of future stock price in view of the effects of abnormal accruals," *Journal of Knowledge Accounting*, vol. 10, no. 1, pp. 67–90, 2019.
[33] A. Onan and M. A. Tocoglu, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.
[34] S. Korukoğlu, A. Onan, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing and Management*, vol. 53, no. 4, pp. 814–833, 2017.
[35] M. Toaar, B. Ergen, and Z. Cmert, "Waste classification using AutoEncoder network with integrated feature selection method in convolutional neural network models," *Measurement*, vol. 153, p. 107459, 2019.
[36] S. Wang, J. Chen, and W. Guo, "Structured learning for unsupervised feature selection with high-order matrix factorization," *Expert Systems with Applications*, vol. 140, no. Feb., pp. 1–11, 2020.