Research on Financial Fraud Identification Based on Machine Learning Algorithm

Yihan Wang[1], Tianyu Zhang[2], Yue Xie[3], Shiyu Cui[4]
826960941@qq.com[1], 246413182@qq.com[2], 205427646@qq.com[3], 2216114159@qq.com[4]

Soochow university[1], Soochow university[2], Soochow university[3], Soochow university[4]

Abstract. In recent years, many cases of financial fraud have occurred at home and abroad. The repeated cases of financial fraud not only bring serious property losses to investors, but also seriously affect the credit mechanism of the capital market and slow down the process of healthy development of market economy, so how to efficiently and accurately identify enterprises with financial fraud is a hot topic of academic research. In this paper, a comprehensive identification model is formed by using decision trees in machine learning algorithms as well as artificial neural networks for model construction training. The results show that the accuracy of the two algorithmic models improves after importance selection, with 74.2% and 90%, respectively. It is finally concluded that the importance selection is beneficial to improve the accuracy of the financial fraud identification model, and the artificial neural network can demonstrate better identification results in the algorithm selection.

Keywords: financial falsification; machine learning; financial indicators; importance selection

1 Introduction

Since entering the 21st century, the continuous occurrence of financial fraud of listed companies has, to a certain extent, undermined the stability of social economy. From the Enron incident in the United States at the beginning of the 21st century to the German payment giant Wirecard, there are financial frauds in China such as Xiangyirongtong and Swertia Island, financial frauds at home and abroad keep The financial frauds at home and abroad have been occurring continuously, and the means of implementing financial frauds are abundant, which has affected the global economic order to a certain extent and is not conducive to the healthy development of the economy. From the viewpoint of formal system, at present, with the gradual improvement of market economy and the adjustment and optimization of industrial structure, the accounting system regulations in China are not yet sound, which provides opportunities for accounting fraud. And with the advent of the era of big data analysis, the techniques of financial falsification are becoming more and more concealed. Previously, people mainly used manual logical relationships between statement items to study financial data falsification and look for contradictions to find financial data falsification problems, but this way not only requires more experience of auditors, but also is time-consuming and laborious. Since the number of listed companies in China is growing exponentially nowadays, combining machine learning algorithms can build financial data falsification models more efficiently, analyze or predict large samples of financial data, and use the computer's ability to quickly process massive amounts of information to achieve

MSIEID 2022, December 09-11, Chongqing, People's Republic of China

DOI 10.4108/eai.9-12-2022.2327705

financial data identification pairing. Therefore, this paper mainly uses machine learning algorithms to carry out research on financial data falsification identification models, and seeks to find a suitable financial data falsification identification model so as to improve the effectiveness and accuracy of auditing.

2 Construction of financial fraud identification model under machine learning algorithm

2.1 Data pre-processing

In this paper, there are 3300 independent and dependent variables in the original data. In order to make the model analysis better, three data optimization steps, namely missing value checking, outlier processing and normalization processing, are added in this paper.

In order to facilitate data processing, this paper adopts zero-mean normalization as the normalization process, which is a process of normalizing data as decimals between zero and mean, transforming abstract and complicated expressions into pure values, which is beneficial to the model output results. The formula algorithm is as follows.

$$X^{*} = \frac{X_{\mathrm{i}} - X_{\mathrm{min}}}{X_{\mathrm{max}} - X_{\mathrm{min}}} \tag{1}$$

```
def transform_to_1(data):
new_data =(data- datamin())/(data.max()- datamin0)#
return new_data
```

After the raw data are processed, the training and test sets are randomly divided in a 7:3 ratio. After data pre-processing and importance selection, the dataset is ready for model training. The dataset is fed into three different models, namely decision tree, random forest and artificial neural network, for training. The main codes for the dataset input to each model are.

```
data = pd.read_excel(’./data2/Summary table of data statistics.lsx'converters={3:str})
```

2.2 Construction of financial fraud model based on decision tree algorithm

2.2.1 Model principle

Decision tree is a kind of tree structure with inverted shape, in which the samples are classified as the same category in each sub-node through different classification conditions and different attribute judgments from the main node through each branch node.

How to choose the decision conditions for the current node is very important. Different decision conditions will result in different data contained in the sub-nodes resulting from the application of this decision condition, which means that the tree structure is also completely different. Therefore to quickly find the most suitable decision condition to make the most efficient decision tree, then a new concept - Entropy (Entropy) should be introduced. The formula for information entropy is.

$$Entropy = -\sum_{k=1}^{N} P_k \log_2(P_k) \tag{2}$$

P: probability of different categories at the current node.

Information entropy is a tool to measure the uncertainty within a node, representing is the richness of the sample categories under the branch, uncertainty increases with the increase of information entropy, the larger the information entropy the more evenly the samples are classified. Our purpose of applying different decision conditions is to distinguish different categories of samples in different nodes as much as possible, thus ensuring the consistency of information within a single node, i.e., the entropy is as small as possible. So a new concept, the gain of entropy, can be introduced again. The entropy gain is the sum of the entropy of the previous layer minus the entropy of the current layer.

$$G\text{（}D\text{）} = H(X) - \sum_{V=1}^{V} \frac{|D^V|}{|D|} H(D)^V \tag{3}$$

2.2.2 Model tuning

In the parameter adjustment of the decision tree, information entropy is used as the feature selection criterion in this paper. In order to determine the best division point in each feature division point, the parameter of feature division criterion is set to "best". In order to solve the problem of overfitting, the maximum depth of the decision tree is set to 7 layers. In order to prevent the training set from having too many samples in a certain category, which would cause the decision tree model to be overly biased towards that category during training, the balanced algorithm is used to calculate the weights and thus specify the weights of each category in the sample.

```
model = DecisionTreeClassifier(criterion='entropy'

class_weight='blanced'max_depth=7

splitter="best"

random_state=2)
```

2.3 Construction of financial fraud model based on artificial neural network algorithm

Artificial neural network is an algorithm that imitates the way the human brain processes information, so it has brain-like information processing capability. From the viewpoint of information processing, the model established by using mathematical and physical means is called artificial neural network.

In this paper, the artificial neural network model is constructed by determining the number of neurons in the hidden layer through a grid search method, with 150 neurons in the first layer and 200 neurons in the second layer. The activation function of the hidden layer (Relu) is selected by default. Relu is a common activation function expressed in the following form.

$$f(x) = \max(0, x) \tag{4}$$

As can be seen from the expression, the Relu function is a function that expresses the maximum value, while setting the seed used by the random number generator to 67.

```
params ={hidden_layer_sizes': (150200)'activation':['relu'],'solver':
['lbfgs'],'random_state':[67]}

model = MLPClassifier(**params)
```

3 Identification results of financial fraud model based on machine learning algorithm

3.1 Financial fraud identification model based on feature selection and decision tree algorithm

The results of the decision tree model are analyzed and judged. The accuracy and AUC values of the decision tree model in the training set reached 83.33% and 71.43%, respectively, with a recall rate of 50% and an F1 score of 57.14%. The decision tree model has 83.33% accuracy, which is a good performance. However, due to the F1 score of 57.14% and 50% recall rate, it shows that this model is not effective in identifying financial frauds.

The decision tree model on the test set had 74.2% accuracy (9.13% decrease compared to the training set) and 70.3% recall (20.3% increase compared to the training set). In decision tree model effect recognition, the performance of recall on the test set versus the training set fluctuates widely and has low referability. The accuracy of the decision tree model on the test set decreases, and it can be initially judged that the decision tree model is generally effective in the application of financial fraud identification.

3.2 Recognition results of financial fraud model based on feature selection and artificial neural network algorithm

The accuracy of the artificial neural network model in financial falsification recognition is 88% on the training set, with an AUC value of 88.74%, a recall rate of 84.62%, and an F1 score of 84.62%.

The artificial neural network model has an accuracy of 90% on the test set (a 2% increase over the training set performance) and a recall of 88.8% (a 4.18% increase over the training set). The artificial neural network had the highest accuracy of the two models, with satisfactory performance and a recall rate of 88.8%, much higher than the decision tree model.

4 Conclusion

This paper has referred to and studied a large amount of literature in the process of research, summarized and found that it is still difficult to identify financial fraud today, and that identifying data is more complex under the influence of the big data environment. During the learning process, it was found that the deficiency in the use of combined financial and computer methods, that is, most of the use of a single algorithm for identification and analysis, this paper aims to use a combination of multiple algorithms for in-depth learning.

The formation of financial fraud is hidden and requires a period of time, according to the model identification output, there may be fraudulent behavior of the relevant company stock code, the relevant personnel can carry out a deep investigation of whether they have financial fraudulent behavior. The financial fraud identification model has narrowed the scope of review to a certain extent. Compared with the traditional financial fraud identification, the use of model identification can carry out daily screening of financial fraud, avoiding the turmoil caused by sudden bursts of mines, and improving the efficiency of the work of relevant personnel, effectively avoiding financial fraud.

References

[1] BolognaJ.Theoneminutefraudauditor[J].1989,8(1):29-31.

[2] LiXingGe,PeiDanPing. A study on the motives, means and identification methods of financial fraud in listed companies[J]. National Circulation Economy, 2021(33): 160-162. DOI:10.16834/j.cnki.issn1009-5292.2021.33.024.

[3] Huang S. C. The old yardstick can't measure the new economy--On the deterioration and salvation of accounting information relevance [J]. Contemporary Accounting Review,2018,11(04):1-23.

[4] Liang Gonglord. Research on early warning of corporate financial fraud identification[D]. Sichuan University,2021.Xie Xiaoying,Sun Yandong. Research on the influence factors of accounting fraud based on fraud triangle theory[J]. Finance and Accounting Communication,2014(27):3-7.

[5] Liu, Jin. The application and prospect of machine learning in financial fraud prediction[J]. Journal of Wuhan Shipbuilding Vocational Technology College,2021,20(03):155-159.

[6] CALDERONTG,GREENBP.SignalingFraudbyUsingAnalyticalProcedures[J].TheOhioCPAJournal,1994,53(4): 27-38.

[7] Zhang Chi,Guo Yuan,Li Ming. A review of artificial neural network model development and applications[J]. Computer Engineering and Applications,2021,57(11):57-69.

[8] Qin Jiangping,Zhang Man. Research on financial early warning of listed companies--analysis of data based on Xinjiang[J]. Finance and Accounting Newsletter,2011(36):83-85.

[9] Hu, A. N.. Can education make us healthier--a comparative analysis of urban and rural areas based on the 2010 China Integrated Social Survey[J]. China Social Science,2014(05):116-130+206.

[10] BreimanL.RandomForests[J].MachineLearning,2001,45(1).