Financial Prediction Model Based on Data Mining Technology

Yu Ren*

shenxuan1404@163.com

Shandong Institute of Commerce and Technology 250103, Shandong Provice, China

Abstract. It is difficult for traditional audit methods to find various financial problems hidden in massive data. With interdisciplinary thinking, this paper uses data mining technology, fault diagnosis theory based on expert knowledge and financial audit theory, and uses data mining technology in financial forecasting to build three single financial models, namely, BP neural network, support vector machine and random forest model, Through experimental analysis, the most suitable model for use in financial auditing and the support vector machine model have been identified, which can improve efficiency and reduce risks for future financial auditing work.

Keywords： Financial Audit, Risk Prediction, Data Mining, Support Vector Machine

1.Introduction

With the continuous development of China's social economy, especially after the outbreak of the COVID-19 in 2020, the normal operation order of the whole society was disrupted, the market consumption demand was depressed, the majority of small and micro enterprises stopped production, and many medium and large enterprises delayed the resumption of work, which brought great impact to the domestic economy.

Based on the market share of some listed companies being divided by rising small and medium-sized enterprises, business performance has significantly declined, financial conditions have deteriorated, and financial difficulties are gradually being faced.It first appeared in the 1950s and was mainly used to analyze the business databases of various units to discover valuable knowledge. The Z-score model proposed by Altman (1968) [2] can be said to be groundbreaking. The article proposes that a company is a comprehensive entity, and various financial ratio indicators are related. The impact of different financial indicators on a company's financial risk is also different. It is necessary to summarize different warning indicators and obtain a discriminant value to evaluate whether the company will fall into financial difficulties overall. Odom and Shard (1990) [3] first used artificial neural networks to establish warning models, and their research results showed that this model outperforms traditional warning models based on methods such as logistic regression. Sumaira et al. (2019) [4] aimed to compare the accuracy of traditional financial distress prediction models, with sample companies from an emerging capital market such as Pakistan. The research method involves constructing a multivariate discriminant analysis model for financial distress and financial stability companies, and then comparing the prediction accuracy to test the prediction

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334648

accuracy of the entire sample before, during, and after the financial crisis. Nur et al. (2020) [5] used 100 financial failure samples from the manufacturing sector of the Indonesian Stock Exchange to test various models for predicting the occurrence of financial failures in companies. The accuracy of analyzing artificial neural network models, logistic regression, and multiple discriminant analysis was compared. The research results show that the prediction performance of the neural network model is superior to the other two methods. Mu Yen Chen (2011) [6] pointed out that if executives intentionally whitewash financial statements, there is no possibility of financial distress being observed in the short or long term. Since the promotion of support vector machines at the end of the last century, their excellent classification performance has been highly praised by many scholars. C Chen et al. (2020) [7] used non-financial information to improve traditional support vector machine based warning models, grouping data according to the size of the company.

In summary, the application of data mining technology in the field of auditing has received widespread attention from researchers, but most of it still remains in the theoretical exploration stage. Developing a convenient, efficient, and intelligent auditing tool meets the needs of practical auditing work. This article mainly focuses on the use of data mining techniques in financial statement audit analysis models.

2.Overview of relevant concepts

Data mining: Data Mining refers to the process of discovering hidden information in a database that has potential value once conditions are ripe through certain means [7]. The discovered information should be identifiable, understandable, and easy to use; The discovered information only needs to be able to achieve the required decision support [8].

Data mining is a data analysis method that goes deeper than regular data analysis. Data mining is the process of obtaining useful information through deep analysis from a large amount of data.

Data mining process: There is no standard regulation regarding the process of data mining, but there is not much difference in the general steps.

(1) Data collection

Data collection is the process of abstracting the feature information required for data analysis based on previously defined analysis objects, and then collecting the information into a database through certain means.

(2) Data preprocessing

Data preprocessing is mainly divided into three small steps: data integration, data cleaning, and data transformation [10]. For the purpose of this article, it is to classify and store the collected financial information into a database according to certain classification standards [11]. If the data analysis process is completed manually, the importance of data transformation is not obvious, but if the data analysis is completed by computers, data transformation is a very important step [12].

(3) Data analysis

Data analysis in data mining refers to the process of selecting appropriate analysis tools based on the characteristics of data information in a data warehouse, and ultimately obtaining useful analysis information. Data analysis is usually divided into two categories: one is predictive, such as classification and regression; The other type is descriptive, such as association and clustering [13].

Decision Tree Algorithm: Decision trees are used to split samples by selecting the elements that maximize the "difference" between each group during each decision. That is to say, based on the characteristics of the analyzed data, logical classification is carried out to objectively judge the sample elements, and then intuitively display the importance of each element in the final decision [14].

Model building algorithm: Assuming the financial data of 34 companies were extracted for data verification, the calculation method for the per capita expenditure and total per capita expenditure of account i in enterprise m is as follows：

$$f_{m,n,i,j} = \frac{S_{m,n,i}}{N_{m,n,i}} (i = 1,2,...,40 \quad m = 1,2,...,34 \quad j = 1,2,3 \quad n = 1,2,3) \tag{1}$$

$$F_{n,i,j} = \frac{\sum_{m=1}^{34} S_{m,n,i}}{\sum_{m=1}^{34} Q_{m,n,j}} (i = 1,2,...,40 \quad m = 1,2,...,34 \quad j = 1,2,3 \quad n = 1,2,3) \tag{2}$$

F represents the per capita expenditure of enterprise m in the nth year; S represents the annual expenditure; F represents the total per capita expenditure of subject i for 34 enterprises.

To identify the enterprise with the highest degree of financial data anomaly, equation (3) will be used to calculate the degree of difference and identify the enterprise with abnormal fluctuations and significant differences from other enterprises [15].

$$E_{m,n,i,j} = E^{*}_{m,n,i,j} * \frac{f_{m,n,i,j} - F_{n,i,j}}{F_{n,i,j}} \tag{3}$$

$$E^{*}_{m,n,i,j} = \frac{f_{m,n,i,j} - M_{m,i,j}}{M_{m,i,j}} \tag{4}$$

$$M_{m,i,j} = \frac{\sum_{n=1}^{3} f_{m,n,i,j}}{3} \tag{5}$$

Among them, E * represents the degree of difference within the subject itself; M is the average value of the subject for 3 years.

After obtaining the difference value, sort it and select the top 20% universities with significant differences by default (this proportion can also be adjusted by auditors) as the key audit targets.

3.Model construction based on data mining

Selection of data mining methods: In this paper, BP neural network, support vector machine and random forest are selected as recognition models. The sample data consists of 1195 items and 35 eigenvalues, which belong to non-linear, small sample size, and high eigenvalues samples. These three algorithms well correspond to the characteristics of the sample in this paper, so these three algorithms are selected to build the Audit risk identification model.

Model construction process:

(1) Data collection. After determining the sample object, collect the sample object data from relevant web portals, platforms and databases, mainly by keyword search, web crawler, etc.

(2) Data feature filtering. According to the existing literature on Audit risk, select the feature variables suitable for classification to avoid redundancy and complexity and improve the performance of the classifier.

(3) Sample data preprocessing. Sample data preprocessing is the use of certain technical means to standardize data that does not meet experimental specifications. Common means include attribute construction, data discretization, etc; Data reduction achieves sample data compression by reducing useless features and retaining them. The commonly used method is numerical reduction.

(4) Data grouping processing. Use random functions to group the sample data into two parts: the training set and the testing set, and generate corresponding label sets respectively. The former is used for model learning and training, while the latter is used for model testing.

(5) Build a single Audit risk identification model.

(6) Result collection. After the training and testing of the training set, calculate the recognition accuracy of the training set, sample set, and overall sample set.

(7) Analyze calculation errors and experimental results. By comparing and analyzing the results of model testing with expected values, adjusting algorithm parameters in a timely manner, continuously optimizing model performance, and ultimately achieving optimal classification and recognition model performance.

4.Experimental results

Analysis of BP neural network results:

Table 1. BP neural network recognition model recognition results table

Identifying objects	Correct quantity	Number of samples	Recognition accuracy
Training set	697	797	87.45%
Sample as a whole	1039	1195	86.94%
Test set	342	398	85.93%

As shown in Table 1, when using the learned recognition model to identify the test set data, although the accuracy rate of the comparison training set data has declined, it still has an

accuracy rate of 85.93%, which indicates that the Audit risk recognition model has an accuracy rate of 85.93% when facing unfamiliar information。

Analysis of the results of random forest recognition model:

Table 2. Recognition results of random forest recognition model

Identifying objects	Correct quantity	Number of samples	Recognition accuracy
Training set	651	797	81.68%
Sample as a whole	966	1195	80.84%
Test set	315	398	79.14%

As can be seen from Table 2, similar to the BP neural network recognition model, the recognition accuracy of random forest recognition model training set, sample as a whole and test set also shows a downward trend, but the accuracy of the three is different from that of the BP neural network model, and the overall accuracy is 6.1 percentage points lower than that of the BP neural network model. There is a significant difference in performance, and algorithm improvement is still needed.

Analysis of Support Vector Machine Model Results:

Table 3. Support Vector Machine Recognition Model Recognition Results Table

Identifying objects	Correct quantity	Number of samples	Recognition accuracy
Training set	728	797	91.47%
Sample as a whole	1089	1195	91.12%
Test set	361	398	90.70%

After 1195 data tests, as shown in Table 3, the accuracy rate in the training set sample is 91.47%. In the overall sample, the accuracy is 91.12%. In the test set sample, the accuracy is 91.45%.

Comparative analysis of models

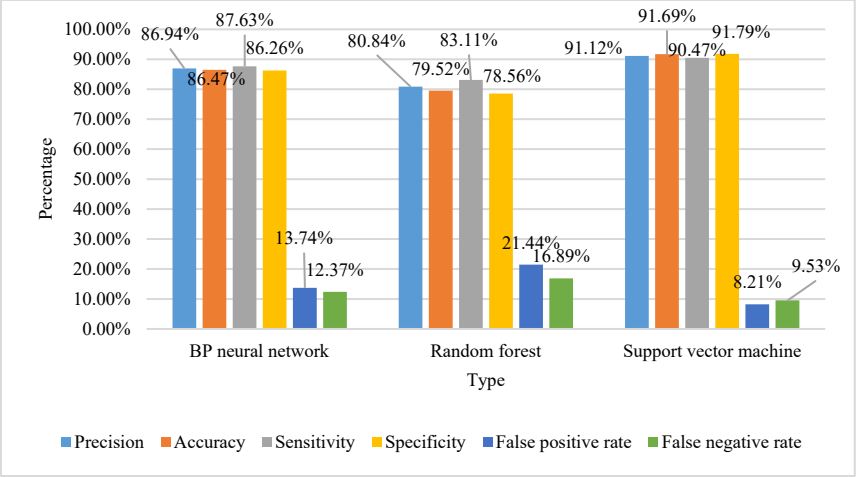

Figure 1. Comparison of Three Recognition Models

From the above table, we can see that the accuracy of the three models is more than 80%, which indicates that these models have a good effect on Audit risk identification. Among them, the accuracy of the support vector machine model is 10.18% higher than that of the random forest model and 4.18% higher than that of the BP neural network model.

5.Conclusions

At present, the market environment is constantly changing, and corporate finance is facing a more diverse and complex audit environment. In this context, this article introduces data mining technology to establish recognition models, uses D-S evidence theory to fuse the model results, and constructs three recognition models. Through experimental analysis of three single recognition models, the support vector machine recognition model has the highest accuracy, reaching 91.12%.

References

[1]Erkossa T , Laekemariam F , Abera W , et al. Evolution of soil fertility research and development in Ethiopia: From reconnaissance to data-mining approaches[J]. Experimental agriculture, 2022(58-):58-58.

[2]Altman E I. Financial Ratios, Discriminant Analysis And The Prediction Of Corporate Bankruptcy[J]. Journal of Finance, 1968, 23(4): 589-609.

[3]Odom, M. and Sharda, R. A Neural Network for Bankruptcy Prediction[J]. International Joint Conference on Neural Networks, 1990(2): 163-168.

[4]Sumaira Ashraf, Elisabete G Félix, Zélia Serrasqueiro. Do Traditional Financial Distress Prediction Models Predict the Early Warning Signs of Financial Distress?[J]. Journal of Risk and Financial Management, 2019(4).

[5]Triasesiarta Nur, Rosinta Ria Panggabean. Accuracy of Financial Distress Model Prediction: The Implementation of Artificial Neural Network, Logistic Regression, and Discriminant Analysis[A], Proceedings of the 1st Borobudur International Symposium on Humanities, Economics and Social Sciences (BIS-HESS 2019)[C], 2020, 402-406.

[6]Mu-Yen Chen, Predicting corporate financial distress based on integration of decision tree classification and logistic regression[J]. Expert Systems with Applications,2011(09):11261-11272.

[7]C Chun Chen, C Da Chen, Donald Lien. Financial distress prediction model: The effects of corporate governance indicators[J]. Journal of Forecasting, 2020(08): 238-252.

[8]Yurtay Y , Ayanolu M . Dashboard application model in supplier evaluation by using artificial immune system and data mining methods[J]. Pamukkale University Journal of Engineering Sciences, 2021, 27(2):162-172.

[9]Gao B , Balyan V . Construction of a financial default risk prediction model based on the LightGBM algorithm[J]. Journal of Intelligent Systems, 2022, 31(1):767-779.

[10]Kim J H . A Study on Estimation and Prediction of Vector Time Series Model Using Financial Big Data (Interest Rates)[J]. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(5):309-316.

[11]Munoz-Izquierdo N , Laitinen E K , Camacho-Minano M , et al. Does audit report information improve financial distress prediction over Altman's traditional Z-Score model?[J]. Journal of International Financial Management & Accounting, 2020, 31(1):65-97.

[12]Ragab Y M , Saleh M A . Non-financial variables related to governance and financial distress prediction in SMEs–evidence from Egypt[J]. Journal of Applied Accounting Research, 2022, 23(3):604-627.
[13]Tyagi S , Qu B . An Intelligent Internet of Things aided Financial Crisis Prediction Model in FinTech[J]. IEEE Internet of Things Journal, 2021, PP(99):1-1.
[14]Podhorska I , Vrbka J , Lazaroiu G , et al. Innovations in Financial Management: Recursive Prediction Model Based on Decision Trees[J]. Marketing and Management of Innovations, 2020(3):276-292.
[15]Betancourt B , Boyd N , Chandler R , et al. Modelling and prediction of financial trading networks: an application to the New York Mercantile Exchange natural gas futures market[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2020, 69(1):195-218.