Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 1314798, 12 pages
https://doi.org/10.1155/2022/1314798

Research Article

A Study on Early Warning of Financial Indicators of Listed Companies Based on Random Forest

Zilin Wang

College of Professional Study, Northeastern University, Boston 360 Huntington Ave, Boston, MA 02115, USA

Correspondence should be addressed to Zilin Wang; wang.zilin@northeastern.edu

Received 8 July 2022; Revised 24 July 2022; Accepted 17 August 2022; Published 20 September 2022

Academic Editor: Wen-Tsao Pan

Financial crises can have a negative impact on business operations, and in serious cases, they directly affect the survival and growth of a company. Therefore, the study of financial early warning based on financial indicators is particularly important. However, there are still some shortcomings in the current research on financial early warning, for example, it still evaluates the scoring method or only uses a single model to participate in the construction of financial early warning algorithm. In view of the above problems, this study will mainly use the random forest method combined with the decision tree algorithm to study the financial early warning problem of listed companies in China. Firstly, this paper uses the literature review method to analyse the relevant literature and generate the financial indicator system for this study. Subsequently, by collecting the financial data of A-share listed companies in China from 2013 to 2018 as the research object, the importance ranking of financial indicators was generated by using random forest modelling after data preprocessing. On this basis, CART decision tree modelling was applied to generate financial indicator early warning determination rules and analyse them. The results of the study show the importance ranking of financial indicators and the six financial warning rules based on the CART decision tree. Through this research, it is expected to achieve the objective of providing early warning for the risk of financial crisis and to provide constructive financial warning solutions for relevant stakeholders.

1. Introduction

The establishment of early warning indicators for financial risks and the exploration of related models have always been a priority in corporate financial management. As economic globalisation progresses, competition between enterprises intensifies, and the market becomes saturated, as well as the risks to the survival and development of enterprises increase. If an enterprise is not prudent enough in the internal control of financial risks, it may lead to an avalanche of financial crises. In many cases, it is the internal financial situation that has led to the downfall of a company. Therefore, early warning analysis, alerting, and control of the financial situation have become the key to the financial management of many companies.

There are many kinds of methods and models for conducting financial early warning. This paper will use the random forest method to rank the importance of a number of financial indicators on the screened financial data of listed companies in China. In addition, this paper will use the CART algorithm to analyse the factors influencing the generation of financial crises of listed companies in China based on the ranking of the importance of financial indicators. Based on the results of the final financial warning rule construction, the paper will be summarised and evaluated in order to obtain a reasonable and effective financial warning solution that is practically meaningful to the relevant stakeholders of the listed companies in China.

2. Theory and Methodology of Financial Early Warning

2.1. Theory of Financial Early Warning. In general, the financial situation of a company is a key lifeline for its sustainability. As a result, early warning of financial crises is a key element of the financial management process.

In the early stage of financial early warning research, Fitzpatrick [1] was a pioneer in introducing quantitative analysis to financial risk early warning, with his innovative classification of financial indicators and the related study of a sample divided into insolvency and noninsolvency data sets. At the same time, he used a single ratio of corporate financial data as the independent variable factor and combined it with an innovative method of multiratio fusion analysis. Beaver [2] then identified better financial warning indicators such as gearing and return on assets. This led to the development of an effective early warning model using a fusion of multiple ratios.

Subsequently, around 1980, Altman [3] designed the Z-value model, which was very new at the time and applied it to the field of financial early warning. The Z-score was generated by selecting factors as discriminatory variables to determine the financial condition (insolvency or otherwise) of a company. A few years later, Altman et al. [4] developed the ZETA model after nearly six years of research, collecting and studying financial data from over fifty insolvent companies at the time. Later, in order to overcome the shortcomings inherent in linear discriminant models, Martin [5] made an innovative use of logistic models in his study on predicting bank failures.

Around the 1990s, Zmijewski [6] followed the logistic model made by Martin and developed the complex Probit model. In addition, the study of financial early warning based on neural network technology is also worthy of attention. Lapedes and Farber [7] first applied neural network techniques to early warning and analysis of credit risk in banking practice. Several years later, Wilson and Sharda [8] also used neural network techniques in their study of corporate operating insolvency risk, building models with an accuracy of 97%. At the same time, they compared various models used to identify corporate insolvency and finally found that the neural network-based technique had some advantages.

At the beginning of the 21st century, Breiman [9] first proposed the random forest algorithm based on the basic machine learning algorithm decision trees. Since then, Richard et al. [10] has argued that the random forest algorithm has high classification accuracy, as well as the advantage of high flexibility in unsupervised learning, regression, and classification. In terms of innovation, Bernard et al. [11] proposes an innovative dynamic random forest algorithm (DRF), which is a random forest induction algorithm developed based on adaptive decision tree induction. By guiding decision tree induction to enrich existing decision trees as much as possible to enhance the sampling rate, satisfactory prediction rates are eventually obtained. Based on previous research, Hapfelmeier [12] et al. concluded that the random forest algorithm can achieve good prediction rates for this after a large number of experiments related to indicator data that need to be integrated and complex indicator correlations. In addition, in terms of neural network-based technology, Yang [13] et al. applied a forward neural network model (with a three-layer model of the output layer, hidden layer, and output layer) for early warning of financial crises of enterprises, and the results surface that the method is effective. Thereafter, Liu and He [14] optimized and built an artificial network financial early warning model.

In the last three years, research on financial early warning using data mining methods has remained popular. In 2020, Liu [15] et al. proposed a financial early warning model based on the AdaBoost strong classifier and selected 1350 groups of enterprise financial data for classification. The experimental results showed that the accuracy of the AdaBoost-based strong classifier was higher than that of the BP neural network-based weak classifier. In 2021, Jia [16] constructed a corporate financial crisis early warning model based on time series and the random forest algorithm and proposed an improved K-fold random forest algorithm, whose model accuracy improved by 1.54% compared with the traditional random forest model. In 2022, Wang [17] et al. applied the Wilcoxon rank sum nonparametric test and principal component analysis for feature engineering and used the logistic regression model to study the financial early warning of forestry listed companies in Shanghai and Shenzhen A-shares in China, and the accuracy of their model reached more than 80%.

It is clear from the above literature that financial early warning is still highly dependent on the construction of models and that the key to the construction of each model is many input financial indicators. At the same time, we can also see that random forest has significant advantages in the selection of indicators for the models. Therefore, this study aims to combine the random forest algorithm and the CART tree model in order to analyse the factors influencing the financial crisis of the listed companies in China.

2.2. Approach to Financial Early Warning. Financial early warning is a key task in the financial management of an enterprise. In the financial practice of enterprises, financial early warning usually requires the selection and setting of indicators in advance of the relevant financial data. Based on these predefined sets of indicators and according to certain judgment rules, the occurrence of financial crises is monitored and predicted in advance. The use of computer technology (e.g., decision trees, neurons, and random forests) is therefore almost always an important part of the process of detecting and forecasting financial indicators.

3. Design of the Financial Early Warning Method Based on Random Forest

3.1. Random Forest Based Indicator Selection

3.1.1. The Idea of a Random Forest. A random forest is the result of multiple optimisation of decision trees, both in terms of the integration of decision tree forests with the idea of rich randomisation. The basic principle is that, firstly, the basic elements of each random forest are both mutually unrelated decision trees $t_1(x)$ and $t_2(x)$, $t_3(x)$, ..., $t_n(x)$. Since they are called decision trees, they may contain both imprecise and efficient binomial trees and precise and inefficient multinomial trees. We then introduce our dataset

into these randomly generated decision trees to determine the classification, which we refer to as the act of building a random forest classifier. At the same time, we vote on the classification results of the relevant datasets from the many randomly generated decision trees to obtain the final classification results. In short, the purpose of building a random forest classifier is to determine which class an input dataset belongs to. The purpose of voting on the classification results is to find the classification that is chosen most often. Here, we define the total data set as $S(x) = \{S(x_{a1}, x_{a2}, x_{a3}, \ldots, x_{aF})\}_{a=1}^{N} S(x) = \{S(x_{a1}, x_{a2}, x_{a3}, x)\}_{aFa=1}^{N}$. Here, there are N subdata in the $S(x)$ dataset, and each dataset corresponds to F features as x_{a1}, x_{a2}, ..., x_{aF} etc.

Usually, the discriminatory F features $\{x_{a1}, x_{a2},..., x_{aF}\}$ is by calculating the degree of uncertainty of the sources between the different features (also known as information entropy). There are generally three decision tree methods from front to back: ID3, C4.5, C5.0, CART, whose corresponding calculated values are the information gain value, the information gain ratio value, and the Gini index value, respectively. By calculating these information entropy values, we can obtain the optimal splitting attribute at a node. If the relevant attribute value meets the discrete condition, then forking can continue. The relevant formula is as follows.

The information entropy of the total sample $S(x)$ is judged to be directly related to the purity of the source data in information theory by Shannon as follows:

$$\text{Entropy}(S(x)) = -\sum_{i=1}^{c} P_i \log_2 P_i. \tag{1}$$

In the above formula P_i is the proportion of samples of type i to the total sample $S(x)$ of the total sample. Since for the F kinds of characteristics x_{a1} , x_{a2} ,..., x_{aF} for decision bifurcation, the total sample $S(x)$ is divided into a total of k parts. The resulting calculated information entropy values as well as the information gain values are as follows:

$$\text{Entropy}(S(x_{an})) = -\sum_{j=1}^{k} \frac{|S_j|}{|S|}, \text{Entropy}(S_J), \tag{2}$$

$$\text{Gain}(S(x), x_{an}) = \text{Entropy}(S(x)) - \text{Entropy}(S(x_{an})). \tag{3}$$

Thereafter, the information gain rate discriminant, which improves on the information gain as follows, is judged with even better accuracy due to the penalty factor of information attached to its denominator.

$$\text{GainRation}(S(x), x_{an}) = \frac{\text{Gain}(S(x), x_{an})}{\text{SplitEntropy}(S(x), x_{an})}, \tag{4}$$

$$\text{SplitEntropy}(S(x), x_{an}) = -\sum_{i=1}^{a} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}. \tag{5}$$

The Gini coefficient is a discriminant of information purity invented after the information gain rate and has higher discriminant efficiency because it is free from the calculation of the log-log function, which is as follows:

$$\text{Gini}(S(x)) = 1 - \sum_{i}^{c} P_i^2, \tag{6}$$

$$\text{Gini}(S(x), x_{an}) = \sum_{j=1}^{k} \frac{|S_J|}{S} \text{Gini}(S_j). \tag{7}$$

As the three information purity discriminants above have less influence on the final result, the Gini coefficient will be chosen as the criterion for discriminating at the random forest nodes. Finally, the random forest ends when the feature attributes have been exhausted, the decision tree has been classified to its maximum depth, the Gini coefficient has reached a previously determined threshold, and the number of data sets at the end has reached a previously given value.

3.1.2. Indicator Screening Design for Random Forests. Due to the ease of operation of the random forest, the process design consists of only an input layer and an output layer. The input data contains the number of classes F of features, the total sample set $S(x)$, the size of all decision trees Tree, the depth of the tree h, and related parameters, as well as the filtering end algorithm. The output layer contains the results of the random forest feature selection and the model built by the random forest. The specific random forest process has seven steps as follows:

Step 1: for the total data set $S(x)$, expect to generate a random forest with a total of j trees $j = 1$: nTree

Step 2: using bootstrap-based bagging sampling, repeat sampling with put-back and selection of a training set of dataset size X

Step 3: select F signs at the nodal forks of the decision tree, filter and find the best features, and then divide the data set in this way

Step 4: generate all decision trees accordingly, completing the algorithm

Step 5: calculate the probability that a given uncertain sample a and is classified as N in the test set in which test learning is performed

$$P\left(\frac{N}{a}\right) = \left(\frac{1}{n\text{Tree}}\right), \sum_{i=1}^{k} h_i\left(\frac{N}{a}\right). \tag{8}$$

Step 6 : obtaining the classification error while selecting the best category N after voting

Step 7 : return the classification results generated by the random forest and the model

3.1.3. Evaluation of Indicator Screening Results for Random Forests. The process of constructing a financial early warning model using the random forest algorithm generates two pieces of data, the OOB (out-of-bag) error rate and the AUC value under the ROC curve. The smaller the out-of-bag error rate or the higher the AUC value, the more effective the random forest model will be.

3.2. CART-Based Decision Tree Model Construction

3.2.1. The Idea of a CART Decision Tree. The core of the CART decision tree algorithm is the selection of features for the original dataset and the pruning of the decision tree after forking.

For a dataset $S(x)$ CART decision tree T constructed, a total of N feature categories are covered in the dataset, and Pj is defined as the data with feature category j as a percentage of the total dataset $S(x)$ of the probability, which can be given by the following formula:

$$gini\left(S\left(x_j\right)\right) = 1 - \sum_{i=1}^{n} P_j^2. \tag{9}$$

Sort the data set at the node $S(x)$. Classify into subsets $S(x_1)$ and subsets $S(x_2)$, whose Gini index is calculated as follows:

$$gini\left(S(x_1), S(x_2)\right) = \frac{|S_1|}{S} gini\left(S_1\right) + \frac{|S_2|}{S} gini\left(S_2\right). \tag{10}$$

After selecting the best feature indicator and the value of the feature indicator at the bifurcation node by the above formula, the data set is $S(x)$. If the Boolean value of a piece of data is true, the data will be placed in the left leaf node of the bifurcation tree at that node, and if it is false, it will also be placed in the right node. By calculating and placing the data in this order, a CART decision tree can eventually be constructed.

3.2.2. Design of the CART Decision Tree. The data set is now defined $S(x)$. A total of N data exist for each category of indicators. The algorithm consists of a total of input and output layers. Among the data that should be input are the total data set $S(x)$ and relevant threshold conditions. The specific process steps are as follows:

Step 1: enter the relevant dataset data and associated threshold conditions.

Step 2: calculate the total data set at the node $S(x)$ of the overall Gini index, meanwhile, calculate the corresponding sample eigenvalues K based on different sample features K.Subsequently, this is followed by dividing the datasets $S(x_1)$ and $S(x_2)$ according to the Boolean case of the sample dataset with K features and the overall eigenvalue K,and then calculating the Gini index after the division.

Step 3: after all features with K and the feature value of k, the feature with the smallest Gini index and its eigenvalue are selected and used as the node segmentation indicator to assign the data.

Step 4: repeat steps 1 to 3 until the stop building condition is met.

Step 5: return the classification results from the CART decision tree and the model.

3.2.3. Evaluation of the Results of the CART Decision Tree. As the CART decision tree is constructed, the corresponding confidence values are calculated. In layman's terms, the confidence level reflects how trustworthy the tested data features in the total data set are compared to the tested values. It can be interpreted as both the accuracy and reliability of the algorithm. The higher the confidence level of an algorithm or rule, the better the prediction of the system.

4. Example Analysis of Financial Early Warning Based on Random Forest

4.1. Selection of the Sample. The aforementioned method of determining the financial crisis of the listed companies in China has been described in detail at this stage of the consensus of the academic community and the security industry. Therefore, this paper adopts the indicator of "whether a listed company is on the ST warning board" as the basis for judging whether a company is in financial crisis. Companies on the ST warning board (including ST companies and ∗ ST companies) are considered to be in financial crisis, while companies not on the ST warning board are considered to be in financial health. Generally speaking, a company will only be placed on the ST Alert if its net asset value per share is lower than its value per share in the market, and if the company's annual reports for two consecutive financial periods show consecutive losses (i.e., the net profit of the company's financial statements for two consecutive years is less than the total cost of ownership).

In this paper, the relevant financial data of all A-share listed companies from 2013 to 2018 were collected from the Wind database, and the data of all ST companies (148 companies in total) in these six years were extracted and sorted. Subsequently, in order to avoid the impact of data imbalance (both too few ST companies) on the subsequent experiments, 302 non-ST companies in the A-share market from 2013 to 2018 were randomly selected to jointly construct a data pool of A-share listed companies in China (a total of 450) in this paper.

4.2. Determination of Financial Indicators. By combining the statistics of the indicator results of the financial early warning indicator system in the literature reviewing process (check Table1) and the analysis of specific financial indicators of A-share listed companies in China, we finally identified a total of 24 financial indicators to be used in the random forest model. We finally determined a total of 24 financial indicators (check Table 2) of A-share listed companies in China to participate in the construction of the random forest model, and the specific indicators are listed in the following table.

Table 1: Actual financial indicators used.

Financial indicators	Variables
Gross margin of sales	Grossprofitmargin
Net sales margin	Netprofitmargin
Return on net assets	roe_avg
Total return on assets	roa2
Current ratio	Current
Quick ratio	Quick
Total equity/liabilities attributable to shareholders of the parent company	Equitytodebt
All debt/EBITDA	Tltoebitda
Net cash flows from operating activities/total liabilities	Ocftodebt
Net profit (year-on-year growth)	Yoyprofit
Total assets (year-on-year growth)	yoy_assets
Inventory turnover rate	Invturn
Accounts receivable turnover ratio	Arturn
Current asset turnover ratio	Caturn
Fixed asset turnover rate	Faturn
Total asset turnover ratio	assetsturn1
Cash received from sales of goods and services/operating income	Salescashintoor
Net cash flow per share	Cfps
Corporate free cash flow	Fcff
Gearing ratio	Debttoasset
Equity attributable to shareholders of the parent company/all invested capital	Equitytototalcapital
Earnings per share (basic)	eps_basic
Earnings per share (diluted)	eps_diluted
Net assets per share	Bps

4.3. Data Standardisation of Financial Indicators. Firstly, since the missing values account for a small proportion of the total data and are numeric, we can usually fill in missing values with a mean value. In addition, since the data are in line with normal distribution. The principle of missing value filling proposed by Anderson [18] et al. is that "under normal distribution, the sample mean is the best possible value to be estimated." Therefore, in order to ensure the integrity of the experimental data and the reasonableness of the results as far as possible, filling the mean value is the best solution at this stage.

At the same time, as the final financial indicators identified in this paper have a large number of data lines and the data vary greatly in order of magnitude, a data standardisation operation should be performed prior to data analysis. If such raw data with widely different data characteristics are directly operated, it will cause the final results to be biased towards data-based indicators in order to lead to poor model building results.

There are many methods of data normalisation, such as extreme value normalisation, Z-score normalisation, normalisation, and so on. Each of them has its own corresponding advantages. Due to the more constrained and effective nature of polar normalisation, we will now focus on polar normalisation. Extreme value normalisation, also known as extreme value difference normalisation, is the process of deflating data to between 0 and 1 by varying it by equal proportions. For data columns $x_1, x_2, x_3, \ldots, x_n$, extreme value normalisation is performed to produce a normalised data column $y_1, y_2, y_3, \ldots, y_n$, and the transformation formula is as follows:

$$y_i = \frac{x_i - \min\limits_{1 \le j \le n}\{x_j\}}{\max\limits_{1 \le j \le n}\{x_j\} - \min\limits_{1 \le j \le n}\{x_j\}}. \tag{11}$$

The raw data will fall between 0 and 1 when transformed by polar normalisation while eliminating its original dimensional limits and the effects of different orders of magnitude and facilitating the subsequent conduct of the test.

4.4. Results and Analysis of Random Forest Modelling

4.4.1. Optimization of Parameters for Random Forest Modelling. Due to the advantage of random forest in being able to handle a large number of features well, this paper puts all the financial indicators identified above into the random forest algorithm to construct the corresponding financial warning model. The 24 identified financial indicators such as gross sales margin, net sales margin, return on net assets, return on total assets, and current ratio are used as independent variables, while whether the listed company is placed on the ST warning board (hereinafter referred to as ST, with non-ST company value being 0 and ST company value being 1) is used as the dependent variable for prediction. On top of this, 70 per cent of the sample data set was divided into training samples, and the remaining 30 per cent was divided into test samples to participate in the construction of the random forest model.

Subsequently, we obtained the in-bag error rate of the model corresponding to each mtryvalue by traversing all mtry (which refers to both the number of random variables

Table 2: Statistics table of the financial warning indicators.

Indicator	Wu [22]	Zhang [23]	Zhou [24]	Meng [25]	Zhang [26]	Yang [27]	Wu [28]	Song [29]	Yang [30]	Lu [31]	Fu [32]	Wang [33]	Feng [34]	Chen [35]	Wang [36]	Counting
Inventory turnover ratio	√	√	√	√	√	√	√	√	√	√		√	√	√	√	14
Current ratio	√		√	√	√	√	√	√	√	√	√	√	√		√	13
Accounts receivable turnover ratio	√		√	√	√	√	√	√	√	√		√	√	√	√	13
Gearing ratio		√	√	√	√	√		√		√	√	√	√	√	√	12
Quick ratio	√		√	√	√	√	√	√	√	√			√		√	11
Total assets growth rate	√	√	√	√		√	√		√			√	√	√	√	11
Total assets turnover ratio	√		√	√	√			√	√	√		√	√	√	√	11
Return on net assets			√		√	√		√	√	√			√	√	√	9
Fixed assets turnover ratio			√	√	√		√		√			√	√		√	8
Current assets turnover ratio				√	√	√	√	√		√		√			√	8
Return on total assets				√		√	√	√			√	√	√		√	8
Interest cover multiplier	√						√		√				√	√	√	6
Cash ratio		√	√				√	√		√					√	6
Net profit growth rate			√	√		√				√				√	√	6
Gross sales margin				√	√		√	√		√						5
Return on total assets			√		√				√					√		4
Earnings per share			√		√		√			√						4
Net assets per share			√		√					√					√	4
EBIT debt ratio					√			√	√	√						4
Current liabilities to total liabilities			√		√			√			√					4
Operating profit margin						√			√	√					√	4
Total net asset margin							√			√			√		√	4
Growth rate of revenue from main business	√									√		√				3
Cash flow from operating activities per share				√	√				√							3
Net sales margin					√		√			√						3
Percentage of long-term debt	√						√	√								3
Profit margin from main business	√											√				2
Growth rate of shareholders' equity	√												√			2
Cash ratio from sales		√		√												2
Operating cash flow growth rate		√								√						2
Interest payment multiplier			√									√				2
Cash current liabilities ratio			√						√							2
Shareholder equity ratio				√								√				2
Shareholder equity to debt ratio					√								√			2
Operating cash debt ratio					√								√			2
Equity multiplier					√			√								2
Current assets to total assets ratio					√			√								2
Cost margin						√	√									2
Operating income growth rate						√							√			2

Table 2: Continued.

Indicator	Wu [22]	Zhang [23]	Zhou [24]	Meng [25]	Zhang [26]	Yang [27]	Wu [28]	Song [29]	Yang [30]	Lu [31]	Fu [32]	Wang [33]	Feng [34]	Chen [35]	Wang [36]	Counting
Net asset growth rate						√								√		2
Net profit total operating income ratio							√	√								2
Net cash flow from operating activities operating income ratio							√	√								2
Total operating income							√	√								2
Net cash flow from operating activities total liabilities ratio							√	√								2
EBITDA								√			√					2
Operating profit growth rate									√				√			2
Equity ratio										√					√	2

```
> print (err)
 [1] 0. 043258226 0. 010354549 0. 004983789 0. 004343646 0. 004224496 0. 004666905 0. 004177031 0. 003686504 0. 004289552 0. 004221901 0. 004164435 0. 003835454
[13] 0. 004105213 0. 003911085 0. 003952671 0. 004096224 0. 004241499 0. 003981544 0. 004461810 0. 004099474 0. 004251103 0. 004229795 0. 004324168 0. 004775342
[25] 0. 004663714 0. 004272235 0. 004845236 0. 005055579 0. 005583973 0. 005767815 0. 005177871
```

Figure 1: Results of mtry value traversal for random forest parameter search.

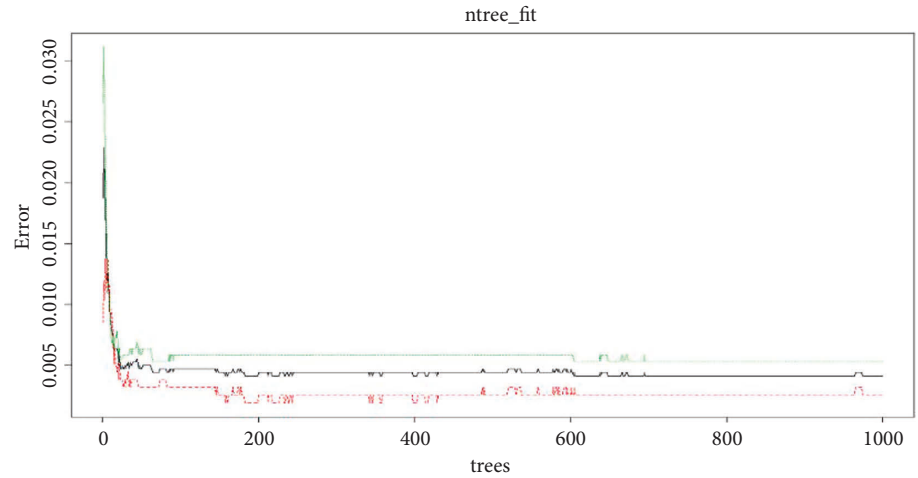

Figure 2: Results of ntree value traversal for random forest parameter search.

in the process of constructing a random forest) in the range of 1 to 24. As shown in Figure 1, the eighth modelled in-bag error rate of the parameter search reaches the lowest value, i.e., when the mtry value is 8, the in-bag error rate is 0.0036, which is the lowest value. At the same time, the in-bag error rate for subsequent iterations increases sequentially, so the optimal value of mtry for this model is 8.

Similarly, we modelled and traversed a range of ntrees (i.e., the number of decision trees in the random forest construction process) again at the lowest in-bag error rate mtry value (both 8) above to obtain a graph of the change in the in-bag error rate of the model as the ntree value changed. As shown in Figure 2, the in-bag error rate tends to stabilise when the ntree value is greater than 800.

Through the parameter search process, we found that either too high or too low mtry values will affect the final random forest prediction. At the same time, a high ntree value will cause a rapid increase in model complexity and affect the computational efficiency, while a low ntree value will directly affect the model performance. The result of the above parameter search is that the mtry value is 8 and the ntree value is 800, and the error rate of the random forest model constructed in this paper is the lowest and the best.

4.4.2. Random Forest Model Results. Using the final optimisation results of the above algorithm for the mtry and ntree values, we can determine that the best prediction efficiency of the random forest is achieved when the mtry value of the proposed model is 8, while the ntree value is 800 and is therefore modelled with these optimal parameters. Figures 3–5 show the final results of random forest modelling.

At the same time, the results of the R-language random forest operations in Figure 3 show that when the random

```
        OOB estimate of  error rate: 8.41%
Confusion matrix:
    0  1 class.error
0 192 15  0.07246377
1  11 91  0.10784314
```

Figure 3: Results of R code for random forest.

forest-based financial early warning model is constructed with the settings described above, the out-of-bag data error rate of the final model is 8.41%. From the perspective of the out-of-bag error rate, this random forest model has excellent prediction results.

The final ranking of the metrics generated according to the random forest algorithm takes two forms, one of which, the left-hand graph in Figure 4, is composed according to mean decrease accuracy (MDA). This evaluation scale is generated based on the out-of-bag error rate (OOB). Under this measure, the horizontal coordinate indicates the extent to which the prediction rate of the constructed model decreases when a variable is replaced with a random number. Thus, if the value of an indicator is larger in this indicator ranking form, it indicates that the level of the indicator is more important. Secondly, the right-hand panel of Figure 4 is constructed on the basis of the mean decrease Gini. This evaluation scale is based on the Gini coefficient. The horizontal coordinate indicates the differential impact of a feature on the observed training values at the nodes of the random decision tree when it is replaced. Similarly, a larger value of the horizontal coordinate of a feature indicator indicates that the feature is more important.

One of the evaluation rules based on mean decrease accuracy is based on the out-of-bag error rate (OOB ERROR), which embodies the core bagging algorithm idea of the random forest algorithm. It ensures that the distribution

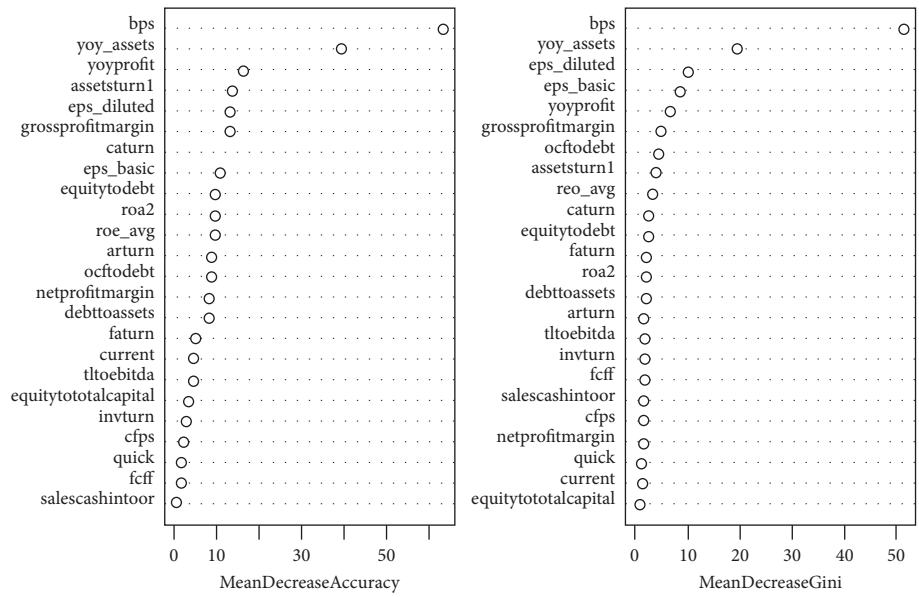

Figure 4: Ranking the importance of indicators in random forests.

of the data after feature replacement is infinitely close to the original (with both put-back resampling). In contrast, the mean decrease Gini-based evaluation rule is based on the binomial distribution idea of the CART algorithm, which is more incompatible with the overall idea of the random forest algorithm. Therefore, the importance ranking of financial indicators based on mean decrease accuracy is more scientific. The results of the specific indicator importance ranking are as follows: (1). net assets per share; (2). total assets (year-on-year growth rate); (3). net profit (year-on-year growth rate); (4). total asset turnover ratio; (5). earnings per share (diluted).

Figure 5 shows the ROC plots for the final random forest financial early warning model of this paper. As can be seen, the AUC value of the financial early warning model constructed in this paper is 0.909, which is at a fairly high value level. Therefore, from the perspective of the ROC curve, this model has a very high predictive performance.

Taken together, the financial early warning model constructed in this paper has a low out-of-bag error rate (OOB ERROR = 8.41%) as well as a very high AUC value (AUC = 0.909).

The OOB rate for the risk evaluation of real estate projects using the random forest algorithm by Li and Shenjiang [19] et al. is 10.53%. The OOB rate for the financial failure prediction of the listed companies using the random forest algorithm by Zhou [20] et al. is 26.37%. Zhang [21] used the shuffle-based random forest technology to establish an AUC value of 0.8666, while he used the embedded-based random forest technology to establish an AUC value of 0.8404.

In addition, among the latest financial early warning research results in the past three years, Liu [15] et al. used a financial early warning model based on the AdaBoost strong classifier with an accuracy of 96%, which was higher than the accuracy of the weak classifier model based on the BP neural

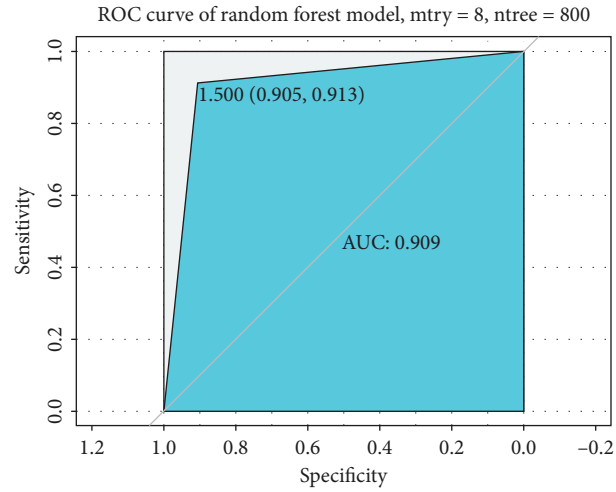

Figure 5: Plot of ROC results for random forest.

network (91.54%). Jia [16] used an improved K-fold random forest algorithm based on time series for financial early warning with an accuracy of 90.327%. The final accuracy of Wang and Lu[17] financial early warning modelling using a logistic model and a PCA method based on the Wilcoxon rank sum nonparametric test for processing financial data of A-share listed forestry companies was 86.7%.

Therefore, the OOB of the financial early warning model of the listed companies in this paper is 8.41%, and the AUC of this model is 0.909, which can be concluded that the model possesses a good predictive effect.

4.5. Discovery of Financial Warning Rules Based on CART Decision Tree Construction. Through the screening process for the listed companies' financial data columns after the standardisation process above, we imported these data into

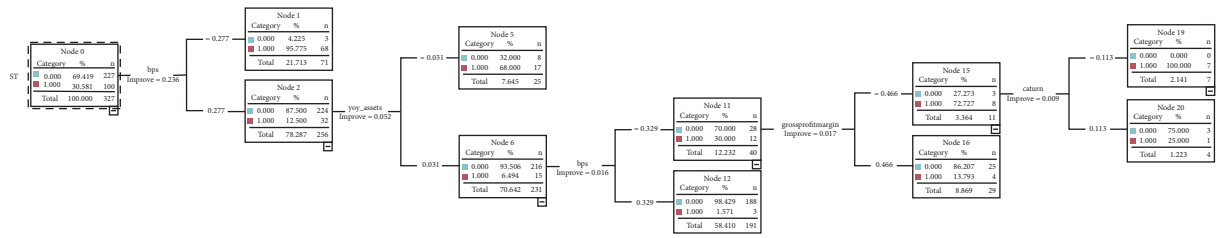

Figure 6: CART decision tree diagram.

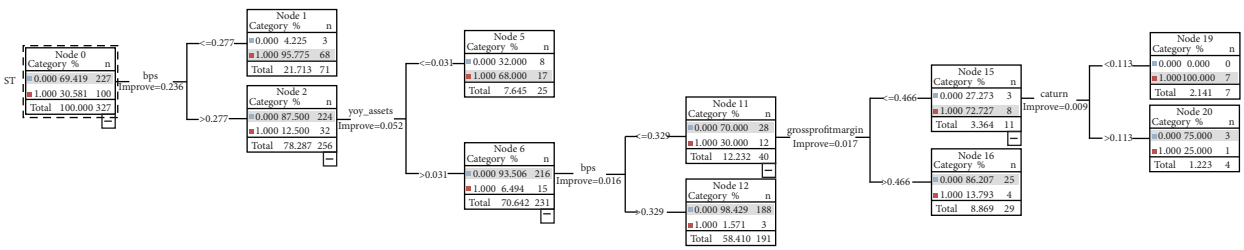

Figure 7: CART decision tree diagram.

the SPSS modeler software for rule building based on the CART decision tree, and after computing, the CART tree diagram was drawn as shown in Figure 6, in which the four financial indicators of net assets per share, total assets (year-on-year growth rate), gross sales margin, and current assets turnover ratio were retained down to participate in the construction of the CART decision tree (check Figure 7).

The early warning rules based on the CART decision tree and their confidence values are shown in Figure 6. There are three rules for determining that a company is in financial crisis (i.e., the company is ST), and at the same time, there are three rules for determining that a company is financially healthy (i.e., the company is not ST).

The rules for determining when an enterprise is in financial crisis are analysed as follows: firstly, an enterprise is considered to be in financial crisis when its net assets per share (normalised by the extreme values described above and the same for all the indicators below) are less than or equal to 0.277. The confidence level for this rule was 95.8%, and 71 data items were used for training. Secondly, a company was identified as being in financial crisis when its net assets per share were greater than 0.277, but its total assets (year-on-year growth rate) were less than or equal to 0.031. The confidence level of this rule is 68%, and there are 25 data items involved in the training; thirdly, when the net asset per share of an enterprise is greater than 0.277 and its total assets (year-on-year growth rate) are greater than 0.031, but its net asset per share is less than or equal to 0.329, its gross sales margin is less than or equal to 0.466, and its current asset turnover ratio is less than or equal to 0.113, then the enterprise is considered to be in financial crisis. This rule has a confidence level of 100% and a total of seven data items were involved in the training.

The rules for determining the financial health of an enterprise are as follows: first, when the net asset per share of an enterprise is greater than 0.277 and less than or equal to 0.329, and its total assets (year-on-year growth rate) are greater than 0.031, gross sales margin is less than or equal to 0.466, and current asset turnover ratio is greater than 0.113, then the enterprise is considered to be financially healthy. The confidence level of this rule is 75%, and a total of 4 data items are involved in the training. Secondly, when the net asset per share of a company is greater than 0.277 and less than or equal to 0.329, and the total assets (year-on-year growth rate) are greater than 0.031 and the gross sales margin is greater than 0.466, then the company is considered to be a financially healthy company. The confidence level of this rule is 86.2%, and a total of 29 data items are involved in the training. Thirdly, when the net assets per share of a company are greater than 0.329 and the total assets (year-on-year growth rate) are greater than 0.031, then the company is considered to be a financially healthy company. The confidence level of this rule was 98.4%, and a total of 191 data items were involved in the training.

5. Conclusion

In view of the actual situation of the financial data of the listed companies in China, we constructed a financial early warning model based on the random forest algorithm and the CART decision tree algorithm. The method uses the random forest algorithm to rank the importance of many financial indicators, which can significantly improve the predictive effect of subsequent modelling and is in line with the principles underlying the establishment of financial warning models. In addition, the quantitative analysis of the filtered indicators using the CART decision tree algorithm to determine the decision thresholds for specific financial warning indicators can significantly improve the operability of the model. The results of the study show that the method has an excellent predictive effect. At the same time, the method can provide the listed companies' management and

investors with a prediction model that can quantify the financial situation of the listed companies.

Data Availability

The Wind China Listed Companies Dataset was used to support the findings of this study.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] A. Fitzpatrick, *Comparison of Ratios of Successful Industrial Enterprises with Those of Failed Firms*, Certified Public Accountant, New York, NY, USA, 1932.

[2] W. H. Beaver, "Financial ratios as predictors of failure, empirical research in accounting: selected studies," *Journal of Accounting Research*, vol. 4, no. 4, pp. 71–111, 1966.

[3] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[4] E. I. Altman, R. G. Haldeman, and P. Narayanan, "ZETATM analysis A new model to identify bankruptcy risk of corporations," *Journal of Banking & Finance*, vol. 1, no. 1, pp. 29–54, 1977.

[5] D. Martin, "Early warning of bank failure: a logit regression approach," *Journal of Banking & Finance*, vol. 1, no. 3, pp. 249–276, 1977.

[6] M. E. Zmijewski, "Methodological issues related to the estimation of financial distress prediction models," *Journal of Accounting Research*, vol. 22, no. 22, pp. 59–82, 1984.

[7] A. Lapedes and R. Farber, *Nonlinear Signal Processing Using Neural Networks: Prediction and System modelling*, San Diego, CA, USA, 1987.

[8] R. L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision Support Systems*, vol. 11, no. 5, pp. 545–557, 1994.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] C. D. Richard, C. Edwards Thomas, H. Beard Karen et al., "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, 2007.

[11] S. Bernard, S. . Adam, and L. Heutte, "Dynamic random forests," *Pattern Recognition Letters*, vol. 33, no. 12, pp. 1580–1586, 2012.

[12] A. Hapfelmeier, T. Hothorn, K. Ulm, and C. Strobl, "A new variable importance measure for random forests with missing data," *Statistics and Computing*, vol. 24, no. 1, pp. 21–34, 2014.

[13] B. Yang, H. Ji, J. Xu, and J. Wen, "BP neural network in the application of enterprise financial crisis early warning," *Forecasting*, no. 2, pp. 49–54, 2001.

[14] H. Liu and G. J. He, "Research on early warning of business failure of listed companies based on artificial neural network method," *Accounting Research*, vol. 2, pp. 42–46, 2004.

[15] Y. Liu, "A financial early warning model based on Adaboost strong classifier," *Modern Business*, no. 31, pp. 187-188, 2020.

[16] L. Jia, *Research on the Application of K-fold Random forest Algorithm in enterprise Financial Crisis Early Warning*, 2021.

[17] J. P. Wang and P. G.. Lu, "Research on financial early warning of listed forestry companies based on PCA and Logisitc regression," *China Forestry Economy*, no. 4, p. 5, 2022.

[18] A. B. Anderson, A. Basilevsky, D. P. J. Hum, P. H. Rossi, J. D. Wright, and A. B. Anderson, *Missing Data: A Review of the Literature, Handbook of Survey Reasearch*, Academic Press, New York, NY, USA.

[19] X. Li and Shenjiang, "Risk assessment of real estate project based on random forest," *Journal of Engineering Management*, vol. 33, no. 6, pp. 144–149, 2019.

[20] X. Zhou, Z. Peng, and Y. Zeng, "Research on financial failure early warning of listed companies based on Stochastic Forest model," *Journal of Hunan University of Arts and Science (Natural Science Edition)*, vol. 29, no. 2, pp. 17–20, 2017.

[21] J. Zhang, *Research on Hybrid Random forest Financial Crisis Early Warning Model*, Software Guide, 2021.

[22] S. N. Wu and X. Y. Lu, "A study on the prediction model of financial distress of listed companies in China," *Economic Research*, vol. 6, pp. 46–55, 2001.

[23] J. Zhang, *Research on the Early Warning of Financial Risk of China's Real Estate Development Enterprises Based on Cash Flow*, Taiyuan University of Technology, Shanxi Province, 2015.

[24] X. Zhou, Z. Peng, and Y. Zeng, "Research on early warning of financial failure of listed companies based on random forest model," *Journal of Hunan Academy of Arts and Sciences (Natural Science Edition)*, vol. 29, no. 2, pp. 17–20, 2017.

[25] J. Meng, "Application of random forest model in financial failure early warning," *Statistics and Decision Making*, no. 4, pp. 179–181, 2014.

[26] L. Zhang, L. Zhang, Y. B. Chen, and W. L. Teng, "Application of data mining method based on information fusion in company financial early warning," *China Management Science*, vol. 23, no. 10, pp. 170–176, 2015.

[27] L. H. Yang, Q. Chen, and M. Deng, "Research on financial risk evaluation of LD group and its early warning," *Finance and Accounting Monthly*, vol. 23, no. 35, pp. 72–79, 2017.

[28] Q. H. Wu, X. H. Tang, and Y. Lin, "Identification and early warning of financial crisis of GEM listed companies," *Finance and Accounting Monthly*, no. 2, pp. 56–64, 2020.

[29] Y. Song and H. X. Li, "Research on financial early warning of bond issuing enterprises based on decision tree integration," *Journal of Finance and Accounting*, vol. 7, no. 6, pp. 45–50, 2020.

[30] G. Yang, Y. Zhou, and L. Sun, "The early warning method of enterprise financial risk based on Benford-Logistic model," *Quantitative Economic and Technical Economics Research*, vol. 36, no. 10, pp. 149–165, 2019.

[31] Di Lu and G. Wang, "Research on early warning of financial risks of agricultural listed companies--based on factor analysis method and cluster analysis method," *Friends of Accounting*, no. 24, pp. 79–83, 2019.

[32] L. Fu, "External shocks, leverage ratio and financial risk early warning model," *Friends of Accounting*, no. 11, pp. 27–30, 2019.

[33] Q. Wang and F. Ye, "Research on early warning of financial distress of ST companies under the new normal - panel data of financial reports based on C5.0 algorithm," *Finance and Accounting Communication*, vol. 21, no. 23, pp. 107–111, 2018.

[34] F. Nannan, "Early warning model establishment and analysis of corporate financial crisis," *Friends of Accounting*, vol. 17, no. 9, pp. 113–115, 2018.

[35] F. Chen and J. Wu, "A comparative study of financial crisis early warning models for small and medium-sized enterprises--a comparison based on factor analysis and logistic regression model," *Finance and Accounting Communication*, no. 5, pp. 106–108, 2017.

[36] Yi Wang and Z. Yao, "Construction and comparison of financial early warning systems of listed manufacturing companies--based on data mining techniques," *Finance and Accounting Monthly*, vol. 13, no. 21, pp. 49–55, 2016.