Financial Risk Analysis of Enterprises Based on Random Forest Algorithm

Yuan Yan[1], Zhe Yang[2], Shuai Zhang[3]
yyvvbb@163.com[1], 18663225702@163.com[2], 760357225@qq.com[3]

[1]State Grid Information and Communication Industry Group CLP Puhua Information Technology Co., Ltd[1],

[1]State Grid Information and Communication Industry Group CLP Puhua Information Technology Co., Ltd[2],

[1]State Grid Information and Communication Industry Group CLP Puhua Information Technology Co., Ltd[3]

Abstract. As an important part of China's beverage and food industry, the brewing industry plays a significant role in the national economy in terms of its socioeconomic benefits. This paper uses 36 listed enterprises in the brewing industry in 2020 as a research sample, classifies the financial quality of brewing enterprises by "financial leverage coefficient", selects three different combinations of financial indicators, and constructs a financial risk model using the random forest algorithm to determine the financial risk of brewing enterprises, and finds that the random forest model has a high prediction accuracy, and uses The model was found to have high prediction accuracy, and the model was used to analyze the influence of financial indicators related to profitability, solvency, and operating capacity on the financial risk of brewing enterprises.

Keywords: finance; random forest; risk

1 Introduction

Foreign scholars have researched financial risk since the 1930s, and focused on constructing financial risk early warning models, among which the more influential ones are: univariate model, Z-Score model and ZETA model, Logistic model, Probit model, cash flow model, neural network model, and hybrid model.

However, with the rapid development of the economy and the emergence of more and more listed enterprises in China, domestic scholars are paying more and more attention to the financial risk of enterprises, and more and more studies on the definition, causes and early warning of financial risk are being improved [1].

The random forest algorithm is a randomly selected classifier based on the CART tree model. RandomForest is a machine learning algorithm that uses the CART tree model as the base classifier, randomly selects the number of samples, feature variables and classifiers, constructs multiple decision trees, generates a forest using the Bagging algorithm set, and finally decides the category of sample data based on the voting results of all decision trees [2]. Considering the

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

DOI 10.4108/eai.9-12-2022.2327664

advantages of random forest algorithm with high accuracy, insensitive to outliers and missing values, and able to handle large amount of data, this paper will use this classification algorithm for learning and analysis.

2 Importance ranking of financial indicators

2.1 The principles of constructing financial index system

The data should be collected through the annual reports of enterprises, authoritative databases and other means to obtain the financial indicators that can be quantified. Principle of comprehensiveness. The selected financial indicators should cover as many aspects as possible, avoiding the omission of important financial indicators, and constructing a multi-level and multi-dimensional comprehensive financial indicator system.

2.2 Performance comparison of financial indicators

Since the financial quality of brewing companies is classified by the "financial leverage coefficient" in this paper, the "financial leverage coefficient" (X26) is excluded from the feature subset when comparing the prediction performance of different feature subsets. Let the original 25 financial indicators be the feature subset T0, the 12 financial indicators based on Pearson correlation coefficient be the feature subset T1, and the 11 financial indicators based on random forest be the feature subset T2, as follows [6].

Table 1 Feature subsets and included feature indicators

Subset of features	Screening Principle	Include feature indicators
T0	—	X1:X25
T1	Pearson correlation coefficient	X1, X7, X8, X11, X12, X13, X16, X17, X18, X21, X22, X23, X25
T2	Random Forest	X6, X7, X9, X10, X11, X12, X14, X15, X17, X18, X25

A random forest model was built using the randomForest package in R language with parameters set to the default state, and three different subsets of feature variables were used for learning and forecasting, resulting in the performance of financial indicators as shown in Table 2.

Table 2 Comparison of the performance of three financial indicators based on random forest

Subset of accompanying features	Total accuracy	Accuracy of companies with poor financial quality	Financial quality normal corporate accuracy
T0	91.67	100.00	66.67
T1	86.11	96.30	55.56
T2	94.44	100.00	77.78

According to Table 2, it can be seen that the prediction accuracy of all three feature subsets for enterprises with poor financial quality reaches more than 95%, among which the prediction accuracy of T0 and T1 feature subsets for enterprises with poor financial quality reaches 100%. However, the prediction accuracy of the three feature subsets for enterprises with normal financial quality varies greatly, among which T2 has the greatest prediction accuracy for enterprises with normal financial quality, reaching 77.78%, followed by T0, reaching 66.67%, and finally T1, reaching 55.56%. The total prediction accuracy of T2, T0, and T1 feature subsets was 94.44%, 91.67%, and 86.11%, respectively, influenced by the prediction accuracy of financial quality normal enterprises. Overall, the prediction results of the original feature subset T0 have been more accurate, but the prediction accuracy of the feature subset filtered based on the random forest algorithm has still increased to a certain extent, on the contrary, the prediction accuracy of the feature subset filtered based on the Pearson correlation coefficient has decreased rather than increased.

3 Construction of a financial risk model based on random forest algorithm

3.1 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.

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

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.

$$G（D）= H(X) - \sum_{V=1}^{V} \frac{|D^{V}|}{|D|} H(D)^{V} \tag{2}$$

3.2 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)
```

4 Conclusion

In this paper, 26 financial indicators are selected from six aspects: profitability, solvency, growth, operating capacity, cash flow and financial leverage of listed enterprises. Considering the small number of brewery companies that were judged as "*ST", this paper uses the indicator "financial leverage coefficient" to classify the financial status of brewery companies. The subset of features is constructed based on Pearson correlation coefficient and random forest respectively, and the prediction accuracy of the subset of features is detected by applying random forest algorithm, and it is found that the original data set already has better prediction accuracy, but the prediction accuracy of the subset of features constructed based on random forest is higher, while the prediction accuracy of the subset of features based on Pearson correlation coefficient is reduced.

In this paper, based on the feature subset constructed based on random forest, the ntree and mtry parameters of the random forest model were adjusted, and it was found that when the ntree parameter was set to about 700, the OOB error of the model was stable and unchanged, while the AUC value of the model was maximum when the mtry value was set to 1.

The financial status of enterprises is one of the most important concerns of managers, investors, creditors and other stakeholders. In this paper, we used the random forest algorithm to analyze the financial risk of brewing enterprises and optimized the parameters of the random forest model to obtain the following conclusions [7].

(1) This paper takes listed enterprises in the brewing industry as research samples, collects relevant financial indicators, uses Pearson correlation coefficients and random forest principles to screen financial indicators, finds that financial indicators related to solvency and growth capacity have a greater impact on the financial status of brewing enterprises, and subsequently explores the performance differences of comparing different combinations of financial indicators for financial risk prediction.

(2) In this paper, by compiling the principles and advantages and disadvantages of random forest, combined with R language software, we use random forest model to construct financial risk model, and find that the prediction accuracy of random forest is still high even under the default parameters, and with the subset of features based on random forest, the prediction accuracy will be improved to a certain extent, and at the end, the parameters of random forest are optimized to obtain the optimal ntree and mtry values[10].

(3) This paper uses the random forest algorithm to predict the financial risk of brewing enterprises. Although the prediction rate of enterprises with poor financial quality is high, even reaching 100%, it may be caused by the small sample size, and the prediction rate of enterprises with normal financial quality reaches more than 60% on average, the accuracy of model prediction is not particularly high, and further improvement of the model is needed if a higher accuracy rate is required.

(4) Although the random forest algorithm performs better in the financial prediction of brewing enterprises, this paper does not investigate other empirical methods, and further research is needed if we want to compare the accuracy of different research methods for predicting the financial risk of brewing enterprises.

References

[1] Fitzpatrick P J . A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies. 1932.

[2] Beaver W.Financial Ratios As Predictors Of Failure[J]. Journal of Accounting Research, 1966, 4 (1):71-111.

[3] Altman E.Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. 1968.

[4] Daniel, Martin. Early warning of bank failure: A logit regression approach[J]. Journal of Banking & Finance, 1977.

[5] Ohlson J . Financial Ratios and The Probabilistic Prediction Of Bankruptcy[J]. Journal of Accounting Research, 1980, 18(1):109-131.

[6] Aziz.Lawson. Bankruptcy Prediction-An Investigation of Cash Flow Based Model [J]. Journal of Management Studies, 1988, 25 :419-437.

[7] Sharda, "A neural network model for bankruptcy prediction," 1990 IJCNN International Joint Conference on Neural Networks, 1990, pp. 163-168 vol.2,

[8] Mcclean S. A data mining approach to the prediction of corporate failure[J]. Knowledge-Based Systems, 2001, 14 (3–4):189-195.

[9] Liu, E.L., Tang. On financial risk management [J]. Journal of Beijing Business School, 1989(1):5.

[10] Hou Yu. On the construction of early warning mechanism for financial risks of small and medium-sized enterprises [J]. Journal of Shanxi University of Finance and Economics, 2012, 34(S1):167-168.