Research on Financial Distress Prediction Models of Chinese Listed Companies in Pharmaceutical Manufactures Based on Machine Learning

Jian Ke(✉) and Qiqi Wang

Newhuadu Business School, Fuzhou 350100, Fujian, China
wangqiqi_2022mba@nbs.edu.cn

Abstract. The pharmaceutical manufacturing industry, as a strategic emerging industry based on technology support, has received more and more investors' attention, especially after the outbreak of the new crown epidemic. However, its industry characteristics, such as a long R&D cycle, sizeable upfront investment, and uncertain market returns in the later stage, make these companies vulnerable to financial distress if they do not focus on early warning monitoring of corporate financial and non-financial data. The purpose of this study is to build and evaluate machine learning models for financial distress prediction, including random forest (RF), decision tree (DT), logistic regression (LR), and support vector machine (SVM). The forecasting results of the above models were compared and analyzed to build more accurate forecasting models. The machine learning models were constructed using 156 financial and non-financial data sets containing 26 listed pharmaceutical manufacturing companies in China from 2016 to 2021.

Keywords: financial distress · random forest · svm · decision tree · logistic regression

1 Introduction

The current global economic development is facing significant uncertainties. The negative impact of the new crown epidemic on the economy and the banking turmoil in the United States and Europe make the global economic and financial situation very challenging. Financial risks are passing from the financial market to the real economy, and the impact on the real economy is becoming more and more prominent. As a strategic new industry based on technology support, the financial sustainability of the pharmaceutical manufacturing industry is one of the most critical factors in maintaining national and social stability.

Faced with a more severe external environment, if pharmaceutical manufacturing enterprises cannot monitor and warn their financial data in a timely and effective manner, they can easily fall into financial distress. On the contrary, if an enterprise can establish a financial distress early warning model based on real-time monitoring of

K. Hemachandran et al. (Eds.): ICAID 2023, AHIS 9, pp. 544–552, 2023.
https://doi.org/10.2991/978-94-6463-222-4_59

big data, catch early warning signals and take practical actions before financial distress occurs, then the enterprise can effectively avoid falling into financial distress and achieve high-quality and sustainable development. Financial distress early warning models help companies avoid financial distress and play a crucial role in the management activities of other stakeholders (banks, insurance companies, creditors, investors, and regulators). Therefore, there is an urgent need to design a big data-based, multi-dimensional, and intelligent financial distress early warning model for real-time monitoring by companies, investors, creditors, and regulators in complex situations.

With the rapid development of information technology, artificial intelligence, 5G technology, data mining, machine learning, and other information technologies becoming increasingly mature, building an effective early warning model for financial distress has become possible. In the face of massive financial data mining and analysis, big data technology and artificial intelligence have full advantages. Big data technology can combine fuzzy logic with inference results to meet the requirements of computer processing network data, build the corresponding model and improve data processing efficiency. Compared with the traditional statistical methods, artificial intelligence technology can simulate the human brain and has strong learning ability, which can quickly find the non-linear problems in financial management and deal with them properly quickly, and provide a reliable reference basis for management to make decisions.

After reviewing a large amount of related literature, scholars have started to use machine learning techniques for financial distress early warning to obtain more accurate financial distress early warning models. Different models have advantages and disadvantages. Studies using logistic regression, support vector machine, random forest, and decision tree methods have received more attention, so these four models are selected for comparative analysis in this paper.

The remainder of the paper is organized as follows. Section 2 reviews the research related to financial distress forecasting. Research methodology: Sect. 3 describes the data collection and modeling for forecasting financial distress. In the next section, the predictive performance of all our models is described, analyzed, and compared. Finally, in Sect. 5, we extrapolate our findings using the best model for financial distress forecasting.

2 Literature Review

There are many different views on the definition of financial distress, both nationally and internationally. In Beaver's [1] (1966) study, bankruptcy, default on preferred dividends, and default on debts are defined as financial distress, and Altman [2] (1968) defines financial distress as "an enterprise that enters legal bankruptcy. Qi Gu and Shulian Liu [3] (1999) define financial distress as "an economic phenomenon in which a firm is unable to pay its debts or expenses as they fall due. Changjiang Lv [4] (2004) considers financial distress as a dynamic and continuous process and state, which is not a temporary situation but has the characteristics of continuity and recurrence. In this paper, we quote Changjiang Lu's view that the current ratio is less than 1 for at least two consecutive years to define financial distress.

And the prediction of financial distress has always been an area of interest for most researchers. In the 20th century, scholars used financial ratios as the basis for constructing financial early warning models through univariate and multivariate analysis. Beaver

(1966) first used statistical methods to build univariate financial models; in 1968, Altman proposed building financial early warning models using multivariate discriminant methods, and the Z-Score model was born. 1980. Ohlson [5] first used Logistic Regression Model to build financial distress early warning model. In 1999, Jing Chen [6] attempted to construct a financial early warning model using univariate and multivariate analysis. Wu, S. N. and Lu, H. Y. (2001) [7] compared the prediction effects of multivariate discriminant analysis, linear probability model, and logistic model. Shouhua Zhou et al. [8] introduced cash flow indicators into the early warning model for the first time and constructed an F-model. This model has been improved relative to Altman's Z-Score model.

3 Research Methodology

This section explains driving modeling, the algorithms used, how the model is evaluated, the dataset used for simulation, and the financial distress early warning indicator system. The schematic representation of the steps involved in this study is shown in Fig. 1. First, the collected data are preprocessed and divided into test set training sets. Then the data divided into training sets are brought into the four models in Fig. 1 for training. Next step, the data from the test set are used to test the model effects, and the model with the best prediction is selected.

3.1 Data Collection

Our sample contains 156 financial and non-financial data for 26 companies listed in the pharmaceutical manufacturing industry in China. The initial data show an unbalanced mix of financially healthy and distressed companies. There are 13 financially distressed companies compared to 57 economically healthy companies. We selected these 13 financially distressed companies as our sample, marked "1". These companies

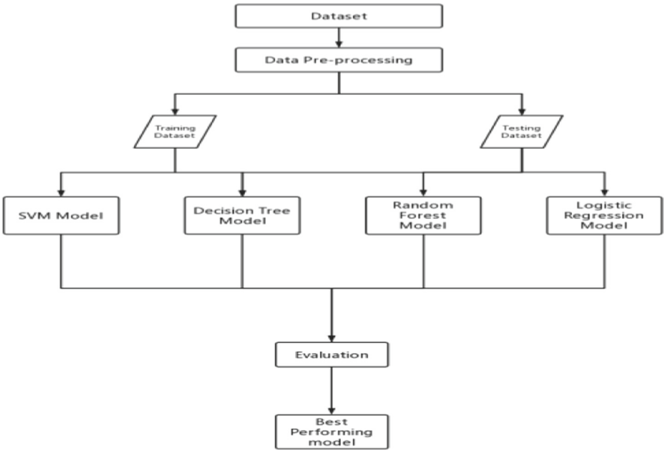

Fig. 1. Schematic diagram showing the steps of research

have had a current ratio of less than 1 for two consecutive years between 2018 and 2019. to create a balanced sample, we selected the same number of healthy listed SOE companies from our database as a control sample, marked as “0”.

3.2 Construction of Financial Distress Early Warning Indicator System

In terms of indicator selection, this paper selects indicators from multiple dimensions based on the commonly used financial risk warning indicators. It combines the characteristics of Chinese pharmaceutical manufacturing enterprises, such as significant R&D investment, long R&D cycle, and challenges to guarantee expected returns, to make the indicator system reflect the overall situation of pharmaceutical manufacturing enterprises as much as possible. The data are obtained from CSMAR database. The selected financial indicators include five aspects such as solvency, operating capacity, and profitability, and considering that the non-financial indicators can objectively reflect the actual working situation of the company, the non-financial indicators selected in this paper include four aspects, such as management governance ability, R&D innovation of listed companies, internal control and social responsibility, totaling 52 alternative indicators.

Subsequently, we first removed the indicators with missing values greater than 15% in this paper according to the actual situation of the data. Then, we conducted significance tests and correlation analyses for the remaining indicators separately to achieve the purpose of screening indicators with significant differences and avoiding excessive redundancy in specific data dimensions, which affects the efficiency and quality of the model. Finally, the Shapiro-Wilk and Mann-Whitney U tests in SPSS were used to screen the financial indicators. After the above data processing, we finally selected 23 economic indicators, as shown in the Table 1. In addition to the traditional financial indicators, this paper also selects innovative indicators, such as the ratio of capitalized R&D investment to R&D investment, which can be used to measure a company’s R&D investment output ratio.

3.3 Driving Modeling

In the stage of modeling financial distress prediction data, four prevalent data mining classifier models, namely support vector machine, random forest, decision tree, and logistic regression, were trained and tested. We collected financial and non-financial data for the first two and last two years of listed companies in the pharmaceutical manufacturing industry labeled as financially distressed. The dataset was divided into training and testing sets in the ratio of 70:30.

Support vector machine (SVM) emerged in the 1990s as a supervised machine learning classifier. SVM is particularly suitable for data with many variables because, in high-dimensional space, the data are “broken up”, so it is easier to separate them by a hyperplane. In our study, the support vector machine classifier is constructed using Scikit Learn with linear kernel, polynomial kernel, radial kernel, and sigmoid kernel, respectively. Comparing the above results reveals that the radial kernel has the best prediction results. Further, the optimal combination of parameters is selected for optimal

Table 1. CSMAR DATA BASE FINANCIAL RATIOS

Variable Description					
X1	Current Ratio	X9	Return on Total Assets Ratio	X17	Whether the directors and supervisors have financial background
X2	Equity Ratio	X10	Rate of Return on Investment	X18	Number of R&D personnel as a percentage
X3	Total Current liabilities/Total liabilities	X11	Total Cash Recovery Rate	X19	Capitalized R&D investment /R&D investment
X4	Receivables Turnover Ratio	X12	Operating Cash Flow to Operating Profit Ratio	X20	Whether internal controls are effective
X5	Inventory Turnover	X13	The Rate of Capital Preservation and Appreciation	X21	Whether to refer to GRI
X6	Account Payable Turnover Rate	X14	Dual Role of the Board Chairman	X22	Whether to disclose the protection of shareholders' rights and interests
X7	Fixed Assets Turnover	X15	Managerial Ownership	X23	Are the auditors from the Big 4
X8	Total Assets Turnover	X16	Board Size		

adjustment by 10-fold cross-validation. Then the confusion matrix is computed by comparing the actual and predicted categories of the whole data, and finally, the evaluation results are obtained from the confusion matrix.

The decision tree is a classification and prediction method often used in data mining. The basic principle of the algorithm is:

- Classify the overall data according to the classification conditions specified by the algorithm
- Produce a decision node
- Still, follow the rules of the algorithm classification
- Repeat the above operation at each decision node until the sort can not continue

Since the decision tree considers the information of y when splitting the nodes, it is more "intelligent", not affected by noise variables, and suitable for high-dimensional data. The idea and prototype of decision trees were formed in the 1960s and matured in the 1980s. In the field of statistics, it is represented by the classic work Classification and Regression Trees by Breiman, Friedman, Stone, and Olshen (1984), referred to as the CART algorithm. Quinlan (1979, 1986) proposed the ID3 algorithm in the computer field, which later evolved into the C4.5 and C5.0 algorithms. In this study, the CART decision tree algorithm is used to specify the maximum depth of the decision tree as 3. We use the Gini index to select the optimal partition. We choose the best cost complexity parameter by 10-fold cross-validation, ccp_alpha, to optimize the model.

Random Forest: Based on bagging, only some variables (e.g., m variables) are randomly selected as candidate splitting variables at each node of the decision tree when breaking, e.g., m variables are randomly chosen as candidate splitting variables at one node (the remaining (p-m) variables are not used), and (possibly different) m variables are randomly selected as candidate splitting variables again at the next node, and so on. m variables at the next node as candidate splitting variables, and so on: this is then done for each decision tree in the random forest. In our study, we built the model using Scikit Learn with full features set to 23 financial and non-financial indicators in the input dataset, selected the optimal hyperparameters max_features by 10-fold cross-validation with stratified random grouping, and finally trained the model using this parameter and drew the variable importance graphs and ROC curves. The ROC curve and the area under the curve AUC can be used to visualize the prediction effect of random forest.

Logistic regression is one of the classical algorithms in statistics. It is equivalent to a linear combination ($\theta^{T}x$) of features deflated by a Sigmoid function, which compresses the target value to within [0,1]

The prediction form of the logistic regression model is shown as (1). After the compression of the Sigmoid function, $h_{\theta}(x)$ represents the probability that the sample is classified as 1. In general, the classification threshold of logistic regression is 0.5. If the output value of the model $h_{\theta}(x) > 0.5$, the model will classify the sample as category 1; if the output value of the model $h_{\theta}(x) < 0.5$, the model will classify the sample as category 0.

$$h_{\theta}(x) = g\left(\theta^{T}x\right) = \frac{1}{1 + e^{-\theta^{T}x}} \tag{1}$$

The logistic regression model is suitable for dichotomous problems and can see the effect of different indicators on the final prediction results. The model has better interpretability, so we selected the logistic regression model as a candidate for the financial distress identification model.

3.4 Evaluation

In this paper, precision, recall, F1-score and total correctness are selected as the model evaluation metrics based on the common evaluation metrics of machine learning. Precision is the ratio of correctly predicted positive to the sample of predicted positive cases; recall is the proportion of correctly predicted positive to the actual positive; F1-score is the summed average of precision and recall. Area Under Curve (AUC) is the area under the ROC curve. We use the AUC value as the evaluation criterion of the model. Because in many cases, the ROC curve does not clearly indicate which classifier is more effective, and as a value, the classifier with a larger AUC is more effective. The ROC curve is called the receiver operating characteristic curve, which is a curve based on a series of different dichotomous classifications, with the actual positive rate (sensitivity) as the vertical coordinate and false positive rate (1-specificity) as the horizontal coordinate. The AUC is a performance measure of a learner's strengths and weaknesses.

4 Result Analysis

As shown from Table 2, the four models have a high overall correct rate, which are all above 80%. From the comparison of the results in Table 2. It indicates that all four models have considerable prediction ability for financial distress. And the above four models are, in order of superiority, the random forest model, decision tree model, support vector machine model, and logistic regression model. Therefore, through the investigation of this paper, it is found that the random forest model is a better choice for early financial warning of companies.

Figure 2 shows the ROC curves of the four models and their AUC values. The larger the AUC value, the better the prediction effect. The highest AUC value indicates that the random forest has the best prediction effect on the comparison with the other three models.

Random forest, as an integrated learning algorithm based on Bagging, can handle very high dimensional data; it can also derive feature importance after the training is completed. By analyzing the feature importance, we can measure the contribution of each input feature to the prediction result of the model and also visualize the relevance of the feature to the target. The more significant the contribution, the more importance should be given to the feature. The feature importance plot of the random forest model in this paper is shown in Fig. 3. This figure shows that the variable X1 (current ratio) indicator ranks the highest. And current ratio as an important indicator to measure the solvency of enterprises should indeed be paid attention to by enterprises.

Table 2. Comparison of variable classification prediction results

		Logistic	SVM	Decision Tree	Random Forest
precision	0	0.81	0.90	0.95	0.96
	1	0.87	0.81	0.85	1.00
recall	0	0.93	0.79	0.83	1.00
	1	0.68	0.91	0.96	0.96
F1	0	0.87	0.84	0.89	0.98
	1	0.76	0.86	0.90	0.98
accuracy		0.83	0.85	0.89	0.98

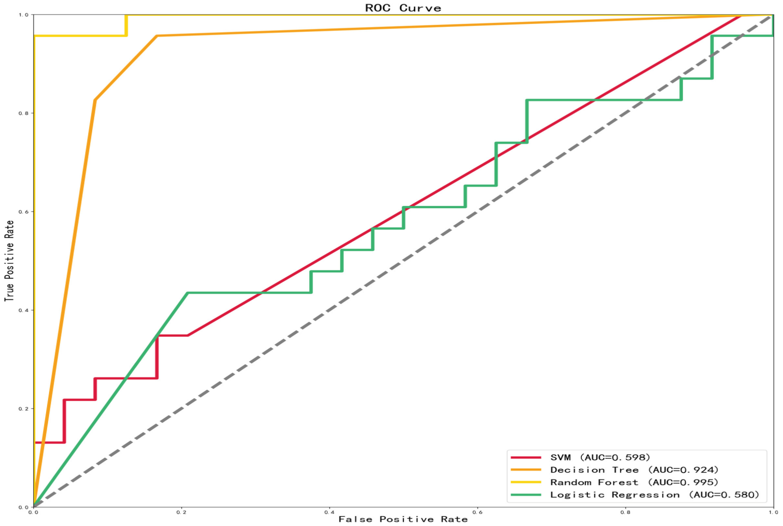

Fig. 2. False Positive Rat

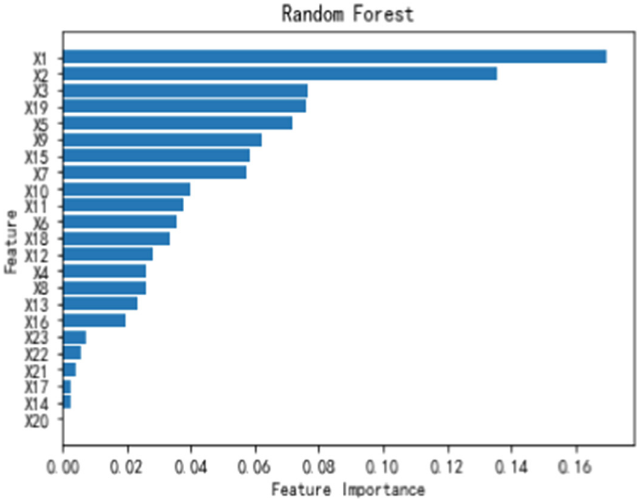

Fig. 3. Feature Importance of Random Forest

5 Conclusion

This paper discusses using SVM, decision trees, random forests, and logistic regression models to solve the corporate financial distress prediction problem. By comparing the results of the four models, we find that random forest has the best prediction effect. The top four features contributing to the model, as shown in the feature importance graph generated by the random forest model, are the current ratio, equity ratio, total current liabilities/total liabilities, and capitalized R&D investment/R&D investment. The results supported by the test sample indicate that the random forest model is an effective method for predicting financial distress in China, and the random forest has an advantage over the traditional linear model. For further research the random forest can depth optimize the calculating to obtain better results, and it can also extend the predicting function to other industries.

References

1. Beaver W. H., "Financial ratios as predictors of failure," .Journal of Accounting Research , 1966(4): 71-111.
2. Altman E., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,". Journal of Finance, 1968(23): 589-609.
3. Qi Gu and Shulian Liu , "Analysis and countermeasures of investment behavior of enterprises in financial crisis," Chinese Journal of Accounting Research,1999(10):28-31.
4. Changjiang Lv, Lily Xu, Lin Zhou, "Comparative Analysis of Financial Distress and Financial Insolvency of Listed Companies," Chinese Journal of Economic Research,2004(08):64-73.
5. Ohlson J. S., "Financial ratios and the probabilistic prediction of bankruptcy," Journal of Accounting Research,1980(19):109-131.
6. Jing Chen, "Empirical analysis of early warning of financial deterioration of listed companies," Chinese Journal of Accounting Research,1994(04):31-38.
7. Shinong Wu. and Xianyi Lu, "Study on Chinese Listed Companies Financial Distress Prediction Models," Chinese Journal of Economic Research,2001(06):46–55.
8. Shouhua Zhou , Jihua Yang , Ping Wang ,"On Early Warning Analysis of Financial Crisis—F-score Model," Chinese Journal of Accounting Research,1996(08):8-11.