Original Research Article

Corporate Financial Distress Prediction: A Hybrid Tracking Model Approach

ABSTRACT

The purpose of this study was to build a highly accurate corporate financial distress tracking and prediction model based on hybrid machine learning technology. The research data were from Taiwan Economic Journal, and the research subjects were enterprises with financial distress risk announced in September 2022. In consideration of enterprise features, this study excluded the finance and insurance industries. The research period was three years (2019, 2020, and 2021) before the distress announcement. This study matched enterprises with financial distress and enterprises without financial distress (normal enterprises) at a ratio of 1:1 for each year. The sample size for each year included 374 enterprises with financial distress and 374 enterprises without financial distress. This study applied several machine learning technologies. At first, important variables were screened by applying artificial neural networks (ANNs). Next, prediction models were built based on decision tree C5.0 and random forest (RF) and were compared. According to the empirical result, the ANN-RF model provided a higher accuracy.

Keywords: tracking model approach; machine learning; financial distress prediction; artificial neural network; C5.0; random forest

1. INTRODUCTION

Corporate financial distress affects the rights, interests, and livelihoods of shareholders and employees. Any financial distress of large and super-large enterprises not only affects shareholders, employees, suppliers, and customers, but also negatively affects the state, region, and even the world. For example, amid the subprime mortgage crisis triggered in the U.S. during the second half of 2008, Lehman Brothers announced bankruptcy, Indy Mac Bank was taken over by the Federal Deposit Insurance Corporation of the U.S. government due to financial distress, and American International Group (AIG) applied to the Federal Reserve Board (FRB) for emergency funding of US$85 billion. This not only struck a heavy blow against the U.S. economy but also triggered a global financial crisis. The global economy, including Taiwan's, was seriously hit. In Taiwan, many companies and factories closed, causing huge unemployment and heavy investment losses (Chen and Shen, 2020; Jan, 2021). The 2008 global financial crisis shows that even powerful international enterprises may encounter financial distress and must be constantly alert to their financial conditions (Woodlock and Dangol, 2014).

If rumors arise that enterprises are encountering business distress, then society can be destabilized, and the entire economic environment could pay a heavy price. The occurrence of the aforementioned events not only directly harmed the rights and interests of stakeholders but also put a heavy cost upon the whole society. If the management of enterprises can identify risk warnings or such problems as early as possible, then they can take related measures to prevent the occurrence or deterioration of distress. To this end, effective financial distress prediction is very important (Chen and Jhuang, 2018). Dirman (2022) proposed the importance of sound corporate governance to corporate financial distress prevention. Bankruptcy prediction and credit risk assessment are the two most pressing problems in the finance field (Elhoseny, Metawa, Sztano, and El-hasnony, 2022). Corporate financial distress has a wide range of influences, and therefore effective financial distress prediction and prediction models are increasingly important (Chen and Shen, 2020; Jan, 2021; Elhoseny, Metawa, Sztano, and El-hasnony, 2022; München, 2022).

2. LITERATURE REVIEW

One feature of financial distress is that adverse conditions encountered by institutions may generate adverse effects on their capability to fulfill their commitments and may lead to bankruptcy (München, 2022). Financial distress prediction models can be used for many purposes, including supervising a company's solvency, assessing loan and bond default risk, and pricing credit derivatives and other credit risk bearing securities (Yan, Chi, and Lai, 2020). Beaver (1966) predicted the possibility of the occurrence of corporate financial distress through financial ratio analysis, such as measuring the ratios between profitability, liquidity, and solvency. That study considered that cash flow/total liabilities can provide the highest capability of discriminating between default and non-default companies. Altman (1968) determined companies that are legally bankrupt, taken over, or determined as reorganized pursuant to bankruptcy law as companies with financial distress. Altman constructed the discriminant function, Z =1.2X1+1.4X2+3.3X3+0.6X4+1.0X5, where X1 is working capital/total assets, X2 is retained earnings/total assets, X3 is earnings before interest and taxes/total assets, X4 is the market value of equity/book value of total debt, and X5 is sales/total assets. Moyer (1977) found that Altman's prediction model cannot be applied to all terms.

Many prior studies on financial distress prediction and prediction modeling after the 1960s applied multivariate analysis, multiple regression analysis, stepwise regression analysis, and logistic regression. For example, Ohlson (1980) used Logit models to predict financial distress, Zmijewski (1984) used Probit models to predict financial distress, while Shumway (2001) used discrete time hazard models to build financial distress warning models (Chen and Jhuang, 2018). However, the stringent hypothesis of traditional statistics (such as linearity, normality, and independence between prediction variables) limits the application of these models in practice (Jan, 2021). The application of machine learning has further promoted studies on financial distress prediction and prediction modeling and improved financial distress prediction accuracy (Chen, 2011; Kim, and Upneja, 2014; Geng, Bose, and Chen, 2015; Sun, Fujita, Chen, and Li,2017; Jan and Hsiao, 2018; Chen and Jhuang, 2018; Chen and Shen, 2020; Gregova, Valaskova, Adamko, Tumpach, and Jaros, 2020; Jan, 2021; Elhoseny, Metawa, Sztano, and El-hasnony, 2022).

3. MATERIALS AND METHODS

This study applied artificial neural networks (ANNs), C5.0, and random forest (RF) to build financial distress prediction models. These machine learning algorithms have many strengths and a strong classification function, which are described below.

3.1 ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural network (ANN) is widely applied in classification and prediction studies. The greatest strength of ANN is that it can process non-linear data and address the weakness of multiple regression analysis in which many hypotheses must be proposed. In addition, both qualitative variables and quantitative variables can be used as input or output variables (Rumelhart et al., 1986). The network structure of ANN generally consists of three layers of neurons: the input layer, hidden layer, and output layer. Data are inputted from the neurons at the input layer and then transferred to the neurons at the hidden layer; lastly, data are outputted by the neurons ait the output layer. The number of neurons at the input layer is generally the number of variables. Only one hidden layer is required to handle general problems. There is no criterion for the number of neurons at the hidden layer. The number of neurons at the output layer is the number of variables expected to be obtained.

3.2 C5.0

C5.0 is an improvement of ID3 (Quinlan, 1986). The decision tree C5.0 consists of two parts. The first one is the classification criterion. The complete decision tree is built based on the calculation of the gain ratio, as expressed in Eq. (1).

$$Gain\ Ratio(S,A)\frac{\text{Information}\,Gain(S,A)}{Entropy(S,A)} \quad (1)$$

In Eq. (1), Information Gain is used to calculate the earnings of the dataset before and after the test and is expressed in Eq. (2). Information Gain is defined as "information before the test" minus "information after the test". Entropy in Eq. (1) is used to calculate impurity, which is called chaos here and is used to calculate the chaos in the dataset. When the chaos in the dataset reaches the highest level, the value of chaos will be 1. Therefore, smaller chaos in the dataset after the test will result in a larger Information Gain, which will be more favorable for building the decision tree.

$$Gain\ Ratio(S,A) = Entropy(S) - \sum v \in values(A) \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

Decision trees are appropriately pruned based on the criterion of Error Based Pruning (EBP) to improve classification accuracy. EBP evolved from Pessimistic Error Pruning (PEP). Both pruning approaches were proposed by Quinlan. The main concept of EBP is to decide on the error rate and to calculate the error rate of each node to further decide the node that contributes to the rise of the error rate of the decision tree. This node will then be appropriately pruned to improve the accuracy of the decision tree. The overfitting problem of the training set will also be prevented.
Strengths of C5.0 can be summarized as follows: 1) fewer training operations are generally required for estimation; 2) repeated partitions of more than two subsets are allowed; and 3) a C5.0 model is easier to understand than other types of models and the rules inferred from a C5.0 model have intuitive explanations.

3.3 RANDOM FOREST (RF)

Random forest (RF) (Breiman, 2001) was developed from bootstrap aggregating (bagging) and provides better classification results than other classification methods and user-friendly operation interfaces. Bootstrap aggregating (bagging) is an ensemble learning method. It uses bootstrap samples in data to build each classifier in the collective classifier and then decides the final prediction result based on a simple majority vote (Pardo and Sberveglieri, 2008). Therefore, the RF strategy is to select the tree classifier consisting of many random variable numbers (mtry).
The quantity of variable numbers (mtry) is used to partition the best node to obtain the best random variable combination. The quantity of variable numbers is smaller than the original total number of variables. Each tree classifier is based on the random vector of independent samples, and the standard random forest consists of prediction tree classifiers with the same partition (Breiman, 2001).
RF has two main control variables: the number of tree classifiers in the random forest (n_{tree}) and the number of variables in a random variable combination of the tree classifier node (mtry). The following describes the calculation process of random forest.
n_{tree} bootstrap samples are extracted from the original training sample combination.
The best random variable combination of m_{try} variables is used as the node of tree classifiers to develop each bootstrap sample into an unpruned tree classifier.
The test dataset is inputted into the classification model built in the first two steps, and the final test result is decided based on a simple majority vote.
During the bootstrap sample extraction process, samples not extracted are called out-of-bag (OOB) samples. The error rate of the model (also called OOB error rate) can be estimated based on OBB samples as the basis for parameter selection. In addition, the random forest can rank input variables by importance according to the change in OBB error rates.
The advantages of the random forest are listed as follows. (1) A large number of input variables can be processed. (2) Each tree in the random forest is independent. (3) Training data and features to be used by each tree are decided at random. (4) The importance of variables can be assessed when the class is decided. (5) Each tree can run in parallel in the training or prediction phase. (6) A highly accurate classifier can be generated.

3.4 SAMPLING AND VARIABLE SELECTION

3.4.1. Data Sources

The research data were from Taiwan Economic Journal (TEJ), and the research objects were enterprises with financial distress risk announced in September 2022. In consideration of enterprise features, this study excluded the finance and insurance industries. The research period was three years (2019, 2020, and 2021) before the distress announcement. Moreover, enterprises with financial distress and enterprises without financial distress (normal enterprises) were matched at a ratio of 1:1 for each year. The sample size for each year included 374 enterprises with financial distress and 374 enterprises without financial distress.

3.4.2. Variable Definitions

(1) Dependent variable
This study discriminated enterprises with financial distress according to the financial distress conditions revealed by Taiwan Economic Journal (TEJ): bankruptcy, reorganization, bounced checks, bank runs, bailouts, takeovers, CPAs

having doubts about their continued operations, negative net values, delisting, financial strain, and work stoppage. The value is 1 in case of financial distress or 0 otherwise.

(2) Independent variables
This study selected 14 financial indicators commonly used in financial prediction studies as research variables, which are listed in Table 1.

Table 1. Research variables

No.	Variable	Description
X1	ROA	[Net income + interest expense × (1 - tax rate)]/Average total assets
X2	Net profit margin before tax	Income before tax/Net sales
X3	Operating expense ratio	Operating expense/Net sales
X4	R&D expense ratio	R&D expense/Net sales
X5	Net value per share	(Assets - Liabilities)/Number of common shares
X6	Revenue growth ratio	(Revenue for the current year - Revenue for the previous year)/Revenue for the previous year
X7	Current ratio	Current assets/Current liabilities
X8	Quick ratio	Quick assets/Current liabilities
X9	Debt ratio	Total liabilities/Total assets
X10	Times interest earned	EBIT/Interest expense
X11	Accounts receivable turnover	Net sales/Average accounts receivable
X12	Inventory turnover	Cost of goods sold/Average inventory
X13	Total assets turnover	Net sales/Total assets
X14	Operating cash flow	Cash flow from operating activities - Taxes and interests paid - Investment income - Income tax on dividends paid

3.5 RESEARCH PROCESS

Data of the 14 input variables in this study were from TEJ. At first, sample data of enterprises with financial distress were mixed with sample data of enterprises without financial distress (normal enterprises) at the ratio of 1:1 for each of the three years (T-1, T-2, and T-3). The sample size for each year included 374 enterprises with financial distress and 374 enterprises without financial distress. Next, two datasets were extracted at the ratio of 8:2 as the training and test subsets, respectively. To reduce the complexity of prediction models, this study used the financial indicators for Term T-1 as input variables and inputted them into the ANN model to screen relatively important financial indicators. In terms of prediction modeling, this study applied decision tree C5.0 and Random Forest (RF), which have been widely used in previous classification studies to build and compare prediction models. After the optimal model was identified based on the data of Term T-1, tracking and prediction models were built based on the data of Term T-2 and Term T-3. Fig. 1. shows the research design and process.

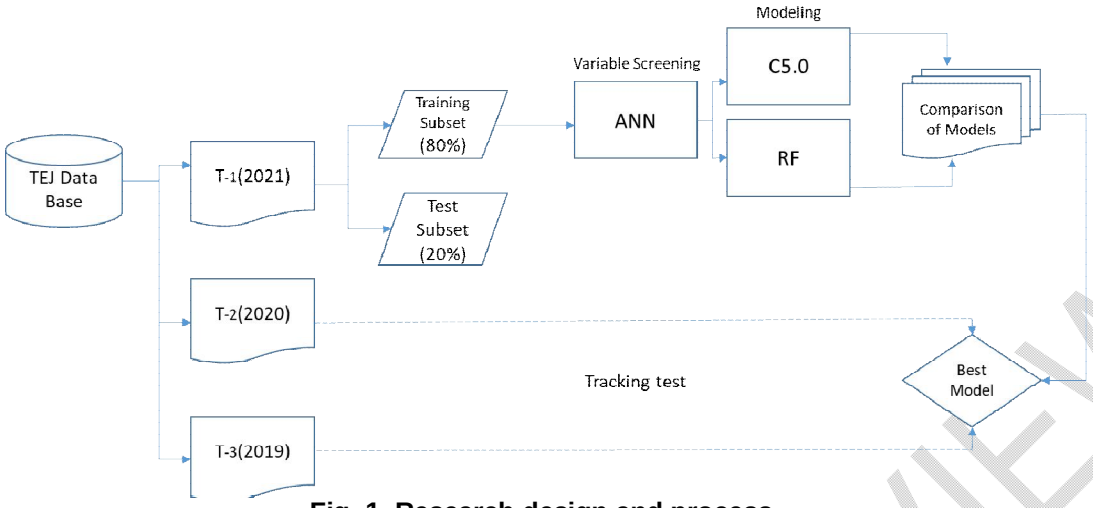

Fig. 1. Research design and process

4. RESULTS

4.1 SCREENING IMPORTANT VARIABLES BY ANN

This study applied ANN to screen important variables from the 14 research variables. Table 2 lists the important variables in sequence: X9 (debt ratio); X1 (ROA); X10 (times interest earned); X12 (inventory turnover); X13 (total assets turnover); X6 (revenue growth ratio); X3 (operating expense ratio); X14 (operating cash flow); X2 (net profit margin before tax); X8 (quick ratio).

Table 2. ANN screening result

Variable		Importance
X9	Debt ratio	0.57
X1	ROA	0.15
X10	Times interest earned	0.08
X12	Inventory turnover	0.05
X13	Total assets turnover	0.04
X6	Revenue growth ratio	0.03
X3	Operating expense ratio	0.03
X14	Operating cash flow	0.02
X2	Net profit margin before tax	0.02
X8	Quick ratio	0.01

4.2 MODELING AND TEST

After important variables were screened by ANN, corporate financial distress prediction models were built based on C5.0 and Random Forests (RF). In financial distress prediction, the impact of Type II error (misjudging enterprises with financial distress as being enterprises without financial distress) will be more serious than that of Type I error (misjudging enterprises without financial distress as being enterprises with financial distress). Therefore, in addition to accuracy, this study also included a Type II error rate in comparison. Table 3 lists the results for Term T-1.

Table 3. Accuracy comparison for Term T-1

Term T-1	Model	Accuracy	Type II Error Rate
Training	ANN-C5.0	75.82%	28.26%
	ANN-RF	92.93%	3.26%
Test	ANN-C5.0	72.12%	30.86%
	ANN-RF	84.85%	8.6%

In terms of the training subset, the ANN-RF model has a higher accuracy (92.93%) than the ANN-C5.0 model (75.82%). The ANN-RF model has a lower Type II error rate (3.26%) than the ANN-C5.0 model. In terms of test subset, the ANN-RF model has an accuracy of 84.85% and a Type II error rate of 8.6%, while the ANN-C5.0 model has an accuracy of 72.12% and a Type II error rate of 30.86%. When compared, the ANN-RF prediction model is superior to the ANN-C5.0 model in terms of both the training subset and test subset.

To identify signs of corporate financial distress as early as possible, this study carried out a tracking test for the data of Term T-2 and Term T-3 in the ANN-RF model. The data of Term T-2 and Term T-3 were preprocessed in the same way as the data of Term T-1: enterprises were matched at the ratio of 1:1, and 80% and 20% of data were extracted as training and test subsets, respectively. Table 4 lists the tracking test results for Term T-2 and Term T-3.

Table 4. Tracking test results for Term T-2 and Term T-3 by ANN-RF

Term	Subset	Accuracy	Type II Error Rate
Term T-2	Training	86.27%	15.61%
	Test	84.42%	21.9%
Term T-3	Training	62.86%	41.41%
	Test	60.49%	44.10%

Table 4 lists the test results by the ANN-RF tracking model. As listed in the table, for Term T-2 the ANN-RF model had a prediction accuracy of nearly 85% in terms of training and test subsets and a Type II error rate of 15.61% and 21.9% in terms of training and test subsets. For Term T-3, the ANN-RF model had a prediction accuracy of 62.86% and 60.49% in terms of training and test subsets and a Type II error rate of 41.41% and 44.10% in terms of training and test subsets.

5. DISCUSSION

This study screened important variables from the 14 research variables by ANN. According to the findings, the five most important variables were debt ratio, ROA, times interest earned, inventory turnover, and X13 total assets turnover. In terms of prediction models, the ANN-RF model was superior to the ANN-C5.0 model in terms of accuracy or Type II error rate.

According to the test result of the ANN-RF tracking model, the prediction performance for Term T-1 was higher than that for Term T-2. Lastly, the prediction performance for Term T-3 declined significantly.

6. CONCLUSION

Corporate financial distress will affect the rights, interests, and livelihoods of shareholders and employees. Financial distress of large and super-large enterprises not only affects shareholders, employees, suppliers, customers, and other stakeholders, but also causes negative effects on the state, region, and even the world. The purpose of this study was to establish a highly accurate corporate financial distress tracking and prediction model based on hybrid machine learning technology. This study applied several machine learning technologies. At first, important variables were screened by applying artificial neural networks (ANNs). Next, prediction models were built based on decision tree C5.0 and random forest (RF) and were compared.

This study screened important variables from the 14 research variables by ANN. According to the results, the five most important variables were debt ratio, ROA, times interest earned, inventory turnover, and X13 total assets turnover. In a study on corporate financial distress, attention needs to be paid to the selection of research variables.

In terms of the performance of prediction models, the ANN-RF model was superior to the ANN-C5.0 model in terms of accuracy or Type II error rate. According to the test results of the ANN-RF tracking model, the prediction performance for Term T-1 was higher than that for Term T-2, and the prediction performance for Term T-3 declined significantly. This is normal and meets the expectation.

The research findings herein offer a reference for academic research on corporate financial distress, the audit process and audit reports of CPAs and auditors, credit rating agencies, securities analysts, and investors. This study also suggests applying other algorithms of machine learning or deep learning in subsequent studies to implement financial distress prediction and prediction modeling.

REFERENCES

1. Altman EI. Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. J Finance. 1968;23:589-609. doi.org/10.2307/2978933

2. Angela Dirman. The Effectiveness of Good Corporate Governance Implementation against Financial Distress Conditions with Intellectual Capital as Moderating Variable Asian Journal of Economics, Business and Accounting.2022;22(23):37-49.
3. Beaver WH. Financial Ratios as Predictors of Failure. J. Account. Res. 1966;4:71-111. doi.org/10.2307/2490171
4. Breiman L. Random forests. Mach Learn. 2001;45(1):5－32. http://doi.org/10.1023/A:1010933404324
5. Chen MY. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. Expert Syst. Appl. 2011;38:11261-11272. doi.org/10.1016/j.eswa.2011.02.173
6. Chen SD, Jhuang S. Financial distress prediction using data mining techniques, ICIC Express Letters, Part B: Applications. 2018;9(2):131-136.
7. Chen SD, Shen ZD. Financial Distress Prediction Using Hybrid Machine Learning Techniques Asian Journal of Economics, Business and Accounting, 2020;1-12. DOI:10.9734/ajeba/2020/v16i230231
8. Douglas DRM. The effect of financial distress on capital structure: The case of Brazilian banks Quarterly Review of Economics and Finance 86 2022; 296–304.
9. Elhoseny, Metawa, Sztano, Ibrahim M El-h. Deep Learning-Based Model for Financial Distress Prediction Annals of Operations Research 2022 https://doi.org/10.1007/s10479-022-04766-5
10. Geng R, Bose I, Chen X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. Eur. J. Oper. Res. 2015;241:236–247. doi:10.1016/j.ejor.2014.08.016.
10. Gregova E, Valaskova K, Adamko P, Tumpach M, Jaros J. Predicting Financial Distress of Slovak Enterprises: Comparison of Selected Traditional and Learning Algorithms Methods. Sustainability 2020, 12, 3954, doi:10.3390/su12103954
11. Jan Cl. Financial Information Asymmetry: Using Deep Learning Algorithms to Predict Financial Distress. Symmetry 2021;13(3):443. doi.org/10.3390/sym13030443
12. Jan CL, Haiso D. Detection of Fraudulent Financial Statements Using Decision Tree and Artificial Neural Network. ICIC Express Letters, Part B: Applications. 2018;9(4):347-352
13. Kim SY, Upneja A. Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Econ. Model. 2014;36:354–362. doi:10.1016/j.econmod.2013.10.005.
14. Moyer RC. Forecasting financial failure: A re-examination. Financ. Manage. 1977;23(4):11-17.
15. Ohlson JA. Financial ratios and the probability prediction of bankruptcy. J. Account. Res. 1980;18(1):109-131.
16. Pardo M, Sberveglieri G. Random forests and nearest shrunken centroids for the classification of sensor array data. Sens. Actuat. B Chem.2008;131:93–99.
17. Quinlan JR. Introduction of decision trees. Mach Learn. 1986;1(1):81–106. doi.org/10.1007/BF00116251
18. Rumelhart DE, Hinton DE, Williams RJ. Learning Internal Representations by Error Propagation in Parallel Distributed Processing, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1986;1 MIT Press, Cambrige, Massachusetts
19. Shumway T. Forecasting bankruptcy more accurately: A simple hazard model, J. Bus. 2001;74(1):101-124.
20. Sun J, Fujita H, Chen P, Li H. Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. Knowledge-Based Systems, 2017;120:4–14.
21. Woodlock P, Dangol R. Managing Bankruptcy and Default Risk. Journal of Corporate Accounting & Finance. 2014;26:33–38. doi:10.1002/jcaf.22002.
22. Yan D, Chi G, Lai KK. Financial Distress Prediction and Feature Selection in Multiple Periods by Lassoing Unconstrained Distributed Lag Non-linear Models. Mathematics 2020;8: 1275. doi:10.3390/math8081275.
23. Zmijewski ME. Methodological issues related to the estimation of financial distress prediction models. J. Account. Res. 1984;22:59-82.