Improving financial distress prediction using machine learning: a preliminary study

Guo Dong Hou[1], *Dong Ling* Tong[1*], *Soung Yue* Liew[1], and *Peng Yin* Choo[1]

[1]Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia

Abstract. Financial distress is when a company faces significant difficulties meeting its financial obligations and maintaining profitability, leading to bankruptcy, insolvency, and severe economic losses. Therefore, early warning for companies at risk of financial distress is vital for business stakeholders to take timely corrective actions and avoid adverse outcomes. Existing financial distress predictions often rely on historical datasets, incorporating various indicators collected through varied methodologies and experts' opinions. The challenge arises in discerning which indicators are pivotal for predicting corporate distress, as their influence and relevance may vary. This study proposed a machine learning framework to eliminate variations of different experts' knowledge when selecting pivotal indicators. Data containing 4006 companies and 204 indicators was extracted from CSMAR. The Chi-square test is employed to select significant indicators. The correlation of these selected indicators is modeled using the C4.5 decision tree. Results showed that this selected feature set is closely aligned with those obtained when utilizing all features in the data. A thorough comparison of the indicators selected by the expert revealed notable distinctions. Features chosen by the Chi-square test are related to financial ratios and also exhibit a pronounced focus on societal attention, shareholding concentration, and market dynamics.

1 Introduction

Financial distress prediction is a critical and dynamic area of research that holds immense significance for various stakeholders in the financial landscape. As businesses navigate through economic uncertainties, the ability to anticipate and provide early warning signals for companies facing financial challenges becomes imperative.

In recent years, numerous statistics and machine learning methods have been used to study factors contributing to financial distress, aiming to provide an early detection model to avert significant business losses. As financial scenarios grow more intricate, optimizing statistics and machine learning techniques becomes crucial [1]. A pivotal area for such enhancement lies in feature selection. Past research predominantly leaned on expertise or theoretical financial knowledge to pinpoint crucial influencing factors [2]. However, this method posed three challenges:

*Corresponding author: tongdl@utar.edu.my

(1) Individual biases could influence the features chosen, potentially skewing the true relationships within the data.

(2) These features might lean towards specific financial aspects, limiting a comprehensive evaluation of risk indicators in the companies.

(3) The chosen features might adhere to linear patterns, rendering them inadequate for machine learning models.

To address these challenges, this study examines the features selected through the Chi-square test and expert financial knowledge to identify distinct focal areas of these methods. The relationships of these chosen features are then modeled using decision trees.

This paper is organized as follows. Section 2 details the dataset and the proposed method. Section 3 discusses the results, and Section 4 concludes the study and outlines the potential future work.

2 Materials and methods

2.1 Datasets

Ten (10) datasets from the CSMAR platform (https://data.csmar.com) encompassing 4,006 Chinese listed companies for 2020 were obtained. These datasets are sourced from 10 financial-related categories, as depicted in Table 1. Out of these companies, 52 were identified as financially distressed, indicated by the ST (special treatment) label in 2022 [3], while the remaining 3,954 operated under normal conditions. Upon merging these datasets, a total of 294 indicators were obtained.

Table 1. Summary of the datasets.

Dataset	Number of Features
Basic Company Information	16
Corporate governance mechanisms and industry competitive intensity	24
Analyst focus	4
Media attention	4
Audit information	7
Dividend distribution	11
Equity shareholders and institutional investors	42
Stock trading information	12
Financial statement	59
Analysis of financial indicators	115
Merged data	**294**

2.2 Proposed workflow

Fig. 1 shows the proposed workflow. This research aims to pinpoint characteristics associated with companies that have experienced financial distress using decision tree modeling. The merged data comprised 294 indicators of which 30 indicators overlapped, 31 were textual context, and 32 had > 60% of their values as NULL. It's crucial to address data duplication and missing data, as its presence can skew the results in favor of a more optimistic portrayal of the tree model. Hence, a preprocessing step was undertaken to eliminate duplicated features and features with textual content and those with NULL values.

Additionally, there exists a significant imbalance in the dataset, with only 1.3% representing distressed companies and the remaining 98.7% being normal companies. To counteract this imbalance during the feature selection phase, the SMOTE-Tomek technique was implemented.

Following these steps, the refined data containing 7908 samples and 204 features. The Chi-square test was then employed to discern significant features, and the prediction power of these features was modeled using a decision tree. To further solidify the financial relevance of these features, their validity was cross-validated using a feature set recognized by a finance expert.

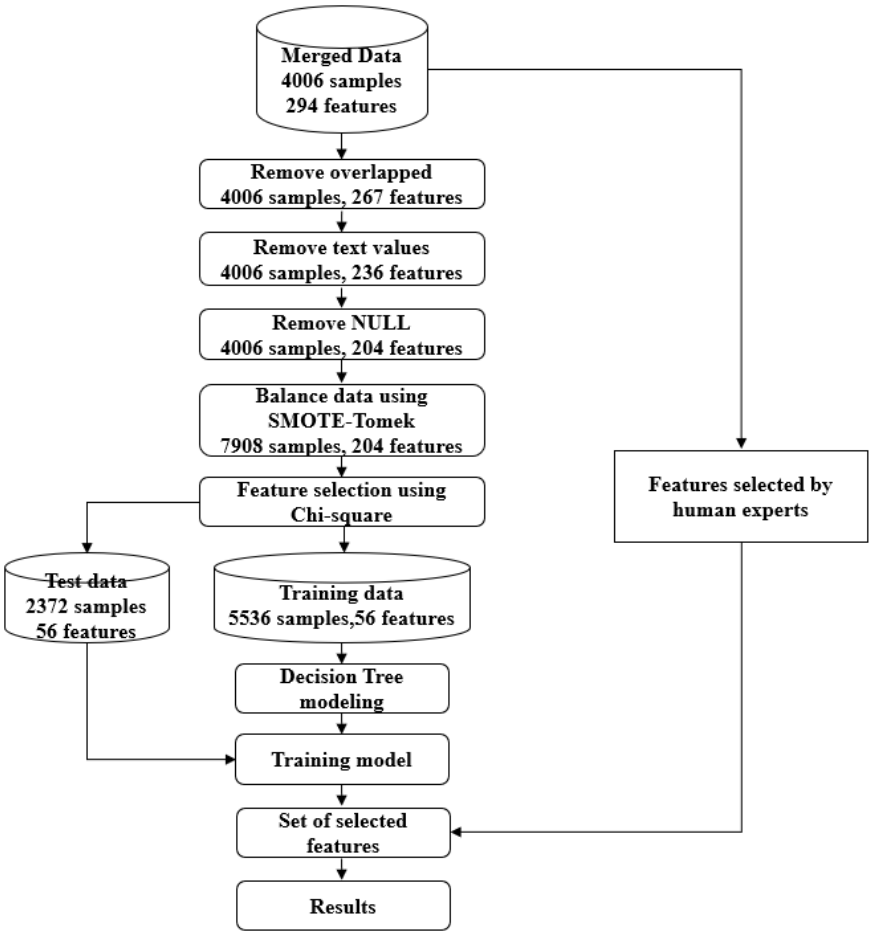

Fig. 1. Proposed workflow.

(a) SMOTE-TOMEK Link

SMOTE-Tomek link [4] is a hybrid sampling technique designed to address class imbalance issues in the data. This technique amalgamates 2 distinct methods: SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links. SMOTE is an oversampling approach that leverages the K-nearest neighbor algorithm to produce synthetic samples for the minority class, thereby augmenting its representation in the dataset. Conversely, the Tomek link is an

under-sampling method that targets the samples produced by SMOTE and eliminates those that are near one another. This hybrid technique enhances the quality of the minority class while simultaneously removing some instances that may be considered noisy or ambiguous.

(b) Chi-square

Chi-square (χ^2) [5] is a statistical test used to determine if there is a significant association between two categorical variables.

The formula for the Chi-square test statistic is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

where O_i is the observed frequency in each cell, and E_i is the expected frequency in each cell.

The Chi-square test is commonly used in hypothesis testing to determine whether a significant association exists between the variables under consideration. If the calculated Chi-square statistic significantly differs from what would be expected by chance, it suggests a relationship between the variables. The significance of the variables is determined using a p-value with an alpha of 0.05. Table 2 shows the features selected by Chi-square.

Table 2. Features selected by Chi-square.

No.	Features	Chi-Square Value	p-value
1	Loss or not	1977.80	0.00E+00
2	Analyst Attention	3806.13	8.82E-207
3	Ratio of independent directors	4616.21	1.19E-90
4	Dividend payout ratio A	4783.17	5.39E-15
5	Dividend payout ratio C	4783.17	5.86E-15
6	Dividend payout ratio before tax	4797.89	1.76E-295
7	Dividend payout ratio after tax	4797.89	1.76E-295
8	**Dividend Payout Number**	4830.79	2.44E-18
9	Dividend payout ratio B	4872.68	1.63E-16
10	Degree of Separation of Powers	6116.30	2.15E-15
11	**Price-earnings ratio (PE) A**	6179.59	1.88E-27
12	**P/E Ratio (PE) B**	6179.59	1.88E-27
13	**Price-to-Earnings Ratio (PE) TTM**	6179.59	1.88E-27
14	Notes receivable, net	6321.73	2.82E-05
15	Return on Net Assets Growth Rate A	6466.48	1.51E-03
16	Net Profit Growth Rate A	6545.59	4.82E-02
17	**Long-term loans**	6601.43	4.70E-22
18	Net Profit Growth Rate B	6861.87	2.59E-02
19	Management's shareholding	7227.86	2.98E-10
20	Herfindahl-Hirschman Index_A	7306.25	7.05E-211
21	Total long-term liabilities	7326.71	1.15E-18

22	**Interest Coverage Multiple A**	7349.59	2.52E-09
23	**Interest Coverage Multiple B**	7349.59	2.52E-09
24	Number of R&D personnel	7363.81	4.44E-100
25	R&D investment as a percentage of operating revenue	7371.97	1.94E-92
26	**Short-term borrowings**	7396.27	4.63E-10
27	Audit Fees	7396.82	5.97E-156
28	Media Attention	7463.88	1.44E-209
29	R&D expense ratio	7542.53	1.43E-02
30	Herfindahl-Hirschman Index_D	7586.35	7.54E-221
31	Herfindahl-Hirschman Index_B	7586.35	7.54E-221
32	Herfindahl-Hirschman Index_C	7586.35	7.54E-221
33	Amount of R&D investment	7640.67	3.73E-02
34	Bank borrowing ratio	7647.00	6.79E-06
35	Fund Shareholding Ratio	7690.03	7.46E-06
36	Enterprise size	7694.02	3.39E-165
37	Proportion of the listed company's control owned by the actual controller	7770.08	5.16E-20
38	Diluted earnings per share	7779.06	2.32E-87
39	Basic earnings per share	7780.19	1.95E-84
40	Yearly Opening Price	7799.21	3.59E-46
41	**Yearly Closing Price**	7802.20	6.91E-38
42	Proportion of Shares Held by Other Institutions	7815.03	7.22E-11
43	Percentage of ownership of listed companies by actual controllers	7819.53	1.78E-04
44	Number of Shares Held by Other Institutions	7851.03	2.17E-02
45	Book-to-Market Ratio	7894.07	5.99E-08
46	Shareholding Concentration5	7896.00	2.16E-02
47	Total Accrued Profit	7896.07	8.36E-06
48	Non-debt tax shield	7898.00	3.38E-02
49	Annualized stock return without reinvestment of cash dividends	7904.01	7.54E-05
50	Annual stock return considering reinvestment of cash dividend	7904.01	1.62E-04
51	Value of Shares Traded	7904.80	2.83E-02
52	Shareholding of major shareholders	7906.00	2.98E-02
53	Of which: Interest expense (finance costs)	7908.00	2.16E-05
54	Current liabilities ratio	7908.00	4.62E-02
55	Non-current liabilities ratio	7908.00	4.54E-02
56	Financial Expense Ratio	7908.00	1.97E-02

Features highlighted in **boldface** are also selected by human experts.

(c) Decision tree

The C4.5 algorithm, pioneered by Quinlan [6], is a machine-learning approach that leverages information gain calculations to craft a decision tree. This method begins by designating the most suitable attribute from the dataset as the root node for tree construction. As the process continues, the tree expands, with each iteration incorporating a newly identified optimal attribute. For classification tasks, the Gini impurity is often used [7]. The Gini impurity for a node is calculated as follows:

$$Gini(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \quad (2)$$

where t is the node, c is the number of classes, and $p(i|t)$ is the proportion of samples of class i in node t.

In the pursuit of determining the prime attribute for tree bifurcation, the algorithm computes the information content for both the labeled target attribute (i.e., ST or Normal in our study) and individual split attributes. Subsequently, these computed values are compared, and the attribute demonstrating the greatest information gain becomes the subsequent node for further splitting.

2.3 Financial ratios selected by human experts

Financial ratios are a way to evaluate the company's financial performance and identify potential financial problems. It can be calculated by the indicators corresponding to the following 6 aspects: Liquidity Ratios, Profitability Ratios, Solvency Ratios, Efficiency Ratios, Market Ratios, and Operating Performance Ratios [8-10]. Table 3 summarizes the indicators selected by human experts.

Table 3. Indicators selected by human experts.

Financial Ratios	Description	Indicators in the data
Liquidity Ratios [11]	A company's ability to meet its short-term obligations and cover immediate financial needs.	Current ratio, Quick ratio, Net profit, Short-term borrowings, Long-term loans
Profitability Ratios [12]	A company's ability to generate profits relative to its revenue and resources.	Operating income, Operating costs, ROA(A), ROA(B), ROA(C), ROA(TTM), ROE(A), ROE(B), ROE(C), ROE(TTM), Gross operating margin, Gross operating margin TTM
Solvency Ratios	A company's ability to meet its long-term debt obligations.	Total liabilities, Total owners' equity, Interest coverage multiple A, Interest coverage multiple B
Efficiency Ratios [13]	How effectively a company is utilizing its resources to generate sales and manage its assets.	Inventory turnover A, Inventory turnover B, Inventory turnover C, Inventory turnover D, Inventory turnover TTM, Accounts receivable turnover A, Accounts receivable turnover B, Accounts receivable turnover C, Accounts receivable turnover D, Accounts receivable turnover TTM
Market Ratios	Provide insights into investor perceptions and the valuation of a company's stock in the market.	Price-earnings ratio (PE) A, P/E ratio (PE) B, Price-to-Earnings ratio (PE) TTM, Dividend payout number, Yearly closing price

Operating Performance Ratios	A company's core business operations and efficiency in generating operating profits.	Total profit, Total asset turnover A, Total asset turnover B

3 Results

3.1 Classification performance

Table 4 shows the prediction results of the model.

Table 4. Prediction results.

Data	Precision	Recall	F1-Score	Accuracy	Computational time
All features	96.65%	96.63%	96.63%	96.55%	30.65s
56 features selected by Chi-Square	96.19%	96.17%	96.17%	96.21%	4.5s

Using the entire features in the data, the decision tree achieved an accuracy of 96.55% and using the Chi-square-selected features, the tree achieved an accuracy of 96.21%. The sensitivity of the model is 96.63% for the entire features and 96.17% for the 56 features selected by Chi-square. In other words, out of 3,954 distressed samples, >96% of the samples were correctly classified. This indicates that the Chi-square algorithm can identify statistically significant features for identifying companies experiencing financial distress. Moreover, using the 56 features selected by Chi-square, the computational time for decision tree modeling was significantly reduced to 4.5 seconds when compared to modeling the entire features in the decision tree. This can subsequently improve the efficiency of detecting financial distress companies.

High recall gained from using the entire features in the data (96.63%) and the 56 features selected by Chi-square (96.17%) indicate that these models are capable of predicting distressed companies. The F1-Score of using decision trees on these 2 datasets are 96.63% and 96.17%, respectively, suggesting that the tree has achieved a good balance between minimizing false positives and false negatives. In other words, the tree is effective at making accurate normal companies' predictions (low false negatives) while also capturing a large proportion of actual distressed companies' instances (low false positives).

3.2 Financial implication

Out of the 56 indicators selected by Chi-Square and the 39 indicators highlighted by the expert, 9 were found to be common between the 2 sets (see Table 5). These overlapped indicators offer insights into 3 distinct facets of a company's financial position: Solvency, Interest rate, and Price-earnings ratio. These aspects have been widely recognized as crucial determinants of financial risk in previous studies [14-16].

Table 5. Overlapped features selected by both Chi-square and human experts.

No.	Indicators	No.	Indicators
1	Dividend Payout Number	6	Interest Coverage Multiple B
2	Yearly Closing Price	7	Price-earnings ratio (PE) A
3	Short-term borrowings	8	P/E Ratio (PE) B

4	Long-term loans	9	Price-to-Earnings Ratio (PE) TTM
5	Interest Coverage Multiple A		

Among the 47 non-overlapping features from Chi-square, 19 were related to one or more of the aspects in the following Table 6.

Table 6. Non-overlapping features selected by Chi-square.

Aspects	Indicators
Social Visibility	Analyst attention, Media attention
Dividend Payout Behavior	Dividend payout ratio before tax, Dividend payout ratio after tax, Dividend payout ratio A, Dividend payout ratio B, Dividend payout ratio C
Shareholding Concentration Behavior	Management's shareholding, Shareholders occupy, Ownership proportion, Control proportion, Shrcr5, Fund hold proportion, Other hold shares, Other hold proportion
Market Influence during the Pandemic	HHI_A, HHI_B, HHI_C, HHI_D (Herfindahl-Hirschman Index [17])

These 19 non-overlapping features that Chi-square selected can be systematically categorized into 4 distinct aspects, as delineated in Table 5:

(1) Social Visibility: When a listed company encounters financial distress, its predicament often garners heightened social scrutiny, manifesting increased attention from analysts and the media.

(2) Dividend Payout Behavior: Companies that reduce or eliminate dividends may be perceived as grappling with financial challenges. Such decisions can subsequently precipitate a downturn in their stock prices.

(3) Shareholding Concentration Behavior: The potential divestment of substantial holdings within a company may serve as a red flag, hinting at underlying financial distress or even bankruptcy concerns.

(4) Market Influence during the Pandemic: Four (4) of the indicators associated with the Hirschman Index surfaced prominently in the Chi-Square analysis, predominantly due to market volatility induced by the global pandemic spanning 2020 to 2022. The Herfindahl-Hirschman Index (HHI) is a pivotal metric for assessing market concentration, commonly employed in economic evaluations and antitrust oversight [18-19]. The dataset's reflection of the pandemic's market impact is evident through these indicators. Absent the pandemic's influence, these specific metrics might have been less prominently highlighted for discerning financial risks.

4 Conclusion

This study aims to identify key indicators for forecasting corporate distress. Data sourced from 10 financial-related categories in the CSMAR platform were merged and balanced using SMOTE-Tomek links. A set of 56 features was then chosen via the Chi-square test. These features were subsequently modeled using a decision tree. The results showed that the selected indicators are highly associated with distressed companies. A high sensitivity of 96.17% was achieved, indicating that these indicators are strong predictors for predicting distressed companies. Notably, the Chi-square-selected features outperformed the entire feature set regarding computational time, enhancing the efficiency of financial distress detection.

Furthermore, juxtaposing these results with the indicators earmarked by human experts revealed an intriguing divergence. Specifically, the Chi-square algorithm pinpointed 19 distinct features, which could be categorized into 4 unique dimensions: Social visibility, Dividend payout behavior, Shareholding concentration behavior, and Market influence during the Pandemic. These dimensions, exclusively unearthed by the algorithm, shed light on subtle nuances that transcend conventional expert-driven selections. This underscores the algorithm's potential to augment both the accuracy and granularity of financial risk evaluations, heralding a more comprehensive approach to corporate distress prediction. In the future, it is essential to conduct additional research to comprehend the potential links between a broader range of indicators and financial distress. Additionally, there is a need to assess the effectiveness of the proposed method in handling larger datasets. In future research, we can use different algorithms to evaluate the performance of features selected by Chi-square to enhance the robustness of financial distress early warning models.

References

1. S. Theußl, F. Schwendinger, K. Hornik, J. Stat. Softw., **94**(15), 1-64 (2020)
2. X.V. Vo, J. Econ. Dev., **17**(1), 41-49 (2015)
3. E. I. Altman, J. Finance, **23**(4), 589-609 (1968)
4. G. E. Batista, R. C. Prati, M. C. Monard, ACM SIGKDD Explorations Newsletter, **6**(1), 20–29 (2004)
5. K. Pearson, Breakthroughs in Statistics: Methodology and Distribution, Springer New York, 11-28 (1992)
6. J. R. Quinlan, Machine learning, **1**, 81–106 (1986)
7. L. Breiman, Classification and regression trees, Routledge, (2017)
8. M. Z. J. Vochozka, Naše more, **63**(3), 227-236 (2016)
9. H. F. V. P. Assous, Complexity, **2022**(1), 1-15 (2022)
10. K. Y. A. Emre Çelik, J. Eur. Real Estate Res., **15**(2), 192-207 (2021)
11. A. N. Farohah, D. D. Dahruji, Jurnal Ekonomi Syariah, **8**(1), 86–98 (2023)
12. Y. Sari, N. Nofinawati, S. Batubara, F. Alfadri, Journal of Sharia Banking, **l**(1), 10 (2020)
13. H. Yanīkkaya, Y. U. Pabuçcu, ISRA Int. J. Islamic Finance, **9**(1), 43–61 (2017)
14. M. Sherris, J. Risk Insur., **73**(1), 71-96 (2006)
15. X. P. X. Zhang, Front Public Health, **9**, 756977 (2021)
16. T. Kang, Managerial Finance, **30**(11), 30-44 (2004)
17. M. M. S. K. B. Bogamuwa, K. L. W. Perera, Int. J, Account. Bus. Fin., **8**(2), 82–105 (2022)
18. M. Naldi, M. Flamini, Economic Papers: A journal of applied economics and policy, **37**(3), 344–362 (2018)
19. T. O. Kvålseth, Contemp. Econ., **16**(1), 51–60 (2021)