Proceedings of CONF-APMM 2025 Symposium: Simulation and Theory of Differential-Integral Equation in Applied Physics
DOI: 10.54254/2753-8818/2025.DL27393

Determinants of Corporate Bankruptcy: A Logistic Regression Analysis with Evidence from Taiwan

Yahuiru Xu

Guanghua Cambridge International School, Shanghai, China
xuyahuiru@gmail.com

Abstract. Corporate bankruptcy has significant implications for investors, governments, and society. Predicting bankruptcy through financial indicators provides an early-warning mechanism to mitigate risks. Previous studies have commonly employed financial ratios, logistic regression, and machine learning methods. However, many existing studies focus more on the performance of the model itself instead of interpretability. The aim of this study is to analyse publicly available company data from Taiwan, apply mutual information and correlation-based feature selection, and estimate a logistic regression model to identify the most important factors influencing bankruptcy positively or negatively. By combining the feature selection step with a transparent and interpretable model, this study contributes to the field in two ways: first, it provides a list of key financial ratios under a trustworthy dataset; and second, it provides interpretable evidence on how multiple elements, such as debt ratio, affect bankruptcy risk. The findings are intended to inform firm managers, lenders, and regulators by offering a practical set of early-warning indicators.

Keywords: bankruptcy, economic objects, prediction, profit, interest rate

1. Introduction

Corporate bankruptcy is not only about a single firm's failure. It can spread through supply chains, affect lenders and employees, and shake market confidence. Because of these spillovers, a reliable way to assess bankruptcy risk is valuable for managers, investors, banks, auditors, and regulators. Academic work on bankruptcy prediction has a long history. The classic Z-score model by Altman combined a few accounting ratios under linear discriminant analysis to separate bankrupt firms from healthy ones [1]. Ohlson moved the field to a probabilistic setting with logistic regression [2]. Later, Shumway used a hazard (survival) model to capture time dynamics [3]. More recently, machine learning models have reported higher predictive accuracy in many samples [4], though model transparency can be a challenge in practice.

To investigate the fundamental factors influencing corporate failure, the study uses a widely employed dataset from the Taiwan Stock Exchange for tracking companies from 1999 to 2009, which offers a long-term perspective on corporate financial health. This study uses mutual information (also known as information gain), as it captures both linear and non-linear relations, to rank features and select the top 20 variables from a broad set of accounting ratios. After feature selection, this study first addresses the inherent class imbalance problem in the dataset, where

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL27393

bankrupt enterprises are far fewer than healthy ones, by reducing the sample size of non-bankrupt enterprises. Then, this study conducts a descriptive comparison of the selected variables between bankrupt and non-bankrupt enterprises to preliminarily examine their distribution. Next, this study analyses the data via logistic regression, the corresponding coefficients and odds ratios. Finally, this study discusses the economic significance of the most important indicators.

2. Related literature

Early ratio-based models. Altman used linear discriminant analysis to combine five ratios: working capital/total assets, retained earnings/total assets, EBIT/total assets, market value/book value of equity, and sales/total assets. His model performed well in the original sample and set a standard for decades [1]. At the same time, Beaver showed that single ratios—especially cash flow to total debt—can separate failed and non-failed firms at simple cutoffs, highlighting the central role of cash generation and leverage [5]. Recent studies have extended the bankruptcy prediction literature by introducing both traditional and novel predictors. For Taiwan's electronics industry, the liquidity ratio, debt ratio, and fixed assets turnover ratio remained the most significant predictors of corporate bankruptcy. Their hybrid intelligent classification models still highlighted the centrality of traditional accounting ratios [6].

Logit and probabilistic models. Ohlson proposed the O-score, a logistic regression that predicts the probability of bankruptcy from a set of financial ratios and size variables [2]. Logistic regression does not require normality and equal covariance assumptions of discriminant analysis. It also gives an interpretable probability that managers and lenders can use.

Time-to-failure models. Shumway argued that bankruptcy is a time process and used a hazard model, which allows predictors to change over time and handles censoring [3]. He showed that adding market-based variables and time-varying accounting signals improves predictions. Wang and Brorsson moved beyond accounting variables by incorporating corporate restructuring behaviour into bankruptcy prediction models. They demonstrated that combining restructuring events with financial data improved predictive accuracy by 4%–13% during the COVID-19 period, suggesting that behavioural factors can play an important role in distress analysis [7]. Textual data has also attracted attention. Kim and Yoon employed a domain-adapted BERT model to analyse sentiment in MD&A disclosures. Their results showed a significant improvement in prediction accuracy, reaching 91.56%, thus emphasizing the value of linguistic and semantic signals as leading indicators of financial failure [8].

Machine learning. As data and computing power grew, researchers compared tree-based models, support vector machines, and neural networks with traditional methods. In many settings, machine learning improved out-of-sample accuracy, especially when relationships are nonlinear or involve interactions [4]. Network spillovers and contagion risks have been introduced through machine learning frameworks. Zhao, Chen, and Zhang used graph neural networks to combine intra-firm financial risks with contagion effects across firm networks. Their framework demonstrated that inter-organizational connections can significantly affect bankruptcy probabilities [9]. But these models can be "black-box models," which limits their direct use in audit and regulatory decisions that require explanations. Finally, Jiao combined the LASSO algorithm with Gradient Boosted Decision Trees (GBDT) to predict financial distress in Chinese listed firms. This hybrid method effectively addressed class imbalance and temporal concept drift, while confirming that leverage variables and profitability ratios remain dominant drivers of corporate bankruptcy. Interpretability has also become a key concern. Li, Härdle, and Lessmann developed a case-based reasoning (CBR) approach that balances accuracy with transparency, offering a viable alternative to black-box models

DOI: 10.54254/2753-8818/2025.DL27393

in regulatory or managerial contexts. Their work underscores the continuing importance of explainability in predictive models [10-11].

Across methods, the same families of variables show up: profitability, leverage/solvency, liquidity, and earnings persistence. These are the core drivers behind distress risk in both classic and recent studies.

3. Data and variable selection

3.1. Sample and target

The study uses the data from https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction/data. The sample is derived from firm-level accounting ratio data and includes several years of financial indicators of listed companies in Taiwan, labelled with a dual bankruptcy variable (Bankrupt?), where 1 indicates bankruptcy and 0 indicates non-bankruptcy. The dataset contains 6,819 observation records and approximately 95 candidate variables. Due to the relatively small number of bankruptcy samples, the dataset suffers from a severe class imbalance problem. To address this, the samples are balanced before modelling.

3.2. Variable description

The candidate variables cover multiple financial indicators such as profitability, debt-paying ability, operational efficiency, cash flow, and financing structure. In Section 4.1 of this article, the top 20 most relevant indicators selected based on mutual information are presented.

4. Method

The empirical process and technical details of this paper will be explained step by step below for reproduction and verification.

4.1. Step 1: balance the data

Due to the low proportion of bankrupt samples, not handling them may lead the model to be biased towards the predicted non-bankrupt ones. Common processing methods include under-sampling, over-sampling, and weighted loss. In this paper, under-sampling is adopted in the exploration stage, and the results using class weights are compared in the robustness test. The goal of the 1:1 under-sampling is to enable the model to better learn the characteristics of bankrupt firms.

4.2. Step 2: calculate mutual information

To measure the correlation between each feature and the bankruptcy label, mutual information is used. Mutual information can capture linear and nonlinear correlations without relying on the assumption of normality, and is suitable for screening indicators that have a strong relationship with the target variable but may be nonlinear. Discretize each numerical feature, calculate its mutual information score with the binary target, and sort them in descending order.

Based on the balanced data, the top 20 indicators most closely associated with corporate bankruptcy are obtained using R. These indicators span multiple dimensions of financial performance, as listed below:

1.ROA(C) before interest and depreciation before interest

DOI: 10.54254/2753-8818/2025.DL27393

2.ROA(A) before interest and % after tax
3.ROA(B) before interest and depreciation after tax
4.Continuous interest rate (after tax)
5.Net Value Per Share (A)
6.Persistent EPS in the last four seasons
7.Per-share net profit before tax
8.Interest expense ratio
9.Total debt / Total net worth
10.Debt ratio (%)
11.Net worth / Assets
12.Borrowing dependency
13.Net profit before tax / Paid-in capital
14.Retained earnings / Total assets
15.Net income / Total assets
16.Net income / Stockholders' equity
17.Liability / Equity
18.Degree of financial leverage (DFL)
19.Interest coverage ratio (interest expense / EBIT)
20.Equity / Liability

4.3. Step 3: calculate the correlation between features

Highly correlated features can cause multicollinearity, which undermines the stability of coefficient estimates and complicates their interpretation. In this study, the Pearson correlation matrix (in absolute values) is calculated to identify highly correlated pairs with $r > 0.9$.

4.4. Step 4: eliminate features that are highly correlated and have little mutual information

For each pair of variables with a correlation greater than 0.9, the variable with the smaller mutual information score is removed. This procedure ensures that predictors with higher information content are retained, thereby reducing redundancy without discarding potentially important variables at random. This step reduces multicollinearity, simplifies the model, and preserves the predictive information.

4.5. Step 5: retain the first 20 or the remaining features after elimination

After the redundancy removal processing in the fourth step, if the remaining variables exceed 20, the top 20 will be selected based on the mutual information score. If there are fewer than 20, all the remaining variables will be retained for the next step of modelling.

4.6. Step 6: logistic regression

A logistic regression model is established using the retained feature set: the bankruptcy label is taken as the binary dependent variable and the selected features as the independent variables. Coefficients, standard errors, and p-values are estimated for each variable. For ease of comparison, all independent variables are standardized before modelling (mean 0, standard deviation 1). The model report includes coefficient signs, statistical significance (p-values), and odds ratios ($OR = \exp(\beta)$).

DOI: 10.54254/2753-8818/2025.DL27393

The advantage of logistic regression lies in its strong interpretability and the ease of mapping coefficients to risk through odds ratios.

5. Results

$$logit\left(p_i\right)=\ln\left(\frac{p_i}{1-p_i}\right)=\ \beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\beta_5X_{5i} \tag{1}$$

The logistic regression confirms several clear patterns:

The coefficient for return on assets is negative and highly significant. An odds ratio of 0.12 indicates that higher profitability substantially lowers the odds of bankruptcy. The debt ratio shows a positive and significant coefficient. The odds ratio of 6.51 suggests that higher leverage sharply raises the likelihood of bankruptcy. The coefficient for interest coverage is negative and significant. An odds ratio of 0.39 implies that better coverage reduces the odds of bankruptcy. As for internal funds, the variable is negative and significant, with an odds ratio of 0.24, showing that firms with stronger internal capital and a history of profits are less likely to fail. The coefficient of Persistent EPS (4Q avg) is negative and significant. With an odds ratio of 0.46, sustained earnings lower bankruptcy risk. Table 1 presents the detailed estimation results.

Table 1. Results of logistic regression

Variable	Coefficient(β)	Std.Error	p-value	Odds Ratio (Exp(β))	VIF
Intercept	-1.872	0.541	0.001**	-	-
ROA(Return on Assets)	-2.154	0.602	0.000**	0.12	1.45
Debt Ratio	+1.873	0.417	0.000**	6.51	2.12
Interest Coverage	-0.947	0.393	0.015**	0.39	1.89
Retained Earnings/TA	-1.423	0.511	0.004**	0.24	1.73
Persistent EPS (4Q avg)	-0.786	0.292	0.008**	0.46	1.32

6. Conclusion

This study ranks and filters the accounting ratios with mutual information and correlation, and estimates a logistic model to explain bankruptcy status among firms. Profitability is negative and highly significant, showing that more profitable firms are much less likely to fail. Leverage (Debt ratio) is positive and significant, indicating that higher debt exposure sharply raises bankruptcy risk. Interest coverage has a negative and significant effect, implying that firms with a stronger ability to meet interest obligations face lower bankruptcy risk. Retained earnings/TA also reduce bankruptcy odds, confirming the protective role of accumulated profits. Finally, the negative and significant persistent EPS (4Q avg) suggests that sustained earnings lower the bankruptcy risk.

However, this analysis relies solely on accounting ratios and the sample is limited to listed firms in Taiwan, which may restrict generalisation to other circumstances.

Future work may address these limitations by incorporating additional variables, such as country-year macroeconomic variables, and by comparing logistic regression with more complex machine learning approaches to see the consistency of factor effects.

DOI: 10.54254/2753-8818/2025.DL27393

References

[1] Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589–609.
[2] Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109–131.
[3] Shumway, T. (2001) Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business, 74(1), 101–124.
[4] Barboza, F., Kimura, H., and Altman, E. (2017). Machine Learning Models and Bankruptcy Prediction. Expert Systems with Applications, 83, 405–417.
[5] Beaver, W.H. (1966). Financial Ratios as Predictors of Failure. Journal of Accounting Research, 4, 71–111.
[6] Chen, Y.-S., Lin, C.-K., Lo, C.-M., Chen, S.-F., and Liao, Q.-J. (2021). Comparable studies of financial bankruptcy prediction using advanced hybrid intelligent classification models to provide early warning in the electronics industry. Mathematics, 9(20), 2622.
[7] Wang, X., & Brorsson, M. (2024). Augmenting bankruptcy prediction using reported behavior of corporate restructuring. arXiv: 2401.14901. Retrieved from https: //arxiv.org/abs/2401.14901.
[8] Kim, A., & Yoon, S. (2023). Corporate bankruptcy prediction with domain-adapted BERT. arXiv: 2312.03194. Retrieved from https: //arxiv.org/abs/2312.03194.
[9] Zhao, Y., Chen, L., & Zhang, H. (2022). Combining intra-risk and contagion risk for enterprise bankruptcy prediction using graph neural networks. arXiv: 2202.03874. Retrieved from https: //arxiv.org/abs/2202.03874.
[10] Li, W., Härdle, W. K., & Lessmann, S. (2022). A data-driven case-based reasoning in bankruptcy prediction. arXiv: 2211.00921. Retrieved from https: //arxiv.org/abs/2211.00921.
[11] Jiao, Z. (2022). Application of LASSO Algorithm and GBDT Algorithm in Predicting Financial Distress of Companies. Informatica, 46(3), 357–368.