CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17826

Data-Driven Strategies for Early Detection of Corporates' Financial Distress

Donato Riccio[1], Giuseppe Bifulco[2], Paolone Francesco[3], Andrea Mazzitelli[4], Fabrizio Maturo[4]

[1]Machine Learning Engineer, Student at the Master's in Data Science at the University of Campania Luigi Vanvitelli, Caserta, Italy, [2]Department of Economics, Management, Institutions, University of Naples Federico II, Naples, Italy, [3]Faculty of Economic and Legal Sciences, Universitas Mercatorum, Rome, Italy, [4]Faculty of Technological and Innovation Sciences, Universitas Mercatorum, Rome, Italy.

How to cite: Riccio, D.; Bifulco, G.; Paolone, F.; Mazzitelli, A.; Maturo, F. 2024. Data-Driven Strategies for Early Detection of Corporates' Financial Distress. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17826

Abstract

Scholars have taken a keen interest in predicting corporate crises in the past decades. However, most studies focused on classical parametric models that, by their nature, can consider few predictors and interactions and must respect numerous assumptions. Over the past few years, the economy has faced a severe structural crisis that has resulted in significantly lower income, cash, and capital levels than in the past. This crisis has led to insolvency and bankruptcy in many cases. Hence, there is a renewed interest in research for new models for forecasting business crises using novel advanced statistical learning techniques. The study shows that using tree-based methods and hyper-parameters optimization leads to excellent results in terms of accuracy. This approach allows us to automatically consider all possible interactions and discover relevant aspects never considered in past studies. Furthermore, we employ SHAP (SHapley Additive exPlanations) to enhance the explainaibility of our model. This line of research provides fascinating results that can bring new knowledge into the reference literature."

Keywords: *corporate crises; financial distress; statistical learning*

1. Introduction

The first forecasting models to diagnose the state of corporate health date back to the 1960s and 1970s. These models were based on balance sheet indexes, which are still valid tools for preventing the company's state of health. The most used in the literature were profitability (ROE, ROI, ROS, Turnover), liquidity, and capital-financial solidity (e.g. debt ratio). The reference economic and business doctrine furnishes numerous contributions to developing forecasting

models based on the combined observation of performance indicators (Altman, 1968; Altman et al., 2013; Altman and Hotchkiss, 2006; Jones and Hensher, 2004; Shumway, 2001).

In recent years, the economy has gone through a deep, sometimes irreversible, structural crisis, where the settling of income, cash, and capital levels has been significantly lower than in the past. In many circumstances, this crisis has led to insolvency and bankruptcy. Therefore, there is an urgent need to determine warning signs to promptly and effectively activate a recovery process before business continuity is compromised. To this end, there is a growing interest in combining quantitative models based on refined statistical learning techniques that, considering large volumes of data, contemplate the temporal aspect of the balance sheet data, the dynamic element of the context variables, and the interactions of these data over time. Accordingly, there is a revitalised attraction in research for new models for forecasting business crises using novel advanced statistical learning techniques. The study reveals that using XGBoost (Chen, 2016) and hyper-parameter optimisation leads to superior predictive power results. While ensemble methods are often considered black-box models, we employ SHAP (SHapley Additive exPlanations) (Lundberg, 2017) to interpret their predictions, ensuring accuracy and explainaibility.

There are often large datasets available regarding sample size and the number of variables in the business field. However, having many statistical units is undoubtedly a great advantage when the goal is to create a classifier but, on the other hand, having a large number of variables available leads to the so-called curse of dimensionality, which leads to different methodological problems such as the choice of the model, the sparsity of data, the concentration of distances and, above all, multicollinearity. All these aspects make classical parametric approaches obsolete.

In statistics, tree-based classifiers have numerous advantages: they overcome the problems related to the curse of dimensionality, significantly improve performance both in terms of precision and reducing the estimates' variability, offer a dynamic interpretative key to the determinants of a phenomenon, and do not require particular assumptions to be respected. Moreover, tree-based classifiers automatically consider all possible interactions and uncover relevant aspects that may have yet to be considered.

Our study shows that integrating the latest machine learning techniques in this area significantly benefits prediction and explainability. Since these techniques have been little used in the business field to predict the state of crisis, further studies are recommended, introducing context variables and spatial analysis.

2. Material and Methods

The data used for the study were collected from the AIDA dataset, which is a database created and distributed by Bureau van Dijk S.p.A. It contains the balance sheets, personal data, and product information of active and failed Italian capital companies, excluding banks, insurance companies, and public bodies.

The sample selection criteria are based on Legislative Decree 139/2015 and aim to identify Italian non-financial companies that can be categorised as medium or large based on at least two of three thresholds. These criteria exclude small businesses, which are defined as companies that do not exceed two of the following limits at the balance sheet closing date: a net equity total of 4 million EUR, net sales revenue of 8 million EUR, and an average number of 50 employees during the fiscal year. The final sample size of medium to large Italian non-financial companies that meet these specified benchmarks is 37,369.

The research utilises a comprehensive set of financial variables from 2015, 2016, and 2017 to analyse Italian non-financial companies. These variables include annual sales revenue in millions of EUR, EBITDA in millions of EUR, net profit in millions of EUR, total assets in millions of EUR, and the number of employees. Ratios such as EBITDA to sales percentage, debt to EBITDA percentage, and invested capital turnover are included to assess profitability and financial efficiency. The analysis also considers the net equity in millions of EUR, short and long-term debt ratios, liquidity ratios, current ratios, coverage indices of immovable assets, net financial position in millions of EUR, and debt-to-equity ratios. Furthermore, it evaluates financial autonomy from third parties, financial charges on turnover, interest coverage, the cost of borrowed money, bank debts to turnover, and various profitability ratios like Return on Equity (ROE), Return on Sales (ROS), Return on Investment (ROI), and Return on Assets (ROA). Tax liabilities, both short-term and long-term, are also included. Additional variables account for differences between years for these metrics, capturing changes over time. Geographical variables are included based on the province (e.g., Brescia, Milano, Roma, Torino, other) and the sector of activity coded according to NACE (e.g., 25, 28, 41, etc.), with additional categories for other sectors. These variables enable a detailed analysis of the financial health, performance, and geographical and sectoral distribution of the companies in the sample. Considering the differences between years, there are a total of 181 variables in the dataset. The dataset, consisting of 434 instances, was divided into a training set containing 347 instances (80%) and a test set with 87 instances (20%). The training and test sets were balanced regarding the target variable.

The research employs an undersampling technique on the majority class, stratified according to the NACE sector codes. This approach ensures a balanced representation across different sectors by reducing the size of the majority class to match that of the minority class. After the undersampling process, the dataset achieved equilibrium with 217 observations for each class,

labelled '0' and '1', resulting in a balanced dataset. The dataset, consisting of 434 instances, was divided into a training set containing 347 instances (80%) and a test set with 87 instances (20%). The training and test sets were balanced regarding the target variable.

We employed XGBoost (Chen, 2016), a robust machine learning algorithm that uses gradient boosting frameworks for predictive modelling, which is renowned for its performance and speed in classification tasks. The target variable is labeled 1 if the company started filing a procedure for bankruptcy in 2018 or 2019, and 0 otherwise.

For hyperparameter tuning, we utilized Bayesian optimization (Bergstra, 2015), a probabilistic model-based approach for global optimization, running 100 evaluations to efficiently converge on the optimal set of parameters. To validate our model, we implemented a 10-fold cross-validation technique.

3. Results and Conclusions

Without hyperparameter tuning, the initial model achieved a mean ROC AUC score of 0.8990 using 10-fold cross-validation, demonstrating the algorithm's strong predictive capabilities out-of-the-box. After conducting thorough hyperparameter tuning to optimise the model's performance, we obtained an improved mean ROC AUC score of 0.9199, indicating a substantial enhancement in the model's ability to discriminate between bankrupt and non-bankrupt companies. Furthermore, the test set AUC score is 0.9413, ensuring good predicting performance on unseen data.

The SHAP summary plot in Figure 1 visualizes the impact of various financial metrics on the model's prediction across many observations. Positive SHAP values (red) indicate features that positively influence the model's outcome, whereas negative values (blue) show a decreasing effect. Key influencers include net income, return on equity, and the interest coverage ratio, which vary across different data points, showing both strong positive and negative effects on the model's output, highlighting their critical role in determining financial health and stability.

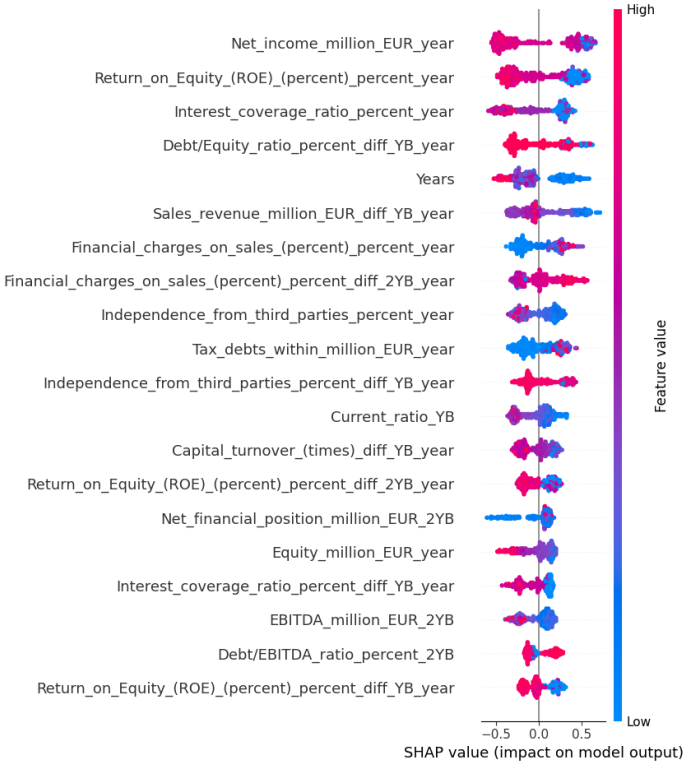

Figure 1. Global explanation SHAP plot.

Figure 2 and Figure 3 illustrate how various financial indicators influence the predicted probability of a company's failure. For a financially healthy company. Features like high net income, positive changes in interest coverage ratio, and return on equity significantly decrease the risk of failure. A large negative financial position increases the likelihood of failure, demonstrating the impact of financial health on company stability. Notably, a negative ROE raises the probability. A negative net income and increases in tax debts within the year also contribute positively.

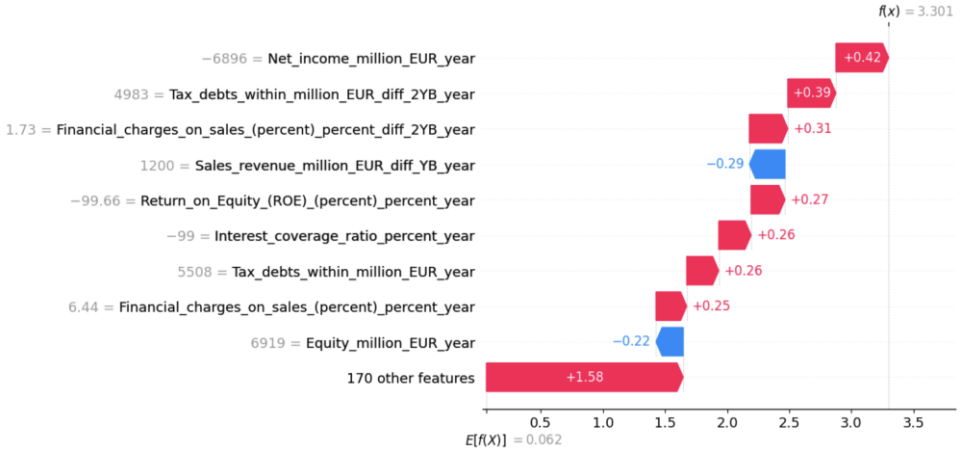

Figure 2. Local Explanation for a Financially Distressed Company.

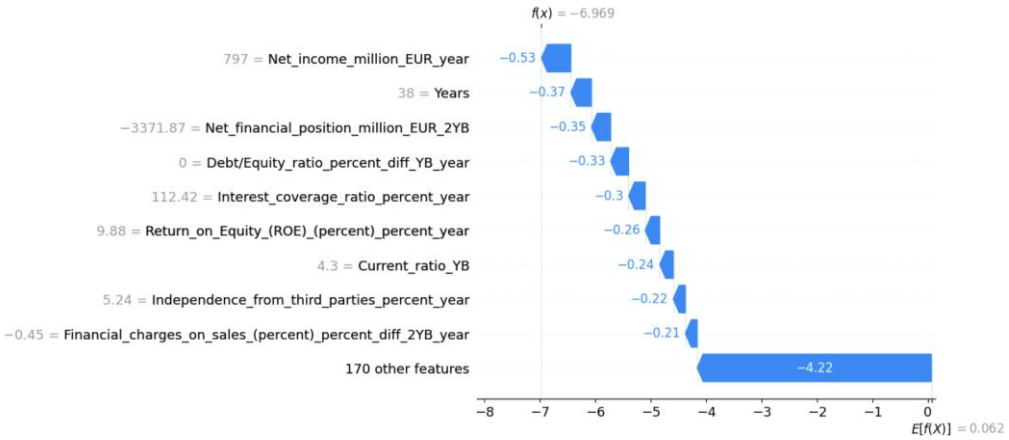

Figure 3. Local SHAP Explaination for a financially healthy company.

The research findings reveal the significant potential of advanced machine learning techniques, particularly tree-based ensemble methods like XGBoost, in predicting corporate financial distress. By leveraging a comprehensive set of financial variables and employing rigorous hyperparameter optimisation, our model achieves superior performance compared to traditional parametric approaches. Moreover, by integrating SHAP, our work ensures that our model maintains explainability, allowing for a clear understanding of the key factors driving the predictions. The SHAP summary plot and local explanations provide valuable insights into the impact of various financial indicators on a company's risk of failure. This combined approach improves predictive performance and offers an explainable approach to corporate crisis forecasting, bridging the gap between accuracy and interpretability. This line of investigation promises compelling insights that have the potential to enrich the current literature. Further

insights and developments on the topic will provide exciting results in generalising to more countries and sectors.

Funding

This research is funded as part of a project supported by the Mercatorum University of Rome under grant code *12-FIN/RIC 2023*.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589-609.

Altman, E. I., Danovi, R., & Falini, A. (2013). Z-Score models' application to Italian companies subject to extraordinary administration. Bancaria, 4, 24-37.

Altman, E. I., & Hotchkiss, E. (2006). Corporate Financial Distress and Bankruptcy (3rd ed.). John Wiley & Sons.

Jones, S., & Hensher, D. A. (2004). Predicting firm financial distress: A mixed logit model. Journal of Accounting and Public Policy, 23(6), 467-487.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. Journal of Business, 74(1), 101-124.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. Computational Science & Discovery, Volume 8, Number 1.