CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21934

Anticipating corporate distress: an interpretable XGBoost early-warning model

Chiara Cornacchione [1], Claudio Lupi[1], Margherita Smarra[1]
[1]Department of Economics, University of Molise, Italy.

How to cite: Cornacchione, C.; Lupi, C.; Smarra, M. (2026). Anticipating corporate distress: an interpretable XGBoost early-warning model. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 26-28 June 2026. https://doi.org/10.4995/CARMA2026.2026.21934

Abstract

This study contributes to the literature on corporate crisis prediction by developing an early-warning model based on financial and non-financial data from small Italian manufacturing firms. The XGBoost algorithm is employed to identify early indicators of distress and to assess predictive performance over one- and three-year horizons. Model accuracy is evaluated through confusion matrix metrics and ROC/AUC analysis, while interpretability is examined via SHAP values. The results indicate higher sensitivity over the one-year horizon, with EBITDA, number of employees, EBT, and equity emerging as key determinants of distress risk. This approach transforms complex data into interpretable signals, helping managers and stakeholders understand which factors indicate a firm's potential crisis.

Keywords: *financial distress prediction; XGBoost; machine learning; SHAP analysis.*

1. Introduction

Corporate failure prediction has been part of the accounting and financial literature since the mid-twentieth century. Classic theories conceptualize bankruptcy as the observable outcome of a gradual deterioration of a company's economic and financial condition, visible in accounting data. A significant body of the literature has been devoted to the development of corporate failure prediction models. The earliest contributions in this field employed classic statistical methods. Fitzpatrick (1932) was among the first to compare financial ratios of failed and non-failed firms, laying the foundations for subsequent studies, such as Beaver (1966), who employed univariate discriminant analysis to assess the predictive ability of individual ratios. However, the awareness of the multidimensional nature of bankruptcy led to the adoption of multivariate models, most notably Altman's (1968) Z-score, which demonstrated the predictive capabilities of several financial ratios combined into a single synthetic score. Lo-

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

gistic and probit models later gained relevance, particularly with Ohlson's (1980) O-score and Zmijewski's (1984) probit analysis.

These approaches constitute the "classical paradigm" (Balcaen & Ooghe, 2006) used, in the majority of the literature as a benchmark to evaluate the performance of more sophisticated methods. Since the 1990s, machine learning models, and more recently, deep learning techniques, have been extensively applied in this field of research as tools to predict firm bankruptcy, due to their ability to detect complex relationships between variables. Techniques such as decision trees, support vector machines, and neural networks are now widely employed, often in combination with ensemble methods, to improve accuracy and robustness (Dasilas & Rigani, 2024). These tools respond to a growing interest in advanced managerial support systems capable of identifying early warning signals and timely detecting the risk of failure. The academic interest in bankruptcy prediction aligns with the policy and regulatory focus on crisis prevention, as reflected in recent legal interventions in the field of insolvency (EU Directive 2019/1023 and the Italian Code of Business Default and Crisis, CBDC). Ensuring business continuity is crucial not only for shareholders and creditors, but also for the entire economic ecosystem.

2. Data and methodology

The present study aims to contribute to bankruptcy prediction research by evaluating the predictive capability of variables across two distinct time horizons. Beyond assessing model accuracy, the analysis identifies the most relevant variables and provides a financial interpretation of their role in bankruptcy dynamics.

The research is based on data available in AIDA (Bureau van Dijk), focusing on an initial sample of 289,389 manufacturing firms, identified using the first two digits (codes 10-33) of the division of the ATECO classification. For each firm, a set of financial and non-financial variables was selected for the period from 2015-2024. Financial variables were used to construct the ratios employed in the analysis and to compute the absolute and percentage year-on-year variations. Since the computation of variations requires data from the previous year, the final data span 2016–2024.

The selection of the variables and ratios employed in the predictive models was conducted in accordance with the established guidelines outlined by the National Council of Chartered Accountants and Accounting Experts (CNDCEC, 2019) and the National Foundation of Accountants (FNC, 2019), as well as on the basis of the relevant academic literature (see, e.g., Dasilas and Rigani, 2024).

At this stage, the dataset included both solvent and distressed firms. In this study, companies are classified as being in financial distress if they enter into formal insolvency proceedings.

Article in Press

This information is available in the AIDA database. Particular attention was paid to the official opening date of the proceedings, which served as the objective reference point for determining the firm's state of financial distress. In particular, the initial year of the proceedings was used to define two dichotomic dependent variables, d1 and d3, which identify companies one year and three years prior to financial distress, respectively. By considering a three-year time horizon, this study aims to extend the analysis beyond the more commonly examined short-term setting, contributing to research on medium to long-term bankruptcy prediction.

The dependent variables were used to construct two separate samples hereafter referred to as d1-set and d3-set. A comprehensive data cleaning process ensured the reliability of both samples. Initially, firms that were not active in 2015 or were less than 1 year old were eliminated from the sample. In a second step, in consideration of the typical class imbalance of bankruptcy prediction datasets, the cleaning process was carried out separately for the positive and negative observations, in order to maximize the number of positive cases in both sets.

For the positive observations (d1=1 or d3=1), variables presenting more than 50% of missing values across all observations were excluded from the dataset. Subsequently, a further cleaning step was applied at the firm-year level, resulting in the removal of any observation containing even a single missing value among the remaining variables. The negative observations were cleaned from those variables that were not available for the positive cases, and from all the observations that included missing values, and then merged with the other observations. Finally, those variables which presented a correlation higher than 0.90 were removed to enhance the interpretation of model outputs.

In light of the very low incidence of insolvency among medium and large firms, the analysis was restricted to small firms, defined as those with fewer than 50 employees and annual turnover not exceeding €10 million.

The data cleaning process resulted in two datasets: 1,128,698 observations in the d1-set and 958,281 observations in the d3-set, both containing the 22 variables described in Table 1:

Table 1. Variables employed in the analysis.

Variables	Description
Age	Year of balance sheet – Year of establishment +1
Asset turnover	Sales/Total asset
ATECO	Dummy variable: the first two digits of the ATECO division
Current ratio	Total current asset/Total current debts
Debt equity ratio	Total debts/ Equity
EBITDA to revenue	EBITDA/Sales
EBITDA	Earnings before interest, taxes, depreciation, amortization
EBT	Earnings before taxes

Article in Press

Employees	Number of employees
Equity	Book value of equity
Interest cover ratio	EBIT/Interest expenses
Interest expenses	Book value of interest expenses
Interest to EBITDA	Interest expenses/EBITDA
Interest to revenue	Interest expenses / Sales
Legal form	Dummy variable: firm's legal form
Liquidity	Working capital / Total asset
Productivity	EBIT / Total asset
Retained earnings	Book value of retained earnings
Sales	Book value of sales
Total debts	Book value of total debts
Wages	Wages and salaries / Number of employees
Working capital	Current Asset – Current Liabilities

As shown in Table 1, relative and absolute variations were excluded due to insufficient data availability.

Both final datasets are characterized by a prominent class imbalance, reflecting the low incidence of corporate failure. In the d1-set, the minority class accounts for 1.91% of the observations, while in the d3-set the proportion of positive cases is 1.69%. Each dataset was randomly split into training and test sets, using a 75/25 split rule: 75% for model training, and 25% for testing.

The technique applied to the datasets was the Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016), noted for its scalability and efficiency in handling structured data and dealing with imbalanced samples, and widely adopted in the literature as an alternative to traditional statistical approaches for capturing non-linear relationships among variables (Zedda, 2024). Its implementation required the tuning of several hyper-parameters, which were selected through a 5-fold cross-validation procedure performed on the training set, using the mean AUC-PR across folds as the selection criterion. The model was trained in R (R Core Team, 2025) using the xgboost package (Chen et al., 2026).

The Random Under Sampling Boosting (RUSBoost) (Seiffert et al., 2010), a hybrid sampling/boosting algorithm, was also tested using Random Forest (Breiman, 2001) as a weak learner. However, due to its comparable performance with XGBoost and its higher computational cost, it was not included in the final version of this analysis. RUSBoost was performed using the R package ebmc (Hsiang Hao, 2022).

Article in Press

3. Results

The discriminatory power of the XGBoost was evaluated using the test sample metrics in Table 2. The results reveal better performance in the one-year horizon, while classification ability declines slightly as the prediction horizon extends, suggesting that the predictive power of the variables weakens over time.

Table 2. Performance metrics of XGBoost on the test sample

	d1-set	d3-set
Sensitivity	0.79109462	0.70064166
Specificity	0.83262038	0.75866388
Balanced accuracy	0.81185750	0.72965277
AUC	0.897382	0.8107927

In order to address the lack of interpretability typical of machine learning models (Crosato et al. 2023) and to assess the contribution of each variable to the prediction, SHAP (SHapley Additive exPlanations, Lundberg & Lee, 2017) values, expressed on a log-odds scale, were computed using the R package shapviz (Mayer, 2025). These values are visualized in three distinct graphs (Figures 1-3), based on a random subset of the original dataset to enhance readability.

The first aspect analysed was the average contribution of each feature to the model's prediction, measured using the mean absolute SHAP value and displayed in a bar plot (Figure 1). Higher values indicate greater importance.

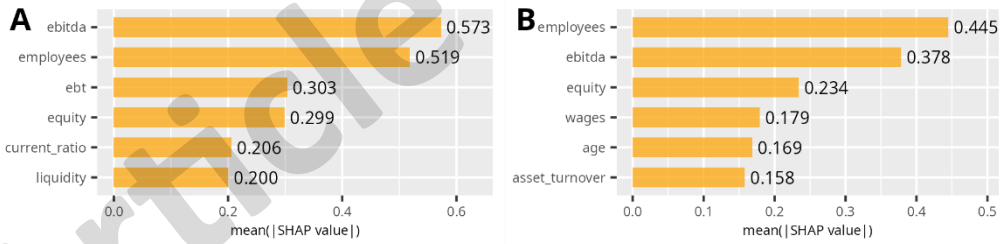


Figure 1. Bar plots for feature importance (Mean |SHAP|) for the d1-set (panel A) and d3-set (panel B)

The bar plot in Panel A of Figure 1 refers to the d1-set, and it shows that the EBITDA plays a central role in the classification, followed by firm size (employees), EBT, and capital structure (equity). In the three-year ahead prediction, shown in Panel B, the most important variables are similar, with a slight difference in their order. The fourth most important variable is a measure of per capita wages and salaries, calculated as the value of wages and salaries divided by the number of employees in the company. Other variables present a lower contribution.

Article in Press

The results shown in Figure 1 are confirmed by the beeswarm plots (Figure 2). The figure presents the SHAP summary plots, which illustrate both the importance of the predictors and the direction of their effect on the model's predictions. Each point represents an observation, while the colour scale reflects the magnitude of the feature's value.

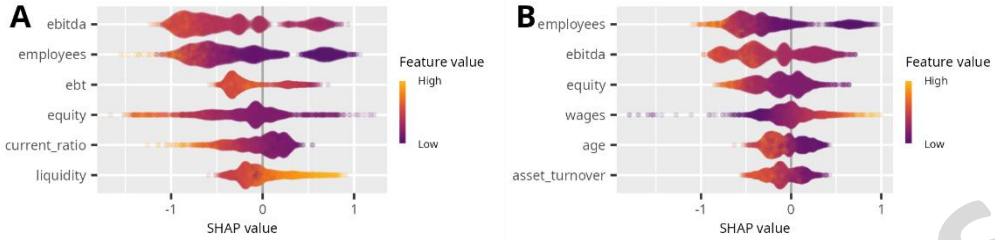


Figure 2. Beeswarm plots for thed1-set (panel A) and d3-set (panel B)

Finally, beeswarm plots are supplemented by SHAP dependency plots (Figure 3), which illustrate how feature values affect the model's prediction. Panel A refers to the one-year time horizon, while panel B refers to the three-year prediction.

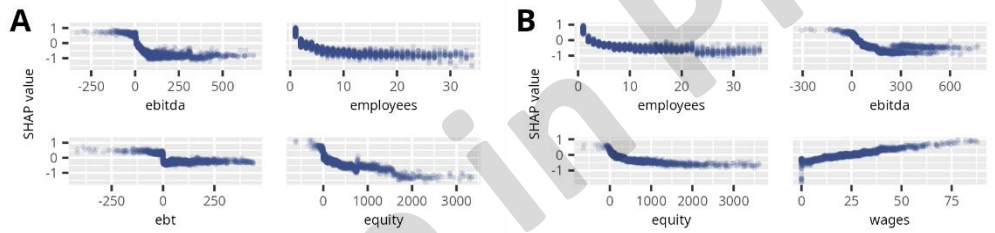


Figure 3. Dependence plots of the four most relevant features of d1-set (A) and d3-set (B)

EBITDA is a key indicator of performance, widely used in practice to evaluate firms' financial health (Grant & Parker, 2002). The scatterplots show that negative values of EBITDA are associated with a higher probability of distress, whereas positive values reduce the probability of default. This suggests that profitable companies are more stable.

Similarly, the number of employees, a measure of the firm size, is a critical determinant of bankruptcy. In detail, it is evident that smaller firms have a higher probability of default, which decreases as the number of employees increases. These results align with the systemic fragility of Italian small enterprises, which are characterized by limited managerial capacity, undercapitalization, and significant reliance on external bank financing.

An interesting behaviour can be observed for EBT (Earning Before Tax) which shows a clear threshold effect around zero, with higher probabilities of default in correspondence of negative values, indicating that firms generating persistent losses represent an early signal of dete-

Article in Press

riorating financial conditions. Conversely, positive values correspond to negative SHAP values, suggesting that profitable firms are associated with a lower likelihood of distress.

Another relevant predictor is equity, defined as the residual interest in the assets of the entity after deducting all its liabilities. The graphs show that values around zero or above are associated to a lower probability of failure, whereas negative values, resulting from operating losses, represent a potential threat to business continuity and indicate a higher probability of entering formal insolvency proceedings. This finding is consistent with the guidelines issued by the CNDCEC (2019).

Over the three-year time horizon (Panel B), the behaviour of most variables remains broadly similar, even though the date of the procedure is further away. The variable wages, which replaces EBT, indicate that higher per capita wages and salaries are associated with an increased probability of default. This behaviour may be related to the rigidity of labour costs in the short term, associated with a higher risk of bankruptcy.

Overall, the interpretation of these results is influenced by the quality and availability of data. Reliance on accounting data is both a strength and a weakness. On the one hand, financial statements are widely available and standardized; on the other hand, financial data are often complex, heterogeneous, and prone to errors, requiring careful cleaning processes to prevent potential distortions and to improve the predictive performance of the model.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance, 23*(4), 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *The British Accounting Review, 38*(1), 63–93. https://doi.org/10.1016/j.bar.2005.09.001

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research, 4*, 71–111. https://doi.org/10.2307/2490171

Breiman, L. (2001), "Random forests," *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 785-794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., Cortes, D. (2026). *xgboost: Extreme Gradient Boosting.* R package version 3.1.3.1, <https://github.com/dmlc/xgboost>

Article in Press

CNDCEC. (2019). Crisi d'impresa: gli indici di allerta. Consiglio Nazionale dei Dottori Commercialisti e degli Esperti Contabili.

Crosato, Liberati, and Repetto (2023) Lost in a black box? Interpretable machine learning for assessing Italian SMEs default, Applied Stochastic Models in Business and Industry, 39:829–846, https://doi.org/10.1002/asmb.2803

Dasilas, A., & Rigani, A. (2024). Machine learning techniques in bankruptcy prediction: A systematic literature review. Expert systems with applications, 255, 124761, https://doi.org/10.1016/j.eswa.2024.124761

Directive (EU) 2019/1023 of the European Parliament and of the Council of 20 June 2019 on preventive restructuring frameworks, on discharge of debt and disqualifications, and on measures to increase the efficiency of procedures concerning restructuring, insolvency and discharge of debt, and amending Directive (EU) 2017/1132 (Directive on restructuring and insolvency).

Fitzpatrick, P. J. (1932). A Comparison of Ratios of Successful Industrial Enterprises with Those of Failed Firms. *Certified Public Accountant*, *12*, 598-605

FNC. (2019). Relazione sul governo societario contenente programma di valutazione del rischio di crisi aziendale. Fondazione Nazionale dei Commercialisti.

Grant, J., & Parker, L. (2002). EBITDA!. *Research in accounting regulation*, 15, 205-212.

Hsiang Hao, C. (2022). *ebmc: Ensemble-Based Methods for Class Imbalance Problem*. doi:10.32614/CRAN.package.ebmc, https://doi.org/10.32614/CRAN.package.ebmc, R package version 1.0.1, https://CRAN.R-project.org/package=ebmc

Italian Code of Business Default and Crisis (CBDC). Legislative Decree No. 14 of 12 January 2019

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.doi: 10.48550/arXiv.1705.07874

Mayer, M. (2025). *shapviz: SHAP Visualizations*. doi:10.32614/CRAN.package.shapviz https://doi.org/10.32614/CRAN.package.shapviz, R package version 0.10.3, https://CRAN.R-project.org/package=shapviz

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, 18*(1), 109–131, https://doi.org/10.2307/2490395

R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 40*(1), 185-197. doi: 10.1109/TSMCA.2009.2029559.

Zedda (2024) Credit scoring: Does XGBoost outperform logistic regression? A test on Italian SMEs, Research in International Business and Finance, 70(Part b) 102397, https://doi.org/10.1016/j.ribaf.2024.102397

Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59–82. https://doi.org/10.2307/2490859

Article in Press