CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17761

Prediction of SMEs Bankruptcy at the Industry Level with Balance Sheets and Website Indicators

Carlo Bottai [1], **Lisa Crosato**[2], **Caterina Liberati**[1]

[1]Department of Economics, Management and Statistics, Università di Milano-Bicocca, Italy, [2]Department of Ecoomics, Ca' Foscari Universty of Venice, Italy.

How to cite: Bottai, C.; Crosato, L.; Liberati, C.; 2024. Prediction of SMEs bankruptcy at the Industry level with balance sheets and website indicators. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17761

Abstract

This paper addresses the importance of industry-specific models for SMEs bankruptcy prediction, building on earlier research finding larger predictive accuracy and enhanced temporal stability. Using Italian data, we propose separate bankruptcy prediction models for a few industries based on balance sheet data and explore the predictive power of SMEs' website HTML code structure. Our findings suggest that website data can serve as a valid complementary source for bankruptcy prediction, with different performances across sectors. We observe a certain degree of sectoral heterogeneity in the importance of balance sheet indicators and website structure, calling for an industry-tailored approach in bankruptcy prediction models.

Keywords: *website data, HTML code, SMEs, supervised learning.*

1. Introduction

The importance of exploring industry-specific variations in bankruptcy models was emphasized in earlier studies by Altman (1973), Altman and Izan (1984) and Platt and Platt (1990, 1991). They advocated for the use of industry-normalized company ratios as primary indicators in early-warnings for bankruptcy, asserting that industry-relative ratios effectively control for industry differences, resulting also in more temporally stable models. More recent research identified differences in variables influencing financial distress based on the technological level of the industry (Madrid-Guijarro et al., 2011). Bragoli et al. (2022) further proved the benefits of incorporating industrial variables into bankruptcy models, particularly in forecasting performance within the manufacturing sector.

An alternative approach in the literature has been the development of prediction models tailored to specific industries (Rikkers and Thibeault 2011, Ciampi and Gordini, 2013). Models estimated on miscellaneous industries have been found to underperform when applied on a single industry (a case study for retail can be found in He and Kamath, 2006). Several works have focused on single sectors, aiming to uncover sector-specific factors influencing SME failure. This is based on the assumption that variables effective in one industry may not be applicable to others. For example, the construction sector is unique due to the extended duration of construction projects (Wang et al., 2024). Compared to other manufacturing industries, the food sector appears more susceptible to productivity shocks and less impacted by bank credit on default risk (Aleksanyan and Huiban, 2016).

Some studies have highlighted limitations of financial ratio variables, noting that they may reflect a company's recorded book value rather than its true value. Moreover, financial ratios may not encompass all information related to financial distress. Typically, financial ratios perform better in the manufacturing sector compared to retail, hospitality, and construction sectors. Concerns have also been raised about the potential impact of deliberate managerial actions distorting the financial situation (Serrano-Cinca et al. 2014, da Silva Mattos and Shasha, 2024). Consequently, researchers have started integrating other variables to enhance the predictive power of models, including non-financial characteristics and economic conditions of companies and industries. Some studies have explored the incorporation of website data, albeit on a limited sample basis (Blázquez et al., 2018, Crosato et al., 2021, Crosato et al., 2023).

Our paper contributes to this literature in two ways. Firstly, we present separate models predicting bankruptcy in six sectors at the 1-digit level of the NACE classification using balance sheet data from Aida, the Bureau Van Dijk (BvD) database describing Italian companies. Secondly, we investigate whether the HTML code structure of Small and Medium Enterprises' (SMEs) websites aids in predicting bankruptcy within the same sectors. The rationale behind utilizing website data is that financially sound firms would likely continue updating and maintaining their websites, while distressed firms might cease or neglect such activities.

Our results indicate that website data serve as a valid complementary source for bankruptcy prediction, with varying performances across sectors. We observe a certain degree of variability in the correct classification also when using balance-sheet data, although to a less extent. The variables selected by the stepwise algorithm also change, indicating sectoral heterogeneity in the importance of balance sheet indicators.

2. Data description

To proceed in our investigation, we combine information about each Italian SME (including any small business with NACE codes from C to N, except K, and incorporated by 2017) from two sources.

The data about each firm's characteristics (number of employees, industry, geographical location and website URL) and balance sheets are from Aida by BvD and refer to the year 2018, for a total of 853,124 Italian SMEs. We define as defaulted by 2019 ($y = 1$) the 23,872 SME whose balance sheet was available up to 2018 and not in 2019; i.e., 2.8 per cent of the sample. Instead, we consider as survived to 2019 ($y = 0$) any SME lastly observed later than 2018 or that is still ‘active’ with no ongoing default procedure, according to BvD (829,252 SMEs).

Data about each firm’s website are from the Wayback Machine by the Internet Archive (IA). For each SME, we pick the URL of its corporate website, if present on Aida; we download the snapshot closest to mid-June 2018, if archived by IA; and preserve as correct websites only those on which we detect the VAT number, phone number, postal code, or full address of the corresponding enterprise (see Blázquez et al., 2018 and Bottai et al., 2022 about this methodology).

To work on SMEs with both balance sheets and website information, our sample was reduced to 152,559 SMEs, of which 1.3 per cent defaulted by 2019. As shown in Fig. 1–3, firm size seems making some difference in terms of default probability. On the contrary, there is little evidence of industry- or location-specific effects, apart from exceptions like the manufacturing sector.

For each downloaded website, we build several size variables, capturing different aspects of the complexity of the website’s homepage. Moreover, we extract the HTML tags from its code and count the number of times each website is used on the web-page.

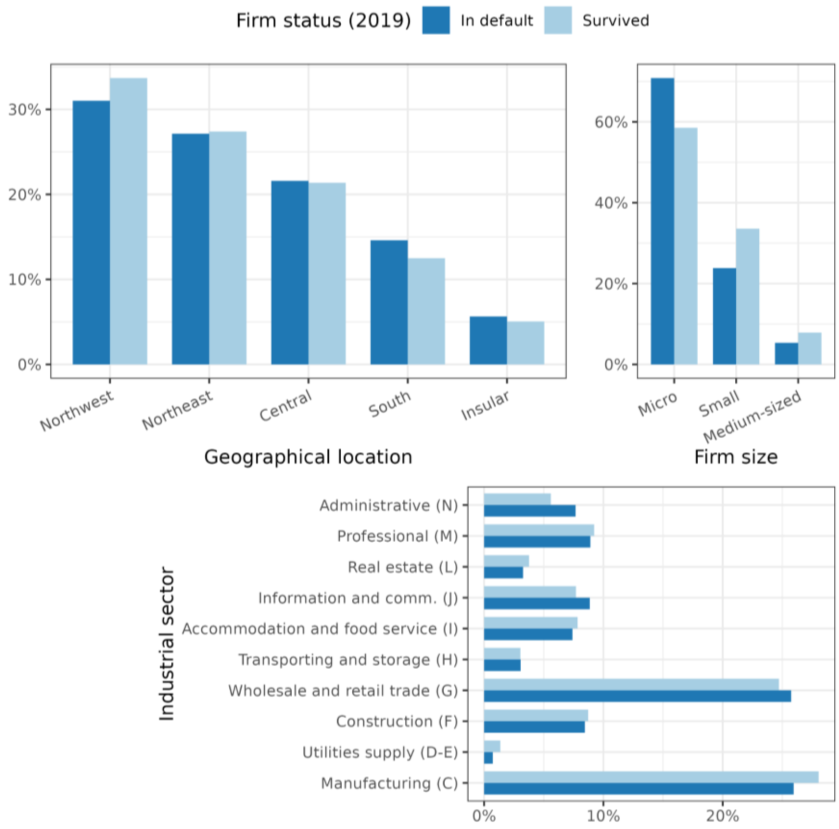

Figure 1. Percentage distributions of defaulted and survived SMEs by location, size and industry. Values computed on the 152,559 instances composing the final sample.

3. Results

We selected the NACE sections that had at least 1% of bankruptcies in 2019. These sectors include Manufacturing (C), Construction (F), Wholesale and retail trade; repair of motor vehicles and motorcycles (G), Transportation and storage (H), Information and communication (J), and Administrative and support service activities (N). Our dependent variable is default in 2019, whereas all predictors refer to the year 2018. For each sector, we divided the companies into two groups: training set (70%) and test set (30%). The training set was balanced to match the number of survived companies with the number of bankrupt ones, while the testing set remained unbalanced in favor of survived companies. Subsequently, we estimated stepwise logistic regressions separately on balance sheet and HTML indicators. We chose logistic regression because it is a well-known method that offers a straightforward interpretability.

Classification performances vary significantly from one sector to another, using either set of proposed variables (Table 1). Specifically, the overall classification measured by the geometric mean between sensitivity and specificity is very good for sectors C and G using both balance sheets and HTML code indicators. It is still fair in sectors H, J, and N when using balance sheet indicators and low, but still larger than 50%, in sectors F, J, and N when using HTML variables.

Table 1. Classification performances based on balance sheets or HTML indicators by industry (geometric mean of sensitivity and specificity).

Indicators	**C**	**F**	**G**	**H**	**J**	**N**
Balance sheets	0.783	0.783	0.759	0.731	0.683	0.712
HTML	0.780	0.540	0.750	0.591	0.538	0.532

Considering that HTML variables are available in real-time and completely free and accessible, the results in sectors C and G are noticeable. A probable reason behind the lower metrics within the remaining sectors lies in how companies in those sectors utilize their websites. For example, SMEs in sectors F and H most likely make poor use of their websites, hence they do not invest in the necessary technology. In other words, the limited discriminant power of HTML variables might be due to a general low level of website quality. The same reasoning can be extended to sectors J and N where, on the contrary, a high level of website quality is to be expected due to the need for well-functioning websites.

The balance sheet variables included by the model in each sector are collected in Table 2, where it can be seen that no variables are selected in all sectors. Out of the total 27 stepwise-selected variables, one is present in 4 sectors (ROE), four are present in three sectors (asset turnover, solvency ratio, profit and sales), eight in two sectors, and the remaining fourteen in one sector only. Therefore, most variables prove useful for discrimination in only one sector or the other. Notice that the set of discriminant variables for the manufacturing sector is the largest one (10 variables), followed by the sets for sectors G and N (9), H (8), and F and J (5 variables only).

In conclusion, these results highlight the importance of applying sector-specific prediction models. While working within sectors may have the disadvantage of a small number of failed companies to work with, considering a global model with dummy variables identifying sectors has the drawback of not highlighting sector-specific variables. Thus, they do not provide guidance to financial intermediaries on the aspects to focus on when evaluating firms in a sector. Furthermore, in a global model, classification metrics are usually reported at the aggregate, and not sector-specific, level. Further research should investigate in a similar fashion the role of HTML features, as well as estimate more complex models possibly combining the two types of variables. Augmenting HTML code with textual analysis could also improve classification metrics.

Table 2. Significant balance sheet variables (5% level) selected by the stepwise logistic regressions in each industry

	Industry					
Variable	**C**	**F**	**G**	**H**	**J**	**N**
Inventories	X					
Asset turnover	X			X		X
Asset	X					
Net working capital	X				X	
Tangible fixed assets	X					
Personnel cost	X		X			
EBITDA/Sales	X			X		
ROS	X					
Added value per employee	X			X		
Solvency ratio	X	X	X			
Cost of production services		X				
Total debts			X			
Profit		X	X			X
Revenue from sales		X	X			X
ROE		X	X	X	X	
Shareholder funds			X	X		
Financial fixed assets			X			
Long-term payable due to banks (yes)			X		X	
Intangible fixed assets				X		
Current assets				X		X
Added value				X		X
Short-term debt over the total debt					X	
EBIT					X	
Current ratio						X
EBITDA						X
Raw consumable materials and goods for resale						X
Wages						X
n. of relevant variables	10	5	9	8	5	9

References

Aleksanyan, L., & Huiban, J. P. (2016). Economic and financial determinants of firm bankruptcy: evidence from the French food industry. *Review of Agricultural, Food and Environmental Studies, 97,* 89-108.

Altman, E., 1973, Predicting railroad bankruptcies in America, Bell Journal of Economics and Management Science, 184-211.

Altman, E. and H. Izan, 1984, Identifying corporate distress in Australia: An industry relative analysis, Working paper (New York University).

Blázquez, D., Domènech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival? Online Information Review 42(6), 956–970. doi:10.1108/OIR-11-2016-0321.

Bottai, C., Crosato, L., Domènech, J., Guerzoni, M., & Liberati, C. (2022). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In Proceedings of the 2022 ACM Conference on Information Technology for Social Good, 338–344. New York, NY: Association for Computing Machinery. doi:10.1145/3524458.3547246.

Bragoli, D., Ferretti, C., Ganugi, P., Marseguerra, G., Mezzogori, D., & Zammori, F. (2022). Machine-learning models for bankruptcy prediction: do industrial variables matter?. *Spatial Economic Analysis*, *17*(2), 156-177.

Ciampi, F., & Gordini, N. (2013). Small enterprise default prediction modeling through artificial neural networks: an empirical analysis of i talian small enterprises. *Journal of Small Business Management*, *51*(1), 23-45.

da Silva Mattos, E., & Shasha, D. (2024). Bankruptcy prediction with low-quality financial information. *Expert Systems with Applications*, *237*, 121418.

He, Y., & Kamath, R. (2006). Business failure prediction in retail industry: an empirical evaluation of generic bankruptcy prediction models. *Academy of Accounting and Financial Studies Journal*, *10*(2), 97.

Madrid-Guijarro, A., Garcia-Perez-de-Lema, D. and van Auken, H. (2011), An analysis of non-financial factors associated with financial distress, Entrepreneurship & Regional Development, Vol. 23 Nos 3-4, pp. 159-186.

Platt, H. D., & Platt, M. B. (1990). Development of a class of stable predictive variables: the case of bankruptcy prediction. *Journal of Business Finance & Accounting*, *17*(1), 31-51.

Platt, H. D., & Platt, M. B. (1991). A note on the use of industry-relative ratios in bankruptcy prediction. *Journal of Banking & Finance*, *15*(6), 1183-1194.

Serrano-Cinca, C., Fuertes-Callén, Y., Gutiérrez-Nieto, B., & Cuellar-Fernández, B. (2014). Path modelling to bankruptcy: causes and symptoms of the banking crisis. *Applied Economics*, *46*(31), 3798-3811.

Wang, J., Li, M., Skitmore, M., & Chen, J. (2024). Predicting Construction Company Insolvent Failure: A Scientometric Analysis and Qualitative Review of Research Trends. *Sustainability*, *16*(6), 2290.