Journal of Corporate Finance 94 (2025) 102830

Contents lists available at ScienceDirect

Journal of Corporate Finance

journal homepage: www.elsevier.com/locate/jcorpfin

Machine Learning for the Unlisted: Enhancing MSME Default Prediction with Public Market Signals

Alessandro Bitetto [a], Stefano Filomeni [b,*], Michele Modina [c]

[a] Carlo Cattaneo University - LIUC, School of Economics and Management, Italy
[b] University of Essex, Essex Business School, Finance Group, Colchester, UK
[c] University of Molise, Department of Economics, Italy

ARTICLE INFO

Editor: R. Iyer

JEL classification:
C52
C53
D82
D83
G21
G22

Keywords:
Credit risk
Distance to Default
Machine learning
Market information
Probability of Default
Shapley
XAI

ABSTRACT

This paper contributes to the growing body of research on private firms, particularly private firm accounting. We explore the economic factors that drive improvements in the default prediction of unlisted private firms using peers' market-based information. Specifically, we examine how the market-based default probability of a peer firm can provide valuable insights into the often noisy accounting data of private firms. Our analysis delves deeply into these economic issues to uncover essential insights. To address our research question, we utilize a granular proprietary dataset of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) that are required to disclose their financial statements publicly. We propose a novel public–private firm mapping approach to investigate whether incorporating peers' market-based information improves the accuracy of default predictions for private unlisted firms. Our mapping approach matches the market information of listed firms with private firms through a data-driven clustering technique using Neural Network Autoencoder. This method enables us to link the Merton Probability of Default (PD) of public peers to the corresponding private firms within the same cluster. We then apply five statistical techniques – linear models, multivariate adaptive regression splines, support vector machines, k-nearest neighbours and random forests – to predict corporate default among private firms, comparing model performance with and without the inclusion of Merton's PD estimated using peers' market-based information. To assess the contribution of each predictor, we employ Shapley values. Our results demonstrate a significant improvement in default prediction for unlisted private firms when incorporating peers' market-based information, confirming that the noisy accounting data of private firms alone hinders accurate default prediction. Furthermore, our findings highlight the importance for banks to broaden the scope of information used in credit risk assessments of private firms. These results have important policy implications for financial institutions and policymakers, providing a tool to mitigate the challenges posed by the noisy information disclosure of MSMEs while ensuring more accurate credit risk assessments.

1. Introduction

Evaluating the credit risk of non-listed MSMEs, which often lack transparency that hampers access to credit, presents significant challenges. These firms experience information asymmetries, which regulators try to mitigate by requiring financial disclosures. However, these disclosures are often unreliable, particularly in private firms, where distorted and noisy accounting data can hinder

* Corresponding author.
E-mail addresses: abitetto@liuc.it (A. Bitetto), stefano.filomeni@essex.ac.uk (S. Filomeni), michele.modina@unimol.it (M. Modina).

https://doi.org/10.1016/j.jcorpfin.2025.102830
Received 19 April 2024; Received in revised form 15 May 2025; Accepted 26 May 2025
Available online 11 June 2025

accurate default prediction (Beuselinck et al., 2023). This is especially important because MSMEs represent a large segment of the corporate sector in Europe. The lack of reliable financial reporting makes it difficult for lenders to assess credit risk, which is critical in maintaining the stability of the banking sector.

To address these challenges, this study introduces a novel approach by applying market data from comparable listed firms as a proxy for missing market signals in MSMEs. By integrating market indicators, which are less prone to manipulation than accounting data, this study aims to reduce information asymmetries and enhance the accuracy of creditworthiness evaluations.

Unlike larger companies, MSMEs face higher default risks and significant information opacity (Burgstahler et al., 2006). While MSMEs rely more on soft, relationship-based information for credit access (Berger and Udell, 2002), the increasing dominance of large banking conglomerates has limited the effectiveness of traditional relationship banking (Filomeni et al., 2021). This underscores the need for alternative, hard information-based methodologies for assessing credit risk. Given their importance in many economies, developing credit models tailored to MSMEs is critical for minimizing both expected and unexpected losses.

This paper develops a hybrid credit risk model for MSMEs that combines accounting data with market data from listed firms, i.e., peers. Using a dataset of 10,136 unlisted Italian MSMEs, we employ advanced statistical techniques (e.g., random forests, multivariate regression) to predict default. The novel contribution lies in estimating Merton's Probability of Default (PD) using market data from matched listed companies via a data-driven clustering approach, avoiding any assumptions based on size, industry, or number of employees.[1] This novel mapping process between unlisted and listed firms contributes significantly to our methodology. By leveraging this mapping, we gain deeper insights into private firms' risks, not only in terms of PD but also through various market indicators. Our results are robust across alternative market measures, such as stock price volatility and market leverage, consistent with Campbell et al. (2008).

Our findings reveal a unique economic channel through which the PD of public firms can predict the default likelihood of matched private firms. This approach uncovers the economic drivers behind improved default predictions for unlisted private firms using market data from their peers. Market data provides insights into the noisy accounting data of private firms, and due to the higher risk of MSMEs, it more effectively captures their corporate default risk. Market data indeed responds more quickly to changes in borrowers' creditworthiness than accounting measures, reflecting common risk factors between public and matched private firms and capturing aggregate risk not accounted for in firm-specific measures.

This study makes two main contributions. First, it addresses challenges in private firm financial reporting by implementing predictive models with enhanced explainability, overcoming the impact of noisy accounting data. Recent studies have applied Machine Learning (ML) models to economic problems (Mullainathan and Spiess, 2017; Akbari et al., 2021; Avramov et al., 2021; Olson et al., 2021), with Kim et al. (2020) surveying ML applications in credit default prediction. Linear classification models, such as LDA or logistic regression (Shumway, 2001; Altman and Sabato, 2007; Bauer and Agarwal, 2014; Tian et al., 2015), show lower predictive accuracy than non-linear models like Random Forest (RF) or Boosted Trees (BT) (Zhu et al., 2019; Barbaglia et al., 2021). However, most studies focus solely on performance improvements over linear models without exploring input variable relevance and their effect on predictions. Moscatelli et al. (2019) attempt to explain the overall importance of input variables, while Albanesi and Vamossy (2019) and Barbaglia et al. (2021) emphasize explaining individual predictions. This paper extends this line of research (eXplainable Artificial Intelligence) by implementing both non-linear parametric and non-parametric ML algorithms, offering not only default predictions but also advanced techniques like Permutation Feature Importance (Fisher et al., 2018) to evaluate variable relevance and Shapley Additive Explanations (Lundberg et al., 2020) to explain how each variable contributes to a single observation's predicted probability of default. Additionally, to the best of our knowledge, we are the first to introduce a novel clustering technique using Artificial Neural Networks to map financial ratios and compare unlisted MSMEs with listed companies.

Second, our hybrid credit scoring models, combining market and accounting information, outperform models relying on only one type of data. While previous studies have applied the Merton model to private firms (Rikkers and Thibeault, 2009; Andrikopoulos and Khorasgani, 2018; Falkenstein et al., 2000; Filomeni et al., 2024), our study is the first to incorporate peer market information into credit risk modelling for unlisted companies. We demonstrate that adding Merton's PD measure to a multivariate regression model, already incorporating accounting data, enhances corporate default prediction accuracy.

Our findings are consistent with prior research but offer key methodological improvements. Unlike Falkenstein et al. (2000) and Rikkers and Thibeault (2009), who rely on industry-wide market averages or discounted cash flow methods to estimate the market value of unlisted firms, we adopt a data-driven "comparable approach". Specifically, we estimate the market value of unlisted firms by matching them with listed counterparts through a data-driven clustering technique. This approach overcomes several limitations of the KMV model for private firms (Falkenstein et al., 2000) and the cash flow-based valuation method (Rikkers and Thibeault, 2009). Compared to Andrikopoulos and Khorasgani (2018), our study introduces a more sophisticated and flexible methodological framework for predicting unlisted firms defaults. While their approach estimates market-informed default probabilities using linear regression to project Merton-KMV EDFs from listed peers onto accounting ratios of unlisted SMEs, our method employs an autoencoder-based clustering technique to match unlisted firms with listed peers in a non-linear latent space. Additionally, our framework enhances interpretability through SHAP values and Permutation Feature Importance, whereas the prior study relies solely on ROC-based validation with logistic models. Overall, our methodology provides stronger predictive performance, greater model transparency, and a more robust foundation for practical credit risk assessment. Our methodology is further supported by corporate finance literature on equity valuation of private firms (Andrikopoulos and Khorasgani, 2018; Baker and Ruback,

[1] Hence, we argue that our modelling approach for evaluating the market risk of MSMEs is not prone to estimation or misspecification error. Instead, we argue that this is the only feasible modelling approach for capturing firms' market risk for which no market data exists.

1999; Alford, 1992; McCarthy, 1999). In addition, we leverage a unique dataset of 10,136 Italian MSMEs collected from 113 cooperative credit banks, rather than relying on a single financial institution as in Rikkers and Thibeault (2009), which enhances the representativeness, robustness, and external validity of our results across both manufacturing and service sectors.

This research has important policy implications for banks, as it demonstrates that incorporating market data into hybrid credit scoring models can improve forecasting accuracy for unlisted MSMEs, helping banks make better-informed lending decisions. These findings contribute to forward-looking financial risk management frameworks (Breden, 2008; Rodriguez Gonzalez et al., 2018; Bitetto et al., 2023b,a) aimed at addressing challenges related to MSMEs' noisy financial disclosures and improving credit risk assessment accuracy.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature, Section 3 discusses the data, Section 4 presents the econometric methodology, Sections 5 and 6 present the empirical results and robustness tests, and Section 7 concludes.

2. Literature review

Much of the literature on corporate credit risk has historically focused on accounting-based models (Beaver, 1966; Altman, 1968; Ohlson, 1980), which use financial ratios from firm balance sheets to assess creditworthiness.[2] These models are particularly prevalent for private firms due to the lack of market data and suggest that adding relevant accounting variables improves the forecasting ability of corporate default.[3] However, limitations such as multicollinearity among ratios, heterogeneity in accounting standards (Stickney and Weil, 1997), and increased ease of manipulation (Beuselinck et al., 2023) prompted researchers to explore alternative or complementary approaches to improve the accuracy of creditworthiness evaluation.

More recent studies have highlighted the value of integrating non-accounting data, including legal, audit, and bank relationship information, into default prediction frameworks (Altman et al., 2010; Bhimani et al., 2010; Dierkes et al., 2013; Fiordelisi et al., 2014). These sources, especially those capturing soft information from bank-firm interactions (Gropp and Guettler, 2018; Liberti and Petersen, 2017), have been shown to improve the predictive performance for private firms, which typically lack publicly traded securities and, therefore, market data.[4]

In parallel, a distinct strand of literature has applied market-based models, such as Merton's Distance to Default (DD), to listed firms (Bharath and Shumway, 2008; Byström, 2006; Vassalou and Xing, 2004). These studies consistently find that market-implied measures capture forward-looking risk not always reflected in accounting data (Agarwal and Taffler, 2008; Hillegeist et al., 2004; Doumpos et al., 2015). While such models have traditionally been applied to public firms, a growing interest exists in extending them, or using peer market data as a proxy, to estimate the credit risk of private firms. This approach remains underexplored but has shown promising results in enhancing default predictive accuracy for listed companies when integrated into hybrid frameworks (Hernandez Tinoco and Wilson, 2013; Das et al., 2009; Hernandez Tinoco et al., 2018).

Our work builds on this literature by proposing a structural market-based approach that leverages peer market data to model credit risk for private firms, addressing the gap between purely accounting-based predictions and market-informed models. Indeed, most default prediction models either rely only on accounting data (e.g., financial ratios from balance sheets), or they rely only on market data (e.g., stock prices or CDS spreads), but very few successfully integrate the two, especially when dealing with private firms that do not have their own market data. Our work fills this gap by proposing a way to bring market-based insights, via peer firm market data, into the default prediction of private firms, and thereby blending the strengths of both approaches, i.e., forward-looking nature of market models in addition to detailed but often retrospective fundamentals from accounting models. This is important because private firms typically lack direct market signals, so using peer market data as a proxy is both novel and practically valuable.

3. Data

We use two sources of information for our analysis: a proprietary one, consisting of granular information on 10,136 Italian unlisted micro-, small-, and mid-sized enterprises (MSMEs), and a public one, comprising data on comparable publicly listed companies, hereinafter referred to as the peers.

3.1. MSME data

We exploit a unique and disaggregated dataset on an unbalanced panel sample of 10,136 firms and 113 cooperative credit banks for a total of 19,743 firm–year observations over the period 2012–2014. Specifically, we consider firms with fewer than 250 employees and revenue of no more than 50 million. Our sample of unlisted firms had established credit relationships with cooperative banks with loans issued prior to 2012. We selected a subset of 22 financial ratios out of 30, removing the ones showing a high partial correlation with many other ratios. Therefore, some ratios with only mild correlation to one other ratio are still retained because the models we use for the predictions are robust to multicollinearity. Tables 1 and A.1 in the Appendix report the

[2] For a broader overview of classic accounting-based models and their evolution, see Edminster (1972), Blum (1974), Grice and Ingram (2001), Pindado et al. (2008), Louzada et al. (2016)

[3] Additional studies on default prediction for small or private firms include Peel et al. (1986), Keasey and Watson (1987), Calabrese et al. (2016), Mselmi et al. (2017).

[4] Studies such as Foglia et al. (1998), Norden and Weber (2010), Volk (2012), Qian et al. (2015) explore relational banking data – such as loan officer discretion, account activity, and multi-bank relationships – as predictors of default risk.

Table 1
List of input variables for MSMEs dataset.

Variable	Description	Mean	St.Dev.	Min	5th perc	Median	95th perc	Max
1 - Oth Reven on Reven	Other revenues on revenues	0.03	0.05	0	0	0.01	0.19	0.19
2 - Deprec on Costs	Depreciation on costs	0.06	0.08	0	0	0.03	0.26	0.34
3 - Pay to Bank on Assets	Payables to banks on current assets	0.83	1.5	0	0	0.47	2.73	11.25
4 - Cashflow on Reven	Cash flow on revenues	0.08	0.08	0.01	0.01	0.06	0.26	0.41
5 - Fixed Asset Cov	Fixed asset coverage	1.15	1.99	0.07	0.07	0.57	4.89	11.17
6 - Labour Cost on Reven	Labour cost on revenues	0.56	0.32	0	0	0.61	1.03	1.03
7 - ST Pay on Due to Bank	Short-term payables on amounts due to banks	2.05	2.46	0.16	0.21	1	9.49	9.49
8 - Tot Debt on ST Debt	Total debt on short-term debts	2.3	2.04	1	1	1.67	5.79	13.35
9 - Tot Debt on Net Worth	Total debt on net worth	7.92	10.2	0.35	0.48	3.73	36.5	41.94
10 - Pay to Suppl on Net Worth	Payables to suppliers on Net worth	2.69	3.48	0.04	0.12	1.01	13.01	13.01
11 - Pay to Suppl on Tot Debt	Payables to suppliers on Total debt	0.4	0.22	0.02	0.07	0.36	0.84	0.84
12 - Inventory Duration	Inventory on revenues x 365	0.78	1.09	0.02	0.03	0.5	2.68	7.16
13 - Quick Ratio	Current assets less inventory on current liabilities	1.41	1.1	0.04	0.22	1.18	3.42	6.54
14 - Debt Burden Index	Financial interest on EBITDA	0.4	0.38	0.01	0.02	0.23	1	1
15 - Fin Int on Reven	Financial interest on revenues	0.02	0.02	0	0	0.02	0.08	0.1
16 - Fin Int on Added Val	Financial interest on added value	0.08	0.07	0.01	0.01	0.05	0.25	0.25
17 - Net Worth on LT Eqt/Pay	Net worth on long-term equity and payables	0.49	0.31	0.05	0.06	0.48	1	1
18 - Net Worth on NW+Invent	Net worth on net worth and inventories	0.64	0.3	0.07	0.1	0.7	1	1
19 - ROA	Return on Assets	0.02	0.07	−0.1	−0.1	0.01	0.17	0.27
20 - ROD	Return on Debt	0.03	0.01	0	0	0.02	0.05	0.05
21 - Working Cap Turnover	Revenues on net working capital	2.3	2.25	0.25	0.55	1.77	5.78	18.32
22 - Turnover	Revenues on total assets	1.16	0.74	0.07	0.2	1.02	2.84	3.17

Table 2
List of control variables for MSMEs dataset. Percentages refer to proportions over the entire sample.

Variable	Target										
FIRM SIZE		Large	Medium	Micro	Small	TOTAL					
	0	2.4%	9.1%	54.7%	27.1%	93.2%					
	1	0.2%	0.8%	3.8%	2%	6.8%					
	TOTAL	2.6%	9.9%	58.5%	29%						
DUMMY INDUSTRY		Manufacturing	Services	TOTAL							
	0	33.1%	60.1%	93.2%							
	1	2.2%	4.6%	6.8%							
	TOTAL	35.3%	64.7%								
INDUSTRIAL SECTOR		Food & Accommodation	Energy supply	Entertainment	Information & communication	Manufacturing	Professional, scientific and technical	Real estate	Trade	Transportation	TOTAL
	0	4.5%	1.3%	1.4%	4.1%	33.1%	8.7%	6.9%	29.2%	4%	93.2%
	1	0.6%	0.1%	0.1%	0.2%	2.2%	0.6%	0.7%	2.1%	0.3%	6.8%
	TOTAL	5%	1.4%	1.4%	4.4%	35.3%	9.3%	7.6%	31.3%	4.3%	
REGION		Central	Islands	North-east	North-west	South	TOTAL				
	0	1.4%	2.3%	52%	26.2%	11.4%	93.2%				
	1	0.1%	0.3%	3.6%	1.8%	0.9%	6.8%				
	TOTAL	1.5%	2.6%	55.6%	28%	12.3%					
FIRM TYPE		Enterprises	SEO	Small business	TOTAL						
	0	75.7%	3.4%	14.1%	93.2%						
	1	5.9%	0.1%	0.8%	6.8%						
	TOTAL	81.6%	3.5%	14.9%							

complete list of variables with descriptions and statistics and their pairwise correlations, respectively. The target variable we want to predict is a binary flag indicating whether the firm defaulted (1) or not (0). In our context, the default flag is assigned when the client becomes insolvent within the last 12 months following loan disbursement, with a past due of at least 180 days. Specifically, we utilize historical financial data from the defined period to predict defaults occurring in the future. It is crucial to emphasize that the embedding process relies exclusively on historical data available up to the point of prediction. We do not incorporate any future default information into the training data, thereby ensuring the integrity of our predictive model and avoiding any lookahead bias. Moreover, we control for additional categorical variables, describing both time-invariant characteristics of our unlisted firms, such as the region to which the firm belongs and industry and time-varying characteristics, such as the size of the firm and its level of funding risk. Table 2 reports the list of control variables used in the analysis and their distribution across the two target classes.

3.2. Peers data

We select a panel of 40 Italian listed firms, evenly distributed across the manufacturing and services sectors. The choice of the peers follows a mapping of the most representative firms by size, industry and number of employees, to match the characteristics of the MSMEs. We collect accounting figures from Orbis database, developed by Bureau Van Dijk (a Moody's analytics company), by matching the VAT[5] code for each given peer firm[6] The accounting figures are used to reconstruct and match or proxy the 22 financial ratios of the MSMEs dataset. Moreover, daily stock prices are collected from the Refinitiv Eikon database and are used to compute the annual asset volatility of comparable publicly listed companies. Table A.2 in the Appendix reports the statistics for the 22 variables as well as for the volatility, total assets and total liabilities used as inputs in the Merton's model formula, as described in Section 4.1.

3.3. Representativeness of the sample

In order to quantitatively prove the representativeness of the selected Peers group and the adequacy of the MSMEs sample, we start by comparing the distribution of the 22 financial ratios. Fig. A.1 in the Appendix depicts the comparison of the 22 variables between the two datasets, including the Kolmogorov–Smirnov p-value for testing the alternative hypothesis that the samples come from different populations. Results show that, for most variables, distributions for MSMEs and Peers are similar showing that the selected peers are adequately representative of our sample of unlisted micro-, small-, and mid-sized enterprises (MSMEs).

Furthermore, we conduct a validation of our matching process, examining how closely the characteristics of MSMEs align with those of their listed Peers. This involves a more detailed analysis of the financial structure of the firms to ensure that the matched samples are indeed comparable. Fig. A.2 represents the distribution of the bank debt level of the MSMEs compared to the Peers group. The density curves show how the level of debt varies between MSMEs and listed companies. Although MSMEs have a more concentrated distribution at lower debt levels, MSMEs tend to have similar debt-to-asset ratios to the Peers group. The overlap of the two areas shows no significant differences in debt profiles. Both MSMEs and listed companies appear to have a uniform distribution. The Kolmogorov–Smirnov p-value (0.24852) also suggests that there are no significant differences between the two groups. Therefore, the presence of a capital structure with similar debt levels suggests that the MSMEs in our sample exhibit the same financial vulnerability as the peer group of listed companies.

Finally, we incorporate the credit relationship dimension into our analysis, exploring how variations in credit dependency between MSMEs and listed firms may impact default likelihood and overall risk assessment. To assess the impact of credit relationships with cooperative banks, particularly since our dataset predominantly consists of firms with established credit relationships and to explore how these relationships influence financial health and risk exposure, we run an analysis only for firms with very few banking relationships for robustness. In particular, to test the robustness of the results, we evaluate the default prediction procedure described in Section 4.3 on a subset of the MSMEs sample, including only the ones that have a credit relationship with a single bank, for a total of 7,155 firms. The analysis conducted shows that our results hold because the predictive performances are comparable to the ones on the full sample, as shown in Table C.6 in Appendix C.

4. Methodology

This paper aims to assess the impact of market information, i.e., Merton's probability of default (PD), in predicting the corporate default risk of unlisted firms, in addition to accounting-based measures. Our analysis can be summarized into three steps. Firstly, we match each MSME to one or a group of peers and evaluate its firm-wise PD. Section 4.1 recalls how the PD is evaluated using Merton's model, and Section 4.2 describes the peers-to-firm matching procedure, consisting of a low dimensional representation of the 22 variables space and its subsequent clustering. Secondly, we predict corporate default by calibrating different classification models, both using financial ratios as predictors (baseline) and including the PD (extended). Section 4.3 shows the calibration of the models and the differences in models' performance between the baseline and extended cases. Then, we investigate which predictor most strongly contributes to predicting corporate default using feature importance techniques. Section 4.4 reports the estimation of the contribution of each variable to the predicted class (default or non-default) for both the baseline and extended cases. Lastly, a set of robustness tests is performed to further confirm the stability of the results.

4.1. Estimation of the Merton model

We estimate the Merton model (Merton, 1974) of corporate default risk for our sample of MSMEs. According to the Merton model, corporate default occurs when a company cannot pay off its debts or when the current market value of its assets falls below the market value of its liabilities. For this reason, the market value of the MSME's equity is treated as a call option on the asset

[5] Value Added Tax

[6] The database construction process played a crucial role in making such an empirical analysis possible, despite being time-consuming due to the required manual input of proprietary micro-level data, integrated adequately with additional accounting data collected from Orbis database.

value of the MSME with a strike price equal to the market value of its debt.[7] The MSME asset value process follows a Geometric Brownian motion as shown in Eq. (1) below:

$$dA_t = rA_t dt + \sigma_A A_t dz \tag{1}$$

where A_t is the firm's market value of assets and σ_A is the volatility of assets. r is the one-year maturity risk-free rate of return, which we choose to be the yield of the 1-year maturity domestic government bond with 1-year maturity.[8] Since the market value of equity is treated as a call option, the company's equity value E_t at maturity (which is the end of each yearly period in our model) is priced as shown below:

$$E_t = rA_t\Phi(d_1) - Le^{-rT}\Phi(d_2) \tag{2}$$

where A_t is the firm's assets and L is the firm's liabilities (assumed to be constant for each yearly period). T is the time to maturity which in our model is equal to one year ($T = 1$), r is the risk-free interest rate with one-year maturity (the 1-year government bond rate) and Φ is the cumulative standard normal distribution function. Since default is treated as a European call option, then the values d_1 and d_2 are given by the following formulas:

$$d_1 = \frac{\ln A_0/L + (r + \sigma_A^2/2)T}{\sigma_A\sqrt{T}} \tag{3}$$

$$d_2 = d_1 - \sigma_A\sqrt{T} \tag{4}$$

According to the assumptions of the model, the value of the firm's equity is a function of the value of the firm's assets and time, so it follows from Ito's lemma that:

$$\sigma_E = \frac{A}{E}\left(\frac{dE}{dA}\right)\sigma_A \tag{5}$$

where σ_A is the volatility of assets and σ_E the volatility of firms' equity value. Solving Eqs. (3) to (5) allows to evaluate A and σ_A which are the inputs for the calculation of the Distance to Default (DD) measure, given in Eq. (6):

$$DD = \frac{\ln A_0 + (r + \sigma_A^2/2)T - \ln L}{\sigma_A\sqrt{T}} \tag{6}$$

The resulting Probability of Default (PD) is given in Eq. (7) below:

$$PD = \Phi(-DD) \tag{7}$$

where DD is the Distance to Default measure given in Eq. (6).

4.2. Matching unlisted firms with peers

Since no market data is available for our sample of unlisted MSMEs, we proxy the market volatility of the assets of unlisted MSMEs with that of their comparable publicly-listed companies. As for the latter, the market value of assets is computed as the daily product of their share price multiplied by the number of shares outstanding. Our implicit assumption made for the estimation of Merton's Probability of Default (PD) and Distance to Default (DD) is that those MSMEs which operate in the same industry sectors and have similar balance sheet behaviour with our Italian peers share the same risk profile and belong to the same (market) risk class of the latter.[9] To render the matching procedure as accurate as possible, we opt for a novel clustering approach: the original input ratios are mapped into a lower dimensional space on which clustering techniques can be applied more reliably and robustly. Then, we find the optimal number of clusters in the MSME dataset and assign each peer to the most similar cluster by minimizing the average distance from all firms in the cluster.

Given that the high number of variables of the MSME dataset can affect the clustering algorithm, we apply several dimensionality reduction techniques to obtain a condensed representation of the original data, hereinafter referred to as "embedding". The main idea is to find a function $f : \mathbb{R}^p \mapsto \mathbb{R}^k$, with $k \ll p$, that can project the original high-dimensional data $X \in \mathbb{R}^p$, e.g., $p = 22$ for MSMEs, into a low-dimensional one $E = f(X) \in \mathbb{R}^k$ with the objective of preserving the mutual distance between points, both locally and globally (Gracia et al., 2014). The embedding E has the same number of observations of X but fewer variables. The inverse of f, $f^{-1} : \mathbb{R}^k \mapsto \mathbb{R}^p$ can be used to project back the embedding E to get the reconstruction $\hat{X}$ of the original X and the reconstruction error (RE) can be defined as follows:

$$RE = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2$$

[7] For modelling issues, it is assumed that the market value of debt (or liabilities) is equal to the book value (or accounting value) of total liabilities of the MSME. Moreover, the market value of debt (liabilities) is assumed to remain constant during each yearly period.

[8] We obtain yearly time series data for the 1-year domestic government bond yield for the period covering 2009 to 2014. The yearly Italian government bond yield data are downloaded from Thomson Reuters.

[9] Our assumption on the (market) risk classes goes back to the (Modigliani and Miller, 1958) risk class assumption according to which firms with similar characteristics and balance sheet data belong to the same 'risk class'.

where N is the total number of observations, and x_{ij} and $\hat{x}_{ij}$ are the ith row and jth column elements of X and $\hat{X}$, respectively. The reconstruction error decreases as k increases and can be used to find the optimal value of k, as a trade-off of keeping both RE and k small enough. The use of reconstruction error is widely accepted in the literature as a proxy for information preservation (Bengio et al., 2013; Hinton and Salakhutdinov, 2006). A low reconstruction error indicates that the model has effectively captured the key patterns in the input data, thereby supporting its use for downstream tasks such as clustering.

To further validate that the local geometry and neighbourhood structure of the original data were maintained, we compute Trustworthiness and Continuity scores (Kaski et al., 2003; Venna and Kaski, 2006) for each embedding. Trustworthiness $\mathcal{T}$ quantifies the extent to which points are neighbours in the low-dimensional embedding were also neighbours in the original high-dimensional space. It is defined as:

$$\mathcal{T}(n_b) = 1 - \frac{2}{N n_b (2N - 3n_b - 1)} \sum_{i=1}^{N} \sum_{j \in U_{n_b}(i)} (r(i,j) - n_b)$$

where n_b is the number of nearest neighbours considered, $r(i,j)$ is the rank of point j in the list of neighbours of i in the original space X, $U_{n_b}(i)$ is the set of points that are among the n_b-nearest neighbours of i in the embedding space E, but not in X.

Continuity C measures the extent to which the local structure of the embedding space reflects the local structure of the original space. Specifically, it penalizes cases where points that were neighbours in the original space are not neighbours in the embedding space. It is defined as

$$C(n_b) = 1 - \frac{2}{N n_b (2N - 3n_b - 1)} \sum_{i=1}^{N} \sum_{j \in V_{n_b}(i)} (\hat{r}(i,j) - n_b)$$

where $\hat{r}(i,j)$ is the rank of point j in the list of neighbours of i in the embedded space E and $V_{n_b}(i)$ is the set of points that are among the n_b-nearest neighbours of i in X, but not in E. Both metrics range from 0 to 1, with values closer to 1 indicating better preservation.

Given the panel structure of our MSME data, we evaluate the embedding both at the firm–year level, getting different low-dimensional coordinates for each firm–year pair, and at the firm level, getting a single shared low-dimensional coordinate for each firm–year pair. The former approach evaluates a time-variant embedding, whereas the latter estimates an "average" embedding on the trend of each firm over the years. Thus, in the former, we have $p = 22$ input variables, in the latter we have $p = 22 \times 3$ variables, as we reshape the dataset to have variables-year pairs as new input variables. We tested three different dimensionality reduction techniques on both types of dataset: Robust Principal Component Analysis (RobPCA) (Candes et al., 2009), Auto-Encoder with Multilayer Perceptron (AE) (Kramer, 1991) and Auto-Encoder with Long-Short Term Memory (AE-LSTM) (Cho et al., 2014). RobPCA builds the embedding by creating a linear combination of the original variables, where each combination is the new coordinate. AE is a particular architecture of Artificial Neural Networks that mixes linear combinations of variables with their non-linear transformations to overcome the linearity hypothesis of RobPCA. AE-LSTM is an extension of AE that incorporates auto-regressive terms to account for the time dependence of variables. In this way, AE-LSTM treats each batch of observations of the same firm over the years as a single input. Each neural network has a set of hyper-parameters that must be defined before the calibration, e.g., the number of layers and neurons in each layer. We find the optimal value of hyper-parameters by means of Bayesian Optimization with a 5-fold Cross-Validation[10] performance estimation.

Being the embeddings evaluated, we use them to perform the clustering of the data testing the k-means (MacQueen, 1967) and Gaussian mixture model (Day, 1969) with both the Euclidean distance and Cosine Similarity. We test different numbers of clusters and we find the optimal value based on the Davies–Bouldin index (Davies and Bouldin, 1979) and Silhouette coefficient (Rousseeuw, 1987). The lower the former and the closer to 1 the latter, the better the clustering. After selecting the best clustering, we apply the embedding function f on the peers' dataset to have the same low-dimensional representation and to assign each peer to the closest cluster. The closeness is intended to be the minimum average Euclidean distance from all observations within the cluster.

Although the optimal low dimension ($k = 6$) of the embedding resulted in a better performance for the clustering, data cannot be visualized. Therefore, we use another dimensionality reduction technique that is better suited for visualization rather than for clustering. Thus, by applying the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018; McInnes et al., 2018), we can visualize the clusters in a 3-dimensional space.

As the cluster evaluated in the embedding space, we explore the effects of the grouping on the original 22 variables, allowing us to provide an explanation for the segmentation of the firms. In particular, we evaluate the differences in distribution for each original variable, using the ANOVA test while controlling for the group defined by the cluster labels. For every variable, we perform the one-way ANOVA test followed by the post-hoc Dunn test, controlling the family-wise error rate (FWER) as suggested in Holm (1979). Then, we count how many times each cluster significantly differs from the others, identifying which variable accounts for the most overall difference and which cluster shows the highest deviation from the others.

Finally, we provide each MSME observation with its respective PD. As described in Section 4.1, PD can be calculated using Eq. (7) after evaluating DD with Eq. (6), with the total assets A, total liabilities L, and assets volatility σ_A. We evaluate the PD with two approaches. In the first (named *average-PD*), we evaluate the average $\bar{A}_{ct}$, $\bar{L}_{ct}$ and $\bar{\sigma_{A}}_{ct}$ over all peers in the same cluster $c = 1 \dots C$, where C is the optimal number of clusters, for each year t and use them to evaluate the average DD. So, we have $C \times 3$ different DD

[10] In the k-fold Cross-Validation, k models are calibrated on $k - 1$ folds and the performances on the kth fold are then averaged.

values, one for each year-cluster pair. The DD is then matched with each MSME observation by year-cluster. In the second approach (named *pointwise-PD*), we use the ith MSME firm's A_{it} and L_{it} at time t and the average year-cluster peers' $\bar{\sigma}_{A_{ct}}$ to have a firm–year level DD.

4.3. Prediction of default

After assigning the PD to all our unlisted MSMEs, we calibrate five different models to predict the binary target: 1 for defaulted firms and 0 otherwise. Each model is calibrated using the set of 22 variables (*baseline*) and with the addition of the PD (*extended*). First, we inspect the distribution of each input variable with respect to the target variable. Fig. A.3 in the Appendix shows similar behaviour of the input variables for both subsets of defaulted and non-defaulted firms, meaning that the overall relation between each predictor and the target is weak because there is no clear polarization in the distributions. Thus, we expect low prediction performances when using classical linear models because they estimate coefficients that should discriminate between the 0s and the 1s, considering the average of the distribution of input variables. Moreover, the true relationship between input and target variables may be non-linear. Therefore, we opt for a non-linear and piecewise model, the Multivariate Adaptive Regression Spline (MARS) (Friedman et al., 2009), that estimates multiple polynomial relationships in different partition intervals of each input variable. Therefore, the model can be seen as an ensemble of sub-models estimated in each combination of partitions into which the input variables can be divided. For example, suppose we split the input domain into quartiles of each variable. In that case, MARS estimates a polynomial function for observations whose input variables are in the lowest quartile of the corresponding distributions, and so on, for all possible variable–quartile combinations. As MARS is a parametric algorithm, meaning that we have to define a structure of each estimation function, e.g., polynomial.

To further investigate the non-linear relationships in the data, we also employ three additional non-parametric algorithms: Support Vector Machine with a Radial Basis Function kernel (SVM-RBF), k-Nearest Neighbours (k-NN), and Random Forest (RF). SVM (Cortes and Vapnik, 1995) constructs a hyperplane or set of hyperplanes in high-dimensional space to separate classes. The RBF kernel allows SVM to capture complex, non-linear relationships by mapping the input data into a higher-dimensional space where classes may become linearly separable. This makes SVM particularly effective when the decision boundary is highly non-linear, as it can adapt to intricate patterns in the data without requiring explicit feature engineering. k-NN (Cover and Hart, 1967) is an instance-based learning algorithm that classifies a data point based on the majority class among its k nearest neighbours in the feature space. Unlike parametric models, k-NN does not make assumptions about the underlying data distribution, making it highly flexible for capturing local patterns and non-linear relationships. Random Forest (Breiman, 2001) is an ensemble of decision trees that partition the input domain with nested binary splitting to maximize the discrimination of all target values. Each branch of the tree contains a set of hierarchical rules, e.g., values of a certain variable greater or less than a fixed threshold, so that (possibly) all observations satisfying each chain of rules have the same target value, i.e., 0 or 1. The estimation function of RF is then a combination of rules that can approximate non-linear relationships between input and target variables. Nonetheless, we use a regularized linear model, i.e., Elastic-Net, as a benchmark.

As noted in Section 3, the presence of a few variables with moderate correlation will not affect the models' performances because the ensemble nature of MARS and RF and the regularization feature of Elastic-Net are suitable to deal with multicollinearity. SVM is less sensitive to multicollinearity because it focuses on maximizing the margin between classes rather than relying on the coefficients of individual features. The RBF kernel further mitigates the impact of multicollinearity by mapping the data into a higher-dimensional space where features may become less correlated. Instead, the k-NN algorithm does not rely on feature coefficients, making it inherently robust to multicollinearity, as it uses distance metrics to classify data points based on their neighbours, which reduces the influence of correlated variables. Each model has a set of hyper-parameters that must be defined before the calibration. For example, MARS requires the maximum degree of polynomials to be fitted, RF requires the number of decision trees to be estimated. We find the optimal value of hyper-parameters by means of Bayesian Optimization with a 5-fold Stratified Cross-Validation[11] performance estimation. For the Bayesian Optimization, we proxy the objective function with a Random Forest surrogate model, and for the acquisition function, used to balance exploitation and exploration, we use the Upper Confidence Bound (UCB), combining the mean and uncertainty of the surrogate model predictions. Then, for the initial space-filling, we use 5 points for MARS and 30 points for the remaining models. The maximum number of iterations is 6 for MARS and 40 for the remaining models.

Given the imbalanced nature of the data (1s are 7% of all samples), as described in Section 3, we use the F1-score as a class-specific performance metric to highlight the importance of predicting the rarest 1-labelled targets, the Matthew's Correlation Coefficient (MCC) as a balanced measure of classification performance and the Area Under the Precision–Recall Curve (PRAUC as an overall performance metric. In particular, F1-score is the harmonic mean of precision and recall and emphasizes the correct prediction of the minority class (1s in this case). The Area Under the Precision–Recall Curve provides a robust evaluation of model performance across different probability thresholds, especially for imbalanced datasets where the ROC-AUC can be misleading. Matthew's Correlation Coefficient is a single metric that considers true positives, true negatives, false positives, and false negatives, providing a balanced classification performance evaluation. It ranges from −1 (perfect inverse prediction) to +1 (perfect prediction), with 0 indicating a random classifier. MCC is particularly effective for imbalanced datasets because it is less influenced by class distribution and provides a more comprehensive assessment of model performance. Moreover, each model has been calibrated with

[11] Stratification is performed with respect to both target variable and control variables when included.

the additional constraint of weights for each observation, i.e., penalizing the prediction error on 1s more than the error on 0s. Both F1-score and weighting help the calibration procedure to prevent overfitting to a certain extent, allowing the model to have a good generalization power.[12]

We also test both undersampling and oversampling techniques, testing different percentages of balancing the minority class, i.e., 30% and 50%. In particular, we tested the Adaptive Synthetic Sampling (ADASYN) oversampling algorithm, as it can handle class imbalance effectively and usually improves classifier performance by focusing on difficult classes (He et al., 2008). Results from all models showed only a negligible improvement in the performance (no more than 1.5%), proving that the weighted calibration can account for the class imbalance properly. For the undersampling technique, we used stratified sampling (based on Region, Firm Size and Industry) to test the aforementioned level of balancing. Five different repetitions have also been used to account for randomness in the sampling procedure. Results showed similar performance for all models, with up to 2% improvement. The slight improvement compared with the oversampling technique is still due to the synthetic nature of the sample generated by the ADASYN algorithm.

Furthermore, we include control variables in both *baseline* and *extended* cases to assess the models' robustness to time-invariant (static) and time-varying (dynamic) characteristics of the observations. In particular, we test the static Dummy Industry, Firm Type, Industrial Sector and Region described in Table 2 and the dynamic Firm Size and Funding Risk from the Italian National Institute of Statistics[13] (ISTAT) and Bank of Italy, which evaluates the likelihood of a firm facing difficulties in obtaining or repaying financing. It reflects the firm's financial health, creditworthiness, and ability to meet its debt obligations. It is a regional and annual indicator.

Finally, we investigate the persistence of target values over time, i.e., we examine the impact of clients who changed their outcome over the years, both from defaulted to recovered and vice-versa. Table A.3 in the Appendix reports the number of clients that changed over time. To assess the impact of this phenomenon, we compare the distribution of the input variables subject to clients' behaviour, and we calibrate the models both on the entire dataset and on the dataset where we remove the clients that changed the outcome over the years. We find that models' performances are not affected by the inclusion of target-switching clients, resulting in the robustness of our results to this phenomenon. Fig. A.4 in the Appendix shows the distribution of relative changes over the years of each input variable split by clients' behaviour, i.e., clients that are persistent over time and clients that do not exhibit such a behaviour.

4.4. Importance of variables

We explore which input variable contributes the most to each model prediction, focusing on the changes when the PD is added. For this reason, we evaluate the predictive power of the variables using two state-of-the-art techniques for feature importance: Permutation Feature Importance (PFI) and Shapley Additive Explanations (SHAP). PFI evaluates the importance of the jth variable by comparing the performance, e.g., F1-score, of the model that predicts the observations used for the calibration against the performance of the model that predicts the same observations where the values of the jth column are shuffled (Fisher et al., 2018). In this way, the correlation between the jth variable and all the others is broken, thus removing the influence of that variable on the model predictions. If the change in performance is negligible, the jth variable is unimportant for the model. SHAP is based on Shapley values, a method from coalitional game theory which provides a way to fairly distribute the payout among the players by computing the average marginal contribution of each player across all possible coalitions (Shapley, 1953; Osborne and Rubinstein, 1994). SHAP, proposed by Lundberg et al. (2020), uses Shapley values to evaluate the difference in the predicted value of a single observation, by comparing the prediction of all possible combinations of variables that include the jth variable against the ones that do not. The differences are then averaged, and the positive or negative change in the prediction is used for variable importance. For example, if the model predicts the probability of default, SHAP evaluates, for a single observation, which variable contributed most in increasing or decreasing the final probability. In this way, exploiting the additive property of Shapley values, it is possible to estimate the impact of all variables on the final predicted value for every single observation. PFI provides a global measure of importance, measuring the impact of all observations together. Moreover, it measures the changes in global performance. On the other hand, SHAP provides a local measure of importance, measuring the impact of variables for every observation. However, taking the average of the absolute values of each observation's SHAP, it is still possible to get a global measure of the average importance of the variables. Instead, taking the average of the Shapley values rather than their absolute value provides an average effect of each variable on the predictions. Both techniques are described in detail in Appendix A.

5. Results

5.1. Matching unlisted firms with peers

As described in Section 4.2, we first find the embedding that minimizes the Reconstruction Error. Table 3 reports the optimal embedding dimension k, the reconstruction error of the different algorithms, the R^2 and the Trustworthiness $\mathcal{T}$ and Continuity $\mathcal{C}$ metrics. In our context, in analogy with the classical R^2, we compute the RSS term as the Reconstruction Error given by the

[12] This means that the model has similar performances on both data used for calibration and unseen observations.

[13] https://www.istat.it/sistema-informativo-6/banca-dati-territoriale-per-le-politiche-di-sviluppo/

Table 3
Results of dimensionality reduction. Reconstruction Error and its proportion with the average absolute value of the input data is reported for all methods as well as R^2. In our context, in analogy with the classical R^2, we compute the RSS term as the Reconstruction Error given by the embedding and the TSS term as the total variance contained in the original data. Trustworthiness and Continuity metrics are also reported. Values close to 1 mean good preservation of the information. Values in parentheses are the optimal number of neighbours used to evaluate the metric.

Input level	Rows	Columns	Method	Input dimension	Embedding dimension	Reconstruction error (% of Avg Abs Input)	R^2	$\mathcal{T}(n_b)$ (n_b)	$C(n_b)$ (n_b)
Firm–year	Firm–year pairs	Variables	AE	19,743 × 22	19,743 x 6	**0.1418 (20%)**	**98%**	**0.96 (6)**	**0.95 (7)**
			RobPCA	19,743 × 22	19,743 × 9	0.2033 (30.6%)	95.70%	0.9 (5)	0.88 (5)
Firm (batch of years)	Firms	Variables	AE-LSTM	10,136 × 22	10,136 × 10	0.2138 (31.8%)	94.60%	0.89 (6)	0.85 (5)
Firm	Firms	Variables-year pairs	AE	10,136 × 66	10,136 × 32	0.2391 (35.9%)	91.30%	0.72 (12)	0.69 (11)
			RobPCA	10,136 × 66	10,136 × 15	0.3857 (58%)	84.80%	0.51 (8)	0.88 (9)

embedding and the TSS term as the total variance contained in the original data which represents a proxy for how much intrinsic information within the data is preserved in the transformation. The optimal number of neighbours for $\mathcal{T}$ and C is selected via bootstrap, selecting the value that leads to the smallest variability in the metric. The embedding resulting from AE with the firm–year level approach performed best, showing the lowest reconstruction error, the highest R^2 and both $\mathcal{T}$ and C close to 1. Methods evaluated with a firm-level approach performed worse and will not be included in the following analysis.

We tune the hyper-parameters of each neural network by means of Cross-Validation, and we calibrate the models with the optimal parameters on the entire dataset to have a single model to be used for the evaluation of the embeddings. Here, optimality is intended as a trade-off between the minimized reconstruction error and both model complexity, i.e., the dimension of the embedding space, and information loss, i.e., the Trustworthiness and Continuity scores. We generate elbow graphs for both the information loss metrics and the reconstruction error, plotting these metrics against the dimensionality of the embedding. By examining the elbow points across all metrics, we identify a consistent dimensionality that minimizes information loss while avoiding unnecessary complexity.

For the AE model we tune the layers' structure l_S (both the number of layers and neurons), the size of the bottleneck layer l_B, the activation functions act_S and act_B used in l_S and l_B, respectively, the number of epochs n_E and batch size s used during the training. For the AE-LSTM model we tune the recurrent blocks' structure l_S (both number of layers and number of neurons), the size of the bottleneck layer l_B, the type of recurrent unit $type_{\text{rec}}$ used in all recurrent blocks, the α share of L^1 and L^2 regularization for the weights in each block, the number of epochs n_E and batch size s used during the training. $type_{\text{rec}}$ can be $LSTM$ for Long-Short Term Memory or GRU for Gated Recurrent Unit. Table C.4 in the Appendix reports the best parameters from the tuning of each model on the MSME dataset with the 22 ratios.

Then, we look for the optimal number C of clusters. Table 4 reports the performance of the clustering on each low-dimensional embedding as well as the comparison with the clusters found in the original high-dimensional data. We select $C = 5$ clusters identified in the AE embedding. Regardless of the method, clusters evaluated on the embeddings always show better clustering performances than their counterparts evaluated on the original set of variables. Moreover, we apply the UMAP algorithm to visualize the clusters in a 3-dimensional space. Fig. 1 depicts the five clusters for all observations (small points) as well as the matched peers (bold spheres), showing a good separation, even if there is small overlapping between the yellow and green cluster and few blue peers are mapped close to the red ones. We recall that the embedding function f is estimated only on the MSMEs dataset and then the peers' embedding is evaluated by applying f, as an out-of-sample set. UMAP, instead, is calibrated on MSMEs and peers datasets as a whole and is not directly influenced by the estimated embeddings and clusters: it only represents an "optimal" visualization of the high-dimensional data in the three-dimensional space. Here optimal is intended as the best way to exacerbate the distance between points that are dissimilar and to reduce the gap between points with similar features. Therefore, from Fig. 1 we can conclude that the good clustering performance found on low-dimensional embeddings is corroborated by the visual representation of UMAP technique. On the contrary, Fig. B.5 in the Appendix shows the UMAP projection of original high-dimensional data, clustered with the optimal number of clusters found for the embedding, i.e., five, where the clusters are clearly overlapping, and most of the peers are misplaced. It is worth pointing out that the figure does not show the actual distribution of the data, which actually lies in a 22-dimensional space, and the clustering has been performed in the 6-dimensional space of the AE embedding. Therefore, the distances between the plotted points cannot be used in a proper clustering algorithm. Furthermore, we compare the distribution of the original 22 variables of MSMEs with the average of the assigned peers in each cluster. Figs. B.6 and B.7 in the Appendix show that the average value of the peers is within the inter-quantile range of the MSMEs for the majority of the variables and clusters, both in the original 22-dimensional space and in the 6-dimensional embedding space. Finally, the clusters are used to assign the PD to each MSME firm, matching year-cluster pairs for asset volatility.

We provide an economic interpretation of the matching procedure by exploring the resulting segmentation in the original variable space. As described in Section 4.2, we perform one-way ANOVA for each original variable, grouped according to the evaluated clusters, and record the number of significant differences in distribution for every cluster compared to others. Furthermore, we

Table 4
Results of clustering. Davies–Bouldin index and Silhouette coefficient are reported for clusters evaluated on both embedding and original data, as well as the optimal clustering technique and distance metric, in parentheses. Davies–Bouldin is a positive number, the smaller the better the separation between the clusters. Silhouette coefficient is bounded between −1 and 1, where −1 means overlapping of clusters and 1 perfect separation. Only the top two results for each method are reported.

Method	Clusters	Dimension		Algorithm (Distance)		Davies–Bouldin		Silhouette	
		Original	Embedding	Original	Embedding	Original	Embedding	Original	Embedding
AE	**5**	22	6	k-Means (Euclidean)	GMM (Euclidean)	0.36	**0.08**	0.45	**0.91**
	4					0.43	0.11	0.37	0.59
RobPCA	4	22	9	k-Means (Cos. Simil.)	GMM (Cos. Simil.)	0.53	0.13	0.41	0.72
	3					0.71	0.2	0.32	0.44
AE-LSTM	3	22	10	k-Means (Euclidean)	GMM (Euclidean)	0.8	0.2	0.07	0.21
	2					0.93	0.32	−0.13	0.06

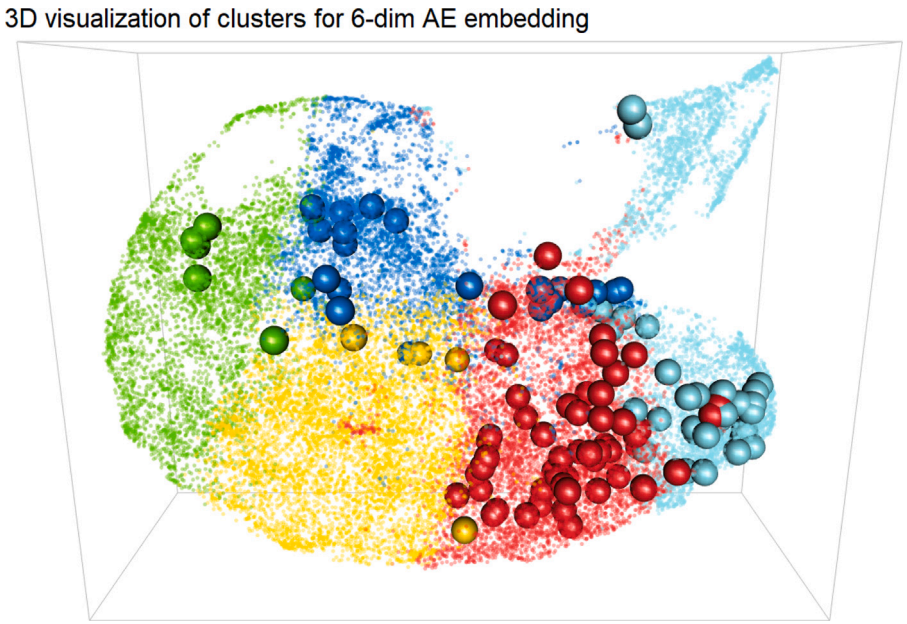


Fig. 1. 3D visualization of five clusters for the 6-dimensional AE embedding. Visual embedding is evaluated with the UMAP algorithm. Small points are MSMEs observations, bold spheres are peers' observations.

utilize box-plot comparisons (see Fig. B.8 in the Appendix) to visualize how each cluster is differentiated from the others across each variable, providing a clearer economic interpretation.

The financial symptoms of each group and their corresponding clusters are outlined below. To enhance understanding of the management implications of this study, we analyse the findings from each cluster in relation to the primary causes of failure identified in the literature, along with possible interventions to mitigate the likelihood of default.

Cluster 1: This group is predominantly characterized by a high level of bank debt (indicated by a high Debt Burden Index and high Financial Interest on Revenues) and modest economic performance (high Fixed Asset Coverage, low Net Worth on Equity, low Return on Assets). The observed poor asset management and underutilization of resources indicate financial fragility—one of the most common precursors to bankruptcy (Kucher et al., 2020). Consequently, effective cost management emerges as a key strategy for mitigating short-term bankruptcy risk, while long-term strategies should focus on optimizing asset utilization and enhancing operational efficiency.

Cluster 2: The second group exhibits high Payables to Sales alongside moderate debt dependence. Enhanced cost management practices could foster improvements in profitability, thereby increasing key profitability ratios. Such advancements may ultimately yield greater value creation for shareholders.

Cluster 3: This cluster encompasses firms with significant debt reliance and low capitalization levels. Insufficient equity is frequently identified as a critical internal factor leading to bankruptcy, as it restricts operational sustainability and complicates

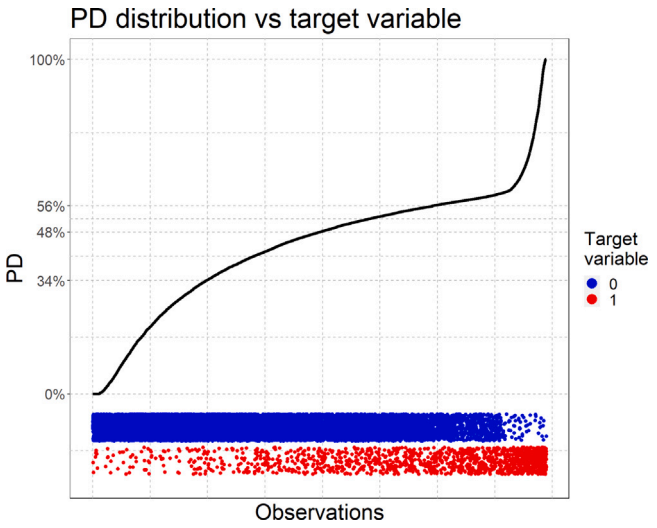


Fig. 2. Distribution of PDs compared with the corresponding target values. y-axis reports quartiles of PD values.

access to alternative financing sources (Carter and Auken, 2006; Ooghe and De Prijcker, 2008; Kucher et al., 2020; Mayr et al., 2021). To mitigate financial risk, a reevaluation of debt strategies is essential.

Cluster 4: Firms in this cluster demonstrate strong economic performance and balanced equity and financial structures. Consequently, no specific symptoms indicative of potential default are apparent. The primary implication for these firms is the opportunity to reinvest excess liquidity into growth initiatives to sustain positive long-term performance.

Cluster 5: This final group comprises firms characterized by high Payables to Assets, moderate Operating Margins, and balanced cash flows. These companies can be classified as slow movers, illustrating that even with a sound financial position, their economic performance remains lackluster (Kucher et al., 2020; Mayr et al., 2021). The potential for stagnation due to conservative financial strategies could erode their competitive advantage over time.

Each cluster exhibits unique characteristics that allow for the identification of specific financial and economic symptoms. Understanding these relationships enables the development of targeted strategies to address vulnerabilities, ultimately improving the overall financial health and stability of the companies involved.

5.2. Prediction of default

Being the PD assigned, we calibrate the prediction models. The following results refer to the PDs evaluated with the *pointwise-PD* approach described in Section 4.2 because it performed better than the *average-PD* one, although the findings described below still remain robust. Fig. 2 shows the distribution of PDs compared with the corresponding target values. PD seems to be a reliable indicator of the outcome of the target variable.

We tune the parameters of each model with the Stratified Cross-Validation, and we calibrate the models with the optimal[14] parameters on the entire dataset, to have a single model[15] to be used for feature importance evaluation. In particular, for the Elastic-Net model we tune the α parameter that represents the share of L^1 and L^2 regularization, for the MARS model we tune the degree d of the polynomial functions; for SVM-RBF model we tune the regularization cost C and the scale of RBF kernel σ; for the k-NN model we tune the number of neighbours k; and for the Random Forest model we tune the number of trees n_{tree}, the number n_{var} of variables randomly sampled as candidates at each split and the minimum size s of observation in each node. Table C.5 in the Appendix reports the best parameters from the tuning of each model on both the dataset with and without Merton's PD. In Table 5 we report the performance on the entire dataset and the average performance on validation folds for each model. We also compare the models trained with the 22 ratios only and the ones with the addition of PD. Random Forest is the top-performing model with good performances, followed by the k-NN model, both capturing the different local separations of the data, as discussed in Section 4.3. Nevertheless, all models show an improvement in class-specific performance, i.e., F1-score for the defaulted class, and on the PRAUC when the PD is included as a predictor. The statistical significance of the relative differences in performances $\Delta\%$ between the *baseline* and *extended* setting is evaluated using a permutation test (Efron and Tibshirani, 1994), performed by randomly shuffling the predicted classes between the two models and recalculating the performance metric for each permutation. This process is repeated 5,000 times to create a null distribution of the performance metric under the assumption that there is no significant difference between the models. The p-value is then calculated as the proportion of permutations where the observed

[14] Optimal parameters are selected according to the maximum cross-validation F1 score.

[15] In the k-fold Cross-Validation, k models are calibrated on $k-1$ folds and the performances on the kth fold are then averaged.

Table 5
F1-score, Precision–Recall AUC and Matthews Correlation Coefficient for all considered models calibrated on a dataset with input variables only and with the addition of PD. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. *Δ%* is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)			PRAUC (Cross-Val)			MCC (Cross-Val)			PD Feature Imp. Rank (Relative %)	
	Baseline	With PD	Δ%	Baseline	With PD	Δ%	Baseline	With PD	Δ%	PFI	SHAP
Elastic-net	30.7%	31%	1%**	21.7%	21.9%	1%***	25.6%	25.9%	1%*	1	1
	(29.8 ± 1.7%)	(30.3 ± 1.1%)		(20.8 ± 0.8%)	(20.8 ± 1.4%)		(24.5 ± 1%)	(25 ± 1%)		(60.7%)	(38.1%)
MARS	43.6%	44.1%	1%*	40.7%	41.1%	1%*	37.7%	38%	1%	1	1
	(42 ± 1.2%)	(42.1 ± 1.5%)		(39.9 ± 1%)	(39.5 ± 0.9%)		(35.8 ± 1.5%)	(36.6 ± 1.1%)		(38.5%)	(32%)
SVM-RBF	51.6%	53.1%	3%*	36.6%	37.7%	3%***	45.8%	47.2%	3%**	1	2
	(49.7 ± 1.3%)	(51.8 ± 1.3%)		(35.7 ± 1.7%)	(36.6 ± 0.7%)		(43.6 ± 0.8%)	(44.4 ± 1.3%)		(20.1%)	(16.6%)
k-NN	70.5%	72.6%	3%**	74%	76.2%	3%***	66.9%	69%	3%**	1	2
	(67.5 ± 1.2%)	(70.8 ± 1.4%)		(70.6 ± 1.6%)	(73.5 ± 1.3%)		(64.7 ± 0.8%)	(65.8 ± 1.2%)		(19.2%)	(12.2%)
Random	88.4%	95%	7.5%*	93.3%	96.3%	3.2%***	87%	93.5%	7.5%*	1	1
Forest	(84.1 ± 0.9%)	(91.9 ± 0.8%)		(88.2 ± 0.9%)	(91.1 ± 1.2%)		(82.9 ± 1.3%)	(90.9 ± 1.4%)		(25.9%)	(24.6%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

difference is equal to or more extreme than the actual difference. If the *p*-value is below the significance threshold, the observed difference is considered statistically significant. Tables C.7 and C.8 in the Appendix report the results of the models with Static and Dynamic controls for fixed effects, respectively, showing the stability of performances and the resulting robustness of the models. Fig. C.9 in the Appendix shows Precision–Recall curves of all models with no fixed effects only.

5.3. Importance of variables

We explore the feature importance for all models. PFI and SHAP are evaluated on a model calibrated with input variables and with the addition of PD. Fig. 3 shows the PFI of the Random Forest model, where the changes of the F1-score are normalized to sum up to 100%. PD is the second most important variable, slightly below the financial interest on revenues. Figs. 4(a) and 4(b) show the effect of input variables on the predicted probabilities[16] of the Random Forest model, for each observation predicted as 1 and 0, respectively, by the means of SHAP. The colour of the points ranges from red, meaning that the observation has a low value for the specific variable, to blue, meaning high values for the same variable. The position on the horizontal axis represents the contribution of the variable in increasing or decreasing the predicted probability of each observation. Values in the left column report the average absolute change in predicted probability across all observations, along with normalized values in parentheses. PD is one of the top two most important variables, and we can check the expected impact on the predicted probability: for defaulted observations, high values of PD (blue) result in a major increase in probability, whereas for non-defaulted observations, low values of PD (red) result in a significant decrease of probability. The accounting variables, as well as the PD, exhibit the expected effect on the predicted probability, e.g., lower return on assets (ROA) and working capital turnover increase the predicted probability, whereas lower financial interests decrease the latter. Figs. 5(a) and 5(b) show the average signed effect of input variables on the predicted probabilities for all observations predicted as 1 and 0, respectively. In both cases, PD is one of the top two most important variables, increasing the predicted probability for defaulted observations while reducing the probability for non-defaulted observations. We see that PFI and SHAP agree on the importance of PD, supporting its added value as measured by the increase in model performance. Although both techniques lead to the same conclusion, it is worth noting the complementary contribution to model interpretability: PFI provides a synthetic overall measure of the relative importance of the variables, whereas SHAP offers insights on the magnitude and the direction of the effect of the variables on each observation, similarly to the explanation of linear regression coefficients. Figures from D.10 to D.12 in the Appendix report the PFI and SHAP variable importance for k-NN, leading to similar results, supporting the relevance of the addition of PD as a predictor. We omit plots for Elastic-net, MARS and SVM-RBF because of the poor performance. Nevertheless, similar results still hold.

To further investigate the non-linear relationships and possible overlapping classes in the data we evaluate SHAP Scatter Plots, that are able to highlight the relationship between individual feature values, their corresponding SHAP values, and the true class labels. In these plots, the x-axis represents the actual values of a specific feature, while the y-axis represents the evaluated SHAP values for the feature itself, indicating how changes in the feature influence the model's output. Points are coloured according to the true class labels, allowing for a clear distinction between how the feature affects predictions for each class. Fig. 6 reports the SHAP scatter plots of the most important variable for every model in the *baseline* setting. Although no strong non-linearity emerges from the plots, it is clear how Elastic-net, MARS and SVM-RBF struggle in separating the overlapping classes, whereas k-NN and, in particular, Random Forest are able to better partition the complex structure.

[16] All five classification models predict probabilities in $[0, 1]$. If the probability is above 0.5, the observation is classified as defaulted (1), non-defaulted (0) otherwise.

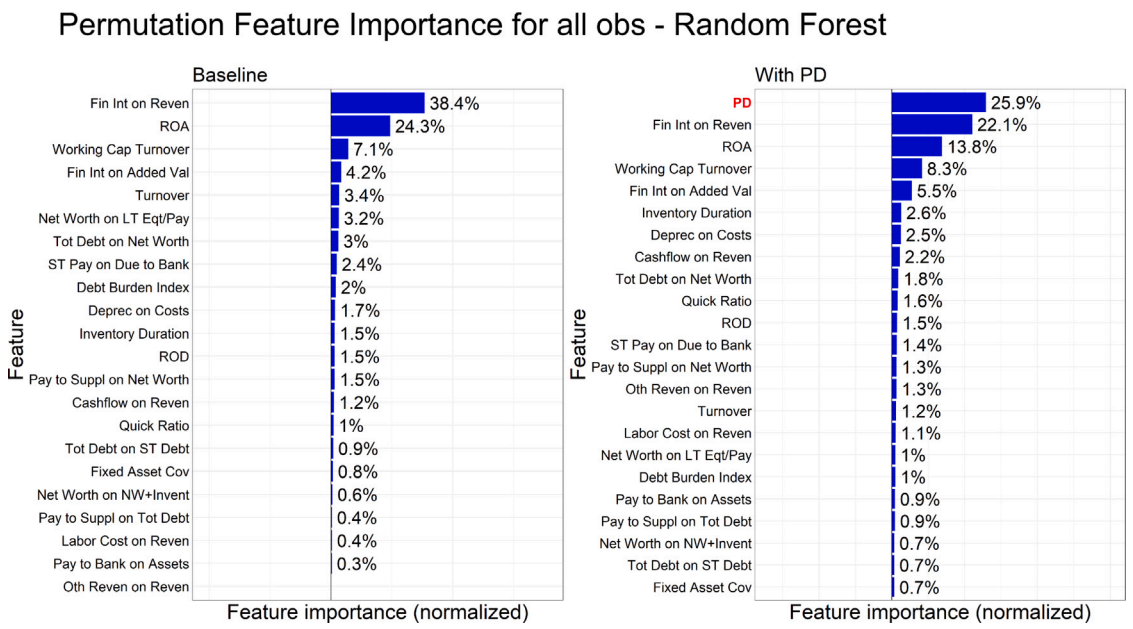


Fig. 3. Permutation Feature Importance for Random Forest model, comparing variable importance of model calibrated with input variables and with the addition of PD. Normalized changes in the F1-score are used to rank the variables.

6. Additional robustness checks

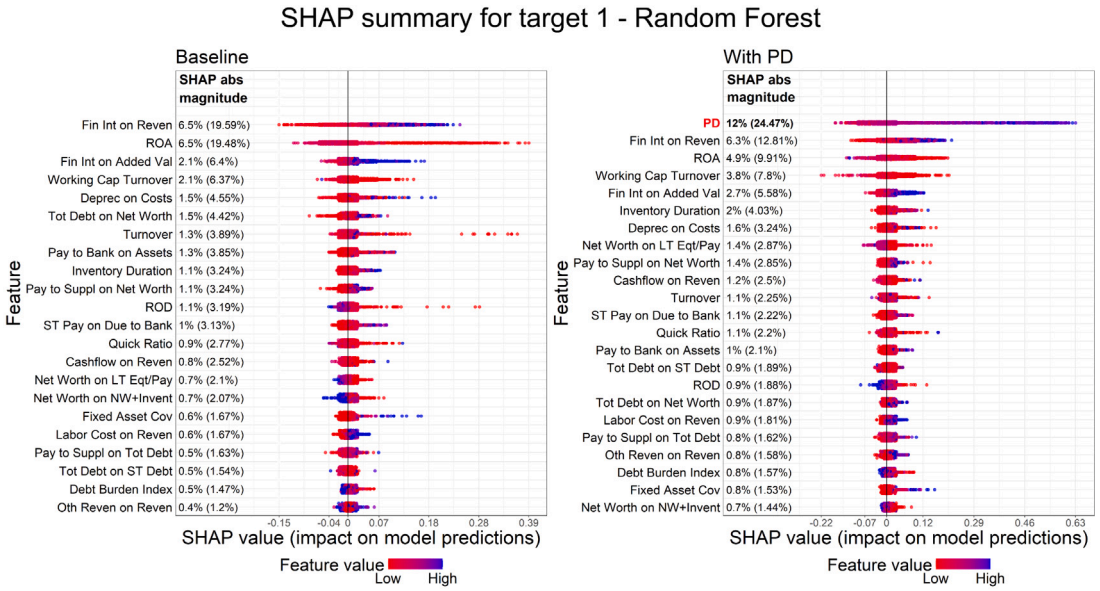
6.1. Information loss in the embedding

There is an inherent risk of information loss when employing dimensionality reduction techniques to generate low-dimensional embeddings. This occurs because the process of compressing high-dimensional data into a lower-dimensional space may discard some of the original information, mainly if the embedding space is not sufficiently expressive or if the reduction process prioritizes certain features over others. However, the goal of dimensionality reduction is not to preserve all information but rather to retain the most relevant and discriminative information for downstream tasks, such as clustering or classification (Hinton and Salakhutdinov, 2006).

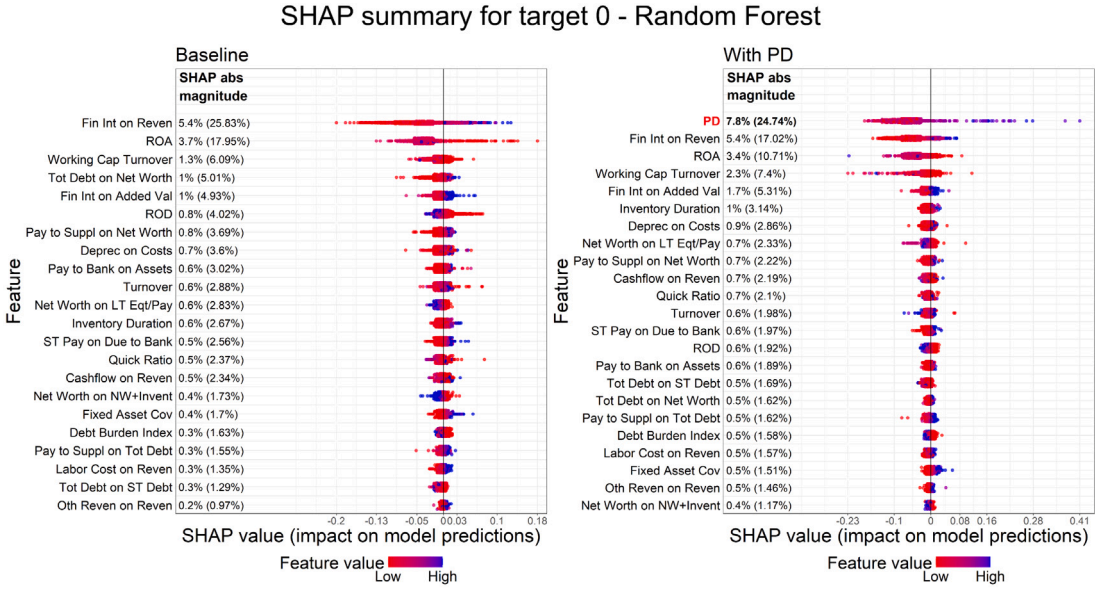
Specifically, we train the same five prediction models to predict the default flag using only the low-dimensional embeddings as inputs. Table C.9 in the Appendix reports the F1-score performance for the five prediction models when using the original 22 variables and the embeddings evaluated with the AE, RobPCA and AE-LSTM techniques. The maximum performance degradation due to the use of low-dimensional representation is around 7% for the selected AE technique for the top two-performing models. This indicates that the AE-derived representation retains most of the relevant information, supporting its suitability for our clustering purposes. The performances of RobPCA and AE-LSTM are worse than those of the AE models, consistently across each prediction model. Moreover, we notice that the two linear models, Elastic-net and MARS show an increase in performance (although they have very low F1-scores) when using the AE embedding, benefiting from the non-linear information compression of the dimensionality technique.

6.2. Placebo test for PD

To ensure the mapped PD reflects firm-specific risk rather than cluster-level noise, we conducted a placebo test by randomly reassigning PDs within clusters, re-training the models and evaluating model performance. Table C.10 in the Appendix compares the F1-score, Precision–Recall AUC and Matthews Correlation Coefficient of the models trained on the described settings. To further control for randomness, we repeated the random assignment five times, using a different starting seed each time. Results show that there is no significant improvement in performance compared to the *baseline* setting. Moreover, the feature importance relevance of the randomized PD confirms that the variable makes no contribution to the prediction performance, both in terms of ranking and relative importance, for all models. This analysis complements our PFI analysis in Section 4.4, which similarly demonstrated the PD's importance post-hoc by evaluating the loss of performance when shuffling the variable. Together, these tests mitigate concerns about unobservable factors or sector-wide signals that may be driving the results.



(a) Defaulted clients.

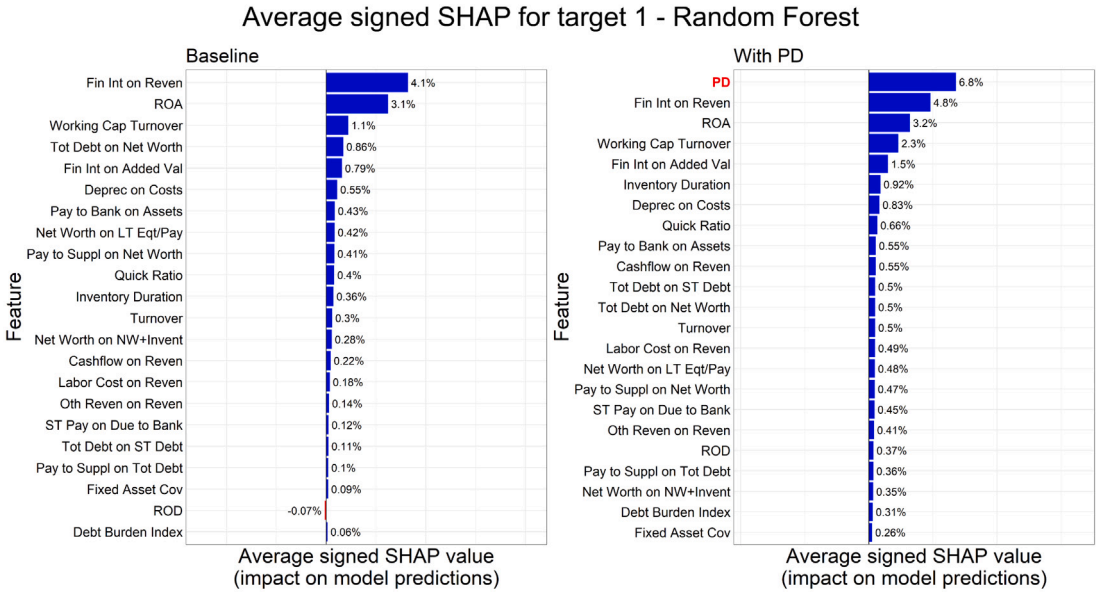


(b) Non-defaulted clients.

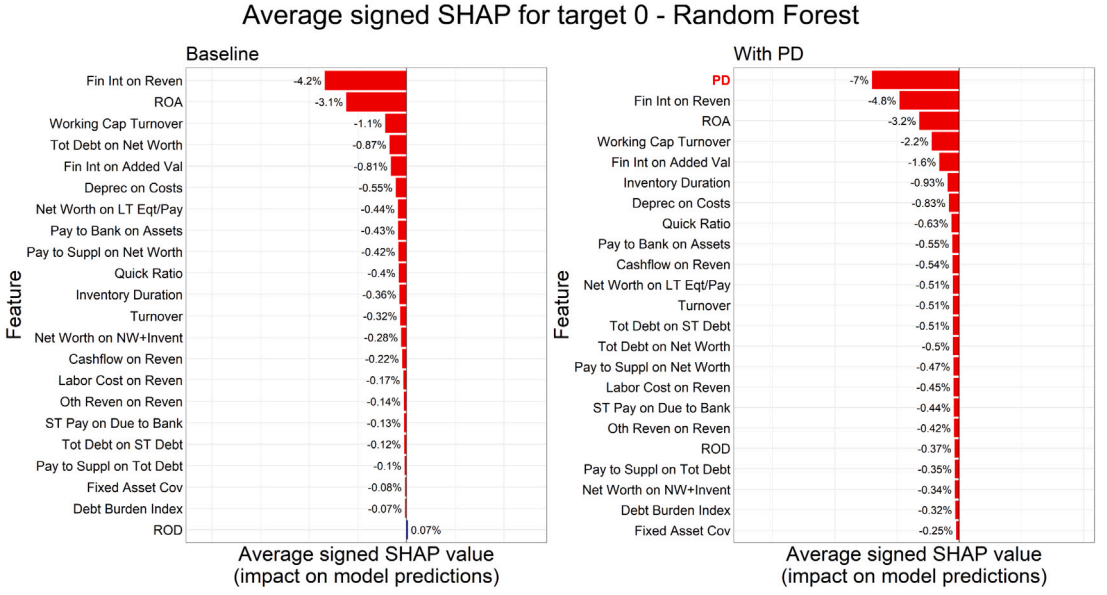
Fig. 4. SHAP effects on predicted probability for Random Forest model and defaulted (top) and non-defaulted (bottom) observations only, comparing variable importance of model calibrated with input variables and with the addition of PD. The colour of the points ranges from red, meaning that the observation has a low value for the specific variable, to blue, meaning high values for the same variable. The position on the horizontal axis represents the contribution of the variable in increasing or decreasing the predicted probability of each observation. Values on the left column report the average absolute change in predicted probability over all observations and the normalized values in parentheses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.3. Stability over time

To address potential overfitting to the 2012–2014 period, we conducted out-of-time validations using both a "rolling window" (train on year t, test on $t + 1$) and a holdout sample (train on 2012–2013, test on 2014). The models are then cross-validated by sampling from the defined year(s). Table C.11 in the Appendix compares the F1-score of the models trained on the two described settings. Results show that both the improvement in performance and the feature importance relevance in the two settings are



(a) Defaulted clients.



(b) Non-defaulted clients.

Fig. 5. SHAP average signed effect for Random Forest model and defaulted (top) and non-defaulted (bottom) observations only, comparing variable importance of model calibrated with input variables and with the addition of PD. Bars report the average effect of input variables on the predicted probabilities for all observations predicted as 1 and 0, respectively.

comparable to those in the *extended* setting in Table 5. Therefore, our findings confirm that performance gains are not period-specific. This stability underscores the model's capacity to generalize across macroeconomic conditions.

6.4. Disentangling PD contribution

To rigorously disentangle whether the Merton PD's predictive power stems from unique firm-specific risk signals or merely correlates with observable characteristics like sector, size, region, etc. as in Tables C.7 and C.8 in the Appendix, we conducted a three-way 'horse race' analysis comparing: (a) a baseline model with only the mapped PD (no controls for fixed effects), (b) our original specification combining PD with fixed effects, and (c) a novel residualized PD model where we first purge the PD of its linear

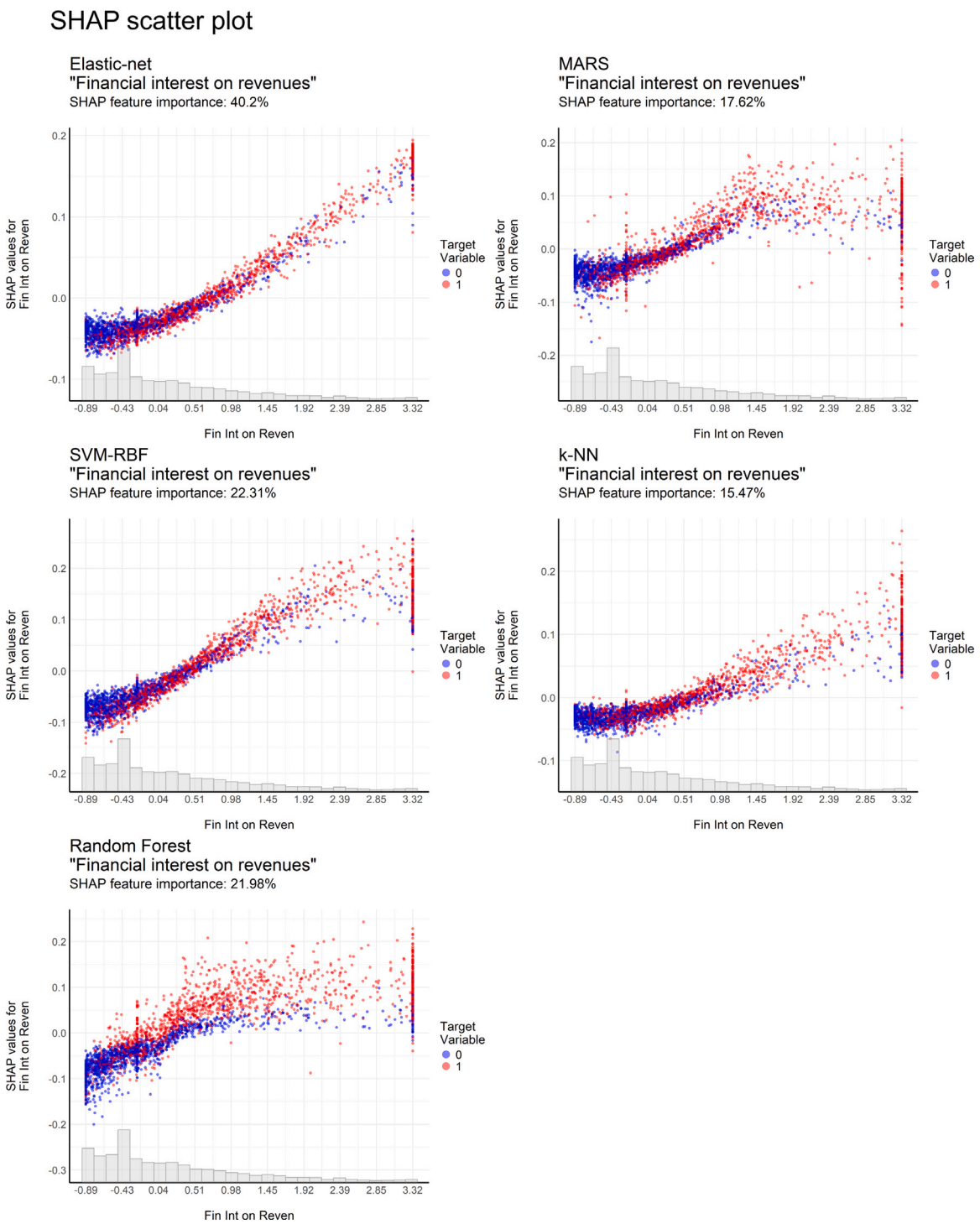


Fig. 6. SHAP scatter plots of the most relevant feature for all models. Value on *x*-axis represent the feature values and distribution, value on the *y*-axis are the corresponding evaluated SHAP values. Points are coloured according to the true class labels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dependence on observables. In particular, for approach (c), first we regressed the raw PD values on fixed effects, then we extracted the regression residuals (orthogonalized PD) and finally we replaced the original PD with these residuals in the full model (while excluding fixed effects). Table C.12 reveals that the performance gain from adding fixed effects to PD (positive Δ% from (a) to (b))

nearly offsets the loss from using residualized PD without fixed effects (a negative Δ% from (a) to (c)), demonstrating that fixed effects primarily capture the observable-correlated portion of PD's predictive power. The residualized PD retains substantial standalone predictive power, confirming that the majority of PD's value derives from firm-specific risk factors orthogonal to sector/size/region. This decomposition proves PD's role as an independent source of information beyond observable proxies.

6.5. Industry-specific controls

Matching private firms to public counterparts to estimate default risk may simply capture industry-specific information rather than firm-specific characteristics. Therefore, a prediction model might overfit industry trends and miss peculiar and sensitive signs of risk specific to the private firm itself. This can be problematic for firms that deviate from the industry norms.

To address this concern and validate that our models capture both firm-specific and industry-level characteristics, we conduct additional experiments by incorporating industry-level measures into our analysis. Specifically, we include two industry-level variables: European βs by Sector, provided by A. Damodaran,[17] a measure of systematic risk at the industry level, and the Industrial Production Index[18] provided by the Italian National Institute of Statistics (ISTAT), a macroeconomic indicator reflecting industry-level production trends. We train all prediction models in additional settings where we include both the proposed PD and the two industry-level indicators separately. Additionally, we evaluated feature importance using the same techniques to assess whether the relevance of the PD remains stable and to determine the contribution of the added industry-level indicators.

Table C.13 in the Appendix compares the F1-score of the models trained on the described settings. Results show that the performance improvement due to the inclusion of both PD and the considered industry-level indicators, both in absolute values and relative differences, is comparable to the ones in the *extended* setting with no controls in Table 5. Therefore, our findings are robust across both settings, with no significant improvement in predictive accuracy when industry-level measures are added. This suggests that our original model already captures the relevant firm-specific characteristics necessary for accurate default risk prediction without relying heavily on industry-level trends. Moreover, the relevance of the proposed PD as a feature is comparable to the one in Table 5, both in ranking and relative importance and for all models. On the other side, the two added indicators show low relevance in all models, both in ranking and relative importance. This reinforces our conclusion that the model primarily relies on firm-specific characteristics rather than industry-level trends.

6.6. Market information

As a further robustness check, to assess the validity of the market information provided by the mapping, we fit the models using the assets volatility and market leverage as additional variables instead of Merton's PD, as suggested in Campbell et al. (2008). Table C.14 in the Appendix reports the F1-score of the models with the two alternative measures of market information, showing similar performance of Merton's PD. The feature importance, both in ranking and relative contribution, for PFI and SHAP prove that the results are aligned with the ones obtained with the PD in Table 5. This provides further empirical evidence that our results are robust to alternative measures of market information, such as stock price volatility and market leverage, consistent with Campbell et al. (2008).

In addition, we replicate our analysis by controlling for the current leverage of the given unlisted firm, i.e., bank finance, as different leverage levels may influence the probability of a company defaulting on its obligation, consistent with well-known studies in the theoretical corporate finance literature (Andrade and Kaplan, 1998). Table C.15 in the Appendix reports the F1-score of the models with both Merton's PD and the variable "Bank Finance", again showing a similar performance of the model with Merton's PD alone. The feature importance for both PFI and SHAP, in terms of ranking and relative contribution, prove that the results are aligned with those obtained with the PD in Table 5, whereas Bank Finance is contributing little to the predictive performance.

6.7. Aggregated risk measures

Furthermore, the added value of the MSMEs-peers matching is evaluated by comparing the impact of the firm-wise Merton's probability of default with the average PD provided by Cerved, an Italian credit rating agency. As for the previous tests, we replicate the analysis while replacing Merton's PD with both the sectorial and geographical Cerved PD. Cerved average PD ranges from 4% to 10%, whereas Mertons' PD ranges from 0% to 100%. Table C.16 in the Appendix reports the F1-score of the models when replacing Merton's PD with the "Average sectorial PD" and "Average geographical PD". Both lower F1-score increases and smaller feature importance ranking and relative contribution prove that aggregated measures of default are not as powerful as firm-wise ones in capturing the credit riskiness of a firm.

Finally, to test both the effect of alternative measures of default riskiness and aggregated indicators, we apply the same routine explained above and replace Merton's PD with the market implied volatility of FTSE 100 and FTSE MIB Italian indexes. Table C.17 in the Appendix reports the F1-score of the models when replacing Merton's PD with the "Market Volatility Implied Index" variable. Again, results show that average market implied volatility cannot adequately capture the uniqueness of each firm's risk profile.

[17] https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datahistory.html
[18] http://dati.istat.it/Index.aspx?DataSetCode=DCSC_INDXPRODIND_1

7. Conclusions

By exploiting a unique and proprietary dataset comprising 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) operating with 113 cooperative banks over the period 2012–2014, this paper investigates the role of market information in predicting corporate default for unlisted firms by digging into the underlying economic drivers of the improvement in default prediction of unlisted private firms that comes from the peers' market-based information. To address our research question, we exploit a granular proprietary dataset of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) required to publicly disclose their financial statements and we propose a novel public–private firms' mapping approach to investigate whether peers' market-based information enhances the accuracy of default prediction for private unlisted firms. Specifically, our novel mapping approach matches the market information of listed firms with that of private firms by following a data-driven clustering by means of Neural Networks Autoencoder. This process allows us to map the Merton's Probability of Default (PD) of public peers to the private firms belonging to the same cluster. We then adopt five statistical techniques, namely linear models, multivariate adaptive regression spline (MARS), support vector machine with radial basis function kernel (SVM-RBF), k-Nearest Neighbours (k-NN) and random forest (RF) to predict corporate default at the private firm-level, and compare the performance of the models with and without the inclusion of the Merton's Probability of Default (PD) estimated using the peer's market-based information. Finally, we make use of Shapley values to assess the contribution of each predictor. The status of the bank's clients is predicted using five statistical models.

Our results provide novel evidence that market information represents a crucial indicator in predicting the corporate default of unlisted firms. Indeed, we show a significant improvement in model performance, both on class-specific (F1-score for defaulted class) and overall metrics (Area Under the Precision–Recall curve and Matthew's Correlation Coefficient) when using market information in credit risk assessment, in addition to accounting information. Moreover, by taking advantage of global and local variable importance techniques, we prove that the increase in performance is effectively attributable to market information, highlighting its relevant effect in predicting corporate default. Our results, therefore, confirm that private firms are characterized by noisy accounting data that, if considered alone, prevent accurate default prediction.

Our study makes important inferences for policy implications. Indeed, our findings shed new light on the opportunity for banks to potentially integrate their hybrid credit scoring methodologies with market information for credit risk assessments that capture commonality between matched public and private, to increase the accuracy of forecasting corporate defaults for unlisted firms. Thus, the results of this paper could be beneficial for forward-looking financial risk management frameworks (Breden, 2008; Rodriguez Gonzalez et al., 2018) to mitigate issues related to the noisy information disclosure of MSMEs while reaching accurate credit risk assessment.

Future extensions stemming from this work could involve not only applying alternative prediction models to provide further evidence on the importance of market information in predicting corporate default of unlisted firms but also testing the impact of synthetic information extracted using the dimensionality reduction technique when replacing the original financial ratios.

CRediT authorship contribution statement

Alessandro Bitetto: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Stefano Filomeni:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Michele Modina:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to the Editor Rajkamal Iyer and to the anonymous referee for their valuable insights and suggestions.

Appendix A. Dataset

See Figs. A.1–A.4 and Tables A.1–A.3.

Table A.1
Correlation matrix of input variables for MSMEs. Legend is below:
1 is 'Oth Reven on Reven', 2 is 'Deprec on Costs', 3 is 'Pay to Bank on Assets', 4 is 'Cashflow on Reven', 5 is 'Fixed Asset Cov', 6 is 'Labour Cost on Reven', 7 is 'ST Pay on Due to Bank', 8 is 'Tot Debt on ST Debt', 9 is 'Tot Debt on Net Worth', 10 is 'Pay to Suppl on Net Worth', 11 is 'Pay to Suppl on Tot Debt', 12 is 'Inventory Duration', 13 is 'Quick Ratio', 14 is 'Debt Burden Index', 15 is 'Fin Int on Reven', 16 is 'Fin Int on Added Val', 17 is 'Net Worth on LT Eqt/Pay', 18 is 'Net Worth on NW+Invent', 19 is 'ROA', 20 is 'ROD', 21 is 'Working Cap Turnover', 22 is 'Turnover'.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	1																					
2	0.19***	1																				
3	0.11***	0.38***	1																			
4	0.16***	0.52***	0.2***	1																		
5	−0.05***	−0.18***	−0.16***	−0.01**	1																	
6	−0.1***	−0.14***	−0.13***	−0.38***	−0.08***	1																
7	−0.04***	−0.13***	−0.28***	−0.06***	0.14***	0.05***	1															
8	0.1***	0.31***	0.55***	0.2***	−0.11***	−0.16***	−0.36***	1														
9	0.03***	−0.09***	0.06***	−0.21***	−0.2***	0.1***	−0.02**	0.03***	1													
10	−0.01	−0.19***	−0.11***	−0.26***	−0.13***	0.11***	0.17***	−0.15***	0.82***	1												
11	−0.11***	−0.32***	−0.4***	−0.13***	0.21***	0	0.55***	−0.52***	−0.03***	0.29***	1											
12	0.21***	0	−0.04***	−0.01	0.06***	−0.09***	−0.07***	0.11***	0.05***	0.01**	−0.1***	1										
13	−0.04***	0.05***	−0.12***	0.16***	0.18***	−0.09***	−0.16***	0.35***	−0.15***	−0.21***	−0.19***	−0.17***	1									
14	0	−0.06***	0.04***	−0.11***	−0.02***	0.2***	−0.09***	0.03***	0.12***	0.08***	−0.06***	0.21***	−0.07***	1								
15	0.25***	0.37***	0.45***	0.23***	−0.17***	−0.2***	−0.29***	0.46***	0.1***	−0.07***	−0.42***	0.31***	0	0.26***	1							
16	0.05***	0.02**	0.25***	−0.09***	−0.1***	−0.09***	−0.27***	0.31***	0.2***	0.11***	−0.25***	0.24***	−0.05***	0.36***	0.6***	1						
17	−0.05***	0.02**	−0.21***	0.24***	0.28***	−0.12***	0.33***	−0.33***	−0.57***	−0.47***	0.34***	−0.08***	0.06***	−0.05***	−0.27***	−0.38***	1					
18	−0.02***	0.29***	0.12***	0.36***	0.04***	−0.16***	0.05***	0.04***	−0.45***	−0.47***	−0.07***	−0.4***	0.31***	−0.18***	0	−0.23***	0.43***	1				
19	−0.08***	−0.04***	−0.12***	0.52***	0.19***	−0.29***	0.08***	−0.09***	−0.21***	−0.16***	0.14***	−0.16***	0.15***	−0.31***	−0.2***	−0.26***	0.31***	0.23***	1			
20	−0.01	0.06***	0.13***	0.06***	−0.14***	−0.08***	−0.3***	0.22***	0.02**	−0.08***	−0.32***	0.02***	0.11***	0.18***	0.52***	0.46***	−0.21***	−0.02***	−0.03***	1		
21	−0.15***	−0.02**	0.34***	0	−0.08***	0.02***	0.02**	0.03***	−0.06***	−0.06***	0	−0.28***	−0.17***	−0.1***	−0.16***	−0.11***	0.04***	0.13***	0.13***	0.1***	1	
22	−0.17***	−0.4***	−0.25***	−0.25***	0.2***	0.09***	0.18***	−0.26***	0.02**	0.14***	0.3***	−0.3***	−0.07***	−0.11***	−0.47***	−0.2***	0.08***	−0.09***	0.25***	−0.11***	0.41***	1

Table A.2
List of input variables for peers dataset.

Variable	Description	Mean	St.Dev.	Min	5th perc	Median	95th perc	Max
1 - Oth Reven on Reven	Other revenues on revenues	0.03	0.08	0	0	0.01	0.08	0.93
2 - Deprec on Costs	Depreciation on costs	0.08	0.11	0	0	0.05	0.31	0.72
3 - Pay to Bank on Assets	Payables to banks on current assets	−1.48	9.64	−90.58	−10.6	0.19	2.74	16.84
4 - Cashflow on Reven	Cash flow on revenues	−3.4	41.62	−526.34	−0.31	0.05	0.26	0.71
5 - Fixed Asset Cov	Fixed asset coverage	14.52	137.75	−0.19	0.49	1.13	2.86	1727.34
6 - Labour Cost on Reven	Labour cost on revenues	−0.06	10.79	−125.98	−0.05	0.69	1.43	37.86
7 - ST Pay on Due to Bank	Short-term payables on amounts due to banks	26.86	81.79	0.51	0.97	4.31	100	924.29
8 - Tot Debt on ST Debt	Total debt on short-term debts	1.73	1.05	1.01	1.07	1.38	3.37	7.28
9 - Tot Debt on Net Worth	Total debt on net worth	2.42	9.67	−72.91	0.24	1.61	6.72	68.4
10 - Pay to Suppl on Net Worth	Payables to suppliers on Net worth	0.71	2.28	−6.31	0.04	0.35	1.84	17.8
11 - Pay to Suppl on Tot Debt	Payables to suppliers on Total debt	0.28	0.17	0.02	0.04	0.25	0.59	0.75
12 - Inventory Duration	Inventory duration	0.79	1.15	0	0	0.5	2.26	7.13
13 - Quick Ratio	Quick ratio	1.25	1.07	0.09	0.3	1	2.47	9.41
14 - Debt Burden Index	Debt burden index	0.28	3.07	−16.8	−1.65	0.16	1.58	30.5
15 - Fin Int on Reven	Financial interest on revenues	3.2	38.5	0	0	0.02	0.39	486.94
16 - Fin Int on Added Val	Financial interest on added value	−0.19	3.05	−28.69	0	0.07	0.7	5.86
17 - Net Worth on LT Eqt/Pay	Net worth on long-term equity and payables	0.62	0.48	−3.82	0.25	0.7	0.96	0.99
18 - Net Worth on NW+Invent	Net worth on net worth and inventories	0.75	0.37	−2.92	0.42	0.76	1	2.35
19 - ROA	Return on Assets	0	0.09	−0.48	−0.18	0.01	0.09	0.2
20 - ROD	Return on Debt	0.11	0.31	−0.15	−0.04	0	1	1
21 - Working Cap Turnover	Working capital turnover	2.18	5.61	0	0.13	1.25	5.43	69.26
22 - Turnover	Turnover normalized by Total Assets	0.8	0.46	0	0.1	0.78	1.64	2.12
Total Assets	Total Assets (EUR Mln)	201.85	329.77	4.91	9.45	72.93	775.71	1621.96
Total Liabilities	Total Liabilities (EUR Mln)	66.82	243.67	0	0	7.42	118.94	1742.64
Volatility	Assets Volatility	0.52	0.82	0.01	0.04	0.21	2.31	4.18

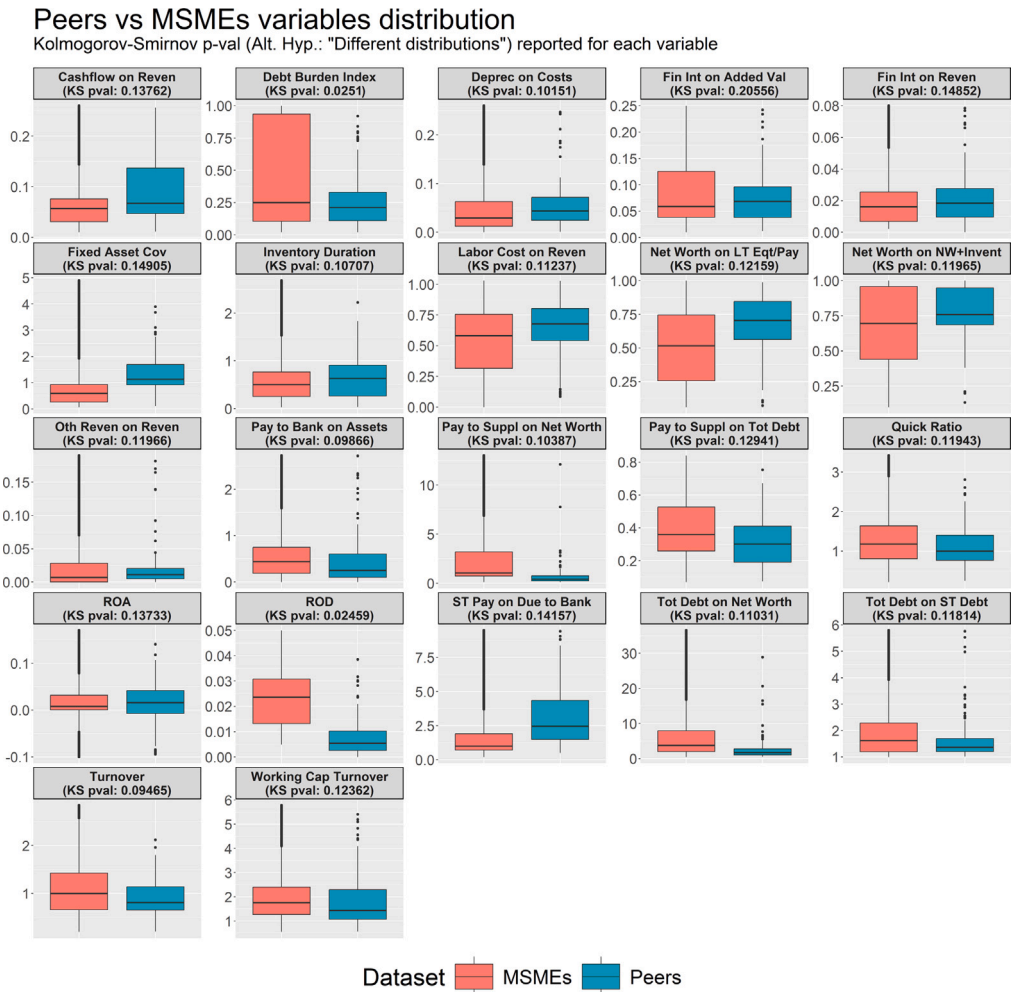


Fig. A.1. Distribution of input variables for Peers and MSMEs. For every variable, the Kolmogorov–Smirnov *p*-value is evaluated when testing the alternative hypothesis of samples coming from different populations.

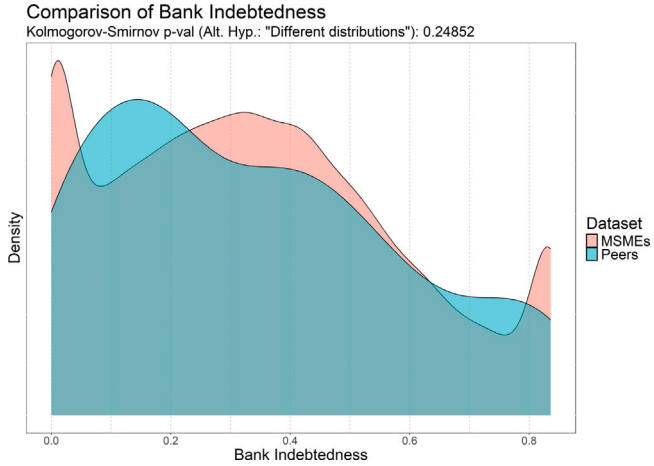


Fig. A.2. Distribution of the bank debt level of the MSMEs compared to the Peers group. The Kolmogorov–Smirnov *p*-value is evaluated when testing the alternative hypothesis of samples coming from different populations.

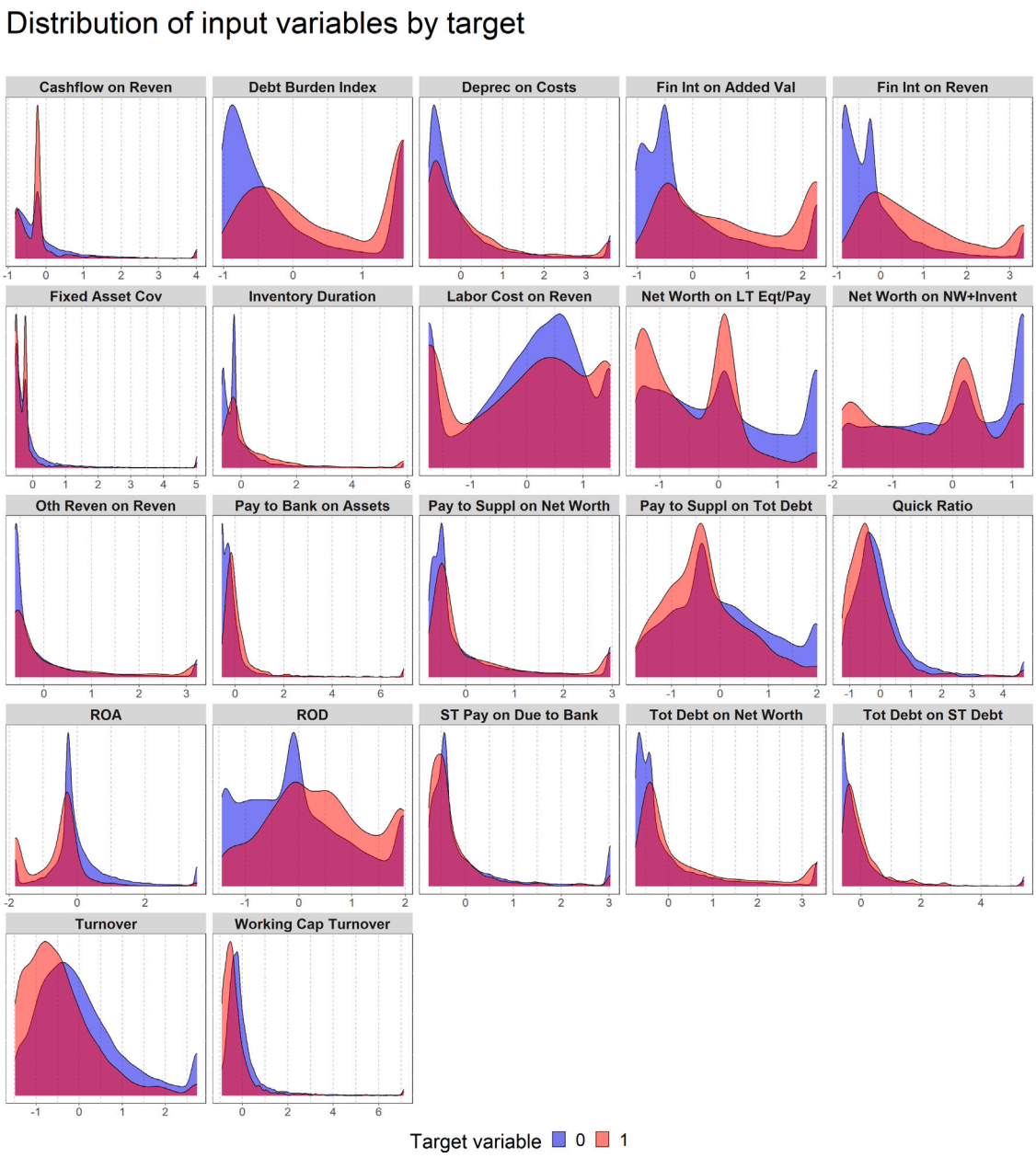


Fig. A.3. Distribution of input variables for MSMEs split by target variable.

Table A.3
Distribution of clients that are persistent over time, i.e., the target is always 0 or 1, compared with clients that move from 0 to 1 and vice-versa.

Target	Total clients	Total banks
0	17,943	9228
1	876	446
0 (0->1)	388	388
0 (1->0)	74	74
1 (0->1)	388	
1 (1->0)	74	
Total	19,743	10,136

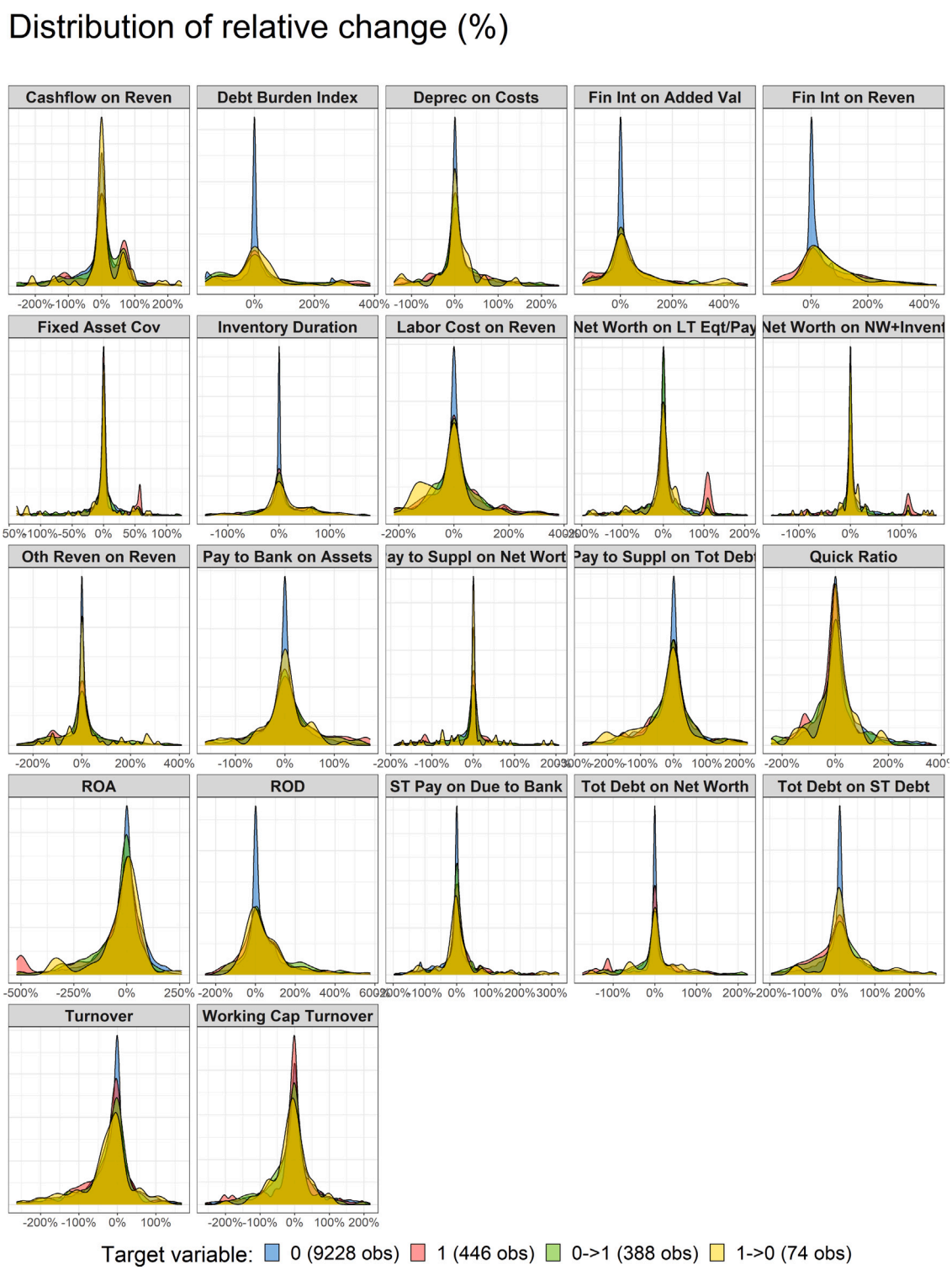


Fig. A.4. Distribution of relative changes over the years of each input variable divided by clients' behaviour. Blue and red distributions represent the clients with persistent targets of 0 and 1, respectively; green and yellow distributions represent the clients that moved from 0 to 1 and vice-versa, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix B. Matching unlisted firms

See Figs. B.5–B.8.

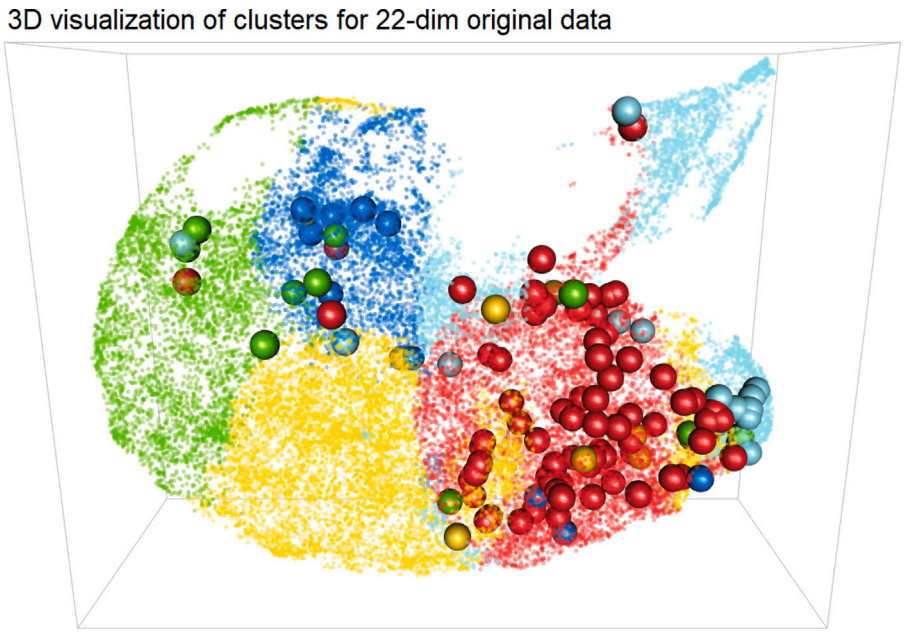


Fig. B.5. 3D visualization of five clusters for the 22-dimensional original data. Visual embedding is evaluated using the UMAP algorithm. Small points are MSMEs observations, bold spheres are peers' observations.

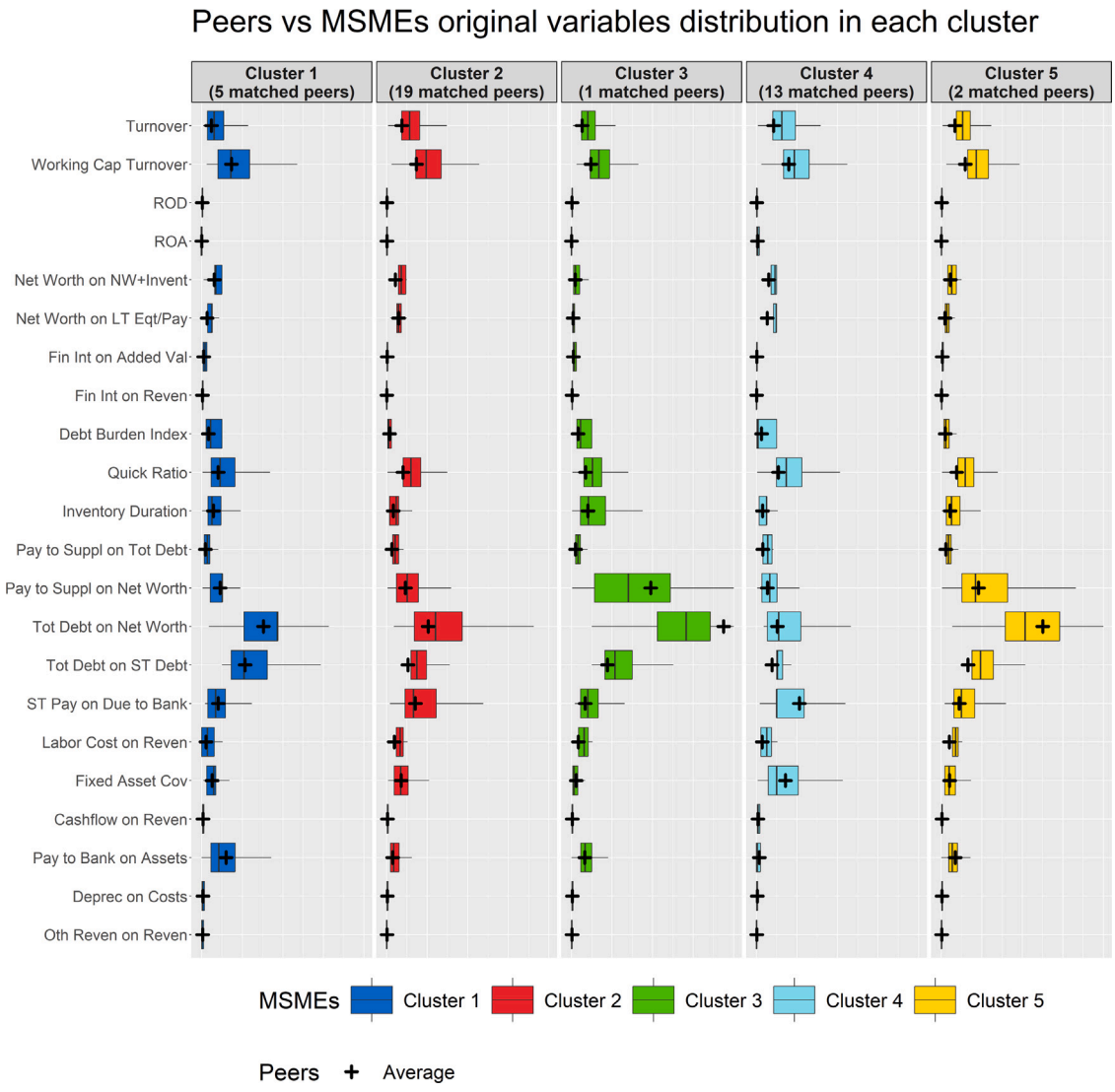


Fig. B.6. Comparison of the distribution of input variables for MSMEs with an average of the assigned peers in each cluster. Colours reflect the ones of Fig. 1, and bold crosses represent the peers' average.

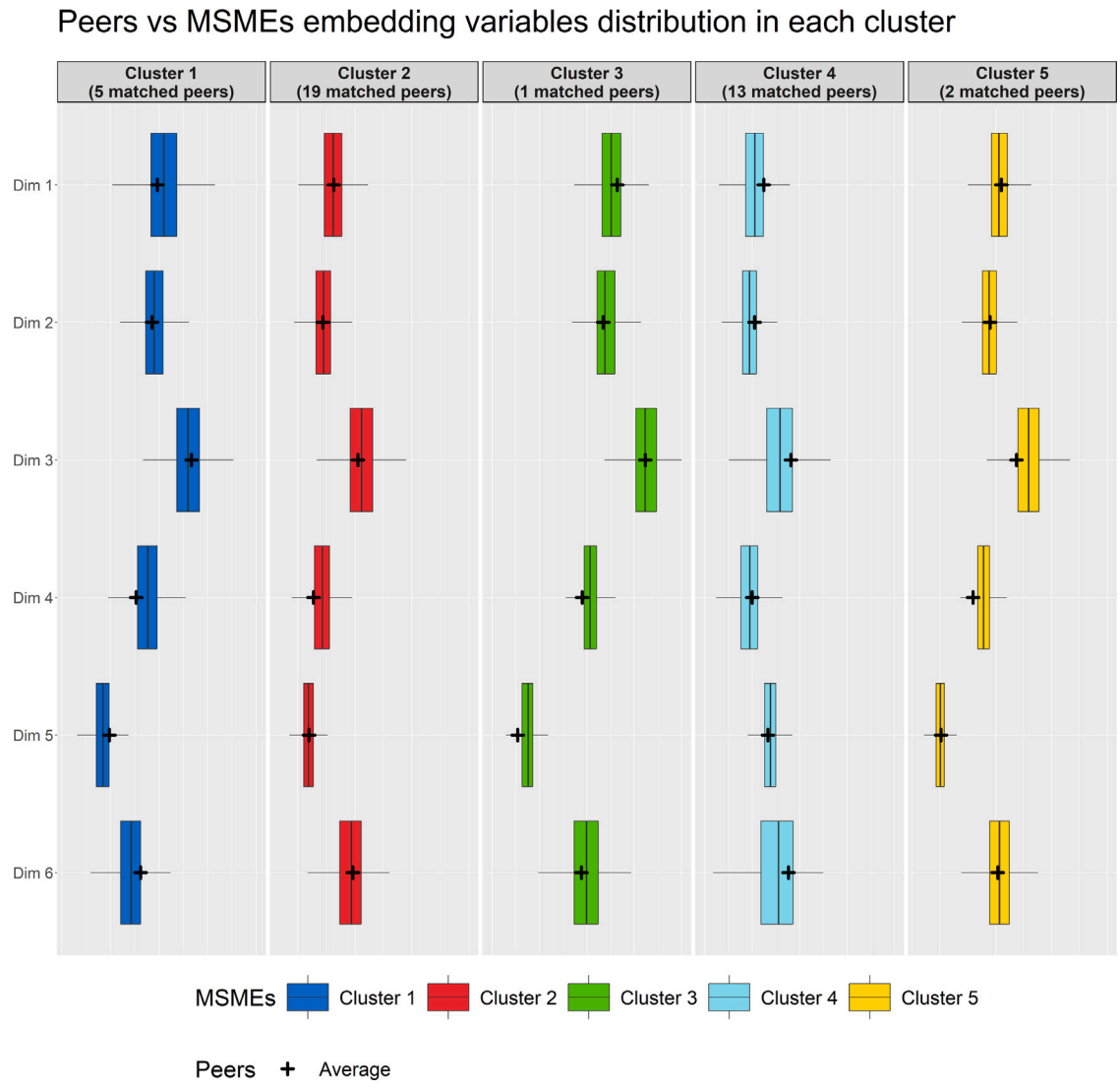


Fig. B.7. Comparison of the distribution of embedding variables for MSMEs with an average of the assigned peers in each cluster. Colours reflect the ones of Fig. 1, and bold crosses represent the peers' average.

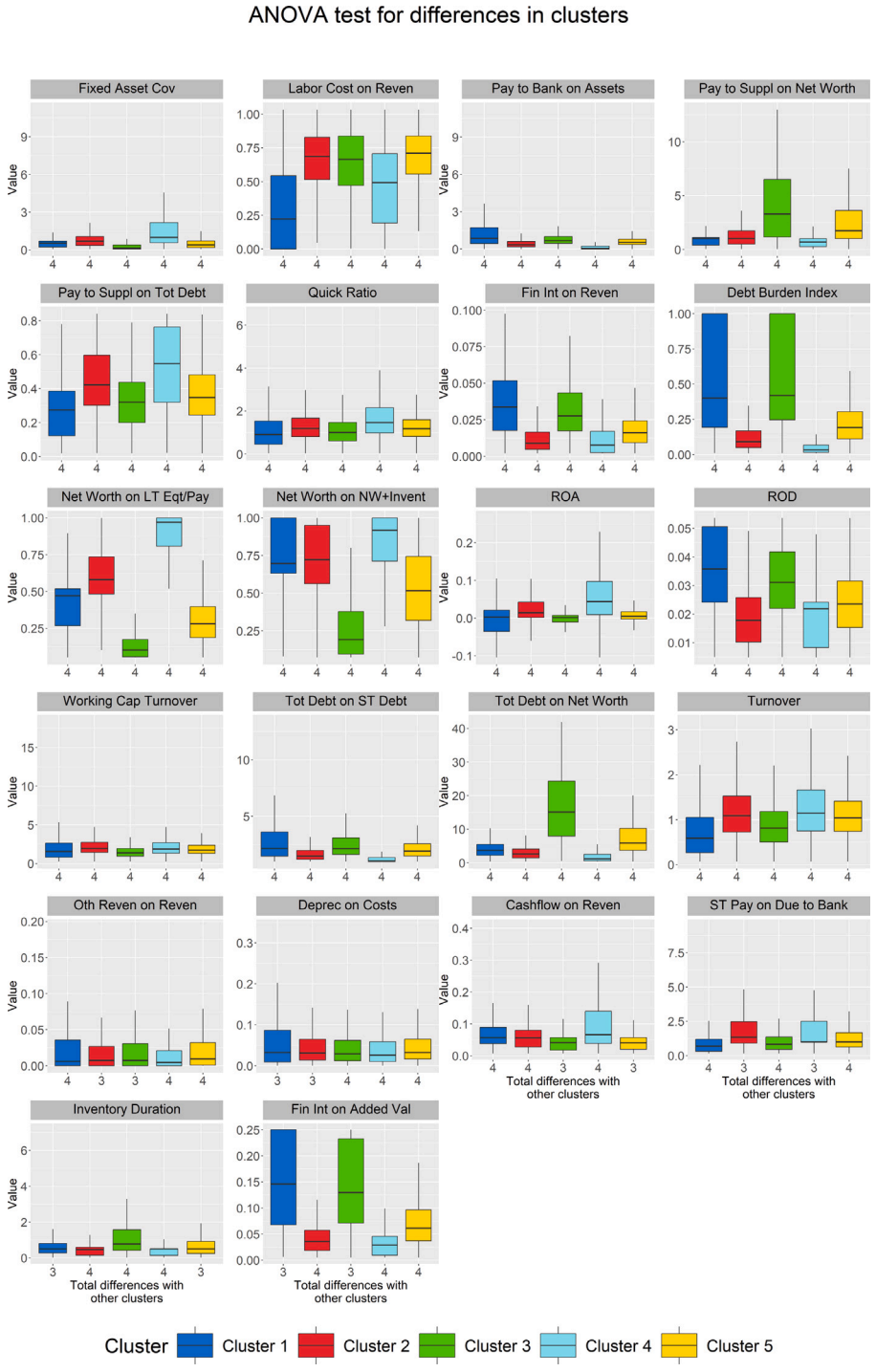


Fig. B.8. Box-plot comparison for original variables grouped by clusters evaluated on the embedding space. Values on *x*-axis report the total of significant differences for the cluster from the others, evaluated via the post-hoc Dunn test.

Appendix C. Prediction of default

See Tables C.4–C.17 and Fig. C.9.

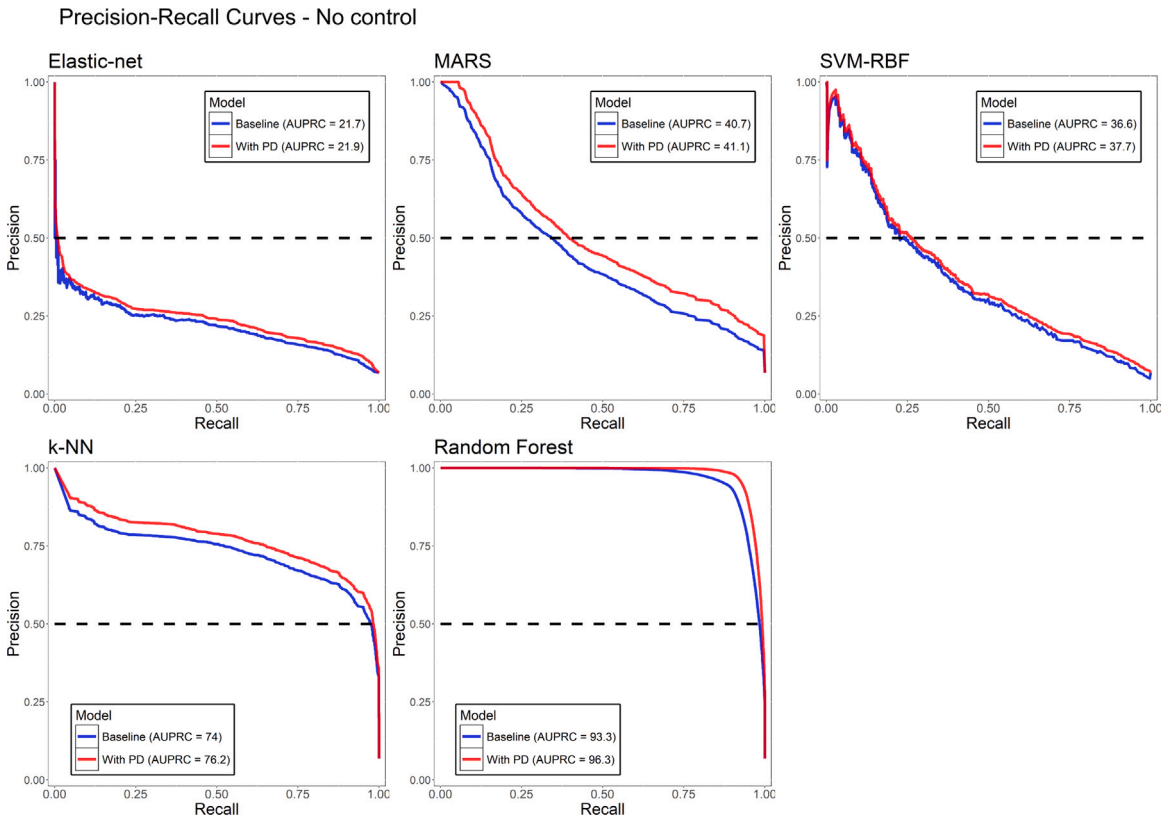


Fig. C.9. Comparison of Precision–Recall curves for all models calibrated with input variables only and with the addition of PD, with no controls for fixed effects. Area Under the Precision–Recall curve is reported in the legend.

Table C.4
Optimal hyper-parameters from the tuning of neural networks. Tuning is performed with Cross-Validation, and the minimum reconstruction error is used as the optimality criterion. For the AE model we tune the layers' structure l_S (both the number of layers and neurons), the size of the bottleneck layer l_B, the activation functions act_S and act_B used in l_S and l_B, respectively, the number of epochs n_E and batch size s used during the training. For the AE-LSTM model we tune the recurrent blocks' structure l_S (both number of layers and number of neurons), the size of the bottleneck layer l_B, the type of recurrent unit $type_{\text{rec}}$ used in all recurrent blocks, the α share of L^1 and L^2 regularization for the weights in each block, the number of epochs n_E and batch size s used during the training. $type_{\text{rec}}$ can be $LSTM$ for Long-Short Term Memory or GRU for Gated Recurrent Unit.

Algorithm	Hyper-parameters
AE	$l_S = [20, 16, 14, 12], l_B = 6, act_S = \tanh, act_B = \text{ReLU}, n_E = 500, s = 500$
AE-LSTM	$l_S = [55, 31, 17], l_B = 10, \alpha = 0.2, type_{rec} = \text{GRU}, n_E = 500, s = 100$

Table C.5
Optimal hyper-parameters from the tuning of each model on both the dataset with and without the Merton's PD. Tuning is performed with Stratified Cross-Validation, and the maximum cross-validation F1 score is used as the optimality criterion. For the Elastic-Net model, we tune the α parameter that represents the share of L^1 and L^2 regularization; for the MARS model we tune the degree d of the polynomial functions; for SVM-RBF model we tune the regularization cost C and the scale of RBF kernel σ; for the k-NN model we tune the number of neighbours k; and for the Random Forest model, we tune the number of trees n_{tree}, the number n_{var} of variables randomly sampled as candidates at each split and the minimum size s of observation in each node.

Algorithm	Hyperparameter range	Baseline	With PD
Elastic-Net	$\alpha \in [0, 1]$	$\alpha = 0.4$	$\alpha = 0.06$
MARS	$d \in \{1, \dots, 6\}$	$d = 2$	$d = 4$
SVM-RBF	$C, \sigma \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$	$C = 100,\ \sigma = 0.001$	$C = 1000,\ \sigma = 0.0001$
k-NN	$d \in \{1, \dots, 100\}$	$k = 27$	$C = 1000,\ k = 31$
Random Forest	$n_{\text{tree}} \in \{50, \dots, 500\}, n_{\text{var}} \in \{1, \dots, 21\}, s \in\in \{1, \dots, 200\}$	$n_{\text{tree}} = 262, n_{\text{var}} = 10, s = 40$	$n_{\text{tree}} = 371, n_{\text{var}} = 9, s = 55$

Table C.6
F1-score, Precision–Recall AUC and Matthews Correlation Coefficient for all considered models on a subset of the MSMEs sample, including only the ones that have a credit relationship with a single bank, for a total of 7155 firms. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via permutation test. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)			PRAUC (Cross-Val)			MCC (Cross-Val)			PD Feature Imp. Rank (Relative %)	
	Baseline	With PD	$\Delta\%$	Baseline	With PD	$\Delta\%$	Baseline	With PD	$\Delta\%$	PFI	SHAP
Elastic-net	31%	31.5%	1.8%**	21.6%	21.3%	−1.2%***	25.4%	25.5%	0.2%*	1	1
	(30.3 ± 1.6%)	(30.1 ± 1.1%)		(20.5 ± 0.7%)	(21 ± 1.4%)		(24.6 ± 1%)	(24.8 ± 1%)		(61.4%)	(39.3%)
MARS	43.5%	43.9%	1%*	41.5%	40.4%	−2.5%*	37%	38.4%	3.6%	1	1
	(41.6 ± 1.3%)	(42.7 ± 1.4%)		(39.7 ± 0.9%)	(38.9 ± 0.9%)		(35.7 ± 1.5%)	(36.4 ± 1%)		(39.6%)	(30.6%)
SVM-RBF	51.6%	53%	2.9%*	37.1%	38.4%	3.7%***	45.1%	46.1%	2.2%**	1	2
	(50.4 ± 1.4%)	(52 ± 1.4%)		(36 ± 1.5%)	(35.8 ± 0.7%)		(42.3 ± 0.8%)	(44.6 ± 1.4%)		(20.8%)	(16.2%)
k-NN	69.9%	72.2%	3.3%**	73.2%	77.4%	5.8%***	67.1%	67.3%	0.3%**	1	2
	(68.5 ± 1.1%)	(71.5 ± 1.3%)		(70.6 ± 1.7%)	(74 ± 1.4%)		(63.9 ± 0.8%)	(66 ± 1.3%)		(19%)	(12.3%)
Random	87%	93.1%	7%*	91.2%	94.3%	3.5%***	85.9%	93.4%	8.8%*	1	1
Forest	(84.4 ± 0.8%)	(90.1 ± 0.8%)		(86.2 ± 0.8%)	(91 ± 1.3%)		(83.8 ± 1.4%)	(92 ± 1.5%)		(24.9%)	(24.5%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.7
Comparison of F1-score for all considered models calibrated on a dataset with input variables only and with the addition of PD and with or without Static controls for fixed effects. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. *Δ%* is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

	Control	Algorithm	F1 (Cross-Val)			PD Feature Importance Rank (Relative %)	
			Baseline	With PD	Δ%	PFI	SHAP
Static	Dummy Industry	Elastic-net	31.1% (30.1 ± 1.8%)	31.5% (30.8 ± 1.2%)	1.5%**	1 (58.1%)	1 (39%)
		MARS	44.3% (42.7 ± 1.2%)	44.6% (42.7 ± 1.5%)	0.7%*	1 (37.1%)	1 (30.6%)
		SVM-RBF	52.6% (50.5 ± 1.4%)	54% (52.6 ± 1.3%)	2.7%*	1 (20%)	2 (15.8%)
		k-NN	71.3% (68.5 ± 1.1%)	73.7% (71.8 ± 1.3%)	3.3%**	1 (8.8%)	2 (12%)
		Random Forest	89.4% (85.7 ± 0.8%)	96.3% (93.3 ± 0.7%)	7.7%*	1 (27.3%)	1 (24.1%)
	Firm Type	Elastic-net	31.2% (30.2 ± 1.8%)	31.5% (30.7 ± 1.1%)	1.1%**	1 (61.8%)	1 (38.2%)
		MARS	44.2% (42.6 ± 1.2%)	44.6% (42.9 ± 1.5%)	0.8%*	1 (38.4%)	1 (32.4%)
		SVM-RBF	52.3% (50.6 ± 1.4%)	53.7% (52.4 ± 1.2%)	2.6%*	1 (21.1%)	2 (16.8%)
		k-NN	71.8% (68.7 ± 1.2%)	73.5% (72.2 ± 1.5%)	2.4%**	1 (8.9%)	2 (11.9%)
		Random Forest	89.9% (85.2 ± 0.9%)	96.7% (93.3 ± 0.9%)	7.6%*	1 (26.8%)	1 (24.5%)
	Industrial Sector	Elastic-net	31.3% (30.2 ± 1.6%)	31.5% (30.7 ± 1.1%)	0.5%**	1 (63.2%)	1 (39.4%)
		MARS	44% (42.5 ± 1.2%)	44.8% (42.5 ± 1.5%)	1.8%*	1 (39.6%)	1 (30.2%)
		SVM-RBF	52.1% (50.5 ± 1.4%)	53.9% (52.4 ± 1.3%)	3.3%*	1 (21.2%)	2 (16.4%)
		k-NN	71.6% (68.4 ± 1.1%)	73.9% (71.7 ± 1.4%)	3.2%**	1 (9.4%)	2 (12.2%)
		Random Forest	90% (85.4 ± 1%)	96.8% (93.2 ± 0.8%)	7.5%*	1 (26.3%)	1 (23.4%)
	Region	Elastic-net	31.1% (30.3 ± 1.7%)	31.4% (30.6 ± 1.1%)	0.9%**	1 (57.2%)	1 (35.9%)
		MARS	44.3% (42.4 ± 1.3%)	44.6% (42.6 ± 1.5%)	0.7%*	1 (36.2%)	1 (31.9%)
		SVM-RBF	52.5% (50.4 ± 1.3%)	54.1% (52.5 ± 1.4%)	3%*	1 (18.9%)	2 (16%)
		k-NN	71.4% (68.7 ± 1.3%)	73.4% (72 ± 1.5%)	2.9%**	1 (8.8%)	2 (12.5%)
		Random Forest	89.3% (85.5 ± 0.8%)	96.9% (93.1 ± 0.9%)	8.5%*	1 (25.8%)	1 (25.4%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.8
Comparison of F1-score for all considered models calibrated on a dataset with input variables only and with the addition of PD and with or without Dynamic controls for fixed effects. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

	Control	Algorithm	F1 (Cross-Val)			PD Feature Importance Rank (Relative %)	
			Baseline	With PD	$\Delta\%$	PFI	SHAP
Dynamic	Firm Size	Elastic-net	31.5% (30.5 ± 1.7%)	31.5% (31.2 ± 1.1%)	0.1%**	1 (60.9%)	1 (40.1%)
		MARS	44.3% (43.2 ± 1.3%)	44.6% (43.1 ± 1.6%)	0.6%*	1 (37.3%)	1 (34%)
		SVM-RBF	52.3% (51 ± 1.3%)	53.9% (53.1 ± 1.3%)	3.1%*	1 (19.9%)	2 (17%)
		k-NN	72.5% (68.3 ± 1.3%)	74.7% (71.8 ± 1.4%)	2.9%**	1 (8.8%)	2 (12.4%)
		Random Forest	90.3% (85.1 ± 0.9%)	96.6% (93.4 ± 0.8%)	7%*	1 (25.2%)	1 (25.9%)
	Funding Risk	Elastic-net	31.1% (30.5 ± 1.9%)	31.4% (31 ± 1.1%)	0.8%**	1 (62.5%)	1 (36%)
		MARS	44.8% (42.9 ± 1.3%)	45.1% (43.3 ± 1.6%)	0.6%*	1 (38.8%)	1 (31.4%)
		SVM-RBF	53% (50.7 ± 1.3%)	54.3% (53.2 ± 1.4%)	2.5%*	1 (20.1%)	2 (16.6%)
		k-NN	72.1% (69 ± 1.2%)	73.8% (72.4 ± 1.3%)	2.3%**	1 (9.4%)	2 (12.7%)
		Random Forest	90.5% (85.4 ± 1%)	96.8% (93.2 ± 0.8%)	6.9%*	1 (24.6%)	1 (24.7%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.9
Comparison of F1-score for all considered models when predicting the default flag with the original 22 variables (first column) and with the embedding evaluated through the AE, RobPCA and AE-LSTM techniques. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions.

Algorithm		Embedding					
	22 features	AE		RobPCA		AE-LSTM	
	F1	F1	$\Delta\%$	F1	$\Delta\%$	F1	$\Delta\%$
Elastic-net	30.7% (29.8 ± 1.7%)	32.1% (31.7 ± 1.6%)	4.7%*	28.5% (27.8 ± 1.6%)	−7.3%*	29.9% (28.6 ± 1.7%)	−2.8%*
MARS	43.6% (42 ± 1.2%)	45.9% (43.5 ± 1.2%)	5.2%**	40.2% (38.3 ± 1.2%)	−7.7%**	40.7% (40 ± 1.1%)	−6.6%**
SVM-RBF	51.6% (49.7 ± 1.3%)	48.6% (46.5 ± 1.3%)	−5.8%**	47.6% (45.4 ± 1.4%)	−7.8%**	46.4% (45.3 ± 1.3%)	−10.1%**
k-NN	70.5% (67.5 ± 1.2%)	65.8% (63.7 ± 1.2%)	−6.7%*	65.2% (61.4 ± 1.3%)	−7.5%*	64.4% (62 ± 1.2%)	−8.7%*
Random Forest	88.4% (84.1 ± 0.9%)	82.4% (79 ± 1%)	−6.8%*	81.7% (77.9 ± 0.9%)	−7.6%*	81.5% (77.6 ± 0.9%)	−7.8%*

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.10
Comparison of F1-score, Precision–Recall AUC and Matthews Correlation Coefficient for all considered models when predicting the default flag with the original variables only and when including a randomized PD. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "Baseline" and "Random PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)			PRAUC (Cross-Val)			MCC (Cross-Val)			PD Feature Importance Rank (Relative %)	
	Baseline	Random PD	Δ%	Baseline	Random PD	Δ%	Baseline	Random PD	Δ%	PFI	SHAP
Elastic-net	30.7% (29.8 ± 1.5%)	30.7% (29.9 ± 1.6%)	0%***	21.7% (21 ± 1.7%)	21.7% (21.1 ± 1.6%)	0.2%*	25.6% (24.9 ± 1%)	25.6% (25 ± 0.9%)	0%***	18 (3.4%)	13 (2.2%)
MARS	43.6% (42.7 ± 1.6%)	43.8% (42.9 ± 1.7%)	0.4%***	40.7% (39.1 ± 0.7%)	40.8% (39.2 ± 0.7%)	0.1%*	37.7% (35.6 ± 1.3%)	37.8% (35.7 ± 1.2%)	0.3%	19 (1.1%)	15 (2.9%)
SVM-RBF	51.6% (49.2 ± 1.2%)	51.7% (49.2 ± 1.1%)	0.1%	36.6% (34.8 ± 1.5%)	36.7% (34.9 ± 1.5%)	0.2%**	45.8% (43.4 ± 1.6%)	45.8% (43.4 ± 1.5%)	0%***	16 (2.3%)	16 (1.6%)
k-NN	70.5% (66.8 ± 1.2%)	70.6% (67 ± 1.1%)	0.1%***	74% (69.6 ± 0.8%)	74% (69.6 ± 0.8%)	0%*	66.9% (63.3 ± 1.3%)	67% (63.4 ± 1.4%)	0.1%	16 (1%)	15 (2.7%)
Random Forest	88.4% (83.8 ± 0.7%)	88.6% (84.2 ± 0.7%)	0.2%*	93.3% (91.3 ± 1.3%)	93.5% (91.5 ± 1.3%)	0.2%***	87% (81.8 ± 0.7%)	87.3% (81.9 ± 0.7%)	0.4%	14 (2%)	15 (1.5%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.11
Comparison of F1-score for all considered models when predicting the default flag with the original variables only and when including the PD and when the train-test procedure is evaluated in a year-on-year rolling window or training on 2012–2013 and testing on 2014. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)						Feature Importance Rank (Relative %)			
	Rolling window			2012–2013 vs 2014			PFI		SHAP	
	Baseline	With PD	Δ%	Baseline	With PD	Δ%	Rolling window	2012–2013 vs 2014	Rolling window	2012–2013 vs 2014
Elastic-net	30.7% (29.8 ± 1.5%)	30.9% (29.5 ± 1.2%)	0.5%*	30.7% (29.8 ± 1.5%)	30.7% (30.7 ± 1.3%)	0.2%**	1 (61.8%)	1 (61.4%)	1 (37.9%)	1 (38%)
MARS	43.6% (42.7 ± 1.6%)	43.7% (42.3 ± 1.2%)	0.1%***	43.6% (42.7 ± 1.6%)	44.2% (44.2 ± 1.3%)	1.4%	1 (38.4%)	1 (38.2%)	1 (31.9%)	1 (31.8%)
SVM-RBF	51.6% (49.2 ± 1.2%)	52.4% (52.4 ± 1.7%)	1.5%**	51.6% (49.2 ± 1.2%)	53.3% (53.3 ± 1.5%)	3.3%***	1 (19.8%)	1 (20.1%)	2 (16.8%)	2 (16.4%)
k-NN	70.5% (66.8 ± 1.2%)	71.8% (71.8 ± 1.1%)	1.9%*	70.5% (66.8 ± 1.2%)	72.7% (72.7 ± 1.1%)	3.1%*	1 (18.8%)	1 (19.2%)	2 (12.1%)	2 (12.1%)
Random Forest	88.4% (83.8 ± 0.7%)	92.7% (92.7 ± 1%)	4.8%**	88.4% (83.8 ± 0.7%)	94.7% (94.7 ± 1%)	7.1%*	1 (25.5%)	1 (26.2%)	1 (24.8%)	1 (24.6%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.12
Comparison of F1-score only for k-NN and Random Forest models calibrated on a dataset with the addition of PD, with the PD and Static or Dynamic controls for fixed effects and with the PD "residualized" on the controls. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "With PD" and "With PD and Fix. Eff." and "With PD" and "With PD Residualized" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions.

	Control	Algorithm	F1 (Cross-Val)					
			With PD	With PD and Fix. Eff.	$\Delta\%$	With PD	With PD Residualized	$\Delta\%$
Static	Dummy Industry	k-NN	72.6% (70.8 ± 1.4%)	73.7% (71.8 ± 1.4%)	1.5%**	72.6% (70.8 ± 1.4%)	71.4% (69.9 ± 1.4%)	−1.6%**
		Random Forest	95% (91.9 ± 0.8%)	96.3% (93.3 ± 0.8%)	1.4%*	95% (91.9 ± 0.8%)	93.2% (90.6 ± 0.9%)	−1.9%*
	Firm Type	k-NN	72.6% (70.8 ± 1.4%)	73.5% (72.2 ± 1.4%)	1.3%**	72.6% (70.8 ± 1.4%)	72.1% (70.5 ± 1.5%)	−0.6%**
		Random Forest	95% (91.9 ± 0.8%)	96.7% (93.3 ± 0.8%)	1.8%*	95% (91.9 ± 0.8%)	93.4% (89.8 ± 0.7%)	−1.6%*
	Industrial Sector	k-NN	72.6% (70.8 ± 1.4%)	73.9% (71.7 ± 1.4%)	1.7%**	72.6% (70.8 ± 1.4%)	71.8% (70.1 ± 1.4%)	−1.1%**
		Random Forest	95% (91.9 ± 0.8%)	96.8% (93.2 ± 0.8%)	1.9%*	95% (91.9 ± 0.8%)	93.7% (90.6 ± 0.8%)	−1.3%*
	Region	k-NN	72.6% (70.8 ± 1.4%)	73.4% (72 ± 1.4%)	1.2%**	72.6% (70.8 ± 1.4%)	72.2% (70.4 ± 1.3%)	−0.6%**
		Random Forest	95% (91.9 ± 0.8%)	96.9% (93.1 ± 0.8%)	2%*	95% (91.9 ± 0.8%)	92.7% (89.7 ± 0.8%)	−2.4%*
Dynamic	Firm Size	k-NN	72.6% (70.8 ± 1.4%)	74.7% (71.8 ± 1.4%)	2.8%**	72.6% (70.8 ± 1.4%)	70.6% (68.3 ± 1.4%)	−2.8%**
		Random Forest	95% (91.9 ± 0.8%)	96.6% (93.4 ± 0.8%)	1.7%*	95% (91.9 ± 0.8%)	93.6% (89.5 ± 0.8%)	−1.4%*
	Funding Risk	k-NN	72.6% (70.8 ± 1.4%)	73.8% (72.4 ± 1.4%)	1.7%**	72.6% (70.8 ± 1.4%)	71.3% (69 ± 1.3%)	−1.8%**
		Random Forest	95% (91.9 ± 0.8%)	96.8% (93.2 ± 0.8%)	1.9%*	95% (91.9 ± 0.8%)	93.8% (90.2 ± 0.9%)	−1.2%*

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.13
Comparison of F1-score for all considered models when predicting the default flag with the original variables only and when including the PD and two industry-level controls: European βs by Sector and Italian Industrial Production Index. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. $\Delta\%$ is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)						Feature Importance Rank (Relative %)					
	Sector β			Ind. Prod.			PFI			SHAP		
	Baseline	With PD and Sector β	$\Delta\%$	Baseline	With PD and Ind. Prod.	$\Delta\%$	PD	Ind. Prod.	Ind. Prod.	PD	Sector β	Ind. Prod.
Elastic-net	30.9% (30 ± 1.5%)	31.2% (30.6 ± 1%)	1.1%**	30.4% (29.4 ± 1.7%)	30.6% (29.9 ± 1.1%)	0.8%**	1 (63%)	6 (7.9%)	10 (5.6%)	1 (36.8%)	8 (4.6%)	9 (3.6%)
MARS	44% (42.4 ± 1.3%)	44.4% (42.3 ± 1.4%)	1%*	43% (41.4 ± 1.2%)	43.6% (41.6 ± 1.6%)	1.4%*	1 (36.6%)	6 (4.1%)	11 (4%)	1 (30.9%)	10 (6.2%)	10 (3.8%)
SVM-RBF	52.1% (50 ± 1.3%)	53.5% (52.1 ± 1.4%)	2.7%*	51% (49.1 ± 1.2%)	52.4% (51.1 ± 1.2%)	2.8%*	1 (20.9%)	9 (2.6%)	8 (2%)	2 (16.9%)	7 (3.4%)	10 (1.7%)
k-NN	70.9% (68 ± 1.1%)	73.1% (71.3 ± 1.3%)	3.1%**	69.7% (66.6 ± 1.1%)	71.7% (70 ± 1.5%)	2.9%**	1 (18.9%)	10 (1.1%)	8 (1.2%)	2 (11.7%)	7 (1.3%)	13 (1.6%)
Random Forest	88.9% (84.6 ± 0.9%)	95.7% (92.5 ± 0.8%)	7.7%*	87.2% (82.9 ± 0.9%)	93.9% (90.8 ± 0.9%)	7.7%*	1 (25.3%)	9 (5.1%)	10 (2.5%)	1 (24.6%)	10 (3.5%)	11 (2.7%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.14
Comparison of F1-score for all considered models when predicting the default flag with the original variables only and when replacing the PD with two alternative measures for market information: Volatility and Leverage. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. *Δ%* is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)						Feature Importance Rank (Relative %)			
	Volatility			Leverage			PFI		SHAP	
	Baseline	With Volatility	Δ%	Baseline	With Leverage	Δ%	Volatility	Leverage	Volatility	Leverage
Elastic-net	30.7% (29.8 ± 1.7%)	30.8% (29.9 ± 1.2%)	0.1%**	30.7% (29.8 ± 1.7%)	30.7% (29.7 ± 1.1%)	0.1%*	1 (60%)	1 (61.2%)	1 (37.8%)	1 (37.8%)
MARS	43.6% (42 ± 1.2%)	44.3% (42.7 ± 1.5%)	1.5%*	43.6% (42 ± 1.2%)	44.4% (43.7 ± 1.5%)	1.8%***	1 (38.7%)	1 (39.2%)	1 (32%)	1 (32.5%)
SVM-RBF	51.6% (49.7 ± 1.3%)	52.5% (52.5 ± 1.3%)	1.8%**	51.6% (49.7 ± 1.3%)	51.8% (50.6 ± 1.2%)	0.5%***	1 (20.5%)	1 (20%)	2 (16.3%)	2 (16.3%)
k-NN	70.5% (67.5 ± 1.2%)	71.2% (71.2 ± 1.3%)	1%*	70.5% (67.5 ± 1.2%)	70.6% (70.6 ± 1.5%)	0.1%*	1 (19.1%)	1 (19.4%)	2 (12.4%)	2 (12.2%)
Random Forest	88.4% (84.1 ± 0.9%)	92.7% (92.7 ± 0.7%)	4.9%***	88.4% (84.1 ± 0.9%)	94.2% (94.2 ± 0.9%)	6.5%	1 (26.3%)	1 (25.6%)	1 (24.2%)	1 (24.8%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.15
Comparison of F1-score for all considered models when predicting the default flag with the original variables only and when including both Merton's PD and the "Bank Finance" variable. Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. *Δ%* is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)			Feature Importance Rank (Relative %)			
	PD and Bank Finance			PFI		SHAP	
	Baseline	With PD and Bank Finance	Δ%	PD	Bank Finance	PD	Bank Finance
Elastic-net	30.7% (29.1 ± 1.6%)	30.9% (29.9 ± 1.1%)	0.6%**	1 (61.6%)	5 (7.3%)	1 (38.3%)	5 (5.1%)
MARS	41.9% (40.7 ± 1.2%)	43.4% (43.4 ± 1.5%)	3.7%	1 (38.3%)	5 (4.6%)	1 (32.3%)	5 (3.4%)
SVM-RBF	49.7% (49 ± 1.4%)	52% (52 ± 1.3%)	4.7%*	1 (20.2%)	6 (3%)	2 (16.5%)	8 (3.3%)
k-NN	69.2% (65.6 ± 1.1%)	71.8% (71.8 ± 1.4%)	3.7%	1 (18.9%)	6 (2.9%)	2 (12.2%)	7 (1.8%)
Random Forest	87.9% (82.4 ± 1%)	94.5% (94.5 ± 0.8%)	7.6%**	1 (26%)	7 (1%)	1 (24.8%)	6 (1.2%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.16
Comparison of F1-score for all considered models when predicting the default flag with the original variables only and when replacing Merton's PD with the "Average sectorial PD" and "Average geographical PD". Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. *Δ%* is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)						Feature Importance Rank (Relative %)			
	PD Sector Avg.			PD Geogr. Avg.			PFI		SHAP	
	Baseline	With PD Sector Avg.	Δ%	Baseline	With PD Geogr. Avg.	Δ%	PD Sector Avg.	PD Geogr. Avg.	PD Sector Avg.	PD Geogr. Avg.
Elastic-net	30.7% (29.8 ± 1.7%)	30.9% (30.9 ± 1.1%)	0.6%*	30.7% (29.8 ± 1.7%)	31.3% (31.3 ± 1.1%)	1.8%*	4 (14.9%)	5 (6.5%)	6 (10.5%)	7 (5.1%)
MARS	43.6% (42 ± 1.2%)	44.1% (43.1 ± 1.6%)	1.1%**	43.6% (42 ± 1.2%)	43.7% (42.2 ± 1.6%)	0.3%***	4 (9.4%)	6 (7.7%)	6 (8.3%)	5 (3.2%)
SVM-RBF	51.6% (49.7 ± 1.3%)	52% (52 ± 1.3%)	0.8%***	51.6% (49.7 ± 1.3%)	52.2% (52.2 ± 1.4%)	1.2%*	6 (5%)	6 (3%)	5 (3.3%)	6 (3.3%)
k-NN	70.5% (67.5 ± 1.2%)	73.1% (73.1 ± 1.4%)	3.7%***	70.5% (67.5 ± 1.2%)	70.7% (70.7 ± 1.3%)	0.3%**	6 (4.8%)	6 (1.9%)	5 (1.2%)	7 (1.2%)
Random Forest	88.4% (84.1 ± 0.9%)	93.5% (93.5 ± 0.8%)	5.7%*	88.4% (84.1 ± 0.9%)	92.9% (92.9 ± 0.8%)	5.1%***	6 (5.2%)	5 (7.3%)	6 (3.7%)	5 (8.9%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Table C.17
Comparison of F1-score for all considered models when predicting the default flag with the original variables only and when replacing Merton's PD with the market implied volatility of FTSE 100 and FTSE MIB Italian indexes "Market Volatility Implied Index". Values refer to the performance of the model calibrated on the entire dataset. Values in parentheses refer to the average performance of validation folds of Cross-Validation. *Δ%* is the relative improvement in performance between the "Baseline" and "With PD" settings. The statistical significance of the difference is evaluated via a permutation test with 5000 repetitions. The rank of importance for PD variable is reported in the last two columns for both Permutation Feature Importance and SHAP techniques. Values in parentheses refer to the relative normalized importance.

Algorithm	F1 (Cross-Val)			Feature Importance Rank (Relative %)	
	Mkt. Impl. Vol. Index			Mkt. Impl. Vol. Index	
	Baseline	With Mkt. Impl. Vol. Index	Δ%	PFI	SHAP
Elastic-net	30.7% (29.8 ± 1.7%)	31% (29.3 ± 1.1%)	1.1%	7 (0.7%)	8 (3.8%)
MARS	43.6% (42 ± 1.2%)	44.6% (43 ± 1.5%)	2.2%*	9 (2%)	8 (5.4%)
SVM-RBF	51.6% (49.7 ± 1.3%)	51.9% (51.9 ± 1.2%)	0.7%***	9 (5.3%)	10 (2.8%)
k-NN	70.5% (67.5 ± 1.2%)	71.7% (72.2 ± 1.4%)	1.8%***	7 (0.9%)	10 (1.2%)
Random Forest	88.4% (84.1 ± 0.9%)	89.7% (89.7 ± 0.9%)	1.5%**	7 (1.2%)	7 (1%)

Notes: The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

Appendix D. Feature importance

Explainability capabilities of all models PB have been compared using Permutation Feature Importance (PFI) and Shapley Additive Explanations (SHAP). The change in models' performances and the probability correlated to each predictor has been explored to understand the sign of the effect on each class of the target variable.

PFI evaluates the importance of each variable by computing the gain in the model's prediction error after shuffling the feature's values. A feature is considered relevant for model's prediction if the prediction error increases after permuting its values, otherwise, if model error remains unchanged, its contribution is not important. As proposed by Fisher et al. (2018), the algorithm for a generic model f can be defined as:

Algorithm 1: Permutation Feature Importance

Input: Trained model f, feature matrix X, target vector y, performance metric $P(y, f)$

Estimate the original model performance $P_{\text{orig}} = f(y, X)$;
foreach *feature* $j = 1, \ldots, p$ **do**
  Generate feature matrix X_{perm} by permuting feature j in the data X;
  Estimate $P_{\text{perm}} = f(y, X_{\text{perm}})$ based on the predictions of the permuted data;
  Evaluate $\text{PFI}_j = P_{\text{perm}}/P_{\text{orig}}$. Alternatively, the difference can be used: $\text{PFI}_j = P_{\text{perm}} - P_{\text{orig}}$;
  return PFI_j;
end
Sort features by descending PFI

Shapley values represent the marginal contribution of each feature to the prediction of a given data point. The feature values, for instance, x, behave like players in a game where the prediction is the payout. As described in Shapley (1953), the Shapley value Φ_j of a feature value x_j is defined using a value function val of actors in S and represents its contribution to the prediction, weighted and summed across all possible coalitions:

$$\Phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{x_j\}) - val(S))$$

where S denotes a subset of features, x represents the feature values of the instance of interest and p the number of features and $val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in S:

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Estimating the Shapley values for more than a few features becomes computationally infeasible since all possible coalitions of feature values need to be considered with and without feature j. A Monte-Carlo sampling was proposed by Strumbelj and Kononenko (2014):

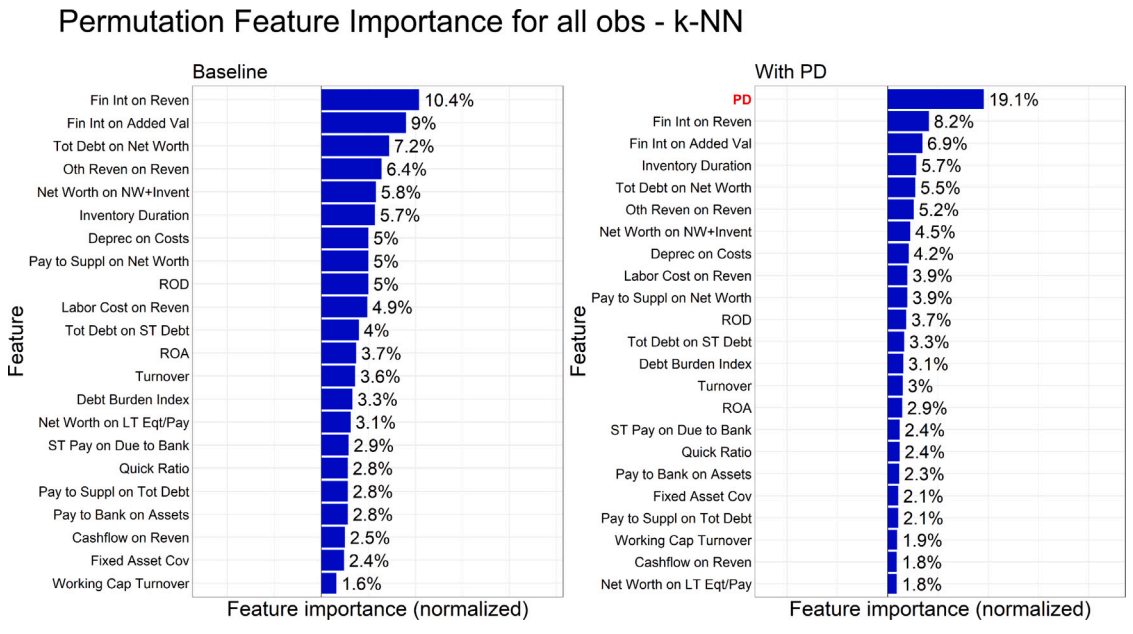
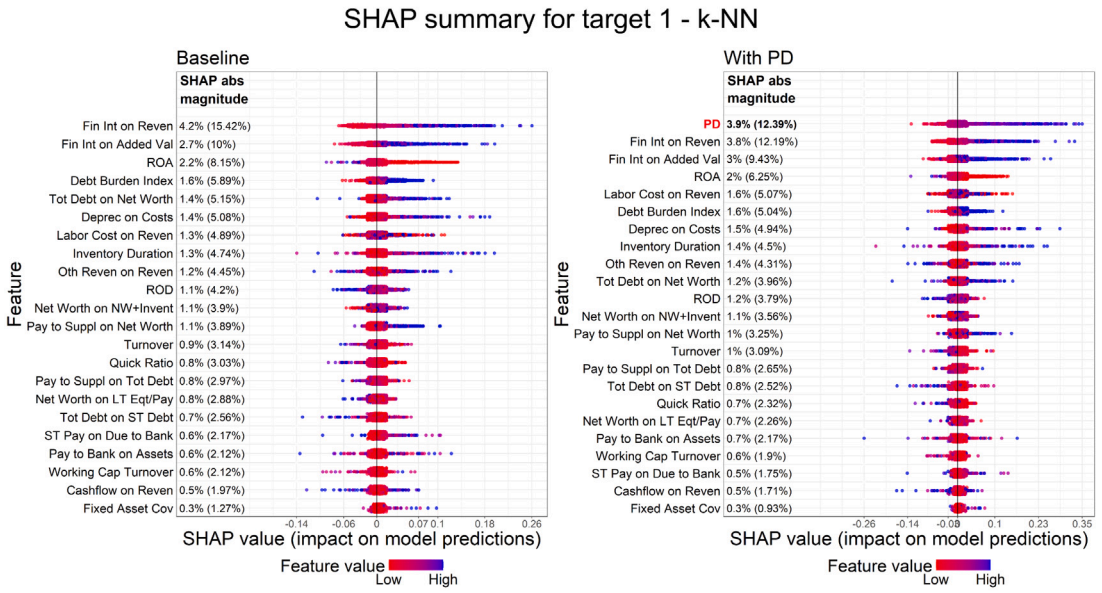
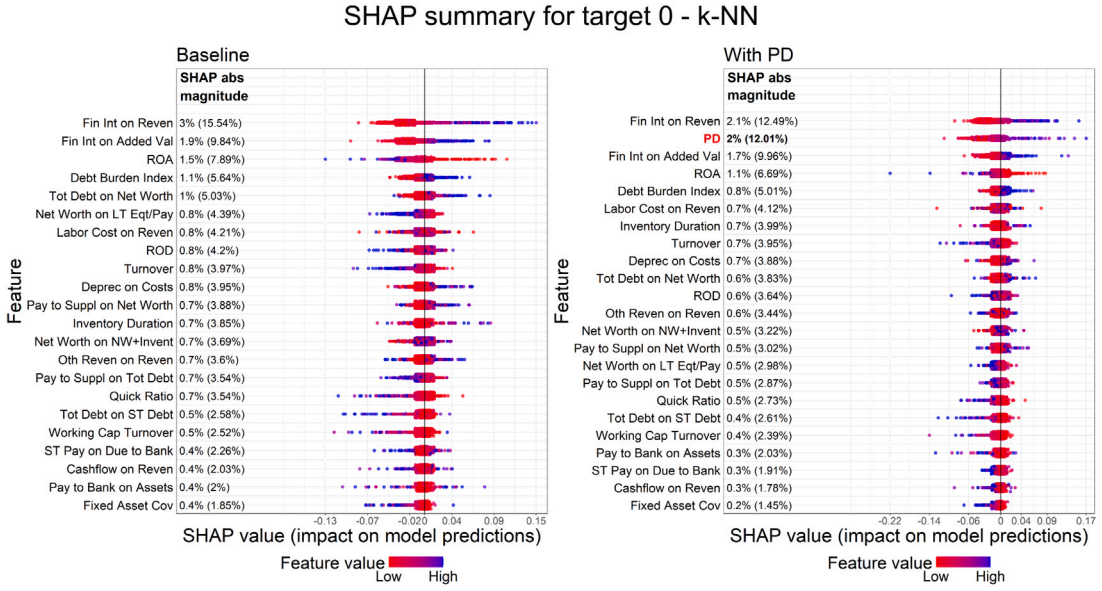


Fig. D.10. Permutation Feature Importance for k-NN model, comparing variable importance of model calibrated with input variables and with the addition of PD. Normalized changes in the F1-score are used to rank the variables.

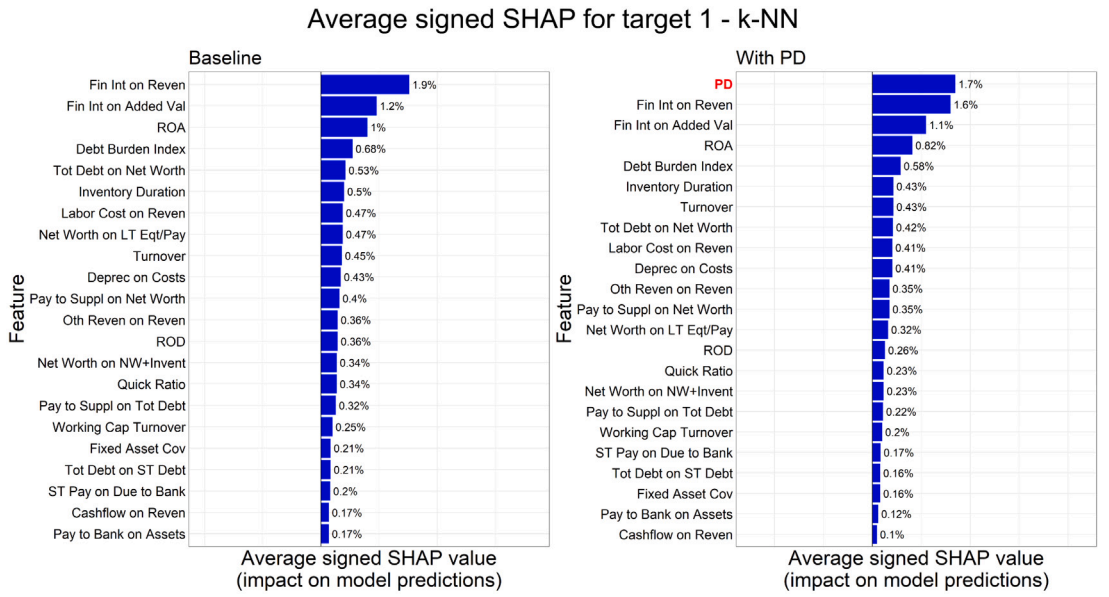


(a) Defaulted clients.

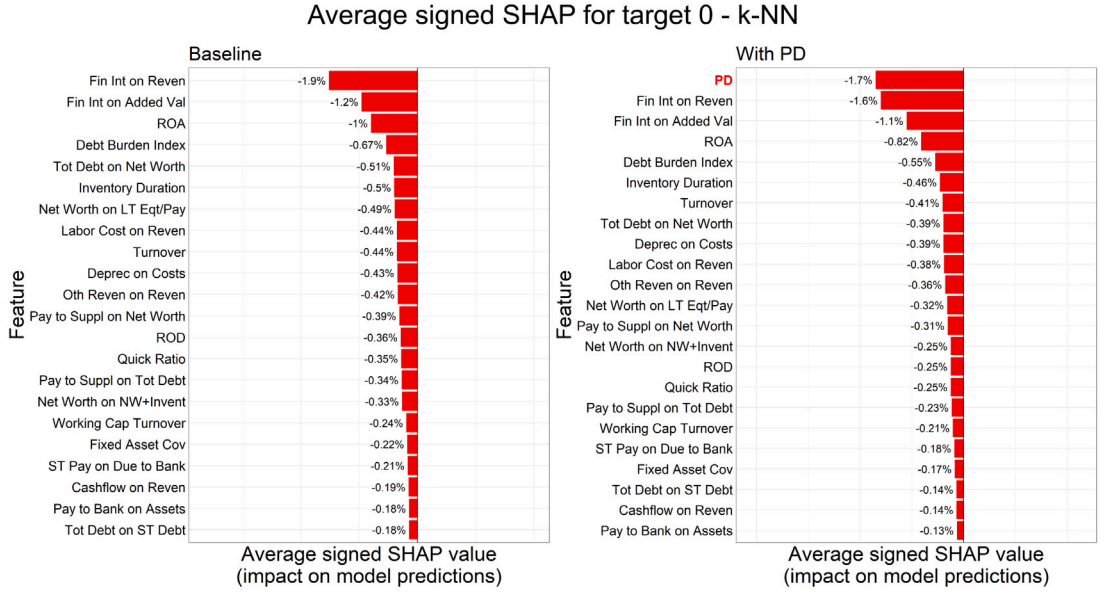


(b) Non-defaulted clients.

Fig. D.11. SHAP effects on predicted probability for k-NN model and defaulted (top) and non-defaulted (bottom) observations only, comparing variable importance of model calibrated with input variables and with the addition of PD. The colour of the points ranges from red, meaning that the observation has a low value for the specific variable, to blue, meaning high values for the same variable. The position on the horizontal axis represents the contribution of the variable in increasing or decreasing the predicted probability of each observation. Values on the left column report the average absolute change in predicted probability over all observations and the normalized values in parentheses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Defaulted clients.



(b) Non-defaulted clients.

Fig. D.12. SHAP average signed effect for k-NN model and defaulted (top) and non-defaulted (bottom) observations only, comparing variable importance of model calibrated with input variables and with the addition of PD. Bars report the average effect of input variables on the predicted probabilities for all observations predicted as 1 and 0, respectively.

$$\hat{\Phi}_j = \frac{1}{M}\sum_{m=1}^{M}(\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m))$$

where $\hat{f}(x_{+j}^m)$ represents the prediction for the instance of interest x but with a random permutation of features (taken from a random data point z) except for jth feature. The vector x_{-j}^m is identical to x_{+j}^m, but the value for feature j is randomized as well from the sampled z. The algorithm for a generic model f can be defined as:

Algorithm 2: Shapley value

Output: Shapley value for the value of the j-th feature
Input : Number of iterations M, instance of interest x, feature index j, data matrix X, and machine learning model f

foreach $m = 1, \dots, M$ **do**
- Draw random instance z from data matrix X;
- Choose a random permutation o of the feature values;
- Order instance x: $x_O = (x_{(1)}, \dots, x_{(j)}, \dots, x_{(p)})$;
- Order instance z: $z_O = (z_{(1)}, \dots, z_{(j)}, \dots, z_{(p)})$;
- Construct two new instances:
  - With feature j: $x_{+j} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots, z_{(p)})$
  - Without feature j: $x_{-j} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots, z_{(p)})$
- Compute marginal contribution: $\Phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$;
- **return** Φ_j^m;

end

Compute Shapley value as the average: $\Phi_j(x) = \frac{1}{M}\sum_{m=1}^{M}\Phi_j^m$

This procedure must be repeated for each feature of interest to get all the Shapley values. Among the advantages of Shapley values over the other methods, in the first place, there is the efficiency property, i.e., the difference between prediction and average prediction is fairly distributed among features.

Figures from D.10 to D.12 report the PFI and SHAP variable importance for k-NN model, calibrated with input variables and with the addition of PD as a predictor.

Data availability

Data will be made available on request.

References

Agarwal, V., Taffler, R., 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. J. Bank. Financ. 32, 1541–1551.

Akbari, A., Ng, L., Solnik, B., 2021. Drivers of economic and financial integration: A machine learning approach. J. Empir. Financ. 61, 82–102.

Albanesi, S., Vamossy, D.F., 2019. Predicting consumer default: A deep learning approach. National Bureau of Economic Research Working Paper 26165.

Alford, A.W., 1992. The effect of the set of comparable firms on the accuracy of the price-earnings valuation method. J. Account. Res. 30, 94–108.

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Financ. 23, 589–609.

Altman, E., Sabato, G., 2007. Modelling credit risk for SMEs: Evidence from the US market. J. Bus. 43, 332–357.

Altman, E., Sabato, G., Wilson, N., 2010. The value of non-financial information in small and medium-sized enterprise risk management. J. Credit. Risk 6.

Andrade, G., Kaplan, S.N., 1998. How costly is financial (not economic) distress? Evidence from highly leveraged transactions that became distressed. J. Financ. 53 (5), 1443–1493. http://dx.doi.org/10.1111/0022-1082.00062.

Andrikopoulos, P., Khorasgani, A., 2018. Predicting unlisted SMEs' default: Incorporating market information on accounting-based models for improved accuracy. Br. Account. Rev. 50, 559–573.

Avramov, D., Li, M., Wang, H., 2021. Predicting corporate policies using downside risk: A machine learning approach. J. Empir. Financ. 63, 1–26.

Baker, M., Ruback, R.S., 1999. Estimating industry multiples. Working Paper Harvard University Cambridge.

Barbaglia, L., Manzan, S., Tosetti, E., 2021. Forecasting Loan Default in Europe with Machine Learning*. J. Financ. Econ..

Bauer, J., Agarwal, V., 2014. Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. J. Bank. Financ. 40, 432–442.

Beaver, W., 1966. Financial ratios as predictors of failure. J. Account. Res. 4, 71–111.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35 (8), 1798–1828. http://dx.doi.org/10.1109/TPAMI.2013.50.

Berger, A., Udell, G., 2002. Small business credit availability and relationship lending: The importance of bank organisational structure. Econ. J. 112 (477), F32–F53.

Beuselinck, C., Elfers, F., Gassen, J., Pierk, J., 2023. Private firm accounting: the European reporting environment, data and research perspectives. Account. Bus. Res. 53 (1), 38–82. http://dx.doi.org/10.1080/00014788.2021.1982670.

Bharath, S., Shumway, T., 2008. Forecasting default with the merton distance to default model. Rev. Financ. Stud. 21, 1339–1369.

Bhimani, A., Gulamhussen, M.A., Lopes, S.D.R., 2010. Accounting and non-accounting determinants of default: an analysis of privately-held firms. J. Account. Public Policy 29, 517–532.

Bitetto, A., Cerchiello, P., Mertzanis, C., 2023a. Measuring financial soundness around the world: A machine learning approach. Int. Rev. Financ. Anal. 85, 102451. http://dx.doi.org/10.1016/j.irfa.2022.102451, URL https://www.sciencedirect.com/science/article/pii/S105752192200401X.

Bitetto, A., Cerchiello, P., Mertzanis, C., 2023b. On the efficient synthesis of short financial time series: A dynamic factor model approach. Financ. Res. Lett. 53, 103678. http://dx.doi.org/10.1016/j.frl.2023.103678, URL https://www.sciencedirect.com/science/article/pii/S1544612323000521.

Blum, M., 1974. Failing company discriminant-analysis. J. Account. Res. 12, 1–25.

Breden, D., 2008. Monitoring the operational risk environment effectively. J. Risk Manag. Financ. Instit. 1, 156–164.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Burgstahler, D., Hail, L., Leuz, C., 2006. The importance of reporting incentives: Earnings management in European private and public firms. Account. Rev. 81, 983–1016.

Byström, H., 2006. Merton unraveled: A flexible way of modeling default risk. J. Altern. Investments 8, 39–47.

Calabrese, R., Marra, G., Osmetti, S., 2016. Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. J. Oper. Res. Soc. 67, 604–615.

Campbell, J.Y., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. J. Financ. 63 (6), 2899–2939.

Candes, E.J., Li, X., Ma, Y., Wright, J., 2009. Robust principal component analysis?

Carter, R., Auken, H.V., 2006. Small firm bankruptcy. J. Small Bus. Manag. 44 (4), 493–512. http://dx.doi.org/10.1111/j.1540-627X.2006.00187.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-627X.2006.00187.x.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 1724–1734.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. http://dx.doi.org/10.1007/BF00994018.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13 (1), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964.

Das, S., Hanouna, P., Sarin, A., 2009. Accounting-based versus market-based cross-sectional models of CDS spreads. J. Bank. Financ. 33, 719–730.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 224–227.

Day, N.E., 1969. Estimating the components of a mixture of normal distributions. Biometrika 56 (3), 463–474. http://dx.doi.org/10.1093/biomet/56.3.463, arXiv:https://academic.oup.com/biomet/article-pdf/56/3/463/635460/56-3-463.pdf.

Dierkes, M., Erner, C., Langer, T., Norden, L., 2013. Business credit information sharing and default risk of private firms. J. Bank. Financ. 37, 2867–2878.

Doumpos, M., Niklis, D., Zopounidis, C., Andriosopoulos, K., 2015. Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms. J. Bank. Financ. 50, 599–607.

Edminster, R., 1972. An empirical test of financial ratio analysis for small business failure prediction. J. Financ. Quant. Anal. 7, 1477–1493.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman & Hall.

Falkenstein, E., Boral, A., Carty, L., 2000. RiskCalcTM for private companies: Moody's default model. Moody' s Invest. Serv..

Filomeni, S., Bose, U., Megaritis, A., Triantafyllou, A., 2024. Can market information outperform hard and soft information in predicting corporate defaults?. Int. J. Finan. Econom. 29 (3), 3567–3592.

Filomeni, S., Udell, G.F., Zazzaro, A., 2021. Hardening soft information: does organizational distance matter? Eur. J. Financ. 27 (9), 897–927.

Fiordelisi, F., Monferrà, S., Sampagnaro, G., 2014. Relationship lending and credit quality. J. Financ. Serv. Res. 46, 295–315.

Fisher, A., Rudin, C., Dominici, F., 2018. Odel class reliance: Variable importance measures for any machine learning model class, from the 'Rashomon' perspective. URL http://arxiv.org/abs/1801.01489.

Foglia, A., Laviola, S., Reedtz, P.M., 1998. Multiple banking relationships and the fragility of corporate borrowers. J. Bank. Financ. 22, 1441–1456.

Friedman, J., Hastie, T., Tibshirani, R., 2009. The Elements of Statistical Learning. Springer.

Gracia, A., Gonzalez, S., Robles, V., Menasalvas, E., 2014. A methodology to compare dimensionality reduction algorithms in terms of loss of quality. Inform. Sci. 270, 1–27.

Grice, J., Ingram, R., 2001. Tests of the generalizability of Altman's bankruptcy prediction model. J. Bus. Res. 54, 53–61.

Gropp, R., Guettler, A., 2018. Hidden gems and borrowers with dirty little secrets: Investment in soft information, borrower self-selection and competition. J. Bank. Financ. 87, 26–39.

He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 1322–1328. http://dx.doi.org/10.1109/IJCNN.2008.4633969.

Hernandez Tinoco, M., Holmes, P., Wilson, N., 2018. Polytomous response financial distress models: The role of accounting, market and macroeconomic variables. Int. Rev. Financ. Anal. 59, 276–289. http://dx.doi.org/10.1016/j.irfa.2018.03.017, URL https://www.sciencedirect.com/science/article/pii/S1057521918302114.

Hernandez Tinoco, M., Wilson, N., 2013. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. Int. Rev. Financ. Anal. 30, 394–419.

Hillegeist, S., Keating, E., Cram, D., Lundstedt, K., 2004. Assessing the probability of bankruptcy. Rev. Account. Stud. 9, 5–34.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507. http://dx.doi.org/10.1126/science.1127647, arXiv:https://www.science.org/doi/pdf/10.1126/science.1127647.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6 (2), 65–70, URL http://www.jstor.org/stable/4615733.

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., Castrén, E., 2003. Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics 4 (1), 48. http://dx.doi.org/10.1186/1471-2105-4-48.

Keasey, K., Watson, R., 1987. Non-financial symptoms and the prediction of small company failure: A test of Argenti's hypotheses. J. Bus. Financ. Account..

Kim, H., Cho, H., Ryu, D., 2020. Corporate default predictions using machine learning: Literature review. Sustainability 12 (16).

Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 37, 233–243.

Kucher, A., Mayr, S., Mitter, C., Duller, C., Feldbauer-Durstmüller, B., 2020. Firm age dynamics and causes of corporate bankruptcy: age dependent explanations for business failure. Rev. Manag. Sci. 14, 633–661.

Liberti, J., Petersen, M., 2017. Information: Hard and soft. Rev. Corp. Financ. Stud. 8, 1–41.

Louzada, F., Ara, A., Fernandes, G., 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. Surv. Oper. Res. Manag. Sci. 21, 117–134.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 2522–5839.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297.

Mayr, S., Mitter, C., Kücher, A., Duller, C., 2021. Entrepreneur characteristics and differences in reasons for business failure: evidence from bankrupt Austrian SMEs. J. Small Bus. Entrep. 33 (5), 539–558. http://dx.doi.org/10.1080/08276331.2020.1786647.

McCarthy, E., 1999. Pricing IPOs: Science or science fiction? J. Account. 188, 51–58.
McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv E-Prints, arXiv:1802.03426.
McInnes, L., Healy, J., Saul, N., Grossberger, L., 2018. UMAP: Uniform manifold approximation and projection. J. Open Source Softw. 3 (29), 861.
Merton, R.C., 1974. On the pricing of corporate debt: The risk structure of interest rates. J. Financ. 29, 449–470.
Modigliani, F., Miller, M., 1958. The cost of capital, corporate finance, and the theory of investment. Am. Econ. Rev. 3 (48), 261–297.
Moscatelli, M., Narizzano, S., Parlapiano, F., Viggiano, G., 2019. Corporate default forecasting with machine learning. Bank Italy Work. Pap. 1256.
Mselmi, N., Lahiani, A., Hamza, T., 2017. Financial distress prediction: The case of French small and medium-sized firms. Int. Rev. Financ. Anal. 50, 67–80. http://dx.doi.org/10.1016/j.irfa.2017.02.004, URL https://www.sciencedirect.com/science/article/pii/S1057521917300236.
Mullainathan, S., Spiess, J., 2017. Machine learning: An applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.
Norden, L., Weber, M., 2010. Credit line usage, checking account activity, and default risk of bank borrowers. Rev. Financ. Stud. 23, 3665–3699.
Ohlson, J., 1980. Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. 18, 109–131.
Olson, L.M., Qi, M., Zhang, X., Zhao, X., 2021. Machine learning loss given default for corporate debt. J. Empir. Financ. 64, 144–159.
Ooghe, H., De Prijcker, S., 2008. Failure processes and causes of company bankruptcy: a typology. Manag. Decis. 46 (1-2), 223–242. http://dx.doi.org/10.1108/00251740810854131.
Osborne, M.J., Rubinstein, A., 1994. A Course in Game Theory. MIT Press.
Peel, M., Peel, D., Pope, P., 1986. Predicting corporate failure - some results for the UK corporate sector. Omega 14, 5–12.
Pindado, J., Rodrogues, L., de la Torre, C., 2008. Estimating financial distress likelihood. J. Bus. Res. 61, 995–1003.
Qian, J., Strahan, P.E., Yang., Z., 2015. The impact of incentives and communication costs on information production and use: Evidence from bank lending. J. Financ. 70, 1457–1493.
Rikkers, F., Thibeault, A.E., 2009. A structural form default prediction model for SMEs, evidence from the Dutch market. Multinatl. Financ. J. 13, 229–264.
Rodriguez Gonzalez, M., Basse, T., Kunze, F., 2018. Early warning indicator systems for real estate investments: Empirical evidence and some thoughts from the perspective of financial risk management. ZVersWiss 107, 387–403.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. 53–65.
Shapley, L.S., 1953. A value for n-person games. pp. 307–317, Contributions to the Theory of Games 2.28.
Shumway, T., 2001. Forecasting bankruptcy more accurately: a simple Hazard model. J. Bus. 74, 101–124.
Stickney, C.P., Weil, R.L., 1997. Financial Accounting: An Introduction to Concepts, Methods, and Uses. Dryden Press Series in Accounting.
Strumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41. 3 647–665.
Tian, S., Yu, Y., Guo, H., 2015. Variable selection and corporate bankruptcy forecasts. J. Bank. Financ. 52, 89–100.
Vassalou, M., Xing, Y., 2004. Default risk in equity returns. J. Financ. 59, 831–868.
Venna, J., Kaski, S., 2006. Local multidimensional scaling. Neural Netw. 19 (6), 889–899. http://dx.doi.org/10.1016/j.neunet.2006.05.014, URL https://www.sciencedirect.com/science/article/pii/S0893608006000724. Advances in Self Organising Maps - WSOM'05.
Volk, M., 2012. Estimating probability of default and comparing it to credit rating classification by banks. Econ. Bus. Rev. 14, 299–320.
Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K., 2019. A study on predicting loan default based on the random forest algorithm. Procedia Comput. Sci. 162, 503–513.