ISSN: 2281-1346

UNIVERSITÀ DI PAVIA
Department of Economics and Management

DEM Working Paper Series

Understanding corporate default using Random Forest: The role of accounting and market information

Alessandro Bitetto
(University of Pavia)

Stefano Filomeni
(University of Essex)

Michele Modina
(University of Molise)

 # 205 (10-21)

Via San Felice, 5
I-27100 Pavia

economiaweb.unipv.it

Understanding corporate default using Random Forest: The role of accounting and market information

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

[a] *University of Pavia, Italy*
[b] *University of Essex, Essex Business School, Finance Group, Colchester (UK)*
[c] *University of Molise, Italy*

Abstract

Recent evidence highlights the importance of hybrid credit scoring models to evaluate borrowers' creditworthiness. However, the current hybrid models neglect to consider the role of public-peer market information in addition to accounting information on default prediction. This paper aims to fill this gap in the literature by providing novel evidence on the impact of market information in predicting corporate defaults for unlisted firms. We employ a sample of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) that borrow from 113 cooperative banks from 2012–2014 to examine whether market pricing of public firms adds additional information to accounting measures in predicting default of private firms. Specifically, we estimate the probability of default (PD) of MSMEs using equity price of size-and industry- matched public firms, and then we adopt advanced statistical techniques based on parametric algorithm (Multivariate Adaptive Regression Spline) and non-parametric machine learning model (Random Forest). Moreover, by using Shapley values, we assess the relevance of market information in predicting corporate credit risk. Firstly, we show the predictive power of Merton's PD on default prediction for unlisted firms. Secondly, we show the increased predictive power of credit risk models that consider both the Merton's PD and accounting information to assess corporate credit risk. We trust the results of this paper contribute to the current debate on safeguarding the continuity and the resilience of the banking sector. Indeed, banks' hybrid credit scoring methodologies that also embed market

*Corresponding author
Email address: alessandro.bitetto@unipv.it (Alessandro Bitetto)

information prove to be successful to assess credit risk of unlisted firms and could be useful for forward-looking financial risk management frameworks.

Keywords: Default Risk, Distance to Default, Machine Learning, Merton model, SME, PD, SHAP, Autoencoder, Random Forest, XAI

JEL: C52, C53, D82, D83, G21, G22

---

1. Introduction

What are the factors affecting corporate default risk? The aim of banks' core business, as in their intrinsic nature, is to perform accurate assessment of borrowers' capability to repay their debt. This activity is performed by collecting information about a given borrower from different sources. The type of information a bank should use when assessing credit risk has been a matter of concern for policy makers since inaccurate credit risk measurement could threaten the stability of the banking sector and undermine the pivotal intermediation role played by banks in the economy. In this regard, banks' need to implement reliable credit risk models to timely and precisely forecast business failure is imperative to reach appropriate lending decisions and, eventually, to engage in corrective action.

When focusing on the predictions of default risk of micro-, small- and mid-sized enterprises (MSMEs), it is crucial to adopt a credit risk assessment model that takes into account their peculiarities in order to provide a reliable prediction of default. Indeed, MSMEs have specific characteristics which are not similar to those of larger firms on which the existent literature on default prediction modeling has mainly been based (Norden and Weber, 2010, Peel et al., 1986, Hol, 2007). In this regard, MSME lending suffers from more severe agency problems, exhibits higher default risk, has lower accounting quality and is more informationally opaque (Burgstahler et al., 2006). Given the importance of MSMEs for market economies, it is imperative to implement credit assessment models specifically addressed to MSMEs with the objective to minimize expected and unexpected losses as accurately as possible. To this purpose, the objective of this paper is to develop a credit risk model for MSMEs that takes into account, in addition to accounting measures, market information obtained from comparable publicly listed companies. Motivated by the findings of the relevant literature on the assessment and forecasting of corporate default risk, we exploit a unique and proprietary dataset comprising over 10,136 Italian firms and their 113 co-operative banks over the period 2012–2014 to estimate multivariate forecasting models on the incidence of corporate default by using both market and accounting information. Given the unlisted nature of our Italian micro-, small, and mid-sized enterprises (MSMEs), we estimate the Merton's Probabil-

ity of Default (PD) based on market information obtained from those publicly listed and deemed as comparable by a data-driven clustering approach, avoiding any a-priori assumption of mapping by size, industry and number of employees[1].

The paper contributes to the literature along two dimensions. Firstly, our hybrid credit scoring models, which use a combination of market and accounting information, provide better default predictions for unlisted firms when compared with the respective predictive power of models which only use accounting or market information. Although several papers have applied Merton model to private companies (Rikkers and Thibeault, 2009, Andrikopoulos and Khorasgani, 2018, Falkenstein et al., 2000), our study, to the best of our knowledge, represents the first attempt to introduce a new credit risk modeling approach that encompasses market information for predicting corporate defaults of unlisted companies. Indeed, we show that the estimated Merton's Probability of Default (PD) credit risk measure has incremental predictive power on corporate default when added into a multivariate predictive regression model which already includes accounting information. A possible economic interpretation of this finding relies on the fact that the nature of MSMEs as quite risky companies with a lot of time variation and sudden shifts in their risk profile (Islam and Tedford, 2012) leads market information to better capture the corporate default risk of these firms. In fact, market data respond more quickly to new information about borrowers' creditworthiness when compared to more sluggishly-responding accounting measures. In support of this claim, Islam and Tedford (2012) finds that SMEs usually face three types of risks: operational, occupational and economic. The first involves the loss of production and its capability, the second comprises the risks associated with employees' health, safety and well-being, and the third is affected by financial penalties resulting from the first two as well as compensation claims and damage to reputation. Monetary factors and accounting measures alone can ignore many issues that impact the long-term competitive advantage of the company and the reasons that can lead to business demise. Furthermore, there are multiple internal and external causes contributing to failure and none of them seem

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

to dominate in leading the firm to default. The former encompass poor business competencies, high cost pressure, fraud, poor quality of products and services and private domain, whereas the latter involve government policies, natural disaster, global economic downturns and increased industry competition (Kucher et al., 2020). As explained in Islam (2008), all these risks can cause loss of market share and eventually put the organization out of business. Following the advice of Virdi (2005) on the importance for MSMEs of managing risk in a professional and structured way, our hypothesis relies on the importance of market volatility in capturing the effects of the aforementioned risks. In particular, employees' dissatisfaction and strikes are usually reported by media and may affect the firm's reputation causing shocks on their stock prices. Similarly, internal policy, low level of enterprise culture and dubious choices of management board can reduce the stakeholder and shareholder trust resulting in stock downturn. Moreover, shortage of goods and machinery breakdown and national or international government policies can lead to analogous effects. Although the data we use are annual, short-term shocks or rare events can still be captured by asset volatility, regardless of the overall price trend as well. We therefore conclude that the assets volatility can be a reliable proxy for several types of risk and can be accurately mapped through a panel of representative peers spanning over different industrial sectors and firm sizes.

The second dimension involves the implementation of predictive models and their explainability. In the last years, many works on the application of Machine Learning (ML) model to economic problems have been published (Mullainathan and Spiess, 2017, Akbari et al., 2021, Avramov et al., 2021, Olson et al., 2021). Specifically, Kim et al. (2020) report a survey on ML applications to credit default prediction. Generally, linear classification models, such as linear discriminant analysis (LDA) or logistic regression (Shumway, 2001, Altman and Sabato, 2007, Bauer and Agarwal, 2014, Tian et al., 2015), show lower prediction ability rather than non-linear and non-parametric models, such as Random Forest (RF) or Boosted Trees (BT) (Zhu et al., 2019, Barbaglia et al., 2021). However, most of these studies restrict their analysis on the increase of performances compared to linear models and do not investigate the relevance of the input variables and their effects on the predictions. An attempt on the explanation of the overall importance of the input variables

can be found in Moscatelli et al. (2019). Moreover, Barbaglia et al. (2021), Albanesi and Vamossy (2019) show an increasing attention on local explanation of predictions, i.e. understanding how each variable's value can impact the predicted outcome, justified by the need of a more deep investigation of the complex non-linear correlations captured by the advanced models. Our work contributes to this new stream of research (usually referred to eXplainable Artificial Intelligence) because we implement both a non-linear parametric and non-parametric ML algorithms and we go beyond the prediction of corporate defaults and implement an advanced methodology that involves the use of two state-of-the-art techniques so to evaluate the importance of the variables on the predictions: Permutation Feature Importance (Fisher et al., 2018) explains the overall variables' relevance, whereas Shapley Additive Explanations (Lundberg et al., 2020) provide the contribution of each variable's values to the predicted probability of default for a single observations. In addition, we implement a sophisticated clustering technique that, to the best of our knowledge, is the first application of Artificial Neural Networks to compress the information of financial ratios so to map unlisted MSMEs to a pool of listed ones.

The studies of Falkenstein et al. (2000), Rikkers and Thibeault (2009) and Andrikopoulos and Khorasgani (2018) are closely related to our work, providing evidence that hybrid models which incorporate both market and accounting information have significant predictive power on corporate defaults of unlisted firms. Despite our findings are supportive of the results reached in those studies, our study differs in several aspects. Firstly, in contrast with Falkenstein et al. (2000) and Rikkers and Thibeault (2009), we derive the market value of unlisted firms by collecting market data from comparable publicly-listed companies that operate in the same industry matched in terms of size, industry, and number of employees by a data-driven approach ("comparable approach") rather than by KMV's private firm model which makes use of the industry average market value of equity. The latter has the drawback of being exposed to considerable variation over time (Falkenstein et al., 2000) or by the present value methodology of cash flows for valuing a company (Rikkers and Thibeault, 2009, Chen and Liao, 2005, J.F. et al., 2001). The rationale behind choosing a "comparable approach" stems from the benefits highlighted by the extant literature in the field of

corporate finance when adopting a comparability method in equity valuation of private unlisted companies using industry-level data (Andrikopoulos and Khorasgani, 2018, Baker and Ruback, 1999, Alford, 1992, McCarthy, 1999). Secondly, differently from Rikkers and Thibeault (2009), we exploit a unique proprietary dataset comprising granular data on a sample of 10,136 Italian unlisted MSMEs operating within 113 Italian co-operative credit banks over the period 2012-2014, rather than by relying on data collected from a single bank. Lastly, our sample comprises Italian unlisted companies operating in the manufacturing sector, as in Andrikopoulos and Khorasgani (2018), and also in the service industry.

One policy implication resulting from our findings is that banks can potentially integrate their hybrid credit scoring methodologies with market information for credit risk assessments, with the purpose of increasing the accuracy of forecasting corporate defaults for unlisted firms. This would allow banks to expand the spectrum of information used in credit risk measurements helping them to enhance their internal hybrid credit scoring by including both accounting and market information on the credit quality of a given borrower. Thus, results reported in this paper could be very helpful for forward-looking financial risk management frameworks (Breden, 2008, Rodriguez Gonzalez et al., 2018).

The remainder of this paper is organized as follows. Section 2 briefly reviews the relevant literature. Section 3 discusses the data and Section 4 presents the econometric methodology. Section 5 illustrates the empirical results and Section 6 concludes.

2. Determinants of corporate default: a review of the literature

Up to date, most of the literature on corporate credit risk modeling has focused on both accounting-based approaches and structural market-based models.

2.1. Accounting-based models

The power of accounting-based techniques to predict default risk has been already widely explored by the extant literature. Such methodologies include all the statistical techniques that elaborate quantitative information about a borrower, i.e. financial ratios and statement data, into a

numerical score reflecting the credit quality and predictive of the default probability of the borrower itself (Beaver, 1966, Altman, 1968, Ohlson, 1980, Edminster, 1972, Blum, 1974, Grice and Ingram, 2001, Pindado et al., 2008, Louzada et al., 2016). In this regard, there are several studies investigating the efficacy of accounting-based credit scores to predict future corporate default. Calabrese et al. (2016) show that accounting-based information has significant in-sample and out-of-sample forecasting power on the timing of bankruptcy of Italian MSMEs over the period 2006-2011. Beaver (1966) examines whether financial ratios predict subsequent business bankruptcy. Altman and Sabato (2007) analyze the most relevant variables in forecasting the company's future credit quality and construct a default prediction model. Peel et al. (1986) expand the variable set used in forecasting default and find that non-conventional variables computed from UK companies' annual reports and statements can significantly enhance the predictive power of more conventional models. Keasey and Watson (1987) implement a similar model to predict small business failure. Charalambous et al. (2004) apply logistic regressions and neural networks to develop bankruptcy forecasting models and assess the power of cash flow variables to forecast UK corporate default. Lin et al. (2012) test a number of different accounting-based risk models to predict UK small business bankruptcy. Overall, these models seem to suggest that the more relevant accounting variables are added to the model, the better its corporate default forecasting power becomes. These models remain the most widely used methodology for the prediction of default of unlisted companies, although they exhibit some disadvantages. The main disadvantages are the fact that ratios might correlate with each other affecting the estimates, and when using comparative ratio analyses, differences between firms in terms of methods of accounting or methods of operations have to be taken into account (Stickney and Weil, 1997). In this regard, we attempt to overcome those disadvantages by adopting a structural form model approach, i.e., the Merton's model, whose principle is that a firm defaults if the company's asset value falls below the default boundary. Moreover, the models we use can deal with multicollinearity between input variables and thus provide robust estimations.

However, another stream of literature suggests that credit risk models which incorporate both

accounting and non-accounting information, i.e. creditors' legal action and audit reports, have a higher predictive power of corporate default risk (Altman et al., 2010). In this regard, Bhimani et al. (2010) provide further support to the findings in Altman et al. (2010) by showing that, for a large sample of privately-owned Portuguese firms, accounting and non-accounting information is a significant predictor of corporate default. Along the same lines, Dierkes et al. (2013), by examining the forecasting performance of credit models on the default risk of privately-owned small firms, find that non-accounting business information improves default forecasting performance when added into the information variable set of default predictive models that use only accounting information. Fiordelisi et al. (2014) find that, especially for small firms, bank-firm relationship information is a significant determinant of corporate default risk (Volk, 2012, Qian et al., 2015). In a similar analysis aimed at assessing corporate default risk of Italian firms, Foglia et al. (1998) show that multiple bank-firm relationships increase the likelihood of default. Moreover, Norden and Weber (2010) find that borrowers' checking account activity and credit line usage are significant early warning signals of default for a sample of German firms. By using loan-level data from a large Chinese bank, Qian et al. (2015) show that delegation of authority to line units increases the predictive power of internal ratings on borrowers' credit risk. Greater accountability to loan officers increases soft information production and affects the effort of the loan officer in evaluating a borrower's credit risk. This increased production of soft information is likely to be hardened into the bank's internal ratings which result in scores having a stronger effect on the terms of loan contracts and predictive power on ex-post loan outcomes (Gropp and Guettler, 2018, Liberti and Mian, 2009, Liberti and Petersen, 2017, Brown et al., 2012).

2.2. Market-based models

In addition to the aforementioned accounting-based models, structural market-based models in the form of Merton-type models have been applied to predict corporate bankruptcy. Specifically, the Merton's Distance to Default (DD) model, which is based on observable equity market data, has been extensively used for the estimation and prediction of corporate default risk for listed firms in the US equity market (Bharath and Shumway, 2008, Byström, 2006, Vassalou and Xing, 2004).

Moreover, other studies have compared the corporate default predictive power of accounting- and market-based models. Overall, these empirical studies have extensively shown the incremental predictive information content of the Merton model on corporate default predictions with respect to alternative models primarily based on accounting ratios like the Altman (1968) z-score model (Agarwal and Taffler, 2008, Altman, 1968, Bauer and Agarwal, 2014, Das et al., 2009, Doumpos et al., 2015, Hillegeist et al., 2004). Among these, Das et al. (2009) show that the accounting-based models of corporate default perform similarly to the market-based models of credit risk assessments when it comes to predict the distress risk of an international sample of 2860 quarterly Credit Default Swap (CDS) spreads, while Doumpos et al. (2015), when forecasting the bankruptcy risk of a panel of European listed firms for the period 2002-2012, find that the inclusion of the Merton's market-based Distance-to-Default (DD) measure into the information variable set which is composed only by accounting-based financial ratios, adds significant predictive information content. In a similar vein, Tinoco and Wilson (2013) find that corporate default risk can be more accurately predicted when models simultaneously incorporate accounting, stock-market and macroeconomic information, rather than using different types of information in isolation. However, those studies do not integrate, but rather compare, accounting with market information in forecasting corporate default.

3. Data

We use two sources of information for our analysis: a proprietary one, consisting of granular information on over 10,136 Italian unlisted micro-, small, and mid-sized enterprises (MSMEs)s, and a public one, comprising data on comparable publicly listed companies, i.e., 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 less than 250 employees and revenue at most of 50 million. We selected a subset of 22 financial ratios out of 30 removing the ones showing

high partial correlation with many other ratios. Therefore, some ratios with not so low correlation with at most one other ratio are still kept because the models we use for the predictions are robust to multicollinearity. Tables 1 and A.1 in the Appendix report the complete list of variables with description and statistics and their pairwise correlations, respectively. The target variable we want to predict is a binary flag indicating whether the firm defaults (1) or not (0). In our context, the flag of default is assigned when the client becomes insolvent in the last 12 months following loan disbursement, and with a past due of at least 180 days. Moreover, we control for additional categorical variables, describing time-invariant characteristics of our unlisted firms, such as the region to which the firm belongs, firm size and industry. Table 2 reports the list of control variables used in the analysis and their distribution over the two target classes.

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 - Labor Cost on Reven	Labor 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 duration	0.78	1.09	0.02	0.03	0.5	2.68	7.16
13 - Quick Ratio	Quick ratio	1.41	1.1	0.04	0.22	1.18	3.42	6.54
14 - Debt Burden Index	Debt burden index	0.4	0.38	0.01	0.02	0.23	1	1
15 - Fin Int on Reven	Financial inrerest on revenues	0.02	0.02	0	0	0.02	0.08	0.1
16 - Fin Int on Added Val	Financial inrerest 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	Working capital turnover	2.3	2.25	0.25	0.55	1.77	5.78	18.32
22 - Turnover	Turnover normalized by Total Assets	1.16	0.74	0.07	0.2	1.02	2.84	3.17

3.2. Peers data

We select a panel of 40 Italian listed firms, evenly distributed in manufacturing and services sector. We collect accounting figures from Orbis database, developed by Bureau Van Dijk (a

Table 2
List of control variables for MSMEs dataset.

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%							

Moody's analytics company), by matching the VAT code for each given peer firm[2]. 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 Refinitiv Eikon database and are used to compute the annual assets 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. Figure A.1 in the Appendix depicts the comparison of the 22 variables between the two datasets, showing that the selected peers are adequately representative of our sample of unlisted micro-, small, and mid-sized enterprises (MSMEs).

4. Methodology

The aim of this paper is to assess the impact of market information, i.e., the Merton's probability of default (PD), in predicting 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

[2]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, properly integrated with additional accounting data collected from Orbis database.

following the 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 variables as predictors (baseline) and including the PD (extended). Section 4.3 shows the calibration of the models and the differences of models' performance between the baseline and extended cases. Lastly, we investigate which predictor contributed the most to predict corporate default, by the means of 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.

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, the corporate default takes place when the company is unable to pay off its debts, or when the current market of assets falls below the market value of liabilities. For this reason, the market value of equity of the MSME is treated as a call option on the asset value of the MSME with strike price equal to the market value of debt[3]. The MSME asset value process follows a Geometric Brownian motion as shown in Equation (1) below:

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

where A_t is the firms market value of assets and σ_A is the volatility of assets. r is 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[4]. Since the market value of equity is treated as a call option, the company's equity value at maturity (which is the end of each yearly period in our model), the company's equity E_t at maturity (at the end of each yearly period in our model) is priced as shown

[3]For modeling 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.

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

below:

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

where A_t is firm's assets and L is firm's liabilities (which are assumed to be constant for each yearly period). T is the time to maturity which in our model is equal to one year equal to 1 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 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 Equations (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 Equation (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 Equation (7) below:

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

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

4.2. Matching unlisted firms with peers

Since there are no market data available for our sample of unlisted MSMEs, we proxy the market value of equity of unlisted MSMEs with those of their comparable publicly-listed companies. As for the latter, the market value of equity 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 the Merton's Probability of Default (PD) and Distance-to-Default (DD) is that those MSMEs which have similar size, number of employees, and industry sector with our Italian peers share the same risk profile and belong to the same (market) risk class of the latter[5]. In order to render the matching procedure as accurate as possible, we opt for a clustering approach: we find the optimal number of clusters in the MSME dataset and then we assign each peer to the most similar cluster by minimizing the average distance from all firms in the cluster.

Finally, we provide each MSME observation with its respective PD. As described in Section 4.1, PD can be calculated with Equation (7) after evaluating DD with Equation (6), using the total assets A, the total liabilities L and the assets volatility σ_A. We evaluate the PD with two approaches. In the first (named *average-PD*), with evaluate the average $\bar{A}$, $\bar{L}$ and $\bar{\sigma_A}$ over all peers in the same cluster for each year and use them to evaluate the average DD. So, we have $kx3$ different DD values, one for each year-cluster pair, where k is the optimal number of clusters. The DD is then matched with each MSME observation by year-cluster. In the second approach (named *pointwise-PD*), we use each MSME observation's A and L and the average year-cluster peers $\bar{\sigma_A}$ to have a firm-year level DD.

4.3. Prediction of default

After assigning the PD to all our unlisted MSMEs, we calibrate three different models to predict the binary target, (1) for defaulted firm and (0) otherwise. Each model is calibrated with 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. Figure A.2 in the Appendix

[5]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'.

shows similar behavior of the input variables for both subset 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 in the entire distribution of input variables. Moreover, the true relationship between input and target variable 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. So, the model can be seen as an ensemble of sub-models that are estimated in each combination of partitions in which input variables can be divided. For example, if we split the input domain into quartiles of each variable, 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, we test also a non-parametric model, the Random Forest (RF) (Breiman, 2001). RF is an ensemble of decision trees that partition the input domain with nested binary splitting aiming to maximise 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 noticed in Section 3, the presence of few variables with moderate correlation won't 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. 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 the means of Bayesian Optimization with a 5-fold Stratified Cross-Validation[6] performance estimation. Given the imbalanced nature of the data, as described in Section 3, we use the F1-score as a class-specific performance metric, so to highlight the importance of predicting the rarest 1-labelled targets, and the Area Under the Receiver Operating Characteristic Curve (AUC) as an overall performance metric. Moreover, each model has been calibrated with 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[7]. Furthermore, we include control variables in both *baseline* and *extended* case in order to assess models' robustness to time-invariant characteristics of the observations. Finally, we investigate the persistence of target values over time, i.e. we examine the impact of clients that 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. In order to assess the impact of this phenomenon, we compare the distribution of the input variables subject to clients' behavior, 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. Figure A.3 in the Appendix shows the distribution of relative changes over the years of each input variable splitted by clients' behavior, i.e., clients that are persistent over time and clients that do not exhibit such a behavior.

4.4. Importance of variables

We explore which input variable contributed most in each model predictions, 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 i-th variable by comparing the performance, e.g. F1-score, of the model that predicts the observations used for

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

[7]This means that the model has similar performances on both data used for calibration and new observations

the calibration against the performance of the model that predicts the same observations where the values of the *i*-th column are shuffled (Fisher et al., 2018). In this way the correlation between the *i*-th 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 *i*-th variable is not important 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 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 of the predicted value of a single observation, comparing the prediction of all possible combinations of variables that include the *i*-th variable against the ones that don't. The differences are then averaged and the positive or negative change in the prediction is used as 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 of a global performance. SHAP, on the other way, provides a local measure of importance, measuring the impact of variables for every single observation. However, taking the average of the absolute values of each observation's SHAP, it is still possible the 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. Appendix C illustrates both techniques in details.

5. Results

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 hold robust. Figure 1

shows the distribution of PDs compared with the corresponding target values. PD seems to be a reliable indicator for the outcome of the target variable.

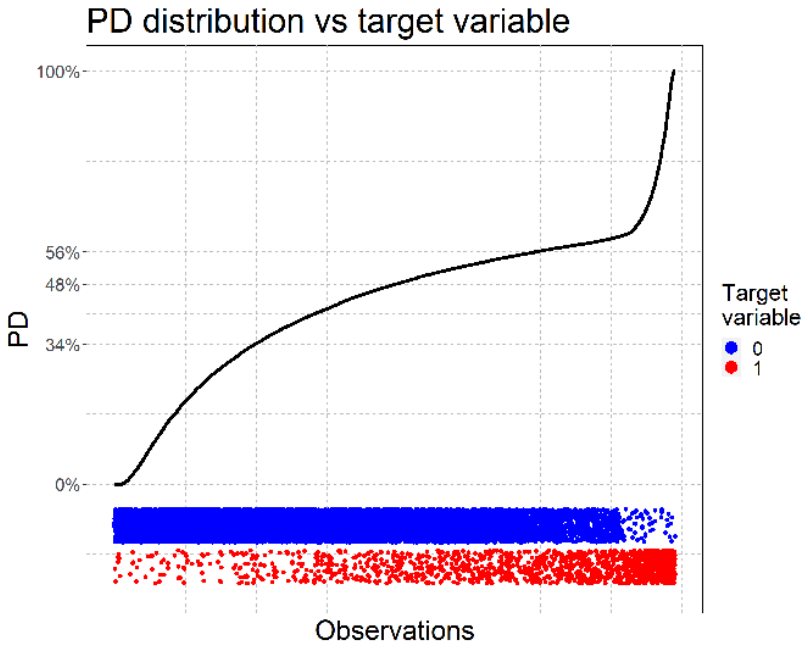


Fig. 1. Distribution of PDs compared with the corresponding target values. y-axis reports quartiles of PD values.

We tune the parameters of each model with the Stratified Cross-Validation and we calibrate the models with the optimal parameters on the entire dataset, so to have a single model[8] to be used for feature importance evaluation. In Table 3 we report performance on cross-validation folds and on the entire dataset for each model as well a comparison between the models trained with the input variables only and the ones with the addition of PD. Random Forest is the only model with good performances, being able to capture the different local separation of the data, as discussed in Section 4.3. Nevertheless, all models show an improvement on class-specific performance, i.e. F1-score for the defaulted class, and on the AUC when the PD is included as predictor. Table B.4 in the Appendix reports the results of the models with controls for fixed effects, showing stability of performances and a resulting robustness of the models. Finally, Figures from B.4 to B.9 in the Appendix show the comparison of F1-score and ROC curves of all models and fixed effects.

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

Table 3
F1-score and AUC for Elastic-Net, MARS and Random Forest calibrated on dataset with input variables only and with the addition of PD. Values refer to performance of model calibrated on the entire dataset. Values in parenthesis refer to average performance of validation folds of Cross-Validation.

	F1 (Cross-Val)		AUC (Cross-Val)	
Algorithm	Baseline	With PD	Baseline	With PD
Elastic-Net	30.7% (30.1±1.7%)	35.1% (35.1±1.5%)	79.8% (79.6±0.6%)	82% (81.7±0.8%)
MARS	36% (33.8±1.4%)	40% (37.5±0.6%)	82.5% (81.7±0.6%)	84.2% (82.8±0.8%)
Random Forest	89.5% (85.1±1.7%)	95.8% (91.4±1.2%)	89.8% (85.4±1.1%)	96.1% (91.7±0.7%)

Finally, we explore the feature importance for all models. PFI and SHAP are evaluated on model calibrated with input variables and with the addition of PD. Figure 2 shows the PFI of Random Forest model, where the changes of F1-score are normalized. PD is the second most important variable, slightly below the financial interest on revenues. Figures 3a and 3b show the effect of input variables on the predicted probabilities[9] of Random Forest model, for each observation predicted as 1 and 0, respectively. The color of the points ranges from red, meaning that the observation has 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 reports the average absolute change in predicted probability over all observations and the normalized values, in parenthesis. PD is on 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 of probability, whereas for non-defaulted observations low values of PD (red) result in a major 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. Figures

[9] All three 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.

4a and 4b 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 on the top two most important variables and increases 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 already measured with the increase of performance of the models. Although both techniques lead to the same conclusion, it is worth noting the complementary contribution to model interpretability: PFI provides a synthetic cumulative measure of the relative importance of the variables, whereas SHAP provides 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 C.10 to C.15 in the Appendix report the PFI and SHAP variable importance for Elastic-Net and MARS models, leading to similar results, supporting the relevance of the addition of PD as a predictor.

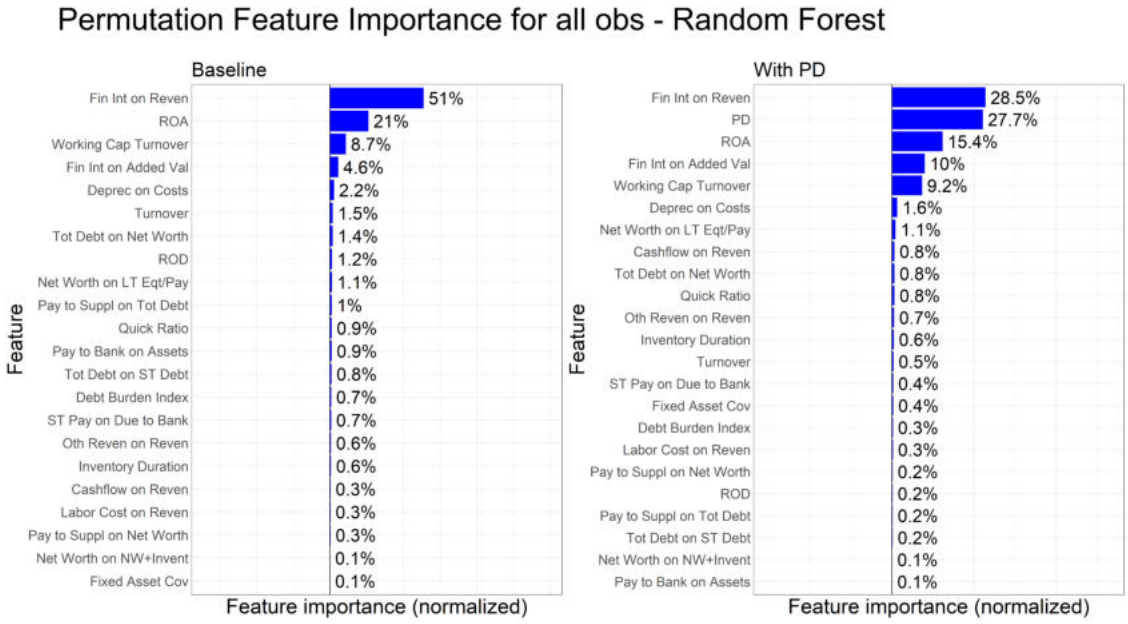


Fig. 2. Permutation Feature Importance for Random Forest model, comparing variable importance of model calibrated with input variables and with the addition of PD. Normalized changes of F1-score are used to rank the variables.

6. Conclusions

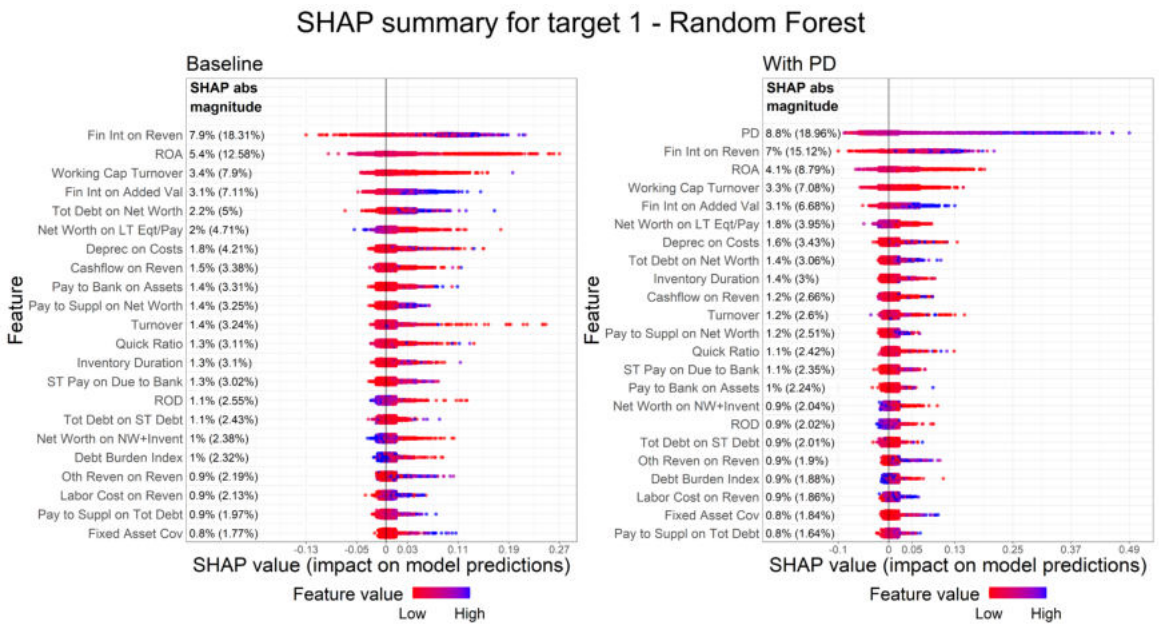
By exploiting a unique and proprietary dataset on a sample of 10,136 Italian micro-, small, and mid-sized enterprises (MSMEs) operating with 113 cooperative banks over the period 2012–2014, this paper investigates the role of market information in predicting corporate default for unlisted firms. The status of bank's clients is predicted by the means of three empirical models, i.e., logistic Elastic-Net, Multivariate Adaptive Regression Spline (MARS), and Random Forest (RF). We calculate the proxy of market-based Merton's PD credit risk measure by using market data of comparable publicly-listed companies to proxy for the asset price volatility of our unlisted firms. Specifically, the matching procedure between unlisted and their comparable publicly-listed firms is implemented by the means of dimensionality reduction and clustering technique. Moreover, we further evaluate each variable importance in predicting corporate default through the use of Shapley values.

Our results provide novel evidence that market information represents a crucial indicator in predicting corporate default of unlisted firms. Indeed, we show a significant improvement of the model performance, both on class-specific (F1-score for defaulted class) and overall metrics (Area Under the Curve) 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.

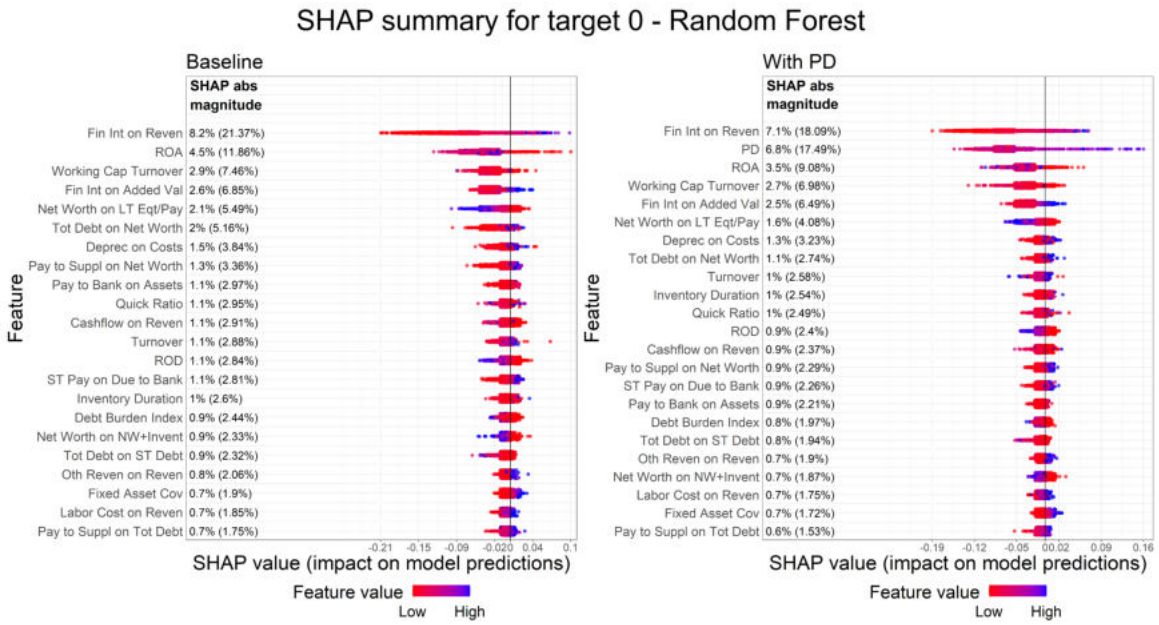
Our study makes important inferences for policy implications. Indeed, our findings shed new light on the opportunity for banks to potentially integrate their hybrid credit scoring methodologies with market information for credit risk assessments, with the purpose of increasing the accuracy of forecasting corporate defaults for unlisted firms. Thus, the results of this paper could be very helpful for forward-looking financial risk management frameworks (Breden, 2008, Rodriguez Gonzalez et al., 2018).

Future extensions stemming from this work could involve not only applying alternative prediction models so to provide further evidence on the importance of market information to predict

corporate default of unlisted firms, but also testing the impact of synthetic information extracted with the dimensionality reduction technique when replacing the original financial ratios. Testing different clustering technique and exploring the distribution of the clusters could also lead to new insights on clients' behavior and their connections with the market, through the mapping with their publicly-listed peers.

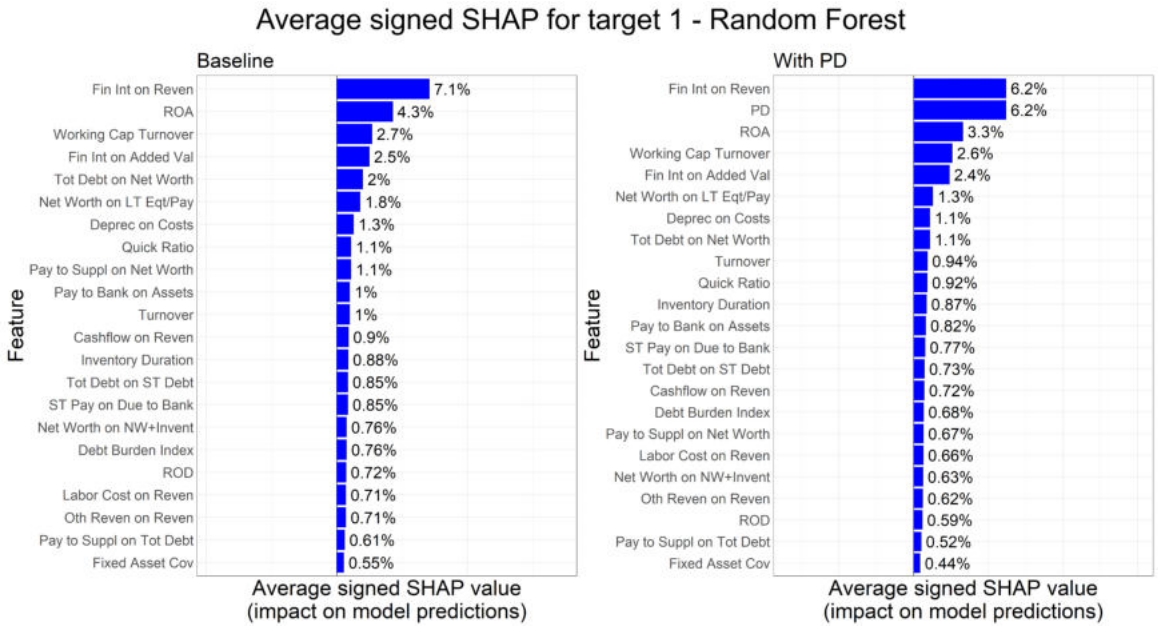


(a) Defaulted clients.

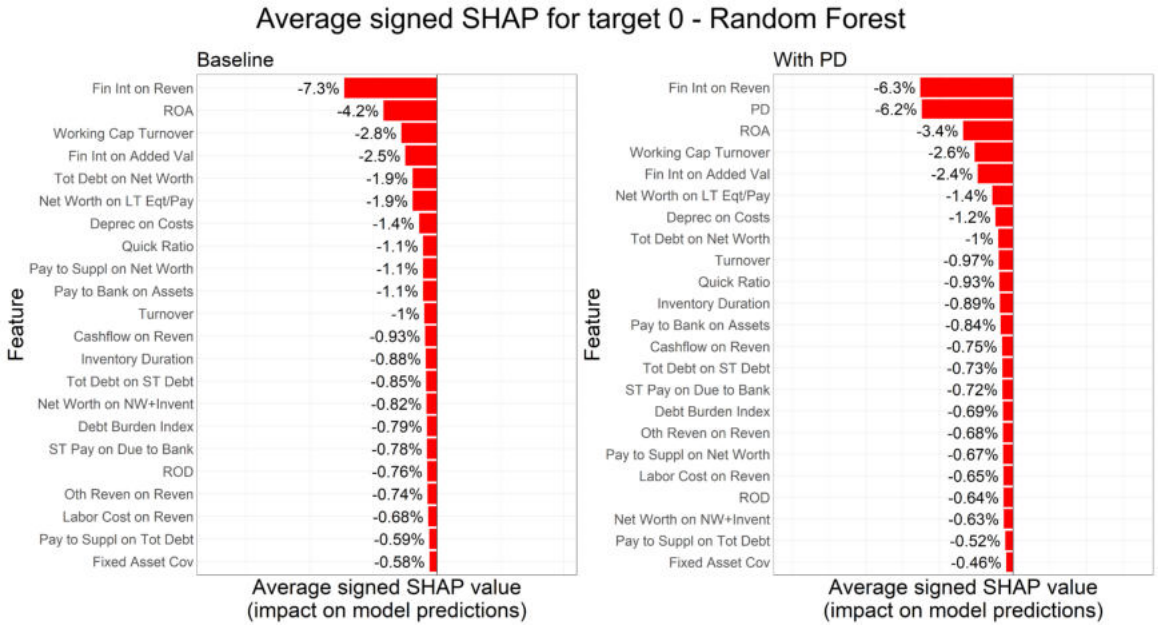


(b) Non-defaulted clients.

Fig. 3. 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 color of the points ranges from red, meaning that the observation has 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 reports the average absolute change in predicted probability over all observations and the normalized values, in parenthesis.



(a) Defaulted clients.



(b) Non-defaulted clients.

Fig. 4. 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.

References

Agarwal, V. and Taffler, R., 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance*, 32:1541–1551.

Akbari, A., Ng, L., and Solnik, B., 2021. Drivers of economic and financial integration: A machine learning approach. *Journal of Empirical Finance*, 61:82–102.

Albanesi, S. and 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. *Journal of Accounting Research*, 30:94–108.

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23:589–609.

Altman, E. and Sabato, G., 2007. Modelling credit risk for smes: Evidence from the us market. *The Journal of Business*, 43:332–357.

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

Andrikopoulos, P. and Khorasgani, A., 2018. Predicting unlisted smes' default: Incorporating market information on accounting-based models for improved accuracy. *The British Accounting Review*, 50:559–573.

Avramov, D., Li, M., and Wang, H., 2021. Predicting corporate policies using downside risk: A machine learning approach. *Journal of Empirical Finance*, 63:1–26.

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

Barbaglia, L., Manzan, S., and Tosetti, E., 07 2021. Forecasting Loan Default in Europe with Machine Learning*. *Journal of Financial Econometrics*.

Bauer, J. and Agarwal, V., 2014. Are hazard models superior to traditional bankruptcy prediction approaches? a comprehensive test. *Journal of Banking and Finance*, 40:432–442.

Beaver, W., 1966. Financial ratios as predictors of failure. *Journal of Accounting Research*, 4: 71–111.

Bharath, S. and Shumway, T., 2008. Forecasting default with the merton distance to default model. *Review Financial Studies*, 21:1339–1369.

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

Blum, M., 1974. Failing company discriminant-analysis. *Journal of Accounting Research*, 12: 1–25.

Breden, D., 2008. Monitoring the operational risk environment effectively. *Journal of Risk Management in Financial Institutions*, 1:156–164.

Breiman, L., 2001. Random forests. *Machine Learning*, 45:5–32.

Brown, M., Westerfeld, S., Schaller, M., and Heusler, M., 2012. Information or insurance? on the role of loan officer discretion in credit assessment. *MoFiR Working Paper*, 67.

Burgstahler, D., Hail, L., and Leuz, C., 2006. The importance of reporting incentives: Earnings management in european private and public firms. *The Accounting Review*, 81:983–1016.

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

Calabrese, R., Marra, G., and Osmetti, S., 2016. Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society*, 67:604–615.

Charalambous, C., Neophytou, E., and Charitou, A., 2004. Predicting corporate failure: empirical evidence for the uk. *European Accounting Review*, 13:465–497.

Chen, T. and Liao, H., 2005. A multi-period corporate credit model-an intrinsic valuation approach.

Das, S., Hanouna, P., and Sarin, A., 2009. Accounting-based versus market-based cross-sectional models of cds spreads. *Journal of Banking and Finance*, 33:719–730.

Dierkes, M., Erner, C., Langer, T., and Norden, L., 2013. Business credit information sharing and default risk of private firms. *Journal of Banking and Finance*, 37:2867–2878.

Doumpos, M., Niklis, D., Zopounidis, C., and Andriosopoulos, K., 2015. Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from european listed firms. *Journal of Banking and Finance*, 50:599–607.

Edminster, R., 1972. An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis*, 7:1477–1493.

Falkenstein, E., Boral, A., and Carty, L., 2000. Riskcalctm for private companies: Moody's default model. *Moody's Investor Service*.

Fiordelisi, F., Monferrà, S., and Sampagnaro, G., 2014. Relationship lending and credit quality. *Journal of Financial Services Research*, 46:295–315.

Fisher, A., Rudin, C., and 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., and Reedtz, P. M., 1998. Multiple banking relationships and the fragility of corporate borrowers. *Journal of Banking and Finance*, 22:1441–1456.

Friedman, J., Hastie, T., and Tibshirani, R., 2009. The elements of statistical learning. *Springer*.

Grice, J. and Ingram, R., 2001. Tests of the generalizability of altman's bankruptcy prediction model. *Journal of Business Research*, 54:53–61.

Gropp, R. and Guettler, A., 2018. Hidden gems and borrowers with dirty little secrets: Investment in soft information, borrower self-selection and competition. *Journal of Banking and Finance*, 87:26–39.

Hillegeist, S., Keating, E., Cram, D., and Lundstedt, K., 2004. Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9:5–34.

Hol, S., 2007. The influence of the business cycle on bankruptcy probability. *International Transactions in Operational Research*, 14:75–90.

Islam, A. and Tedford, D., 2012. Risk determinants of small and medium-sized manufacturing enterprises (smes) - an exploratory study in new zealand. *Journal of Industrial Engineering International*, 8.

Islam, M. A., 2008. Risk management in small and medium-sized manufacturing organization in new zealand. *PhD Thesis, Department of Mechanical Engineering. The University of Auckland*.

J.F., W., J.A., S., and B.A., J., 2001. Takeovers, restructuring, and corporate governance. *Pearson*.

Keasey, K. and Watson, R., 1987. Non-financial symptoms and the prediction of small company failure: A test of argenti's hypotheses. *Journal of Business Finance and Accounting*.

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

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

Liberti, J. and Mian, A., 2009. Estimating the effect of hierarchies on information use. *Review Financial Studies*, 22:4057–4090.

Liberti, J. and Petersen, M., 2017. Information: Hard and soft. *Review of Corporate Finance Studies*, 8:1–41.

Lin, S., Ansell, J., and Andreeva, G., 2012. Predicting default of a small business using different definitions of financial distress. *The Journal of the Operational Research Society*, 63:539–548.

Louzada, F., Ara, A., and Fernandes, G., 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21:117–134.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I., 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:2522–5839.

McCarthy, E., 1999. Pricing ipos: Science or science fiction? *Journal of Accountancy*, 188:51–58.

Merton, R. C., 1974. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29:449–470.

Modigliani, F. and Miller, M., 1958. The cost of capital, corporate finance, and the theory of investment. *American Economic Review*, 3(48):261–297.

Moscatelli, M., Narizzano, S., Parlapiano, F., and Viggiano, G., 2019. Corporate default forecasting with machine learning. *Bank of Italy Working Paper*, 1256.

Mullainathan, S. and Spiess, J., 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Norden, L. and Weber, M., 2010. Credit line usage, checking account activity, and default risk of bank borrowers. *Review of Financial Studies*, 23:3665–3699.

Ohlson, J., 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18:109–131.

Olson, L. M., Qi, M., Zhang, X., and Zhao, X., 2021. Machine learning loss given default for corporate debt. *Journal of Empirical Finance*, 64:144–159.

Osborne, M. J. and Rubinstein, A., 1994. A course in game theory. *MIT press*.

Peel, M., Peel, D., and Pope, P., 1986. Predicting corporate failure - some results for the uk corporate sector. *Omega*, 14:5–12.

Pindado, J., Rodrogues, L., and de la Torre, C., 2008. Estimating financial distress likelihood. *Journal of Business Research*, 61:995–1003.

Qian, J., Strahan, P. E., and Yang., Z., 2015. The impact of incentives and communication costs on information production and use: Evidence from bank lending. *Journal of Finance*, 70: 1457–1493.

Rikkers, F. and Thibeault, A. E., 2009. A structural form default prediction model for smes, evidence from the dutch market. *Multinational Finance Journal*, 13:229–264.

Rodriguez Gonzalez, M., Basse, T., and 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.

Shapley, L. S., 1953. A value for n-person games. *Contributions to the Theory of Games 2.28*, pages 307–317.

Shumway, T., 2001. Forecasting bankruptcy more accurately: a simple hazard model. *The Journal of Business*, 74:101–124.

Stickney, C. P. and Weil, R. L., 1997. Financial accounting: An introduction to concepts, methods, and uses. *Dryden Press Series in Accounting*.

Strumbelj, E. and Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems 41.3*, pages 647–665.

Tian, S., Yu, Y., and Guo, H., 2015. Variable selection and corporate bankruptcy forecasts. *Journal of Banking and Finance*, 52:89–100.

Tinoco, M. H. and Wilson, N., 2013. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, 30:394–419.

Vassalou, M. and Xing, Y., 2004. Default risk in equity returns. *Journal of Finance*, 59:831–868.

Virdi, A. A., 2005. Risk management among smes–executive report of discovery research. *The Consultation and Research Centre of the Institute of Chartered Accountants in England and Wales*.

Volk, M., 2012. Estimating probability of default and comparing it to credit rating classification by banks. *Economic and Business Review*, 14:299–320.

Zhu, L., Qiu, D., Ergu, D., Ying, C., and Liu, K., 2019. A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, 162:503–513.

Appendix A. Dataset

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 'Labor 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 - Labor Cost on Reven	Labor 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 inrerest on revenues	3.2	38.5	0	0	0.02	0.39	486.94
16 - Fin Int on Added Val	Financial inrerest 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.

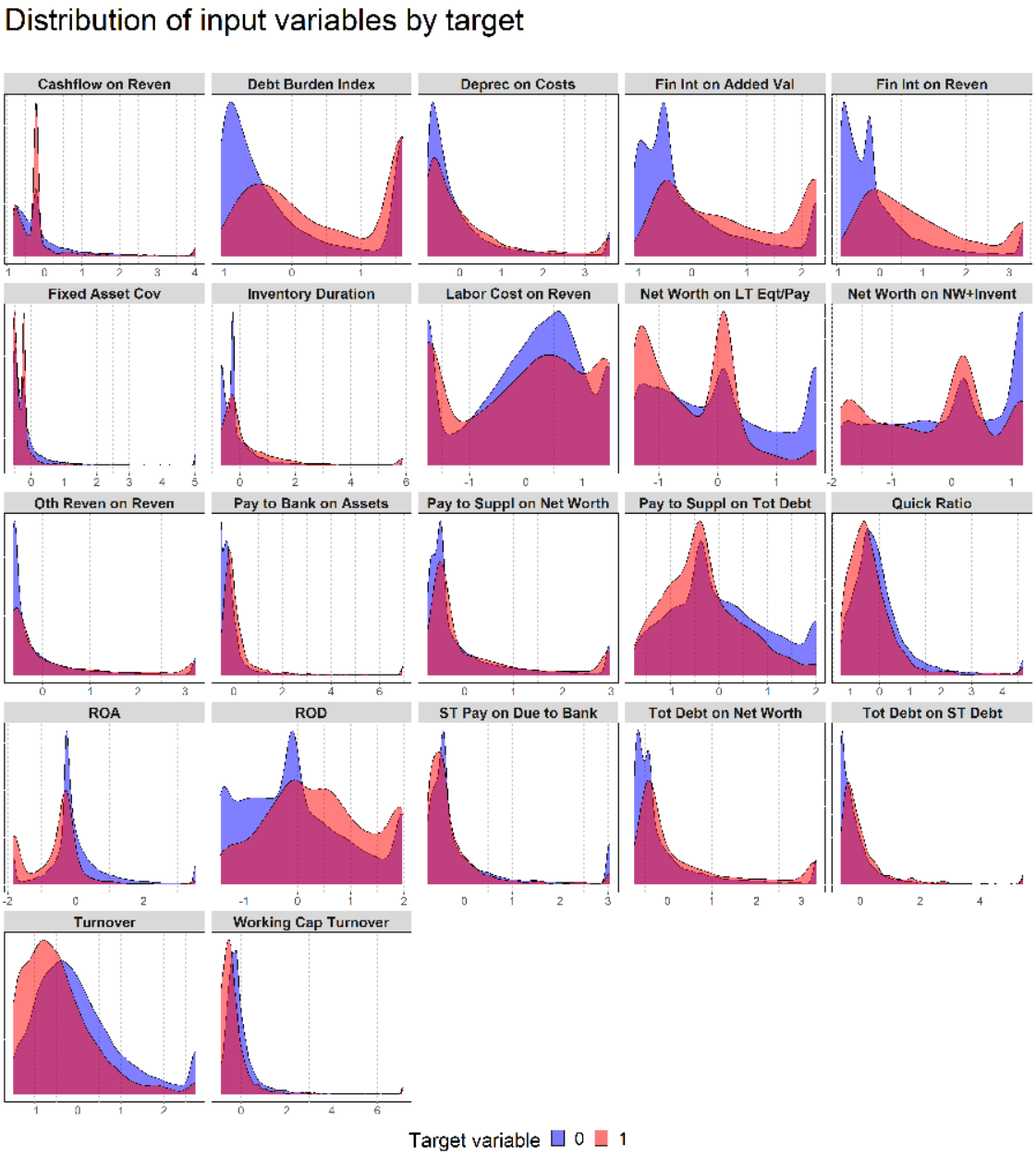


Fig. A.2. Distribution of input variables for MSMEs splitted by target variable.

Table A.3
Distribution of clients that are persistent over time, i.e. 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	9,228
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

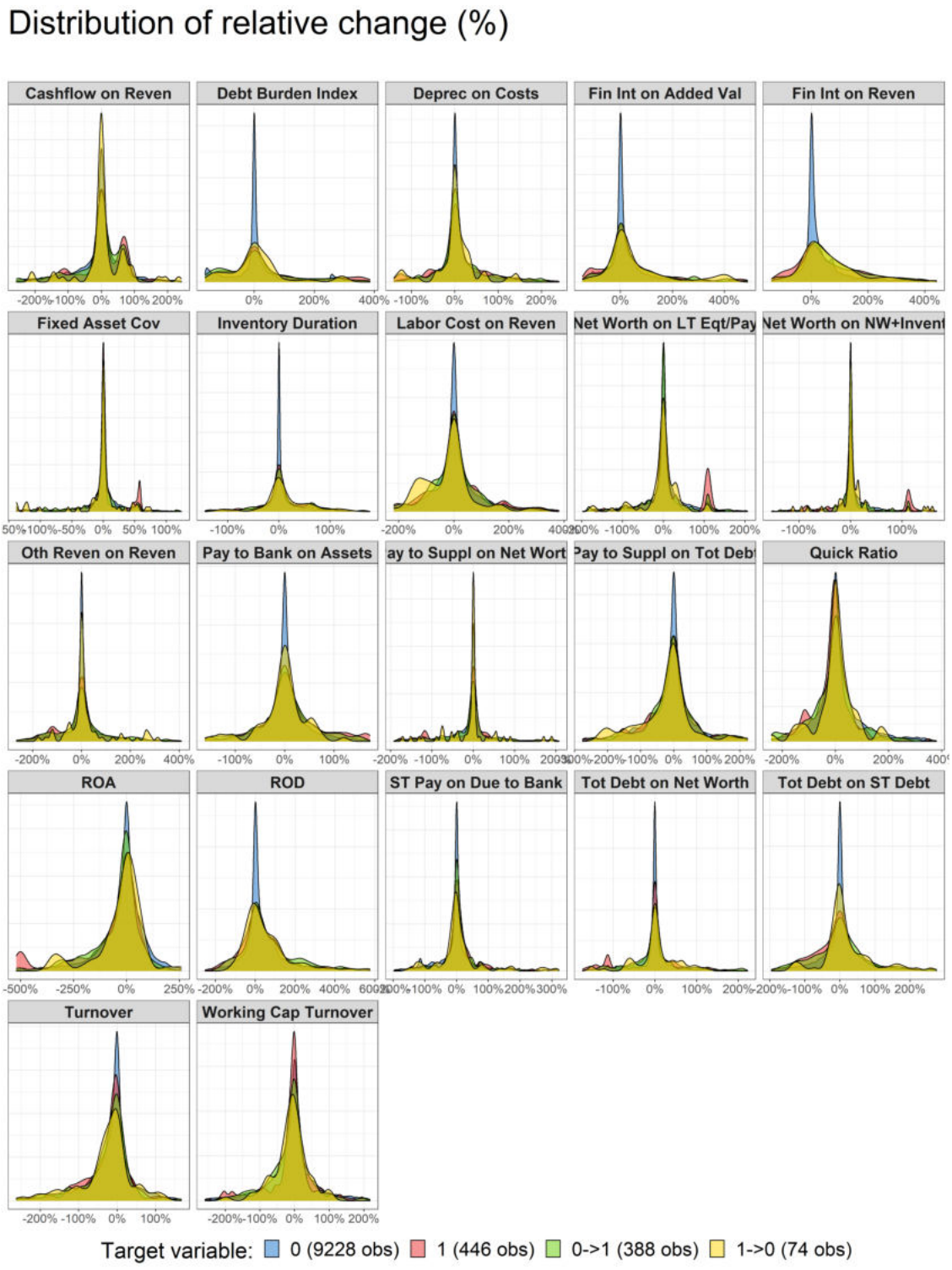


Fig. A.3. Distribution of relative changes over the years of each input variable divided by clients' behavior. Blue and red distributions represent the clients with persistent target of 0 and 1, respectively, green and yellow distributions represent the clients that moved from 0 to 1 and vice-versa, respectively.

Appendix B. Results

Table B.4
F1-score and AUC for Elastic-Net, MARS and Random Forest calibrated on dataset with input variables only and with the addition of PD and with or without controls for fixed effects. Values refer to performance of model calibrated on the entire dataset. Values in parenthesis refer to average performance of validation folds of Cross-Validation.

		F1 (Cross-Val)		AUC (Cross-Val)	
Control	Algorithm	Baseline	With PD	Baseline	With PD
No control	Elastic-Net	30.7% (30.1±1.7%)	35.1% (35.1±1.5%)	79.8% (79.6±0.6%)	82% (81.7±0.8%)
	MARS	36% (33.8±1.4%)	40% (37.5±0.6%)	82.5% (81.7±0.6%)	84.2% (82.8±0.8%)
	Random Forest	89.5% (85.1±1.7%)	95.8% (91.4±1.2%)	89.8% (85.4±1.1%)	96.1% (91.7±0.7%)
Dummy Industry	Elastic-Net	30.7% (30.7±1.3%)	35.1% (35±3%)	79.8% (79.5±1%)	82% (81.8±1.3%)
	MARS	34.2% (34.4±1.8%)	38.8% (37.5±2.8%)	82.4% (81.9±0.4%)	83.8% (83.2±1.2%)
	Random Forest	90.5% (87.3±1.9%)	95.9% (93.4±2.8%)	90.8% (87.6±1%)	96.2% (93.7±1.6%)
Firm Size	Elastic-Net	30.9% (30.8±0.9%)	35.3% (35.3±1.4%)	79.9% (79.8±1.6%)	82.5% (82.4±1.2%)
	MARS	37.3% (35.4±0.8%)	41.3% (39.3±2.3%)	83.4% (82.2±1.2%)	84.5% (83.3±1.3%)
	Random Forest	90.7% (88.5±2.7%)	96% (91.3±1.9%)	91% (88.8±1.5%)	96.3% (91.6±1.6%)
Firm Type	Elastic-Net	30.8% (30.8±1.2%)	35.4% (35.1±1.7%)	79.8% (79.6±1.2%)	82.2% (82.1±1.3%)
	MARS	36.2% (34.6±2.1%)	40.5% (37.7±3.3%)	82.9% (81.8±1.3%)	84.7% (82.8±1.4%)
	Random Forest	89.5% (87±1.4%)	96.1% (91.5±2.8%)	89.8% (87.3±1%)	96.4% (91.8±1.2%)
Industrial Sector	Elastic-Net	31.3% (31.3±1.7%)	35.4% (34.9±1.5%)	80.1% (80±2%)	82.3% (82±1.6%)
	MARS	34.3% (33.8±2%)	40.3% (36.9±2.8%)	82.4% (81.9±1.7%)	84.6% (82.3±1.5%)
	Random Forest	93.4% (90.2±1.6%)	97.3% (94.7±2%)	93.6% (90.4±1.5%)	97.6% (95±1.4%)
Region	Elastic-Net	30.9% (30.7±1.6%)	35.1% (35±2.6%)	79.8% (79.6±1.9%)	82.1% (81.9±2.1%)
	MARS	34.3% (34.1±1.5%)	37% (36.6±2.7%)	82.4% (82±2.3%)	83.8% (83.1±2.2%)
	Random Forest	92.4% (89.5±1.4%)	97.5% (95.3±2.7%)	92.7% (89.8±2.2%)	97.8% (95.5±2.2%)

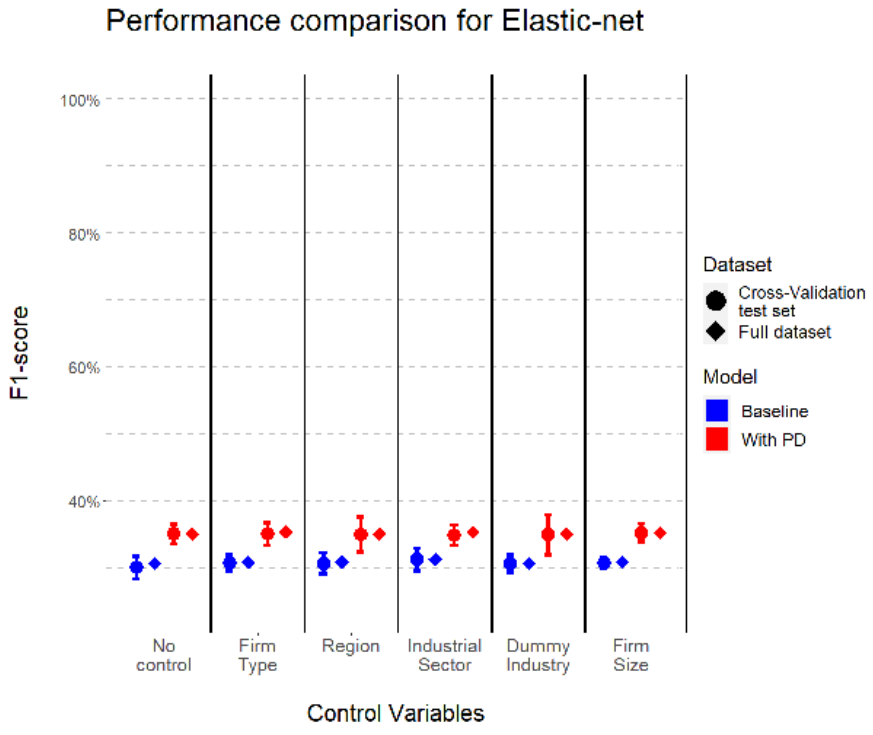


Fig. B.4. Comparison of F1-score for Elastic-Net model for models calibrated with input variables only and with the addition of PD, as well as with or without controls for fixed effects.

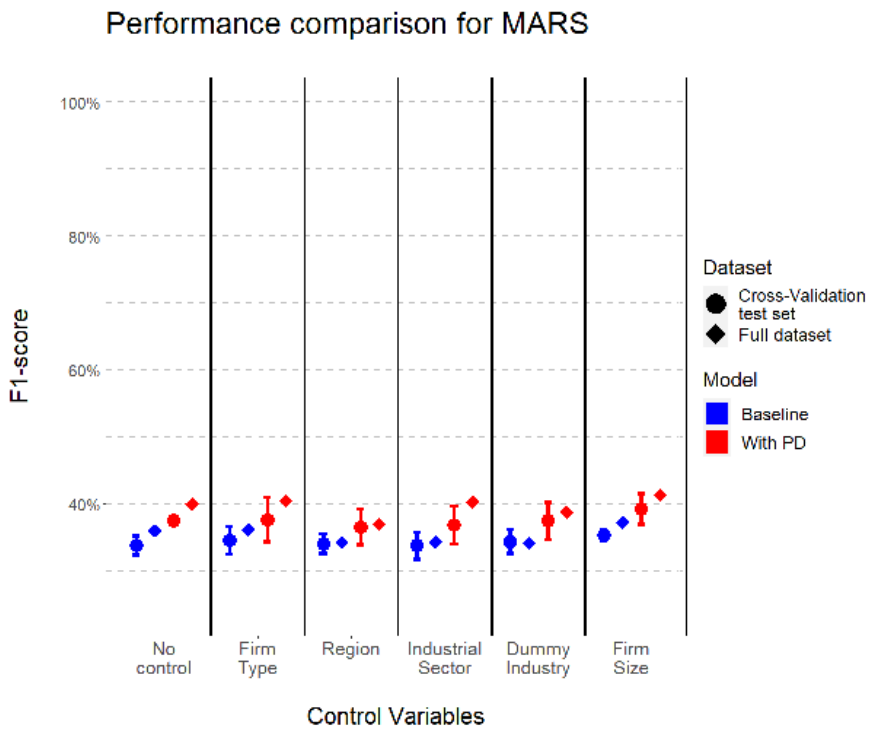


Fig. B.5. Comparison of F1-score for MARS model for models calibrated with input variables only and with the addition of PD, as well as with or without controls for fixed effects.

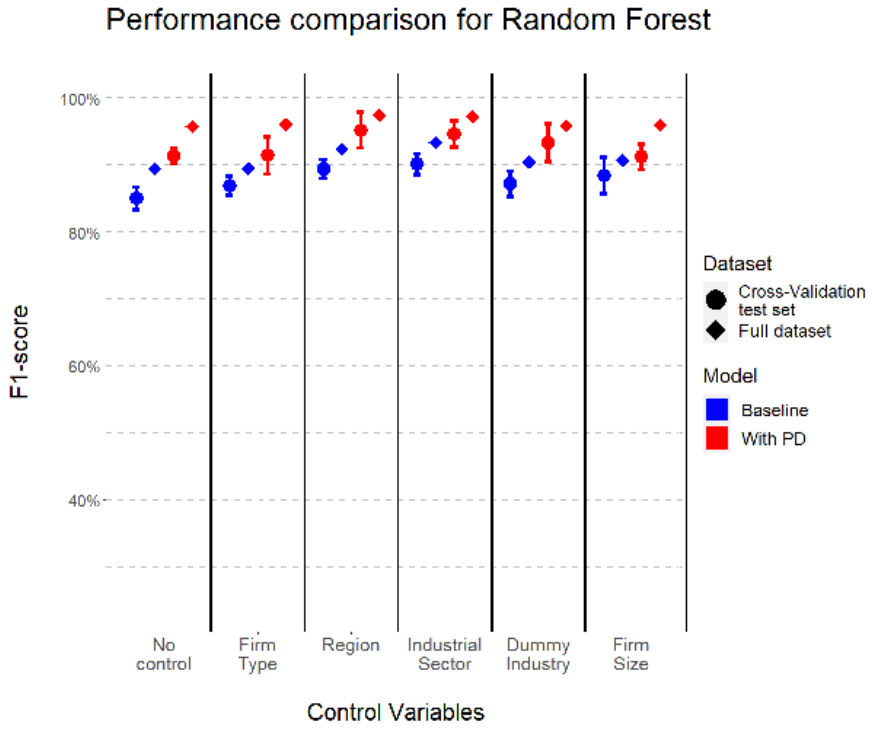


Fig. B.6. Comparison of F1-score for Random Forest model for models calibrated with input variables only and with the addition of PD, as well as with or without controls for fixed effects.

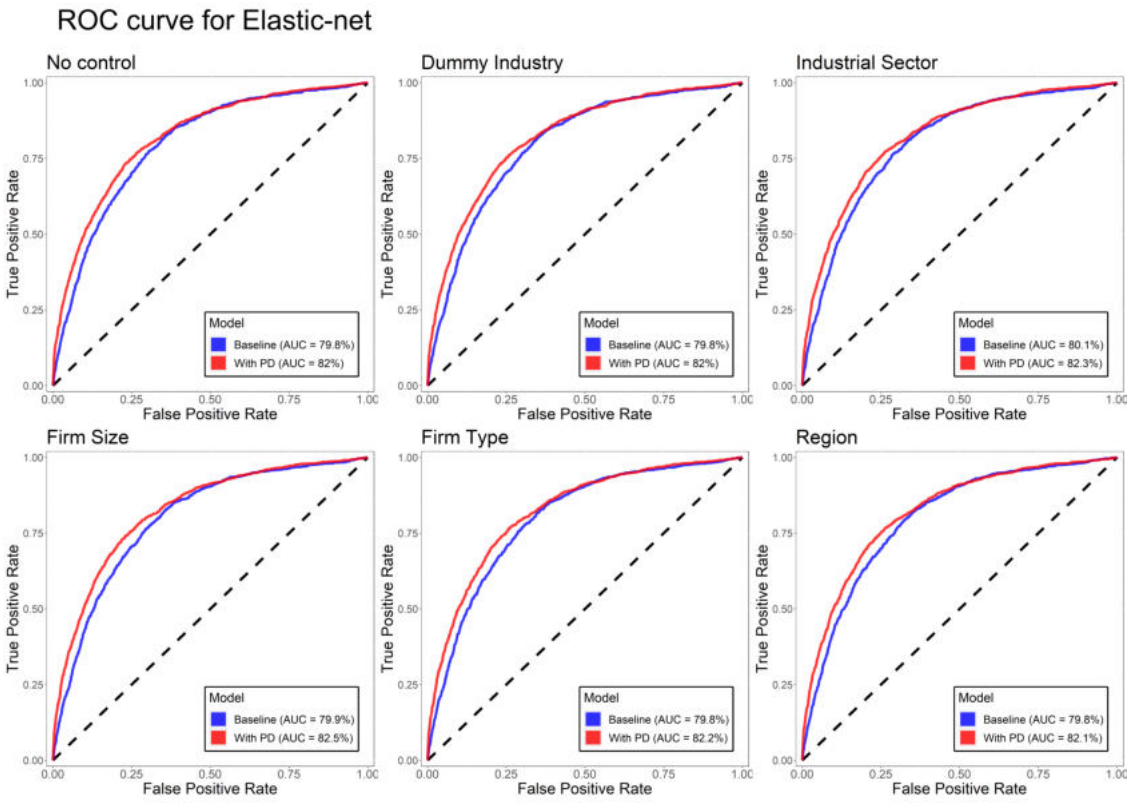


Fig. B.7. Comparison of ROC curves for Elastic-Net model for models calibrated with input variables only and with the addition of PD, as well as with or without controls for fixed effects.

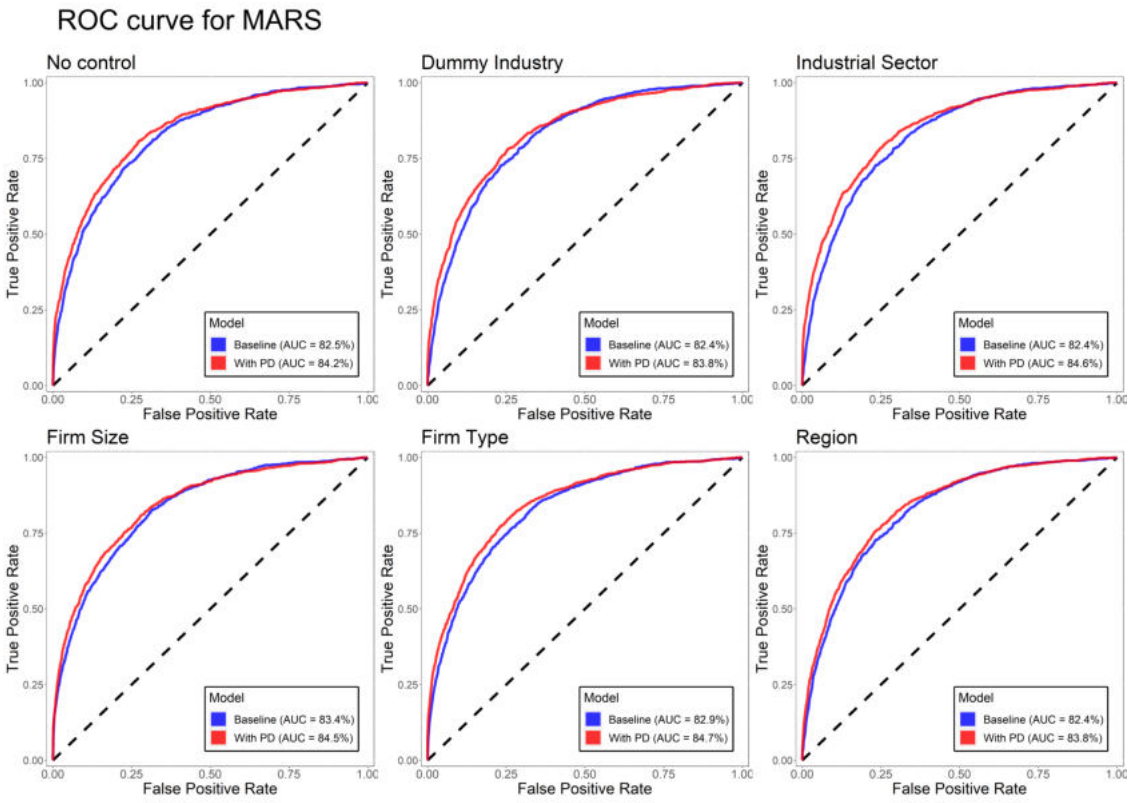


Fig. B.8. Comparison of ROC curves for MARS model for models calibrated with input variables only and with the addition of PD, as well as with or without controls for fixed effects.

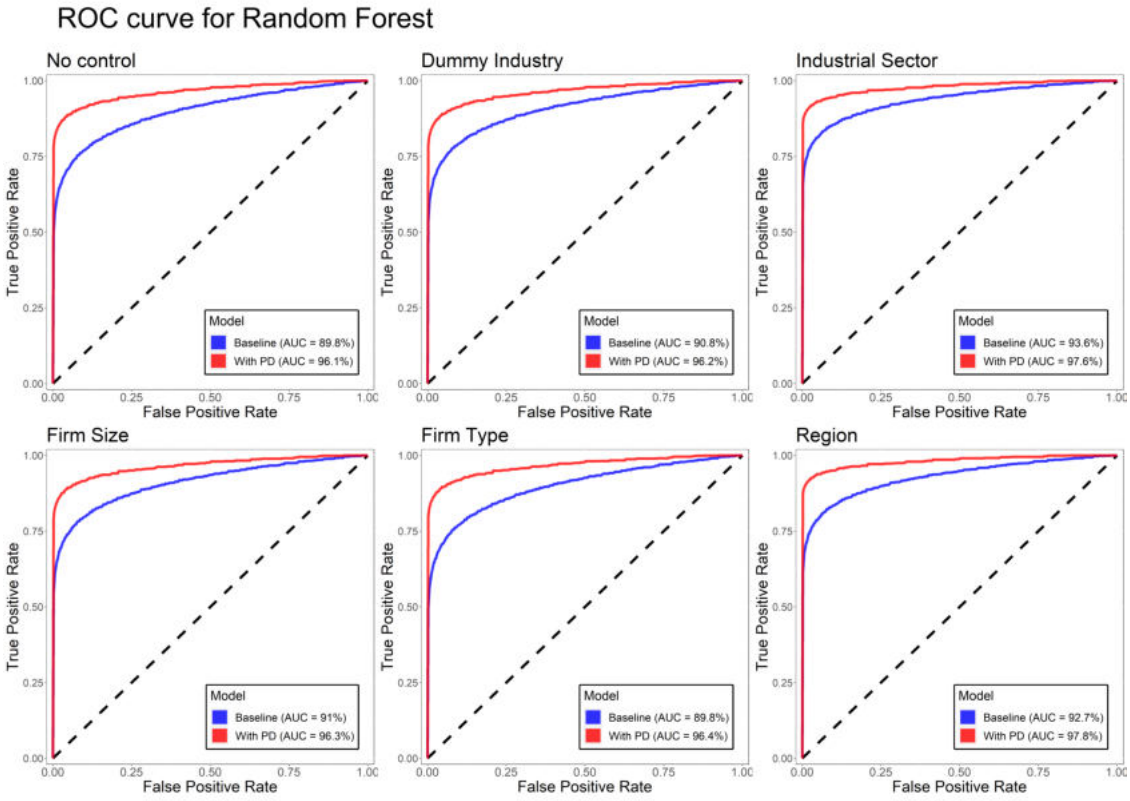


Fig. B.9. Comparison of ROC curves for Random Forest model for models calibrated with input variables only and with the addition of PD, as well as with or without controls for fixed effects.

Appendix C. Feature importance

Explainability capabilities all models PB have been compared using Permutation Feature Importance (PFI) and Shapley Additive Explanations (SHAP). The change in models' performances and in the probability correlated to each predictor has been explored in order 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 model's prediction error after shuffling 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 by means of 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):

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

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 j-th 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 $m = 1, \ldots, M$ **do**

- Draw random instance z from data matrix X;
- Choose a random permutation o of the feature values;
- Order instance x: $x_O = (x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)})$;
- Order instance z: $z_O = (z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)})$;
- Construct two new instances:
  - With feature j: $x_{+j} = (x_{(1)}, \ldots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$
  - Without feature j: $x_{-j} = (x_{(1)}, \ldots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$
- 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 needs to be repeated for each feature of interest in order to get all the Shapley values. Among the advantages of Shapley values over the other methods, in first place there

is the efficiency property, i.e., the difference between prediction and average prediction is fairly distributed among features.

Figures from C.10 to C.15 report the PFI and SHAP variable importance for Elastic-Net and MARS models, calibrated with input variables and with the addition of PD as a predictor.

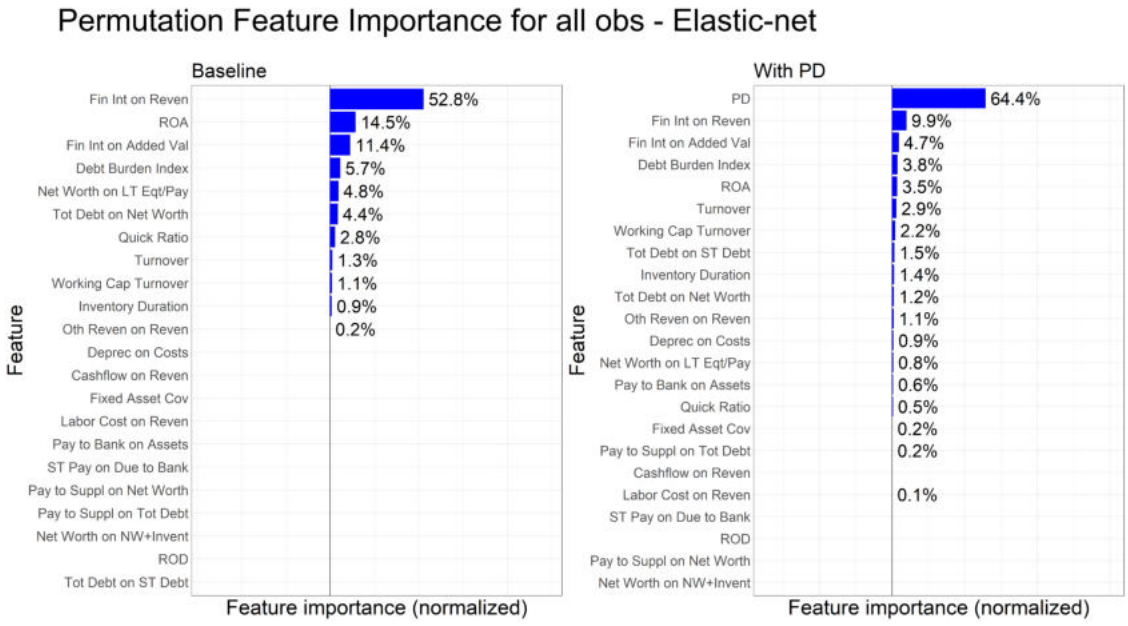
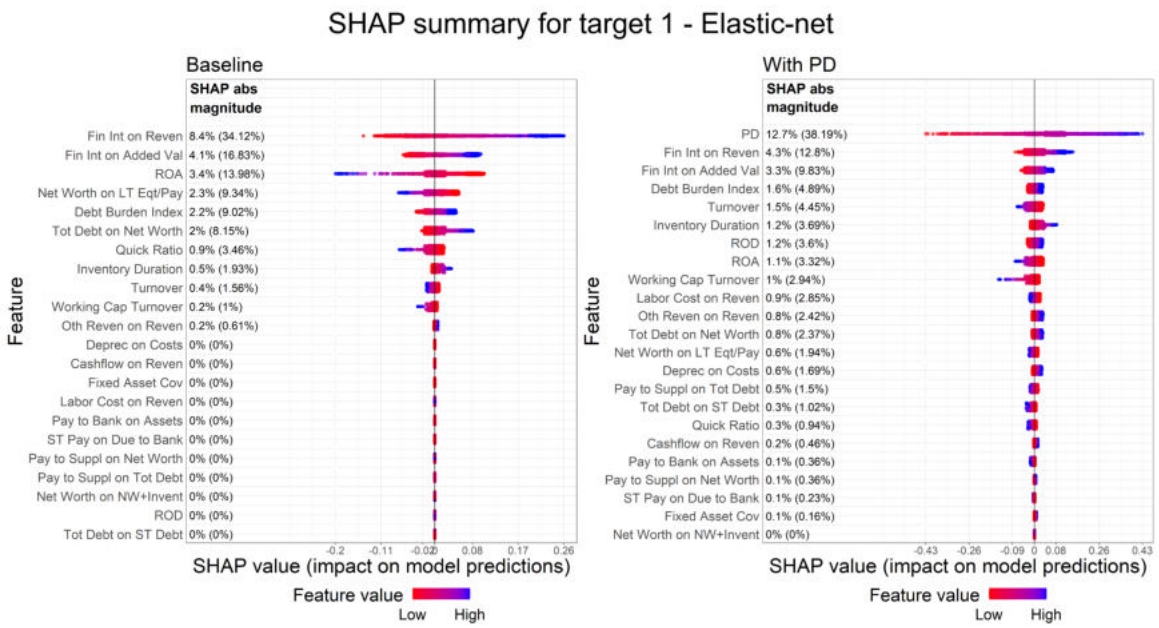
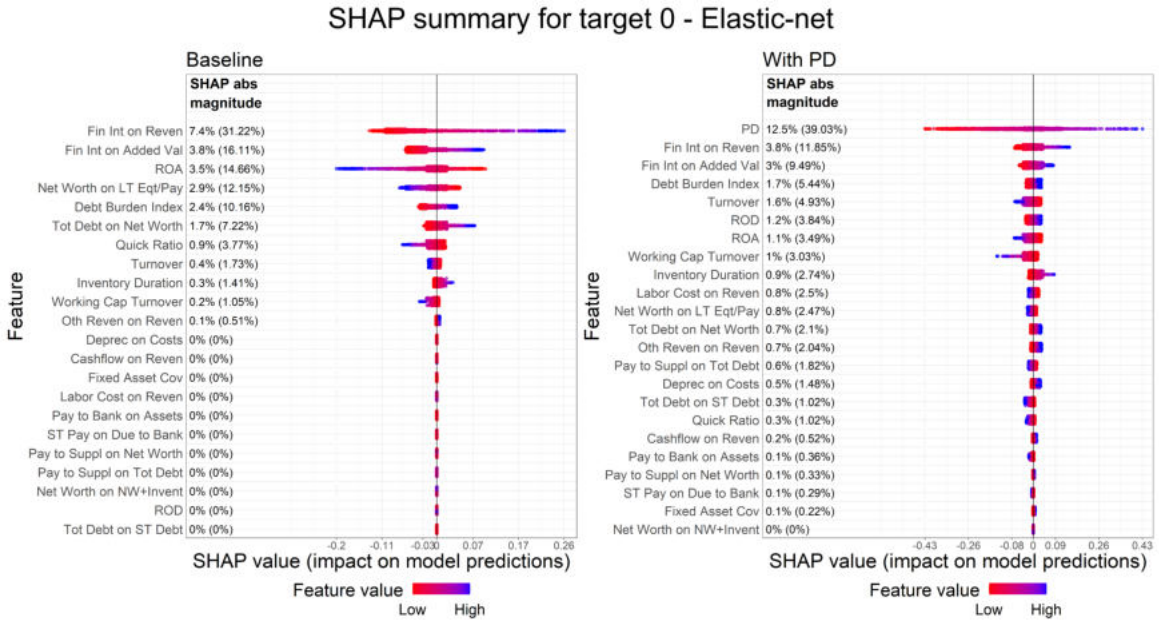


Fig. C.10. Permutation Feature Importance for Elastic-Net model, comparing variable importance of model calibrated with input variables and with the addition of PD. Normalized changes of F1-score are used to rank the variables.

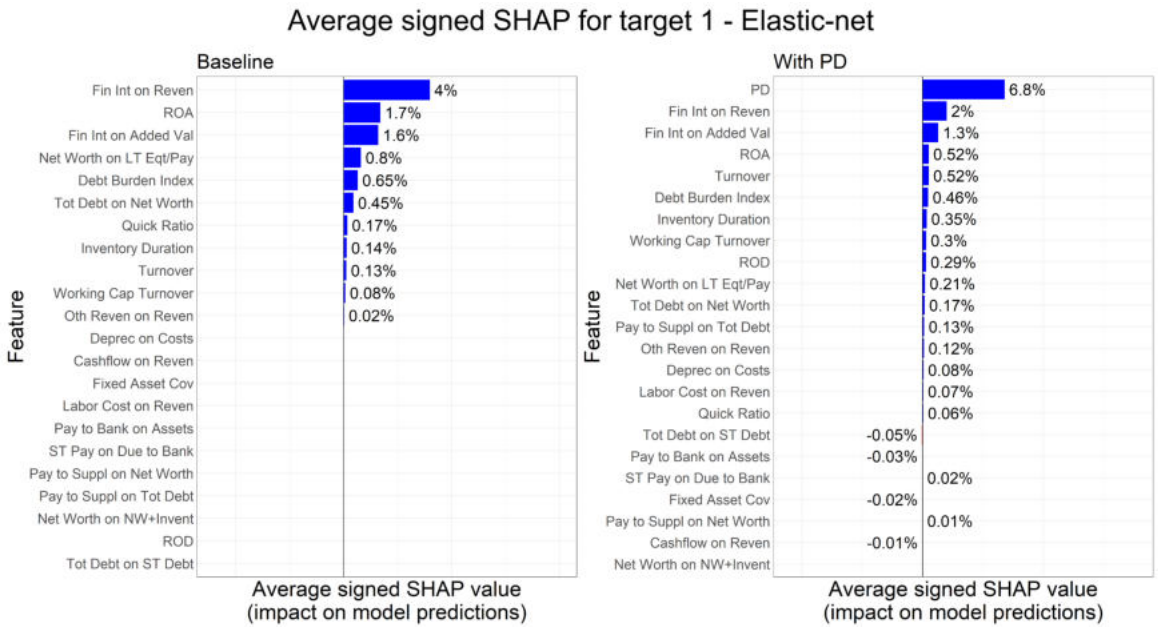


(a) Defaulted clients.

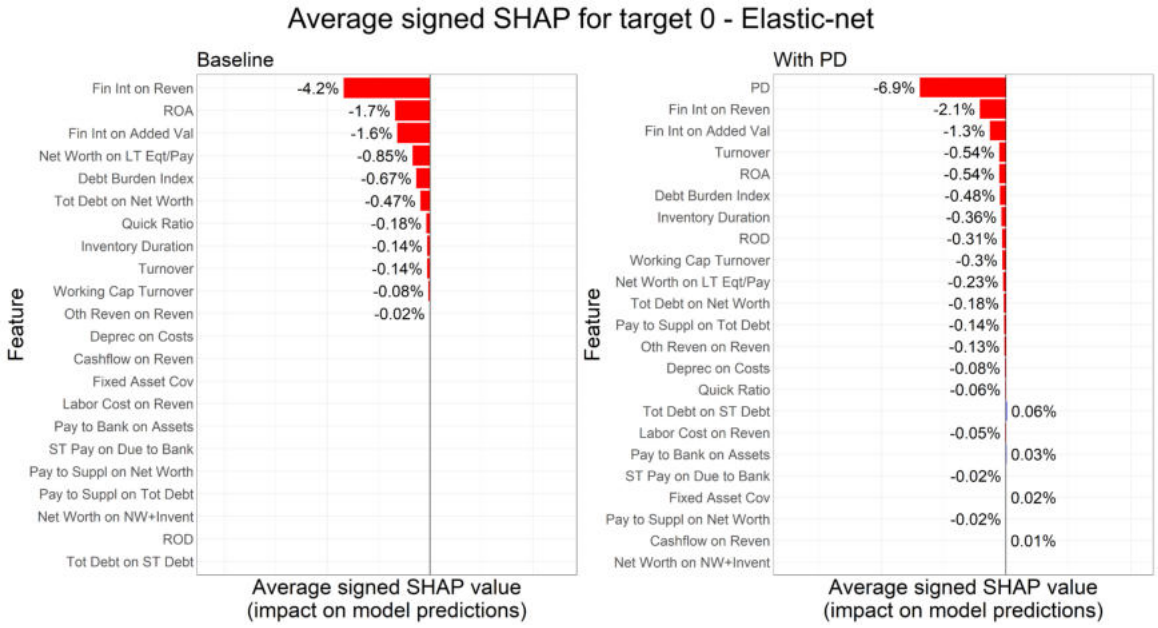


(b) Non-defaulted clients.

Fig. C.11. SHAP effects on predicted probability for Elastic-Net 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 color of the points ranges from red, meaning that the observation has 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 reports the average absolute change in predicted probability over all observations and the normalized values, in parenthesis.



(a) Defaulted clients.



(b) Non-defaulted clients.

Fig. C.12. SHAP average signed effect for Elastic-Net 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.

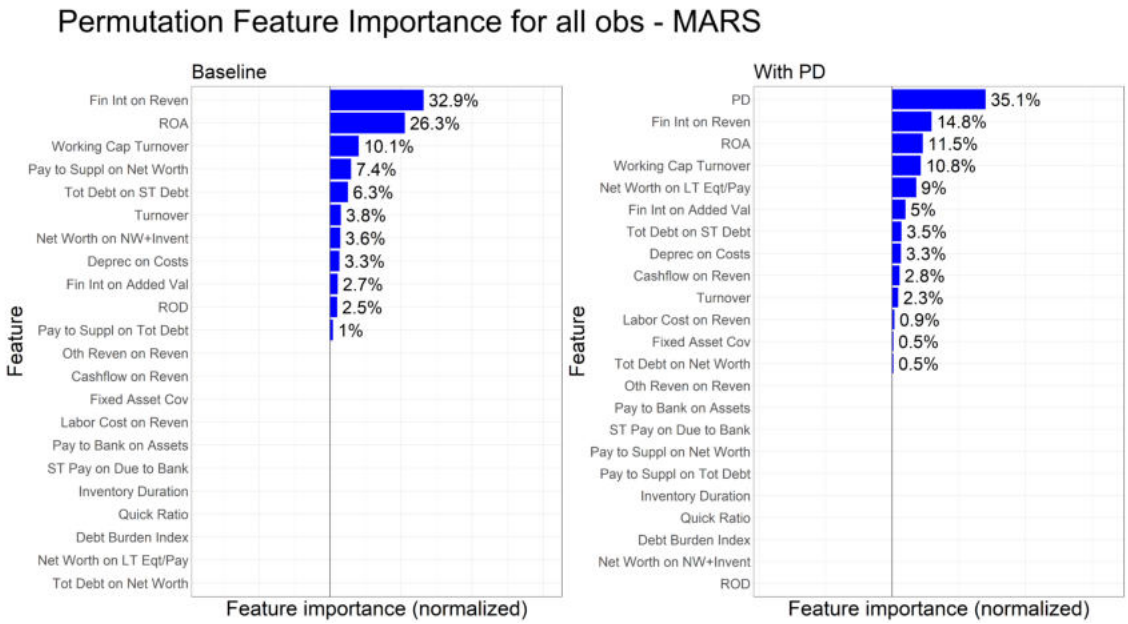
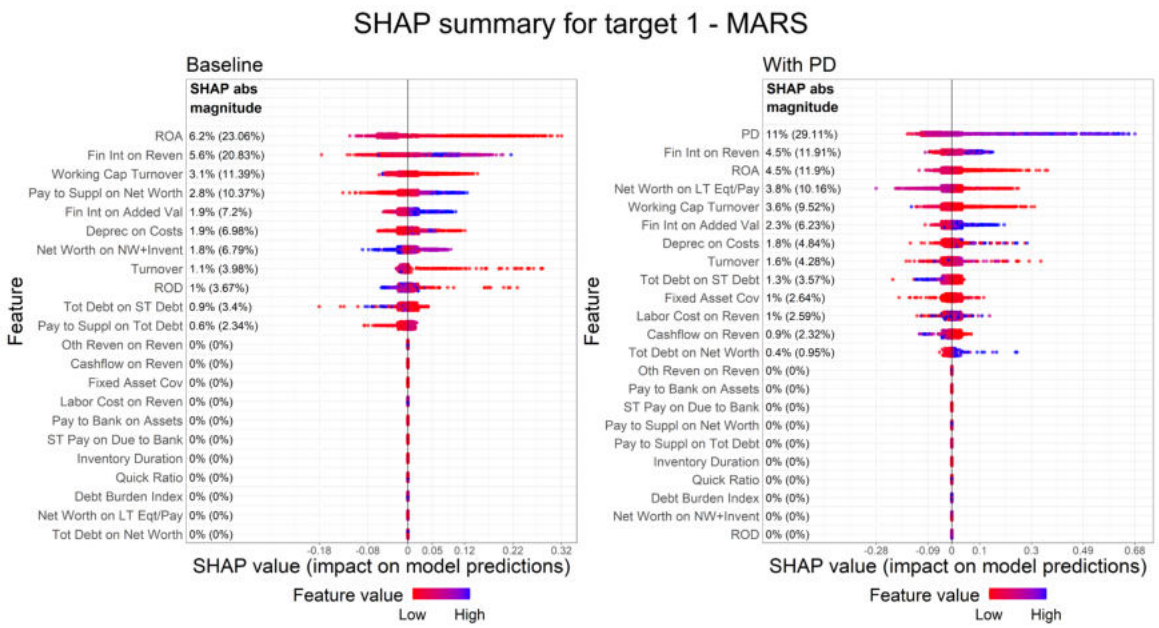
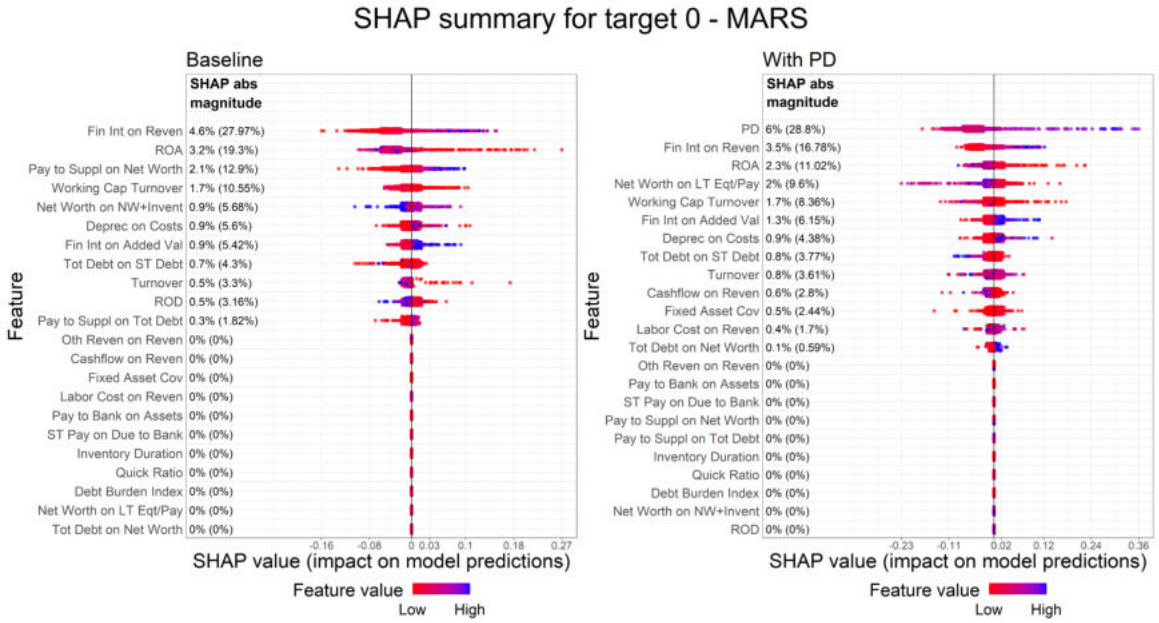


Fig. C.13. Permutation Feature Importance for MARS model, comparing variable importance of model calibrated with input variables and with the addition of PD. Normalized changes of F1-score are used to rank the variables.

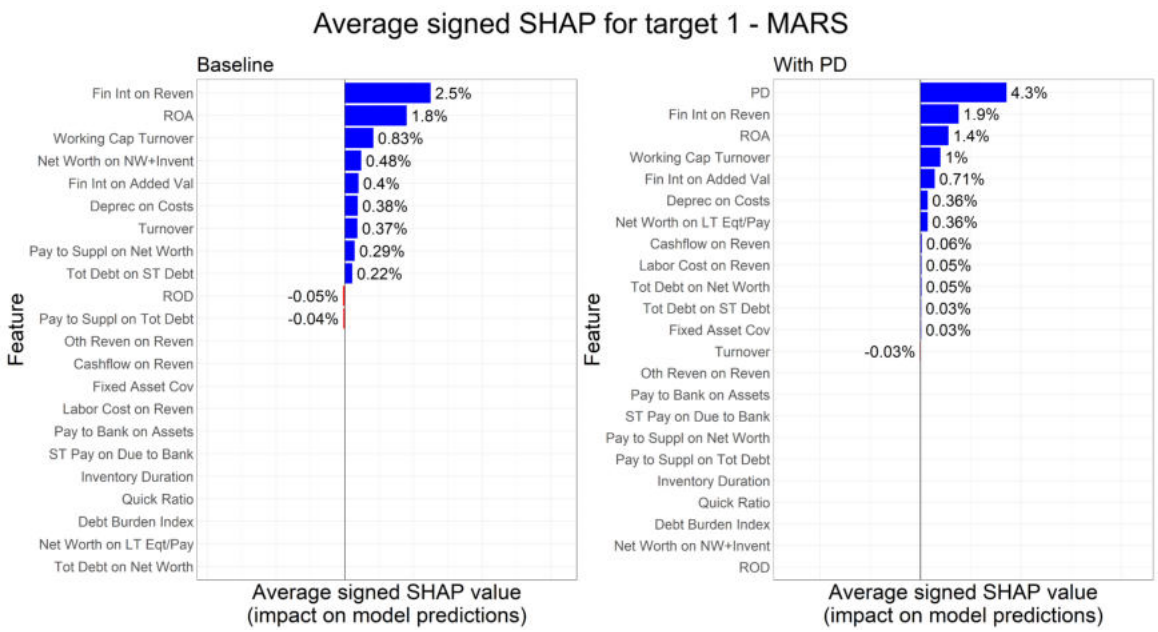


(a) Defaulted clients.

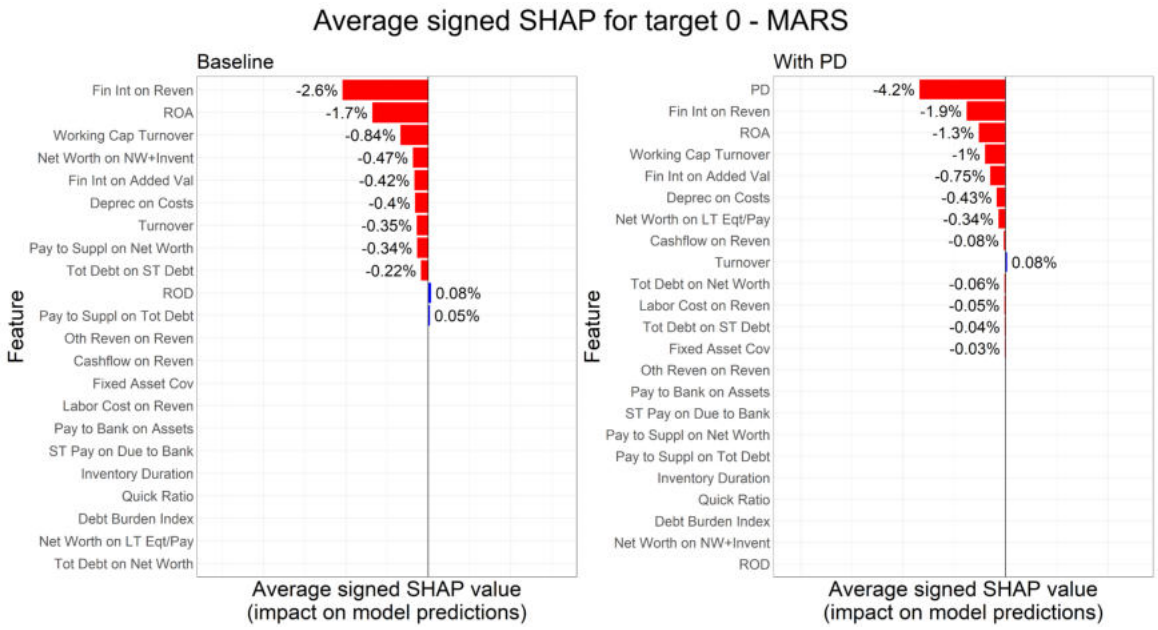


(b) Non-defaulted clients.

Fig. C.14. SHAP effects on predicted probability for MARS 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 color of the points ranges from red, meaning that the observation has 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 reports the average absolute change in predicted probability over all observations and the normalized values, in parenthesis.



(a) Defaulted clients.



(b) Non-defaulted clients.

Fig. C.15. SHAP average signed effect for MARS 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.