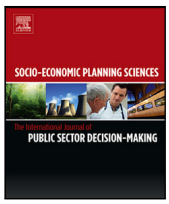
Socio-Economic Planning Sciences 90 (2023) 101746

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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



Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses

Alessandro Bitetto [a,*], Paola Cerchiello [a], Stefano Filomeni [b], Alessandra Tanda [a], Barbara Tarantino [a]

[a] University of Pavia, Department of Economics and Management, Italy
[b] University of Essex, Essex Business School, Finance Group, Colchester, UK

ARTICLE INFO

JEL classification:
C52
C53
D82
D83
G21
G22

Keywords:
Credit rating
SMB
Historical random forest
Machine learning
Relationship banking
Invoice lending

ABSTRACT

In this paper, we compare two different approaches to estimate the credit risk for small- and mid-sized businesses (SMBs), namely a classic parametric approach, by fitting an ordered probit model, and a non-parametric approach, calibrating a machine learning historical random forest (HRF) model. The models are applied to a unique and proprietary dataset comprising granular firm-level quarterly data collected from a European investment bank and an international insurance company on a sample of 464 Italian SMBs over the period 2015–2017. Results show that the HRF approach outperforms the traditional ordered probit model, highlighting how advanced estimation methodologies that use machine learning techniques can be successfully implemented to predict SMB credit risk, i.e. when facing high asymmetries of information. Moreover, by using Shapley values, we are able to assess the relevance of each variable in predicting SMB credit risk.

1. Introduction

The determination of corporate credit ratings and credit risk is a key topic that has alimented the academic debate both theoretically and empirically and has extreme relevance for the industry and the regulatory and supervisory bodies in the financial system as it is a tool to ensure the allocational efficiency of financial markets and intermediaries [3–5].

The determination of corporate ratings is particularly challenging for small- or mid-businesses (SMBs) that are mostly unlisted. SMBs indeed represent a large segment of the corporate market in several economies, especially in the European context, and they are generally subject to relevant asymmetries of information that make the estimation of accurate credit ratings more difficult.

In this context, many scholars have investigated the opportunity to use alternative sources of information, also including soft information derived from intensive relationship banking [6–8], whose importance also has been acknowledged by regulators.[1] Soft information, however, might not be always effective in improving lending activity by banks [9] and it cannot be easily transferred in especially complex organizations [10–15] and this advocates the need to find alternative solutions to exploit hard information to obtain to more accurate credit ratings also for informationally opaque SMBs.

Over time, the technological and methodological advancements in the models determining credit risk and credit rating allowed the inclusion of sophisticated AI techniques to improve credit rating accuracy, which however might suffer from limited explainability. Except for a few studies implementing alternative methodologies [16–18], the literature has been mainly focused on the types of information a financial intermediary should use in assessing SMB credit risk. To date, limited evidence is provided on the opportunity to employ Explainable AI methodologies to estimate SMB credit ratings and few studies perform a comparison of classic parametric and machine learning (ML) approaches.

To fill this gap, this paper has the objective to provide a comparison between two alternative approaches to estimate SMBs credit ratings: on one side, we use a classic parametric approach, namely an ordered

* Correspondence to: Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia, Italy.
E-mail address: alessandro.bitetto@unipv.it (A. Bitetto).

[1] Indeed, since the introduction of the internal ratings-based (IRB) approach, banks are allowed to include qualitative soft information when assessing corporate credit risk [1,2].

https://doi.org/10.1016/j.seps.2023.101746
Received 15 February 2023; Received in revised form 11 September 2023; Accepted 24 October 2023
Available online 30 October 2023

probit model; on the other we employ a non-parametric approach, namely a machine learning Historical Random Forest (HRF).

Our objective is to provide validation of ML techniques in the estimation of SMB credit ratings, especially in the presence of limited information. In this regard, we add to the existing studies by testing and comparing the performance of parametric versus non-parametric methodologies. However, differently from the extant literature, this paper is the first one that applies a Historical Random Forest (HRF) approach, that is – so far – the only approach able to treat the dynamic nature of time series. By doing so, hence, we are able to compare traditional and ML techniques in the "pooled" and "historical" dimensions. Moreover, we further contribute by assessing the relevance of each variable to predict SMB credit risk, through the use of Shapley values.

To reach our research objective, we employ a unique and proprietary dataset comprising granular firm-level data on a panel of Italian SMBs over the period 2015–2017 for which we match financial and economic data from the Bureau van Dijk Orbis database.

Our research question represents a matter of concern to policymakers since inaccurate credit risk measurement could threaten the stability of the banking sector, undermining the pivotal intermediation role played by banks in the economy in efficiently allocating resources to the most promising businesses. This assumes even greater relevance in light of the current COVID-19 crisis. Indeed, in periods of financial distress, an accurate credit risk assessment would allow banks to better forecast ex-ante corporate default probability.

The remainder of the paper is structured as follows. In Section 2 we review the existing literature, highlighting the novelty and the contribution of our approach. In Section 4 we present the empirical methodology, that relies on the comparison of a traditional probit model and of the historical random forest, chosen as the most appropriate ML technique in this specific framework. In Section 3 we describe the dataset construction, the sources used and the key variables employed in the analyses. In Section 5 we present and discuss our main results and compare the two methods employed. Finally, in Section 6, we discuss the limitations of our studies and provide hints for future research, while in Section 7 we draw our conclusions.

2. Related literature and contribution

Within the existing literature, the application of alternative methodologies for estimating SMB credit ratings, such as data mining techniques, tree-based methodology, AI [19–21] or other hybrid methods [22,23] have become relatively widespread ([24] for a detailed discussion). More recently, the latest wave of digitalization in financial markets, i.e., Fintech, has contributed to unprecedented technological development and an increase in the number and variety of new statistical methodologies applied to the financial sector. Indeed, banks have started to explore the implementation of advanced estimation techniques for SMB credit risk evaluation, although the adoption of machine learning and AI algorithms is still not fully permitted by regulators [25–27]. As a matter of fact, machine learning (ML) techniques can introduce biases in lending behaviour at the risk of financial inclusion and may entail issues related to consumer protection, ethics, privacy, and transparency in the eyes of supervisors and policymakers [28,29]. Indeed, ML results can be harder to interpret and explain to the various stakeholders [30–32]. Therefore, SMB credit rating estimation has gained renewed attention lately, also thanks to the availability of new statistical techniques and different data sources that complement the basic information available on SMBs to reach a more accurate assessment of SMB credit risk.

On the one hand, we start from the existing literature and follow a path of continuity with [16–18,33] in terms of comparison between two types of default forecasting techniques, i.e., statistical (parametric approach) and ML models (non-parametric approach). Moscatelli et al. [17], using data on financial and credit behavioural indicators for Italian non-financial firms, present better forecasting performance with the employment of ML models, although this gain is minimal when high-quality information, i.e., credit behavioural features, is added to training data and becomes negligible if the dataset is small. Overall, their results suggest that the ML-based credit allocation rule results in lower credit losses for lenders. [16] apply Random Survival Forests to compare their relative performance to a standard logit model and find that, while the latter outperforms the former in terms of out-of-sample accuracy, the opposite holds for in-sample accuracy. More recently, an array of machine learning methods has been compared to logistic regression by [18]. The findings of the authors suggest that ML models perform especially well when information is limited. [33] offer a theoretical framework for the correct comparison of the two alternative models fitting through a modified resampling scheme.

On the other hand, we depart from the existing studies and provide a novel contribution to this stream of literature along three dimensions. Firstly, we extend [17] data comparison in terms of model discriminatory power by making use of granular micro-level data collected from a European investment bank and an international insurance company. Secondly, while previous studies have applied static credit scoring models to analyse the key determinants of firm credit ratings, we apply a static (or pooled) and a historical modelling framework. Specifically, historical models are introduced to analyse persistence in credit rating and compare the predictive power of the two approaches, i.e., ordered probit and Historical Random Forest (HRF). Thirdly, the lack of explainability in models with high prediction performance, i.e. ML models, has been addressed with an innovative model-agnostic interpretation approach of results known as SHAP (SHapley Additive exPlanations). Specifically, as reported in previous works [16], while permutation feature importance helps in making comparisons among features easily, it does neither show how much each feature weights nor identify the impact of features with medium permutation importance. In this regard, the Shapley explainer is crucial to correctly understanding the positive or negative contribution of a feature value to the difference between the actual and the mean prediction. This contribution extends the notion of permutation feature importance and SHAP to a pooled and historical setting for an ordered probit model and Historical Random Forest (HRF) approach.

3. Data

The dataset employed in this paper is derived by a proprietary dataset on 464 SMBs made available by a European investment bank which plays a leading role in the niche of revolving trade receivables' securitization programmes.[2]

We gain access to securitization variables (hereinafter referred to as SEC variables) from the investment bank, including the credit rating. The credit rating is provided by an insurance company in the process of the securitization programme and includes not only balance sheet information but also private information (soft information) that the

[2] The European investment bank represents the investment bank of a large multinational European banking group, with total assets of 646 billion euros, a total market capitalization of about 50 billion euros and subsidiaries in twelve central-eastern European and Mediterranean countries. In the home country, the group has 14 affiliated banks and about 4500 branches covering a market share of about 15% in the loan and deposit markets. In particular, the investment bank offers products and services in the home country through a network of Corporate Offices, coordinated by Territorial Areas, and relationship structures dedicated to Financial Institutions. Abroad, the investment bank is present in more than 20 countries, supporting the cross-border activities of both the bank's national and international clients, with a specialized foreign network consisting of Branches, Representative Offices and Subsidiaries that carry out corporate and investment banking activities. It can be assumed, therefore, that possible bank-specific idiosyncratic issues that may characterize the single banking organization are mitigated by the representativeness and geographical scope of the bank.

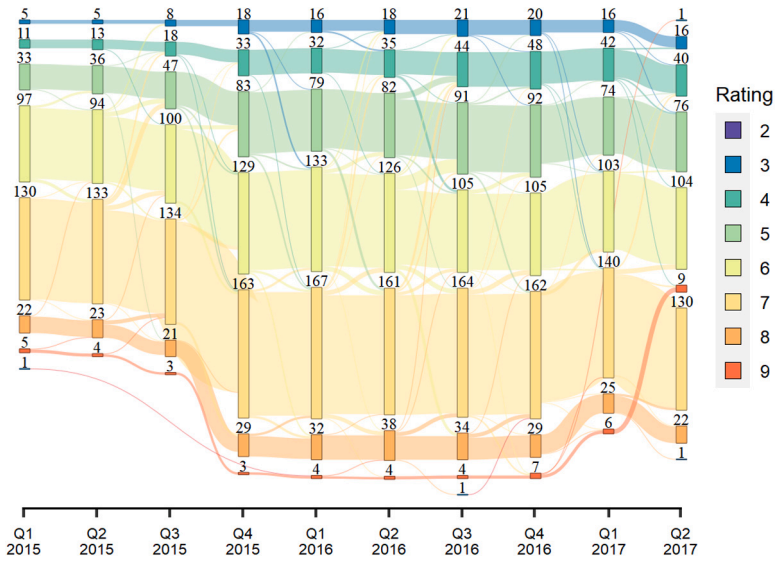


Fig. 1. Rating evolution over time.

insurance company is able to collect via different methods, including onsite inspections and special investigation teams. Overall the final rating assigned is the result of different sources of data and knowledge acquired by the insurance company (i.e., partnerships, registered payment defaults, credit reference agencies, accounting data, payment performance data, and network of risk information).

The insurance rating assigned to each SMB is categorized on a numeric scale ranging from 2 to 9 according to the given firm credit risk. The higher the number, the worse the credit rating. Credit rating evolution over time is shown in Fig. 1, highlighting an overall persistent behaviour for all classes of risk. SEC data is available for a total of 10 quarters (from Q1 2015 to Q2 2017).

Besides, we complement the information obtained by the European investment bank, by collecting accounting information by matching the company on the Orbis database, developed by Bureau Van Dijk (a Moody's Analytics company). Balance sheet and economic data are retrieved for the same period, but show an annual frequency, due to the nature of the balance sheet information (hereinafter referred to as BS). Annual values are hence repeated over all quarters of each year to mimic the frequency of the SEC variables.

We furthermore collected NACE Rev. 2 Codes, to classify firms' main sector (NACE) and main division (Industry). Geolocalization variables have been extracted through Google Maps API and have been linked to each SMB in the dataset to control for unobserved heterogeneity in the given SMB's industry and location. Table 1 reports the definition of the variables used in the empirical analysis. The variables are chosen according to past empirical literature that identifies the main drivers of firms' creditworthiness and that influence companies' capital structure. For instance, firms with higher profitability, higher revenues, better efficiency scores, stronger asset or employment growth, and higher capital ratios tend to be more robust to shocks and have a lower probability of default [34–36]. Table 2 shows the main descriptive statistics.

The final dataset consisted of 464 firms and 21 variables, 6 SEC and 15 BS, and was then treated according to an unbalanced panel data structure, resulting in 3009 rows. The final dataset is the result of a data cleaning process, removal of missing observations[3] Dummy and categorical variables distribution by each rating is reported in Table 3. The geographical distribution of SMBs is shown in Fig. 2. It clearly appears that we have SMBs spread all over the territory under of the country under investigation, moreover the majority of the Manufacturing firms is located in the North while wholesale and retail trade ones are more homogeneously diffused on the territory.

[3] We removed firms with more than 10% of missing variables and firms with null securitization data. The remaining missing values were imputed through the strategy reported in Appendix B and the removal of highly correlated variables before running the empirical analysis according to the value of the Variance Inflation Factor (VIF).

4. Methodology

With the aim of comparing traditional parametric models with Machine learning methods, we describe the chosen methodologies employed in the paper. First, we clarify that, given the longitudinal nature of the data, a comparison of models has been performed along two dimensions: a *pooled (or static)* versus a *historical* framework tested both in a *parametric* and a *non-parametric* context.

In the pooled setting the target rating at time t is regressed on BS and SEC variables at the same time t, whilst in the historical setting both target and independent variables at time s, with $s < t$, are added as additional regressors.

The target rating in our framework can take any integer value from 2 to 9, where 2 is the score associated with the lowest credit risk. The target variable has been firstly modelled through the following pooled ordered probit model:

$$y_{it} = \mathbf{X}_{it}\beta + \alpha_i + \epsilon_{it},$$

where $y_{it} \in [2, 9]$ is an observed index of credit quality for the ith firm at tth quarter, $i = 1, \dots, N$ and $t = 1, \dots, T$, $\mathbf{X}_{it}$ indicates a vector $1 \times k$, where $k = 21$, of explanatory variables for ith firm at time t, β is a $k \times 1$ vector of unknown parameters to be estimated, α_i is a firm-specific and time-invariant component and ϵ_{it} is the disturbance term which is assumed to be normally distributed.

Given the ordinal nature of the target variable, an ordered probit has been selected as the parametric model and the Random Forest, including its historical version, as the non-parametric one.

Several studies pointed out that rating changes tend to exhibit serial correlation [37,38] and that agencies seem to be slow to react to new information [39]. Therefore, the model has been extended to the historical framework, adding the delta values of the dependent variable, that is the difference between the current quarter (or year) and the previous one. The resulting model can be interpreted as a first-order Markov process and, following [40–42], is defined as:

$$y_{it} = X_{it}\beta + \Delta y_{it}\gamma + \alpha_i + \epsilon_{it},$$

Table 1
List of final variables.

Variable	Description
Rating	Rating score, 3 means low risk
Purchase	Accounting of Cash and Credit purchases
Current liabilities	Company's debts or obligations that are due to be paid to creditors within one year
Delinquency	Dummy variable equal to 1 if the firm misses a scheduled payment on an invoice, otherwise equal to 0
EBIT	Company's net income before income tax expense and interest expenses are deducted
Collections	Amount of invoices currently sold to the bank
Liquidity	Company's ability to pay off current debt obligations without raising external capital
Outstanding	Amount of securitization transactions in which the borrowing firm is involved, expressing its economic exposure in logarithmic scale (base 10)
Turnover	Annual sales volume net of all discounts and sales taxes in logarithmic scale (base 10)
LT Debt	Debt with maturities greater than 12 months
New Receivables	Monetary amount of receivables sold to the bank with respect to a given borrowing firm at the current invoices' transfer
Profit Margin	Percentage of sales turned into profits
Profit per employee	Net Income for the past twelve months (LTM) divided by the current number of Full-Time Equivalent employees
ROA	Net income divided by total assets
ROE	Fiscal year net income divided by total equity
Solvency_A	Firm's capacity to meet its long-term financial commitments
Tangibles	Assets that have a physical value
Working Capital	Difference between a company's current assets and its current liabilities
Delinquency Severe	Dummy variable equal to 1 if Delinquency is larger or equal than +2 standard deviations from the mean of all clients
Delinquency 90	Dummy variable equal to 1 if Scaduto90 (i.e., payments overdue by more than 90 days evaluated on average by ID) is larger than 0, otherwise equal to 0
NACE	Statistical Classification of Economic Activities in the European Community
Region	Geographical macro-areas

Table 2
Main descriptive statistics of numerical variables employed.

Variable	Mean	Stdev	Median	Minimum	Maximum
Rating	5.1091	1.2443	5	2	9
Purchase	1.4914	0.9555	1.3062	0.0168	6.4811
Current liabilities	0.5480	0.1996	0.5457	0.0383	2.0324
Delinquency	0.0162	0.1043	0	0	1
EBIT	0.0485	0.0881	0.0400	−1.4438	0.6867
Collections	2.7146	90.7025	0.7687	0	5520.7044
Liquidity	0.0104	0.0078	0.0091	0.0008	0.1594
Outstanding	4.1590	1.9485	4.6758	0	7.1786
Turnover	4.5325	0.8341	4.4351	2.8520	6.9362
LT Debt	0.0911	0.1035	0.0540	0	0.5163
New Receivables	0.2060	0.2355	0.1634	0	1
Profit Margin	0.0240	0.0643	0.0174	−0.7288	0.5611
Profit per employee	0.0047	0.0502	0.0004	−0.0178	1
ROA	0.0318	0.0887	0.0210	−0.3528	1.9188
ROE	0.0822	0.6385	0.0831	−13.7168	9.7300
Solvency_A	0.2834	0.1843	0.2468	−0.7866	0.9333
Tangibles	0.2477	0.1931	0.2121	0	0.9797
Working Capital	0.1372	0.2414	0.1195	−1.7193	1.0661

where $\Delta y_{it} = y_{it} - y_{i(t-1)}$ indicates the ith firm difference of rating between two consecutive quarters, γ represents the parameters linked to that difference.

Both pooled and historical version of model have been implemented using *R* package `oglmx` [43].[4]

Random forest (RF), introduced by [44], is a non-parametric learning method based on the ensemble of decision trees, which represents one of the state-of-the-art machine learning methods for prediction and classification [45]. The pooled version of the model uses the classic implementation of RF where the target variable y_{it} of the ith firm at quarter t is predicted by the dependent variable X_{it} as described in the probit model. Given the ordinal nature of the target variable, the classification version of RF has been considered. The historical version, HRF, makes use of flexible summary functions that capture the time-dependency for each variable, without the explicit use of lags or deltas. A detailed description of the algorithm is reported in Appendix A.1 and the model has been implemented by the *R* package `htree` [46].

In order to evaluate the performance of both models and to select the optimal subset of variables of the probit model, a set of evaluation metrics has been taken into consideration. First of all the confusion matrix has been used to assess the accuracy of the prediction of each rating class and the F_1-score was selected as an aggregated metric.[5]

Validation of model performances and train/validation set splitting of the data have been evaluated with a variable-length rolling-window temporal approach.[6]

[4] https://cran.r-project.org/web/packages/oglmx/index.html; last access 29th May 2023.

[5] The difference of performances on train and validation set must be minimized when tuning the hyperparameters so that overfitting can be avoided. Therefore, a weighting adjustment on the F_1-score has been selected among the following:

- $F_{1\,\text{ratio}} = F_{1\,\text{test}} + \frac{F_{1\,\text{test}}}{\Delta F_{1\,\text{train-test}}}$
- $F_{1\,\text{harmonic}} = \frac{2}{\frac{1}{F_{1\,\text{test}}} + \frac{1}{\Delta F_{1\,\text{train-test}}}}$
- $F_{1\,\text{cross-entropy}} = -F_{1\,\text{test}}^{\gamma}\log(1 - F_{1\,\text{test}}) - (1 - \Delta F_{1\,\text{train-test}})^{\gamma}\log(\Delta F_{1\,\text{train-test}}), \gamma \geq 1$

The most efficient weighting resulted to be $F_{1\,\text{cross-entropy}}$ with $\gamma = 4$.

[6] In particular, given that the maximum number of available quarters is 10 and the non-constant number of total quarters for each firm, a validation set of the 2 most recent quarters and a train set of all remaining quarters have been chosen. As the minimum number of available quarters for each firm is 7 and the minimum number of observations in each train set has been fixed to 10, the final number of folds used in the cross-validation is 4.

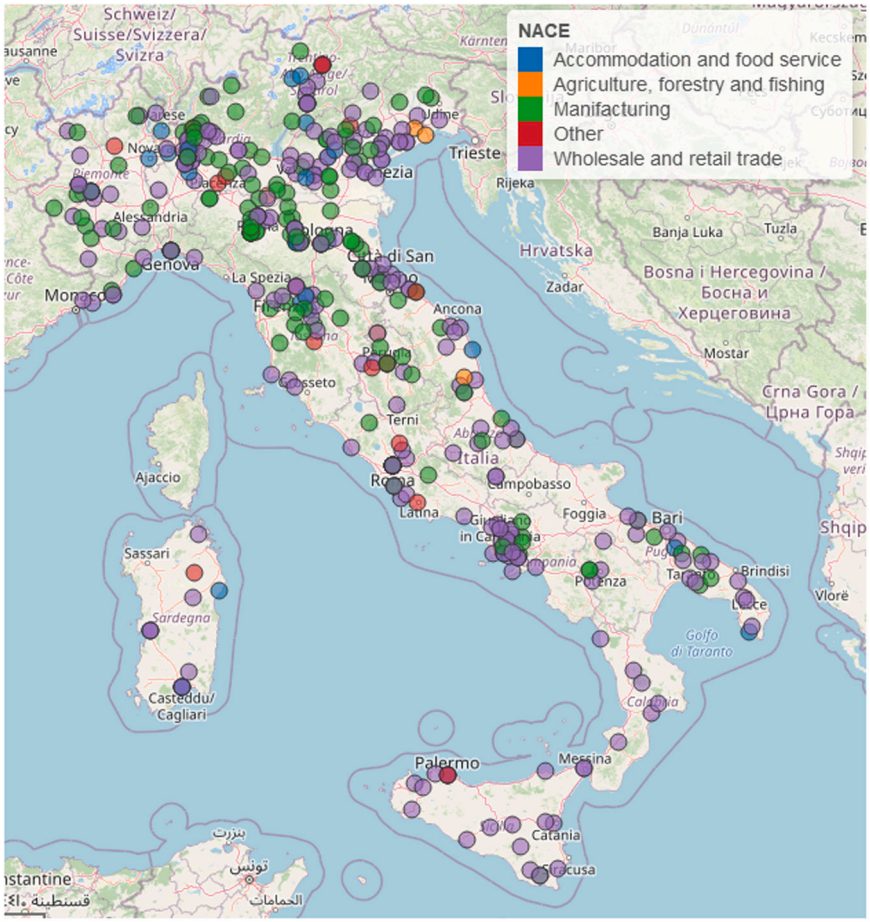


Fig. 2. Geographical distribution of firms, generated with *R package* `mapview`.

A comparison of all the fitted models has been considered in terms of variables' predictive power, using two relevant state-of-the-art techniques: Permutation Feature Importance (PFI) and SHAP values. Both methods aim to estimate the importance of each variable determining the most relevant ones for the prediction.

In the PFI the importance of each feature is evaluated by computing the gain in the model's prediction error after shuffling the feature's values. A feature is considered relevant for the model's prediction if the prediction error increases after permuting its values, otherwise, if the model error remains unchanged, its contribution is not important.

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 (see Appendix A.2 for further details). 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. It is important to remark that the SHAP values have been computed for this multiclass problem in order to investigate, for each class, how the predictors bring up or down the probability of belonging to a certain class, compared to the average probability of this class for the full data.

5. Empirical analysis

In this section, a comparison of the classification performance of models along three dimensions is presented. As introduced in Section 4, models have been distinguished according to *pooled* versus *historical* framework, *parametric* versus *non-parametric* approach and *BS* vs *SEC* set of predictors. Model evaluation has been made in terms of macro-averaged F_1-score on both in-sample and out-of-sample predictions.

5.1. Model evaluation

A set of evaluation metrics has been used in order to obtain an optimal combination of hyperparameters (RF model) and variables (PB model). Respectively, F_1 cross-entropy[7] metric has been maximized during the cross-validation phase to avoid overfitting; Akaike Information Criterion (AIC) has been minimized during best subset selection to obtain a stable function of predictors.

Table 4 shows the optimal hyperparameters set and the selected set of predictors with reference to both the BS and SEC sets of predictors. Based on the shown model architecture, a summary of classification performance with regards to both the training and validation sample is shown in Table 5.

F_1-score with regards to RF (historical) can be highlighted as the lowest ones for both BS and SEC set of predictors, even if the distinction is more evident for the latter set of variables. Predictors' history seems to be necessary for a correct classification of insurance credit risk. Historical RF outperforms the other estimated models, with around 90%–70% F_1-score respectively on train and test set for BS set of variables and 70%–50% F_1-score for SEC set of variables. As expected given the lower number of predictors, lower performance is reported by the SEC set of variables.

5.2. Model explanation

In this section, the explainability capabilities of both HRF and PB have been compared using PFI and SHAP values. In the first case,

[7] See note 5.

Table 3
Frequency table of categorical variables employed.

Variable	Description	Value	Frequency
Delinquency Severe	Payment behaviour (missed scheduled payment, 0=no, 1=yes)	0	2791
		1	125
Delinquency 90	Payment behaviour (payments overdue by more than 90 days, 0 = no, 1 = yes)	0	2043
		1	873
NACE (reference category = Manufacturing)	Wholesale and retail trade; repair of motor vehicles and motorcycles	0	2217
		1	1056
	Accommodation and food service activities	0	1500
		1	1773
	Agriculture, forestry and fishing	0	2953
		1	320
	Other	0	3194
		1	79
Region (reference category = North-East)	North-West	0	2194
		1	1079
	Center	0	2357
		1	916
	South+Islands	0	2583
		1	690

the predictive power of each feature has been evaluated whilst, in the second one, more complex relationships have been investigated through SHAP values. According to classification performance, feature importance figures with reference to the best statistical model (historical RF) for the two sets of variables have been reported in Fig. 3.

With regards to PFI, relative importance has been computed as the difference between the original and the permuted F_1-score then averaged and normalized over the sum of the absolute values of all the obtained permutation metrics. This procedure results in a range of values between 0%–100%, with a negative score when a random permutation of a feature's value results in a better performance metric and high importance score when a feature is more sensitive to random shuffling, i.e., it is more "important" for prediction. In the process of selecting the most important predictors, the features are considered, individually, in terms of relative importance ranking and, on an aggregated level, in terms of the total percentage of relative importance carried by the features in the top position. Related figures are presented on a macro-level (aggregated for all Rating classes).

PFI helps to easily make comparisons between features but it does not tell how each feature matter and does not allow the identification of the impact of features with medium permutation importance. The Shapley explainer is crucial to correctly understand why a model predicts a given class for a given ID on a given period (single row-prediction pair), since it goes through the input data, row-by-row and feature-by-feature, changing its values to identify how the base prediction differs holding all else equal for that row and, as a consequence, explains how this prediction was reached. The contribution of each variable towards the single row prediction compared to the base prediction for the full data set is called Shapley value (*phi*). On a multiclass perspective, SHAP will output a separate matrix for each class prediction for the given row in order to understand how, for each class, the predictors bring down or up the probability of belonging to that specific class. The Shapley values of each feature have been aggregated in two ways based on the average contribution computed by the feature and grouped according to rating classes with the aim of investigating how each feature impacts, on average, the predicted probability of each class compared to the average probability of this class for the full dataset.

Starting from the BS set of variables, Fig. 3(a) shows the importance ranking in terms of PFI for historical RF model. Turnover can be observed in the top position with 22% of relative importance value, followed by Solvency_A (13%), Profit Margin (9%) and Working Capital (8.8%) with a lower order of magnitude. On an aggregated level, the four previous features represent almost the 50% of relative feature importance over the total of 13 considered predictors. Turnover, Solvency_A and Profit Margin are strictly related to the risk profile of the investigated firms, since high values for size, liquidity and profitability measures represent a signal of solid financial and operational performance, increasing the probability of belonging to low-risk classes. Specifically, high liquidity implies a better ability of the company to meet its short-term obligations on time, resulting in lower debt and, consequently risk; associated with a healthy profile, the efficiency of the management and the annual sales volume as signals of firm expansion and consolidated business model. Regarding the key indicator of the financial solvency of the company, i.e., Working capital, an increment of this metric implies a positive impact on the probability of belonging to high-risk classes. Higher long and short-term financial obligations reflect higher debt and, consequently, higher risk. In the end, it can be concluded that quantitative variables are more important than qualitative ones, i.e. NACE and Region, since each dummy has a frequency that affects its importance value.

Furthermore, Shapley values (Fig. 3(b)) confirm previous results, highlighting the high average contribution of Turnover, together with Solvency_A, Purchase and ROA. It can be noticed that these features report higher SHAP value with respect to high-risk Rating class 6.

Following the same computational procedure for the SEC set of variables, it can be noticed that a relevant role is played by Outstanding (89%), carrying almost all permutation feature importance within historical RF framework (Fig. 3(c)). Outstanding represents a metric of economic exposure of the firms under investigation with respect to securitization transactions in which the borrowing firm is involved. This metric is directly linked to the level of risk reported by each firm.

SHAP results allow us to grasp the contribution of securitization variables on each rating class (Fig. 3(d)). The same conclusions can be reported in terms of the magnitude of feature importance, with Outstanding on the top of all variables with the highest average impact on Rating classes. However, Delinquency and Delinquency 90 report slightly higher metrics compared to the other variables, with higher SHAP value with respect to high-risk Rating class 6. Also, these variables represent metrics of economic exposure like Outstanding, but in terms of missed payments.

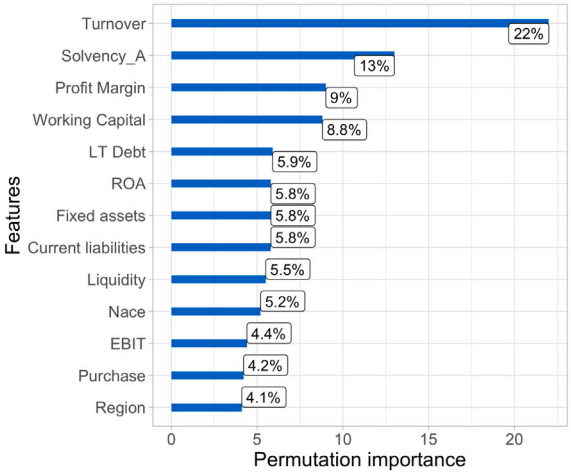
5.3. Assessment of differences and robustness checks

Additional checks have been performed to test the robustness of previous findings, in particular alternative formulation of the target variable. The latter test attempts to reduce the multiclass problem to multiple (or single) binary classification problems (one class vs the others or high-risk vs low-risk class) in order to check the accuracy of results in comparison to the ordinal formulation of the target variable.

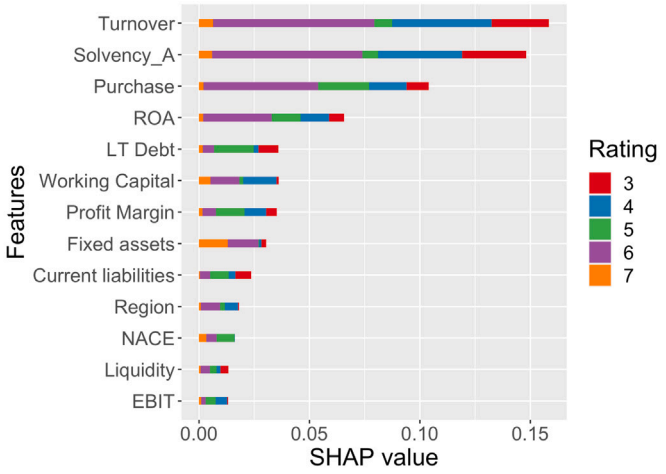
Given the complexity of the classification problem at hand and the subtlety of the different behaviours that the classifiers exhibit, the ordinal scale has been converted to a dichotomous variable. Firstly, a formulation Rating 7 vs ALL has been implemented, resulting in poor performances given the imbalanced nature of the dataset with respect to the tails. Then, the target variable has been defined as

Table 4
Model architecture for BS set of predictors.

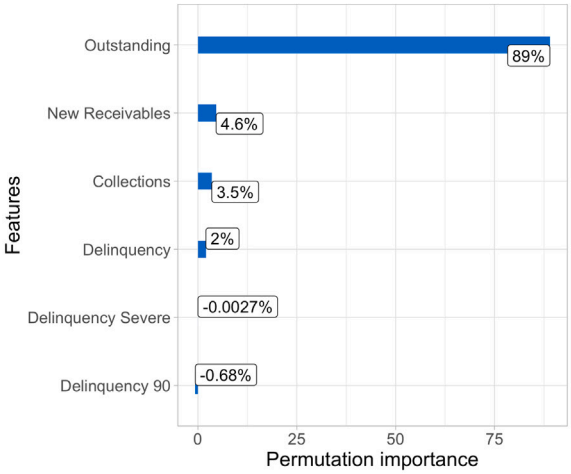
Model	Version	Set	Hyperparameters or Selected set of predictors
RF	Pooled	BS	Mtry = 14; Ntrees = 500; Nodesize = 1
		SEC	Mtry = 5; Ntrees = 10; Nodesize = 100
	Historical	BS	Mtry = 6; Ntrees = 50; Nodesize= 3; Method = "meanw0"
		SEC	Mtry = 4; Ntrees = 141; Nodesize = 89; Method = "mean0"
PB	Pooled	BS	Current liabilities + Liquidity ratio + LT Debt + ROA+ Tangibles + Working Capital + Purchase + Turnover + Region + NACE
		SEC	New Receivables + Outstanding + Delinquency
	Historical	BS	Current liabilities (delta) + Liquidity (delta) + LT Debt (delta) + Working Capital (delta) +Purchase (delta) + EBIT (delta) + Turnover (delta) + Region
		SEC	Collections (delta) + Outstanding (delta) + Delinquency (delta)



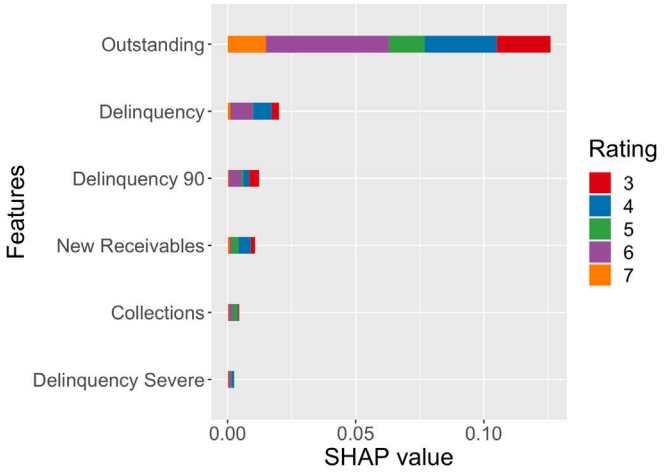
(a) Macro-averaged relative permutation importance with regards to BS set.



(b) SHAP values (average impact of predictors for each class) with regards to BS set.



(c) Macro-averaged relative permutation importance with regards to SEC set.



(d) SHAP values (average impact of predictors for each class) with regards to SEC set.

Fig. 3. Feature importance for RF model (historical). The top row reports results for the BS set, bottom row reports results for the SEC set. The left column reports macro-averaged relative permutation importance, right column reports SHAP values.

High-risk Rating (class 6 and 7) compared to all the other classes, in order to check if the models are able to more accurately price risk and differentiate between lower and higher credit risk borrowers. The descriptive analysis highlights a balanced distribution of observations in the two groups for both the considered set of predictors. Overall, the alternative formulation of the target variable affects positively the SEC case since the classification metrics are slightly higher (+0.1) compared to the ordinal one. For the other set, the performances are almost the same, except for the PB case where the metrics are better with the binary target. The selected set of variables, for the PB model, is the same and the marginal effects of the binary cases reflect exactly the duality into the sign of the partial derivatives for the ordinal case since the threshold that highlights the change of sign is class 6. To summarize, the binary formulation simplifies the classification problem

Table 5
Macro-averaged F_1-score on training and test sample for all sets of predictors.

Model	Version	Sample	F_1-score BS	SEC
RF	Pooled	Train	0.919	0.411
		Test	0.677	0.342
	Historical	Train	0.915	0.748
		Test	0.736	0.552
PB	Pooled	Train	0.477	0.325
		Test	0.466	0.302
	Historical	Train	0.478	0.339
		Test	0.472	0.324

at hand and results in slightly higher performances together with the same explainable conclusions as for individual risk.

6. Limitations and future research

Despite the empirical evidence, the research has also some limitations, that we now discuss. First, as highlighted in the Data Section, we have access to 10 quarters, from Q1 of 2015 to Q2 of 2017. The dataset employed suffers two main issues: (i) the relatively short period; and (ii) the period covered. Because of this, our results can be more easily extended to similar periods, i.e., times of relatively calm market conditions, with no particular distress in terms of business cycle, inflation, or economic uncertainty. Results might change, instead, in times of unstable credit ratings, due to unstable market conditions. This would represent one possible extension of our paper. Namely, future studies could test the capability of HRF to predict credit risk also with rising prices, rising interest rates or external unexpected shocks, e.g., shocks such as COVID-19 or other major business disruptions, also due to extreme weather events, favoured by climate change.

Despite the limitations of the dataset, the latter also represents a value-added to the investigation. Indeed, the dataset obtained by the European investment bank is not publicly available and allows us to take an uncommon point of view, i.e., the financial institution or lender's point of view. Besides this, the European investment bank represents the investment bank of a large multinational European banking group, with total assets of 646 billion euros, a total market capitalization of about 50 billion euros and subsidiaries in twelve central-eastern European and Mediterranean countries. In the home country, the group has 14 affiliated banks and about 4500 branches covering a market share of about 15% in the loan and deposit markets. Our results can be therefore extended to a certain extent to other banks, being our data highly representative of the banking industry. Additionally, differently from the mainstream literature, our study investigates a relatively unexplored set of companies, i.e., small- and mid-sized businesses.

Second, the choice of the machine learning approach. In the paper, we employ only the Historical Random Forest as an ML technique in this paper, due to the limited availability of ML methods able to consider explicitly the panel dimensions. The choice is therefore due to the will to find a suitable replicable technique able to treat the type of dataset we have, but future studies could extend the size of the dataset to make the applicability of neural networks feasible. Other studies could also be devoted to the development of other ML algorithms able to integrate the time/panel dimension of the data.

Third, despite having access to the credit rating issued by the insurance company, we are not able, at this stage, to extract the amount of soft information available to the financial intermediary. The rating, in fact, is also the result of onsite inspections and the long relationship with the customers. Further studies could be devoted to understanding if and how ML methods are also able to treat the soft information embedded in credit ratings for SMBs and whether they are suitable to codify or harden soft information to provide more reliable credit ratings.

7. Conclusions

By employing a unique and proprietary dataset comprising granular firm-level securitization and accounting data on a panel of 464 Italian SMBs over the period 2015–2017, this paper tests two alternative approaches grounded in statistical learning and machine learning frameworks and compares their respective capability in predicting SMB credit risk. Specifically, we compare a classic parametric approach fitting an ordered probit model with a non-parametric one, calibrating a machine learning Historical Random Forest (HRF) approach. Both models are implemented according to a pooled and a historical framework. Moreover, we further assess the relevance of each variable to predict SMB credit risk, through the use of Shapley values.

Our results provide evidence that the Historical Random Forest (HRF) approach outperforms the traditional ordered probit model in assessing SMBs' credit risk. This shows that advanced machine learning methodologies can be successfully adopted by banks to predict SMB credit risk, highlighting the opportunity to complement traditional methods with more advanced estimation techniques that rely on machine learning.

Our research question represents a matter of concern to policymakers since inaccurate credit risk measurement could threaten the stability of the banking sector, undermining the pivotal intermediation role played by banks in the economy. This assumes even greater relevance in light of the latest COVID-19 crisis. Indeed, in periods of financial distress, an accurate credit risk assessment would allow banks to better forecast ex-ante corporate default probability.

This paper paves the way for future and unforeseeable research in this area. Future extensions of this work could involve not only applying alternative machine learning methods but also testing whether the latter could successfully predict and "harden" soft information, thus eventually substituting for the traditional role of relationship banking in small business lending.

CRediT authorship contribution statement

Alessandro Bitetto: Conceived the experiment(s), Conducted the experiment(s), Analysed the results, Reviewed the manuscript. **Paola Cerchiello:** Conceived the experiment(s), Conducted the experiment(s), Analysed the results, Reviewed the manuscript. **Stefano Filomeni:** Conceived the experiment(s), Conducted the experiment(s), Analysed the results, Reviewed the manuscript. **Alessandra Tanda:** Conceived the experiment(s), Conducted the experiment(s), Analysed the results, Reviewed the manuscript. **Barbara Tarantino:** Conceived the experiment(s), Conducted the experiment(s), Analysed the results, Reviewed the manuscript.

Data availability

The data that has been used is confidential.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program "PERISCOPE: Pan European Response to the ImpactS of COvid-19 and future Pandemics and Epidemics", under the Grant Agreement No. 101016233, H2020-SC1-PHE-CORONAVIRUS-2020-2-RTD.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.seps.2023.101746.

References

[1] Bank for International Settlements. International convergence of capital measurement and capital standards: a revised framework. 2006, URL https://www.bis.org/publ/bcbs128.htm.

[2] Cucinelli D, Di Battista ML, Marchese M, Nieri L. Credit risk in European banks: The bright side of the internal ratings based approach. J Bank Financ 2018;93:213–29.

[3] Altman EI. Commercial bank lending: process, credit scoring, and costs of errors in lending. J Financ Quant Anal 1980;813–32.

[4] Louzada F, Ara A, Fernandes GB. Classification methods applied to credit scoring: Systematic review and overall comparison. Surv Oper Res Manag Sci 2016;21(2):117–34.

[5] Blöchlinger A, Leippold M. Are ratings the worst form of credit assessment except for all the others? J Financ Quant Anal 2018;53(1):299–334.

[6] Berger AN, Udell GF. Relationship lending and lines of credit in small firm finance. J Bus 1995;35:1–381.

[7] Claessens S, Krahnen J, Lang WW. The basel II reform and retail credit markets. J Financial Serv Res 2005;28(1–3):5–13.

[8] OECD. Financing SMEs and entrepreneurs: An OECD scoreboard. Special edition: The impact of COVID-19. 2020, URL https://www.oecd.org/industry/smes/SMEs-Scoreboard-2020-Highlights-2020-FINAL.pdf.

[9] Kysucky V, Norden L. The benefits of relationship lending in a cross-country context: A meta-analysis. Manage Sci 2016;62(1):90–110.

[10] Casu B, Chiaramonte L, Croci E, Filomeni S. Access to credit in a market downturn. J Financial Serv Res 2022. http://dx.doi.org/10.1007/s10693-022-00388-x.

[11] Liberti JM, Petersen MA. Information: Hard and soft. Rev Corp Finance Stud 2018;8(1):1–41.

[12] Filomeni S, Udell GF, Zazzaro A. Hardening soft information: does organizational distance matter? Eur J Finance 2021;27(9):897–927. http://dx.doi.org/10.1080/1351847X.2020.1857812.

[13] Filomeni S, Udell GF, Zazzaro A. Communication frictions in banking organizations: Evidence from credit score lending. Econom Lett 2020;195C:109412.

[14] Filomeni S, Bose U, Megaritis A, Triantafyllou A. Can market information outperform hard and soft information in predicting corporate defaults? Int J Finance Econ 2023. http://dx.doi.org/10.1002/ijfe.2840, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ijfe.2840.

[15] Filomeni S, Modina M, Tabacco E. Trade credit and firm investments: empirical evidence from Italian cooperative banks. Rev Quant Financ Account 2023;60(3):1099–141. http://dx.doi.org/10.1007/s11156-022-01122-3.

[16] Fantazzini D, Figini S. Random survival forests models for SME credit risk measurement. Methodol Comput Appl Probab 2009;11:29–45.

[17] Moscatelli M, Narizzano S, Parlapiano F, Viggiano G. Corporate default forecasting with machine learning. Temi Di Discussione (Economic working papers) 1256, Bank of Italy, Economic Research and International Relations Area; 2019.

[18] Moscatelli M, Parlapiano F, Narizzano S, Viggiano G. Corporate default forecasting with machine learning. Expert Syst Appl 2020;161:113567.

[19] Olmeda I, Fernandez E. Hybrid classifiers for financial multicriteria decision making: The case of Bankruptcy prediction. Comput Econ 1997;10(4):317–35.

[20] De Andrés J, Landajo M, Lorca P. Forecasting business profitability by using classification techniques: A comparative analysis based on a Spanish case. European J Oper Res 2005;167(2):518–42.

[21] Lin S-W, Shiue Y-R, Chen S-C, Cheng H-M. Applying enhanced data mining approaches in predicting bank performance: A case of taiwanese commercial banks. Expert Syst Appl 2009;36(9):11543–51.

[22] Ahn BS, Cho SS, Kim C. The integrated methodology of rough set theory and artificial neural network for business failure prediction. Expert Syst Appl 2000;18:65–74.

[23] Hsieh N-C. An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Syst Appl 2004;27(4):623–33.

[24] Falavigna G. Models for default risk analysis: Focus on artificial neural networks, model comparisons, hybrid frameworks. CERIS working paper 200610, Institute for Economic Research on Firms and Growth - Moncalieri (TO) ITALY -NOW-Research Institute on Sustainable Economic Growth - Moncalieri (TO) ITALY; 2006.

[25] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable ai in fintech risk management. Front Artif Intell 2020;3(26).

[26] Bitetto A, Cerchiello P. Initial coin offerings and esg: Allies or enemies? Finance Res Lett 2023;57:104227. http://dx.doi.org/10.1016/j.frl.2023.104227, URL https://www.sciencedirect.com/science/article/pii/S1544612323005998.

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

[28] Bazarbash M. Fintech in financial inclusion: machine learning applications in assessing credit risk. IMF working paper, WPIEA2019109, 2019.

[29] Lee MSA, Floridi L. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds Mach 2021;31:165–91.

[30] Financial Stability Board. Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. 2017, URL https://www.fsb.org/wp-content/uploads/P011117.pdf.

[31] World Bank Group. Credit scoring approaches guidelines. Technical report, 2019, URL https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf.

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

[33] Haerdle W, Mammen E. Comparing nonparametric versus parametric regression fits. Ann Statist 1993;21(4):1926–47. http://dx.doi.org/10.1214/aos/1176349403.

[34] Altman EI, Sabato G. Modelling credit risk for SMEs: Evidence from the US market. Abacus 2007;43(3):332–57.

[35] Dainelli F, Giunta F, F Cipollini. Determinants of SME credit worthiness under Basel rules: the value of credit history information. PSL Q Rev 2013;66(264):21–47.

[36] Corazza M, Funari S, Gusso R. Creditworthiness evaluation of Italian SMEs at the beginning of the 2007–2008 crisis: An MCDA approach. North Am J Econ Finance 2016;38:1–26.

[37] Carty LV, Fons JS. Measuring changes in corporate credit quality. J Fixed Income 1994;4(1):27–41.

[38] Gonzalez F, Haas F, Johannes R, Persson M, Toledo L, Violi R, Wieland M, Zins C. Market dynamics associated with credit ratings. A literature review. Occasional paper 16, European Central Bank; 2004.

[39] Odders-White E, Ready M. Credit ratings and stock liquidity. Rev Financ Stud 2006;19:119–57.

[40] Contoyannis P, Jones A, Rice N. The dynamics of health in the British household panel survey. J Appl Econometrics 2004;19:473–503.

[41] Wooldridge J. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. J Appl Econometrics 2005;20:39–54.

[42] Greene W, Hemsher D. Modeling ordered choices: A primer and recent developments. Working paper 26, New York University, Leonard N. Stern School of Business, Department of Economics; 2008.

[43] Carroll N. Estimation of ordered generalized linear models. 2018, URL https://CRAN.R-project.org/package=oglmx.

[44] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[45] Capitaine L, Genuer R, Thiébaut R. Random forests for high-dimensional longitudinal data. Stat Methods Med Res 2021;30(1):166–84.

[46] Sexton J. Historical tree ensembles for longitudinal data. 2018, URL https://CRAN.R-project.org/package=htree.

Alessandro Bitetto is a post-doc researcher interested in Dimensionality Reduction techniques for the construction of synthetic indexes. He works on mixed frequency spatio-temporal data by the means of neural networks, namely recurrent networks on time dimension and graph neural networks on the spatial dimension. He makes use of feature explainability techniques.

Paola Cerchiello is Full Professor of Statistics. She mainly focuses on methodological statistics and data analysis: she is currently working on text data models, systemic risk, financial technologies (fintech), big data analysis, ordinal variables, spatio-temporal models. She collaborates with the Bank of Italy on a Big Data analysis project.

Stefano Filomeni is a Lecturer in the Finance Group of Essex Business School (University of Essex). He holds a PhD in Economic Sciences issued by the Marche Polytechnic University. His research focuses on the areas of banking, including bankruptcy reforms, Islamic banking and bank lending.

Alessandra Tanda is Associate Professor of Financial Institutions at the Department of Economics and Management, University of Pavia. She holds a PhD in Financial Markets and Institutions issues by Cattolica University (Milan). Hermain research interests include FinTech, the financial structure of firms, green finance and corporate governance.

Barbara Tarantino is a PhD student in Electronic, Computer Science and Electrical Engineering. She holds a Master's Degree in Economics, Finance and International Integration from the University of Pavia. She has a background that includes mathematics, statistics and economics. Her research interests relate to Biostatistics, Bayesian statistics, SEM, Explainable AI and Fintech.