ISSN: 2281-1346

UNIVERSITÀ DI PAVIA
Department of Economics and Management

DEM Working Paper Series

Machine Learning and Credit Risk: Empirical Evidence from SMEs

Alessandro Bitetto
(Università di Pavia)

Paola Cerchiello
(Università di Pavia)

Stefano Filomeni
(University of Essex)

Alessandra Tanda
(Università di Pavia)

Barbara Tarantino
(Università di Pavia)

 # 201 (02-21)

Via San Felice, 5
I-27100 Pavia

economiaweb.unipv.it

Machine Learning and Credit Risk: Empirical Evidence from SMEs

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

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

Abstract

In this paper we assess credit risk of SMEs by testing and comparing a classic parametric approach fitting an ordered probit model with a non-parametric one calibrating a machine learning historical random forest (HRF) model. We do so by exploiting a unique and proprietary dataset comprising granular firm-level quarterly data collected from a large European bank and an international insurance company on a sample of 810 Italian small- and medium-sized enterprises (SMEs) over the time period 2015-2017. Our results provide novel evidence that a dynamic Historical Random Forest (HRF) approach outperforms the traditional ordered probit model, highlighting how advanced estimation methodologies that use machine learning techniques can be successfully implemented to predict SME credit risk. Moreover, by using Shapley values for the first time, we are able to assess the relevance of each variable in predicting SME credit risk. Traditionally, credit risk evaluation of informationally-opaque SMEs has relied on soft information-intensive relationship banking. However, the advent of large banking conglomerates and the limits to successfully "harden" and transmit soft information across large banking organizations, challenge the traditional role of relationship banking, urging the need to evaluate SME credit risk by implementing alternative methodologies mostly based on hard information.

Keywords: Credit Rating, SME, Historical Random Forest, Machine Learning, Relationship Banking, Soft Information

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

*Corresponding author
Email address: stefano.filomeni@essex.ac.uk (Stefano Filomeni)

1. Introduction

Determining corporate credit ratings is a well-known topic theoretically and empirically, both in the financial academic literature and in the industry (Altman, 1980, Louzada et al., 2016, Blöchlinger and Leippold, 2018). Within this topic, the corporate martket can be viewed as being composed of different segments. Among the latter, SMEs represent a large segment of the corporate market in several economies. As such, SME credit ratings have recently drawn the attention of academics and policy makers. This attention has led to a fervent debate on how to reach accurate estimation of SME credit risk. In this debate, the key features of SMEs are their informational opaqueness, greater perceived risk, and reliance on soft information-intensive relationship banking (OECD, 2020, Berger and Udell, 1995, Claessens et al., 2005). In this regard, the importance to incorporate soft information in SME credit risk assessment has been acknowledged by regulators. As a matter of fact, regulators have introduced the internal ratings-based (IRB) approach, that allows banks to include qualitative soft information when assessing corporate credit risk (Bank for international settlements, 2006, Cucinelli et al., 2018). However, the successful implementation of the internal ratings-based (IRB) approach is challenged by the presence of severe communication frictions; the latter limit the successful "hardening" and transmission of soft information across large banking organizations (Stein, 2002, Liberti and Petersen, 2018, Filomeni et al., 2020a, Filomeni et al., 2020b), challenging the traditional role of relationship banking. These communication frictions are even exacerbated when banks engage in M&A activity, that leads to the creation of large banking conglomerates mostly relying on transactional (rather than relationship) banking (Ferri and Pesic, 2017, Berger et al., 2005). This has spurred us to investigate alternative methodologies to evaluate SME credit risk, mostly based on hard information.

Except for a few studies implementing alternative methodologies (Fantazzini and Figini, 2009, Moscatelli et al., 2019), the literature has been mainly focused on the types of information a financial intermediary should use in assessing SME credit risk. This occurs at the expense of testing the performance of advanced statistical and machine learning techniques. Indeed, the high predictive capability of advanced methodologies (mostly based on hard information) would challenge

the role of soft information and mitigate those communication frictions that hamper the successful "hardening" and transmission of soft information.

To fill this gap, this paper tests two alternative approaches grounded in both statistical learning and machine learning, and compares their respective capability in predicting SME 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.

Our objective is to provide an alternative methodology that allows to reach accurate SME credit risk evaluation, by overcoming issues related to the transmission of soft 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 dynamic Historical Random Forest (HRF) approach. Moreover, we further contribute by assessing the relevance of each variable to predict SME credit risk, through the use of Shapley values.

By way of preview, our results provide novel evidence that a dynamic Historical Random Forest (HRF) approach outperforms the traditional ordered probit model in assessing SME credit risk. This highlights how advanced estimation methodologies, based on machine learning techniques and mostly on hard information, can be successfully implemented in predicting SME credit risk.

To reach our research objective, we employ a unique and proprietary dataset comprising granular firm-level data on a panel of 810 Italian SMEs over the time period 2015-2017. Particular relevance is attributed to SME credit ratings. The latter are assigned to SMEs by an insurance company in the context of a revolving trade receivables securitization program initiated by a large European investment bank in favour of some of its most valuable corporate clients. Indeed, SME credit ratings are firstly produced by the insurance company and then used by the bank. In this way, the latter can assess the credit risk of the acquired portfolio of securitized trade receivables originated by its valuable corporate clients. Securitization data are matched with accounting information on our sample of 810 Italian SMEs retrieved from Orbis database. The below-explained analogy between insurance and banking SME credit ratings makes the former suitable for the pur-

pose of our study. Indeed, both banking and insurance SME credit ratings are based both on hard and soft information. On the one hand, the former are based on relationship-intensive soft information collected directly and indirectly through continuous and personal bank-firm interactions. On the other hand, the latter are based on both proprietary soft information (i.e., client information, special investigation teams) and private and publicly-available hard information (i.e., partnerships, registered payment defaults, credit reference agencies, accounting data, payment performance data, network of risk information).

Our research question represents a matter of concerns to policy makers, 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 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. In Section 3 we present the empirical methodology. In Section 4 we describe data. In Section 5 we present and discuss our results. Finally, in Section 6, we conclude.

2. Related literature and our contribution

Within the existing literature, the application of alternative methodologies for estimating SME credit ratings, such as data mining techniques, tree based methodology, AI (Lin et al., 2009, Olmeda and Fernandez, 1997, De Andrés et al., 2005) or other hybrid methods (Ahn et al., 2000, Hsieh, 2004) have become widespread (Falavigna, 2006 for a detailed discussion). More recently, the latest wave of digitalization in financial markets, i.e., Fintech, has contributed to an 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 SME credit risk evaluation, although the adoption of machine learning and AI algorithms is still not fully permitted by regulators (Bussmann et al., 2020). As a matter of fact, machine learning techniques can introduce biases in lending behavior

at the risk of financial inclusion and may entail issues related to consumer protection, ethics, privacy, and transparency in the eyes of supervisors and policy makers (Bazarbash, 2019). Indeed, machine learning can be harder to interpret and explain to the various stakeholders (Financial Stability Board, 2017, World Bank Group, 2019). Therefore, SME credit rating estimation has gained a renewed attention lately, also thanks to the availability of new statistical techniques and different data sources that complement the basic information available on SMEs with the aim to reach a more accurate assessment of SME credit risk.

On the one hand, we start from the existing literature and follow a path of continuity with Moscatelli et al. (2019) and Fantazzini and Figini (2009) in terms of comparison between two types of default forecasting techniques, i.e., statistical (parametric approach) and machine learning (ML) models (non-parametric approach). Moscatelli et al. (2019), using data on financial and credit behavioral 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 behavioral features, is added to training data and becomes negligible if the dataset is small. Overall, their results suggest that ML-based credit allocation rule results in lower credit losses for lenders. Fantazzini and Figini (2009) 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.

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 Moscatelli et al. (2019) data comparison in terms of model discriminatory power by making use of granular micro-level data collected from a large European bank and an international insurance company. Secondly, while previous studies have applied static credit scoring models to analyze the key determinants of firm credit ratings, we apply a static and dynamic modelling framework. Specifically, dynamics are introduced to analyze 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 (Fantazzini and Figini, 2009), 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 understand 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 static and dynamic setting for an ordered probit model and Historical Random Forest (HRF) approach.

3. Methodology

Given the longitudinal nature of the data, a comparison of models has been performed along two dimensions: a *static* versus a *dynamic* framework and a *parametric* versus a *non-parametric* approach. In the static setting the target rating at time t is regressed with balance sheet and securitization variables at the same time t, whilst in the dynamic setting both target and independent variables at time s, with $s < t$, are added as additional regressors. Given the ordinal nature of the target variable, an ordered probit has been selected as parametric model and the Historical Random Forest as non-parametric one. The impact of standalone balance sheets and securitization variables has been further evaluated aside of the set including all variables, adding a third dimension to the comparison analysis.

The target rating has been firstly modeled through the following static ordered probit model:

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

where $y_{it} \in [2,9]$ is an observed index of credit quality for the i-th firm at t-th quarter, $i = 1,\ldots,N$ and $t = 1,\ldots,T$, $\mathbf{X}_{it}$ indicates a vector $1 \times k$, where $k = 21$, of explanatory variables for i-th 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.

Several studies pointed out that rating changes tend to exhibit serial correlation (Carty and Fons,

1994, Gonzalez et al., 2004) and that agencies seem to be slow to react to new information (Odders-White and Ready, 2006). Therefore, the model has been extended to the dynamic framework, adding the lagged values of the dependent variable. The resulting model can be interpreted as a first-order Markov process and, following Wooldridge (2005), Contoyannis et al. (2004) and Greene and Hemsher (2008), is defined as:

$$y_{it} = X_{it}\beta + y_{i(t-1)}\gamma + y_{i0}\delta + \alpha_i + \varepsilon_{it},$$

where $y_{i(t-1)}$ indicates the i-th firm rating in the previous quarter, γ represents the parameters linked to rating in the previous quarter, y_{i0} is the first available firm rating, at time $t = 0$. Both static and dynamic version of model have been implemented using *R* package `oglmx` (Carroll, 2018).

Random forest (RF), introduced by Breiman (2001), is a non-parametric learning method based on the ensemble of decision trees, which represents one of the state-of-the-art machine learning method for prediction and classification (Capitaine et al., 2019). The static version of the model uses the classic implementation of RF where the target variable y_{it} of the i-th 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 used.

The first approaches dealing with longitudinal and clustered data involved tree-based methods (Segal, 1992, Hajjem et al., 2014, Sela and Simonoff, 2012) and are based on the idea of iterating between fixed and random part and estimating the parameters via Expectation Maximization (EM) algorithm. All these approaches represent semi-parametric fixed effects model in which the non-parametric part is evaluated through RF. The main contribution of this paper regards the application of an innovative random forest algorithm based on historical trees, suitable for longitudinal data and implemented in the *R* package `htree` (Sexton, 2018).

Let us consider longitudinal data with n individuals, the i-th individual having n_i observations over time. Specifically, data is assumed to be of the form

$$z_{ij} = (y_{ij}, t_{ij}, x_{ij}),$$

with $i = 1,\ldots,n$ and $j = 1,\ldots,n_i$, where y_{ij} represent the response of the i-th individual at time t_{ij} and x_{ij} the vector of predictors at time t_{ij}. The method applies both with regular and irregular sampling in time, i.e., the number of observations can be different for each subject. HRF estimates the response variable y_{ij} using the concurrent observations and all preceding observations of the i-th individual at (but not including) time t_{ij}. Node splitting follows the standard approach of RF, e.g. minimizing the Gini impurity or the Cross-Entropy for classification or Root Mean Square Error for regression, except for the fact that the number of observations of an historical predictor will vary according to i and j. In particular, a summary function first transforms all previous values of a predictor and is denoted as $s(\eta;\bar{z}_{ijk})$ where η represent parameters of the function and $\bar{z}_{ijk}$ denotes the set of historical values of the k-th element of z_{ij}. Then, node splitting is done by minimizing the following expression over the vector (k,μ,c,η):

$$\text{argmin} \sum_{(ij)\in\text{Node}} (y_{ij} - \mu_L I(s(\eta;\bar{z}_{ijk}) < c) - \mu_R I(s(\eta;\bar{z}_{ijk}) \geq c))^2,$$

where $I(\cdot)$ is the indicator function, μ_L and μ_R are the weighted number of samples reaching node in the left and right split, respectively and c is the splitting threshold. Setting $n_{ij}(\eta)$ as the number of observation of the i-th individual in the time window $[t_{ij} - \eta_1, t_{ij} - \eta_2]$, the set of possible value of η_1 and η_2 is determined by the difference in time between within-individual successive observations in the data, both provided by the user or selected among the quantiles of the corresponding distribution. When a split is attempted on a historical predictor, a sample of this set is taken upon which the best split is selected and the size of this set can be defined by the user as well. The summary function can be defined in different ways, according to its set of parameters η. For example:

- "frequency"

$$s(\eta;\bar{z}_{ijk}) = \sum_{h:t_{ij}-\eta_1 \leq t_{i_h} < t_{ij}} I(z_{ihk} < \eta_2)$$

- "normalized frequency"

$$s(\eta;\bar{z}_{ijk}) = \sum_{h:t_{ij}-\eta_1 \leq t_{i_h} < t_{ij}} \frac{I(z_{ihk} < \eta_2)}{n_{ij}(\eta)}$$

- "average"

$$s(\eta;\bar{z}_{ijk}) = \sum_{h:t_{ij}-\eta_1 \leq t_{i_h} < t_{ij}} \frac{z_{ihk}}{n_{ij}(\eta)}$$

If $n_{ij}(\eta) = 0$, the summary function is set to zero.

In order to evaluate the performance of both models, to optimize the hyperparameter of HRF 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. Moreover, 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 $\gamma = 4$.

Validation of model performances and train/validation set splitting of the data have been evaluated with a variable-length rolling-window temporal approach. In particular, given that the maximum number of available quarters is 10 and the variable amount 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 a minimum number of observations in each train set has been fixed to 2, the final number of folds used in the cross-validation is 4.

HRF hyperparameters tuning has been performed by means of a Bayesian Optimization approach through *R* package `rBayesianOptimization` (Yan, 2016).

A comparison of explanatory power of all combination of models, framework and set of variables has been added to the predictive power one using two relevant state-of-the-art techniques: Permutation Feature Importance (PFI) and SHAP values. The aim of both methods is 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 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):

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

where $\hat{f}(x^m_{+j})$ 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^m_{-j} is identical to x^m_{+j}, 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. 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.

3.1. Statistical assessment of differences

After the evaluation of the quality of a multiple learned classifiers and consequent classification of new samples with unknown class labels, the statistical comparison of classifiers is needed to

assess the statistical differences between the results obtained by different algorithms in different instances of problems, datasets, etc. The typical sequence of analysis involves, firstly, using a test that compares simultaneously all the considered algorithms in order to test the presence of any algorithm that behaves differently. Then, if the null hypothesis is rejected (i.e., if the results show globally significant differences), the next step is analyzing which pairwise combinations are different by implementing post hoc tests.

Firstly, classical non-parametric Friedman test (Friedman, 1937) has been implemented. Sometimes observations do not meet measurement requirements and in order to avoid assumptions about the underlying populations, nonparametric statistical tests would be appropriate, like Friedman's test. The latter represents an alternative nonparametric procedure to the parametric two-way analysis of variance and it is used to detect differences in treatments across multiple test attempts. The computational procedure involves ranking each row together, ordering the rows values in decreasing order and calculate the average rank for each column. To compare two columns, the formula is the following:

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6N}}}$$

where R_i is the average rank obtained from the Friedman test for the i-th column, k represents the number of columns and N the number of blocks sets both used for comparison purposes. The underlying idea is to compare the accuracy of different classifiers using different data; as a consequence, columns will be the classifiers and rows the datasets.

Then, the corresponding post hoc tests for Friedman have been implemented, correcting p-values for multiple testing (Bergmann and Hommel's correction procedure). The latter applies a correction based on a list of possible hypothesis testing and amplifies the test power by considering only exhaustive sets of hypothesis (i.e., hypothesis that can be simultaneously true).

4. Data

Data used in this paper can be divided into two main categories: securitization (SEC) vs accounting (BS) data. Data on securitization transactions have been collected from a large European

bank which plays a leading role in the niche of revolving trade receivables' securitization programmes. Data on accounting figures have been collected from Orbis database, developed by Bureau Van Dijk (a Moody's analytics company), by matching the VAT code for each given borrowing firm[1]. Among securitization data, a peculiar role is played by a measure considered as the expression of a given borrower's credit quality: the credit rating. The credit rating, attributed by the insurance company to each SME, analyses the given SME's financial health and creditworthiness to predict its default risk based on both proprietary soft information (i.e., client information, special investigation teams) and private and publicly-available hard information (i.e., partnerships, registered payment defaults, credit reference agencies, accounting data, payment performance data, network of risk information)[2]. Therefore, insurance rating provides an objective and quantifiable means by which a company's degree of credit risk can be assessed. More specifically, the credit rating, i.e., our target variable, is a factor variable with eight categories, ranging from 2 till 9. Through this rating, the insurance company is capable of distinguishing between high-risk and low-risk clients by assigning, respectively, high and low insurance ratings. The insurance rating assigned to each SME is categorized in a numeric scale ranging from 2 to 9 according to the given firm credit risk. The higher the number, the worst the credit rating. Credit ratings evolution over time is showed in Fig. 1, highlighting an overall persistent behavior for all classes of risk. Proprietary credit rating data have been linked to Orbis' firm-level balance sheet statements and profit and loss accounts for the analyzed time period. More specifically, the dataset consists of 34 variables for 810 Italian firms collected from Q1 2015 to Q2 of 2017. In particular, 6 numerical securitization variables (hereinafter referred as SEC) were provided by the Insurance Company with quarterly frequency and 28 (25 numerical and 3 categorical) balance-sheet variables (hereinafter referred as BS) were collected by Orbis platform with annual frequency. Annual values are

[1]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.

[2]Scores are not permanent and can be affected by different factors. There are several ways to increase low scores and possibly lower premiums. To begin, a consumer will benefit by improving his or her credit rating and paying bills on time, as well as reducing debt. Also, limiting the number of insurance claims filed over a certain period can help boost an insurance score.

repeated over all quarters of each year. Furthermore, Nace Rev. 2 has been used 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 SME present in the dataset to control for unobserved heterogeneity in the given SME's industry and location. Tab. A.5 reports the definition of the variables used in the empirical analysis and some descriptive statistics.

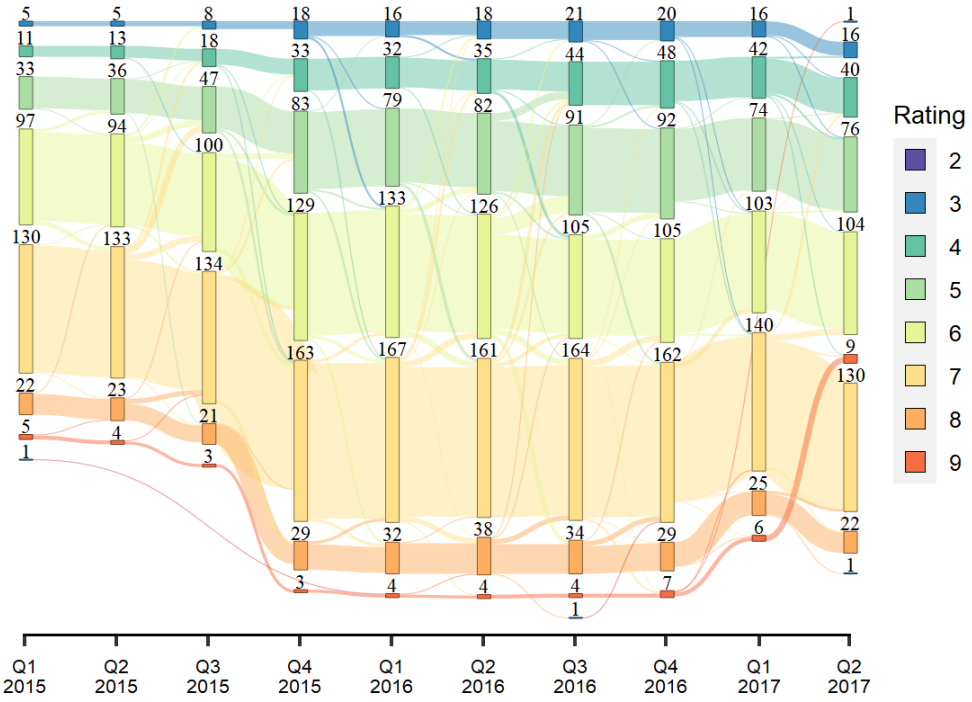

Fig. 1. Rating evolution over time.

4.1. Dataset construction and cleaning

The initial raw dataset contained a panel data of 810 firms in the time period that spans from the first quarter of 2015 to the second quarter of 2017. Data were then cleaned, checked for outliers, redundant data and missing values.

Firstly, SMEs with too many missing values were excluded from the analysis, specifically those SMEs with a large number of null observations with respect to Orbis variables and with null securitization data were not considered in the analysis, given the different industrial sector, size and number of employees of the firms (we removed 254 firms with more than 10% of missing variables and 92 firms with null securitization data). Moreover, not all firms were available for

the entire time period because some entered the insurance company portfolio after Q1 2015 and some left before Q2 2017. Therefore, given the high percentage of missing, some quarters of SEC variables were excluded as well. However, missing values (aka NA) for the cleaned dataset are still present, although with a low incidence (less than 10%). Details on the strategy used for handling NA are reported in Appendix D.

Secondly, data cleaning involved checking for holes in credit ratings. In this regard, the exploratory data analysis resulted in 1,484 missing values for the target variable (including also zero values, which are meaningless). In the overall dataset, the percentage of missing is 18.3%, with highest percentages of missing values showed at the beginning and at the end of the considered time period, i.e., the first three quarters of 2015 and the first two quarters of 2017. Moreover, some gaps within the time series were founded and replaced with recovered additional data.

Thirdly, the distribution of the time difference between consecutive years has been investigated in order to understand if computing the average to recover additional missing data was an appropriate procedure. In this regard, the stability of the distributions proved that this approach was a suitable one to follow.

Fourthly, the variables have been normalized with respect to a different set of features, according to the securitization or accounting nature of the specific data in order to obtain a normalized set of predictors between 0 and 1. However, some extreme values have been deliberately left in the dataset to reflect the extreme characteristics of some SMEs with respect to the normalized range and to avoid having a too lean dataset in terms of number of observations.

Lastly, outliers have been removed based on inter-quantile range (α–quantile and $(1-\alpha)$–quantile) but, in order to keep values of variables with small variance, if the distance between maximum and minimum value was less than an arbitrary value of tolerance, then no outlier has been removed.

The whole dataset cleaning process resulted in a final dataset comprising 534 Italian SMEs in the time period that spans from the first quarter of 2015 to the second quarter of 2017, from our original panel data composed of 810 Italian SMEs.

4.2. Additional variables and transformation

Dataset has been augmented with additional variables created from the original ones and, given the different size of companies, both BS and SEC variables have been normalized. Moreover categorical variables have been converted into dummies, dropping the $n-th$ level in order to avoid multicollinearity. In particular:

- adding log10 of annual *Turnover* and quarterly *Outstanding*;
- adding annual ratio *Delinquency* and *Delinquency90* and *Outstanding*;
- adding quarterly ratio *Outstanding_Invoices* and *Outstanding_Portfolio* of *Outstanding* and *InvoicesCount* and *PortfolioCount*, respectively. Both ratios, being still an average amount of money, are further normalized by annual *Purchase*;
- adding annual binary dummy *Liquidity Tension*, 1 if *Collectionperioddays* is greater than *Creditperioddays*, 0 otherwise;
- adding annual binary dummy *Dummy_Delinquency*, 1 if *Delinquency* is greater of equal then the average *Delinquency* of all firms, 0 otherwise;
- adding annual binary dummy *Delinquency Severe*, 1 if *Delinquency* is greater of equal then the average *Delinquency* of all firms with a two standard deviations confidence, 0 otherwise;
- adding regional dummy *Region* aggregating each city into 4 geographical macro-areas: North-East, North-West, Center and South+Islands;
- normalizing all BS variables by *TotalAsset*;
- normalizing all SEC variables by *Outstanding*.

Finally, correlation among the BS and SEC variables has been checked and 9 variables have been removed according to a Variance Inflation Factor (VIF) value above 5. A complete list of variables description and corresponding statistics is reported in Tab. 1. Dummy and categorical

Table 1
List of final variables.

Variable	Description	Mean	Stdev	Median	Minimum	Maximum	Removed due to VIF
Rating	Rating score, 2 means low risk	5.1091	1.2443	5	2	9	
Purchase	Accounting of Cash and Credit purchases	1.4914	0.9555	1.3062	0.0168	6.4811	
Current liabilities	Company's debts or obligations that are due to be paid to creditors within one year	0.5480	0.1996	0.5457	0.0383	2.0324	
Current ratio	Comparison a firm's current assets to its current liabilities	0.0136	0.0089	0.0121	0.0008	0.1920	x
Delinquency	Dummy variable equal to 1 if the firm misses a scheduled payment on an invoice, otherwise equal to 0	0.0162	0.1043	0	0	1	
EBIT	Company's net income before income tax expense and interest expenses are deducted	0.0485	0.0881	0.0400	-1.4438	0.6867	
Fixed assets	Long-term tangible piece of property or equipment that a firm owns and uses in its operations	0.3316	0.2169	0.3045	0.0014	0.9833	x
Collections	Amount of invoices currently sold to the bank	2.7146	90.7025	0.7687	0	5520.7044	
Liquidity	Company's ability to pay off current debt obligations without raising external capital	0.0104	0.0078	0.0091	0.0008	0.1594	
Outstanding	Amount of securitization transactions in which the borrowing firm is involved, expressing its economic exposure in logarithmic scale (base 10)	4.1590	1.9485	4.6758	0	7.1786	
Turnover	Annual sales volume net of all discounts and sales taxes in logarithmic scale (base 10)	4.5325	0.8341	4.4351	2.8520	6.9362	
LT Debt	Debt with maturities greater than 12 months	0.0911	0.1035	0.0540	0	0.5163	
Asset Turnover	Sales revenue divided by capital employed	0.0753	0.1769	0.0412	0.0008	3.5299	x
New Receivables	Monetary amount of receivables sold to the bank with respect to a given borrowing firm at the current invoices' transfer	0.2060	0.2355	0.1634	0	1	
Outstanding_Invoices	Amount of securitization transactions in which the borrowing firm is involved divided by total number of invoices and annual Purchase	0.1332	0.3392	0.0002	0	1	x
Outstanding_Portfolio	Amount of securitization transactions in which the borrowing firm is involved divided by total number of portfolios and annual Purchase	0.1454	0.3464	0.0017	0	1	x
Profit Margin	Percentage of sales turned into profits	0.0240	0.0643	0.0174	-0.7288	0.5611	
Profit per employee	Net Income for the past twelve months (LTM) divided by the current number of Full-Time Equivalent employees	0.0047	0.0502	0.0004	-0.0178	1	
ROA	Net income divided by total assets	0.0318	0.0887	0.0210	-0.3528	1.9188	
ROCE	Company's earnings before interest and tax (EBIT) divided by its capital employed	0.0782	0.2240	0.0710	-7.3081	0.8548	x
ROE	Fiscal year net income divided by total equity	0.0822	0.6385	0.0831	-13.7168	9.7300	
Solvency_A	Firm's capacity to meet its long-term financial commitments	0.2834	0.1843	0.2468	-0.7866	0.9333	
Tangibles	Assets that have a physical value	0.2477	0.1931	0.2121	0	0.9797	
Working Capital	Difference between a company's current assets and its current liabilities	0.1372	0.2414	0.1195	-1.7193	1.0661	
Delinquency Severe	Dummy variable equal to 1 if delinquency_outstandingpost is larger of 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						
Liquidity Tension	Dummy variable equal to 1 if Collectionperioddays (number of days it takes to turn accounts receivable into cash) is larger than Creditperioddays (number of days that a customer is allowed to wait before paying an invoice), otherwise equal to 0						x
NACE	Statistical Classification of Economic Activities in the European Community						
Industry	Industrial classification variable reflecting main division within main section of NACE						x
Region	Geographical macro-areas						

variables distribution by each rating is reported in Tab. 2. Geographical distribution of firms is showed in Fig. 2.

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 3,009 rows.

Table 2
Dummy and categorical variables distribution by each rating.

Variable		Rating 2	3	4	5	6	7	8	9
NACE									
Manifacturing		20%	23%	28%	34%	35%	31%	18%	0%
Wholesale and retail trade; repair of motor vehicles and motorcycles		57%	61%	54%	53%	54%	59%	82%	100%
Accommodation and food service activities		22%	8%	15%	10%	7%	8%	0%	0%
Agriculture, forestry and fishing		2%	7%	1%	1%	2%	1%	0%	0%
Other		1%	6%	3%	2%	2%	3%	0%	0%
REGION									
North-East		24%	35%	42%	38%	26%	20%	8%	0%
North-West		61%	35%	28%	23%	26%	25%	31%	0%
Center		6%	20%	16%	16%	19%	29%	12%	25%
South+Islands		9%	10%	15%	22%	30%	26%	49%	75%
DUMMY									
Delinquency Severe	0	74%	90%	92%	98%	100%	100%	100%	100%
	1	26%	10%	8%	2%	0%	0%	0%	0%
Delinquency 90	0	47%	29%	60%	74%	82%	82%	43%	25%
	1	53%	41%	40%	26%	18%	18%	57%	50%
Liquidity Tension	0	76%	69%	68%	54%	50%	56%	59%	100%
	1	24%	31%	32%	46%	50%	44%	41%	0%

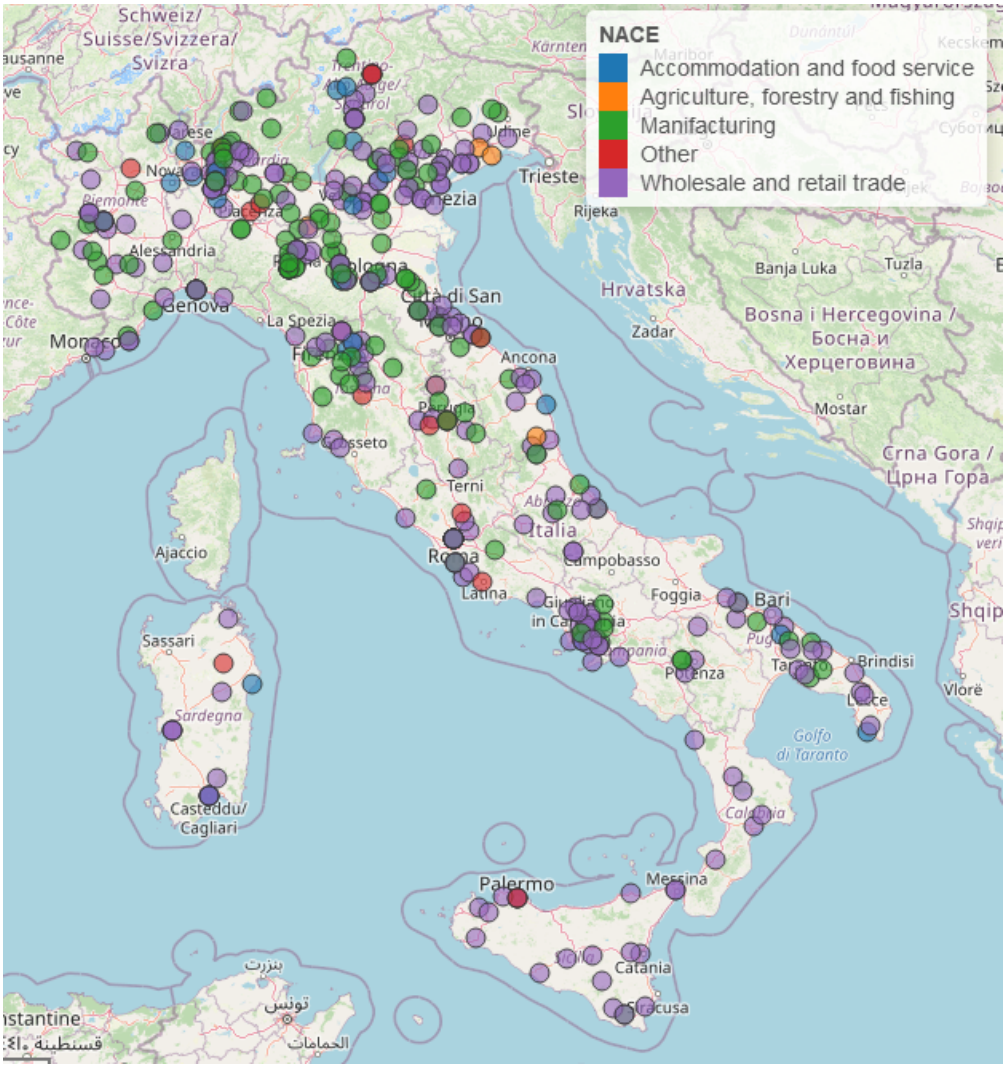

Fig. 2. Geographical distribution of firms.

5. Empirical Analysis

In this section, comparison of classification performance of models along three dimension is presented. As introduced in section (3), models have been distinguished according to *static* versus *dynamic* 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. At the end, an optimal set of predictors, combining both *BS* and *SEC* set, has been identified according to feature importance evaluated on the last dimension of comparison analysis. Based on the latter, classification performance has been presented.

5.1. Model evaluation

A set of evaluation metrics has been used (Appendix B) in order to obtain an optimal combination of hyperparameters (HRF model) and variables (PB model). Respectively, F_1 cross-entropy metric has been maximized during cross-validation phase to avoid ovefitting; Akaike information criterion (AIC) has been minimized during best subset selection to obtain a stable function of predictors.

Tab. B.6 shows the optimal hyperparameters set and the selected set of predictors with reference to SEC set 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 Tab. 3. F_1-score with regards to HRF and dynamic PB can be highlighted as the lowest ones compared to the case without temporal dependence. Previous rating states, together with predictors' history, seem to be necessary for a correct classification of insurance credit risk. Dynamic PB outperforms the other estimated models, with 80% - 70% F_1-score respectively on train and test set. Tab. B.7 reports the proposed models for the BS set of predictors, for both dynamic and static version.

A first difference between SEC and BS set that can be pointed out is the number of predictors, lower in the former with respect to the latter; this could influence the analysis resulting into different conclusions for this set of variables. Analyzing the classification performance on BS set in Tab. 3, HRF outperforms the other models in terms of overall classification performance with

an F_1-score of 90% for train set and 70% for test set. If compared with SEC case, the persistence of rating history leads to more accurate prediction in the PB case but does not have a significant influence in the accuracy of HRF predictions.

Based on the results in terms of variable importance and best subset selection aimed at reducing the high dimensional feature space and obtaining an optimal group of features, a final model has been implemented by combining both BS and SEC set of variables. Before fitting the models, preliminary analysis has been implemented in order to check for correlation and collinearity. The results report significant correlation between Outstanding and Turnover, that leads to the removal of Outstanding since a measure of firm's financial exposure has already been selected with Delinquency and having a metric for firms' size in terms revenues seems to be useful in the determination of credit rating. Even if the regional and industrial classification variables have a significant effect on the target only for one specific category and the impact cannot be confirmed in terms of importance, both regional and industrial classification variables have been kept in order to have an insight about the economic framework. To summarize, the following variables have been selected for the final set: *Collections, New Receivables, Delinquency, Turnover, Solvency_A, Working Capital, LT Debt, Current liabilities, Liquidity, NA- CE, Region*. In terms of classification performance (3), HRF outperforms the other models, with good performance on a macro level (90%-70% F_1-score on train and test set). Comparing the final model with the BS set of variables, no relevant differences can be reported since the performances are almost the same. This could be due to the high numerosity of variables for the BS set that allows to reach the performances of the optimal set. To conclude, the autoregressive behavior of the models is necessary to reduce the misclassification cost.

Table 3
Macro-averaged F_1-score on training and test sample for all set of predictors.

Model	Version	Sample	F_1-score BS	SEC	BS + SEC
HRF	Static	Train	0.919	0.4105	0.9611
		Test	0.677	0.3417	0.6986
	Dynamic	Train	0.9154	0.7480	0.9014
		Test	0.7361	0.5519	0.7326
PB	Static	Train	0.4634	0.4579	0.4609
		Test	0.4529	0.4284	0.4543
	Dynamic	Train	0.8148	0.7901	0.799
		Test	0.7407	0.7346	0.74487

5.2. Model explanation

In this section, explainability capabilities of both HRF and PB have been compared using PFI and SHAP values together with marginal effects. On one side, the change in probability correlated to each predictor has been explored in order to understand the sign of the effect on each class of the target variable; on the other side, more complex relationships have been investigated through SHAP values. As a result, the most relevant features, in terms of relative importance, have been selected in order to implement the final credit-scoring model. According to classification performance, feature importance figures with reference to the best statistical models for the three set of variables have been reported in Appendix C. As classifiers based on probit model do not seem to work better than random choice (i.e. accuracy metrics less than 50%), results in terms of feature importance are meaningless.

With regards to PFI, relative importance has been computed as 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 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 total percentage of relative importance carried by the features in top position. Related figures are presented on a macro-level (aggregated for all Rating classes) and distinguished according to time dependence. The latter distinction has been carried out when both models (static and dynamic version) report accuracy metrics higher than 50% on test set. Otherwise, only one case has been analysed.

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 time 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 feature and grouped according to rating classes with the aim of investigating how each feature impacts, on average, on the predicted probability of each class compared to the average probability of this class for the full dataset. Given the best performance of Random forest algorithm as an ensemble of historical classification trees and the slow computational procedure in calculating the Shapley values, figures with reference to only dynamic HRF model have been reported in Appendix C for comparison with PFI.

Starting from BS set of variables, Fig. C.3 shows the importance ranking in terms of PFI for both static and dynamic version of the HRF model. Relative importance values slightly vary if the autoregressive behavior is considered in the model or not. With reference to the static version of the model, Turnover can be observed in the top position with 47% of relative importance value, followed by Solvency (17%), Working Capital (7%), LT Debt (6%) and Liquidity (5%) with lower order of magnitude. On an aggregated level, the previous features represent almost the 90% of relative feature importance over the total of 20 considered predictors. Considering autoregressive behavior of the target and predictors, the LagRating shows a slightly negative relative importance of 1%, carrying no positive effect on model prediction error if compared to balance sheet informations in the top of the graph. Importance of considerable order is reported by the same previous set of BS variables, with the addition of Profit Margin (10%). Quantitative variables are more important than qualitative ones, i.e. NACE and Region, since each dummy has a frequency that affects its importance value. Opposite behaviour is reported from Probit model (Fig. C.4), where autoregressive behaviour seems to carry 90% of relative importance on model prediction error. As

a result, the other variables reports negligible relative importance scores.

Furthermore, for PB model, marginal effects can be analyzed in order to investigate the change in probability when the predictor variable increases by one unit. According to Tab. B.9 , which reports the PB models distinguished according to autoregressive behavior, it can be noticed that Rating class 6 represents a threshold for change of sign of partial derivatives, allowing the interpretation of results by distinguishing between low-risk (3,4,5) and high-risk (6,7). Regarding the key indicators to the financial solvency of the company, i.e., Current liabilities, LT Debt and Working capital, an increment of these metrics implies 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. Together with debt, Working capital has the same effects on rating classes because, being computed as the difference between shares and the sum of trade credit and payables, the positive sign could reflect the weight of trade payables in the short term that, in the case of SMEs, is particularly high and could result in higher risk. Furthermore, high working capital is flagged as having liquidity issues, since a company is not effectively reallocating capital into higher growth. Contrarily, Liquidity, ROA, Tangibles, Collections e Turnover show negative marginal effects in correspondence of the riskiest classes, since high values for liquidity, profitability and size measures represent a signal of solid financial and operational performance. As a consequence, a rise in these metrics is associated with higher probability of belonging to low-risk classes. Specifically, high liquidity implies 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 categorical variables, regional classification seems to have a significant impact on the predictive power, since the marginal effects show a high level of risk in Southern Italy, possibly caused by the different economic context compared to Northern Italy. Belonging to some specific classes (3-4-5) does not imply temporal persistence of those rating classes over time; on the contrary, belonging to all classes (except Rating 7) seems to increase the probability of having

a rating score of Rating 6, given probably the high numerosity of the latter. In addition, all the lagged classes show positive impact on the riskiest Rating 7. However, given the complexity of the classification problem at hand, defining the target variable as binary (5.3) allows to understand that there exists a persistence in belonging to low-risk and high-risk classes over time.

Furthermore, Shapley values (Fig. C.5) confirm previous results, highlighting high average contribution of Turnover, together with Solvency, Profit margin, Working capital and Liquidity. As expected, heterogeneous contribution is carried by aforementioned features with respect to Rating classes, with highest impact on the largest ones (i.e. Rating class 4,5 and 6).

Following the same computational procedure for the SEC set of variables, it can be noticed that a relevant role is played by LagRating (97%) within HRF modelling framework, followed by slightly positive scores of Outstanding and New Receivables in the determination of rating score (Fig. C.6). Securitization variables reports negligible contribution if compared to time dependence. Same conclusions could be reported for Probit model (Fig. C.7).

Delinquency and Outstanding represent metrics of economic exposure of the firms under investigation; the former with respect to missed payments and, the latter, to securitization transactions in which the borrowing firm is involved. These metrics are directly linked to the level of risk reported by each firm. On the other side, New Receivables measures trade balance credits in terms of volume and lengthening of deadlines. Higher position of trade credit could reflect liquidity drainage, i.e., less investment availability.

As mentioned before, the partial derivatives (Tab. B.10) highlight class 6 as threshold for change of sign, and, as a consequence, an increase in New Receivables and Delinquency results into a positive effect on the probability of belonging to the riskiest rating classes. The opposite behavior is showed by Outstanding and Collections. With regards to the lagged dummies of the target variable, same conclusions as for the BS set can be extracted.

Shap results allow to grasp individual contribution of securitization variables (Fig. C.8). Delinquency and Delinquency 90 report the highest average impact on Rating classes, being relatively important in HRF classifier.

Among the combined set of variables, Turnover shows a predominant role with a relative importance of about 30% on a macro level, followed by Solvency, Working Capital, LT Debt and Liquidity, reaching a total relative importance of 70% for the dynamic case and 80% for the static one (Fig. C.9). Within the optimal set of variables, the time contribution is in the lowest position, with slightly negative PFI; the selected predictors enable to better differentiate between classes without allowing for the persistence of credit history. The BS set of variables overcomes, in terms of PFI, the SEC predictors. On the contrary, PB model highlights 92% contribution of time dependence (Fig. C.10). In line with the previous results, the partial derivatives of Tab. B.11 highlight conclusions already mentioned for the distinct set of variables (BS and SEC); it is worth noticing the high levels of significance for all the marginal effects.

The Shapley values report, overall, the same importance ranking for the selected set of predictors; Turnover, Working Capital, Liquidity, LT Debt and Current liabilities show the highest magnitude in terms of average impact by feature and class (Fig. C.11). The magnitude of the features' effect is smaller for SEC set compared to BS one; the latter has relevant impact on the classes with largest number of observations (4,5,6). Specifically, the combination of Current liabilities, Liquidity, Solvency and Turnover plays a significant role for the identification of the extreme classes, bringing up or down the probability of belonging to that specific class.

5.3. Assessment of differences and robustness checks

According to the methodology presented in 3.1, statistical comparison of classifiers has been implemented to assess significant differences between the results obtained in the previous section. Firstly, macro-weighted balanced accuracy obtained by the previous algorithms in the three different datasets (i.e., BS, SEC and BS+SEC) has been imported; then, differences have been tested on algorithm (further divided based on autoregressive behavior) and dataset level. Since results obtained from Friedman test show globally significant differences on algorithm level, the next step involves analyzing which pairwise combinations are different. The p-value matrix generated when doing all the pairwise comparisons show significant differences at 0.05 level for the PB model based on different time dimension, highlighting the temporal component as statistically significant

discriminant between algorithms (Tab. 4).

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 behaviors that the classifiers exhibit, the ordinal scale has been converted to dichotomous variable. Firstly, a formulation Rating 7 vs ALL has been implemented, resulting into poor performances given the imbalanced nature of the dataset with respect to the tails. Then, the target variable has been defined as 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 a risk and differentiate between lower and higher insurance risks. 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 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 ordinal case, since the threshold that highlights the change of sign is class 6. Summarizing, the binary formulation simplifies the classification problem at hand and results in slightly higher performances together with same explainable conclusions as for individual risk.

Table 4
Corrected p-value matrix using Bergmann and Hommel's correction procedure generated when doing all the pairwise comparisons.

		PB		HRF	
		Static	Dynamic	Static	Dynamic
PB	Static		**0.02**	0.52	0.21
	Dynamic	**0.02**		0.12	0.52
HRF	Static	0.52	0.12		0.52
	Dynamic	0.21	0.52	0.52	

6. Conclusions

By employing a unique and proprietary dataset comprising granular firm-level securitization and accounting data on a panel of 810 Italian SMEs over the time period 2015-2017, this paper tests two alternative approaches grounded in statistical learning and machine learning frameworks and compares their respective capability in predicting SME 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 static and a dynamic framework. Moreover, we further assess the relevance of each variable to predict SME credit risk, through the use of Shapley values.

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

Our research question represents a matter of concerns to policy makers, 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 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.

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.

References

Ahn, B. S., Cho, S. S., and Kim, C., 2000. The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems With Applications*, 18: 65–74.

Altman, E. I., 1980. Commercial bank lending: process, credit scoring, and costs of errors in lending. *Journal of Financial and Quantitative Analysis*, pages 813–832.

Bank for international settlements, 2006. International convergence of capital measurement and capital standards: a revised framework. *Bank for international settlements*.

Bazarbash, M., 2019. Fintech in financial inclusion: machine learning applications in assessing credit risk.

Berger, A. N. and Udell, G. F., 1995. Relationship lending and lines of credit in small firm finance. *Journal of business*, pages 351–381.

Berger, A. N., Miller, N. H., Petersen, M. A., Rajan, R. G., and Stein, J. C., 2005. Does function follow organizational form? evidence from the lending practices of large and small banks. *Journal of Financial Economics*, 76:237–269.

Blöchlinger, A. and Leippold, M., 2018. Are ratings the worst form of credit assessment except for all the others? *Journal of Financial and Quantitative Analysis*, 53(1):299–334.

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

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J., 2020. Explainable ai in fintech risk management. *Frontiers in Artificial Intelligence*, 3:26.

Capitaine, L., Genuer, R., and Thiébaut, R., 2019. Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*.

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

Carty, L. and Fons, J., 1994. Measuring changes in corporate credit quality. *Special report, Moody's*.

Claessens, S., Krahnen, J., and Lang, W. W., 2005. The basel ii reform and retail credit markets. *Journal of Financial Services Research*, 28(1-3):5–13.

Contoyannis, P., Jones, A., and Rice, N., 2004. The dynamics of health in the british household panel survey. *Journal of Applied Econometrics*, 19:473–503.

Cucinelli, D., Di Battista, M. L., Marchese, M., and Nieri, L., 2018. Credit risk in european banks: The bright side of the internal ratings based approach. *Journal of Banking & Finance*, 93:213–229.

De Andrés, J., Landajo, M., and Lorca, P., 2005. Forecasting business profitability by using classification techniques: A comparative analysis based on a spanish case. *European Journal of Operational Research*, 167(2):518–542.

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.

Fantazzini, D. and Figini, S., 2009. Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11:29–45.

Ferri, G. and Pesic, V., 2017. Bank regulatory arbitrage via risk weighted assets dispersion. *Journal of Financial Stability*, 33(C):331–345. doi: 10.1016/j.jfs.2016.10.006.

Filomeni, S., Udell, G. F., and Zazzaro, A., 2020a. Hardening Soft Information: Does Organizational Distance Matter? *European Journal of Finance*. doi: https://doi.org/10.1080/1351847X.2020.1857812.

Filomeni, S., Udell, G. F., and Zazzaro, A., 2020b. Communication frictions in banking organizations: Evidence from credit score lending. *Economics Letters*, 195C(109412).

Financial Stability Board, 2017. Artificial intelligence and machine learning in financial services: Market developments and financial stability implications.

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`.

Gonzalez, F., Haas, F., Johannes, R., Persson, M., Toledo, L., Violi, R., Wieland, M., and Zins, C., 2004. Market dynamics associated with credit ratings. a literature review. *Occasional Paper 16, European Central Bank*.

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

Hajjem, A., Bellavance, F., and Larocque, D., 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84:1313–1328.

Hsieh, N.-C., 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, 27(4):623 – 633.

Liberti, J. M. and Petersen, M. A., 2018. Information: Hard and Soft. *The Review of Corporate Finance Studies*, 8(1):1–41.

Lin, S.-W., Shiue, Y.-R., Chen, S.-C., and Cheng, H.-M., 2009. Applying enhanced data mining

approaches in predicting bank performance: A case of taiwanese commercial banks. *Expert Systems with Applications*, 36(9):11543 – 11551.

Louzada, F., Ara, A., and Fernandes, G. B., 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117 – 134.

Moscatelli, M., Narizzano, S., Parlapiano, F., and 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.

Odders-White, E. and Ready, M., 2006. Credit ratings and stock liquidity. *Review of Financial Studies*, 19:119–157.

OECD, 2020. Financing smes and entrepreneurs: An oecd scoreboard. special edition: The impact of covid-19.

Olmeda, I. and Fernandez, E., 1997. Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. *Computational Economics*, 10(4):317–335.

Segal, M. R., 1992. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87:407–418.

Sela, R. J. and Simonoff, J. S., 2012. Re-em trees: A new data mining approach for longitudinal data. *Machine Learning*, 86:169–207.

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

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

Stein, J. C., 2002. Information production and capital allocation: Decentralized versus hierarchical firms. *Journal of Finance*, LVII(5):1891–1921.

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

Wooldridge, J., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20:39–54.

World Bank Group. Credit Scoring Approaches Guidelines. Technical report, 2019.

Yan, Y., 2016. rbayesianoptimization: Bayesian optimization of hyperparameters. URL `https://CRAN.R-project.org/package=rBayesianOptimization`.

Appendix A. List of raw variables

Table A.5
List of initial variables.

Variable	Type	Missing %	Minimum	Maximum	Mean	St Dev	Unique values	Source	Frequency	Action
Industry	Cat	0%					22			
NACE	Cat	0%					11			
City	Cat	0%					375			
Purchase_2015	Num	4%	242,000	6,038,375,000	157,916,200	468,969,700		Orbis	Annual	
Purchase_2016		3%	236,000	6,277,094,000	154,080,200	452,602,000				
Purchase_2017		6%	18,000	6,497,610,000	174,069,100	508,482,300				
Collectionperioddays2015		2%	0	264	52	42				
Collectionperioddays2016		0%	0	290	51	43				
Collectionperioddays2017		1%	0	265	50	40				
Creditperioddays2015		2%	0	171	52	29				
Creditperioddays2016		0%	0	174	53	32				
Creditperioddays2017		1%	0	557	57	45				
Current liabilities2015		2%	116,078	1,762,623,000	65,453,870	180,990,000				
Current liabilities2016		0%	142,224	1,716,557,000	64,732,280	176,089,600				
Current liabilities2017		1%	142,548	1,746,051,000	67,123,170	175,378,800				
Current ratio2015		2%	0	19	1	1				
Current ratio2016		0%	0	9	1	1				
Current ratio2017		2%	0	7	1	1				
EBIT2015		2%	-97,272,000	423,930,000	6,082,113	29,011,610				
EBIT2016		0%	-73,706,000	402,693,000	6,155,327	28,772,780				
EBIT2017		1%	-114,775,000	410,921,000	6,409,824	30,289,210				
Fixed assets2015		2%	2,039	3,255,230,000	72,978,520	271,180,300				
Fixed assets2016		0%	500	3,422,961,000	74,910,600	280,711,700				
Fixed assets2017		1%	500	4,578,240,000	79,002,090	314,053,100				
Liquidity2015		2%	0	16	1	1				
Liquidity2016		0%	0	7	1	1				
Liquidity2017		1%	0	6	1	1				
LT Debt2015		2%	0	743,361,000	13,335,680	51,390,820				
LT Debt2016		0%	0	516,854,000	13,735,920	49,286,210				
LT Debt2017		1%	0	1,378,198,000	16,004,280	75,481,730				
Asset Turnover2015		2%	0	353	8	22				
Asset Turnover2016		1%	0	308	7	16				
Asset Turnover2017		2%	0	231	7	15				
Profit Margin2015		2%	-40	25	2	4				
Profit Margin2016		0%	-67	29	2	6				
Profit Margin2017		2%	-73	56	2	7				
Profit per employee2015		2%	-69,693	566,486	18,088	43,258				
Profit per employee2016		1%	-89,917	480,993	17,743	41,255				
Profit per employee2017		2%	-76,612	273,365	16,557	33,493				
ROA2015		2%	-25	23	3	4				
ROA2016		0%	-35	31	3	5				
ROA2017		2%	-35	47	3	6				
ROCE2015		2%	-731	84	8	35				
ROCE2016		1%	-120	85	8	15				
ROCE2017		2%	-364	81	6	28				
ROE2015		2%	-529	973	11	53				
ROE2016		1%	-309	95	9	26				
ROE2017		3%	-837	94	3	62				
Solvency_L2015		2%	-0	91	28	18				
Solvency_L2016		0%	-10	92	29	18				
Solvency_L2017		1%	-79	93	29	20				

Variable	Type	Missing %	Minimum	Maximum	Mean	St Dev	Unique values	Source	Frequency	Action
Tangibles2015	Num	2%	0	3,041,447,000	46,639,800	186,760,700	Orbis	Annual		
Tangibles2016		0%	0	3,257,302,000	49,759,500	199,723,300				
Tangibles2017		1%	0	4,388,377,000	53,987,410	240,358,800				
TotalAsset2015		2%	238,723	4,807,100,000	138,682,700	425,440,600				
TotalAsset2016		0%	305,390	5,641,500,000	141,434,300	446,524,100				
TotalAsset2017		1%	297,820	6,122,933,000	148,336,500	468,131,000				
Turnover2015		2%	661,365	8,315,389,000	226,622,000	683,864,800				
Turnover2016		0%	250,000	8,688,413,000	231,283,200	697,866,600				
Turnover2017		1%	250,000	8,896,700,000	243,088,200	720,414,600				
Working Capital2015		2%	-470,089,000	401,200,000	-1,101,782	53,508,490				
Working Capital2016		0%	-526,333,000	437,500,000	-1,357,159	57,046,520				
Working Capital2017		1%	-532,052,000	417,400,000	-1,516,843	57,449,010				
EBITDA_2015		89%	-6,573,000	381,351,000	12,709,390	51,550,450				Removed
EBITDA_2016		89%	-6,708,000	417,812,000	13,301,330	55,816,760				Removed
EBITDA_2017		89%	-9,458,000	425,655,000	13,246,350	55,360,140				Removed
Gearing2015		22%	0	998	158	173				Removed
Gearing2016		18%	0	993	151	163				Removed
Gearing2017		20%	0	987	158	172				Removed
Interestcover2015		15%	-65	980	34	101				Removed
Interestcover2016		9%	-87	743	33	82				Removed
Interestcover2017		12%	-81	841	38	102				Removed
Solvency_L2015		14%	0	99	35	24				Removed
Solvency_L2016		14%	0	99	35	24				Removed
Solvency_L2017		16%	1	100	36	25				Removed
InvoicesCount_03_2015	Num	100%					Insurance	Quarterly		Removed
InvoicesCount_03_2016		0%	0	15,442	232	1,263				
InvoicesCount_03_2017		0%	0	33,175	246	1,631				
InvoicesCount_06_2015		100%								Removed
InvoicesCount_06_2016		0%	0	17,894	249	1,402				
InvoicesCount_06_2017		0%	0	21,945	220	1,280				
InvoicesCount_09_2015		100%								Removed
InvoicesCount_09_2016		0%	0	20,218	229	1,325				
InvoicesCount_12_2015		0%	0	20,375	240	1,349				
InvoicesCount_12_2016		0%	0	20,781	243	1,399				
Collections_03_2015		7%	0	3,767,527	51,407	267,665				
Collections_03_2016		0%	0	4,549,522	70,702	322,386				
Collections_03_2017		7%	0	6,827,631	101,783	458,061				
Collections_06_2015		3%	0	4,913,739	80,929	378,060				
Collections_06_2016		0%	0	3,131,328	48,427	229,664				
Collections_06_2017		8%	0	6,838,858	95,551	440,323				
Collections_09_2015		2%	0	5,525,285	74,247	366,442				
Collections_09_2016		0%	0	6,368,652	66,993	378,023				
Collections_12_2015		0%	0	5,685,703	71,191	357,369				
Collections_12_2016		0%	0	8,502,823	93,058	497,841				
Delinquency90_032015		7%	0	574,535	2,482	25,639				
Delinquency90_032016		0%	0	987,342	6,087	55,590				
Delinquency90_032017		7%	0	802,891	3,897	40,401				
Delinquency90_062015		3%	0	1,053,269	3,206	45,242				
Delinquency90_062016		0%	0	792,409	3,152	33,447				
Delinquency90_062017		8%	0	461,054	2,584	22,676				
Delinquency90_092015		2%	0	1,184,993	3,745	46,382				
Delinquency90_092016		0%	0	902,860	3,357	39,019				
Delinquency90_122015		0%	0	1,461,948	3,511	54,612				
Delinquency90_122016		0%	0	653,700	2,520	30,138				

Variable	Type	Missing %	Minimum	Maximum	Mean	St Dev	Unique values	Source	Frequency	Action
New Receivables_03_2015		7%	0	3,212,298	70,264	277,914				
New Receivables_03_2016		0%	0	3,153,487	62,513	258,509				
New Receivables_03_2017		0%	0	3,371,893	63,498	262,418				
New Receivables_06_2015		3%	0	2,869,105	67,241	278,122				
New Receivables_06_2016		0%	0	2,968,038	72,552	295,227				
New Receivables_06_2017		0%	0	3,854,462	68,858	286,216				
New Receivables_09_2015		2%	0	4,501,308	74,487	321,610				
New Receivables_09_2016		0%	0	5,217,448	74,585	324,415				
New Receivables_12_2015		0%	0	3,560,522	83,456	311,808				
New Receivables_12_2016		0%	0	5,279,336	93,226	378,014				
Outstanding_03_2015		7%	0	12,163,100	394,611	1,317,095				
Outstanding_03_2016		0%	0	14,320,530	367,375	1,296,520				
Outstanding_03_2017		5%	0	14,515,050	443,584	1,521,907				
Outstanding_06_2015		3%	0	10,712,350	327,825	1,160,684				
Outstanding_06_2016	Num	0%	0	14,682,840	390,294	1,443,342	Insurance	Quarterly		
Outstanding_06_2017		7%	0	14,497,480	479,817	1,526,787				
Outstanding_09_2015		2%	0	11,777,300	326,774	1,163,738				
Outstanding_09_2016		0%	0	14,052,470	363,672	1,343,489				
Outstanding_12_2015		0%	0	14,598,040	379,470	1,346,056				
Outstanding_12_2016		0%	0	15,085,840	383,250	1,444,328				
PortfolioCount_03_2015		7%	0	10	2	2				
PortfolioCount_03_2016		0%	0	13	2	2				
PortfolioCount_03_2017		0%	0	12	2	2				
PortfolioCount_06_2015		3%	0	10	2	2				
PortfolioCount_06_2016		0%	0	12	2	2				
PortfolioCount_06_2017		0%	0	12	2	2				
PortfolioCount_09_2015		2%	0	10	2	2				
PortfolioCount_09_2016		0%	0	11	2	2				
PortfolioCount_12_2015		0%	0	13	2	2				
PortfolioCount_12_2016		0%	0	11	2	2				

Appendix B. Performance

Table B.6
Model architecture for SEC set of predictors.

Model	Version	Hyperparameters or Selected set of predictors
HRF	Static	Mtry = 5; Ntrees = 10; Nodesize = 100
	Dynamic	Mtry = 4; Ntrees = 141; Nodesize = 89; Method = "mean0"
PB	Static	New Receivables+Outstanding+Delinquency
	Dynamic	Collections+Outstanding+Delinquency+LagRating

Table B.7
Model architecture for BS set of predictors.

Model	Version	Hyperparameters or Selected set of predictors
HRF	Static	Mtry = 14; Ntrees = 500; Nodesize= 1
	Dynamic	Mtry = 6; Ntrees = 50; Nodesize= 3; Method = “meanw0”
PB	Static	Current liabilities + Liquidity ratio + LT Debt + ROA+ Tangibles + Working Capital + Purchase + Turnover + Region + NACE
	Dynamic	Current liabilities + Liquidity + LT Debt + Working Capital + Purchase + EBIT + Turnover + Region + LagRating

Table B.8
Model architecture for BS+SEC set of predictors.

Model	Version	Hyperparameters or Selected set of predictors
HRF	Static	Mtry = 5; Ntrees = 500; Nodesize= 1
	Dynamic	Mtry = 5; Ntrees = 50; Nodesize= 3; Method = “freqw”
PB	Static	Collections + New Receivables + Delinquency + Turnover + Solvency_A + Working Capital + LT Debt + Current liabilities + Liquidity
	Dynamic	Collections + New Receivables + Delinquency + Turnover + Solvency_A + Working Capital + LT Debt + Current liabilities + Liquidity + LagRating

Table B.9
Table of PB marginal effects for BS variables.

Model	Historical	Variables	Marginal effects				
			y = 3	y = 4	y = 5	y = 6	y = 7
PB	Static	Current liabilities	-0.2871 (****)	-0.5251 (****)	-0.1829 (****)	0.7660 (****)	0.2292 (****)
		Liquidity	1.1073 (ns)	2.0251 (ns)	0.7056 (ns)	-2.9543 (ns)	-0.8837 (***)
		LT Debt	-0.3299 (****)	-0.6034 (****)	-0.2102 (****)	0.8802 (***)	0.2633 (***)
		ROA	0.3262 (****)	0.5966 (****)	0.2079 (****)	-0.8702 (****)	-0.2603 (****)
		Tangibles	0.0233 (ns)	0.0425 (ns)	0.0148 (ns)	-0.0620 (ns)	-0.018 (ns)
		Working Capital	-0.0569 (****)	-0.1042 (****)	-0.0363 (****)	0.1520 (****)	0.0455 (****)
		acquisti	0.0131 (***)	0.0240 (**)	0.0083 (**)	-0.0349 (**)	-0.010 (**)
		Turnover	0.0584 (****)	0.1068 (****)	0.0372 (****)	-0.1558 (****)	-0.0466 (****)
		R1	0.0168 (*)	0.0295 (*)	0.0091 (*)	-0.0431 (*)	-0.0123 (*)
		R2	0.0057 (ns)	0.0102 (ns)	0.0033 (ns)	-0.0149 (ns)	-0.0043 (ns)
		R3	-0.0233 (**)	-0.0464 (**)	-0.0210 (*)	0.0675 (**)	0.0232 (**)
		N1	-0.0279 (ns)	-0.0540 (ns)	-0.0227 (ns)	0.0785 (ns)	0.0260 (ns)
		N2	-0.0225 (ns)	-0.0405 (ns)	-0.0136 (ns)	0.0591 (ns)	0.0175 (ns)
		N3	-0.0042 (ns)	-0.0079 (ns)	-0.0029 (ns)	0.0116 (ns)	0.8485 (ns)
		N4	0.0116 (ns)	0.0199 (ns)	0.0055 (ns)	-0.0291 (ns)	0.6509 (ns)
	Dynamic	Current liabilities	-0.0509 (****)	-0.3914 (****)	-0.2904 (****)	0.7069 (****)	0.0259 (****)
		Liquidity	0.2106 (ns)	1.6164 (ns)	1.1993 (ns)	-2.9192 (ns)	-0.1071 (ns)
		LT Debt	-0.0582 (****)	-0.4471 (****)	-0.3317 (****)	0.8074 (****)	0.0296 (****)
		Working Capital	-0.0114 (***)	-0.0878 (**)	-0.0651 (**)	0.1586 (**)	0.0058 (**)
		acquisti	0.0018 (*)	0.0142 (*)	0.0105 (*)	-0.0257 (*)	-0.0009 (*)
		ebit	0.0409 (**)	0.3143 (**)	0.2332 (**)	-0.5676 (**)	-0.0208 (**)
		Turnover	0.0099 (****)	0.0757 (****)	0.0562 (****)	-0.1368 (****)	-0.005 (****)
		R1	0.0020 (ns)	0.0151 (ns)	0.0107 (ns)	-0.0269 (ns)	-0.0009 (ns)
		R2	0.0023 (ns)	0.0169 (ns)	0.0118 (ns)	-0.0299 (ns)	-0.0011 (ns)
		R3	-0.0034 (*)	-0.0274 (ns)	-0.0231 (ns)	0.0518 (ns)	0.0021 (ns)
		LagRating_4	-0.0157 (****)	-0.1522 (****)	-0.2306 (****)	0.3613 (****)	0.0371 (****)
		LagRating_5	-0.0314 (****)	-0.2405 (****)	-0.3465 (****)	0.5251 (****)	0.0933 (****)
		LagRating_6	-0.1026 (****)	-0.4008 (****)	-0.3555 (****)	0.6173 (****)	0.2417 (****)
		LagRating_7	-0.0264 (****)	-0.2293 (****)	-0.5007 (****)	-0.2166 (****)	0.9730 (****)

Table B.10
Table of PB marginal effects for SEC variables.

Model	Version	Variables	Marginal effects				
			y = 3	y = 4	y = 5	y = 6	y = 7
PB	Static	New Receivables	-0.052 (**)	-0.046 (**)	-0.0132 (**)	0.0727 (**)	0.0384 (**)
		Outstanding	0.0242 (****)	0.0217 (****)	0.0062 (****)	-0.0341 (****)	-0.0180 (****)
		Delinquency	-0.2319 (****)	-0.2080 (***)	-0.0593 (****)	0.3267 (****)	0.1725 (****)
	Dynamic	Collections	0.0042 (ns)	0.0164 (ns)	0.0084 (ns)	-0.0271 (ns)	-0.0021 (ns)
		Outstanding	0.0039 (****)	0.0158 (****)	0.0081 (****)	-0.0259 (****)	-0.0019 (****)
		Delinquency	-0.0542 (**)	-0.2143 (**)	-0.1103 (**)	0.3519 (**)	0.0269 (**)
		LagRating_4	-0.0342 (****)	-0.1718 (****)	-0.1851 (****)	0.3291 (****)	0.0621 (****)
		LagRating_5	-0.0656 (****)	-0.2709 (****)	-0.2998 (****)	0.4692 (****)	0.1670 (****)
		LagRating_6	-0.1835 (****)	-0.4003 (****)	-0.2866 (****)	0.4951 (****)	0.3755 (****)
		LagRating_7	-0.0541 (****)	-0.2563 (****)	-0.4291 (****)	-0.2386 (****)	0.9769 (****)

Table B.11
Table of PB marginal effects for BS+SEC variables.

Model	Historical	Variables	Marginal effects y = 3	y = 4	y = 5	y = 6	y = 7
PB	Static	Delinquency	-0.1739 (****)	-0.2928 (****)	-0.0834 (****)	0.4206 (****)	0.1295 (****)
		Turnover	0.0659 (****)	0.1110 (****)	0.0316 (****)	-0.1594 (****)	-0.0491 (****)
		Working Capital	-0.0884 (****)	-0.1488 (****)	-0.0424 (****)	0.2138 (****)	0.0658 (****)
		LT Debt	-0.3978 (****)	-0.6697 (****)	-0.1908 (****)	0.9619 (****)	0.2963 (****)
		Current liabilities	-0.2948 (****)	-0.4963 (****)	-0.1414 (****)	0.7129 (****)	0.2196 (****)
		Liquidity	1.9809 (***)	3.3348 (***)	0.9503 (***)	-4.7904 (***)	-1.4757 (***)
	Dynamic	Delinquency	-0.0399 (**)	-0.2439 (**)	-0.1881 (**)	0.7069 (**)	0.4012 (**)
		Turnover	0.0141 (****)	0.0861 (****)	0.0482 (****)	-2.9192 (****)	-0.1416 (****)
		Working Capital	-0.0179 (****)	-0.1095 (****)	-0.0613 (****)	0.8074 (****)	0.1802 (****)
		LT ebt	-0.0850 (****)	-0.5190 (**)	-0.2905 (**)	0.1586 (**)	0.8537 (**)
		Current liabilities	-0.0639 (*)	-0.3902 (*)	-0.2184 (*)	0.0257 (*)	0.6418 (*)
		Liquidity	0.4997 (**)	3.0508 (**)	1.077 (**)	-0.5676 (**)	-5.0185 (**)
		LagRating_4	-0.0211 (****)	-0.1522 (****)	-0.1881 (****)	0.3613 (****)	0.0385 (****)
		LagRating_5	-0.0386 (****)	-0.2405 (****)	-0.2978 (****)	0.5251 (****)	0.0959 (****)
		LagRating_6	-0.1111 (****)	-0.4008 (****)	-0.3212 (****)	0.6173 (****)	0.2354 (****)
		LagRating_7	-0.0331 (****)	-0.2293 (****)	-0.4708 (****)	-0.2166 (****)	0.9537 (****)

Appendix C. Feature importance

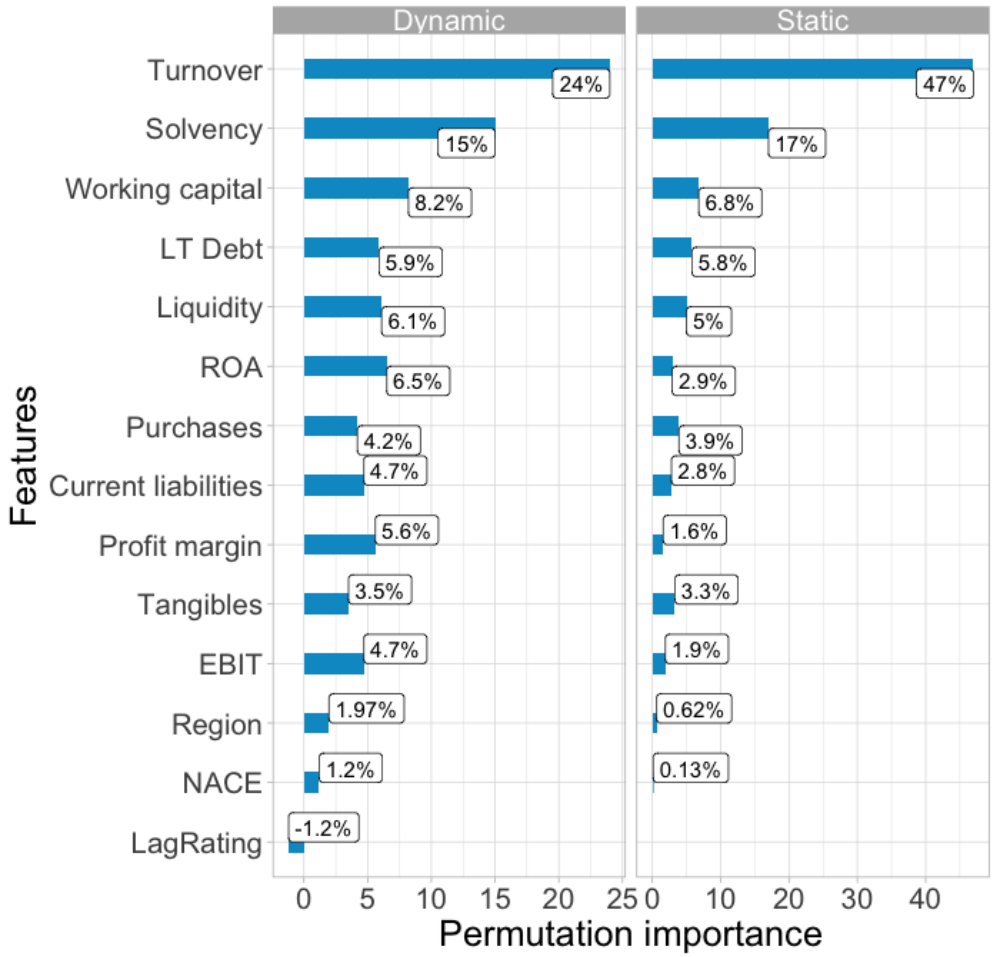

Fig. C.3. Macro-averaged relative permutation importance for HRF model for BS set (dynamic and static case).

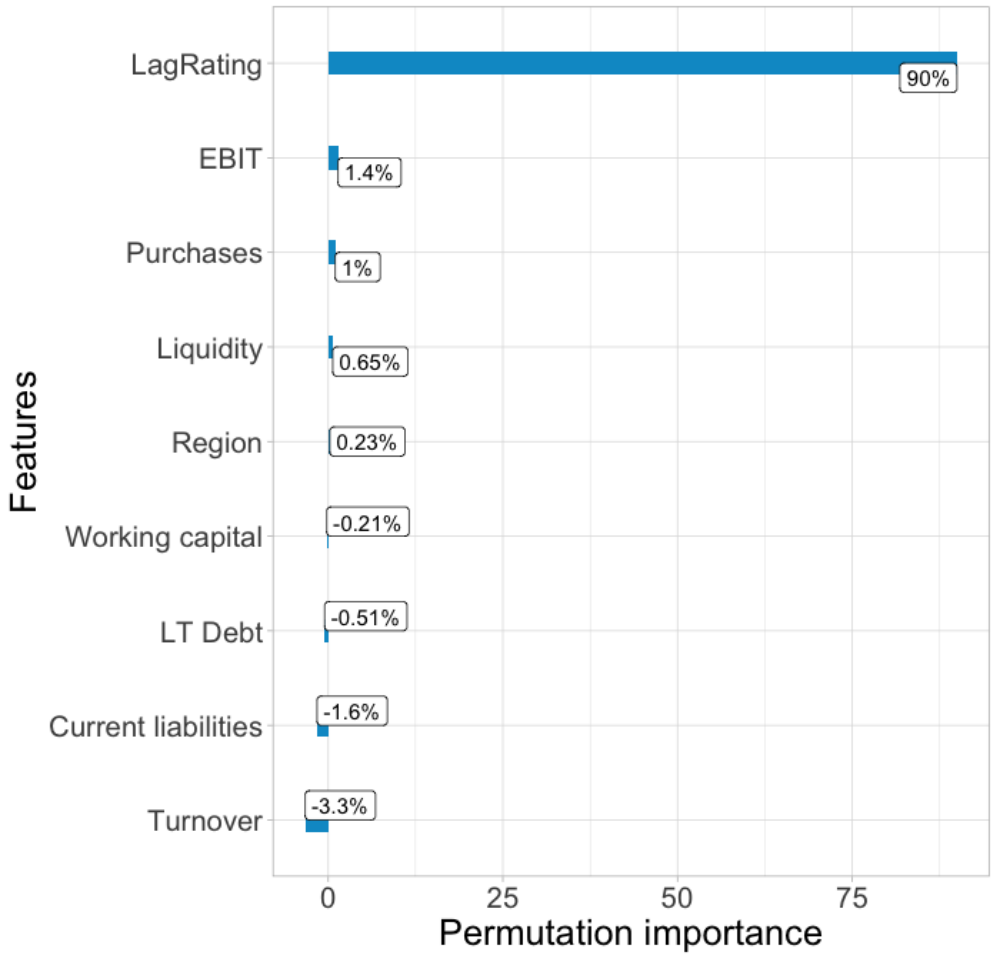

Fig. C.4. Macro-averaged relative permutation importance for PB model for BS set (dynamic case).

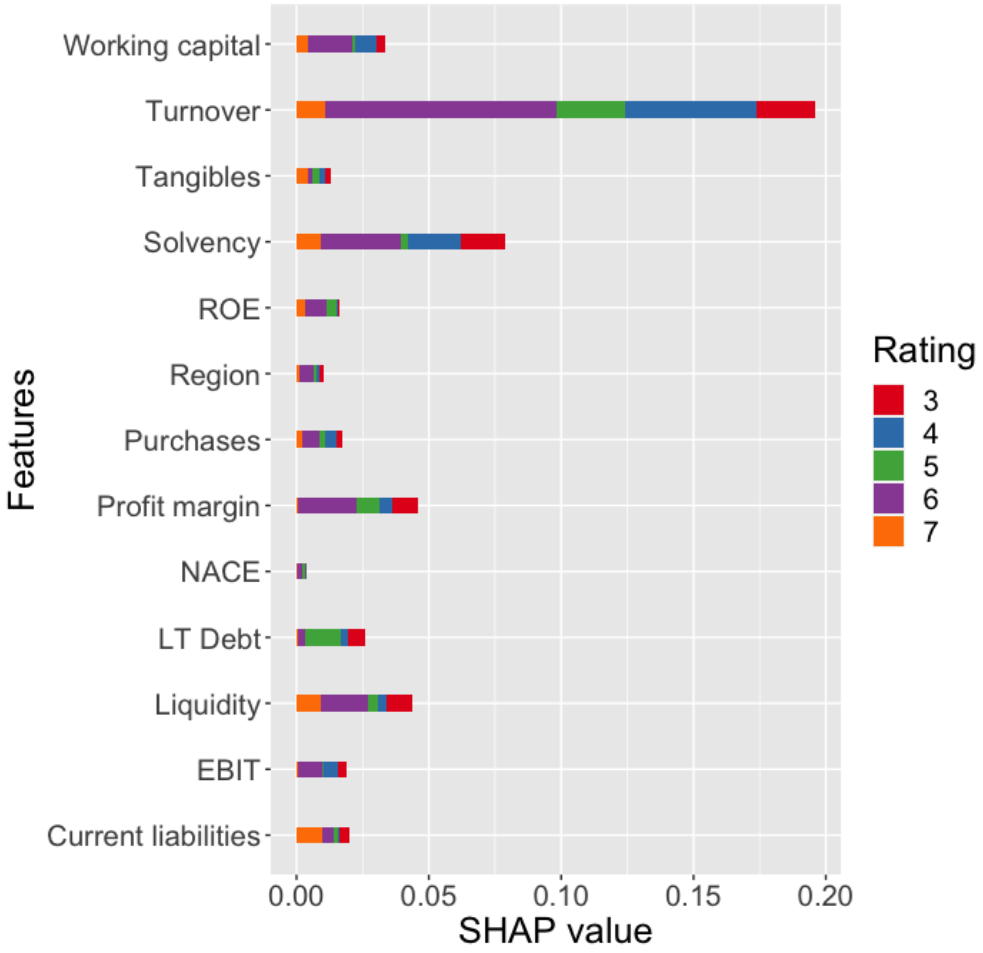

Fig. C.5. SHAP value (average impact of predictors for each class) for dynamic HRF model with regards to BS set.

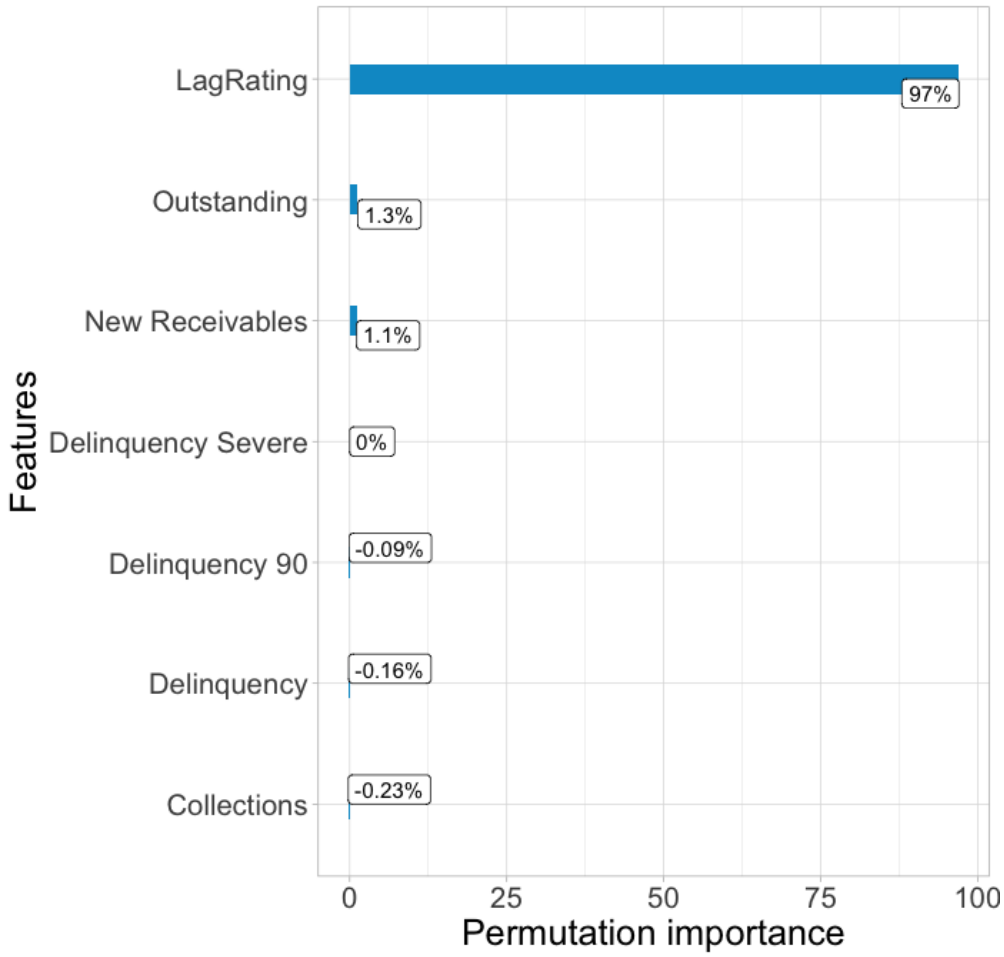

Fig. C.6. Macro-averaged relative permutation importance for HRF model for SEC set (dynamic case).

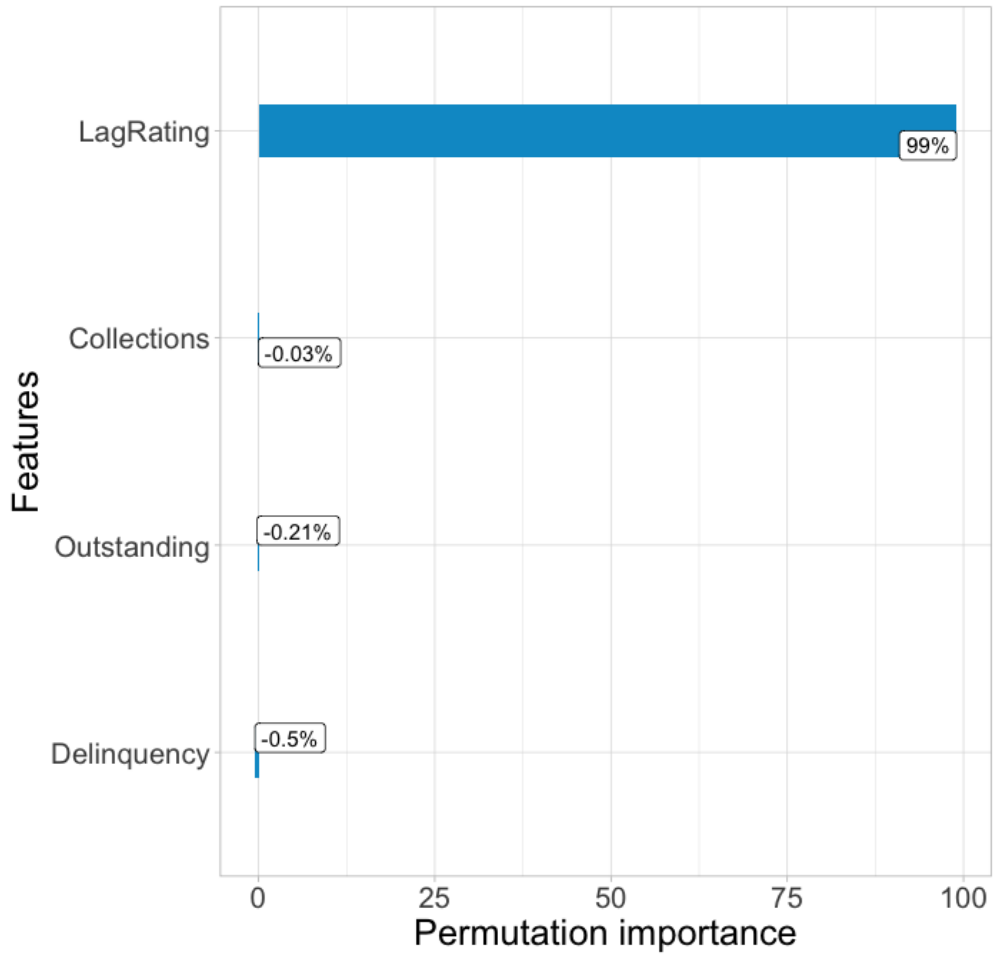

Fig. C.7. Macro-averaged relative permutation importance for PB model for SEC set (dynamic case).

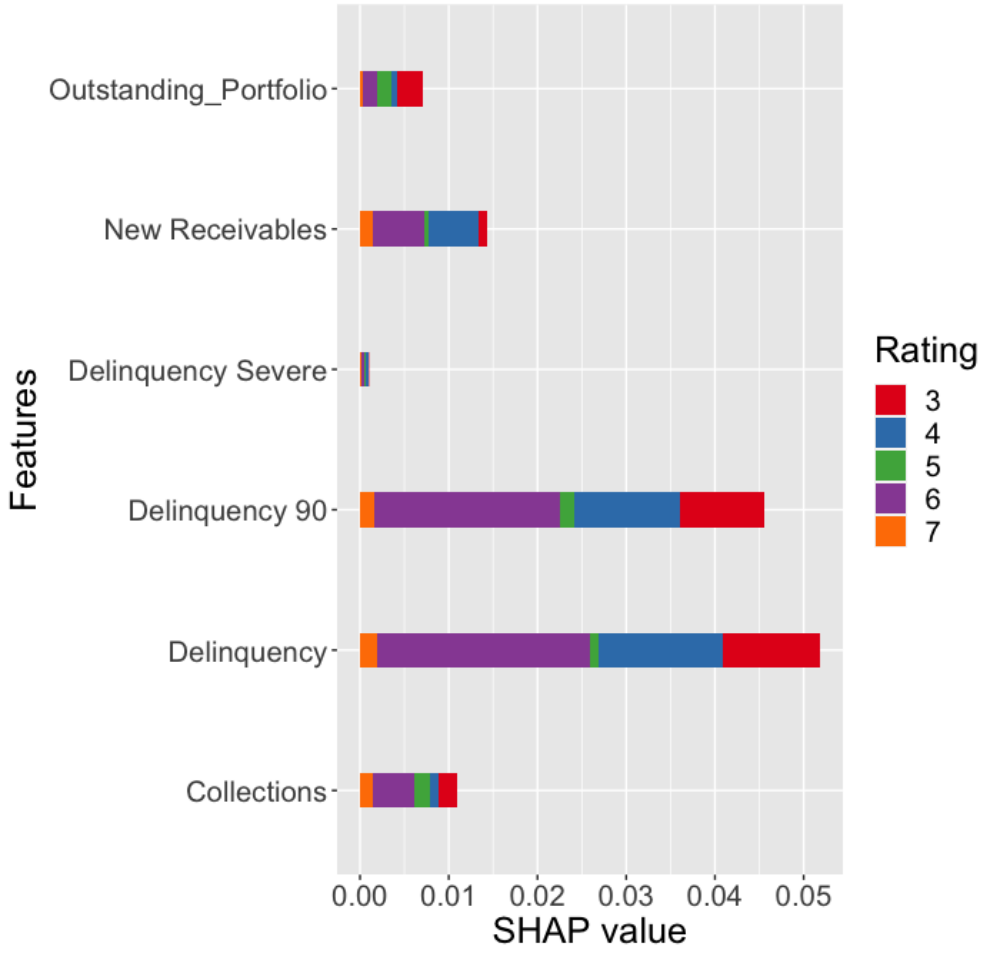

Fig. C.8. SHAP value (average impact of predictors for each class) for dynamic HRF model with regards to SEC set.

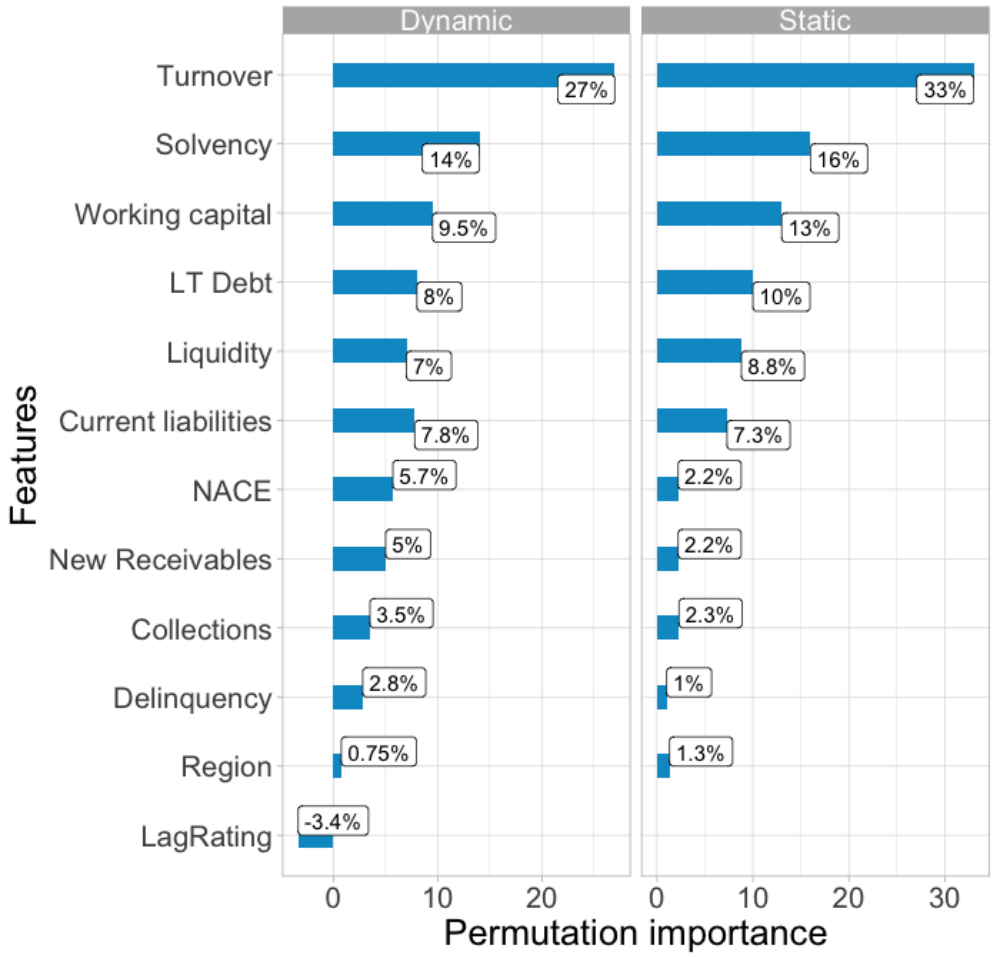

Fig. C.9. Macro-averaged relative permutation importance for HRF model for SEC+BS set (dynamic and static case).

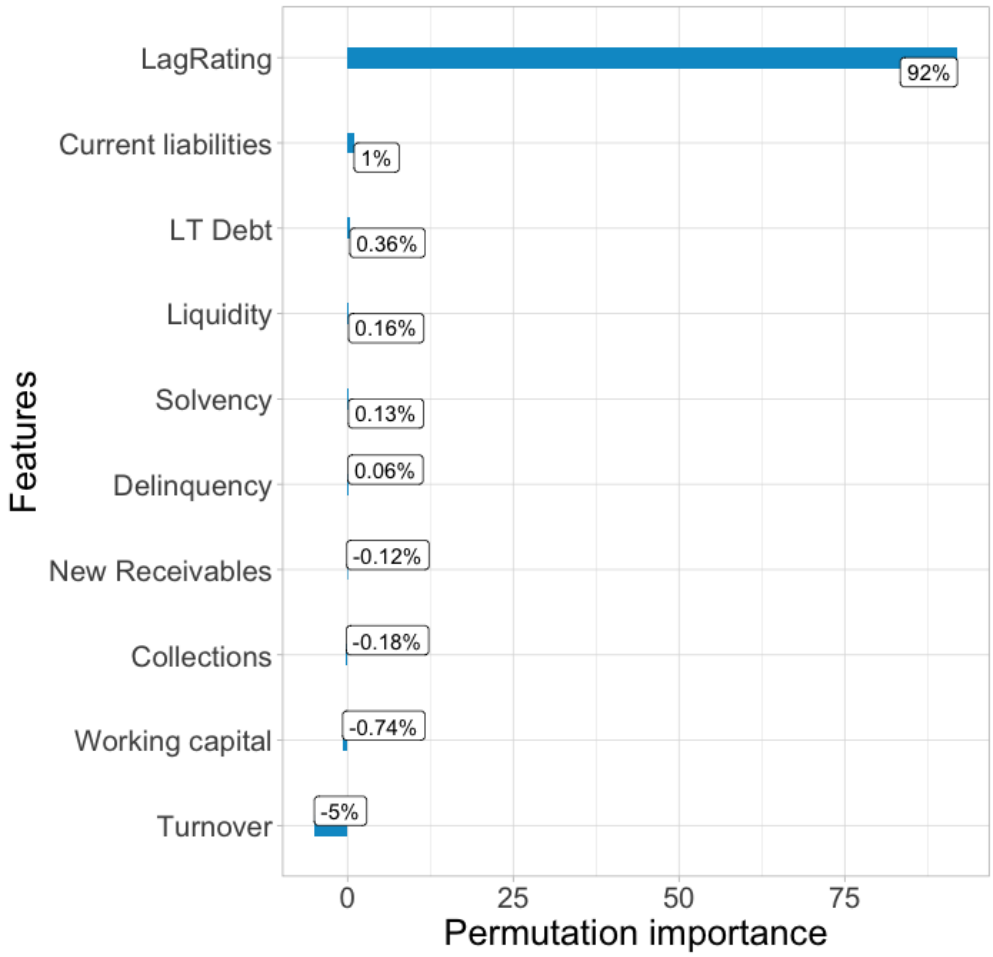

Fig. C.10. Macro-averaged relative permutation importance for PB model for SEC+BS set (dynamic case).

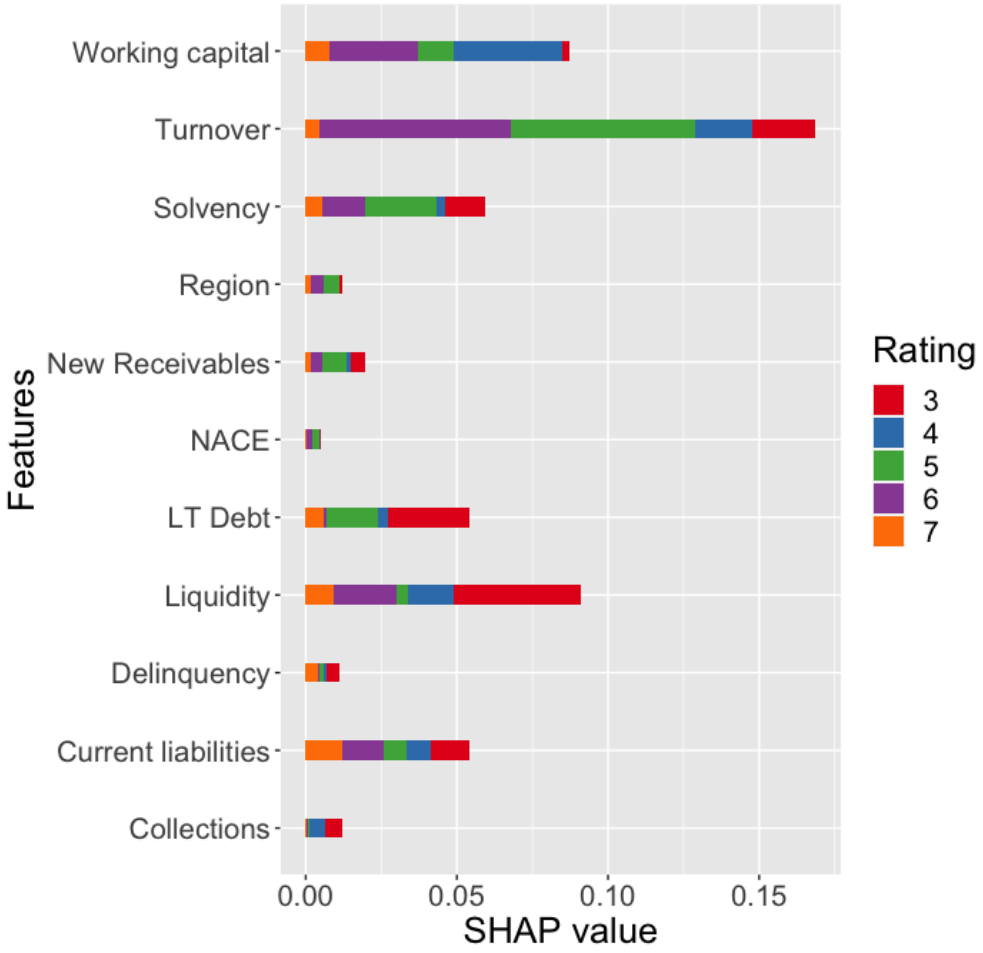

Fig. C.11. SHAP value (average impact of predictors for each class) for dynamic HRF model with regards to BS+SEC set.

Appendix D. Missing Values handling

In order to impute missing values for BS variable the following approach was used:

- for each BS variable, evaluate the average percentage increase/decrease of consecutive years:

$$\Delta_{t+1,t} = \frac{1}{N}\sum_{i=1}^{N}\frac{BS_i^{t+1}}{BS_i^t} - 1,$$

  where $t = 2015, 2016$ is the reference year and N it the total number of observations.

- impute missing value for t-th year given the $(t+1)$-th year by:

$$BS_i^t = \frac{BS_i^{t+1}}{1+\Delta_{t+1,t}}, \quad t = 2015, 2016$$

  for single missing year and impute value for t-th year given the $(t+2)$-th year by:

$$BS_i^t = \frac{BS_i^{t+2}}{(1+\Delta_{t+1,t})(1+\Delta_{t+2,t+1})}$$

  for double consecutive missing years.

Missing values for leading or trailing quarters of 2015 and 2017, respectively, are allowed for SEC variables given the unbalanced panel nature of the data.