BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning

by Giuseppe Cascarino, Mirko Moscatelli and Fabio Parlapiano

March 2022

Number 674

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning

by Giuseppe Cascarino, Mirko Moscatelli and Fabio Parlapiano

Number 674 – March 2022

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

EXPLAINABLE ARTIFICIAL INTELLIGENCE: INTERPRETING DEFAULT FORECASTING MODELS BASED ON MACHINE LEARNING

by Giuseppe Cascarino, Mirko Moscatelli and Fabio Parlapiano*

Abstract

Forecasting models based on machine learning (ML) algorithms have been shown to outperform traditional models in several applications. The lack of an easily interpretable functional form, however, is a major challenge for their adoption, especially when a knowledge of the estimated relationships and an explanation of individual forecasts are needed, for instance due to regulatory requirements or when forecasts are used in policy making. We apply some of the most established methods from the eXplainable Artificial Intelligence (XAI) literature to shed light on the random forest corporate default forecasting model in Moscatelli et al. (2019) applied to Italian non-financial firms. The methods provide insight into the relative importance of financial and credit variables to predict firms' financial distress. We complement the analysis by showing how the importance of these variables in explaining default risk changes over time in the period 2009-19. When financial conditions deteriorate, the variables characterized by a more complex relationship with financial distress, such as firms' liquidity and indebtedness indicators, become more important in predicting borrowers' defaults. We also discuss how ML models could enhance the accuracy of credit assessment for those borrowers with less developed credit relationships such as smaller firms.

JEL Classification: G2, C52, C55, D83.
Keywords: explainable artificial intelligence, mode-agnostic interpretability, artificial intelligence, machine learning, credit scoring, fintech.
DOI: 10.32057/0.QEF.2022.0674

Contents

* Bank of Italy, Directorate General for Economics, statistics and research.

1 INTRODUCTION[1]

Machine Learning (ML) algorithms often yield better forecasting performance than standard statistical models. Researchers and practitioners exploited this property in many fields, especially when many predictors are available, predictors' relationship with outcomes is non-linear, and interactions between predictors are necessary for accurate forecasting.[2] Despite their better performance, forecasting models based on ML often lack an easily interpretable functional form. This limitation constrains their use when explainability is required (Ertel, 2017). In this regard, eXplainable Artificial Intellingence (XAI) is a new branch of the ML literature that tries to improve our capacity to explain and interpret ML algorithms.[3]

In a nutshell, given a model based on ML, XAI methods can be employed to identify the set of rules and principles used by the model to make its forecasts and are extremely helpful in auditing it and testing its ability to provide consistent results in a variety of applications.[4] XAI methods, therefore, enhance our understanding of 'black-box' ML models (Doshi-Velez and Kim, 2017; Molnar, 2019). These methods help, for instance uncover issues that may hinder the accuracy of the model when applied to out-of-sample observations (Ribeiro et al., 2016) by testing the consistency of the prediction rules of the model with prior knowledge on the phenomenon of interest (*e.g.*, known facts and causal relations). XAI methods may also reveal problems in the dataset used for the model's training, such as its partial or biased representation of the population of interest or circumstances under which the model may learn inaccurate forecasting rules. Explainability is also helpful to verify whether a model suggests unethical decisions, due to the use of impermissible information such as gender, ethnicity, or religion, either directly or by using proxy variables.[5]

We apply recently developed XAI methods to the corporate default forecasting models described in Moscatelli et al. (2019), namely the Random Forest model (RDF) and the standard logit model. In our application, logit is a benchmark and we apply post-hoc, model-agnostic XAI methods, *i.e.* methods applied after the training of the model (post-hoc) and that can explain any type of model irrespective of their structure and complexity (model-agnostic), to assess: i) the importance of each variable in determining the predictions of the model; ii) the

[1] The authors wish to thank Alessio De Vincenzo, Emilia Bonaccorsi Di Patti, Giuseppe Cappelletti and Francesco Columba for their comments.

[2] Machine learning algorithms are designed to perform tasks (Mitchell, 1997), such as predicting the value of a variable, by learning decision rules from the data and improving automatically through experience. In contrast to standard statistical models, potential constraints to the learning process are overcome by making minimal or no assumptions on the structure of the data generating process and the model's parameters. These features allow ML algorithms to explore complex patterns such as non-linearities, non-monotonicity, and interactions amongst predictors.

[3] The concept of interpretability and explainability are often used interchangeably: "Interpretability is the degree to which a human can understand the cause of a decision" (Miller 2018). In some applications, knowing the 'why' of a specific prediction can help learn about the problem, the data, and why a model might fail. The need for interpretability may not arise in a low-risk environment, *e.g.* when a forecasting error does not have serious consequences. However, getting good predictions (the what) may not always be enough. The modeler must explain how it came to the prediction (the why). For additional references on the concept of explainability, see also Miller (2018) and Arrieta et al. (2020).

[4] In some applications, however, explainability is less crucial or explanations are not supposed to be disclosed. The first case occurs when the model has a purely predictive task and incorrect predictions have no significant impact, while the changes needed to obtain explainability may come at the expense of predictive performance (*e.g.*, in a shopping recommendation system). In other cases, one may want the model's inner working to remain opaque for a certain audience. Fore example, we may need to protect intellectual property, or we may be concerned that interpretability would allow users to manipulate the information provided to obtain a favourable outcome.

[5] For instance, the insurance sector in the EU is required to comply with a gender-neutral pricing policy, despite a possible predictive relationship between gender and risk exposure which could be exploited by ML algorithms (https://ec.europa.eu/commission/presscorner/detail/en/IP_12_1430).

relationships between the variables and the model's estimated probabilities of default, and iii) how the contribution of each variable changes across observations, thereby explaining the determinants of individual predictions. We also estimate the RDF model using a one-year rolling window approach over the 2009-2019 period and employ XAI methods to evaluate changes in the relevant predictors of firms' default through crises and recovery times.

Existing literature increasingly applied ML models to credit scoring, both for consumer and corporate default forecasts. Most such applications result in better reported predictive performance than traditional models, such as logistic regressions (see, among others, Chakraborty and Joseph, 2017; Fuster et al., 2020; Khandani et al., 2010; Barboza et al., 2017; Bacham and Zhao, 2017; Fantazzini and Figini, 2009; Moscatelli et al., 2019, Alonso and Carbo, 2020). In addition, new digital technologies facilitated the use of ML models, allowing financial intermediaries to collect and use larger information sets (both in terms of variables and observations) for the training of the models. In contrast to traditional models, such as the logit model, machine learning models can empirically detect non-linearities in the relationships between the outcome variable and its predictors and detect interaction effects among the latter. In this regard, Moscatelli et al. (2019) show that ML models consistently outperform traditional statistical models employed in credit scoring when the training dataset is large enough. They also show that a decision rule based on the predictions of the ML model results in a greater share of credit allocated to ex-post relatively safer borrowers compared to the logit model, which suggests that the transition from traditional to ML systems may have consequences for credit allocation. In turn, the greater accuracy of default forecasts can provide advantages to banks in terms of both lower credit losses and regulatory capital savings (Alonso and Carbó, 2021).

However, against these upsides, ML models are "black boxes" and pose new challenges to institutions operating in the regulated financial services sector and their supervisors.[6] The difficulty of linking predicted default probabilities to borrowers' characteristics limits the accountability of ML models. The lack of interpretability of individual outcomes, for instance, in assessing prospective borrowers' default risk, poses challenges for the accountability and auditing of estimated default probabilities against expert knowledge or the borrower's actual characteristics. Moreover, from the regulatory perspective, the explainability of credit decisions is a consumers' right in itself that supervisors need to ensure. In the USA the right to explanations for adverse decisions was introduced in 1975 with the Equal Credit Opportunity Act (ECOA), in order to contrast the discrimination of credit applicants based on their color, religion or other non-creditworthiness-related information. In the EU, credit applicants' rights are more limited; for instance, prospective borrowers are, under certain circumstances, entitled to receive meaningful information about the logic of automated decision-making. In addition, non-legislative bodies have put forward explainability among ML systems' desirable requirements (AI HLEG, 2019; EBA, 2020; ROFIEG, 2019; Dupont et al., 2020). Trustworthy ML systems demand, *inter alia*, the possibility to understand and track their decision-making process from both a technical perspective and a societal one. This possibility rests on providing a suitable and timely explanation adapted to the various stakeholders' expertise (e.g. non-specialist, regulator or researcher). The recognition of a right to an explanation seems to have significantly affected the pace of adoption of ML in credit scoring, conditioning its adoption to the development of appropriate explanation techniques.

Our analysis suggests that XAI methods can be of substantial help in revealing the decision rules behind a complex RDF model to forecast corporate default. In particular, we show that (a) default forecasts based on the RDF model exploit the available information set more broadly, assigning greater importance to indicators that

[6] Financial firms' supervisors are challenged in several ways by the adoption of forecasting models based on ML: on the one hand, they are confronted with machine learning models adopted by financial intermediaries in many aspects of their business processes (from business development to risk management), which would eventually need to be validated and monitored; on the other hand, they may use machine learning models for surveillance (SupTech) and other internal processes.

display non-linear or non-monotonic relationships with the outcome variable than the logit model. In contrast, the benchmark logit model mainly uses distress signals from a smaller set of indicators, that display an almost linear association with the insolvency. Furthermore, (b) RDF forecasts display stability over time in the importance assigned to key predictors (such as credit indicators) and in their estimated relationship with the probability of default; this notwithstanding, (c) when financial conditions change, liquidity ratios also give a relevant contribution to predictions. Finally, our research suggests that the joint use of complex models based on ML and XAI techniques can also advance our understanding of the determinants of financial risks . ML and XAI can help us pin down several features of the relationship between credit risk and its determinants (such as threshold or cliff effects). Such knowledge can become helpful even in designing traditional statistical models or the specification of predicting variables.[7]

Our work is related to a growing literature applying XAI methods to credit risk models. Chen et al. (2018) provide an interpretable model for credit risk assessment that is as accurate as other black-box neural network models. Bussmann et al. (2020) and Ariza-Garzón et al. (2020) propose a novel XAI model, built on similarity network models applied to the Shapley values of individual predictions, to improve the explainability of credit risk scores assigned by a peer to peer lending platform. Finally, Visani et al. (2020) use a relatively novel XAI method, LIME, to interpret a credit scoring model based on gradient boosted trees and develop a method to assess the stability of the resulting explanations.

The rest of the paper is organized as follows. Section 2 describes the XAI methods. Section 3 describes the dataset and the predictive models used to forecast non-financial firms defaults. Section 4 presents the results obtained from applying the explanation methods to the forecasting model. Section 5 concludes.

2 XAI METHODS AT A GLANCE

This section provides a brief overview of the XAI methods used in our work (for a technical description, see Appendix 2). First, we illustrate the permutation variable importance method, which is used to assess the relevance of different variables for the predictions made by the model. Second, we describe the Accumulated Local Effects plot (ALE plot), which visually represents the relationship between the individual variables and the estimated forecasting outcomes. Finally, we introduce the Shapley values method, which is used to compute the contribution of the characteristics of an individual observation to its predicted outcome.

2.1 Measuring variable importance: the permutation variable importance method

The permutation variable importance method provides a quantitative assessment of the importance of individual predictors within a forecasting model, allowing to answer the question of which variables have the biggest impact – on average – on the predicted outcomes.[8] In turn, the method yields information about the structure of the model and the potential problems that statistical errors (or noise) in a variable could cause on its predictions.

The importance of a variable is gauged by comparing the predictions obtained from the forecasting model applied to two different datasets, namely the original test dataset and a permuted test dataset where the values of

[7] The usage of predicting variables inferred from ML models to adjust the functional form of traditional statistical models is discussed, for example, in Dumitrescu et al. (2021).

[8] The method was originally introduced by Breiman (2001) in his seminal paper on random forests and then extended in several works, including Altman et al. (2010), Datta et al. (2016), and Fisher et al. (2018).

the variable of interest have been reshuffled randomly. The reshuffling preserves the variable of interest's marginal distribution, breaking at the same time its relationship with the other variables and the outcome. As a result, the greater the importance of the variable, the more diverging the two sets of predictions will be.

Variable importance (i.e. the difference between the two sets of predictions) can be computed using several metrics. Some of the most used ones include the extent to which the permutation reduces the model's forecasting power (accuracy) and the contribution of a variable the forecasts' variability, i.e. the extent to which the permutation changes the model's predictions. The joint use of these metrics returns information about those variables that give a low contribution to the accuracy of forecasts but whose variation induces a significant change in the predictions, a signal of potential overfitting.[9]

The contribution of a variable to the accuracy can be computed as the difference between the out-of-sample AUC of the predictions obtained from the original dataset and the permuted dataset.[10] Higher values for this metric indicate that the predictive power of the model strongly depends on the variable of interest, as its permutation greatly decreases the AUC. A variable's contribution to the forecasts' variability is computed as the average absolute difference in the predictions, for the same firm, between the original dataset and the permuted dataset. Higher values for this metric indicate that individual predictions are sensitive to changes in the variable of interest.

2.2 The relationship between predictors and the outcome: Accumulated Local Effects plot

A visual representation of the relationship between each variable and the predictions of the model can be obtained using the Accumulated Local Effects (ALE) plot (Apley and Zhu, 2019). The method can give insights into the existence of non-linear or non-monotonic relationships, which could be accounted for in a variety of contexts. For instance, a possible application of the method may include credit loss provisioning, where banks and supervisors can identify some early warning indicators and related trigger values to be used as proxies of firms' non-performing status; dependency plots can help to identify if and for which values of a certain variable the empirically estimated effect changes significantly due to cliff or threshold effects.[11]

Usually, the effect of a specific characteristic X on the model's outcome depends on the values of all the other variables. Thus, there is no single prediction for each value of X. Therefore, the task is to properly define and calculate some expected prediction at each given value of X, accounting for the statistical dependence between variables. Unfortunately, it can be shown that basic approaches to the problem -- such as calculating an average prediction drawing from the other predictors' unconditional joint distribution or taking a local average of observations sharing the same value of X -- suffer from severe shortcomings when predictors show statistical dependence.[12]

[9] A model is said to overfit when its forecasting performance on a dataset is substantially worse than the performance on the training dataset used to fit the model. In contrast, when a model does not fit the training data accurately, this is known as underfitting, and the model is likely to have a large bias since it is not complex enough in terms of the features or the type of model being used.

[10] The AUC is a common measure of the discriminatory power of a binary classification model. It captures the model's ability to assign higher probabilities to positive outcomes compared to negative ones. For example, in the case of credit risk assessment, a random model that does not discriminate between sound and distressed firms has a 0.5 AUC, while a perfect model has an AUC of 1.

[11] ECB (2017), "Guidance to banks on non-performing loans", European Central Bank.

[12] A first approach, also known as the Partial Dependence Plot by Friedman (2001), requires taking an average drawing from the unconditional joint distribution of the other variables. However, this is likely to return unrealistic results: the stronger the correlation between predictors, the more likely the difference between the unconditional joint distribution and the

The Accumulated Local Effects (ALE) plot (Apley and Zhu, 2019) solves these issues. For a given variable X of interest, the ALE function is computed as the cumulative sum of the average differences in the predictions obtained by varying X in a small neighborhood around each value, keeping all other variables fixed. Therefore, the method identifies the effects of a small change in X for each value of X, while keeping all other factors (almost) unchanged. By summing these local effects, the overall effect of a change in X can be obtained. In order to facilitate a comparison of the effects of different variables, the average effect is usually subtracted so that the average effect shown in the plot equals zero.

2.3 Explaining predictions for individual observations: Shapley values

The Shapley Values method (Štrumbelj and Kononenko, 2014) aims at measuring the contribution of each observation's characteristic to the prediction returned by the model.[13] Although the Shapley Values method employs a rather complex computing methodology (see Appendix 2), the underlying idea is straightforward: Given an observation of interest, the method computes the marginal contribution of a variable to the prediction as the difference between the model's prediction for that observation and the average prediction that the model would return, for that observation, if the value of the variable were unknown. This marginal contribution is called the Shapley Value of the variable. In order to take variables' correlation into account, the method estimates the marginal contribution of the predictor over all the possible subsets of the other variables. The sum of all variables' Shapley Values is equal to the difference between the model's prediction for the observation of interest and the model's average (or baseline) prediction, i.e. the best prediction possible if the values of all the variables for the observation were unknown.

The use of the Shapley Values method entails several advantages: it enables identifying the drivers of a specific prediction, allowing to integrate the prediction in a braoder decision process or to justify the decision made. Furthermore, understanding the rules behind a prediction increases humans' acceptance of the model. The possibility to explain which characteristics of a given observation determined its prediction is particularly important – for instance – in credit scoring, where customers are entitled to obtain an explanation for why they were denied credit and where predictions made by models are often integrated with expert judgment.

3 DATA AND DEFAULT FORECASTING MODEL

Following Moscatelli et al. (2019), we estimate the probabilities of default (PD) of Italian NFCs with a wide set of financial and credit behavioral indicators drawn from the Company Accounts Data System (provided by Cerved) and the Italian Credit Register (CR). Our data spans the 2009-2020 time frame: the years 2019 and 2020

conditional one. For example, some variables are strongly correlated in our dataset since they involve common quantities (e.g. the cash to total assets or the cash to short term ratios). Thus, finding a firm for which the indicators are not correlated is implausible. A second approach draws from the subset of observations where the variable of interest has the same value. This approach avoids extrapolation problems at the cost of not isolating the effect of X on the outcome: as long as the predictors are not independent, the local average prediction is, in fact, also strongly affected by the values that other variables tend to assume in correspondence of that particular value of X. Therefore, these two basic solutions are both unsatisfactory (see also Apley and Zhu, 2019).

[13] The concept of Shapley Values originated in the field of game theory (Shapley, 1953): in cooperative games, the Shapley Values measure how the payout should be fairly distributed among the players in such a way that each player receives a share proportional to his contribution, being the unique distribution satisfying some properties which are meaningful also in the context of explainability (see Joseph, 2019 and Štrumbelj and Kononenko, 2014).

are used as training (or estimation) and test (or validation) dataset respectively for the main analysis, while the remaining years are used in a time-series application. We employ firms' credit status, a dummy equal to one for non-performing or defaulted borrowers, as the dependent variable. Our definition of default is based on a system-wide assessment of the borrower's credit exposure: a firm is classified as 'non-performing' at the end of the year if the ratio of non-performing credit to total credit has been greater than 5 per cent for at least three months over the course of the year.[14]

A wide set of firm-level predictors is employed to train the credit risk model. We calculate twenty-four financial indicators relating to profitability, financing structure, debt sustainability and the maturity of assets. Credit behavioral indicators include eight variables providing a measure of a firm's financial flexibility, such as the proportion of drawn to granted bank credit and the occurrence of delinquencies within a firm-bank credit relationship. We also consider firms' descriptive indicators, such as the economic sector and the geographical area. A full description of the variables is given in Table A1. Descriptive statistics on the dependent variable and predictors are provided for both the training and test dataset in Tables A2-A5.

We estimate one-year-ahead PD using financial information available with a time lag of 12 months and credit behavioral indicators available with a time lag of two months; indeed this is the usual delay with which such information is available from Cerved and the Italian Credit Register. The PD is estimated only for firms that are not already in default.

We estimate firms' PDs using both the Random Forest (RDF) algorithm (Breiman, 2001; Hastie et al, 2009), and the standard logistic regression model (Logit)[15], which we consider as a benchmark. The RDF has been used extensively in credit risk modelling, both in academic and industry applications. However, since its predictions combine hundreds or thousands of different forecasts, it is very difficult to trace back the contributing factors of an individual prediction and hence explain the reasons behind a decision. The estimation of the model follows the approach of Moscatelli et al. (2019). In particular, grid search and five-fold cross validation maximizing AUC are used to find the optimal hyper-parameters of the model, namely (*i)* the number of variables selected at each split and (*ii)* the minimum number of observations in each leaf.

4 XAI METHODS AT WORK

4.1 The importance of corporate default predictors

To investigate the importance that each variable has in predicting firms' financial distress, we apply the permutation variable importance method (section 2.1) to the set of predictors utilized by the RDF model. We also apply the same method to the Logit model in order to compare how variables' importance differs between the two models. The results are shown in Figure 1. Using the contribution to out-of-sample accuracy as the preferred metric for assessing feature importance, we found that RDF and logit differ quite substantially in the weight given to financial relative to credit behavioral indicators. While both models attribute high importance to a firm's total and short term credit drawings, the accuracy of RDF relies on a wider set of predictors, including financial indicators

[14] The status of non-performing loans recorded in the Italian Credit Register includes three stages of impairiment: 90 days past-due, unlikely to pay and bad loans. Our definition of default is aligned to the one used by the Bank of Italy's In-house Credit Assessment System (BI-ICAS) and by Moscatelli et al. (2019).

[15] In which the relationship between our indicators X and future default Y is assumed to be of the form $Y = \Lambda(X^T\beta)$ where $\Lambda(\mathrm{a}) = \frac{\exp(\mathrm{a})}{1+\exp(a)}$ is the logistic function.

(such as the liquidity ratios: cash to short-term debt and cash to total assets). In contrast, the accuracy of the logit model is mostly dependent on the credit behavioral indicators.

This difference is key since credit behavioral indicators might not be available for some borrowers (due to their thin or even inexistent lending relationships), especially for the smaller or younger ones; hence ML models can provide more opportunities for a wider credit assessment and access to credit. Moreover, RDF's ability to exploit a larger set of financial information might come as an advantage for a more comprehensive credit quality assessment of borrowers: for instance, due to their lower frequency and different time horizon, financial variables might 'stabilize' otherwise highly seasonal point-in-time information such as a firms' drawings from credit lines.[16] For example, a trade credit indicator (*Receivables Turnover*) and profitability indicators (EBITDA margin and turnover) appear to have significantly more importance for the RDF model.

A second insight from the use of the permutation method stems from differences in the ranking of variables, both under the contribution to accuracy or impact on predictions criteria. For instance, the occurrence of credit delinquencies in the past (*NPL*, a dummy equal to 1 if the firm has deteriorated loans)[17] appears as the fifth most important variable for the accuracy of the RDF model, but it has a very low impact on predictions (in terms of average change in probabilities). This finding is related to the distribution of the variable itself: its value is zero in most cases (see Table A2), therefore it has limited impact on the predictions for the majority of the observations; however, in the cases when past delinquencies are detected, this information feeds the model with a very accurate signal about a firm's distress.

The permutation method also allows to assess the potential overfitting of a model. One symptom of overfitting is when variables that do not contribute significantly to a model's accuracy have a significant impact on its forecasts. Both the RDF and the logit model do not seem to outweigh variables with low accuracy; the value added to total assets ratio (*Value Added to Total Assets*) might be the only exception, since it shows a low contribution to accuracy while it ranks high by impact on firms' PDs.

4.2 The relationships between firms' characteristics and default

The ALE plots method (Figure 2) shows that for the RDF it is non-linear for all variables and non-monotonic in the cases of liquidity (Cash to Total Assets) and debt sustainability ratios (interest expenses to cash flow, Interest Expenses to Cashflow).[18] This latter evidence explains why financial ratios are more important for the RDF model, which unlike the logit captures non-monotonic relationships.

Another important insight is the consideration that the usual assumption of fixed marginal effects underpinning standard credit scoring models is strongly misleading in approximating the empirical relationship between predicting variables and defaults. Most variables display threshold effects, that is, a sharp increase in the average PD for values above a certain threshold. For instance, when a firm approaches the maximum usage of its credit facilities its predicted PD level increases more rapidly, as can be seen from the blue line in Figure 2 (last two panels). Similarly, when cash buffers fall below 25 per cent of short-term financial liabilities, the PD increases sharply.

[16] Financial variables are collected annually and record year-end firms' assets, liabilities and 12-month economic performances. Credit behavioral indicators are collected monthly and record a borrower's month-end credit exposure.

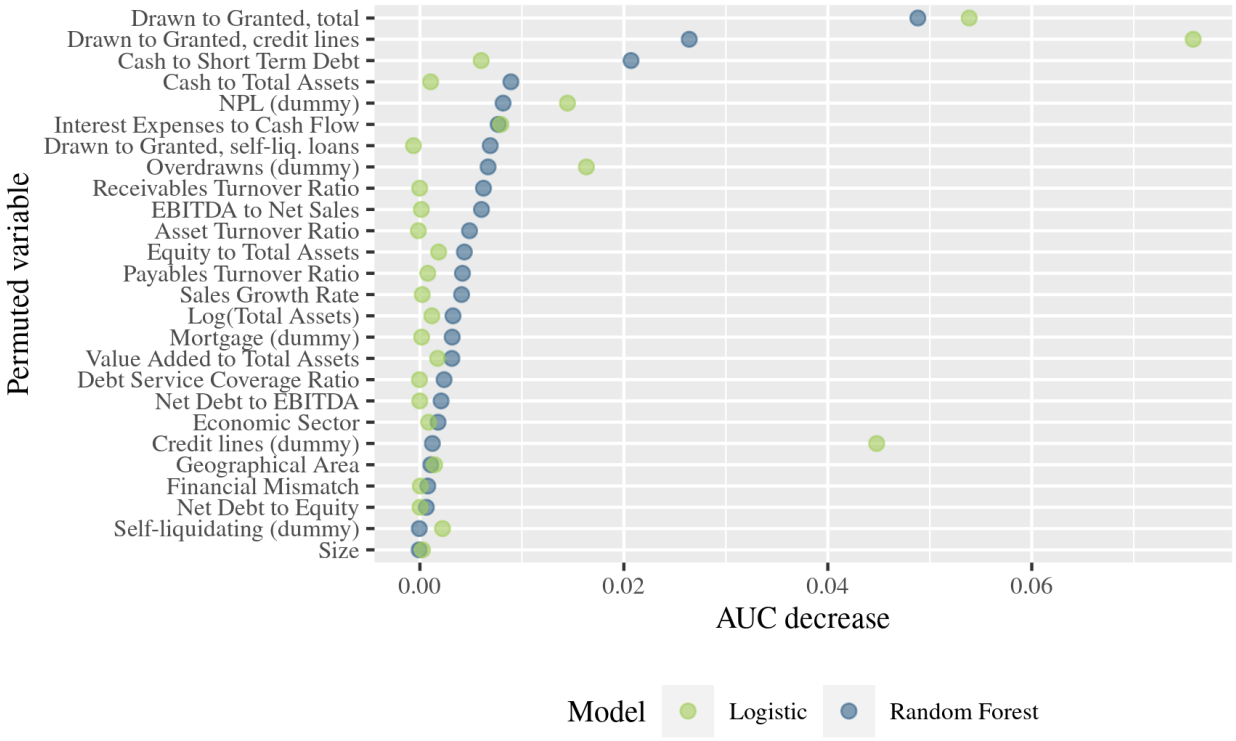
[17] Deteriorated loans are identified if these loans account for less than the 5 per cent overall exposure threshold that identifies firms in distress.

[18] The full set of ALE plots is reported in the Appendix.

Visual inspection of ALE plots may also turn helpful in the design of standard statistical models: for instance, the observation of non-linearities may suggest the inclusion of non-linear terms or threshold effects in the specification of a logistic regression model in order to increase its flexibility.

Figure 1: Variable importance (permutation method)

Panel A - AUC decrease



Panel B - Mean probability difference

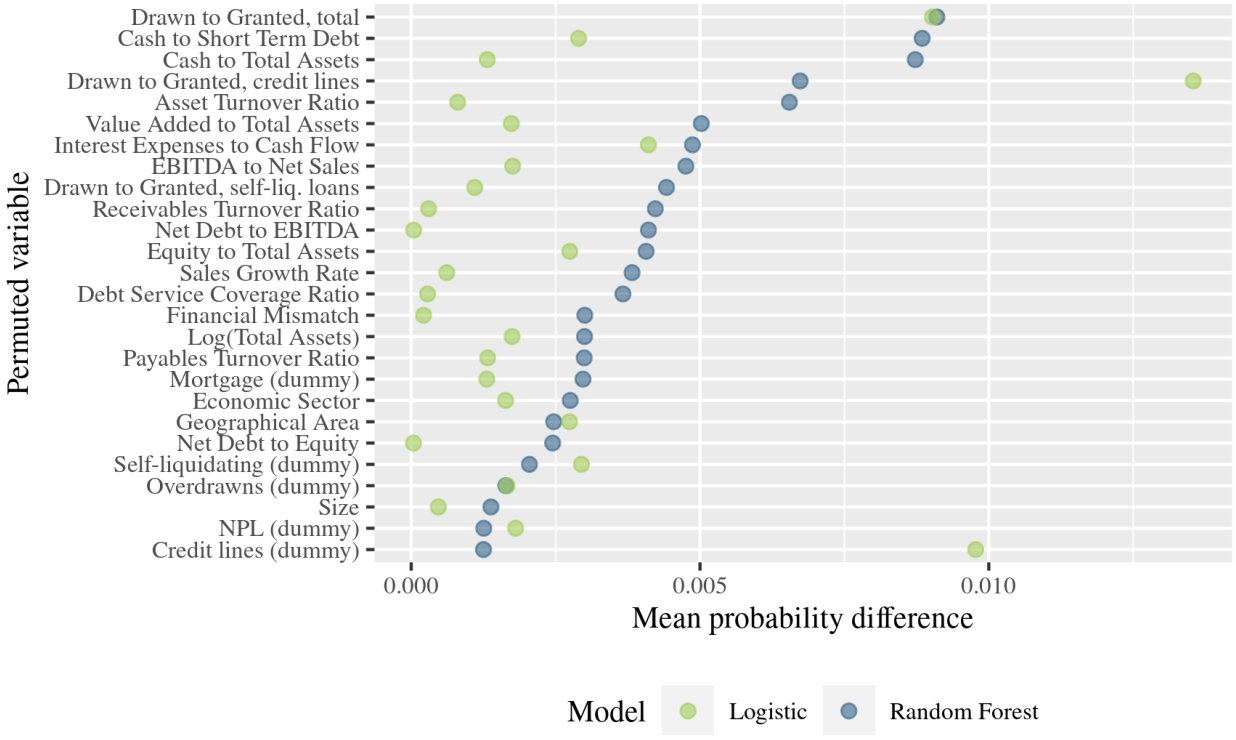
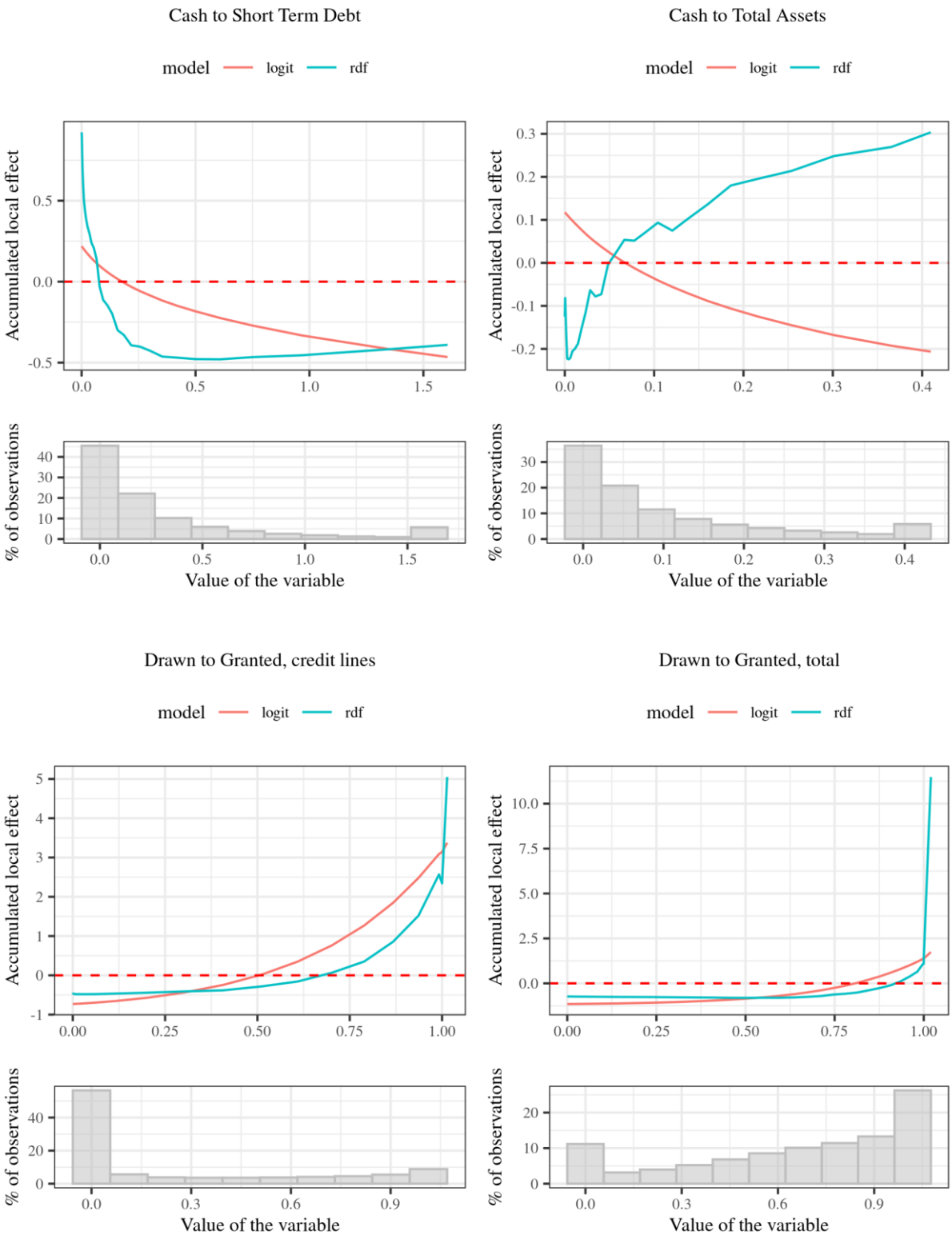


Figure 2. ALE plots of the first four variables by importance [(1)]



Notes: (1) The ALE is represented as a difference from the mean effect, so that the mean effect is 0.

4.3 The contribution of a firm's characteristics to its PD

In this section we apply the Shapley values method to explain which factors contribute to determining individual borrowers' estimated PD and how. As an example, we select three observations corresponding to borrowers for which the RDF model estimates a PD equal to 0.4, 1.5 and 5 percent respectively, placing them in different risk classes. Table 1 reports the value of the predictors for each observation, along with the means and medians from the overall sample. The contributions of individual variables to the predicted PDs are measured by the Shapley values, which are reported in Table 1 and represented graphically in Figure 3. We also report, for the same borrowers, the Shapley values for the PDs predicted by the logistic regression model, in order to show how the underlying structure of the two models differs for the same borrowers.

The main results are the following. The most important variables in terms of global importance (as shown in Figure 1), i.e. the credit behavioral indicators, are also among the main drivers for the estimated PDs of the three selected observations. However, the magnitude of the local effect may differ substantially between observations: for instance, the Shapley values for the riskier borrower clarify that for this subject the estimated PD is increased mostly by the high share of cash flow absorbed by interest expenses (*Interest Expenses to Cash Flow*) as well as the amount of drawn credit over granted (*Drawn to Granted total*, *Drawn to Granted credit lines*).

In addition, in the case of RDF it is possible to observe non-monotonicity for some variables. For instance, in the logit regression model, firms' credit risk relates negatively with the sales growth rate (SALES_GWT), with fast-growing firms' being assessed as less risky. However, in the RDF model the sales growth variable has a positive Shapley value both for the first two borrowers, which have a negative rate of -50%, and for the riskier one, for which the growth rate is positive and very high (+80%). This result is consistent with the ALE plots of the variable *Sales Growth Rate* (see Appendix 3), which shows an almost linear negative relationship being estimated by the logit and a "U-shaped" relation for the RDF. A possible reason behind this non-monotonicity is the fact that, for debtors with volatile economic performance, very high growth rates often appear because of a base effect, i.e. when previous year's sales were particularly low. The model is therefore likely capturing the fact that, for borrowers showing weakness in other financial and economic indicators, fast-growing sales result from the volatility of business conditions, rather than the firms' positive economic perspectives.

Overall, the use of Shapley values allows us to understand how the variables contribute to the predicted PD for a specific borrower and, consequently, which are those that influence the most the choice of classifiyng a borrower in low or high-risk classes.

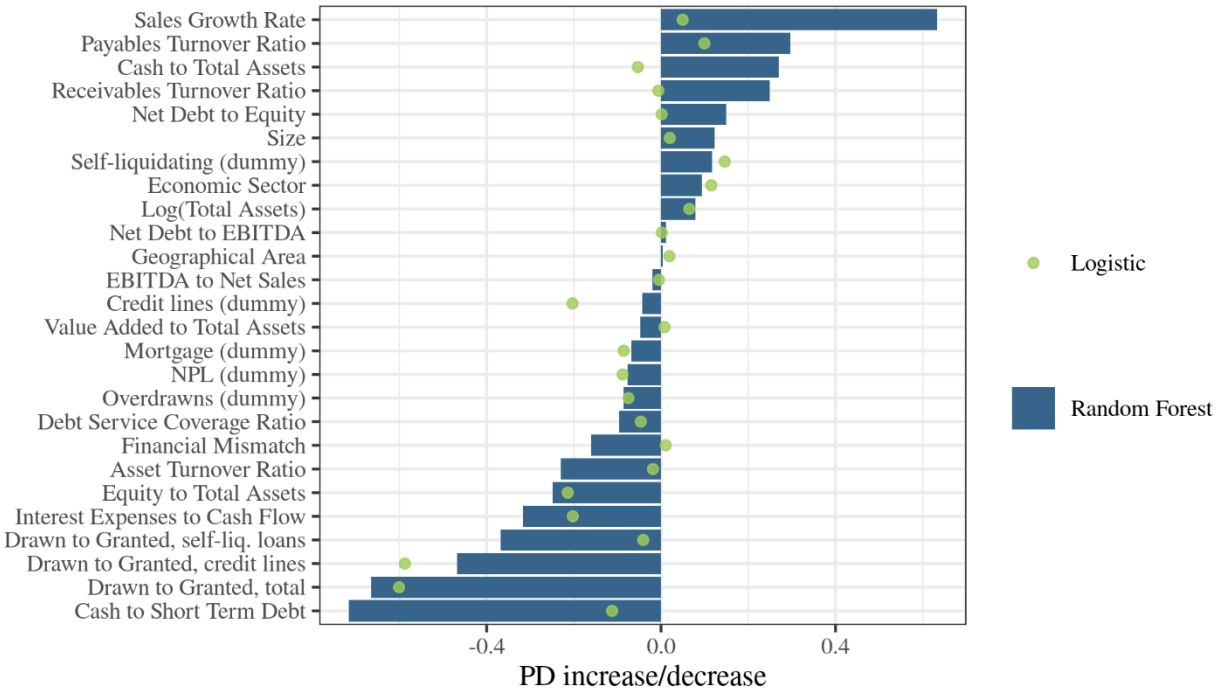
Table 1. Feature values and Shapley values for selected borrowers

	Feature values					Random Forest Shapley values (%)			Logistic Shapley values (%)		
	Mean PD: 1.9%	Median PD: 0.8%	Obs. with RDF $\widehat{PD}_i =$ 0.4%	1.5%	5%	Obs. with RDF $\widehat{PD}_i =$ 0.4%	1.5%	5%	Obs. with RDF $\widehat{PD}_i =$ 0.4%	1.5%	5%
Log(Total Assets) [(1)]	7.1	7.0	8.2	6.1	4.8	0.08	-0.14	-0.05	0.06	-0.15	-0.59
EBITDA to Net Sales	0.15	0.09	0.13	0.08	0.06	-0.02	-0.14	-0.25	0.00	0.04	0.15
Receivables Turnover Ratio	187.0	114.7	394.8	95.3	32.4	0.25	-0.19	-0.21	-0.01	0.02	0.04
Payables Turnover Ratio	77.8	54.0	263.3	0.0	0.0	0.30	0.00	-0.09	0.10	-0.15	-0.24
Debt Service Coverage Ratio	22.4	5.8	97.3	5.4	3.0	-0.10	-0.10	0.05	-0.05	0.01	0.03
Sales Growth Rate	0.08	0.03	-0.48	-0.49	0.75	0.63	0.88	0.33	0.05	0.14	-0.28
Asset Turnover Ratio	1.10	1.02	0.48	1.30	1.75	-0.23	0.05	0.26	-0.02	0.04	0.15
Value Added to Total Assets	0.30	0.24	0.21	0.37	0.11	-0.05	0.06	0.17	0.01	-0.10	0.21
Net Debt to Equity	12.6	20.5	20.5	20.5	20.5	0.15	0.07	0.07	0.00	0.00	0.01
Financial Mismatch	-0.18	-0.17	-0.43	-0.17	-0.14	-0.16	-0.09	0.00	0.01	0.00	0.00
Equity to Total Assets	0.29	0.24	0.57	0.18	0.16	-0.25	-0.08	-0.01	-0.21	0.08	0.21
Net Debt to EBITDA	6.82	0.24	0.24	0.24	0.24	0.01	-0.02	0.01	0.00	0.00	0.01
Interest Expenses to Cash Flow	0.34	0.13	0.01	0.26	1.00	-0.32	0.01	0.73	-0.20	-0.26	0.84
Cash to Short Term Debt	0.32	0.11	0.39	0.14	0.04	-0.72	-0.41	0.33	-0.11	0.04	0.36
Cash to Total Assets	0.10	0.05	0.14	0.09	0.04	0.27	0.08	-0.08	-0.05	-0.04	0.11
Drawn to Granted, total	0.66	0.75	0.44	1.00	0.98	-0.66	0.73	0.69	-0.60	0.90	1.30
Drawn to Granted, credit lines	0.26	0.01	0.00	0.00	0.98	-0.47	-0.38	1.71	-0.59	-1.84	3.25
Drawn to Granted, self-liq. loans	0.18	0.00	0.00	0.00	0.00	-0.37	-0.29	-0.22	-0.04	-0.11	-0.15
Geographical Area			3	2	2	0.00	-0.11	-0.21	0.02	-0.14	-0.21
Economic Sector			C	E	E	0.09	0.05	0.05	0.12	0.04	0.09
Size			2	1	1	0.12	-0.05	-0.15	0.02	0.01	0.03
Mortgage (dummy)			1	1	0	-0.07	-0.33	0.34	-0.09	-0.25	0.23
NPL (dummy)			0	0	0	-0.08	-0.10	-0.13	-0.09	-0.14	-0.25
Overdrawns (dummy)			0	0	0	-0.09	-0.10	-0.06	-0.07	-0.09	-0.17
Credit lines (dummy)			1	0	1	-0.04	0.19	0.00	-0.20	2.28	-1.43
Self-liquidating (dummy)			1	0	0	0.12	-0.18	-0.06	0.15	-0.37	-0.59
[(1)] Total Assets (th. euros):	1,192	1,068	3,758	454	124						

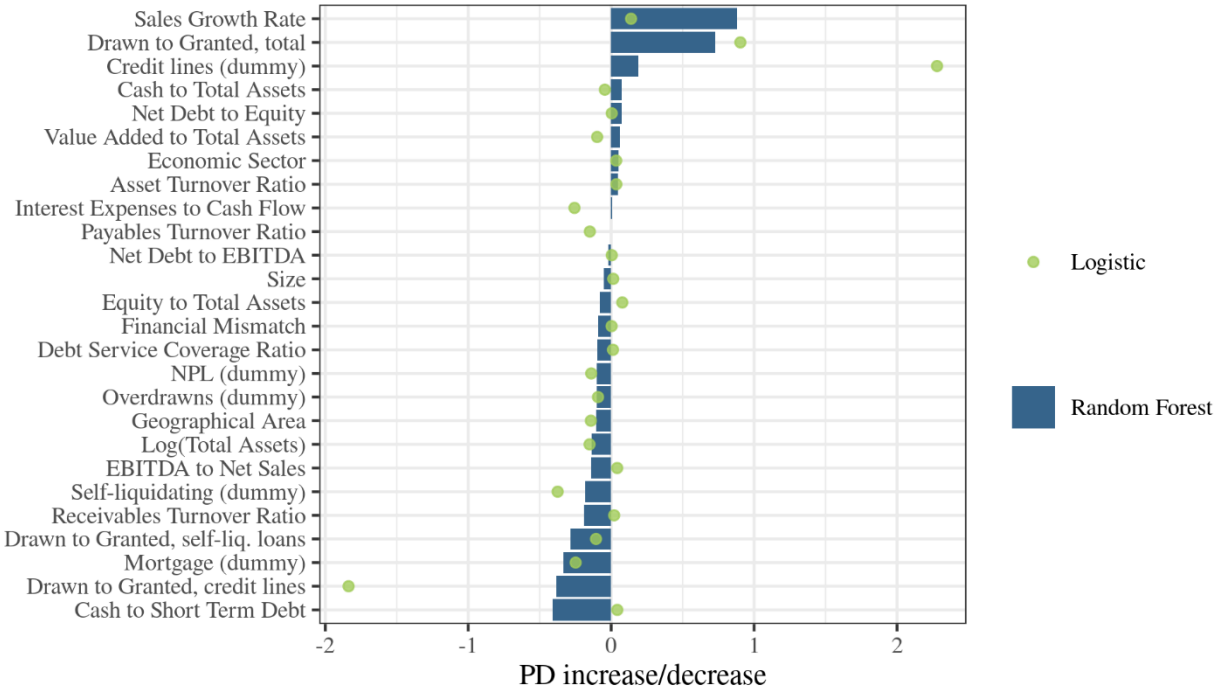
Source: our elaboration on Cerved and the CR database.

Figure 3. Shapley values of the selected observations

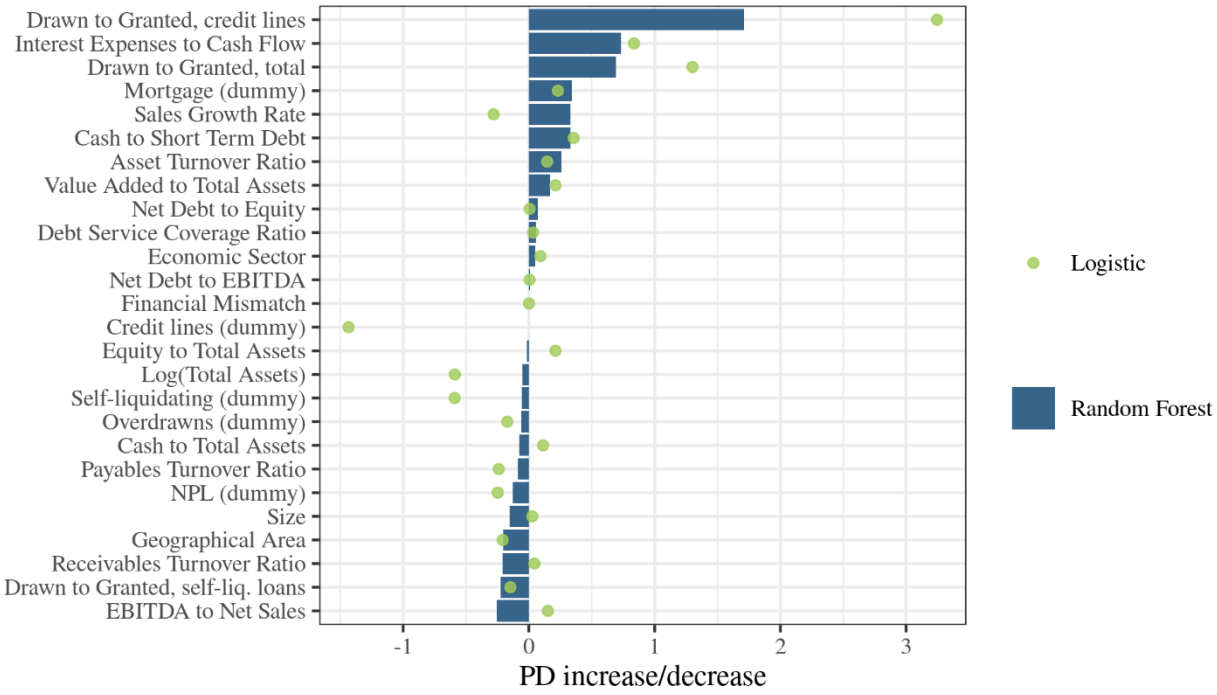
Observation with $\widehat{PD}_i = 0.4\%$



Observation with $\widehat{PD}_i = 1.5\%$



Observation with $\widehat{PD}_i = 5\%$

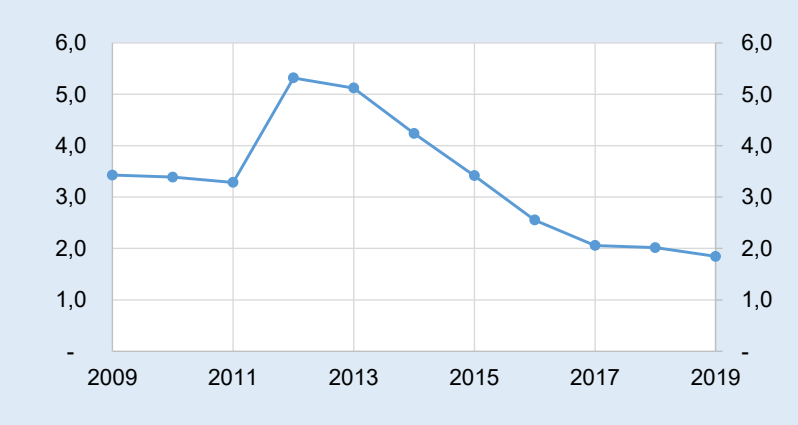


4.4 Assessing the stability of the RDF model

While several works have shown that the RDF model provides more accurate default forecasts, , less is known about the reasons for such better outcomes or regarding the stability of the models' over time. In Moscatelli et al. (2019) it is argued that the RDF adapts rapidly to changes in the structure of the relationships between predicting variables and future default. Such type of changes are likely to occur during economic downturns, when historical parameters estimated by standard credit risk models may be 'broken' by economic shocks. The lack of explainability, however, prevents clear identification and monitoring of these changes in the structure of credit risk models. In this section we estimate RDF default forecasting models over the 2009-2019 period using a one-year rolling window approach, and we apply the global explainability methods to analyze the dynamic of the relative importance of the different variables and of their relationship with the probability of default.

Our dependent variable displays a peak in the aftermath of the European Sovereign debt crisis, where the corporate default rate increased by about 2 percentage points year-on-year (Figure 4); we therefore utilize the years 2012-13 as our reference 'crisis' period.

Figure 4. Corporate default rate



Source: our calculation based on Credit Register data. Note: The corporate default rate is the ratio between the number of new non-performing and the *in-bonis* borrowers at the beginning of each year.

Table 2 reports the time-varying variable importance metric for the RDF model (only the top 10 most important variables are included in the graph). As expected, regardless of the macro-financial conditions, the importance of credit behavioral indicators in predicting corporate defaults firmly overcomes that of balance-sheet indicators; this is the case for both the total and the short-term drawn to granted credit ratios, which show a higher importance with respect to the other indicators in every year of the analysis. Moreover, their importance appears stable other time. Some balance-sheet indicators increase instead their relevance in predicting defaults depending on the financial conditions. For instance, the liquidity ratios, the cash reserves to short-term debt ratio and the liquid assets to total assets ratio have a higher importance during the crisis period when corporate default peaked, with the former having the greatest change in annual importance among all indicators.

Table 2: AUC decrease variable importance over time

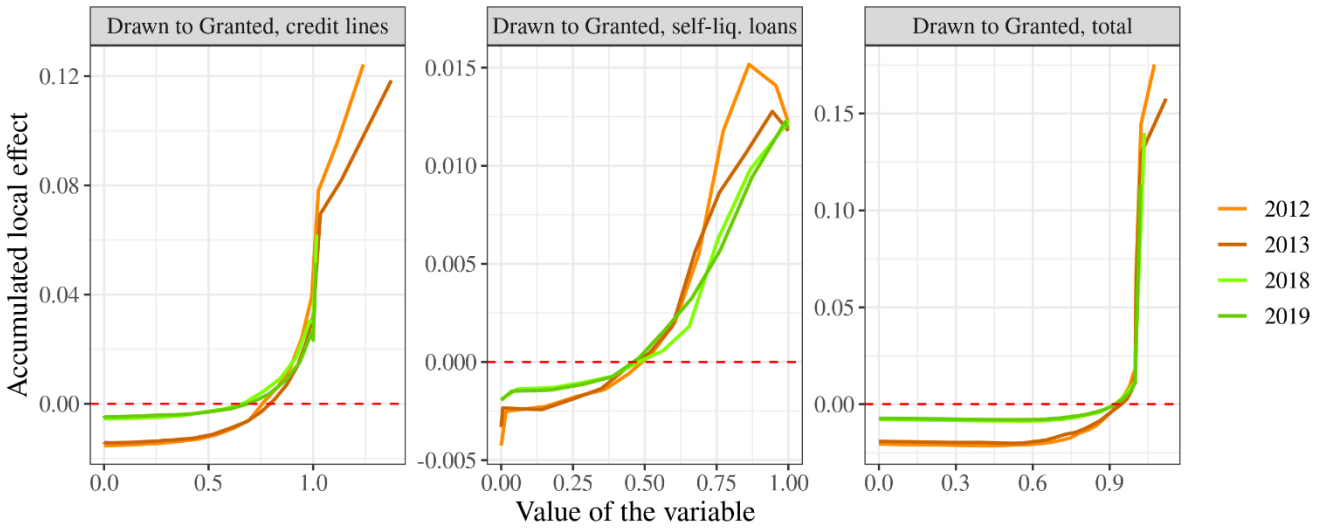
Variable	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	average	coef. var.
Cash to Short Term Debt	0,8%	1,4%	1,7%	0,9%	1,1%	1,2%	1,6%	2,9%	2,3%	3,0%	2,4%	1,8%	0,43
Cash to Total Assets	0,5%	0,7%	0,8%	0,4%	0,5%	0,7%	0,8%	1,3%	1,2%	1,1%	1,0%	0,8%	0,36
Drawn to Granted Credit, credit lines	3,4%	3,5%	3,8%	4,1%	3,4%	3,2%	3,4%	3,0%	3,3%	3,5%	2,9%	3,4%	0,09
Drawn to Granted Credit, self-liquid. loans	0,8%	0,7%	0,8%	0,8%	0,6%	0,6%	0,7%	1,2%	1,2%	1,2%	0,8%	0,8%	0,26
Drawn to Granted Credit, total	2,9%	4,2%	5,0%	4,6%	4,7%	4,7%	4,1%	4,3%	4,7%	4,5%	5,1%	4,4%	0,12
NPL (dummy)	2,5%	2,0%	1,4%	1,1%	1,3%	1,2%	1,0%	0,9%	0,8%	0,6%	0,8%	1,2%	0,43
Overdrawns (dummy)	0,7%	0,7%	0,7%	0,8%	0,6%	0,7%	0,6%	0,6%	0,6%	0,4%	0,7%	0,6%	0,13
Interest Expenses to Cash Flow	0,3%	0,4%	0,5%	0,7%	0,6%	0,7%	0,8%	0,8%	1,0%	0,8%	0,6%	0,7%	0,29
Receivables Turnover Ratio	0,5%	0,6%	0,7%	0,5%	0,6%	0,7%	0,7%	1,1%	1,0%	0,8%	0,6%	0,7%	0,25
Asset Turnover Ratio	0,5%	0,4%	0,7%	0,6%	0,5%	0,5%	0,8%	1,2%	1,2%	1,1%	0,7%	0,7%	0,38

Source: our calculation. Notes:The coefficient of variability is the ratio between the standard deviation and the average of variable importance computed over the 2009-19 period.

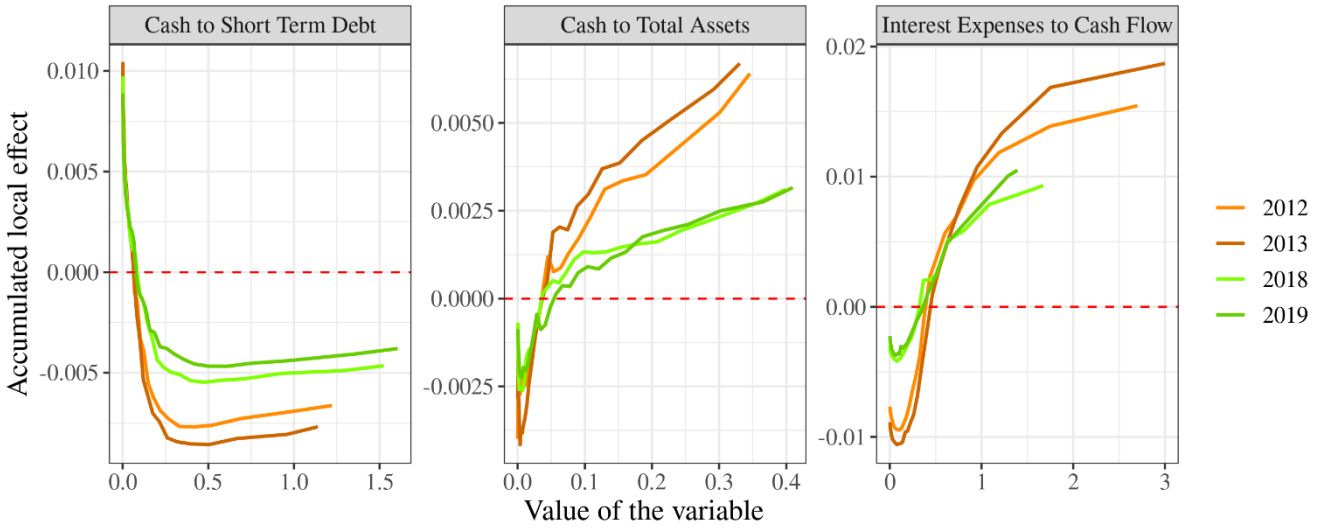
Figure 5 shows the estimated relationships between the most important predictive variables and firms' default for different periods: the 2012-13 years (when the default rate peaked) and the 2018-19 years (when it was at its lowest). Overall, both credit and financial indicators have a similar estimated relationship with firms' default in both crises and recovery periods (as can be seen from the similar shapes of the curves), suggesting a relatively stable structure of the model. However, for financial indicators the relationships with firms' default seem more sensitive to changes in the financial condition; for instance, the effect of liquidity ratios during the crisis period appear more pronouced than during the economic recovery.

Figure 5: ALE plots across years

Panel A – Credit register variables



Panel B – Cerved variables



5 Conclusions

The use of machine learning models in the financial sector has been increasing rapidly in the recent years (IMF, 2021). Their higher forecasting accuracy compared to traditional approaches may prove useful in key lending processes (such as loan allocation and regulatory capital calculation), allowing to reduce capital absorption and credit losses. At the same time, some advantages can be foreseen for borrowers; for instance, more accurate risk-sensitive pricing of credit may be turned into interest rate savings for some borrowers, or the exploitation of larger

datasets (unstructured data or less standard credit risk indicators) may allow the credit risk assessment of a broader spectrum of borrowers.

The adoption of models based on ML, however, is often limited by the complexity of their functional form, as their "black box" labelling makes clear. The explainability of forecasting models is a crucial property which is needed to assess the validity of forecasts in new samples, to comply with legal frameworks (in particular, ensuring the fairness of conditions amongst borrowers) and to use the predictions into a wider decision process that may involve expert judgement. In credit risk applications, the explainability of forecasts is particularly relevant both for the financial intermediaries involved in the screening of prospective borrowers, and for the regulators, which require fairness and transparency in decisions based on automated algorithms.

This paper reviews some of the most established methods from the eXplainable Artificial Intelligence literature and applies them to the RDF and the logit models developed in Moscatelli et al. (2019) to predict corporate defaults of Italian NFCs.

Results point to differences in the structure of these models. While the RDF model relies on a wider set of predictors to make forecasts, including lower frequency financial indicators, the logit model attributes greater importance to the distress signals coming from a limited number of credit behavioral indicators. This is not a subtle difference, and it may entail several benefits for borrowers. First, the RDF model allows for a more comprehensive credit quality assessment based on a larger information set. Second, the higher importance attributed to financial ratios helps to stabilize rating grades which could otherwise be highly sensitive to credit indicators such information on firms' credit drawings, which could fluctuate seasonally in some business sectors.In the context of the Covid-19 crisis, the set of policy measures adopted in many jurisdiction to avoid that firms' liquidity distress could turn into solvency issues (namely the guarantee schemes on loans and the moratorium on loans to SMEs) might have reduced the usual information content related to credit behavioral indicators. From this perspective, the capacity of models based on ML to rely more heavily on a broader information set might provide more robust forecasts. Third, some borrowers (especially the smaller ones whose credit relatioships are not reported in the Credit Register or those without a credit history) may still be able to obtain an assessment of their credit risk and access, therefore, credit funding.

Despite the convenience of using and interpreting linear relationships, our analysis indicates that the empirical associations between firms' financial and credit indicators and their defaults are rather complex, i.e. non-linear and non-monotonic. This complexity also explains why the logit model assigns higher importance to a few key indicators that show a more standard behavior, without exploiting the entire information set. For some values of a limited set of variables, we show possible tresholds or cliff effects whose knowledge might be helpful in the assessment of prospective borrowers when their 'values' fall outside or close to a specific threshold.

Finally, we find that the importance of different predicting variables varies between crisis and recovery years.

References

AI HLEG - High-Level Expert Group on Artificial Intelligence. (2019). Policy and investment recommendations for trustworthy AI. Report published by the European Commission.

Alonso, A., and Carbó, J. M. (2020). Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost. Banco de Espana, Documentos de Trabajo, N.º 2032.

Alonso, A. and Carbó, J. M. (2021). Understanding the performance of machine learning models to predict credit default: a novel approach doe supervisory evaluation. Banco de Espana, Documentos de Trabajo, N.º 2105.

Altman, A., Toloşi, L., Sander, O., Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

Apley, D.W., Zhu, J., (2019). Visualizing the effects of predictor variables in black box supervised learning models.

Ariza-Garzón, M.J., Arroyo, J., Caparrini, A., Segovia-Vargas, M., (2020). Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending, *IEEE Access*.

Bacham, D., & Zhao, J. (2017). Machine learning: challenges, lessons, and opportunities in credit risk modeling. Moody's Analytics Risk Perspectives, 9, 30-35.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J. (2020). Explainable AI in Fintech Risk Management. *Front. Artif. Intell.* 3:26. doi: 10.3389/frai.2020.00026

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., Wang, T. (2018). An Interpretable Model with Globally Consistent Explanations for Credit Risk, arXiv:1811.12615v1

Datta, A., Sen, S., Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (pp. 598-617). IEEE.

Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dumitrescu, E. I., Hué, S., & Hurlin, C. (2021). Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds.

Dupont, L., Fliche, O., Yang, S. (2020). Governance of Artificial Intelligence in Finance, ACPR - Banque de France Discussion Document.

EBA - European Banking Authority (2020). EBA eport on big data and advanced analytics. EBA report.

Ertel,W. (2017). Introduction to Artificial Intelligence, 2nd ed,. Springer.

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. Methodology and computing in applied probability, 11(1), 29-45.

Fisher, A., Rudin, C., Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the" rashomon" perspective. arXiv preprint arXiv:1801.01489, 68.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A. (2020). Predictably Unequal? The Effects of Machine Learning on Credit Markets, working paper, https://ssrn.com/abstract=3072038

Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

IMF (2021). Powering the Digital Economy Opportunities and Risks of Artificial Intelligence in Finance. International Monetary Fund, Departmental papers, DP/2021/024.

Joseph, A. (2019). Shapley regressions: a framework for statistical inference on machine learning models, Bank of England, Staff Working Paper No.784.

Miller, T. (2018). Explanation in Artificial Intelligence: Insights from the Social Sciences, arXiv:1706.07269v3.

Mitchell, T. (1997). Machine Learning, McGraw Hill.

Molnar, C., (2019). Interpretable machine learning, https://christophm.github.io/interpretable-ml-book/.

Moscatelli, M., Narizzano, S., Parlapiano, F., Viggiano, G. (2019). Corporate default forecasting with machine learning (No. 1256). Bank of Italy, Economic Research and International Relations Area.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

ROFIEG - Expert Group on Regulatory Obstacles to Financial Innovation (2019). 30 recommendations on regulation, innovation and finance. Final report to the European Commission

Shapley, L. S. (1953). A value for n-person games. Contributions to the Theory of Games, 2(28), 307-317.

Štrumbelj, E., and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems, 41(3), 647-665.

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2020). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. Journal of the Operational Research Society, 1-11.

Appendix 1

We describe the financial and credit behavioral indicators used to predict the default of non-financial firms.

Table A1. Description of indicators used by the credit scoring model to predict the default of non-financial firms.

Variable	Description
Geographical Area	Categorical variable identifying the geographical region where the firm operates (North-East, North-West, Center, South and Islands).
Economic sector	Categorical variable identifying the economic sector of the firm, according to ATECO classification.
Cash to Short Term Debt	Liquidity ratio that measures a firm's ability to pay off short-term debt obligations with cash and cash equivalents.
Cash to Total Assets	Ratio between cash and liquid assets to total assets. It measures a firm's liquidity and how easily it can service debt and short-term liabilities if the need arises.
Drawn to Granted Credit, credit lines	Drawn amount to granted amount of uncommitted short term loans. Financial flexibility ratio: it measures the percentage of available uncommitted short term loans that the firm is actually using.
Drawn to Granted Credit, self-liquid. loans	Drawn amount to granted amount on self-liquidating loans. Financial flexibility ratio: it measures the percentage of available self-liquidating short term loans that the firm is actually using.
Drawn to Granted Credit, total	Drawn amount to granted amount of credit. Financial flexibility ratio: it measures the percentage of available credit that the firm is actually using. It refers to all the different types of loans.
Debt Service Coverage Ratio	Ratio of debt sustainability. It is defined as the amount of cash flow available over interest expenses and annual principal payments on financial debt.
Credit lines (dummy)	Dummy equal to 1 if the firm has uncommitted short term loans, and 0 otherwise.
NPL (dummy)	Dummy equal to 1 if the firm has deteriorated loans, and 0 otherwise[19].
Overdrawns (dummy)	Dummy equal to 1 if the firm has a drawn amount greater than the granted amount, and 0 otherwise.
Self-liquidating (dummy)	Dummy equal to 1 if the firm has self-liquidating, and 0 otherwise.
EBITDA to Net Sales	Operating profitability ratio, that measures how much earnings the company is generating before interest, taxes, depreciation, and amortization, as a percentage of revenue.

[19] Although the dataset contains only firms not already in default, some firms may still have non-performing loans if they are below the materiality threshold required in the default definition.

Equity to Total Assets	Ratio between equity and total assets. Used to assess a company's financial leverage
Financial mismatch	Ratio of the mismatch (difference) between short-term liabilities and short-term assets and total assets. A negative value of the ratio (short term liabilities > short term assets) indicates that the firm has enough short-term assets to meet its short-term liabilities.
Interest Expenses to Cash Flow	Ratio that indicates the firm's ability to pay interest from its generated cash flow.
Log(TotalAssets)	Natural Logarithm of Total Assets. Measures the size of the firm.
Mortgage (dummy)	Dummy variable equal to 1 if long term loans are more than 90 per cent of total loans.
Net Debt to EBITDA	Debt sustainability ratio, gives an indication as to how long a company would need to operate at its current level to pay off all its financial debt.
Net Debt to Equity	Measure of a firm's financial leverage, calculated by dividing its net liabilities by stockholders' equity.
Payable Turnover Ratio	Commercial Debt/(NetRevenues - Operational Value Added)Ratio that measures the efficiency with which a company collects its receivables or the credit it extends to customers
Receivables Turnover Ratio	AccountsReceivable/NetRevenues. Efficiency and liquidity ratio that relates the firm debt towards its suppliers to its revenues net of variable costs.
Sales Growth Rate	Yearly growth rate of net sales. Measures a company's growth in a specific year, as well as the stability of a firm's performance.
Size	Categorical variable identifying the size of the firms, as defined by the European Commission (micro, small, medium, large).
Asset Turnover Ratio	Net Sales/ Total Assets. Efficiency ratio that measures a firm's ability to generate sales from its assets.
Value Added to Total Assets	Ratio between economic value added and total assets. It is a ratio that measures the firm's ability to generate value from its assets.

Table A2. Training dataset descriptive statistics (numerical variables)

dataset	variable	n.	mean	median	std	q10	q25	q75	q90
CERVED	Log(Total Assets)	327.410	7,1	7,0	1,3	5,4	6,1	8,0	9,0
CERVED	EBITDA to Net Sales	327.410	15%	8%	21%	-1%	4%	18%	45%
CERVED	Receivables Turnover Ratio	327.410	202,1	117,7	305,9	0,0	26,8	215,3	427,1
CERVED	Payables Turnover Ratio	327.410	82,5	57,5	99,3	0,0	0,0	125,4	209,0
CERVED	Debt Service Coverage Ratio	327.410	21,6	5,6	31,9	0,3	1,5	23,3	93,0
CERVED	Sales Growth Rate	327.410	8%	3%	29%	-21%	-6%	17%	44%
CERVED	Asset Turnover Ratio	327.410	1,1	1,0	0,8	0,1	0,5	1,5	2,2
CERVED	Value Added to Total Assets	327.410	29%	23%	24%	3%	10%	42%	66%
CERVED	Net Debt to Equity	327.410	14,0	23,0	10,4	0,4	1,6	23,0	23,0
CERVED	Financial Mismatch	327.410	-17%	-16%	27%	-55%	-36%	0%	17%
CERVED	Equity to Total Assets	327.410	28%	23%	21%	5%	10%	42%	62%
CERVED	Net Debt to EBITDA	327.410	7,3	0,2	13,2	0,2	0,2	5,6	39,3
CERVED	Interest Expenses to Cashflow	327.410	0,4	0,1	0,5	0,0	0,0	0,5	1,4
CERVED	Cash to Short Term Debt	327.410	31%	11%	44%	0%	2%	39%	97%
CERVED	Cash to Total Assets	327.410	10%	5%	12%	0%	1%	15%	30%
CR	Drawn to Granted, total	327.410	65%	75%	34%	2%	42%	97%	100%
CR	Drawn to Granted, credit lines	327.410	27%	1%	37%	0%	0%	56%	94%
CR	Drawn to Granted, self-liq. loans	327.410	19%	0%	32%	0%	0%	34%	76%
CR	Mortgage (dummy)	327.410	41%						
CR	NPL (dummy)	327.410	1%						
CR	Overdrawns (dummy)	327.410	1%						
CR	Credit lines (dummy)	327.410	79%						
CR	Self-liquidating (dummy)	327.410	50%						

Notes: Own calculation based on Cerved and the national Credit Register data. The training dataset refers to default year 2019.

Table A3. Training dataset descriptive statistics (categorical variables)

variable	value	n.	perc.
Geographical Area	North-East	108.215	33,1%
	North-West	82.640	25,2%
	Center	71.591	21,9%
	South and Islands	64.774	19,8%
	n/a	190	0,1%
Economic Sector	A	5.072	1,5%
	B	4.083	1,2%
	C	40.621	12,4%
	D	75.789	23,1%
	E	163.238	49,9%
	F	38.607	11,8%
Size	Micro	225.410	68,8%
	Small	77.943	23,8%
	Medium	19.244	5,9%
	Large	4.813	1,5%

Notes: Own calculation based on Cerved and the national Credit Register data. The training dataset refers to default year 2019

Table A4. Test dataset descriptive statistics (numerical variables)

dataset	variable	n.	mean	median	std	q10	q25	q75	q90
CERVED	Log(Total Assets)	330.093	7,1	7,0	1,3	5,4	6,1	8,0	9,0
CERVED	EBITDA to Net Sales	330.093	15%	9%	21%	-1%	4%	18%	45%
CERVED	Receivables Turnover Ratio	330.093	187,0	114,7	266,3	0,0	26,1	208,0	405,2
CERVED	Payables Turnover Ratio	330.093	77,8	54,0	93,4	0,0	0,0	118,4	199,2
CERVED	Debt Service Coverage Ratio	330.093	22,4	5,8	33,2	0,4	1,5	24,0	97,3
CERVED	Sales Growth Rate	330.093	8%	3%	28%	-20%	-6%	17%	44%
CERVED	Asset Turnover Ratio	330.093	1,1	1,0	0,8	0,1	0,5	1,6	2,2
CERVED	Value Added to Total Assets	330.093	30%	24%	25%	3%	10%	43%	68%
CERVED	Net Debt to Equity	330.093	12,6	20,5	9,2	0,4	1,5	20,5	20,5
CERVED	Financial Mismatch	330.093	-18%	-17%	27%	-56%	-37%	0%	17%
CERVED	Equity to Total Assets	330.093	29%	24%	21%	5%	11%	43%	63%
CERVED	Net Debt to EBITDA	330.093	6,8	0,2	12,2	0,2	0,2	5,5	36,2
CERVED	Interest Expenses to Cashflow	330.093	0,3	0,1	0,4	0,0	0,0	0,4	1,3
CERVED	Cash to Short Term Debt	330.093	32%	11%	47%	0%	2%	40%	102%
CERVED	Cash to Total Assets	330.093	10%	5%	12%	0%	1%	15%	30%
CR	Drawn to Granted, total	330.093	66%	75%	34%	2%	43%	98%	100%
CR	Drawn to Granted, credit lines	330.093	26%	1%	36%	0%	0%	53%	93%
CR	Drawn to Granted, self-liq. loans	330.093	18%	0%	31%	0%	0%	32%	75%
CR	Mortgage (dummy)	330.093	43%						
CR	NPL (dummy)	330.093	1%						
CR	Overdrawns (dummy)	330.093	1%						
CR	Credit lines (dummy)	330.093	78%						
CR	Self-liquidating (dummy)	330.093	49%						

Notes: Own calculation based on Cerved and the national Credit Register data. The test dataset refers to default year 2020

Table A5. Test dataset descriptive statistics (categorical variables)

variable	value	n.	perc.
Geographical Area	North-East	108.821	33,0%
	North-West	83.644	25,3%
	Center	71.657	21,7%
	South and Islands	65.719	19,9%
	n/a	252	0,1%
Economic Sector	A	5.216	1,6%
	B	4.126	1,2%
	C	40.708	12,3%
	D	76.745	23,2%
	E	165.162	50,0%
	F	38.136	11,6%
Size	Micro	224.766	68,1%
	Small	80.415	24,4%
	Medium	19.875	6,0%
	Large	5.037	1,5%

Notes: Own calculation based on Cerved and the national Credit Register data. The test dataset refers to default year 2020

Appendix 2

We describe the step-by-step procedures for constructing the explainability methods introduced in section 2. Let n be the number of observations in the dataset, X the variable of interest, Z the set of all the other variables, and $f(\,)$ the predictive function of the model that, given input observations (X, Z), returns the predictions $\hat{Y} = f(X, Z)$.

Permutation variable importance

Let (X^p, Z) be the permuted version of (X, Z) with respect to X, meaning that the values of X have been randomly shuffled across all observations. Then:

1. Predictions $\hat{Y} = f(X, Z)$ are obtained from the forecasting model applied to the dataset (X, Z).
2. Predictions $\widehat{Y^p} = f(X^p, Z)$ are obtained from the forecasting model applied to the dataset (X^p, Z).
3. Variable importance for X is defined as a dissimilarity function $d(\hat{Y}, \widehat{Y^p})$ between $\hat{Y}$ and $\widehat{Y^p}$; in our case, we use as dissimilarity functions:
   - the difference in AUC: $d\left(\hat{Y}, \widehat{Y^p}\right) = AUC\left(\hat{Y}\right) - AUC\left(\widehat{Y^p}\right)$
   - the average absolute difference in individual predictions: $d\left(\hat{Y}, \widehat{Y_p}\right) = \frac{1}{n}\sum_{k=1}^{n} |\widehat{Y_k} - \widehat{Y_k^p}|$

Accumulated Local Effects plot

In order to obtain an expected prediction function that only depends on the value of the variable of interest X, the Accumulated Local Effects (ALE) plot (Apley and Zhu, 2019) method first estimates the partial derivative function with respect to X by taking discrete differences in the average conditional prediction, and then obtains the average expected prediction by cumulative summation of these partial differences. The algorithm can be described as follows:

1. Define a number of intervals $l > 0$ over which the support of X will be split. This number determines the trade-off between the robustness and the granularity of the estimate[20].
2. The percentiles $\{x^{(0)}, x^{(1)}, \dots, x^{(l)}\}$ of X are computed; they define the l intervals $I(i) = [x^{(i-1)}, x^{(i)}]$, each containing $\frac{n}{l}$ observations.
3. For each interval $I(i)$, the average local effect is computed as:

$$LE(\,I(i)\,) = \frac{1}{|I(i)|} \cdot \sum_{(X_k, Z_k):\, X_k \in I(i)} \left[\, f\left(x^{(i-1)}, Z_k\right) - f\left(x^{(i)}, Z_k\right) \right]$$

4. The ALE function of a value $x \in X$ is computed as the accumulated local effects of all intervals up to the one containing x:

$$ALE(x) = \sum_{I(i)\ s.t. x < x^{(i)}} LE(\,I(i)\,)$$

[20] We choose for our application $l = 30$, which allows us to have robust estimates with a good granularity (the default of the R IML package is $l = 20$)

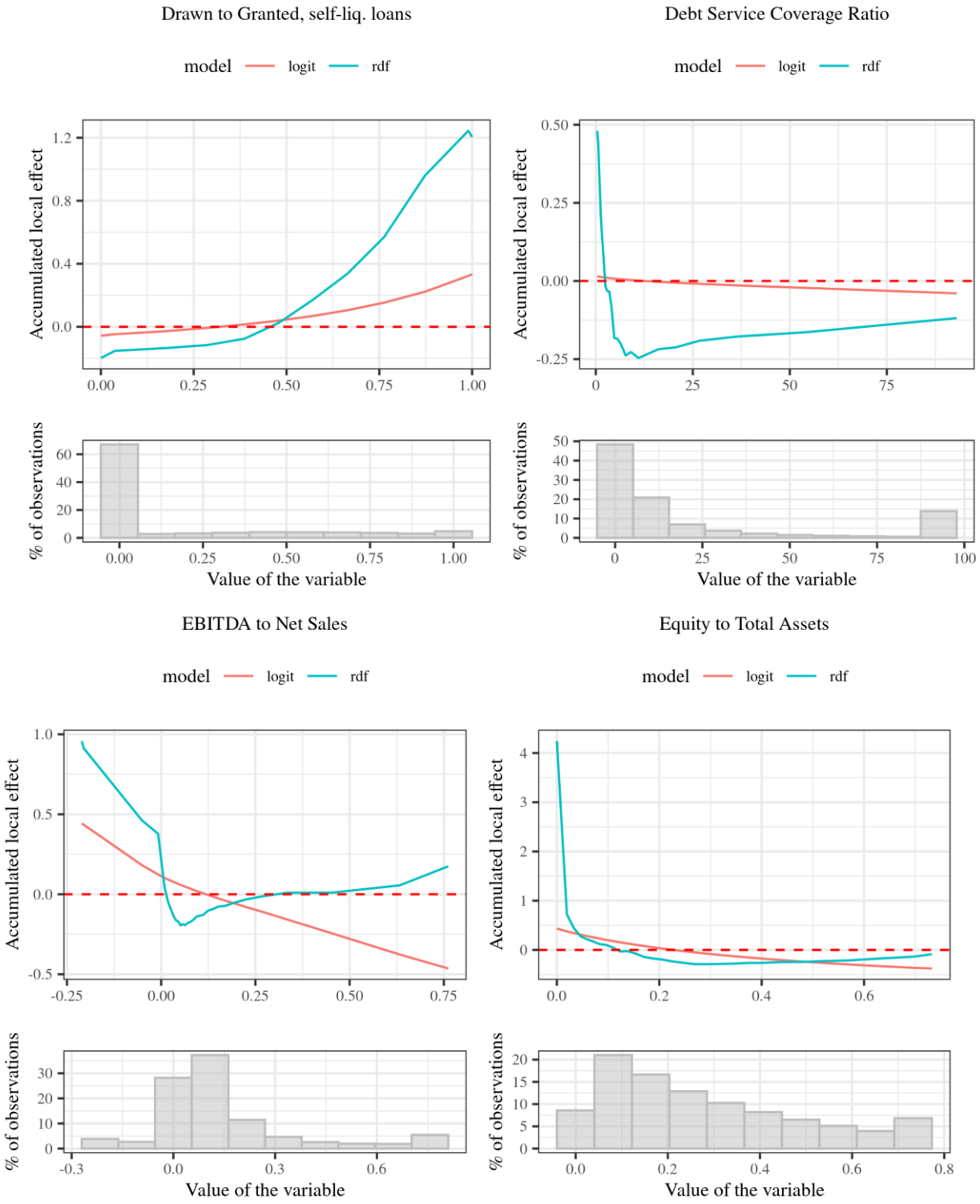
The mean of the ALE function is then subtracted, so that the average effect across all data is 0. The ALE plot is obtained as the plot of the ALE function.

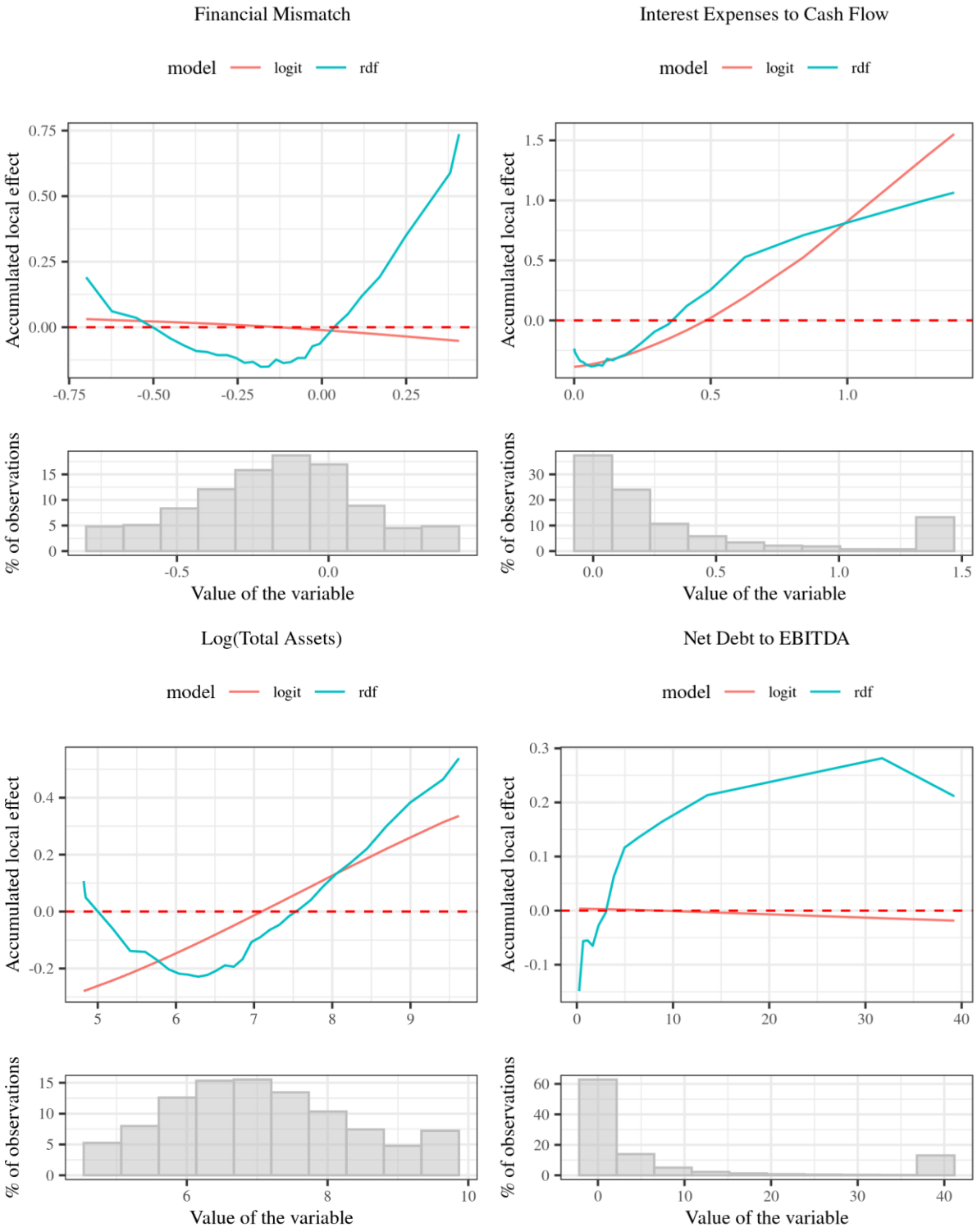
Shapley Values

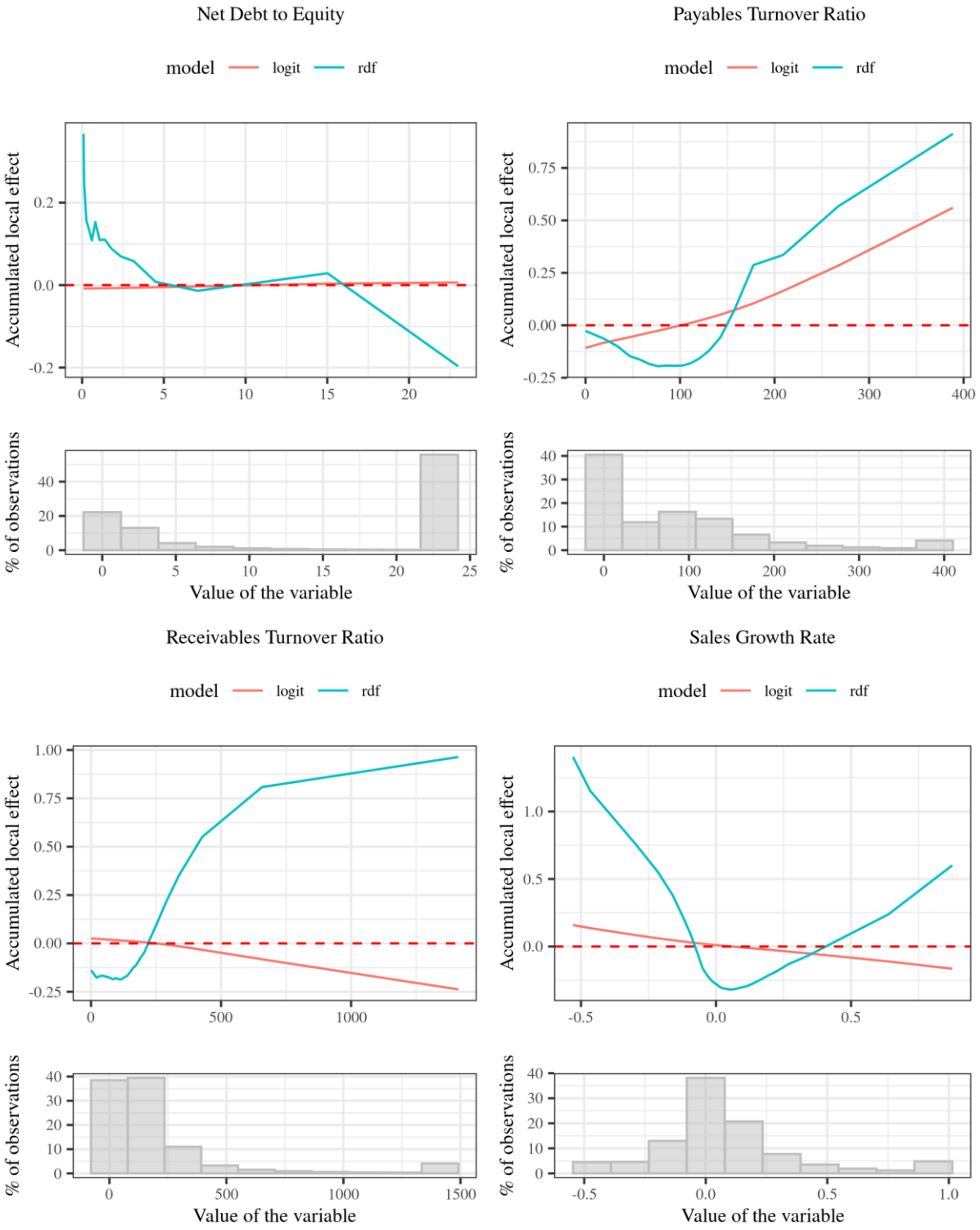
Let (X_i, Z_i) be the observation we want to explain. The Shapley value of X with respect to observation (X_i, Z_i) is obtained by repeating a large number of times the following procedure and averaging the marginal contributions obtained:

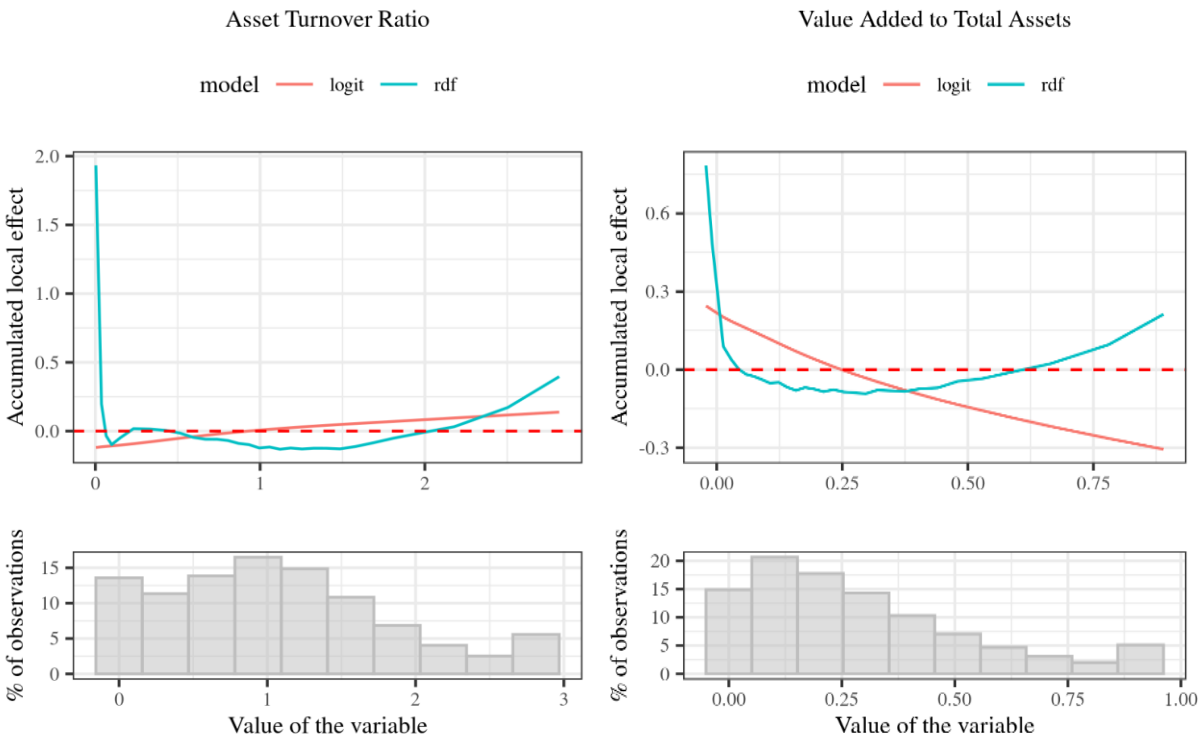
1. An observation (X_r, Z_r) is randomly sampled across all the observations of the dataset.
2. A subset of features $Z^{(1)} \subseteq Z$ is randomly selected; let $Z^{(2)} = Z - Z^{(1)}$ be the set of the remaining features.
3. The forecasting model is applied to the synthetic observation $(X_i, Z_i^{(1)}, Z_r^{(2)})$, obtaining the prediction $\widehat{Y_{+i}} = f(X_i, Z_i^{(1)}, Z_r^{(2)})$.
4. The forecasting model is applied to the synthetic observation $(X_r, Z_i^{(1)}, Z_r^{(2)})$, obtaining the prediction $\widehat{Y_{+r}} = f(X_r, Z_i^{(1)}, Z_r^{(2)})$. Note that the difference with the previous synthetic observation is that the value of X is taken from observation r instead of observation i.
5. The marginal contribution is computed as $\widehat{Y_{+i}} - \widehat{Y_{+r}}$.

Appendix 3 – ALE plots




Financial Mismatch
model logit rdf
Accumulated local effect
0.75
0.50
0.25
0.00
-0.75 -0.50 -0.25 0.00 0.25
% of observations
15
10
5
0
-0.5 0.0
Value of the variable
Interest Expenses to Cash Flow
model logit rdf
Accumulated local effect
1.5
1.0
0.5
0.0
0.0 0.5 1.0
% of observations
30
20
10
0
0.0 0.5 1.0 1.5
Value of the variable
Log(Total Assets)
model logit rdf
Accumulated local effect
0.4
0.2
0.0
-0.2
5 6 7 8 9
% of observations
15
10
5
0
6 8 10
Value of the variable
Net Debt to EBITDA
model logit rdf
Accumulated local effect
0.3
0.2
0.1
0.0
-0.1
0 10 20 30 40
% of observations
60
40
20
0
0 10 20 30 40
Value of the variable


Net Debt to Equity
model logit rdf
Accumulated local effect
0.2
0.0
-0.2
0 5 10 15 20
% of observations
40
20
0
0 5 10 15 20 25
Value of the variable
Payables Turnover Ratio
model logit rdf
Accumulated local effect
0.75
0.50
0.25
0.00
-0.25
0 100 200 300 400
% of observations
40
30
20
10
0
0 100 200 300 400
Value of the variable
Receivables Turnover Ratio
model logit rdf
Accumulated local effect
1.00
0.75
0.50
0.25
0.00
-0.25
0 500 1000
% of observations
40
30
20
10
0
0 500 1000 1500
Value of the variable
Sales Growth Rate
model logit rdf
Accumulated local effect
1.0
0.5
0.0
-0.5 0.0 0.5
% of observations
40
30
20
10
0
-0.5 0.0 0.5 1.0
Value of the variable


Asset Turnover Ratio
model
logit
rdf
Accumulated local effect
2.0
1.5
1.0
0.5
0.0
0
1
2
% of observations
15
10
5
0
0
1
2
3
Value of the variable
Value Added to Total Assets
model
logit
rdf
Accumulated local effect
0.6
0.3
0.0
-0.3
0.00
0.25
0.50
0.75
% of observations
20
15
10
5
0
0.00
0.25
0.50
0.75
1.00
Value of the variable