Explainable Artificial Intelligence: Interpreting credit scoring models based on machine learning

Mirko Moscatelli

Bank of Italy

May 13, 2022

Overview

Motivation

- Machine Learning (ML) studies models that improve automatically their performance on a task learning from data. In several applications, ML models have shown to achieve superior predictive performance in comparison to traditional approaches; thus, it has being increasingly used to make decisions that have a significant impact in people's lives.

- ML models, however, are often *black boxes*, intrinsecally unable to explain their general decision logic and the reason behind their individual predictions. This de facto limits their widespread adoption in several areas, and creates risks in those where it is applied.

- The field of eXplainable Artificial Intelligence (XAI) explores methods and techniques that enhance human understanding of AI and ML systems, enabling a wider and more informed use of ML models.

What do we mean by explainable?

An exact definition of explainable is elusive; a significant subfield of the XAI research focuses on providing the tools to answer a number of questions about a predictive model, such as:

- What are the most "important" variables for the model?
- How are predictors linked with the output (how strong is their effect, is it linear/nonlinear/nonmonotonic, how do they interact, ...)?
- What determined a specific prediction?
- Etc.

Why is explainability desirable?

1. Model diagnostics and improvement of predictive ability:
   - consistency with prior knowledge
   - detect bias in the sample
   - understand the consequences of a distribution shift
   - decrease the probability of data leakage (performance overestimation)
   - increase robustness against adversarial attacks
   - overall, model debugging

2. Reasons that transcend predictive accuracy:
   - learning and causality
   - legal constraints
   - ensure fairness and avoid discrimination
   - integration in a wider decisional process
   - increase humans' acceptance of decisions and trust in results

Explainability and credit scoring

- When actions based on the predictions of a model affect natural or legal persons, they are entitled to understand the reason of the decision. The need for explainability therefore naturally arises in numerous areas where ML is applied or applicable. One of the main examples is credit scoring.

- The explainability of credit decisions has become a consumers' right in itself that regulators and supervisors need to insure (US: Equal credity opportunity act; US: Fair credit reporting act; EU: General data protection regulation; EU: Policy and investment recommendations for trustworthy artificial intelligence; EU: 30 recommendations on regulation, innovation and finance; EBA: Report on big data and advanced analytics; BdF: Governance of artificial intelligence in finance; ...).

The work in one slide

1. We build a large firm-level dataset containing financial, credit behavioral and descriptive indicators, as well as our target variable i.e. the financial default (ratio of non-performing credit to total credit drawn from the banking system for each firm greater than 5 percent).

2. We train, using the previous dataset, a random forest model for forecasting financial default (and a logit model for benchmark).

3. We apply methods from the XAI literature to interpret the trained model, showing how it is possible to get a good insight on which variables are most important for the model, how predictions of the model change when a variable changes, and what are the main reasons behind selected individual predictions.

Overview

Data

- Target variable: financial distress of a borrower firm. A firm is classified as being in distress at the end of the year if the ratio of non-performing credit to total credit drawn from the banking system by the firm has been greater than 5 per cent for at least one month.

- The covariates are 24 economic and financial indicator covering profitability, financing structure, debt sustainability, asset type and credit behavior of the firm, as well as descriptive characteristics such as economic sector, geographical area and size of the firm. They are drawn from the balance sheets of the firms provided by Cerved (BS data) and from the Italian Credit Register (CR data).

- The train dataset has BS data referred to 2016, CR data referred to October 2017, and target variable referred to December 2018. The test dataset is shifted a year later: BS data ref. to 2017, CR data ref. to October 2018, and target variable ref. to December 2019.

Predictive model

- We train a random forest model to predict the future distress of non-financial firms. The random forest model, introduced in Breiman (2001), is nowadays one of the most popular and best performing models in the machine learning field.

- A grid search is used to find the optimal hyper-parameters of the model, the number of variables selected at each split and the minimum number of observations in a leaf, maximizing the AUC with the use of five-fold cross validation.

- A drawback of the random forest model is its low interpretability: since the model is an aggregation of hundreds or thousands of different trees, it is very difficult to understand how the predictions are determined.

Overview

Explainability classification

1. Pre-modelling description of data.
2. Explainable modelling.
3. Post-hoc model-specific methods.
4. Post-hoc model-agnostic methods (*we focus on this*).
   - Post-hoc: applied after the training of the model.
   - Model-agnostic: can be applied to any model - logistic regression, random forest, svm, neural network - irrespective of its functional form.

Post-hoc model-agnostic explainability

- Applicable to any pre-developed model: it just assumes that there is a predictive function f, that we can interrogate, that given an input x gives an output y in return.
- Allows for a greater flexibility in the choice of the specification of the predictive model and an easier comparison between different models (compared to model-specific explainability).
- Typically works by perturbing the data (there are several ways to do it, depending on which is the need) and seeing what effect it has on the predictions.

Several model-agnostic techniques have been (and are being) developed, which allow to explain both the overall decision logic of the model and the reason behind individual predictions.

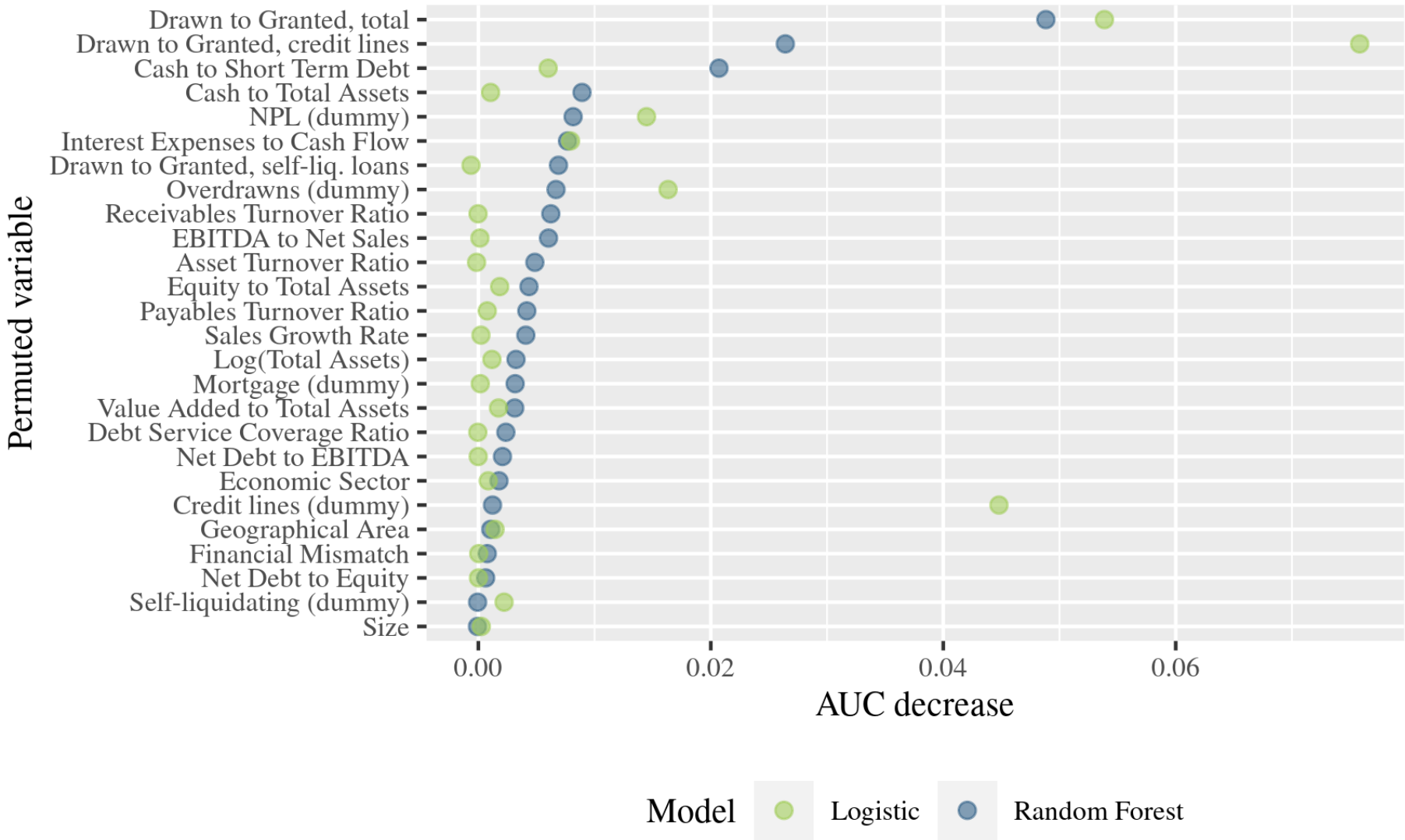
Overview

Permutation variable importance

- *Aim* - Understand which variables are most important for the model. The definition of 'important' can vary depending on which aspect of the model you want to understand.

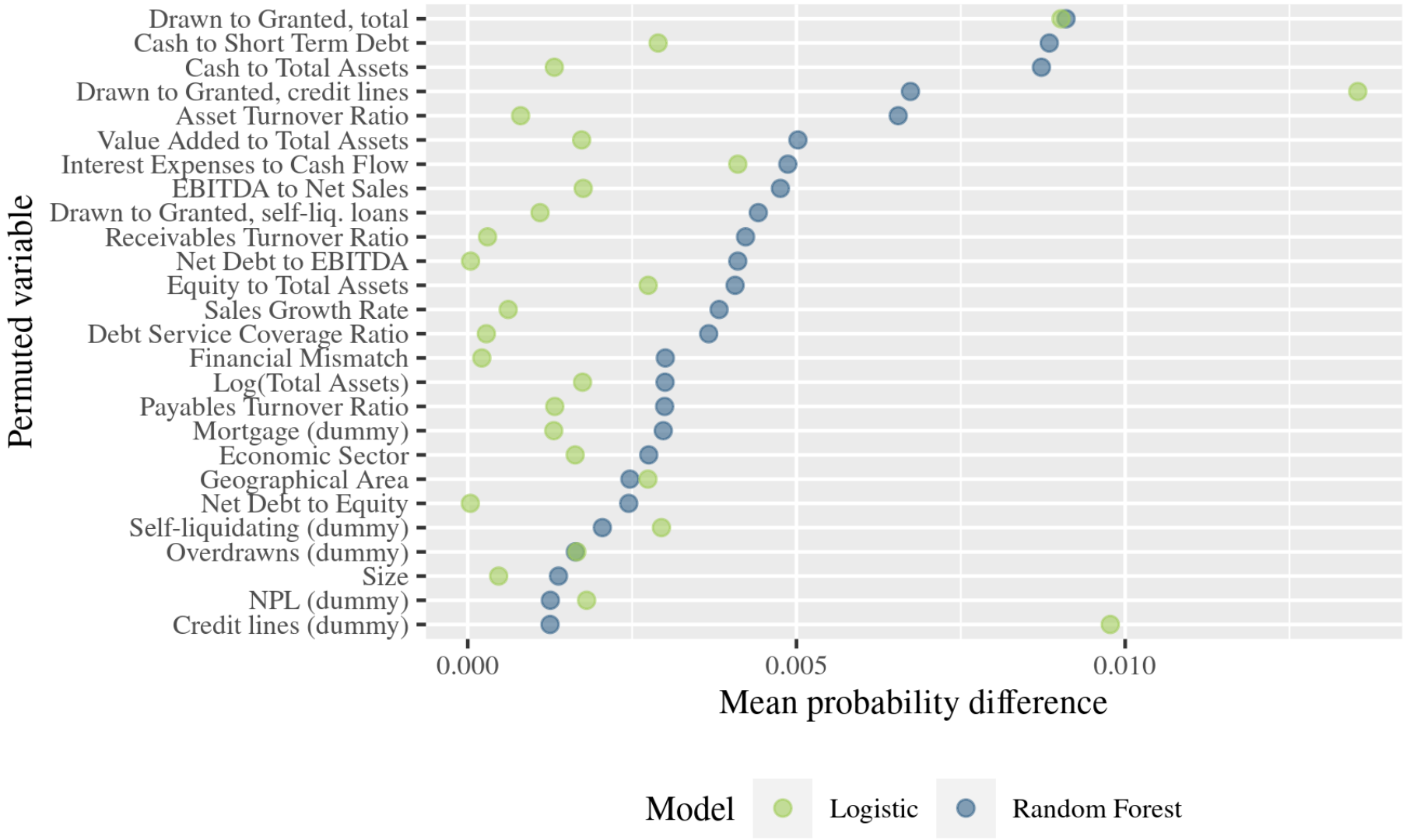
- *How it works* - The variable of interest is randomly permuted across the test dataset (breaking thus its relationship with the other variables and the outcome), and a new set of predictions is computed. Variable importance is defined as a chosen measure of dissimilarity between the original predictions and the new predictions obtained on the permuted test dataset (e.g. the difference between the two AUCs).

Permutation variable importance



Measure of interest: AUC difference

Permutation variable importance



Measure of interest: average difference in the estimated probabilities

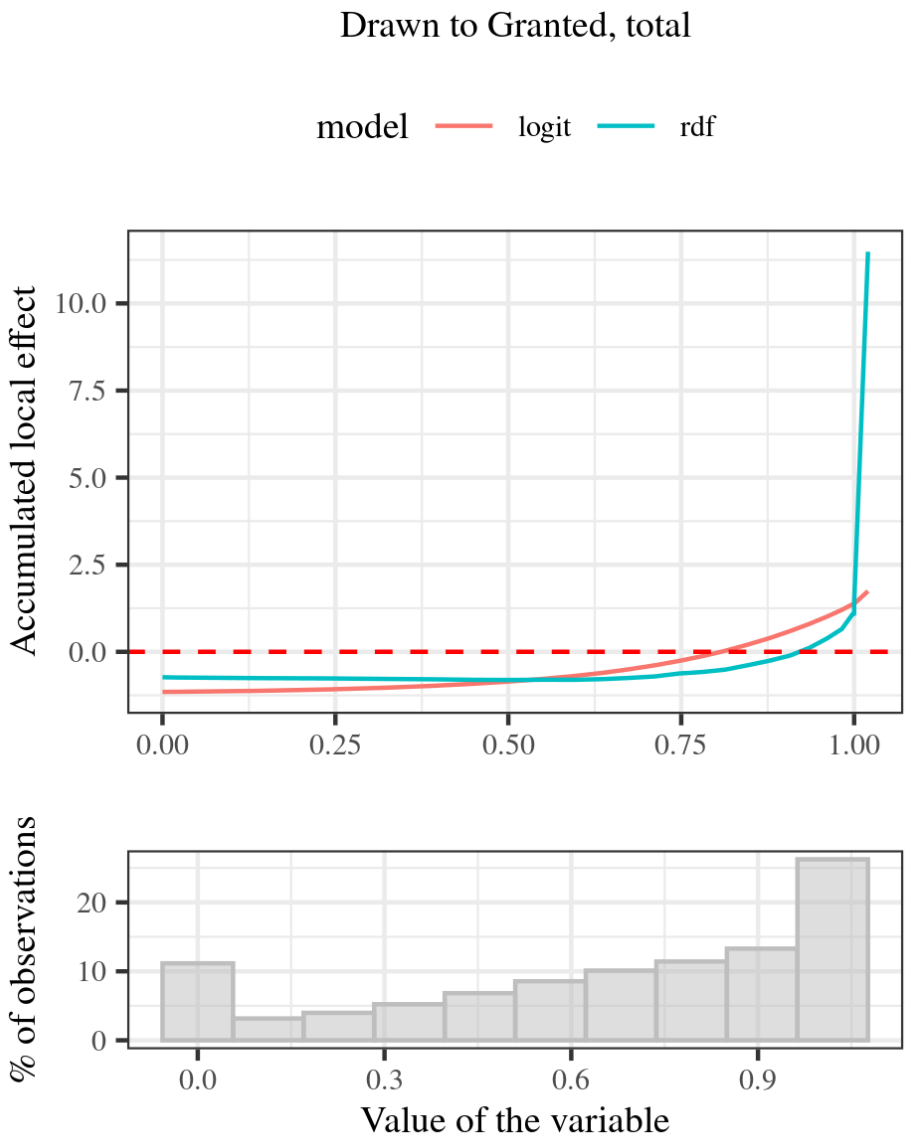
Dependency plots

- *Aim* - Understand the relationship between a variable and the predictions of the model.

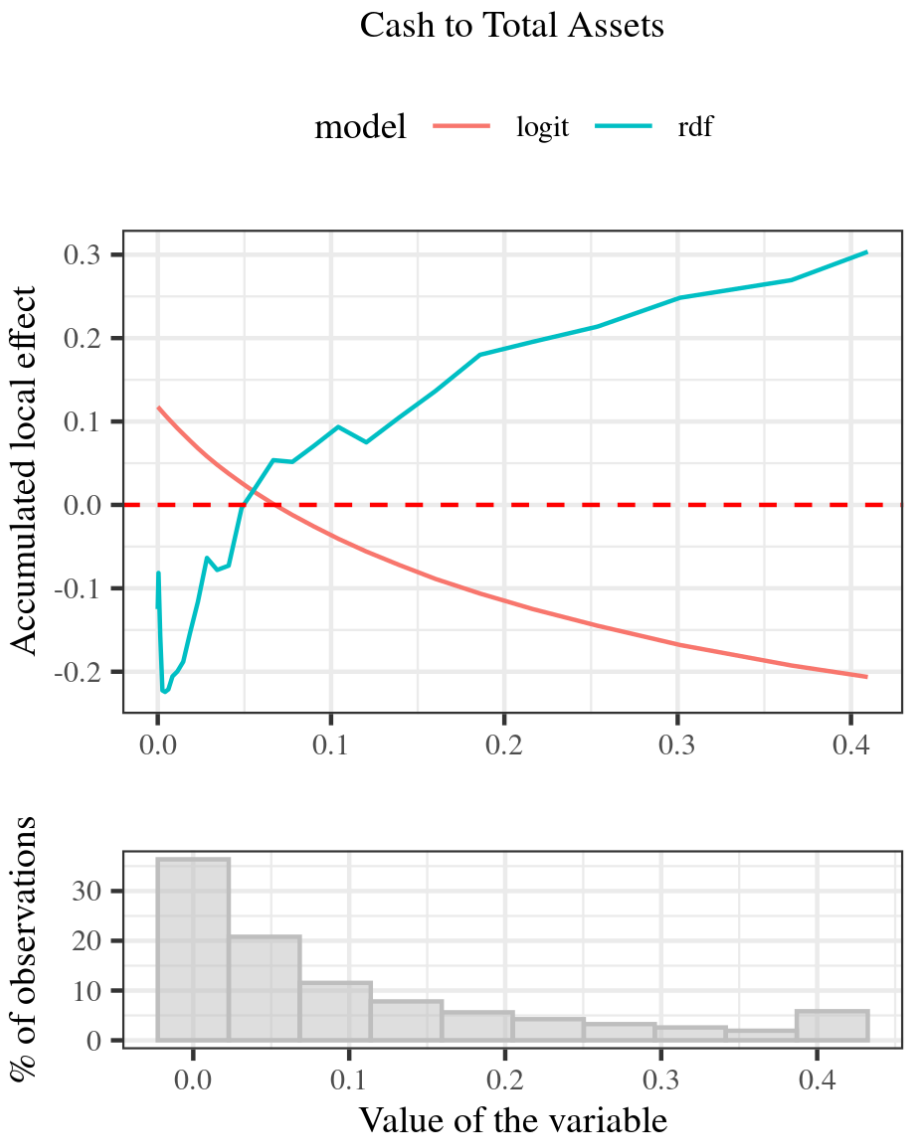
- *How it works* - The general idea is to shift the variable of interest while keeping the others fixed (*ceteris paribus*), and see how the prediction varies; different ways of doing so generate different types of plot (with different underlying assumptions):
  - Partial Dependence Plot (PD plot)
  - Individual Conditional Expectation plot (ICE plot)
  - Accumulated Local Effects plot (ALE plot)

We focus on the ALE plot, a dependency plot developed in the recent years that is able to overcome the main drawback of the most common PD plot, which is being potentially biased if the predictors are not independent.

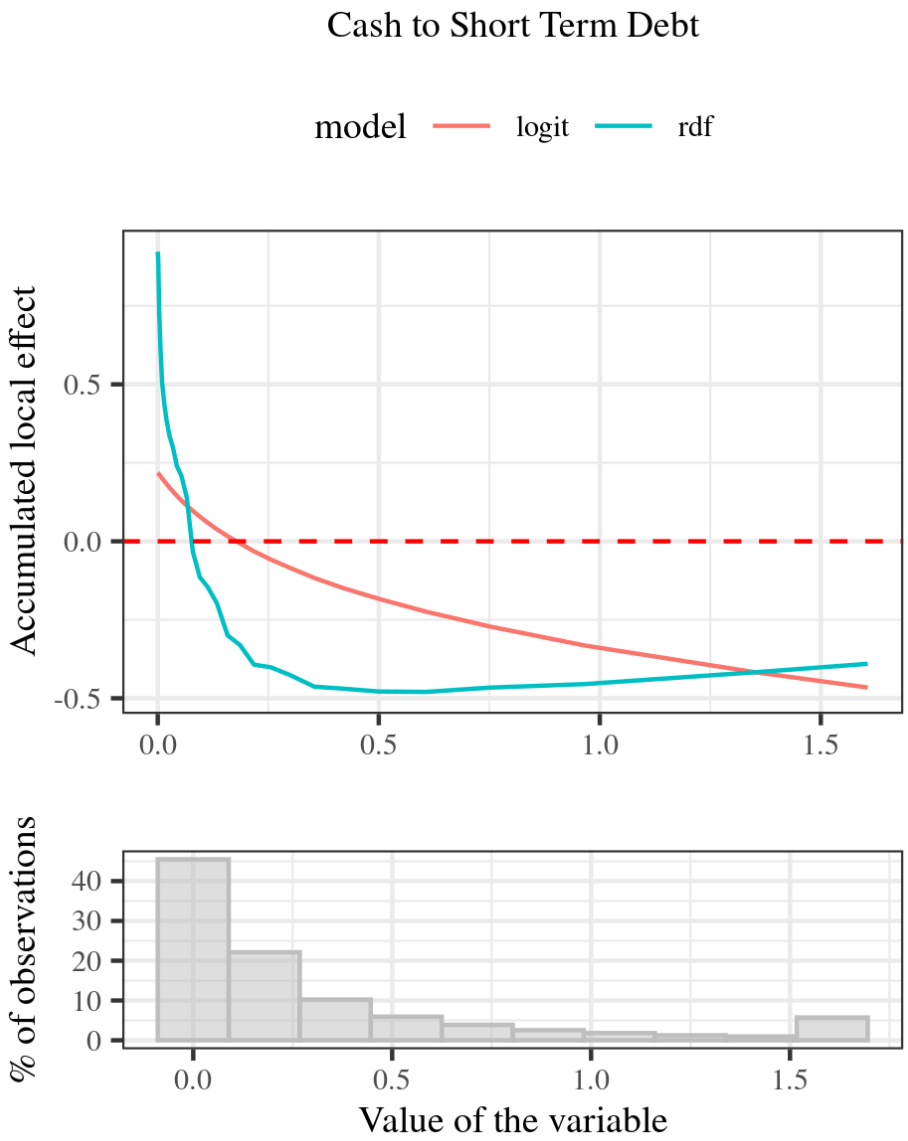
ALE plot


Drawn to Granted, total
model
logit
rdf
Accumulated local effect
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
% of observations
20
10
0
0.0
0.3
0.6
0.9
Value of the variable

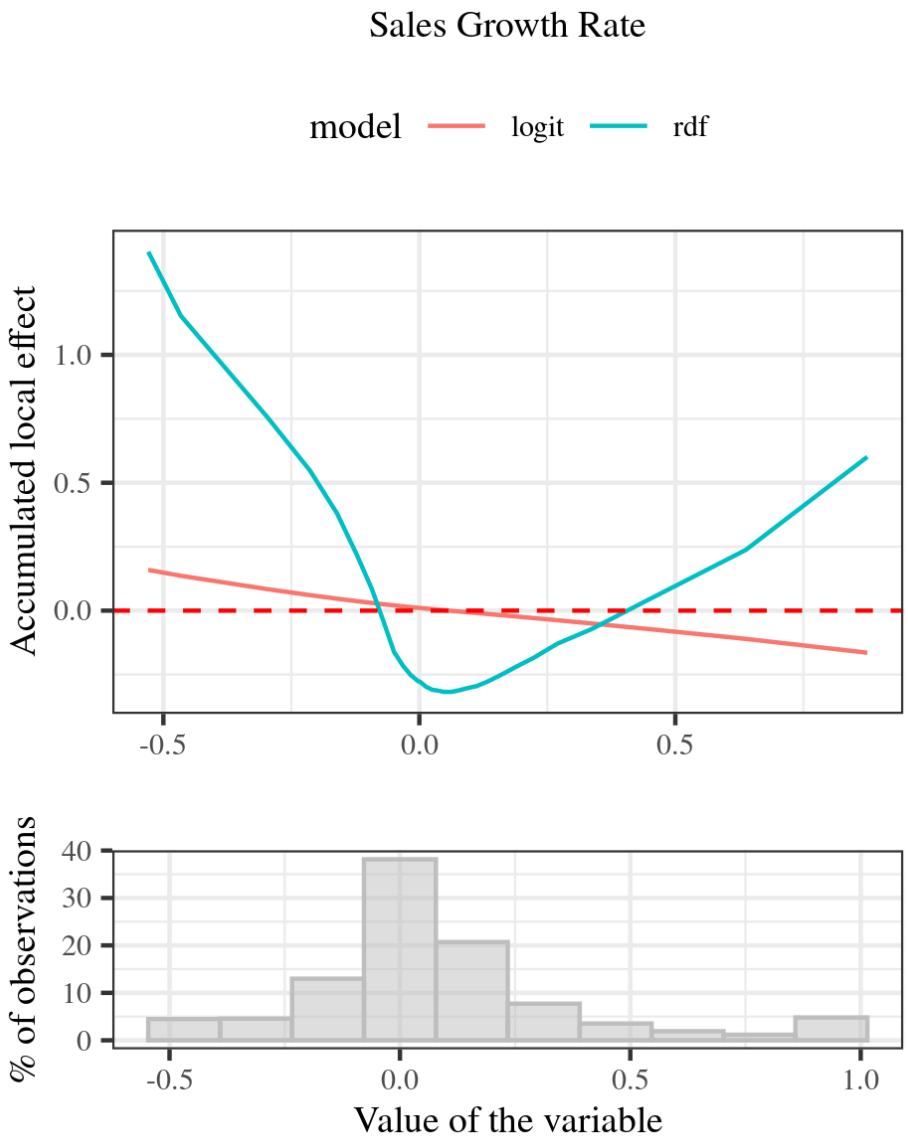
ALE plot


Cash to Total Assets
model
logit
rdf
Accumulated local effect
0.3
0.2
0.1
0.0
-0.1
-0.2
0.0
0.1
0.2
0.3
0.4
% of observations
30
20
10
0
0.0
0.1
0.2
0.3
0.4
Value of the variable

ALE plot


Cash to Short Term Debt
model
logit
rdf
Accumulated local effect
0.5
0.0
-0.5
0.0
0.5
1.0
1.5
% of observations
40
30
20
10
0
0.0
0.5
1.0
1.5
Value of the variable

ALE plot


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

Shapley values

- *Game theory concept:* How to fairly distribute a payout among a group of players in such a way that each player receives according to his contribution toward obtaining it. The Shapley Values are the only payout division satisfies some 'reasonable' axioms (efficiency, symmetry, dummy player, additivity).

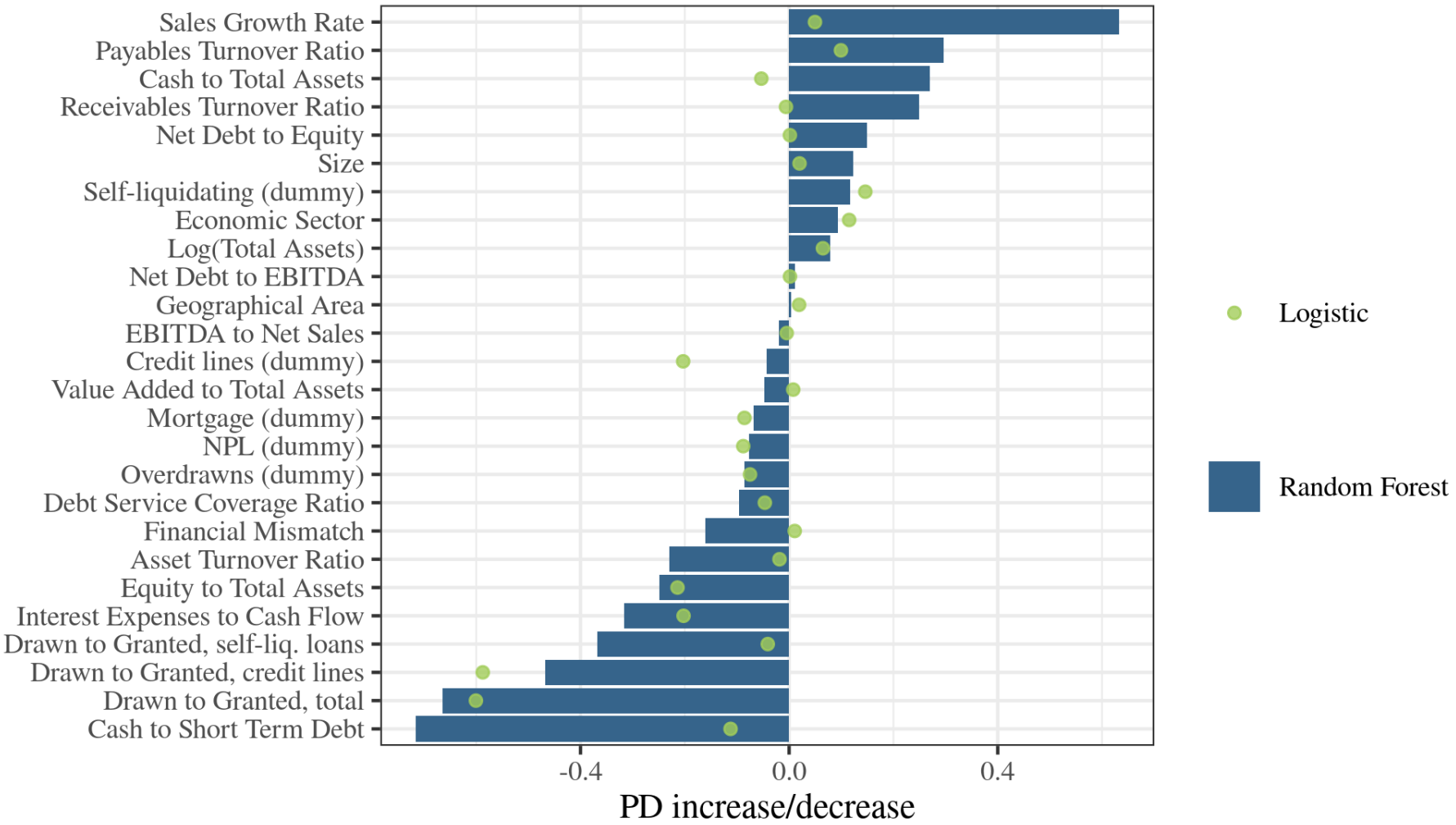
- *Application in XAI:* For a given observation, distribute the difference between the average prediction of the model and the specific prediction among the predictors, according to their contribution.

Shapley values

- *Aim*: Understand the contribution of each variable to the prediction obtained for a given observation.

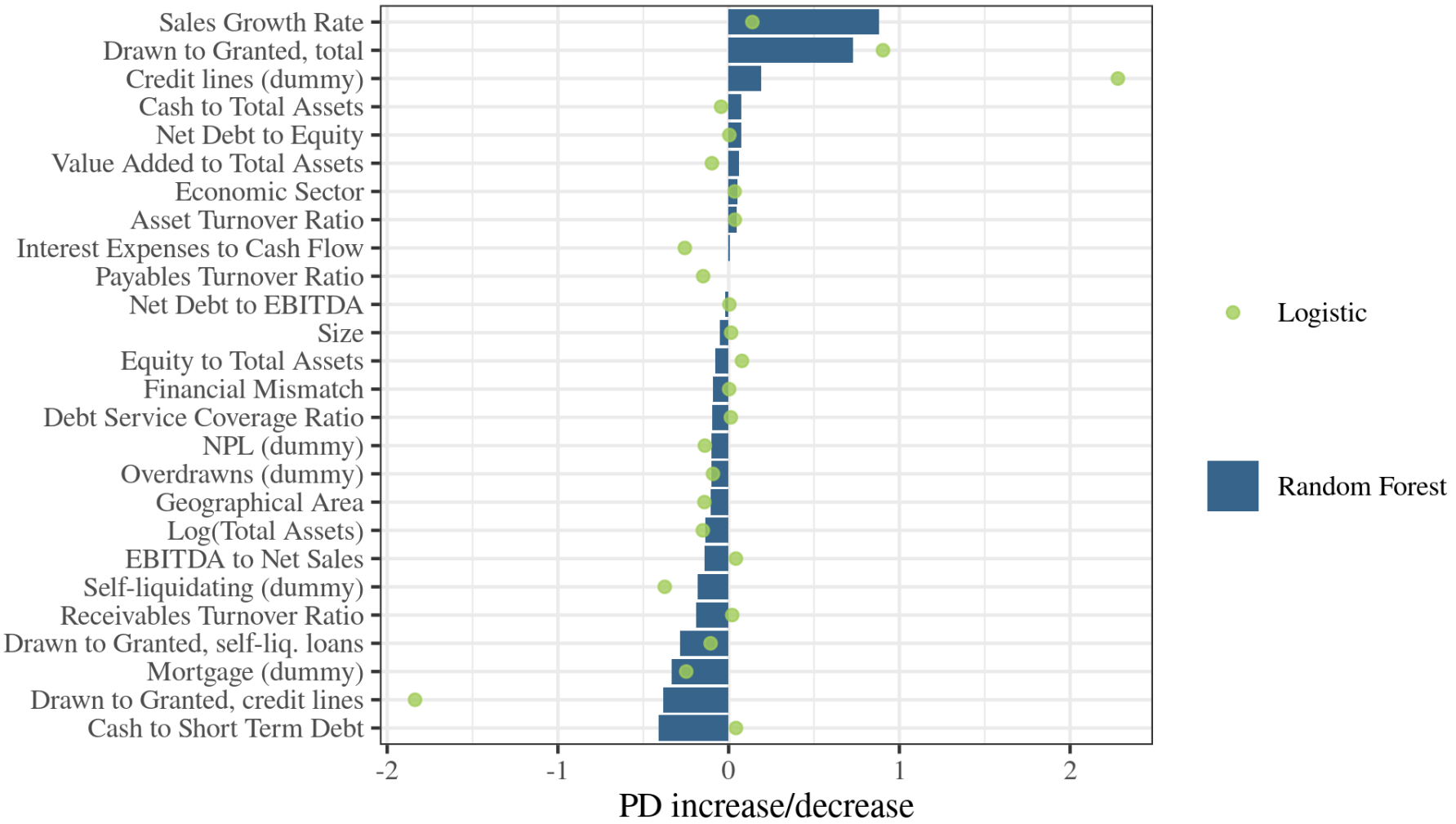
- *How it works*: Given the value x_j of variable j for the observation x of interest, measure the average "marginal difference" that knowing x_j implies on the prediction over all possible subsets S of feature values of $x \setminus x_j$ (i.e. $\mathbb{E}(\hat{Y}|S \cup x_j) - \mathbb{E}(\hat{Y}|S)$).

Shapley values



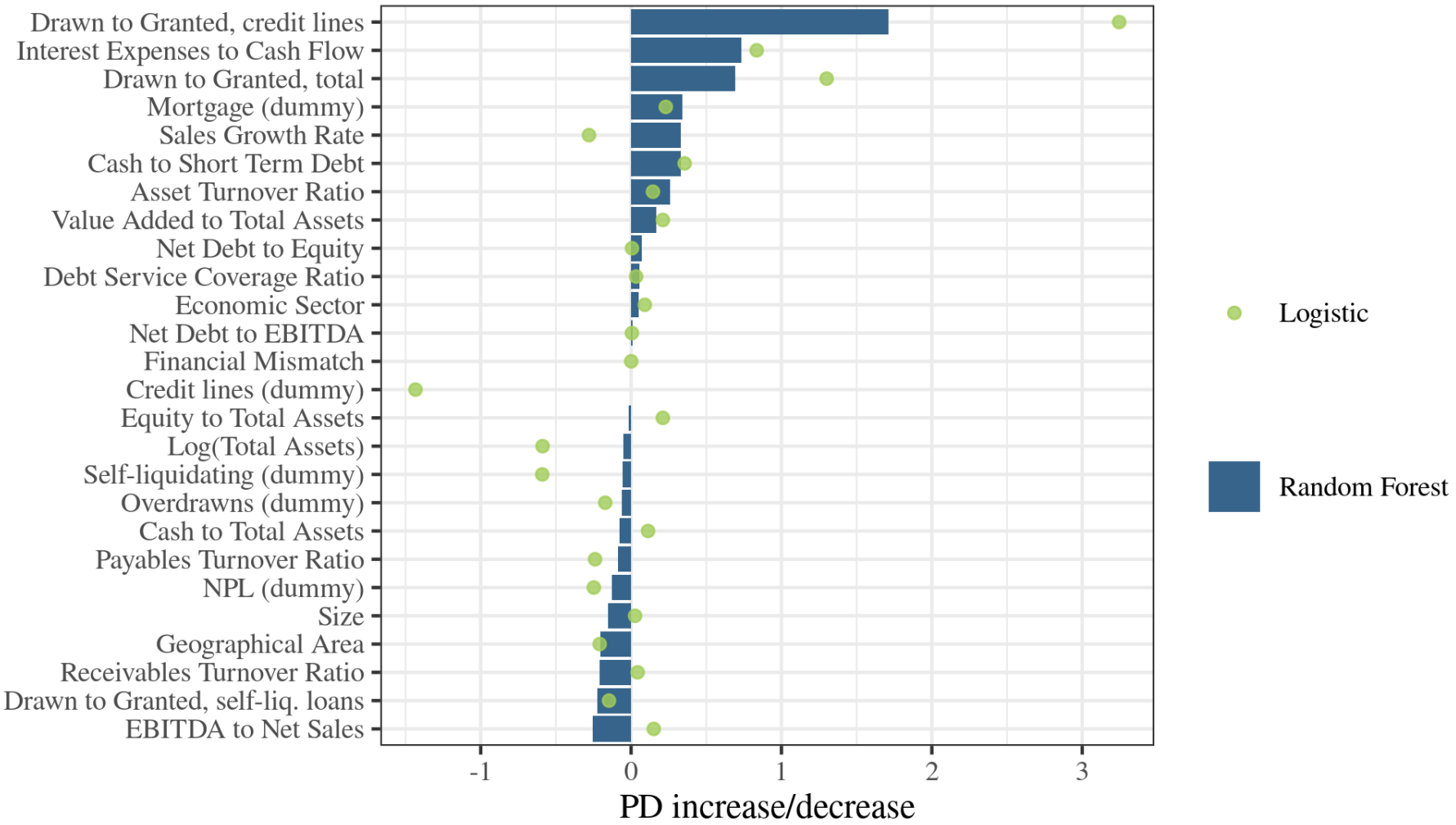
Shapley values for the 0.4% PD observation

Shapley values



Shapley values for the 1.5% PD observation

Shapley values



Shapley values for the 5% PD observation

Other methods

Other methods to explain individual predictions (that we tried and discarded):

- *LIME*: Works by training an interpretable model that mimics the behavior of the model in a neighbourhood of an observations of interest.
- *Counterfactual explanations*: Show observations close to the one of interest for which the model makes a significantly different prediction.

Analysis of importance over time

Are the results stable over time? Do macro-financial conditions matter?

- Credit behavioral indicators (total and the short-term drawn to granted credit ratios) are always more important than balance-sheet indicators.
- The importance of balance-sheet indicators have greater variability over the years, and is higher in non-crisis periods.

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.

Overview

Conclusions/Summary

- Machine Learning models are becoming more and more popular due to their better performance than traditional approaches in various areas. Their widespread adoption, however, is limited by their inability to explain their decision logic.

- Credit scoring is one application where explainability is strongly demanded both from the institutions using it to screen applicants, which need to verify if models have external validity and therefore predictive power for unseen applicants, and from the public and the regulators, which require robustness, fairness and transparency in decisions based on automated algorithms.

Conclusions/Summary

- In this work we describe and apply some of the most popular and established model-agnostic methods from the XAI literature, a rapidly growing branch of AI that studies how to make the logic and the predictions of complex AI models understandable to humans.

- The use of these methods allows us to get a very good insight regarding which variables are most important for the random forest model, how predictions of the model change when a variable changes, and what are the main reasons behind selected individual predictions.

Thank you!

Overview

Framework

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

ALE plot

First, a number I of intervals must be defined, which defines a trade-off between the robustness and the granularity of the estimate. Then:

1. The percentiles $x^{(0)}, x^{(1)}, \ldots, x^{(I)}$ of the variable of interest are computed; they define the I intervals $I(i) = [x^{(i-1)}, x^{(i)})$, each containing $\frac{n}{I}$ observations.
2. For each interval $I(i)$, the average local effect is computed as:

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

3. The ALE function for a value x o X is computed as the accumulated local effects of all intervals up to the one cointaining x:

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

The mean of the ALE function across all data is then subtracted so that the average effect 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 (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$(X_i, Z_i^{(1)}, Z_r^{(1)})$, obtaining the prediction $\hat{Y}_{+i} = f(X_i, Z_i^{(1)}, Z_r^{(1)})$.
4. The forecasting model is applied to the synthetic observation$(X_r, Z_i^{(1)}, Z_r^{(1)})$, obtaining the prediction $\hat{Y}_{+r} = f(X_r, Z_i^{(1)}, Z_r^{(1)})$.
5. The marginal contribution is computed as $\hat{Y}_{+i} - \hat{Y}_{+r}$.

Overview

Additional material - description of the variables

- 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.

Additional material - description of the variables

- 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.
- NPL (dummy): Dummy equal to 1 if the firm has deteriorated loans.

Additional material - description of the variables

- Overdrawns (dummy): Dummy equal to 1 if the firm has a drawn amount greater than the granted amount.
- Self-liquidating (dummy): Dummy equal to 1 if the firm has self-liquidating.
- EBITDA to Net Sales: Operating profitability ratio; how much earnings the company is generating before interest, taxes, depreciation, and amortization, as a percentage of revenue.
- Equity to Total Assets: Equity/Total Assets (financial leverage).
- Financial mismatch: [short-term liabilities - short-term assets] / total assets.

Additional material - description of the variables

- Interest Expenses to Cash Flow: 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% of total loans.
- Net Debt to EBITDA: Debt sustainability ratio, how long a company would need to operate at its current level to pay off all its financial debt.
- Receivables Turnover Ratio: AccountsReceivable/NetRevenues. Efficiency and liquidity ratio that relates the firm debt towards its suppliers to its revenues net of variable costs.

Additional material - description of the variables

- Sales Growth Rate: Yearly growth rate of net sales.
- 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.
- 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.