Corporate Default Forecasting with Machine Learning

Mirko Moscatelli

Bank of Italy

October 21, 2019

Overview

The paper in one sentence

We compare traditional statistical models with machine learning models based on ensemble decision trees, namely Random Forest and Gradient Boosted Trees, in the task of corporate default forecasting.

The work belongs to the Suptech framework, that is, the use of innovative technology by central banks and supervisory agencies to support their institutional activities.

The importance of default forecasting

Default forecasting is of key importance for several financial players:

1. Financial institutions (e.g. screening potential borrowers, setting the terms of new loans).
2. Investors (e.g. bond princing and portfolio management).
3. Macroprudential authorities (e.g. surveillance of aggregate default risk).

Motivation of the work

1. Investigate how default forecasting can benefit from the use of machine learning models, that are able to fully exploit large dataset by capturing complex non-linear interactions between economic, financial and credit variables.

2. Compare, along several performance measures, the traditional approach for default forecasting based on standard statistical models (logistic regression and linear discriminant analysis) with a machine learning approach based on ensemble trees models (random forest and gradient boosted trees).

Empirical approach in brief

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 statistical models (linear discriminant analysis, logistic regression, penalized logistic regression) and machine learning models (random forest, gradient boosted trees) using the previous dataset.

3. We use the trained models to compute predictions (expected default probabilities) on previously unseen data.

4. We compare the accuracy of the predictions along several dimensions (AuROC, credit allocation, ECB backtesting, variable importance).

Overview

The forecasting framework

$$Y = f(X) + \varepsilon$$

- $Y \in \{0, 1\}$ is the variable that we want to predict.
- X is a set of variables that are supposed to be informative about the behavior of Y.
- f is the function describing the relationship between X and Y.
- ε is an error term that accounts for all the influences on Y not measured (or only partially measured) by X.

The forecasting framework

Usually we don't know f exactly, so we must use a model (e.g. logistic regression) to estimate it. Let $\hat{f}$ be the estimation of f. We can measure the expected squared error of the predictions as:

$$Err(X) = \mathbb{E}[(Y - \hat{f}(X))^2].$$

The error can be decomposed as:

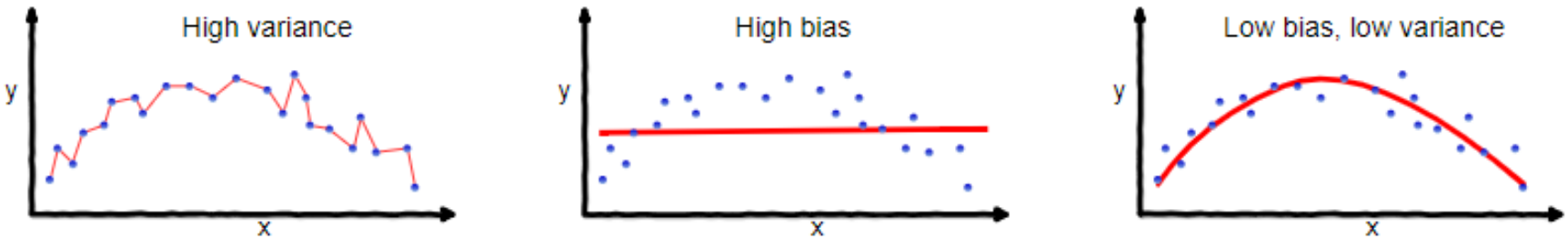
$$Err(X) = \left(\mathbb{E}[\hat{f}(X)] - f(X)\right)^2 + \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}[\hat{f}(X)]\right)^2\right] + \sigma_\varepsilon^2$$

$$Err(X) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Irreducible error arises from the fact that X doesn't completely identify Y. Reducible error ($\text{Bias}^2 + \text{Variance}$), on the other hand, arises from the fact that we don't know f and we are estimating it using $\hat{f}$.

The forecasting framework

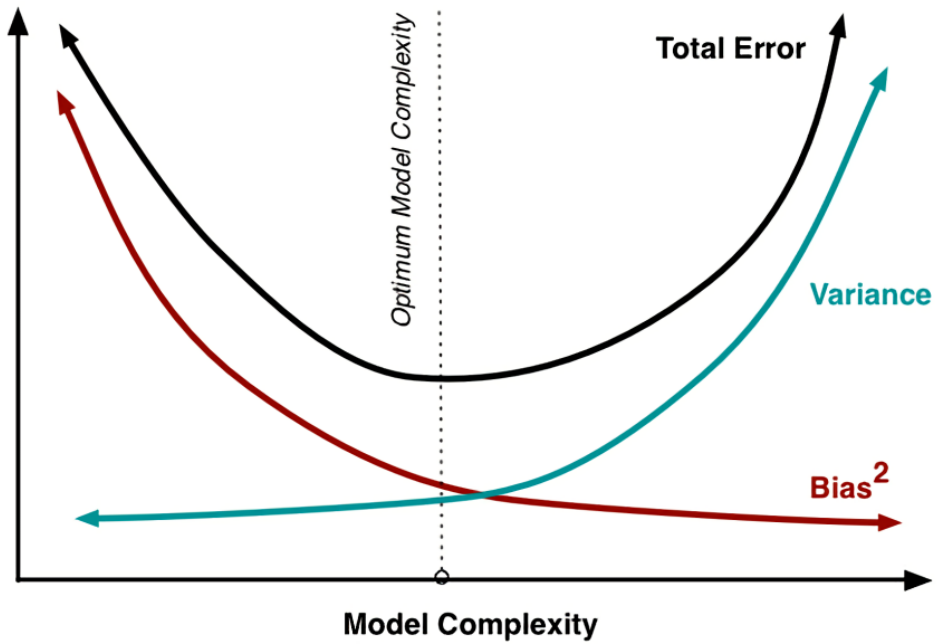
- High bias means that the model is missing the relevant relations between the predictive variables and the outcome (underfitting).
- High variance means that the model captures the random noise in the training data, rather than just the underlying relations between predictive variables and the outcome (overfitting).



Picture from towardsdatascience.com

The forecasting framework

Traditional statistical models typically rely on strong assumptions regarding the structural relationships between variables (e.g. $f(X) = g(X\beta)$) and, thus, tend to have a low variance but an high bias. Machine learning models, on the other hand, do not make strong assumption about the form of f, and try to achieve the best balance between bias and variance so as to maximize forecasting accuracy.



Picture from Fortmann-Roe 2012

Pros and cons of machine learning models

1. Pros:
   - Automatically capture non linear and non monotonous relationships between the predictors and the outcome variable.
   - Automatically capture relevant interaction effects between the predictors.
   - Typically more accurate out-of-sample forecasts.
   - Very weak assumptions on the structure of the data generating process.

2. Cons:
   - Less familiar for people coming from an econometrical background.
   - Less transparent than parametrical models (*arguably*).
   - Very weak assumptions on the structure of the data generating process.

Overview

Data

- Extensive dataset covering the years 2011-2018, including about 250.000 yearly firm-observations.
- Our target variable, the financial default, reflects a ICAS[1] system-wide definition of non-performing of a borrower: the ratio of its non-performing credit over its total credit greater than 5 per cent.
- Our predictive variables are financial and credit behavioral indicators for Italian non-financial firms (Credit Register, Balance sheet, Descriptive indicators). 38 variables in total, 26 after variable selection.

[1] Bank of Italy's In-house Credit Assessment System, whose purpose is to assess the credit risk of loans used as collateral in Eurosystem monetary policy operations.

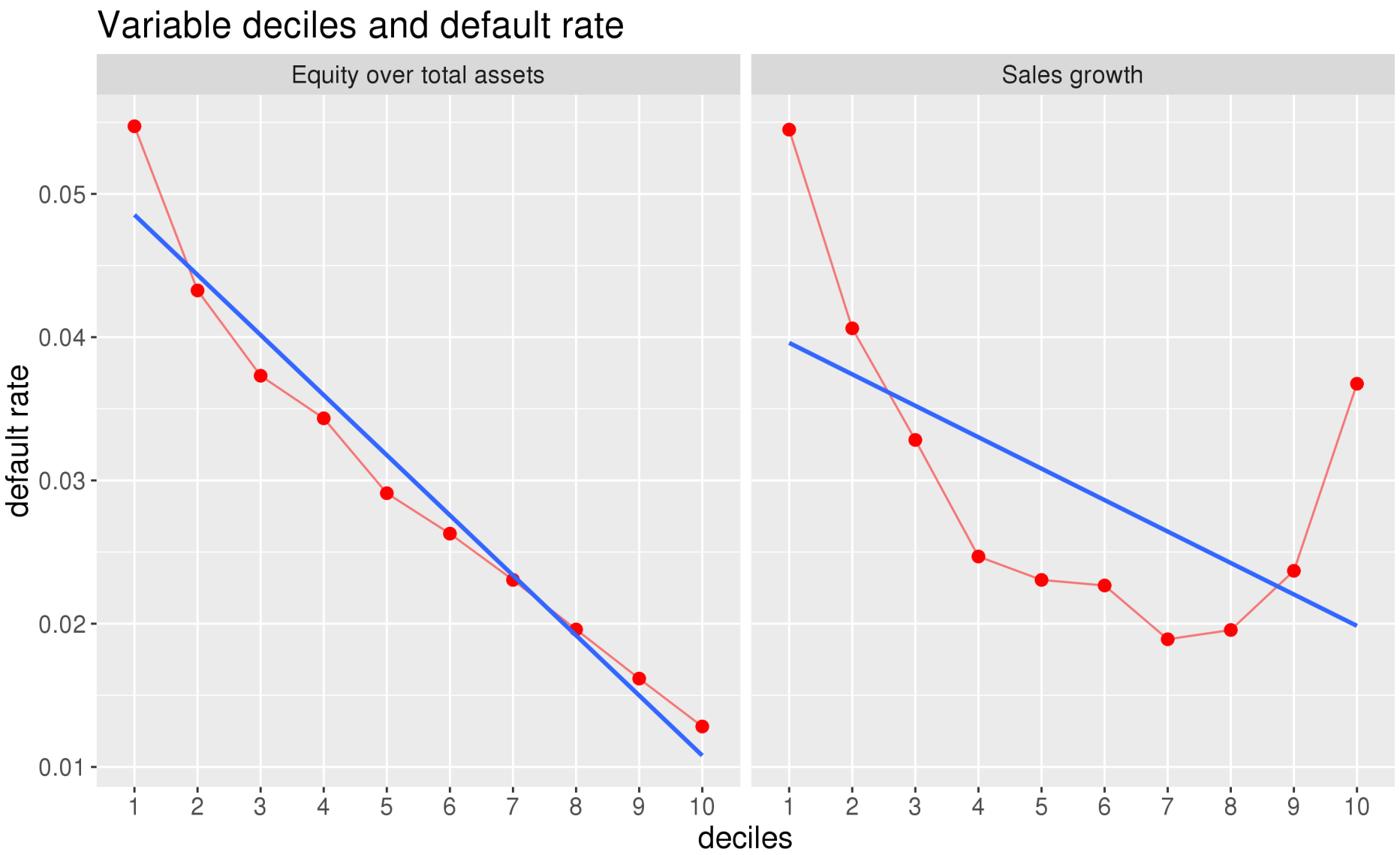
Variable selection

With the aim of not disadvantaging statistical models using variables completely non-linear or with an excessive correlation between them, we adopt a variable selection procedure often used when training traditional statistical models of credit scoring:

1. Using univariate logistic regression, we drop variables having an AuROC lower than 55 per cent.
2. Using the Kolmogorov-Smirnov test, we drop variables having insignificant differences in the distributions between the default and non-default groups.
3. From the variables satisfying the previous points, we iteratively drop variables having a correlation > 0.7, using the lower univariate AuROC as the drop discriminant.

This, however, results is a variable selection *far from optimal* for machine learning models.

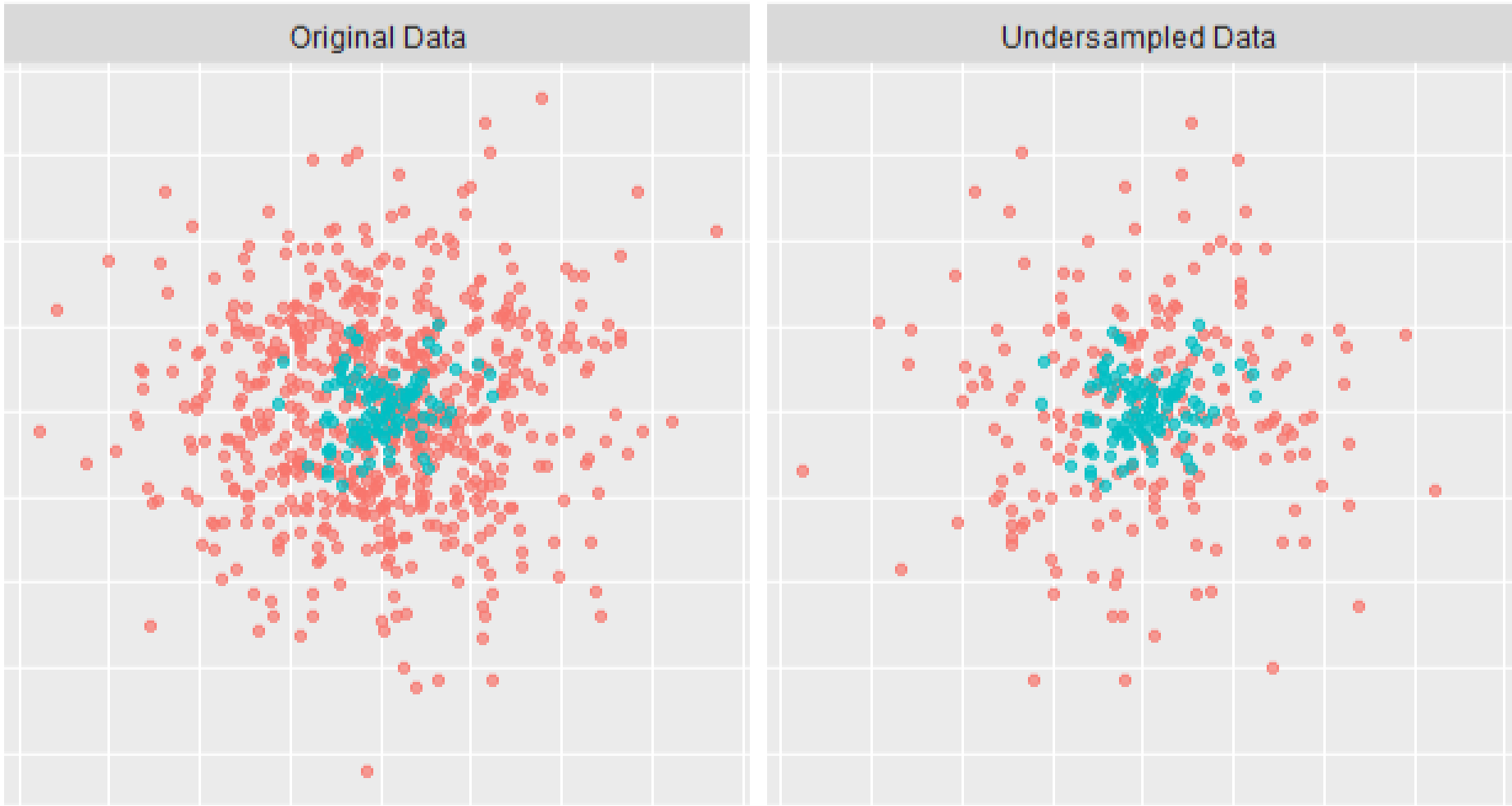
Variable selection



Imbalanced learning

- When one of the target classes is underrepresented in a dataset, the dataset is said to be *imbalanced*.
- Having few instances of a class often means that the estimation model tend to always predict the majority class, is unable to understand the characteristics of the minority class and performs poorly.
- A common strategy for dealing with imbalanced classification tasks is to undersample the majority class in the training set before learning the model, following the assumption that in the majority class there are typically many redundant observations.
- Since our dataset is strongly imbalanced (only about 3% of the firms are defaulting), we use undersampling to balance it.

Imbalanced learning



Picture from towardsdatascience.com

Imbalanced learning

- However, artificially rebalancing the training dataset violates the assumption that training and test dataset follow the same underlying distribution.

- This produces a bias in the out-of-sample posterior probabilities. However, since the probability that an observation is in the balanced training dataset is independent of the predicting variables x given the class y, we can exactly quantify the bias using the bayes rule and algebraically correct for it :

$$p_u = \frac{\beta \cdot p_b}{\beta \cdot p_b - p_b + 1},$$

where p_b is the raw (biased) probability, β is the number of defaulted firms over the number of non-defaulted firms, and p_u is the final (unbiased) probability.

Competing models

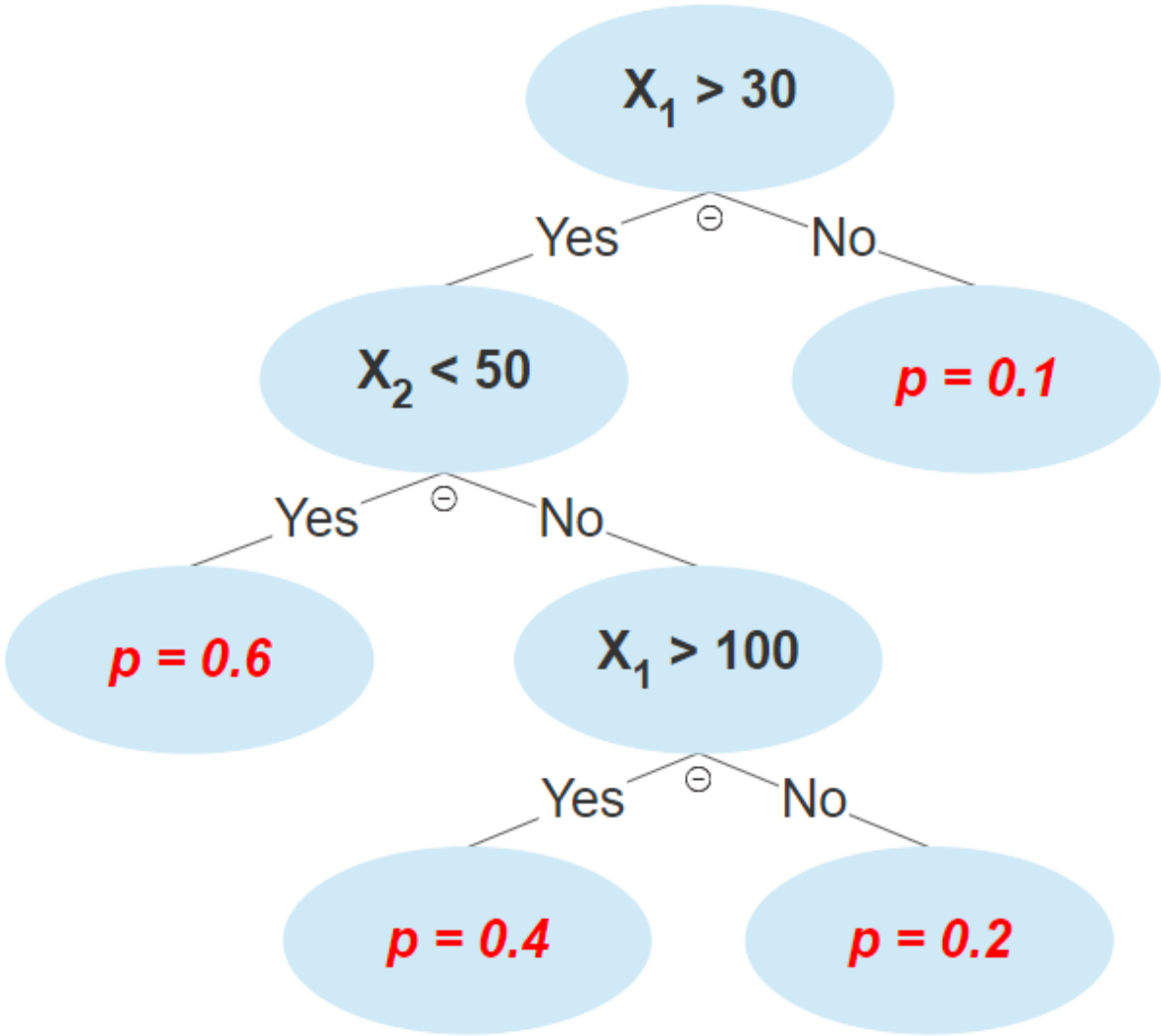
1. Linear Discriminant Analysis (LDA).
2. Logistic Regression (LOG).
3. Penalized Logistic Regression (PLR), i.e. the elastic net regularization of the logistic regression.
4. Random Forest (RDF).
5. Gradient Boosted Trees (GBT).

Machine learning models

- Random Forest and Gradient Boosted Trees.
- The building blocks for both of them are Classification Trees, partition algorithms that recursively select a variable X_i and a threshold t for that variable such that the resulting subsets (or branches) $\{X_i < t\}, \{X_i \geq t\}$ best separate defaulters from non-defaulters.
- Classification trees typically have low bias but high variance, and, thus, overall accuracy isn't great.

Machine learning models

An example of a classification tree:



Machine learning models

To improve forecasting accuracy, a large set of trees is grown and the final prediction is obtained as the average of the predictions stemming from individual trees. Random Forest and Gradient Boosted Trees differ in how they generate the set of trees. In particular:

- In *Random Forest*, trees are grown using boostrapped samples of the dataset and selecting, at each partition, the best split using only a randomly selected subset of the variables.

- In *Gradient Boosted Trees*, trees are grown recursively using a learning from mistakes approach, where at each step classification errors from the previous trees are used as the target variable to grow the next tree.

Overview

Evaluation metrics

- Our main measure of performance is the popular AuROC, that we choose because i) differently from accuracy, it deals well with situations where there is a skewed sample distribution, and ii) differently from measures such as precision, recall and F_1 score, it allows to assess the model independently of the choice of an arbitrary threshold.

- We compute the AuROC of the models on three dataset: one containing only financial indicators, one containing financial and credit behavioral indicators, and one containing financial and credit behavioral indicators but with a limited number of observations.

- Moreover, in the paper we also compute the AuROC across different cluster of firms, according to their sector and their size.

Evaluation metrics

To analyze the economic impact of the use of machine learning models, we also compare the models across:

- Credit allocation, examining the amount of credit that each model would allocate, the number of borrowers gaining access to credit and the default rate a lender would record using that credit rating model.
- The ECB backtesting, which aims to assess how closely estimated probabilities of default match realized defaults using a binomial-style test.
- The importance that each model gives to individual variables, measured as the decrease of AuROC that a model incurs if that variable is randomly permuted in the test dataset.

AuROC

AuROC when financial and credit behavioral indicators are available:

Year	LDA	LOG	PLR	RDF	GBT
2012	83.8%	**84.0%**	84.0%	**84.6%**	84.7%
2013	83.2%	**83.3%**	83.3%	**84.2%**	84.4%
2014	81.1%	**81.6%**	81.6%	**82.5%**	82.7%
2015	82.8%	**82.9%**	82.9%	**84.4%**	84.6%
2016	82.9%	**83.0%**	83.0%	**84.1%**	84.0%
2017	82.9%	**83.1%**	83.1%	**84.1%**	84.2%

AuROC

AuROC when only financial indicators are available:

Year	LDA	LOG	PLR	RDF	GBT
2012	73.7%	**73.9%**	73.9%	**76.6%**	76.3%
2013	73.7%	**73.9%**	73.9%	**77.2%**	77.3%
2014	72.2%	**72.3%**	72.4%	**74.4%**	73.9%
2015	73.7%	**73.7%**	73.7%	**76.1%**	76.0%
2016	72.6%	**72.6%**	72.6%	**75.3%**	75.3%
2017	73.0%	**73.0%**	73.0%	**75.7%**	75.4%

AuROC

AuROC when financial and credit behavioral indicators are available (restricted dataset):

Year	LDA	LOG	PLR	RDF	GBT
2012	82.5%	**82.6%**	83.5%	**83.9%**	83.5%
2013	82.8%	**82.9%**	82.9%	**83.0%**	83.2%
2014	80.5%	**80.8%**	80.7%	**81.1%**	80.8%
2015	82.6%	**82.7%**	82.7%	**83.3%**	83.4%
2016	82.6%	**82.7%**	82.7%	**82.7%**	82.3%
2017	82.4%	**82.6%**	82.6%	**82.5%**	82.2%

Credit allocation

Fixed granted amount:

% of the total granted	Allocated Amount	Method	Number of firms	Defaulted amount	Default rate	Default rate - diff. wrt LOG
1%	4,237	LDA	8,247	29	0.68%	0.12%
		LOG	7,724	23	0.56%	-
		PLR	7,777	24	0.56%	0.00%
		RDF	5,172	10	0.24%	-0.32%
		GBT	5,029	12	0.26%	-0.29%
5%	21,183	LDA	19,741	147	0.69%	-0.12%
		LOG	18,713	172	0.81%	-
		PLR	18,951	169	0.79%	-0.01%
		RDF	13,070	106	0.50%	-0.31%
		GBT	13,398	71	0.33%	-0.48%
10%	42,366	LDA	29,702	593	1.38%	0.06%
		LOG	28,349	564	1.33%	-
		PLR	28,633	498	1.17%	-0.16%
		RDF	20,022	232	0.54%	-0.79%
		GBT	20,479	349	0.81%	-0.52%
20%	84,732	LDA	49,826	1,761	2.03%	-0.03%
		LOG	48,745	1,788	2.06%	-
		PLR	49,300	1,802	2.08%	0.02%
		RDF	34,894	835	0.97%	-1.09%
		GBT	36,458	990	1.14%	-0.92%

Credit allocation

Fixed probability threshold:

Threshold	Method	Allocated amount	Number of firms	Default rate	Allocated amount - % diff. wrt LOG	Default rate - diff. wrt LOG
0.4%	LDA	4,123	8,060	0.76%	-46.0%	0.00%
	LOG	7,635	12,637	0.75%	-	-
	PLR	7,207	12,154	0.74%	-5.6%	-0.02%
	RDF	8,495	8,836	0.17%	11.3%	-0.59%
	GBT	16,264	15,338	0.28%	113.0%	-0.47%
1.5%	LDA	46,772	49,157	1.08%	-9.9%	0.00%
	LOG	51,893	53,133	1.12%	-	-
	PLR	50,455	52,290	1.10%	-2.8%	-0.01%
	RDF	75,830	54,098	0.70%	46.1%	-0.42%
	GBT	106,319	74,532	1.02%	104.9%	-0.10%
5.0%	LDA	244,409	151,450	2.62%	-0.4%	0.00%
	LOG	245,391	151,822	2.62%	-	-
	PLR	242,660	151,730	2.62%	-1.1%	0.00%
	RDF	256,324	152,906	2.41%	4.5%	-0.21%
	GBT	244,626	152,154	2.37%	-0.3%	-0.25%

ECB backtesting

		2012					2013				
CQS	*Treshold*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*
CQS1-2	*0.1%*	0.4%	0.5%	0.5%	0.0%	0.0%	1.0%	0.6%	0.6%	0.0%	0.0%
CQS3	*0.4%*	0.6%	0.7%	0.7%	0.4%	0.6%	0.4%	0.4%	0.5%	0.2%	0.2%
CQS4	*1%*	1.3%	1.4%	1.4%	1.1%	1.6%	0.6%	0.7%	0.7%	0.5%	0.6%
CQS5	*1.5%*	2.3%	2.5%	2.5%	2.3%	3.1%	0.9%	1.1%	1.0%	0.8%	1.2%
CQS6	*3%*	4.4%	4.5%	4.5%	4.5%	4.9%	1.7%	1.9%	1.8%	1.5%	2.0%
CQS7	*5%*	9.0%	9.0%	9.0%	8.9%	8.1%	3.4%	3.5%	3.5%	3.1%	3.6%
		2014					2015				
CQS	*Treshold*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*
CQS1-2	*0.1%*	0.5%	0.7%	0.7%	0.0%	0.0%	4.8%	2.7%	2.0%	0.5%	NA
CQS3	*0.4%*	1.0%	1.1%	1.0%	0.2%	0.5%	0.9%	0.8%	0.8%	0.1%	0.1%
CQS4	*1%*	1.4%	1.6%	1.6%	0.7%	1.4%	0.9%	0.9%	0.9%	0.3%	0.5%
CQS5	*1.5%*	2.1%	2.3%	2.3%	1.5%	2.7%	0.9%	1.1%	1.1%	0.6%	1.0%
CQS6	*3%*	3.3%	3.4%	3.3%	3.2%	4.1%	1.7%	1.8%	1.8%	1.4%	1.9%
CQS7	*5%*	5.4%	5.6%	5.6%	6.3%	6.4%	2.8%	2.8%	2.8%	3.1%	3.5%
		2016					2017				
CQS	*Treshold*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*
CQS1-2	*0.1%*	0.0%	6.4%	7.9%	0.0%	NA	12.5%	2.0%	1.8%	0.4%	0.0%
CQS3	*0.4%*	1.1%	1.0%	1.0%	0.0%	0.1%	0.5%	0.5%	0.4%	0.1%	0.2%
CQS4	*1%*	0.8%	0.8%	0.8%	0.3%	0.6%	0.7%	0.7%	0.7%	0.3%	0.6%
CQS5	*1.5%*	1.0%	1.1%	1.1%	0.8%	1.1%	1.0%	1.0%	1.1%	0.7%	1.2%
CQS6	*3%*	1.8%	1.9%	1.8%	1.6%	2.2%	1.7%	1.7%	1.6%	1.7%	2.2%
CQS7	*5%*	3.6%	3.5%	3.5%	3.5%	3.9%	3.7%	3.7%	3.7%	3.6%	3.7%

Variable importance (AuROC decrease)

Excluded variable	LDA	LOG	PLR	RDF	GBT
IE_CASHFLOW	4.5%	3.7%	3.7%	4.5%	5.4%
LOG_ASSETS	3.2%	3.1%	2.9%	1.7%	2.6%
DSCR	1.9%	2.9%	2.9%	2.2%	2.5%
EQ_TA	2.9%	3.0%	3.0%	1.2%	1.6%
TURNOVER	2.9%	2.9%	2.8%	1.4%	1.6%
CASH_ST_DEBT_S	2.4%	2.9%	2.8%	1.0%	2.3%
AREA_CVD	0.9%	0.9%	0.9%	0.3%	0.5%
ATECO_CVD	0.6%	0.6%	0.6%	0.4%	0.3%
PAYABLES_TURNOVER	0.4%	0.3%	0.3%	0.7%	0.6%
EBITDA_MARGIN	0.4%	0.3%	0.3%	0.3%	0.7%
DIM_CVD	0.6%	0.6%	0.5%	0.0%	0.0%
PFN_EBITDA	0.3%	0.3%	0.3%	0.3%	0.4%
SALES_GWT	0.0%	0.0%	0.0%	0.5%	0.9%
VA_TA	0.2%	0.2%	0.2%	0.4%	0.4%
FIN_MISMATCH	0.2%	0.1%	0.1%	0.5%	0.4%
CASH_TA	0.1%	0.2%	0.2%	0.2%	0.3%
RECEIVABLES_TURNOVER	0.1%	0.1%	0.1%	0.4%	0.3%
PFN_PN	0.0%	0.0%	0.0%	0.3%	0.2%

Main results

1. In general, models based on machine learning have a greater ability to predict default than traditional statistical models. This advantage is greater when the information available for the estimation is of lower quality (like the one available to external credit analysts).
2. A credit allocation rule based on borrowers' estimated default probabilities results in a larger supply of credit and, at the same time, a lower realized default rate when using PDs obtained from machine learning models.
3. The improvements appear to be due to the capacity of machine learning models to exploit complex relationships between predictors and the default outcome: indicators presenting a non-linear relationship with the default outcome are more important for machine learning models than for statistical ones.

Overview

Summary

1. We compare traditional statistical models and machine learning models with respect to the task of default forecasting.

2. Machine learning models seem to outperform traditional models, in particular when the information available for the estimation is of lower quality.

3. The improvements appear to be due to the ability of machine learning models to automatically capture interactions between variables as well as non linear/non monotonous relationships with the default outcome.

Next steps

1. Develop an optimal framework to train the Random Forest model by i) improving the variable selection and the hyper-parameter tuning, and ii) updating the default predictions periodically.

2. Work on *understanding* the Machine Learning models, using measures developed in the recent literature that help to interpret and explain their underlying logic, so as to increase the trust in them.

3. Aggregate the default probabilities by policy-relevant bank/firm clusters to obtain accurate and timely estimations of the aggregate default rates.

Thank you!