Loan Default Analysis in Europe: Tracking Regional Variations using Big Data

Luca Barbaglia

European Commission, Joint Research Centre (JRC), Ispra, Italy

February 18, 2019

Abstract. We analyse 20 million residential mortgages in seven European countries. The loans are observed over time, thus constituting a big multivariate data set with approximately 400 million observations. We model the occurrence of a default using loan-level information about the conditions at which the loan was granted and the borrower's economic situation. We associate each loan to a European province (NUTS 3 statistical regional) and control for local economic conditions. We compare the performance of three machine learning methods in predicting the default event, namely Logistic Regression, Gradient Boosting and Random Forest. We run this comparison at country and regional level. We find that the most important variables are the loan originator and the date at which the loan was issued. Conversely, other loan specific or borrower specific information, like interest rate or income, are less relevant. These results are robust across the regions of analysis, although our methodology captures across-country differences in variable importance magnitudes.

Keywords: Big Data, European Datawarehouse, Loan Default, Machine Learning, NUTS Statistical Regions

JEL classification: C55, D14, R11

1 Introduction

The empirical mortgage literature has investigated the drivers of loan default behavior, especially after the 2008 sub-prime crisis. The delinquency status of a loan might be linked, for instance, to the interest rate value applied, to the loan-to-value ratio, to the employment status of the borrower or to the macro-economic conditions specific to the borrower's geographical area (Elul et al., 2010, Campbell and Cocco, 2015). Many factors might influence simultaneously the decision of a borrower to default, but only few of these factors can be tuned or controlled by policy makers. A better understanding of the drivers of loan default behavior might help policy makers to identify which leverages to address in order to reduce delinquency cases, thus reducing the costs caused by inefficient allocation of resources.

A large strand of the existing literature has explored the US case, while few studies focus on Europe. The US and European credit systems differ significantly with respect to market structure and macroeconomic conditions, and the findings about borrower's behavior in the US might not apply to Europe. For instance, Europe faced a sovereign debt crisis in 2011 which caused severe credit restrictions in many European countries, which might cause an European borrower to behave differently than a US counterpart. The main reason why few studies focus on Europe is the lack of consistent and reliable data for the various European countries. The European Datawarehouse (ED) data set partially fills this gap, giving researchers an opportunity to explore the European credit market.

The ED data set provides us a large amount of information about the loan conditions and borrower's demographics. Consider, for instance, the ED table about residential mortgage-backed securities (RMBS) in Europe. This table reports micro-level data for around 20 million loans, whose information is updated when the corresponding pool is re-submitted to the ECB by the responsible bank. In total, it sums up to approximately 400 millions rows, which will continue to grow given that ED is regularly updated. Not only we have many rows for each table, but also many columns are available. RMBS contains around 60 columns, many of which are categorical variables with up to 20 categories (e.g., loan purpose). Additional information about the health of the economy in a specific region and or in a specific moment in time can be incorporated, thus further increasing the size of the data set.

The large number of factors that can affect loan default dynamics create a highly complex system to study. Factors might interact between each other: the probability to default might be impacted by a factor, but the size or sign of the effect may depend on another variable's level. Moreover, the majority of the literature constraints the relationship between default behavior and other factors to be linear. This choice simplifies the analysis but might miss-specify effects which can potentially show significant non-linearities (Sirignano et al., 2018).

We tackle the difficulties created by the size of the data and by the complexity of the non-linear relations

among variables by exploiting some of the major algorithms provided by the machine learning literature. Typically these models do not involve complex mathematical operations (e.g., matrix inversions) which might result hard to scale or even unfeasible with big data. Machine learning algorithms combine multiple operations by making some simplifying assumptions or by aggregating (e.g., averaging over multiple decision trees in Random Forest). The simple basic structure of these models allows us to exploit the loan-level information available in ED and possibly attain a more accurate description of the default behavior in Europe.

Section 2 makes an overview on the literature on loan default analysis and display how this work enter the existing literature. The data and the models are discussed in Section 3 and 4, respectively. Section 5 discusses the results of the country-level analysis. Finally, Section 6 concludes and highlight possible future line of research.

2 Literature Review

Economic theory has formulated a number of hypotheses to explain the default of a loan. In the two last decades, the sharp increase in the number of default and personal bankruptcies attracted the attention not only of policy makers, but also the attention of researchers trying to understand the drivers of such phenomenon. A standard reference in the literature is the incomplete-market model by Eaton and Gersovitz (1981), where the interest rate applied on the loan is a key factor considered by the borrower when choosing whether to default or not. The interest rate is a function of the amount borrowed and reflects the probability of default, thus compensating the lenders for the losses they will suffer (see Livshits, 2015 for a review). A large strand of the literature has focused on default occurrences specifically for residential mortgages in the US market. Cunningham and Capone Jr (1990), Deng (1997), Elul et al. (2010) and Campbell and Cocco (2015) are some example of default analysis using macroeconomic and loan-level information. Cunningham and Capone Jr (1990) use a multinomial logistic regression model to study the default behavior of fixed-rate and adjustable-rate loans in time of volatile interest rates and house prices. Deng (1997) adopts a proportional hazard framework to evaluate mortgage default, where the hazard function includes time-varying covariates. Campbell and Cocco (2015) solve a dynamic model of household mortgage decisions including labor income, house price, inflation, and interest rate risk.

The above models use classical option models or standard econometric techniques to analyse the default behavior. The recent increased availability of data and the contemporaneous advances in statistical tools to analyse large data sets have pushed researchers to explore the applicability of machine learning techniques to study the residential credit markets. Among other works, Feldman and Gross (2005) apply the Classification and Regression Tree (CART) algorithm to study a set of approximately $3,000$ mortgages in Israel using loan's

and borrower's information. Fitzpatrick and Mues (2016) find that boosted regression trees and random forests outperform penalised regressions in analysing the default behavior of $300,000$ Irish mortgages. More recent works increased the number of loans under analysis. For instance, Sirignano et al. (2018) train a deep-learning model on a 120 million mortgage data set of US household, including loan-level information, as well as zip code level macroeconomic variables. Also Fuster et al. (2018) study millions of US mortgages using the logistic (linear and non-linear) and random forest models: among other variables, they include as regressors non-financial borrower's information (e.g., ethnicity) and analyse whether the application of machine learning techniques might favour a demographic group against the others.

All the above works focus on US large data sets or on smaller non-US data sets. to our knowledge, no work analyses the default behavior in Europe using a large multi-country data set. This work aims at bridging this gap by analysing the behavior of millions of default in residential mortgages across seven European countries, namely Belgium, France, Ireland, Italy, The Netherlands, Portugal and Spain. In terms of complexity of the model and of explanatory variables analysed, our work is close to Sirignano et al. (2018), while with respect to the methodology, we extend Fitzpatrick and Mues (2016) by comparing penalised logistic regression, gradient boosting and random forest models on a large multi-country European data set. The comparison of different machine learning models on a big data set from many European countries allows us to investigate which drivers are most important in describing the default occurrence and whether there exist relevant differences in the default behavior across NUTS3 European regions.

Finally, notice that machine learning methods and large data set have already been explored in various economic studies. For instance, Khandani et al. (2010) construct a non-linear non-parametric forecasting method using tree-based machine learning techniques to study consumer credit-risk time series. They claim to significantly improve the classification rates of credit-card-holder delinquencies and defaults, thus inducing a large reduction of the associated financial losses. Among other works, we highlight Mullainathan and Spiess (2017) who propose a review on machine learning methods with an econometric perspective and show the possible applications and associated challenges if used to study economic problems.

3 Data

The ED is a centralized platform that collects information on loans for 10 European countries as part of the liquidity operations of the European Central Bank (ECB). The program started in January 2013 and requires banks to report information on the structure and performance of their securitized loan portfolios in a detailed and standardized format in order for an Asset-Backed Security (ABS) to be eligible as collateral in Eurosystem refinancing operations. The type of loans covered by ED ranges from residential mortgages, car leasing, small-medium enterprise loans and credit cards. The data set contains (static) information measured

Table 1: Default Rate by Asset Country

Country	Distinct Loans	Distinct Default	Default Rate (%)
Belgium	1,898,699	10,586	0.56
France	5,589,915	13,080	0.23
Ireland	387,880	11,921	3.07
Italy	2,136,487	70,901	3.32
The Netherlands	5,054,017	30,045	0.59
Portugal	727,367	29,549	4.06
Spain	2,122,939	76,655	3.61

at the time the loan was granted, such as the total amount of loan or the gross income of the borrower at the time the loan was originated. In addition, the data set reports (dynamic) information about the performance of the loan, updated on a monthly basis when the bank submit their reports to the ECB.

We use the loan-level data on RMBS provided by the ED. Notice that the same analysis could be carried out using the SMEs, credit card, car leasing and consumer finance tables from ED (see Ertan et al., 2017, for a description of the data set). Table 1 reports the number of distinct loans by country in ED, the number of distinct loans that ever appeared in the data set as defaulted and the corresponding default rate. The distribution of the loan is not proportional to the population of the country. Indeed, countries like Italy and Spain present fewer unique loans than a smaller country like the Netherlands. Nevertheless, the default rate does not substantially differ across countries: ranging from the 0.23% of France to the 4.06% of Portugal. Importantly, the default rate is low for all countries (below 5%), thus indicating a very unbalanced data set.

We clean the ED data set in the following steps. (i) We remove all missing or duplicated rows. (ii) We rebuild the the NUTS2 and NUTS3, if not already present, by searching for unique correspondences with the partial zip code provided in ED (ED code AR129 for RMBS). (iii) For each static variable, we substitute non-coherent entries within groups by the mode of the variable within the group (e.g., substitute primary income at the loan origination by individual loan). (iv) We remove all NUTS2 and NUTS3 with less than 100 observations. (v) We keep only countries in the euro area (i.e., remove all Scandinavian countries and the UK) and only loans denominated in euro. (vi) For each numeric variable, we remove the first and last percentiles. After all the above steps, the cleaned data set that we analyse in the application consists of 162 million observations.

We flag one loan as defaulted if the number of months in arrears at the corresponding pool cut-off date (ED code AR170) is larger than three (i.e., more than 90 days of delinquency). We rely on the number of

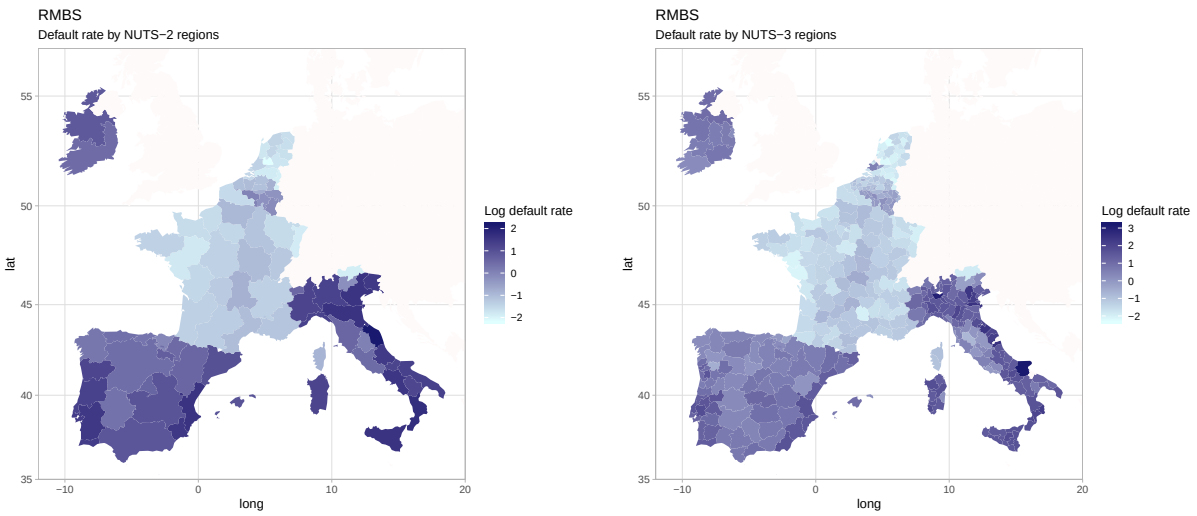

Figure 1: RMBS Log Default Rate by NUTS2 (left) and by NUTS3 (right).

months in arrears and not on the account status provided provided in the ED data set (ED code AR 166) in order to give a uniform and conservative definition of default across loan issuers and countries. When a loan is flagged as defaulted, we remove all following updates of the loan status and exclude the possibility of returning to the performing status.

Exploratory analysis

Figure 1 represents the log-transformed default rates on residential mortgages by NUTS2 and NUTS3. Across country variation is still evident: countries are easily distinguishable, thus indicating the presence of strong country-level effects. We also notice a clear intra-country heterogeneity. The regions in the east coast of Spain show higher default rates than other Spanish regions. For instance, the Valencian Community (NUTS2 ES52) presents a 4.89% default rate, 128 basis point higher than the national average. The range of intra-country variation can be very large for certain countries. For instance, in Italy the default rate ranges from the 0.01% of the South Tyrol (NUTS2 ITH1) to the 8.51% of the Marches (NUTS2 ITI3).

In order to explain the occurrence of default, we can exploit the variables present in ED containing borrower's or loan's information. For instance, the borrower's employment status when the loan was granted is one of the available categorical variables. Figure 2 reports the frequency with which each category of the borrower's employment status appears in the data set. The distribution of the borrower's employment status seems to be quite similar across the data set: in all countries, the category "Employed" represents more than 50% of the borrowers, the category "Self-employed" around 10%. Interestingly, the fraction of unemployed

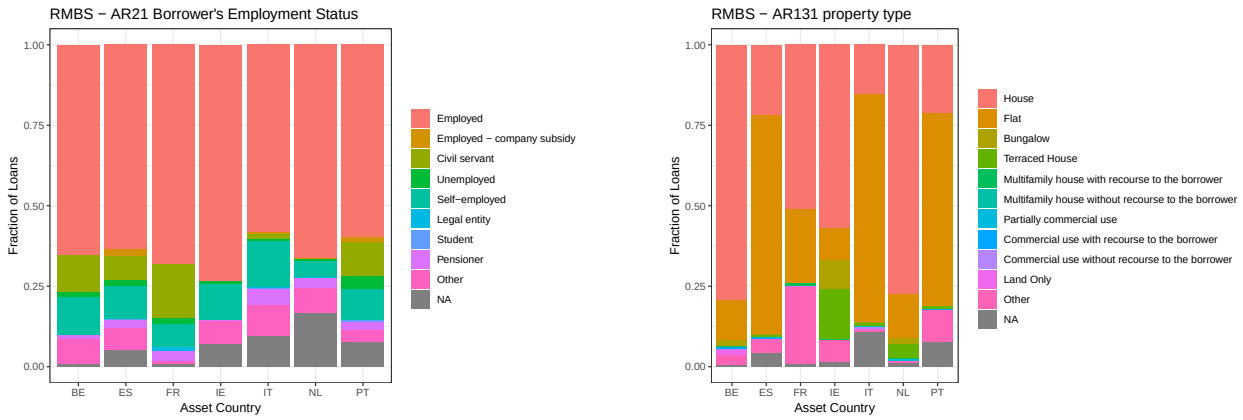

Figure 2: RMBS Borrower's Employment Status (left) and Property Type (right) by Country.

borrowers who were granted a mortgage is higher for Portugal than for other countries. Other variables could show larger differences across countries. For instance, Figure 2 reports the property type financed by the mortgage. In Belgium, France, Ireland and The Netherlands the largest fraction of mortgages was granted for houses, whereas in Spain, Italy and Portugal it was for flats.

Table 2 reports the features extracted from ED used as explanatory variable in explaining the occurrence of default. Notice that dates will be encoded as categorical variables in our modeling exercise. Based on the ED data, we add some additional features the *seniority* of the loan computed as time since its origination and the observed default rate by NUTS3 in the previous year as additional features. Moreover, we include the *average employment rate* and the *average GDP* by NUTS2 with a one year lag downloaded from Eurostat. The choice of explanatory variable is in line with, among others, Cunningham and Capone Jr (1990); Deng (1997); Campbell and Cocco (2015); Sirignano et al. (2018), where similar features are studied.

4 Model and Methods

In this section, we describe the different methods used to model default occurrences, namely the Logistic Regression, the Gradient Boosting and the Random Forest. The Logistic Regression models a binary dependent variable using a linear combination of the independent variables and it represents a standard econometric tool for binary classification purposes. The Gradient Boosting and the Random Forest are tree methods, respectively based on boosting and bagging from the machine learning literature, and they are able to capture non-linear interactions among the explanatory variables. In the last part of the section, we discuss the tools used to interpret the results.

Table 2: RMBS Features.

Feature	Source	Type	Description
Loan information			
Originator	ED (AR5)	Categorical	Lender that advanced the original loan
Loan Origination	ED (AR55)	Date	Quarter of original loan advance
Original Loan Balance	ED (AR66)	Numeric	Loan balance at loan origination (inclusive of fees)
Interest Rate Type	ED (AR107)	Categorical	Interest rate type
Current Interest Rate	ED (AR109)	Numeric	Interest rate at the pool cut-off date
Property Type	ED (AR131)	Categorical	Property type of the underlying asset
Original Loan-to-value	ED (AR135)	Numeric	Loan-to-value at loan origination
Log Valuation Amount	ED (AR136)	Numeric	Logarithmic property value as at loan origination
Current Loan-to-value	ED (AR141)	Numeric	Loan-to-value at pool cut-off-date
Seniority	ED (computed)	Numeric	Loan seniority at the pool cut-off date
Borrower information			
Borrower's Employment	ED (AR21)	Categorical	Employment status of the applicant at loan origination
Log Primary Income	ED (AR26)	Numeric	Logarithmic primary borrower underwritten gross annual income
NUTS3	ED (computed)	Categorical	NUTS3 where the property is located
NUTS information			
Lagged Default Rate	ED (computed)	Numeric	Default rate by NUTS3 lagged 1 year
Unemployment Rate	Eurostat	Numeric	Employment rate by NUTS2 lagged 1 year
GDP	Eurostat	Numeric	GDP by NUTS2 lagged 1 year

4.1 Logistic Regression

Let the dependent variable y be a categorical variable with two levels, namely 1 if the default occurred and 0 otherwise, and $\boldsymbol{x}$ a J-length vector of covariates associated to that loan. The probability of the default occurrence writes as:

$$P(y = 1|\boldsymbol{x}).$$

The Logistic Regression model the posterior probabilities of the two outcome classes of y as a linear combination of $\boldsymbol{x}$ and it ensures that these probabilities are defined in the $[0, 1]$ range (Hastie et al., 2009, p. 119). The fitted model takes the form:

$$\hat{y} = P\left(y = 1|\boldsymbol{x} = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}+\beta_0}}{1 + e^{\boldsymbol{x}'\boldsymbol{\beta}+\beta_0}}\right), \tag{1}$$

or equivalently:

$$\log\left(\frac{\hat{y}}{1-\hat{y}}\right) = \log\left(\frac{P(y = 1|\boldsymbol{x})}{P(y = 0|\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta} + \beta_0, \tag{2}$$

where β_0 is the intercept and $\boldsymbol{\beta}$ is a J-length vector of coefficients associated to the covariates $\boldsymbol{x}$.

The estimates of the coefficients β_0 and $\boldsymbol{\beta}$ are obtained solving the following likelihood:

$$\sum_{i=1}^{N}\left(y_i(\boldsymbol{x}'\boldsymbol{\beta} + \beta_0) - \log(1 + e^{\boldsymbol{x}'\boldsymbol{\beta}+\beta_0})\right). \tag{3}$$

If the number of coefficients to be estimated $J + 1$ is large relatively to the number of observations N, then solving the standard likelihood might bring low estimation accuracy or even result unfeasible. To avoid such problems, we solve the following penalized likelihood:

$$\min_{\beta_0,\boldsymbol{\beta}} \; -\frac{1}{N}\sum_{i=1}^{N}\left(y_i(\boldsymbol{x}'\boldsymbol{\beta} + \beta_0) - \log(1 + e^{\boldsymbol{x}'\boldsymbol{\beta}+\beta_0})\right) + \lambda\sum_{j=1}^{J}\left(\alpha|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2\right), \tag{4}$$

where λ is a non-negative regularization parameter and $0 \leq \alpha \leq 1$ is the elastic-net parameter (Zou and Hastie, 2005). The elastic-net estimator combines the Lasso and Ridge estimators, which respectively allow for variable selection and shrinkage, via the parameter α. The overall impact of the penalty is controlled by λ: the larger λ, the stronger the variable selection or shrinkage imposed by the elastic-net.

4.2 Gradient Boosting

Boosting methods represent an effective tool from the machine learning literature for classification and regression purposes. They fit *sequentially* a weak classification algorithm on adaptively re-weighted versions of the initial training data. The re-weighting scheme is adaptive in the sense that previously misclassified observations are given a higher weight in the subsequent iteration, whereas correctly classified ones are assigned a lower weight. With this procedure, the boosting classification algorithm focuses more on the observations that are hard to classify than on already correctly classified ones. Several versions of boosting

exists (see Friedman et al., 2000 for a review of the different specifications), here we focus on gradient boosting as presented in Friedman (2001). In particular, we focus on gradient boosting for binary classification purposes which translates into an approximation of additive logistic models built on a Bernoulli likelihood.

The weak classifiers considered at each iteration are decision trees. They partitions the space of all explanatory variables into non-overlapping regions. Each tree can be expressed as $T(\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a set of parameters governing the interactions between the vector of explanatory variables $\boldsymbol{x}$. Being M the number of iterations, the boosted tree model is:

$$f_M^{GB}(x) = \sum_{m=1}^{M} T(\boldsymbol{x}, \boldsymbol{\theta}_m) \tag{5}$$

In a forward stage-wise procedure for additive modeling using gradient boosting, one can estimate the parameters $\boldsymbol{\theta}_m$ at iteration m by solving:

$$\min_{\boldsymbol{\theta}} \sum_{j=1}^{J} (-g_{jm} - T(x_j, \boldsymbol{\theta}_m))^2, \tag{6}$$

where g_{jm} is the gradient of the associated likelihood with respect to the explanatory variable x_j. In words, Equation (6) fits the tree to the negative gradient values by least squares (see Hastie et al., 2009, p. 359, for a detailed presentation of gradient boosting methods).

4.3 Random Forest

Random Forest models build on the *bagging* procedures, also known as *bootstrap aggregation*: bagging is a technique that averages many noisy but approximately unbiased models in order to reduce the variance of an estimated function (Hastie et al., 2009, p. 282). Random Forest models differ from boosting procedures since the latter fit many re-weighted trees in a sequential manner, while the former built many uncorrelated trees on a subset of explanatory variables and observations (Breiman, 2001).

Let B be the number of trees to be considered in the bagging procedure and σ^2 the variance of each tree. If the trees are identically distributed with positive pairwise correlation ρ, the average of the B trees has variance:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{7}$$

If B is large, then the second term in Equation (7) tends to zero, thus making the variance of the tree average depend on the size of ρ. The random forest model randomly selects a sub-sample of explanatory variables when building the trees, thus decreasing the pairwise correlation among the B trees and therefore reducing the variance of the tree average.

The random forest model writes as:

$$f_B^{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} T(\boldsymbol{x}, \boldsymbol{\theta}_b), \tag{8}$$

where $T(\boldsymbol{x}, \boldsymbol{\theta}_b)$ is the tree characterized by the set of parameters $\boldsymbol{\theta}_b$ in iteration $b = 1, \ldots, B$. A small number of explanatory variables selected in each tree will reduce the pairwise correlation between trees, thus reducing the variance of the average.

4.4 Assessment Metrics and Variable Importance

We build a number of out-of-sample metrics in order to compare the performance of the above models. We measure the classification performances computing true and false positive rates:

$$\text{TPR}(\hat{\boldsymbol{y}}, y) = \frac{\#\{i \ : \ \hat{y}_i = 1 \ \text{ and } y_i = 1\}}{\#\{i \ : \ y_i = 1\}},$$

$$\text{FPR}(\hat{\boldsymbol{y}}, y) = \frac{\#\{i \ : \ \hat{y}_i = 1 \ \text{ and } y_i = 0\}}{\#\{i \ : \ y_i = 1\}},$$

where $\hat{y}_i$ is the predicted occurrence of default and y_i is the observed outcome in the testing set with $i = 1, \ldots N_{\text{test}}$ and N_{test} the size of the testing data set. The true positive rate (TPR) indicates the proportion of loan default that were correctly predicted, while the false positive rate (FPR) indicates the proportion of loans that were incorrectly classified as defaulted. The plot of TPR against FPR at various thresholds is the Receiver Operating Characteristic (ROC) curve, which is a standard tool for binary classification performance measuring. Given a ROC curve, one can compute the Area Under the ROC Curve (AUC): the larger it is, the higher the classification performance of the model.

For tree-based classification models (e.g., Gradient Boosting and Random Forests), one can build *variable importance* plots. Each time a there is a split in the tree, the reduction in squared error is attributed to the splitting variable as its importance at the that node. This is repeated over all nodes in each tree and accumulated over all trees. The result is standardized on a percentage scale and represents the relative importance of the variable in predicting the default occurrence (Rifkin and Klautau, 2004).

5 Results

In this section we use the Logistic Regression, Gradient Boosting and Random Forest models to analyse the loan-level data from ED. We run the analysis at country and NUTS2 levels (cfr. Section 5.1 and 5.2 respectively). For each region in analysis (country or NUTS2), we partition our data into training, validation and test sets: we take 60% of the distinct loans as the training set, 20% for validation purposes and 20% as a test set. The sampling procedure is stratified such to account for homogeneous proportion of defaulted loans in the three data sets. As reported in Table 1, default occurrences are rare and the data set results highly unbalanced. To accurately account for the probability of the rare default event, we balance the classes either oversampling the most frequent classes or oversampling the less frequent ones. The final probabilities

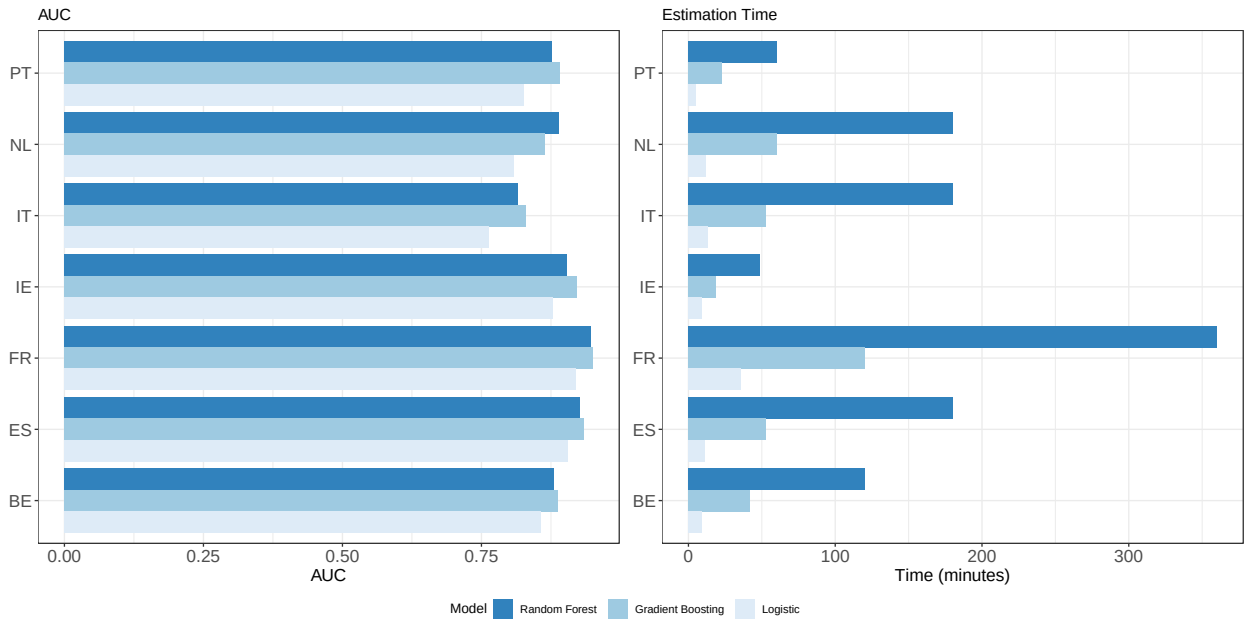

Figure 3: RMBS Performance metrics: AUC (left) and Estimation Time (right) by country and model.

are adjusted to the original sample via a monotonic transformation, which does not alter ordering thus not affecting AUC metrics (more details on class balancing by King and Zeng, 2001). We ran a four-fold cross validation and grid search for the selection of the parameters in various models.

5.1 Country-level results

Figure 3 reports the AUC and the estimation time by country for the Logistic Regression, Gradient Boosting and Random Forest models. The Gradient Boosting attains the highest AUC in all countries, except for The Netherlands where the Random Forest is the preferred model. For all countries, the Logistic Regression is outperformed by the other models[1]. With respect to estimation time, training the Logistic employs on average 14 minutes, the Gradient Boosting 53 minutes and the Random Forest approximately 3 hours[2].

Figure 4 reports the ROC curves for the three fitted models by country. They confirm the results of the AUC analysis: the Logistic is out-performed by the tree-based models and the Gradient Boosting attains the highest performance in the majority of the cases. In particular, we observe that the difference among Gradient Boosting and Random Forest in Belgium, France and The Netherlands is hardly visible and the associated ROCs intersect. Overall the Gradient Boosting seems to achieve the best performance, while, at

[1]The results of the model comparison with AUC are robust to the metric choice. For instance, similar results are obtained with the Gini Index (Hastie et al., 2009, p. 309).

[2]Estimation time for country-level analysis is obtained on a AWS m5d.12xlarge EC2 instance, with 48 CPUs, 192 giga-byte memory and SSD disks.

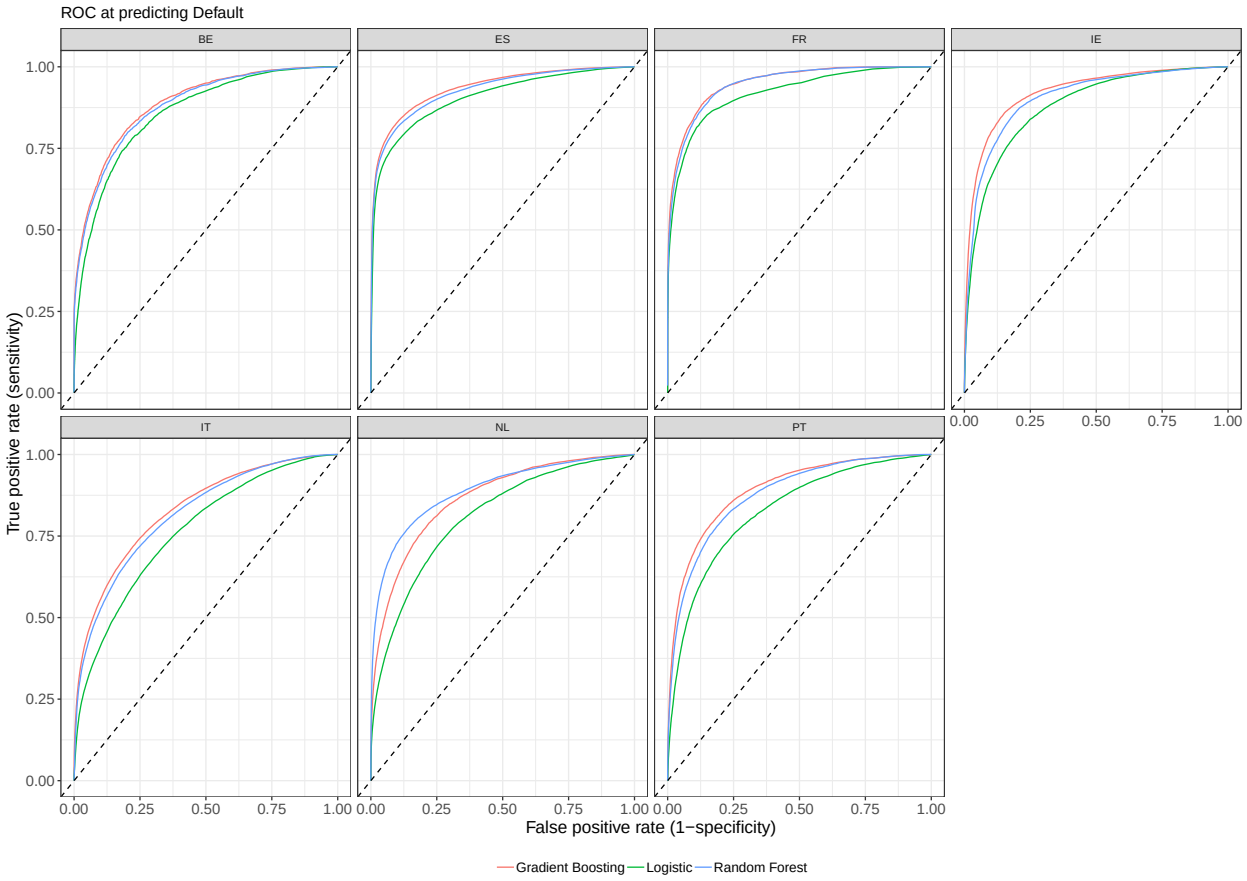

Figure 4: RMBS ROC curves by Asset Country.

the same time, being able to contain the computational complexity.

Variable Importance

Given the better performances of Random Forest and Gradient Boosting over the Logistic, we interpret variable importance estimated with the first two models. Figure 5 reports the percentage variable importance estimated with Random Forest and Gradient Boosting. The two models bring very similar results and the correlation between the percentage variable importance from the models is 0.92. Figure 6 shows the correlation among the variable importance from the Random Forest and the Gradient Boosting models among countries. The computed variable importance percentages not only are similar across models but also across countries: all country correlations are positive and range from the 0.47 of the pair Belgium-France to the 0.97 between Spain and France.

In all countries the most relevant variable is the loan Originator, that is the unique identifier of institution that issued the loan. Its importance ranges from the 78% and 60% of France and Spain respectively, to the

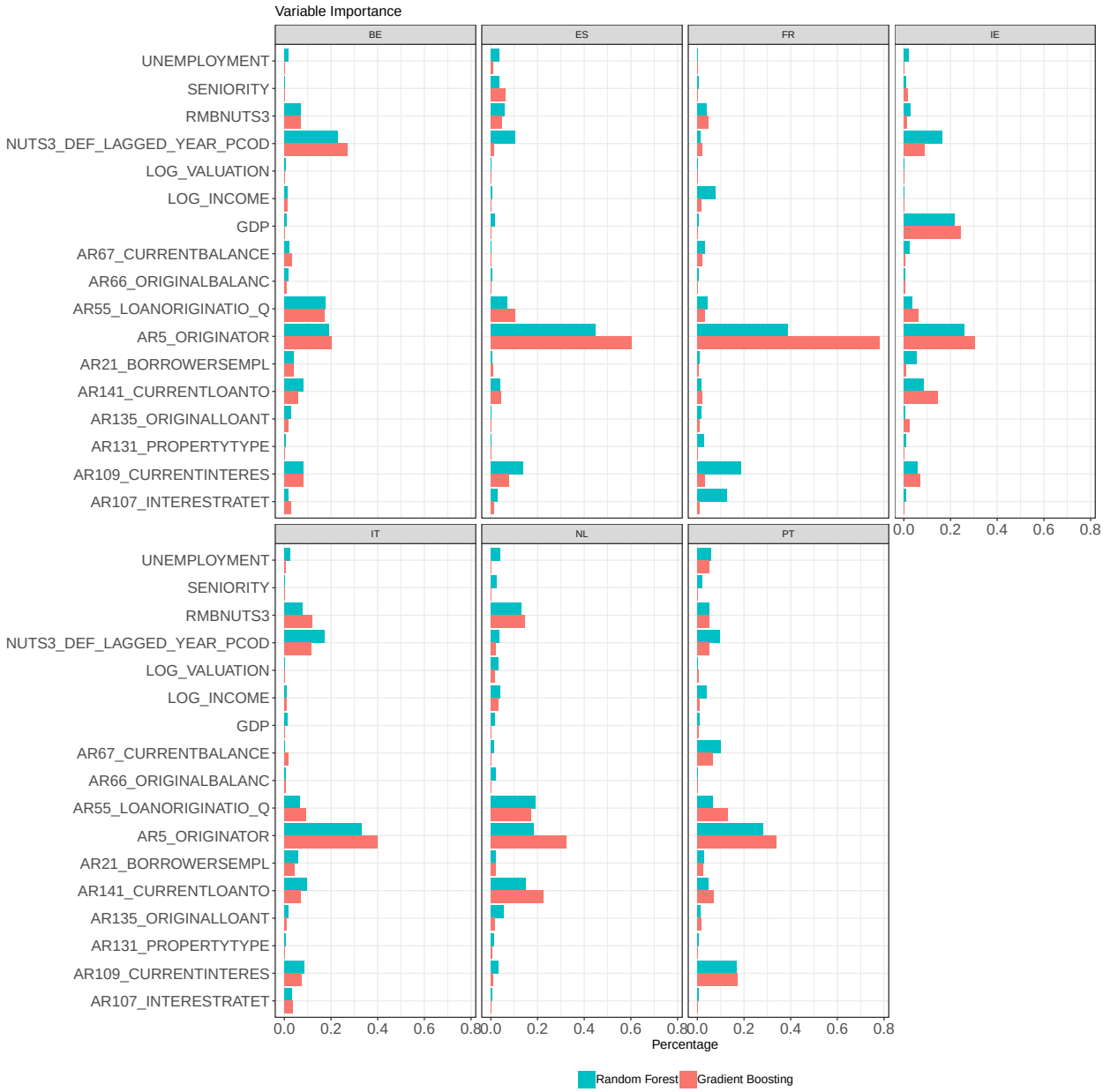

Figure 5: RMBS Variable Importance by Asset Country estimated with Random Forest and Gradient Boosting.

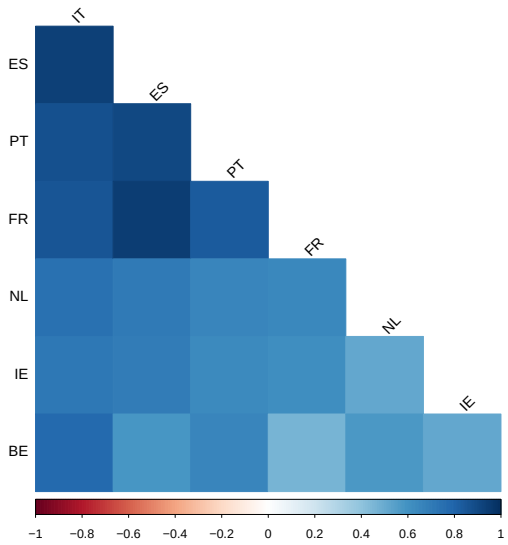

Figure 6: RMBS Correlation among Variable Importance by Asset Country estimated with Random Forest and Gradient Boosting.

20% in Belgium (for Gradient Boosting models). The Originator is the most important variable regardless of the model: the average importance of Originator is 42% and 30% for Gradient Boosting and Random Forest, respectively. Interestingly, the percentage importance of Originator is always larger for Gradient Boosting than for Random Forest models. Another variable that plays a relevant role is the date of loan origination, which has the second highest average importance for Gradient Boosting models (11%). The impact of the origination date is higher for Belgium, The Netherlands and Portugal, where the average importance of the Originator is lower. The interaction of Originator and Loan Origination seems to summarise the most relevant explanatory variables: the issuer and the issuing date are key variables in explaining the default occurrence.

Next to the information about the Originator and the Date Originating, the most relevant variables are geographical variables. Indeed, the combination of NUTS3 and its average lagged default rate show relatively high importance in Belgium, Ireland, Italy and The Netherlands. Their importance implies that geographical information at NUTS3 level is highly relevant in predicting the default occurrence, thus suggesting that NUTS3 region is a key factor to be considered when granting a loan. Conversely, lagged macroeconomic information at NUTS2 level (GDP and Unemployment) do not present high importance, suggesting that economic conditions are not a key factor in explaining default. It is worth noticing that variables containing borrower's related information or loan's specifications are not selected among the most relevant ones, although

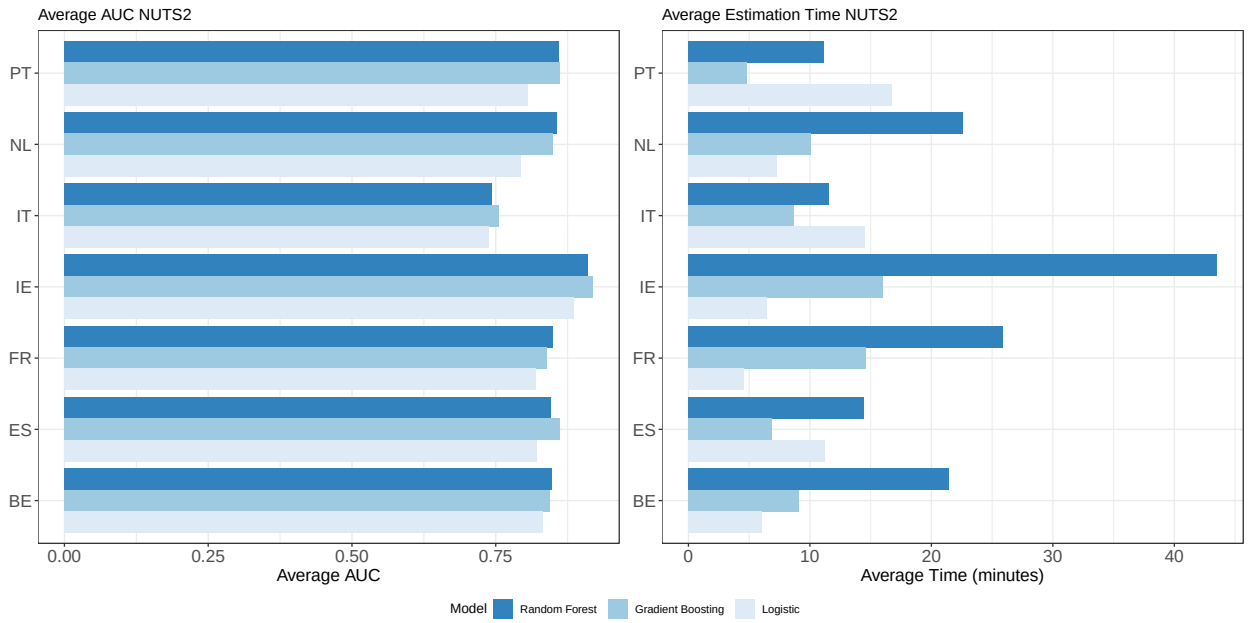

Figure 7: RMBS Performance metrics: Average AUC (left) and Average Estimation Time (right) from NUTS2 estimates.

economic intuition would suggest that they are key factors in explaining the occurrence of a default. For instance, the income of the primary borrower or her employment status at loan origination are among the less important variables, and the same conclusion applies to the initial interest rate and the amount of credit granted. This is quite counter-intuitive since borrower's characteristics and loan's conditions are usually among the most important features under evaluation by the issuer when deciding whether to grant the loan or not. Loan's related information acquire importance if updated at every pool cut-off date, as current interest rate and current loan-to-value.

5.2 NUTS2-level results

We repeat the training of the three models in analysis for each NUTS2 region. Figure 7 reports the out-of-sample AUC and the estimation time for the Logistic Regression, Gradient Boosting and Random Forest models averaged over the NUTS2 regions by country[3]. As for the country-level analysis, the Logistic Regression is outperformed by the tree-based methods considered, whereas there is no clear difference between Gradient Boosting and Random Forest models. However, we notice that the average training time for Gradient Boosting is always shorter than for Random Forest.

[3]Estimation time for NUTS2-level analysis is obtained on a AWS m5d.4xlarge EC2 instance, with 16 CPUs, 64 giga-byte memory and SSD disks.

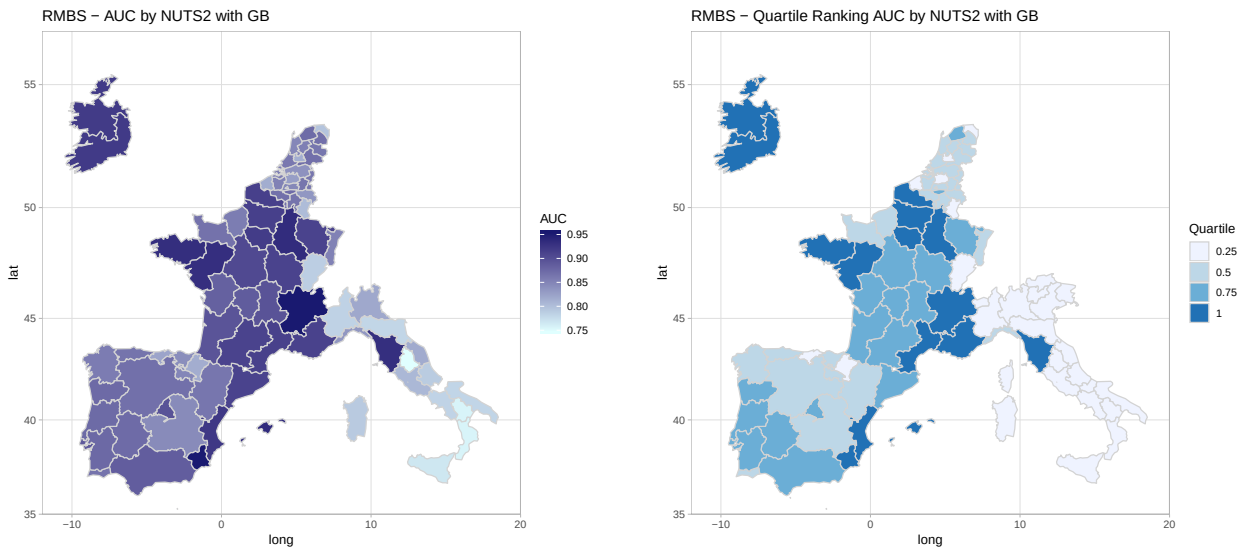

Figure 8: RMBS AUC (left) and associated quartile distribution (right) from Gradient Boosting estimates.

The left panel of Figure 8 plots the AUC from the Gradient Boosting estimates by NUTS2[4]. We observe that there is a large variability in terms of AUC across NUTS2 regions and that this variability does not relate with the default rates in the region: indeed, the comparison between the AUC (cfr. Figure 8, left panel) and the default rate (cfr. Figure 1, left panel) does not highlight any common pattern. The right panel of Figure 8 plots the NUTS2 regions where the AUC is divided into quartiles: we can now clearly identify macro-regions where the trained model attained similar performances. For instance, almost all Italian NUTS2 regions belong to the lowest AUC quartile, Spanish regions are mostly divided between the second and third quartiles, while French regions between the third and top quartiles. The clear presence of macro-areas grouping the NUTS2 regions could be explained by further reasearch and possibly linked to the stronger regional presence of some originators.

The NUTS2-level analysis brings similar results in terms of variable importance. Figure 9 and 10 in Appendix report the mean and the standard deviation of the variable importance estimated on the NUTS2 regions by the Gradient Boosting and Random Forest models. The impact of Origination and Loan Origination is still present as in the country-level analysis, whereas the impact of region-specific variables (e.g., lagged default rate in the NUTS3) has a much smaller magnitude. With respect to standard deviation, the most important variables are associated to a large variability across regions. Moreover, we notice that Italy presents a large standard deviation associated to most variables, thus suggesting the presence of more heterogeneous structure and stronger differences among Italian NUTS2 regions than for other countries.

[4]On the left panel of Figure 8, we remove NUTS2 regions with the lowest 5% of AUC for representation purposes.

6 Conclusions

The late decades witnessed the access to credit opening to a larger and larger audience. The increase in the number of credit lines was followed by a sharp rise in default cases after the 2008 financial crisis. A better understanding of the default drivers is of primary importance for policymakers in order to reduce the societal costs associated to a default and avoid the inefficient allocation of resources. We study the European Datawarehouse (ED) data set, containing information on million individual residential mortgages and on their borrowers across seven European countries, namely Belgium, France, Ireland, Italy, The Netherlands, Portugal and Spain. The information is reported at regional level (NUTS3 statistical region) and we add regional macroeconomic variables as additional controls. We predict the default occurrence using three different methodologies: a standard econometric tool, the Logistic Regression, and two tree-based machine learning technologies, the Gradient Boosting and the Random forest models.

Our results show that the Gradient Boosting attains higher out-of-sample performance in terms of prediction accuracy. For all countries, we find that the most important variables to explain the default occurrence are who issued the loan and when. Also the NUTS region where the loan was issued seems to be a relevant variable in explaining default. Conversely, loan and borrower specific variables (e.g., the value of the loan or the income of the borrower) are less important, even though these features are of primary importance when evaluating the possibility to issue the loan. The country-level and regional-level analysis bring consistent results, which seem to be robust across countries and models.

The importance of originator and the grouping of NUTS2 regions shown in Figure 8 should be investigated by looking in the more details the regional distribution of the originators. Future research should explore the presence of originator specific effects: for instance, table 3 in Appendix shows the originators with the highest proportion of default rates in each country. One could check whether including additional originator specific information would bring similar results in terms of variable importance. Future research could also explore the estimated decision tree to better understand the interactions among variables: for instance, Figure 11 in Appendix plots the estimated decision tree from the Gradient Boosting model for the NUTS2 FR10 (Île de France) and shows the relevant interactions among the originator and the other variables. Finally, future works could investigate whether the application of machine learning methods to study default behavior produce "fair" results or favour borrower with certain characteristics (Burrell, 2016; Fuster et al., 2018).

Appendix

Table 3: Loan percentage and default rate by originator. Top 5 by default rate in the country

Country	Originator	Loan in the Country (%)	Default Rate (%)
BE	DCED6E3F2A7587D9B297E863B559AE9E...	0.04%	14.64%
BE	ING Direct Belgium	1.95%	2.21%
BE	ING Belgium SA	7.65%	1.76%
BE	BNP Paribas Fortis SA/NV	32.36%	1.37%
BE	Axa	10.88%	0.95%
ES	Barclays Bank,S.A.U.(Spain)	0.01%	60.00%
ES	CATALUNYA BANC, S.A.	0.58%	54.76%
ES	0081	1.84%	52.21%
ES	0049	0.78%	52.16%
ES	2074- Caja Terrassa	0.00%	50.00%
FR	SG	0.00%	17.01%
FR	12368	0.92%	6.18%
FR	12328	1.34%	6.14%
FR	18709	0.65%	5.75%
FR	GEMB	1.09%	5.51%
IE	FAB	12.39%	25.18%
IE	KBC Bank Ireland plc	12.12%	20.08%
IE	EBS Ltd.	7.79%	17.63%
IE	UBR	9.50%	16.70%
IE	Bank of Scotland Ireland	7.43%	12.38%
IT	UNICREDIT SPA	0.45%	88.82%
IT	03285	0.04%	76.56%
IT	B.P.E.R.	0.02%	75.64%
IT	05704	0.01%	56.64%
IT	BANCA POPOLARE DI PUGLIA E BASILICATA	0.06%	55.84%
NL	Hypotrust	0.00%	55.88%
NL	Sparck Hypotheken	0.00%	53.33%
NL	856369BCB160DC3B0A8A557A3F3CA8A0	0.12%	46.74%
NL	55C38EBEEB6CDDAC52E6031938800589	0.02%	41.16%
NL	StaalBankiers	0.00%	39.53%
PT	BANCO SANTADER TOTTA. S.A.	0.53%	17.02%
PT	BANCO BPI	10.88%	12.51%
PT	Novo Banco, SA	0.01%	11.90%
PT	Banco Espirito Santo, SA	19.24%	10.95%
PT	Banif	8.10%	10.70%

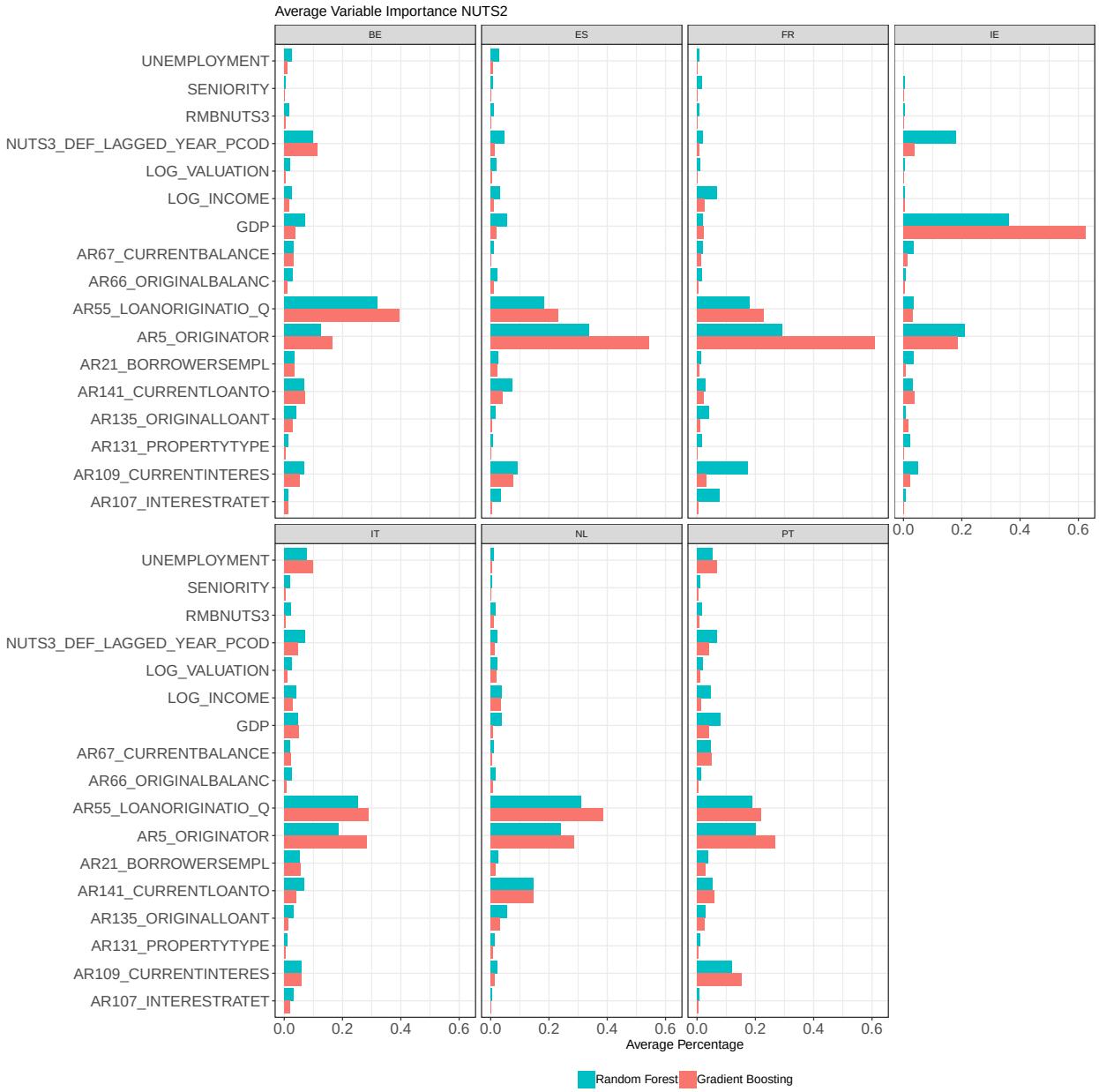

Figure 9: RMBS Average Variable Importance by Asset Country estimated with Random Forest and Gradient Boosting on NUTS2 regions.

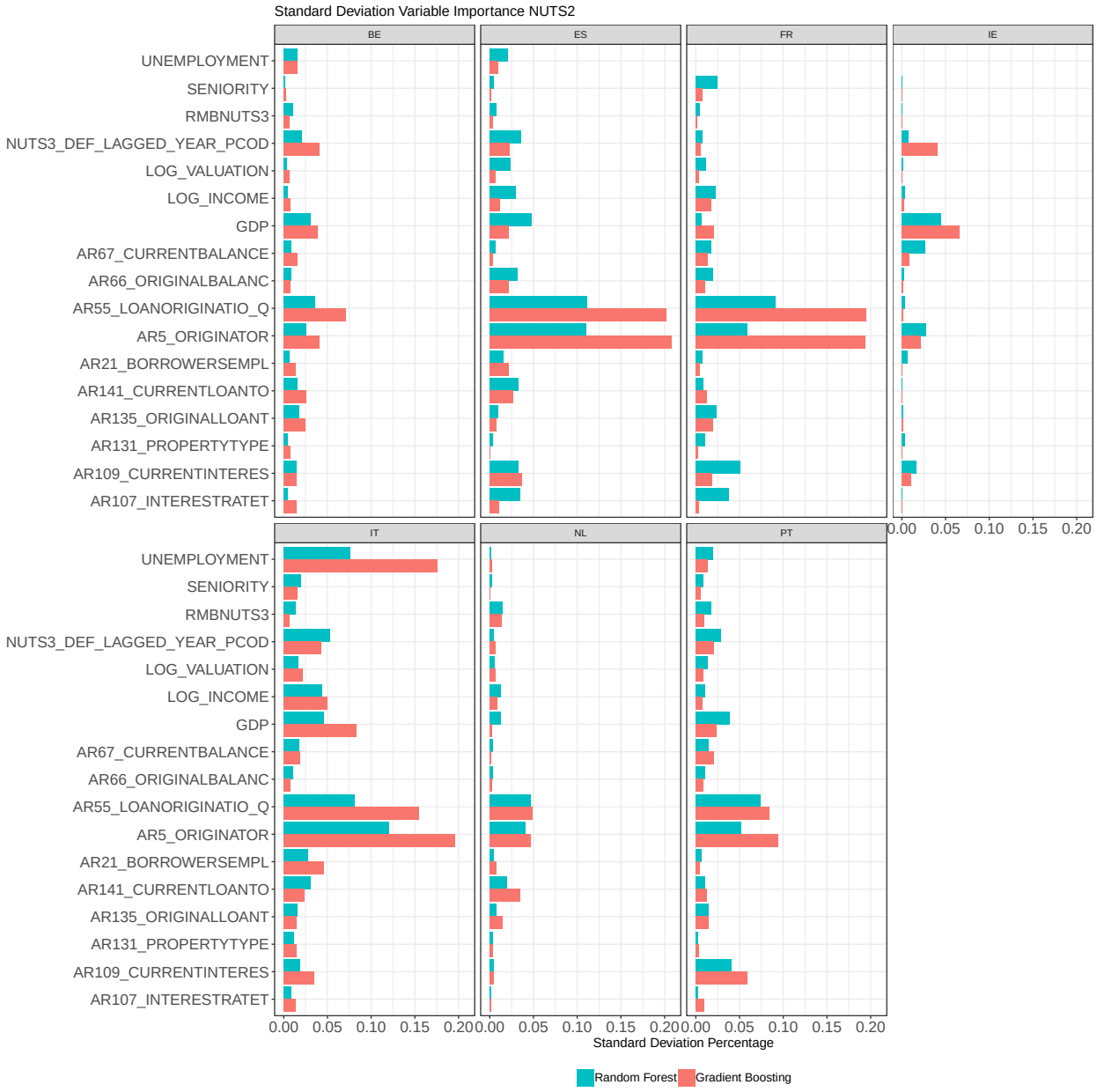

Figure 10: RMBS Standard Deviation Variable Importance by Asset Country estimated with Random Forest and Gradient Boosting on NUTS2 regions.

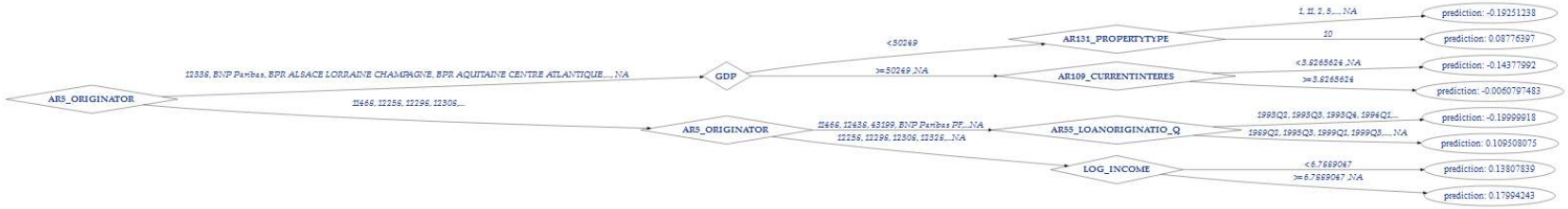

Figure 11: RMBS Decision tree estimated with Gradient Boosting on NUTS2 FR10 (Île de France).

References

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Burrell, J. (2016), "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms," *Big Data & Society*, 3, 1–12.

Campbell, J. Y. and Cocco, J. F. (2015), "A Model of Mortgage Default," *The Journal of Finance*, 70, 1495–1554.

Cunningham, D. F. and Capone Jr, C. A. (1990), "The Relative Termination Experience of Adjustable to Fixed-rate Mortgages," *The Journal of Finance*, 45, 1687–1703.

Deng, Y. (1997), "Mortgage Termination: An Empirical Hazard Model with a Stochastic Term Structure," *The Journal of Real Estate Finance and Economics*, 14, 309–331.

Eaton, J. and Gersovitz, M. (1981), "Debt with Potential Repudiation: Theoretical and Empirical Analysis," *The Review of Economic Studies*, 48, 289–309.

Elul, R.; Souleles, N. S.; Chomsisengphet, S.; Glennon, D. and Hunt, R. (2010), "What Triggers Mortgage Default?" *American Economic Review: Papers & Proceedings*, 100, 490–494.

Ertan, A.; Loumioti, M. and Wittenberg-Moerman, R. (2017), "Enhancing Loan Quality Through Transparency: Evidence from the European Central Bank Loan Level Reporting Initiative," *Journal of Accounting Research*, 55, 877–918.

Feldman, D. and Gross, S. (2005), "Mortgage Default: Classification Trees Analysis," *The Journal of Real Estate Finance and Economics*, 30, 369–396.

Fitzpatrick, T. and Mues, C. (2016), "An Empirical Comparison of Classification Algorithms for Mortgage Default Prediction: Evidence from a Distressed Mortgage Market," *European Journal of Operational Research*, 249, 427–439.

Friedman, J.; Hastie, T. and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting (with discussion and a rejoinder by the authors)," *The Annals of Statistics*, 28, 337–407.

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 1189–1232.

Fuster, A.; Goldsmith-Pinkham, P. and Ramadorai, T. (2018), "Predictably Unequal? The Effects of Machine Learning on Credit Markets," Tech. rep., Working paper available at SSRN: https://ssrn.com/abstract=3072038 or http://dx.doi.org/10.2139/ssrn.3072038.

Hastie, T.; Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

Khandani, A. E.; Kim, A. J. and Lo, A. W. (2010), "Consumer Credit-risk Models via Machine-learning Algorithms," *Journal of Banking & Finance*, 34, 2767–2787.

King, G. and Zeng, L. (2001), "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137163.

Livshits, I. (2015), "Recent Developments in Consumer Credit and Default Literature," *Journal of Economic Surveys*, 29, 594–613.

Mullainathan, S. and Spiess, J. (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Rifkin, R. and Klautau, A. (2004), "In Defense of One-vs-all Classification," *Journal of Machine Learning Research*, 5, 101–141.

Sirignano, J.; Sadhwani, A. and Giesecke, K. (2018), "Deep Learning for Mortgage Risk," Tech. rep., Working paper available at SSRN: https://ssrn.com/abstract=2799443 or http://dx.doi.org/10.2139/ssrn.2799443.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.