Journal of Financial Econometrics, 2021, 1–28
doi: 10.1093/jjfinec/nbab010
Article

Forecasting Loan Default in Europe with Machine Learning*

Luca Barbaglia [1], Sebastiano Manzan[2] and Elisa Tosetti[3]

[1]European Commission, Joint Research Centre, [2]Zicklin School of Business, Baruch College and [3]Ca' Foscari University of Venice

Address correspondence to Luca Barbaglia, European Commission, Joint Research Centre, Via E. Fermi, 2749, 21027 Ispra VA, Italy, or e-mail: luca.barbaglia@ec.europa.eu.

Received April 9, 2020; revised March 31, 2021; editorial decision April 3, 2021; accepted April 8, 2021

Abstract

We use a dataset of 12 million residential mortgages to investigate the loan default behavior in several European countries. We model the default occurrence as a function of borrower characteristics, loan-specific variables, and local economic conditions. We compare the performance of a set of machine learning algorithms relative to the logistic regression, finding that they perform significantly better in providing predictions. The most important variables in explaining loan default are the interest rate and the local economic characteristics. The existence of relevant geographical heterogeneity in the variable importance points at the need for regionally tailored risk-assessment policies in Europe.

Key words: big data, credit risk, loan default, machine learning, regional analysis

JEL classification: C55, D14, R11

Since the 2008 subprime crisis, a growing literature has been trying to identify the factors that have caused a rise in credit defaults and the subsequent financial distress. One explanation focuses on the role played by residential mortgages and points to a shift in the credit supply curve that allowed banks to increase lending to low income, poor credit quality borrowers, and lead to a soar in the house prices. When riskier borrowers started defaulting, the distress propagated from the housing market to the real economy through the banking sector with significant losses of output and jobs. Several studies, mainly focused on the United States, find that areas where lending to subprime borrowers increased significantly between 2002 and 2005 experienced sharp rises in mortgage defaults after the onset of the crisis (see Mian and Sufi, 2009 and Mian, Sufi, and Verner, 2017, for an international comparison). Other studies, however, point at the expectations about house price appreciation

* The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission.

as a key driver of borrowers' and lenders' decisions during the credit expansion and successive slump (Adelino, Schoar, and Severino, 2016). In Europe, between 1999 and 2006, several countries experienced a significant expansion in the credit to households, particularly in that part backed by real estate assets, with a subsequent severe contraction in economic growth and employment, and a decline in house prices. A short period of recovery was followed by a marked deterioration with the onset of the sovereign debt crisis in 2010–2012, when several European countries, like Greece, Portugal, Ireland, and Spain, were unable to repay or refinance their government debt.

Contrary to the United States, there is limited evidence regarding the drivers of mortgage defaults following the Great Recession in Europe. There are significant differences between the residential mortgage market in Europe and in the United States, which are likely to affect both borrowers' behavior and loan performance. First, mortgages in Europe are mostly recourse loans, meaning that the borrower remains personally liable, in case of default, for the difference between the outstanding debt and the sales price of the property. This prevents borrowers from defaulting for strategic considerations, such as in the case of negative equity. However, recent evidence in Gerardi et al. (2018) for the United States shows that the frequency of strategic defaults is low, relative to the alternative explanation that defaults are driven by the inability of the borrower to pay (e.g., due to an income shocks). The recourse nature of mortgages in Europe partly explains the significantly lower default rates relative to the United States, where the vast majority of residential loans are nonrecourse. Another difference with the U.S. mortgage market is that European banks grant loans at lower Loan-to-Value (LTV) ratios, typically limited to 80% of the value of the property.[1] The more conservative approach of European banks is partly explained by the lack of government-sponsored programs to guarantee residential loans. This implies that borrowers contribute a larger equity component to purchase the property, which lowers the risk of default, in particular, for strategic reasons. Finally, European countries are heterogeneous in the type of interest rate that is prevalent; while in Belgium, the Netherlands, and France, the fixed rate mortgage is prevalent, in the remaining countries, the adjustable rate is predominant. Albertazzi, Ongena, and Fringuellotti (2019) discuss the supply and demand factors that drive the choice of the interest rate type based on a European dataset. Floating rates have the effect of boosting credit growth when interest rates are declining. However, borrowers choosing floating rates are exposed to larger periodic payments and higher likelihood to default in periods of increasing interest rates and high inflation (Campbell and Cocco, 2015). On the other hand, Fuster and Willen (2017) analyze adjustable rate mortgages originated in 2005–2006 in the United States and find evidence that the decline in the interest rates had the effect of significantly lowering the default rate after the reset period.

In this article, we use a large dataset of residential mortgages in seven European countries to study the factors driving mortgage default in the period following the Great Recession. Understanding the main drivers of default can help to better predict the occurrence of a loan default and contribute to the design of reliable tools to assess the borrower's creditworthiness by financial institutions. This in turn would improve credit allocation,

[1] Another explanation is that bank capital requirements set a risk threshold for residential loans at 80%. A recent survey on the macroprudential policies curated by the IMF is available at https://www.elibrary-areaer.imf.org/Macroprudential/Pages/Home.aspx.

ultimately leading to new policy designs aimed at preventing default or helping households recovering from its effects. Our dataset includes approximately 12 million unique home-secured mortgages originated between 2000 and 2018, for which we observe the performance over the period from 2013 to 2018. For each loan, we observe loan-specific characteristics (e.g., the interest rate), borrower characteristics (e.g., income and employment status), and information on the geographical location of the asset. One specific feature of our data is that some variables reflect the situation at the time of the origination of the loan (e.g., household income), while other variables are dynamically updated (e.g., interest rates) and can be observed at the time of default. In our default prediction model, we also include proxies for local economic conditions that allow us to control for the overall state of the regional economy and also to account for the aggregate change in the variables that are not dynamically updated. This is particularly important in our setting since household income is only available at the origination date and never updated. Given the strong geographical heterogeneity observed in the default and its drivers, we conducted the analysis at the regional level, using information about the region where the asset is located.

The typical regression specification adopted by the loan default literature is the logistic model, thus constraining the log-odd to have a linear functional form. Although such assumption considerably simplifies the analysis, it may introduce misspecification bias due to neglected nonlinearities and interactions among variables (Sirignano, Sadhwani, and Giesecke, 2018). For instance, while certain loan characteristics, such as interest rate and LTV, are known to be important default drivers, their relevance might depend on other factors, for example, borrower's income or the state of the local economy. Hence, in addition to the logistic regression, in this article, we model the probability of default by adopting boosting techniques from the machine learning (ML) literature. The main characteristic of these models is that they use a large set of "mis-specified" nonlinear models (the so-called "weak learners") and perform model averaging based on some measures of accuracy. To improve the boosting algorithm transparency and to understand the sources of their increased performance relative to that of the logistic regression, in our empirical application, we adopt a set of tools from the recent literature on interpretable ML (IML; see, among others, Ribeiro, Singh, and Guestrin, 2016; Murdoch et al., 2019). These methods allow to measure feature importance and to understand how a specific regressor influences the default prediction.

Turning to our analysis of loan default, our results show that tree-based models significantly outperform the logistic regression in predicting out-of-sample defaults for most countries, thus supporting the usefulness of ML models. In order to validate the robustness of our results, we extend the comparison to a larger set of ML models that include random forests (RFs) and neural networks (NNs). Our results confirm the best performance of boosting models. We also find that the most important factors driving the default occurrence are the current interest rate, the current LTV, and the local economic conditions. The relative importance of these variables changes across country but they are consistently the top variables in terms of explanatory power for all countries. Interestingly, we do not find that borrower's characteristics at origination (income and employment status) are very relevant. Instead, our results indicate that the current interest rate and LTV have a significant impact on default occurrence and also that they are highly nonlinear, thus explaining the better performance of boosting models related to the logistic regression. In addition to differences in the relevance of these factors across countries, we also find significant within-country heterogeneity in their variable importance. For instance, local economic conditions are the most relevant in explaining defaults in the Eastern coastal regions of Spain, while in the rest of the country loan characteristics are more important (in particular the current interest rate). This evidence suggests that macroprudential policies aiming at controlling risks associated with residential lending should also consider the regional heterogeneity of credit markets. Policies should aim at targeting those regions that are experiencing excessive credit growth and a booming housing cycle which might require different policy instruments even within the same country. Furthermore, limiting credit supply in the regions experiencing excessive growth would allow to channel credit toward other regions offering more productive investments.

The rest of the article is structured as follows. Section 1 provides an overview of the literature on loan default analysis and shows how our work relates to the existing literature. Section 2 describes the data, while Section 3 illustrates our econometric approach. Section 4 discusses the results of the analysis at the regional level. Additional material available online presents a Monte Carlo exercise to investigate the properties of our estimators and the extended model comparison of various ML models. Finally, Section 5 concludes and highlights possible directions of future research.

1 Background Literature

There is a substantial empirical literature on the determinants of mortgage delinquency and default, mainly focused on the United States. Cunningham and Capone (1990) use a multinomial logistic regression model to study the default behavior of fixed-rate and adjustable-rate loans in times 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. Bajari, Chu, and Park (2008) study the relative importance of the various drivers behind subprime borrowers decision to default. They point at the role of the nationwide decrease in home prices and increase in borrowers with high payment to income ratios as main drivers of default. Campbell and Cocco (2015) propose a model of mortgage default for the United States in the presence of labor income, house price, inflation, and interest rate risk to show how different shocks contribute to the default decision. One important result from this study is that negative home equity tends to occur when house prices decline in a low-inflation environment and for moderate levels of negative home equity, default is more likely as borrowing constraints bind more tightly on households. Babii, Chen, and Ghysels (2019) analyze the spatial dependence among commercial and residential default using Generalized Autoregressive Score(GAS) models.

In contrast to the large U.S. empirical literature, studies on the drivers of mortgage arrears and default for Europe are limited and typically focused on a single country.[2] Few studies analyze the default behavior across countries in Europe, mostly due to the limited availability of multicountry microdatasets. An exception is the European Community Household Panel (ECHP) that surveys households across Europe and provides information about income and housing conditions. This dataset is used by Diaz-Serrano (2005) to

[2] See, among others, Sanchez-Martinez, Sanchez-Campillo, and Moreno-Herrero (2016) and Aller and Grant (2018) for Spain, Lambrecht, Perraudin, and Satchell (2003) and Aron and Muellbauer (2016) for the UK, and Fitzpatrick and Mues (2016) for Ireland.

examine the determinants of mortgage delinquency across 12 European countries, which is found to be positively associated to income volatility. Duygan-Bump and Grant (2009) examine the effects of adverse shocks that households experience on the likelihood to fall into arrears. They find that adverse events are important, but the extent to which they matter varies across countries and crucially depends on the penalty for defaulting. Jappelli, Pagano, and Di Maggio (2013) study differences in household indebtedness across 11 European countries, showing that higher indebtedness is associated with increased financial fragility, as measured by the sensitivity of household arrears and insolvencies to macroeconomic shocks. Gerlach-Kristen and Lyons (2015) explore the role of 'affordability' and negative equity in explaining mortgage arrears among European households (see also Georgarakos, Lojschová, and Ward-Warmedinger, 2009). Ampudia, van Vlokhoven, and Zochowski (2016) exploit data from the Household Finance and Consumption Survey (HFCS) in Europe to calculate a set of financial burden indicators for households. They calibrate their measures using country-level data on nonperforming loan ratios and estimate a set of stress-test elasticities in response to the interest rate, income, and house price shocks. Although the ECHP and the HFCS represent rich sources of information on household income, housing situation, and outstanding loans, they have the disadvantage to include a very limited number of households which could potentially affect the reliability of results. In addition, in the ECHP survey, arrears are self-reported, and hence likely to be underreported, as also pointed by Duygan-Bump and Grant (2009). On the other hand, loan-level datasets provide information on potentially millions of loans together with demographic information about the borrower that can also be used to investigate these questions.

The studies reviewed above typically employ standard econometric techniques, such as logistic regression or hazard models, to analyze the default behavior of households. Over the past few years, the increasing availability of large and complex datasets as well as the recent developments in new statistical tools has led researchers to a variety of applications of ML techniques to credit risk analysis. Desai, Crook, and Overstreet (1996) provide a first attempt to use neural network models for credit score purposes and obtain comparable results to the logistic regression. Feldman and Gross (2005) apply the Classification and Regression Tree algorithm to study a set of approximately 3000 mortgages in Israel using loan's and borrower's information. Khandani, Kim, and Lo (2010) construct a forecasting model using tree-based ML 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. In a similar study on the default behavior of 300,000 Irish mortgages, Fitzpatrick and Mues (2016) find that boosted regression tree and RF models outperform penalized logistic regression. Sirignano, Sadhwani, and Giesecke (2018) develop a deep learning model of mortgage default risk using data on 120 million U.S. mortgages over the years 1995–2014, including loan-level information, as well as macroeconomic variables at the zip-code level. Butaru et al. (2016) apply ML techniques to predict delinquency by combining information on consumer trade line, credit bureau, and macroeconomic variables for 2009–2013. They find substantial heterogeneity in risk factors, sensitivities, and predictability of delinquency across lenders, implying that no single model applies to all institutions in their data. Fuster, Goldsmith-Pinkham, and Ramadorai (2018) study millions of U.S. mortgages using the logistic and RF models. Among other variables, the authors include as regressors nonfinancial borrower's information (e.g., ethnicity) and analyze whether the application of ML techniques might favor a demographic group against the others. Albanesi and Vamossy (2019) employ a range of tools from deep learning to predict consumer default using data on 1 million U.S. household loans in the period from 2004 to 2015, showing considerable improvements in default prediction relative to traditional approaches. The range of tools available from the most recent ML literature for classification is very wide and its use in the area of credit risk analysis is still largely unexplored, with significant potentials for further study. We refer to Mullainathan and Spiess (2017) for a review of ML methods with an econometric perspective and an indication of possible applications and associated challenges if used to study economic problems, and to Carrasco and Rossi (2016), Medeiros et al. (2019), and Babii, Ghysels, and Striaukas (2020) as some recent examples of ML methods applied to classical macroeconomic forecasting. This article follows the literature on ML models for default forecasting with a novel application to a multicountry European loan-level dataset.

2 Data

European Datawarehouse (ED) is a centralized securitization repository implemented by the European Central Bank (ECB) as part of the loan-level initiative[3] that collects, validates, and distributes standardized loan-level data for several European countries. Through this vehicle, banks provide asset-backed securities as collateral in the ECB refinancing operations. The program started in January 2013 and requires financial institutions to report information on the structure and performance of their securitized loan portfolios in a detailed and standardized format. The dataset includes several loan types, ranging from residential mortgages, credit cards, car loans, and those granted to small- and medium-sized enterprises. The dataset contains (dynamic) information about the performance of each loan, updated at least on a quarterly basis. In addition, the dataset also provides (static) information recorded at the time the loan was originated, such as the loan total amount or the borrower's gross income. So far, ED data have rarely been employed for research, and have never been used for the analysis of the drivers of loan default. We refer to Ertan, Loumioti, and Wittenberg-Moerman (2017) and Van Bekkum, Gabarro, and Irani (2017) for the existing works that exploit ED data for a variety of research purposes.

In this study, we focus on loan data for residential property purchases. This is an important part of household debt, given the high concentration of resources in the housing sector observed before the onset of the financial crisis in several European countries. Data on residential loans provide information about the loan granted, the borrower, and the asset (i.e., the property) underlying each loan at the time of the origination. In particular, loan-level information includes data on the amount of debt at the origination of the loan, the interest rate, and its type (e.g., fixed rate for life, fixed with future periodic resets, or floating), the loan term, the status, and performance of the loan, whether it is performing, defaulted or in arrears, and for how many months it has been in arrears. Borrower-level information includes gross annual income and employment status (e.g., whether the borrower is self-employed or unemployed) at the date of the origination of the loan. Finally, asset-level data contain information about the value of the property underlying each loan, the

[3] See https://www.ecb.europa.eu/paym/coll/loanlevel/html/index.en.html.

type of property (e.g., house, flat, or terraced house), and the (partial) postal code where the property is located.

We collected loan-level information from ED for seven countries, namely Belgium, France, Ireland, Italy, the Netherlands, Portugal, and Spain over the period from January 1, 2013 until December 31, 2018. Although data are available also for Germany, we excluded this country from our analysis because of the low number of mortgages appearing in ED, possibly in line with the low homeownership rate registered for German households (see Voigtländer, 2009). We cleansed the data in various ways. First, we have eliminated the records with missing entries, errors, and duplicated entries. Furthermore, we only kept records for which the total balance of the loan is expressed in Euro. To associate each loan to a specific region, we have linked the postal code of each property to the corresponding geographical region, using the Eurostat Nomenclature of Territorial Units for Statistics (NUTS) classification. Our default risk analysis is carried out separately for each NUTS2 region in the sample. Accordingly, we dropped records for loans that refer to properties located in NUTS2 regions with less than 100 observations or located in overseas territories. We also dropped records in the first and last percentiles of the distribution for all quantitative variable, where percentiles have been calculated separately for each NUTS2 in the sample. In our empirical study, we define a loan as defaulted if it is in arrears for more than 90 consecutive days. One reason for defining default in this way is that it provides a uniform definition of default across Europe, given that the legal definition of default may vary across countries, depending on country-specific regulations. When a loan is classified as defaulted according to our criterion, we remove all updates of the loan status that follows the date of default, thus excluding the possibility that the defaulted loan returns to a performing status. After this cleaning procedure, we obtain a dataset consisting of 162 million observations, with approximately 12 million unique mortgages observed over a time and in 96 NUTS2 regions. The numerosity of the dataset varies largely across NUTS2 regions, ranging from the approximately 20,000 records of the Spanish Ciudad Autónoma de Melilla region to the 4 million records of the French Île-de-France region.[4]

Table 1 reports the number of records by country, the number of unique loans, the default rate, and the percentage of unique loans divided by the 2018 country population obtained from Eurostat. One important remark is that the distribution of loans 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. The default rate also varies substantially across countries, ranging from 0.34% in France to 14.70% for Ireland. Most importantly, the default rate is very low for most countries, indicating the highly unbalanced nature of the default variable.

Table 2 shows loan characteristics at the country level in order to evaluate differences in the residential loan markets across Europe. The median loan maturity at origination is between 18 and 25 years for Belgium, France, and Italy, while the remaining countries have a median term of 30 years or longer as in the case of Portugal. There are also significant differences in interest rate type across countries: in Belgium, France, and the Netherlands mortgages are predominantly of the fixed interest rate type, while in the remaining countries the floating rate prevails. The Netherlands is unique among European countries by

[4] We refer to the Supplementary Appendix for additional data cleaning operations that we performed on ED data.

Table 1 ED loans by country: number of records, number of unique loans, default rate, and percentage of unique loans over the country population in 2018

Country	No. observations	No. of loans	Default rate (%)	Loans by population (%)
Belgium	16,707,359	1,168,744	0.76	10.25
France	50,703,745	4,366,815	0.34	6.49
Ireland	8,208,587	386,535	14.70	7.95
Italy	17,240,131	1,253,678	6.61	2.07
Portugal	55,22,206	342,684	6.62	3.33
Spain	22,065,010	1,871,179	3.15	4.00
Netherlands	26,923,998	2,819,810	0.93	16.36

Notes: The default rate is computed as the number of defaulted loans over the number of distinct loans in the sample.

Table 2 Loan characteristics by country

Country	Term	Interest Rate Type		Interest	LTV	DTI
		Floating	Fixed	only		
Belgium	20.00	0.00	100.00	0.20	80.00	2.24
France	18.00	2.80	97.20	0.10	76.78	2.91
Ireland	30.00	95.00	5.00	6.10	80.00	4.63
Italy	20.08	73.10	26.90	0.00	68.18	4.20
Portugal	34.33	99.30	0.70	0.50	83.33	4.14
Spain	30.00	98.10	1.90	0.00	79.21	5.00
Netherlands	30.00	5.80	94.20	82.70	87.74	2.08

Notes: Summary statistics by country of the median loan term in years (*Term*), the percentage of mortgages that are of the *Floating* and *Fixed* interest rate type, the percentage of loans whose repayment is interest only, and the median LTV and DTI.

having a large percentage of loans that are only interest mortgages. In terms of median LTV ratio, most countries are around 80%, with the highest ratio being the Netherlands with 87.7% and the lowest Italy with 68.18%. Finally, in terms of Debt-to-Income (DTI), the highest leverage is in Spain where households borrow five times their income and the lowest in the Netherlands where the ratio is only 2.08.

We augment the ED dataset with a number of variables that we expect to be relevant in determining loan default. As pointed out by several studies, local economic conditions are likely to play an important role in explaining the differences in credit risk (see, among others, Djeundje and Crook, 2018; Dirick et al., 2019). Accordingly, we integrate ED data with a set of aggregate, regional macroeconomic variables as a proxy for the state of the economy of a region. These include the NUTS2-level change in unemployment rate and gross domestic product (GDP) growth obtained from Eurostat. Furthermore, we calculated the observed NUTS3-level default rate and the average NUTS3-level house price growth, computed using ED data. All variables related to the local economic conditions have been lagged 1 year. This is the shortest possible lag to include in the analysis given that the regional macroeconomic variables provided by Eurostat are only available at the yearly frequency.[5]

Table 3 provides the definitions of the variables included as regressors in our models of loan default. We split the variables into three sets, namely, loan-specific, borrower-specific, and aggregated regional variables. The choice of these variable is in line with the existing literature on loan default prediction, and we refer to Cunningham and Capone (1990), Deng (1997), Campbell and Cocco (2015), and Sirignano, Sadhwani, and Giesecke (2018), among others, where similar features are studied. Among the *loan-specific* characteristics, we include two categorical variables (namely, interest rate type and property type) and two numerical variables observed at the origination of the loan, that is valuation amount (in logs) and the DTI ratio, calculated as the ratio between the total amount of loan and the gross income of the borrower. In addition, we also included three numeric variables that are observed every time the loan information is updated in the dataset, such as the seniority of the loan, computed as number of days since the loan was granted, the current LTV and the current interest rate applied to the loan. With respect to *borrower-specific* variables, we include the borrower gross income and the employment status, where both features are observed only at the time of the origination of the loan.

In our default risk model, we have also included a number of variables calculated at the level of the financial institution issuing the loans. Considering only the loans originated in the seven countries under study, in the ED dataset, there are 601 unique institutions, issuing at least 1000 loans. We expect substantial heterogeneity in the outcome variable and regressors across institutions, due to differences in lending policies, risk perceptions, and risk tolerance. Adding originator fixed effects to our default risk model is likely to create overfitting problems. Rather than controlling for bank-specific effects, we have decided to compute the average LTV ratio, average DTI ratio, and the average balance change of the loan originator in the previous year, and add these variables to our default risk model as additional controllers. These aggregates represent a proxy for the policy adopted by a financial institution at a certain point in time with respect to residential mortgages, and controlling for them allows us to achieve a more accurate description of the borrower's default decision, regardless of who originated the loan.

3 Methods

Assume we observe y_{it}, for $i = 1, \dots, N$ and $t = 1, \dots, T$, a categorical variable associated to the ith loan in quarter t, with $y_{it} = 1$ if default occurs and $y_{it} = 0$ otherwise. In addition, let $\boldsymbol{x}_{it}$ be a K-dimensional vector of covariates associated to the ith loan at time t. Given that in our empirical application both the number of loans, N, and the number of covariates, K, are large, we have decided to adopt a range of methods that are suitable for dealing with high-dimensional, big data. Specifically, we consider three alternative approaches for modeling the risk of default as a function of a set of covariates, namely penalized logistic

[5] As an alternative, we could add in the regression the local economic variables expressed in levels rather than first differences. However, the inclusion of variables, such as GDP and house price would complicate the estimation of the classical logistic regression as the maximum likelihood estimator would not have the standard asymptotic properties in the presence of nonstationary variables (see, e.g., Park and Phillips, 2000).

Table 3 Explanatory variables used to predict the occurrence of a default

Feature	Attribute	Type	Description
Loan-specific variables			
DTI	Static	Numeric	DTI at origination
Interest rate type	Static	Categorical	Interest rate type
Interest rate	Dynamic	Numeric	Interest rate at the pool cutoff date
LTV	Dynamic	Numeric	LTV at pool cutoff-date
Property type	Static	Categorical	Property type of the underlying asset
Seniority	Dynamic	Numeric	Loan seniority at origination (in days)
Valuation amount	Static	Numeric	Property value as at loan origination (in logs)
Borrower-specific variables			
Borrower's employment	Static	Categorical	Employment status of the applicant at origination
Income	Static	Numeric	Borrower gross annual income at origination (in logs)
Regional-specific variables			
Default rate	Dynamic	Numeric	Default rate (%) by NUTS3 lagged 1 year
GDP growth	Dynamic	Numeric	GDP percentage growth by NUTS2 lagged 1 year
House Price growth	Dynamic	Numeric	House price percentage growth by NUTS3 lagged 1 year
Unemployment rate growth	Dynamic	Numeric	Unemployment rate growth by NUTS2 lagged 1 year

regression, Gradient Boosting (GB) and Extreme Gradient Boosting (XGB). In the rest of this section, we first briefly describe these methods[6] and then discuss the range of tools used to assess the performance of our models and interpret results.

3.1 Models

3.1.1 Penalized logistic regression.

Logistic regression is one of the most popular approaches when modeling the probability of loan default as a function of a set of covariates. Let $P(y_{it} = 1|\boldsymbol{x}_{it})$ be the conditional probability of default given the covariates, $\boldsymbol{x}_{it}$. The logistic model assumes that the log of the odds of the conditional probability $P(y_{it} = 1|\boldsymbol{x}_{it})$ is a linear combination of $\boldsymbol{x}_{it}$:

$$\ln\left(\frac{P(y_{it} = 1|\boldsymbol{x}_{it})}{1 - P(y_{it} = 1|\boldsymbol{x}_{it})}\right) = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta}, \tag{1}$$

where β_0 is the intercept and $\boldsymbol{\beta}$ is a K-dimensional vector of coefficients associated to the covariates, $\boldsymbol{x}_i$. Estimates of β_0 and $\boldsymbol{\beta}$ in Equation (1) are traditionally obtained by minimizing the negative log-likelihood. If the number of unknown parameters, $K+1$, is large

[6] We estimated all models using the h2o platform; details about the model implementation are available at http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html. The Supplementary Appendix provides additional details regarding the parameter selection.

relative to the number of observations, N, then solving the standard likelihood problem might lead to low estimation accuracy or even biased results. To avoid these problems, one solution is to add a penalization term to the objective function that shrinks or sets some coefficients to zero. In our application, we adopt the elastic-net penalization that solves the following minimization problem:

$$\min_{\beta_0,\boldsymbol{\beta}} \left\{ -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}[y_{it}(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta}) - \log\left(1 + e^{\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta}}\right)] + \lambda\sum_{k=1}^{K}[\alpha|\beta_k| + \frac{1}{2}(1-\alpha)\beta_k^2] \right\}, \quad (2)$$

where $\lambda \geq 0$ is a regularization parameter and $0 \leq \alpha \leq 1$ is the elastic-net parameter, mixing between ridge ($\alpha = 0$) and lasso ($\alpha = 1$), and allowing for variable selection and shrinkage (Zou and Hastie, 2005). The overall impact of the penalty in Equation (2) is controlled by λ: the larger λ, the stronger the variable selection or shrinkage imposed by the elastic net. In our application, we select α with a grid search over the values $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$: in the majority of the cases, the best model coincides with $\alpha = 0$ which corresponds to a nonsparse solution. We select λ with a search on 100 values similarly to Friedman, Hastie, and Tibshirani (2010).

3.1.2 Gradient tree boosting.

Proposed by the ML literature, GB represents a powerful tool for classification and regression analysis. This method consists of *sequentially* applying a weak classification algorithm to adaptively reweighted versions of the initial training data. The weighting scheme is adaptive in the sense that those observations misclassified by the classifier in the previous step are then given a higher weight, whereas observations that were correctly classified are assigned a lower weight. Thus, as iterations proceed, the boosting classification algorithm focuses more on observations that are difficult to classify than on those that are already classified correctly. In the context of binary classification, boosting can be interpreted as an approximation of additive logistic models built on a Bernoulli likelihood. Several versions of boosting exist, we refer to Friedman, Hastie, and Tibshirani (2000) for a review of various specifications. In our empirical problem, we adopt a GB procedure with decision trees as weak classifiers, following the approach outlined in Friedman (2001). One reason for using decision trees in our empirical context is that, by construction, they allow for interaction among explanatory variables (see Schiltz et al., 2018 for a discussion on this). A decision tree splits the (multidimensional) space of the explanatory variables into a set of nonoverlapping regions. Specifically, a tree can be expressed as follows:

$$T(\mathbf{x}_{it}, \boldsymbol{\Theta}) = \sum_{j=1}^{J}\gamma_j 1\left(\mathbf{x}_{it} \in R_j\right), \quad (3)$$

where $R_1, \ldots, R_J$ represent a partition of the covariates space, $1(\cdot)$ is an indicator function, $\Theta = \left(\gamma_j, R_j\right)_{j=1}^{J}$ are the unknown parameters. Hence, at each interaction $m = 1, \ldots, M$, the boosting algorithm applied to decision trees estimates the unknown parameters, $\boldsymbol{\Theta}$, by solving the following equation:

$$\min_{\boldsymbol{\Theta}_m} \sum_{i=1}^{N} \sum_{t=1}^{T} L\big(y_{it}, f_{m-1}(\mathbf{x}_{it}) + T(\mathbf{x}_{it}, \boldsymbol{\Theta}_m)\big), \tag{4}$$

where $L(\cdot)$ is a loss function and $f_m(\mathbf{x}_{it})$ is the so-called boosted tree model, or tree ensemble:

$$f_m(\mathbf{x}_{it}) = \sum_{\ell=1}^{m} T(\mathbf{x}_{it}, \Theta_\ell). \tag{5}$$

The minimization Equation (4) with boosted tree model in Equation (5) can be solved by fast approximate algorithms. In particular, we apply the so-called GB procedure that achieves minimization by calculating recursively the negative gradient of the loss function, in a "steepest-descent" type algorithm. In our application, we adopt the difference between the outcome and the estimated probability of default as loss function and the Area Under the Curve as stopping criterion. The implementation of the gradient tree boosting algorithm presented above requires fixing a set of meta-parameters relative to the number of regions, J, in Equation (3), and the number of trees, M, in Equation (4). The selection procedure for such meta-parameters is briefly discussed in the Supplementary Appendix. We also refer to Hastie, Tibshirani, and Friedman (2009) for a detailed presentation of GB.

3.1.3 Extreme GB.

Our third specification is the XGB model proposed by Chen and Guestrin (2016). Extreme GB has gained great popularity in the recent ML literature, proving to be one of the most accurate algorithms for classification and forecasting purposes.

Chen and Guestrin (2016) extend the GB model in Equation (4) to include a regularization term:

$$\min_{\boldsymbol{\theta}_m} \sum_{i=1}^{N} \sum_{t=1}^{T} l\big(y_{it}, f_{m-1}(\mathbf{x}_{it}) + T(\mathbf{x}_{it}, \boldsymbol{\theta}_m)\big) + \left(\delta J + \frac{1}{2}\eta \sum_{j=1}^{J} {\gamma_j}^2\right), \tag{6}$$

where $\delta > 0$ is a l_1-penalty on the number of leaves of a tree J as in Equation (3), and $\eta > 0$ is a l_2-penalty on the leaf weights γ_j, for $j = 1, \ldots, J$. In addition, their algorithm departs from an exact greedy algorithm for tree learning in favor of an approximate algorithm that has the advantage of being scalable with big data, as it is the case in this application. The approximate algorithm has two main characteristics. The first is that only percentiles of the feature variables are considered as splitting points, rather than considering all possible values as in the exact case. The second feature of the XGB model is that it handles missing values by assigning them to a tree node, with a data-driven choice of the direction. A detailed discussion of the algorithm and of the regularization term in Equation (6) is provided in Chen and Guestrin (2016).

3.2 Model Performance

We compute a number of out-of-sample metrics in order to compare the performance of the above algorithms. First, we calculate the Receiver Operating Characteristic (ROC) and corresponding Area Under the ROC Curve (AUC), which are the standard tools for measuring binary classification performance.[7] For a given classifier, the AUC values range between 0.5, for a classifier performing as well as a random guess, and 1.0 for the perfect classifier. Although the AUC is widely popular among practitioners for its straightforward interpretation, Hand (2009) states that it has the disadvantage of evaluating a classifier using a weight function that depends on the classifier itself, thus providing an incoherent measure of comparison. Accordingly, Hand (2009) proposes an alternative measure to the AUC, the *H-measure*, which uses a misclassification cost that is fixed and independent of the classifier. Given the highly unbalanced nature of our data, we follow Hand and Anagnostopoulos (2014) and in computing the H-measure we assign a cost to misclassifying a defaulted loan larger than the cost of incorrectly classifying a nondefaulted loan. This is achieved by setting the severity ratio (i.e., the severity of misclassifying a nondefaulted loan relative to the cost of misclassifying a defaulted loan) equal to the reciprocal of the relative class frequencies. Finally, for each algorithm, we compute two additional measures of goodness-of-fit alternative to the AUC and the H-measure. The first is the Brier's score (BS) by Brier (1950) which corresponds to the average squared error of predicted probabilities and observed values, $\mathrm{BS} = 1/N \sum_{i=1}^{N} (y_i - p_i)^2$, where y_i is the observed default occurrence and p_i is the estimated probability of default for the ith loan. The BS ranges between 0 and 1, with smaller values pointing at better classifier performance. The second measure is the Logarithmic score (LS) defined as follows: $\mathrm{LS} = \sum_{i=1}^{N} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$. Larger values of the LS statistic indicate better predictive performance of the classifier.

Given that we estimate the models for each NUTS2 region in our sample, we can exploit our multiple region-specific estimates to perform a statistical comparison between the out-of-sample performance of the different algorithms. To this end, we first apply the Friedman and the Iman–Davenport test statistics that evaluate whether there exists any significant difference in the ranking obtained by comparing a performance metric across models (Friedman, 1940; Iman and Davenport, 1980). If the null hypothesis of equivalence of the rankings of a performance metric across algorithms is rejected in the above tests, then it is possible to compute a pairwise comparison of the ranking of the two algorithms using a post hoc test, as outlined in Garcia and Herrera (2008). The above tests carry out a statistical comparison of overall performance measured across different methods, but do not allow to compare measures calculated at loan-level, such as the BS and LS. Hence, for testing the difference between any two scores, we apply the Diebold and Mariano (1995) (DM) statistic. Specifically, we apply the Lahiri and Yang (2013) modification of the DM statistic to allow the comparison of scoring measures associated to binary outcomes. The fact that we estimate a model for each NUTS2 regions separately allows us to partly account for the heterogeneity present in the data. We refer to Timmermann and Zhu (2019) for a recent discussion on how to account for unobserved heterogeneity using a common factor setup while testing for equal predictive accuracy.

3.3 Interpretable Machine Learning

ML methods have been successful in learning complex relations and predicting future realizations in several applications. However, the difficulty in interpreting what a model has

[7] Despite being a powerful tool to compare model performance, we do not report the ROC curves as we would have to report one plot for each NUTS2 region. We rely on more compact ways to plot the results, which are in line with the ROC curves.

learned has often undermined the credibility of the analysis. The recent work on IML, or Explainable Artificial Intelligence, is an attempt to provide a set of tools to interpret the outcome of ML algorithms and identify the nonlinear interactions present in the data. We refer to Ribeiro, Singh, and Guestrin (2016), Murdoch et al. (2019), and Zhao and Hastie (2019) among others, for more details on IML. One important tool for interpreting results is the calculation of the *variable importance* (for variable importance in boosting models, see Hastie, Tibshirani, and Friedman, 2009, p. 356). In the case of GB and XGB, variable importance is calculated as the reduction in the squared prediction error as a result of the split in the tree and attributed to the splitting variable. We repeat this operation across all nodes in the tree and sum it across the M trees. The resulting quantity is then standardized on a percentage scale and can be interpreted as the relative importance of a variable in predicting default.

To explore the relation between the dependent variable and the explanatory variables, we also produce the so-called Accumulated Local Effect (ALE) plots (Apley, 2016). The ALE averages the prediction changes calculated on small partitions of the variable of interest and then accumulates them over all partitions. If the function is differential, the local effect is computed as the first derivative with respect to the variable of interest over the partition. The uncentered ALE of variable j is computed as follows:

$$\hat{f}^{*}_{j,\mathrm{ALE}}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_{i,j} \in N_j(k)} \left[f(z_{k,j}, x_{i,\, j}) - f(z_{k-1,j}, x_{i,\, j}) \right] \tag{7}$$

for each $x \in (z_{0,j}, z_{K,j})$, where $n_j(k)$ is the number of training observations in the kth partition $N_j(k)$, such that $\sum_{k=1}^{K} n_j(k) = N$, for $k = 1, 2, \ldots, K$, given a fixed number of partitions K. The term within square brackets is the change in predicted output in the kth partition, where all the variables other than the variable of interest $x_{i,\, j}$ are set to their sample average within the kth partition. Equation (7) is then centered such that the centered ALE estimator has a zero mean effect:

$$\hat{f}_{\mathrm{j,ALE}}(x) = \hat{f}^{*}_{\mathrm{j,ALE}}(x)) - \frac{1}{n} \sum_{i=1}^{N} \hat{f}^{*}_{\mathrm{j,ALE}}(x_{i,j})). \tag{8}$$

ALE plots, obtained from Equation (8), serve as alternative to marginal or partial dependence plots (Friedman, 2001), with the advantage of achieving a reliable representation of the effect even in the presence of correlated explanatory variables. An advantage of ALE in the context of a big data application is that they are computationally less intensive related to alternative methods, such as the SHapley Additive exPlanation method proposed by Lundberg and Lee (2017). We refer to Apley (2016) for a detailed presentation of ALE plots.

4 Results

4.1 Exploratory Data Analysis

Figure 1 displays a map of the default rate by NUTS2. It is interesting to observe that regions belonging to the same country have similar default rates, indicating the presence of strong country-level differences, with Portugal, Spain, Italy, and Ireland showing the highest default rates. Significant intracountry heterogeneity is also present. For example, regions

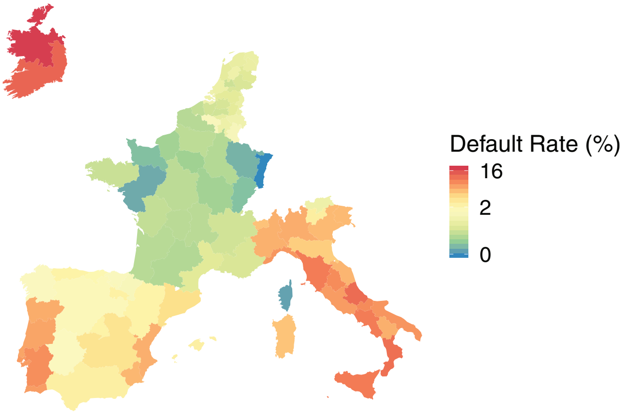

Figure 1 Default rate by NUTS2 (colors in logarithmic scale).

from the East coast of Spain show higher default rates relative to other areas; the default rate for the Valencian Community is 5.61%, 246 basis point higher than that of Northern regions. Similarly, regions in the Center and South of Italy show higher default rates than regions from the North. The range of intracountry variation can be very large for certain countries; for example, in Italy, the default rate ranges from 0.92% in South Tyrol to 11.42% in Abruzzo.

Figure 2 reports the median current interest rate and median current LTV over time: each line represents a NUTS2 region with the color varying depending on the associated country. The interest rate has steadily decreased in all countries, whereas LTV has remained stable, except for the Netherlands where it dropped by approximately 20 percentage points. Both variables show strong across country heterogeneity: the median interest rates are clearly clustered by country, indicating the presence of moderate within-country variability. On the contrary, the LTV observations are much more disperse and suggest that the relative size of the loan can vary largely across regions, especially for Spain and Portugal.

Figure 3 shows the plot of default rate calculated in the years 2013–2018 against the 2008–2013 percentage change in GDP and unemployment, at regional level. It is interesting to observe that relatively higher default rates are associated to negative GDP growth and high unemployment growth. We also note that Portugal, Spain, Italy, and Ireland, having the highest default rates, are also the countries that suffered the sharpest decline in output and employment immediately after the onset of the 2008 Great Recession, indicating that these countries are still experiencing the negative consequences of the economic crisis.

4.2 Model Comparison

In this section, we use the penalized logistic regression, GB and XGB to analyze loan-level data from ED. For each NUTS2 region, we partition our data into training, validation, and test sets, and 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 in order to account for

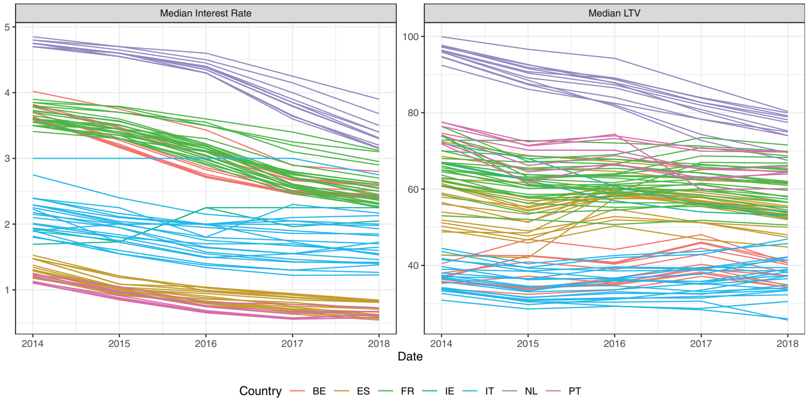

Figure 2 Median interest rate (a) and LTV (b) by NUTS2 regions (colors by country).

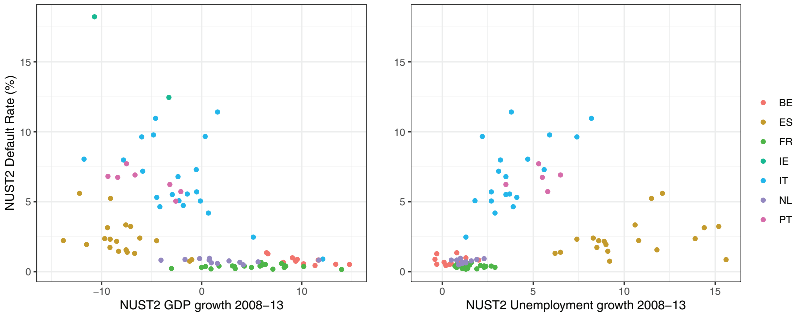

Figure 3 Default rate by NUTS2 against GDP (a) and unemployment (b) growth over 2008–2013 (colors by country).

homogeneous proportions of defaulted loans in the three sets. As reported in Table 1, default occurrences are rare and the dataset results highly unbalanced. To accurately account for the probability of the rare default event, we balance the classes either under-sampling the most frequent classes or oversampling the less frequent ones.[8] The final probabilities are adjusted to the original sample via a monotonic transformation, which does not alter ordering and thus not affecting AUC metrics (more details on class balancing by King and Zeng, 2001). We report the details of the parameter tuning for each model in the Supplementary Appendix.[9] Figure 4 reports the box-plot of the AUC and *H*-measure for the three models by country. Both AUC and *H*-measure indicate that boosting models

[8] A balancing alternative to standard under/over-sampling is the Synthetic Minority Over-sampling Technique proposed by Chawla et al. (2002).

[9] Approximately 9000 models were trained to obtain the results presented in this section. Further details on the models and the parameter tuning are provided in the Supplementary Appendix available on the corresponding author's personal website.

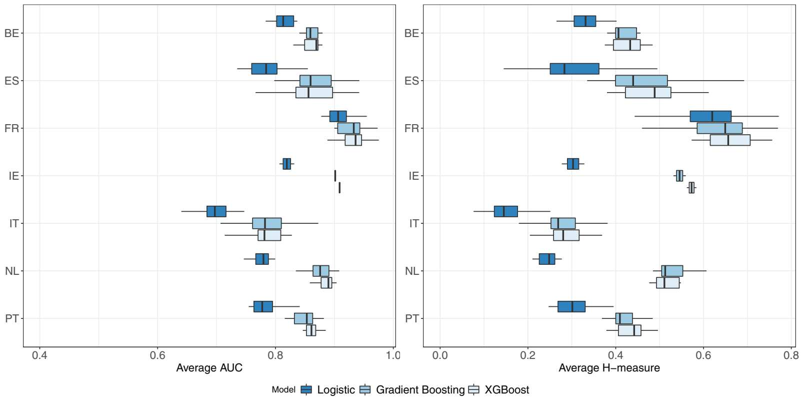

Figure 4 Box plots of the AUC (a) and *H*-measure (b) by country and model.

perform better than the logistic for all countries. When comparing XGB against GB, the *H*-measure is slightly in favor of the XGB.

To better understand for which regions boosting models perform better relative to the conventional logistic regression, Figure 5 reports a map with the *H*-measure of the GB and XGB relative to the *H*-measure of the logistic regression by NUTS2 regions: in blue are denoted the regions where the boosting model outperforms the logistic regression, in gray (ratio of one) the ones where the models achieve the same performance, while in red the regions where the logistic regression performs best. The darker the color the larger the difference relative to the logistic. Overall, we observe a strong prevalence of values larger than one, indicating a better performance of boosting models. GB and XGB perform very similarly, with slightly higher values of the *H*-measure for the latter. The maps show strong country-level effects and clear within-country heterogeneity, suggesting that the model performance can vary largely both across and within each country. Estimating a model for each NUTS2 region allows to capture such regional differences that would not otherwise be observed in country-level models. Ireland, Italy, Spain, Portugal, and the Netherlands present strong evidence of the better performance of the boosting models over the logistic: the best performance is obtained in two Italian regions, Aosta Valley and Sicily, where the boosting models outperform the logistic benchmark by a ratio of 3. In Belgium and France, the difference with the logistic regression is less evident although XGB has better performance in the large majority of Belgian and French regions.

The *H*-measure can be used in statistical testing for the difference in the out-of-sample performance of the three models. Specifically, we rank the three models on the basis of the *H*-measure in each NUTS2 region and obtain the average ranking of each model. Hence, the difference in ranking of the three models can be tested overall by using the Friedman and Iman–Davenport tests, or pairwise using the post hoc test (García et al., 2010). The first column of Table 4 reports the average ranking of the three models across regions, and points at XGB as attaining the best performance. Both the Friedman and Iman–Davenport test statistics reject the null hypothesis of equivalence between the three rankings at the 5% level. Interestingly, the post hoc test indicates that while both boosting models outperform the logistic, there is no significant difference between the rankings of GB and XGB.

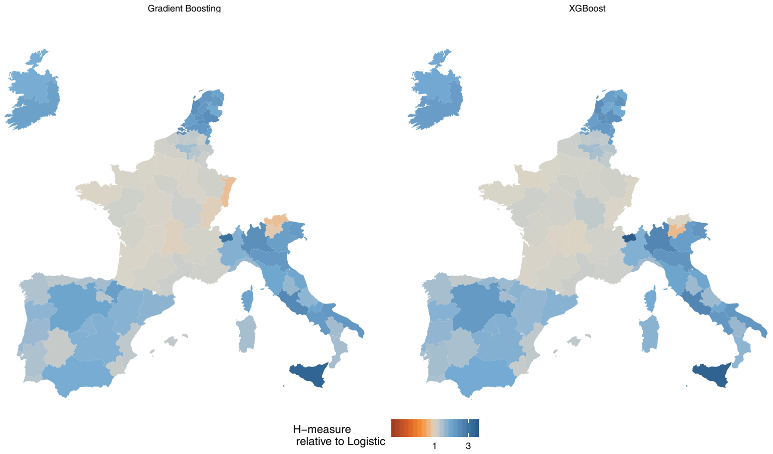

Figure 5 Performance metrics: *H*-measure of GB (a) and XGB (b) related to logistic regression by NUTS2 regions. A value of one (in gray) indicates equivalent performance of the models, values larger than one (blue scale) indicate that boosting models outperform the logistic, while values smaller than one (red scale) indicate that the logistic outperforms boosting models. The darker the color, the larger is the distance from one.

Table 4 Comparison of classifiers using Friedman and Iman–Davenport tests on the *H*-measure rankings under the null hypothesis of ranking equivalence (right) and *p*-values from post hoc test (left)

Model	Average ranking	Pair-wise post hoc test			Friedman and Iman–Davenport tests		
		Logistic	GB	XGB	Test statistics	Value	*p*-value
Logistic	2.78	·	0.00	0.00	Friedman	96.646	0.00
GB	1.82	0.00	·	0.54	Iman–Davenport	96.287	0.00
XGB	1.40	0.00	0.54	·			

The BS and LS can be used to compare the out-of-sample forecasting performance based on the Lahiri and Yang (2013) extension of the Diebold and Mariano (1995) test described in Section 3. Table 5 reports the frequency of rejection of the null hypothesis of equal forecast accuracy of the models when they are compared pairwise. The confidence level is 5% and the *p*-values are corrected with the Benjamini and Hochberg (1995) correction for multiple testing. Boosting models perform significantly better than the logistic in the majority of the cases, with the XGB outperforming the logistic in 90% of the regions based on the LS. Although the BS does not show strong differences between the GB and the XGB, with the latter being preferred to the former only in 59% of the cases, the LS tends to favor the XGB, which is the preferred model in 90% of the cases.

4.3 Extended Model Comparison

The ML literature includes a wide range of models that could potentially provide higher prediction accuracy relative to the tree-based models considered so far (see D'Hondt et al.,

Table 5 Percentage of rejections of the null hypothesis of equivalence in forecast performance against the alternative that model A outperforms model B based on the BS and LS

Score	GB $\succ$ LG	XGB $\succ$ LG	XGB $\succ$ GB
BS	63.54%	68.75%	59.38%
LC	63.54%	90.62%	89.58%

The models A and B are the Penalized Logistic (LG), GB, and XGB. The confidence level is 5% and the *p*-values are corrected with the Benjamini and Hochberg (1995) correction for multiple testing.

2019 and Gu, Kelly, and Xiu, 2020 among others). To evaluate the robustness of our results, we consider three additional ML models, namely, a Naïve Bayes (NB) classifier, a RF, and a NN (we refer to Hastie, Tibshirani, and Friedman, 2009 for a detailed discussion of the models). In addition, we also extend the Logistic model to include nonlinearities by adding the squares of the numerical variables as additional regressors. Our comparison thus includes seven alternative forecasting models: XGB, GB, RF, Penalized Logistic with Power Series (LG-PS), Penalized Logistic (LG), NN, and NB. Although the set of models considered is not exhaustive, it is representative of various ML techniques that have been successfully applied in economics and finance.

Even in this extended comparison, XGB and GB achieve the best performance regardless of the metric considered, and are closely followed by RF. Incorporating nonlinearities to the Logistic model allows the LG-PS to outperform the linear LG, suggesting that capturing nonlinear effects is a key element to achieve higher forecast accuracy. Besides, the performances of NN and NB are quite poor, with the former being associated to a large variability of its performance across regions. Using the Friedman and Iman–Davenport test, we reject the null hypothesis of equivalence of the ranking among all models at the 5% level. Based on a pairwise comparison between model rankings at the 5% significance level, XGB outperforms all other models, except for GB. Although we notice the added value of power series to capture nonlinearities, there is no significant difference among LG and LG-PS rankings. Overall, the analysis of a larger sets of algorithms confirms the results already presented above. First, boosting models, in particular XGB, outperform the other methods in terms of accuracy of the default forecast. Second, the inclusion of nonlinear effects delivers a more accurate forecast. Figures and tables about the extended comparison are available in the Supplementary Appendix. In the next section, we focus on the two best performing models (i.e., GB and XGB) to investigate the most important drivers at explaining the occurrence of a default.

4.4 Default Drivers

The previous results indicate that the ML models have a significant advantage over the logistic regression in out-of-sample forecasts. It is now interesting to understand which factors are driving the performance of these models and examine existing differences across countries. Figure 6 reports the box plot of the variable importance across regions for the two estimated boosting models. As expected, the variable importance for the models is quite similar for the two models, with a correlation coefficient equal to 0.75. Overall, our results point at loan-related variables, namely, current LTV and current interest rate, as the most relevant in predicting the occurrence of a default: both variables have been identified

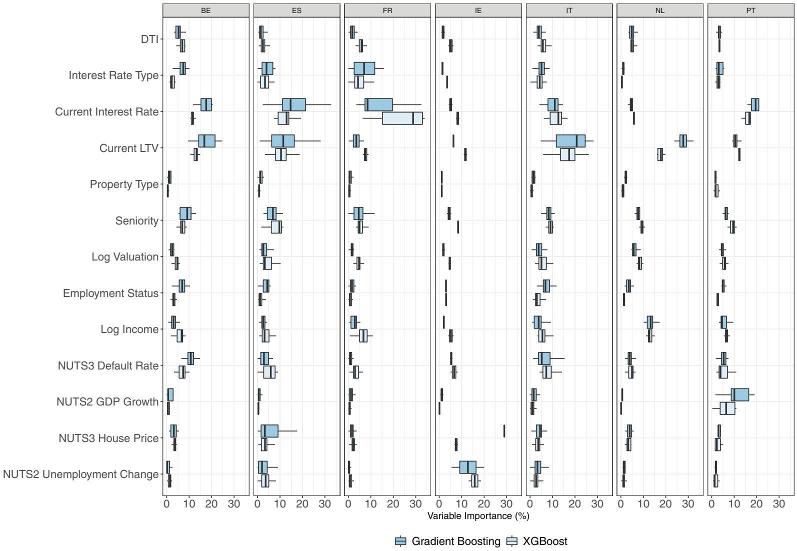

Figure 6 Box plot of the variable importance by country, estimated by GB and XGB.

in the literature as close proxies to the loan's riskiness and their movement can be linked to the default decision (e.g., Elul et al., 2010; Fitzpatrick and Mues, 2016). In addition, the importance of the interest rate seems greater in countries such as Spain, France, Italy, and Portugal where there is a prevalence of floating rate mortgages as shown in Table 2. This result is consistent with the findings of Fuster and Willen (2017) of a relationship between the changes in the interest rate for adjustable-rate mortgages and default rates in the United States.

The geographical differences in variable importance can be better appreciated by looking at Figure 9, which shows the most important variable and group for each region, where groups are defined in Table 3. Looking at the most important variable group, we observe the large prevalence of loan-related variables for all countries. Among these, the bottom graph indicates current interest rate as the most important determinant of default for most regions. Regional macrovariables are also important in Ireland, the Spanish Mediterranean coastline, and the central part of Italy, which are the regions that were particularly affected by the 2008 economic and financial crisis.

Variables containing borrower-related information or loan's characteristics other than current LTV and interest rate are rarely selected among the most relevant ones. For instance, the income of the primary borrower is among the less important variables, except for the Netherlands, where is the second most important regressor in predicting default. Similar conclusions apply to the amount of credit granted or to the borrower's employment status. Although this might seem counterintuitive, notice that the loan- and borrower-specific variables, other than current LTV and interest rate, are measured at the time of the origination of the loan. Looking at the importance of regional variables, while heterogeneous across and within countries, we observe that local economic conditions do matter in

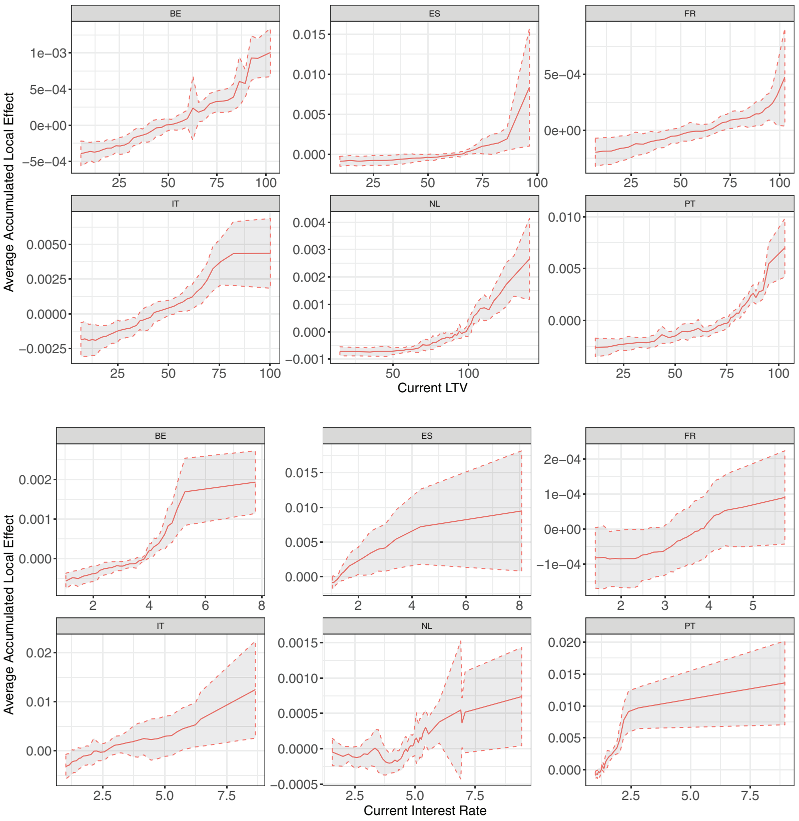

Figure 7 ALE for current LTV (a) and interest rate (b) by country for the XGB estimates. The bands represent the standard deviation of the ALE estimates in the NUTS2 regions of a country.

countries, such as Spain, Ireland, Portugal, and Italy. The box plots reported in Figure 6 provide also an idea of the variability of the variable importance within a country: a high standard deviation for a given country indicates that the variable importance is not homogeneous, but rather varies largely across regions. We observe a within country standard deviation particularly high for many variables in Spain, France, Italy, and Portugal, thus supporting the choice of estimating the model at the regional level rather than producing one single model for each country.

Figure 7 reports the ALE plots for the two most relevant variables, namely, the current LTV (top panel) and the current interest rate (bottom panel). The *y*-axis reports the ALE evaluated at a specific value of the regressor, appearing on the *x*-axis, and keeping all other variables constant. As detailed in Section 3.3, the ALE plots are centered, meaning that a positive value of ALE indicates a rise in the probability of default, and vice versa. The solid line is the average ALE by country and the confidence bands are obtained as the standard deviation from the mean of the ALE estimates in the NUTS2 region.[10] Both current LTV

[10] We exclude Ireland from the ALE plots, since it would consist of only two NUTS2 regions.

and interest rate are positively related to the ALE, indicating that higher values of LTV or interest rate are associated with a higher probability of default, but not always approximately linear as we would assume in a logistic regression model. For instance, the ALE of the LTV in Spain, the Netherlands, and Portugal is quite flat for low LTV values, while it spikes up for LTV values larger than 75. Similarly, the ALE of the interest rate in Belgium and France starts quite flat, grows rapidly for intermediate values of interest rate, and then flattens again. Comparable nonlinear effect have been found in the literature (e.g., Fitzpatrick and Mues, 2016; Sirignano, Sadhwani, and Giesecke, 2018), and support the use of ML methods rather than standard linear approximations. The ALE plots allow us to explore the across-country differences of the effect of one variable on the probability of default. Focusing on the current LTV, we see that in Italy the ALE is positive, thus indicating a rise in the probability of default, for values of LTV larger than 40%, whereas in other countries, the ALE becomes positive only for higher LTV values. For the interest rate, we observe that the ALE of Spain, Italy, and Portugal is positive for values of the interest rate larger than 1%, while in Belgium, France, and the Netherlands it becomes positive only for an interest rate larger than 4%.

We now turn to investigate more in details the role of local economic conditions on loan default. This is very important, given that region-specific variables might be seen as proxies for shocks in local economic conditions that are latent in our dataset. Figure 8 shows the average variable importance of all region-specific variables against the regional GDP growth observed precrisis (2004–2008) and postcrisis periods (2008–2013). One interesting finding is that regions that experienced a sharp drop in output after the onset of the Great Recession are those for which the local economic conditions are the most important in explaining default. This result points at income shocks as a possible explanation of default, where borrowers, for example, lose their jobs and are not able to repay their mortgage anymore. This is somewhat expected, given the nonrecourse nature of mortgages in European countries that reduce the strategic motive for defaulting. Interestingly, Gerardi et al. (2018) find for U.S. data that the "can't pay" motive plays a larger role relative to "won't pay." Our results seem to support these results also for European countries, where income shocks occurring during the Great Recession are likely to have made mortgages not affordable in several regions.

Figure 10 reports the scatter plots of the current LTV and interest rate against the associated ALE, with the color of the dots representing the NUTS2 GDP (left panel) and unemployment (right panel) growth in the period 2008–2013. Pooling all countries together, we observe that the degree of nonlinearity of current LTV and interest rate, already noted in Figure 7, varies largely across NUTS regions and strongly interacts with local economic variables. Regions that suffered a deeper downturn of GDP after the 2008 financial crisis, present a pronounced nonlinear effect with a sharp rise in ALE for high values of current LTV and interest rate. Looking at the interaction with unemployment rate, there is an even clearer distinction between regions that experienced consistent growth in unemployment, where current LTV and current interest rate have large and nonlinear effects on the probability of default, and regions with low unemployment growth, where the ALE of current LTV and current interest rate are mostly flat. Hence, these results indicate that the local economic conditions do not only have a direct impact on loan default, but, most importantly, they have an indirect effect by shaping the impact of loan-specific variables on loan default. ALE plots provide supporting evidence of the regional heterogeneity already

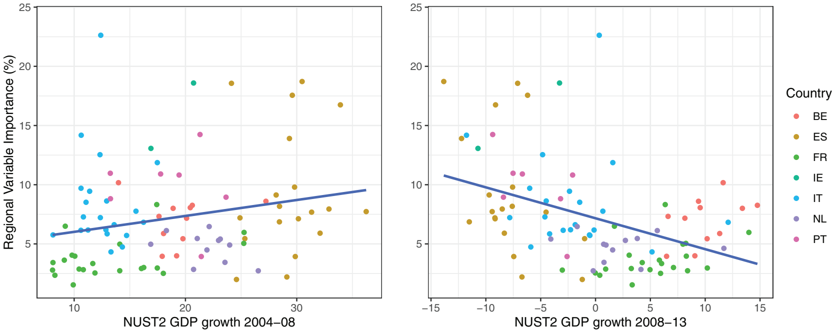

Figure 8 Average Variable Importance of Regional features against NUTS2 GDP growth over 2004–2008 (a) and 2008–2013 (b).

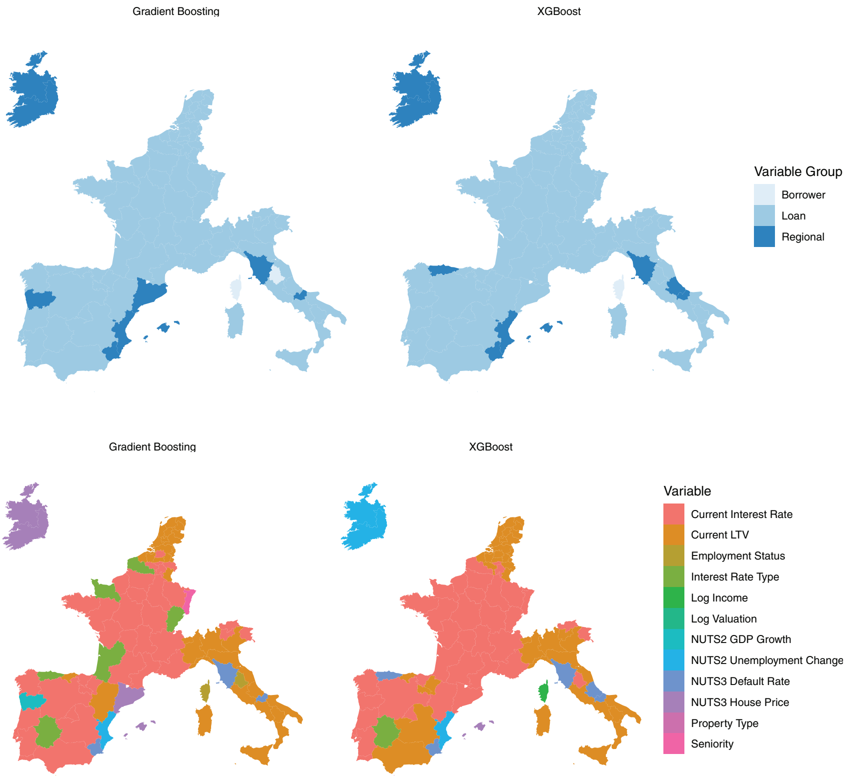

Figure 9 Most important variable group following definition in Table 3 (top) and most important variable (bottom) by region from GB (left) and XGB (right) estimations.

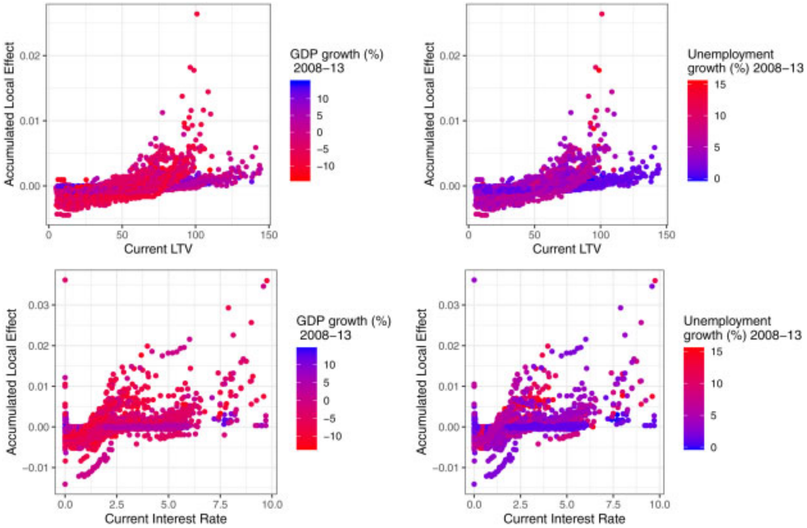

Figure 10 Scatter plot of the current LTV (top) and current interest rate (bottom) and their ALE, with the color of the dot representing the NUTS2 GDP growth 2008–2013 (left) and NUTS2 unemployment rate growth (right) for the XGB estimates.

highlighted in the previous paragraphs and show the importance of allowing for nonlinear interactions among variables when predicting default, being in line with the better performance of ML models against the linear benchmark.

5 Concluding Remarks

A better understanding of the default drivers is of primary importance for policymakers in order to reduce the societal costs associated with a default and avoid the inefficient allocation of resources. We study the ED dataset that provides information on millions of 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 the regional level and we add local macroeconomic variables as additional controls. We predict the default occurrence at the NUTS2 level using different methodologies, including the penalized logistic regression, tree-based ML techniques, and deep learning models.

Our results show that the XGB attains significantly higher out-of-sample performance in terms of prediction accuracy. The most important variables to predict the default occurrence are current interest rate and LTV, together with local economic conditions. Conversely, other loan and borrower-specific variables (e.g., the income of the borrower or its employment status) are less important, even though these features are of primary importance when evaluating the possibility of granting the loan. We show the nonlinear effects of the most important variables on predicting default, which is in line with the better performance of ML methods with respect to the linear logistic benchmark. Our results highlight the presence of across- and within-country heterogeneity in variable importance. In particular, local economic conditions are of primary importance at predicting default in regions that suffered the largest macroeconomic shocks after the 2008 financial crisis, possibly indicating that such variables serve as proxy for shocks to borrower's income, which are unobserved in our dataset.

Overall, in Europe, the main trigger for a residential loan default is the borrower's inability to pay due to income shocks, such as a job loss, or due to an adverse change in economic conditions, such as an increase in interest rates that makes servicing the loan more difficult. We also show the presence of country fixed effects, but also of some intracountry variability in terms of forecast performance and of variable importance. This suggests that regionally tailored risk assessment and policies could potentially achieve more accurate default forecasts and reduce the inefficient allocation of resources to uncreditworthy borrowers. Risk assessment procedures could largely benefit from the application of ML methods, financial institutions could decide whether to grant a loan more conscious of the risk associated to each borrower type and follow more accurately the risk profile evolution of the loan over time. Drawing from the interpretable ML literature, our analysis shows how to identify the more relevant default drivers and explore the nature of their effects, easing the communication of the analysis to a nontechnical audience.

There are some open questions left from our analysis. One relates to the role played by the lending policies of the different financial institutions in determining the loan default across regions. In several European countries, regional saving banks are still major players which have the effect of making the local economy dependent to the lending practices of these banks. Future research might investigate the presence of bank-specific effects and exploring their impact on default behavior at the regional level. A second question could investigate the validity of such loan-level analysis on credit data other than residential mortgages. The ED is a rich source of information about European credit markets and provides also data about car leasing, SME loans, consumer finance, and credit card lines. A better understanding of the default behavior and of the regional differences in these other credit markets could help policy makers to undertake more effective risk-mitigating actions.

Supplementary Data

Supplementary data are available at *Journal of Financial Econometrics* online.

Acknowledgments

We would like to thank participants to the IAAE-2019, CFE-2019, and World Congress 2020 of the Econometric Society for the insightful comments that have greatly improved the article. We are also greatly indebted to the Centre for Advanced Studies of the Joint Research Centre for the support, encouragement, and stimulating environment while the authors were part of the bigNOMICS project.

References

Adelino, M., A. Schoar, and F. Severino. 2016. Loan Originations and Defaults in the Mortgage Crisis: The Role of the Middle Class. *Review of Financial Studies* 29: 1635–1670.

Albanesi, S., and D. F. Vamossy. 2019. *Predicting Consumer Default: A Deep Learning Approach*. Technical report, National Bureau of Economic Research Working paper no. 26165.

Albertazzi, U., S. Ongena, and F. Fringuellotti. 2019. *Fixed Rate versus Adjustable Rate Mortgages: Evidence from Euro Area Banks*. European Central Bank Working paper No. 2322.

Aller, C., and C. Grant. 2018. The Effect of the Financial Crisis on Default by Spanish Households. *Journal of Financial Stability* 36: 39–52.

Ampudia, M., H. van Vlokhoven, and D. Zochowski. 2016. Financial Fragility of Euro Area Households. *Journal of Financial Stability* 27: 250–262.

Apley, D. W. 2016. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv preprint arXiv:1612.08468.

Aron, J., and J. Muellbauer. 2016. Modeling and Forecasting Mortgage Delinquency and Foreclosure in the UK. *Journal of Urban Economics* 94: 32–53.

Babii, A., X. Chen, and E. Ghysels. 2019. Commercial and Residential Mortgage Defaults: Spatial Dependence with Frailty. *Journal of Econometrics* 212: 47–77.

Babii, A., E. Ghysels, and J. Striaukas. 2020. Machine Learning Time Series Regressions with an Application to Nowcasting. arXiv preprint arXiv:2005.14057.

Bajari, P., C. Chu, and M. Park. 2008. *An Empirical Model of Subprime Mortgage Default from 2000 to 2007*. Technical report, National Bureau of Economic Research Working paper no. 14625.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57: 289–300.

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78: 1–3.

Butaru, F., Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique. 2016. Risk and Risk Management in the Credit Card Industry. *Journal of Banking & Finance* 72: 218–239.

Campbell, J., and J. Cocco. 2015. A Model of Mortgage Default. *Journal of Finance* 70: 495–554.

Carrasco, M., and B. Rossi. 2016. In-Sample Inference and Forecasting in Misspecified Factor Models. *Journal of Business & Economic Statistics* 34: 313–338.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research* 16: 321–357.

Chen, T., and C. Guestrin. 2016. "Xgboost: A scalable tree boosting system." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA*, ACM, 785–794.

Cunningham, D. F., and C. A. Capone. 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.

Desai, V. S., J. N. Crook, and G. A. Overstreet Jr. 1996. A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment. *European Journal of Operational Research* 95: 24–37.

D'Hondt, C., R. De Winne, E. Ghysels, and S. Raymond. 2019. Artificial Intelligence Alter Egos: Who Benefits from Robo-investing? arXiv preprint arXiv:1907.03370.

Diaz-Serrano, L. 2005. Income Volatility and Residential Mortgage Delinquency across the EU. *Journal of Housing Economics* 14: 153–177.

Diebold, F. X., and R. S. Mariano. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13: 253–263.

Dirick, L., T. Bellotti, G. Claeskens, and B. Baesens. 2019. Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariates in Mixture Cure Models. *Journal of Business & Economic Statistics* 37: 40–53.

Djeundje, V. B., and J. Crook. 2018. Incorporating Heterogeneity and Macroeconomic Variables into Multi-State Delinquency Models for Credit Cards. *European Journal of Operational Research* 271: 697–709.

Duygan-Bump, B., and C. Grant. 2009. Household Debt Repayment Behaviour: What Role Do Institutions Play? *Economic Policy* 24: 107–140.

Elul, R., N. S. Souleles, S. Chomsisengphet, D. Glennon, and R. Hunt. 2010. What “Triggers” Mortgage Default? *American Economic Review* 100: 490–494.

Ertan, A., M. Loumioti, and R. Wittenberg-Moerman. 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 S. Gross. 2005. Mortgage Default: Classification Trees Analysis. *The Journal of Real Estate Finance and Economics* 30: 369–396.

Fitzpatrick, T., and C. Mues. 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., T. Hastie, and R. Tibshirani. 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., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33: 1.

Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29: 1189–1232.

Friedman, M. 1940. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics* 11: 86–92.

Fuster, A., P. Goldsmith-Pinkham, and T. Ramadorai. 2018. “Predictably Unequal? The Effects of Machine Learning on Credit Markets.” Available at SSRN: https://ssrn.com/abstract=3072038.

Fuster, A., and P. S. Willen. 2017. Payment Size, Negative Equity, and Mortgage Default. *American Economic Journal: Economic Policy* 9: 167–191.

García, S., A. Fernández, J. Luengo, and F. Herrera. 2010. Advanced Nonparametric Tests for Multiple Comparisons in the Design of Experiments in Computational Intelligence and Data Mining: Experimental Analysis of Power. *Information Sciences* 180: 2044–2064.

Garcia, S., and F. Herrera. 2008. An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for All Pairwise Comparisons. *Journal of Machine Learning Research* 9: 2677–2694.

Georgarakos, D., A. Lojschová, and M. Ward-Warmedinger. 2009. “Mortgage Indebtedness and Household Financial Distress. Institute for the Study of Labor (IZA).” IZA Discussion papers 12.

Gerardi, K., K. F. Herkenhoff, L. E. Ohanian, and P. S. Willen. 2018. Can’t Pay or Won’t Pay? unemployment, Negative Equity, and Strategic Default. *The Review of Financial Studies* 31: 1098–1131.

Gerlach-Kristen, P., and S. Lyons. 2015. *Mortgage Arrears in Europe: The Impact of Monetary and Macroprudential Policies*. No. 2015-05. Zürich: Swiss National Bank.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies* 33: 2223–2273.

Hand, D. J. 2009. Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve. *Machine Learning* 77: 103–123.

Hand, D. J., and C. Anagnostopoulos. 2014. A Better Beta for the H-Measure of Classification Performance. *Pattern Recognition Letters* 40: 41–46.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer. https://link.springer.com/book/10.1007/978-0-387-84858-7#authorsandaffiliationsbook.

Iman, R. L., and J. M. Davenport. 1980. Approximations of the Critical Region of the Friedman Statistic. *Communications in Statistics—Theory and Methods* 9: 571–595.

Jappelli, T., M. Pagano, and M. Di Maggio. 2013. Households' Indebtedness and Financial Fragility. *Journal of Financial Management, Markets and Institutions* 1: 23–46.

Khandani, A. E., A. J. Kim, and A. W. Lo. 2010. Consumer Credit-Risk Models via Machine-Learning Algorithms. *Journal of Banking & Finance* 34: 2767–2787.

King, G., and L. Zeng. 2001. Logistic Regression in Rare Events Data. *Political Analysis* 9: 137–163.

Lahiri, K., and L. Yang. 2013. Forecasting Binary Outcomes. In: *Handbook of Economic Forecasting*, Vol. 2, 1025–1106. Amsterdam, Netherlands: Elsevier.

Lambrecht, B. M., W. R. M. Perraudin, and S. Satchell. 2003. Mortgage Default and Possession under Recourse: A Competing Hazards Approach. *Journal of Money, Credit, and Banking* 35: 425–442.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, 4765–4774.

Medeiros, M. C., G. F. Vasconcelos, Á. Veiga, and E. Zilberman. 2019. Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics* 39: 1–22.

Mian, A., and A. Sufi. 2009. The Consequences of Mortgage Credit Expansion: Evidence from the US Mortgage Default Crisis. *Quarterly Journal of Economics* 124: 1449–1496.

Mian, A., A. Sufi, and E. Verner. 2017. Household Debt and Business Cycles Worldwide. *The Quarterly Journal of Economics* 132: 1755–1817.

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

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. 2019. Interpretable Machine Learning: Definitions, Methods, and Applications. arXiv preprint arXiv:1901.04592.

Park, J. Y., and P. C. B. Phillips. 2000. Nonstationary Binary Choice. *Econometrica* 68: 1249–1280.

Ribeiro, M. T., S. Singh, and C. Guestrin. 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, San Francisco California USA*, 1135–1144.

Sanchez-Martinez, M. T., J. Sanchez-Campillo, and D. Moreno-Herrero. 2016. Mortgage Debt and Household Vulnerability: Evidence from Spain before and during the Global Financial Crisis. *International Journal of Housing Markets and Analysis* 9: 400–420.

Schiltz, F., C. Masci, T. Agasisti, and D. Horn. 2018. Using Regression Tree Ensembles to Model Interaction Effects: A Graphical Approach. *Applied Economics* 50: 6341–6354.

Sirignano, J., A. Sadhwani, and K. Giesecke. 2016. *Deep Learning for Mortgage Risk. arXiv preprint arXiv:1607.02470.*

Timmermann, A., and Y. Zhu. 2019. Comparing Forecasting Performance with Panel Data. CEPR Discussion Paper No. DP13746. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3395183.

Van Bekkum, S., M. Gabarro, and R. M. Irani. 2017. Does a Larger Menu Increase Appetite? Collateral Eligibility and Credit Supply. *The Review of Financial Studies* 31: 943–979.

Voigtländer, M. 2009. Why is the German Homeownership Rate so Low? *Housing Studies* 24: 355–372.

Zhao, Q., and T. Hastie.2019. Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics* 39: 1–10.

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