Forecasting Loan Default in Europe with Machine Learning

Luca Barbaglia[a,1], Sebastiano Manzan[b] and Elisa Tosetti[c]

[a]European Commission, Joint Research Centre

[b]City University of New York, Baruch College, United States

[c]Ca' Foscari University of Venice, Italy

European DataWarehouse - Winter Research Seminar

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

Big Data and Economic Forecasting team

BigNOMICS is a project within the JRC's Center of Advanced Studies aimed at exploring the use of Big Data for Economic Forecasting:

- **Gdelt** project: processed indicators of sentiment from world news
- **seismic** data: cultural noise to proxy economic activity
- **news** data: raw texts and speeches
- **loan-level** data: granular credit information ($\rightarrow$ today's talk!)

Check out the recordings of the **II Big Data & Economic Forecasting** workshop to have an overview of our works.

European Commission

Joint Research Centre

EDW works

The economic literature has been already using EDW. Some examples:

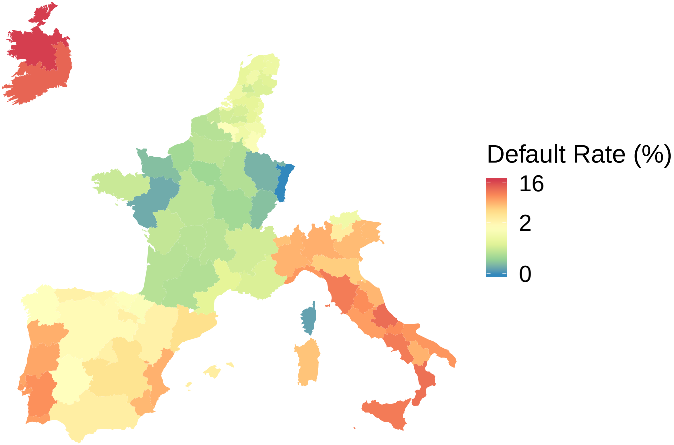
- Ertan et al. (2017): enhancing loan quality through transparency
- Van Bekkum et al. (2017): examine a change in the ECB collateral framework, which significantly lowered the rating requirements
- Billio et al. (2021): link building energy efficiency and default

Jointly with Manzan and Tosetti, our work on **RMBS** is available at:

- Forecasting Loan Default in the EU with Machine Learning (2021) *Journal of Financial Econometrics* **PDF**
- Household Debt and Economic Growth in Europe (2020), *working paper* **PDF**

Forecasting Loan Default with Machine Learning

Millions of mortgages in EU NUTS2 regions



Find the most relevant default **drivers** and provide a default **forecast**

Default prediction

We predict the occurrence of default using information about:

Loan	status, amount, interest rate, loan-to-value (LTV) . . .
Borrower	employment status, income, debt-to-income (DTI)
Local	NUTS3 economic conditions (e.g., GDP growth, . . .)

1. Which **model** among logistic and machine learning performs best?
2. Which **variables** are most relevant to predict default?
3. In which **regions** we predict best?

Literature

Default prediction

We predict the occurrence of default using information about:

Loan	status, amount, interest rate, loan-to-value (LTV) . . .
Borrower	employment status, income, debt-to-income (DTI)
Local	NUTS3 economic conditions (e.g., GDP growth, . . .)

1. Which **model** among logistic and machine learning performs best?
2. Which **variables** are most relevant to predict default?
3. In which **regions** we predict best?

Literature

European DataWarehouse

- Residential Mortgages (RMBS)
- Belgium, France, Ireland, Italy, The Netherlands, Portugal and Spain
- **300 million** records of loans originated from 2000 to 2018
- Default period from 2013 to 2018
- Default: +90 days of delinquencies → **outcome variable**
  (If default, no possibility of returning performing)

Explanatory variables[2]

Loan-level		
Interest Rate Type	static	Categorical
Current Interest Rate	dynamic	Numeric
Property Type	static	Categorical
Log Valuation Amount	static	Numeric
Current Loan-to-value	dynamic	Numeric
Seniority	dynamic	Numeric
DTI	static	Numeric
Borrower-level		
Borrower's Employment	static	Categorical
Log Income	static	Numeric
Regional information		
Lagged Default Rate by NUTS3	dynamic	Numeric
Lagged House Price by NUTS3	dynamic	Numeric
Unemployment Rate Growth by NUTS2	dynamic	Numeric
GDP Growth by NUTS2	dynamic	Numeric

[2] Control for originator-specific effects

Models

We predict the default using many Machine Learning models:

1. Penalized Logistic Regression
2. Penalized Logistic Regression with non-linear effects
3. Naive Bayes
4. Random Forest
5. Gradient Boosting
6. eXtreme Gradient Boosting (XGBoost)
7. Neural Network

Model details

Penalized Logistic Regressions:

- allow for a large number of regressors
- easy to evaluate variable importance by looking at coefficients

RF, **GB** and **XGBoost**:

- Bagging and boosting machine learning techniques
- XGBoost is a penalized and optimized version of GB

Neural Network:

- Capture complex non-linear interactions
- many parameters to tune

Interpret models and default drivers relying on **Explainable AI** techniques.

Model details

Penalized Logistic Regressions:

- allow for a large number of regressors
- easy to evaluate variable importance by looking at coefficients

RF, **GB** and **XGBoost**:

- Bagging and boosting machine learning techniques
- XGBoost is a penalized and optimized version of GB

Neural Network:

- Capture complex non-linear interactions
- many parameters to tune

Interpret models and default drivers relying on **Explainable AI** techniques.

Model details

Penalized Logistic Regressions:

- allow for a large number of regressors
- easy to evaluate variable importance by looking at coefficients

RF, **GB** and **XGBoost**:

- Bagging and boosting machine learning techniques
- XGBoost is a penalized and optimized version of GB

Neural Network:

- Capture complex non-linear interactions
- many parameters to tune

Interpret models and default drivers relying on **Explainable AI** techniques.

Data preparation

- Data cleaning
- Training (60%), validation (20%) and test (20%)
- Stratified sampling
- Rare occurrence of default: balancing via over/under-sampling
- Out-of-sample measures for unbalanced data (e.g., H-measure)

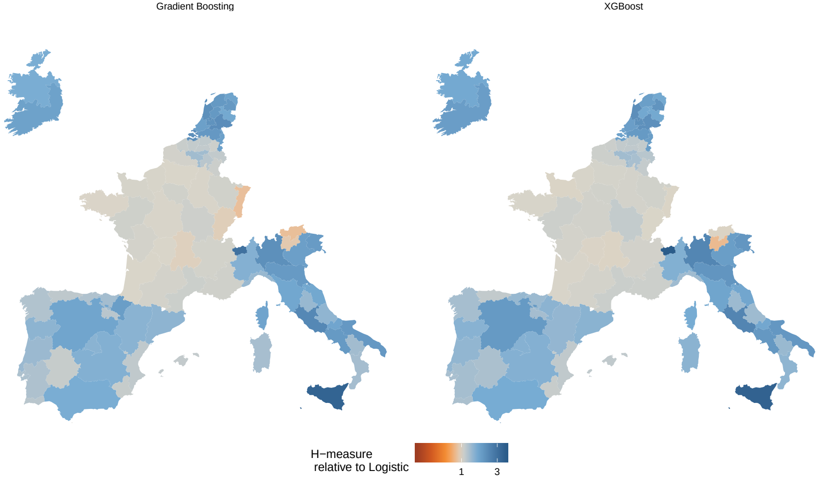
Estimate the default occurrence at **NUTS2** regional level

Data preparation

- Data cleaning
- Training (60%), validation (20%) and test (20%)
- Stratified sampling
- Rare occurrence of default: balancing via over/under-sampling
- Out-of-sample measures for unbalanced data (e.g., H-measure)

Estimate the default occurrence at **NUTS2** regional level

Model performance: H-measure

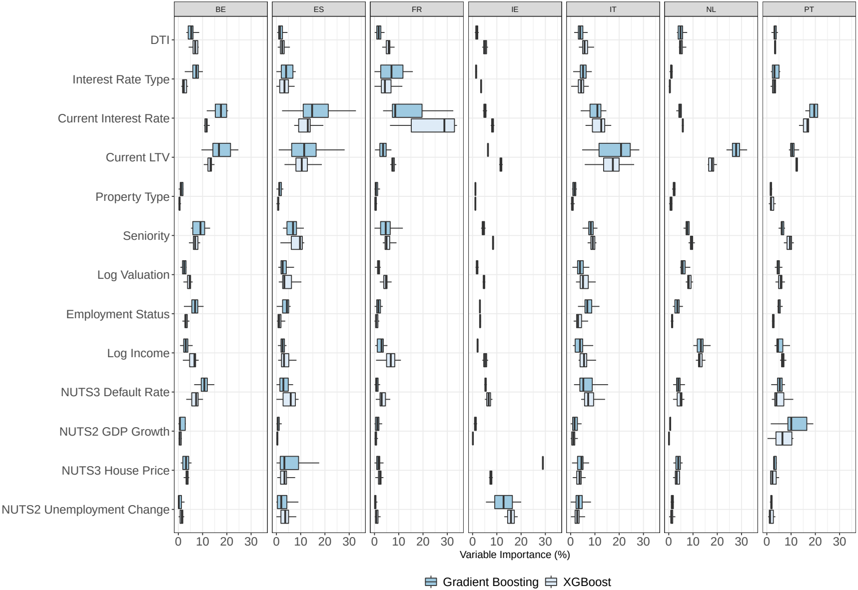


Good forecast performance with boosting methods[3]

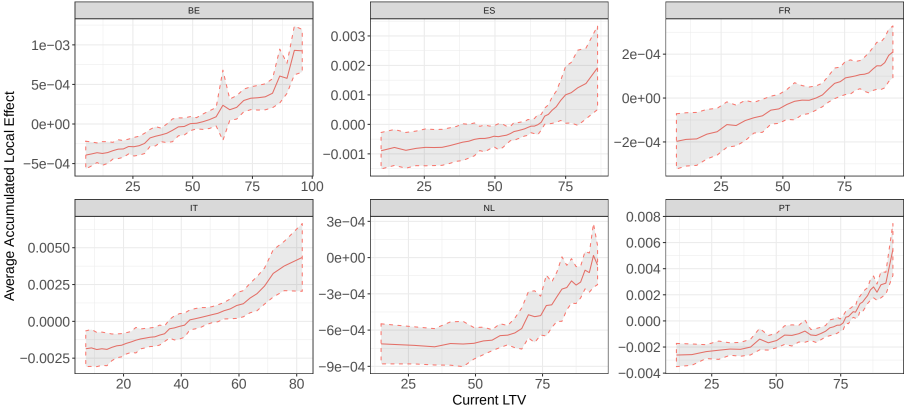
Statistical Tests AUC

[3] Values $> 1 \Rightarrow$ better performance relative to LG

Variable importance

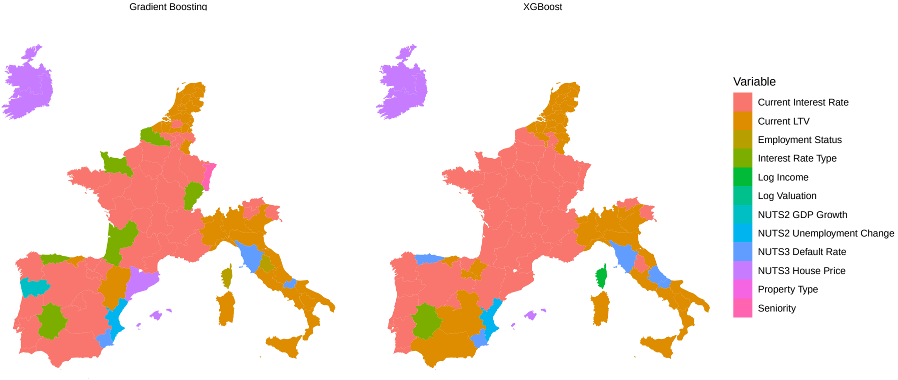


Accumulated Local Effects from XGBoost: LTV



Accumulated Local Effects effects from partial dependence functions without assuming independent explanatory variables

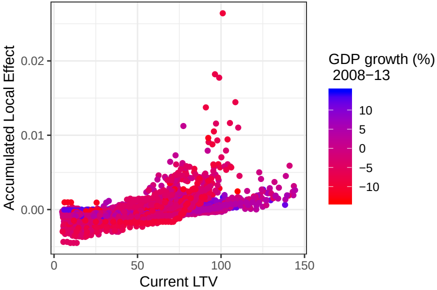
Most Important Variables

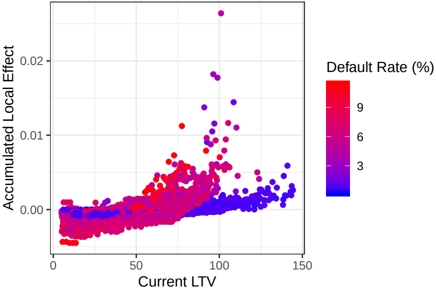


by group

Local effect and macroeconomic variables

Accumulated local effect of LTV against regional GDP growth in 2008-13 and default rate





Non-linear effect of Current LTV in regions that suffered the large GDP drops and high default rate.

Conclusions

Data set on millions of mortgages to analyse default in EU regions

1. Boosting methods have better **out-of-sample performance** in default forecasting than penalized logistic regression
$\rightarrow$ capture **non-linear** effects

2. **Loan-level** variables are the most important at explaining default, and we observe **regional heterogeneity** in default drivers

3. **Regional heterogeneity** $\Rightarrow$ regionally targeted prudential policies

Household Debt and Economic Growth

EDW data to proxy macroeconomic activity

Hypothesis: before the crisis, an expansion of the credit supply and high expectations of rising house prices eased the access to credit

Use EDW data to measure **household indebtedness** at regional level and assess the severity of the 2008 and 2011 crisis in the EU:

- Evaluate EDW as a good proxy for macroeconomic developments against the **Household Finance and Consumption Survey**
- We measure household indebtedness using **DTI** from EDW
- **Credit shock** measure reconstructing interest rates at origination
- Use regional credit shock to explain the severity of the crisis

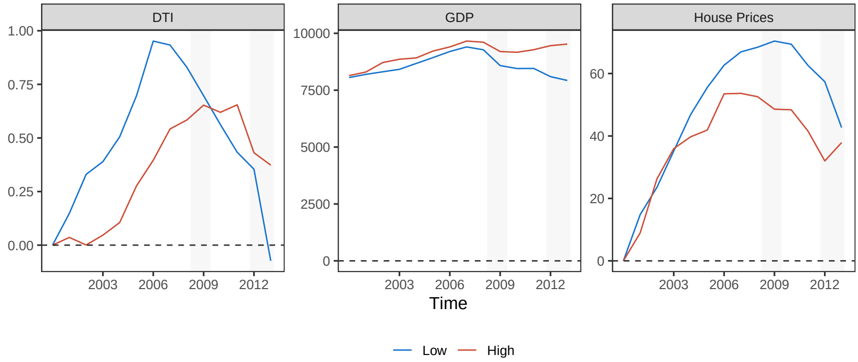
EDW data to proxy macroeconomic activity

Hypothesis: before the crisis, an expansion of the credit supply and high expectations of rising house prices eased the access to credit

Use EDW data to measure **household indebtedness** at regional level and assess the severity of the 2008 and 2011 crisis in the EU:

- Evaluate EDW as a good proxy for macroeconomic developments against the **Household Finance and Consumption Survey**
- We measure household indebtedness using **DTI** from EDW
- **Credit shock** measure reconstructing interest rates at origination
- Use regional credit shock to explain the severity of the crisis

Credit Shock by Income level



Before 2007, low-income households got **more indebted** and expected **higher house prices** relatively to high-income ones

Conclusions

1. Before 2007, credit shock was negative in most European regions indicating that **credit conditions** were significantly relaxed

2. Regions in which household leverage increased rapidly during 2004-06 experienced a severe decline in **output** and **employment**

3. The impact was stronger on **low-** and **middle-income** households

4. The **granular information** in EDW allows to create interesting macroeconomic proxies for the EU regions

References

- Billio, Costola, Pelizzon, Riedel (2021) Buildings Energy Efficiency and the Probability of Mortgage Default: The Dutch Case. *The Journal of Real Estate Finance and Economics*
- Chen, Guestrin (2016) XGBoost: A Scalable Tree Boosting System, in *Proceedings of the 22 Int. Conf. on Knowledge Discovery and Data Mining*
- Ertan, Loumioti, Wittenberg-Moerman (2017) Enhancing Loan Quality through Transparency: Evidence from the European Central Bank Loan Level Reporting Initiative. *Journal of Accounting Research*
- Garcia, Herrera (2008) Statistical Comparisons of Classifiers Over Multiple Data Sets *Journal of Machine Learning Research*
- Lahiri, Yang (2013), Forecasting Binary Outcomes *Handbook of Econ. For.*
- Van Bekkum, Gabarro, Irani (2017) Does a Larger Menu Increase Appetite? Collateral Eligibility and Credit Supply. *The Review of Financial Studies*

These papers are available at:

- Barbaglia, Manzan, Tosetti, "Forecasting Loan Default in Europe with Machine Learning" (2021) *Journal of Financial Econometrics* **PDF**
- Barbaglia, Manzan, Tosetti, "Household debt and Economic Growth in Europe" (2020) *SSRN 3684399* **PDF**

Literature overview

Some relevant works:

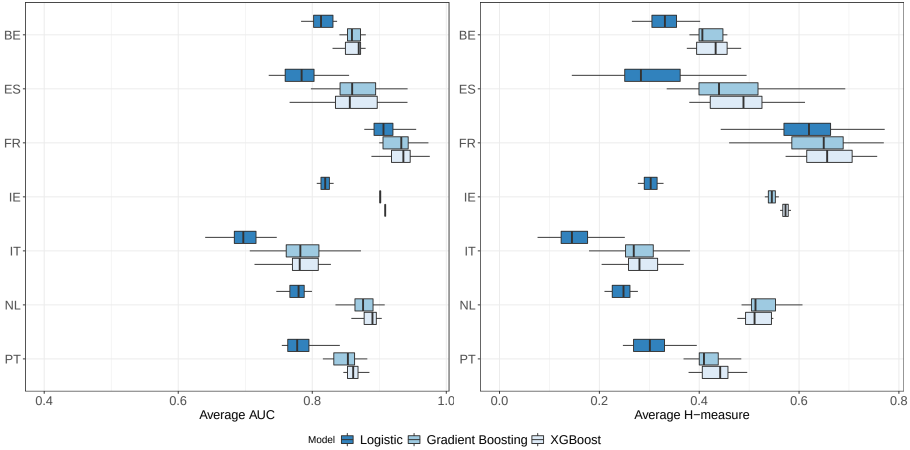
- Sirignano et al. (2018) : deep-learning for mortgage default risk on loan-level information about 120M US mortgages
- Fuster et al. (2018) : compare logistic and random forest on US mortgages including non-financial variables (e.g., ethnicity)
- Fitzpatrick and Mues (2016) : regression tree and random forest on 300,000 Irish mortgages

This paper:

- multi-country loan-level EU data set
- investigate the most important default drivers at regional level
- interpretability of machine learning models for default forecast

main

Model performance



main

Statistical difference of performance rankings

Friedman test H_0 : the rankings are equivalent

Test statistics	Value	p-value
Friedman	96.646	0.00
Iman-Davenport	96.287	0.00

Pairwise test H_0 : the rankings from algorithms i and j are equivalent

	Average ranking	*Pair-wise post hoc test*		
		Logistic	GB	XGBoost
Logistic	2.78	·	0.00	0.00
GB	1.82	0.00	·	0.54
XGBoost	1.40	0.00	0.54	·

main

Statistical difference of default forecast

Compute the Brier's score (BS) and Log score (LS)

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \qquad LS = \sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

with y_i: observed default, p_i: default estimated probability for the i^{th} loan

% of rejections of H_0 of forecast performance equivalence

$\Rightarrow$ prefer model A over model B (i.e., A $\succ$ B)

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

Lahiri and Yang (2013)

main simulations

Statistical difference of default forecast

Compute the Brier's score (BS) and Log score (LS)

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \qquad LS = \sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$$

with y_i: observed default, p_i: default estimated probability for the i^{th} loan

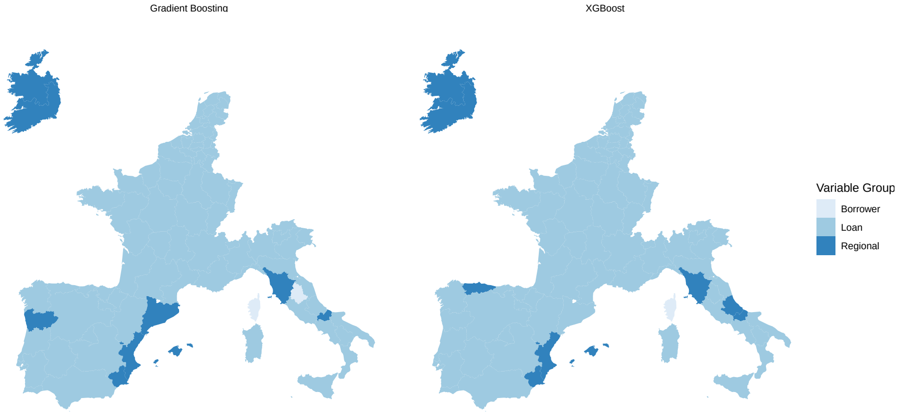
% of rejections of H_0 of forecast performance equivalence
$\Rightarrow$ prefer model A over model B (i.e., A $\succ$ B)

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

Lahiri and Yang (2013)

main simulations

Most Important Variables by Group



main

Explore estimator properties: Monte-Carlo experiment

Assume this DGP $y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$, where:

$$\boldsymbol{x}'_{it}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1,it} + \sum_{j=1}^{K} \beta_{3,j} z_{j,it} + \sum_{j=1}^{K} \beta_{4,j} z_{j,it}^2$$

with $\beta_1 = 0.5$, $\beta_{3,j} = \beta_{4,j} = 1 \ \forall$ j, $x_{1,it} \sim N(0,1)$ and $\boldsymbol{z}_{it} \sim \sigma^2 N(\mathbf{0}, \mathbf{V})$

Four scenarios[4]:

I Low signal, low correlation : $\sigma^2 = 0.01$, $V_{i,j} = 0.01$, for $i \neq j$

II Low signal, high correlation : $\sigma^2 = 0.01$, $V_{i,j} = 0.7$, for $i \neq j$

III High signal, high correlation : $\sigma^2 = 1$, $V_{i,j} = 0.7$, for $i \neq j$

IV High signal, low correlation : $\sigma^2 = 1$, $V_{i,j} = 0.01$, for $i \neq j$

[4] Take $K = 10$ regressors; 2% of non-zero outcomes (highly unbalanced); sample size $n = 10e3, 20e4$ and $10e6$; simulation runs $2e3$.

Monte-Carlo: results relative to Logistic

	N	H-measure		AUC	
		GB	OR	GB	OR
Case I	10e3	0.88	1.76	1.00	1.05
Case I	20e4	14.02	17.13	1.21	1.23
Case I	10e6	526.83	538.67	1.27	1.27
Case II	10e3	1.06	1.86	1.01	1.06
Case II	20e4	15.72	19.20	1.22	1.25
Case II	10e6	598.63	611.46	1.29	1.29
Case III	10e3	3.00	3.87	1.14	1.23
Case III	20e4	50.24	58.28	1.44	1.48
Case III	10e6	1784.84	1843.61	1.51	1.52
Case IV	10e3	1.08	2.40	1.01	1.11
Case IV	20e4	18.00	32.33	1.26	1.37
Case IV	10e6	851.37	1014.56	1.38	1.41

GB better than logistic as the sample grows in all scenarios, except for low-signal and low-correlation (Case I), and close to oracle estimator (OR).

main