BANCO DE ESPAÑA
Eurosistema

USO INTERNO

UNDERSTANDING THE PERFORMANCE OF MACHINE LEARNING MODELS TO PREDICT CREDIT DEFAULT: A NOVEL APPROACH FOR SUPERVISORY EVALUATION

Andrés Alonso and José Manuel Carbó (*)
Financial Innovation Division

12 November, 2020

EBA 9tH RESEARCH WORKSHOP
"New technologies in the banking sector – impacts, risks and opportunities"

(*) The opinions and analyses expressed in this paper are the responsibility of the authors and, therefore, do not necessarily match with those of the Banco de España or the Eurosystem.

DGA INNOVACIÓN FINANCIERA E INFRAESTRUCTURAS DE MERCADO

Motivation

- Recent surveys show that credit institutions are increasingly adopting **Machine Learning (ML)** tools in credit risk management.
  - ➢ Regulatory capital calculation, optimizing provisions, credit-scoring or monitoring.

- While ML usually yield better predictive performance, from a supervisory standpoint it also brings **new challenges:**
  - ➢ Interpretability, biases, data quality, dependency on third-party providers etc.

- Therefore, prior to enter into the risk analysis, it is necessary to **assess the real economic gains** that institutions might get when using ML in credit risk.

Research Question

In this paper we study the performance of several machine learning (ML) models for credit default prediction, and its potential economic impact.

- We use a **unique** and anonymized **database** from a major Spanish bank.
- We compare the statistical performance of **six models**: Logit, Lasso, Classification Tree, Random Forest, XGBoost and Deep Neural Networks.
- We then translate the statistical performance into economic impact by estimating the **savings in regulatory capital** under an IRB approach,
  - Our **benchmark results** show that implementing XGBoost could yield savings from **12.5%** to **17%** with respect implementing Lasso.
  - We believe this estimate would be a **lower bound** of the potential benefits.

Dataset

- An **anonymized** database of consumer credit from **Banco Santander** has been used to conduct this analysis.

- The dataset contains information from **more than 75,000 credit operations** which have been classified into two groups, depending on whether they resulted on default or not.

- Additionally, each operation has a maximum of **370 risk factors** associated to it, whose labels or description have not been provided.

- Around 3.95% of the loans resulted in default.

  ➢ The data has **no temporal dimension,** so we do not know when the loan was granted, and if resulted in default, when it happened.

Predictive Performance

- To assess the predictive performance of the **6 ML models** we will focus on **two tasks**: **classification and calibration**

- These tasks are explicitly mentioned in the validation process of a rating system under the **IRB approach**.
  - Under Basel II guidelines, banks are allowed to use their own estimated probability of default (PD) for the purpose of calculating regulatory capital.

1. **Classification or identification of risk:** discriminating those exposures which are more risky from the rest.

   Suggested metrics: **AUC-ROC** (it is also shown **Recall**).

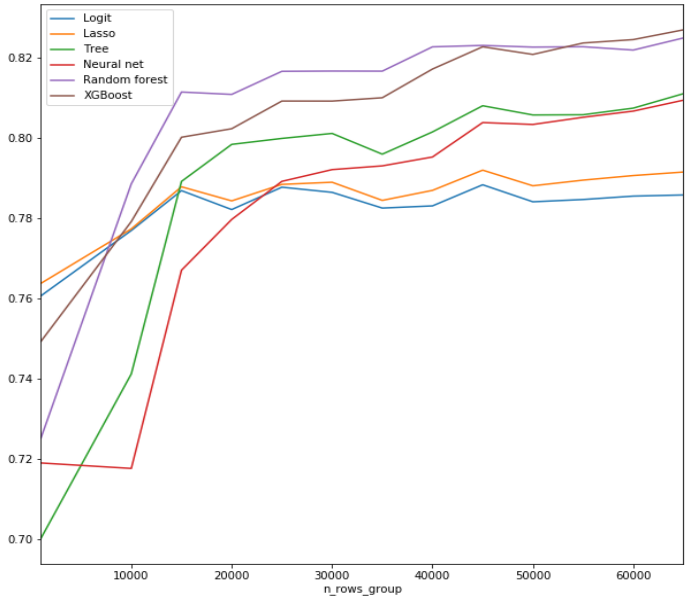
2. **Calibration or quantification of risk:** the risk buckets must be well calibrated, resembling the observed default rate.

   Suggested metrics: **Brier score and reliability curves.**

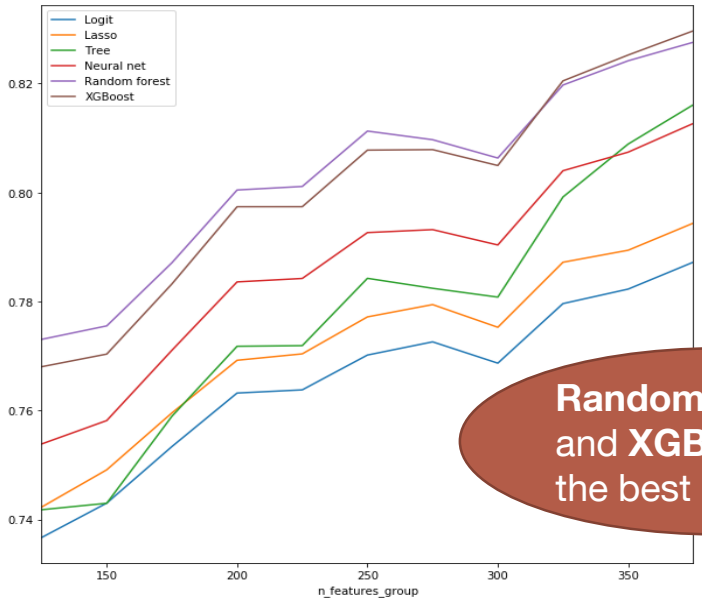
Classification

- We assess the predictive performance under different circumstances:
  - **Different sample sizes**: From 1,000 to 65,000 loans
  - **Different features:** From 100 to 375 features

Simulation of AUC-ROC performance to sample size



Simulation of AUC-ROC performance to number of features

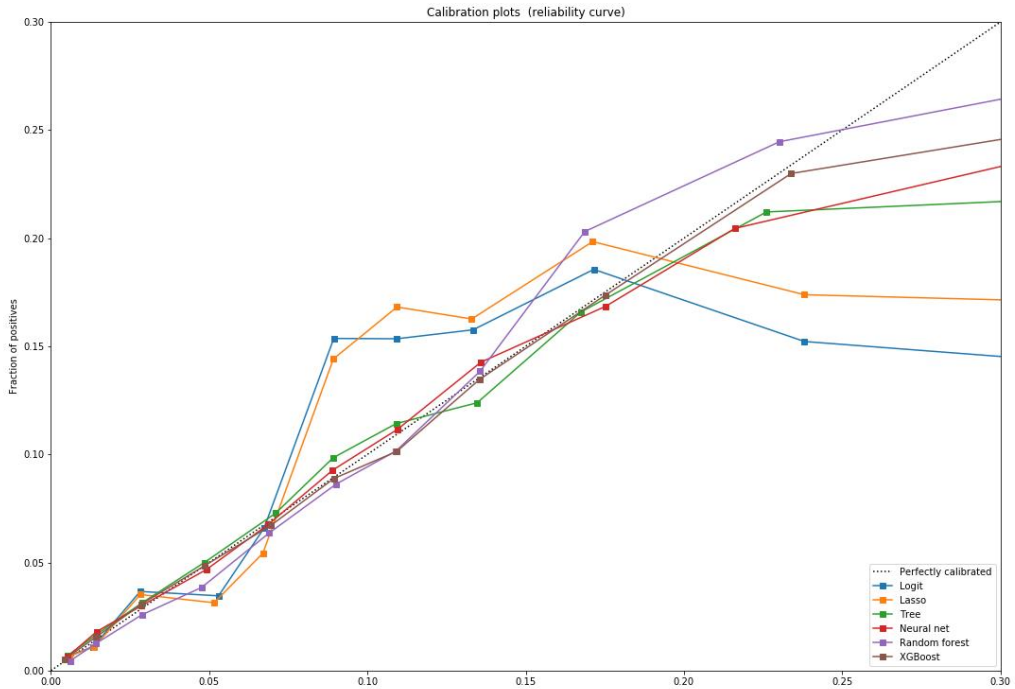


Random Forest and **XGBoost** show the best results.

- **Independently of the amount of available loans or features**, ML models have better performance than Logit or Lasso.

Calibration

- We perform a similar analysis for **Brier score,** which is within a range of **3.3%** and **4**% (XGBoost and Random forest have the best performance
- Differences in Brier Score are so small among models, we propose to build **reliability curves** (45 degrees line represents perfect calibration).



- ✓ **Logit** and **Lasso** are the models further away from the 45º line.
- ✓ **XGBoost** and **Random Forest** are the ones closer to the 45º line.

Link nº loans

Economic impact

- **We compute the savings in terms of regulatory capital** under an IRB approach which could be achieved by using more advanced ML models instead of traditional ones.
  - ➢ We compare savings with XGBoost with respect Lasso

- Basel's risk weighted function for credit risk in the IRB approach is **concave** in the PD (Baena et al, 2015).
  - ➢ If this holds true, a more granular classification of credit ratings should yield a lower overall capital requirement.

✓ Since our data consists of consumer loans, we will use the Basel formulae for **retail exposures** (LGD = 0.45).

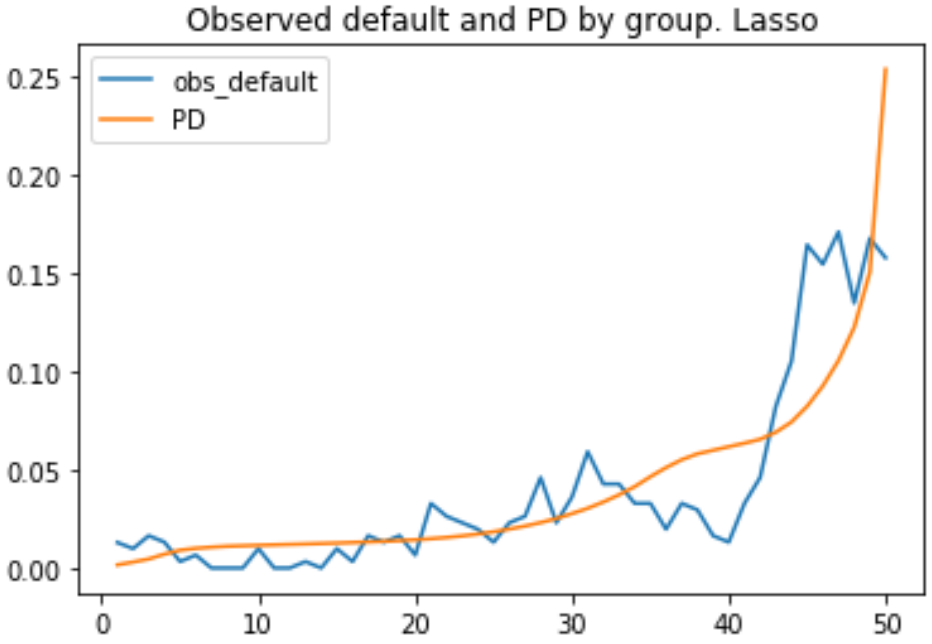
$$Correlation = R = 0.03 \cdot \frac{\left(1-e^{-35 \cdot PD}\right)}{\left(1-e^{-35}\right)} + 0.16 \cdot \left(1 - \frac{\left(1-e^{-35 \cdot PD}\right)}{\left(1-e^{-35}\right)}\right)$$

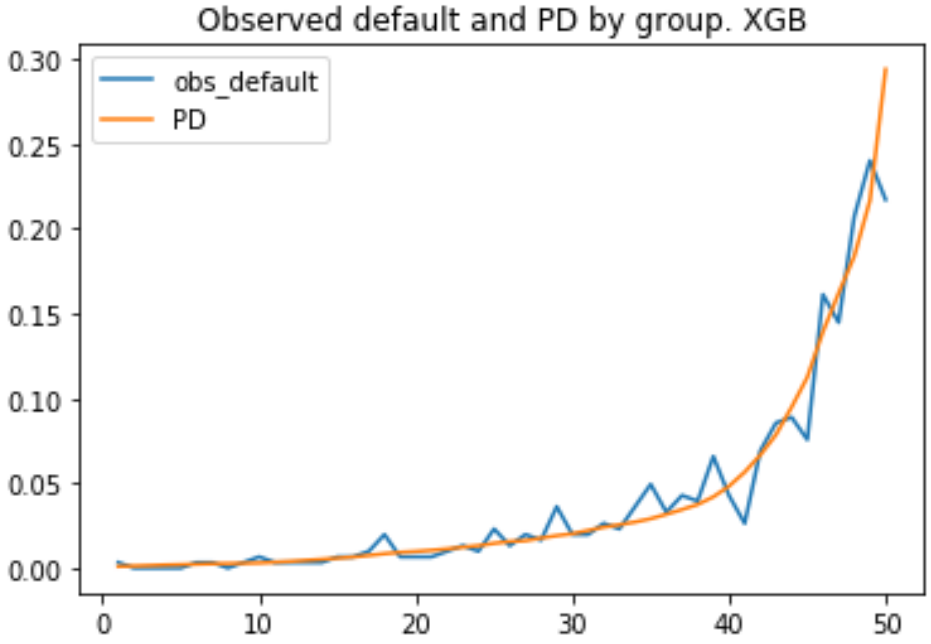
$$Capital\ requirement = K = \left[LGD \cdot N\left[\frac{G(PD)}{\sqrt{(1-R)}} + \sqrt{\frac{R}{1-R}} \cdot G(0.999)\right] - PD \cdot LGD\right]$$

$$RWA = K \cdot 12.5 \cdot EAD$$

Step 1 – Discriminate between risk buckets

- We estimate the PD using both Lasso and XGBoost, and we **order** the predictions proportionally in 50 buckets, from lower to higher values of PD.
- The divergence with the default rate per bucket suggests that a **calibration process needs to be performed**.





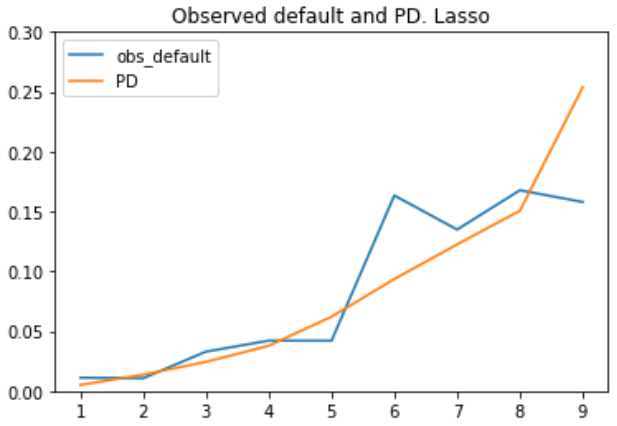
Step 2 – Calibration process

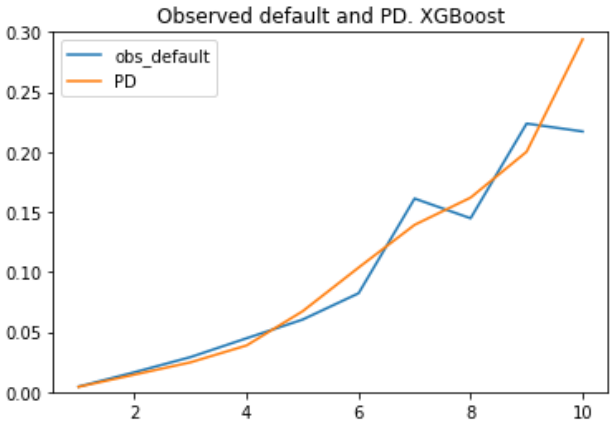
- For these rating buckets to be approved by the supervisor, they must comply with two criteria: **(i) heterogeneity between risk buckets**, and **(ii) homogeneity within risk buckets.**

1. < 1% PD → **AAA**
2. 1%<= PD <= 2% → **AA**
3. 2%<= PD <= 3% → **A**
4. 3%<= PD <= 5% → **BBB**
5. 5%<= PD <= 8% → **BB**
6. 8%<= PD <= 12% → **B**
7. 12%<= PD <= 15% → **CCC**
8. 15%<= PD <= 18% → **CC**
9. 18%<= PD <= 25% → **C**
10. > 25% PD → **D**

Lasso finds 9 buckets

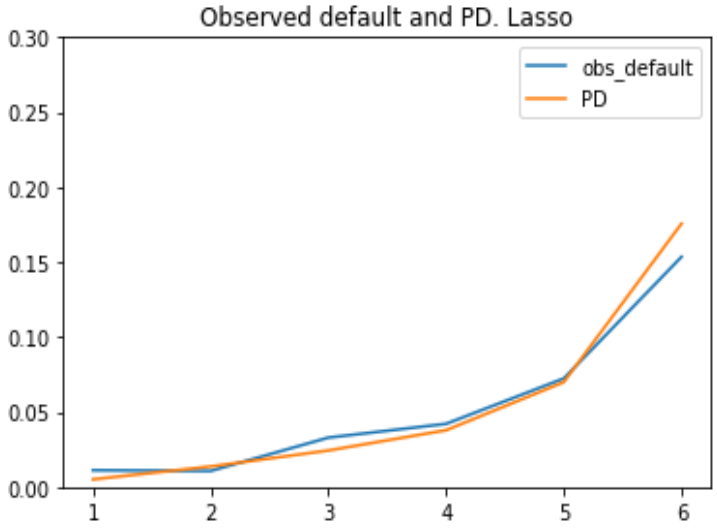
XGBoost finds 10 buckets



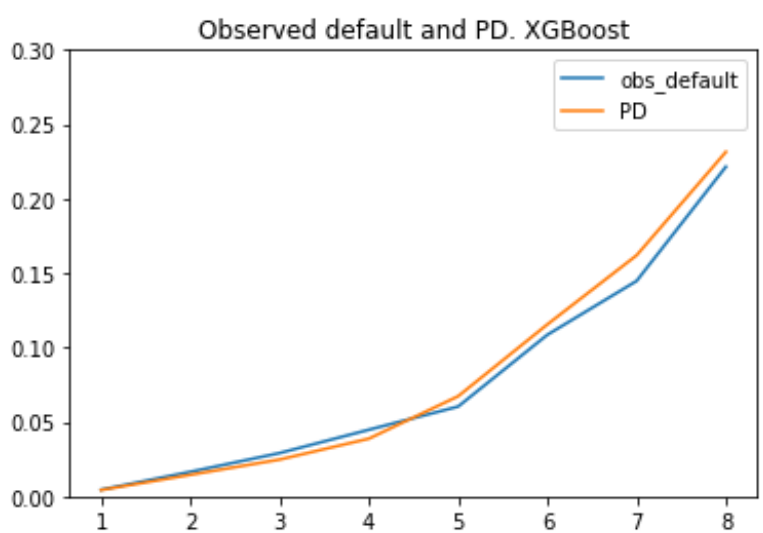


Step 2 – Calibration process

- In order to accomplish the two criteria, **we reduce sequentially the number of buckets**, until we find the first set of ratings for each model which satisfies both criteria.



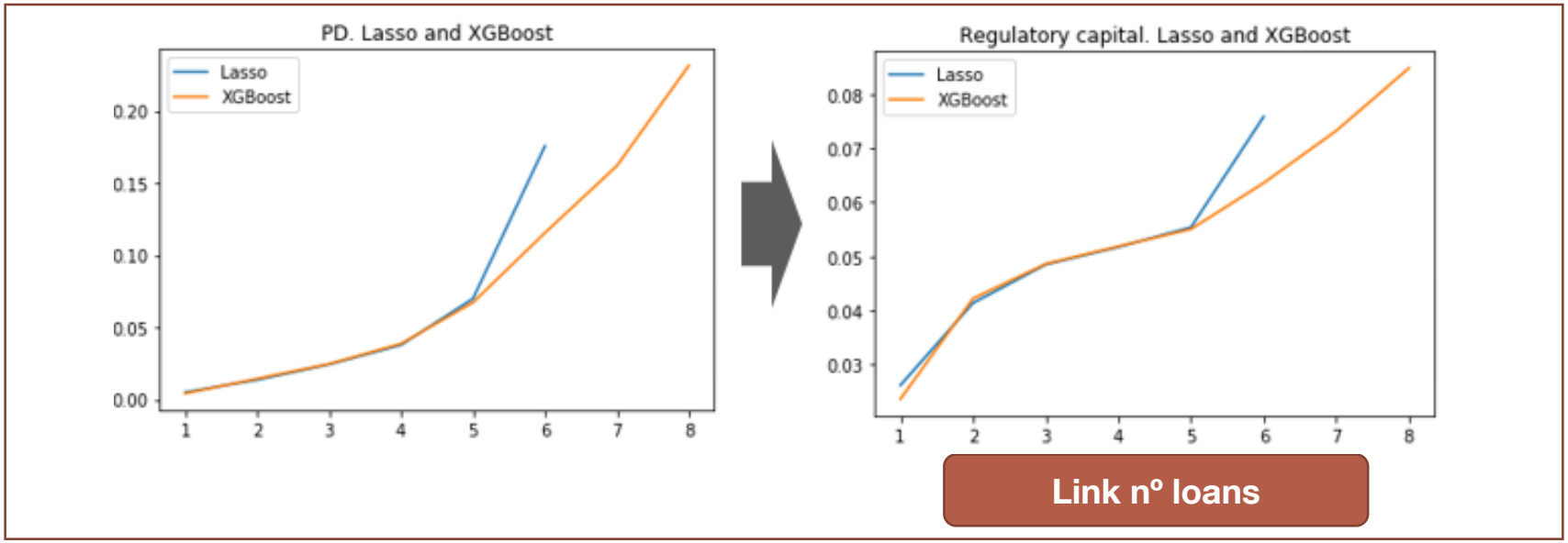
- **Lasso** idenfifies **6** *buckets.*



- **XGBoost** identifies **8** *buckets.*

More granularity!!

Step 3 – Calculation of capital requirements



- **Capital savings from the use of XGBoost are between 12.5% and 17%**
  - XGBoost has more loans below 5% of PD
  - Lasso allocates many loans in groups 5 and 6, while XGBoost is able to differentiate those loans across different groups (5 to 8)

- Jiménez and Saurina (2004) pointed to an inverse relationship between the size of the loan and the PD. Therefore**, we assume that 12.5% is a conservative estimate of the savings in capital requirements.**

CONCLUSIONS

- In this paper we study the performance of several machine learning (ML) models for credit default prediction.

- **Our results** show that ML models perform better than the traditional Logit model, both in classification and calibration terms, showing that **statistically it exists a model advantage** on top of an information advantage (as suggested by Huang et al, 2020).

- Finally, we estimate the **economic impact** of being able to statistically classify and calibrate better by computing the **regulatory capital savings** which could amount to up to 17% in our benchmark exercise.

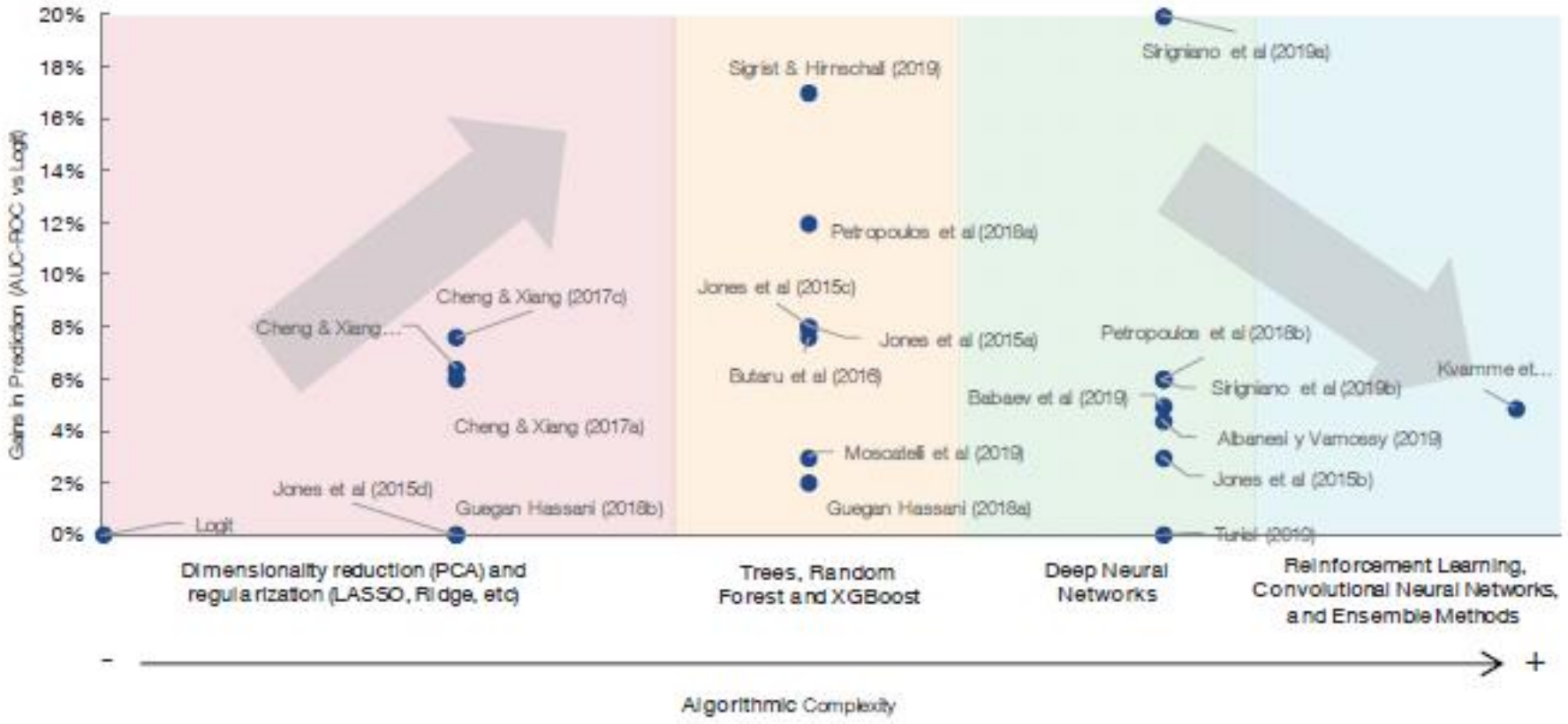
- This is a significant figure that lead us to suggest that **more research is needed to understand the supervisory cost to get a model approval**.

APPENDIX

Literature Review

The dilemma Prediction vs Algorithmic Complexity

(*Alonso y Carbó, 2020*)

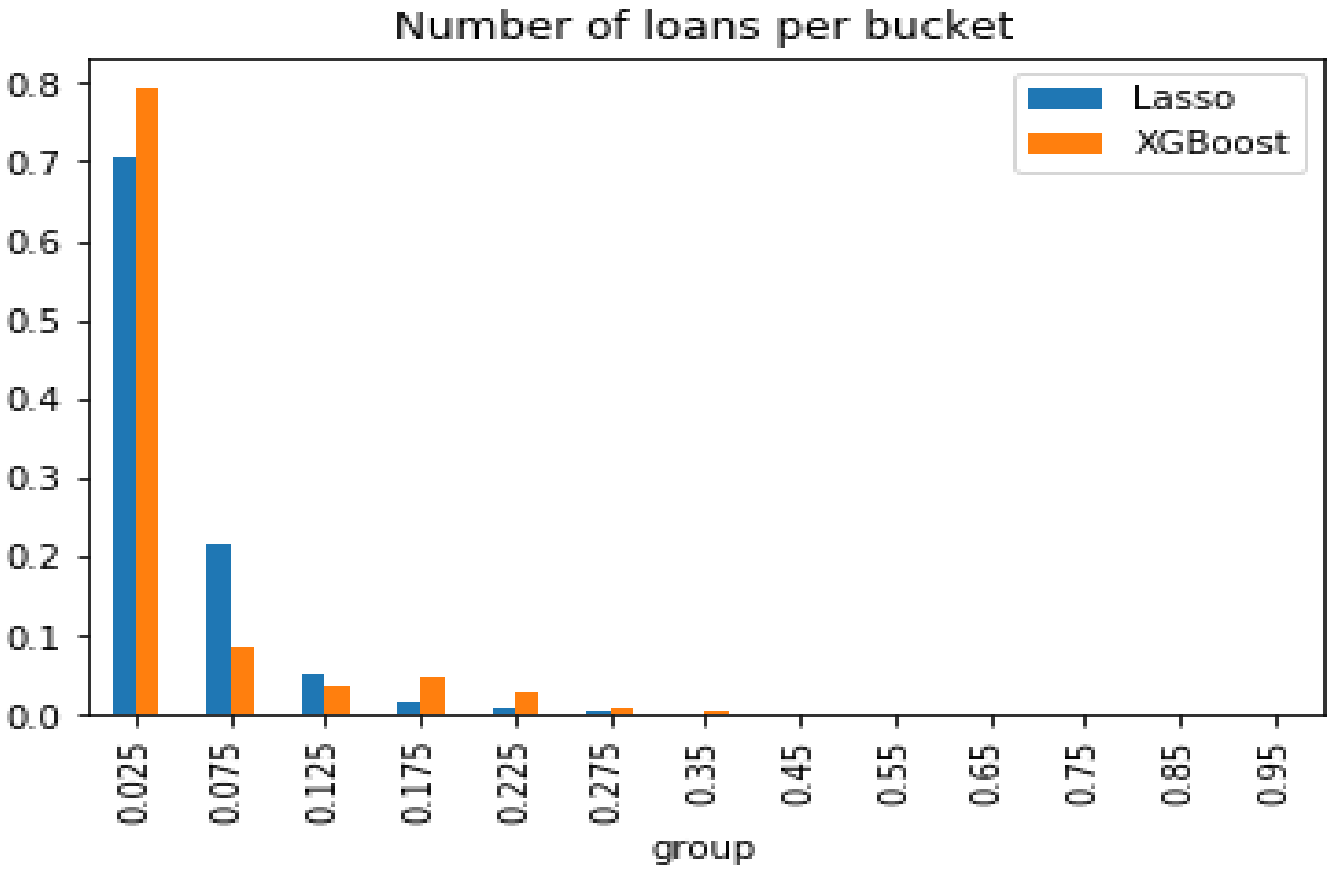


SOURCE: Devised by the authors.

Recall

Method	TPR, Classifier threshold = 10%	TPR, Classifier threshold = 20%	TPR, Classifier threshold = 30%
Logit	33%	6%	1%
Lasso	37%	7%	2%
Tree	49%	18%	4%
Random Forest	55%	9%	2%
XGBoost	55%	24%	8%
Deep learning	52%	16%	2%

Return


Number of loans per bucket
Lasso
XGBoost
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.025
0.075
0.125
0.175
0.225
0.275
0.35
0.45
0.55
0.65
0.75
0.85
0.95
group

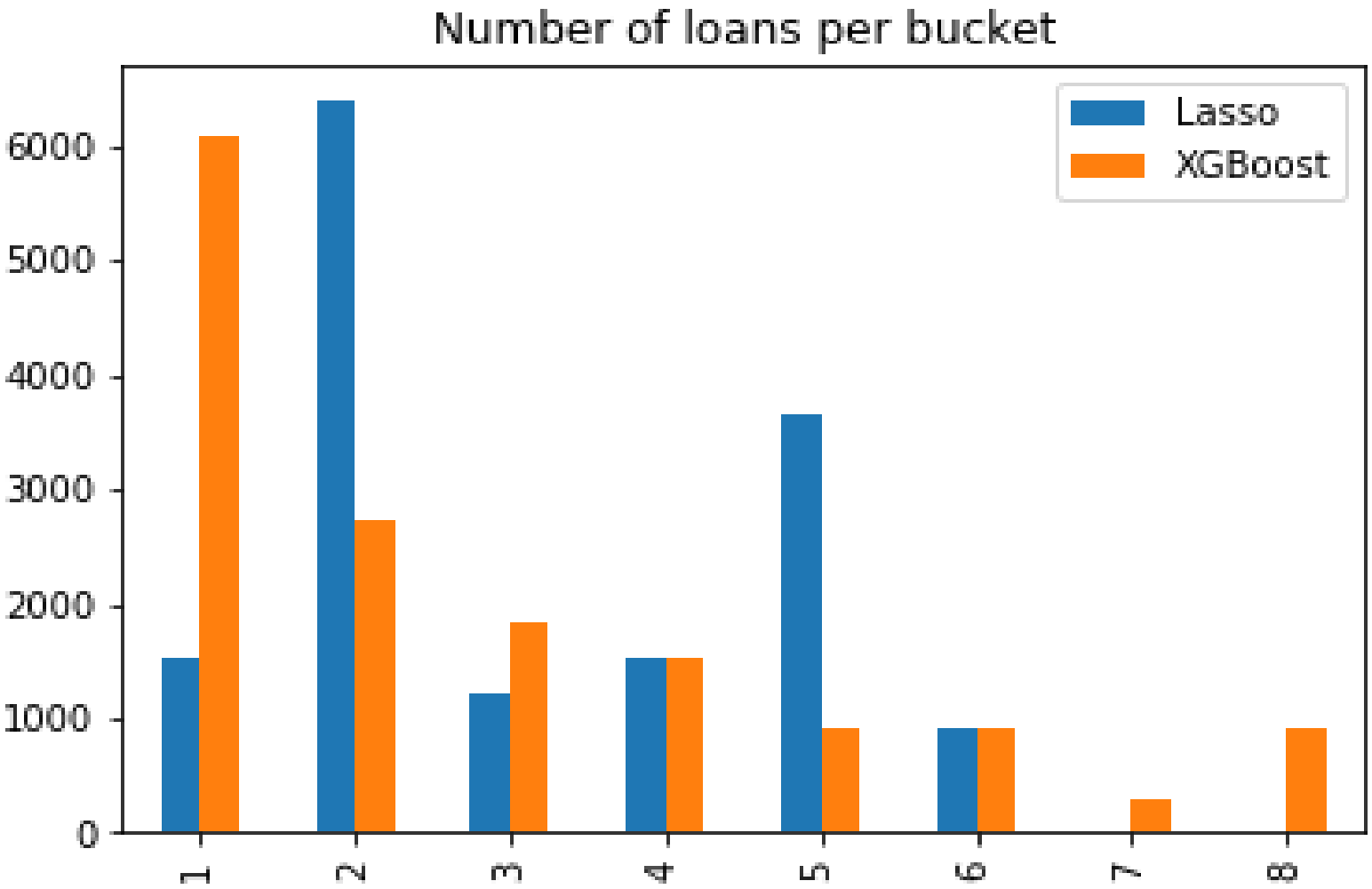

Number of loans per bucket
Lasso
XGBoost
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8

Figure 7a. Reliability curve