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

2021

BANCO DE ESPAÑA
Eurosistema

Documentos de Trabajo
N.º 2105

Andrés Alonso and José Manuel Carbó

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

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

Andrés Alonso and José Manuel Carbó (*)

BANCO DE ESPAÑA

(*) The authors work as senior economists in the Financial Innovation Division of Banco de España, and appreciate the comments received from José Manuel Marqués, Ana Fernández and Sergio Gorjón, as well as all the feedback received in two internal seminar done at the Spanish Central Bank, and the 9th Research Workshop of the European Banking Authority (EBA), specially from our discussant Klaus Duellmann (European Central Bank).
Part of this paper has been used for the realization of the master´s thesis of José Manuel Carbó for the Master in Artificial Intelligence for Financial Markets of Bolsas y Mercados Españoles.

Documentos de Trabajo. N.º 2105

2021

ISSN: 1579-8666 (on line)

Abstract

In this paper we study the performance of several machine learning (ML) models for credit default prediction. We do so by using a unique and anonymized database from a major Spanish bank. We compare the statistical performance of a simple and traditionally used model like the Logistic Regression (Logit), with more advanced ones like Lasso penalized logistic regression, Classification And Regression Tree (CART), Random Forest, XGBoost and Deep Neural Networks. Following the process deployed for the supervisory validation of Internal Rating-Based (IRB) systems, we examine the benefits of using ML in terms of predictive power, both in classification and calibration. Running a simulation exercise for different sample sizes and number of features we are able to isolate the information advantage associated to the access to big amounts of data, and measure the ML model advantage. Despite the fact that ML models outperforms Logit both in classification and in calibration, more complex ML algorithms do not necessarily predict better. We then translate this statistical performance into economic impact. We do so by estimating the savings in regulatory capital when using ML models instead of a simpler model like Lasso to compute the risk-weighted assets. Our benchmark results show that implementing XGBoost could yield savings from 12.4% to 17% in terms of regulatory capital requirements under the IRB approach. This leads us to conclude that the potential benefits in economic terms for the institutions would be significant and this justify further research to better understand all the risks embedded in ML models.

Keywords: machine learning, credit risk, prediction, probability of default, IRB system.

JEL classification: C45, C38, G21.

Resumen

En este artículo estudiamos el rendimiento de diferentes modelos de aprendizaje automático —*machine learning* (ML)— en la predicción de incumplimiento crediticio. Para ello hemos utilizado una base de datos única y anónima de uno de los bancos españoles más importantes. Hemos comparado el rendimiento estadístico de los modelos tradicionalmente más usados, como la regresión logística *(Logit)*, con modelos más avanzados, como la regresión logística penalizada *(Lasso)*, árboles de clasificación y regresión, bosques aleatorios, *XGBoost* y redes neuronales profundas. Siguiendo el proceso de validación supervisora de sistemas basados en calificaciones internas —*Internal ratings-based approach* (IRB)— hemos examinado los beneficios en poder predictivo de usar técnicas de ML, tanto para clasificar como para calibrar. Hemos realizado simulaciones con diferentes tamaños de muestras y número de variables explicativas para aislar las ventajas que pueden tener los modelos de ML asociadas al acceso de grandes cantidades de datos, de las ventajas propias de los modelos de ML. Encontramos que los modelos de ML tienen un mejor rendimiento que *Logit* tanto en clasificación como en calibración, aunque los modelos más complejos de ML no son necesariamente los que predicen mejor. Posteriormente traducimos esta mejoría en rendimiento estadístico a impacto económico. Para ello estimamos el ahorro en capital regulatorio cuando usamos modelos de ML en lugar de métodos tradicionales para calcular los activos ponderados en función del riesgo. Nuestros resultados indican que usar *XGBoost* en lugar de *Lasso* puede resultar en ahorros de un 12,4 % a un 17 %, en términos de capital regulatorio, cuando utilizamos el proceso IRB. Esto nos lleva a concluir que los beneficios potenciales de usar ML, en términos económicos, serían significativos para las instituciones, lo que justifica una mayor investigación para comprender mejor todos los riesgos incorporados en los modelos de ML.

Palabras clave: aprendizaje automático, riesgo de crédito, predicción, probabilidad de impago, modelos IRB.

Códigos JEL: C45, C38, G21.

1. Introduction - Motivation

Recent surveys show that credit institutions are increasingly adopting Machine Learning (ML) tools in several areas of credit risk management, like regulatory capital calculation, optimizing provisions, credit-scoring or monitoring outstanding loans (IIF 2019, BoE 2019, Fernández 2019). While this kind of models usually yield better predictive performance (Albanessi et al 2019, Petropoulos et al 2019)[1], from a supervisory standpoint they also bring new challenges. Aspects like interpretability, stability of the predictions and governance of the models are amongst the most usually mentioned factors and concerns arising from the supervisors when evaluation ML models in financial services (EBA 2017, EBA 2020, BdF 2020). All of them point towards the existence of an implicit cost in terms of risk that might hinder the use of ML tools in the financial industry, as it becomes more difficult (costly) for the supervisor to evaluate these innovative models in order to ensure that all the regulatory requirements are fulfilled. In Alonso and Carbó (2020), we identified a trade-off between predictive performance and the risk of ML models, suggesting a framework to adjust their statistical performance by the models' embedded risk from a supervisory perspective. The absence of transparency and lack of a standardized methodology to evaluate these models is indeed mentioned by market participants when asked about the major impediments that may limit further implementation or scalability of ML technology in the financial industry (IIF 2019, BoE 2019, NP 2020). However, in order to define an adequate regulatory approach it is important to understand not only the risks associated with the use of this technology, but also the tools available to mitigate these risks. Given the novelty and complexity of some of these statistical methods this is not an easy task. Therefore, prior to enter into the risk analysis it could be relevant to ask what will be the real economic gains that credit institutions might get when using different ML models. While there exists an extensive and growing literature on the predictive gains of ML on credit default prediction, any comparison of results from different academic studies carries the caveat of relying on different sample sizes, types of underlying assets and several other differences, like observed frequency of defaults, which would prevent us from having a robust result to be used for this purpose.

In this paper we aim to overcome this limitation by running a simulation exercise on a unique and anonymized database provided by a major Spanish bank. To this extent we compare the performance of a logistic regression (Logit), a well-known econometric model in the banking industry (BIS 2001), with the performance of the following ML models: Lasso penalized logistic regression, Classification And Regression Tree (CART), Random Forest, XGBoost and Deep Neural Networks. As a result, we will compute, firstly, the benefits in

[1] For further references, see next section on literature review.

terms of statistical performance of using ML models from a micro-prudential perspective. Evaluating the macro-prudential effects from the use of ML models is out of the scope of this paper.[2] Finally, we propose a novel approach to translate the statistical performance into actual economic impact of using this type of models in credit default prediction. Assuming the Basel formulas for risk-weighted assets and capital requirements in the Internal Ratings-Based (IRB) approach for retail exposures, as it is in our dataset, we compute the savings in terms of regulatory capital which could be achieved by using more advanced techniques, in particular XGBoost as the most efficient model in this study, compared to a benchmark extensively used in the industry nowadays, such as Lasso.

The fact that we observe potentially significant capital savings due to a better statistical performance of advanced ML tools leads us to conclude that further research is needed in the area of supervisory risks in model evaluation. There seems to be an optimal decision to be taken on model selection, which will not depend only on the predictive performance, but also on the implicit costs observed to get the approval from the supervisor due to the risks embedded in the implementation of this technology.

The paper is organized as follows. Section 2 provides a literature review on the use of ML models for credit default prediction. Section 3 explains the data and the models used in the analysis. Section 4 contains the comparison of the predictive power, in terms of classification and calibration, for the six chosen ML models. In section 5 we show the economic impact of using XGBoost for calculating the risk-weighted assets and regulatory capital requirements. Section 6 concludes.

2. Literature review

There is an extensive empirical literature on the use of ML models for default prediction in credit risk. We have methodically reviewed it in order to find those papers that compare the predictive power of ML models with the predictive power of a logistic regression (Logit) for default prediction of loans. These loans could be either mortgages, corporate loans, or retail exposures. We can separate the literature in different strands, depending on the main ML method used. We consider papers that use tree based methods (either classification and regression trees or ensembles like random forest, boosting or XGBoost), neural networks methods (either deep neural networks or convolutional neural networks), and papers that compare several methods. Among the papers that use mainly tree based methods, one of the first attempts was Khandiani et al (2010), who tested the performance of classification and regression trees for predicting credit card delinquencies using data from a bank´s

[2] Any policy decision should take into account the potential positive impact of using ML and big-data on financial inclusion (see Barruetabeña 2020, Huang et al 2020), as well as the possibility of having negative side effects on social discrimination (Bazarbash 2019, Jagtiani and Lemieux, 2019) due to the better classification performance of ML models (Fuster et al, 2020).

customer base from 2005 to 2009. Butaru et al (2016) also focused on credit card delinquencies prediction, using a large dataset collected by a US financial regulator over a period of 6 years. They found that random forests can have gains of 6% with respect to Logit in terms of recall. Other papers have used ensemble tree based methods to predict corporate loans default. Petropoulos et al (2019), collected data on corporate and SME loans from Greece from 2005 to 2015, and found that random forests can have gains with respect to Logit of 8% in terms of AUC. Sigrist and Hirnschall (2019) used data on SME loans from Switzerland, and showed that a model that combines Tobit and gradient tree boosted can have up to 17% gains over Logit in terms of AUC. And Moscatelli et al (2019) used a dataset of Italian non-financial firms from 2011 to 2017, and found that ensemble tree methods could yield gains over Logit of 2.8% in terms of AUC. On the other hand, there are papers that have used mainly deep learning or convolutional neural networks to predict credit default. Turiel et al (2018) collected loans from Lending Club that covered the 2007 to 2017 period, including consumer loans, corporate loans, etc. They found that deep learning could have gains of 12% in terms of recall over Logit. Sirigniano et al (2018) developed a deep learning model to predict default on a sample of 120 million mortgages from US between 1995 and 2014, and show that deep learning can have gains from 6% to 20% in terms of AUC with respect Logit, depending on the definition of default. Kvamme et al (2018) also collected data on mortgages, specifically more than 20,000 Norwegian mortgages approved from 2012 to 2016. They show that the use of convolutional neural networks can yield gains of 6% in terms of AUC with respect to Logit. Finally, Albanesi and Vamossy (2019) used a combination of deep neural network and gradient boosted trees to predict consumer default form the Experian credit bureau, from 2004 to 2015, and concluded that deep learning performs significantly better than logistic regression, with gains up to 5% in terms of precision. While the majority of the papers reviewed focus on one or few ML models we found some that compare the predictive power of an ample class of ML methods, highlighting Jones (2015) and Guegan and Hassani (2018). All aforementioned papers find that ML models outperform Logit in classification power. This is true regardless of the technique used and the type of underlying asset of the study. However, these results are very heterogeneous.

Our contribution to this literature is that we are able to assess robustly the predictive performance of a wide range of ML methods under different circumstances (different sample sizes, and different amount of explanatory variables, as shown in Section 4) using a unique database on consumer loans granted by a big Spanish bank. Unlike the aforementioned comparisons in the literature, this allows us to test whether the statistical performance comes from an information advantage (associated to the access to big

amounts of data) and/ or model advantage (associated to ML as high end technology) when comparing these innovative tools to traditional quantitative models, as suggested by Huang et al (2020). We find that there exists a model advantage on top of information advantage. Our results are in line with Huang et al (2020), who use a dataset from the Fintech industry in China, but we have used a different approach. They computed the information advantage by discretionally dividing the features into two sets: traditional vs innovative explanatory variables. This way they observed the performance of ML models under both datasets. They found that, within each model, using all the available variables yields better performance, concluding that this is an indication of the existence of information advantage. In our case, instead of discretionary separating the sample into two, we have measured the performance of ML models under both dimensions of the information space MxN, where M represents the number of observations (length) and N the number of features (width). In particular we perform simulations with a random selection of features and sample sizes. This allows us to add statistical robustness to the conclusion that a model advantage exists on top of information advantage, capturing a broader concept of information advantage (and therefore, a finer model advantage definition) when compared to Huang et al (2020).

We also contribute to the literature by assessing the economic impact of using ML for credit default prediction. Khandani et al (2010) and Albanesi and Vamossy (2019) computed the Value Added (VA) as the net savings to lenders of granting credit lines to borrowers based on the (better) predictions of ML models. This method, while useful, has its drawbacks. First, it is limited to the assessment of ML for credit scoring, while we aim to evaluate models in more areas susceptible to implement this technology in the banking industry. Second, it might be considered backward looking metric, as it is estimated using a randomly chosen subset of the loans or credit lines of the outstanding portfolio, assuming that some of them could be granted or cut retrospectively[3]. We, instead, propose to monetize the impact through the comparison of calculated risk-weighted assets and capital requirements under a baseline scenario, using Lasso[4], against a scenario in which the banking institution would have chosen to implement a more statistically efficient model, like XGBoost. We show that the latter scenario can yield savings from 12.4% to 17% in terms of regulatory capital requirements under an IRB approach, depending on the corresponding risk components in the Basel formulae associated to our type of exposure or underlying assets. This is, to the best of our knowledge, the first attempt to measure the impact of using ML methods in terms of regulatory capital savings. This impact could be interpreted as a floor amount,

[3] See section 5.1 for an explanation of this method.

[4] Although in the literature review the comparison in the evaluation has been performed using Logit as benchmark, we assume for this exercise that currently the use of a logistic penalized regression with Lasso is common practice in the banking industry.

since it does not account for the potential benefit of using the model for new business originated. In contrast, while conservative, this estimated amount could be immediately materialized by the credit institution, complementing the exercise that might be additionally carried out through the estimation of the VA.

3. Data collection and ML models

An anonymized database from Banco Santander has been used to conduct this analysis. It contains data from a subset of consumer credits, granted by the aforementioned bank in unspecified dates. This data has been completely and previously anonymized by Banco Santander through an irreversible dissociation process in origin which prevents the possibility of identifying costumers in any way. 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 (features) associated to it, whose labels or description have not been provided. Consequently, the nature of these variables is unknown to us, and they cannot be used to establish the identity of the customers they refer to. Out of 370 features, 105 are binary variables (only two different values), 99 have 3 to 5 different values, 34 have 6 to 10 different values, and 132 have more than 10 values[5]. Around 3.95% of the loans resulted in default, but the data has no temporal dimension, so we do not know when the loan was granted, and if resulted in default, when it happened.[6]

As mentioned in the introduction, we will firstly compare the predictive performance of Logit vs several ML models. In particular, we have chosen Lasso penalized logistic regression, Classification and Regression Trees (CART), Random Forest, XGBoost and Deep Neural Networks[7] because they are amongst the most cited ones in the literature review.[8] We have conducted our analysis using Python and open source libraries like Sklearn and Keras. The hyper-parameters have been chosen according to standard cross-validation techniques, as the purpose of our exercise is neither feature engineering nor optimization, but comparing between correctly calibrated models. We have divided our data into training (80%) and test (20%) set, and we have used a five-fold cross-validation on the training set

[5] Since there are a considerable number of binary variables in our sample, we have performed robustness exercises by removing binary variables with low variance (binary variables which have the same value 80% of the times) and by creating additional dummy variables to account for the binary variables. The main results do not change significantly when performing these transformations.

[6] Therefore, we will focus on the estimation of probabilities of defaults point-in-time, leaving out of the scope of this work any assessment on the impact of macroeconomic variables that could explain observed defaults through-the-cycle.

[7] Our benchmark Neural Network has 5 layers, with 3 hidden units of 300, 200 and 100 neurons. We have selected this architecture after implementing the proper cross-validation and hyper parameter tuning. Our main results are not significantly affected by choosing other variations of Neural Networks.

[8] For an introduction into the functioning of each model, please see WB (2019).

to choose the hyper parameters that maximizes the out-of-sample AUC[9]. As it is common in the literature, input values have been standardized by removing the mean and scaling to unit variance[10].

4. Predictive performance

To assess the predictive performance of the 6 ML models we will focus on two measures: classification and calibration. Classification means the ability of the model to discriminate defaulted loans from those that have been repaid, being able to classify them in different risk buckets. We will use the AUC-ROC *or Area Under the Curve of the Receiving Operating Characteristic* (Fawcett, 2005) in order to measure the discriminatory power (BIS, 2005).[11] On the other hand, calibration refers to the quality of the estimation of the probability by looking, per bucket, at how good the average estimated probability fits the observed default rate. To this purpose we will use the Brier score (BIS, 2005) to measure how precise the estimations are, along with calibration plots, in particular reliability curves, in which we will divide the predictions into groups, and for each one we will compare precisely the average estimated probability of default with the corresponding observed default rate. For both measures, we perform a sensitivity analysis in the two dimensions of the information space (MxN), simulating the impact in the AUC-ROC and Brier score of the models for different sample sizes (M), and for different number of available features (N).

The reason why we have decided to use these two measures to pursue the evaluation of the performance of the ML models is that they are explicitly mentioned in the supervisory process for the validation of IRB systems, which we find to be the most complete framework to understand the potential and limitations of these predictive models when applied in particular to regulatory capital calculation, and generally to credit risk management (Alonso and Carbó, 2020). In this sense, for a supervisor there are two separated phases when evaluating the adequacy of an IRB system. First, a supervisor should carry out an assessment of the design of the rating system. In the calculation of regulatory capital, institutions have to estimate several risk factors, like the Probability of Default (PD), Loss-Given-Default (LGD), or even credit conversion factors or maturity adjustments. As a general rule, institutions have to provide their own estimates of PD and rely on standard values for other risk components. In this paper we will assume this is the case.[12] In this sense, the

[9] Among the hyper parameters, we have chosen the depth of trees for CART and the number of trees and depth of trees for Random Forest and XGBoost. For neural networks, we use *Talos* to choose the optimal number of hidden layers, nodes, activation functions and optimizers.

[10] Our results do not change ostensibly if we use input values without standardization, but standardizing them helps to reduce computing time, especially in the case of deep neural networks.

[11] There are other metrics that evaluate the performance of a classifier, like F1, Gini index, recall, precision and accuracy. We choose AUC because is the most used metric across the papers we reviewed and one of the most popular metrics in the literature (Dastile et al, 2020). We will additionally use recall as a robustness check.

[12] As further explained in Section 5.1.

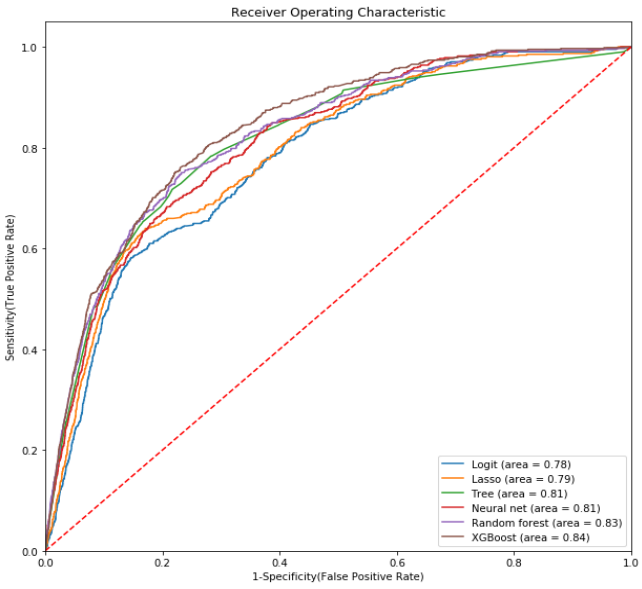
estimation of the PD is a two-folded task. First, institutions are required to identify the risk in different buckets, discriminating those exposures which are riskier from the rest. Secondly, the risk must be quantified. To this purpose, the buckets must be well calibrated, resembling the observed default rate. Once the risk factors are estimated, they will be plugged into an economic model as inputs in order to compute the (un)expected losses, which in the case of minimum capital requirements, comes from the Basel framework.

In sum, to understand the benefits of ML models applied to estimating PDs, it is not enough to evaluate the models in terms of discriminatory power, but we must get a grasp as well on the calibration performance. Once this work is done, supervisors will get deeper into the rating process, which usually includes an investigation on the data sources, privacy of the information and quality of the data sets, technological infrastructure required to put the model into production, and its governance, all subject to the use that each institution gives internally to these models.

4.1. Classification

In this section we use the AUC-ROC to study the discriminatory or classification power of the selected models. As shown in **Figure 1**, this curve plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds.

Figure 1. Comparison of AUC-ROC per model



$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Where TP (true positives) are the loans that, having defaulted, are correctly predicted as such, FN (false negatives) are the loans that, having defaulted, are incorrectly predicted as non-default, FP (false positives) are the loans that did not default but were predicted as default, and TN (true negatives) are the loans that did not default and were correctly predicted as non-default. For each threshold, if a loan has a probability of default higher than such threshold, then we classify it as defaulted. Therefore, the lower the threshold, the higher the TPR and the lower the FPR (upper right of the AUC curve). On the other hand, the higher threshold, the lower the TPR and the higher the FPR (bottom left of the AUC curve). In **Figure 1** we plot the estimated out-of-sample AUC-ROC curves for the six models. The curves show a nonlinear trade-off between TPR and FPR. The discriminatory power is given by the area under the curve. For reference, we plot a dotted 45 degrees line. This line yields an area of 0.5, and it represents a decision rule that categorizes randomly a binary response. The further up from the red dotted area, the more classification power the model would have. In our estimation Logit achieves a 0.78, Lasso 0.79, CART 0.81, Deep Neural Net 0.81, Random Forest 0.83, and XGBoost 0.84. The results are in line with our previous findings on the literature review, which suggest that ML models have better predictive performance than Logit, but deep neural networks do not necessarily outperform tree based methods. These results do not depend on the particular train-test partition we used to train and calibrate our models. In the Appendix we include an exercise in which we show the average AUC for each ML model from 100 simulations with different train-test partitions, and the differences among models remain the same[13].

From a credit institution point of view, the cost of a FP (not granting a loan to a performing counterparty) will presumably have a smaller impact on the benefits than the importance of getting a TP correctly (not granting a loan to a non-performing counterparty). Therefore, model selection rules for credit scoring would usually take into account that the actual cost of having a FN outweighs the opportunity cost of a FP. From an economic point of view both are not equally important, however, standard computational loss functional (like cross-entropy) usually treat both symmetrically. Bearing this in mind, on the traditional ML literature usually it is used the confusion matrix as a 2x2 contingency table to evaluate the performance of the algorithms, visualizing the TP, FP, TN, FN for both the actual and predicted classes. In our case, we propose a separate exercise in which we compare the TPR or recall of each of the ML models due to the fact that, for credit rationing, FN are way more important economically (Abdou and Pointon, 2011). Therefore, it is suggested to prioritize the analysis on the vertical axis of the ROC-AUC chart. In order to compare the

[13] In each simulation we use all the sample and we only change the train-test partition. Differences between the averages AUC of all models are statistically different at 95 confidence interval according to the corresponding Student´s *t*-test.

TPR, we need to specify which threshold we consider to decide when a loan will default or not. We choose from 10% to 30% thresholds, since a loan with an estimated default probability of ca. 10% is associated with speculative grade and 30% corresponds to average default rates observed in ratings at least CCC+ by Standard & Poors and Moody's (Cardoso et al 2013). Both levels therefore might well be representative of early warnings or limits when deciding to grant a loan in any credit scoring system. The results are in **Table 1**.

Table 1: True Positive Rate for different classifier thresholds

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%

Table 2: False Positive Rate for different classifier thresholds

Method	FPR, Classifier threshold = 10%	FPR, Classifier threshold = 20%	FPR, Classifier threshold = 30%
Logit	7%	1%	0.3%
Lasso	7%	1%	0.3%
Tree	8%	2%	0.3%
Random Forest	11%	2%	0.3%
XGBoost	10%	3%	0.3%
Deep learning	9%	2%	0.3%

With a classifier threshold of 10%, the ranking of ML models in terms of TPR is the same as in our benchmark exercise, when we compared ML models in terms of the AUC metric.

XGBoost and Random Forest have the highest TPR, around 55%, followed by Deep Learning, Tree, Lasso and Logit. If we consider a classifier threshold of 20% or 30% instead, then XGBoost continues to be the ML model with highest TPR, but Random Forest falls behind Deep Learning and Tree. In any case, for each possible classifier thresholds, ML models outperform again traditional methods like Logit. From **Table 1** we can see that resulting TPRs are relatively small (never above 60%) even for low classifier thresholds. This happens because in our dataset there are approximately 3.95% of defaulted loans. While it is not a heavily imbalanced dataset, we test the robustness of our results by performing two additional exercises in which we balance our dataset, first by giving more weight in the loss function to observations which defaulted, and second, by performing oversampling techniques. The results are in the Appendix. The main conclusion is that these rebalancing techniques do no change the main results from our benchmark exercise, and the ranking among algorithms remain the same. Taken this into consideration, we continue to work with our original dataset.

For completeness sake, we can also evaluate the performance of ML models in terms of False Positive Rate or FPR. The results are in **Table 2**. Interestingly, Lasso and Logit are, for all three thresholds, the methods with the lowest FPR, although differences in FPR are much smaller than differences in TPR. This indicates that Lasso and Logit tend to have fewer predictions with PD above the initial threshold of 10%, but still, differences in TPR by more advanced ML models offset the small differences in FPR.

We then analyse if the model´s classification power depends on the number of observations and features available. Our aim is to statistically isolate any information advantage due to a better access to big amounts of data from a hypothetical model advantage. We consider information advantage might come from the access to a larger MxN dataset, where M is the number of observations (length) and N the number of features (width). This definition is slightly different to Huang et al (2020), who only considered information advantage the one derived from the use of more features. To this purpose, first we compare the classification performance of each model for different sample sizes. We perform 400 simulations, and for each of them a random number of loans, from 1,000 to 65,000[14], is selected. In **Figure 2** we show the area under the curve of each model for different sample sizes, so that we can find the model with the best classification performance depending on the sample size. Random Forest and XGBoost outperform the rest of the models when 5,000 observations or more are included. It is often believed that algorithmically complex ML methods surpass

[14] Randomly we select a number from 1,000, 5,000, 10,000, 15,000 up to 65,000 (12 groups in total, around 33 simulations per group).

traditional linear models because they can handle a larger amount of data (the so-called, information advantage). But this exercise shows that, given the same amount of data, Random Forest and XGBoost always exceed the discriminatory power of Logit or Lasso thanks to the non-linearity nature of their algorithms, even when a relatively small amount of data (5,000 observations) is used. In this sense, ML techniques offer a model advantage, adding value on top of what traditionally might be understood as *big-data*. All six models experience an increase in their classification performance when more observations are included, but the slope of this gain is smaller for Logit (blue line) and Lasso (yellow line) from 10,000 observations onwards. This means that traditional quantitative models do not benefit as much when more data is available. On the other hand, Logit and Lasso can outperform the rest of models when the sample is 5,000 loans or less. **Figure 2** also shows that Deep Neural Networks only outperform Logit and Lasso when more than 20,000 observations are available. In the Appendix, **Figure 15** shows these simulations for the six models along with their 95% confidence intervals.

Secondly, we compare the classification performance of each model for different number of available features, in order to isolate the second dimension of a potential information advantage[15]. We perform again 400 simulations in which we select all available observations (75,000 loans), but in each simulation we choose a random number of features, from 125 to 375[16]. **Figure 3** shows the AUC-ROC of each of the models for different number of features. The area under the curve increases for every model as we increase the number of features available. The magnitude of the increase is very similar for all the models, except the CART that seems to benefit the most from the inclusion of more features. Again, Random Forest and XGBoost show the best results, since they outperform the rest of the models regardless of the number of features chosen. We also show in **Figure 16** (in the Appendix) these simulations along with their 95% confidence intervals.

[15] Assuming that any information advantage may come from the access to a larger MxN dataset, where M is the number of observations (length) and N the number of features (width).
[16] Randomly we select a number from 125, 150, 175, up to 375 (12 groups in total, around 33 simulations per group).

Figure 2. Simulation of AUC-ROC performance to sample size

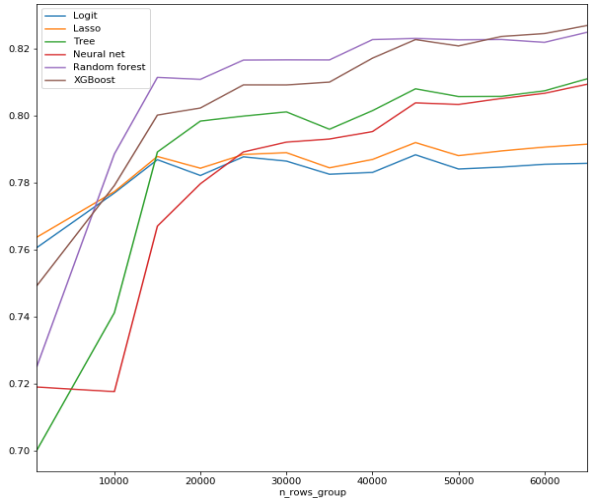
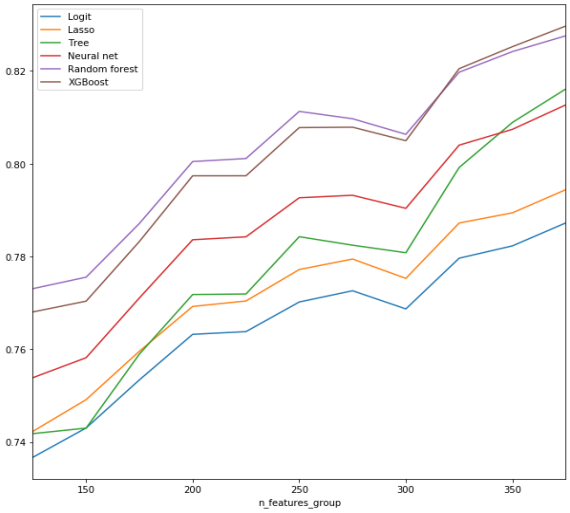


Figure 3. Simulation of AUC-ROC performance to number of features



4.2. Calibration

In this section we will use the Brier score to study the calibration power of the six ML models. We will also use reliability curves in which we will divide the predictions into groups (buckets) depending on their estimated probability of default, and for each group we will compare the average estimated probability of default with the rate of defaulted loans observed over total loans in that group.

The Brier score is a key measure to quantify the accuracy of a probability forecast (BIS, 2005). The formula to compute this metric is:

$$BrierScore = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_{t)}{}^2$$

Where *N* is the number of observations, *f* is the predicted probability of default, and *o* is the class of the observation (1 if default, 0 otherwise). We perform the same two exercises as with classification: we compute the Brier score for different sample sizes (from 1,000 to 75,000) and for different number of features (from 125 to 375). **Figure 4** shows for each model the resulting box plots from 400 simulations with different sample sizes. It can be seen that for most of the simulations and models, the Brier score is within a range of 3% and 4.4%. Differences among models are very small. Brier score values are small and similar among models due to the fact that there are only 3.95% defaulted loans in the whole sample. For illustrative purposes, if we were to consider the whole sample, and we used a model that assigns probability of default equal to zero for all the loans, the Brier score of

that simple model would be 3.95%. Still, we can see in **Figure 4** that the models with the lowest average Brier score are Random Forest and XGBoost. The six models have a Brier score of 3.7% for Logit, 3.6% for Lasso, 3.5% for the CART and Deep Neural Network, and 3.4% for XGBoost and Random Forest when the whole sample is used.

Figure 4. Sensitivity of Brier score to sample size

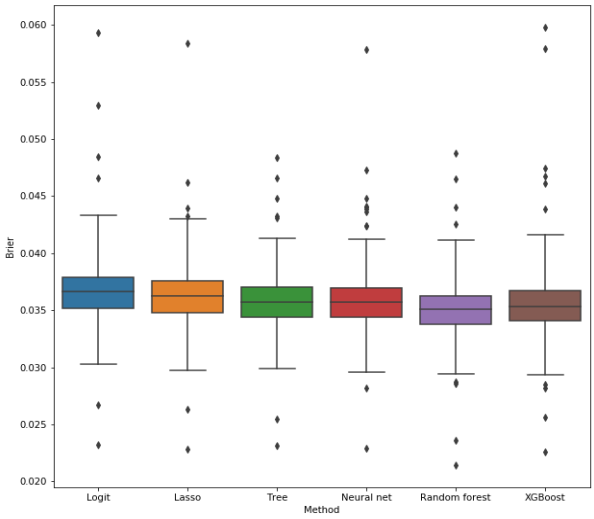
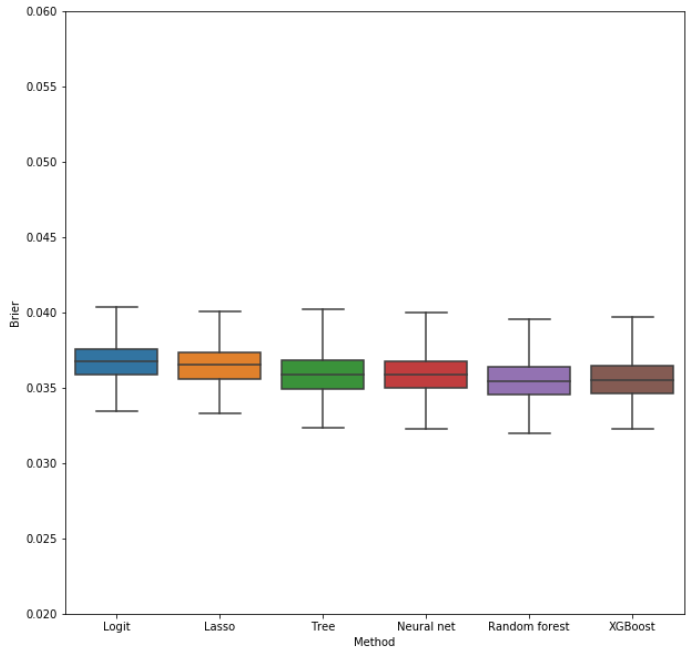


Figure 5 shows for each model the resulting box plots from 400 simulations with different number of features available. It can be seen that for most of the simulations and models, Brier score is within a range of 3.3% and 4%. We observe that Brier Scores are more homogeneous across simulations when changing the number of features than when changing the number of observations. This might be because now each simulation has all observations available, so the amount of defaulted credits is the same across simulations, while in **Figure 4** the percentage of defaulted credits might differ from one simulation to another. In any case, XGBoost and Random Forest also achieve the lowest Brier scores when changing the number of features available, reinforcing the existence of a model advantage from a calibration point of view.

Figure 5. Sensitivity of Brier score to number of features



Since differences in Brier Score are so small among models, we propose the additional calibration exercise. We run 200 simulations for each model, only changing the train-test partitions, with all observations and all features available for each simulation. We have grouped the predictions of each model into 13 buckets[17][18], depending on the estimated probability of default. **Figure 6a** shows reliability curves for each of the models. There, the *x* axis represents the estimated probability of default of each bucket, and the *y* axis has the proportion of defaulted loans over total loans for each bucket. The 45 degrees line represents a perfect calibration. For example, a perfect calibration would imply that a bucket with 20% of estimated probability of default should contain a 20% of defaulted loans. All models seem to perform very similarly for the first two buckets. However, for predicted probabilities above 10%, the performance of the models differs. For predicted probabilities between 10% and 20%, Logit (blue line) and Lasso (yellow line) underestimate the probability of default with respect to the observed default. For estimated probabilities above 20%, all models overestimate the probability of default with respect to the observed default rate. But Logit and Lasso are the models that overestimate the most. On the other hand, XGBoost and Random Forest are the models that are closer to the 45 degrees line. Since these results are based on multiple simulations, we must take into account the variance of

[17] The choice of the number of buckets does not change the ranking between algorithms. In fact, in Figure 7a we do a classification with fewer buckets to focus on smaller probabilities, and the ranking between algorithms remains the same.
[18] The buckets distribution is as follows: Bucket 1 has loans with PD between 0% and 5%, bucket 2 PD between 5% and 10%, bucket 3 PD between 10% and 15%, bucket 4 PD between 15% and 20%, bucket 5 PD between 20% and 25% and bucket 6 PD between 25% and 30%. Buckets seven and above contains intervals of 10% of PD each, up to PD 100%.

the observations. **Figure 16** in the Appendix shows the same results of **Figure 6a** but in 6 subplots (one for each model) in which we display the 95% confidence intervals. This way we can understand better the accuracy of the calibration. It can be seen that for Logit and Lasso, the 45 degrees line lays out of the calibration points' confidence interval from the third bucket (around 12% of probability of default) onwards. The rest of ML models perform better, especially Deep Neural Network, Random Forest and XGBoost, for which the 45 degrees line always lays on the confidence interval for all buckets.

Figure 6a. Reliability curve

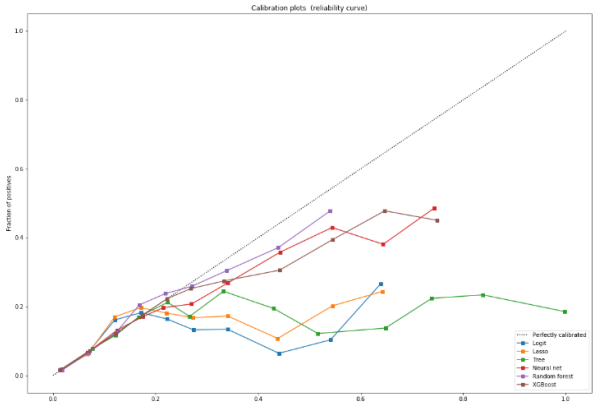
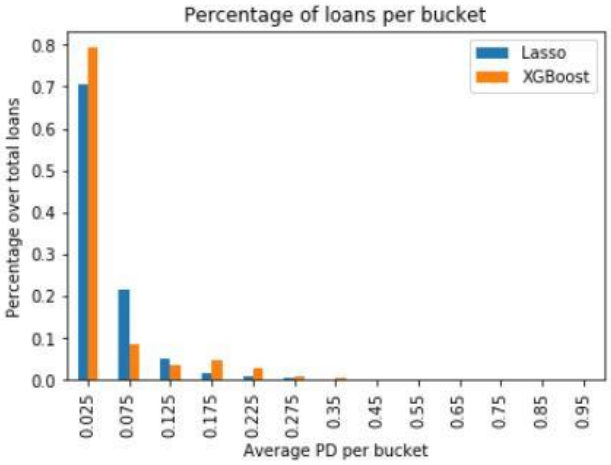


Figure 6b. Distribution of loans



Looking at **Figure 6a** it might seem that the difference in calibration power between XGboost and Lasso or Logit is higher than what their Brier scores suggest. This is explained by the fact that most observations have a probability of default below 10%. In **Figure 6b** we show the amount of credits in each bucket for Lasso and XGBoost. It can be seen that 80% of the credits have probabilities of default below 10%. Therefore, we must acknowledge that XGBoost, a priori, outperforms Lasso and Logit especially for probabilities above 10%, but there are fewer amount of credits in those buckets.

As most of the predictions have a probability of default below 30%, we propose another calibration plot in which we group the predictions into more granular buckets, focusing only in probabilities below this threshold[19]. This way we can assess the performance of the models for lower and more common probabilities in the field of credit default prediction. The results are shown in **Figure 7a.** It can be seen that Lasso and Logit tend to underestimate the probability of default for predictions up to around 3%, then overestimate the probability of default for probabilities from 5% to 7%, underestimate again for

[19] The new bucket distribution is as follows: Bucket 1 has loans with PD between 0% and 1%, bucket 2 contains PD between 1% and 2%, bucket 3 contains PD between 2% and 4%, bucket 4 contains PD between 4% and 6%, bucket 5 contains PD between 6% and 8%, bucket 6 contains PD between 8% and 10%%, bucket 7 contains PD between 1% and 12%, bucket 8 contains PD between 12% and 15%, bucket 9 contains 15% and 20%, bucket 10 contains PD between 20% and 30%, and bucket 11 has up to PD 100%.

probabilities between 10% and 20% (as suggested as well by **Figure 6a**), and overestimate for probabilities above 20%. The rest of ML models are closer to the 45 degrees line. In the appendix we show in **Figures 17-18** the results of **Figure 7a** but in six subplots (one for each model) and with 95% confidence intervals. It confirms that ML models like Deep Neural Network, Random Forest and XGBoost calibrate better. For referential purposes, in **Figure 7b** we show the amount of credits in each bucket for Lasso and XGBoost with this new categorization of buckets.

Taking everything into account, we can conclude that XGBoost and Random Forest clearly outperform the other models, especially in classification, but also in calibration, although this is definitely a more difficult task for all of them. Finally, CART and Deep Neural Network have similar performances, always above Lasso and Logit. The main conclusion of this section is that ML models outperforms Logit both in classification and in calibration, existing a model advantage that can be statistically isolated from an information advantage. Nevertheless, most complex models like Deep Neural Networks, do not necessarily predict better neither in terms of classification nor calibration.

Figure 7a. Reliability curve

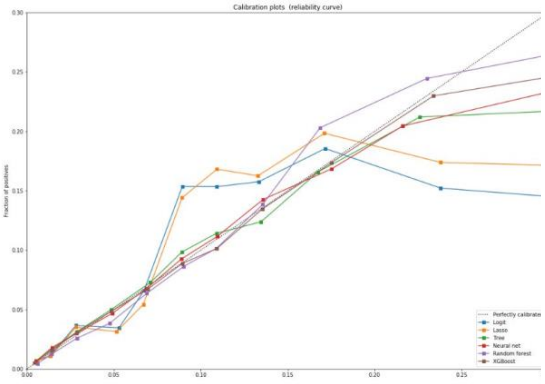
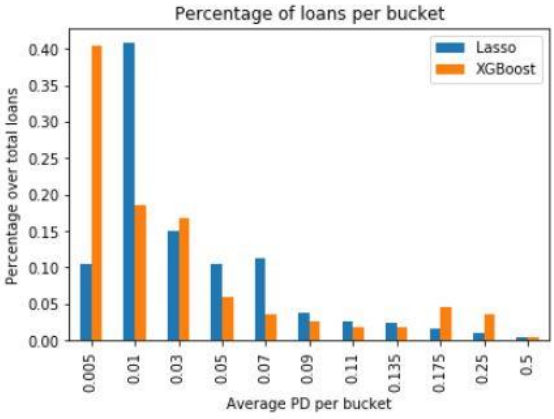


Figure 7b. Distribution of loans



5. Economic impact of using machine learning

In section 4 we found that ML models have better predictive power both in terms of classification and calibration than Logit or Lasso, regardless of the sample size and number of available features. In this section we wonder how to translate this statistical result into a real business metric. In particular, we aim to answer: which is the potential economic impact for a credit institution of using one of these ML models instead of traditional quantitative methods in credit default prediction in real business conditions?

One approach to measure the economic impact of better predictions is finding which loans could have been granted in case of counting with a better predictive model. This means either working out-of-sample, or using a subset of the portfolio and thinking retrospectively. This last approach is followed in Khandani et al (2010) and Albanessi and Vamossy (2019), who estimate the Value Added (VA) of using ML models by comparing the profits with and without forecast. In their model, the savings would be a function of the TP rate, indicating the correct decision not to grant a loan, which would be offset by the opportunity costs due to the lost return on those rejected loans because our model incorrectly expected them to default (FP). In this sense, it could be computed the VA in relative terms, comparing the savings when using a predictive model to the case of using a perfect-foresight strategy.

We understand that this approach, while valuable, has some limitations, as its computation relies on the assumption of working retrospectively on the outstanding portfolio, which means that no institution could anyhow materialize the process in the real world, as the VA might be considered backward looking metric. Therefore, we propose a novel approach to estimate the economic impact of applying ML in credit default prediction, which consists of calculating the potential savings in regulatory capital derived from using ML instead of a more traditional quantitative technique. This approach would complement the VA, and it has the advantage of being implementable after the loan has been granted, so credit institutions would be able to benefit from it immediately in real business conditions. As mentioned before, our dataset consists of consumer loans that have been already granted. These loans represent a credit risk exposure to the institution, with its corresponding cost in terms of regulatory capital. Assuming that the institution follows an IRB approach[20], we can calculate the difference in terms of regulatory capital between using a commonly used model nowadays like Lasso compared to using XGBoost, the model we found to be the most efficient in terms of predictive performance in our dataset. This measure would act as a floor or lower bound in the overall economic impact of using ML, assuming that at least any institution could benefit from reducing the capital requirements on their outstanding credit exposure, on top of which they could add the VA as estimated for instance by Khandani et al (2010), if any institution decides to implement a better predictive model on its new business strategy.

[20] Following CRE 30.42 "*For retail exposures, banks must provide their own estimates of PD, LGD and EAD*". Notwithstanding this, in our exercise we will analyze the impact of calculating the PD as the only risk factor to be estimated, using a standard value for the LGD, and leaving the EAD out of the scope of this work, as mentioned in the following Section.

5.1. Savings on regulatory capital

The pre-crisis (2008) regulatory framework provided credit institutions with a large degree of discretion in determining their capital requirements. This resulted in excessive variability in banks' capital requirements, which ultimately undermined the credibility of the risk-weighted capital framework at the peak of the global financial crisis. As stated in Bastos e Santos et al (2020), the Basel III post-crisis reforms developed by the Basel Committee sought to reduce this variability. To check this prerogative, the authors assess the degree of difference in modelled capital requirements across banks and over time. They observe that those credit institutions whose capital is closer to the minimum Tier1 ratios might be using more precise quantitative models to estimate their risk-weighted assets (RWA).[21]

In this sense, in Baena et al. (2005) it is explained how theoretically statistical models with better predictive power could yield a better outcome in terms of regulatory requirements. They showed that the Basel's risk weighted function for credit risk in the IRB approach is concave in the PD. This implies that the capital requirement for a group of assets increases as its PD increases, but each time less and less. If this holds true, a more granular common in prime portfolios) could affect disproportionally more the RWA than differences in loans with high PD. We are going to test these ideas by performing a step-by-step computation of the capital requirements for our dataset, using both Lasso and XGBoost for estimating the PD.[22]

Before starting the exercise, we summarize the key formulas needed to compute the capital requirements. The Basel framework specifies different formulas depending on the nature of the underlying assets which represent the credit exposure[23]. Since our data consists of consumer loans, we will use the formula of capital requirement *K* for retail exposures, which is calculated as follows (**Equation 1**):

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

Equation 1

[21] This result is consistent with previous findings, which in fact partly led to the Targeted Review of Internal Models (TRIM) back in 2015 from the European Central Bank (ECB), which derived lately in the IRB repair program performed by the European Banking Association (EBA), known as IRB roadmap (EBA, 2019).

[22] We will compute the K function using PD *point-in-time*, although the regulation CRD 2013/36 and CRR 575/2013 requires that the ratings represent a long term assessment of the risk of the underlying loans.

[23] These assets could be corporate, sovereign, bank or retail exposures.

Where *LGD* stands for Loss Given Default[24], *G* is the inverse cumulative distribution function for a standard normal random variable, *PD* is the average probability of default of the portfolio of assets, N stands for the cumulative distribution for a standard normal random variable, and *R* is the correlation. The formula for the correlation *R* is given by:

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

Equation 2

The Basel framework suggests different correlation formulas and values depending on the nature of the assets they refer to. We will use equation (2) for the computation of the correlation in our benchmark exercise, but we will consider additionally other possibilities for illustrative purposes like *R*=0.04 (for revolving retail exposures) up to *R*=0.15 (retail residential mortgage exposures), as mentioned further in this section, in order to account for the uncertain nature[25] of the retail type exposure in our dataset. Finally, the amount of risk weighted asset (*RWA*) can be computed as follows:

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

Equation 3

Where *EAD* is exposure at default measured in euros. We cannot compute this measure, as we don't know which feature corresponds to the outstanding credit balances or Exposures At Default (EAD) [26]. Therefore, we will focus on computing the savings of capital requirement *K* in relative percentage terms, using the number of loans per bucket as a proxy to weigh the size of the exposure.

In **Figure 8** we show the risk weighted assets (RWA) resulted from a series of PDs (from 0% to 20%) for three possible formulations of the RWA formula, **Equation 1:** With the term R as a function of PD as in **Equation 2**, with the term R fixed at 0.04, and with the term R fixed at 0.15. This way it can be seen the concavity of RWA with respect to the PD for different formulations. Notwithstanding this, in our benchmark scenario we consider the

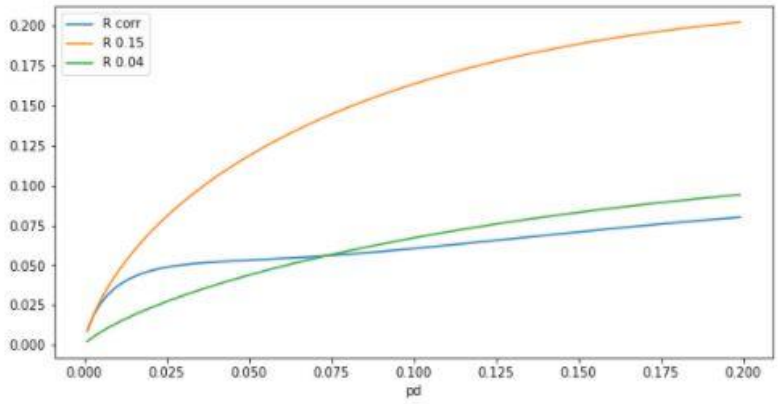
[24] We assume that the bank's estimate for LGD is 0.45 as baseline scenario. This is a standard value for any senior claims on sovereigns, banks, securities firms and other financial institutions that are not secured by recognized collateral (CRE32.6). In any case, different LGD values would not affect our comparison between XGBoost and Lasso, since changes in LGD would affect their capital requirement equally (see Equation 1).

[25] We ignore certain characteristics of the underlying credit, like either if there is any guarantee or collateral or the potential revolving structure of the loans.

[26] As stated before, we do not know the labels of any of our features.

term R as a function of the PD ("R corr" in **Figure 8**). In this scenario, RWA does not display concavity for PD between 5% and 12%, but it is concave and increasing in the rest of the domain of PD. For the other two formulations of RWA (with R fixed at 0.04 and 0.15), RWA is concave in the whole domain of PD. The degree of concavity of RWA in PD will have an important effect on the differences between Lasso's RWA and XGBoost's RWA. As we will see later, over estimating the PD or accumulating many loans in a specific PD interval can have very different repercussions depending on the relationship of the RWA with PD in that interval.

Figure 8. Shape of RWA function subject to parameter R.



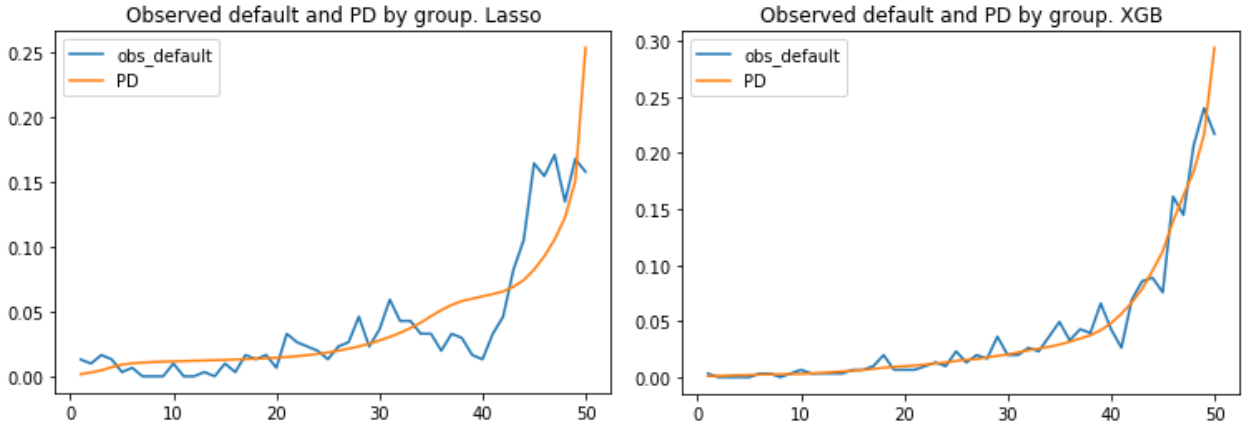
Step 1 – Discriminate between risk buckets.

Out of nearly 75,000 loans we use around 60,000 to train the models and we make predictions over the remaining 15,000 loans[27]. We first rank those 15,000 loans by their perceived credit risk. To this purpose 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[28]. The results are displayed in **Figure 9**, on the left hand side for Lasso and on the right for XGBoost. For both methods we show the average PD (orange line) and the observed default rate (blue line) for all 50 buckets. It can be seen that the estimated probability complies with the desired property of increasing monotonically in order to demonstrate discriminatory power. However, the divergence with the default rate per bucket suggests that a calibration process needs to be performed. This divergence is more significant for Lasso which first tends to overestimate for loans around 1% (from bucket 5 to 20). The distribution of PDs for Lasso around 1% is completely flat, so it accumulates a great mass of loans in a relatively small interval of PD. Then Lasso underestimates the default rate when

[27] Different train-test partitions do not affect the results of this section.
[28] There are around 300 loans per bucket.

PD is around 3% and 4% (which corresponds to buckets 25 to 30 of Lasso), overestimates for PD from 5% to 7% (buckets from 32 to 42 of Lasso), underestimates again for PD from 10% to 20% (buckets from 42 to 48 of Lasso) and finally overestimates when PD is higher than 20% (buckets 49 and 50 of Lasso). These results are in line with our findings in the calibration analysis of section 4.2, while XGBoost seems to adjust better the PD to the default rate in each bucket, its fit is not perfect and therefore needs further calibration as well.

Figure 9. Ranking PDs per model.

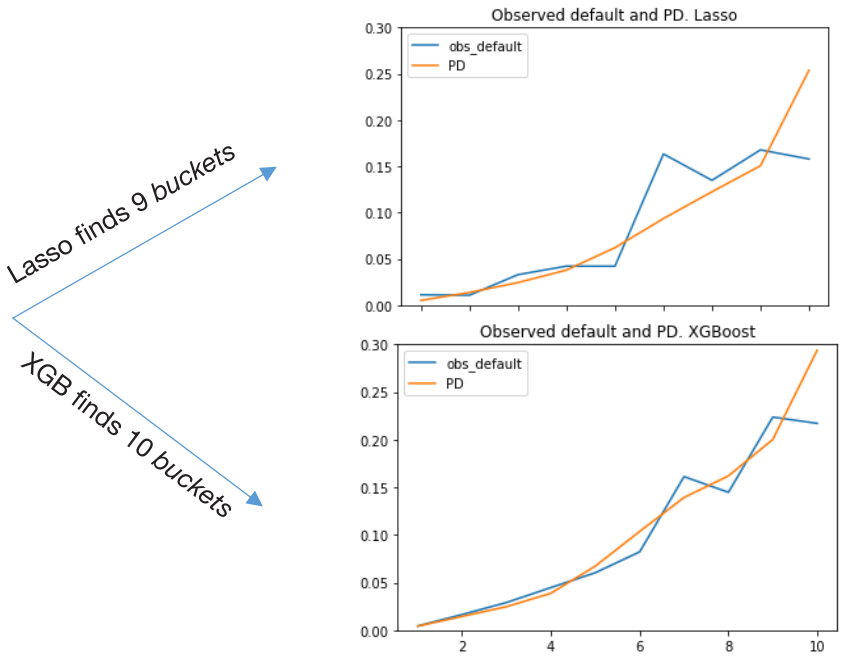


Step 2 – Calibration process.

In order to get approval from a supervisor, the classification resulting from the model must resemble the observed default rate. We propose in **Figure 10** an initial set of 10 rating grades based on the PD estimated, in order to fine-tune the calibration.

Figure 10. Initial distribution in rating buckets.

1. Lower than 1% - **AAA**
2. From 1% to 2% - **AA**
3. From 2% to 3% - **A**
4. From 3% to 5% - **BBB**
5. From 5% to 8% - **BB**
6. From 8% to 12% - **B**
7. From 12% to 15% - **CCC**
8. From 15% to 18% - **CC**
9. From 18% to 25% - **C**
10. Higher than 25% - **D**



For these rating notches to be approved by the supervisor, they must comply with two criteria: (i) risk heterogeneity between buckets, and (ii) risk homogeneity within buckets. This implies that risk categories must be different from each other (in our case, finding a PD which is monotonically increasing fulfils this requirement), while keeping consistency of risk level within each group. In **Figure 10** it is evident that the homogeneity criterion does not apply, as the difference between default rate and PD in each bucket is too high[29].

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 them. We do so by changing the thresholds that determine the buckets, as shown in **Figure 11** and **Table 3**, below:

Figure 11. Final distribution of rating buckets.

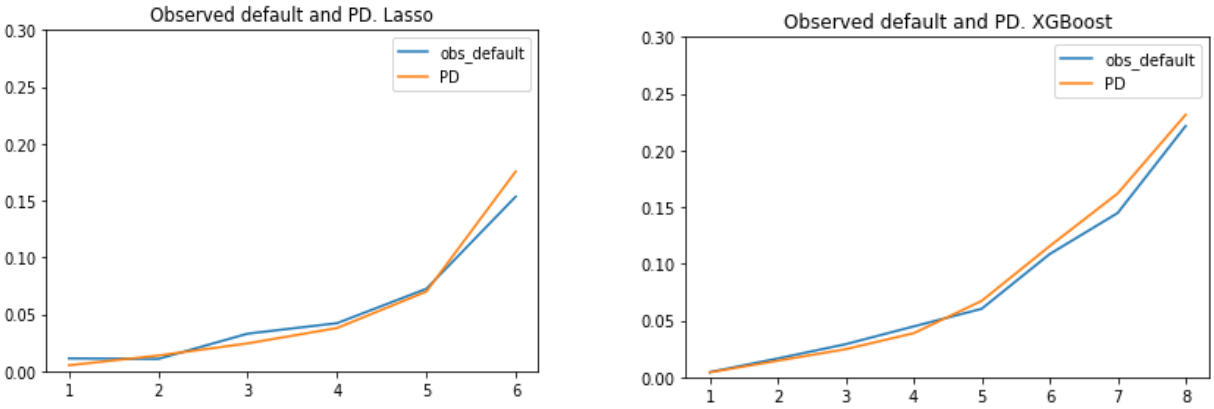


Table 3

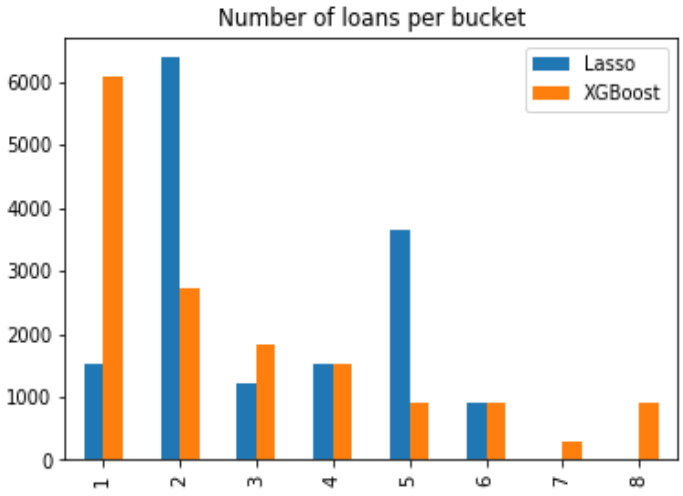
Initial ranking	XGBoost final ranking	Lasso final ranking
1. Lower than 1% - **AAA**	1. Lower than 1% - **AAA**	1. Lower than 1% - **AAA**
2. From 1% to 2% - **AA**	2. From 1% to 2% - **AA**	2. From 1% to 2% - **AA**
3. From 2% to 3% - **A**	3. From 2% to 3% - **A**	3. From 2% to 3% - **A**
4. From 3% to 5% - **BBB**	4. From 3% to 5% - **BBB**	4. From 3% to 5% - **BBB**
5. From 5% to 8% - **BB**	5. From 5% to 8% - **BB**	5. From 5% to 12% - **B**
6. From 8% to 12% - **B**	6. From 8% to 15% - **B**	6. Higher than 12% - **D**
7. From 12% to 15% - **CCC**	7. From 15% to 18% - **CCC**	
8. From 15% to 18% - **CC**	8. Higher than 18% - **D**	
9. From 18% to 25% - **C**		
10. Higher than 25% - **D**		

[29] We set the threshold of the homogeneity criterion in a maximum of 2% of difference between the PD and the default rate.

While Lasso allows us to identify 6 different buckets of riskiness, XGBoost allows a more granular classification, up to 8 buckets. When performing the calibration process, we choose discretionally the parameters (thresholds) that determine the buckets, but bearing in mind that we are restricted by the underlying distribution of PDs of Lasso and XGBoost, and its distance to the observed default rate, as shown in **Figure 9**. In the case of Lasso, the only way to accomplish this risk homogeneity criterion is by merging loans with estimated PD from 5% to 12% into a single bucket (bucket 5 of Lasso), and loans with estimated PD from 12% and higher into another bucket (bucket 6 of Lasso). The difference between PD estimated and the observed default rate for PD higher than 5% (buckets 34 and higher) is considerable. Therefore, the only way to achieve a proper classification of buckets is by merging probabilities between around 5% and 12% (buckets 34 to 47)[30]. Moreover, while the distribution of PD of XGBoost is always increasing, Lasso's distribution presents important flat areas, undifferentiated, which do not allow for further disaggregation (loans around PD = 1% in buckets 5 to 20).

Figure 12 shows the distribution of loans per final rating bucket, according to the thresholds of **Table 3**. The distribution of loans per bucket differ between each model. XGBoost has a more granular and smooth distribution over buckets. It allocates a significantly bigger amount of loans in bucket 1, and then the number of loans per bucket decreases smoothly, stretching overall the distribution of loans between buckets. Lasso, on the other hand, accumulates more loans in buckets 2 and 5. This happens because, as we showed in **Figure 9**, there are parts of the Lasso PD distribution which are completely flat, that is, around 1% and around 5-6%, much flatter than the XGBoost distribution.

Figure 12. Distribution of loans per final rating buckets.



[30] We could vary these thresholds slightly (5.5% instead of 5% and 11% instead of 12%) and still accomplish the criteria, but our main results will remain the same. The PD distribution delivered by Lasso does not depend of any particular the train test partition.

Step 3 – Calculation of capital requirements.

We can now assume that both rating scales would pass the supervisory test, allowing us to calculate the capital requirements per bucket. The capital requirement of each bucket is a function of its average PD, as it was shown in **Equation (1)**. In **Figure 13** we plot on the left hand side the average PD in each bucket for both XGBoost and Lasso, and on the right we plot the corresponding capital requirement. First of all, it can be seen that the higher the PD of a bucket, the higher would be the regulatory capital. But, as we showed in **Figure 8**, the relationship between the regulatory capital and the PD is concave, especially for buckets 1 to 5, where PD for both models is lower than 5%.

If we take the average capital requirement for each bucket (**Figure 13** right), and we weight it by the amount of loans[31] that are in the bucket (see **Figure 12**), capital requirements are 12.4% lower under the XGBoost rating scale than under the Lasso one. As both models have been calibrated, the overall average PD of each one does not significantly differ. Capital savings from the use of XGBoost therefore comes from two sources: the difference in the distribution of loans into buckets between models, and secondly, the difference in the number of buckets found within each model.

First, the difference in the distribution of loans into buckets between models. XGBoost has more loans with estimated probability lower than 5% than Lasso, buckets 1 to 4. In those buckets, the estimated PD for both models is almost the same by construction, since the thresholds were the same (**Table 3**) but Lasso accumulates more loans in bucket 2 than in bucket 1 due to the flat area of Lasso's PD distribution around 1.5% of PD. Roughly, the amount of loans that XGBoost allocates to bucket 1 is allocated by Lasso to bucket 2, and vice versa, since Lasso tends to overestimate the PD in that area (**Figure 9**, buckets 5 to 20). Therefore, the average PD weighted by the number of loans in bucket 1 and bucket 2 is higher in Lasso than in XGB, and this translates into considerable differences in RWA in the first two buckets. Consequently, Lasso allocates less loans in buckets 1 to 4 than XGBoost but needs more regulatory capital. On the other hand, Lasso has many more loans than XGBoost from bucket 5 onwards, i.e., more loans with higher than 5% PD. Independently of how Lasso and XGBoost split those loans, Lasso will charge more weighted capital in that section. Actually, the fact that RWA in our benchmark exercise is

[31] Ideally, we should compute the average capital requirement for each bucket and weight it by the amount of balances of the bucket. But since we have not been given the labels or descriptions of the explanatory variables, we do not know which of the explanatory variables refers to loan balances. Therefore, we weight the average capital requirement for each bucket by the amount of loans of each bucket. As we will show later, we believe that weighting the average capital requirement by the amount of balances will produce larger differences between Lasso's RWA and XGBoost's RWA.

not concave and almost non increasing between 5% and 12% of PD benefits Lasso, since it is allocating a considerable amount of loans in that segment[32].

Secondly, the difference in the number of buckets found within each model. The fact that XGBoost achieves after the calibration process a classification into a larger number of buckets than Lasso, implies, due to the concavity of the RWA function, a difference in capital requirements in its favour. The concavity shown in **Figure 8**, means that, being fixed the remaining inputs in the RWA formula, the following inequality holds true (Baena et al, 2005):

$$K(\lambda \cdot PD_1 + (1-\lambda) \cdot PD_2) > \lambda \cdot K(PD_1) + (1-\lambda) \cdot K(PD_2), \qquad with\ \lambda \epsilon (0,1)$$

Where K(.) represents the regulatory capital requirements function. Following this idea, we force the XGBoost classification into 6 buckets instead of 8, by merging buckets 6, 7 and 8 of XGBoost into a single bucket, in order to match exactly 6 buckets as it is the case of Lasso. In that case, the RWA would be between 0.4% and 1.1% higher (depending on the correlation formula of RWA), being this a rougher classification but which would have respected the supervisory calibration criteria as well. Similarly, we could further split XGBoost buckets and still accomplish the supervisory calibration criteria. If, for instance, we split bucket 1 into two additional buckets, RWA for XGBoost would be between 0.5% and 1.5% lower, depending on the correlation formula of RWA.

These two sources of capital savings would be affected by the concavity of the formula of RWA on PD. We have performed our benchmark exercise under the assumption that the consumer loans of our data would fall into the category of "other retail exposures" according to the Basel framework. Nevertheless, as mentioned before the loans of our dataset could fall into other categories, like "retail exposure with mortgages collateral" or "revolving retail credit exposures". Therefore, we have run a sensitivity analysis considering those possibilities, using their corresponding Basel formulas for the K function, with R=0.04 and R=0.15 (**Figure 8** shows the relationship between RWA and PD with those specifications). The savings in terms of regulatory capital range are 14% and 17% respectively for those two alternative scenarios. These savings are greater than in our benchmark scenario

[32] As we will show below, if we use alternatives of the Basel formula where the RWA is concave and increasing in the whole domain of PD (Figure 8), the difference between Lasso and XGBoost would be even larger.

because, if R is fixed, then RWA would be strictly increasing and concave when PD is higher than 5% as well. Lasso is allocating a considerable amount of loans especially between 5% and 12%, a segment where increases of PD mean big increases of RWA if R is fixed, resulting in a more pronounced effect in capital savings (first source of savings). Also, since the calibration of XGBoost has more buckets, savings with XGBoost are higher when the RWA is concave in the whole domain of PD (second source of savings).

For further clarification, we split each of the 6 buckets of Lasso and the 8 buckets of XGBoost into smaller groups of 304 loans, 50 groups in total. For example, the first 5 groups of 304 loans of Lasso would correspond to Lasso's bucket 1 (1,520 loans, see **Figure 12**), and the first 20 groups of XGBoost would correspond to XGBoost bucket 1 (6,080 loans), etc. To each group we charge the corresponding RWA according to the bucket they belong to. This way we can illustrate how the difference in RWA (stated in relative terms, therefore 100% corresponds to $RWA_{Lasso} - RWA_{XGBoost}$ for the total sample of 15,000 loans) compounds as we further include more loans in the assessment. The resulting chart is in **Figure 14** (left for the case with "R correlation", and right for the case of R = 0.15). With "R correlation", since the formula for RWA is concave and increasing on PD mostly for PDs lower than 5%, the majority of the difference in capital requirement is explained with the first 6,000 loans (roughly bucket 1 for XGBoost and bucket 1 and part of bucket 2 for Lasso). The small decrease at the end is due to the fact that XGBoost distributes the last 1,200 loans into buckets 7 and 8, with higher PD and higher RWA. However, when using R = 0.15, the RWA formula is also concave and increasing on PD for PDs higher than 5%, and we can see a second positive compounding effect between 7,600 – 12,000 loans (up to bucket 5 for Lasso, buckets 3 and 4 of XGBoost), which this time is not compensated by the last 1,200 loans, the ones distributed in buckets 7 and 8 of XGBoost.

Figure 13. Computation of regulatory capital

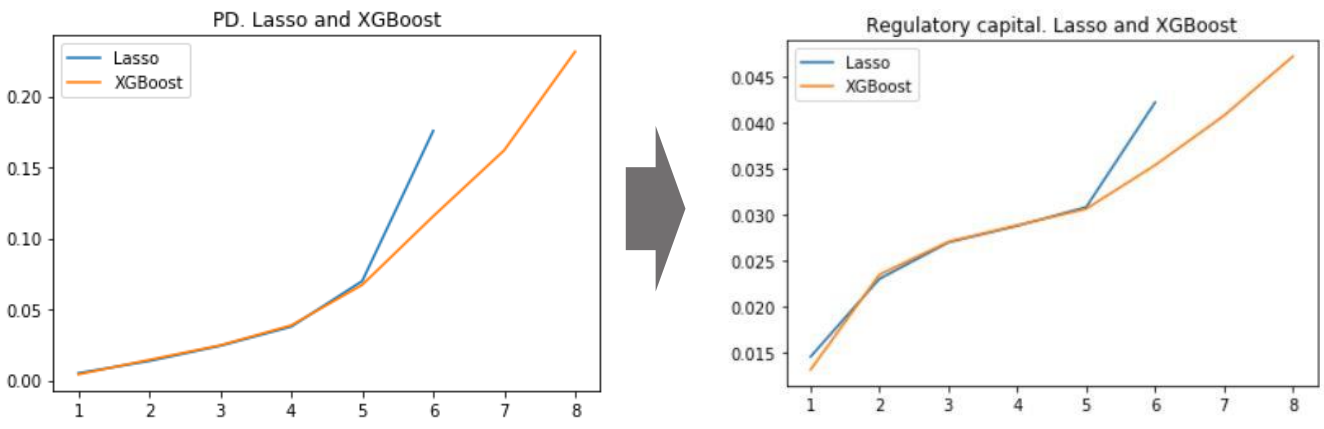
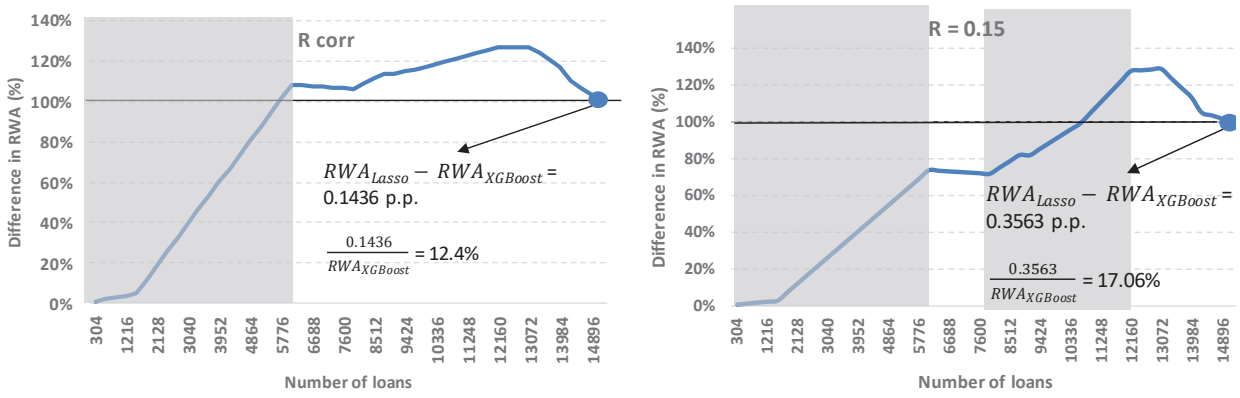


Figure 14. Compounding the difference in RWA



Summarizing, the fact that XGBoost is able to deliver a more granular distribution of loans and a smoother classification of loans per buckets of PD, allows it to deliver those capital savings. Ideally, we should weigh the average capital requirement of each bucket by the loan balance of the bucket. Unfortunately, we do not know which feature of our dataset corresponds with the loan balance. Jiménez and Saurina (2004) pointed to an inverse relationship between the size of the loan and the probability of default because larger loans are more carefully screened. Therefore, we assume that 12.4% is a conservative estimate of the savings in capital requirements, since the number of loans with high probability of default (e.g: PD above 10%) is higher for XGBoost than for Lasso. Moreover, XGBoost could have even smaller partitions for certain buckets, since its distribution of PD is much closer to the observed default than the one of Lasso, which would represent a potential further "within model" advantage when computing regulatory capital, ultimately depending on the assumption on the threshold for the risk homogeneity criterion in each bucket to be stated by the supervisor.

6. Conclusions

While institutions have been using internal models in the context of regulatory capital for a long time, the predominant techniques have not evolved significantly. Multivariate analysis and logistic regressions, like Probit or Logit, are effective tools to predict probability of default (Trucharte et al 2015), being currently common in the industry evolutions like the Lasso penalization. However, nowadays ML tools have the potential to be a game changer, as the technological progress and financial innovation has opened the room for implementing more advanced predictive models, leveraged on big data, advanced analytics and fostered by the push of newcomers into the market, which are implementing these kind of technologies in online platforms (EBA, 2018 and Huang et al, 2020).

In this environment supervisors face the challenge of allowing credit institutions and individuals to benefit from innovation, while at the same time respecting technological neutrality and ensuring compatibility with the prudential regulation and supervisory process. In this article we contribute to the literature in two ways. First, by showing a new way to robustly estimate the existence of a model advantage over an information advantage when using ML for credit risk. We perform our analysis using a unique and anonymized database from a major Spanish bank. Our results show that ML models perform better than the traditional Logit model, both in classification and calibration terms. While calibration is clearly a more difficult task than classification, XGBoost and Random Forest seem to provide the best results in both measures, despite not being the most algorithmically complex models (for instance, when compared to Deep Neural Networks). In order to test the robustness of our results, we perform a sensitivity analysis, simulating how the results would change in case of different number of observations and features, demonstrating that statistically it exists a model advantage on top of an information advantage. Secondly, we contribute by providing a novel approach to measure the economic benefits from using ML models in credit default prediction, through the calculation of the capital savings derived from their use. We simulate the gains in terms of savings that an institution would achieve if they were to use XGBoost compared to a more common Lasso penalized logistic regression. We estimate that these savings could amount to up to 17% of capital requirements in our benchmark exercise (retail exposure), which is a significant figure that lead us to suggest that more research is needed to understand the supervisory cost to get a model approval, based on the risks embedded (as in Alonso and Carbó, 2020). As mentioned before, predictive performance comes at a price, in particular in terms of risk model evaluation, which should be properly quantified too in order to better inform credit institutions and supervisors on the optimal model selection. In this sense, for any policy decision further research is needed on how to integrate macro-prudential effects of an industry wide implementation of ML models in credit risk management.

Bibliography

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), 59-88.

Albanesi, S., & Vamossy, D. (2019). *Predicting consumer default: A deep learning approach* (NBER Working Papers 26165).

Alonso, A., & Carbó, J. M. (2020). *Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost* (Banco de España. Documentos de Trabajo 2032).

Babaev, D., Tuzhilin, A., Savchenko, M., & Umerenkov, D. (2019). ET-RNN: Applying deep learning to credit loan applications. En *Proceedings of the 25 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2183-2190). New York, NY, USA: Association for Computing Machinery.

Bank for International Settlements. (2001). *The Internal Ratings-Based Approach: Supporting Document to the New Basel Capital Accord*.

Bank of England, & Financial Conduct Authority. (2019). *Machine learning in UK financial services*.

Bank for International Settlements. (2005). *Studies on the Validation of Internal Rating Systems* (Basel Committee on Banking Supervision Working Papers 14).

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance*, *72*, 218-239.

Cardoso, V. S., Guimarães, A. L. S., Macedo, H. F., & Lima, J. C. C. O. (2013). Assessing corporate risk: a PD model based on credit ratings. *ACRN Journal of Finance and Risk Perspectives*, *2, 51-58.*

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. *(2002).* SMOTE: Synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16, 321-357.*

Chen, H., & Xiang, Y. (2017). The study of Credit Scoring Model based on Group Lasso*.* En *Procedia Computer Science, 122, 677-684.*

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal, 91, 106263.*

European Banking Authority. *(2017). Report on innovative uses of consumer data by financial institutions.*

European Banking Authority*. (2018). Report on the Prudential Risks and Opportunities arising for Institutions from Fintech.*

European Banking Authority*. (2019). Progress report on the IRB roadmap: Monitoring implementation, reporting and transparency.*

European Banking Authority*. (2020). Report on Big Data and Advanced Analytics (EBA/REP 2020/01).*

Fawcett, T. *(2006).* An introduction to ROC analysis. *Pattern Recognition Letters, 27(8), 861-874.*

Fernández, A. (2019). Inteligencia artificial en los servicios financieros*. Boletín Económico,* 2, art. 7.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2018). *Predictably Unequal? The Effects of Machine Learning on Credit Markets* (mimeo).

García Baena, R., González Mosquera, L., & Oroz García, M. (2005). Aspectos críticos en la implantación y validación de modelos internos de riesgo de crédito. *Estabilidad Financiera, 9, 29-58.*

Guégan, D., & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring. *Journal of Finance and Data Science, 4(3), 157-171.*

Huang, Y., Zhang, L., Li, Z., Qiu, H., Sun, T., & Wang, X. (2020). *Fintech Credit Risk Assessment for SMEs: Evidence from China (IMF Working Papers 20/193).*

Institute of International Finance. (2019). *Machine Learning in Credit Risk Report.*

Jimenez, G., & Saurina, J. (2006). Credit Cycles, Credit Risk, and Prudential Regulation. *International Journal of Central Banking, 2, 65-98.*

Jones, S., Johnstone, D., & Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking and Finance, 56, 72-85.*

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance, 34, 2767-2787.*

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications, 102, 207-217.*

Laurent Dupont, A., Fliche, O., Yang, S., & Pôle Fintech-Innovation, A. (2020). *Governance of Artificial Intelligence in Finance (Banque de France. ACPR Documents de réflexion).*

Moscatelli, M., Narizzano, S., Parlapiano, F., & Viggiano, G. (2019). *Corporate default forecasting with machine learning (Banca d'Italia. Temi di discussione 1256).*

Non-paper - Innovative and trustworthy AI: two sides of the same coin. Position paper on behalf of Denmark, Belgium, the Czech Republic, Finland, France, Estonia, Ireland, Latvia, Luxembourg, the Netherlands, Poland, Portugal, Spain and Sweden on innovative and trustworthy AI. (2020).

Petropoulos, A., Siakoulis, V., Stavroulakis, E., Klamargias, A., & others. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. En *Bank for International Settlements. IFC Bulletins Chapters,* 49.

Santos, E. B. e, Esho, N., Farag, M., & Zuin, C. (2020). *Variability in risk-weighted assets: what does the market think? (BIS Working Papers 844).*

Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quantitative Finance, 19(9), 1449-1459.*

Trucharte Artigas, C., Pérez Montes, C., Cristófoli, M. E., Ferrer Pérez, A., & Lavín San Segundo, N. (2015). Credit portfolios and risk weighted assets: analysis of European banks. *Estabilidad financiera, 29, 63-85.*

Turiel, J. D., & Aste, T. (2019). *P2P Loan acceptance and default prediction with Artificial Intelligence. arXiv preprint arXiv:1907.01800.*

World Bank Group, & International Committee on Credit Reporting. (2019). *Credit Scoring Approaches Guidelines.*

Appendix

Average AUC from 100 simulations

In each simulation we use all data available and we only change the train test partition. The results are in **table 4**. Differences between the averages AUC of all models are statistically different at 95 confidence interval according to the corresponding Student´s t-test. The t-test is based on the following *T statistic*, built under the null hypothesis that two means of the populations are equal.

$$T\ statistic = \frac{Mean_1 - Mean_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where $Mean_1$ and $Mean_2$ are the mean values of each sample, s_1 and s_2 are the standard deviations of the two samples, n_1 and n_2 are the sample sizes of the two samples, and *n-1* are the degrees of freedom. With the T statistic value and the degrees of freedom, we can compute the corresponding *p-values* of every possible comparison of means. Results are in **Table 5.** All *p-values* but one are zero, meaning that the null hypothesis is rejected independently of the statistical significance. For the difference between Tree and Deep learning, the difference is also significant at 0.028.

Table 4: Average AUC

Model	Logit	Lasso	Deep learning	Tree	Random Forest	XGBoost
Average AUC	0.786	0.792	0.811	0.813	0.826	0.837
95% confidence interval	0.784, 0.787	0.791, 0.794	0.809, 0.813	0.812, 0.815	0.825, 0.828	0.835, 0.838
Difference with Logit	0	0.006	0.025	0.027	0.040	0.051

Table 5: p-value associated to each mean difference

Model	Logit	Lasso	Red neuronal	Tree	Random Forest	XGBoost
Logit	**X**					
Lasso	**0**	**X**				
Red neuronal	**0**	**0**	**X**			
Tree	**0**	**0**	**0.0284**	**X**		
Random Forest	**0**	**0**	**0**	**0**	**X**	
XGBoost	**0**	**0**	**0**	**0**	**0**	**X**

Data balancing techniques

Since defaults represent only 3.95% of observations in our data, we perform two data balancing exercises to test the robustness of our results. These balancing techniques are among the most popular ones in the literature (Dastile, 2020). First we scale the calculated loss for each observation by assigning a higher weight in the loss function on the observed defaults. The weight is computed in a way that the statistical losses associated with defaults and not defaults are balanced. The performance of the models in terms of AUC and TPR with classifier threshold 50% are in **Table 6.** We use a classifier threshold of 50% because once we balance the observations, TPR for low classifier thresholds are above 90% for every model.

Table 6: Weighted data set: AUC and True Positive Rate for different classifier thresholds

Method	**AUC**	**TPR, Classifier threshold = 50%**
Logit	78%	73%
Lasso	79%	73%
Tree	82%	74%
Random Forest	81%	66%
XGBoost	83%	75%
Deep learning	81%	72%

We can see that gains in AUC of the ML models with respect to Logit are similar to the ones we observed in the benchmark exercise, show in **Figure 1**. The main difference with this weighted dataset is that the performance of Random Forest is slightly worse than in the benchmark exercise. Regarding TPR, while XGBoost has the best performance again, Lasso and Logit having similar performance to the ML models and Random Forest obtaining a lower TPR.

Secondly, we balance our dataset by oversampling defaults with the Synthetic Minority Oversampling Technique (SMOTE). This is one of the most common methods to solve imbalance problems. It balances the class distribution by generating new examples of the minority class (for more details, see Nitesh Chawla et al 2002). We oversample in a way that we end up with a database with 25% of loans defaulted. The results are in **Table 7.**

Table 7: SMOTE oversampled dataset: AUC and True Positive Rate for different classifier thresholds

Method	AUC	TPR, Classifier threshold = 10%	TPR, Classifier threshold = 20%	TPR, Classifier threshold = 30%
Logit	79%	71%	37%	17%
Lasso	79%	71%	41%	14%
Tree	81%	68%	46%	34%
Random Forest	82%	69%	49%	31%
XGBoost	83%	65%	48%	34%
Deep learning	80%	70%	51%	40%

Again the results in terms of AUC are very similar to the ones of our benchmark exercise. The ranking of the algorithms and the difference of ML models with respect to Logit is the same. Regarding TPR, ML models outperform clearly Logit, especially for thresholds above 20%. The fact that Deep learning is the best performer in terms of TPR with this oversampled dataset stands out.

Figure 15. Simulation of AUC-ROC performance to sample size increase with 95% confidence intervals

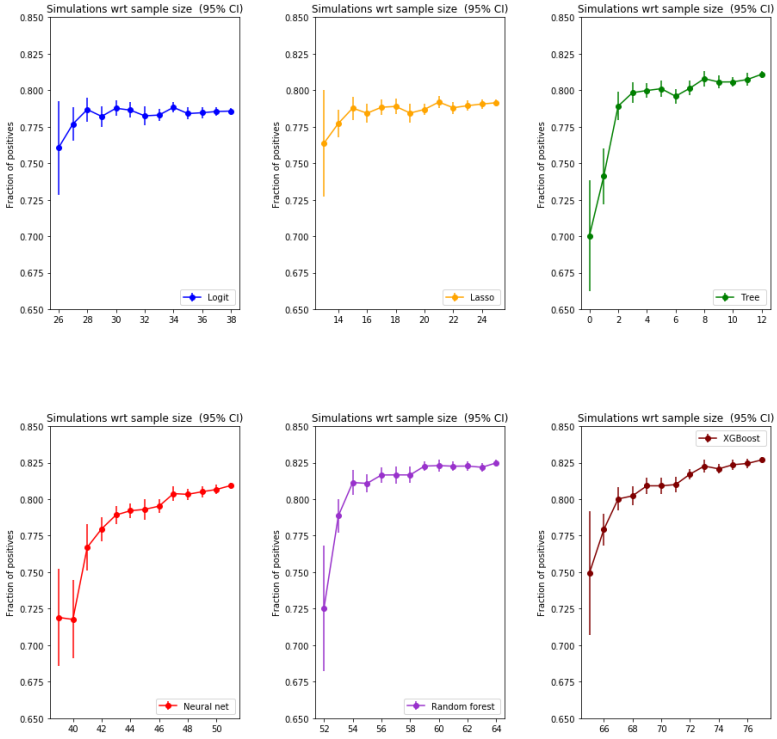


Figure 16. Simulation of AUC-ROC performance to number of features increase with 95% confidence intervals

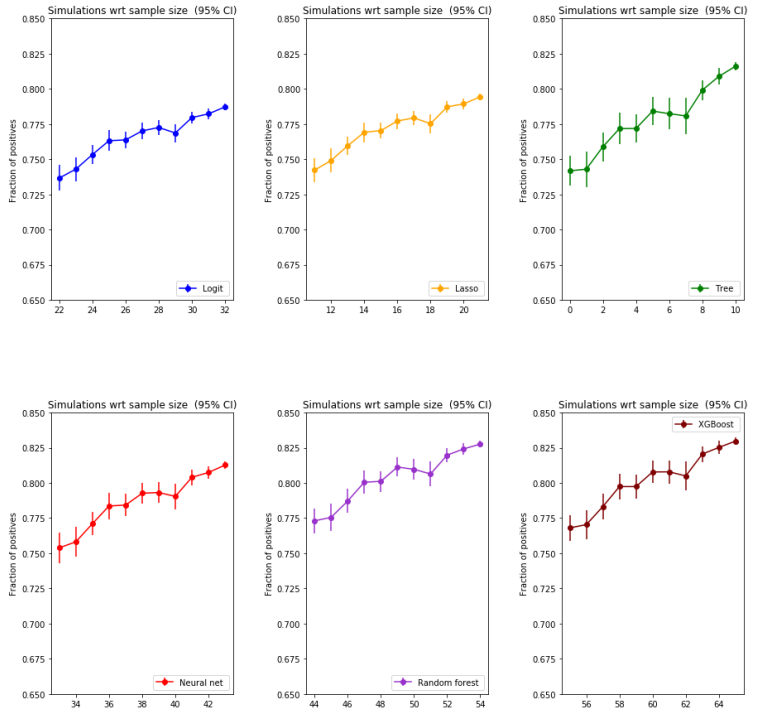


Figure 17. Calibration reliability curve with 95% confidence intervals

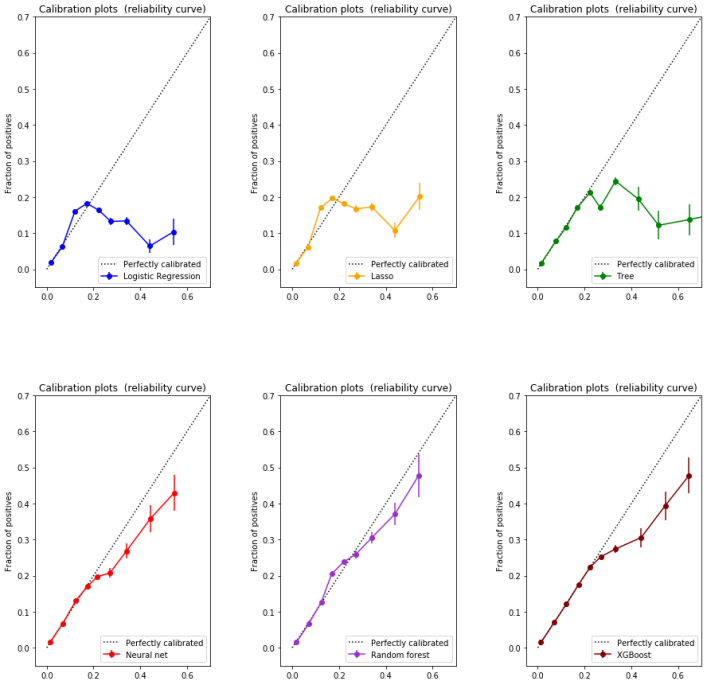
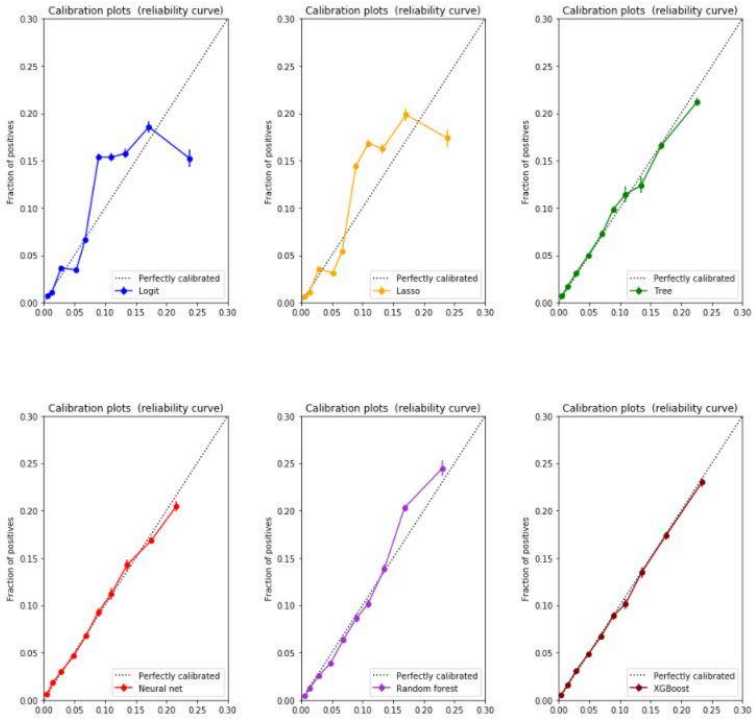


Figure 18. Calibration reliability curve with more granularity and with 95% confidence intervals



BANCO DE ESPAÑA PUBLICATIONS

WORKING PAPERS

1940 MYROSLAV PIDKUYKO: Heterogeneous spillovers of housing credit policy.
1941 LAURA ÁLVAREZ ROMÁN and MIGUEL GARCÍA-POSADA GÓMEZ: Modelling regional housing prices in Spain.
1942 STÉPHANE DÉES and ALESSANDRO GALESI: The Global Financial Cycle and US monetary policy in an interconnected world.
1943 ANDRÉS EROSA and BEATRIZ GONZÁLEZ: Taxation and the life cycle of firms.
1944 MARIO ALLOZA, JESÚS GONZALO and CARLOS SANZ: Dynamic effects of persistent shocks.
1945 PABLO DE ANDRÉS, RICARDO GIMENO and RUTH MATEOS DE CABO: The gender gap in bank credit access.
1946 IRMA ALONSO and LUIS MOLINA: The SHERLOC: an EWS-based index of vulnerability for emerging economies.
1947 GERGELY GANICS, BARBARA ROSSI and TATEVIK SEKHPOSYAN: From Fixed-event to Fixed-horizon Density Forecasts: Obtaining Measures of Multi-horizon Uncertainty from Survey Density Forecasts.
1948 GERGELY GANICS and FLORENS ODENDAHL: Bayesian VAR Forecasts, Survey Information and Structural Change in the Euro Area.
2001 JAVIER ANDRÉS, PABLO BURRIEL and WENYI SHEN: Debt sustainability and fiscal space in a heterogeneous Monetary Union: normal times vs the zero lower bound.
2002 JUAN S. MORA-SANGUINETTI and RICARDO PÉREZ-VALLS: ¿Cómo afecta la complejidad de la regulación a la demografía empresarial? Evidencia para España.
2003 ALEJANDRO BUESA, FRANCISCO JAVIER POBLACIÓN GARCÍA and JAVIER TARANCÓN: Measuring the procyclicality of impairment accounting regimes: a comparison between IFRS 9 and US GAAP.
2004 HENRIQUE S. BASSO and JUAN F. JIMENO: From secular stagnation to robocalypse? Implications of demographic and technological changes.
2005 LEONARDO GAMBACORTA, SERGIO MAYORDOMO and JOSÉ MARÍA SERENA: Dollar borrowing, firm-characteristics, and FX-hedged funding opportunities.
2006 IRMA ALONSO ÁLVAREZ, VIRGINIA DI NINO and FABRIZIO VENDITTI: Strategic interactions and price dynamics in the global oil market.
2007 JORGE E. GALÁN: The benefits are at the tail: uncovering the impact of macroprudential policy on growth-at-risk.
2008 SVEN BLANK, MATHIAS HOFFMANN and MORITZ A. ROTH: Foreign direct investment and the equity home bias puzzle.
2009 AYMAN EL DAHRAWY SÁNCHEZ-ALBORNOZ and JACOPO TIMINI: Trade agreements and Latin American trade (creation and diversion) and welfare.
2010 ALFREDO GARCÍA-HIERNAUX, MARÍA T. GONZÁLEZ-PÉREZ and DAVID E. GUERRERO: Eurozone prices: a tale of convergence and divergence.
2011 ÁNGEL IVÁN MORENO BERNAL and CARLOS GONZÁLEZ PEDRAZ: Sentiment analysis of the Spanish Financial Stability Report. (There is a Spanish version of this edition with the same number).
2012 MARIAM CAMARERO, MARÍA DOLORES GADEA-RIVAS, ANA GÓMEZ-LOSCOS and CECILIO TAMARIT: External imbalances and recoveries.
2013 JESÚS FERNÁNDEZ-VILLAVERDE, SAMUEL HURTADO and GALO NUÑO: Financial frictions and the wealth distribution.
2014 RODRIGO BARBONE GONZALEZ, DMITRY KHAMETSHIN, JOSÉ-LUIS PEYDRÓ and ANDREA POLO: Hedger of last resort: evidence from Brazilian FX interventions, local credit, and global financial cycles.
2015 DANILO LEIVA-LEON, GABRIEL PEREZ-QUIROS and EYNO ROTS: Real-time weakness of the global economy: a first assessment of the coronavirus crisis.
2016 JAVIER ANDRÉS, ÓSCAR ARCE, JESÚS FERNÁNDEZ-VILLAVERDE and SAMUEL HURTADO: Deciphering the macroeconomic effects of internal devaluations in a monetary union.
2017 FERNANDO LÓPEZ-VICENTE, JACOPO TIMINI and NICOLA CORTINOVIS: Do trade agreements with labor provisions matter for emerging and developing economies' exports?
2018 EDDIE GERBA and DANILO LEIVA-LEON: Macro-financial interactions in a changing world.
2019 JAIME MARTÍNEZ-MARTÍN and ELENA RUSTICELLI: Keeping track of global trade in real time.
2020 VICTORIA IVASHINA, LUC LAEVEN and ENRIQUE MORAL-BENITO: Loan types and the bank lending channel.
2021 SERGIO MAYORDOMO, NICOLA PAVANINI and EMANUELE TARANTINO: The impact of alternative forms of bank consolidation on credit supply and financial stability.
2022 ALEX ARMAND, PEDRO CARNEIRO, FEDERICO TAGLIATI and YIMING XIA: Can subsidized employment tackle long-term unemployment? Experimental evidence from North Macedonia.

2023 JACOPO TIMINI and FRANCESCA VIANI: A highway across the Atlantic? Trade and welfare effects of the EU-Mercosur agreement.

2024 CORINNA GHIRELLI, JAVIER J. PÉREZ and ALBERTO URTASUN: Economic policy uncertainty in Latin America: measurement using Spanish newspapers and economic spillovers.

2025 MAR DELGADO-TÉLLEZ, ESTHER GORDO, IVÁN KATARYNIUK and JAVIER J. PÉREZ: The decline in public investment: "social dominance" or too-rigid fiscal rules?

2026 ELVIRA PRADES-ILLANES and PATROCINIO TELLO-CASAS: Spanish regions in Global Value Chains: How important? How different?

2027 PABLO AGUILAR, CORINNA GHIRELLI, MATÍAS PACCE and ALBERTO URTASUN: Can news help measure economic sentiment? An application in COVID-19 times.

2028 EDUARDO GUTIÉRREZ, ENRIQUE MORAL-BENITO, DANIEL OTO-PERALÍAS and ROBERTO RAMOS: The spatial distribution of population in Spain: an anomaly in European perspective.

2029 PABLO BURRIEL, CRISTINA CHECHERITA-WESTPHAL, PASCAL JACQUINOT, MATTHIAS SCHÖN and NIKOLAI STÄHLER: Economic consequences of high public debt: evidence from three large scale DSGE models.

2030 BEATRIZ GONZÁLEZ: Macroeconomics, Firm Dynamics and IPOs.

2031 BRINDUSA ANGHEL, NÚRIA RODRÍGUEZ-PLANAS and ANNA SANZ-DE-GALDEANO: Gender Equality and the Math Gender Gap.

2032 ANDRÉS ALONSO and JOSÉ MANUEL CARBÓ: Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost.

2033 PILAR GARCÍA-PEREA, AITOR LACUESTA and PAU ROLDAN-BLANCO: Raising Markups to Survive: Small Spanish Firms during the Great Recession.

2034 MÁXIMO CAMACHO, MATÍAS PACCE and GABRIEL PÉREZ-QUIRÓS: Spillover Effects in International Business Cycles.

2035 ÁNGEL IVÁN MORENO and TERESA CAMINERO: Application of text mining to the analysis of climate-related disclosures.

2036 EFFROSYNI ADAMOPOULOU and ERNESTO VILLANUEVA: Wage determination and the bite of collective contracts in Italy and Spain: evidence from the metal working industry.

2037 MIKEL BEDAYO, GABRIEL JIMÉNEZ, JOSÉ-LUIS PEYDRÓ and RAQUEL VEGAS: Screening and Loan Origination Time: Lending Standards, Loan Defaults and Bank Failures.

2038 BRINDUSA ANGHEL, PILAR CUADRADO and FEDERICO TAGLIATI: Why cognitive test scores of Spanish adults are so low? The role of schooling and socioeconomic background

2039 CHRISTOPH ALBERT, ANDREA CAGGESE and BEATRIZ GONZÁLEZ: The Short- and Long-run Employment Impact of COVID-19 through the Effects of Real and Financial Shocks on New Firms.

2040 GABRIEL JIMÉNEZ, DAVID MARTÍNEZ-MIERA and JOSÉ-LUIS PEYDRÓ: Who Truly Bears (Bank) Taxes? Evidence from Only Shifting Statutory Incidence.

2041 FELIX HOLUB, LAURA HOSPIDO and ULRICH J. WAGNER: Urban air pollution and sick leaves: evidence from social security data.

2042 NÉLIDA DÍAZ SOBRINO, CORINNA GHIRELLI, SAMUEL HURTADO, JAVIER J. PÉREZ and ALBERTO URTASUN: The narrative about the economy as a shadow forecast: an analysis using Banco de España quarterly reports.

2043 NEZIH GUNER, JAVIER LÓPEZ-SEGOVIA and ROBERTO RAMOS: Reforming the individual income tax in Spain.

2101 DARÍO SERRANO-PUENTE: Optimal progressivity of personal income tax: a general equilibrium evaluation for Spain.

2102 SANDRA GARCÍA-URIBE, HANNES MUELLER and CARLOS SANZ: Economic uncertainty and divisive politics: evidence from the *Dos Españas*.

2103 IVÁN KATARYNIUK, VÍCTOR MORA-BAJÉN and JAVIER J. PÉREZ: EMU deepening and sovereign debt spreads: using political space to achieve policy space.

2104 DARÍO SERRANO-PUENTE: Are we moving towards an energy-efficient low-carbon economy? An input-output LMDI decomposition of CO_2 emissions for Spain and the EU28.

2105 ANDRÉS ALONSO and JOSÉ MANUEL CARBÓ: Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation.