Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation

ANDRÉS ALONSO AND JOSÉ MANUEL CARBÓ

Summary of Banco de España Working Paper no. 2105

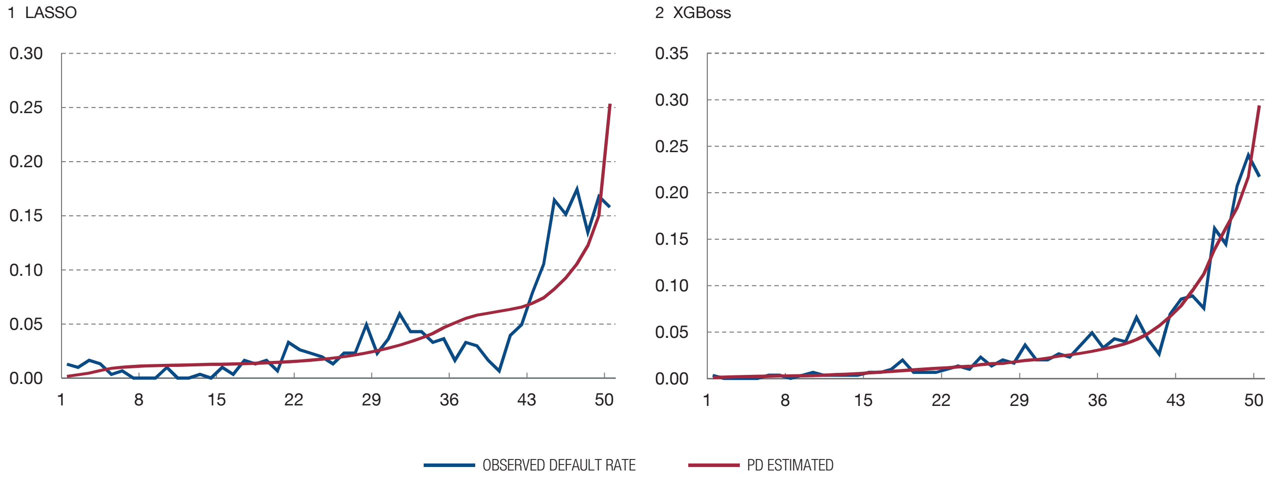
We study the economic impact for financial institutions of using machine learning (ML) models in credit default prediction. We do so by using a unique and anonymized database from a major Spanish bank. We first measure the statistical performance in terms of predictive power, both in classification and calibration, comparing models like Logit and Lasso, with more advanced ones like Trees (CART), Random Forest, XGBoost and Deep Learning. We find that ML models outperforms traditional ones, although more complex ML algorithms do not necessarily predict better. We then translate this into economic impact by estimating the savings in regulatory capital that an institution could achieve when using a ML model instead of a simpler one to compute the risk-weighted assets following the Internal Ratings Based (IRB) approach. Our benchmark results show that implementing XGBoost instead of Lasso could yield savings from 12.4% to 17% in capital requirements, depending on the type of underlying assets.

Recent surveys show that financial 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 (BoE, 2019; Fernández, 2019). While ML models usually yield better predictive performance, from a supervisory standpoint they also bring new challenges, like interpretability of the results, stability of the predictions and governance of the models (EBA, 2020; BdF, 2020). Given the novelty and complexity of some ML models, defining an adequate supervisory model evaluation approach is not an easy task. Therefore, before conducting any model risk analysis, it is essential to understand the real economic gains that financial institutions could realize by using different ML algorithms. While there exists an extensive and growing literature on the predictive gains of ML in credit default prediction, usually the findings are based on different sample sizes and different types of underlying assets, making any conclusion not robust enough. Furthermore, the economic impact of the use of ML in credit default prediction remains understudied.

To tackle this research gap we use a unique and anonymized database provided by one of the most important Spanish banks. We first measure the relative performance of the following ML models, comparing it with a logistic regression (Logit): Lasso penalized logistic regression, Classification And Regression Tree (CART), Random Forest, XGBoost and Deep Neural Networks. To this purpose we calculate the benefits in terms of statistical performance assessing the predictive performance under different circumstances such as different sample sizes and different amount of explanatory variables. This allows us to test whether the better statistical behavior of ML models comes from an information advantage (associated to the access to big amounts of data) or model advantage (associated to ML as high-end technology). We find that ML models outperform Logit both in classification and in calibration, particularly XGBoost, existing a model advantage that can be statistically isolated from an information advantage. Nevertheless, most complex models like Deep Learning (Neural Networks), do not necessarily predict better.

Second, we propose a novel approach to translate this statistical performance into actual economic impact of using ML models in credit default prediction. Taking as a basis the Basel formulas for risk-weighted assets (RWA) and the regulatory capital requirements in the Internal Ratings-Based (IRB) approach, we compute the savings in terms of minimum capital requirements which could be achieved by using more advanced algorithms, in particular XGBoost, compared to traditional techniques like Lasso. We perform a step-by-step computation of the capital requirements for both methods. Out of nearly 75,000 loans in our dataset, we use around 60,000 to train the models and make predictions of the probability of default (PD) over

Figure 1
RANKING PDS PER MODEL



the remaining 15,000 loans.[1] We organise the predictions proportionally into 50 buckets (about 300 loans in each bucket), from lower to higher values of PD. The results are displayed in Figure 1. The discrepancy between the observed default rate (blue line) and the average PD (red line) is greater for Lasso than for XGBoost, as Lasso tends to both overestimate and underestimate the fraction of default.

In order to get the approval from a supervisor, the classification into buckets must comply with two criteria: (i) risk heterogeneity between buckets, and (ii) risk homogeneity within buckets. To meet both criteria, we sequentially reduce the number of buckets. Out of the 50 starting buckets, we end up with six for Lasso and eight for XGBoost. Lasso finds fewer buckets because we are constrained by its underlying PD distribution, which presents important flat areas, undifferentiated, that do not allow further disaggregation (Figure 1 left).

Once we have our final bucket classification for Lasso and XGBoost, we calculate the capital requirements (K) for each bucket, and find that the average K can be up to 17 % lower for XGBoost than for Lasso. These capital savings come from two sources. First, the difference in the distribution of loans in buckets between models. Lasso's PD distribution is particularly flat in areas with low PD (Figure 1), accumulating a disproportionately large amount of loans at around 1.5% of PD. According to the Basel formulas, the K function of a group of loans is mainly concave and increases with the PD of the loans, particularly for low PDs. Second, the difference in the number of buckets found within each model. Since XGBoost's PD distribution (Figure 1 right) fits the observed default better than Lasso's, XGBoost ends up with more buckets in the final rank (eight instead of six). This implies, due to the concavity of the RWA Basel function over the parameter PD, a difference in capital requirements in its favour.

Our results indicate that ML models, due to their better statistical performance, could generate significant savings for financial institutions in terms of regulatory capital requirements compared to traditional statistical models. The magnitude of our results suggests that supervisors need to thoroughly investigate the risks associated with the use of these models, both from a micro and macro-prudential perspective, in order to ease the adoption of this innovation in the market.

[1] Different train-test partitions do not affect the results of this section.

REFERENCES

BdF (2020). “Governance of Artificial Intelligence in Finance,” *Fintech Innovation Hub ACPR. Banque de France.* June 2020.

BoE (2019). “Machine learning in UK financial services.” *Bank of England*

Fernández, Ana (2019). ”Inteligencia artificial en los servicios financieros.” *Boletín Económico* 2/2019. *Artículos Analíticos.* Banco de España

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