Constrained machine learning models for credit default prediction: who wins, and who loses

Abstract

Balancing prediction accuracy of machine learning models with interpretability is crucial, particularly as regulatory demands and supervisory expectations push for transparent decision-making processes. This study investigates the consequences of restricting a machine learning model highly used by financial institutions in credit default prediction to achieve better interpretability, using open source data from Lending Club. In particular, we compare the performance of an XGBoost when monotonicity restrictions are included, against an unconstrained version, assessing potential biases in the estimations, as well as changes in features importance. Our findings suggest that the restrictions push all individual estimations towards the mean. While this effect is not homogeneous, we conclude that in principle it would not harm people's access to credit. Notably, this approach enhances model interpretability while having a relatively small impact in Shapley Share Coefficients (Joseph 2019), which is reassuring since it implies that the model is learning patterns based on a similar range of variables.

Introducing monotonicity restrictions

Most of the restrictions have been derived from universally accepted relationships or based on basic principles of financial theory. To this purpose, Figure 1 shows the SHAP values associated to these selected features to show how monotonicity restrictions correct unreasonable model behavior giving non-monotonic importance to certain variables:

Figure 1. The lower boundary range the borrower's FICO at loan origination

Estimating the impact of monotonicity restrictions on probability of default

We seek to study how the introduction of monotonicity constraints impacts the probability of default (PD) of individuals. Overall, the restrictions push all individuals towards the mean, as shown in Figure 2. However, we observe that the impact is not homogeneous: real-defaults see a higher decrease in their PD than non-defaults. This would explain the underperformance of the restricted model.

Figure 2. The restrictions push all individuals towards the mean.

Impact on explainability using Shapley regressions

SHAP regressions are applied on both XGBoost models. Then, we can compare the lists of important variables ordered according to their Shapley Share Coefficients. The similarity between these ordered lists is captured through the Rank Based Overlap (RBO).

We conclude finding evidence that the introduction of monotonicity constraints has a relatively small impact. This is reassuring since it implies that the model is learning patterns based on a range of variables and does not over rely on a single variable.

Authors & Affiliations

Andres Alonso-Robisco[1], Jose Manuel Carbo[1], Guillermo de Haro[2], Juan Jose Guillen[3]

[1]Banco de España, Madrid, Spain. [2]IE University, Madrid, Spain. [3]Universidad Politecnica de Madrid, Madrid, Spain