Understanding the performance of ML models to predict credit default: A novel approach for supervisory evaluation

Authors  Andrés Alonso and José Manuel Carbó
Discussant  Klaus Düllmann*

EBA Policy Research Workshop
Virtual conference, November 2020

*Any views expressed are those of the author and do not necessarily reflect those of the ECB
## Comments

<table>
<thead>
<tr>
<th></th>
<th>Contribution &amp; Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Comments on how the “supervisory costs of ML models” are assessed</td>
</tr>
<tr>
<td>2</td>
<td>Comments on the calibration measure</td>
</tr>
<tr>
<td>3</td>
<td>Comments on how could the authors make their case more convincing</td>
</tr>
</tbody>
</table>
Contribution

• Explore the …
  – performance of several Machine Learning (ML) methods for credit default prediction and …
  – estimate the respective regulatory capital relief (12.5%-17% of RWA)
• Performance analysis is based on a uniform, rich data set of a bank
• Questions are highly relevant from a policy perspective in light of ongoing discussions on the use of ML for estimating regulatory variables.
• Particularly relevant in this context is a potential trade-off between
  – Higher predictive performance vs.
  – Less transparency on the underlying economic mechanics that render supervisory evaluation more difficult
Rubric

Data
- **Retail portfolio** of one Spanish bank
- **75,000 credit operations** with max 370(!) risk factors
  - But: **no label or description of these factors**
  - **No time dimension**
  - Therefore only PIT estimates and **no macroeconomic variables** captured
- About 4% of loans defaulted
- 80% training sample, 20% test sample
1) Comments on how the “supervisory costs of ML models” are assessed

• Authors mention the trade-off between better predictive performance and higher supervisory costs of ML models as part of their motivation.

• More precisely they see measuring the „economic impact“ (i.e. the RWA reduction) as their contribution to this subject.

• But is the RWA reduction really part of „supervisory costs“ or is it not a justified consequence of a better performance in measuring risks?

• If the „supervisory costs“ are instead rather driven by „interpretability and stability of predictions“ and „governance of the models“ then these aspects would need to be addressed.

• If you look at RWA, then why not also on EL that also affects the solvency ratio?

• What is the intuition behind RWA being lower? Could they also go up?
2) Comments on the calibration measure

• Is the „Brier score“ a clear calibration measure?
• Brier score is a sample estimator of the mean-squared difference of
  – Default indicator variables and
  – default frequency estimates
• Can be interpreted as the residual sum of squares from a non-linear
  regression of the default indicators on the rating.
• Minimising the Brier score is equivalent to maximizing the variance of the
  default frequency estimates which is also achieved by the ROC measure
• Therefore, the Brier score is (also) a measure of discriminatory
  power!
• Solution: Apply alternative measures of calibration, e.g. $\chi^2$ or Hosmer-
  Lemeshow test
3) How could the authors make their case more convincing?

- The viability of a performance analysis depends on that the models in the „horse race“ are representative for their respective class, particularly for those models that perform inferior
  - Need to convince the reader that results would not be different if just better performing models from a respective class had been used

- In this regard more information would be useful
  - on the individual algorithms
  - on the robustness of models under the different algorithms

- Employ statistical tests on equal or superior predictive performance

- Does the lack of macro-economic variables give a systematic disadvantage to regression based models?

- Discuss interpretability of the results for the more advanced algorithms
More information on the individual algorithms

- E.g. how was the logit model chosen? Is it the result of a best subset regression?
- Could model averaging approaches (like e.g. Bayesian model averaging) bring the performance of logit models closer to the more advanced approaches (this is advocated e.g. by Raftery et al. (1997))?  
- Could using Elastic Net instead of Lasso improve the performance (Zou and Hastie (2005))?  
- Why do regression trees perform worse than random forest? Is the XGBOOST a pure tree algorithm? Or a regression tree algorithm?
More information on the robustness of models and statistical tests on predictive performance

• More information on the robustness of models under the different algorithms.
  – How were the models cross-validated (h-fold, bootstrap?).
  – How sensitive are model variables (in particular variable importance) to changes in hyper-parameters and/or the sample?
  – How sensitive is the variable selection of the optimal model to changes in the sample (using e.g. bootstrap or h-fold cross-validation)?

• Statistical tests on equal or superior predictive performance such as suggested by Hansen(2005), White (2000), Diebold and Mariano (1995) could be performed to show the significance of the superior prediction of the more advanced algorithms.
Impact of lack of macro-economic variables and addressing also the interpretability of the results

• Does the lack of macro-economic variables in model setup give a systematic disadvantage to regression based models like logit and lasso? Overall macro-economic variables are found to be an important factor in default prediction models (e.g. Duffie et al. (2007)).

• The authors do not address the issue of interpretability of the results from the more advanced algorithms?
  – How many factors do enter the final models?
  – The balance between prediction performance, model stability (cross-validation) and interpretability of models (variable importance) should be more emphasized in the paper.
Summary

• Measuring the performance of new ML methods relative to „classic“ models is of high policy relevance
  – It affects the trade-off between potentially higher estimation performance coming at the cost of loss of transparency „what is really going on“ in the model
• The contribution of the paper to the question of “supervisory costs of ML models” could be explored further
• The author may consider alternative measures of calibration (eg. Hosmer-Lemeshow test)
• The reader may benefit from more information of the used models, statistical tests, and discussing a potential impact of shortcomings from the data set on the results (no time dimension, no macro variables)
• Recommend reading this paper that is well written and opens a strand of literature that is becoming with technological advances more and more important for supervisors!
References


• **Hansen 2005**: A test for superior predictive ability, Journal of Business and Economic Statistics;


• **White 2000**: A reality check for data snooping, Econometrica;

• **Zou and Hastie 2005**: Regularization and variable selection via the elastic net, Journal of Royal Statistical Society B