Project Review:

« Explainable AI in credit risk management »

Author: Laurent Dupont (Fintech-Innovation Hub, ACPR / Banque de France)

This document is a review of the research paper "Explainable AI in credit risk management" by Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, Jochen Papenbrock (AI_vSSRN_20191218) and of the accompanying code.

Disclaimer

> *The content of this review should in no way be considered as an approval by the ACPR of the method described in the research paper, whether from a technical or process-oriented perspective. Furthermore, any point of view or opinion expressed herein only reflects the author's thinking and not the ACPR's positions.*

Summary of the approach

The approach is quite innovative as it combines black-box ML explanatory method with graph analytics.

Prior research works have embedded network analysis stages in the default prediction itself. For example, they used SVD to analyze latent factors involved in credit scoring on peer-to-peer lending systems. Then a regression model was fitted over the sub-population of interconnected companies – which were thus intuitively subjected to a contagion risk -, and yielded a far better AUC performance than model fitting over the entire population.

By contrast, in this research project the network analysis stage is applied as post-treatment. It consists of a hierarchical clustering of companies in the dataset considered. More precisely, it is comprised of the following steps:

1. An XGBoost algorithm is trained on companies' balance sheet data in order to assess their probability of default over a one-year period.
2. SHAP values are computed on each predictor variable and the score produced by the XGBoost model.
3. An MST (minimum spanning tree) is computed on the vectors composed of the SHAP values for each company.
4. Finally, results are visualized via a dendrogram representation of the MST wherein companies are color-coded either by their predicted (numerical) score or by their (boolean) default status.

Insights gained from the methodology

Two main observations can be drawn from the results presented in this project.

On the one hand, companies having defaulted in the timeframe considered tend to be grouped in the same region of the MST, in other words they have similar values of the predictor variables (and thus their SHAP vectors are close to one another). This result was not a priori obvious, since one could imagine quite heterogeneous profiles for defaulted companies. Here instead, a "typical" profile of companies at risk of defaulting can be outlined and is worth investigating further (see “Characterization of defaulted clusters” and “Counterfactual explanations for defaulted companies” below).

On the other hand, the fact that non-defaulted companies which are close to defaulted ones in the MST carry an increased risk of default is stated in the paper but not demonstrated by evidence. Indeed, their proximity could be an artifact of the XGBoost predictive model rather than due to their actual probability of default, which remains to be demonstrated either by out-of-time testing of the model (see “Backtesting procedure” below) or by a contagion model as in the research previously listed.

Further analysis

Several avenues may be pursued, both as a validation of the method used in the project and as a continuation of the analysis performed.

Characterization of defaulted clusters

The first potential direction for further analysis is to derive meaningful, interpretable (i.e. compact) descriptions of the clusters illustrated in Figure 4. Indeed, clusters appear to be highly homogeneous in the sense that there are "non-defaulted clusters" (most of them) and "defaulted clusters" (the ones dominated by red color).

The defaulted clusters, i.e. those containing a majority of defaulted companies, are particularly interesting: the location of those clusters, in the example given around the bottom-left corner and close to the tree leaves, suggests that their predictor variables exhibit similar levels of contribution.

Two types of analysis could be performed:

- **Characterizing defaulted clusters by their most sensitive features**. In order to do this, one would examine which predictor variables contribute the most to their membership in this cluster and which contribute the least, in other words which variables are the most sensitive in deciding a company's outcome
- **Characterizing defaulted clusters by their shared feature values**. In order to do this, one would examine if the values of some of those predictor variables are themselves close together within the cluster, in other words whether those defaulted companies share similar characteristics.

It might also be interesting to examine the points located at the boundary between a defaulted cluster and a non-default one. Indeed they constitute a very specific kind of decision boundary, since they have very similar sets of most sensitive variables, although not necessarily similar feature values.

Lastly, the proposed hierarchical approach can be compared to other ordination techniques used to visualize high-dimensionality datasets, such as MDS (multi-dimensional scaling) or SOM (self-organizing maps).
It would indeed be interesting to experiment with those other approaches, at least to check the consistency of clusters obtained via the hierarchical clustering method described in the paper.

Backtesting procedure

On the other hand, an out-of-time testing procedure might be implemented in order to test a statement appearing in the conclusions of the paper, namely that ”good” companies that are close to ”bad” ones have a high risk of becoming ”bad” as well. This would require access to additional, post-2016 data pertaining to the ECAI dataset used in the project.

Counterfactual explanations for defaulted companies

The third and final idea for future investigation is to produce counterfactual explanations for why companies defaulted: the closest non-defaulted company in a neighboring cluster (and not in the same cluster since this might yield an outlier) would serve to produce the counterfactual explanation in the following form:

> If feature <shareholders funds plus non-current liabilities divided by fixed assets> had had value x1' instead of x1, and feature <trades payable divided by operating revenue> had had value x2' instead of x2, then the company likely would not have defaulted.

Of course there is no certainty in this kind of counterfactual explanation (hence the term "likely"), since even two companies with the exact same features might have different outcomes (because some market conditions are not captured by the available features, or due to other contingency sources). However, such an approach should provide concrete - and potentially actionable if the levers corresponding to each predictor variable are themselves actionable - insight for managing and mitigating default risk, as opposed to simply predicting future defaults like most ML models aim to do.

Conclusion

This research project constitutes a step forward in building a framework for analyzing financial risks and specifically credit risk. The combination of a predictive ML model, an explanatory method, and a clustering-based visualization technique enables building a data structure which lends itself to exploratory and investigate works, which would constitute an interesting and valuable continuation of this project – even before applying this method to other use csases.

In particular, the analysis sketched in this paper would benefit from being grounded in the data leveraged in this particular use case, for example by explicitly describing profiles of defaulted companies belonging to the same cluster. The increase in predictive power enabled by the methodology could also be assessed by performing back-testing on the “nearly-defaulted” companies, or by evaluating the counterfactual statements derived from it. Both of these further research directions – or a third one we may have not yet considered – would shed light on the value added by the method for the benefit of domain experts tasked with analyzing credit risk or studying financial stability.