

A process for Automatic Feature Generation for credit risk modeling

Regulatory constraints and a sense of conservatism has added barriers to credit institutions' appetite for modeling credit risk using more sophisticated and advanced technical methods. Studies within classification and regression models have developed rapidly during the last decades. However, credit risk modeling cannot utilize these methods, as banks must abide to a continuously growing list of regulations. Therefore, simple, and traditional regression algorithms are the industry standard. In a 2021 survey conduct by Forrester Research and Experian with 600 executives in Europe, more than 1/3 of the respondents stated that the lack of explainability of advanced methods is a barrier for augmenting the adoption of such techniques.

Combining machine learning with traditional modeling methodologies bridges between traditional credit risk models and modern data modeling approaches, without sacrificing the need for further explainability. It is, therefore, a natural step for the credit institutions to optimize the traditional models by incorporating methods such as our process for Automatic Feature Generation, presented in this paper.

In this study we have developed a new way of integrating machine learning methods into a classical scorecard development using decision trees. The study shows how integrating machine learning generated features can boost performance of scorecards and limit multicollinearity issues while still being able to easily explain the model. By using decision trees, hidden relationships between drivers can be found, and an overall better and more robust model can be developed.

Our Automatic Feature Generation has been used in several projects for local bureau models and different client projects within credit risk modeling. The results on the projects show a model performance increase of an average of 15% has been found compared to not using the method. This is measured in modeling exercises across all the credit risk life cycle by comparing Gini coefficients. This could potentially be very valuable for credit institutions. For example, in a fraud detection case, by just adopting new methodology, €2.9M were saved in preventing new fraud cases.

The presentation will go through the underlying theory and methodology behind our Automatic Feature Generation process. Additionally, the business impact of the process along with the added benefit and value will be presented. Including how it has affected explainability and general model performance.