Increasing the interpretability of AI-based credit default models. A rule-extraction proposal based on combining linguistic and unsupervised-fuzzy numeric rules.

In last years, advances in machine learning (ML) have led to the development of complex credit default models, both individually (with a significant focus on neural networks), and in the form of hybrid approaches (e.g. boosting, bagging, and stacking-based models). Although complex AI-based models are more accurate, their limited interpretability is a major concern as financial authorities require institutions to provide detailed explanations of which features are included in risk models and how these are combined to produce credit decisions. In addition, credit risk managers need to understand how the model produces the scores they use, in order to prevent under- or over-confidence on machine credit decisions, which can lead to biased decisions.

To make ML models more explainable, different approaches have been considered to extract explanatory rules from AI-based models, in order to capture the learned knowledge embedded in them while maintaining an adequate accuracy balance. However, as the model becomes more complex, the accuracy of the extracted rules can only be maintained by increasing the number of rules (cardinality) and/or their complexity (conciseness), and/or drastically decreasing their intelligibility. This is particularly critical in presence of continuous-valued inputs as financial attributes included in credit modelling. As an alternative, fuzzy rules are more flexible structures that better capture the complexity of ML internal functions while the resulting rules are closer to natural language.

In this paper, we consider two approaches for extracting fuzzy decision rules from complex AI-based default models (NN and decision forests), including a fuzzy-linguistic 2-tuple approach (prioritising comprehensibility over cardinality) and an fuzzy-numeric k-nn based approach (prioritising accuracy over intelligibility); a pedagogical analysis is considered, so the ML model is considered as a 'black box' without assuming any internal structure, and rule are extracted by mapping inputs to outputs. Results are tested on two real-life credit risk datasets, and compared in terms of accuracy, fidelity, cardinality, and comprehensibility. A fuzzy numeric rule extraction with a post linguistic refinement of labels is found to be a promising approach, producing simple rules that achieve high predictive accuracy for explaining complex AI-based credit risk models.