Interpretable Machine Learning Scorecards

Currently the use of machine learning algorithms is spreading across all the industries. The geopolitical race on the leadership in artificial intelligence is making the regulators worldwide to accept machine learning models in banking. The state governments are influencing the regulators to progress the innovation in banking. The regulators are initiating the research on the current state of the use of machine learning models in their jurisdiction to ensure that the banking industry is not late with the adoption of machine learning models.

Still internal and external validation procedures demand the machine learning models to be interpretable. In this paper several methods of interpretable machine learning are compared including Partial Dependence (PD) Plots, Individual Conditional Expectation (ICE), Local Interpretable Model Agnostic Explanations (LIME). The methods are compared on several credit risk datasets. The interpretability techniques are applied to the different machine learning algorithms including gradient boosting and neural network. The machine learning models are compared with traditional scorecards built on the same datasets.

The paper shows the benefits of machine learning algorithms compared to the traditional scorecards. It also outlines the positive and negative sides of the listed approaches on model interpretability. In overall the authors are concerned that machine learning scorecards can improve the quality of decision making in banking while the interpretability techniques can assure the validation quality.