Interpretable Machine Learning as Decision Support for Processing Fraud Alerts

Black-box machine learning models are increasingly applied in real-world contexts. But how do we know when to trust a prediction, if we do not understand why it was made? In this thesis, we researched whether explanations, in the form of local feature importance, can be useful as decision support for fraud analysts at a major Dutch bank.

To understand whether explanations can be useful in our scenario, we performed a user experiment. As opposed to common intuition, we found no significant difference in performance between tasks for which an explanation was provided and for which it was not shown. It seemed explanations alone are not enough to estimate a prediction’s trustworthiness. Instead, it might be more useful to also know how often a particular rationale resulted in a correct prediction. Following this intuition, we developed a case-based reasoning approach that provides evidence on the trustworthiness of a prediction in the form of a visualization of similar previous instances. Different from previous works, we consider similarity of the model’s explanations rather than similarity of the cases. We showed empirically that our visualization can be useful for processing alerts. Furthermore, an evaluation with fraud analysts confirmed that our approach is perceived useful.

Figure 1: A visualization of cases in which cases with similar explanations are displayed closer together. It can be seen that cases that triggered a false alert, i.e. false positive (FP), are clustered together.

Figure 2: Example of the interactive dashboard for a model trained on a public dataset. The dashboard shows the selected case (left), local feature importance explanation (middle) and the case-based reasoning visualization of cases with similar explanations (right). The color of the case shows the performance of the model: TP (true positive), FP (false positive), FN (false negative).