Human-Interpretable Explanations for Black-Box Machine Learning Models: An Application to Fraud Detection

Vladimir Balayan
Research Data Scientist, Feedzai

In this talk, human-interpretable explanations for black-box Machine Learning (ML) models are analyzed and applied to a fraud detection problem proposed by the fintech company Feedzai that uses ML to prevent financial crime. One of the main Feedzai products is a case management application used by fraud analysts to review suspicious financial transactions flagged by the ML models. Fraud analysts are domain experts without deep ML knowledge, and consequently, current explainable artificial intelligence methods do not suit their information needs. This work was focused on developing a neural network-based framework that jointly learns a decision-making task and associated domain knowledge explanations, proving high-level insights about the model's predictions.