Transparency and Interpretability in Data Science

Introduction **Beyond the Blackbox** **Explanatory Methods**

Approches to Enhance Interpretability

LIME

Local Interpretable Model-Agnostic Explanations (LIME) is a technique that helps us understand why machine learning models make certain decisions. Imagine you have a complex model that acts like a black box, where you input data and get a decision, but you don't know what's happening inside. LIME solves this by creating a simpler model that approximates and explains the decision-making of the complex model, but only around the specific data point you're interested in. It works by approximating the model locally around the prediction of interest.

SHAP

SHapley Additive exPlanations (SHAP) is a unified approach to understand model's output for making machine learning models interpretable. It draws from the game-theoretic concept of Shapley values to assign each feature of a model an importance value for a particular prediction. This approach ensures equity by attributing the predictionfairly among the features. The SHAP value is the average marginal contribution of a feature value over all possible combinations. Most importantly, by allowing interaction term, SHAP captures not just the effect of a single feature but also the interaction effects with other features, presenting a comprehensive picture of the impact on the output.

This website briefly shows my understanding in the course DS-UA 202 Responsible Data Science