Creating Palatable Local Score Explanations for Tree-Segmented Generalized Additive Models

Abstract

When developing credit scores for heterogeneous populations in terms of varying predictive relationships and available features, it can be beneficial for predictive power and for other business reasons to structure the scoring model as a Tree-Segmented Generalized Additive Model (TS-GAM). There, a typically shallow segmentation tree splits the population into more homogeneous, possibly multi-dimensional segments based on interpretable splitter features. Within each segment, the score is modeled by a segment-specific GAM with interpretable input features. Depending on tree depth, such a model can capture pairwise and moderately higher-order interactions between splitter features and GAM features.

While GAM scores can be readily locally explained as sums of interpretable feature contributions by virtue of additivity, it is more challenging to generate interpretable local score explanations for TS-GAM scores because feature contributions can no longer be assumed as additive.

In this presentation we will discuss our research into creating score explanations for TS-GAM scores. We start with an additive score decomposition into within-segment GAM contribution and segment-level contribution, exploiting the hierarchical model structure of TS-GAM. We attribute the GAM contribution directly to the GAM features. We apply the method of Shapley additive explanations (SHAP) to attribute the segment-level contribution fairly to splitter feature contributions. We combine GAM- and splitter feature contributions to generate local score explanations such that interpretable feature contributions sum up to the total score. Instead of applying model-agnostic SHAP with the complex TS-GAM directly, our approach applies Tree SHAP with a simpler shallow regression tree. We will discuss the benefits of our approach in terms of providing meaningful and relevant explanations within the context of more homogeneous segments, as well as overcoming challenges for model-agnostic SHAP when highly correlated features are used in the model.

Finally, we will also discuss the complexity of score explanations that is suitable for different audiences. Explanations that attribute the score to feature contributions as described above are highly detailed and provide valuable information for data scientists. But for models based on more than a handful of features such a level of detail could overwhelm less technical audiences. We will also discuss dimensionality reduction approaches to boil down detailed local score explanations to a smaller number of intuitive dimensions that can simplify explanations for general audiences.

Authors & Affiliations

Dr. Gerald Fahner[1]
[1]FICO, Austin, USA