Supplemental Material for the manuscript titled "Visual Diagnostics of an Explainer Model – Tools for the Assessment of LIME Explanations"

January 3, 2021

1 Extreme Feature Heatmap Scenarios

Two hypothetical examples of feature heatmaps are included in Figure 1. The plots are created with the assumption that LIME is applied to select the top feature out of $p = 4$ features for $n = 10$ cases with $t = 5$ sets of tuning parameter values. Situation 1 is an example where the features selected are consistent across tuning parameter values within a case but vary across cases within a tuning parameter value. This is the ideal situation, because the LIME explanations do not depend on the tuning parameters but do depend on the location of the observation in the feature space. Situation 2 is an example where the selected features vary across tuning parameter values within a case but are consistent across cases within a tuning parameter value. This situation indicates that the features selected by LIME are dependent on the tuning parameters, and the explanations may not be local, because the same feature is chosen regardless of the case. In practice, it is expected that the plot will exhibit a combination of these two situations.

2 Additional Bullet Matching Explanation Scatterplots

Figure 2 includes visual representations of LIME explanation from the *lime* R package and explanation scatterplots for a known match observation in the the bullet example referred to as cases M for two different data simulation methods: 4-quantile-bins (left column) and 4-equal-bins (right column). Both explanations appear to be more faithful to the random forest than those associated with case NM in Section **??**, but the intersections of the bins could be better aligned with the regions containing similar probabilities produced by the random forest. Figure 3 includes explanation scatterplots for bullet example cases M and NM when the kernel density simulation method is used.

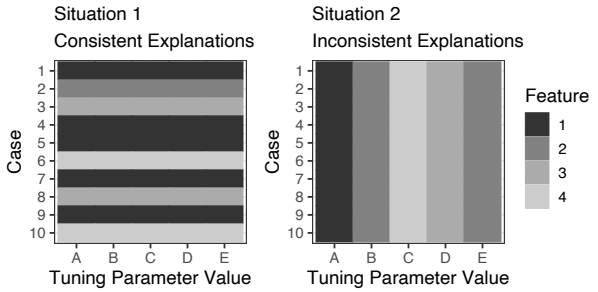

Figure 1: Hypothetical examples of feature heatmaps in two possible situations. Situation 1 is the ideal, because the explanations vary across cases but do not dependent on specific tuning parameter values. Situation 2 suggests global explanations and extreme explanation dependence on the tuning parameter values.

Figure 2: Plots of LIME explanations (first row) and explanation scatterplots (second row) for case M in the bullet test data for two tuning parameter values: 4-quantile-bins (first column) and 4-equal-bins (second column).

Figure 3: Explanation scatterplots for LIME explanations using kernel density simulation for the cases M and NM of the bullet comparison test data.