Supporting Information Corresponding to "Visual Diagnostics of an Explainer Model – Tools for the Assessment of LIME Explanations"

1 Extreme Feature Heatmap Scenarios

Two hypothetical examples of feature heatmaps are included in Figure S1. 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 (left) 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 (right) 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.

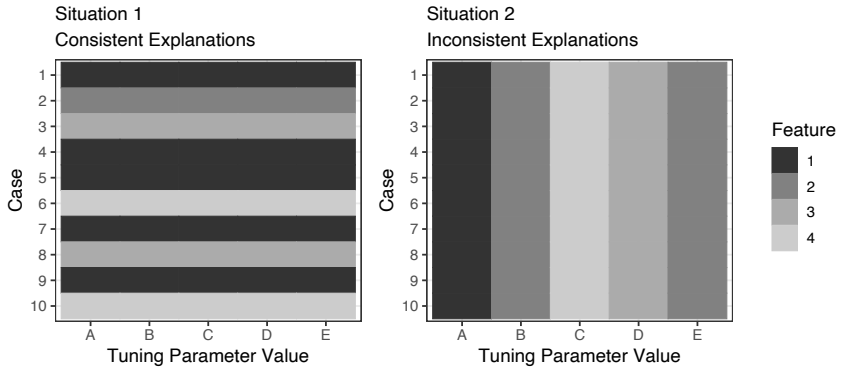

Figure S1: Hypothetical examples of feature heatmaps in two possible situations. (Left) Situation 1 is the ideal, because the explanations vary across cases but do not depend on tuning parameter values. (Right) Situation 2 suggests global explanations and extreme explanation dependence on tuning parameter values.

2 Additional Bullet Matching Explanation Scatterplots

Figures S2 and S3 include visual representations of LIME explanations from the *lime* R package (left) and explanation scatterplots (right) for a known match observation in the the bullet example referred to as case M. The explanations in Figures S2 and S3 are obtained using 4-quantile-bins and Figure 4-equal-bins, respectively. The explanations appear to be faithful to the random forest than those associated with the known non-match (case NM) in Section 4 of "Visual Diagnostics of an Explainer Model – Tools for the Assessment of LIME Explanations". However, the intersections of the bins are still not well aligned with the regions containing similar probabilities produced by the random forest. Figure S4 includes explanation scatterplots using the kernel density simulation method for both cases M and NM from the bullet example.

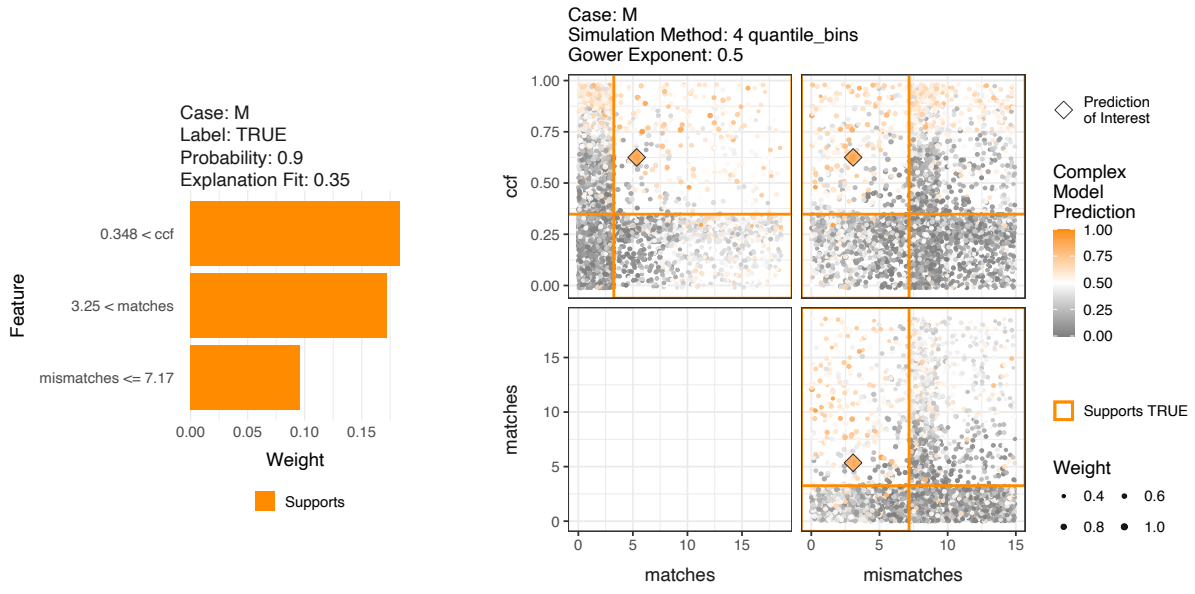

Figure S2: Explanation plot from *lime* R package (left) and explanation scatterplot (right) for case M in the bullet test data for 4-quantile-bins.

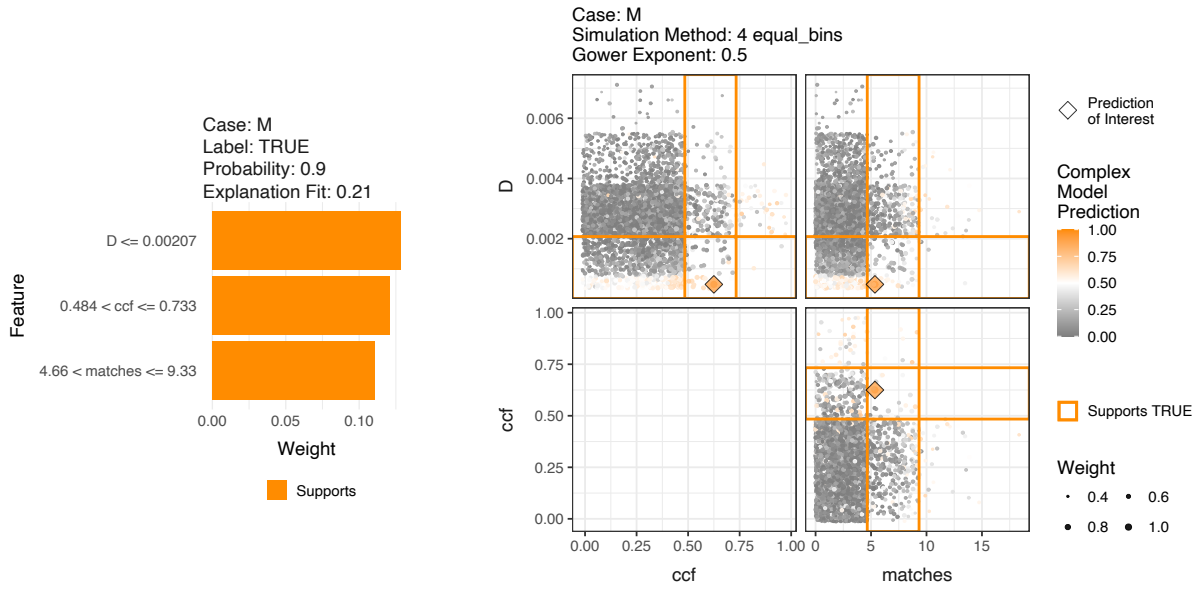

Figure S3: Explanation plot from *lime* R package (left) and explanation scatterplot (right) for case M in the bullet test data for 4-equal-bins.

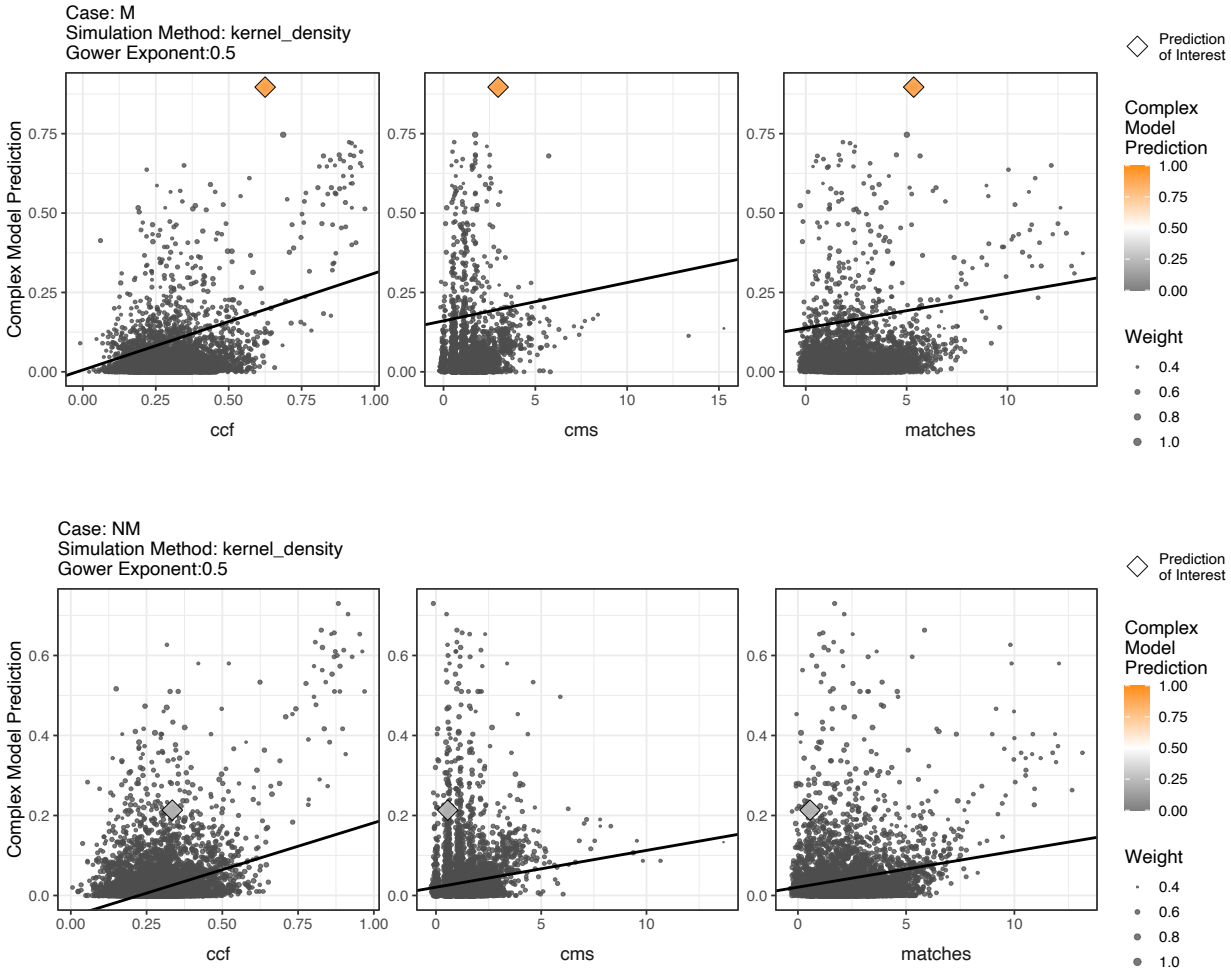

Figure S4: Explanation scatterplots for LIME explanations using kernel density simulation for the cases M (top) and NM (bottom) of the bullet comparison test data.