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

1 Explanation Scatterplots Under Density Simulation Scenarios

The manuscript introduces explanation scatterplots under the default simulation method in the *lime* R package: 4-quantile-bins. The structure of an explanation scatterplot remains the same if any bin based simulation method is used, i.e., any number of quantile or equally spaced bins. However, if the kernel density or normal approximation simulation methods are used, the format of the explanation scatterplot changes. In the density based simulation method scenarios, LIME uses the standardized versions of the predictor variables to fit the explainer model. Thus, the explainer model needs to be represented differently in the explanation scatterplot.

When the kernel density or normal approximation simulation methods are applied, the explanation scatterplot depicts the complex model by plotting the complex model predictions versus a feature selected in LIME the explanation from the simulated data. The explainer model is included as a line on the figure where all features excluding the one plotted on the x-axis are set to the observed values of the prediction of interest. An explanation scatterplot is created for each feature included in the LIME explanation. As with the bin based simulation method, the size of the points represent the weight assigned by LIME.

Figure S1 provides example explanation scatterplots for the sine data prediction of interest when the kernel density simulation method is used. The plots show the relationships between the random forest prediction and the selected features of x_1 and x_2 on the left and right, respectively.

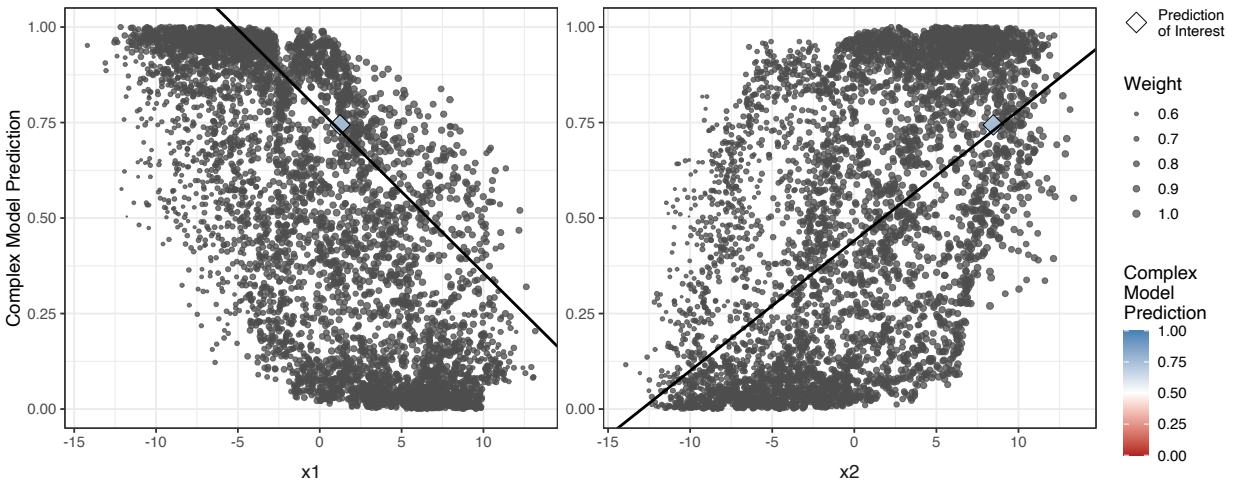

Figure S1: Explanation scatterplots for the sine data prediction of interest with the kernel density simulation method.

2 Extreme Feature Heatmap Scenarios

Two hypothetical examples of feature heatmaps are included in Figure S2. 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.

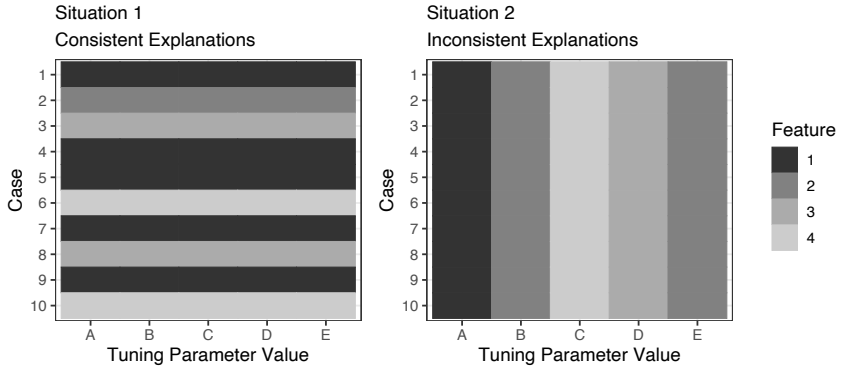

Figure S2: Hypothetical examples of feature heatmaps in two possible situations. (Left) Situation 1 is the ideal, because the explanations vary across cases but do not depend on tuning parameter values. (Right) Situation 2 suggests global explanations and extreme explanation dependence on tuning parameter values.

3 Additional Bullet Matching Explanation Scatterplots

Figures S3 and S4 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 S3 and S4 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) shown in the manuscript. However, the intersections of the bins are still not well aligned with the regions containing similar probabilities produced by the random forest. Figure S5 includes explanation scatterplots using the kernel density simulation method for both cases M and NM from the bullet example.

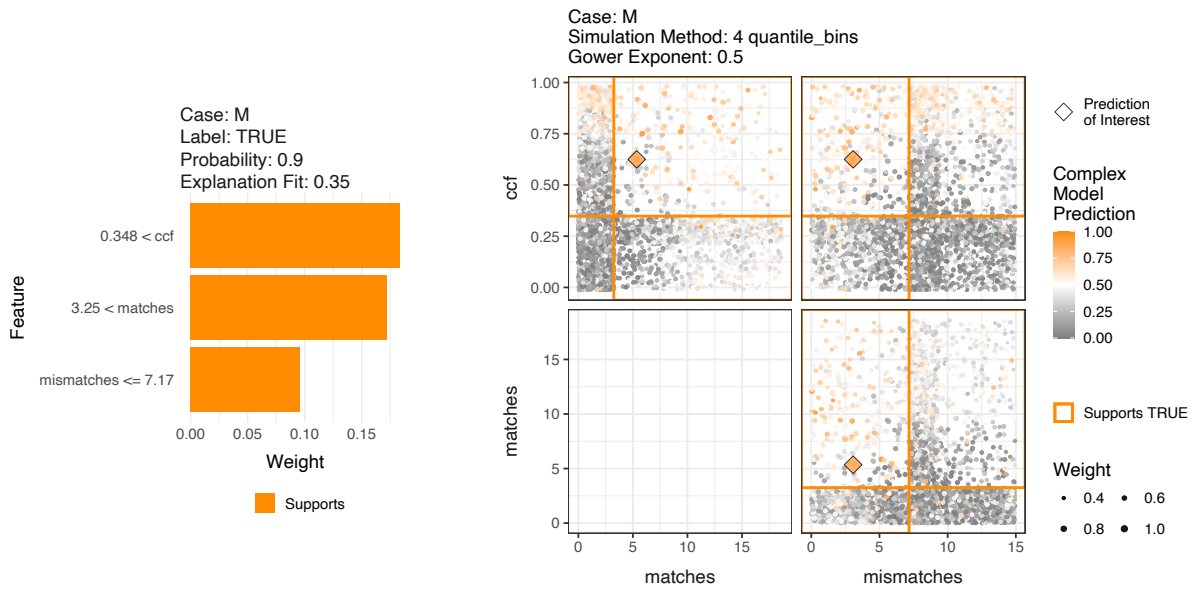

Figure S3: Explanation plot from *lime* R package (left) and explanation scatterplot (right) for case M in the bullet test data for 4-quantile-bins.

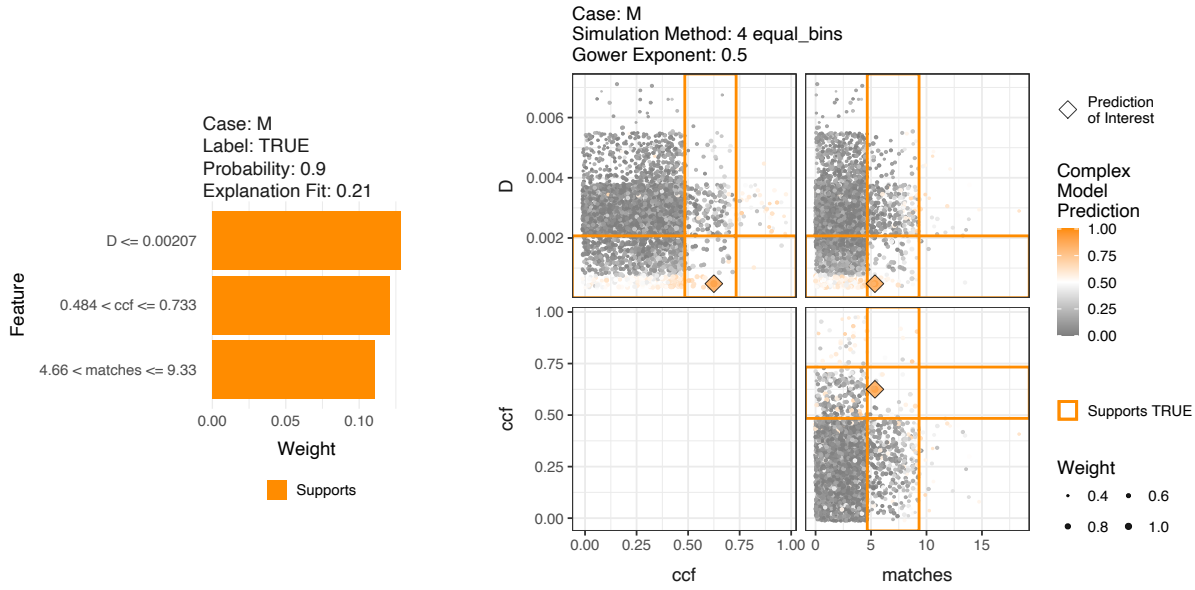

Figure S4: Explanation plot from *lime* R package (left) and explanation scatterplot (right) for case M in the bullet test data for 4-equal-bins.

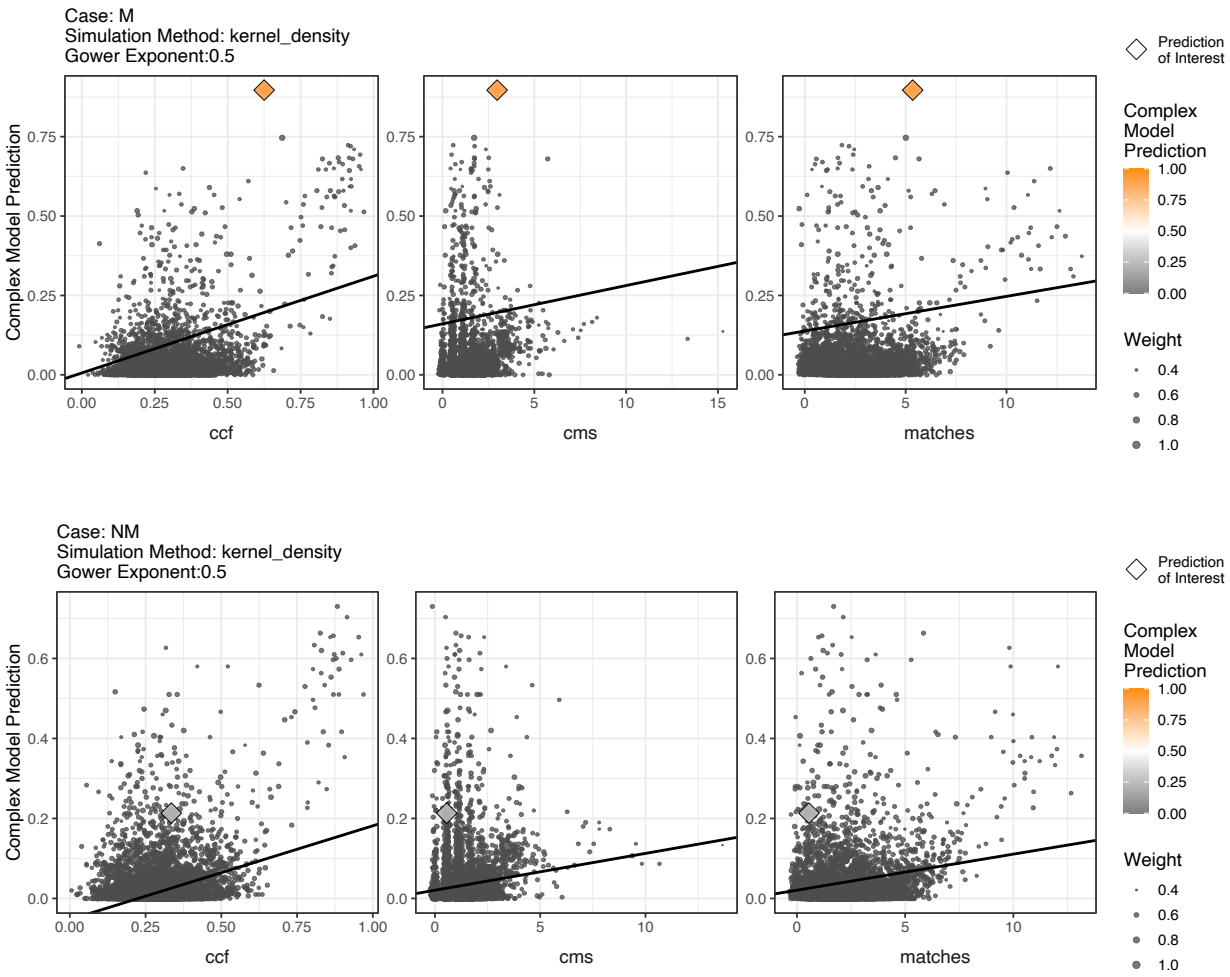

Figure S5: Explanation scatterplots for LIME explanations using kernel density simulation for the cases M (top) and NM (bottom) of the bullet comparison test data.