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

Contents

1 Explanation Scatterplots Under Density Simulation Scenarios 1

2 Extreme Feature Heatmap Scenarios 2

3 Additional Information on the Bullet Matching Data 3
3.1 Training Data 3
3.2 Testing Data 4
3.3 Variable Definitions 4

4 Additional Bullet Matching Explanation Scatterplots 5

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 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.

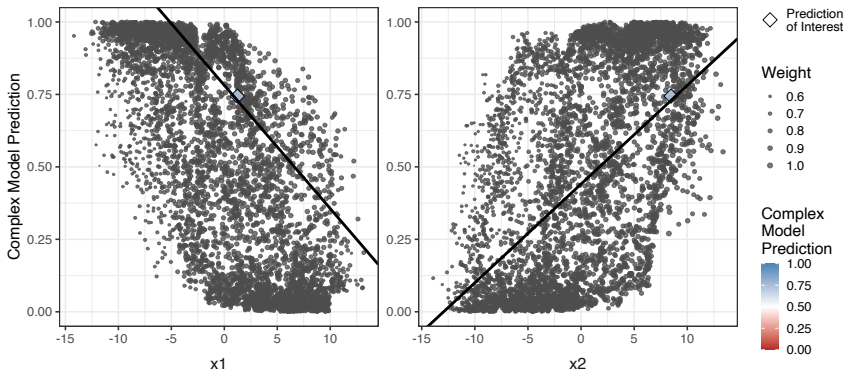

Figure S1: Explanation scatterplots for the sine data prediction of interest with the kernel density simulation method.

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.

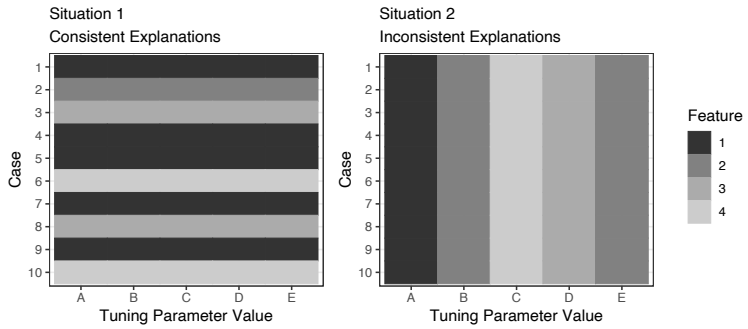

Figure S2: Hypothetical examples of feature heatmaps in two possible situations. (Left) Situation 1 is the ideal, because the explanations vary across cases but do not depend on tuning parameter values. (Right) Situation 2 suggests global explanations and extreme explanation dependence on tuning parameter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

3 Additional Information on the Bullet Matching Data

3.1 Training Data

The bullet matching training data has 83,028 rows and 13 columns that contain comparison features described in Hare et al. [2017] based on high resolution microscopy scans of fired bullets from Hamby sets 173 and 252 [Hamby et al., 2009]. This dataset is created from the x3p scans of bullet land engraved areas available from the NIST Ballistics Toolmark Research Database[1]. It contains comparisons from 408 bullet-land signatures. 12 of the overall 420 lands (6 lands per bullets, 35 bullets in each set) are excluded from the comparison. Six of these lands show so-called "tank rash" - damage to the bullets after it exited the barrel.[2] Another bullet (Bullet E from Hamby 173) is excluded because it could not be matched visually to the barrel it was supposedly from.[3]

The training data used in the manuscript it generated from a raw file of comparison features.[4] The steps taken to create the bullet training data from the raw data involve renaming some variables, selecting the variables of interest for the manuscript, and adjusting the land IDs associated with the signatures.[5]

[1] https://tsapps.nist.gov/NRBTD/

[2] https://github.com/goodekat/LIME-diagnostics-paper/blob/master/LEAscans/tankrash.md

[3] https://github.com/goodekat/LIME-diagnostics-paper/blob/master/LEAscans/bullete.md

[4] https://github.com/goodekat/LIME-diagnostics-paper/blob/master/data/raw/CCFs_withlands.zip

[5] https://github.com/goodekat/LIME-diagnostics-paper/blob/master/code/02-data-preparation.Rmd

3.2 Testing Data

The bullet test data has 364 rows and 13 columns that contains comparison features from test sets 1 and 11 of the Hamby 224 Clone Test Sets. Each test set is arranged as a combination of three bullets: two known bullets and a questioned bullet. Similar to the training set, each bullet has 6 lands. The data contains comparisons of bullet-lands within a set. With three bullets with six lands per set, there are a total of (2 sets) $\times$ (3! bullet comparisons) $\times$ (6^2 land comparisons) = 432 comparisons. However, there are only 364 comparisons in the bullet-test data. This is due to the fact that some of the lands are missing from the data (due to tank rash): land 4 from the unknown bullet in set 1, land 2 from bullet 1 in set 11, and land 4 from the unknown bullet in set 11. The test data set using in the manuscript is generated from the raw versions of the data for set 1 and set 11.[6]

3.3 Variable Definitions

The variables included in the bullet training and testing data sets are defined below. Further descriptions of the comparison features are found in Hare et al. [2017].

Index variables:

- **case**: ID number associated with the bullet-land signature comparison.
- **land_id1**, **land_id2**: IDs describing the two land engraved areas in the comparison. The format is study-barrel-bullet-land.

Predictor (comparison) variables:

- **ccf**: Maximized cross-correlation between two LEA signatures.
- **rough_cor**: Correlation after detrending aligned signatures.
- **D**: Euclidean distance (in millimeters) between two aligned signatures.
- **sd_D**: Standard deviation of the previous measure along the signature.
- **matches**, **mismatches**: Number of matching/non-matching peaks and valleys in the aligned signatures.
- **cms**: Consecutively matching striae is a measure introduced by Biasotti [1959] describing the longest run of matching peaks between two aligned signatures.
- **non_cms**: The number of consecutively non-matching peaks.

[6] `https://github.com/goodekat/LIME-diagnostics-paper/blob/master/data/raw/h224-set1-features.rds.zip` and `https://github.com/goodekat/LIME-diagnostics-paper/blob/master/data/raw/h224-set11-features.rds.zip`

- **sum_peaks**: The depth of peaks measured as the sum of matching peaks between two aligned signatures (in microns).

Response variable:

- **samesource**: Ground truth whether a pair is from the same source (TRUE) or from different sources (FALSE).

4 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 4-equal-bins, respectively. The explanations appear to be more 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.

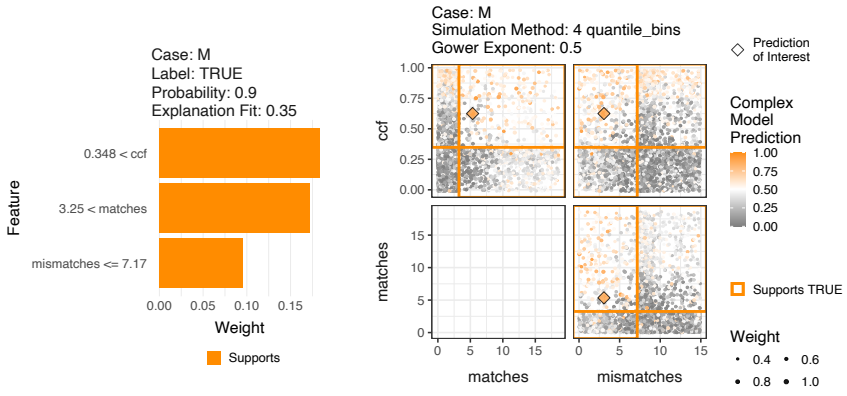

Figure S3: Explanation plot from *lime* R package (left) and explanation scatterplot (right) for case M in the bullet test data for 4-quantile-bins.

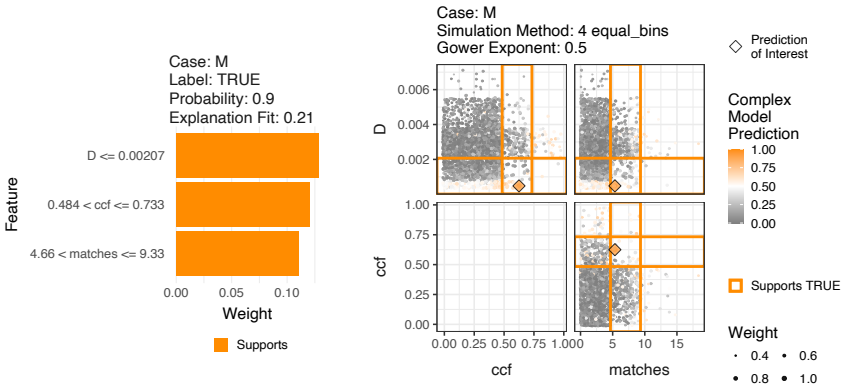

Figure S4: Explanation plot from *lime* R package (left) and explanation scatterplot (right) for case M in the bullet test data for 4-equal-bins.

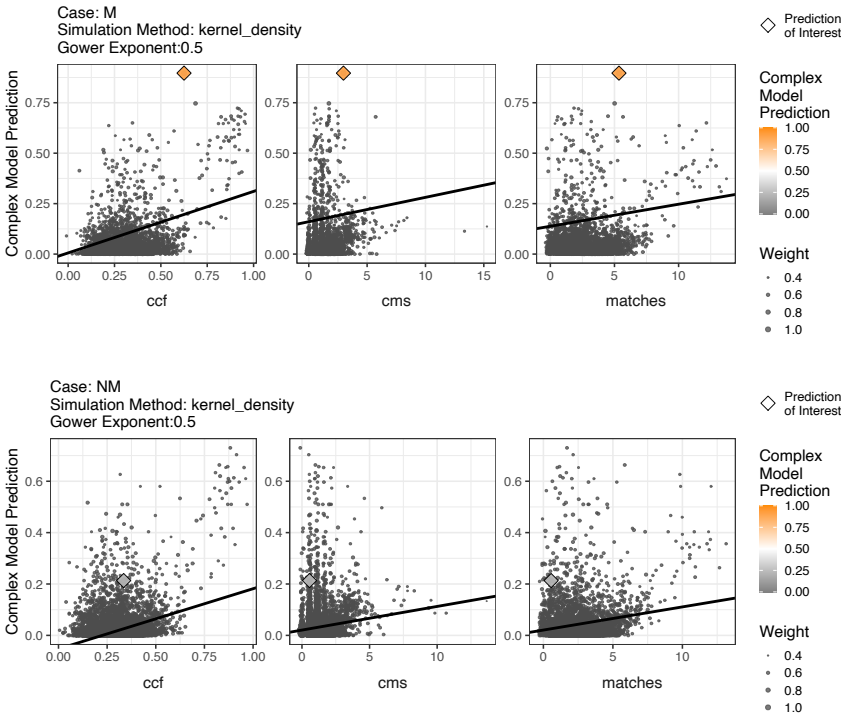

Figure S5: Explanation scatterplots for LIME explanations using kernel density simulation for cases M (top) and NM (bottom) of the bullet comparison test data.

References

Eric Hare, Heike Hofmann, and Alicia Carriquiry. Automatic matching of bullet land impressions. *Annals of Applied Statistics*, 11(4):2332–2356, 12 2017. doi: 10.1214/17-AOAS1080.

James E. Hamby, David J. Brundage, and James W. Thorpe. The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries. *AFTE Journal*, 41(2):99–110, 2009.

Alfred A. Biasotti. A statistical study of the individual characteristics of fired bullets. *Journal of Forensic Sciences*, 4(1):34–50, 1959.