Using LIME to Interpret a Random Forest Model with an Application to Bullet Matching Data

Katherine Goode and Dr. Heike Hofmann

Iowa State University Department of Statistics

ISU Graduate and Professional Student Research Conference

April 10, 2019

Introduction and Objectives

- Random forests are accurate predictive models, but they are difficult to interpret.
- LIME is a method that was designed to provide local interpretations for any predictive model (Ribiero et. al. 2016).
- We want to assess a random forest model fit to a bullet matching dataset to understand cases where the model made incorrect predictions.
- We applied LIME to the random forest, but we found some unsatisfactory results which led us to develop some diagnostic tools for LIME.

Bullet Signature Comparison Data

- To determine if two bullets were fired from the same gun, Hare, Hofmann, and Carriquiry (2017) took high definition scans of bullets from the Hamby study (Hamby et. al. 2009).
- They extracted signatures from the scans of the striations found on the six lands (raised panels) of a bullet.

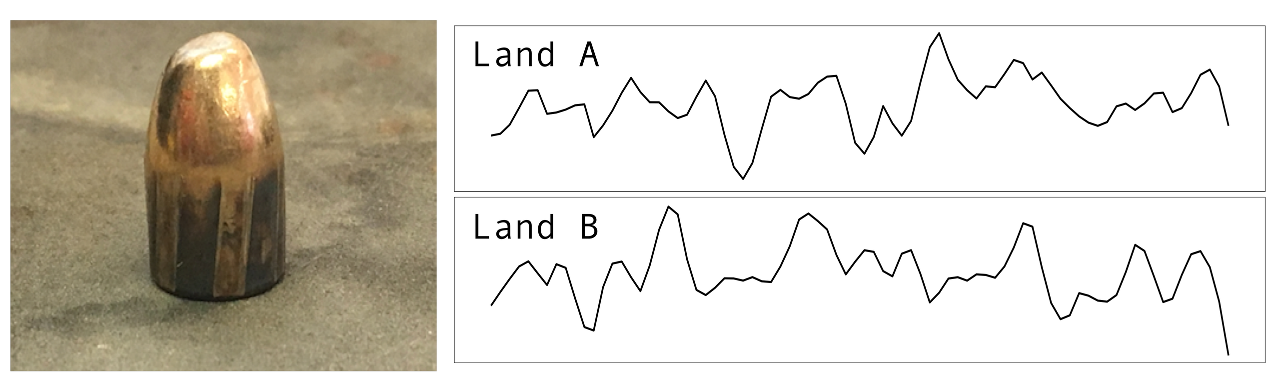

Figure 1: The picture on the left of a bullet shows the alternating land and grove impressions created when the bullet is fired from a gun. The raised portions are referred to as lands. The image on the right shows a representation of the comparison of two signatures obtained from the scans of bullet lands.

- They developed nine numeric features that quantify the similarity between two signatures.
  - e.g. Consecutively Matching Striae (cms): number of consecutive peaks two signatures have in common
- Finally, they fit a random forest to the features to predict whether two bullets were fired from the same gun.

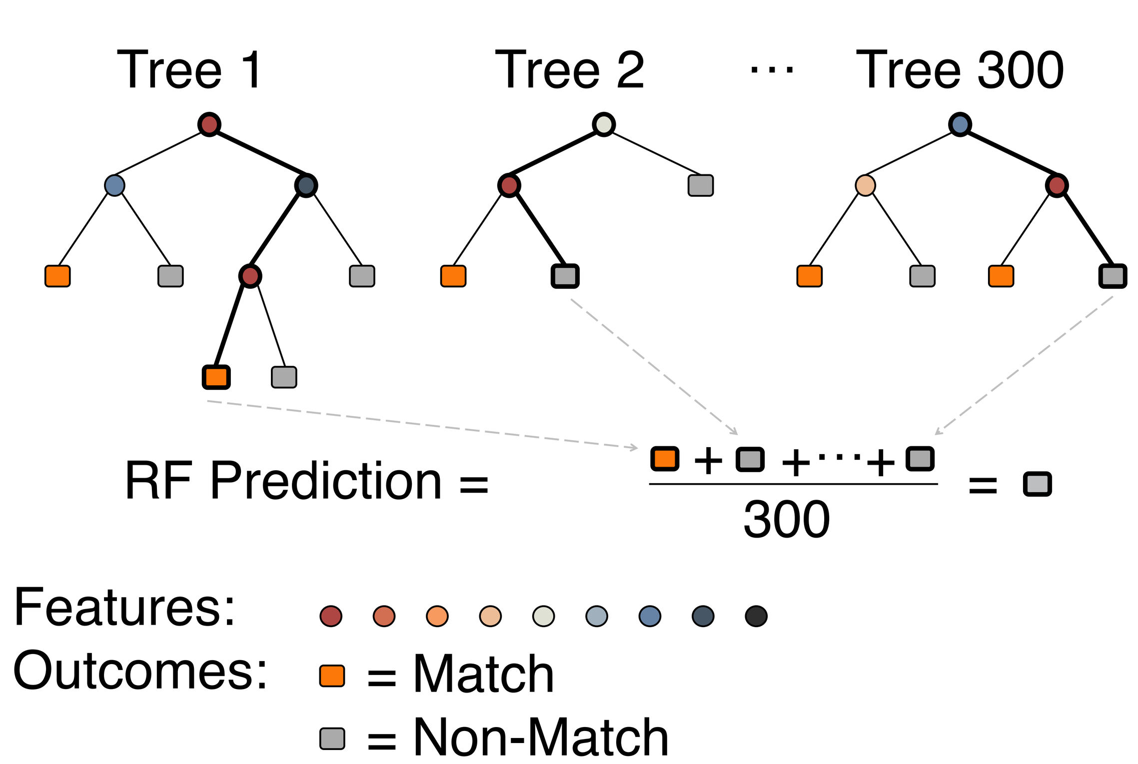

Figure 2: This diagram is a representation of how a prediction is made to determine if two lands are a match using the random forest fit to the bullet matching data. A random forest prediction is obtained by aggregating the results from many classification trees. The model fit by Hare, Hofmann, and Carriquiry used 300 trees. The circles in the trees represent the features chosen by the tree, and the rectangles represent the classification at the end of a path. The bold lines represent the paths corresponding to the observation of interest. In the case depicted, the average of the 300 tree predictions results in a random forest prediction of a non-match.

Application of LIME to Bullet Matching Data

- LIME fits a simple interpretable model in a local region that mimics the complex model.
- Procedure for one prediction of interest:
  1. Simulate data based on observed data (multiple ways to do this)
  2. Apply random forest to simulated data to obtain predictions
  3. Fit a simple interpretable model (such as a linear regression model) that assigns the most weight to observations closest to the prediction of interest
  4. Determine key variables and interpret the simple model to explain the complex model

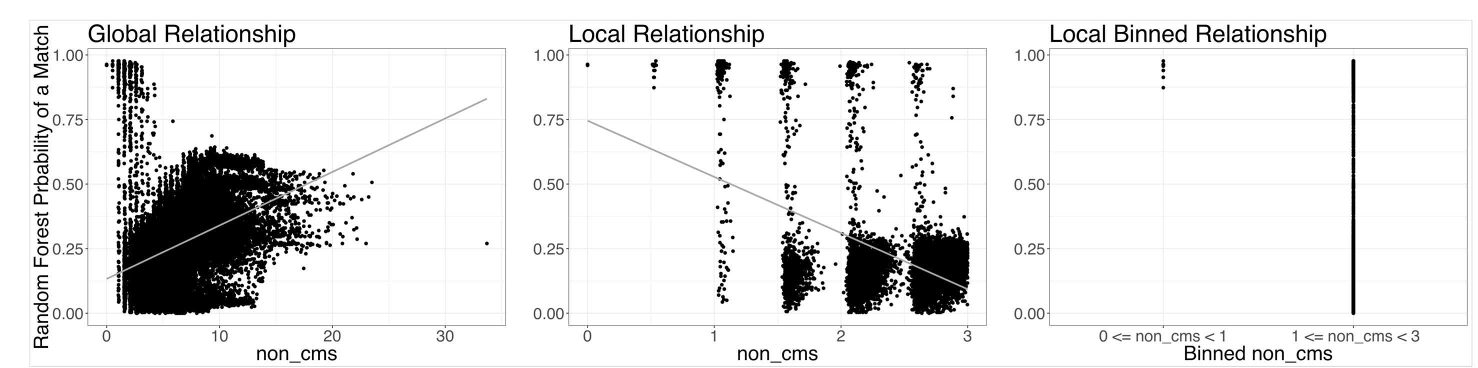

Figure 3: These plots show different views of the random forest prediction probabilities of a match versus the feature non_cms. The plot on the left is a global view of all observations in the bullet data, which shows a complex relationship between the variables, so the linear model is not a good fit. The middle plot shows a local view of the relationship with non_cms restricted between 0 and 3, and the right plot shows the same region but with non_cms divided into two bins. Both of these plots show a simpler relationship, which is the idea that LIME makes use of.

- We used the random forest model from Hare, Hofmann, and Carriquiry (2017) to make predictions on a new set of bullet scans and applied LIME to these predictions using the lime R package (Pederson and Benesty 2018) using the default simulation method.
- It produced some results that contradicted the random forest predictions. We tried the other three simulation methods available in the R package, but they also produced strange results.

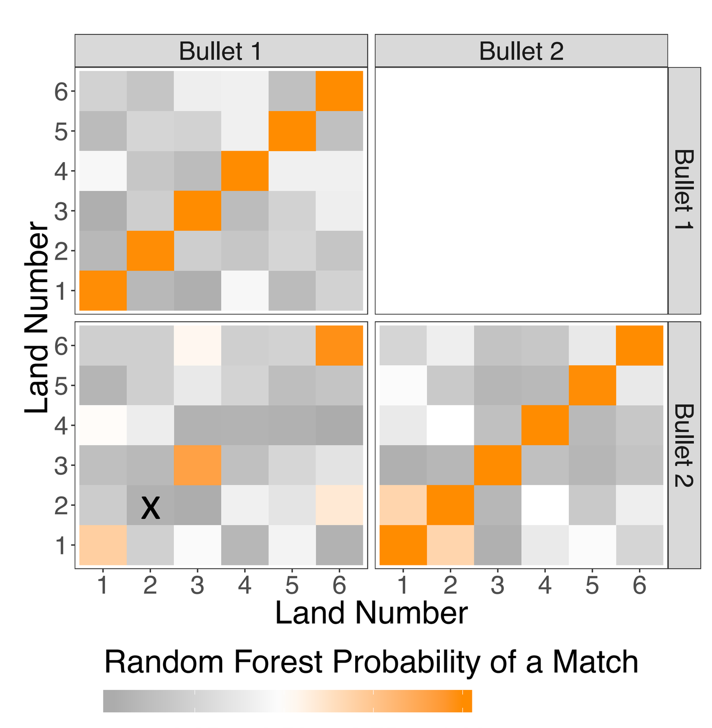

Figure 4: This depicts all pairwise comparisons of six lands from two bullets in the new set fired from the same gun. The color of the tile represents the random forest probability that the lands are a match, and the "x" indicates a prediction where the model is wrong. (The upper right is left blank since the comparisons are the same as the bottom left.)

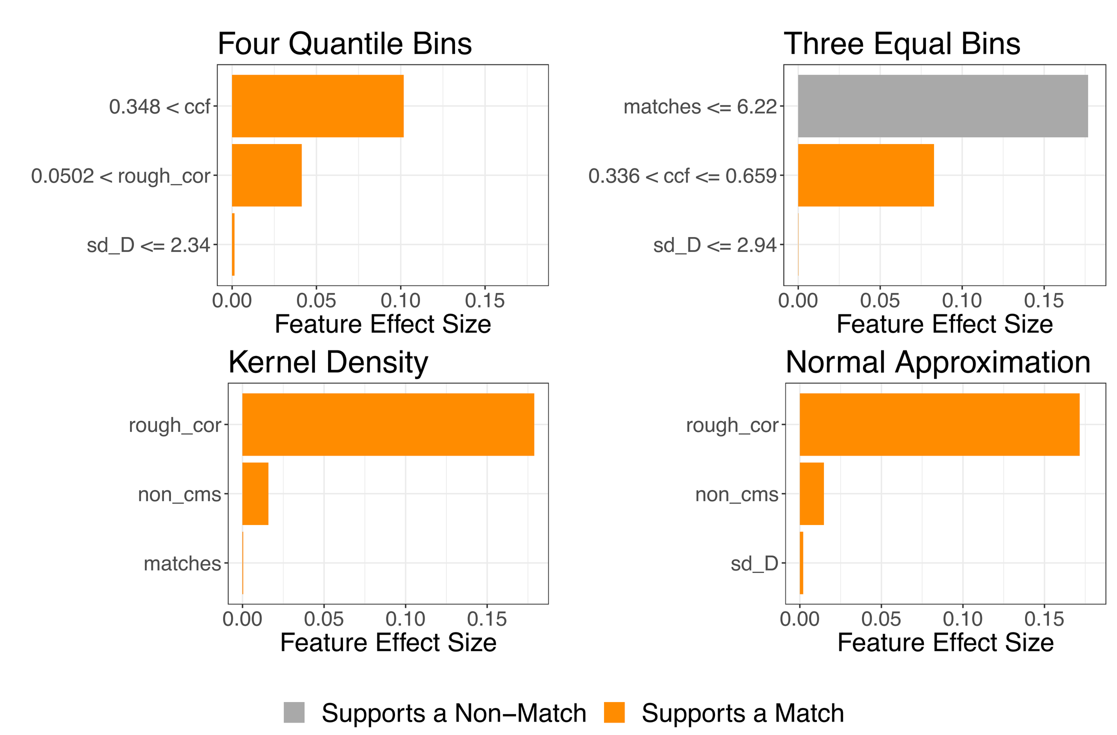

Figure 5: These plots show LIME "explanations" for the random forest prediction marked by the "x" in Figure 4 for four simulation methods. The selected features are on the y-axes, the magnitude of the "importance" of the features are on the x-axes, and the color represents whether the feature supports a match or non-match. The bars supporting a match are a surprise since the random forest predicted a non-match.

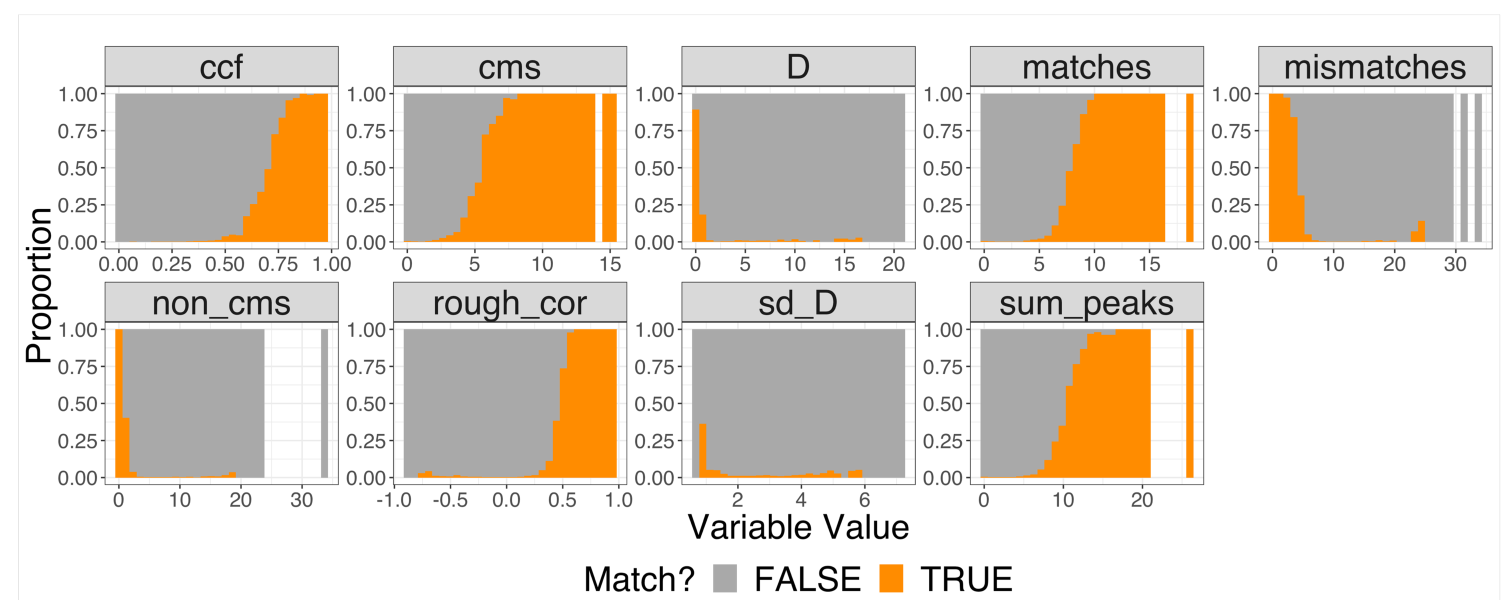

Figure 6: These plots were created from the data used to fit the random forest model for each of the nine model features. Each bar shows the proportions of matches and non-matches in a bin. Key relationships between the features and whether or not a comparison is a match can be seen in these plots.

Diagnostic Tools for LIME

- Due to the contradictory results from simulation methods included in the LIME package, we created two new simulation methods. One uses a regression tree fit to the random forest probability, and the other uses a classification tree applied to the indicator variable of whether or not a comparison is a match.
- To compare the simulation methods, we developed several diagnostic plots.
- After we applied LIME, we computed a mean squared error (MSE) as

$$\frac{\sum_i (\text{random forest prediction}_i - \text{LIME simple model prediction}_i)^2}{\text{number of comparisons}}$$

  and the average of the R^2 values from the simple models fit by LIME, and we compared these across simulation methods .
- For each of the comparisons and bin based simulation methods, we also recorded and compared the most important feature selected by LIME.

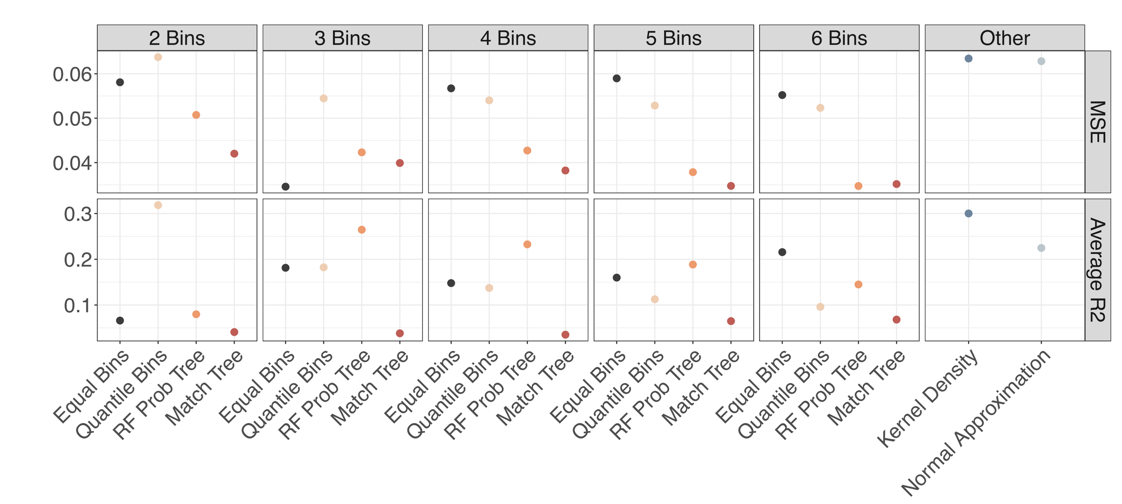

Figure 7: These plots show the MSEs and average R^2 values for the simulation methods used when LIME was applied. The tree based methods typically perform best based on the lowest MSEs, but all methods have low R^2 values suggesting the simple models do not do a good job of capturing the behavior of the complex model.

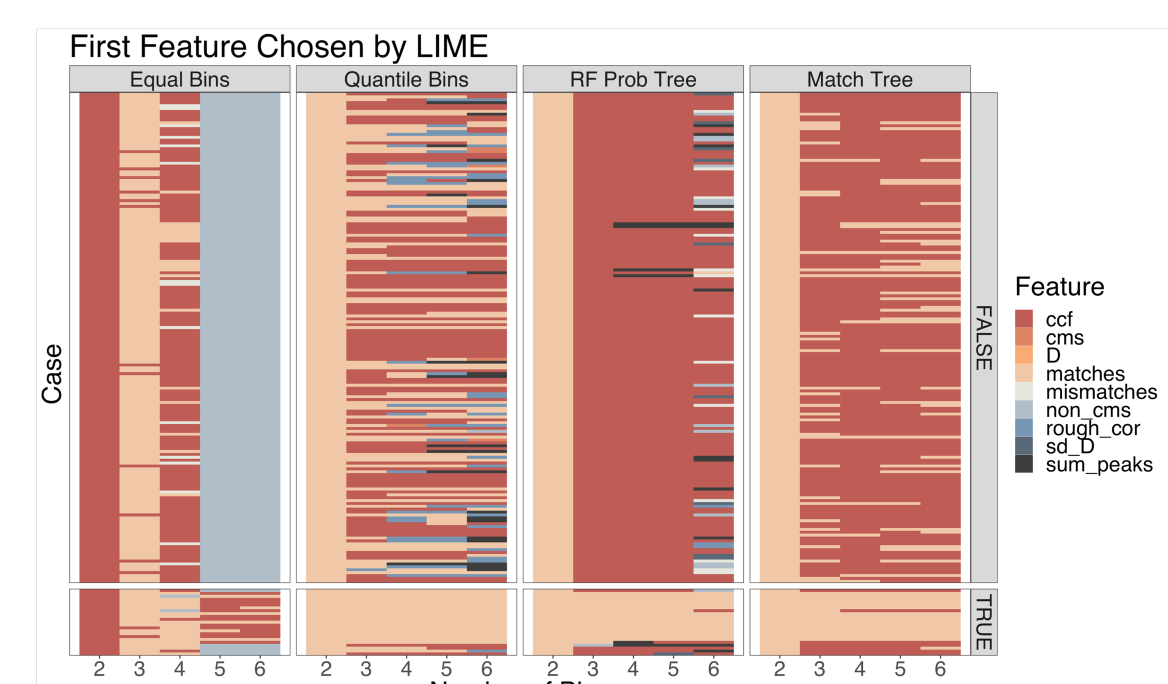

Figure 8: The heatmap shows the most important features (represented by the colors) chosen by LIME for each case in the bullet comparisons and for each of the bin based simulation methods. The cases are separated by the matches and non-matches. The vertical stripes, which can be clearly seen with the equal bins, suggest a dependence between the feature chosen and the number of bins used for the simulation method.

Conclusions and Future Work

- LIME produced explanations that contradicted the random forest model.
- R^2 values were very low for all simulation methods, and there is no obvious best simulation method based on the MSE and R^2 values.
- The most important feature chosen by LIME appears to be dependent on the simulation method.
- We think that the linear regression model used by LIME is too simplistic to capture the trends in the random forest, and we think using a tree for the simple model may produce better results.
- We would like to apply LIME to other random forest models to see if similar trends occur.

References

[1] Hamby, J. E., Brundage, D. J., and Thorpe, J. W. (2009), "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, 99–110.
[2] Hare, E., Hofmann, H., and Carriquiry, A. (2017), "Automatic matching of bullet land impressions," The Annals of Applied Statistics, 11, 2332–2356. https://doi.org/10.1214/17-aoas1080 .
[3] Pedersen, Thomas Lin and Benesty, Michaël (2018). lime: Local Interpretable Model-Agnostic Explanations. R package version 0.4.1. https://github.com/thomasp85/lime
[4] Ribeiro, M., Singh, S., and Guestrin, C. (2016), "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," 1135–1144. https://doi.org/10.1145/2939672.2939778 .