Visual Diagnostics of a Model Explainer: Tools for the Assessment of LIME Explanations from Random Forests

Katherine Goode and Dr. Heike Hofmann

Iowa State University Department of Statistics

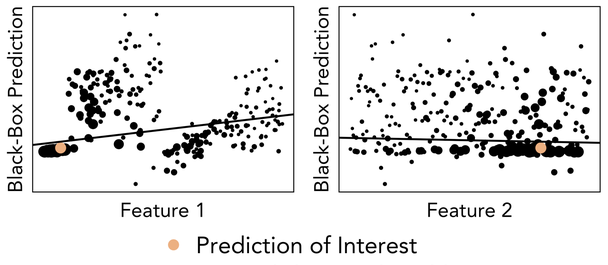
Overview

- Difficult to interpret "black-box" models
- LIME provides "explanations" for black-box model predictions
- Want to assess LIME explanations
- Developed diagnostic visualization tools
- Applied tools to a random forest model fit to a bullet matching dataset

Background on LIME (Ribeiro et al. 2017)

- **LIME:** Local Interpretable Model-Agnostic Explanations
- **Concept:** Approximate relationship between black-box predictions and features near a prediction of interest using an "explainer" model (a "simple" and interpretable model)
- Interpret explainer to select key features

Conceptual Depiction of LIME



Linear regression models weighted by proximity to prediction of interest

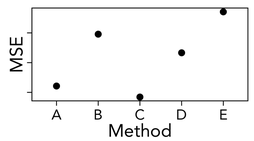
Diagnostic Tools for LIME

- **LIME Assessment Goals:**
  - Simple model approximates complex model well?
  - Local explanation?
  - Comparison of implementation methods (model, distance metric...)
- **Process to obtain values for plots:**
  1. Apply LIME to K predictions
  2. Compute

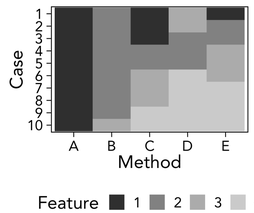
$$MSE = \frac{\sum_{i=1}^{K}(\hat{y}_i^{complex} - \hat{y}_i^{simple})^2}{K-1}$$

  3. Determine top feature chosen by LIME for each of the K predictions
  4. Repeat for M implementation methods

Templates of Diagnostic Plots



MSE Comparison: Lowest MSE suggests best approximation of simple model to complex model

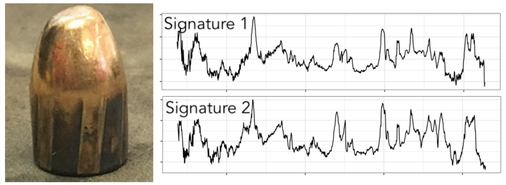


Top Features Comparison: Horizontal stripes suggest consistency across methods and vertical stripes suggest non-local explanations

Bullet Matching Example

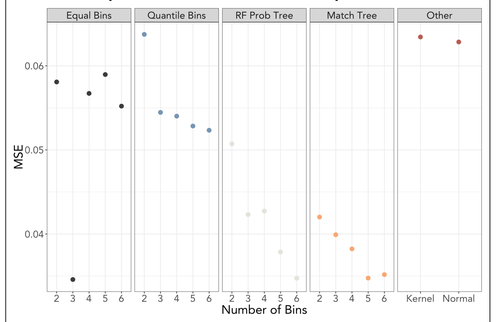
- **Data:** Markings on bullets used to predict whether two bullets were fired from the same gun using a random forest model (Hare, Hofmann, and Carriquiry 2017)
- **Methods:** Applied LIME in R to all observations in testing dataset using several LIME implementation methods

Bullet Comparisons



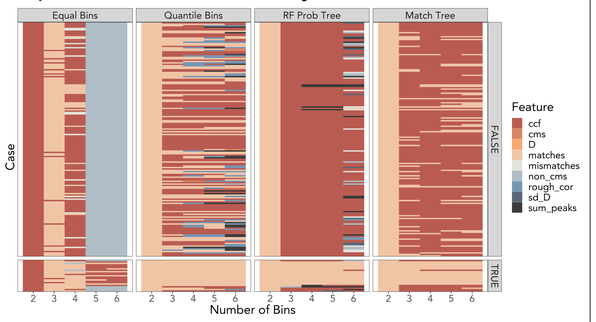
Grooves created when bullet is fired extracted through high definition scans and used to create 'signatures' to identify bullets fired from the same gun

MSE Implementation Comparisons



No one obvious best implementation method and quantile bins do not decrease as the number of bins increases as expected

Top Features Selected by LIME



Clear vertical stripes for many of the methods suggesting LIME has produced global explanations that vary based on the method used

Discussion

- Important to diagnose LIME explanations to see if dependent on implementation methods and if local assumption is met
- How to choose an implementation method?

References

- 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 .
- Pedersen, Thomas Lin and Benesty, Michaël (2018). lime: Local Interpretable Model-Agnostic Explanations. R package version 0.4.1. https://github.com/thomasp85/lime
- 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 .