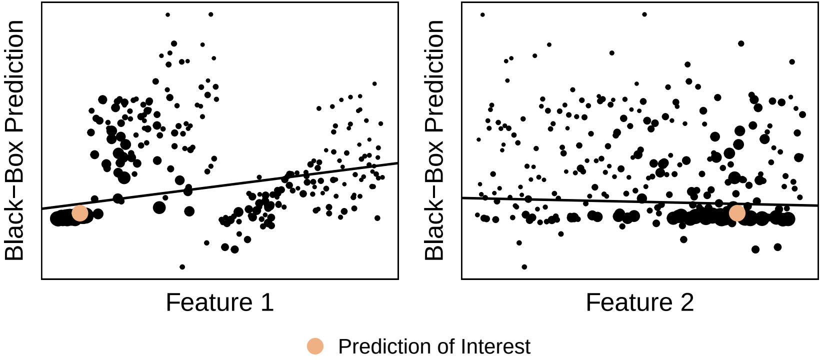
Visual Diagnostics of a Model Explainer: Tools for the Assessment of LIME Explanations from Random Forests

Katherine Goode and Dr. Heike Hofmann
Iowa State University Department of Statistics

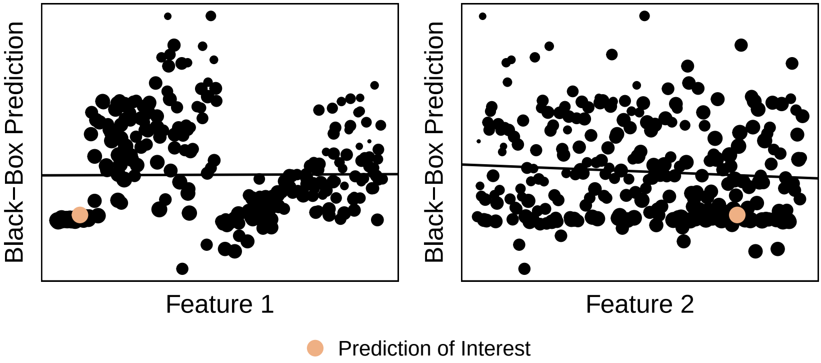
Background on LIME (Ribeiro et al. 2017)

- Local Interpretable Model-Agnostic Explanations
- Provides “explanations” for black-box model predictions to determine if trustworthy



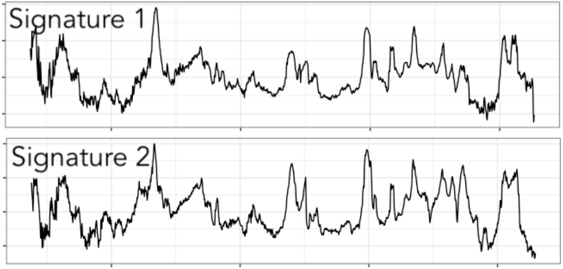
Assessment Goals

- How do we know if LIME explanations are trustworthy?
    - Simple model a good approximation?
    - Explanation local?
    - Explanations consistent across implementation methods (form of explainer model, distance metric, etc.)?



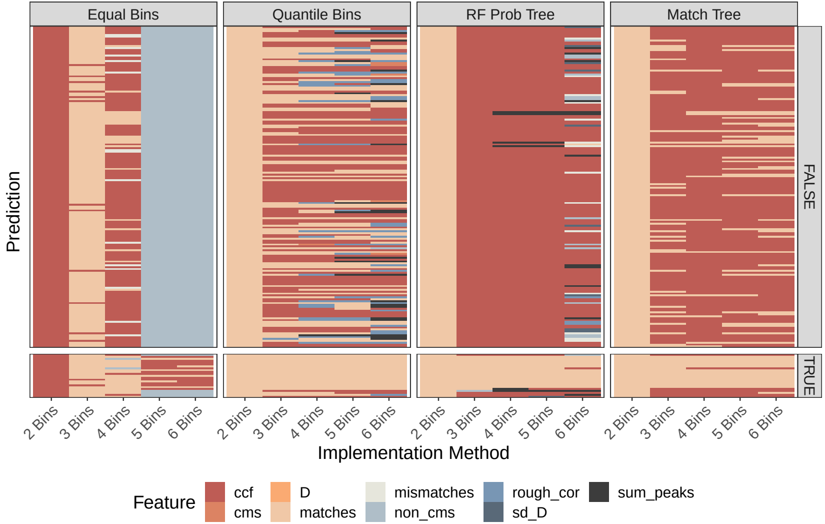
Forensics Bullet Matching Example

Random forest model used to predict if two bullets were fired from the same gun based on markings on bullets (Hare, Hofmann, and Carriquiry 2017)



LIME Diagnostic Plot

Top features selected by LIME applied to all predictions in bullet testing dataset using several implementation methods


Equal Bins
Quantile Bins
RF Prob Tree
Match Tree
FALSE
TRUE
Prediction
2 Bins
3 Bins
4 Bins
5 Bins
6 Bins
Implementation Method
Feature
ccf
cms
D
matches
mismatches
non_cms
rough_cor
sd_D
sum_peaks