

# Visualization Methods for Explainable Machine Learning

## Statistics PhD Oral Prelim

Katherine Goode  
Iowa State University  
May 6, 2020

# Personal Background

## Education

B.A. in Mathematics

- Lawrence University (Appleton, WI)
- Graduated in June 2013

M.S. in Statistics

- University of Wisconsin, Madison
- Graduated in May 2015

Ph.D. in Statistics (in progress)

- Iowa State University
- Started in January 2016

## Teaching

- Teaching assistant at UW Madison
- Lecturer at Lawrence University
- Lecturer at ISU

## Consulting

- AES Statistical Consultant
- NREM Research Assistant

## Internship

- Sandia National Labs: Statistical Sciences Research and Development Intern

# Overview of Talk

1. Background and Overview of Thesis

2. Detailed explanation of Chapter 1

- Visual Diagnostics of a Model Explainer -- Tools for the Assessment of LIME Explanations

3. Plan for Chapter 2

- Explaining Random Forests using Clustering of Trees

4. Ideas for Chapter 3

- Extensions of Neural Network Explanation Tools to Tabular Data Applications

5. Timeline for Completion

6. Discussion Points

# Background and Overview of Thesis

# Explainable Machine Learning

- Machine learning models:
  - Good in prediction problems
  - Many considered "black-boxes" since too complex to directly interpret
- Led to research area of explainable machine learning
  - Goal: provide methods to explain predictions made by black-box models
  - Method overview papers/books: (Gilpin, Bau, Yuan, Bajwa, Specter, and Kagal, 2018; Guidotti, Monreale, Ruggieri, Turini, Pedreschi, and Giannotti, 2018; Ming, 2017; Mohseni, Zarei, and Ragan, 2018; Molnar, 2019).
- European General Data Protection Regulation (GDPR) implemented in 2018 includes a "right to explanation"

Goodman and Flaxman (2016): "It is reasonable to suppose that any adequate explanation would, at a minimum, provide an account of how input features relate to predictions, allowing one to answer questions such as: Is the model more or less likely to recommend a loan if the applicant is a minority?"

# Explainability versus Interpretability

No accepted definitions for explainability and interpretability (Gilpin, Bau, Yuan, et al., 2018; Lipton, 2016; Molnar, 2019; Montavon, Samek, and Müller, 2017; Murdoch, Singh, Kumbier, Abbasi-Asl, and Yu, 2019)

My definitions (implicitly used by Rudin (2018) and Ribeiro, Singh, and Guestrin (2016)):

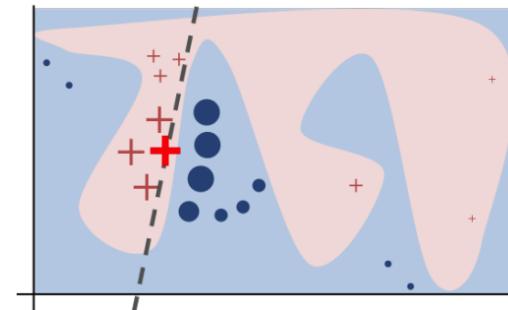
**Interpretability** is the ability to **directly use model parameters** to understand the mechanism of how the model **makes predictions**.

- Linear regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

**Explainability** is the ability to **use the model in an indirect manner** to understand the relationships in the data captured by the model.

- LIME: local interpretable model-agnostic explanations (Ribeiro, Singh, and Guestrin, 2016)



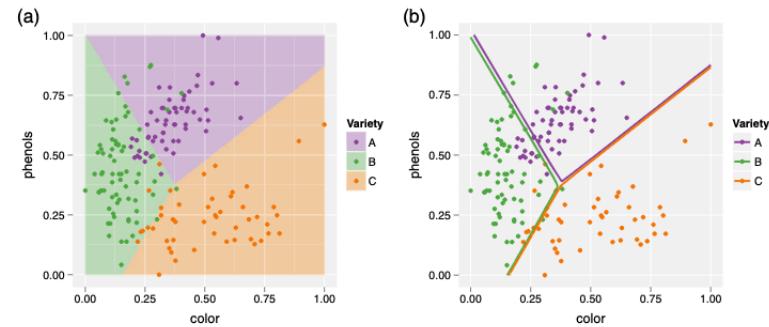
# Model Agnostic Methods

## General Model Visualizations: Strategies for understanding any model

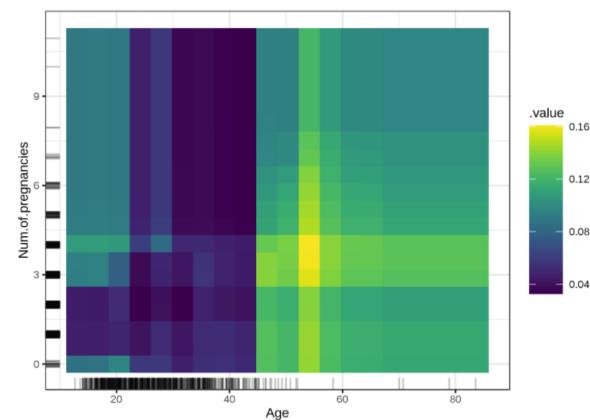
- Removing the blindfold (Wickham, Cook, and Hofmann, 2015)

## Global Methods: Explanation for model as a whole

- Partial dependence plots (Friedman, 2001) and extensions:
  - Interactive partial dependence plots (Krause, Perer, and Ng, 2016)
  - Individual conditional expectation plots (Goldstein, Kapelner, Bleich, and Pitkin, 2013)
  - Accumulated local effect plots (Apley and Zhu, 2016)
  - Feature interaction plots (Friedman and Popescu, 2008; Greenwell, Boehmke, and McCarthy, 2018; Hooker, 2004)
- Global feature importance plots (Fisher, Rudin, and Dominici, 2018; Altmann, Tološi, Sander, and Lengauer, 2010; Casalicchio, Molnar, and Bischl, 2019)
- Global surrogate models (Molnar, 2019)



Model in data-space from Wickham, Cook, and Hofmann (2015).



Partial dependence plot from Molnar (2019).

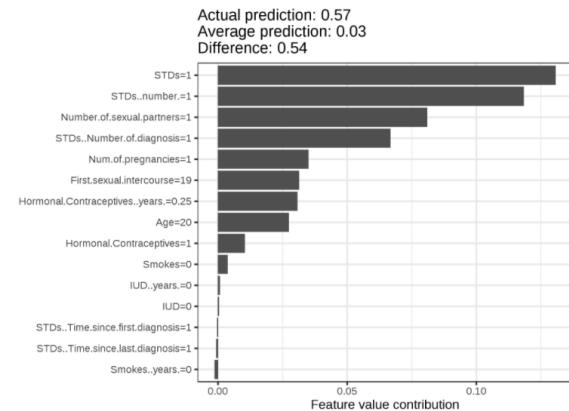
# Model Agnostic Methods

## Local Methods: Explanation for an individual prediction

- Individual conditional importance plots (Casalicchio, Molnar, and Bischl, 2019)
- LIME (Ribeiro, Singh, and Guestrin, 2016)
- Anchors (scoped rules) (Ribeiro, Singh, and Guestrin, 2018)
- Shapely values
- SHAP (Lundberg and Lee, 2017)
- breakDown (Staniak and Biecek, 2018)

## Example Based: Explanations based on examples from the data

- Counterfactual examples (Wachter, Mittelstadt, and Russell, 2017; Martens and Provost, 2014; Looveren and Klaise, 2019; Laugel, Lesot, Marsala, Renard, and Deryniecki, 2017)
- Adversarial examples (Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus, 2013; Goodfellow, Shlens, and Szegedy, 2014; Biggio and Roli, 2018; Su, Vargas, and Sakurai, 2019)
- Prototypes and criticisms
- Influential instances (Koh and Liang, 2017)



Bar chart of shapley values from Molnar (2019).



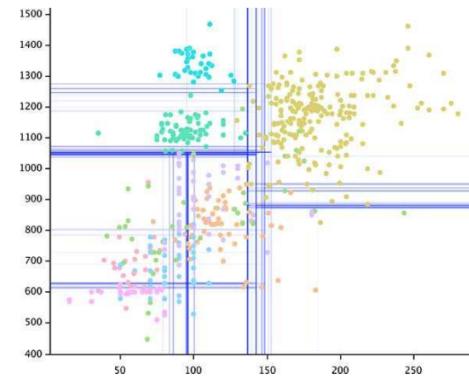
Adversarial example where one pixel change affects prediction from Su (2019).

# Model Specific Methods

## Random Forests

- Random forest impurity based feature importance (Breiman, 2001)
- Sectioned scatterplots (Urbanek, 2008)
- Trace plots of trees (Urbanek, 2008)
- Simplified model (Hara and Hayashi, 2016a)
- Forest floor visualizations (Welling, Refsgaard, Brockhoff, and Clemmensen, 2016)
- Interactive visualizations

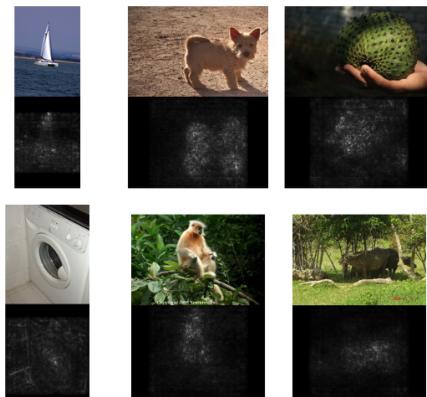
(Beckett, 2018; da Silva, Cook, and Lee, 2017)



Sectioned scatterplot from Urbanek (2008).

## Neural Networks

- Extracting tree structures (Craven and Shavlik, 1996)
- Saliency maps (Simonyan, Vedaldi, and Zisserman, 2013)
- Feature visualization (Olah, Mordvintsev, and Schubert, 2017)
- Grand tours (Li, Zhao, and Scheidegger, 2020)
- Flows (Halnaut, Giot, Bourqui, and Auber, 2020)



Saliency maps from Simonyan, Vedaldi, and Zisserman (2013).

# Assessments of Explainable Machine Learning

## General

Argument against black box model explanations (Rudin, 2018):

- "Explanations must be wrong."
- Explanations may not:
  - be faithful to the original model
  - make sense
  - be detailed enough to understand the "black-box" model
- Debunks accuracy and interpretability trade-off myth
- Use interpretable models for high-stakes decisions

## Method Specific

Assessment of LIME (Laugel, Renard, Lesot, Marsala, and Detyniecki, 2018):

- How to choose a local region?

Assessment of counterfactual examples (Laugel, Renard, Lesot, et al., 2018):

- Issues with unjustified counterfactual examples

Assessment of saliency maps (Kindermans, Hooker, Adebayo, Alber, Schütt, Dähne, Erhan, and Kim, 2017):

- Transformation to input data affects saliency map but not model

# Overview of Dissertation Chapters

## Chapter 1: Visual Diagnostics for LIME

- Discuss importance of assessing LIME
- Suggest the use of visualizations for assessment and provide example visualizations

## Chapter 2: Visualizations for Explaining Random Forests

- Use clustering to identify key tree structures within the random forest
- Improve visualizations for use of trees as global surrogate models

## Chapter 3: Visualizations for Explaining Neural Networks

- Project for my internship with Sandia National Labs
- Application with functional data
- Visualizations for the explaining and understanding the models

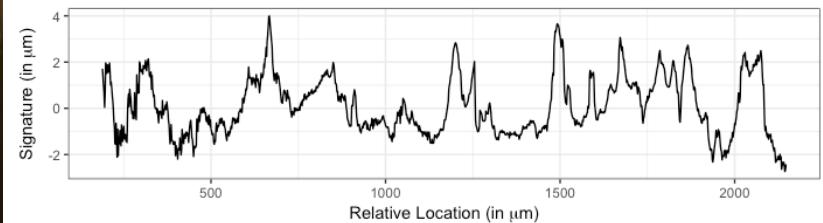
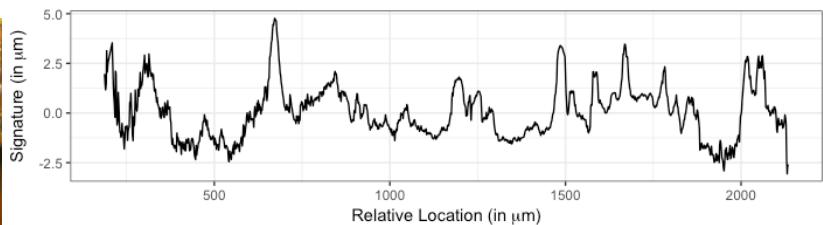
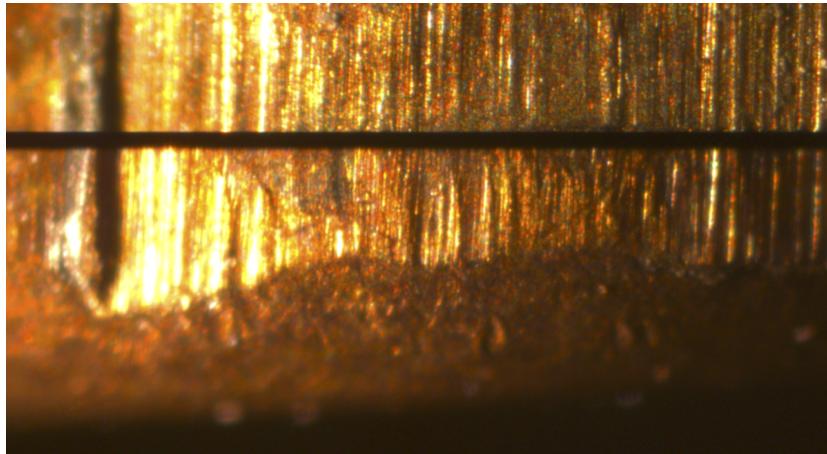
# Chapter 1

Visual Diagnostics of a Model Explainer -- Tools for the  
Assessment of LIME Explanations

# Motivation

Hare, Hofmann, and Carriquiry (2016):

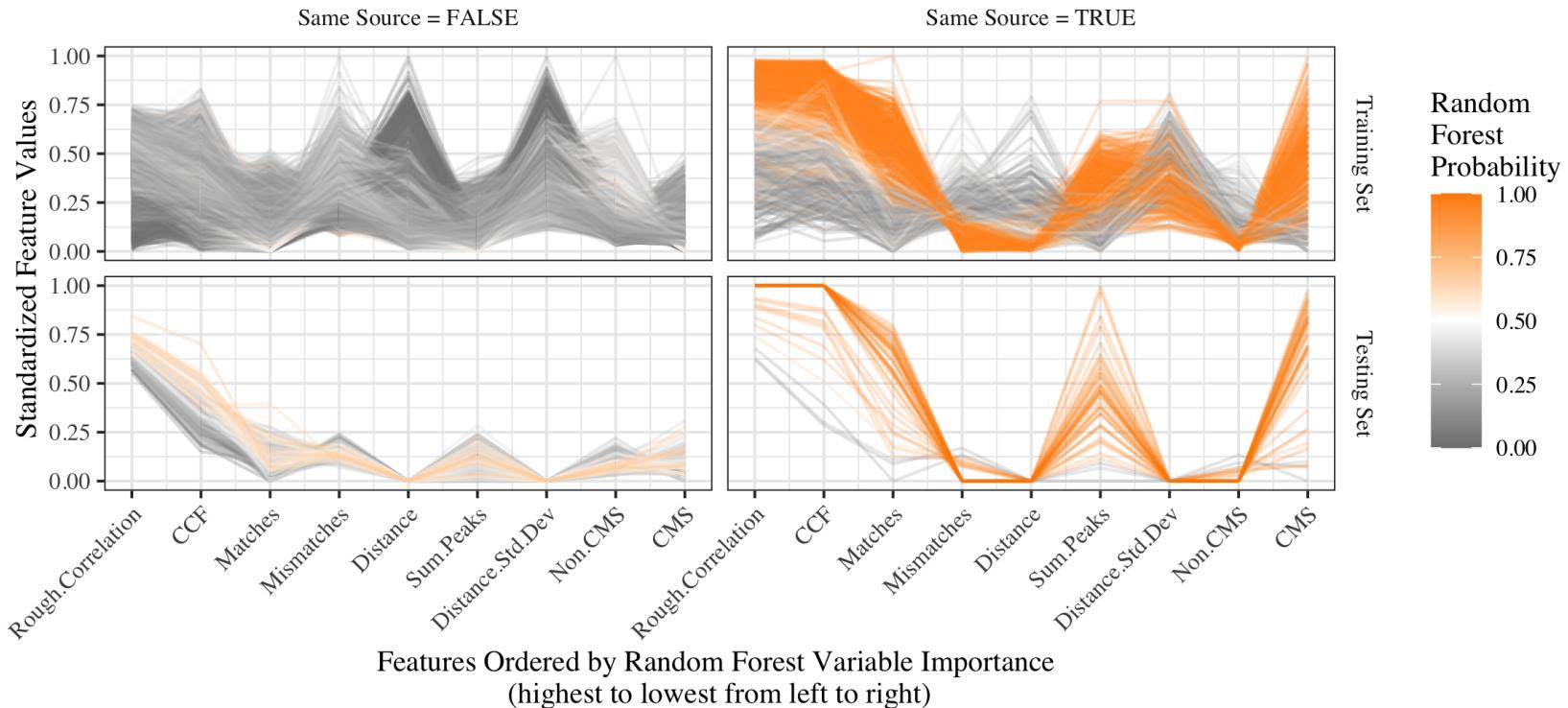
- Want to provide quantitative evidence for whether two bullets were fired from the same gun
- Use high definition scans of striations on bullet lands to extract "signatures"
- Compute similarity features to compare two signatures



# Motivation

Hare, Hofmann, and Carriquiry (2016) approach:

- Random forest model
- 9 signature similarity features
- Returns a score for the comparison of two lands



# Motivation

**Our Original Goal:** Provide explanations for specific signatures comparisons

**Attempt:** Applied LIME

**Result:** Unreasonable explanations (e.g., LIME explanation does not agree with random forest prediction)

**Example:** Known non-match

# Conceptual Depiction of LIME

LIME (Ribeiro, Singh, and Guestrin, 2016):

- Local
- Interpretable
- Model-agnostic
- Explanation

**Concept:** For *one prediction of interest*

- Focus on a neighborhood around the prediction of interest
- Use an inherently interpretable model
- Understand the complex model
- Capture the relationship between the complex model predictions and predictor variables

# Importance of Assessing LIME

Additional layer of complexity:

- Start with a complex model
- LIME uses another model
- End with two models to assess

Questions raised:

- Explainer model a good approximation?
- Appropriate local region?
- Relationship linear in the local region?
- Which tuning parameter settings to use when applying LIME?

# Visualizations for Model Assessment

Claims about LIME (Ribeiro, Singh, and Guestrin, 2016):

- **Interpretability:** Easy to interpret the explainer model to provide meaningful explanations
- **Faithfulness:** Explainer model sufficiently captures the relationship between the complex model predictions and the features in the local region around a prediction of interest
- **Linearity:** Using a ridge regression as the explainer model assumes a linear relationship between complex model predictions and the features
- **Localness:** Explanations are local in regards to a prediction of interest

We suggest visual diagnostics to assess these claims:

- **Diagnostics for individual explanations**
- **Diagnostics for sets of explanations**
- **Diagnostics for comparisons of tuning parameters**

**Note:** We focus on binary response variable and continuous predictor variables

# Sine Example Data

Training data: 500 observations

Testing data: 100 observations

Black-box model: random forest

$$x_1 \sim \text{Unif}(-10, 10)$$

$$x_2 \sim \text{Unif}(-10, 10)$$

$$x_3 \sim \mathcal{N}(0, 1)$$

$$y = \begin{cases} \text{blue} & \text{if } x'_2 > \sin(x'_1) \\ \text{red} & \text{if } x'_2 \leq \sin(x'_1). \end{cases}$$

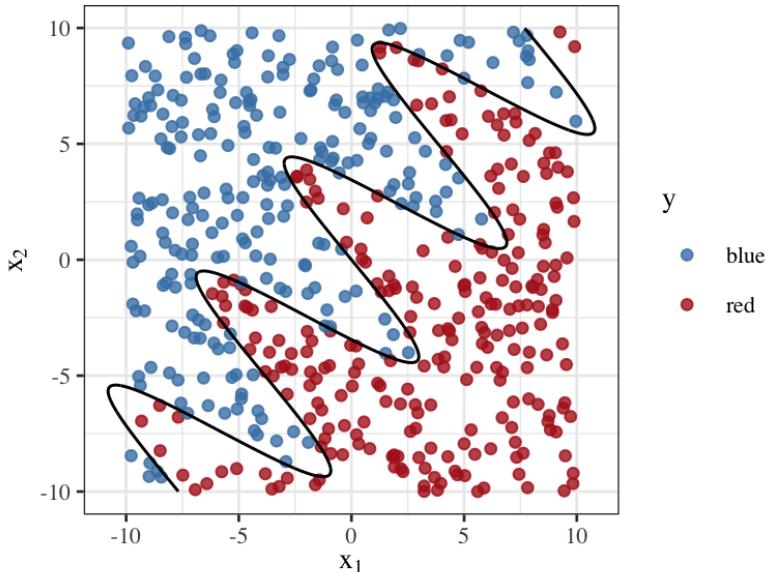
where

$$x'_1 = x_1 \cos(\theta) - x_2 \sin(\theta)$$

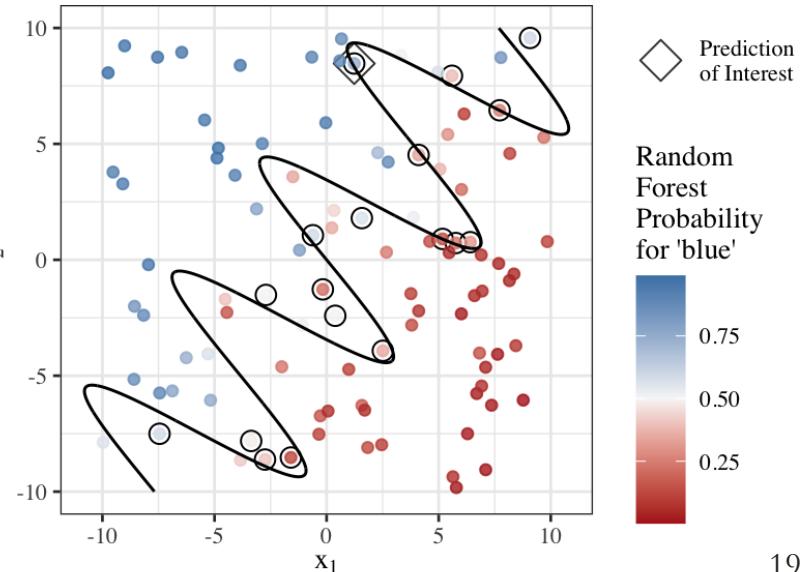
$$x'_2 = x_1 \sin(\theta) + x_2 \cos(\theta)$$

$$\theta = -0.9$$

Training Data



Testing Data



# Set 1 of Visualizations

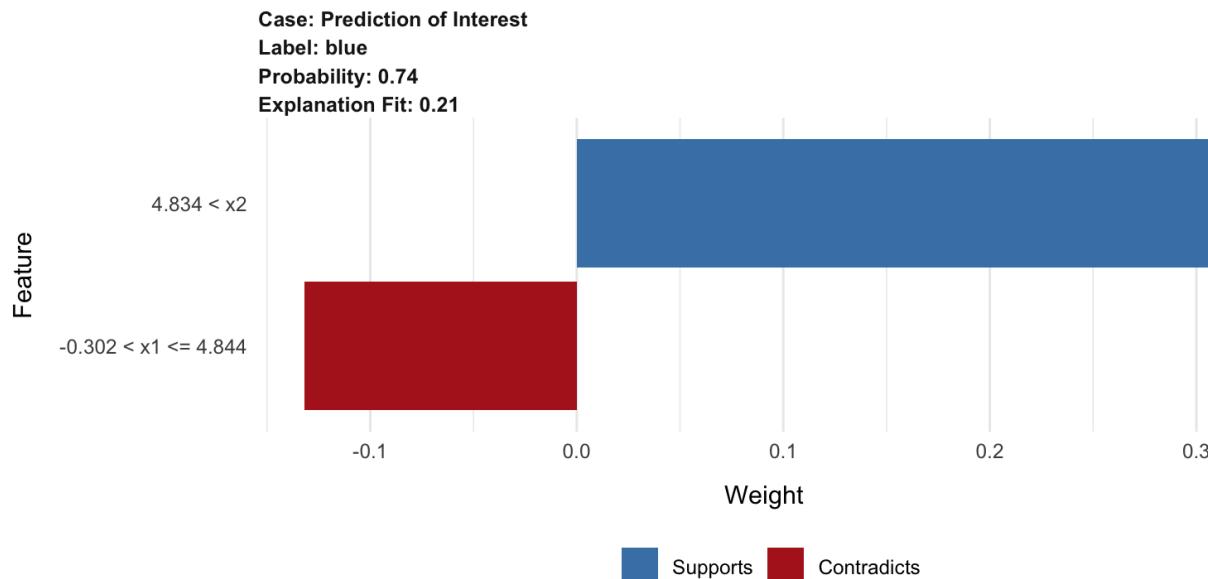
Diagnostics for Individual Explanations

# Diagnostics for Individual Explanations

Applied LIME using *lime* R package (Pedersen and Benesty, 2020)

- To prediction of interest
- Number of features to return in explanation: 2
- Default tuning parameters settings

*lime* R package visualization of the explanation:



# Diagnostics for Individual Explanations

## Step 1a: Data Simulation

- Sample 4999 observations uniformly from 4 quantile bins for each feature in the training data

## Corresponding Diagnostic:

### *Training Data Plot*

- **Axes**: two features selected by LIME
- **Points**: training data
- **Color**: observed response
- **Lines**: 4 quantile bins boundaries



# Diagnostics for Individual Explanations

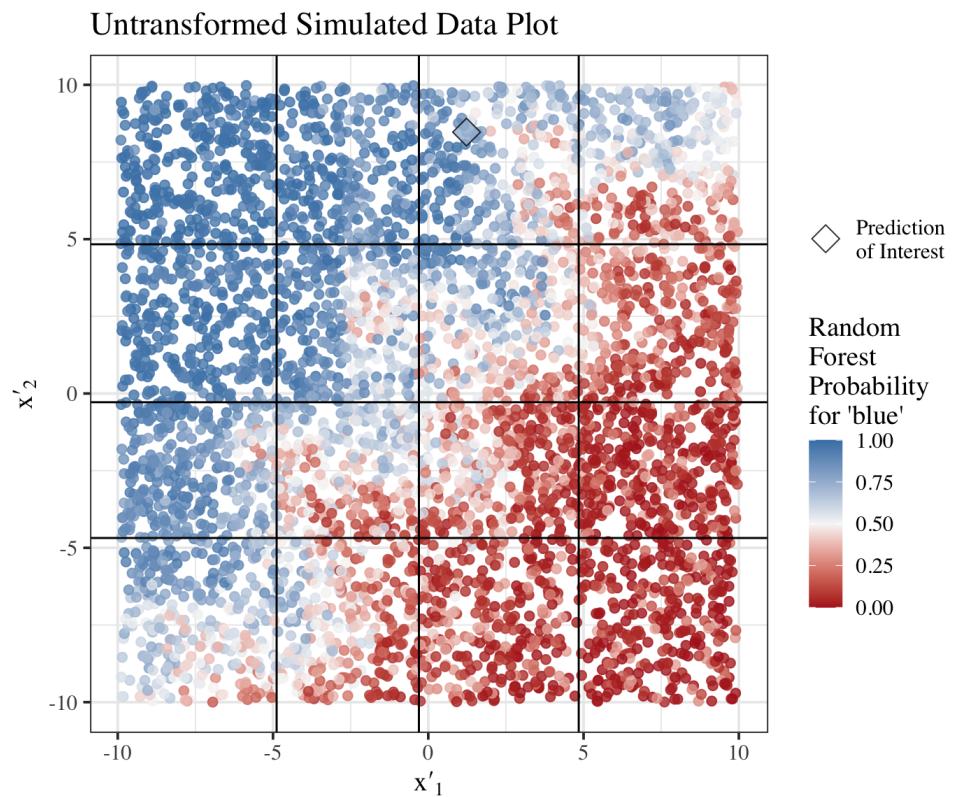
## Step 1b: Complex Model Predictions

- Apply complex model to simulated data to obtain predictions

### Corresponding Diagnostic:

#### *Untransformed Simulated Data Plot*

- **Axes:** simulated data features
  - denoted as  $x'_1$  and  $x'_2$ )
- **Points:** simulated data
- **Diamond:** prediction of interest
- **Color:** random forest prediction (for 'blue')
- **Lines:** 4 quantile bins boundaries



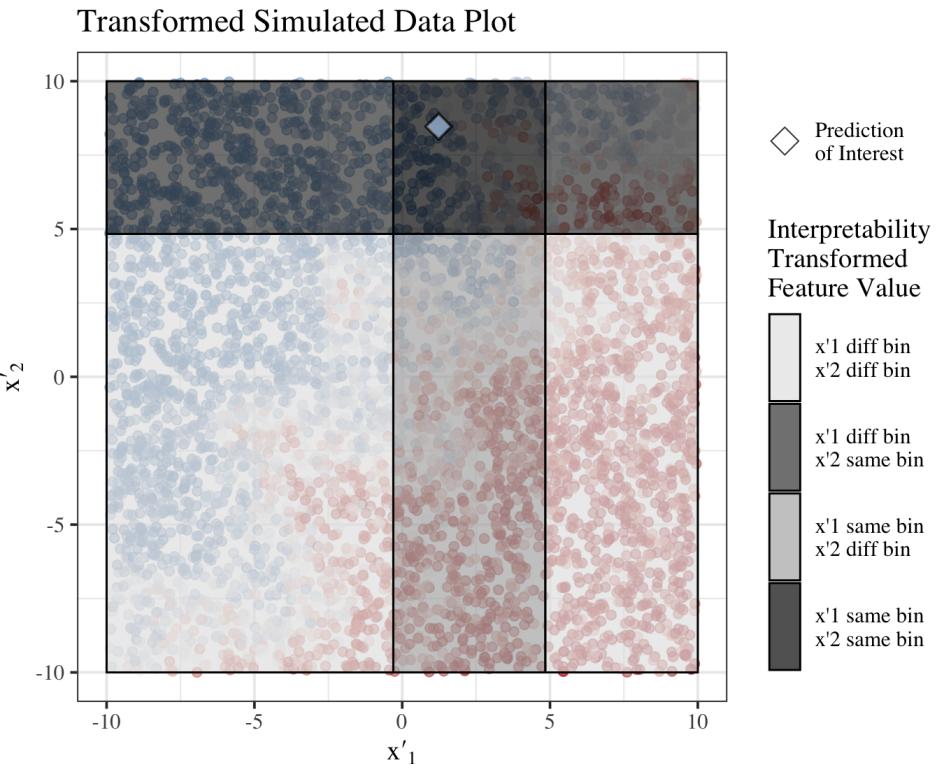
# Diagnostics for Individual Explanations

## Step 1c: Interpretability Transformation

- Convert continuous features to binary variables based on whether the observation falls in the same quantile bin as the prediction of interest or not

Corresponding Diagnostic: *Transformed Simulated Data Plot*

- Axes:** Simulated data features
- Points:** simulated data
- Diamond:** prediction of interest
- Color:** random forest prediction (for 'blue')
- Rectangle Shades:** interpretability transformed feature regions



# Diagnostics for Individual Explanations

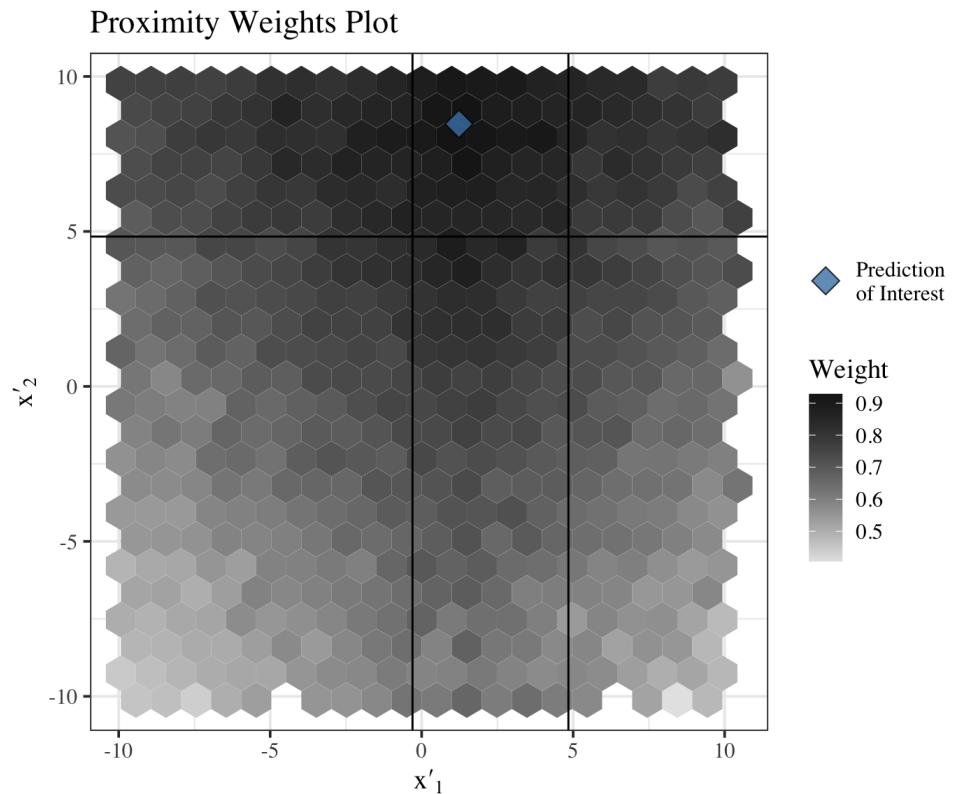
## Step 2a: Assign Weights

- Assign weights to simulated data based on proximity to the prediction of interest (using untransformed feature values)
- Gower distance metric

### Corresponding Diagnostic:

#### *Proximity Weights Plot*

- **Axes:** Simulated data features
- **Rectangle Color:** average weight within hexagon region
- **Lines:** interpretability transformation boundaries
- **Diamond:** prediction of interest



# Diagnostics for Individual Explanations

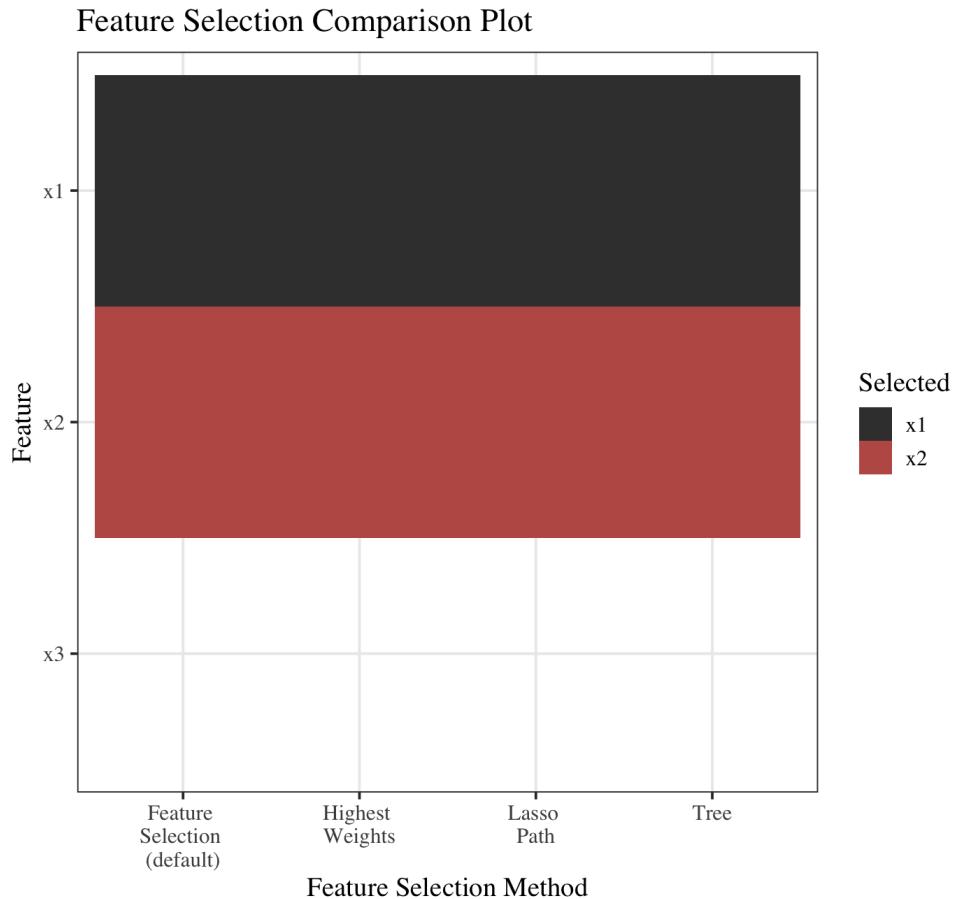
## Step 2b: Feature Selection

- Ridge regression model fit to simulated data
  - Response: complex model predictions
  - Features: interpretability transformed features
- Forward selection (if less than 6 features specified)

### Corresponding Diagnostic:

#### *Feature Selection Comparison Plot*

- **Axes:** Training data features versus feature selection method
- **Tile Color:** indicates if feature selected



# Diagnostics for Individual Explanations

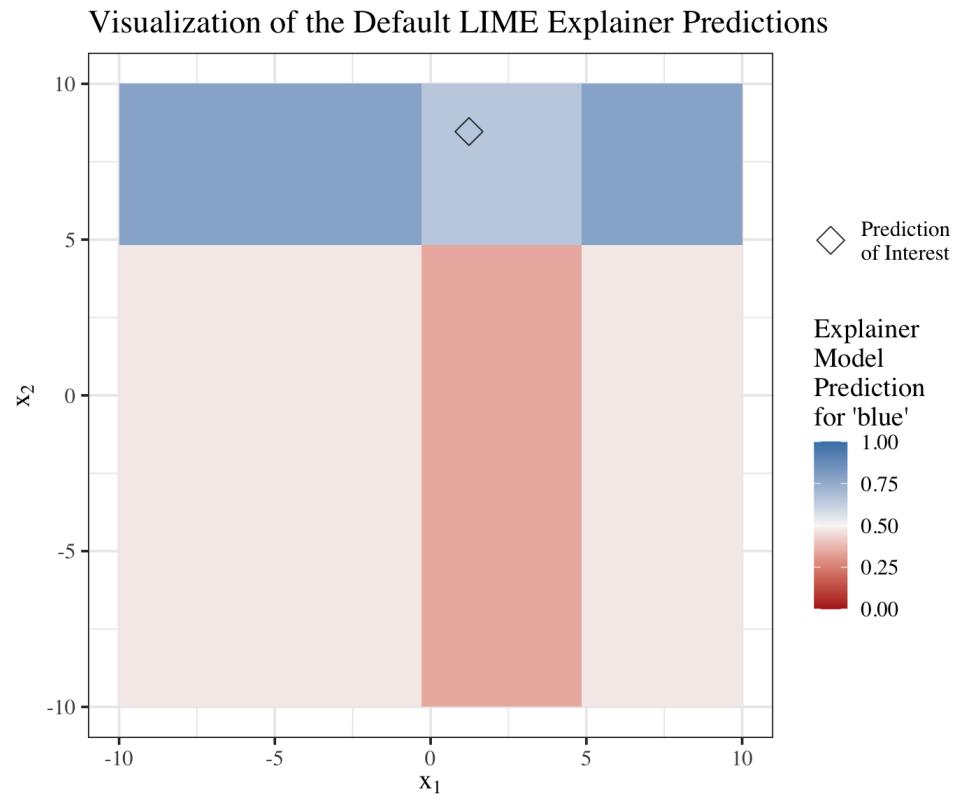
## Step 2c: Fit Explainer Model

- Ridge regression model fit to simulated data
  - Response: complex model predictions
  - Features: selected interpretability transformed features

### Corresponding Diagnostic:

#### *Explainer Model Prediction Plot*

- **Axes**: simulated data features
- **Diamond**: prediction of interest
- **Rectangle**: Interpretability transformed regions
- **Color**: explainer model prediction



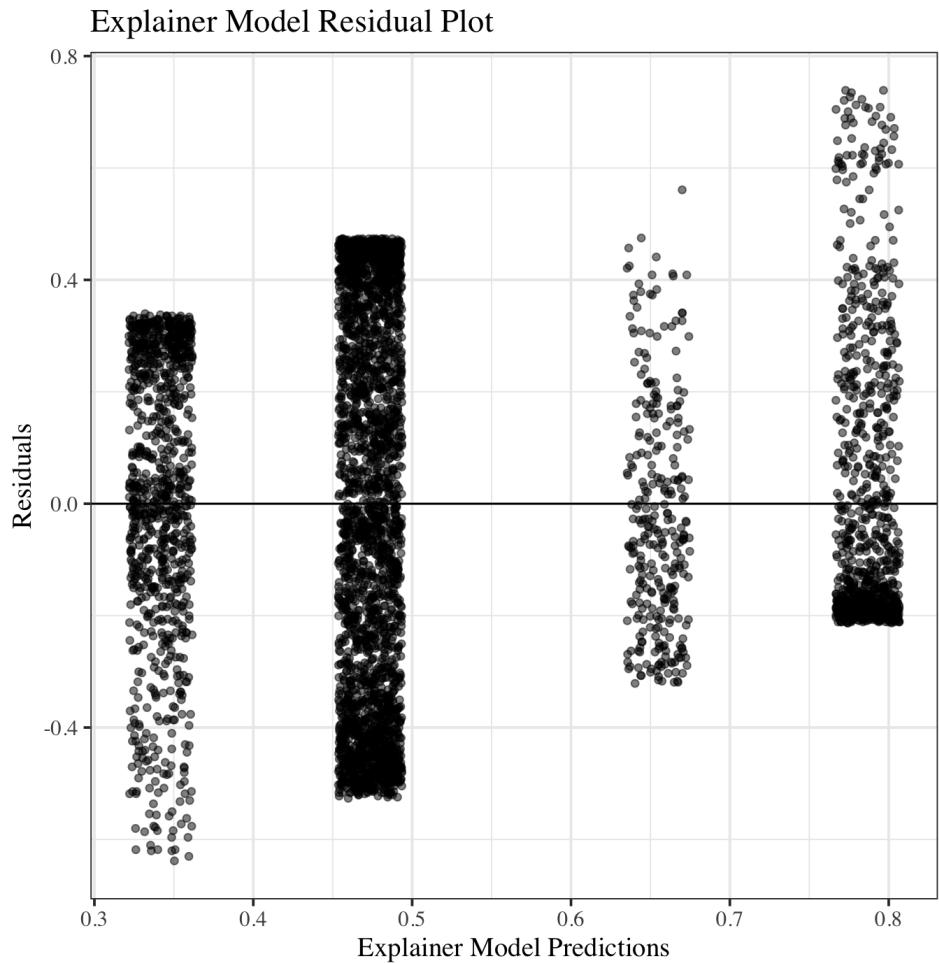
# Diagnostics for Individual Explanations

Step 2c: Fit Explainer Model (continued)

Corresponding Diagnostic:

*Explainer Model Residual Plot*

- Residual plot for the explainer model
- Points have been jittered in the x-direction



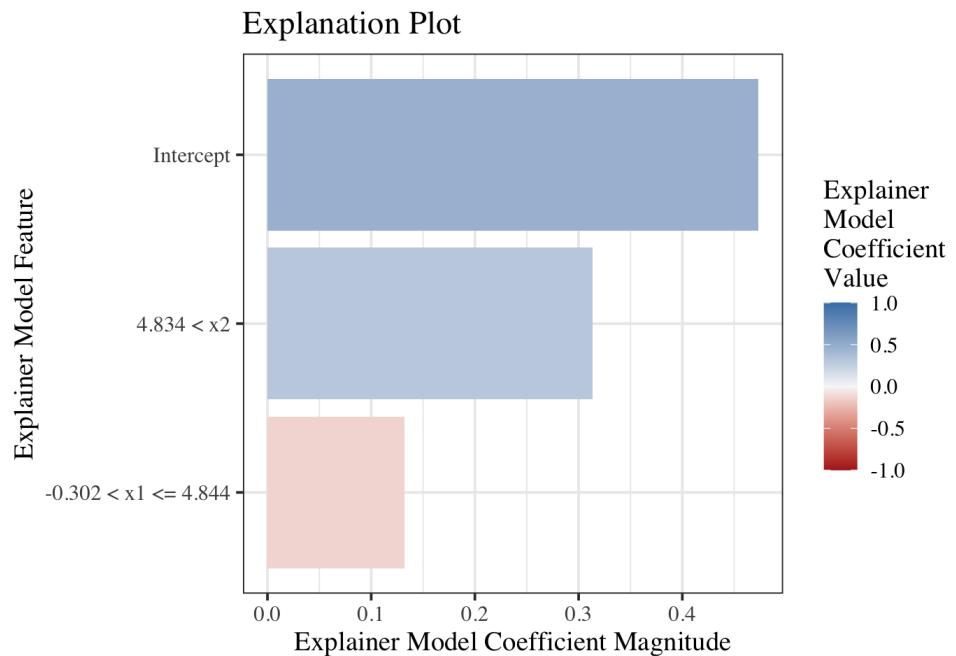
# Diagnostics for Individual Explanations

## Step 3a: Explainer Model Interpretation

- Interpret coefficients of explainer model
- Explains complex model prediction

### Corresponding Diagnostic: *Explanation Plot*

- Adaptation of plot from *lime* R package
- **Axes:** Explainer model (interpretability transformed) feature versus explainer model coefficient
- **Bar Length:** magnitude of explainer model coefficients
- **Bar Color:** value of the explainer model coefficients

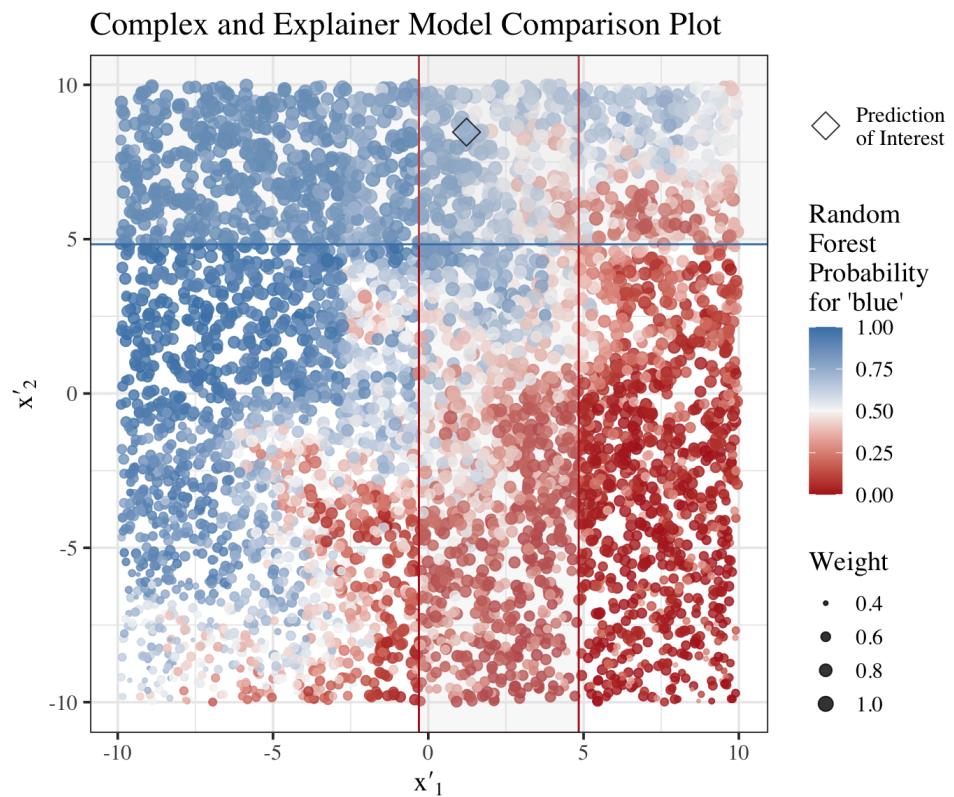


# Diagnostics for Individual Explanations

## Step 3b: Explainer Model Interpretation (continued)

### Corresponding Diagnostic: *Complex and Explainer Model Comparison Plot*

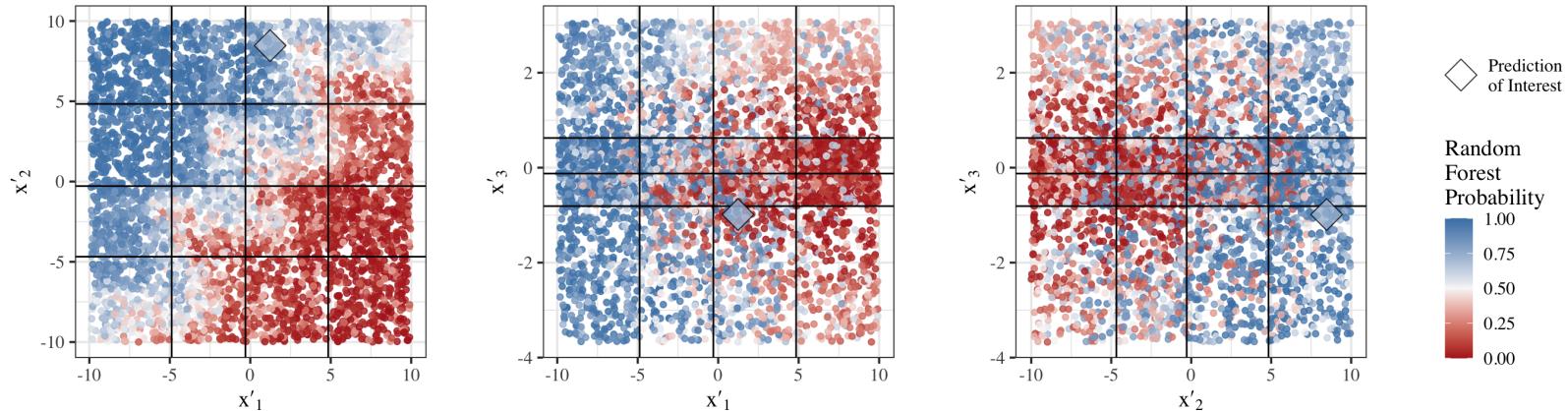
- **Axes:** simulated data features
- **Points:** simulated data
- **Diamond:** prediction of interest
- **Point Color:** random forest prediction (for 'blue')
- **Point Size:** proximity weight
- **Lines:** interpretability transformation boundaries
- **Line Color:** explainer coefficient supports a random forest prediction of 'blue' (blue) or not (red)



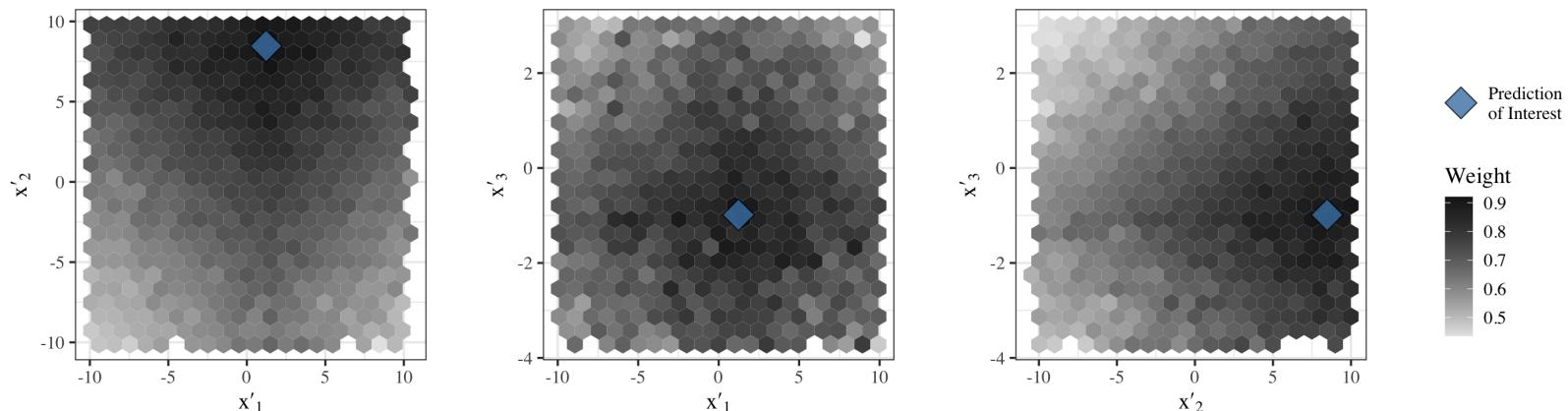
# Diagnostics for Individual Explanations

## Extensions to Higher Dimensions

Untransformed Simulated Data Plots



Proximity Weights Plots



# Set 2 of Visualizations

Diagnostics for Sets of Explanations

# Diagnostics for Sets of Explanations

Applied LIME using *lime* R package (Pedersen and Benesty, 2020)

- All test data observations
- Number of features to return in explanation: 2
- Default tuning parameters settings

Plot of all explanations from *lime* R package:

# Diagnostics for Sets of Explanations

## Explanation Set Plot

Provides an overview of groupings of LIME explanations

- Adaptation to the *lime* plot of all explanations
- **Y-axis:** Interpretability transformed features
- **X-axis:** Observation in the data set
- **Tile Color:** Ridge regression coefficient from the corresponding model

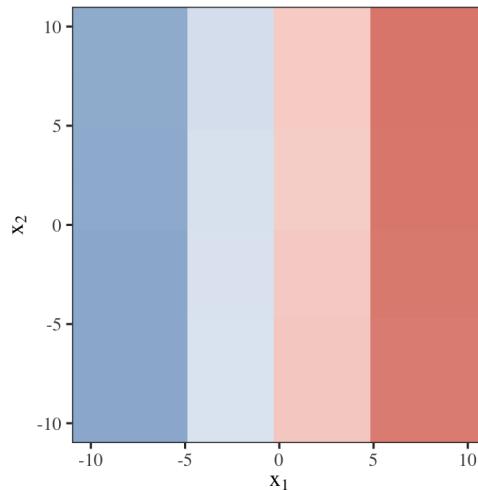
# Diagnostics for Sets of Explanations

## Average Coefficient Plots

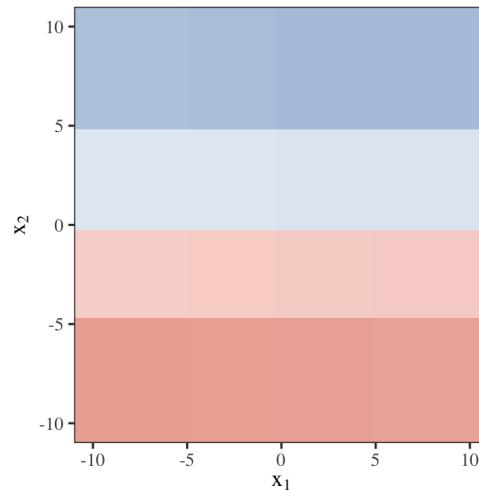
Provides a summary of explainer model coefficients across the set of explanations within quantile bin regions

- **Axes:** Test data features
- **Cells:** Intersections of quantile bins
- **Color:** Average of ridge regression coefficients (or sum of ridge regression coefficients) for observations within a cell

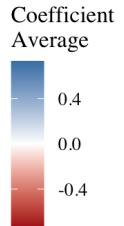
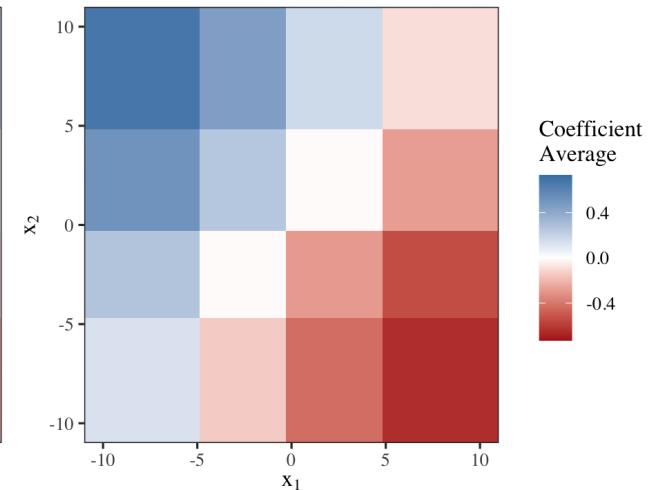
Average of  $x_1$  Coefficients



Average of  $x_2$  Coefficients



Average of  $x_1$  plus  $x_2$  Coefficients

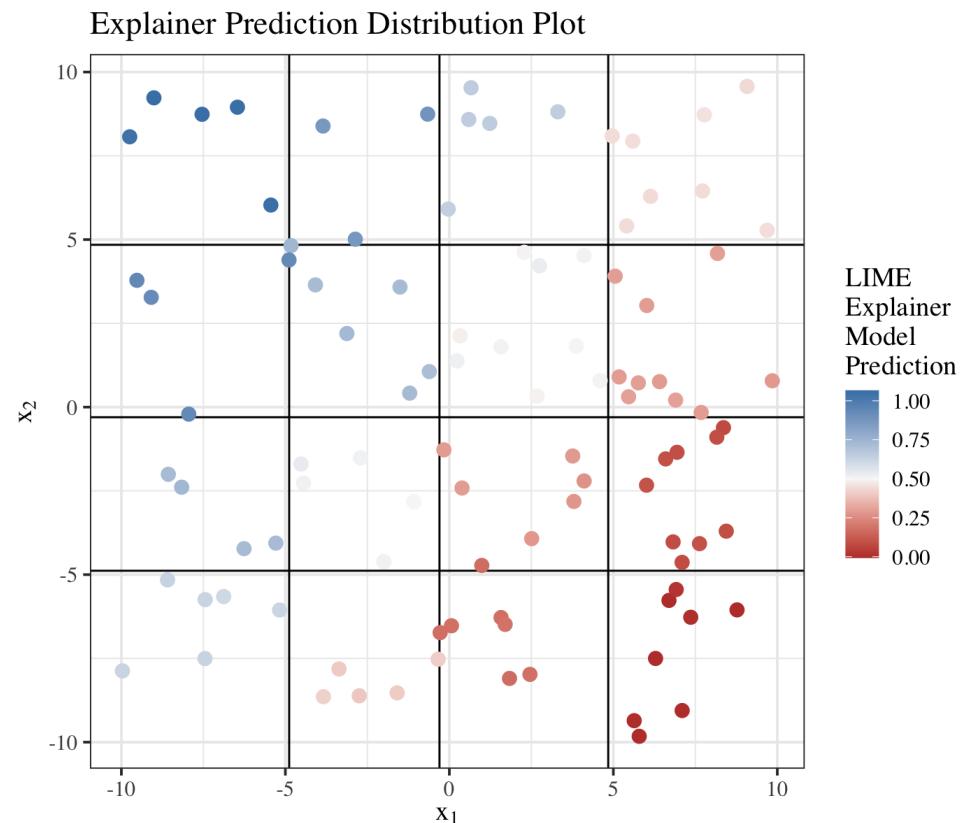


# Diagnostics for Sets of Explanations

## Explainer Prediction Distribution Plot

Shows the relationship between the explainer model predictions and the quantile bins

- **Axes:** Test data features
- **Points:** Test data observations
- **Color:** Explainer model prediction
- **Lines:** Quantile bin boundaries



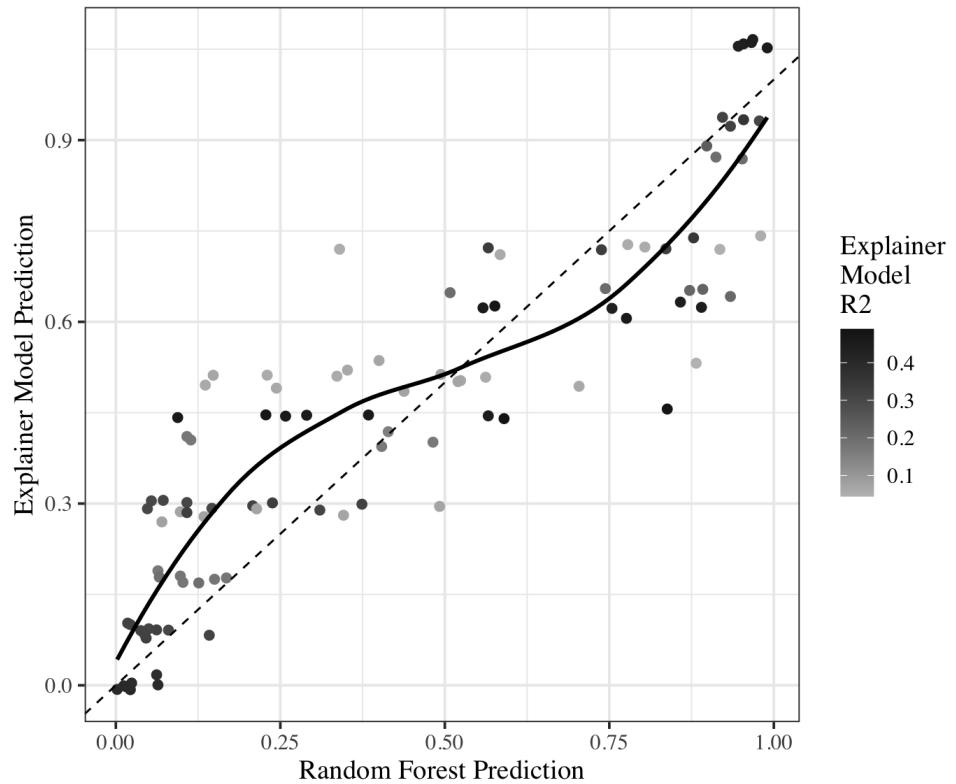
# Diagnostics for Sets of Explanations

## Prediction Comparison Plot

Shows the relationship between the explainer model and complex model predictions

- Y-Axis: Explainer model predictions
- X-Axis: Complex model predictions
- Points: Observations from the test data
- Color: Corresponding explainer model  $R^2$
- Dashed Line: 1-1 line
- .red.[Solid Line]: Loess smoother fit to the points

Prediction Comparison Plot

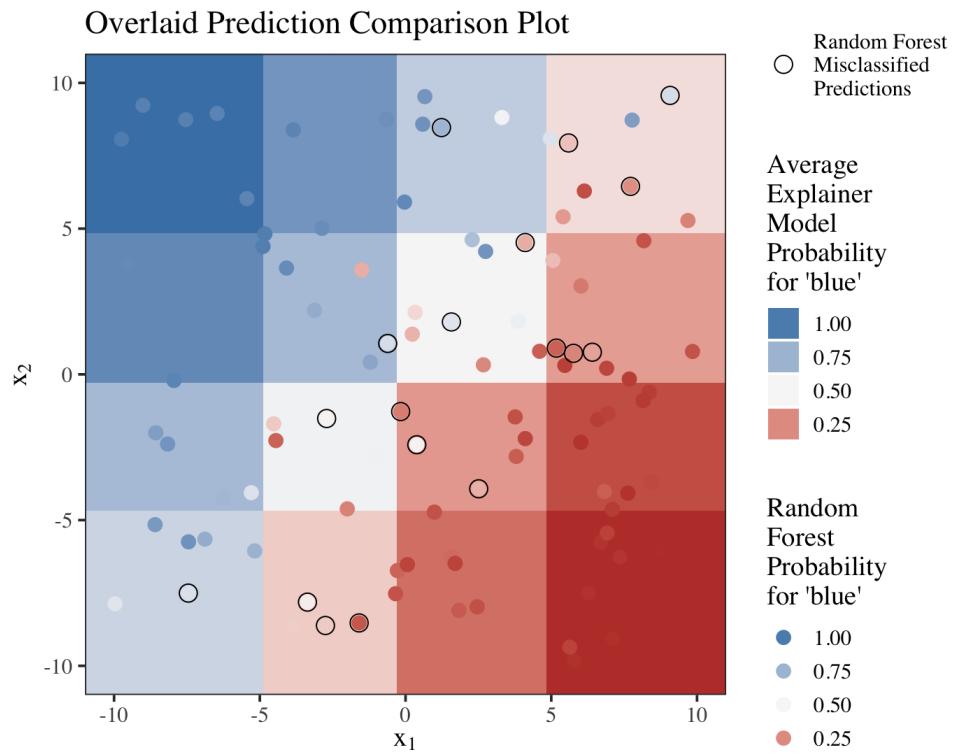


# Diagnostics for Sets of Explanations

## Overlaid Prediction Comparison Plot

Shows the relationship between the average explainer model predictions within a quantile bins cell and the complex model predictions

- **Axes:** Test data features
- **Points:** Test data observations
- **Point Color:** Explainer model prediction
- **Cells:** Intersections of quantile bins
- **Cell Color:** Average explainer model prediction
- **Black Circles:** Identify observations misclassified by the complex model



# Set 3 of Visualizations

Diagnostics for Comparisons of Tuning Parameters

# Diagnostics for Comparisons of Tuning Parameters

## Tuning Parameters

Definition:

- LIME application method options

Examples:

- simulation method
- number of observations to simulate
- distance metric
- feature selection method
- number of features to return

## Questions Raised

- How to select tuning parameters?
- How to assess tuning parameters?

## Assessment of Tuning Parameters

- Check if explanations depend on tuning parameters
- Compute metrics to compare performance

# Diagnostics for Comparisons of Tuning Parameters

Applied LIME using *lime* R package (Pedersen and Benesty, 2020)

- All test data observations
- Number of features to return in explanation: 2
- Multiple applications of LIME for each observation
  - 2 to 6 quantile bins
  - Kernel density simulation

Plot of all explanations from *lime* R package:

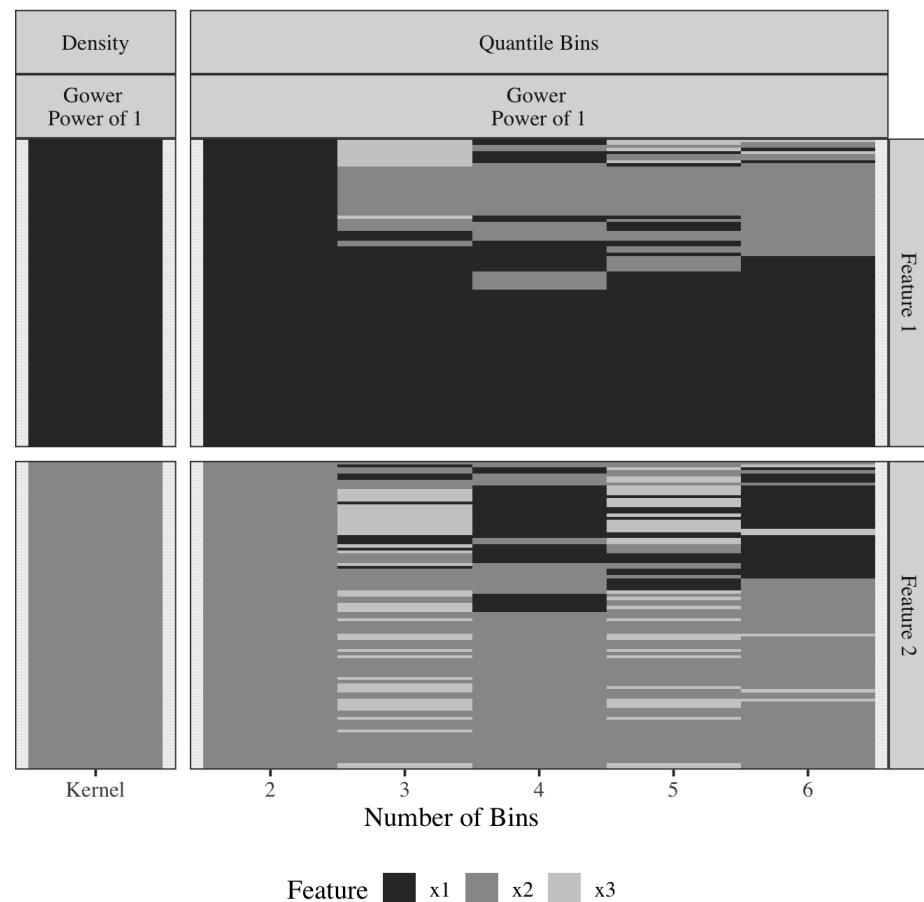
# Diagnostics for Comparisons of Tuning Parameters

## Feature Heatmap Plot

Provides an overview of the features selected by LIME across observations and tuning parameters

- **Y-Axis:** Test data observation
- **X-Axis:** Data simulation method
- **Y-Facet:** LIME feature order selection (first, second, etc.)
- **X-Facet:** Density based or bin based simulation method
- **Color:** Feature selected

## Feature Heatmap Plot



# Diagnostics for Comparisons of Tuning Parameters

## Notation for Assessment Metrics

### Data and Complex Model

For a set of  $E$  explanations

- $\mathbf{X}$  = matrix of observed data to be explained with  $K$  features and  $E$  observations
- $x_e$  = observed feature vector for observation  $e$
- $f$  = complex model
- $f(x_e)$  = complex model prediction for observation  $e$

### Simulated and Transformed Data

For  $x_e$  and set of tuning parameters  $t$

- $\mathbf{X}'_{e,t}$  = LIME simulated dataset with  $K$  features and  $S$  observations
- $x'_{e,t,s}$  = feature vector for simulated data observation  $s$
- $\mathbf{Z}'_{e,t}$  = matrix of interpretability transformed simulated data with  $K$  features and  $S$  observations
- $z'_{e,t,s}$  = interpretability transformed feature vector
- $z_{e,t}$  = interpretability transformed  $x_e$

# Diagnostics for Comparisons of Tuning Parameters

## Notation for Assessment Metrics

### Weights and Explainer Model

For  $x_e$  and set of tuning parameters  $t$

- $\pi_t =$  proximity distance metric
- $\pi_t(x_e, x'_{e,t,s}) =$  weight assigned to  $x'_{e,t,s}$   
which are the proximity between  $x_e$  and  
 $x'_{e,t,s}$
- $g_{e,t} =$  explainer model
- $g_{e,t}(z'_{e,s,t}) =$  explainer model prediction  
for interpretability transformed simulated  
data observation  $s$

### Summary of Indices

- $e =$  explanation ( $e = 1, \dots, E$ )
- $t =$  set of tuning parameters  
( $t = 1, \dots, T$ )
- $s =$  simulated observation ( $s = 1, \dots, S$ )

# Diagnostics for Comparisons of Tuning Parameters

## Metrics for the Assessment of LIME

**Average R Squared:** Average of explainer model  $R^2$  values over a set of explanations with the same tuning parameters to assess the linearity claim where  $R_{e,t}^2$  is the  $R^2$  for  $g_{e,t}$

$$R_{\text{ave}}^2 = \sum_{e=1}^E R_{e,t}^2$$

**Average Fidelity:** Average of fidelity metric introduced in Ribeiro, Singh, and Guestrin (2016) over a set of explanations with the same tuning parameters to assess the explainer model approximation to the complex model

$$\mathcal{L}_{\text{ave}} = \sum_{e=1}^E \mathcal{L}(f, g_{e,t}, \pi_t) = \sum_{e=1}^E \sum_{s=1}^S \pi_t(x_e, x'_{e,t,s}) \left( f(x_e) - g_{e,t}(z'_{e,t,s}) \right)^2$$

**Mean Squared Explanation Error (MSEE):** Comparable to an MSE for comparing complex model predictions to explainer model predictions for each explanation

$$MSEE = \frac{1}{E} \sum_{e=1}^E (f(x_e) - g_{e,t}(z_{e,t}))^2$$

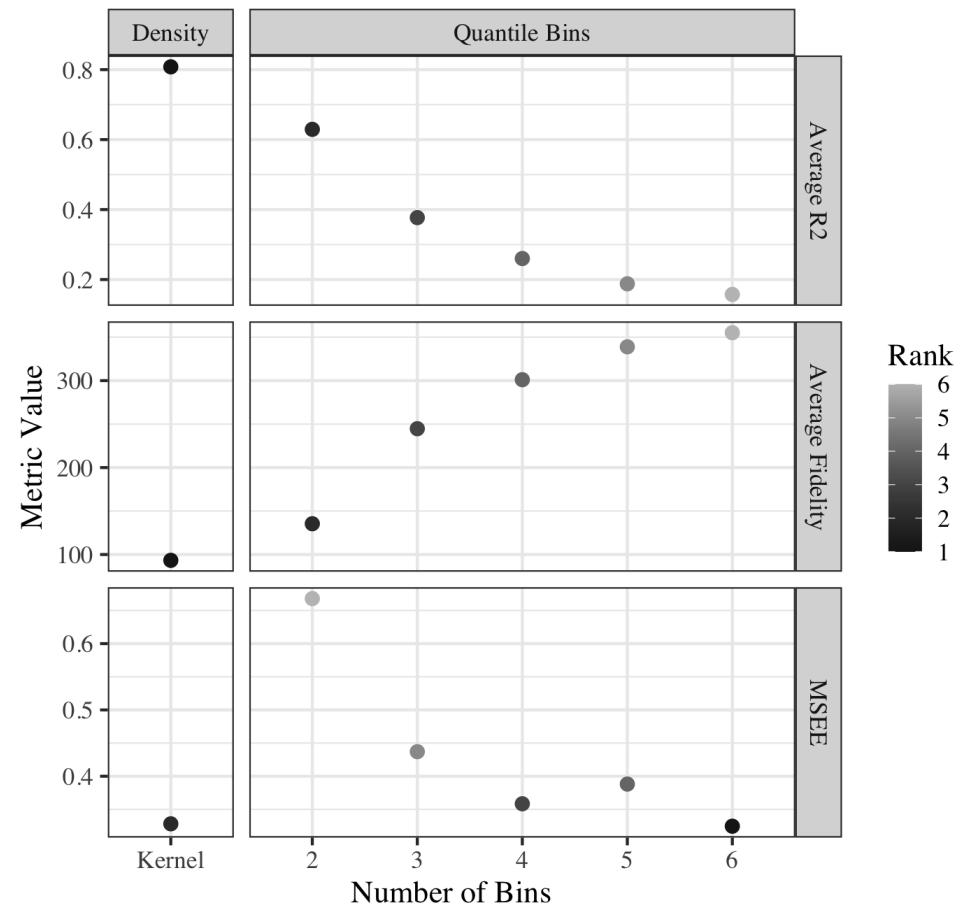
# Diagnostics for Comparisons of Tuning Parameters

## Assessment Metric Plot

Visualization of assessment metrics for comparing tuning parameter performance

- **Y-Axis:** Metric value
- **X-Axis:** Data simulation method
- **Y-Facet:** Metric type
- **X-Facet:** Density based or bin based simulation method
- **Point:** Represents application of LIME to a set of observations
- **Color:** Rank of the application based on metric value (within a metric)

## Assessment Metric Plot



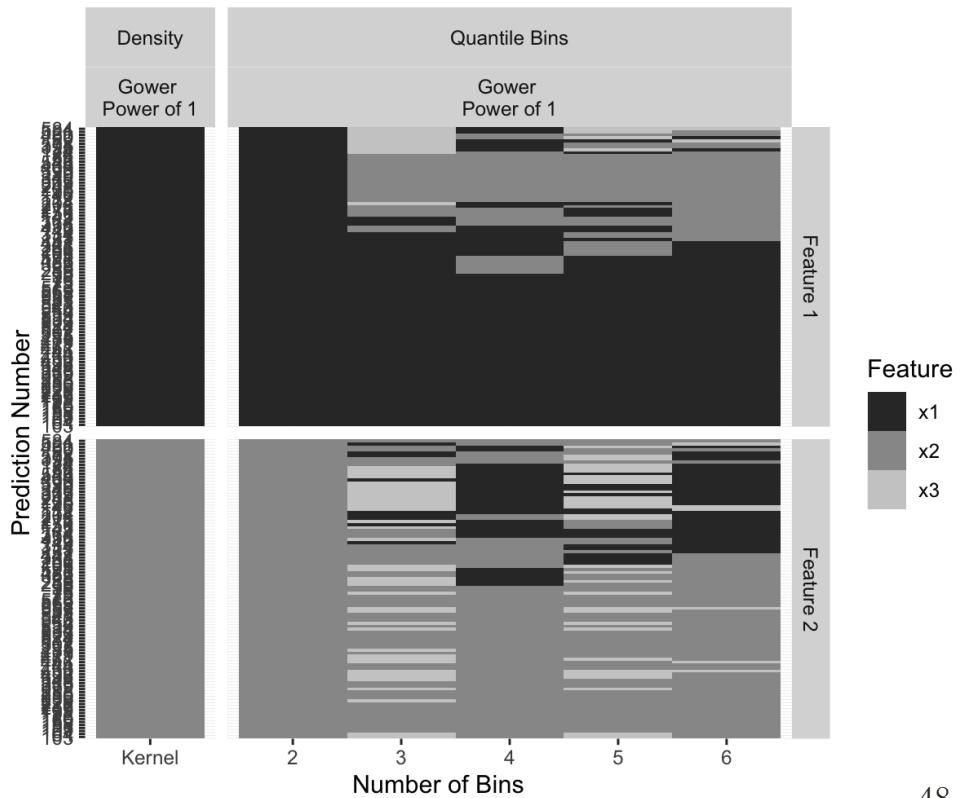
# Implementation and Application

# R package: limeaid

Our R package for visual LIME diagnostics:

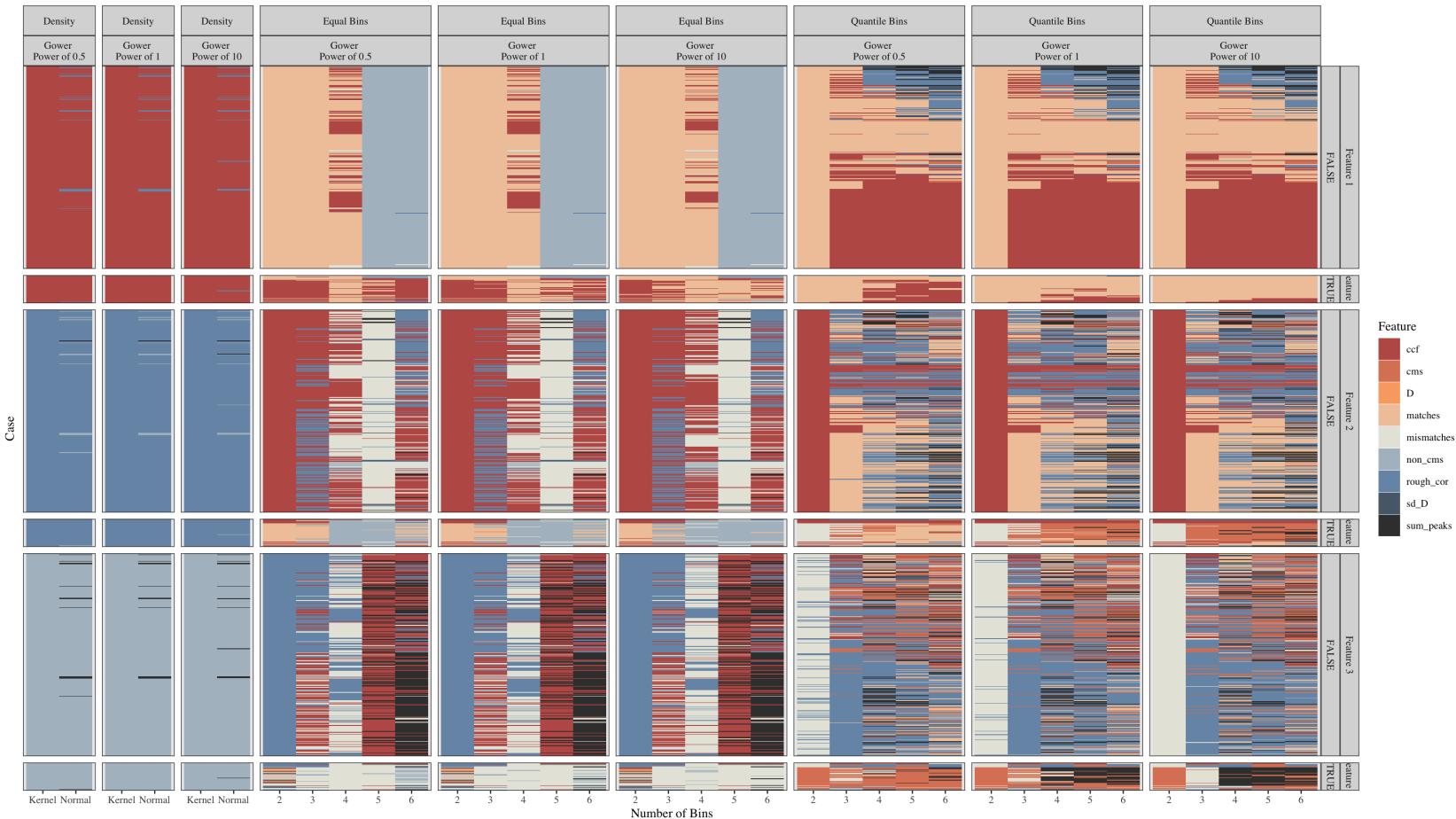
- Available on GitHub
  - [github.com/goodekat/limeaid](https://github.com/goodekat/limeaid)
- Functionality:
  - Applying LIME with multiple tuning parameters
  - Computing assessment metrics
  - Plots:
    - Complex and explainer model comparison plot
    - Feature Heatmap Plot
    - Assessment Metric Plot

```
feature_heatmap(  
  explanations = sine_lime_explain$explain,  
  order_method = "PCA"  
)
```



# Application of Visual Diagnostics to Bullet Random Forest

Feature heatmap of LIME explanation from Hare, Hofmann, and Carriquiry (2016) bullet comparison forest model



# Chapter 2

Explaining Random Forests using Clustering of Trees

# Current Plan

## Motivation

- Still interested in gaining insights for the random forest from (Hare, Hofmann, and Carriquiry, 2016)
- Focus on visualizations of the random forest trees

## Ideas

- Tree as a global explainer model
  - Plot global explainer on top of Urbanek (2008)'s trace plot
- Clustering to identify key trees in random forest
  - Use distance metric to compare trees

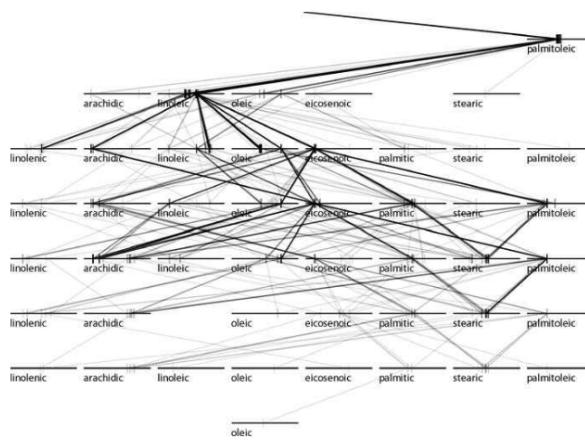
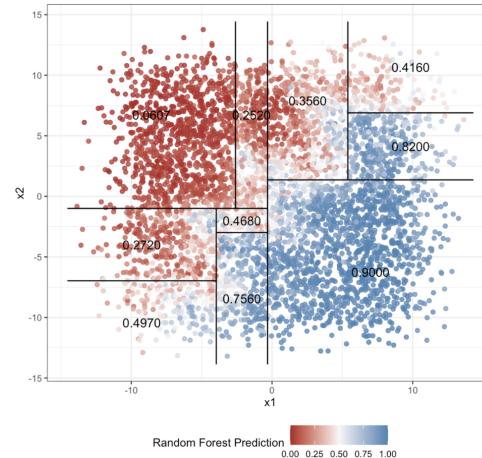


Figure 10.14. Trace plot of 100 bootstrapped trees

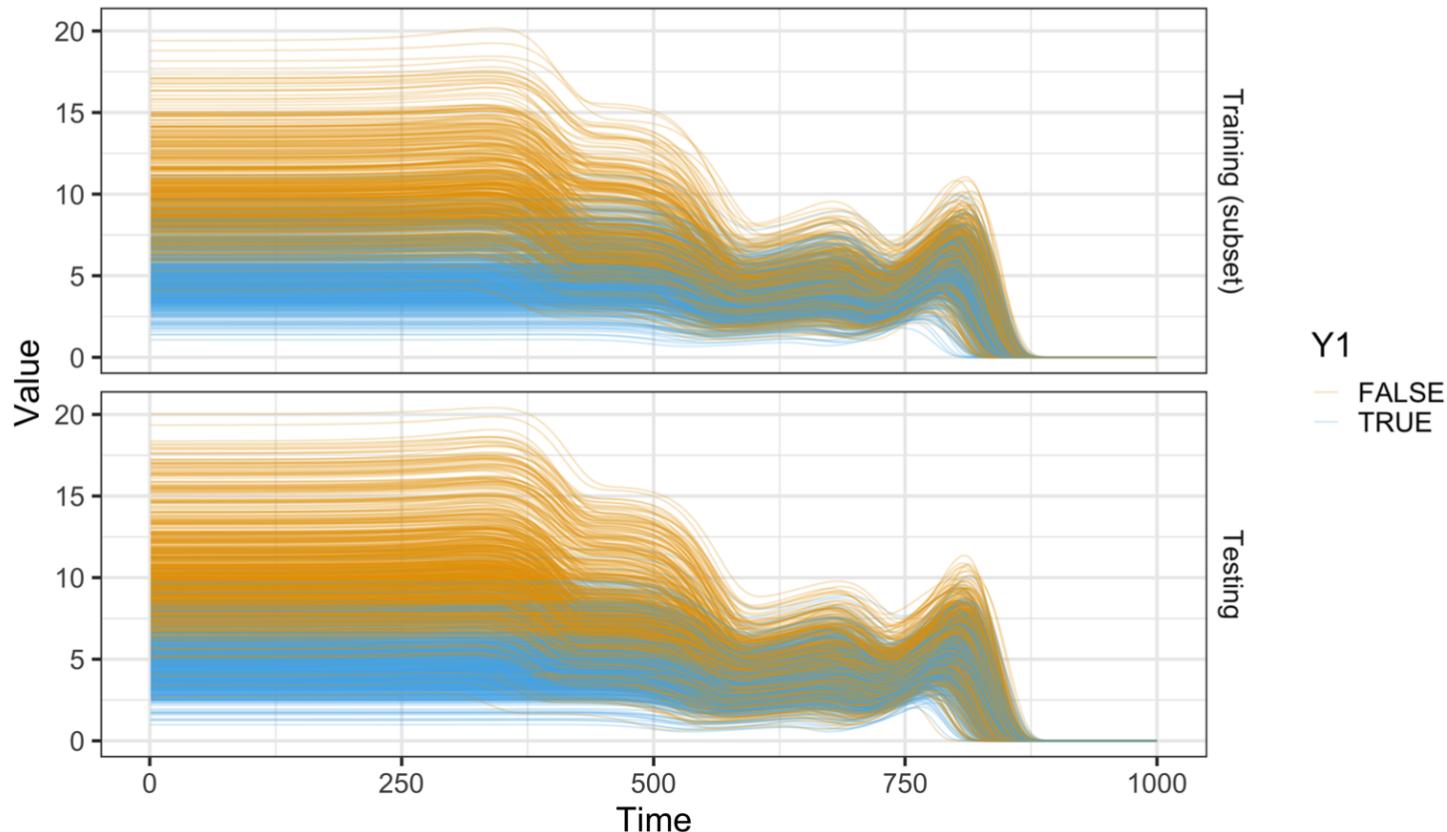
Trace plot from Urbanek (2008).

# Chapter 3

## Extensions of Neural Network Explanation Tools to Functional Data

# Application from Sandia National Labs

Plots of the training (randomly selected subset) and testing data



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

# Current Approach

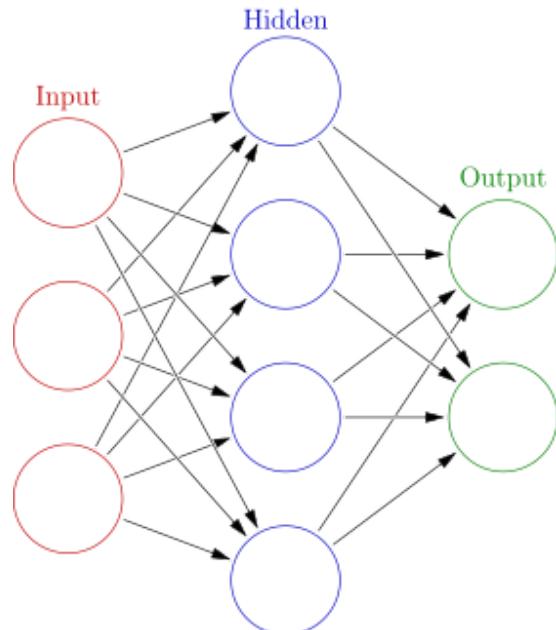
## Feature Visualization

### Concept

- Focus on a "location" in the neural network
- Determine features that maximize the activation function
- Identify example observation that triggers a part of the network

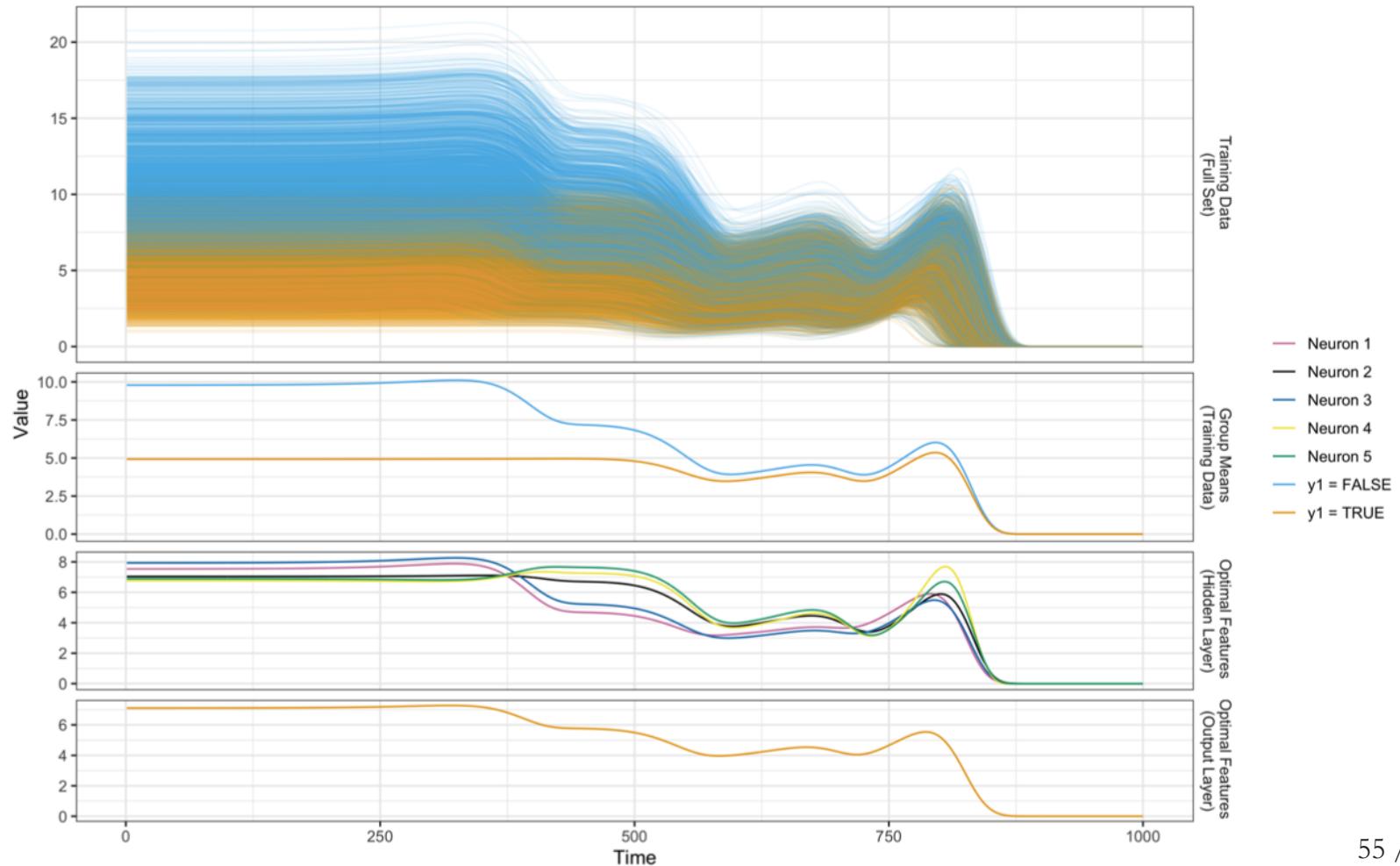
### Process

1. Fit a model
2. Fix estimated parameter values
3. Determine values that maximized activation function at desired "location"



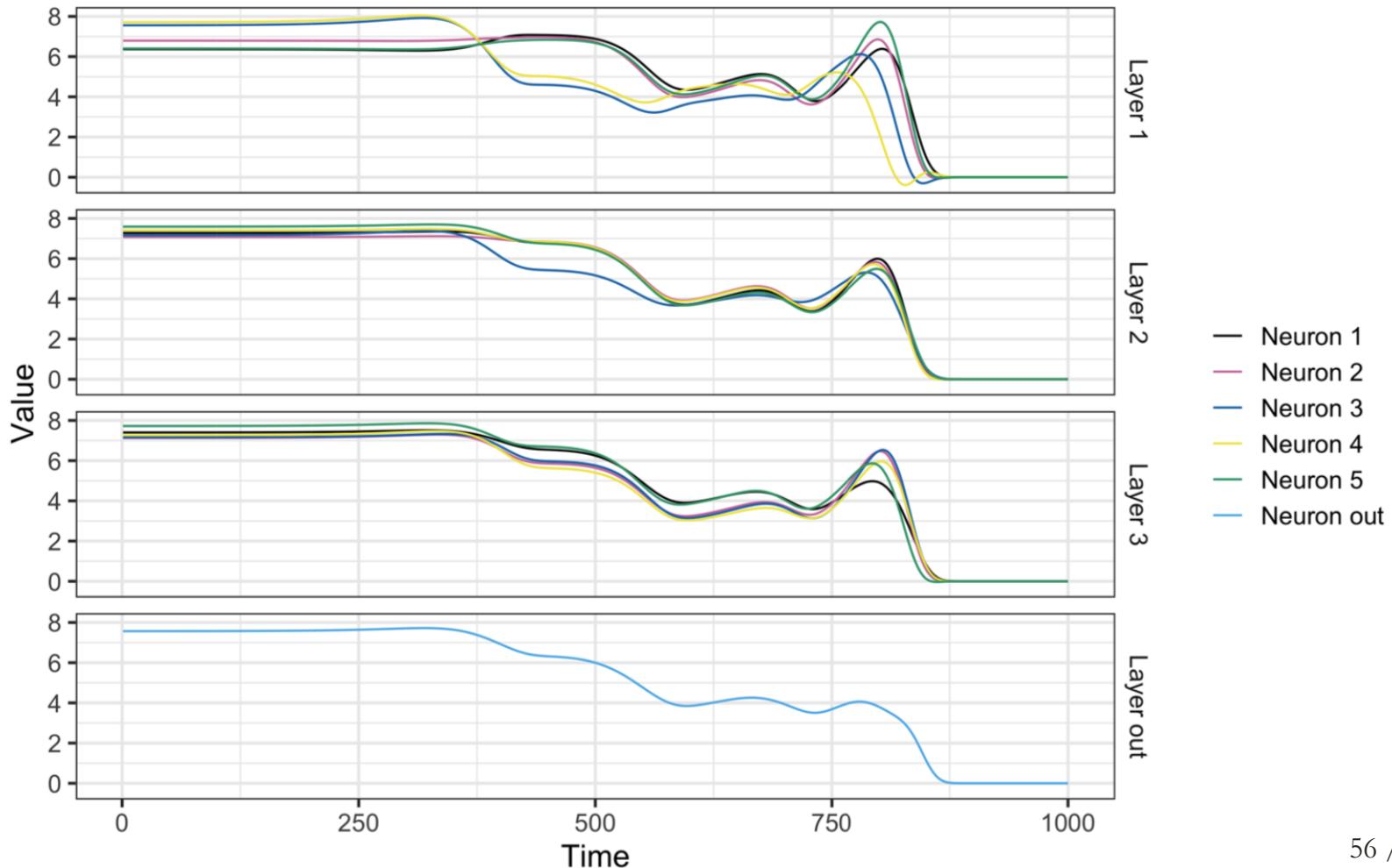
# Feature Visualization

Comparing back-transformed optimal features to the observed data



# Feature Visualization

Plots of the back transformed PCs that optimized the activation functions



# Ideas for Future Work

Feature visualization adjustments

- Visualize the functional principal components

Applications/extensions of other methods

- Permutation feature importance, saliency maps, and partial dependence plots

Visualizations of the paths of an observation through the network

- Example: flow (Halnaut, Giot, Bourqui, et al., 2020)

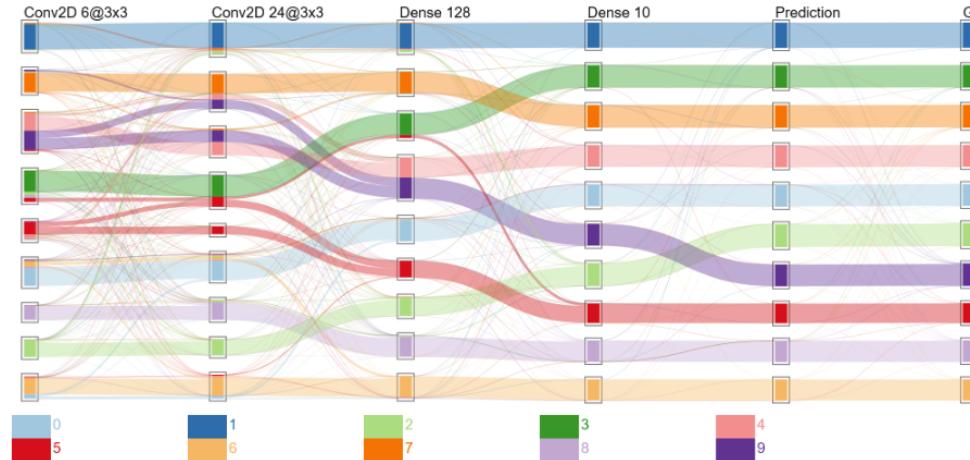


Figure 2: Visualization result on a LeNet5-inspired model evaluated on MNIST.

# Timeline and Discussion Points

# Timeline for Completion

## Summer 2020

- Chapter 1: Finish writing (possibly submit paper)
- Chapter 2: Start on content (if time)
- Chapter 3: Work on content (Sandia internship)

## Fall 2020

- Chapter 1: Submit paper (if not already done)
- Chapter 2: Work on content
- Chapter 3: Write up work

## Spring 2021

- Chapter 1: Review process (if applicable)
- Chapter 2: Write up work
- Chapter 3: Submit paper (if ready)
- Defend dissertation

# Publication of Chapter 1

How to divide up the material from chapter 1 for publication?

Current ideas:

## Paper 1: Survey paper on LIME

- Explain LIME in a statistical context
- Use visualizations to help explain the procedure
- Use diagnostic visualizations to assess LIME
- Highlight issues with LIME
- Use iris and sine data

## Paper 2: Diagnostic plots for LIME

- Motivate assessment of LIME using the bullet matching data (example of high stakes decision using machine learning)
- Demonstrate issues found with LIME explanations using diagnostic plots
- LIME should not be trusted to explain machine learning models when making high stakes decisions

# References

# References

- Altmann, A, L. Tološi, O. Sander, et al. (2010). "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10, pp. 1340-1347. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq134.
- Apley, D. W. and J. Zhu (2016). "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models".
- Beckett, C. (2018). "Rfviz: An Interactive Visualization Package for Random Forests in R". In: *DigitalCommons@USU All Graduate Plan B and otherReports*.
- Biggio, B. and F. Roli (2018). "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning". In: *Pattern Recognition* 84, pp. 317-331. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2018.07.023.
- Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5-32. ISSN: 0885-6125. DOI: 10.1023/a:1010933404324.
- Casalicchio, G, C. Molnar, and B. Bischl (2019). "Visualizing the Feature Importance for Black Box Models", pp. 655-670. ISSN: 2190-5053. DOI: 10.1007/978-3-030-10925-7\_40.
- Craven, M. W. and J. W. Shavlik (1996). "Extracting Tree-Structured Representations of Trained Networks". In: *Advances in Neural Information Processing Systems* 8.
- Fisher, A, C. Rudin, and F. Dominici (2018). "Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective".
- Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5. ISSN: 0090-5364. DOI: 10.1214/aos/1013203451.
- Friedman, J. H. and B. E. Popescu (2008). "Predictive learning via rule ensembles". In: *The Annals of Applied Statistics* 2.3, pp. 916-954. ISSN: 1932-6157. DOI: 10.1214/07-aos148.
- Gilpin, L. H, D. Bau, B. Z. Yuan, et al. (2018). "Explaining Explanations: An Overview of Interpretability of Machine Learning".
- Goldstein, A, A. Kapelner, J. Bleich, et al. (2013). "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation".
- Goodfellow, I. J, J. Shlens, and C. Szegedy (2014). "Explaining and Harnessing Adversarial Examples". In: *arXiv*.
- Goodman, B. and S. Flaxman (2016). "European Union regulations on algorithmic decision-making and a "right to explanation"".
- Greenwell, B. M, B. C. Boehmke, and A. J. McCarthy (2018). "A Simple and Effective Model-Based Variable Importance Measure".
- Guidotti, R, A. Monreale, S. Ruggieri, et al. (2018). "A Survey Of Methods For Explaining Black Box Models".
- Halnaut, A, R. Giot, R. Bourqui, et al. (2020). "Deep Dive into Deep Neural Networks with Flows", pp. 231-239. DOI: 10.5220/0008989702310239.
- Hara, S. and K. Hayashi (2016a). "Making Tree Ensembles Interpretable".
- Hare, E, H. Hofmann, and A. Carriquiry (2016). "Automatic Matching of Bullet Lands". In: *Annals of Applied Statistics*. DOI: <http://adsabs.harvard.edu/abs/2016arXiv160105788H>.
- Hooker, G. (2004). "Discovering additive structure in black box functions", p. 575. DOI: 10.1145/1014052.1014122.
- Kindermans, P, S. Hooker, J. Adebayo, et al. (2017). "The (Un)reliability of saliency methods".
- Koh, P. W. and P. Liang (2017). "Understanding Black-box Predictions via Influence Functions". In: *arXiv*.

# References

- Krause, J. A. Perer, and K. Ng (2016). "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models", pp. 5686-5697. DOI: 10.1145/2858036.2858529.
- Laugel, T. M. Lesot, C. Marsala, et al. (2017). "Inverse Classification for Comparison-based Interpretability in Machine Learning". In: *arXiv*.
- Laugel, T. X. Renard, M. Lesot, et al. (2018). "Defining Locality for Surrogates in Post-hoc Interpretability". In: *CoRR* abs/1806.07498. URL: <http://arxiv.org/abs/1806.07498>.
- Li, M. Z. Zhao, and C. Scheidegger (2020). "Visualizing Neural Networks with the Grand Tour". In: *Distill*. <https://distill.pub/2020/grand-tour>. DOI: 10.23915/distill.00025.
- Lipton, Z. C. (2016). "The Mythos of Model Interpretability".
- Looveren, A. V. and J. Klaise (2019). "Interpretable Counterfactual Explanations Guided by Prototypes". In: *arXiv*.
- Lundberg, S. and S. Lee (2017). "A Unified Approach to Interpreting Model Predictions".
- Martens, D. and F. Provost (2014). "Explaining Data-Driven Document Classifications". In: *MIS Quarterly* 38.1, pp. 73-99. ISSN: 0276-7783. DOI: 10.25300/misq/2014/38.1.04.
- Ming, Y. (2017). "A Survey on Visualization for Explainable Classifiers".
- Mohseni, S. N. Zarei, and E. D. Ragan (2018). "A Survey of Evaluation Methods and Measures for Interpretable Machine Learning".
- Molnar, C. (2019). *Interpretable Machine Learning*.
- Montavon, G. W. Samek, and K. Müller (2017). "Methods for Interpreting and Understanding Deep Neural Networks".
- Murdoch, W. J. C. Singh, K. Kumbier, et al. (2019). "Interpretable machine learning: definitions, methods, and applications". In: *arXiv*. DOI: 10.1073/pnas.1900654116.
- Olah, C. A. Mordvintsev, and L. Schubert (2017). "Feature Visualization". In: *Distill*. <https://distill.pub/2017/feature-visualization>. DOI: 10.23915/distill.00007.
- Pedersen, T. L. and M. Benesty (2020). *lime: Local Interpretable Model-Agnostic Explanations*. <https://lime.data-imaginist.com>, <https://github.com/thomasp85/lime>.
- Ribeiro, M. T. S. Singh, and C. Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135-1144.
- Ribeiro, M. T. S. Singh, and C. Guestrin (2018). "Anchors: High-Precision Model-Agnostic Explanations".
- Rudin, C. (2018). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead".
- Silva, N. da, D. Cook, and E. Lee (2017). "Interactive Graphics for Visually Diagnosing Forest Classifiers in R".
- Simonyan, K. A. Vedaldi, and A. Zisserman (2013). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps".
- Staniak, M. and P. Biecek (2018). "Explanations of model predictions with live and breakDown packages". In: *arXiv* 10.2, p. 395. DOI: 10.32614/rj-2018-072.
- Su, J. D. V. Vargas, and K. Sakurai (2019). "One Pixel Attack for Fooling Deep Neural Networks". In: *IEEE Transactions on Evolutionary Computation* 23.5, pp. 828-841. ISSN: 1089-778X. DOI: 10.1109/tevc.2019.2890858.

# References

- Szegedy, C, W. Zaremba, I. Sutskever, et al. (2013). "Intriguing properties of neural networks".
- Urbanek, S. (2008). "Visualizing Trees and Forests". In: *Handbook of data visualization*. Ed. by A. U. Chun-houh Chen Wolfgang Härdle , pp. 243-266. ISBN: 9783540330363.
- Wachter, S, B. Mittelstadt, and C. Russell (2017). "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR".
- Welling, S, H. H. F. Refsgaard, P. B. Brockhoff, et al. (2016). "Forest Floor Visualizations of Random Forests".
- Wickham, H, D. Cook, and H. Hofmann (2015). "Visualizing statistical models: Removing the blindfold". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8.4, pp. 203-225. ISSN: 1932-1872. DOI: 10.1002/sam.11271.

# Appendix

# Additional Bullet Application Plots

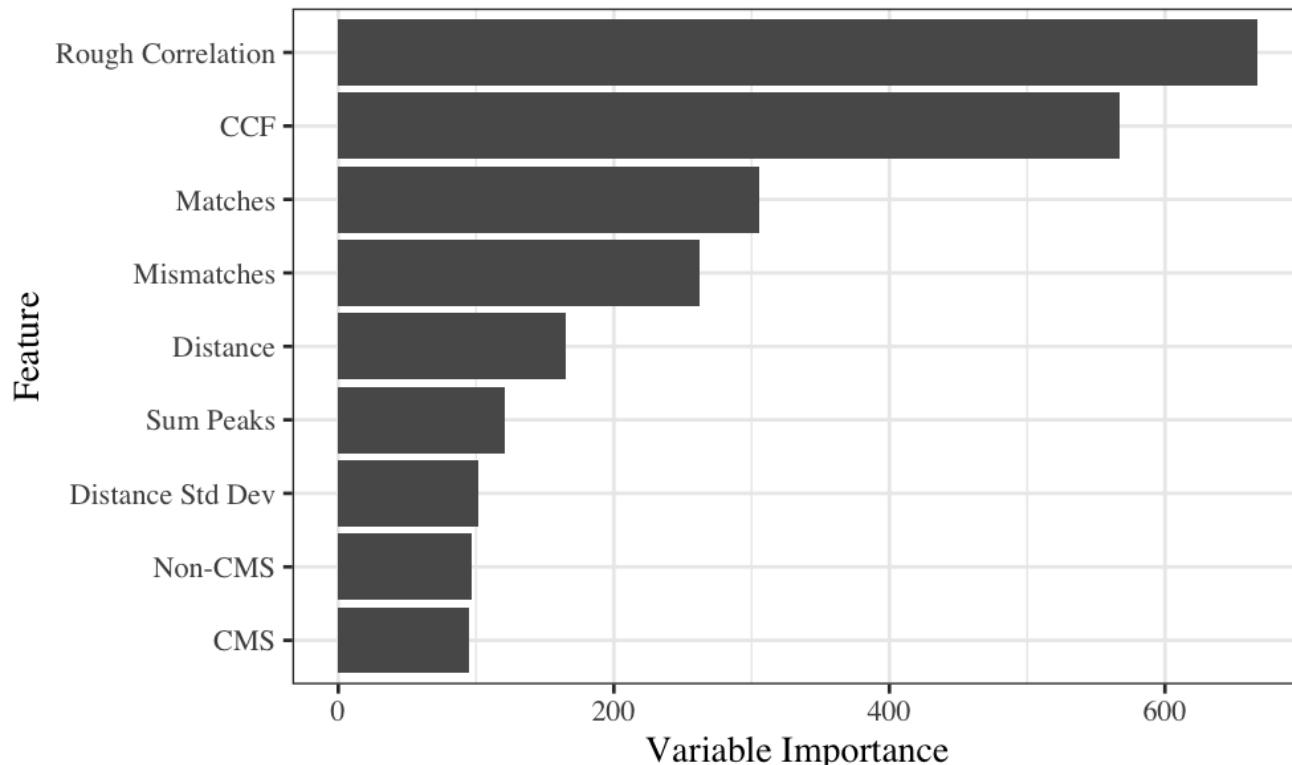
Distribution of the response variable for the random forest features

# Additional Bullet Application Plots

Bivariate visualizations of the random forest training data

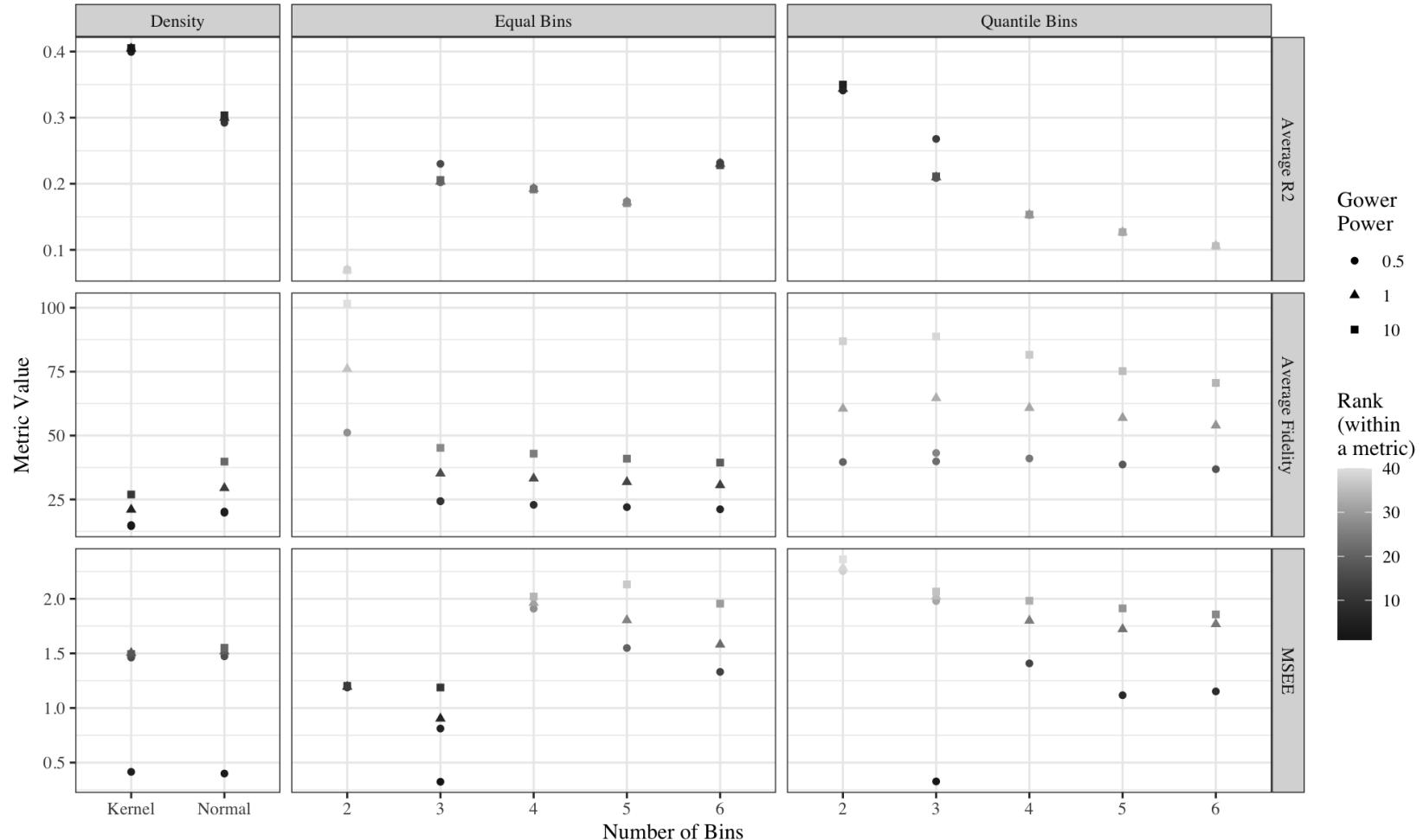
# Additional Bullet Application Plots

Random forest variable importance plot



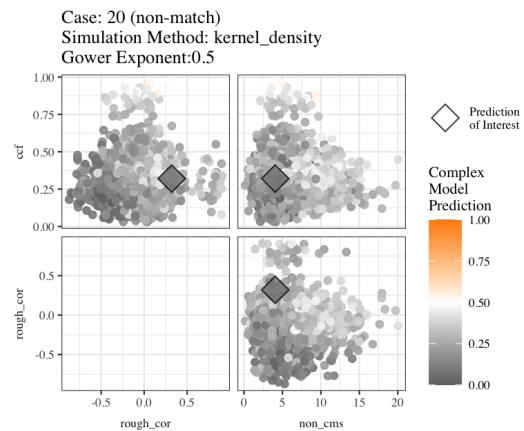
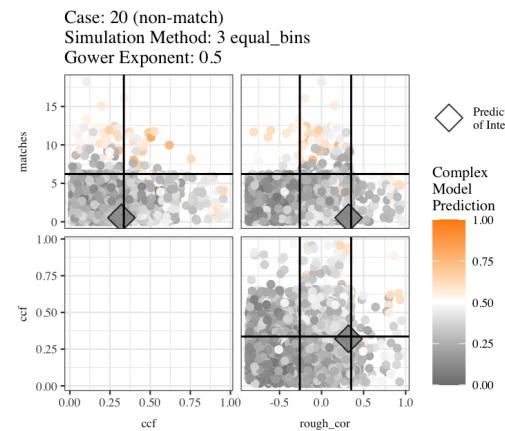
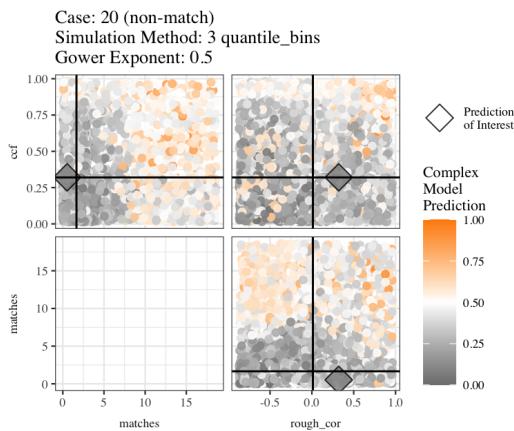
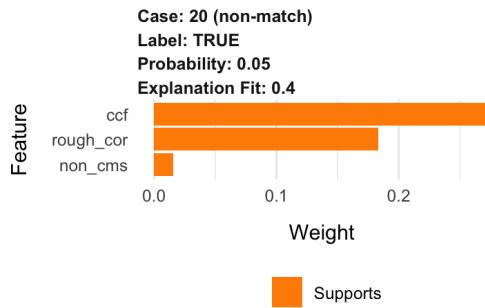
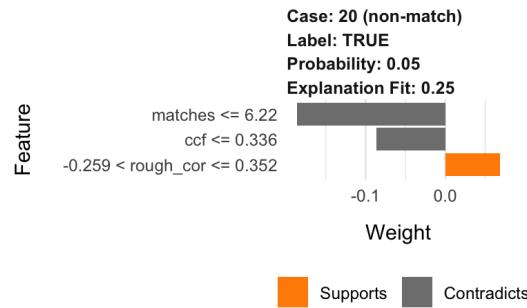
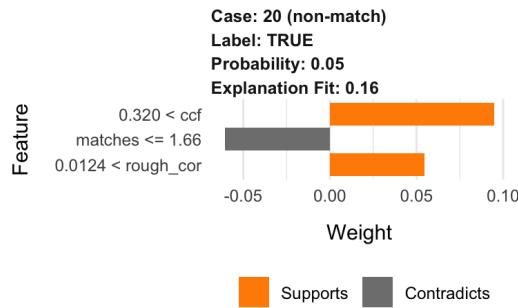
# Additional Bullet Application Plots

Assessment metric plot



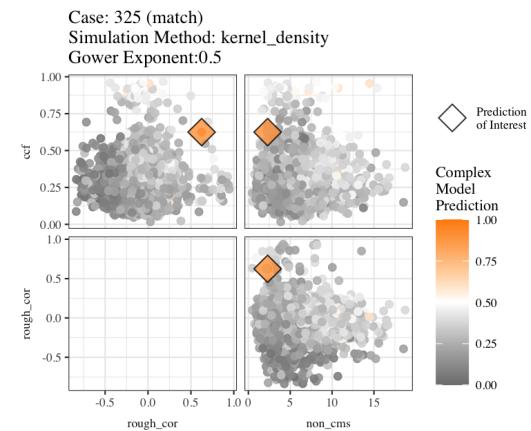
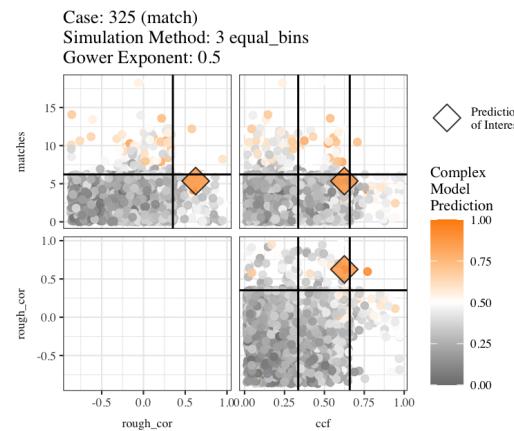
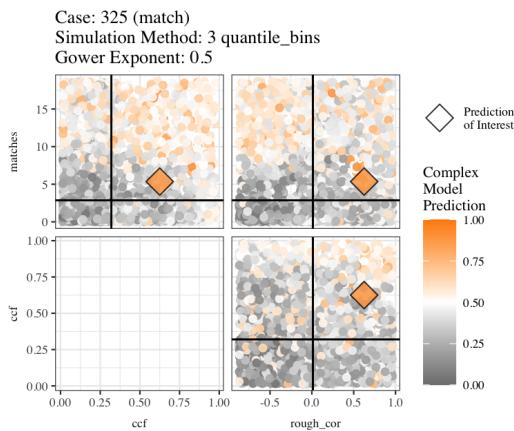
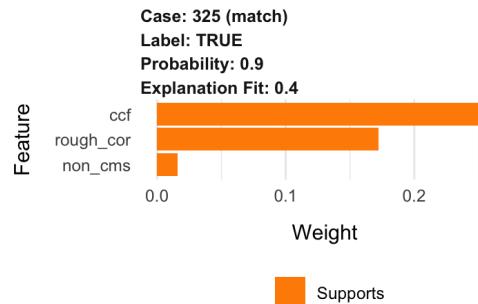
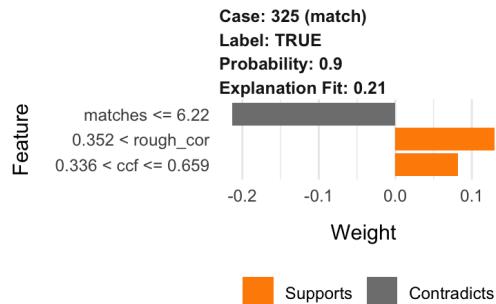
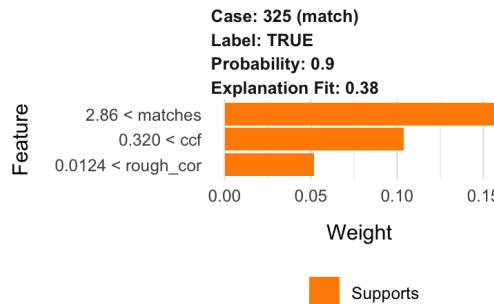
# Additional Bullet Application Plots

Plots of LIME explanations for one case in the test data that is a known non-match



# Additional Bullet Application Plots

Plots of LIME explanations for one case in the test data that is a known match



# Feature Visualization Example Optimization Example

- Response:  $y \in \{0, 1\}$
- Features:

$$PC = (PC_1, PC_2, PC_3, PC_4, PC_5)$$

- Model: 1 hidden layer, 5 neurons
- Neuron  $i$  coefficients ( $i = 1, \dots, 5$ ):

$$(\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \beta_{3,i}, \beta_{4,i}, \beta_{5,i})$$

- Activation Function: logistic

$$\sigma(v) = \frac{1}{1 + e^{-v}}$$

- Feature visualization optimization:

$$\arg \max_{PC} \sigma \left( \hat{\beta}_{0,i} + \hat{\beta}_{1,i} PC_1 + \dots + \hat{\beta}_{5,i} PC_5 \right)$$