Studies on Visualization Methods for Explainable Machine Learning

Statistics PhD Oral Prelim

Katherine Goode
Iowa State University
May 6, 2020

Questions/Notes

Questions:

- How much text is appropriate on a slide during an oral prelim?
- Complete sentences or not for the oral prelim?
- Are the references formatted okay?
- How much time should I spend discussing previous explainable machine learning methods/how many should I mention?
- Title?
- Should I say I'm focusing on "explainable machine learning" or "visualization techniques for explainable machine learning"

Ideas:

- Could turn explainable machine learning methods into a visualization or table to make it more digestible

Notes:

- Remember to send out before prelim
- 50 minute talk

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

Assistantships/Internship

- Lecturer for STAT 101
  - Spring 2016
- AES Statistical Consultant
  - Summer 2016 - current
- NREM Research Assistant
  - Summer 2019

Internship

Sandia National Labs

- Statistical Sciences Research and Development Intern
- December 2019 - current

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

- Many machine learning models are considered "black-boxes", because it is not possible to directly interpret them.
- An area of research in explainable machine learning has developed as a result (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).
- The goal with explainable machine learning is to provide methods to explain predictions made by black-box models.
- The European General Data Protection Regulation (GDPR) implemented in 2018 includes a "right to explanation", which is pertinent to the way in which machine learning models are used in Europe (Goodman and Flaxman, 2016).
    - Goodman and Flaxman (2016) point out, "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

There are no accepted definitions for *explainability* and *interpretability* in the literature (Gilpin, Bau, Yuan, et al., 2018; Lipton, 2016; Molnar, 2019; Montavon, Samek, and Müller, 2017; Murdoch, Singh, Kumbier, Abbasi-Asl, and Yu, 2019; Rudin, 2018). I will use the following definitions.

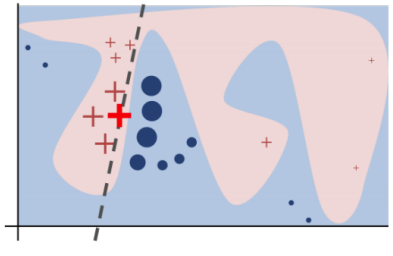
Interpretability is the ability to directly use the parameters of a model to understand the mechanism of how the model makes predictions.

- A linear model coefficient indicates the amount the response variable changes based on a change in the predictor variable when all other predictor variables are held constant.

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

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

- LIME uses a surrogate model to understand the relationship between the complex model predictions and predictor variables in a local region (Ribeiro, Singh, and Guestrin, 2016).



Model Agnostic Methods

General Model Visualizations

- Removing the blindfold (Wickham, Cook, and Hofmann, 2015)
  - Plot the model in dataspace
  - Plot members of a model collection
  - Explore the fitting process

Global Methods

- 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)
- Parallel coordinate plots (Jiang, Liu, and Chen, 2019)
- Global feature importance plots (Fisher, Rudin, and Dominici, 2018; Altmann, Toloşi, Sander, and Lengauer, 2010; Casalicchio, Molnar, and Bischl, 2019)
- Global surrogate models (Molnar, 2019)

Local Methods

- 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

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

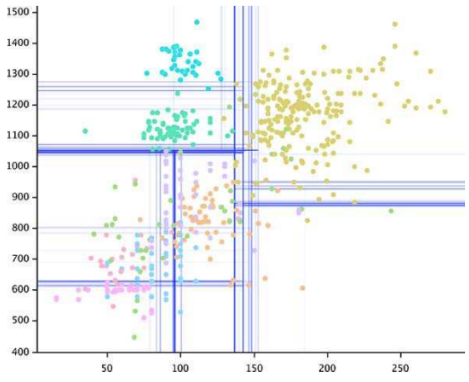
Model Specific Methods

Random Forests

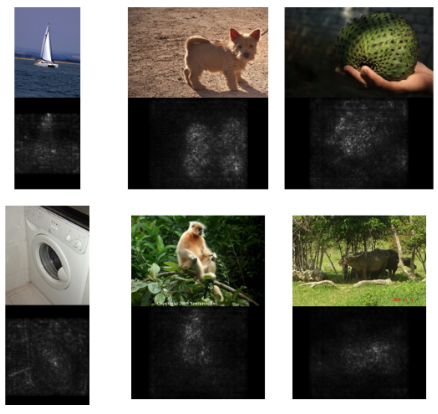
- Random forest impurity based feature importance (Breiman, 2001)
- Sectioned scatterplots (Urbanek, 2008)
- Trace plots of trees (Urbanek, 2008)
- Simplified model (Hara and Hayashi, 2016a)
- Forest floor visualizations (Welling, Refsgaard, Brockhoff, and Clemmensen, 2016)
- Interactive visualizations (Beckett, 2018; da Silva, Cook, and Lee, 2017)

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)



Sectioned scatterplot from Urbanek (2008).



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

Assessment of Explainable Machine Learning Methods

General Assessment

- Rudin (2018) argues against the use of black-box models to make high stakes decisions since all "explanations must be wrong". Rudin (2018) also makes the following claims:
    - The trade-off between accuracy and interpretability is a myth
    - Interpretable models can have just as good accuracy as "black-box" models
    - Explanations may not be faithful to the original model, make sense, or be detailed enough to understand the "black-box" model

Method Specific Assessments

- Laugel, Renard, Lesot, Marsala, and Detyniecki (2018) consider how to appropriately choose a local region for the local surrogate model with the method of LIME.
- Laugel, Renard, Lesot, et al. (2018) consider issues with unjustified counterfactual examples.
- Kindermans, Hooker, Adebayo, Alber, Schütt, Dähne, Erhan, and Kim (2017) show how saliency maps can change when a transformation is applied to the input data that has no effect on the model.

Overview of Dissertation Chapters

Chapter 1

- Focus on the model-agnostic explainer model method of LIME
- Discuss the importance of assessing LIME explanations
- Suggest the use of visualizations for the assessment of LIME and provide many example visualizations

Chapter 2

- Visualizations for the explainability of random forest models
- Use clustering to identify key tree structures within the random forest

Chapter 3

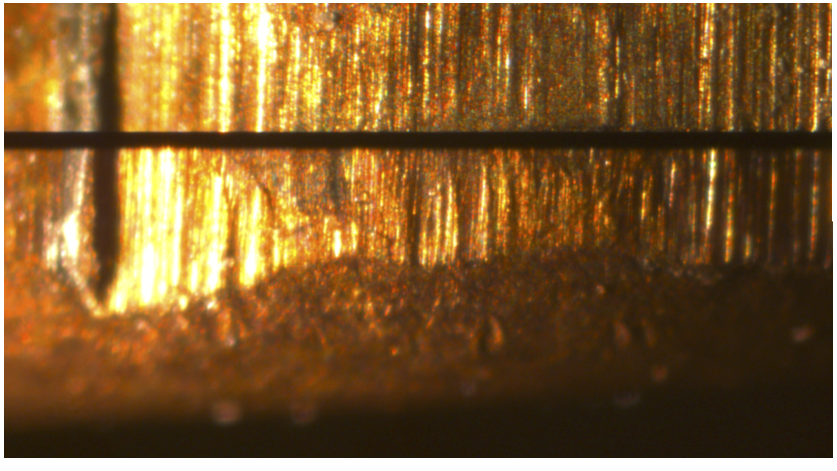
- Visualizations for the explainability of neural networks with functional data

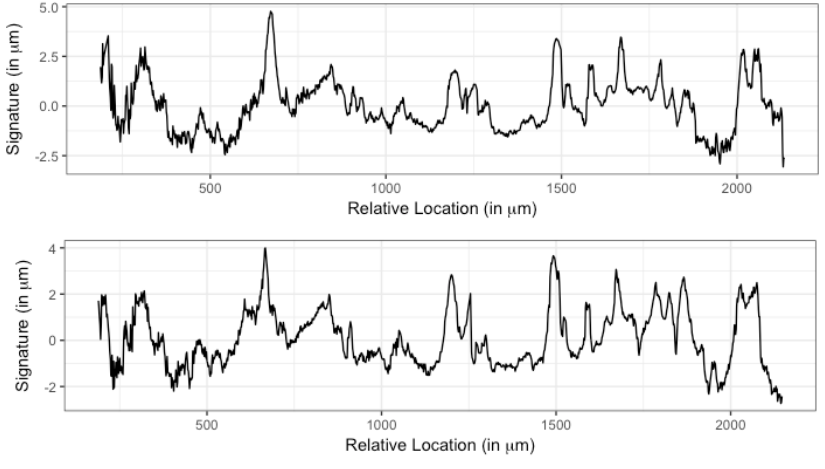
Chapter 1: Visual Diagnostics of a Model Explainer -- Tools for the Assessment of LIME Explanations

Motivation

Hare, Hofmann, and Carriquiry (2016):

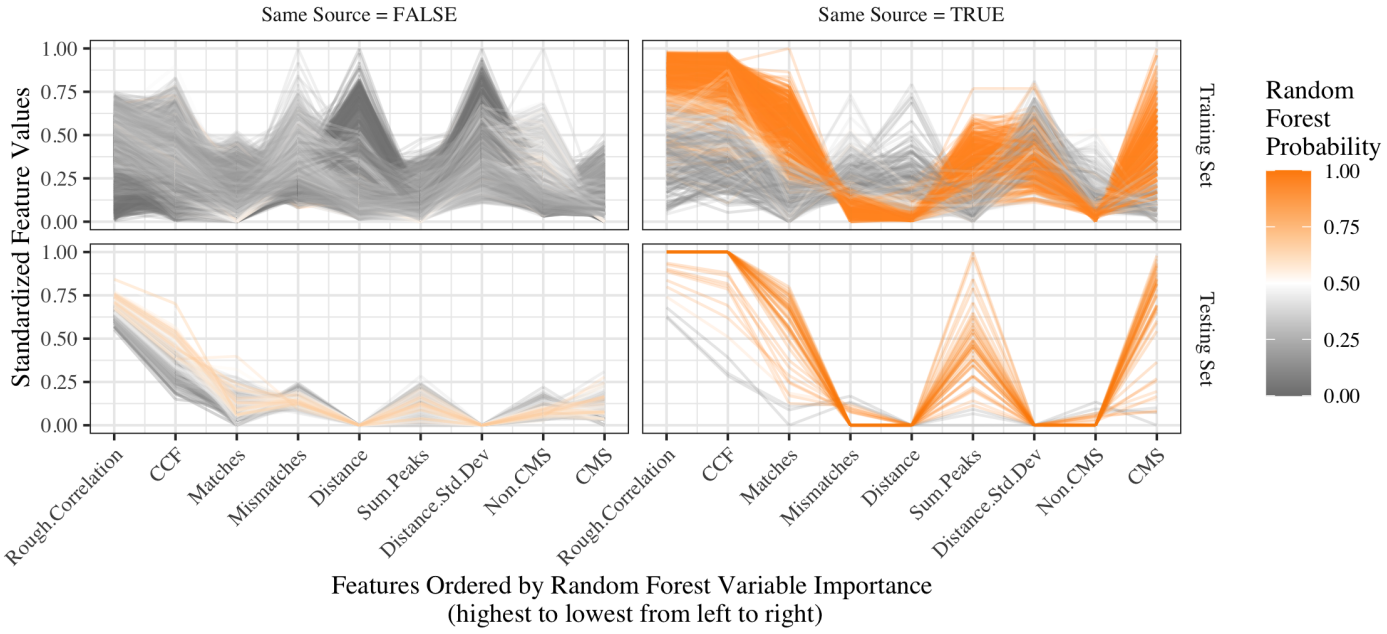
- Interested in providing quantitative evidence to determine whether two bullets were fired from the same gun
- Use high definition scans of striations on bullet lands to extract "signatures"
- Compute similarity features to compare two signatures





Motivation

Hare, Hofmann, and Carriquiry (2016) fit a random forest model with nine signature similarity features.



	FALSE	TRUE	Classification Error
FALSE	81799	21	0.0002567
TRUE	363	845	0.3004967

Motivation

We were interested in providing an explanation for specific signatures comparisons, so we tried applying LIME. However, we found unreasonable results, which led to an assessment of LIME.

Conceptual Depiction of LIME

LIME was introduced by Ribeiro, Singh, and Guestrin (2016).

- **L**ocal
- **I**nterpretable
- **M**odel-agnostic
- **E**xplanation

Concept: Explain the prediction made by a complex model for *one prediction of interest* using an interpretable model focused on a local region near the prediction of interest

Importance of Assessing LIME

LIME adds an additional layer of complexity by using a model that also needs to be assessed, which raises various questions:

- Does the explainer model approximate the complex model well in a local region around the prediction of interest?
- How to determine an appropriate local region?
- Is the relationship linear in the specified local region?

Visualizations for Model Assessment

Ribeiro, Singh, and Guestrin (2016) make the following set of claims regarding the performance of LIME:

- *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 the use of visualizations to assess the claims made by LIME, which we have organized into three categories:

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

Sine Example Data

Training data: 500 observations

Testing data: 100 observations

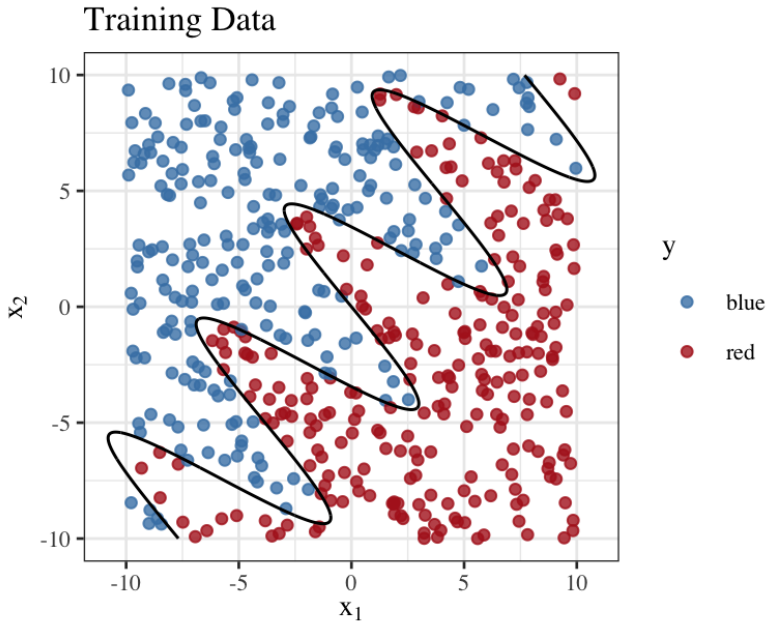
Black-box model: random forest

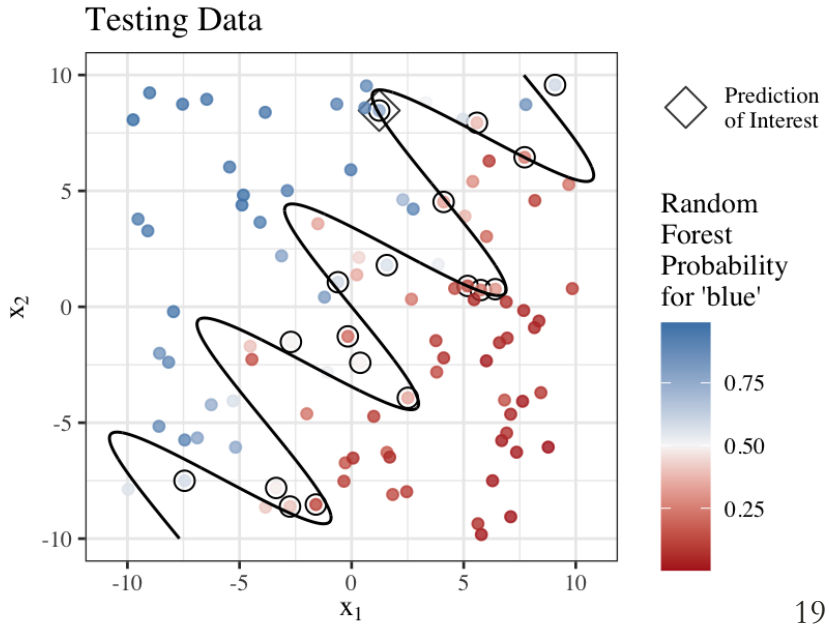
$$x_1 \sim Unif(\text{ -10, 10 })$$
$$x_2 \sim Unif(\text{ -10, 10 })$$
$$x_3 \sim \text{N}(0, 1)$$

$$y = \begin{cases} blue & \text{if } x_2' > \sin(x_1') \\ red & \text{if } x_2' \leq \sin(x_1') \end{cases}.$$

where

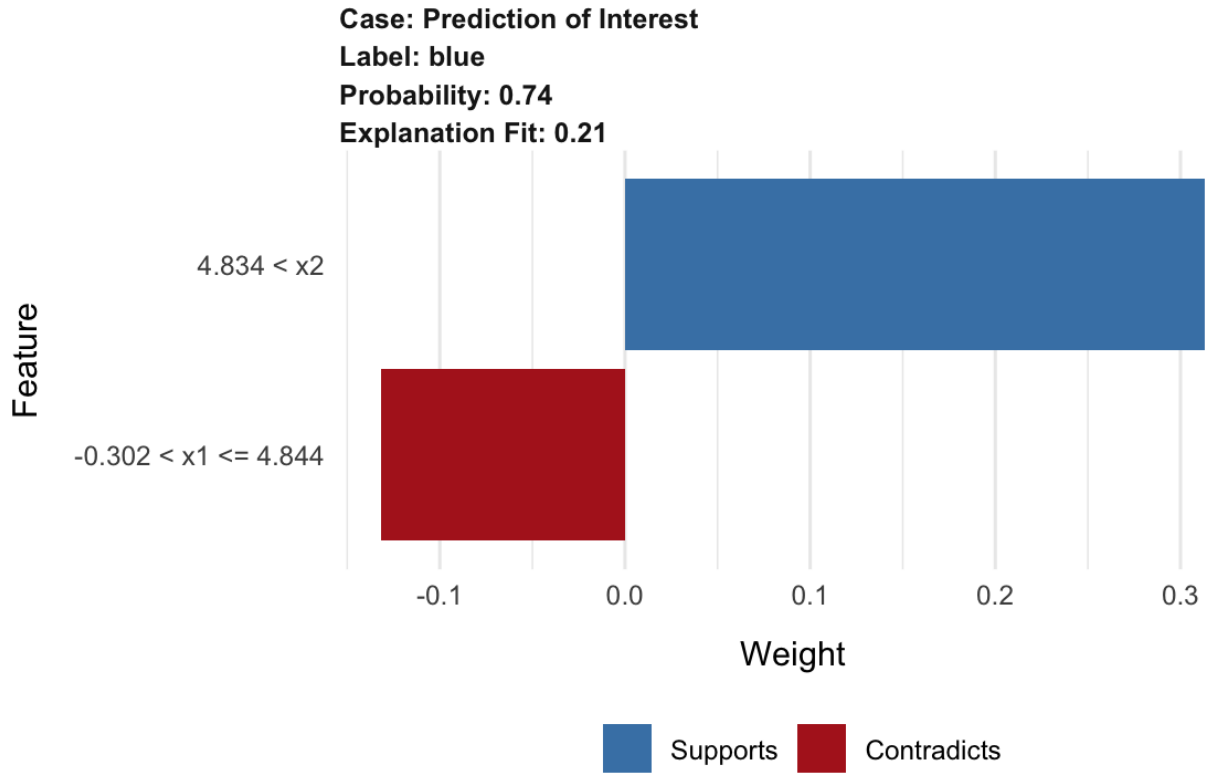
$$x_1' = x_1 \cos(\theta) - x_2 \sin(\theta)$$
$$x_2' = x_1 \sin(\theta) + x_2 \cos(\theta)$$
$$\theta = -0.9$$





Diagnostics for Individual Explanations

Overview visualization of the LIME explanation for the prediction of interest from the *lime* R package



Diagnostics for Individual Explanations

Step 1a: Data Simulation

lime R package default procedure:

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

Corresponding Diagnostic:
Training Data Plot

- Scatterplot of the two features selected by *lime*
- Points are the training data observations colored by the observed response
- Grid of lines represents the boundaries of the 4 quantile bins

Diagnostics for Individual Explanations

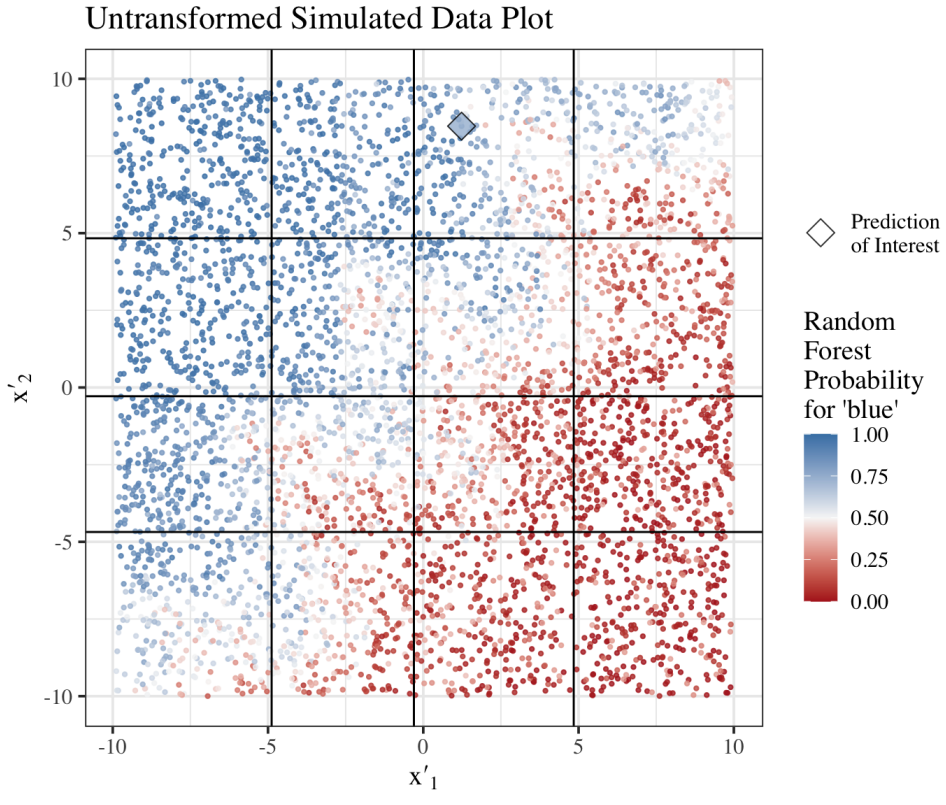
Step 1b: Complex Model Predictions

- Apply complex model to simulated data to obtain predictions

Corresponding Diagnostic:

Untransformed Simulated Data Plot

- Scatterplot of the simulated data features denoted by x'_1 and x'_2
- Points are colored by the random forest predictions (for 'blue')
- Grid of lines represents the boundaries of the 4 quantile bins
- Prediction of interest location indicated by the diamond



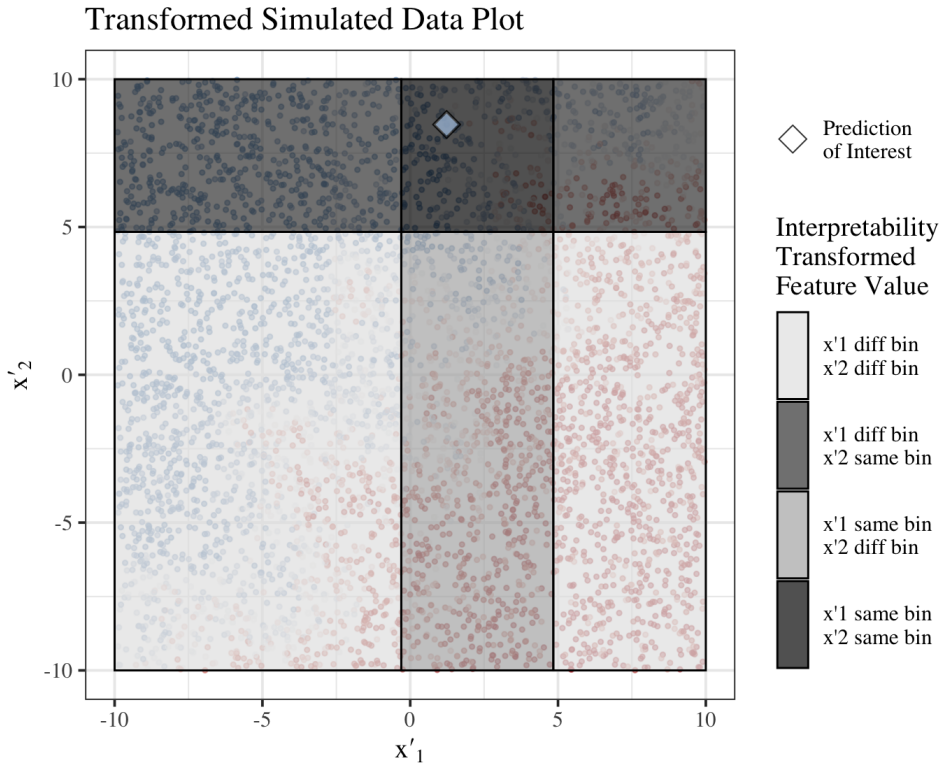
Diagnostics for Individual Explanations

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:
Interpretability Transformed Simulated Data Plot

- Scatterplot of the simulated data features
- Rectangular region shades represent the interpretability transformed feature value
- Prediction of interest location indicated by the diamond



Diagnostics for Individual Explanations

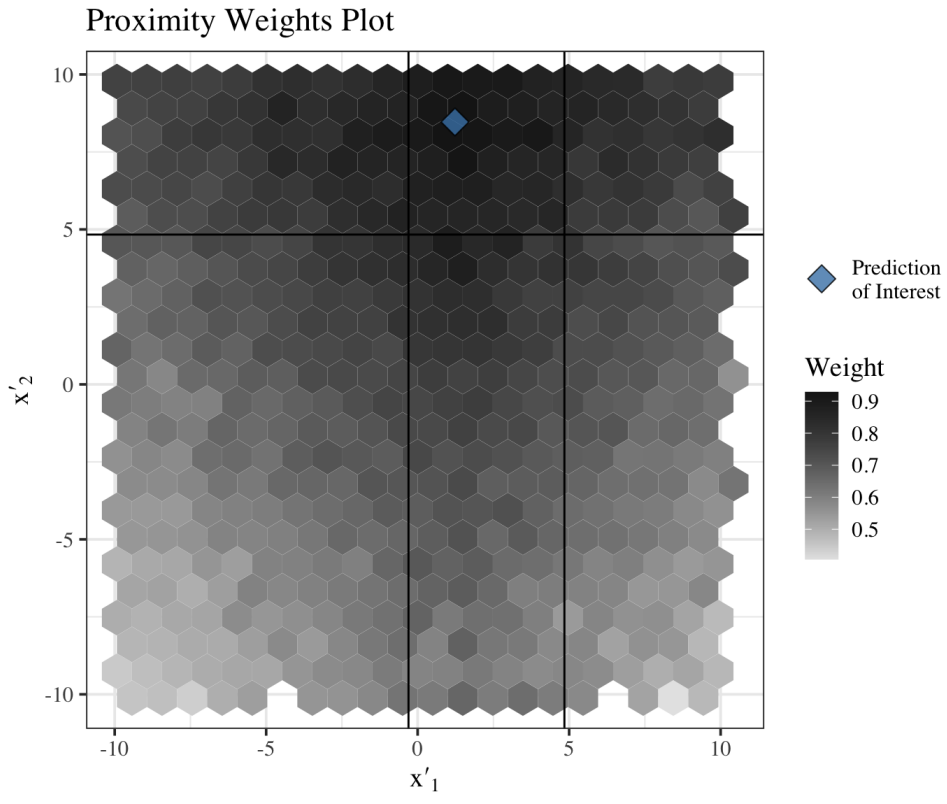
Step 2a: Assign Weights

- Assign weights to simulated data based on proximity to the prediction of interest (using the untransformed feature values)
- *lime* R package uses Gower distance metric by default

Corresponding Diagnostic:

Proximity Weights Plot

- Each hexagon contains the average of the weights of the observations located within the hexagon
- Prediction of interest location indicated by the diamond
- Solid lines are the interpretability transformation boundaries



Diagnostics for Individual Explanations

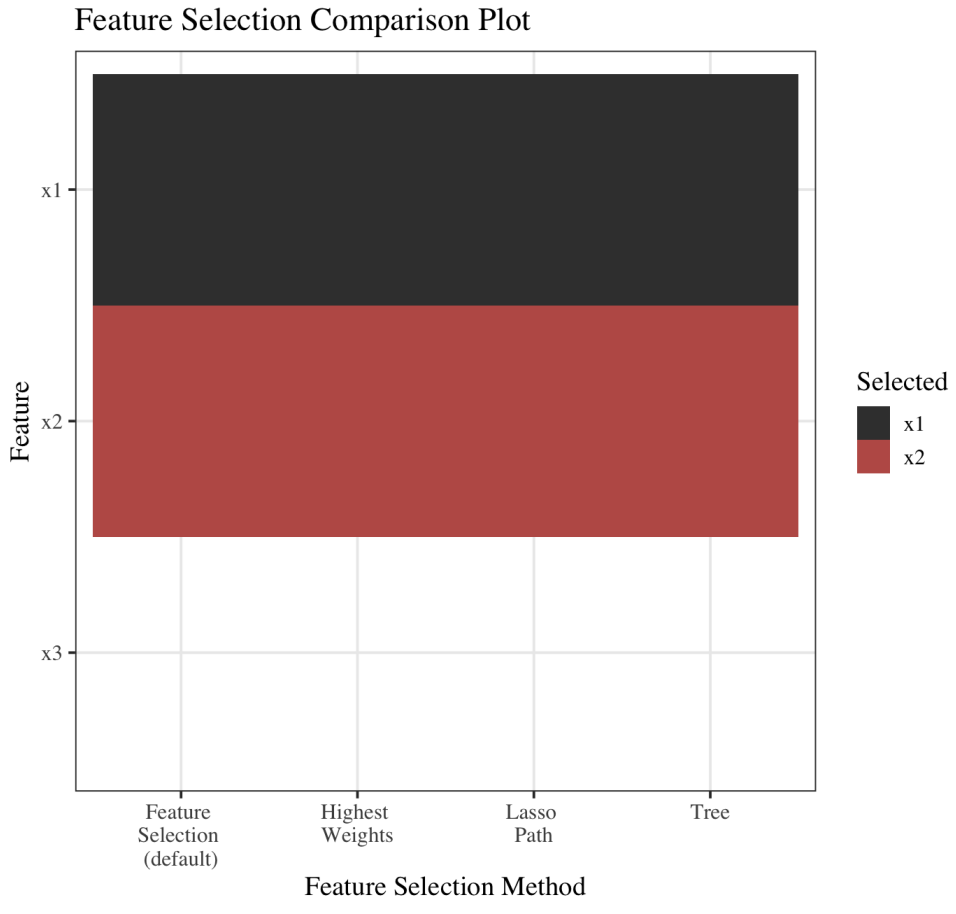
Step 2b: Feature Selection

- Ridge regression model
  - response: complex model predictions from the simulated data
  - features: interpretability transformed simulated data features
- *lime* R package uses forward selection by default for less than 6 features specified (user specifies the number of features)

Corresponding Diagnostic:

Feature Selection Comparison Plot

- Tile plot showing feature selected by various methods in *lime*



Diagnostics for Individual Explanations

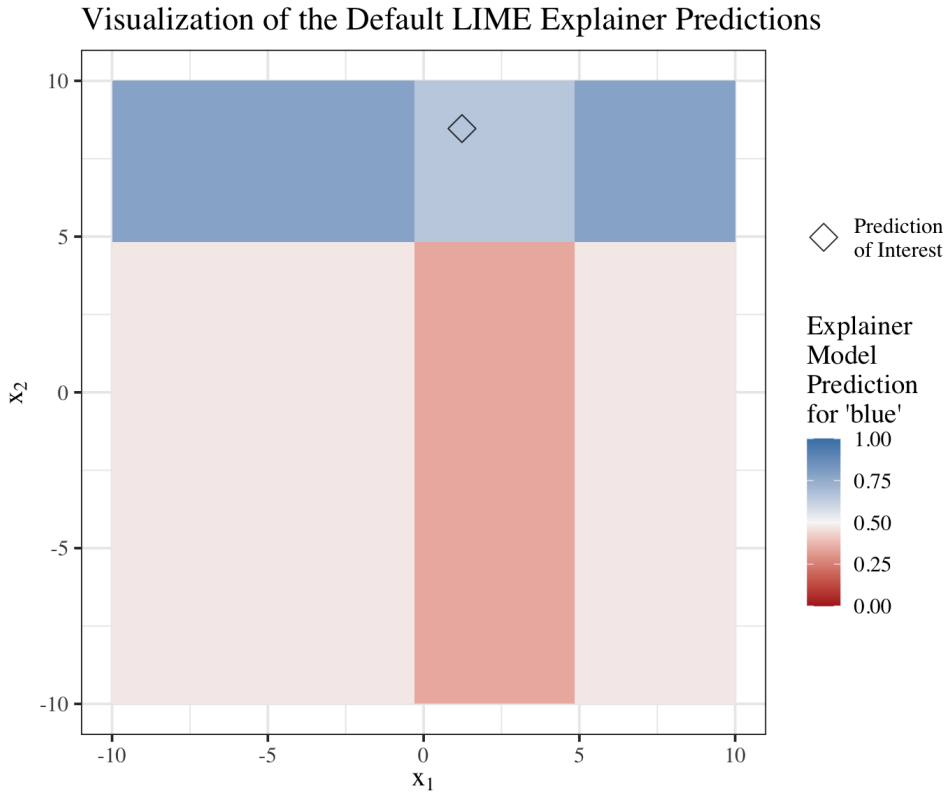
Step 2c: Fit Explainer Model

- Ridge regression model
  - response: complex model predictions from the simulated data
  - features: interpretability transformed simulated data features selected during feature selection

Corresponding Diagnostic:

Explainer Model Prediction Plot

- Color of a rectangular region represents the explainer model prediction for the region



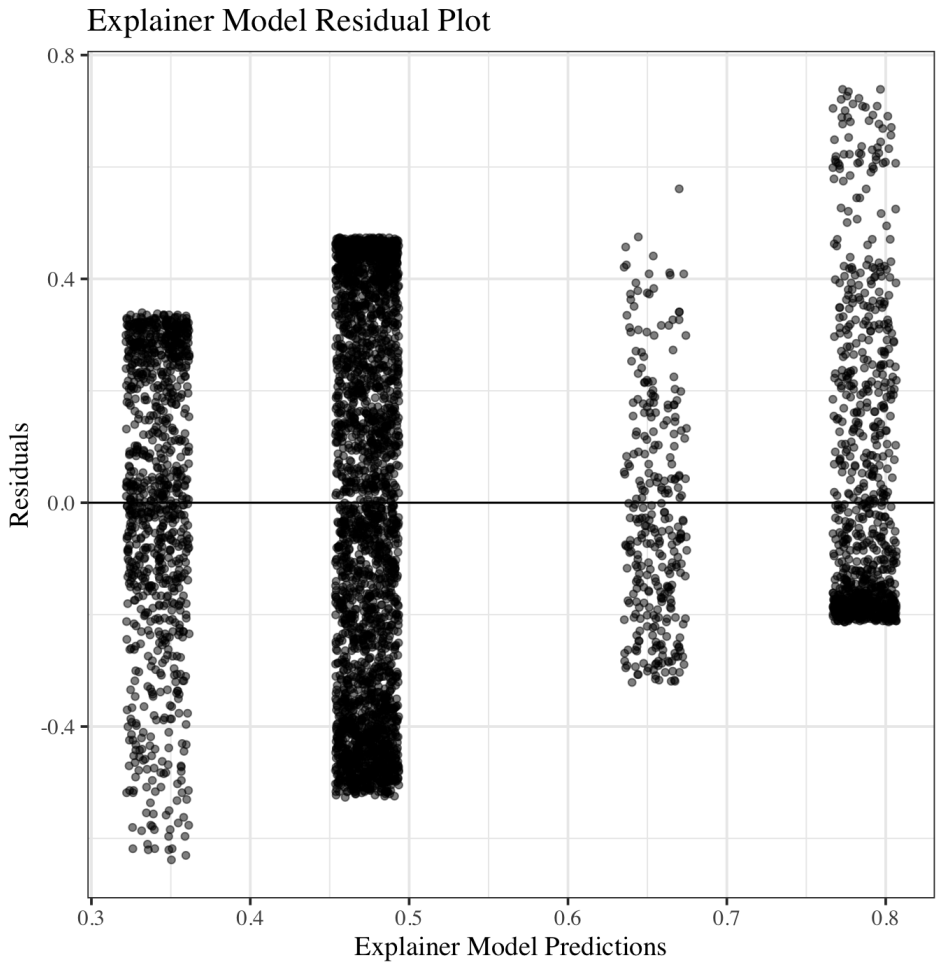
Diagnostics for Individual Explanations

Step 2c: Fit Explainer Model (continued)

Corresponding Diagnostic:

Explainer Model Residual Plot

- Residual plot for the explainer model
- Some jitter has been added to the points in the x-direction



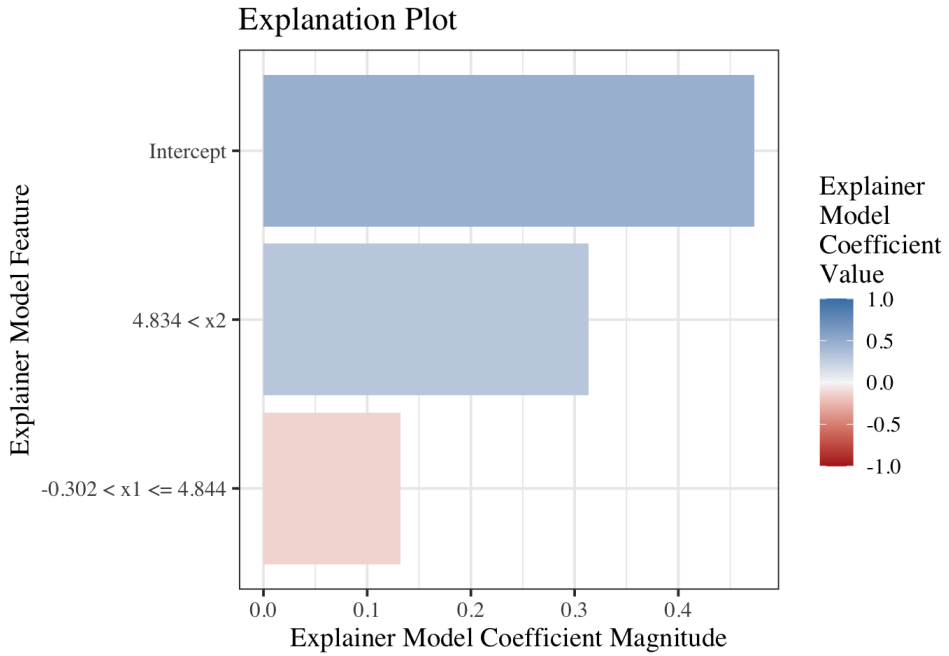
Diagnostics for Individual Explanations

Step 3a: Explainer Model Interpretation

- Interpret the explainer model to explain the complex model prediction

Corresponding Diagnostic: *Explanation Plot*

- Adaptation of the explanation plot from *lime* R package
- Length of bars represents the magnitude of explainer model coefficients
- Color of bars represents the value of the explainer model coefficients



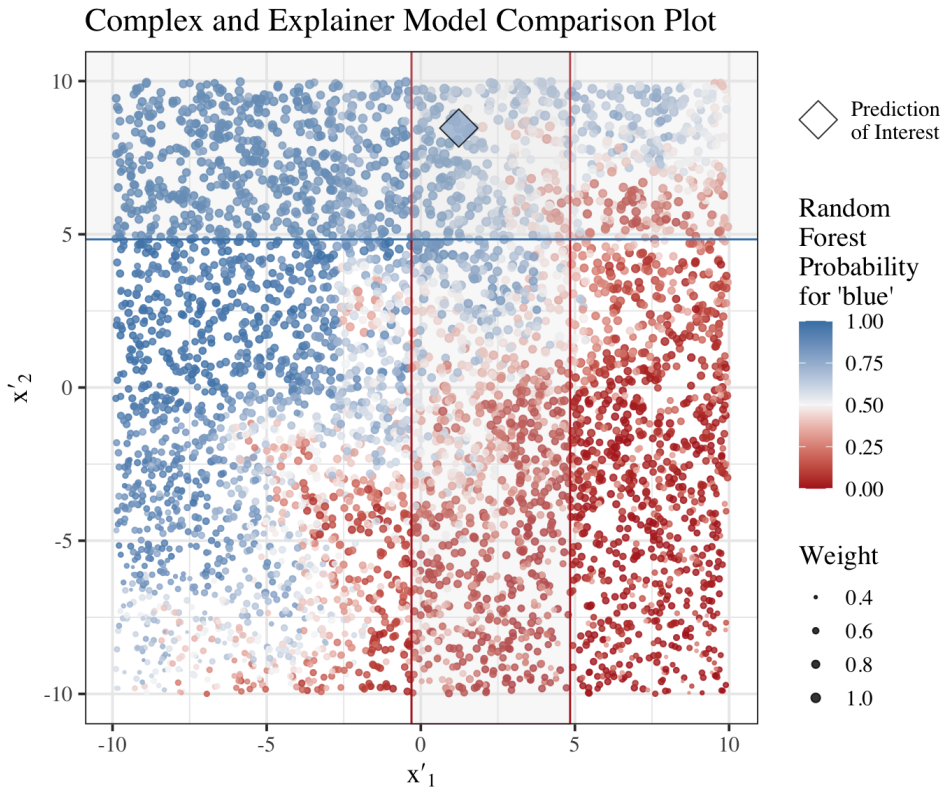
Diagnostics for Individual Explanations

Step 3b: Explainer Model Interpretation (continued)

- Interpret the explainer model to explain the complex model prediction

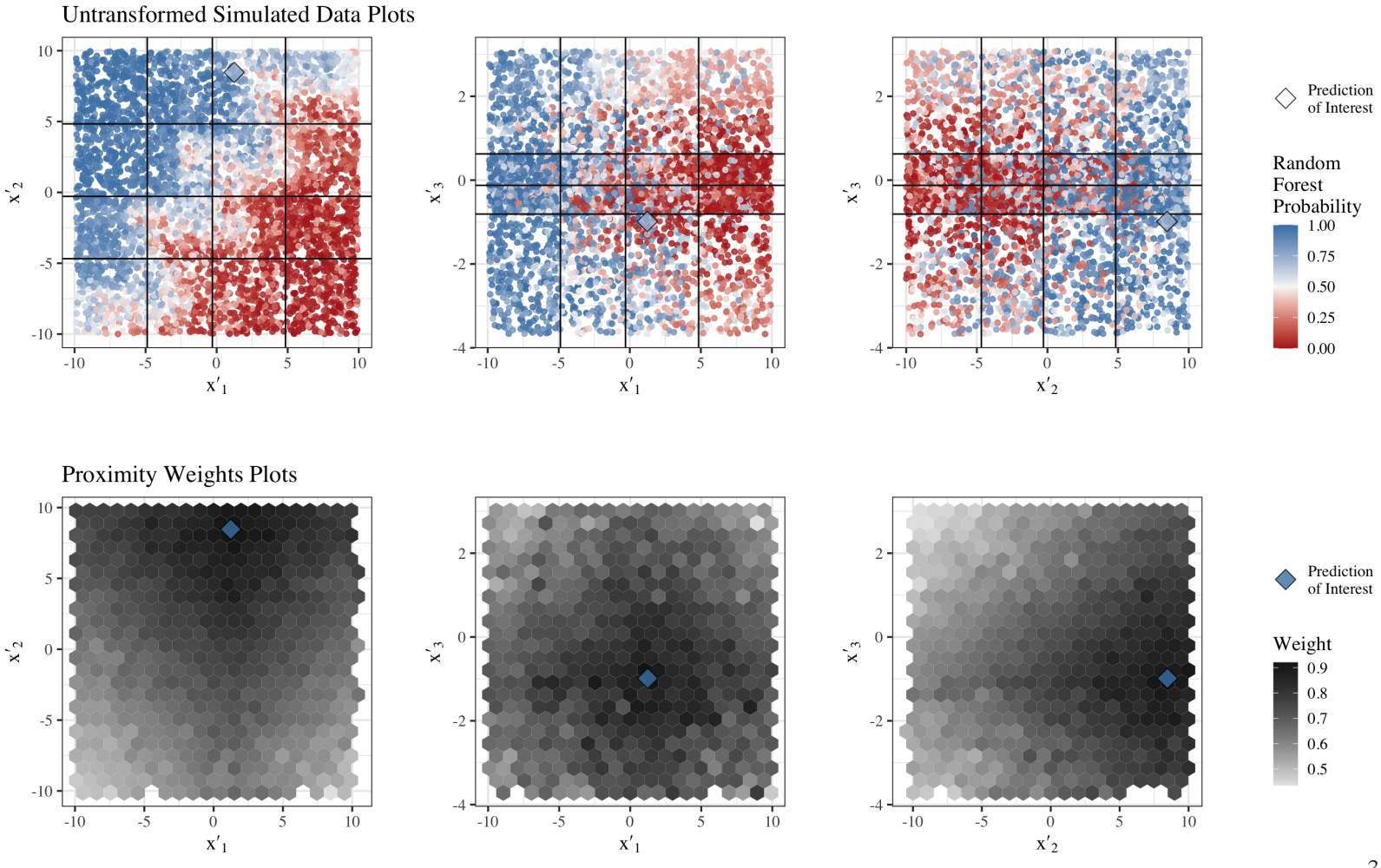
Corresponding Diagnostic: *Complex and Explainer Model Comparison Plot*

- Scatter plot of the simulated data
- Solid lines depict the interpretability transformation boundaries
- Line color represents whether the corresponding explainer model coefficients supports a random forest prediction of 'blue' (blue) or not (red)

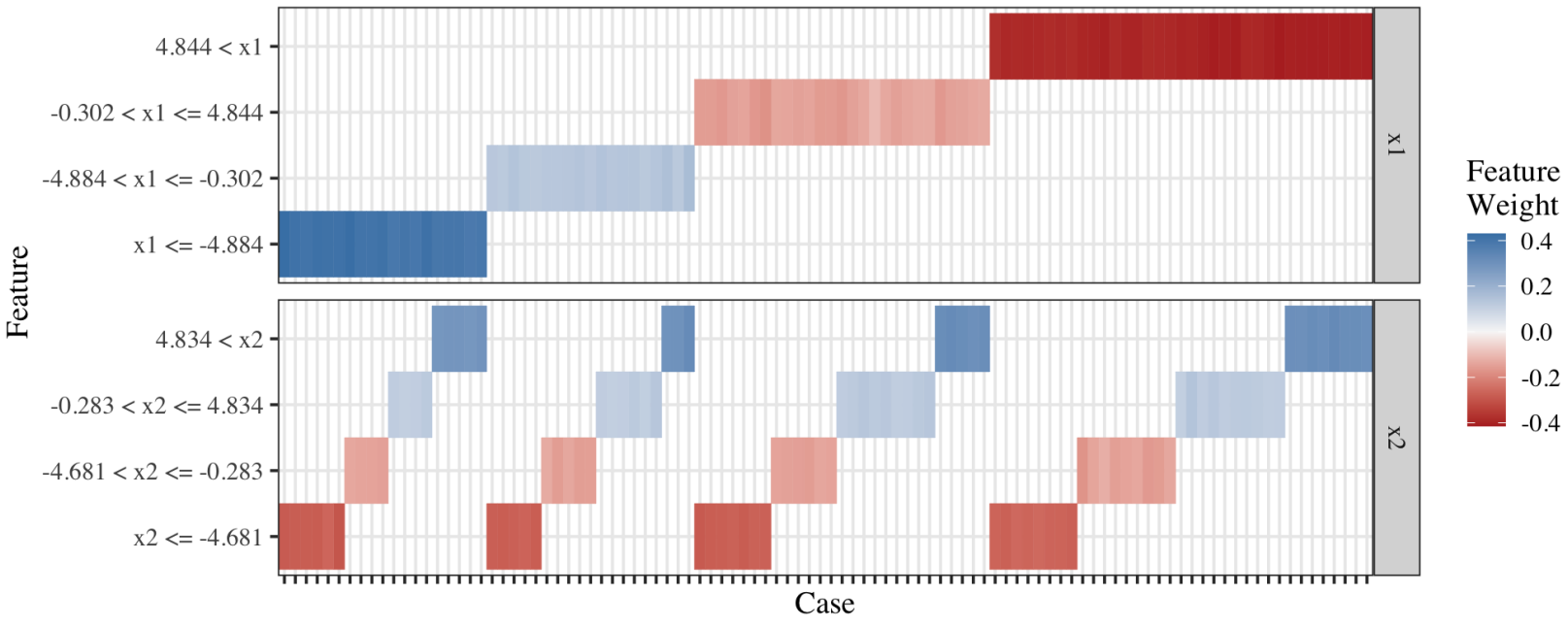


Diagnostics for Individual Explanations

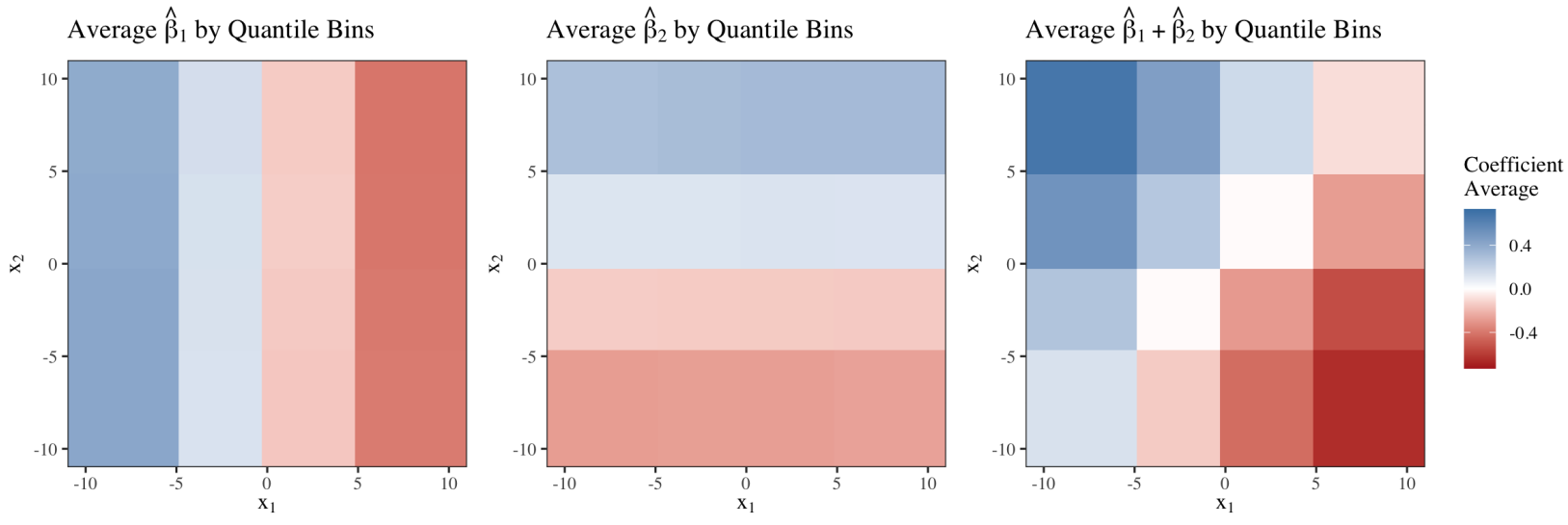
Extensions to Higher Dimensions



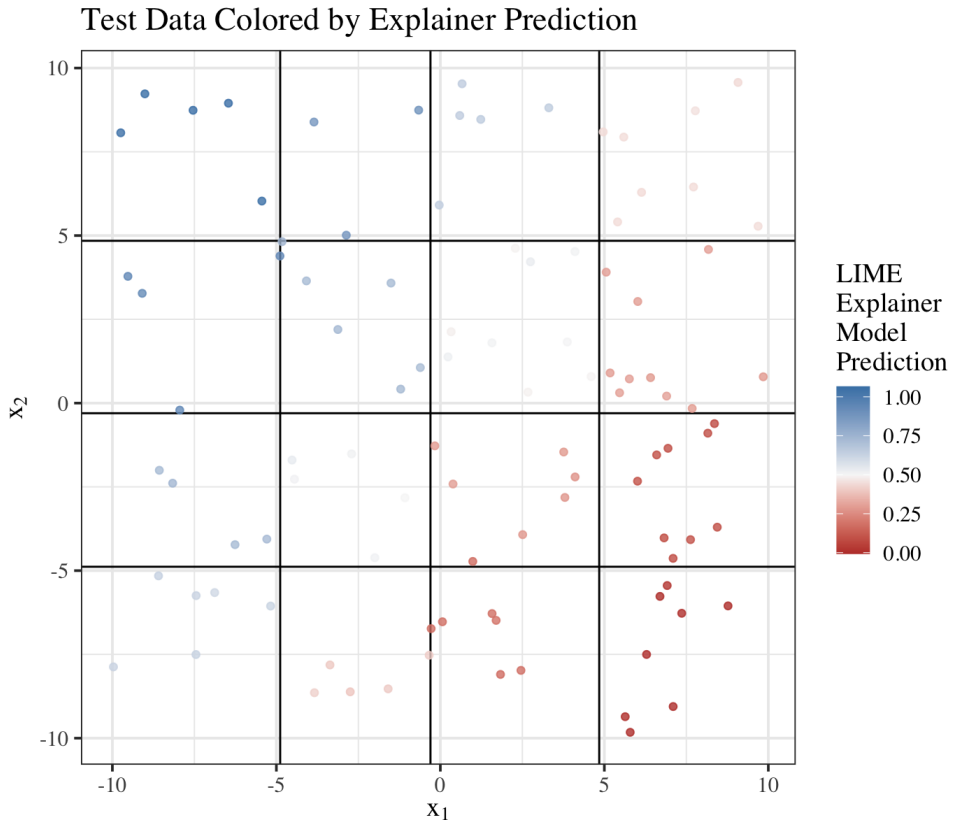
Diagnostics for Sets of Explanations



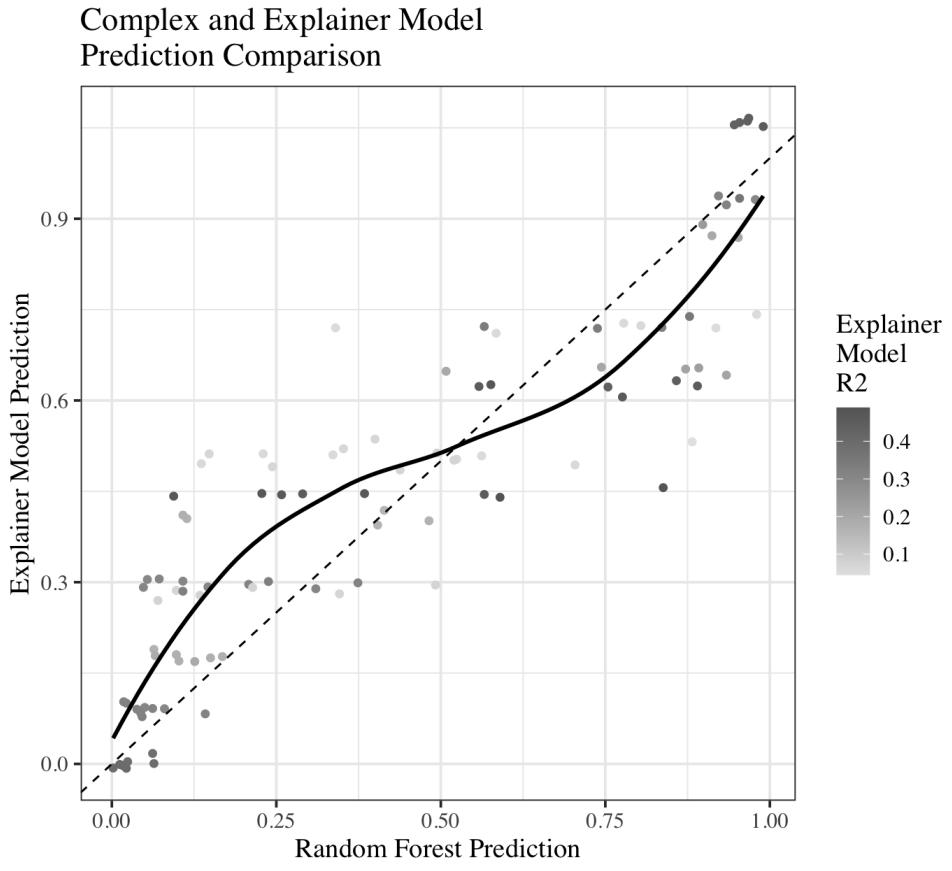
Diagnostics for Sets of Explanations



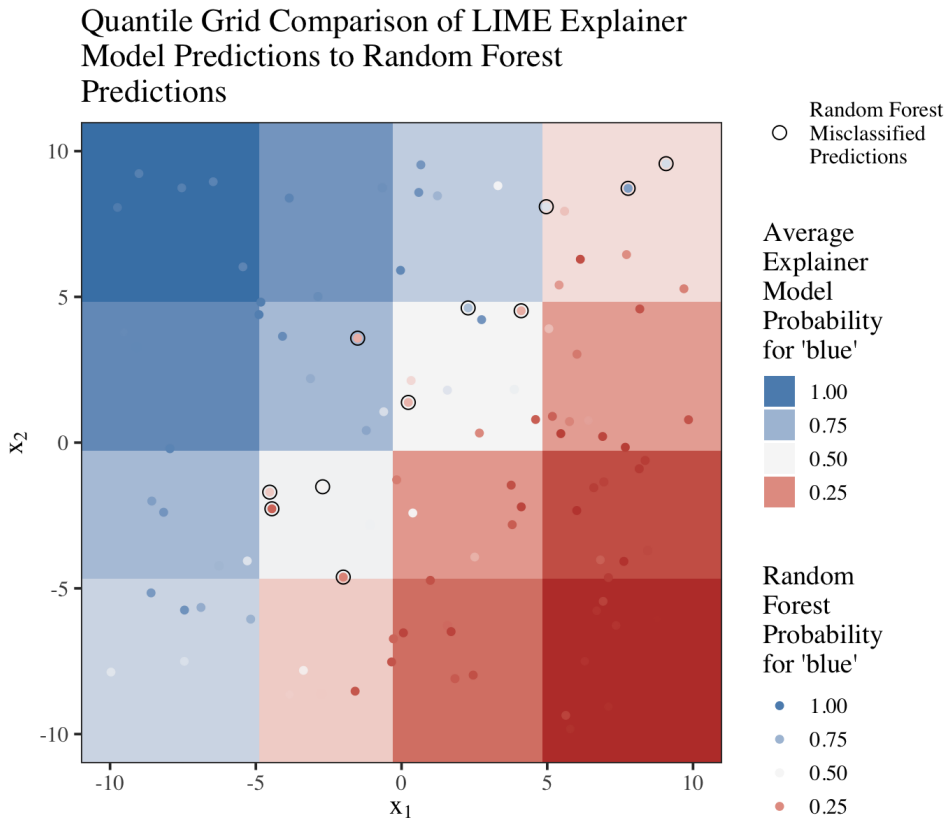
Diagnostics for Sets of Explanations



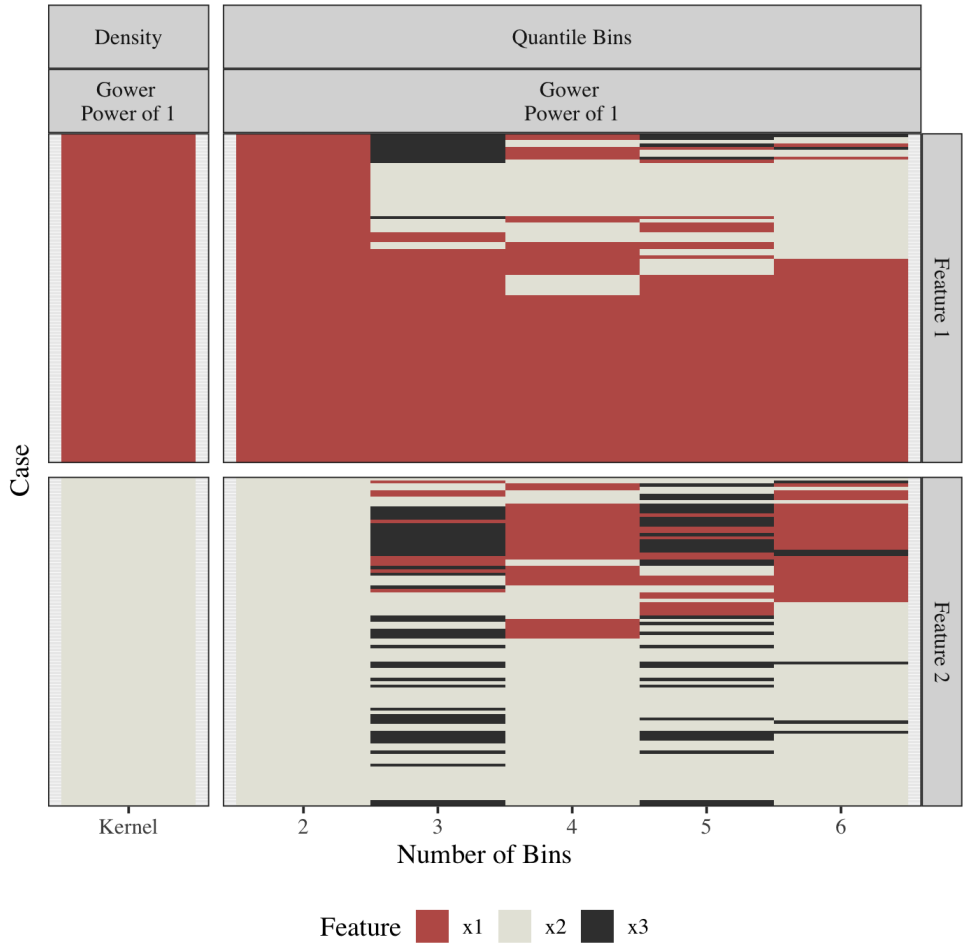
Diagnostics for Sets of Explanations



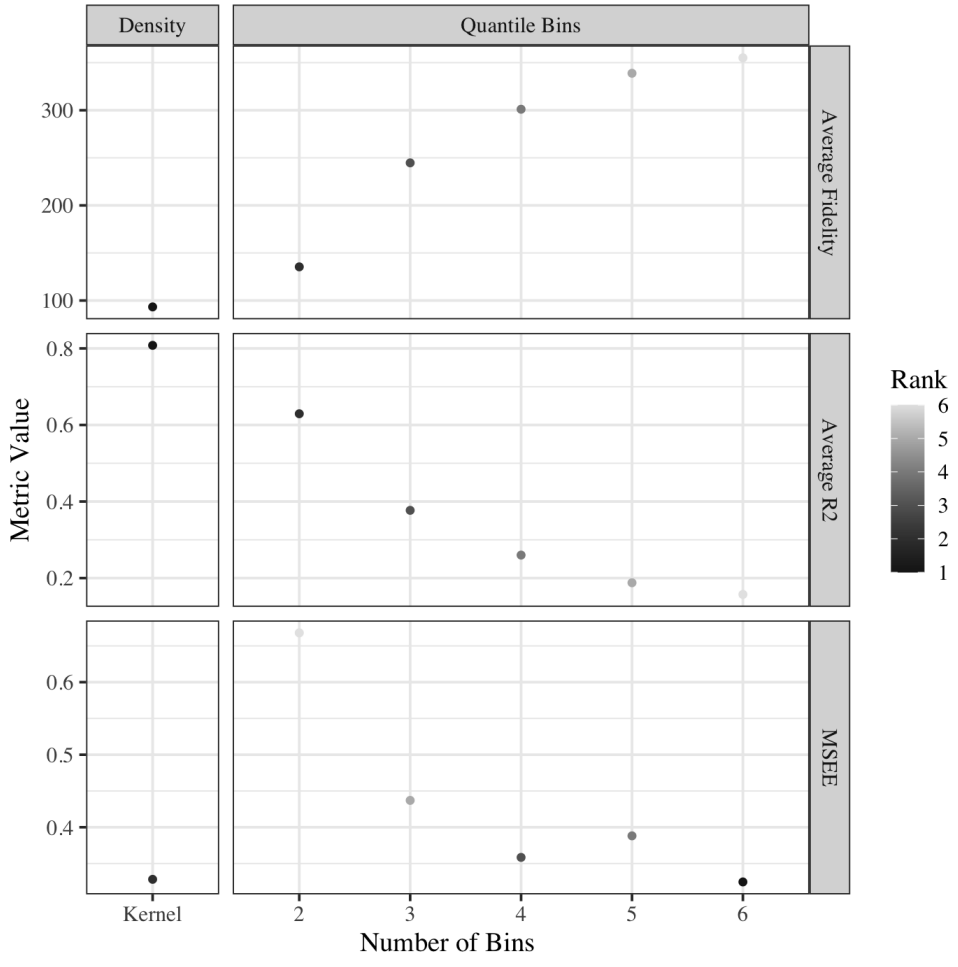
Diagnostics for Sets of Explanations



Diagnostics for Comparisons of Tuning Parameters



Diagnostics for Comparisons of Tuning Parameters



Chapter 2: Explaining Random Forests using Clustering of Trees

Motivation

- Since LIME did not provide reasonable explanations for the bullet matching random forest from the study by (Hare, Hofmann, and Carriquiry, 2016), it is still of interest to visualize the random forest to gain insight into the predictions.
- It is particularly of interest to provide explanations for incorrect predictions.

Current Ideas

Classification Tree as a Global Explainer Model

Current Ideas

Clustering to Identify Key Tree Paths in the Random Forest

Chapter 3: Extensions of Neural Network Explanation Tools to Tabular Data Applications

Motivation

- The majority of research associated with explainable machine learning for neural networks is focused on image data.
- We are working with a Bayesian neural network applied to functional data. It is of interest to use visualizations to provide explanations for the predictions made by the model.

Include figure of functional data here

Current Approaches

Feature Visualization

Current Approaches

Permutation Feature Importance

Ideas for Future Work

Feature visualization adjustments

- Visualize the functional principal components

Develop extensions of other explainable neural network methods that have been developed for image data

- Saliency maps and partial dependence plots

Visualizations of the paths of an observation through the network

- Application/extension of flow such as the figure below from (Halnaut, Giot, Bourqui, et al., 2020)

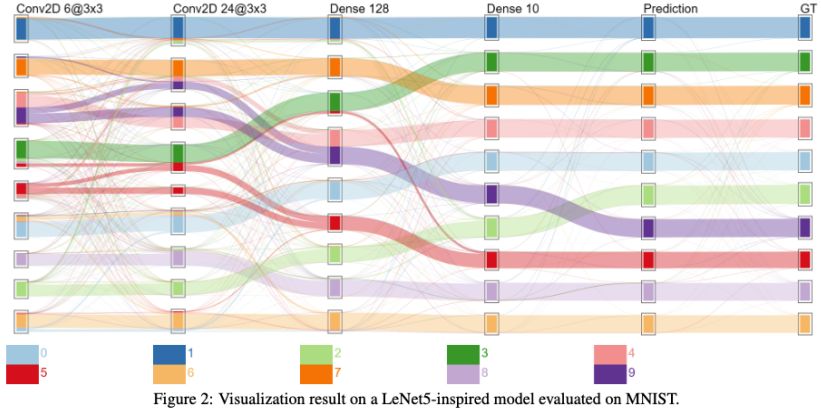


Figure 2: Visualization result on a LeNet5-inspired model evaluated on MNIST.

Timeline for Completion

Discussion Points

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

Need to format this eventually using `start` and `end` options

Altmann, A, L. Toloşi, 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.