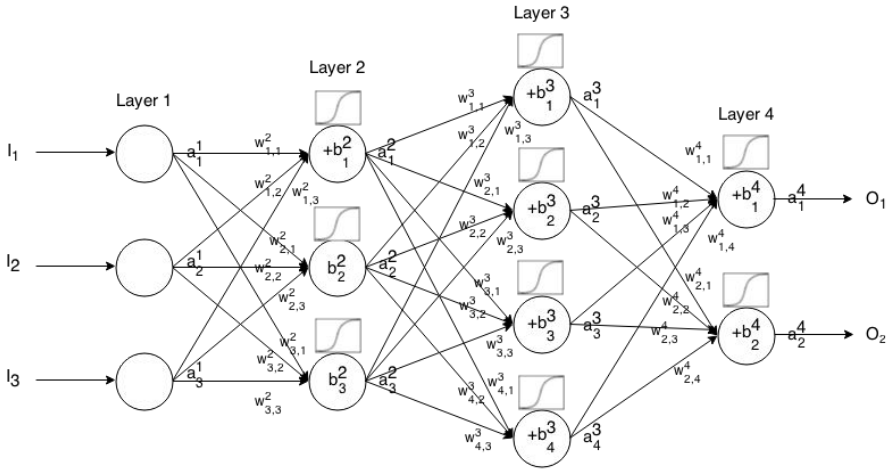
Explaining Black Box Machine Learning Models

Visual Techniques and Statistical Ideas

Katherine Goode and Xiaodan (Annie) Lyu
November 20, 2020
Presentation for ACS Data Analytics team at Autodesk

Background and overview of explainability

What is explainability?



Interpretability and Explainability

No agreed upon definition in the literature, but here is my preferred distinction...

Interpretability = ability to directly use model to understand how predictions are made

Explainability = ability to indirectly use model to understand how predictions are made

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

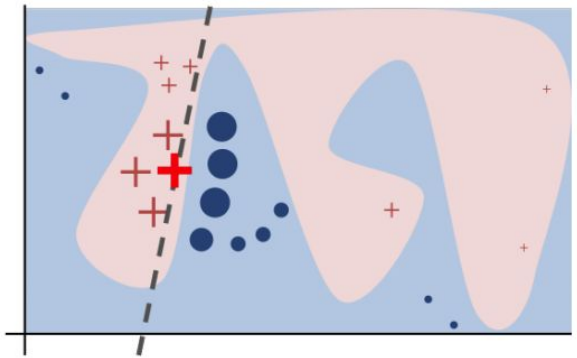


Figure from LIME paper (Ribeiro 2016)

General Data Protection Regulation (GDPR)

General aim (via Wikipedia):

> *"The GDPR's primary aim is to give control to individuals over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU."*

Connection to explainability:

- GDPR implemented in 2018 and 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?"*

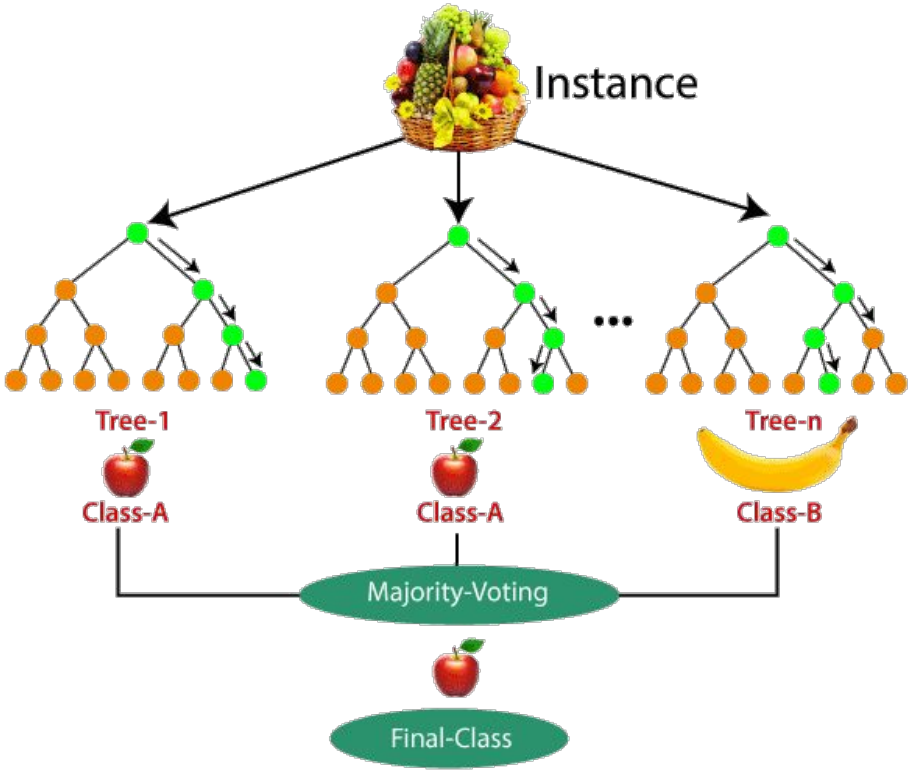
Explanation Audience

Who is the explanation for?

- Data analyst
- Subject matter expert (doctor, scientist, business person)
- Decision maker (client, government official, patient, jury)

Types of Explanations

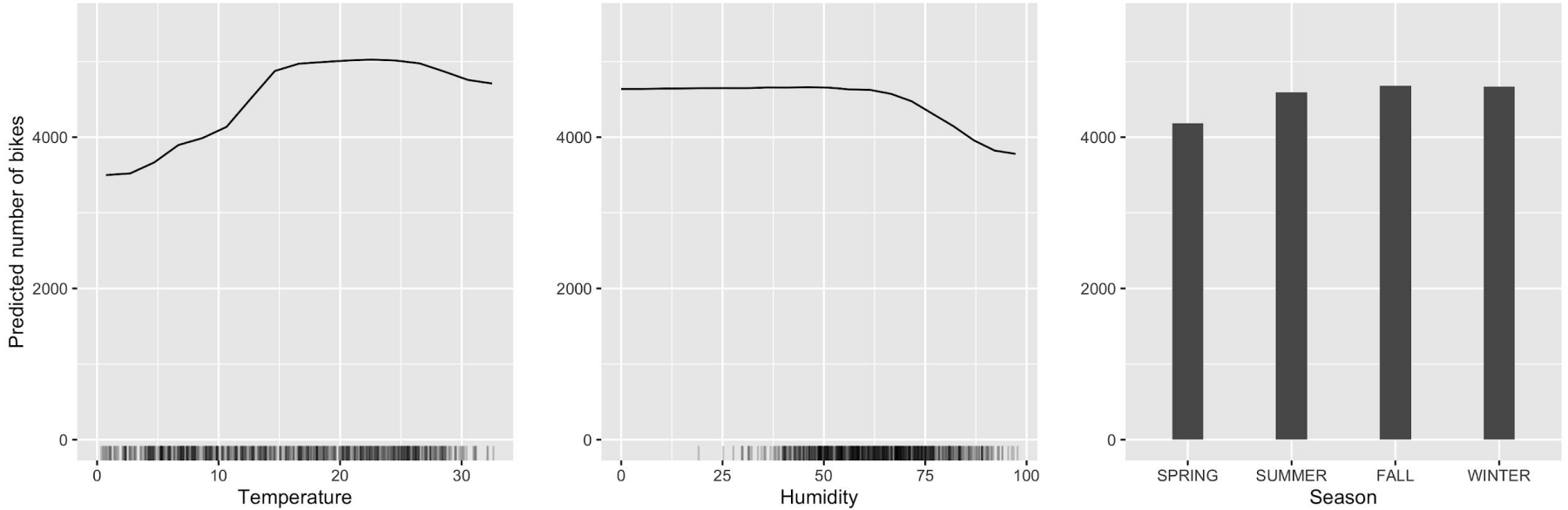
- Model versus case level (aka “global” and “local” explanations)
- Model specific versus model agnostic
- Model development versus post development
- Model structure versus prediction



Model level explanations

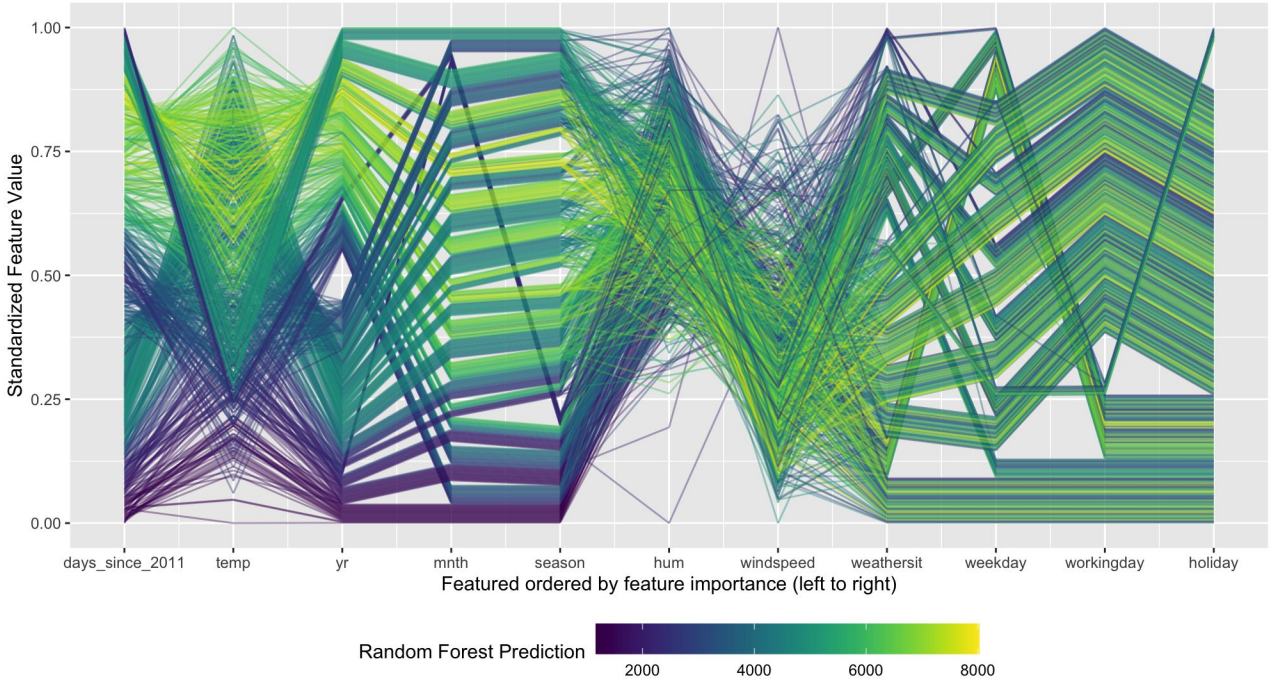
Partial Dependence Plot (PDP)

Depicts average relationship between prediction and feature (averaged over all other features)



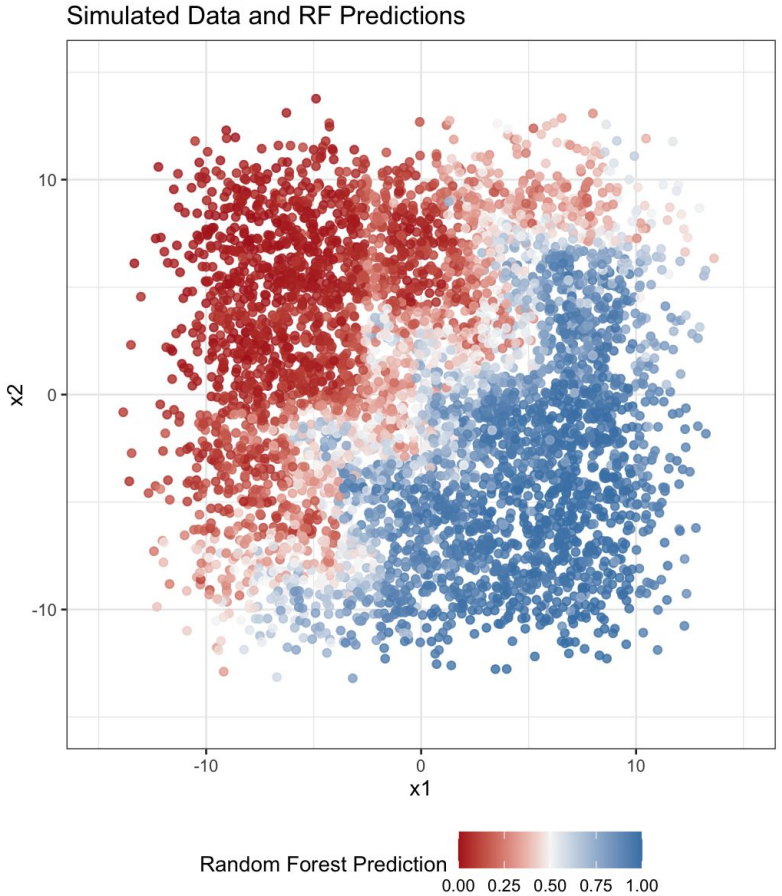
Parallel Coordinate Plot (PCP)

Shows (standardized) feature values and predictions of all observations in one overview figure



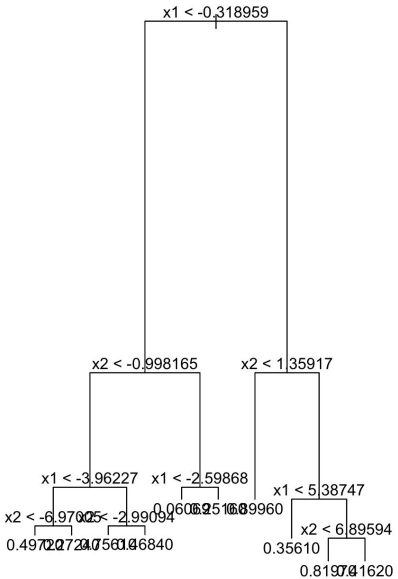
Global Explainer Models

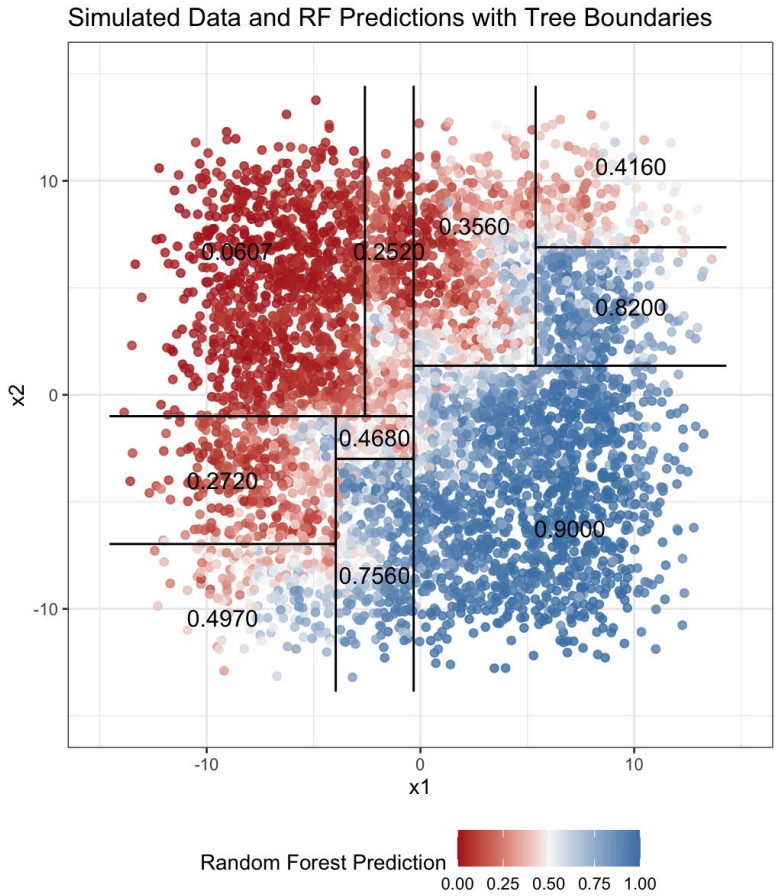
Idea: Use an interpretable model to approximate relationship between complex model predictions and model features



Global Explainer Models

Idea: Use an interpretable model to approximate relationship between complex model predictions and model features



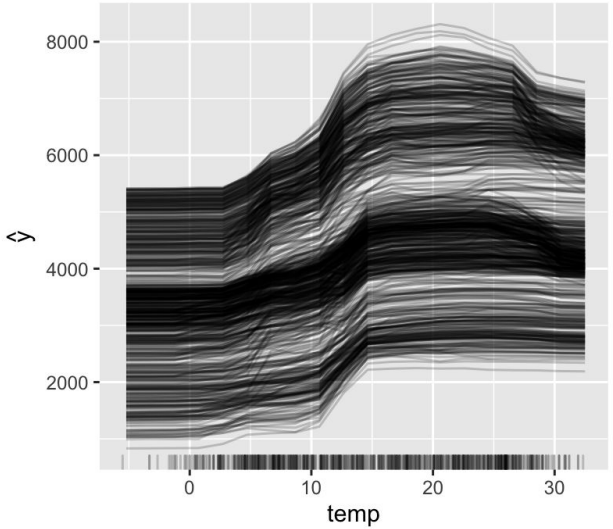


Case level explanations

Local Alternatives to PDPs and PCPs

Individual Conditional Explanation (ICE) Plot

Similar to partial dependence plots but for individual observations



Interactive Parallel Coordinate Plot

Could create an interactive parallel coordinate plot to select observation and investigate further

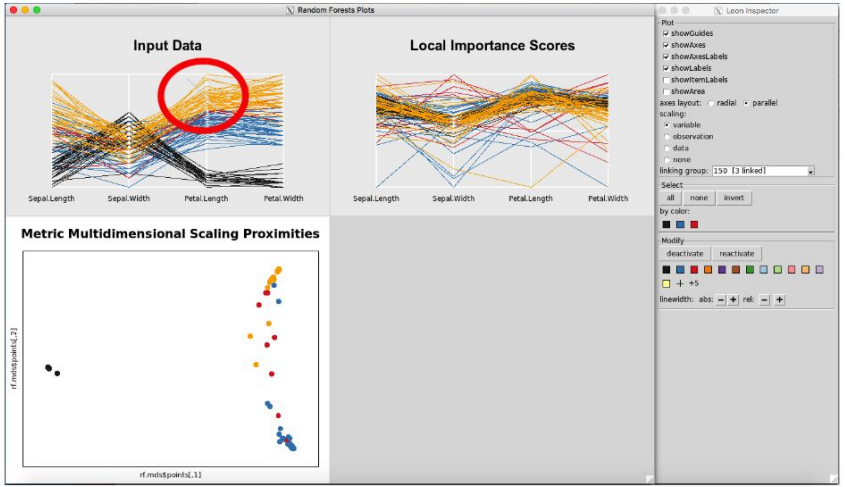
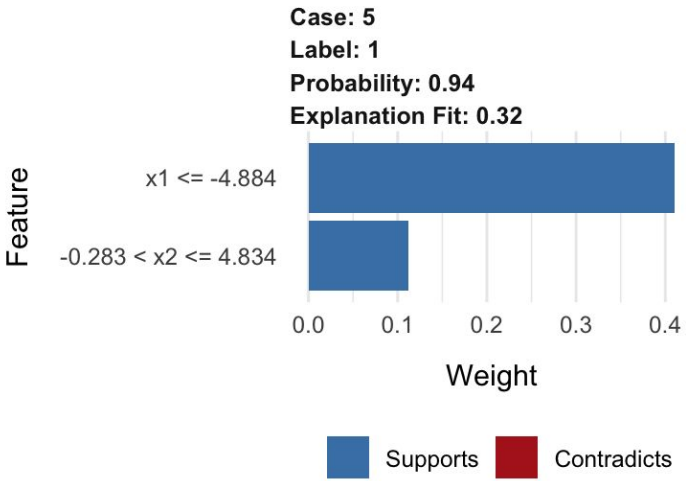
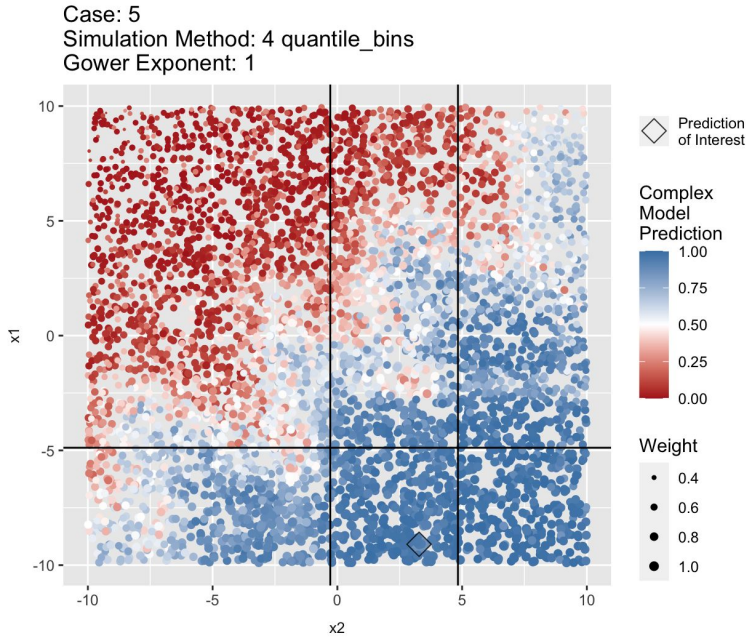


Image sources: (Left) https://christophm.github.io/interpretable-ml-book/ice.html
(Right) https://digitalcommons.usu.edu/gradreports/1335/

Local Explainer Models

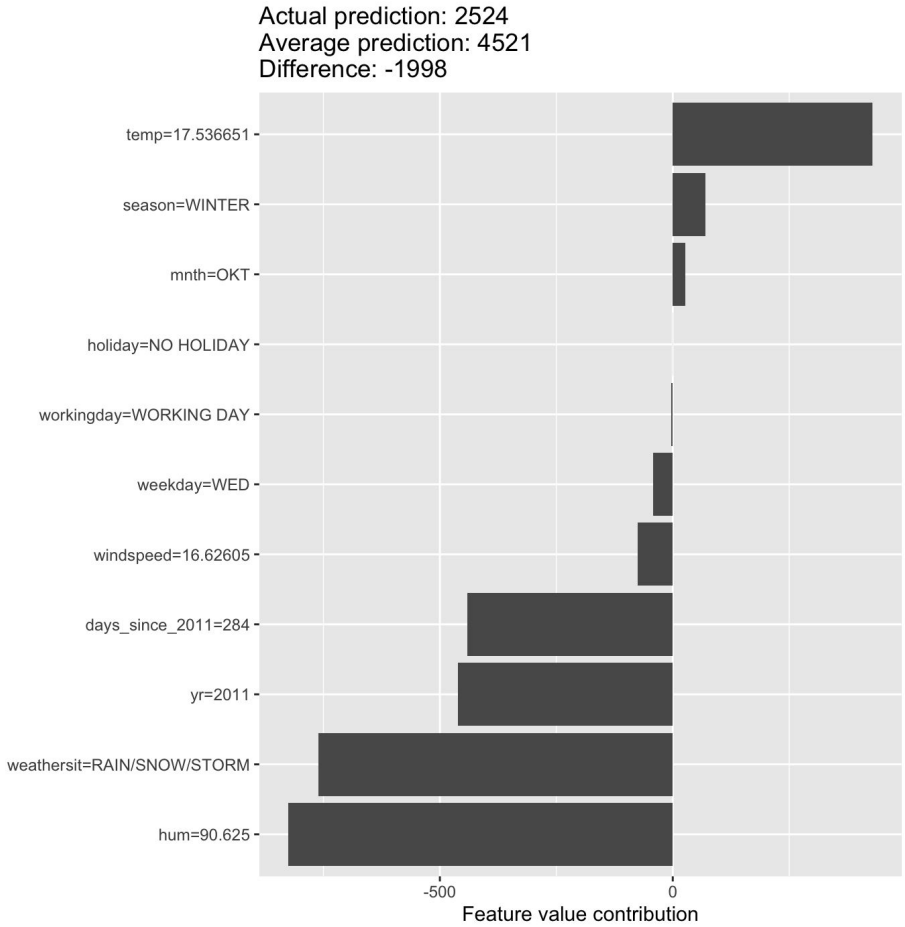
LIME (Local Interpretable Model-agnostic Explanations): Uses linear model to approximate relationship between complex model predictions and features in a local region





SHAP

- Determines contribution of each feature for a specific prediction
- Bars in figure represent: how much a feature contributes to the prediction compared to the average prediction



Model specific methods

Random Forests: Trace Plots

Displays all trees in the forest to show common patterns and variability

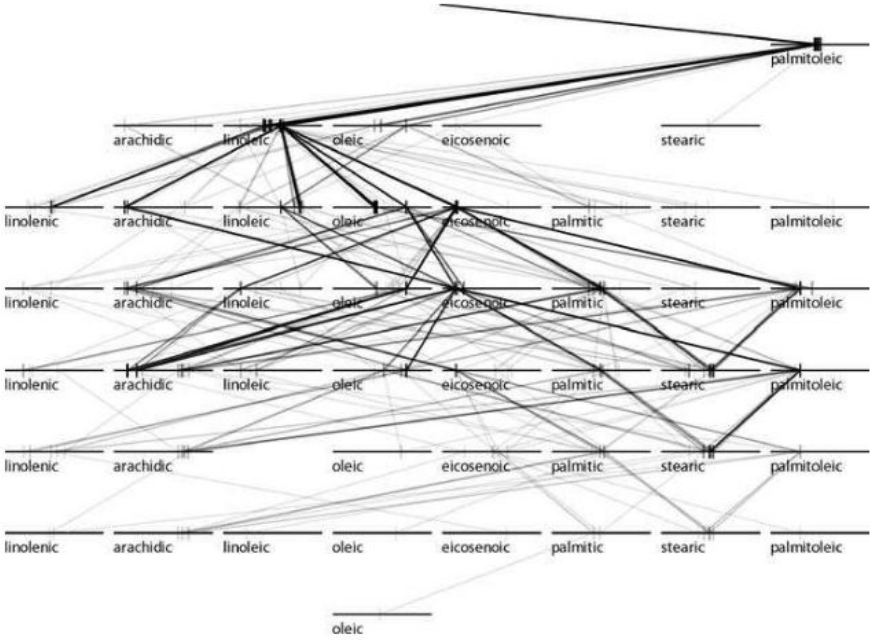


Figure 10.14. Trace plot of 100 bootstrapped trees

Image from Simon Urbank's chapter in
https://haralick.org/DV/Handbook_of_Data_Visualization.pdf

Neural Networks: Saliency Maps

Identifies how much a feature influences a prediction (based on the gradients within a neural network)

Image from https://arxiv.org/abs/1312.6034

Comments on EML

Rudin's paper

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

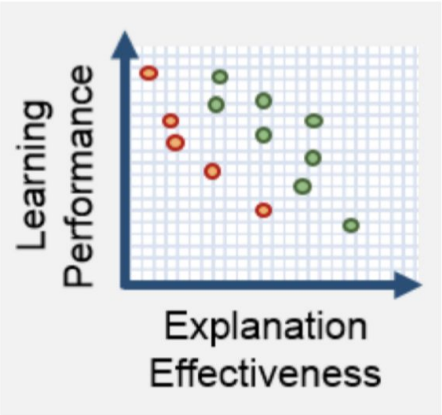
Cynthia Rudin
Duke University
cynthia@cs.duke.edu

Abstract

Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to *explain* black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward – it is to design models that are inherently interpretable. This manuscript clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare, and computer vision.

Some of Rudin's statements:

- "It is a myth that there is necessarily a trade-off between accuracy and interpretability"



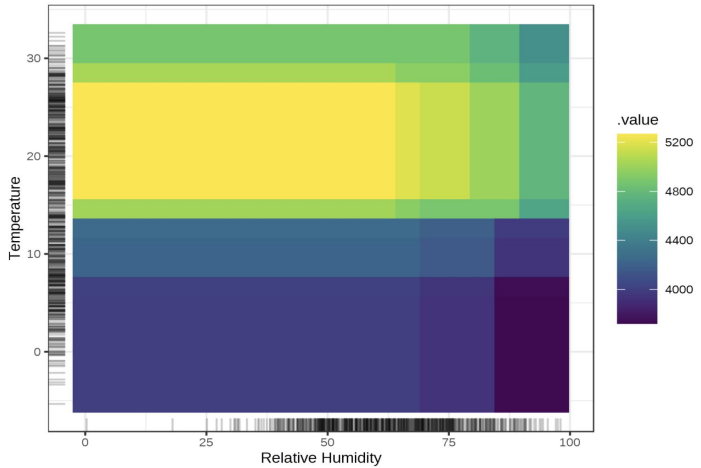
- "Explainable ML methods provide explanations that are not faithful to what the original model computes."
- "Explanations often do not make sense, or do not provide enough detail to understand what the black box is doing."

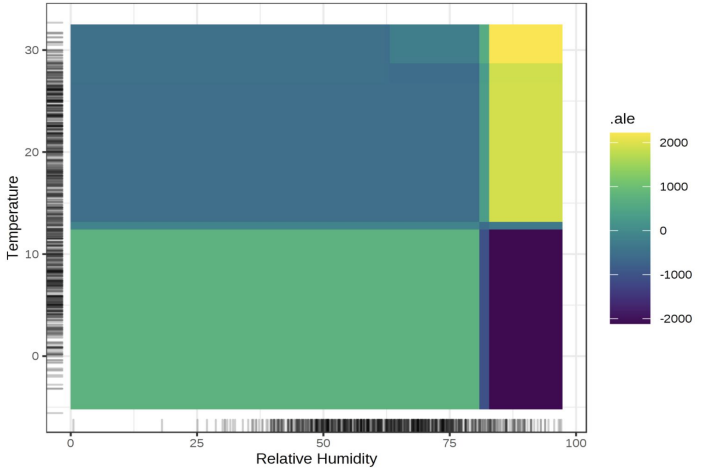
Accounting for Correlation

Correlation between features has often been shown to negatively affect explainability methods

One example....

Accumulated Local Effects (ALE) Plots: Similar to partial dependence plots but account for correlation

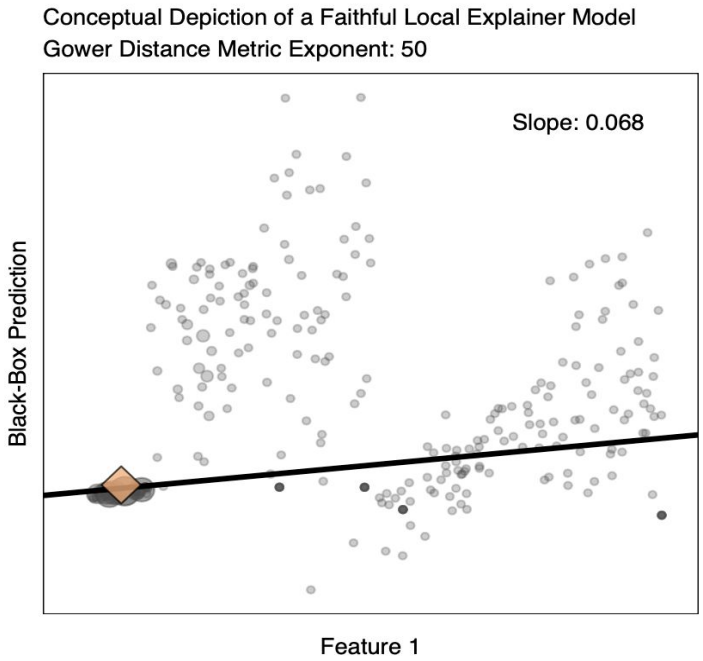


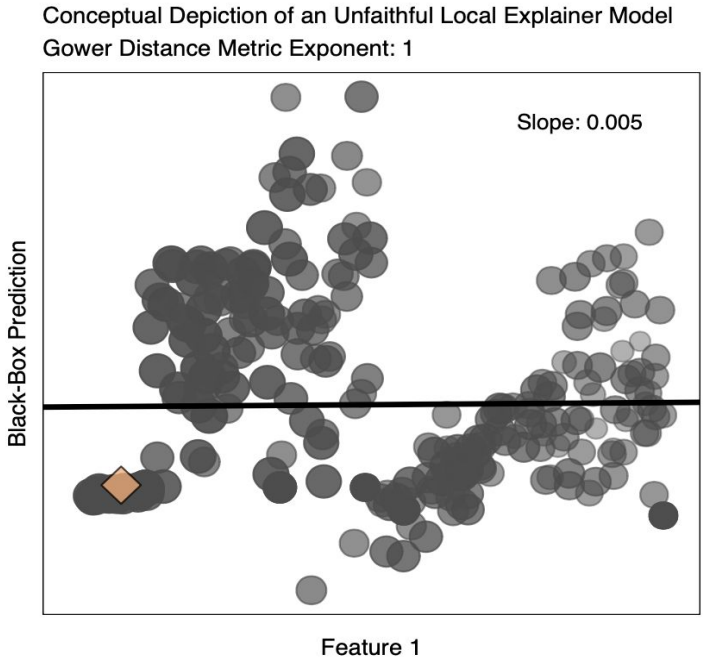


Images from https://christophm.github.io/interpretable-ml-book/ale.html

LIME or Lemon?

Important to assess fit of local explainer model but not often done in practice...use LIME with caution!





So many more methods...

- Interpretable Machine Learning: Online book by Christoph Molner
- Distill: Website with understandable peer reviewed articles about machine learning
- Removing the blindfold: Paper by Hadley Wickham, Di Cook, and Heike Hofmann on general practices for visualizing models
- A few of the many EML overview papers:
  - https://arxiv.org/pdf/1811.11839.pdf
  - https://arxiv.org/pdf/1806.00069.pdf
  - https://arxiv.org/pdf/1802.01933.pdf