



Interpretable Machine Learning





Hello!

I am Anzor Gozalishvili

Lead Machine Learning Engineer & Data Scientist @ [MaxinAI](#)

Email: anzor.gozalishvili@maxinai.com



Table of Contents

- ◇ Simple Machine Learning steps
- ◇ Non-Explained ML Models outputs
- ◇ Interpretability Definition
- ◇ Interpretability Taxonomy
- ◇ Interpretable & Non-Interpretable Models
- ◇ Model-Agnostic Interpretation methods
- ◇ The Future of Interpretability

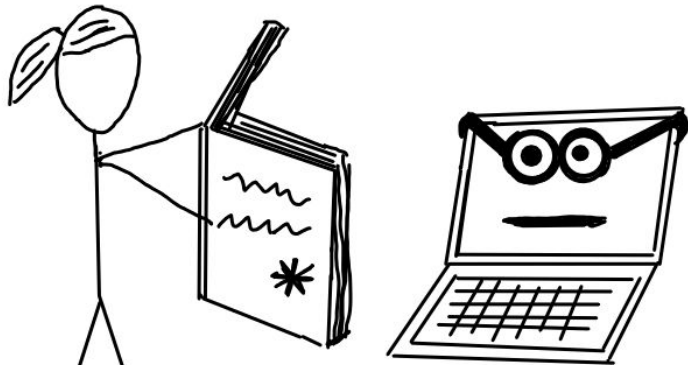


1

Machine Learning?

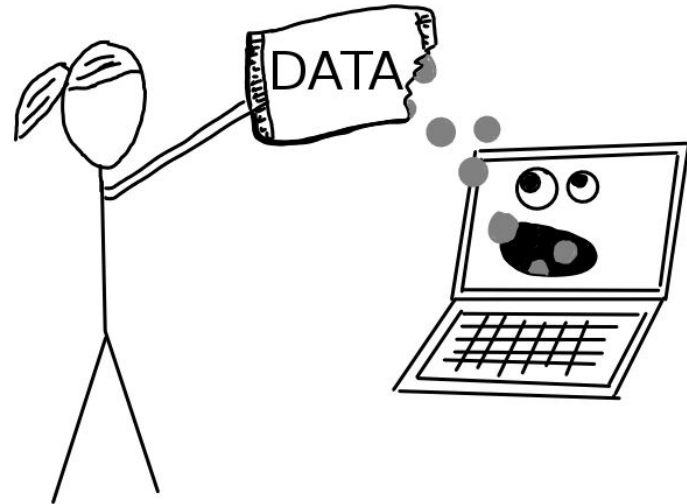


Without Machine Learning



* VERY SPECIFIC INSTRUCTIONS

With Machine Learning



Simple Machine Learning Steps





2

Non-Explained outputs of ML Models

Bias?
Fairness?

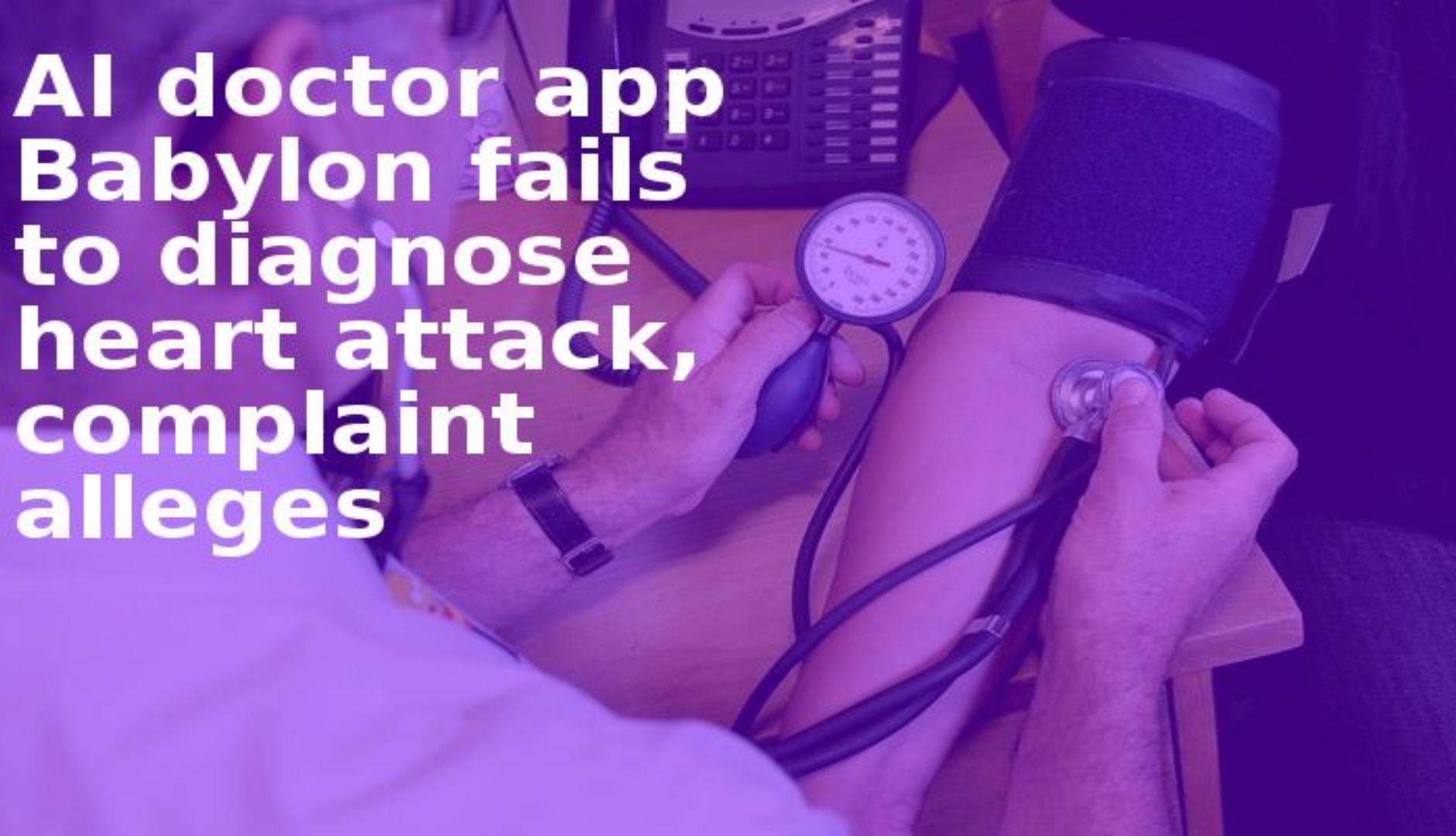




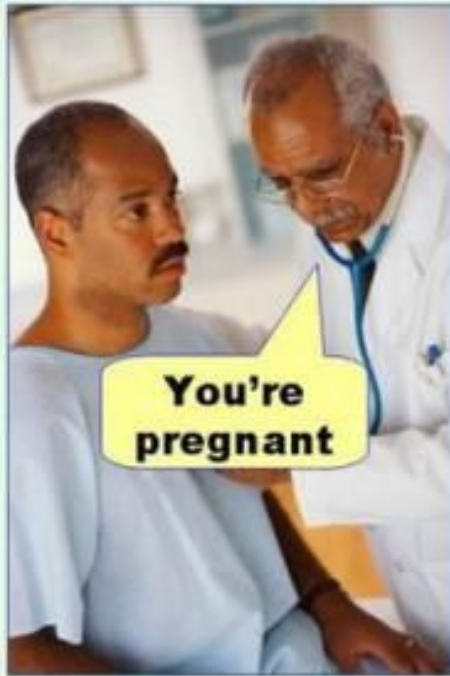
Better to say:

“your loan application was rejected because of lack of sufficient income proof!”

AI doctor app Babylon fails to diagnose heart attack, complaint alleges



Type I error
(false positive)



Type II error
(false negative)





3

Interpretability



“Interpretability is the degree to which a human can understand the cause of a decision”

A cluster of hexagons in various shades of blue and cyan, with some having white outlines, arranged in a non-uniform pattern in the top-left corner.

How to interpret?

- Feature summary statistics
- Feature summary visualization
- Model internals (e.g. learned weights)
- Data points
- Intrinsically interpretable model



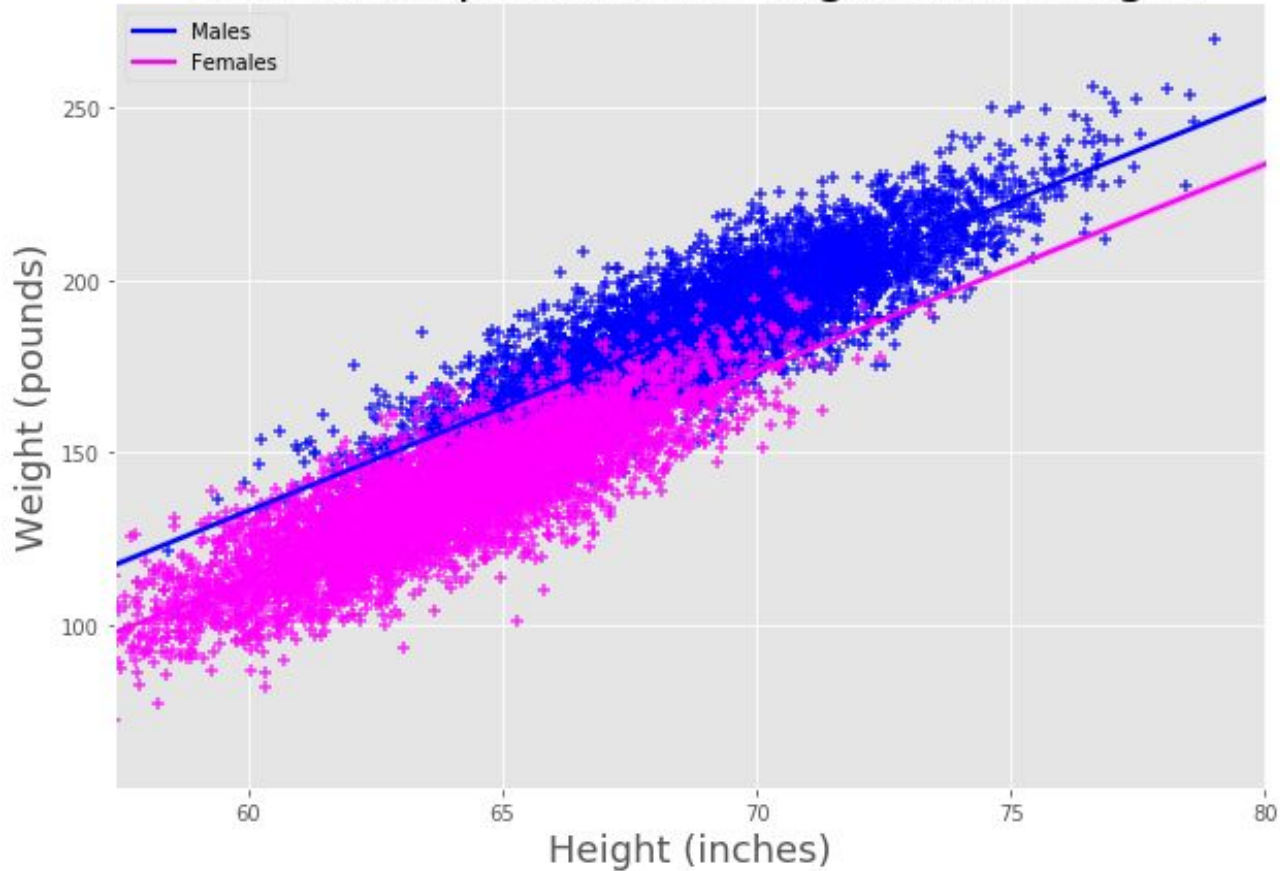


4

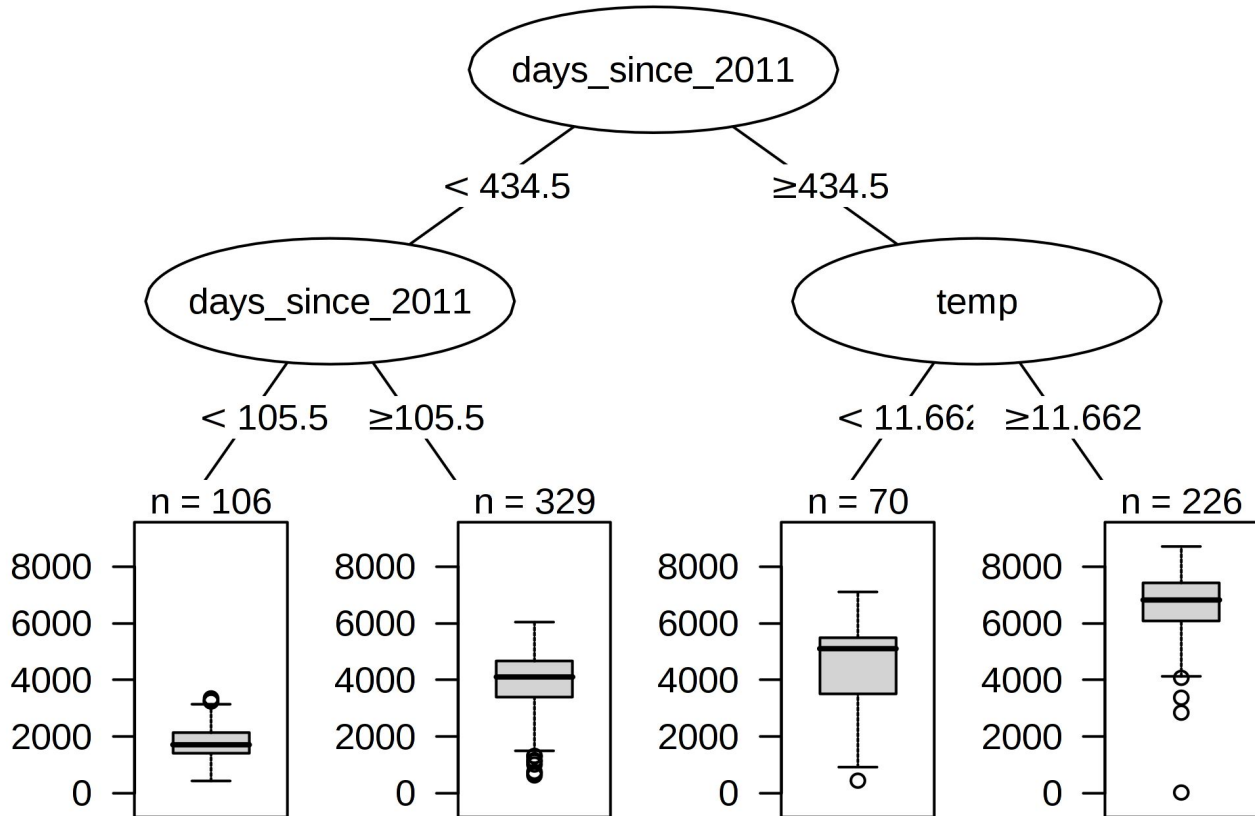
Interpretable Models

Linear Regression

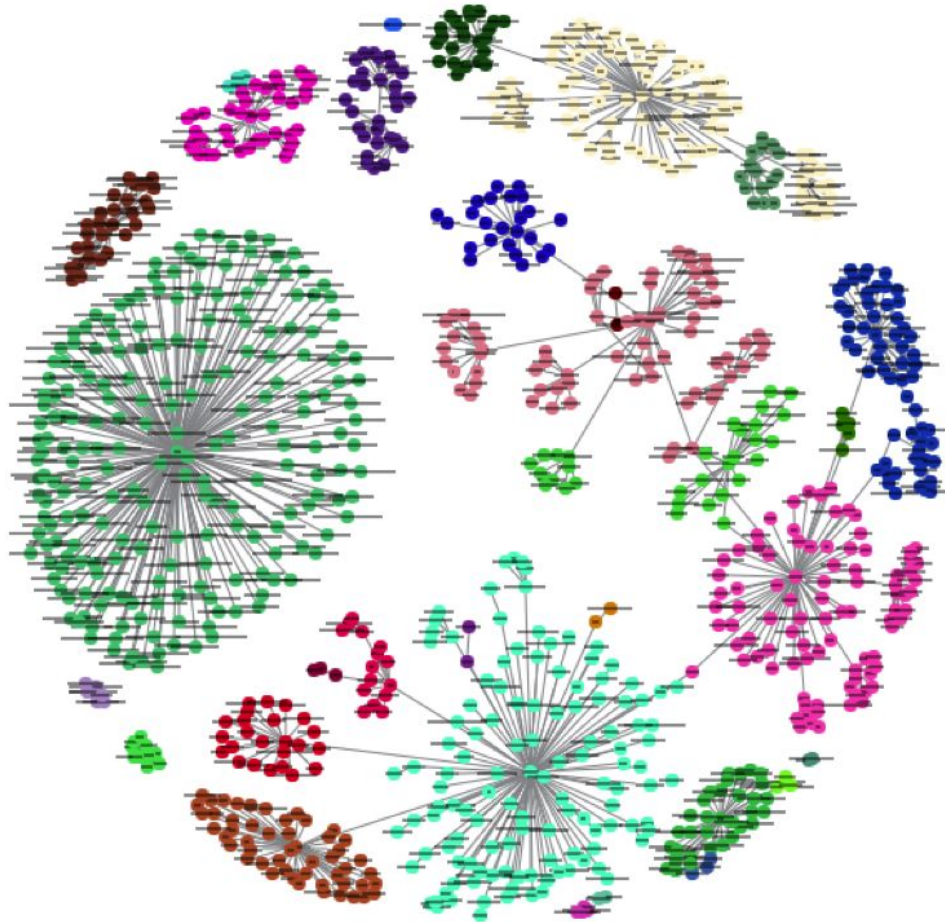
Relationship between Height and Weight



Decision Tree



K-Nearest Neighbours





5

Non-Interpretable Models

Error signal: 0.008

Variation weights: 0.741

Total loss: -0.679



COMPUTER VISION DIGIT RECOGNITION

Epoch: 4

Iteration: 97

Error: 6.107267

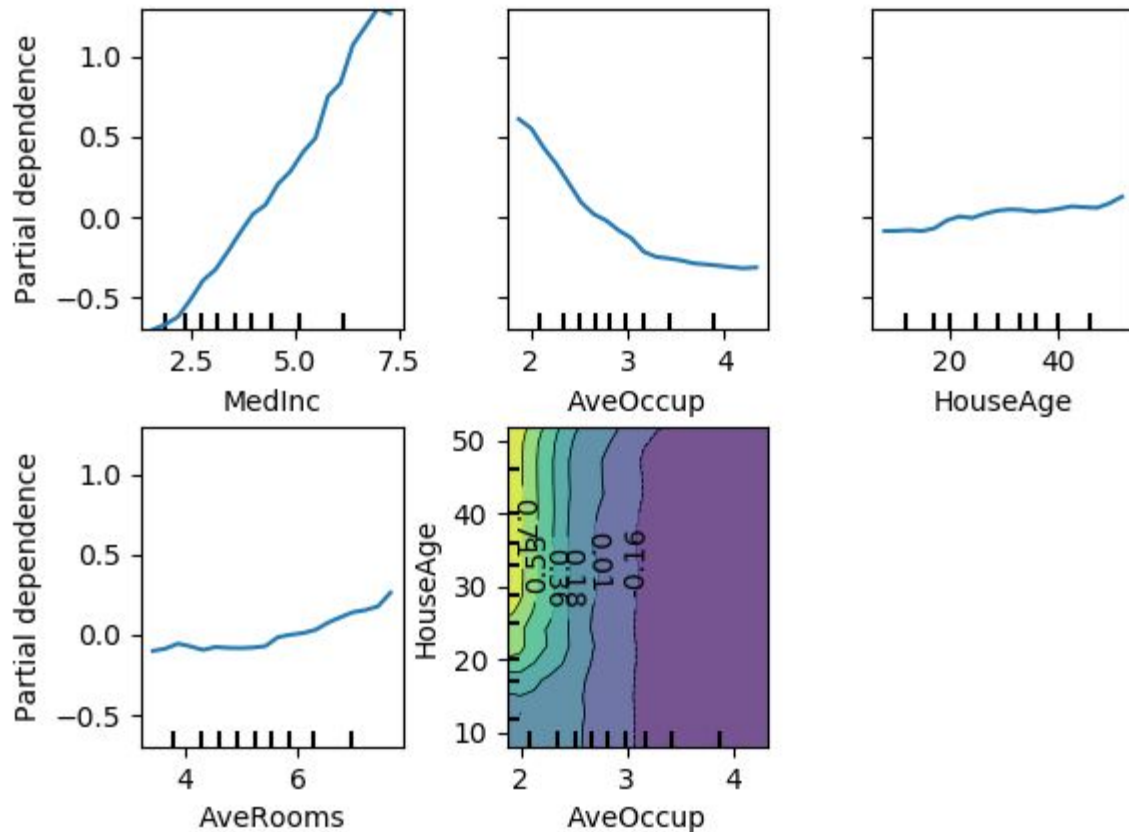


6

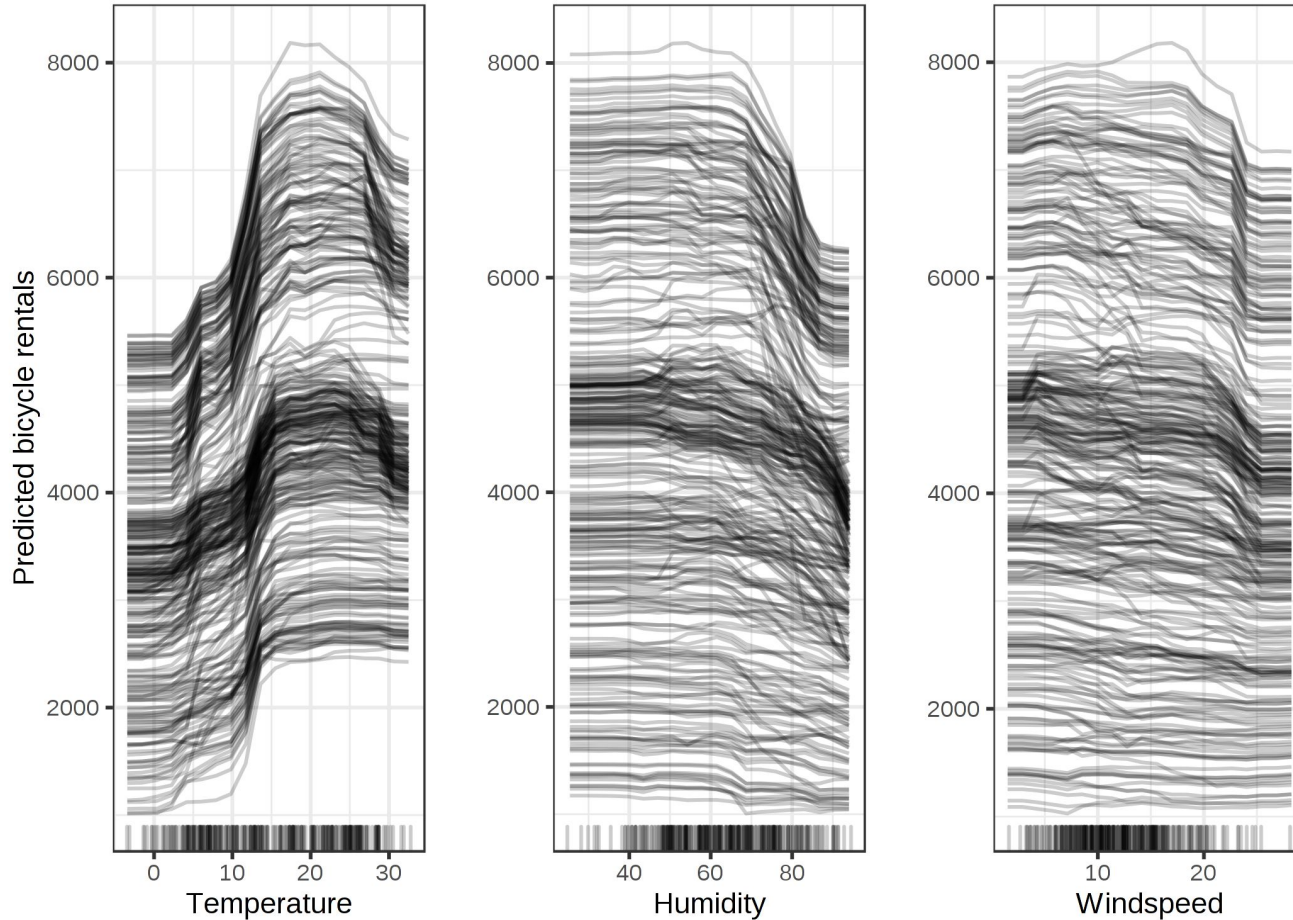
Model-Agnostic Interpretation Methods

Partial Dependence Plot (PPD)

Partial dependence of house value on non-location features for the California housing dataset, with Gradient Boosting

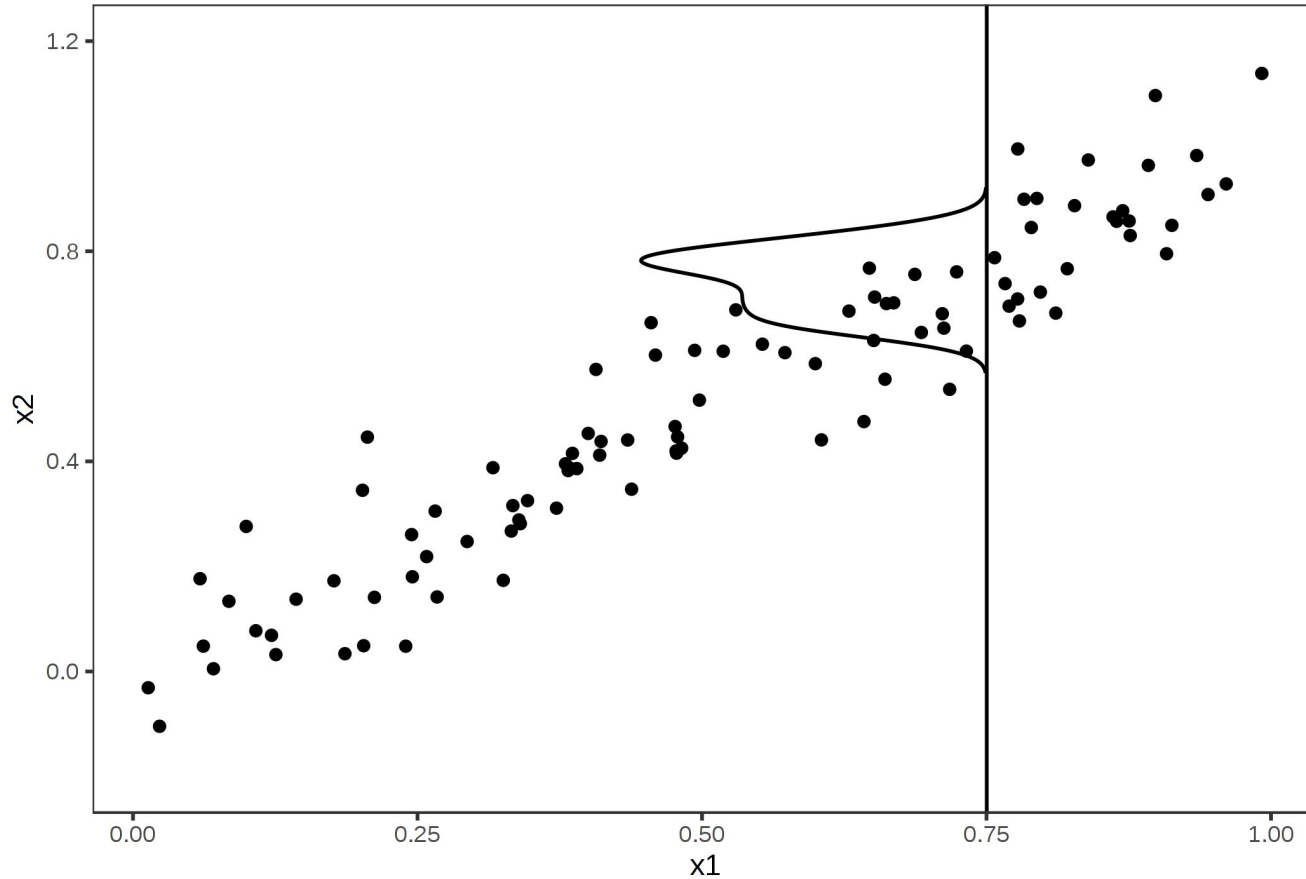


Individual Conditional Expectation (ICE)

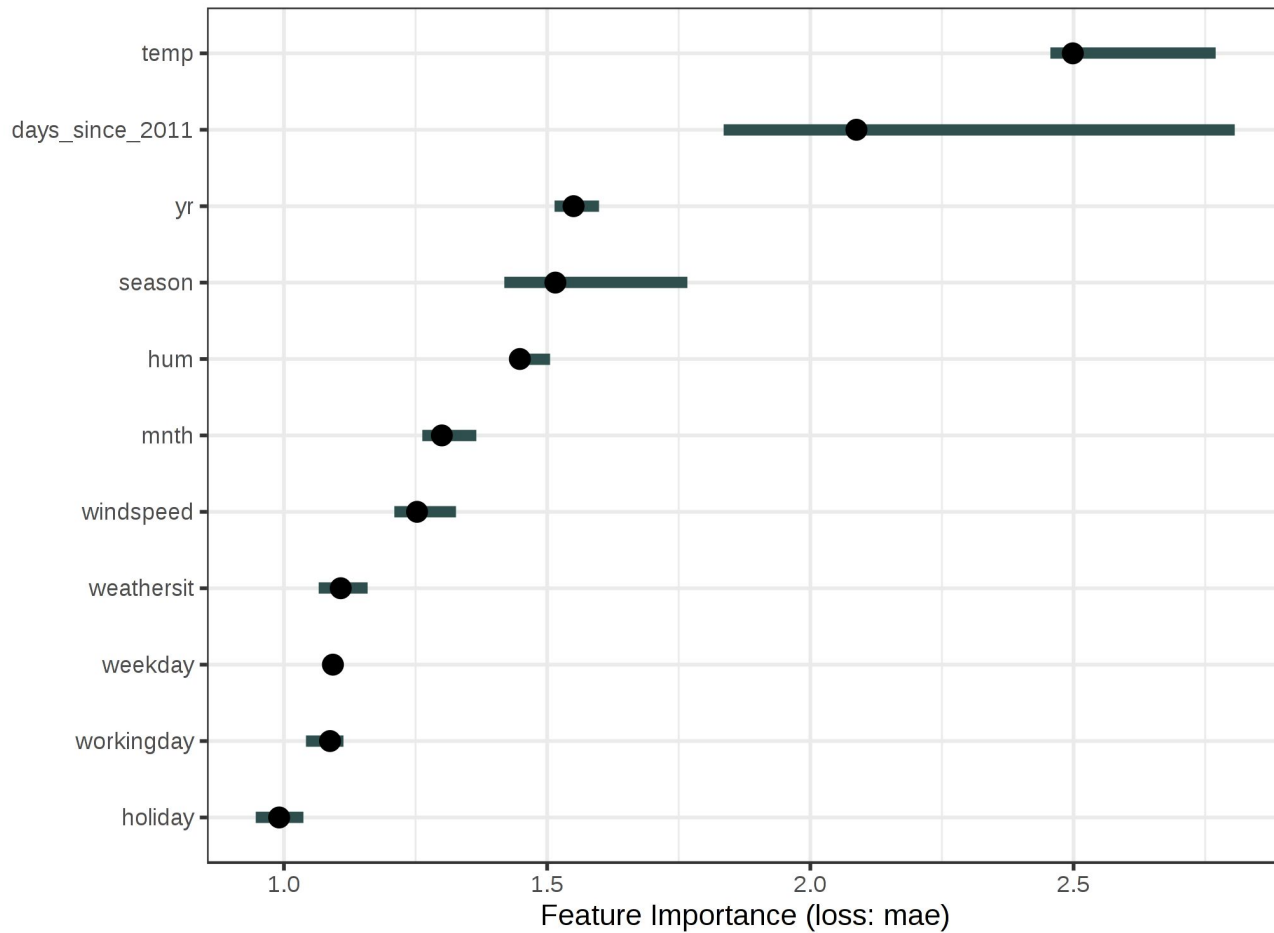


Marginal Plots (M-Plots)

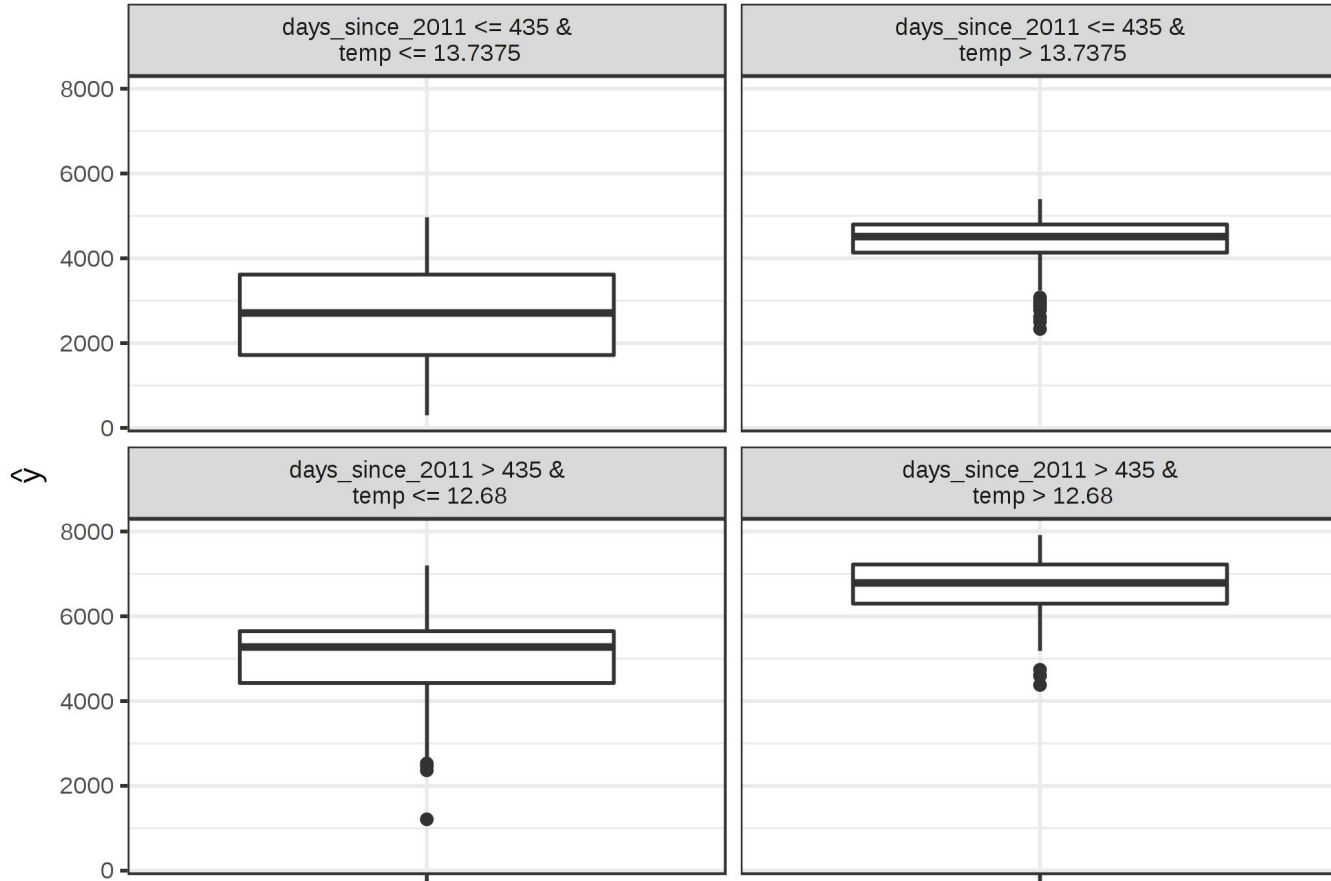
Conditional distribution $P(x_2|x_1=0.75)$



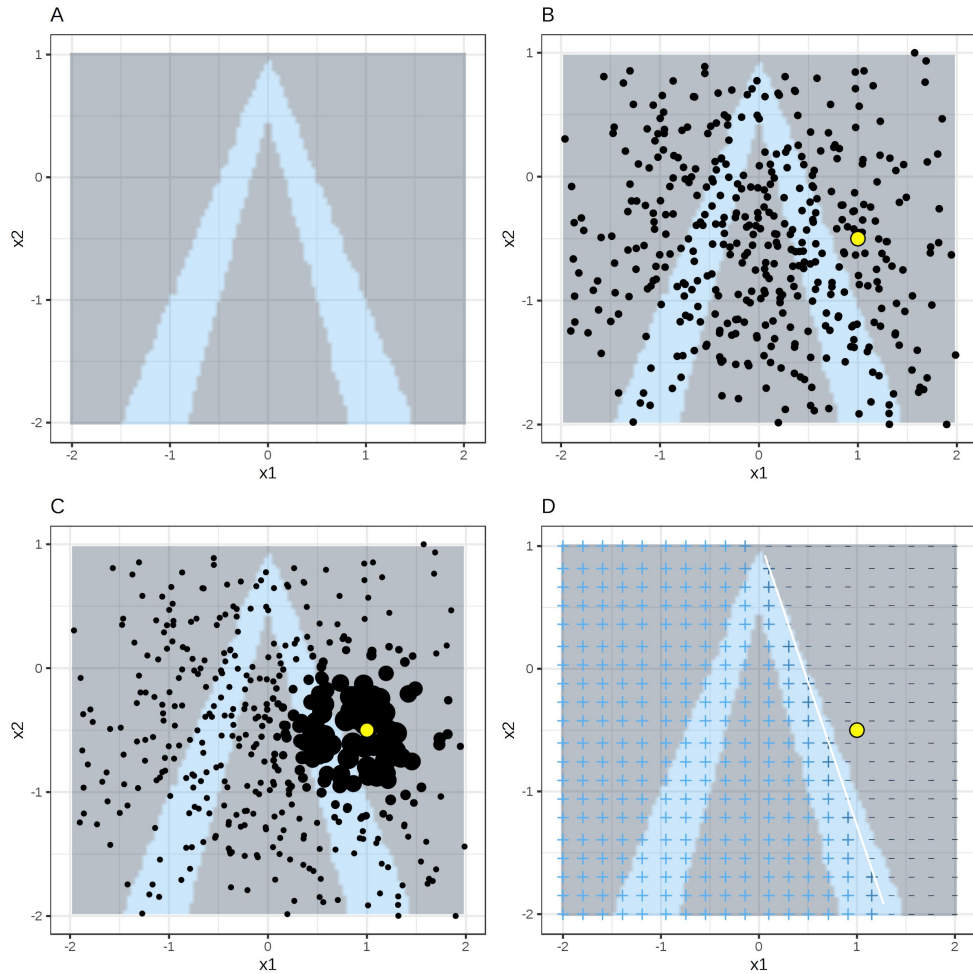
Permutation Feature Importance



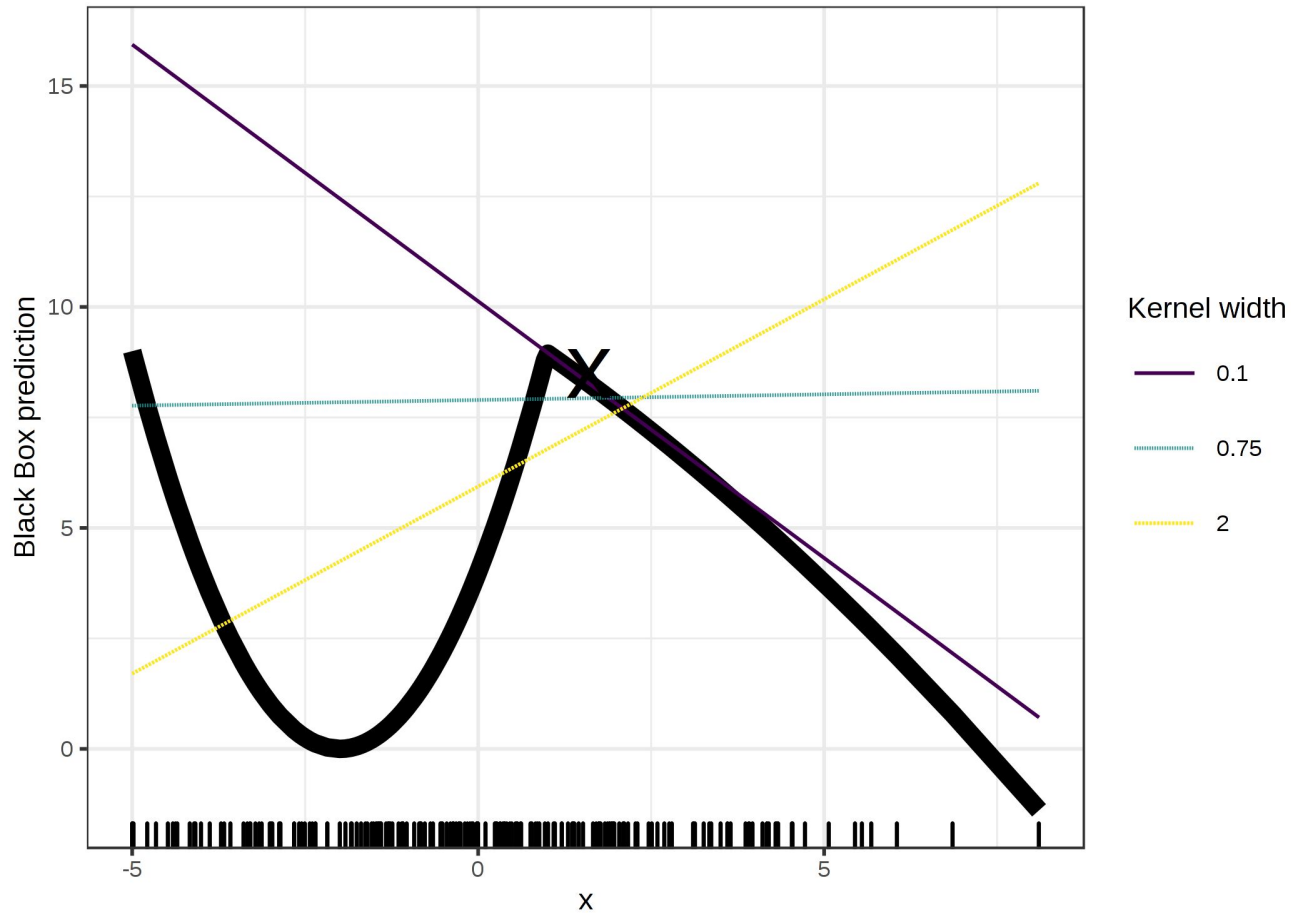
Global Surrogate



Local Surrogate (LIME)



Local Surrogate Kernel



LIME - Text Explainer (ELI5)

y=sci.med (probability 0.576, score 0.621) top features

Contribution?	Feature
+0.972	Highlighted in text (sum)
-0.351	<BIAS>

as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be broken up with sound, or they have to be extracted surgically. when i was in, the x-ray tech happened to mention that she'd had kidney stones and children, and the childbirth hurt less.



Shapley Values

The Shapley value is a solution concept in Cooperative Game Theory.





David and Jacob ate together and they
payed 70\$ in total.





How much they pay usually?

- If **David** is eating alone, he would pay **35**
- If **Jacob** is eating alone, he would pay **45**
-
- If **David** and **Jacob** both eat alone, they would pay **70**





We take all permutations of the 2 participants in sequence and see the incremental payout that each of them has to make.



Consider all Permutations

1. (David, Jacob) – (35, 35)
2. (Jacob, David) – (45, 25)

$$\text{David: } (35 + 25) / 2 = 30$$

$$\text{Jacob: } (45 + 35) / 2 = 40$$

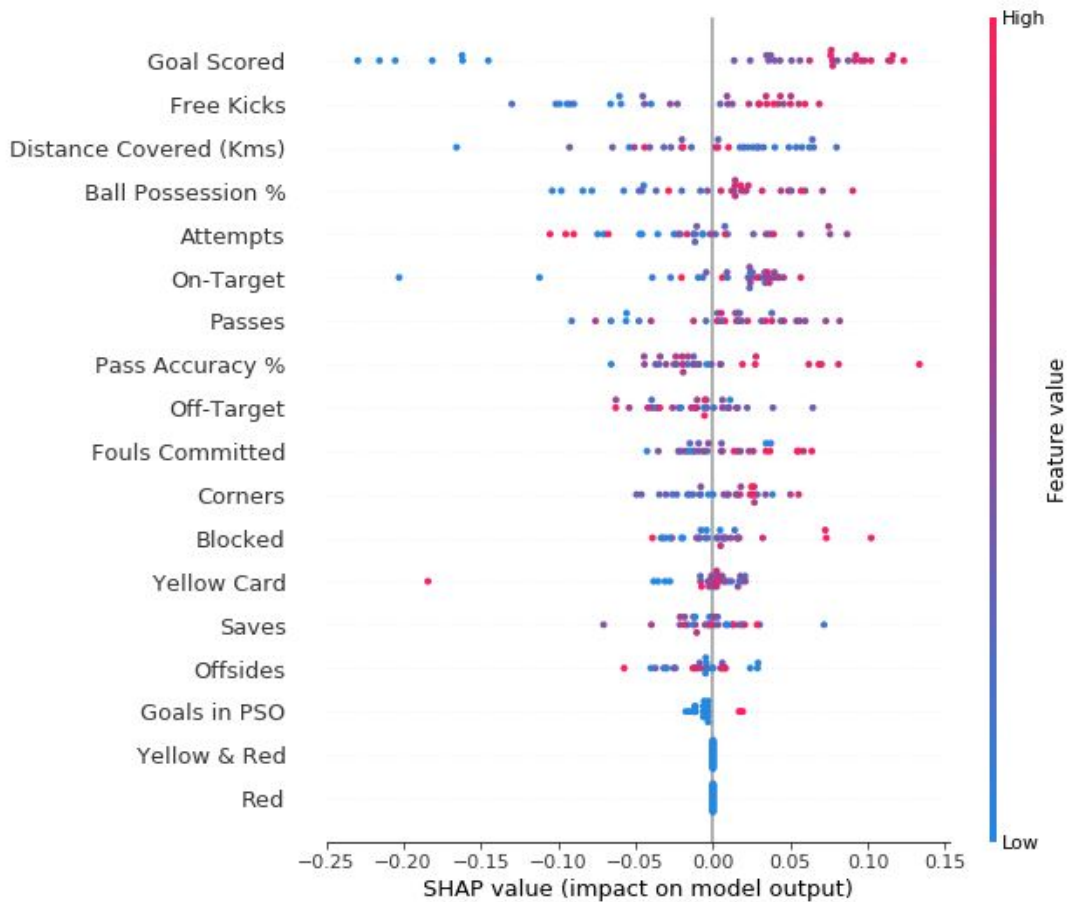


SHAP (Force Plot)

Why my prediction was different from baseline?





SHAP Summary Plot





7

The Future of Interpretability

- 
- The focus will be on model-agnostic interpretability tools.
 - Machine learning will be automated and, with it, interpretability.
 - We do not analyze data, we analyze models.
 - The data scientists will automate themselves.
 - Robots and programs will explain themselves.
 - Interpretability could boost machine intelligence research.
- 



Explainable AI^{BETA} by Google



AI Explanations


Receive a score explaining how each factor contributed to the final result of the model predictions.

What-If Tool

Investigate model performances for a range of features in your dataset, optimization strategies, and even manipulations to individual datapoint values using the What-If Tool integrated with AI Platform.

Continuous Evaluation

Sample the prediction from trained machine learning models deployed to AI Platform. Provide ground truth labels for prediction inputs using the continuous evaluation capability. Data Labeling Service compares model predictions with ground truth labels to help you improve model performance.



Interpretable Machine Learning

A Guide for Making
Black Box Models Explainable



@ChristophMolnar



Thanks!

Any questions?

You can find me at:

- ◇ anzor.gozalishvili@maxinai.com
- ◇ github.com/AnzorGozalishvili
- ◇ facebook.com/anzor.gozalishvili
- ◇ linkedin.com/in/anzor-gozalishvili-481967120/

