# Model Explainability

Robin Lester

# Agenda

- Interpret ML with Python
- Power BI for Data Science interpretability

# Loan Application Decisions

Create a model for loan application acceptance

**Azure Machine Learning**

Fairlearn

Is my model fair?

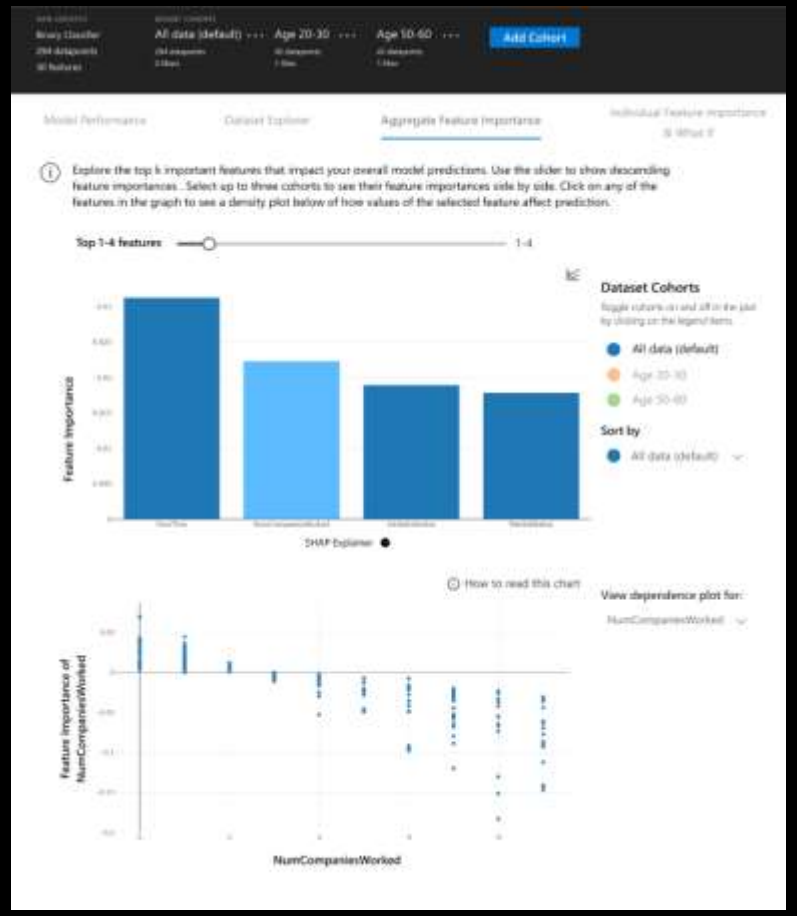
InterpretML

How does it decide who to accept or reject?

# Interpretability

# Understand and debug your model

InterpretML




## Interpret

Glassbox and blackbox interpretability methods for tabular data

### Blackbox models:

**Model formats:**
Python models using scickit predict convention, Scikit, Tensorflow, Pytorch, Keras

**Explainers:**
SHAP, LIME, Global Surrogate, Feature Permutation

## Interpret-community

Additional interpretability techniques for tabular data

### Glassbox Models:

**Model types:**
Linear Models, Decision Trees, Decision Rules, Explainable Boosting Machines

## Interpret-text

Interpretability methods for text data

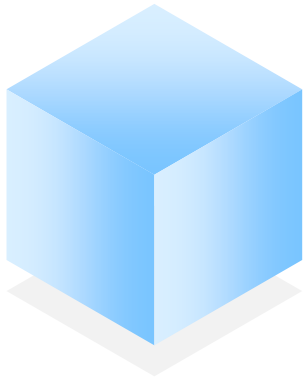
## DiCE

Diverse Counterfactual Explanations

## AzurML-interpret

AzureML SDK wrapper for Interpret and Interpret-community

https://github.com/interpretml



## Glassbox models

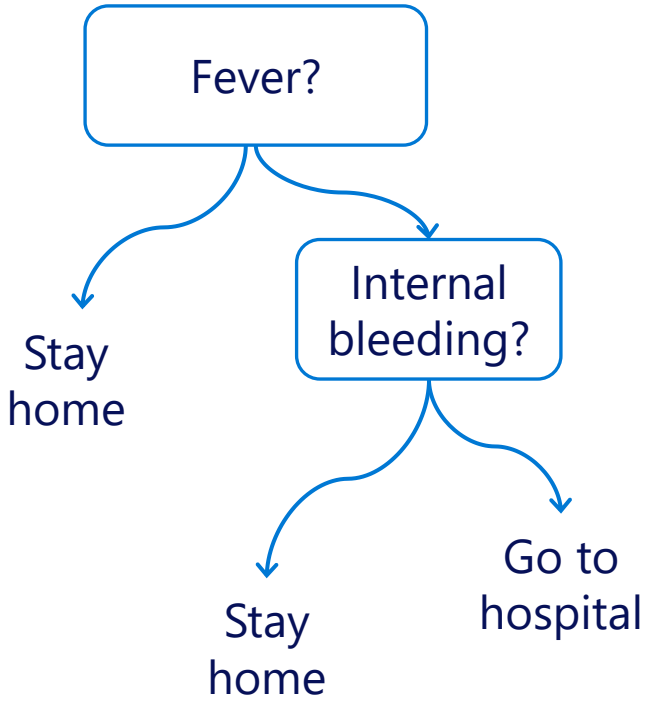
**Models designed to be interpretable. Lossless explainability.**

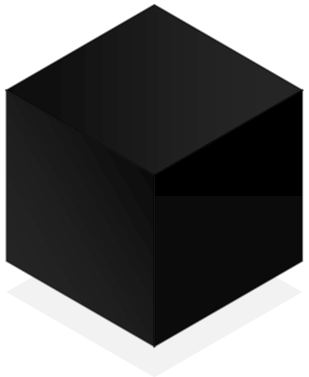
Decision trees

Rule lists

Linear models

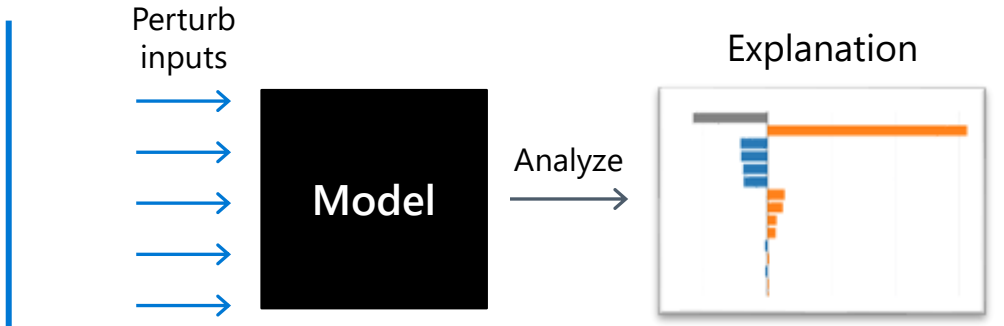
Explainable Boosting Machines





## Blackbox explanations

# Explain *any* ML system. Approximate explainability.




Shap

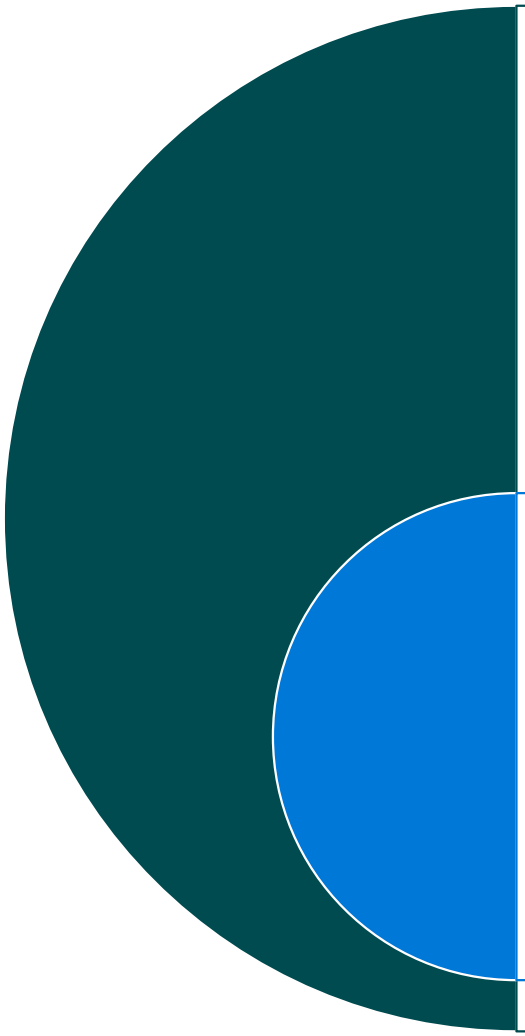
Lime

Partial dependence

Sensitivity analysis

Black-box explainers analyze the relationship between input features and output predictions to interpret models

# InterpretML



## Global point of view

- Feature importance across the global model

## Local point of view

- The marginal contribution of each feature in the data on that instance of prediction

# Local point of view

- SHapley Additive exPlanation
  - Model agnostic
  - Data agnostic

$$y = f(x_1, x_2) = 2x_1 + 3x_2$$

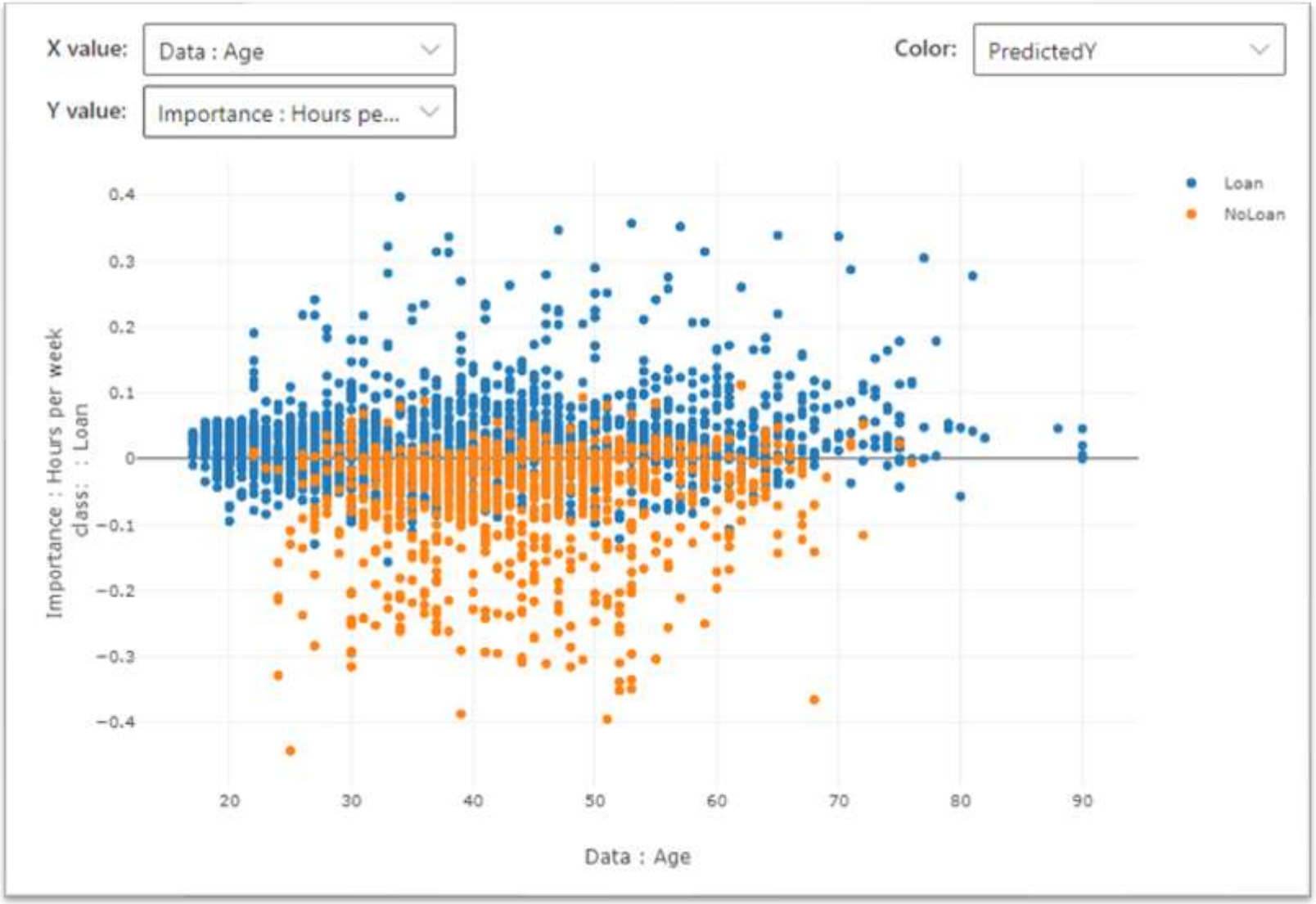
| label | X1 | X2 | IX1 | IX2 |
|---|---|---|---|---|
| 1 | -1 | 1 | -2 | 3 |
| 14 | 1 | 4 | 2 | 12 |
| 1 | 2 | -1 | 4 | -3 |
| 11 | 4 | 1 | 8 | 3 |
| 4 | -1 | 2 | -2 | 6 |
| -1 | 1 | -1 | 2 | -3 |

Contribution of X1 to label is 2 times
Contribution of X1 to label is 3 times

# Demo Interpretability

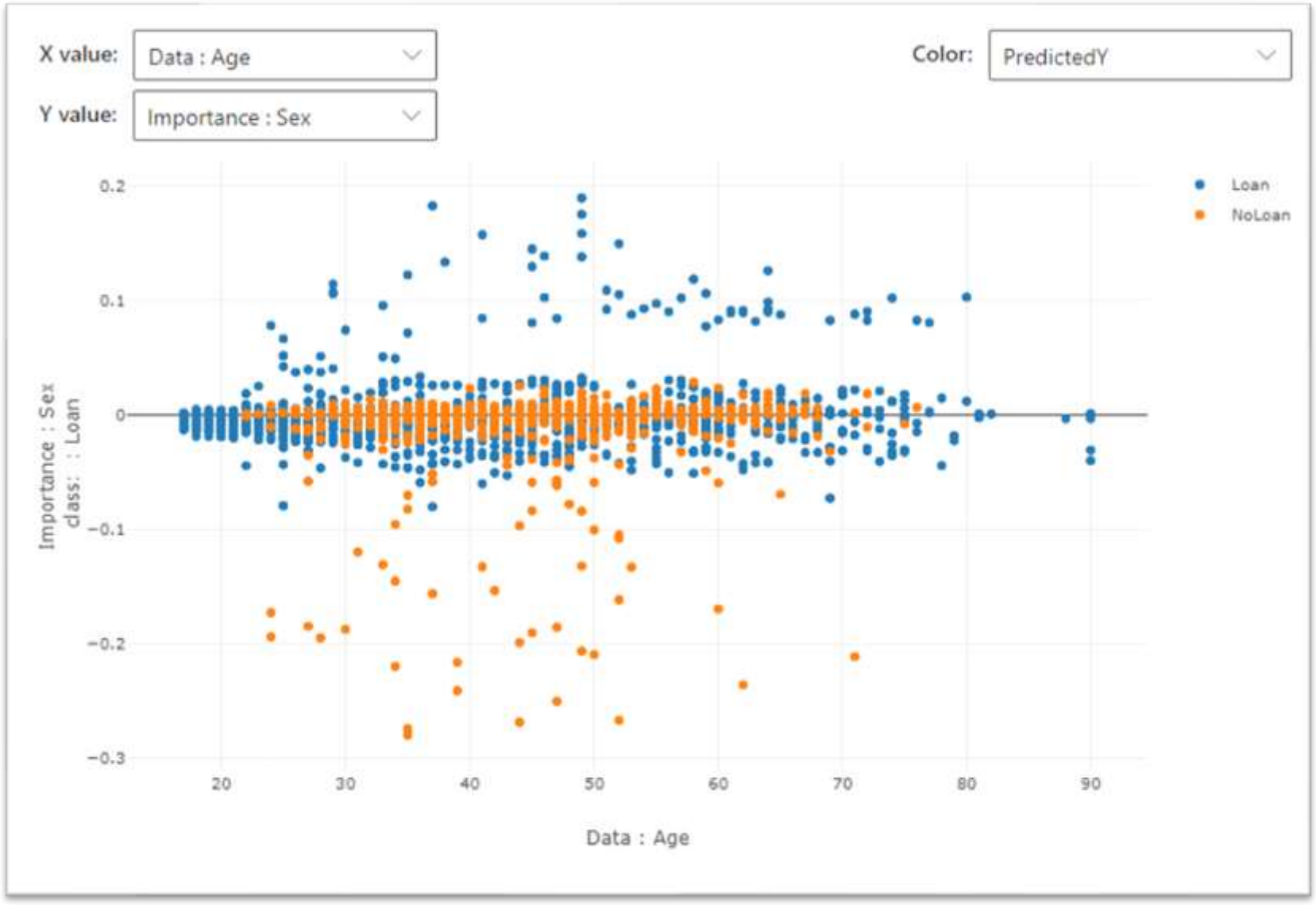
# SHAP importance of ‘Hours Per Week’ against actual age

- The importance of hours per week is less if the age is young or old

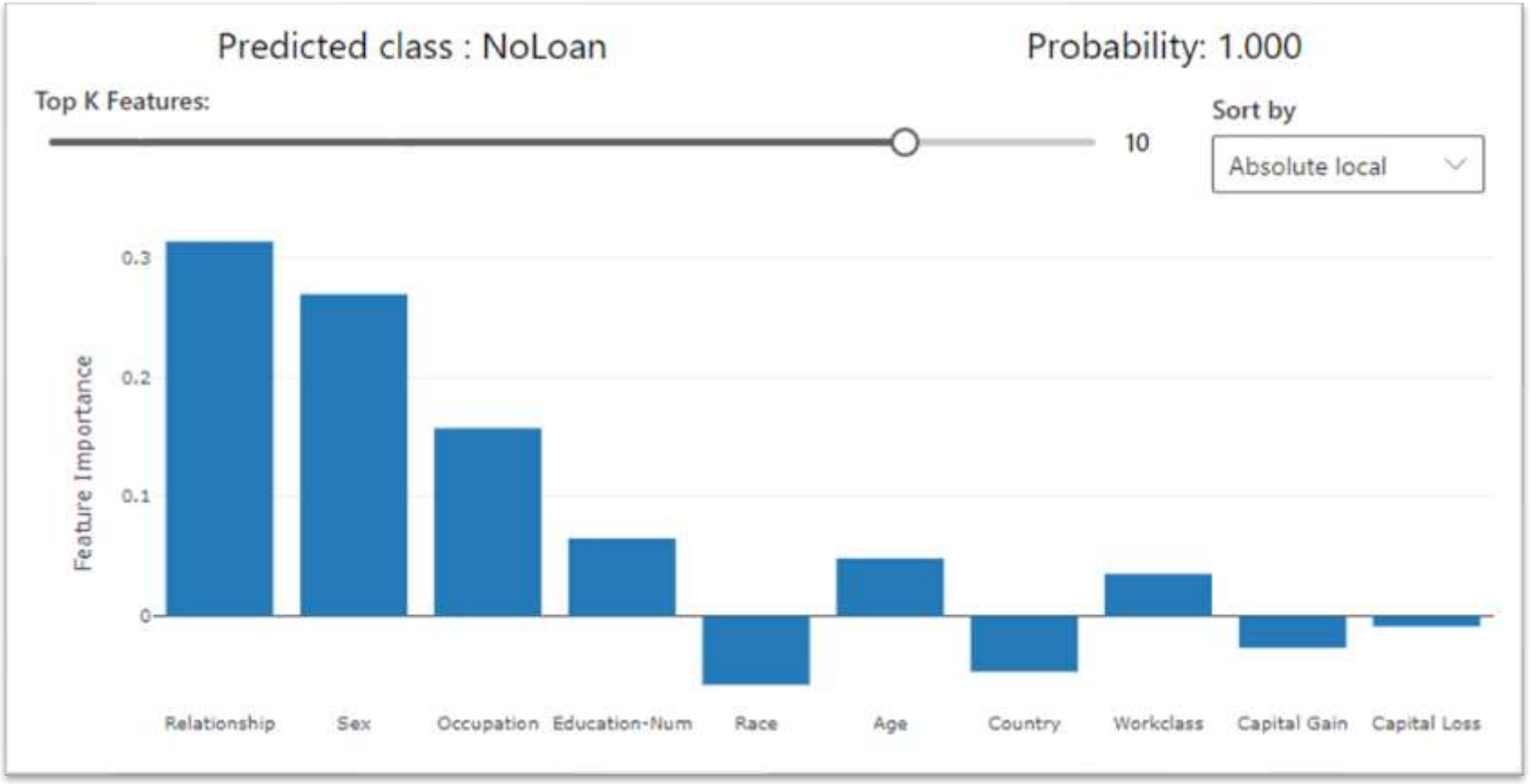


# SHAP importance of ‘Sex’ against actual age

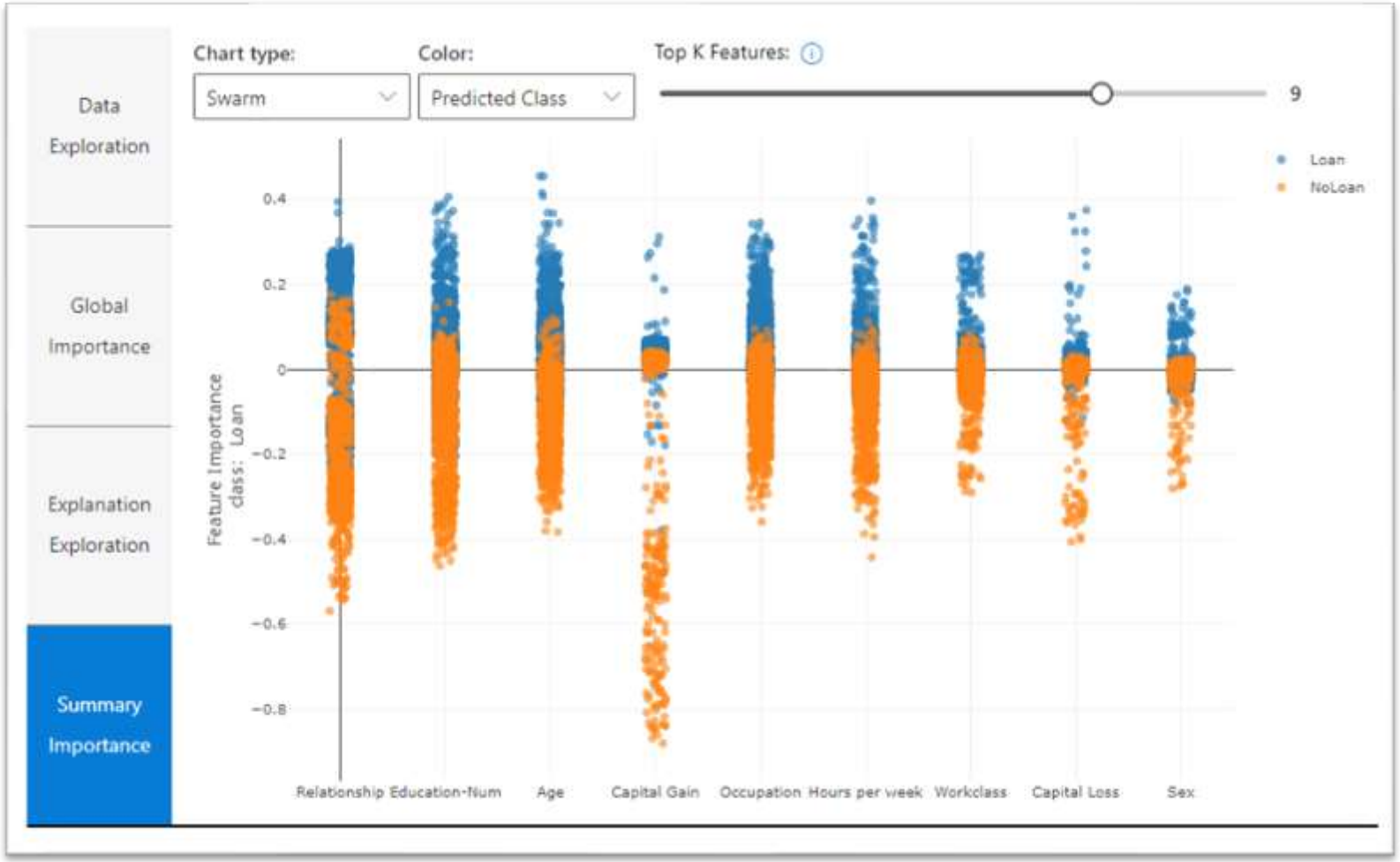
A person's gender is influencing this model showing bias



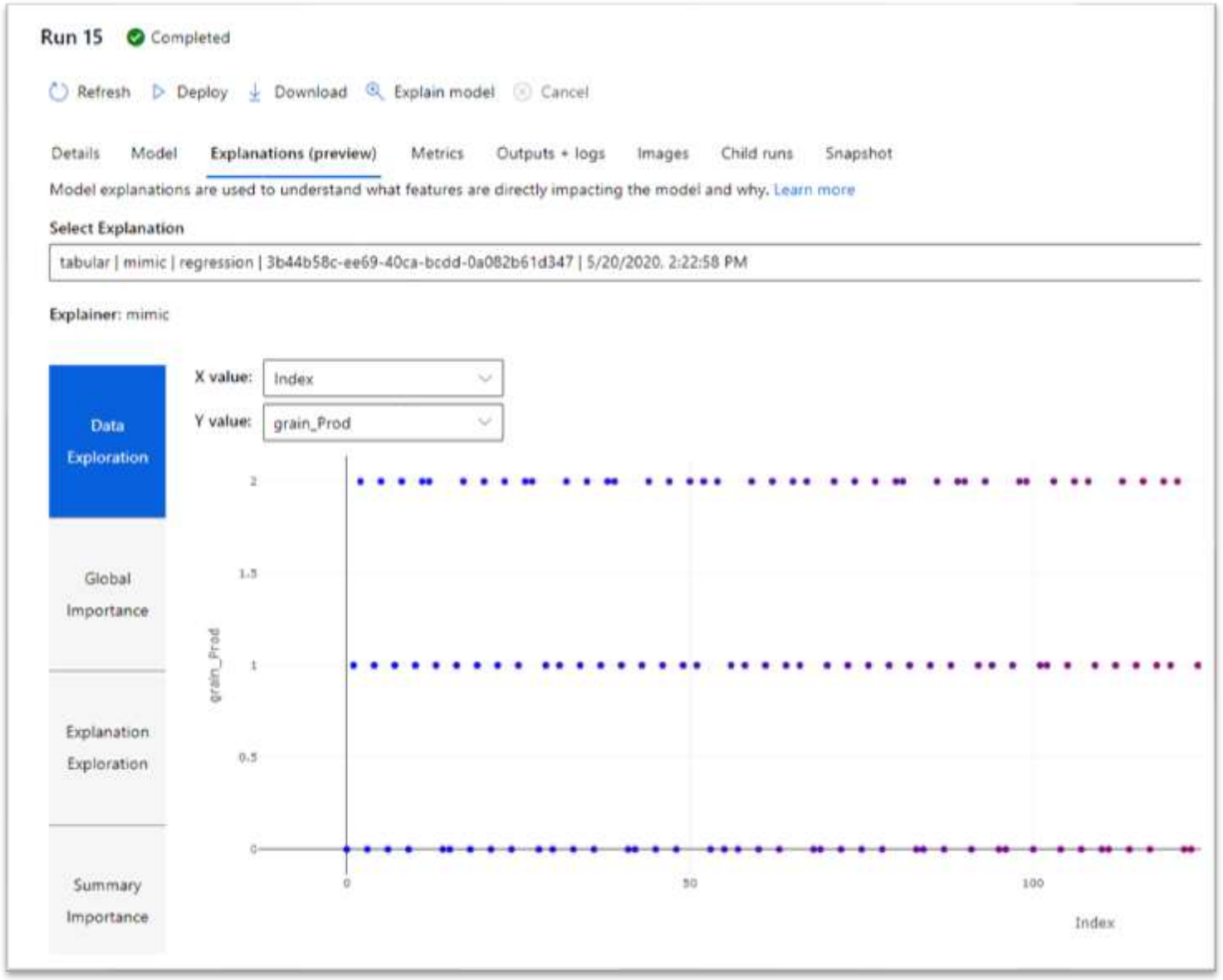
- This female was denied a loan primarily because of her relationship and gender



- Capital Gain can have a big impact on not getting a loan
- ‘Sex’ is influencing loans



# View for experiments



# Power BI for model explainablity and key influencers

**http://aka.ms/powerbiaiworkshop**


Microsoft