Microsoft Azure

Model Explainability and Responsible AI

with Microsoft Azure Machine Learning Services

Jon Tupitza
Sr. Cloud Solution Architect – Data & Analytics – Azure Data Scientist
Jon.Tupitza@Microsoft.com

The Problem: *Working in the Dark*

Machines Learn in the Darkness and Data Scientists Struggle in the Void to Explain Them!

Why Should I Trust You?

- Why does my model predict a particular outcome?
- How can I ensure that the prediction(s) my model made are correct?
- What evidence could I gather to justify my model's prediction(s)?
- How can I find what caused the error if the prediction(s) are ***not*** correct?
- How would my model's predictions change if I altered its input?

The Solution: Let there be Light!

Data Scientists Must be Empowered to Build Interpretable, Explainable and Ethical AI Systems

Responsible AI Principles:

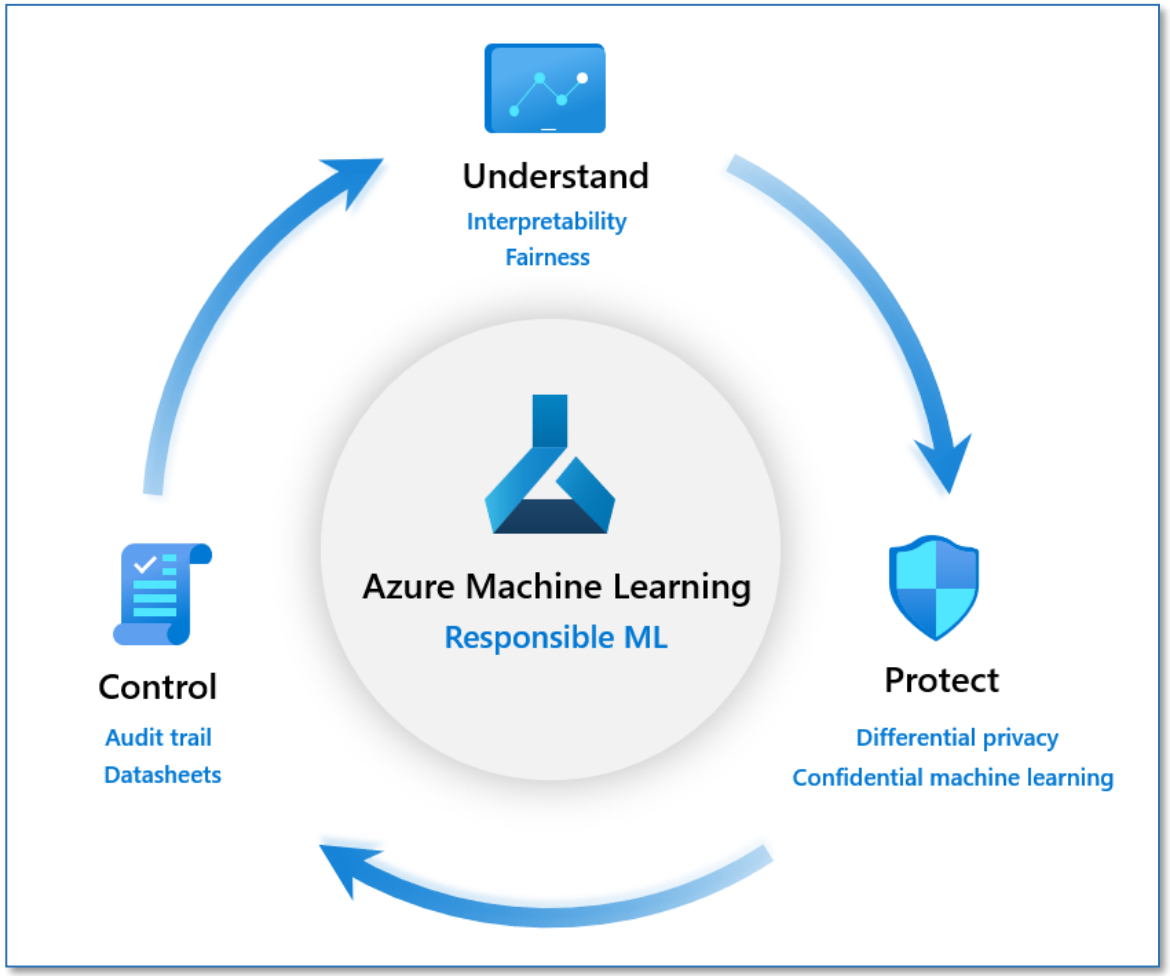
- **Transparency:** AI must be understandable
- **Fairness:** AI must treat people fairly; not reinforce biases or stereotypes
- **Inclusiveness:** AI must not restrict access to opportunities or resources
- **Accountability:** AI must be responsible for its inferences (predictions)
- **Privacy & Security:** AI systems must safeguard all people and their data

Responsible AI: Ethical & Trustworthy AI Systems

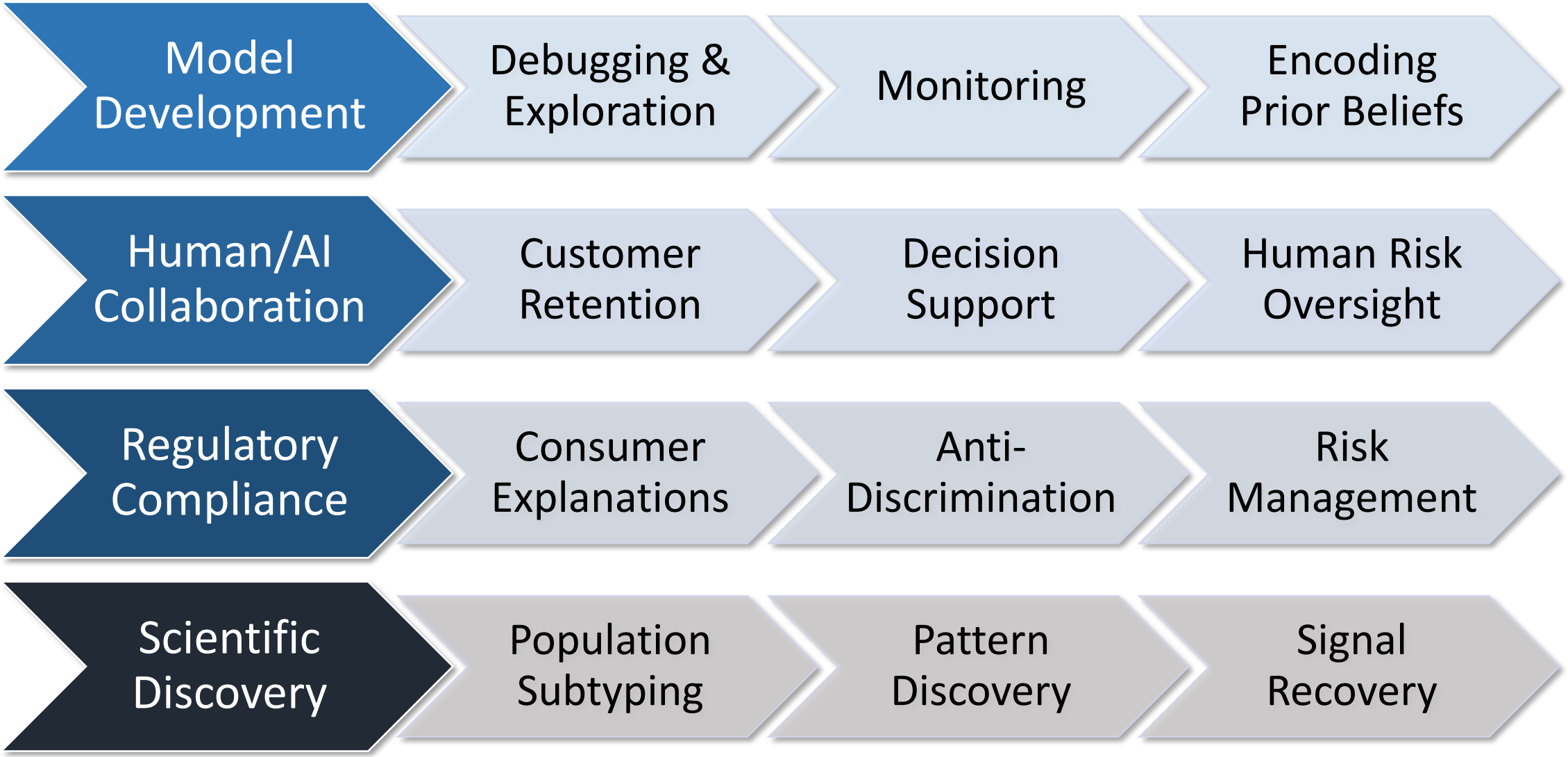
Mitigate Inequity by "Seeing Into" ML Models to Explain How & Why Decisions Were Made

- **Bias in AI Systems Can Result in Unintended Consequences:**
  - Withholding Opportunities, Resources or Information from Groups and/or Individuals
  - Reinforcing Inequity & Stereotypes

- Understand Machine Learning Models:
  - Interpret & Explain Model Behavior
  - Assess & Mitigate Model Unfairness
- Protect People and Their Data
  - Prevent Data Exposure
  - Work with Encrypted Data
- Control the End-to-End ML Process
  - Document the Machine Learning Lifecycle



Explainable AI: Motivations/Practical Applications



Interpretability & Explainability

Understanding ML Models and Their Predictions

Interpretability & Explainability: What's the Difference

Interpreting ML Models versus Explaining How & Why Decisions Were Made

Interpretability

Pertains to the Model
(Global understanding of the Predictor)

- Inspecting a ML model's implementation to understand the decisions it makes, and the algorithms it uses, to comprehend the rationale behind its predictions.
- Humans are more likely to trust a system that justifies its decisions.
- The goal of the model's **creator/owner** is promoting trust in its capabilities.

Explainability

Pertains to the Inferences
(Local understanding of Individual Predictions)

- A post-hoc analysis of a machine learning model's predictions for the sake of explaining why it made the decisions that lead to those inferences.
- The degree to which humans can identify the rationale behind a system's decisions.
- The goal of the model's **subject** is to understand the prediction or behavior.

Interpretability & Explainability: Use Cases

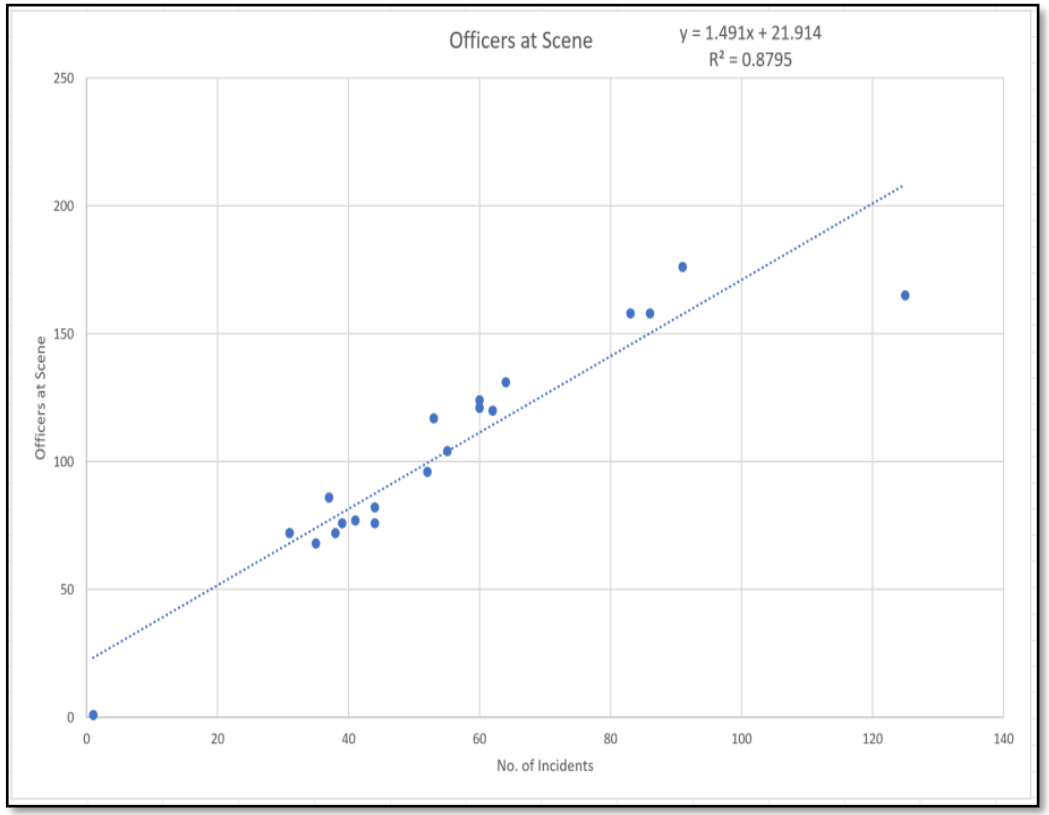
Interpret Machine Learning Models and Explain How & Why Their Decisions Were Made

Explain Predictions	• Build trust by justifying the model's predictions • Assess and mitigate unfair decisions
Assess Model Quality	• Evaluating models to validate goodness-of-fit • Model Development and Debugging
Identify the Most Important Features	• Which features had the most influence on predictions? • How did each feature influence each individual prediction?
Identify Feature Relationships	• What feature interactions had the greatest influence on each prediction?

Machine Learning Models: Basic Algorithms

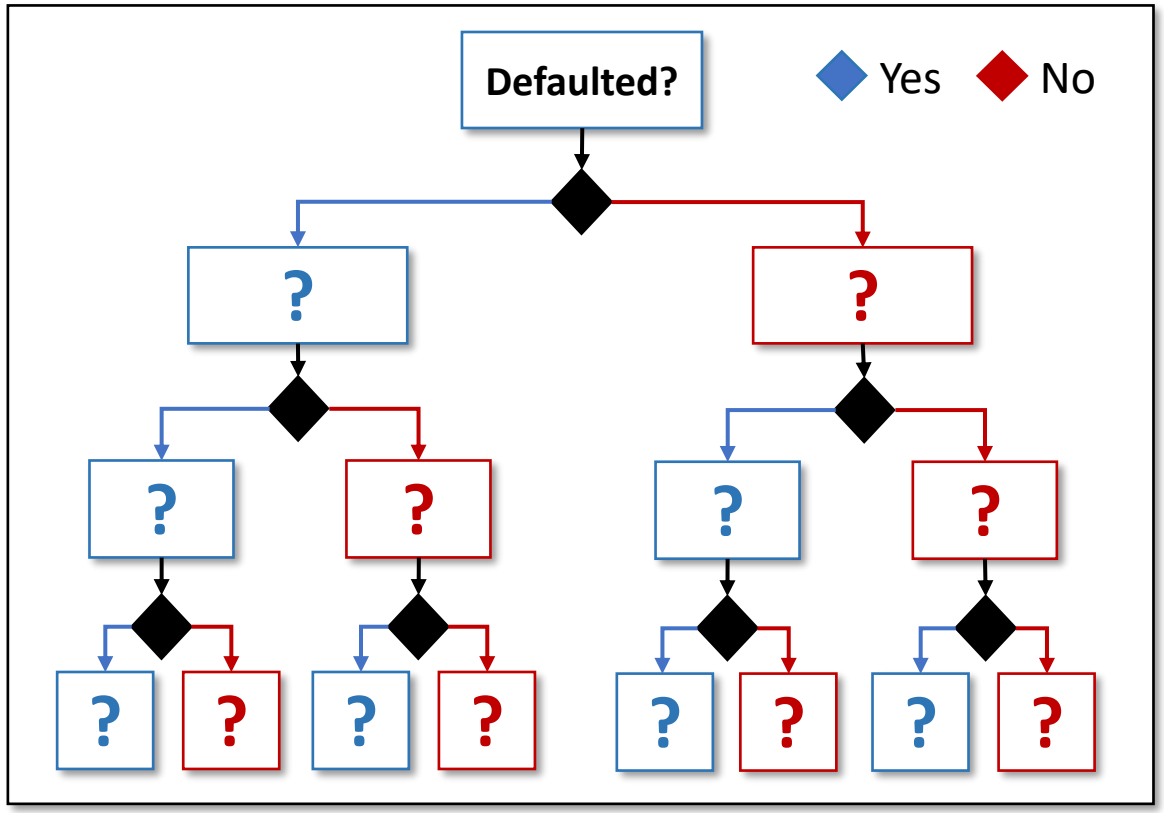
The Type of Question Often Drives the Choice of Modelling Algorithms

Linear Models (e.g., Regression)



How do X and y Correlate?

Tree-Based Models (e.g., Classification)



The outcome of a series of decisions?

Machine Learning Models: Glass Box vs. Black Box

The Choice of Modelling Algorithm Can Profoundly Impact the "Transparency" of a Model

Simple Algorithms Make it Easier to Understand & Interpret Models and to Explain Their Inferences

Complex Algorithms Make it Difficult to Understand & Interpret Models and to Explain Their Inferences

"Glass Box" Models:

- Linear & Logistic Regression
- Decision Trees
- Naïve-Bayes
- K Nearest Neighbors (KNN)

"Black Box" Models:

- Neural Networks
- Support Vector Machines
- Random Forests
- Ensembles

"The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made."

- Christoph Molnar

Interpretability: Intrinsic vs. Post-Hoc Methods

How is ML Model Interpretability Achieved: Before or After Training?

- **Intrinsic:**
  - Typically associated with **Glass-Box models**
  - Restricting the complexity of the machine learning model itself
  - Achieved by way of the models' inherent level of interpretability

- **Post-Hoc:**
  - Typically associated with **Black-Box models**
  - Application of techniques after the model has been trained
  - Achieved by implementing model-agnostic post-hoc explainers

Packages: Interpret, Explain, Mitigate Bias, Protect

Microsoft Azure's Machine Learning Tools for Implementing Responsible AI

InterpretML

- Inspect feature interactions
- Measure feature importance and achieve effective feature selection

FairLearn

- **Build trust** by justifying the model's predictions
- Assess & Mitigate Unfair Decisions
- Identify & Mitigate Unintended Bias

SmartNoise

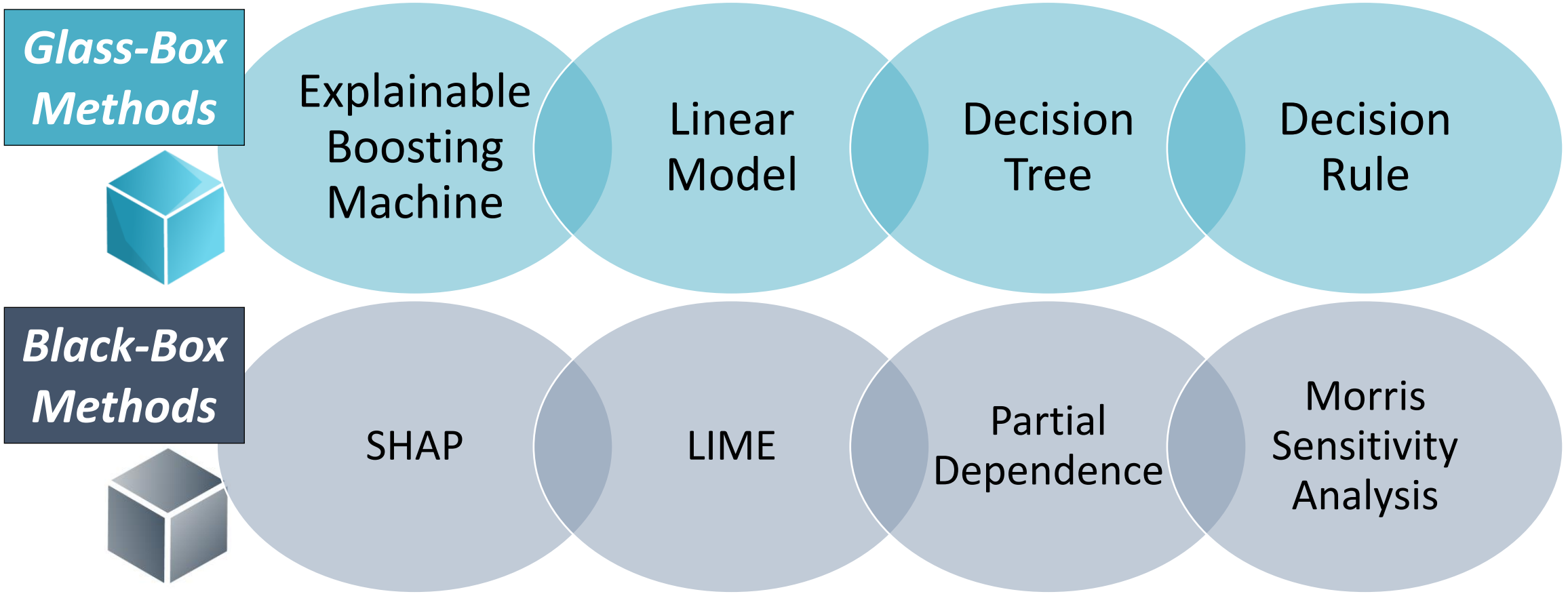
- Implements **differential privacy** by adding noise (randomness) to the data to obfuscate the identifiability of any specific feature observations (tuples).

SEAL

- Implements **homomorphic encryption** to enable training models against encrypted data without accessing the decryption (secret) key.

Interpretability Packages: InterpretML

Microsoft Azure's Machine Learning Tools for Implementing Responsible AI



https://github.com/interpretml/interpret/blob/master/README.md

InterpretML: A Unified Framework for Machine Learning Interpretability" (H. Nori, S. Jenkins, P. Koch, and R. Caruana 2019)

InterpretML: Explainable Boosting Machine

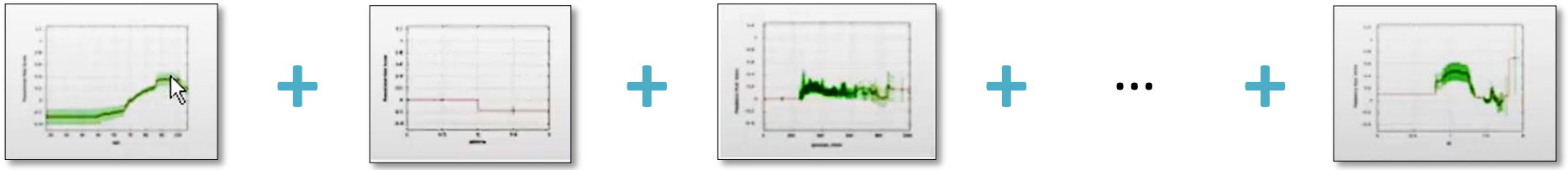
Microsoft Azure's Machine Learning Tools for Implementing Responsible AI

- An intrinsically interpretable (glass-box) model developed at Microsoft Research
- A tree-based, cyclic gradient boosting Generalized Additive Model (GAN)
- Features automatic interaction detection
- As accurate as state-of-the-art techniques like random forests and gradient boosted machines
- Unlike black-box models, EBMs produce lossless explanations
- Extremely compact and fast at prediction time (i.e., memory & CPU efficient)

InterpretML: Explainable Boosting Machine

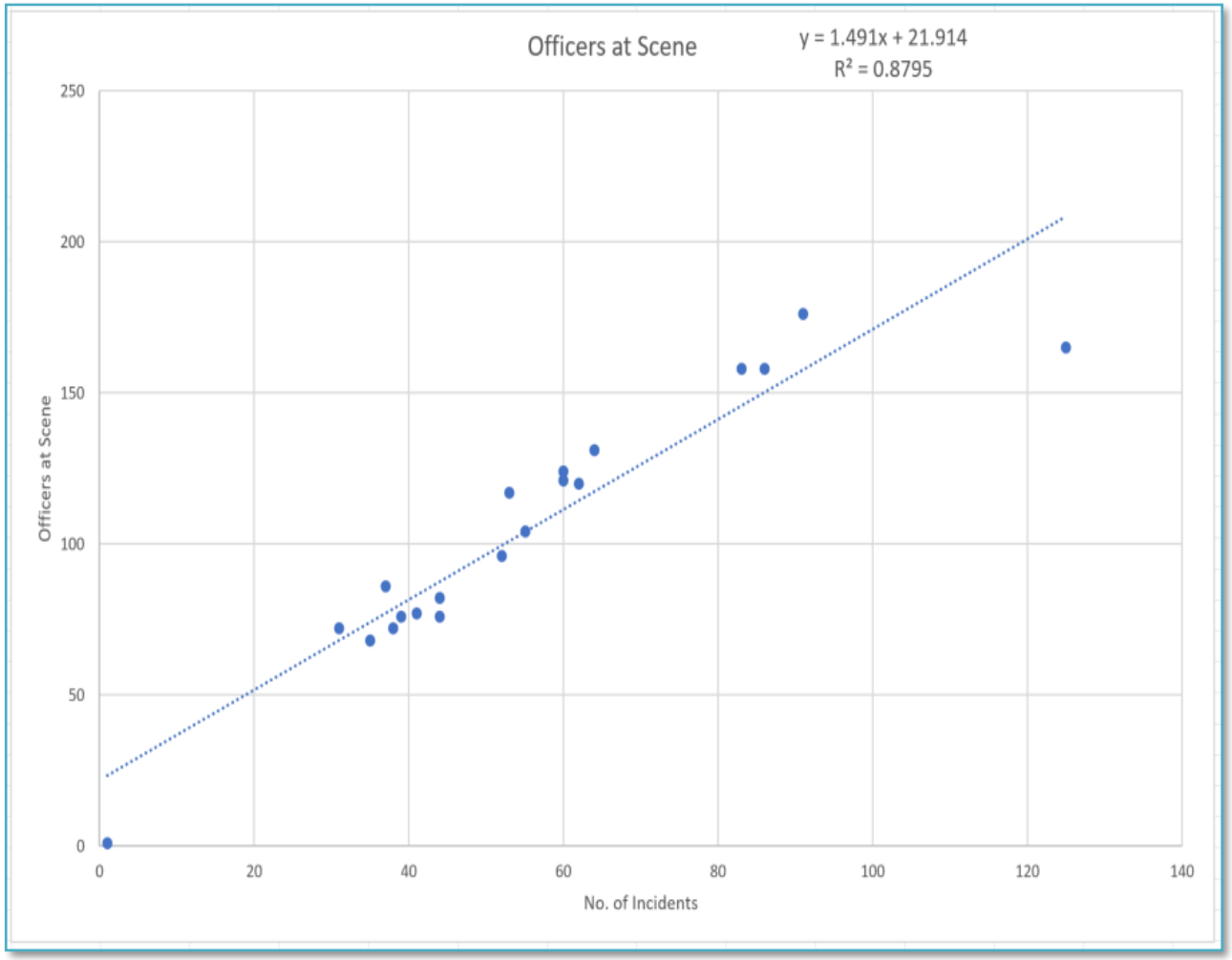
Microsoft Azure's Machine Learning Tools for Implementing Responsible AI

Iteration	$Feature_1$	$Feature_2$	$Feature_3$	...	$Feature_n$
1					
2					
3					
4					
5					
...					
10,000					



InterpretML: Linear Model

Linear Models are So Interpretable Because They Model Linear Relationships



- Linear models can be used to model the dependence of a regression target **y** on some features **X**
- The predicted outcome of a linear model is equal to the weighted sum of its input features

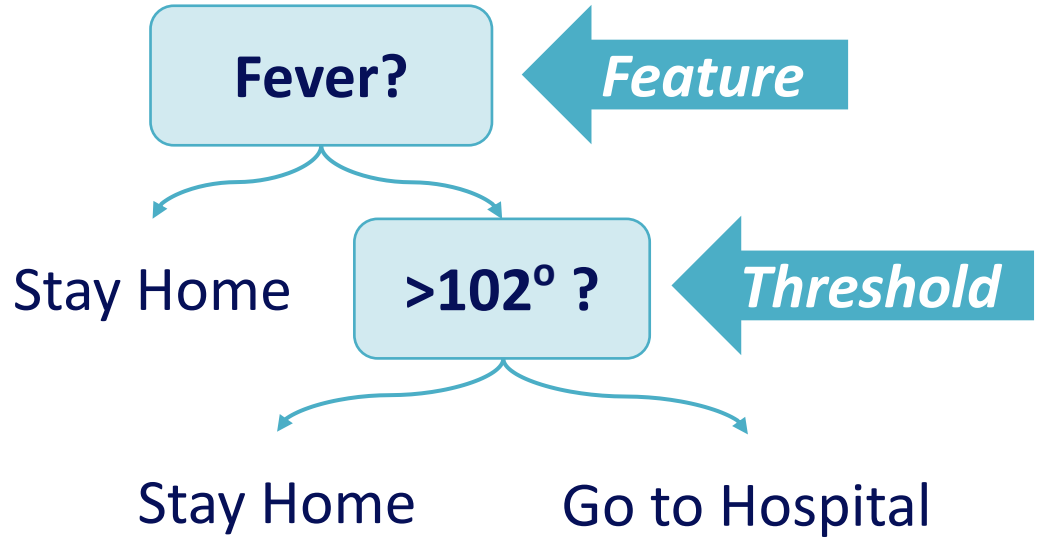
$$y = \beta 0 + \beta 1\, X1 + \ldots + \beta p\, Xp + \epsilon$$

InterpretML: Decision Tree & Decision Rule

Decision Trees and Rules are Perhaps the Most Interpretable Algorithms

Decision Tree

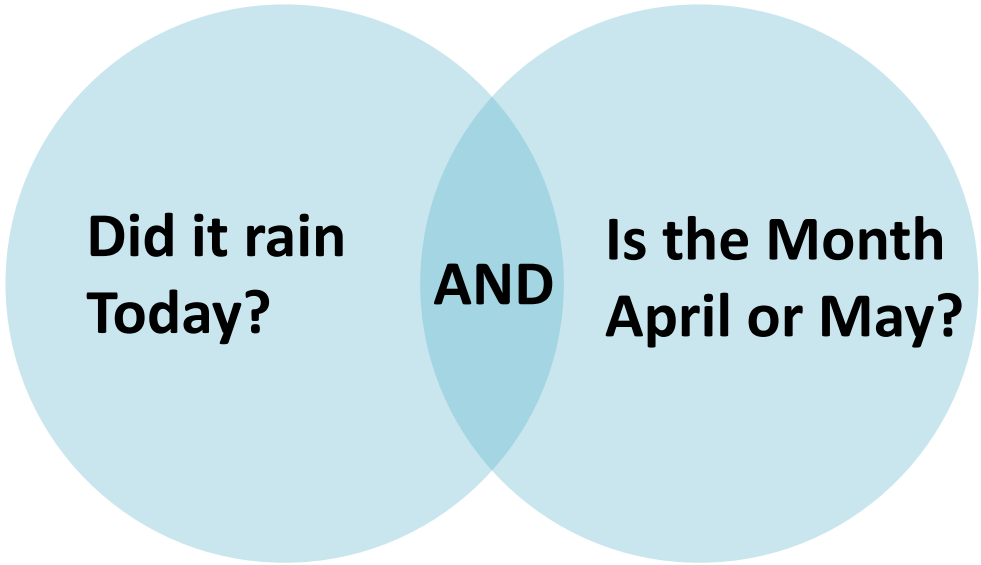
- Creates multiple subsets by splitting the data along threshold values (cutoffs) in influential features
- Performs well where non-linear feature-to-target relationships, and/or feature interactions, exist.



Decision Rule

- Simple IF > THEN statement consisting of a condition (antecedent) and a prediction

Question: Will it Rain Tomorrow? **Yes**

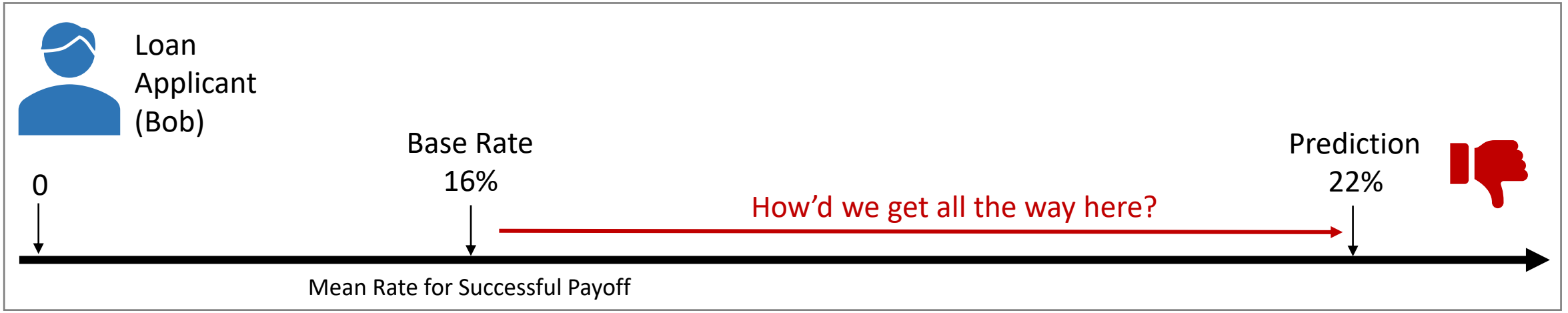


InterpretML: SHAP

Shapley Additive Explanations

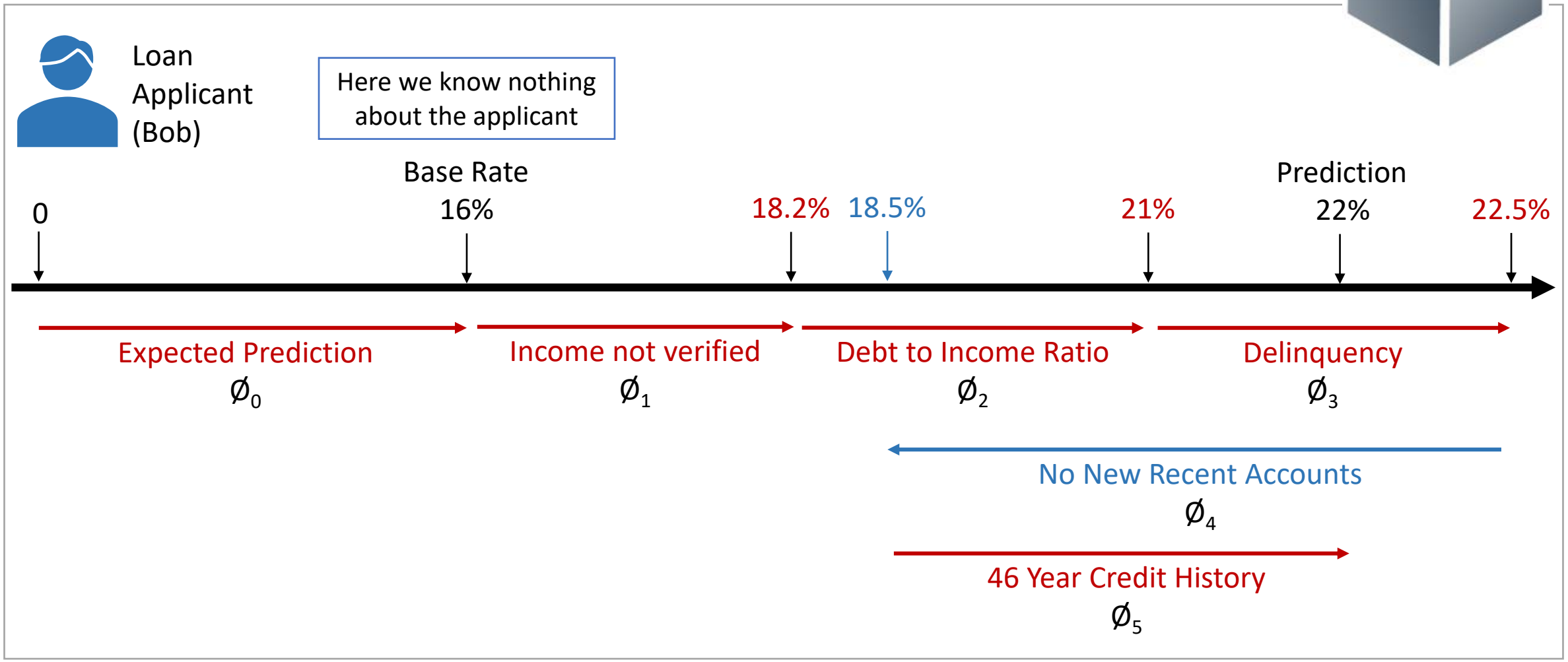
- Based on game theory pioneered by Lloyd Shapley
- Model-Agnostic, but more efficient on specific classes (e.g., Tree Ensembles)
- Measures how much each feature contributed to the prediction as compared to the average prediction

A prediction can be explained by assuming each feature value is a “player” in a game where the prediction is the payout. Shapley values determine how to fairly distribute the payout among the features.



InterpretML: SHAP

Shapley Additive Explanations



InterpretML: LIME

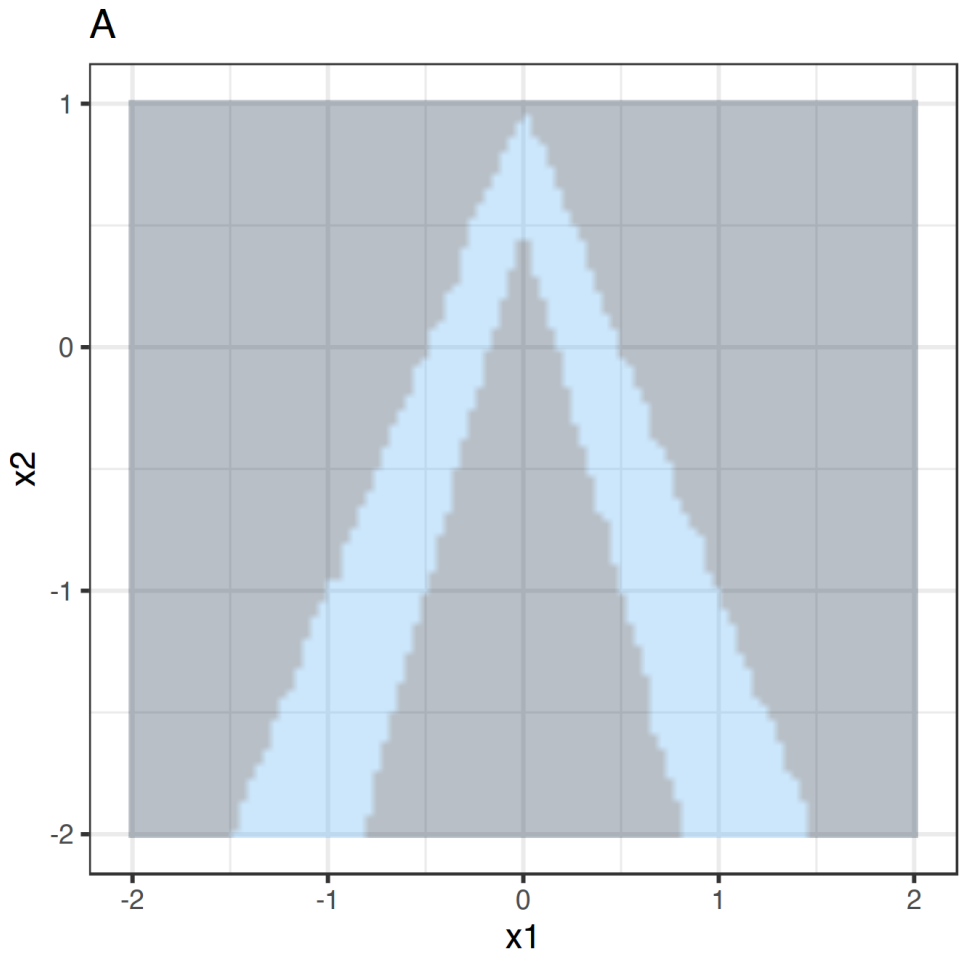
Local Interpretable Model-Agnostic Explanations

- LIME interprets black-box models using local surrogate models
  - Local surrogate models are easily interpretable models that are used to explain the individual predictions of black-box machine learning models.
  - Even a linear model can provide good approximation of black-box model behavior
- LIME fits a simpler (glass-box) model around the local neighborhood of the prediction; focusing on a narrow (local) decision space.
- LIME works by perturbing any individual datapoint and generating synthetic data which gets evaluated by the black-box system and is ultimately used as a training set for the glass-box (surrogate) model.
- LIME is useful for avoiding spurious correlations
- LIME enables interpreting an explanation the same way one would reason about a linear model

InterpretML: Local Interpretable Model-agnostic Explanations by Marco Tulio Ribeiro of Microsoft Research

InterpretML: LIME

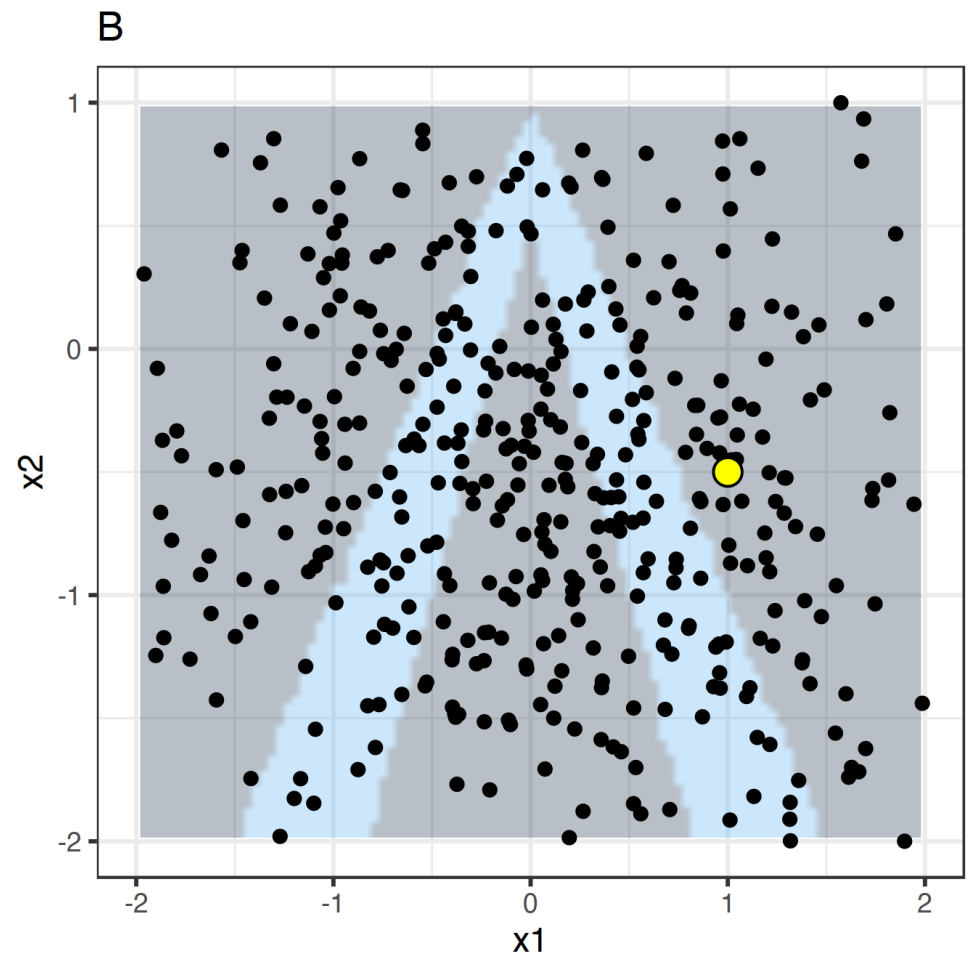
Local Interpretable Model-Agnostic Explanations



- Random Forest model trained for Binary Classification using tabular data.
- Predicted Classes:
  - 1 (Dark)
  - 0 (Light)
- Predictions given for features X_1 and X_2

InterpretML: LIME

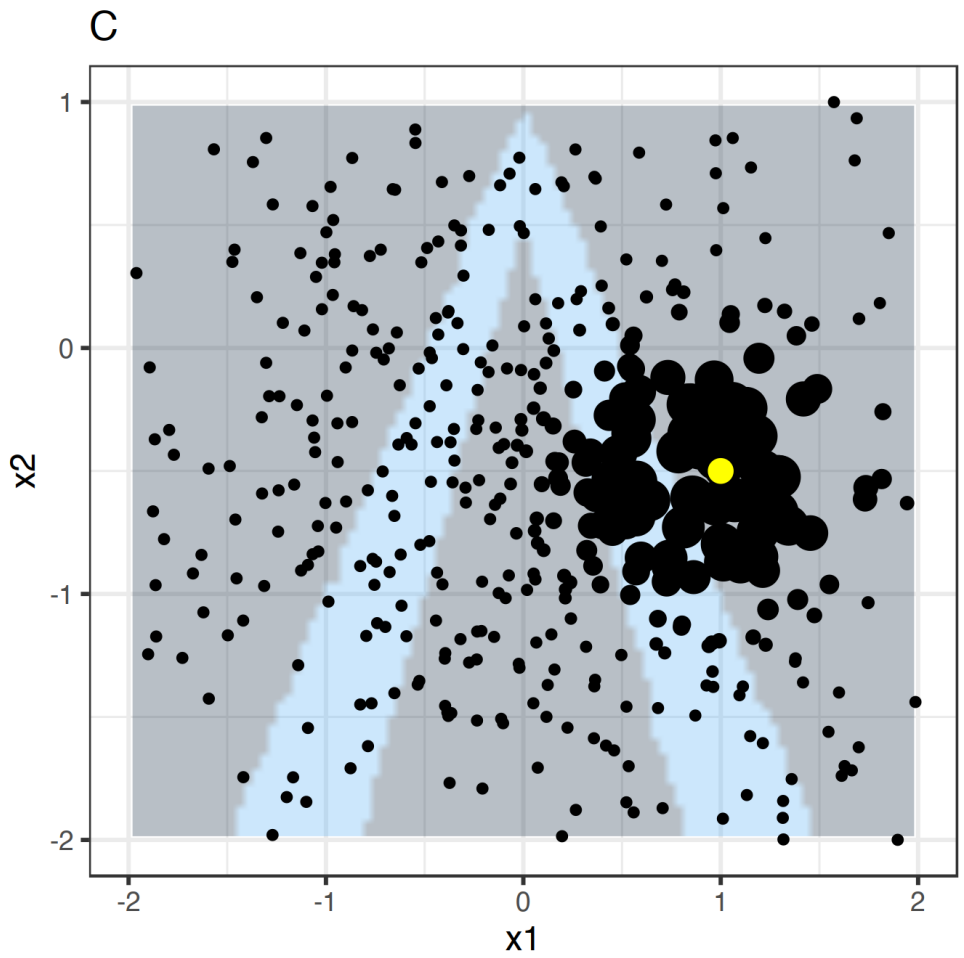
Local Interpretable Model-Agnostic Explanations



- Random Forest model trained for Binary Classification using tabular data.
- Predicted Classes:
  - 1 (Dark)
  - 0 (Light)
- Instance of Interest:
  - *Big Yellow Point*
- Data sampled from a Normal Distribution:
  - *Small Black Points*

InterpretML: LIME

Local Interpretable Model-Agnostic Explanations

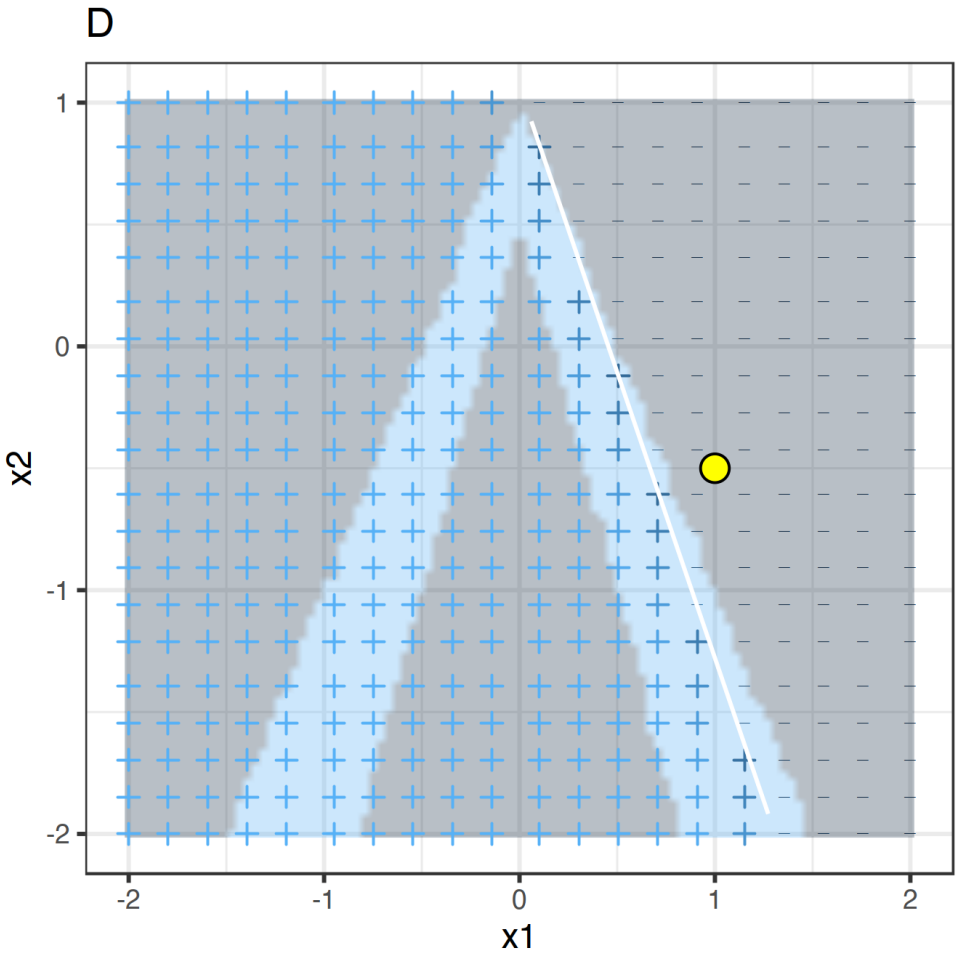


- Random Forest model trained for Binary Classification using tabular data.
- Predicted Classes:
  - 1 (Dark)
  - 0 (Light)

- LIME assigns a higher weight to points that are near the instance of interest

InterpretML: LIME

Local Interpretable Model-Agnostic Explanations



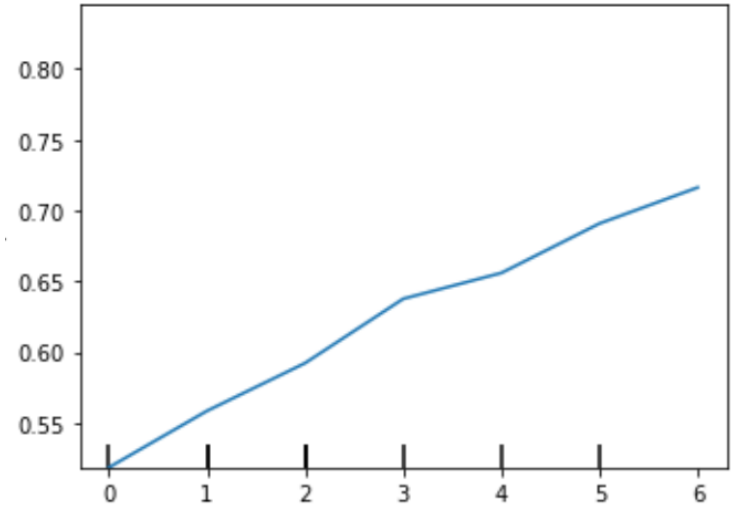
- Random Forest model trained for Binary Classification using tabular data.
- Predicted Classes:
  - 1 (Dark)
  - 0 (Light)
- Signs on the grid show the classifications of the locally learned model from the weighted samples
- The White Line marks the decision boundary (P(class=1) = 0.5)

InterpretML: Partial Dependence Plots

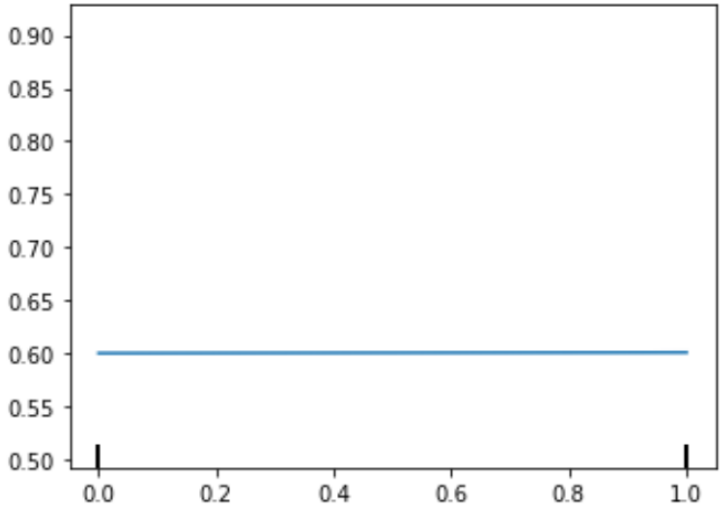
Explaining the Feature Observations that Drove Individual Predictions for an Entire Model

- Illustrates the marginal effect that one or two features have on the predicted outcome of a machine learning model (J.H. Friedman 2001).
- Visualizes the dependence between the target and as set of (one or two) features; revealing if it is linear, monotonic, or more complex.

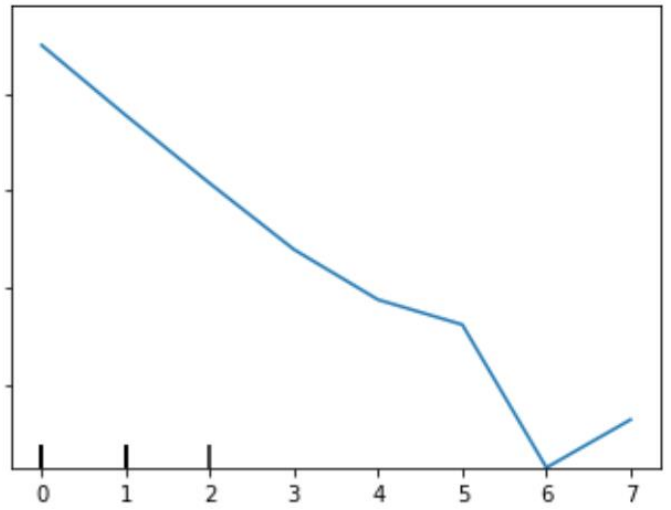
Linear Relationship



Monotonic Relationship



Complex Relationship



Feature Influence: Partial Dependence

A Univariate Method: The degree of each feature's influence is measured before modeling

Iteratively sets all observations in the column to each unique value contained in that column.

Then

Then observe the correlation each value of the column has to the response variable (target).

Original Values

Loan	Income	IsBad
$11,200	$108,000	False
$10,000	$65,000	False
$8,000	$20,000	True
$16,000	$110,000	True
$4,000	$155,000	False

All Values Set to $20k

Loan	Income	IsBad
$11,200	**$20,000**	False
$10,000	**$20,000**	False
$8,000	**$20,000**	True
$16,000	**$20,000**	True
$4,000	**$20,000**	False

All Values Set to $65k

Loan	Income	IsBad
$11,200	**$65,000**	False
$10,000	**$65,000**	False
$8,000	**$65,000**	True
$16,000	**$65,000**	True
$4,000	**$65,000**	False

InterpretML: MSA

Morris Sensitivity Analysis

- Ranks the importance of each feature using a small number of executions.
- Particularly useful for identifying non-influential parameters to determine if they can be safely excluded from further analysis.
- This exclusion step is very important to reduce the size of the problem space when subsequent analytic methods may be computationally expensive.

Feature Selection: Permutation Importance

A Model-based Method: The performance of the model is measured before and after…

Randomly shuffling the values in each column, one-at-a-time, to break the correlation that each column has to the target variable

Then

Measuring the impact the change to each column's influence has upon the model's overall performance according to one of many applicable metrics:

Generates a stack-ranked list of features by their scores

Classification: Accuracy, Precision, or Recall
Regression: MAE, RMSE, RAE, RSE, R^2, Precision, Recall

Original Values

Loan	Income	IsBad
$11,200	$108,000	False
$10,000	$65,000	False
$8,000	$20,000	True
$16,000	$110,000	True
$4,000	$155,000	False

Shuffle Column 1

Loan	Income	IsBad
$8,000	$108,000	False
$4,000	$65,000	False
$16,000	$20,000	True
$10,000	$110,000	True
$11,200	$155,000	False

Shuffle Column 2

Loan	Income	IsBad
$11,200	**$65,000**	False
$10,000	**$155,000**	False
$8,000	**$108,000**	True
$16,000	**$20,000**	True
$4,000	**$110,000**	False

FairLearn: Detecting & Mitigating Unintended Bias

Identifying the Influence of Bias that may be Hidden within otherwise "Innocent" Features

Fairlearn

- Fairness-related harms may arise when a model makes more mistakes for some groups as compared to others.
- The effects of bias can be hidden within innocuous features so excluding obviously biased features isn't enough (e.g., race, sex)
- Fairlearn can be used to assess how different groups are affected and how the observed disparities may be mitigated.

The Fairlearn package has two components:

- ***Metrics*** for assessing which groups are negatively impacted by a model, and for comparing multiple models in terms of various fairness and accuracy metrics.
  - Selection Rate
- ***Algorithms*** for mitigating unfairness in a variety of AI tasks and along a variety of fairness definitions.
  - GridSearch
  - DemographicParity

Resources

Book: Interpretable Machine Learning *by Christoph Molnar*

https://christophm.github.io/interpretable-ml-book

InterpretML

Concept Doc:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability

How-to Doc:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability-aml

Fairlearn

Concept Doc:
https://docs.microsoft.com/azure/machine-learning/concept-fairness-ml

How-to Doc:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-fairness-aml