Microsoft

Responsible ML: Develop and Deploy ML Responsibly

Mehrnoosh Sameki
Senior Program Manager, Azure AI

Why Responsible AI?

7.5.19 [Report] from Capgemini

Microsoft’s AI Principles

Fairness

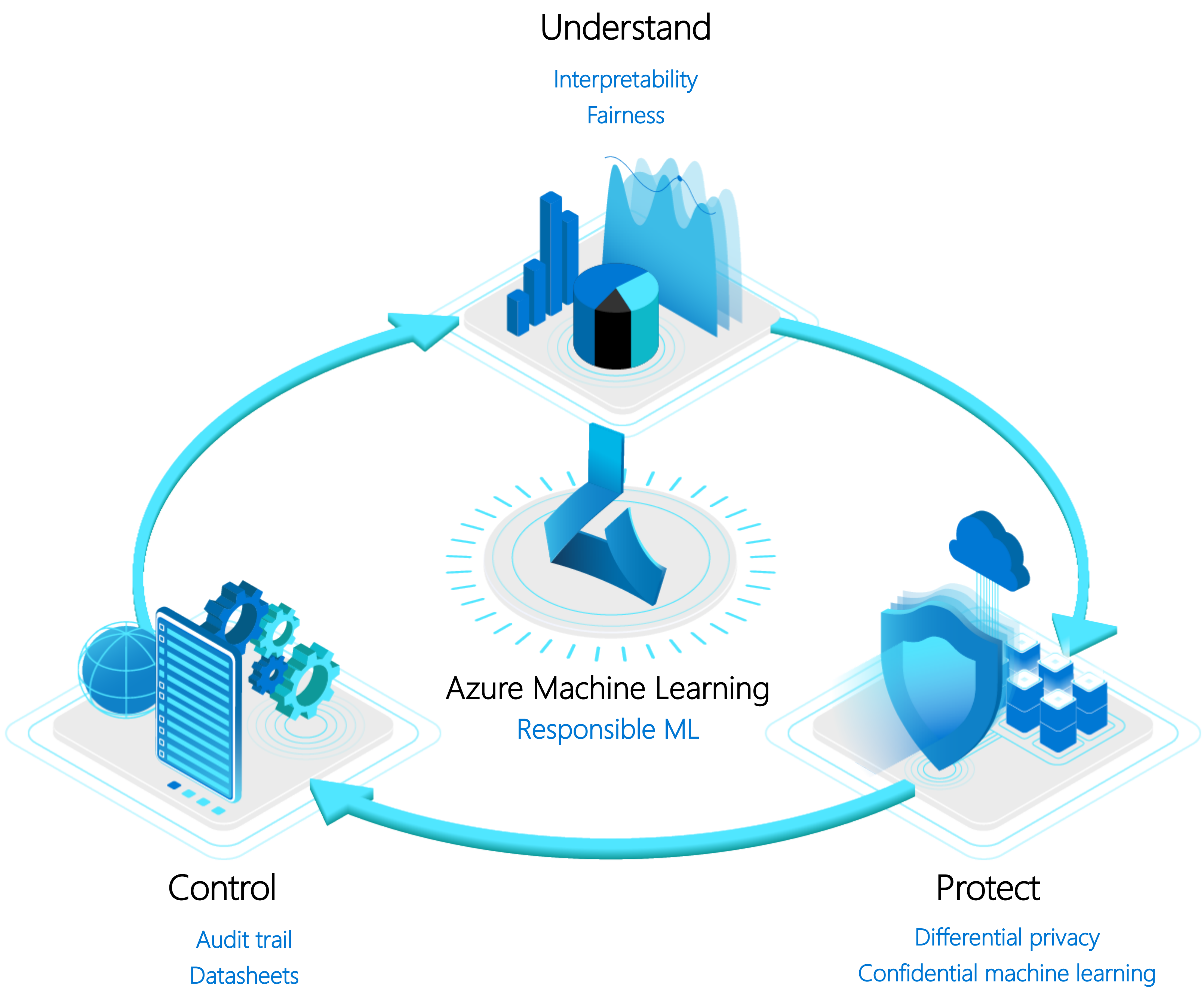
Reliability & Safety

Privacy & Security

Inclusiveness

Transparency

Accountability

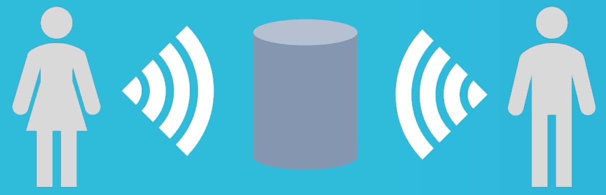

Understand
Interpretability
Fairness
Azure Machine Learning
Responsible ML
Control
Audit trail
Datasheets
Protect
Differential privacy
Confidential machine learning

Fairness

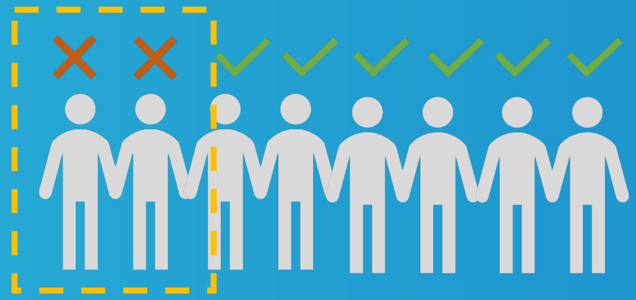
Useful links:
- [AI Show]
- [Tutorial Video]
- Customer [Highlight]

Fairness in AI

There are many ways that an AI system can behave unfairly.



A voice recognition system might fail to work as well for women as it does for men.

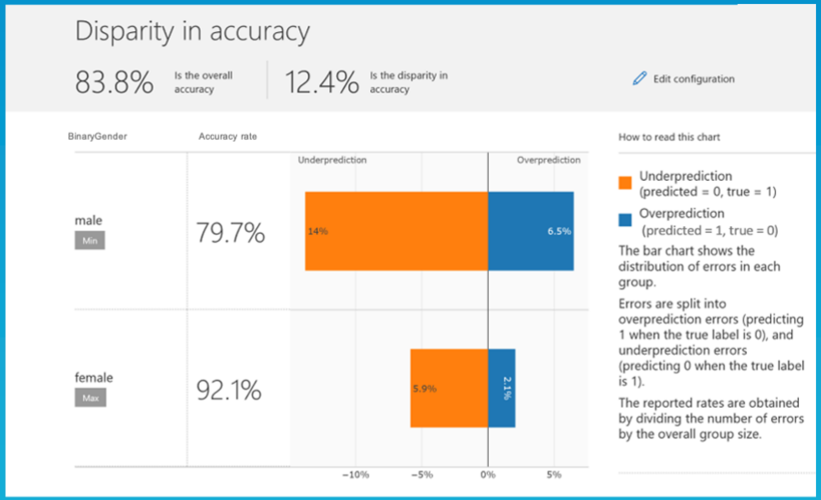


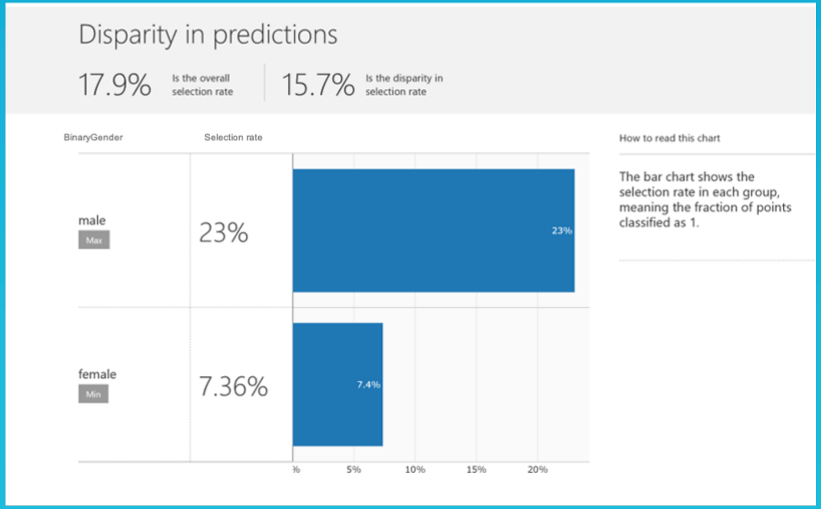
A model for screening loan or job application might be much better at picking good candidates among white men than among other groups.

Avoiding negative outcomes of AI systems for different groups of people

Fairlearn

Assessing unfairness in your model





https://github.com/fairlearn/fairlearn

1 → 2

Fairness Assessment:

Use common **fairness metrics** and an **interactive dashboard** to assess which groups of people may be negatively impacted.

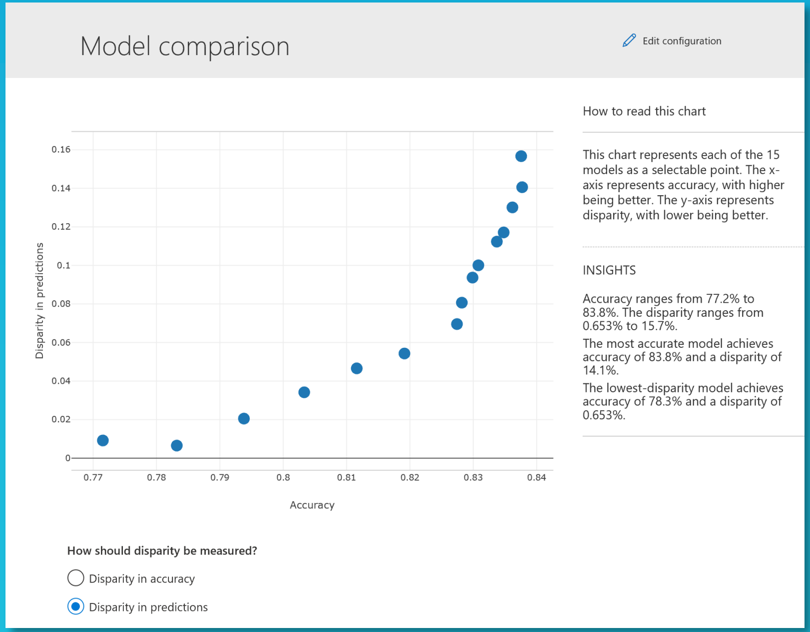
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras

Metrics: 15+ Common group fairness metrics

Model Types: Classification, Regression

Fairness Mitigation:

Use state-of-the-art algorithms to mitigate unfairness in your classification and regression models.



Fairness Assessment

Input Selections

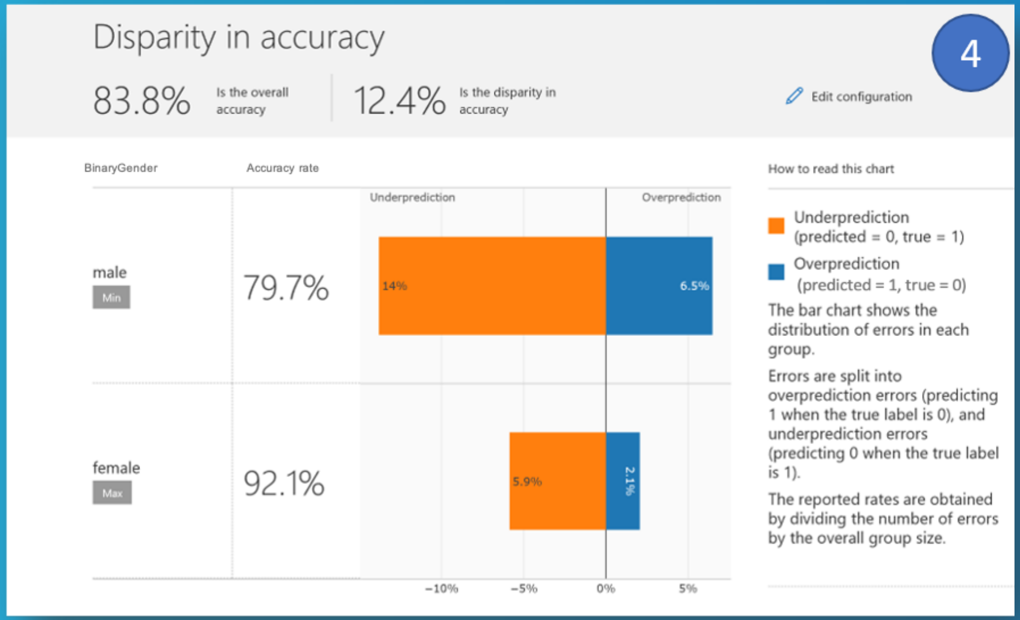
Sensitive attribute
Performance metric

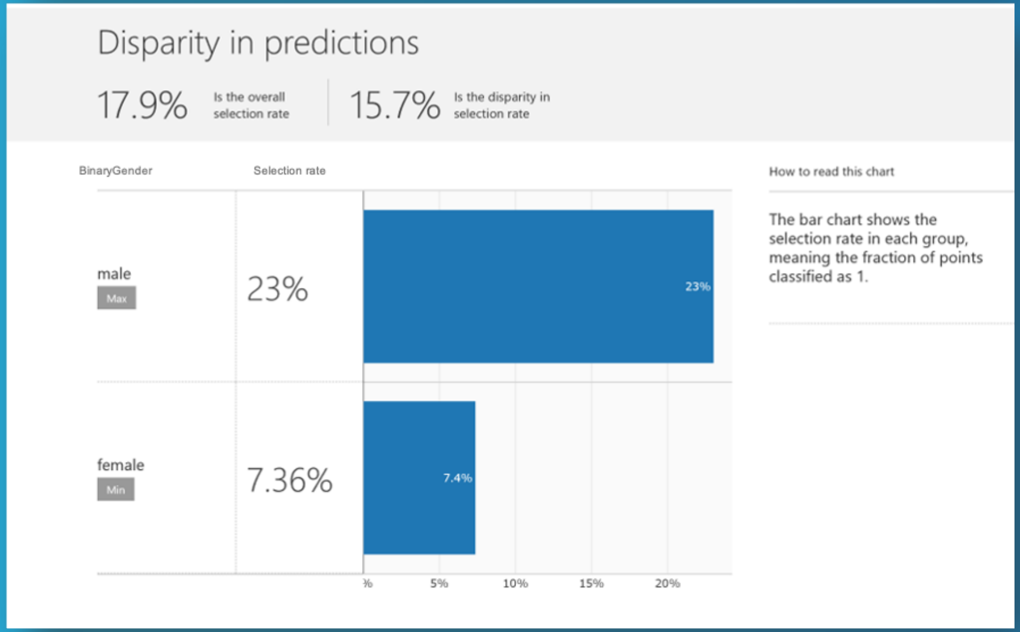
Assessment Results

Disparity in performance
Disparity in predictions

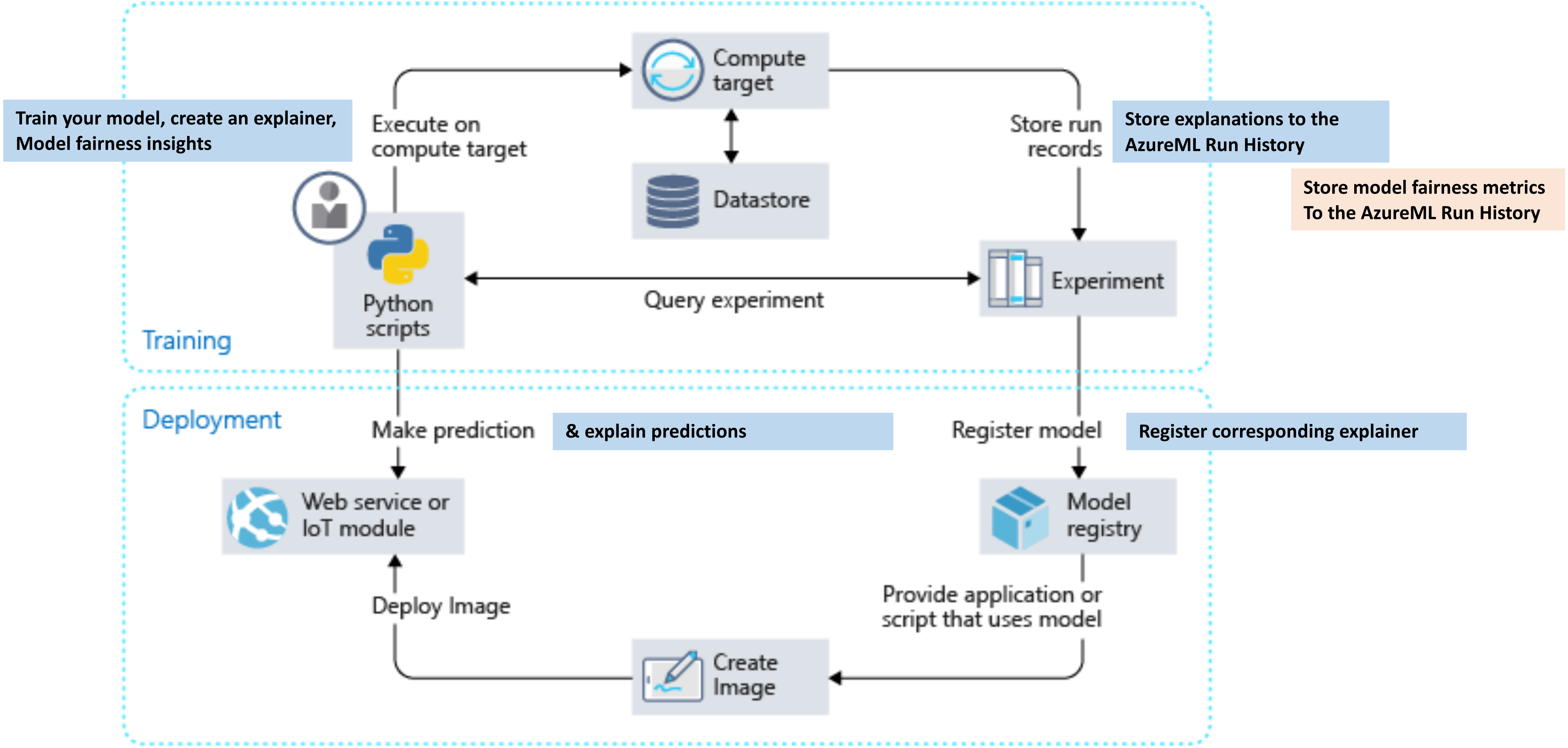
Mitigation Algorithms

Post-processing algorithm
Reductions Algorithm

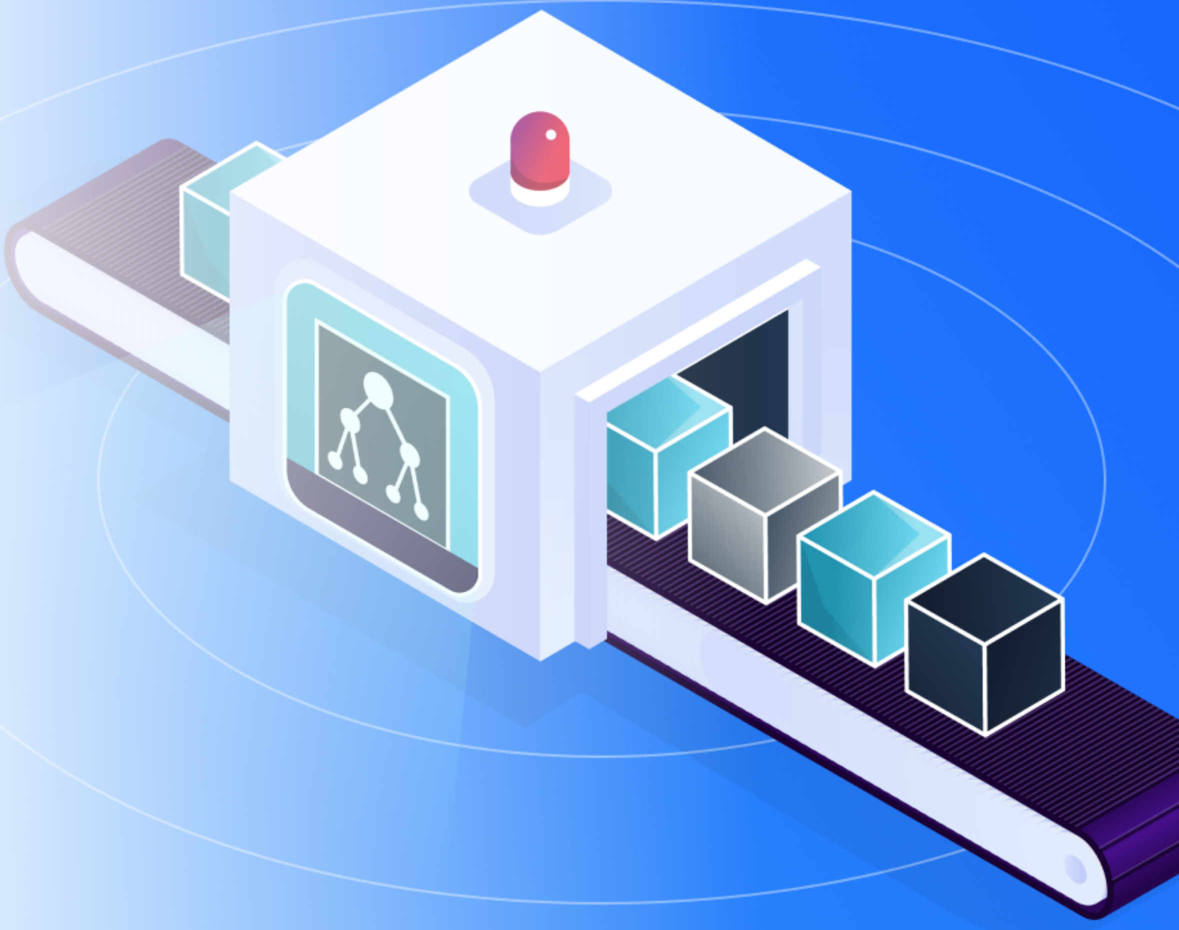




Machine Learning Interpretability in AzureML



Interpretability

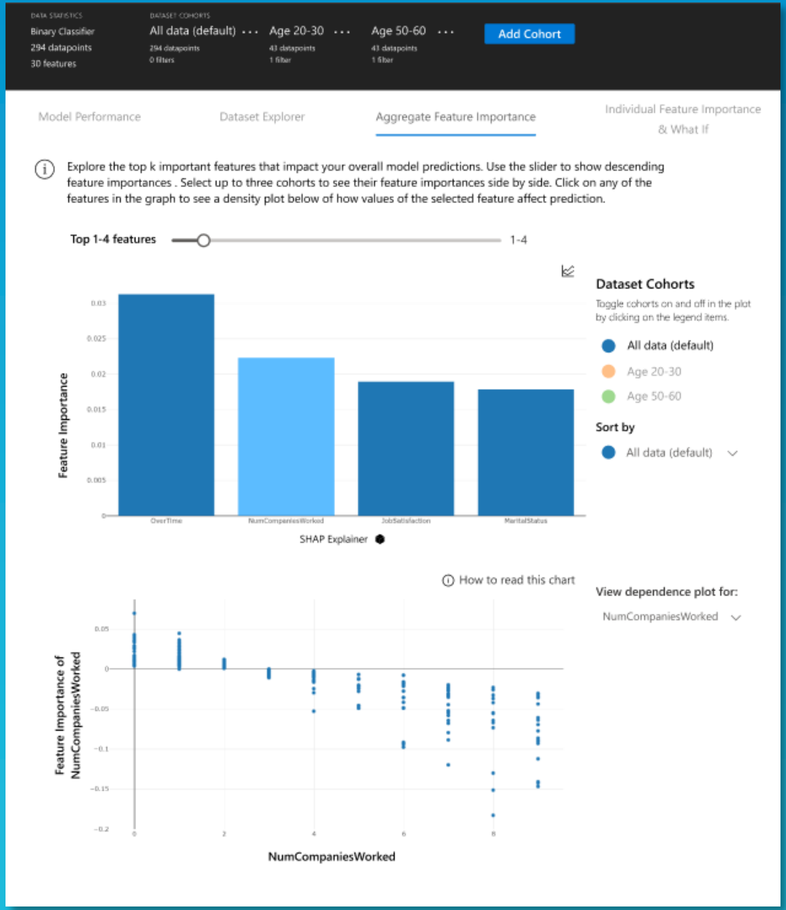


Useful links:

- Tutorial video
- OSS website
- Customer Highlight

InterpretML

Understand and debug your model



Interpret

Glassbox and blackbox interpretability methods for tabular data

Blackbox Models:
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras,

Explainers: SHAP, LIME, Global Surrogate, Feature Permutation

Glassbox Models:
Model Types: Linear Models, Decision Trees, Decision Rules, Explainable Boosting Machines

Interpret-community

Additional interpretability techniques for tabular data

Interpret-text

Interpretability methods for text data

DiCE

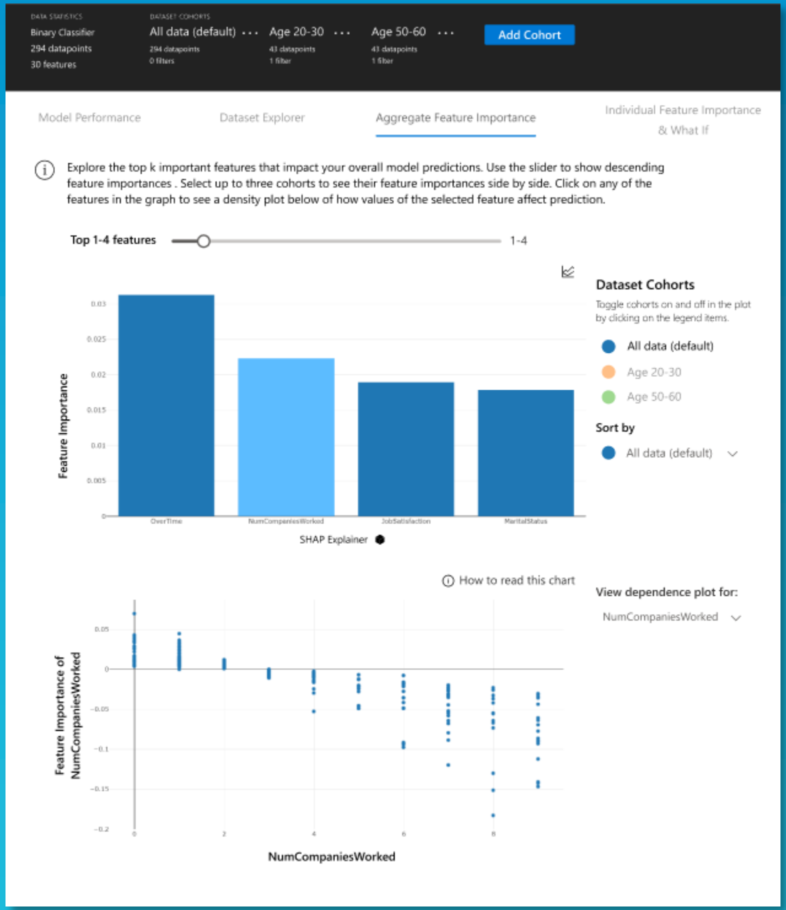
Diverse Counterfactual Explanations

Azureml-interpret

AzureML SDK wrapper for Interpret and Interpret-community

https://github.com/interpretml

InterpretML Understand and debug your model



Interpret
Glassbox and blackbox interpretability methods for tabular data

Interpret-community
Additional interpretability techniques for tabular data

Blackbox Models:
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras,

Explainers: SHAP, LIME, Global Surrogate, Feature Permutation

Glassbox Models:
Model Types: Linear Models, Decision Trees, Decision Rules, Explainable Boosting Machines

Interpret-text
Interpretability methods for text data

DiCE
Diverse Counterfactual Explanations

Azureml-interpret
AzureML SDK wrapper for Interpret and Interpret-community

https://github.com/interpretml

Interpretability Approaches

Glassbox Models

Blackbox Explanations

Glassbox Models

Models *designed* to be interpretable.
Lossless explainability.

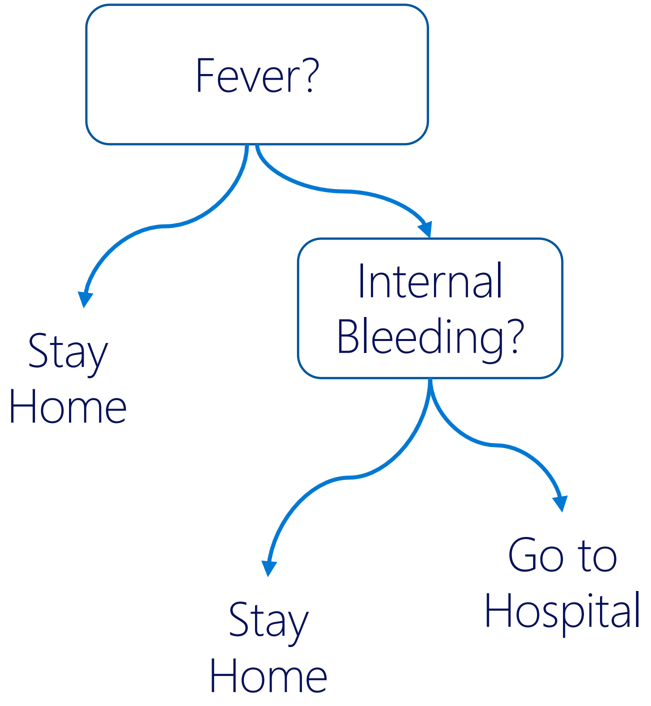
Explainable Boosting Machine

Decision Trees

Rule Lists

Linear Models

....



Blackbox Explanations

Explain *any* ML system.
Approximate explainability.

SHAP

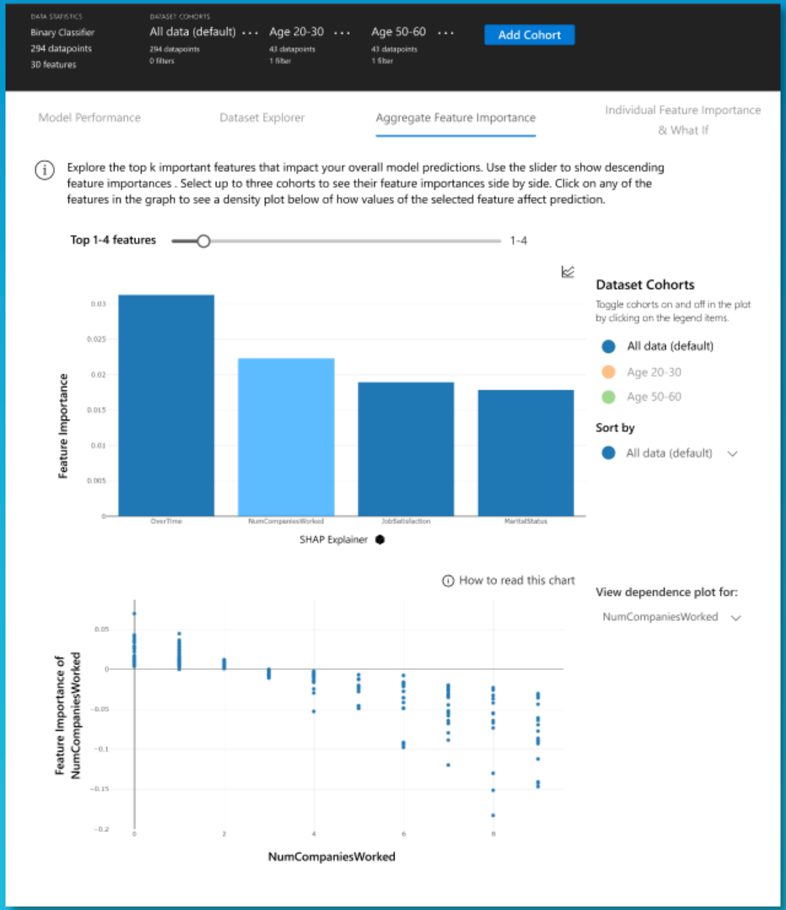
LIME

Partial Dependence

Sensitivity Analysis

InterpretML

Understand and debug your model



Interpret
Glassbox and blackbox interpretability methods for tabular data

Interpret-community
Additional interpretability techniques for tabular data

Blackbox Models:
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras,

Explainers: SHAP, LIME, Global Surrogate, Feature Permutation

Glassbox Models:
Model Types: Linear Models, Decision Trees, Decision Rules, Explainable Boosting Machines

Interpret-text
Interpretability methods for text data

DiCE
Diverse Counterfactual Explanations

Azureml-interpret
AzureML SDK wrapper for Interpret and Interpret-community

https://github.com/interpretml

Loan Application Decisions

Create a model for loan application acceptance

Azure Machine Learning

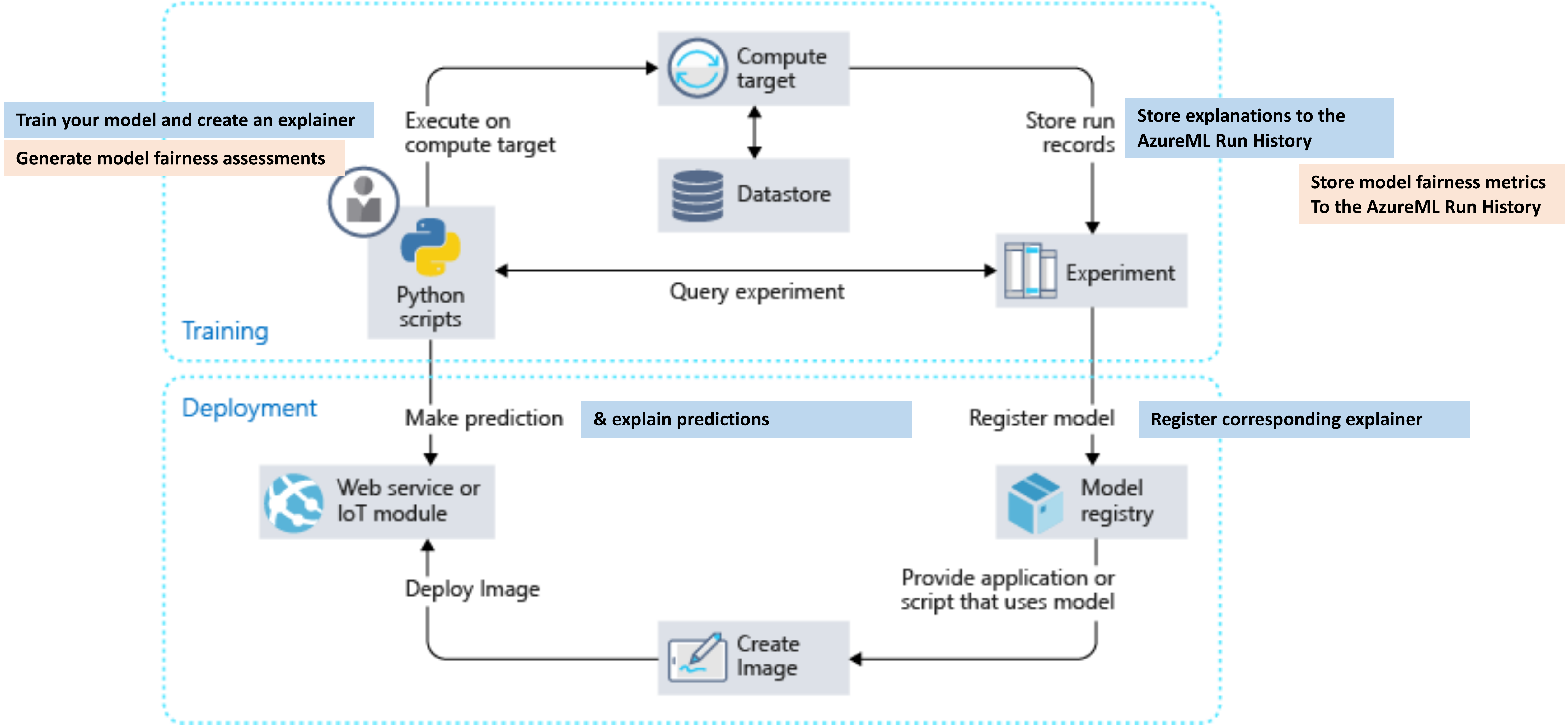
Fairlearn

Is my model fair?

InterpretML

How does it decide who to accept or reject?

Responsible Machine Learning in AzureML



AzureML Responsible ML Resources

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

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