ThaiPy - Bangkok Python Meetup

THU, MAR 11, 6:45 PM GMT+7

Responsible AI toolkits

SeokJin Han

Microsoft

linkedin.com/in/seokjinhan

sehan@microsoft.com

Agenda

- What to consider to deal with Responsible AI
- Why
- What are the approaches
- Some demos and quizzes: Interpret ML, Error Analysis
- Resources
- Join the Community!

Get Hands Dirty

- https://github.com/dem108/explain-ml-models
  - Try git clone and follow README.md to get started

Progression of Responsible AI

Principles

Fairness

Reliability & Safety

Inclusiveness

Privacy & Security

Transparency

Accountability

Practices

AETHER committee

The Partnership on AI

Guidelines for Human-AI Design

Guidelines for Conversational AI

Tools

Homomorphic Encryption

Interpret ML

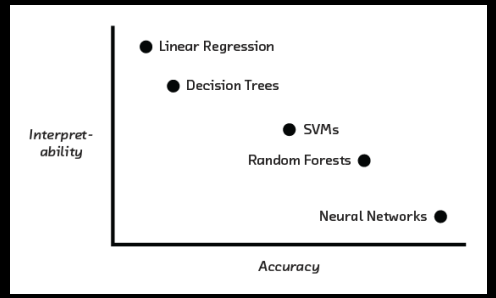
Differential Privacy

Data Drift

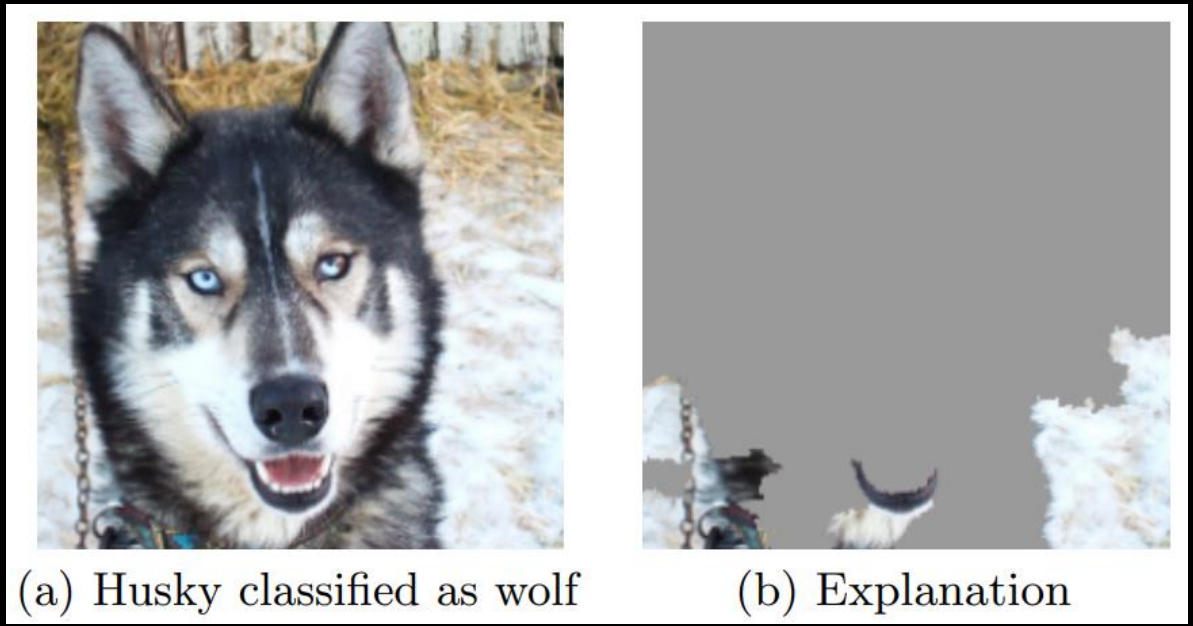
Secure MPC

Why this matters?

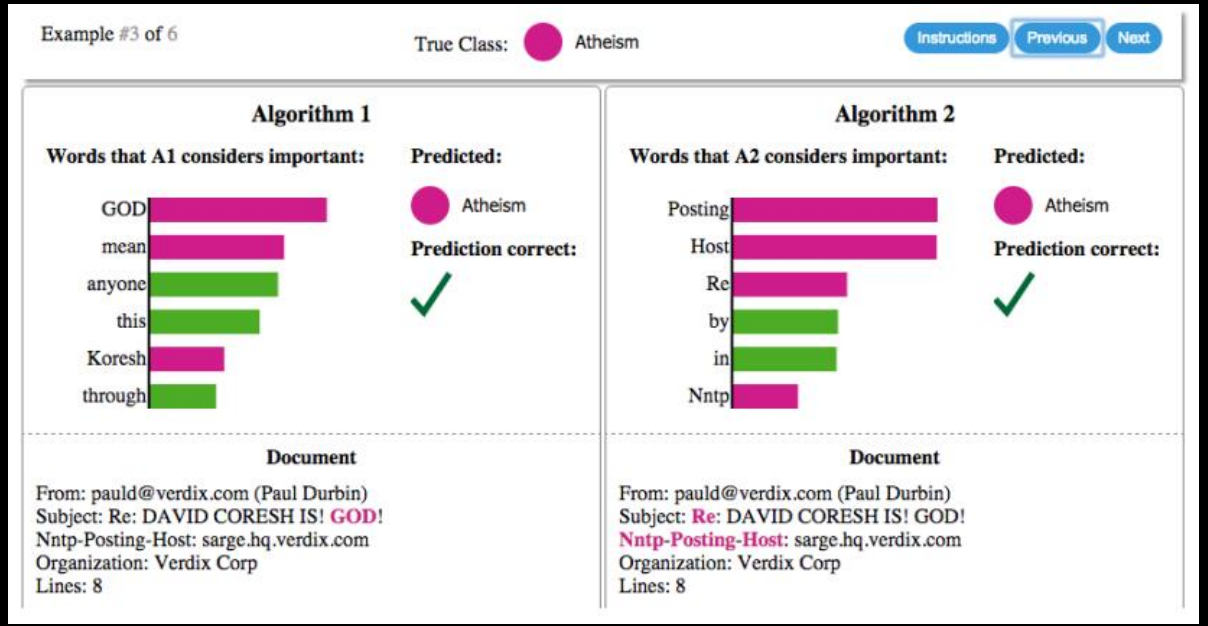
- “If you can't explain it simply, you don't understand it well enough.“ — Albert Einstein
- Accuracy (as a whole) is not enough
- Is my model predicting well, or am I totally mistaken?



Dare to explain?

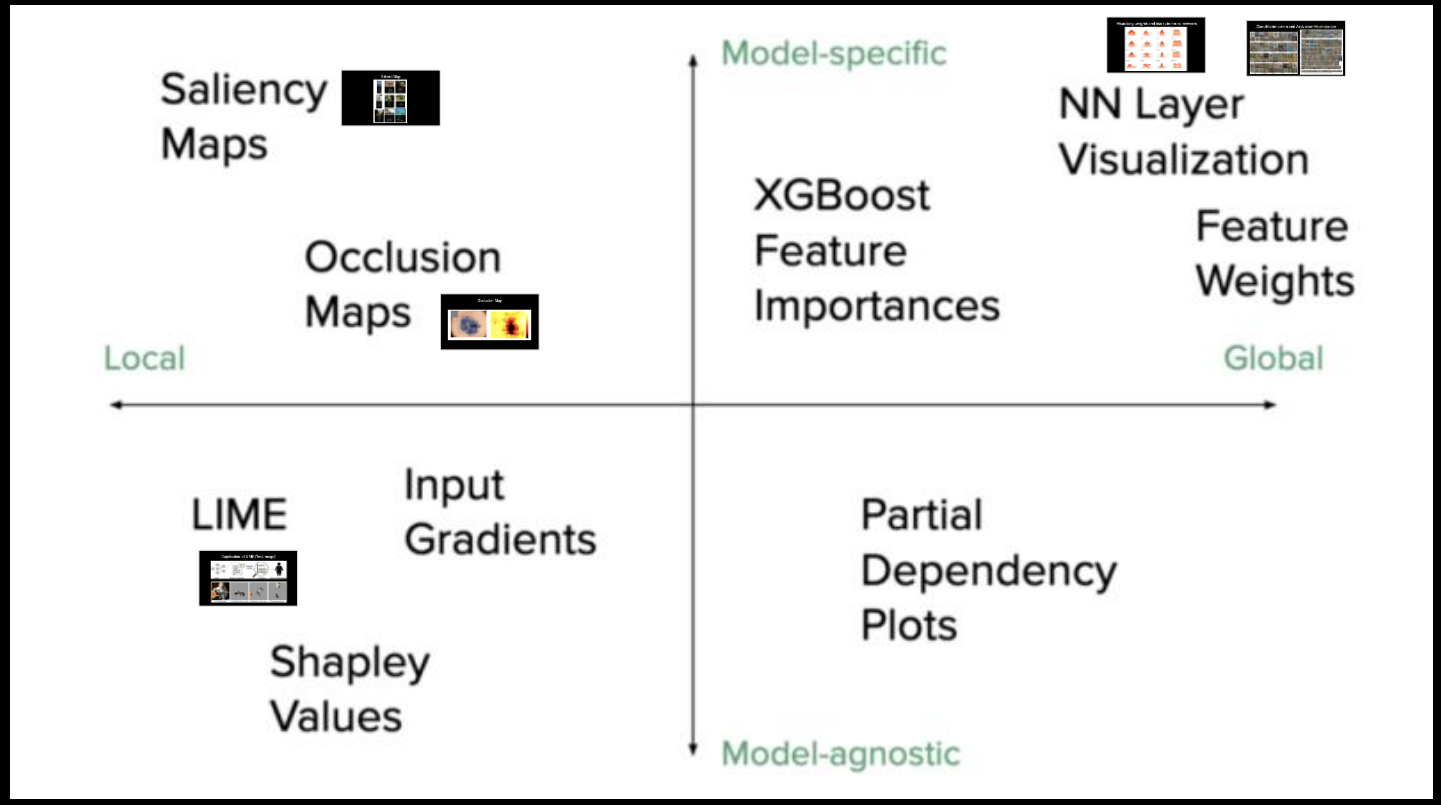


(a) Husky classified as wolf (b) Explanation

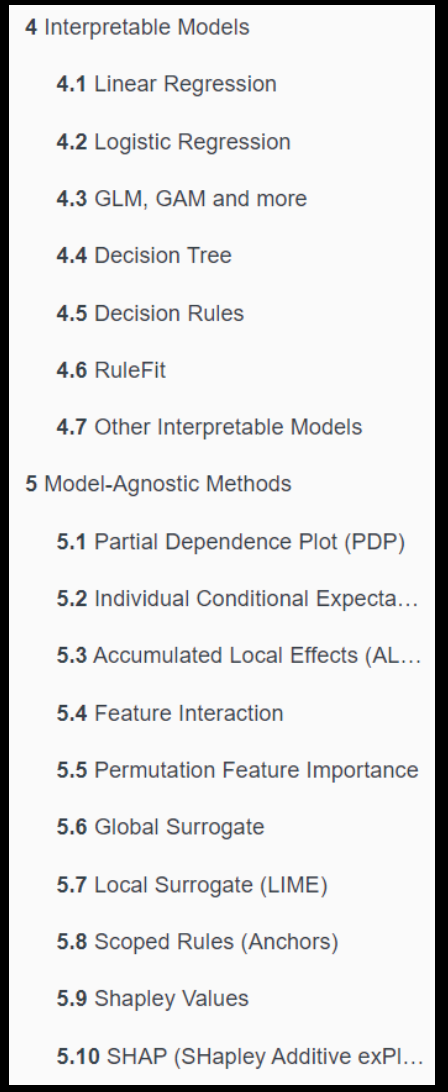


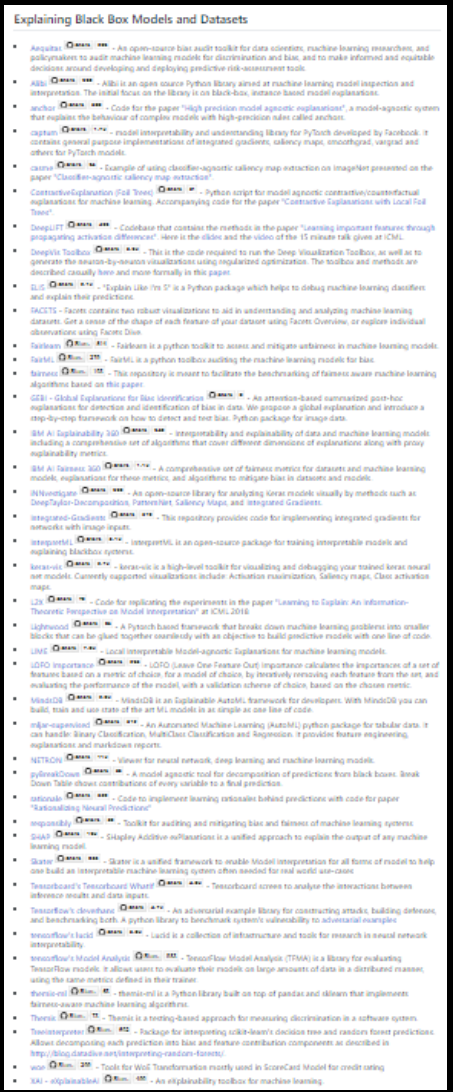
Source: https://arxiv.org/abs/1602.04938, “Why Should I Trust You?”: Explaining the Predictions of Any Classifier

Efforts to explain models

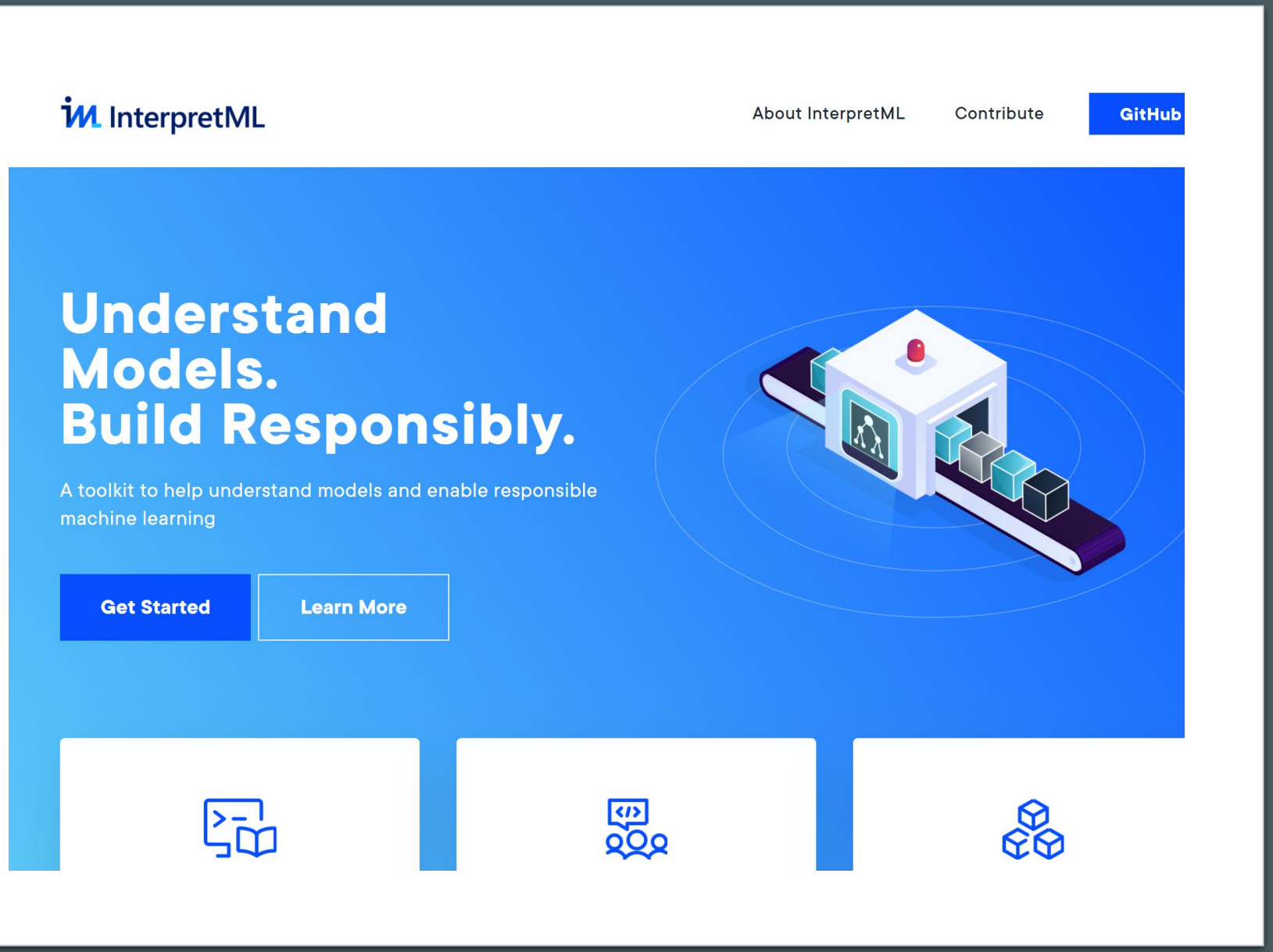


X-axis	**Local**	Explain locally (individual filters in NN etc)
	Global	Explain globally (visualize weights and biases in NN etc)
Y-axis	**Model-agnostic**	Can apply to many model types
	Model-specific	Works for specific model types

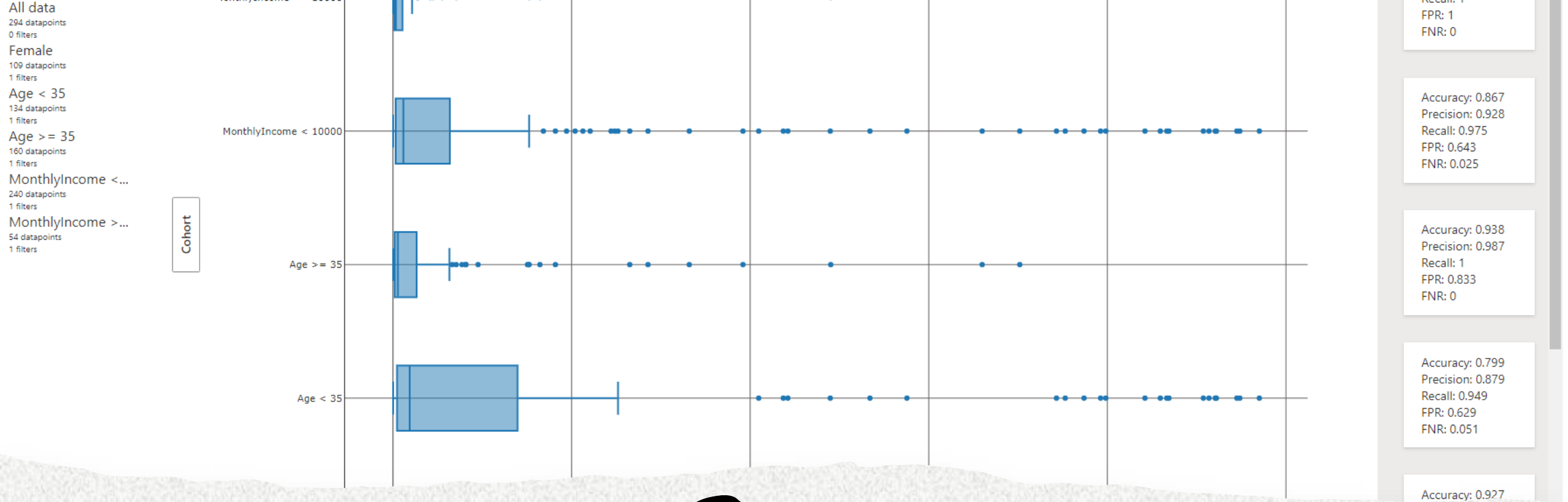




Source: https://github.com/Harvard-IACS/2020-ComputeFest/, https://christophm.github.io/interpretable-ml-book/, https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets

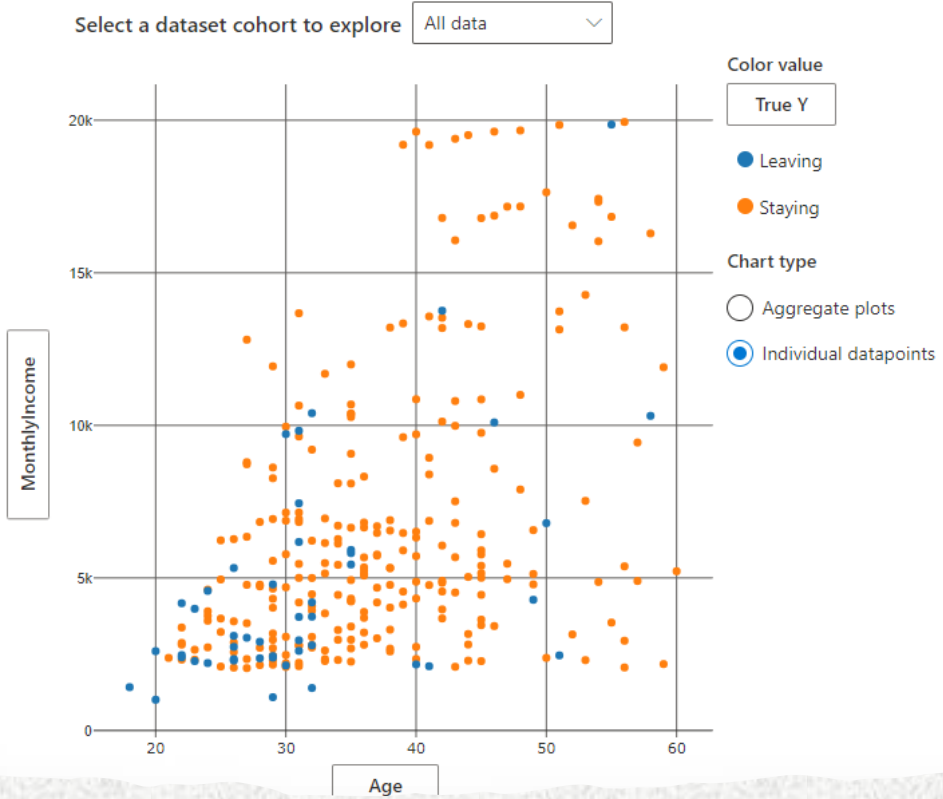


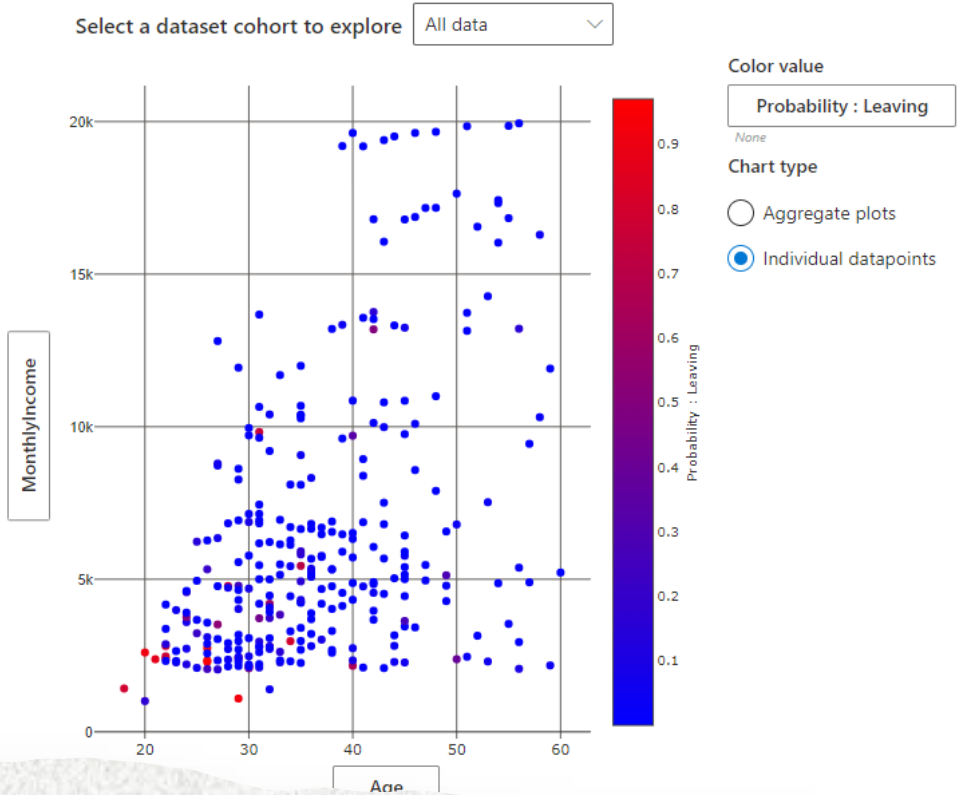
InterpretML



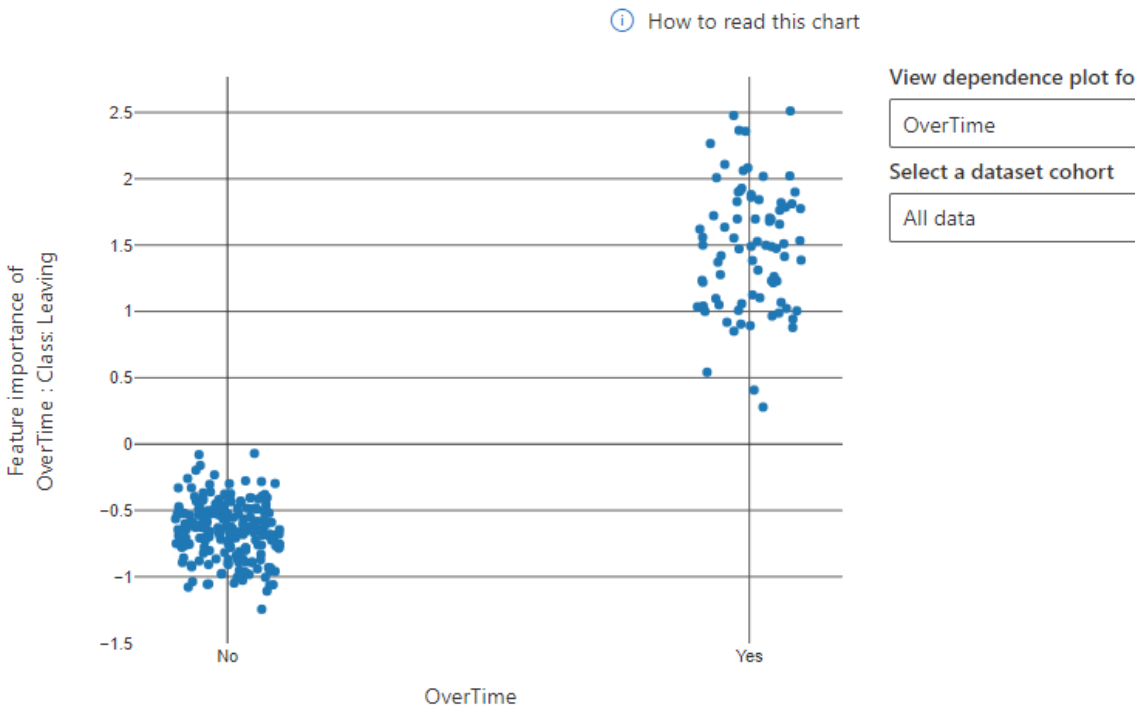
Does Model perform differently per Cohort?

** Disclaimer: Insights from these should be validated with different aspects*

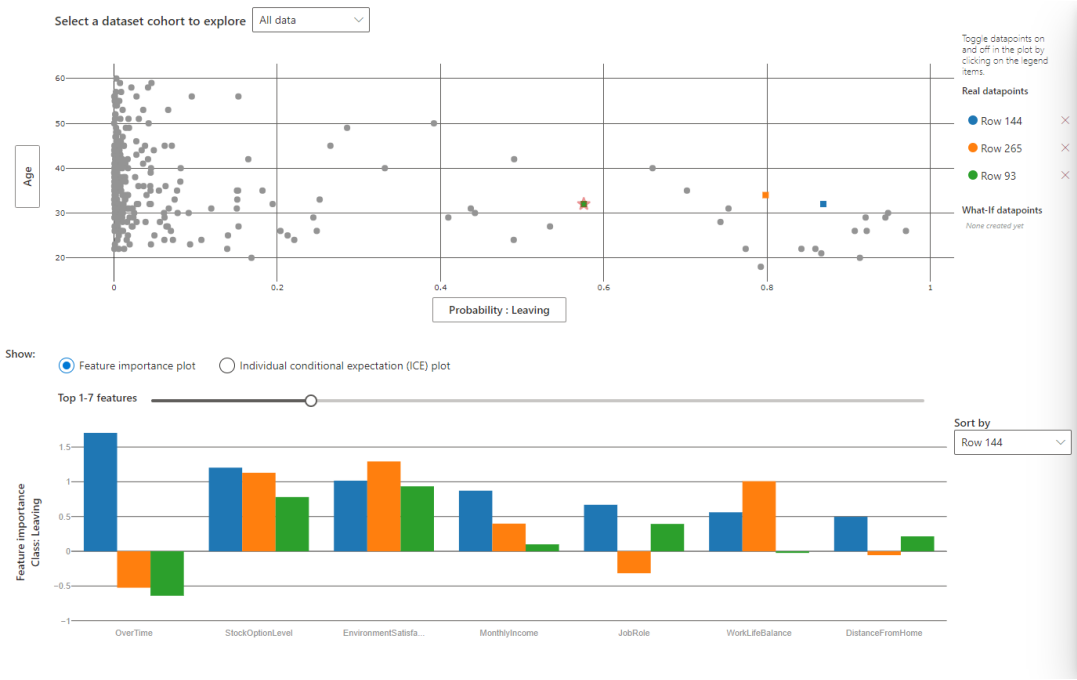


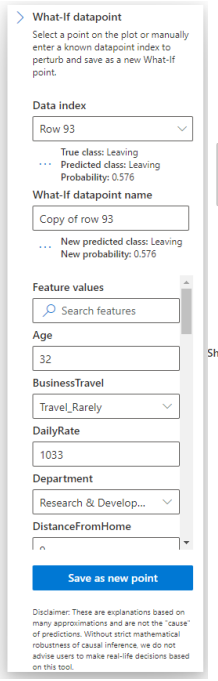


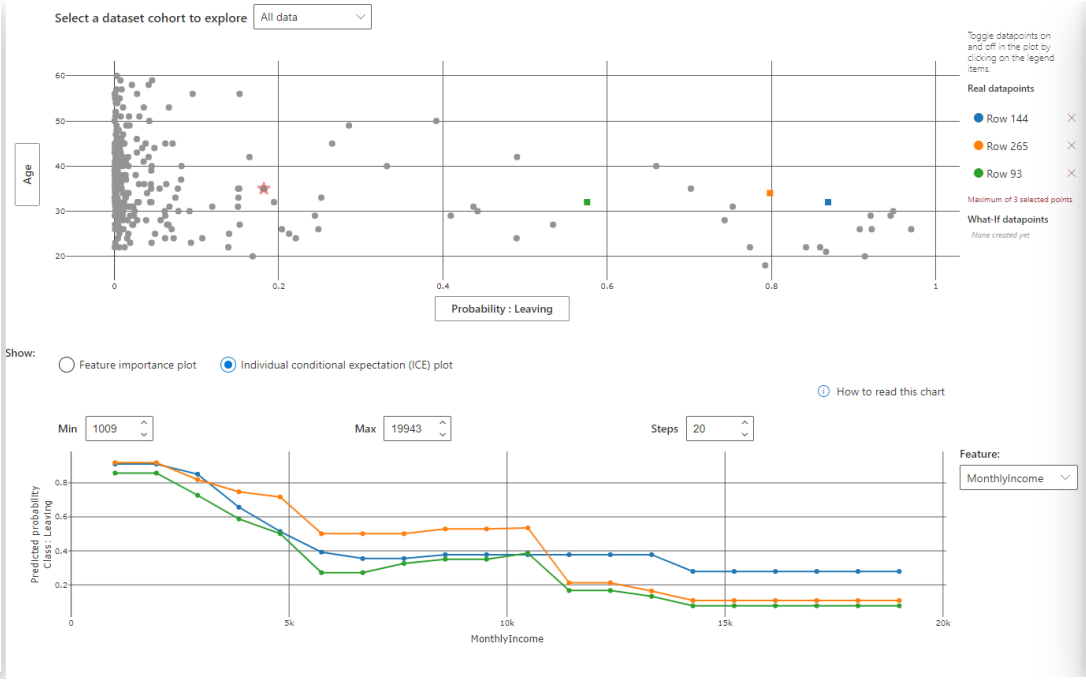
Are they making more money (on average) as they grow older?
When are they actually leaving?
When are they predicted to leave?
Does it change due to Gender?

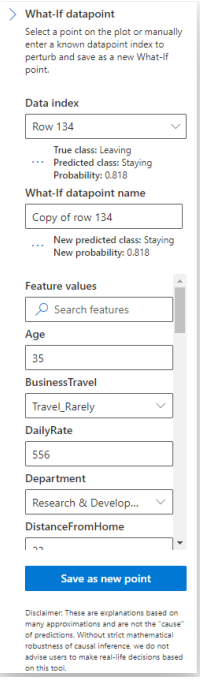


What are the most important features?
How much does Over Time affect attrition?
What about Stock Option Level, Monthly Income?
Does Gender, Age make a difference?





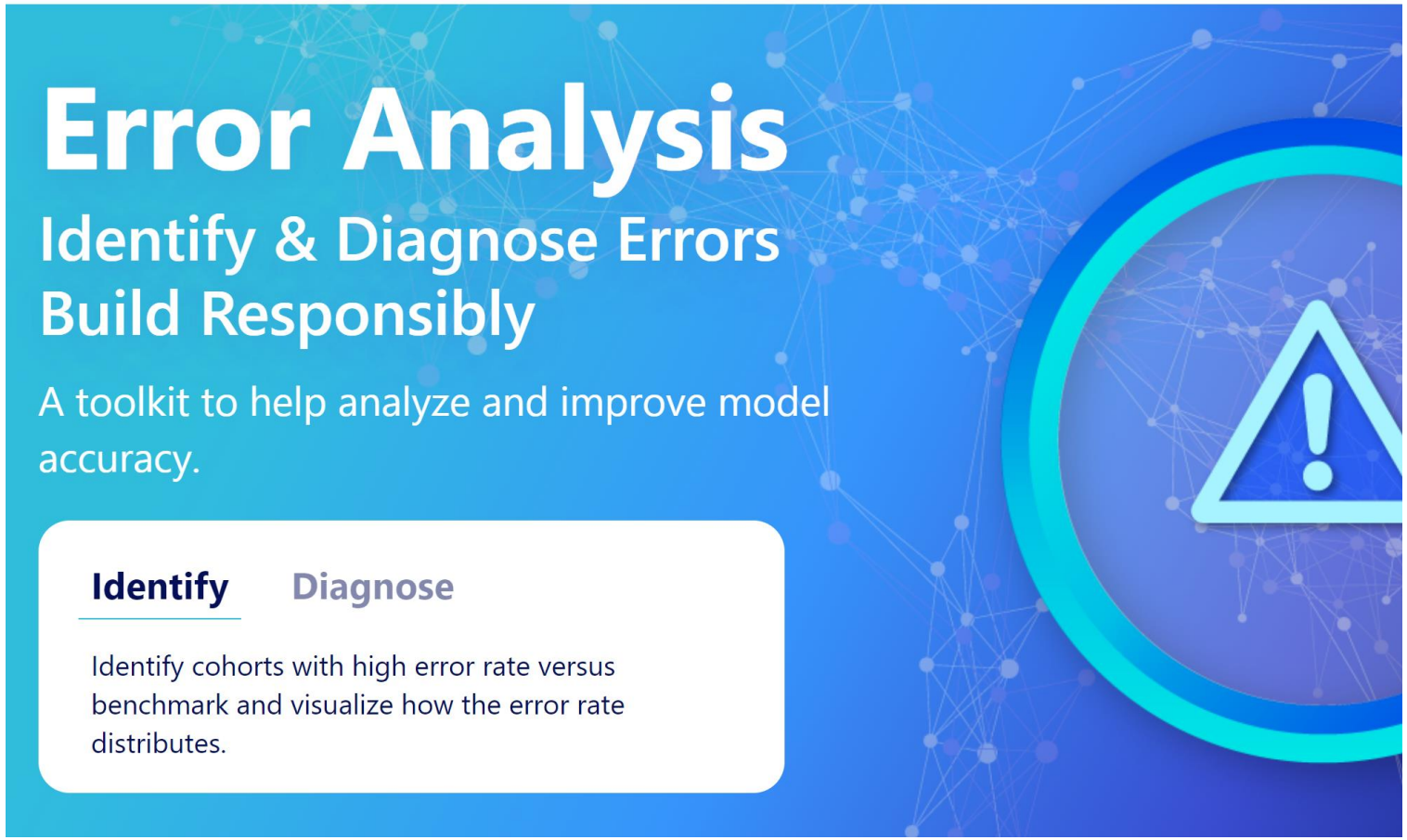




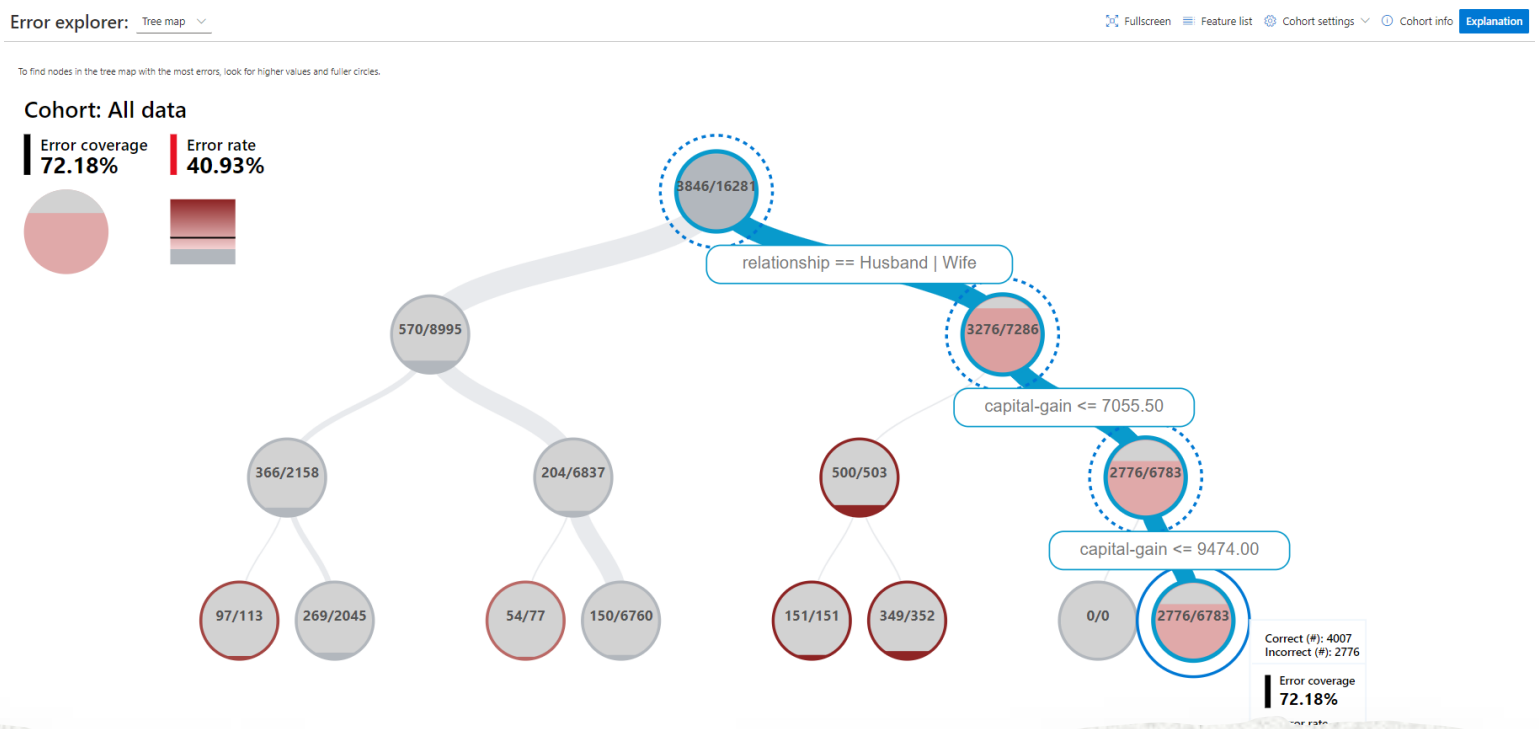
Take 3 employees with similar age, can we tell what is the most importance feature for each person according to the model?

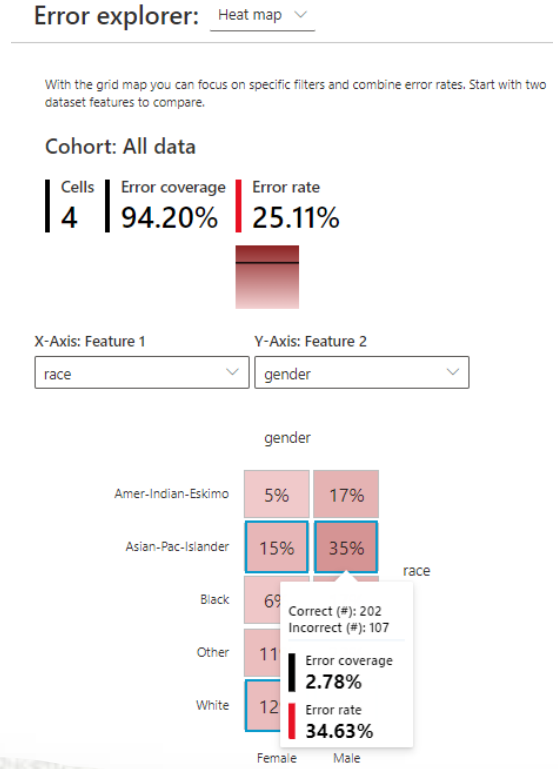
What is the impact of Age, Environment Satisfaction, Monthly Income on the probability of Leaving, and how does the impact differ per person?

Error Analysis



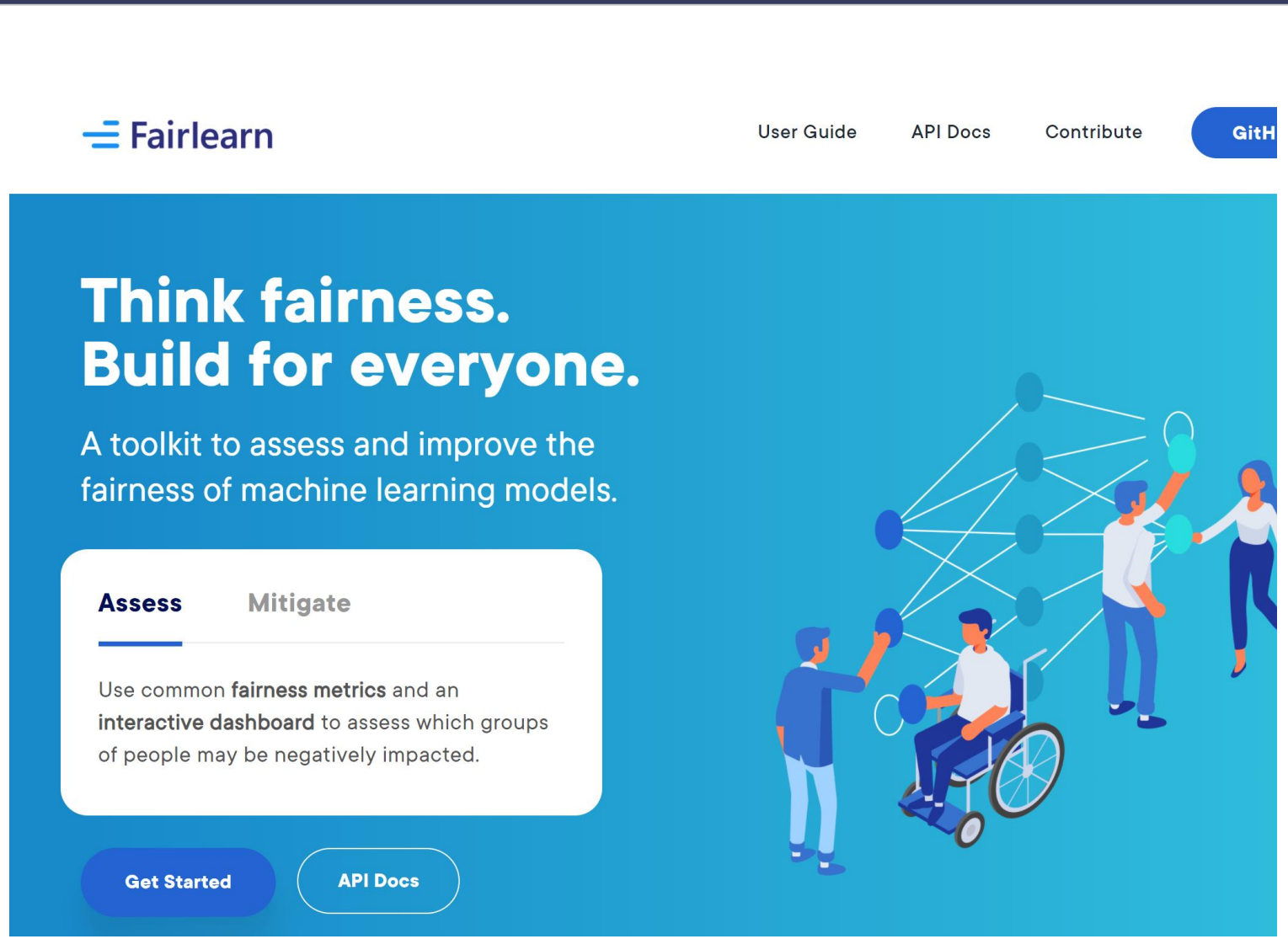
Error Analysis





Which segments cover the most errors?

Which segments have the highest error rates?

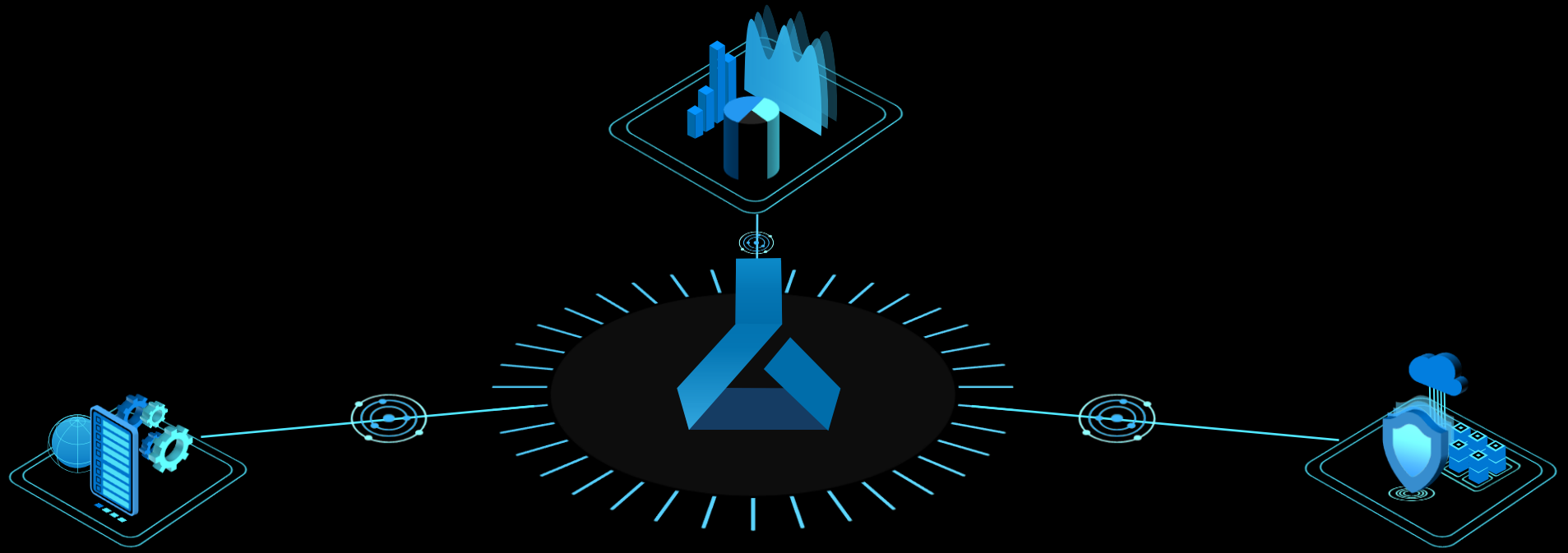

Fairlearn
User Guide
API Docs
Contribute
GitH
Think fairness.
Build for everyone.
A toolkit to assess and improve the fairness of machine learning models.
Assess
Mitigate
Use common fairness metrics and an interactive dashboard to assess which groups of people may be negatively impacted.
Get Started
API Docs

Fairlearn

Understand

Interpretability

Fairness



Control

Audit trail

Datasheets

Azure Machine Learning

Responsible ML

Protect

Differential privacy

Confidential machine learning

Codes, Blogs, Videos

- https://github.com/dem108/explain-ml-models
- Enabling responsible AI development with new open-source capabilities
- https://github.com/microsoft/responsible-ai-widgets
- InterpretML
- Error Analysis
- Build Responsible AI using Error Analysis toolkit – YouTube
- Responsible Machine Learning with Error Analysis – Microsoft Blog
- Model interpretability in Azure Machine Learning
- Responsible AI at Microsoft

Join our community!

Receive regular updates from our community experts and exclusive invites to our community events and workshops.

https://aka.ms/MSSourceTH

Microsoft