OneModel Academy

One AI

SHapley Additive exPlanations (SHAP)

ML Models

Hayley Bresina
One AI Client Enablement

29 May 2024

Topics Covered

- Background & overview of Shapley values
- Introduction to SHAP in machine learning
- Strengths & weaknesses of SHAP as a method of model interpretation
- How to interpret SHAP in One AI

Learning Outcomes

You will:

- Have a clear understanding of Shapley values & how they're adapted to provide interpretability in machine learning models through SHAP
- Understand how SHAP values are used to explain individual model predictions that can be aggregated to larger groupings
- Identify how One AI leverages the strengths & mitigates the weaknesses of SHAP
- Gain practical insights into interpreting SHAP visualizations available in the One AI Results Summary & model storyboards in One Model

Shapley Values Background & Overview

Shapley Values

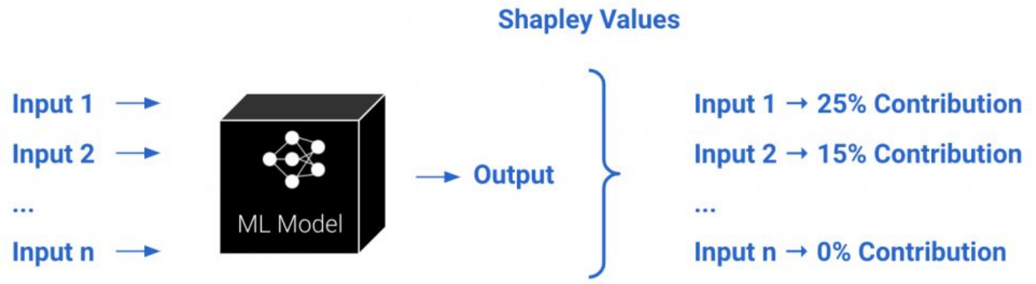
- SHAP builds upon the concept of Shapley Values
- **Shapley values** are a concept from cooperative game theory that has been adapted to machine learning for **model interpretability**
  - Provide a clear, **numerical** way to assign a value or importance to each feature within a predictive model
  - Represents each feature’s average contribution to model predictions across all possible feature combinations
  - Positive Shapley values indicate features that tend to increase predictions; negative values indicate features that tend to decrease predictions
  - Provide insights into **feature importance**

Intro to SHAP in Machine Learning

SHAP

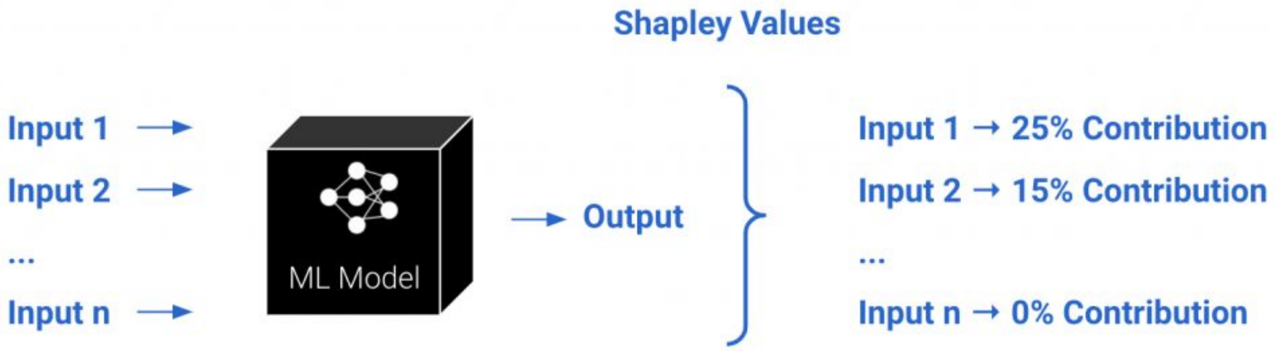
SHAP (**SH**apley **A**dditive ex**P**lanations) is a method used in machine learning to explain individual predictions made by models

- Machine learning models make predictions based on input features
- Explains why a specific prediction was made for a particular instance
- Can be aggregated to provide group insights


Shapley Values
Input 1
Input 2
...
Input n
ML Model
Output
Input 1 → 25% Contribution
Input 2 → 15% Contribution
...
Input n → 0% Contribution

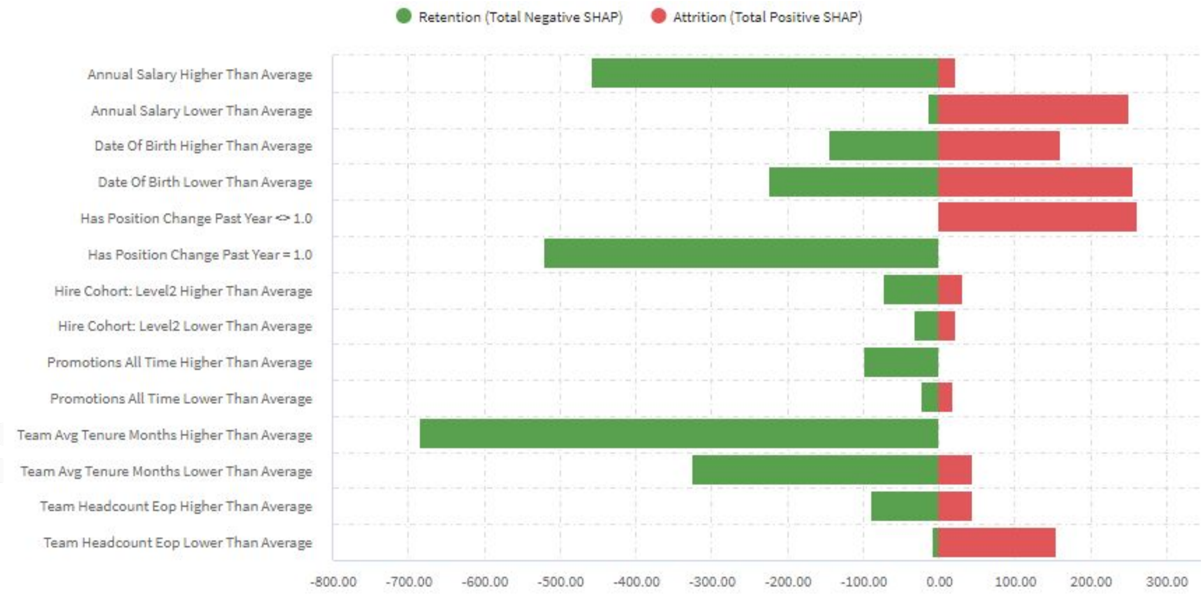
SHAP

- Systematically excludes different features & observes how this impacts predictions
- Considers all combinations of features & their contributions to predictions
- After evaluating the impact of each feature across combinations, SHAP aggregates the results to assign a Shapley value to each feature to fairly distribute importance values


Shapley Values
Input 1 →
Input 2 →
...
Input n →
ML Model
→ Output
Input 1 → 25% Contribution
Input 2 → 15% Contribution
...
Input n → 0% Contribution

SHAP

- SHAP values can be visualized to help interpret which features drive predictions up or down



Strengths & Weaknesses

Strengths & Weaknesses

Strengths

- Promotes ethical AI
- Interpretability
- Feature Importance

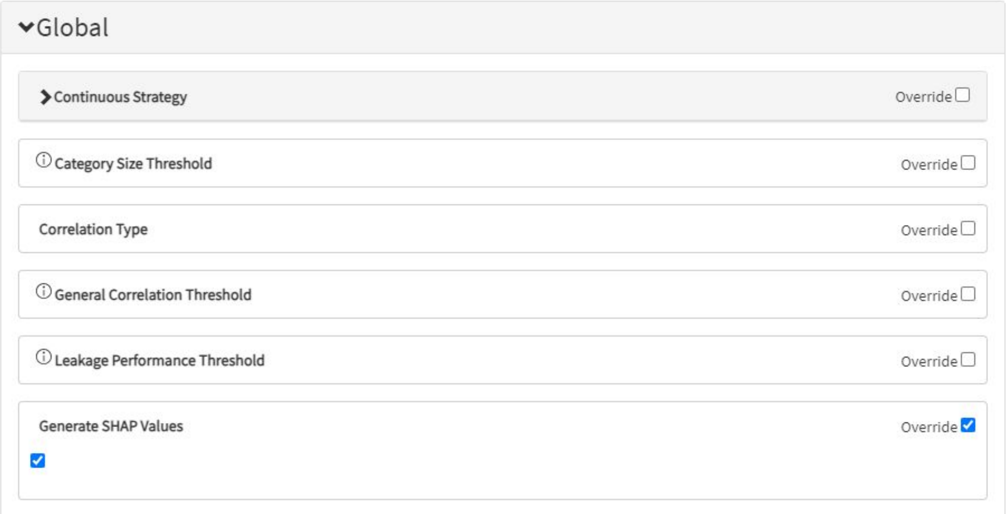
Weaknesses

- Complexity of interpretation
- Time & resources
- Challenge visualizing very high-dimensional data

SHAP in One AI

SHAP in One AI

- Shap values are **not** generated by default
- Can be enabled by model in the global settings



Global

Continuous Strategy — Override □

Category Size Threshold — Override □

Correlation Type — Override □

General Correlation Threshold — Override □

Leakage Performance Threshold — Override □

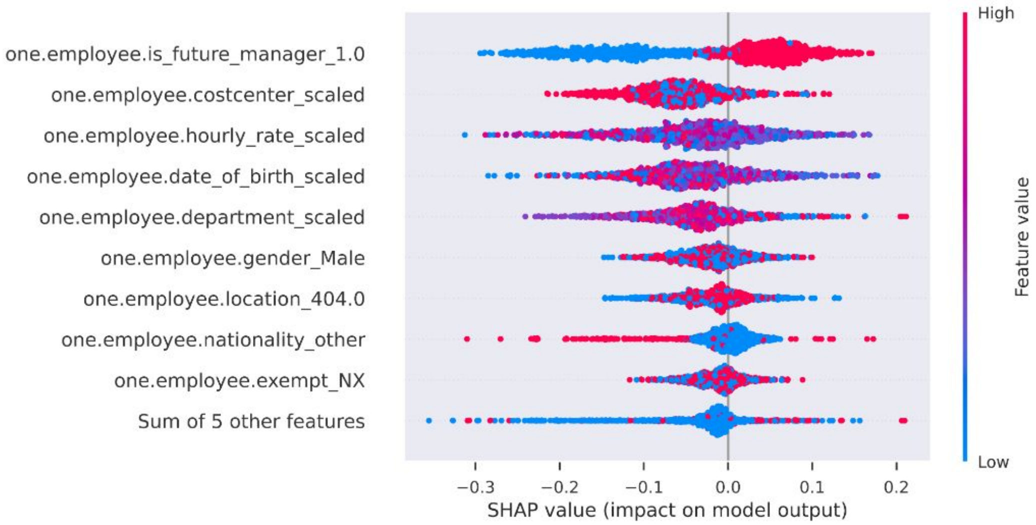
Generate SHAP Values — Override ☑

☑

SHAP Beeswarm Chart

- A feature impact visualization where each SHAP numerical importance for every prediction is plotted as a dot

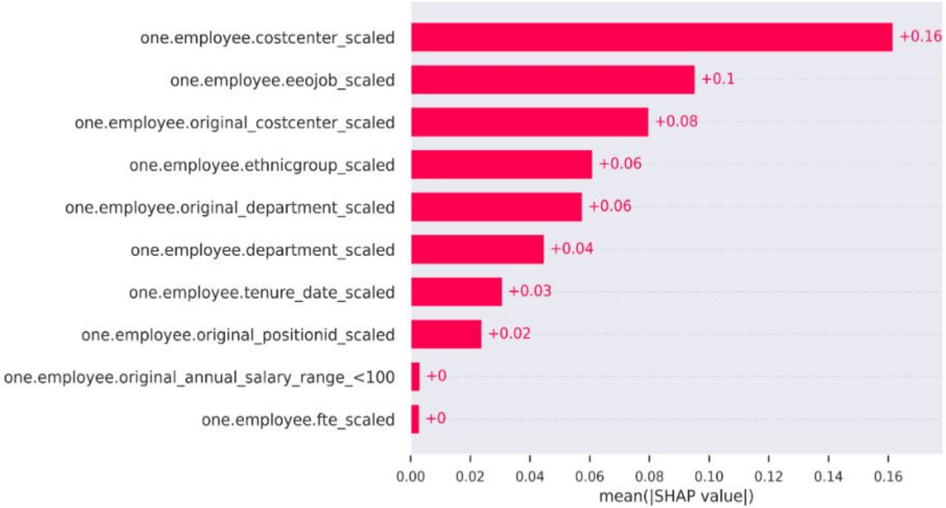
- The horizontal axis indicates how predictive that feature is for that instance & in what direction



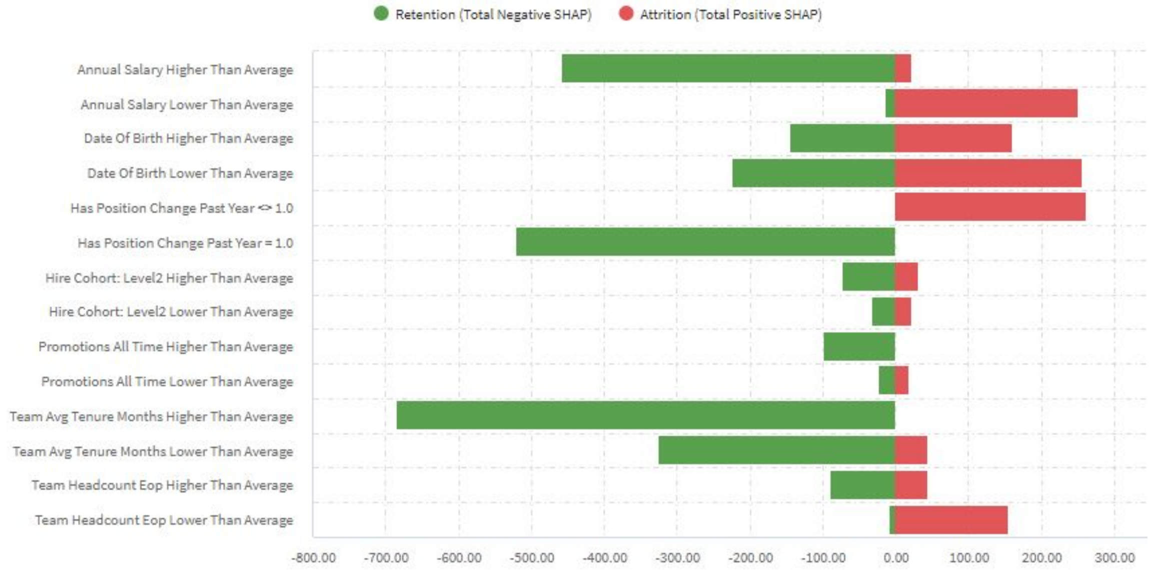
SHAP Average Bar Chart

- Shows the average absolute value of the SHAP values for each feature

- Great indicator of how important the feature was to this set of predictions but not whether the feature made a positive classification more likely



SHAP Values on Storyboard Tiles



- Tables must be configured by Data Engineer & model should be in a deployed status

OneModel Academy

Thanks for watching!

OneAI