NDL Fall 25 GMB #8

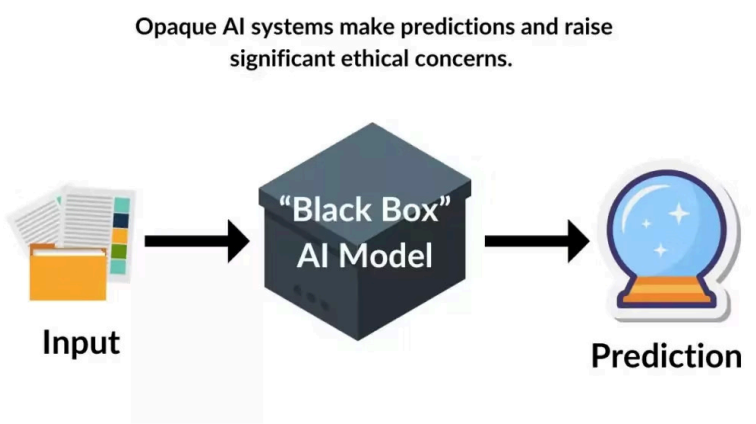
10/26/2025

Attendance: "SHAP"

Why Explain Machine Learning Models?

Complex models (random forests, deep nets) behave like black boxes. Explanations are required to:

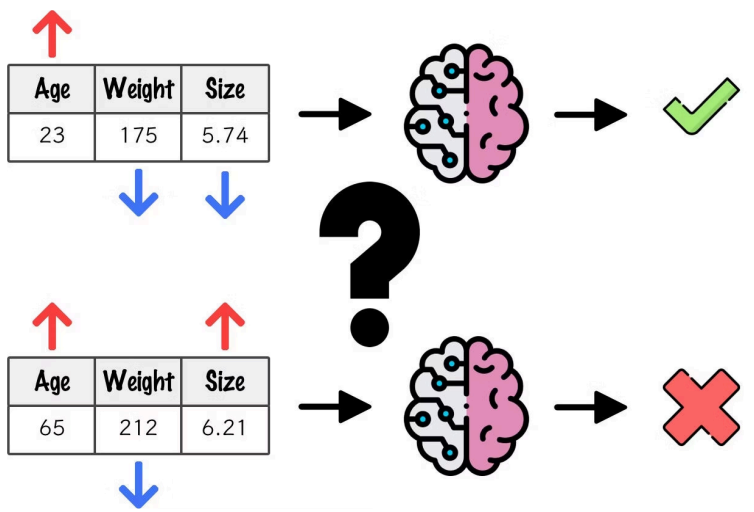
- Build stakeholder trust and justify decisions (e.g., loan denial)
- Diagnose model errors and data issues
- Meet regulatory and ethical requirements



What is SHAP?

SHAP = SHapley Additive exPlanations

- Feature importance method
- Measures how much each input (feature) is helping or hurting the models performance
- Fairly distributes "payout" (prediction) among the features based on contribution



Key Characteristics of SHAP

1

Model-Agnostic

Works with any machine learning model, providing unified explanations regardless of the model's complexity or type.

2

Fair Attribution

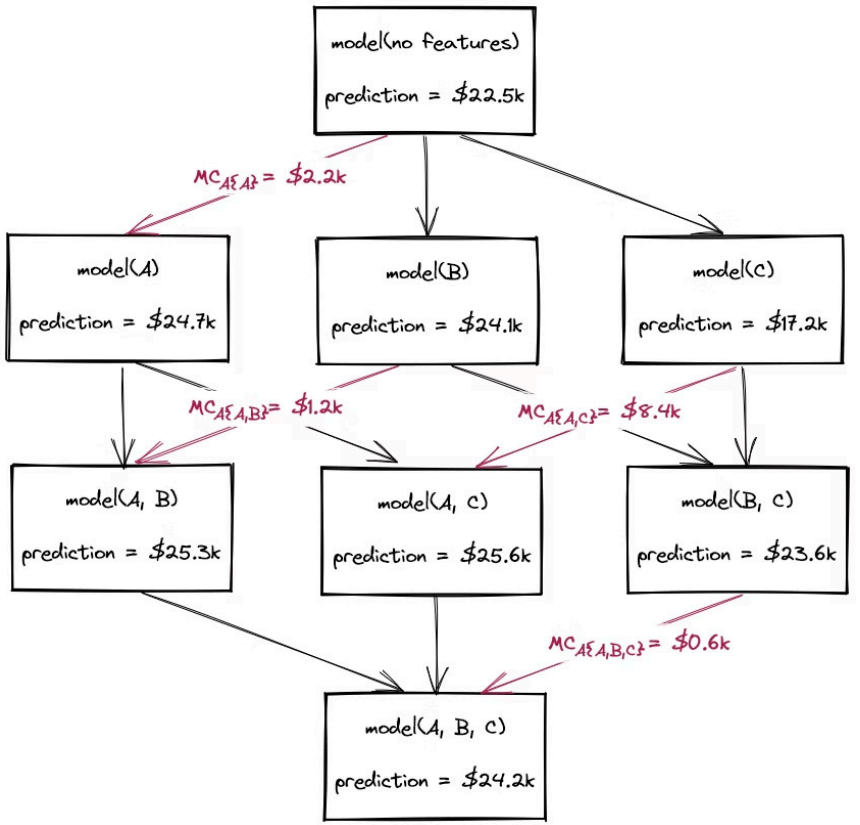
Ensures consistent and fair distribution of feature importance, allocating credit or blame proportional to each feature's contribution.

3

Additive Feature Attribution

The sum of SHAP values for all features equals the difference between the model's prediction and the expected baseline prediction.

How SHAP Works



Shapley Values

For each feature, calculate the average contribution it makes to the prediction over **all possible feature combinations, or coalitions**.

Computational Problems

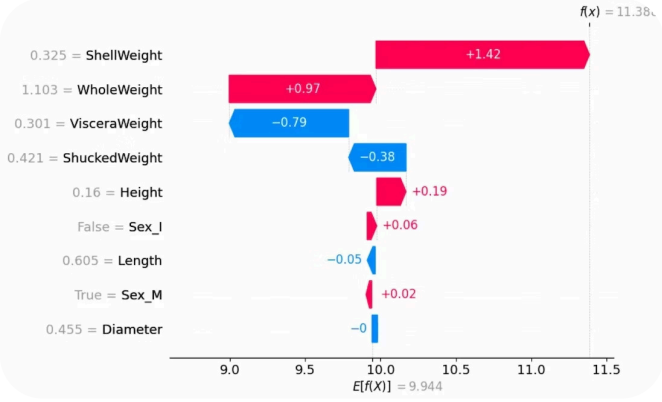
To calculate every Shapley value, you'd need to train the model 2^n times, and SHAP solves this issue.

SHAP Values

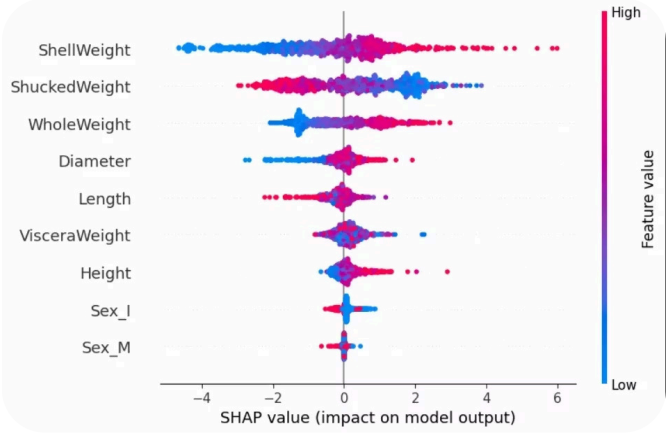
The most common SHAP method uses sampling to estimate Shapley values, using linear regression as a shortcut.

Visualizing SHAP: Common Plots

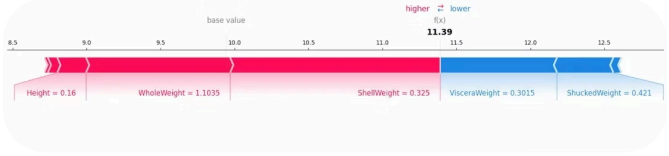
Waterfall Plot



Waterfall Plot



Force Plot



Pros and Cons of SHAP

Pros

- Makes potentially confusing ML models transparent
- Detects possible bias in predictions
- Useful to debug and tune model
- Compatible with any model type

Cons

- Can be slow for big data or complex models
- Struggles with high-dimensional data
- Sensitive to model type and feature correlation

Key Takeaways

- Ability to explain ML models creates value for predictions
- SHAP is a model-agnostic feature importance method that distributes prediction among the features
- There are various effective visualizations for SHAP
- Like PIMP, it can struggle with big data or complex models

Google Colab

Thanks NDL!

1

Next Meeting: 11/2

Topic: Dimensionality Reduction