Program reskilling Data/AI PO
Specific session for SOCIETE GENERALE

5: Interpretability with Shapley values

Eric Benhamou

PSL UNIVERSITÉ PARIS Dauphine UNIVERSITÉ PARIS ENS ÉCOLE NORMALE SUPÉRIEURE MINES ParisTech

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

Outline

- How can we measure the importance of a feature in decision tree
- Introduce Shapley values
- Show on example how to explain models: represent global model structure with local feature importance value

Recall from previous session

- In the previous session, you learned the different AI modeling In this lecture, we learned about supervised learning and clustering in particular to help find similarities between data

- We discuss the concept of distance and the subjectivity of number of clusters

- In the lab, we presented KMeans

Goal of this session

- Learn how to understand the role and impact of features to understand the model

- We will present Shapley values that are the de factor standard for explaining GBDT

Class Machine Learning Example

Data:

Includes every admission to a substance abuse treatment center that is publicly funded by the government

Trying to Predict:

Which admissions are ”first time admissions” and which were “repeat admissions”

Goal:

To predict which individuals are at high risk for relapse upon admission to a substance abuse treatment facility

Measuring Feature Importance

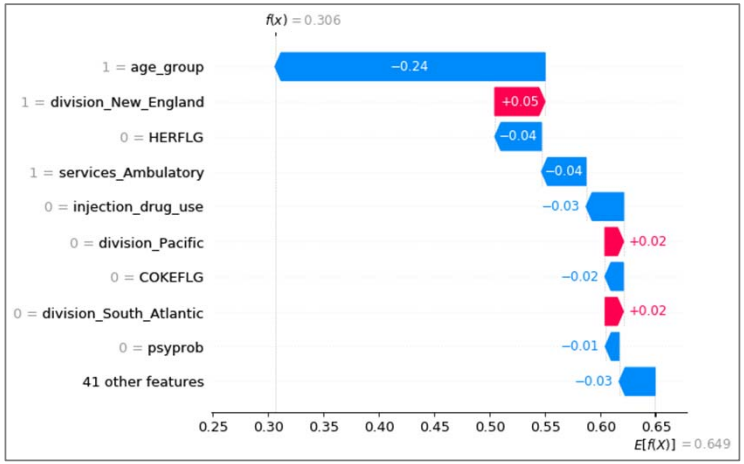
Feature Importance

- Feature importance measures the contribution of each feature/variable to the final prediction
- The higher the feature importance, the higher the impact of that feature on predictions
- There are multiple ways to measure feature importance
- We will cover these methods today:
  - SHapley Additive exPlainers
  - Tree-based model feature importance

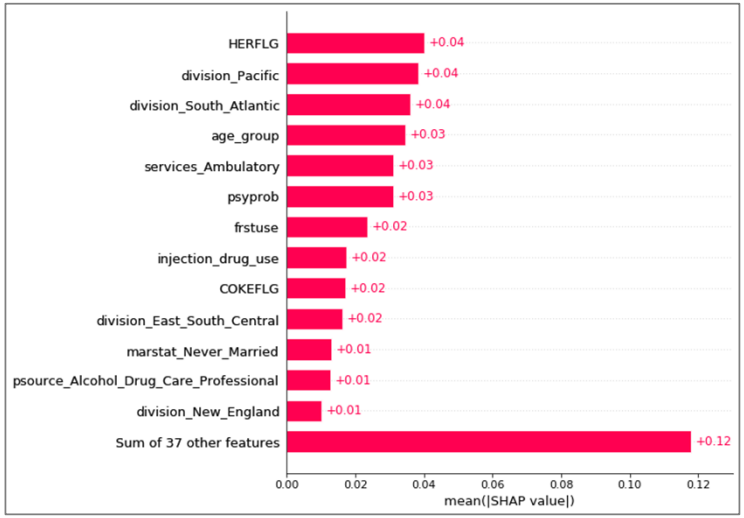
Global vs. Local Feature Importance

- There are two types of feature importance, global and local
- Global feature importance
  - Gives a summary of feature importance across all data points
  - Usually a mean across all observations
- Local feature importance
  - Gives the feature importance for one observation in your data set
  - Shows what features were most important for that specific data point

Local



Global



SHAP Values: The Gold Standard

- SHapley Additive ExPlainers (SHAP) are a model agnostic method for calculating feature importance
- Superior to regression coefficients because feature importance is measured on the same scale regardless of the range of the features
- SHAP values **cannot** determine causality
  - If causality is your goal, check out the econml package
- Several types of SHAP methods exist:
  - Model-agnostic
  - Tree-based
  - Linear
  - Neural networks

SHAP Concept Demonstrated

- **Example:** Three friends go out for a meal and share wine, fries, and pie. It is difficult to know how much they should pay since they all ate different amounts.

People Eating	Cost
Robert eating alone	$80
Alex eating alone	$56
Paul eating alone	$70
Robert and Alex eating together	$80
Robert and Paul eating together	$85
Alex and Paul eating together	$72
Robert, Alex, and Paul all eat together	$90

Take all combinations of each person in order and measure the incremental payout that would have to be made.

- Robert, Alex, Paul – 80, 0, 10
- Alex, Robert, Paul – 56, 24, 10
- Alex, Paul, Robert – 56, 16, 18
- Paul, Robert, Alex – 70, 15, 5
- Paul, Alex, Robert – 70, 2, 18
- Robert, Paul, Alex – 80, 5, 5

Robert: (80 + 24 + 18 + 15 + 18 + 80)/6 = $39.20
Paul: (10 + 10 + 16 + 70 + 70 + 5)/6 = $30.17
Alex: (0 + 56 + 56 + 5 + 2 + 5)/6 = $20.67

All Sum to ~$90

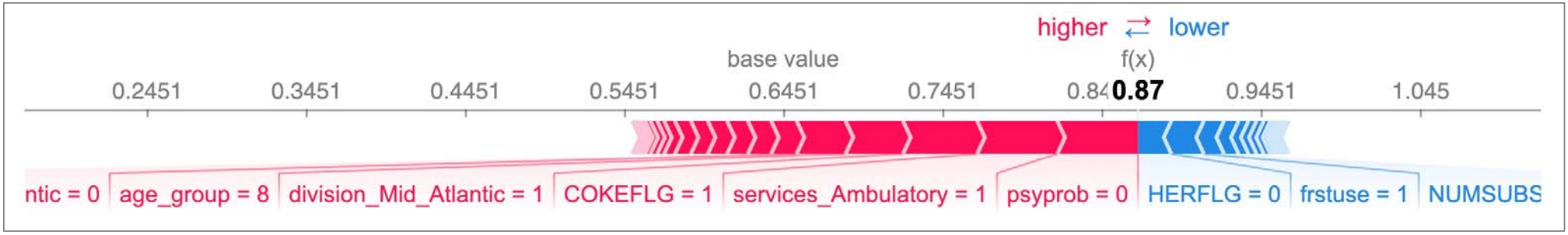
Source: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

SHAP: Detailed Explanation

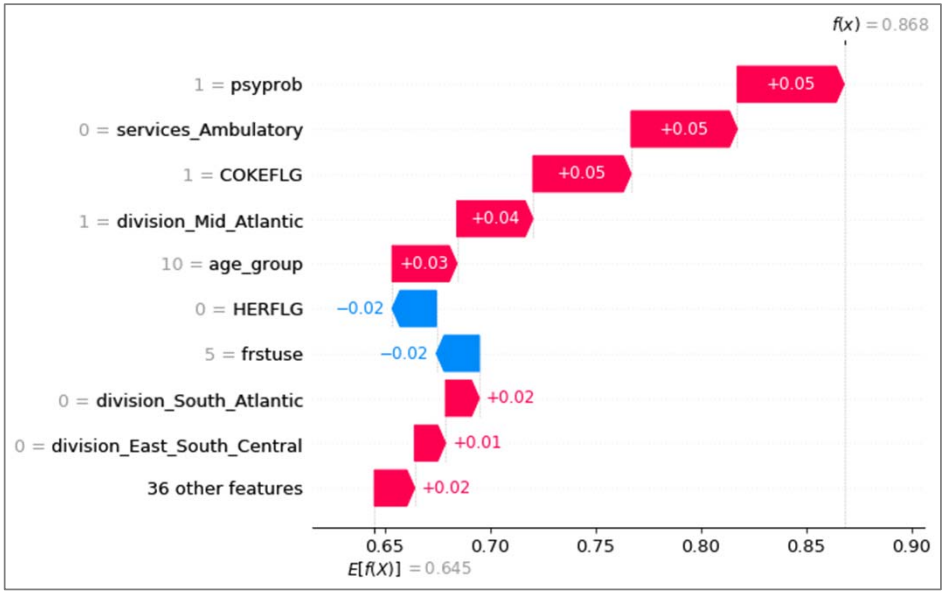
- To get the SHAP value of feature X the package:
  - Retrieves all combinations of the remaining features with X removed
  - Computes the effect on predictions of adding feature X to those subsets
  - Removed features are replaced with the average value across the datasets
  - Aggregates all contributions to find **marginal contributions** of a feature
- The model is **not retrained** for each combination of features

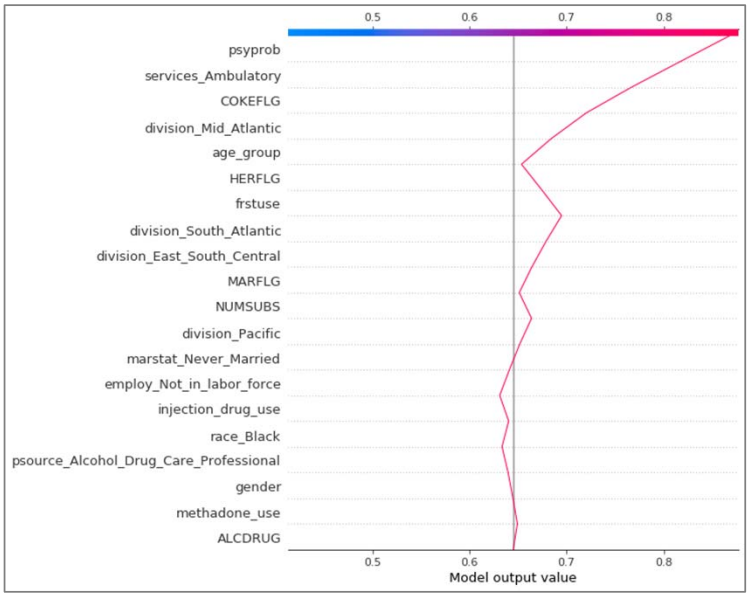
SHAP Example

- For every observation we start with the **base value**, or the average prediction across all observations
- The output value (shown as f(x) below) is the prediction for that observation
- Red areas show features push the output value higher compared to the base value
- Blue areas show features that push the output value lower than the base value
- Space for each feature shows the marginal contribution of that feature
  - For example, not using heroin reduces the likelihood of the instance being a repeat admission and in contrast living in Mid Atlantic increases that likelihood



Waterfall and Decision Plots

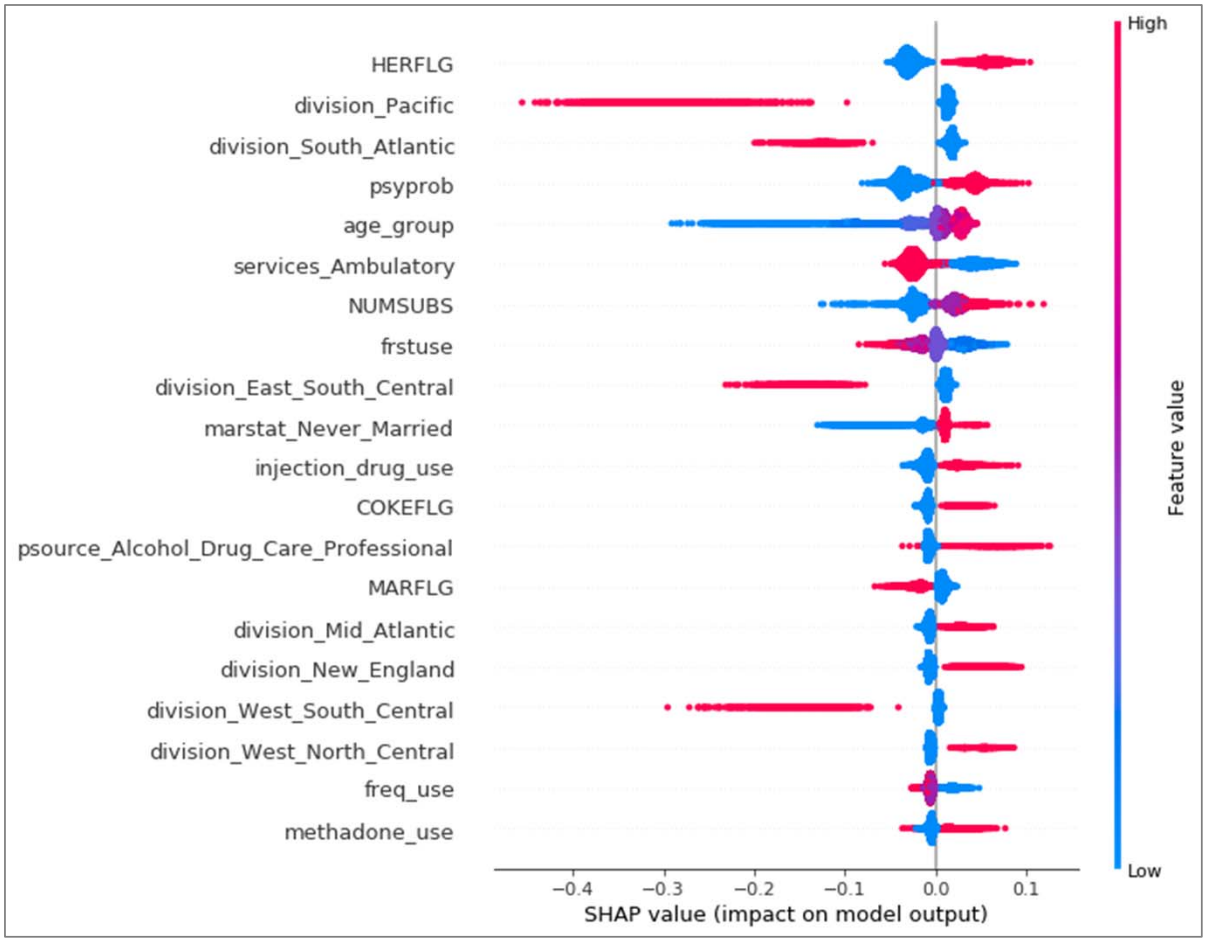




- Plot local SHAP values for each feature
- The actual value is shown at the top
- The average base value is shown at the bottom
- Two ways of displaying the marginal contribution of each feature

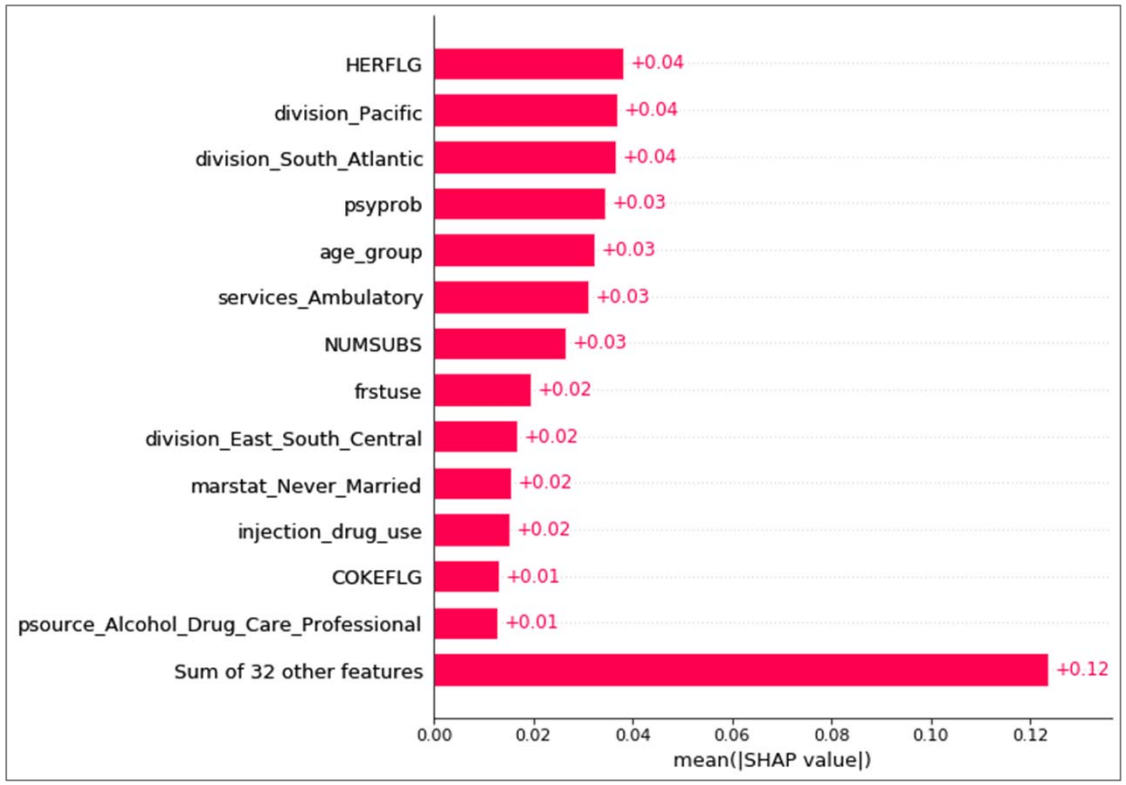
Beeswarm Plot

- Plots local SHAP values for each feature
- Red indicates whether a feature value is high compared to the mean
- Blue indicates a low value
- The SHAP value is shown on the x-axis
- Example: Being from the Pacific Division (1 is the “high value”) for most people is associated with a lower likelihood of a visit being a repeat admission



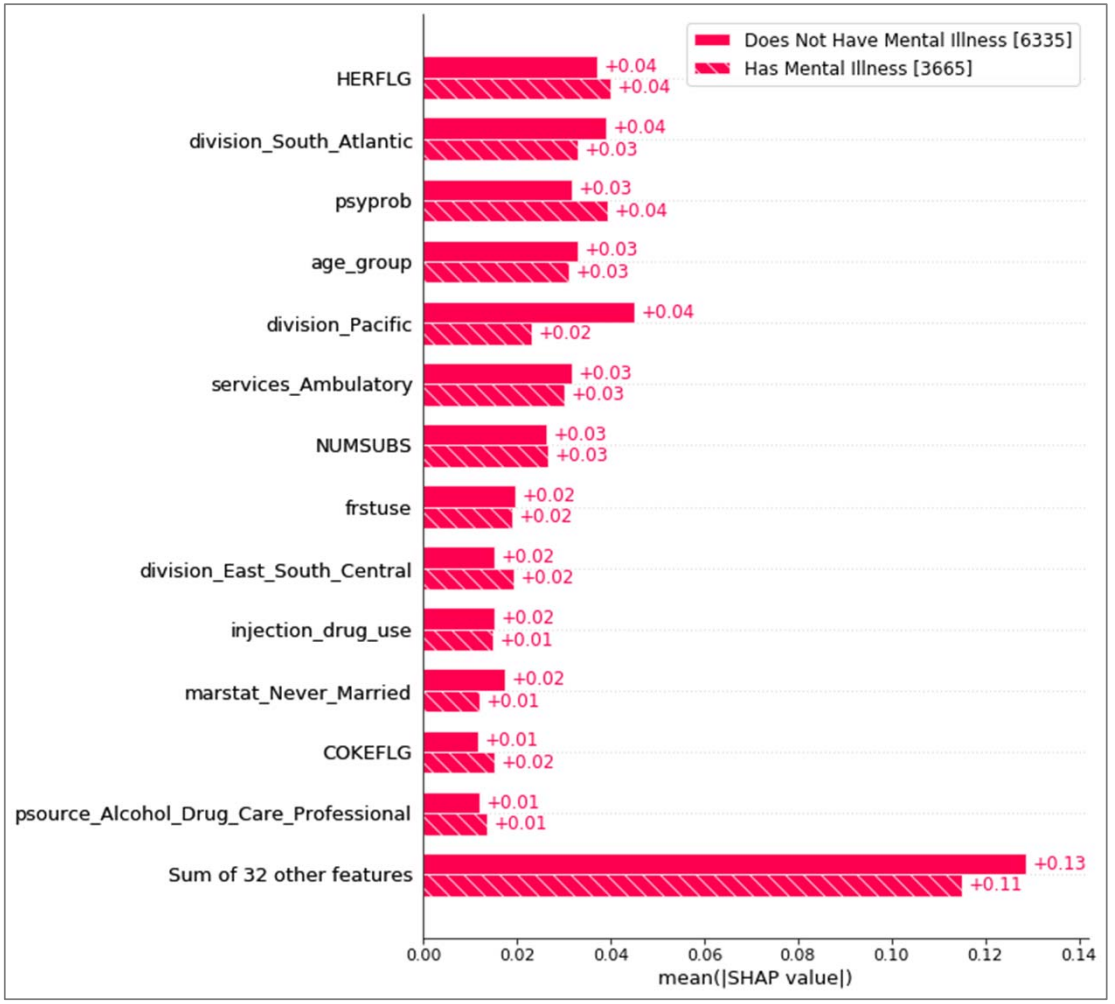
Bar Chart

- This chart shows global feature importance across all people
- Displays the mean SHAP value
  - NOTE: The absolute value is taken of all SHAP values to show importance, regardless of whether it is negative or positive
- Using or not using heroin is the most important feature
- Its marginal importance is 0.04 meaning it on average moves the predicted probability 0.04 percent away from the average base value



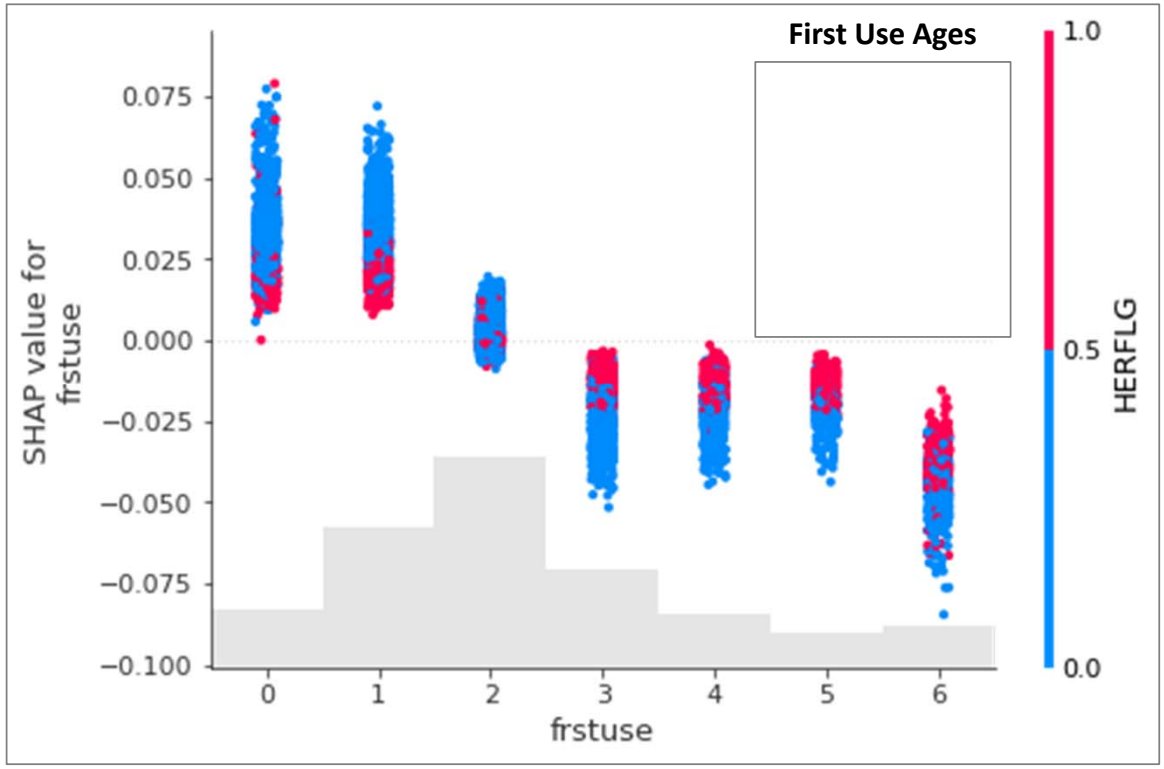
Comparison Bar Chart

- This bar chart shows SHAP values broken out by whether someone has mental illness
- Allows you to compare the most important features across groups
- For example, living in the Pacific Division is 0.02 less important if someone has mental illness
- Useful for identifying model bias across gender, age, ethnicity, and racial groups



Scatter Plots

- Shows the interaction between using heroin and age of first drug use
- If drug use starts before age 18:
  - The “age of first use” variable contributes to higher predictions
  - The contribution of heroin use is mixed
- If drug use starts at 18 or later:
  - The “age of first use” variable contributes to lower predictions
  - The heroin use variable contributes to higher predicted probabilities



Tree-Based Model Feature Importance

- With large data sets, computing SHAP values is often too processing intensive
  - There is currently no Spark function for calculating SHAP values meaning for big data it is often difficult or impossible
- Another option is to calculate feature importance using Gradient Boosted Trees and Random Forest
- This method produces similar but less accurate models than variable selection with SHAP feature importance
- These models by design:
  - Create many decision trees with a subset of available features
  - Produce feature importance by measuring changes in impurity when features are added and removed

Method Pros and Cons

SHAP Methods

Pros	Cons
More accurate than other methods	Highly processing intensive (see LIME for a similar faster method)
Allows calculation of local feature importance	
Has tree-based and linear methods to speed up process of these model types	
Has awesome visualizations for model explainability	
Allows for the detection of bias in machine learning models	

Tree-Based Model Feature Importance

Pros	Cons
Not processing intensive	Not as accurate as SHAP methods
Calculated automatically during training, no need for additional code	Less effective for removing highly correlated variables

What Is Feature Selection?

Review: Overfitting

- When too many features are used, machine learning models often overfit
- When overfitting occurs, the model:
  - "Memorizes" the training data
  - Does not generalize well to unseen data
- This problem can be resolved through:
  - Model tuning
  - Feature selection

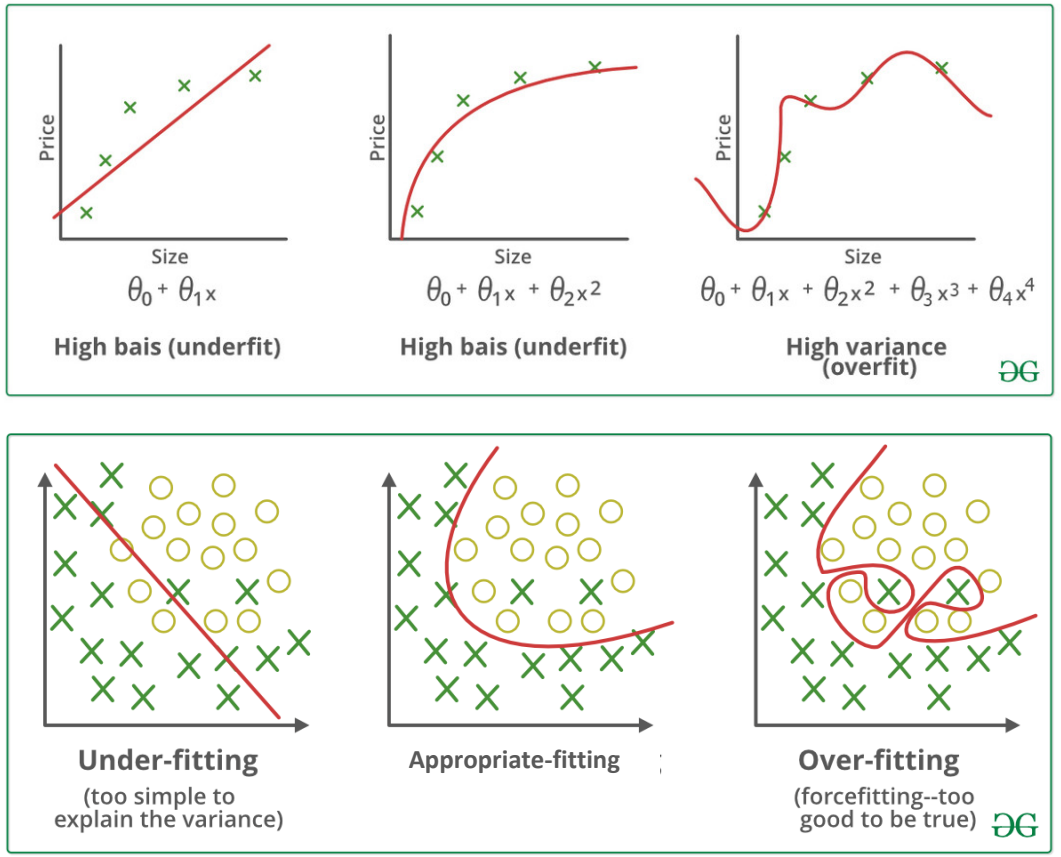


Image Source: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Feature Selection Overview

- Feature selection is a process where you choose the most important features to remain in your model
  - This avoids overfitting and the “memorization” of the training data
- It is also best practice to remove highly correlated features
  - Example: Having percent below the poverty line and median household income (both a measure of income) in the same model
- Goal: Include as few features as possible without allowing model evaluation metrics to drop.

Feature Selection Process

1. Measure feature importance.
2. Iteratively remove features and check for a drop in the f-score. Also keep an eye on precision and recall.
3. If the f-score drops, choose the feature list from the previous model run where the f-score was higher.
4. Include as few features as possible without a drop in f-score.

Gold Standard: Shapley Additive ExPlanations (or SHAP values)

With Big Data: Feature Importance from Tree-Based Models

Feature Importance Process

1. Recommend running both Gradient Boosted Trees and Random Forest for feature selection
   - Models have different calculation methods so will be less biased
2. Export feature importance
3. Find the mean importance for each feature across both models
4. Follow steps 2-4 on the previous slide

	features	importances
52	division_Pacific	0.248781
3	HERFLG	0.093954
53	division_South_Atlantic	0.083624
19	age_group	0.075080
25	psyprob	0.063049
58	services_Ambulatory	0.058877
18	NUMSUBS	0.055542
56	division_West_South_Central	0.030200
51	division_New_England	0.027811
40	psource_Alcohol_Drug_Care_Professional	0.026582
48	division_East_South_Central	0.026302
55	division_West_North_Central	0.022596
1	COKEFLG	0.022016
26	frstuse	0.019661
49	division_Mid_Atlantic	0.018975
31	injection_drug_use	0.018301
2	MARFLG	0.012717
71	livarag_Independent_Living	0.009598
17	ALCDRUG	0.008509
63	marstat_Never_Married	0.008140
21	methadone_use	0.007516
27	freq_use	0.007508
61	marstat_Divorced_or_Widowed	0.006811

Well done!

- You have leant what is Shapley value and how to understand a model

Summary

- In this session, we learned about Shapley values and how to interpret models using shapley values

- This concludes my intervention