

Visual Methods for Finding Human Bias in Machine Learning Models

Jim Box

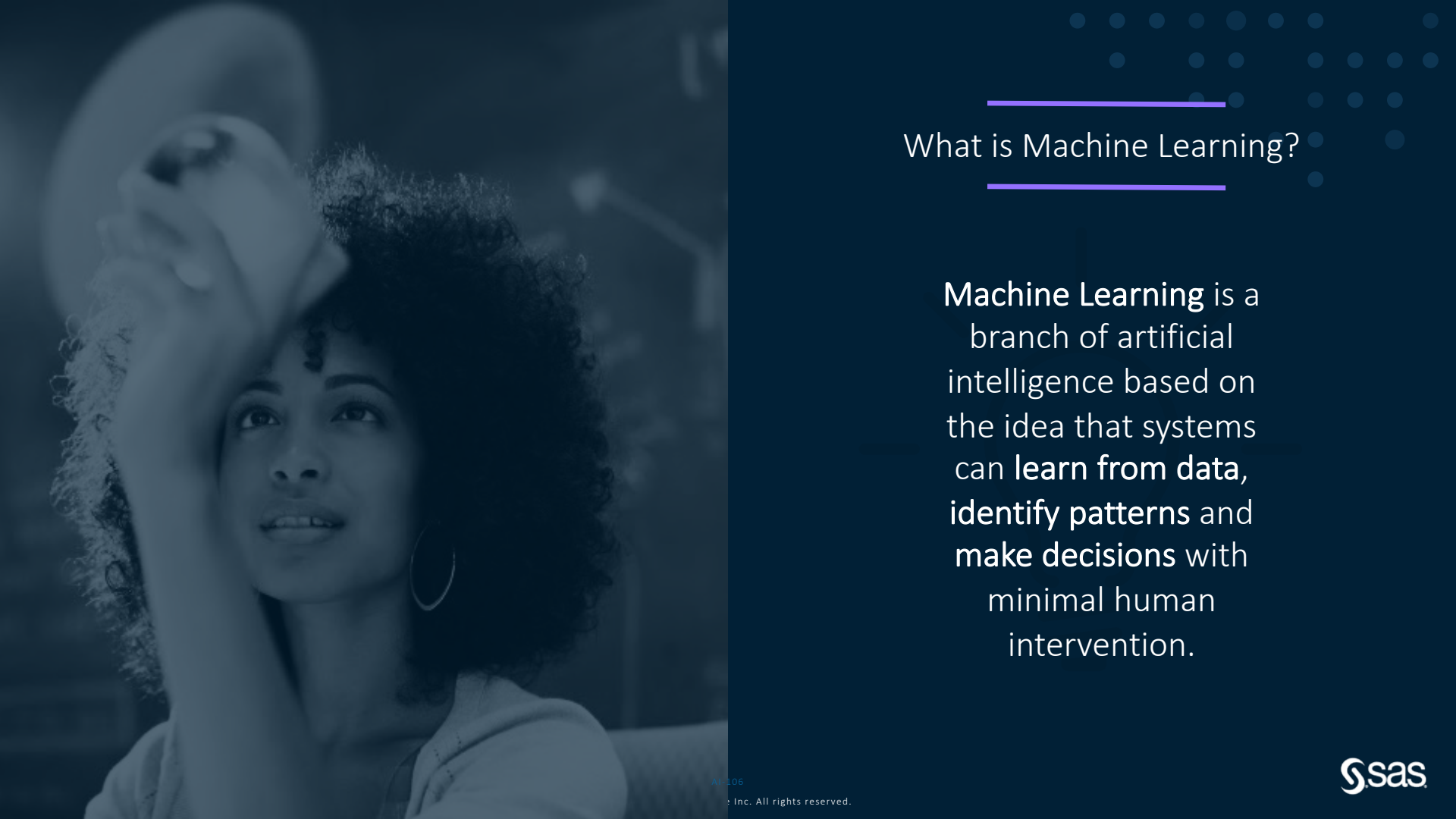
Principal Data Scientist, HLS R&D

SAS Institute

Analytics & Statistics Stream: AS-01

Bias In Machine Learning

Sources and Concerns



What is Machine Learning?

Machine Learning is a branch of artificial intelligence based on the idea that systems can **learn from data**, **identify patterns** and **make decisions** with minimal human intervention.

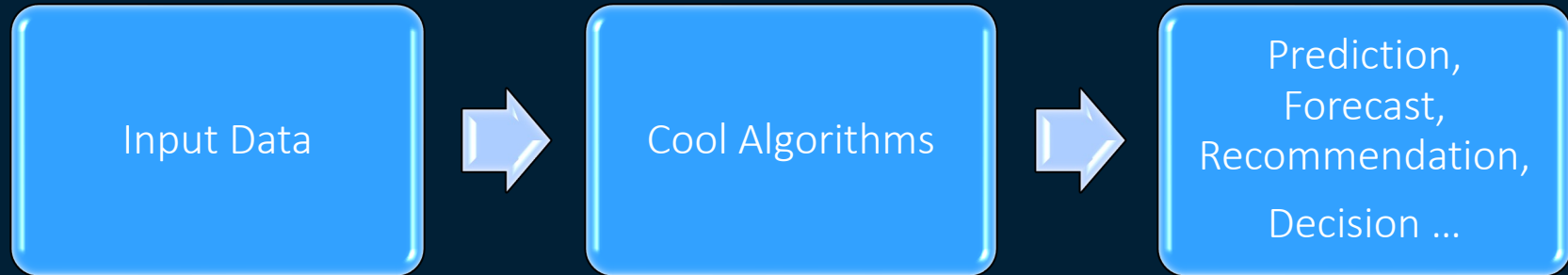
Recognizing the Risks

The Impact on Society

Financial Resources	Criminal Justice	Health Care	HR
 Who gets approved for a loan?	 Who gets a harsher sentence?	 Who gets admitted to a care management program?	 Who gets moved up to the next stage of the hiring process?

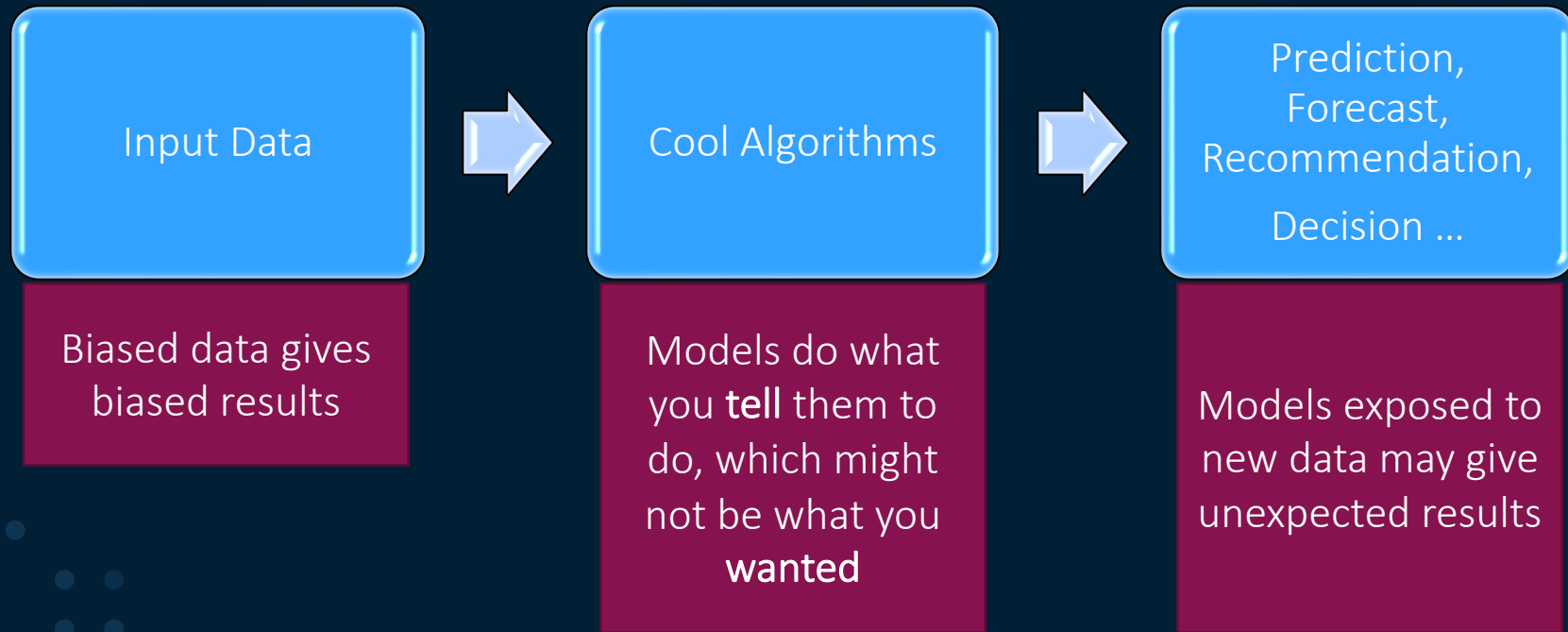
How does it work?

Typical ML Training



What can go wrong?

Beware



Bias In Machine Learning

Detecting Bias in a Model

Detecting Bias

Questions to Ask

Is the data representative?

Is the model accuracy different across different groups?

Are sensitive variables (or their proxies) being used in the models?

Is the Data Representative

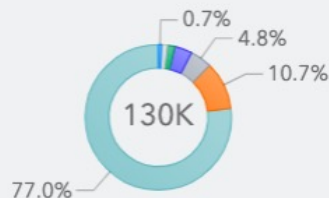
Where did the training data come from

Gender Breakdown



gender
■ M ■ F

Ethnicity Breakdown



ethnicity

Racial/Ethnic Representation

% Asian

1.6%

% American Indian/Alaska
Native

2.3%

% Black

9%

Female Representation

50%

% Caucasian

76%

% Hispanic

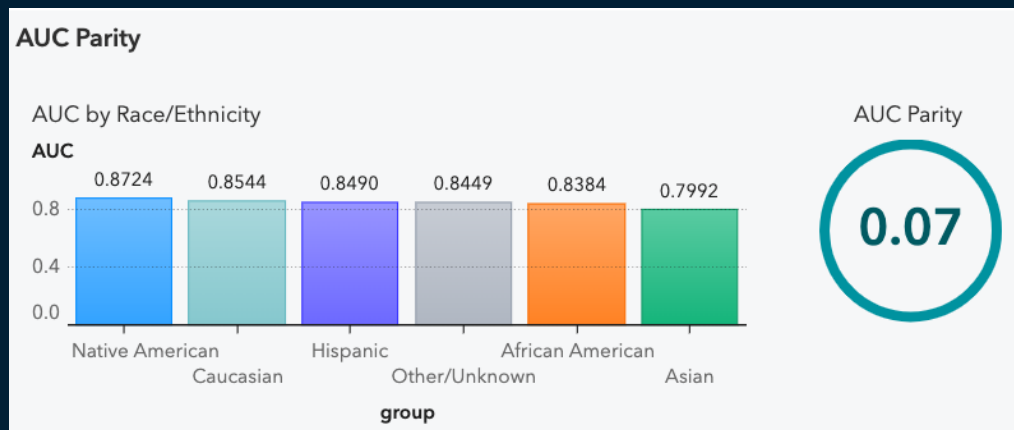
9.7%

% Native Hawaiian/Other Pacific
Islander

0.15%

Model Accuracy Across Groups

AUC Parity



- Area under ROC curve for each group of interest
- AUC Parity is the maximum difference between the groups
- Ideally would be close to 0

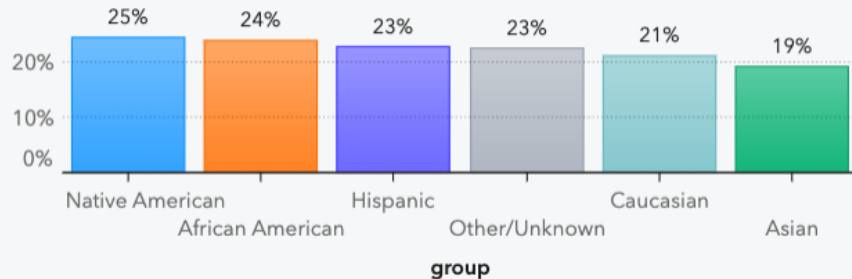
Model Accuracy Across Groups

Predictive Parity

Predictive Parity

Average Predicted Probability for Diabetes (Y=1) by Race/Ethnicity

AVG_P_diabetes_mellitus1



Predictive Parity

0.05

- Maximum difference in Average Predicted Probability across groups of interest

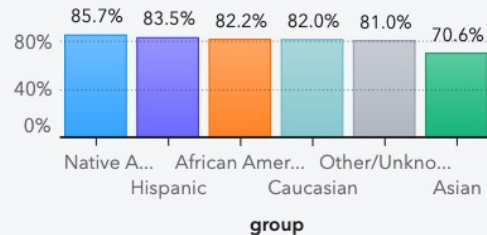
Model Accuracy Across Groups

Equalized Odds

Equalized Odds

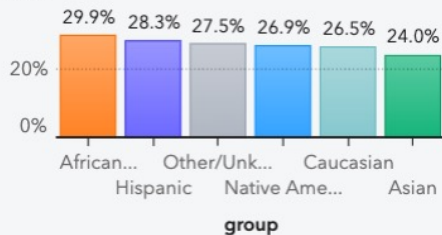
True Positive Rate

TPR



False Positive Rate

FPR



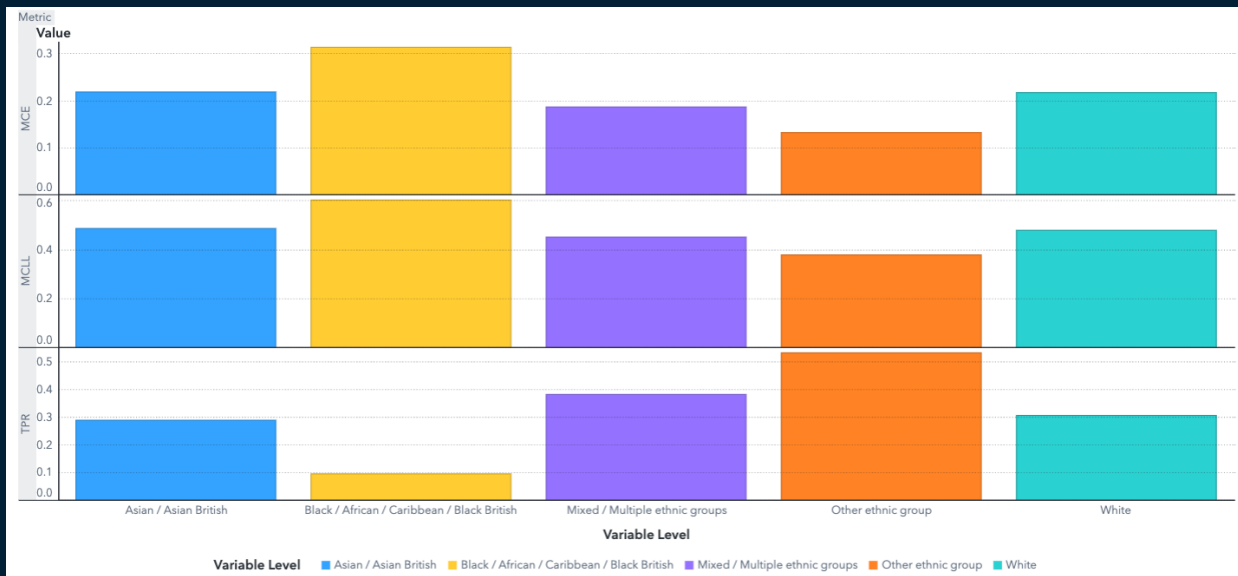
Equalized Odds



- Compare the spread across the True Positive and False Positive Rates
- Take the Maximum

Model Accuracy Across Groups

Other Measures of Interest



- MCE – Misclassification Rate
- Multi-Class Log-Loss (compares predicted to actual)
- TPR – True Positive Rate

Model Accuracy Across Groups

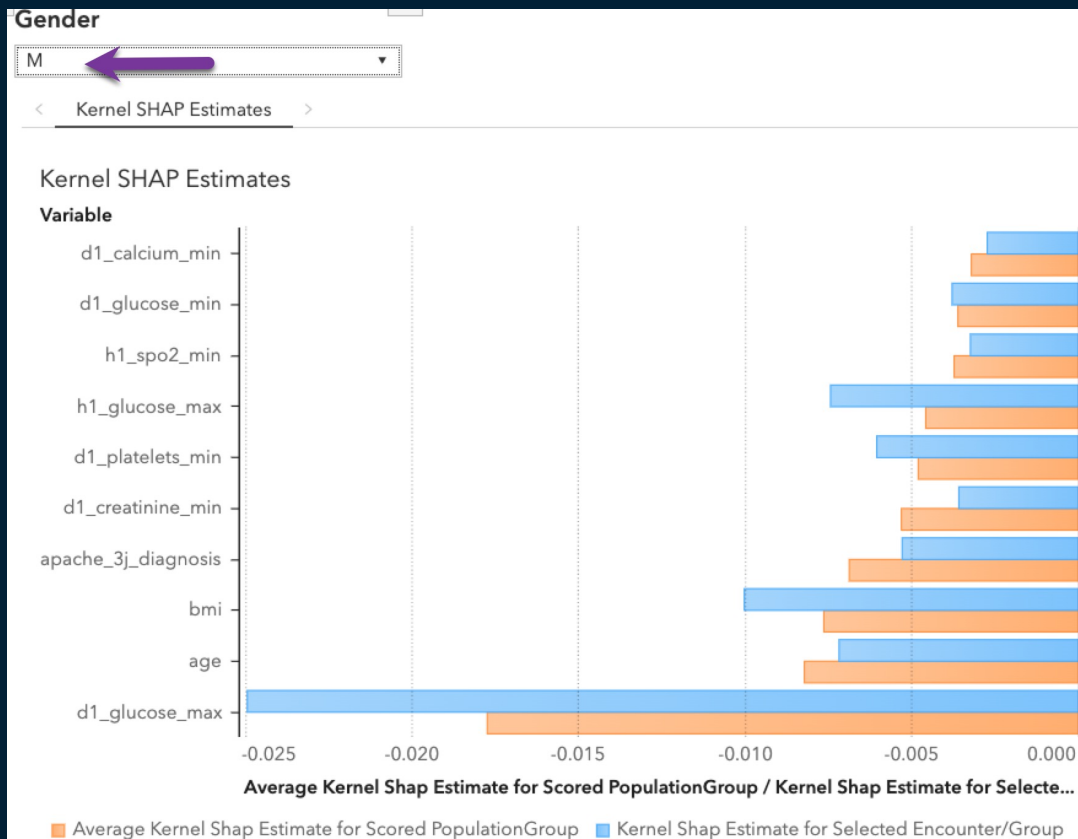
Kernel SHAP Estimates



- Looks at how variables impact the average predicted probability
- Orange Bars: Impact of variable on entire population
- Blue Bars: Impact of variable on selected group

Model Accuracy Across Groups

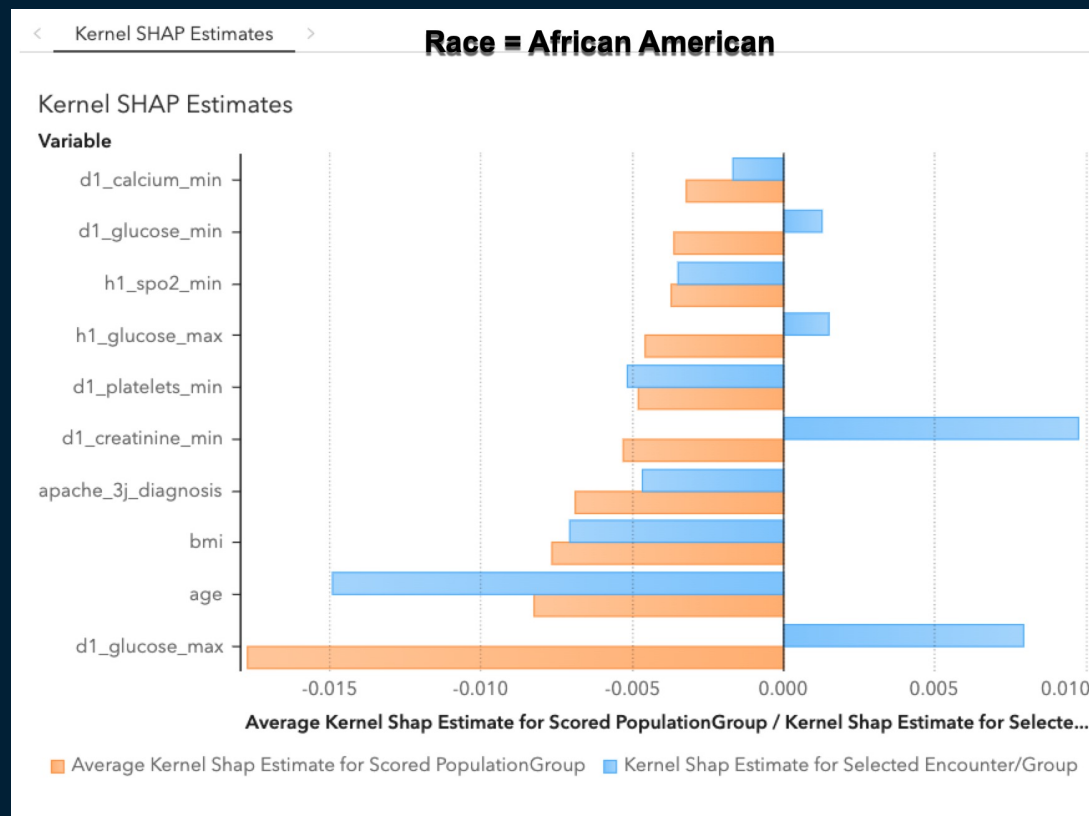
Kernel SHAP Estimates



- Looks at how variables impact the average predicted probability
- Orange Bars: Impact of variable on entire population
- Blue Bars: Impact of variable on selected group

Model Accuracy Across Groups

Kernel SHAP Estimates



Are Sensitive Factors Being Used

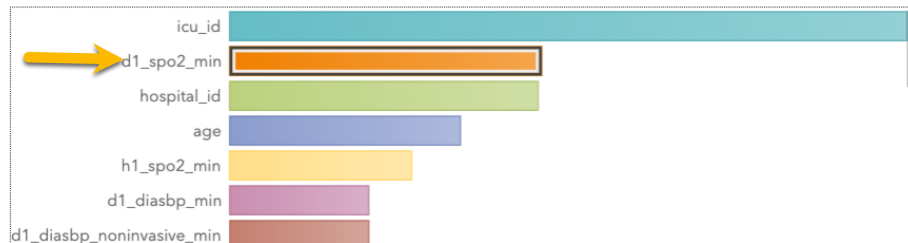
Check the Proxies!

What are the characteristics of ethnicity?

African American is the second most common value representing 10.69% (14K of 130K) of ethnicity. The three most related factors are icu_id, d1_spo2_min, and hospital_id.

Caucasian and African American are much more common than all other ethnicity values, together representing 87.70%.

What factors are most related to ethnicity?



What are the groups based on d1_spo2_min by the chance of ethnicity being African American?

< High Low >

61.84%

If d1_spo2_min is greater than or equal to 96, hospital_id is between 175 and 176, then ethnicity has a 61.84% chance (637 of 1K cases) of being African American.

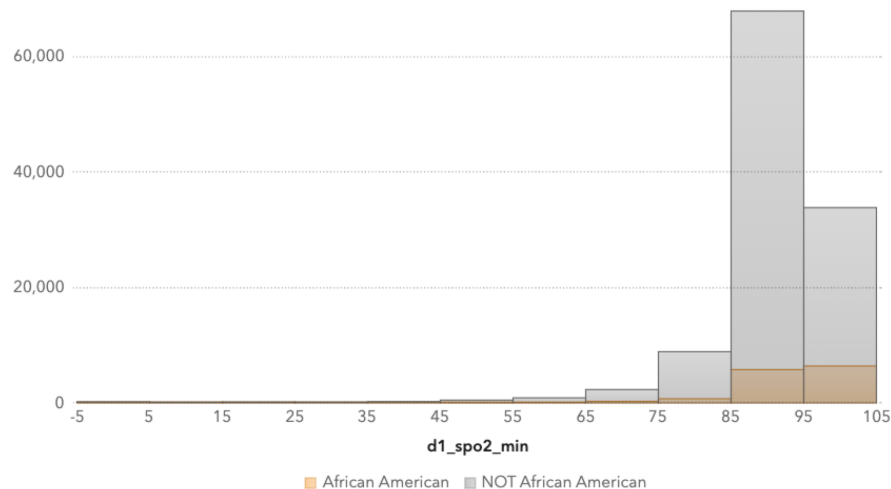
60.33%

If icu_id is between 1.1K and 1.1K, d1_spo2_min is greater than or equal to 95 or Missing, then ethnicity has a 60.33% chance (794 of 1.3K cases) of being African American.

51.30%

If icu_id is greater than or equal to 1.1K, d1_spo2_min is greater than or equal to 98 or

What is the relationship between ethnicity and d1_spo2_min?



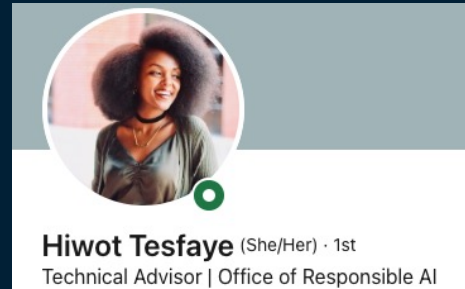
Mitigation

What can we do when we find Bias

- Consult subject matter experts, end users, and affected groups to see if the model process is even appropriate given the data available
- Explore oversampling/undersampling techniques if the problem is due to unbalanced groups in the data
- Consider segmenting the data and making different models for different segments
- Eliminate sensitive variables (and their proxies)

Thanks!

Jim.Box@sas.com



sas.com

