RESPONSIBLE AI

Week 7: Replication Project 3 – Bias Mitigation

Meira Gilbert and Nandita Rahman

TODAY'S OBJECTIVES

- Types of bias in model development
- Mitigating fairness issues in models through pre-, in-, and post-processing
- Re-learning and re-deploying models
- Explainability methods
- Replication Project Overview

- **STRICT 5 minute presentations today!**

Replication Part 01 Review

EDA Expectations and Review

- EDA Section should be same quality or better as a DSC80 project
- If you do not produce a decision/takeaway from the analysis, it's the same impact **as if you never performed the analysis at all**
- The purpose is to understand the context of your question
  - "Can we predict whether a person will have high utilization? Can we improve upon the bias/fairness of our initial classifier?"

Replication Part 01 Review

Advice for EDA

- Always start by visualizing data distributions.
- **Nulls are more than “null”**
- Justify your decisions for handling nulls and outliers.
  - Are you imputing? Why? (DSC80 has the answer!)
  - You must **state the assumptions that are supported by the data** (normality, reasoning for missingness) to determine which action to take.
  - **Tell a story. Use the notebook format to your advantage.**

Overall Comments

Remember what you learned about distributions, charts, correlation, etc. The quality of output for this EDA section should be at the level of a DSC 80 report.

If you don’t produce a decision / takeaway from the analysis you conduct, it’s the same impact as if you never performed the analysis at all!

Use notebooks effectively – they’re not just code files. Use multiple cells, add a bunch of written sections/comments to guide the reader (this includes yourself a few weeks into the future) through what you’re doing and your takeaways.

READING PRESENTATION #1

Model Cards for Model Reporting (Mitchell et al.)

READING PRESENTATION #2

Fairness Through Awareness
(Dwork et al.)

BIAS DETECTION WITHIN AI SYSTEMS.

Types of Algorithmic Bias

Historical bias: misalignment between the world as-is and the values or objectives required from the ML model;

Representation bias: under-representation or failure for a population to generalize for groups in population;

Measurement bias: choosing and utilizing features/labels that are noisy proxies for real-world quantities;

Aggregation bias: inappropriate combination of heterogeneous, distinct groups into a single model;

Evaluation bias: use of inappropriate performance metrics or the testing / external benchmark that does not represent the entire population; and

Deployment bias: inappropriate use or interpretation of model in a live environment.

Model bias is commonly thought to occur **primarily** in the data collection step.[1]

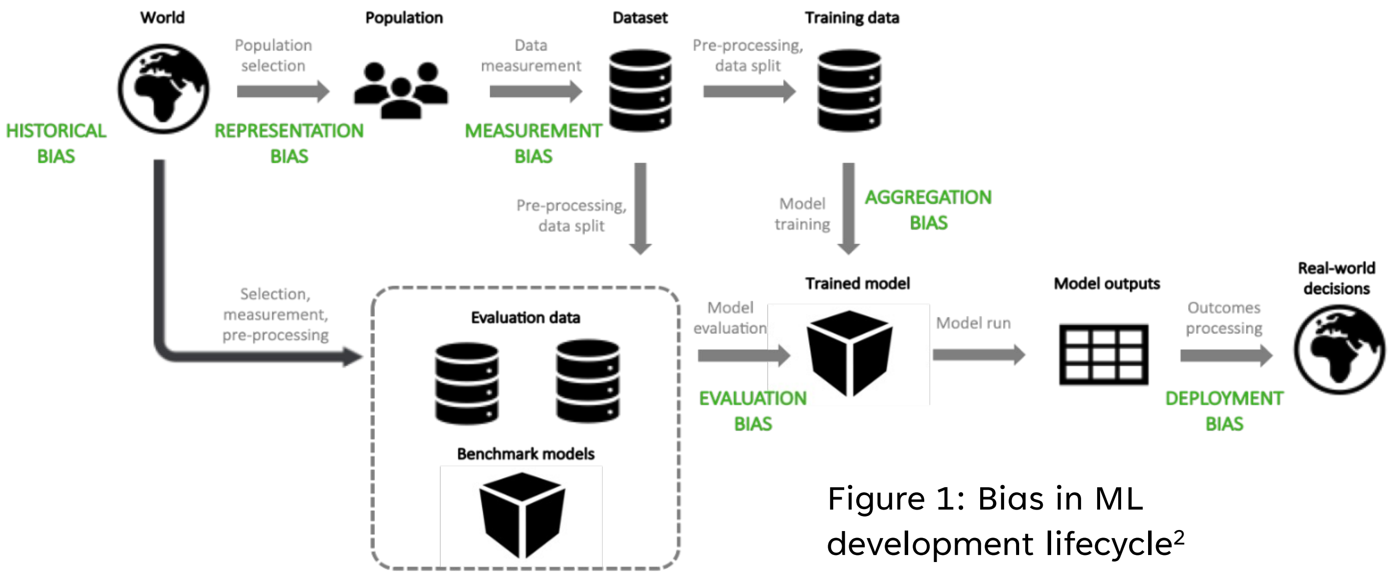


Figure 1: Bias in ML development lifecycle[2]

Recent research suggests that bias can be introduced at **any stage** in the machine learning model life cycle (see Figure above).[2]

"ZOOMING OUT" - *HEALTH EXPENDITURE USE CASE*

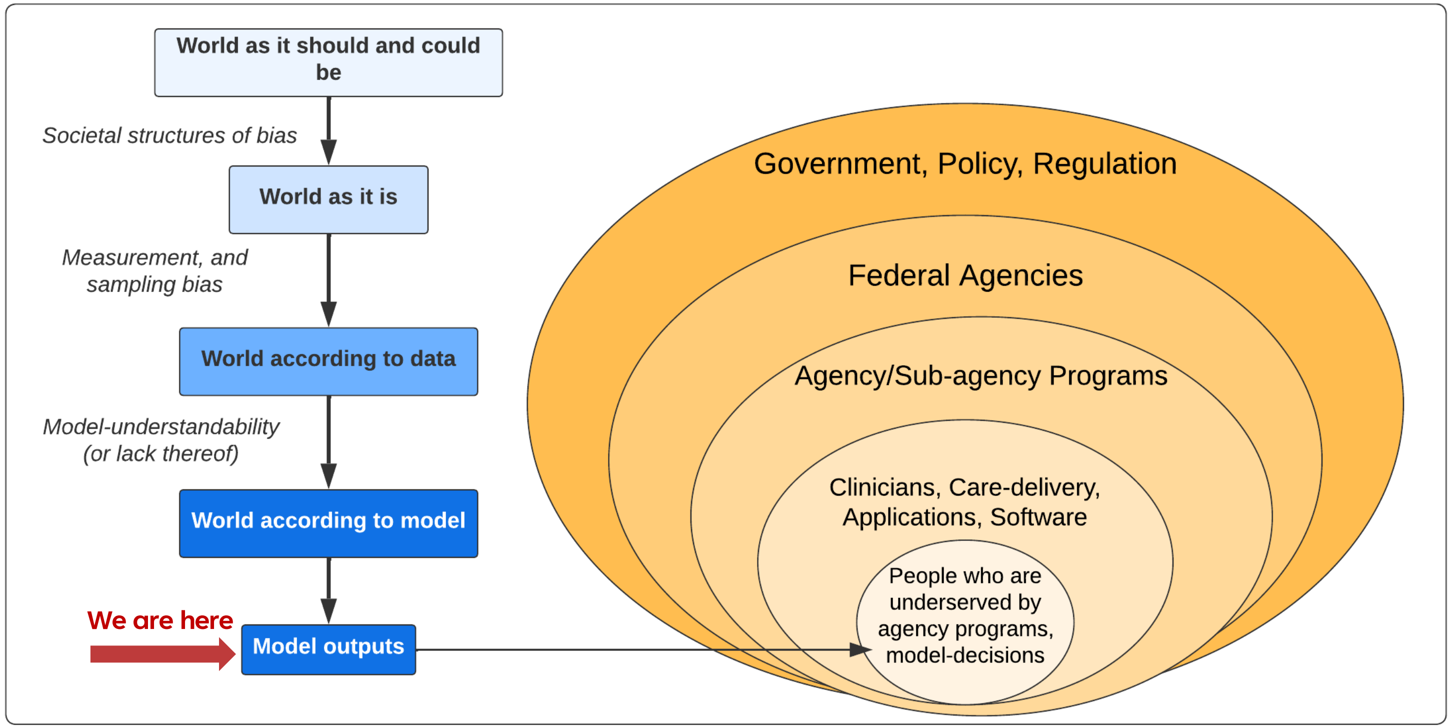


Figure above: Illustration of the types of bias that can enter a model's decision-space (stages of data and algorithm use), resulting in model-outputs that affect individuals downstream. *(Adapted from: Mhasawadade 2020, Mitchell 2020)*

Bias and Algorithmic Fairness

If classifiers tied to model predictions are biased or not well-understood, this could lead to differentials in delivered action.

Composite scores for classifiers such as '**utilization**' may indicate higher risk for certain groups over others – but is this a true representation of the real world?

Therefore, it is important to understand **different types of bias** and which de-biasing methods are best suited for your model.

Overrepresentation of a particular race in the training data will return more opportunity to learn fine-grained information about that group compared to others, despite **not explicitly including race as a feature.**

How would this affect Non-White beneficiaries who could have risk factor predictors that could inform a model to prioritize additional care? **How could it do the opposite?**

SOURCES OF BIAS[1,2,3] IN AI AND HEALTH DATA – OPTIONAL SLIDE

Bias within AI in health can often hold co-existing sources both (technical and non-technical)[3,4].

(Taken from literature) are example "factors," that have shown evidence of introducing bias during model development and use[1].

We prioritized factors that demonstrated contributing to:

Negatively impacting group(s) disproportionately relative to general population[4,5]	Posed a barrier for **advancing health equity**[6,7] within historically marginalized group(s) or communities of concern[9,10]

Stages Where Bias/Disparate Impact Can Arise During Development and Use of AI in Health

STAGE

Problem Selection

DESCRIPTION

Is disparate impact or bias the result of a field lacking cultural diversity, funding or being understudied?[1,11]

FACTORS

- Racial Bias[12]
- Gender Bias [13]
- Workforce Diversity [11]
- Lack of research funding [1,14]

Data Considerations

How can the type, dimension, collection-method, and representation in data lead to bias and disparate impact in communities of concern? [15,16]

- Data Privacy[17]
- Social media data[1]
- Electronic health records[18]
- Population-specific data loss/missingness[1,16,17]

Outcome Definition

What social factors are often overlooked when developing AI-targets or outcomes for the general population?[21] How do they delay access and quality of care to underserved populations?[22]

- Billing codes/Label noise [23,24]
- Access to care program[25]
- Predictors used to appropriate funding[25]
- Healthcare costs[26]

Algorithm Development

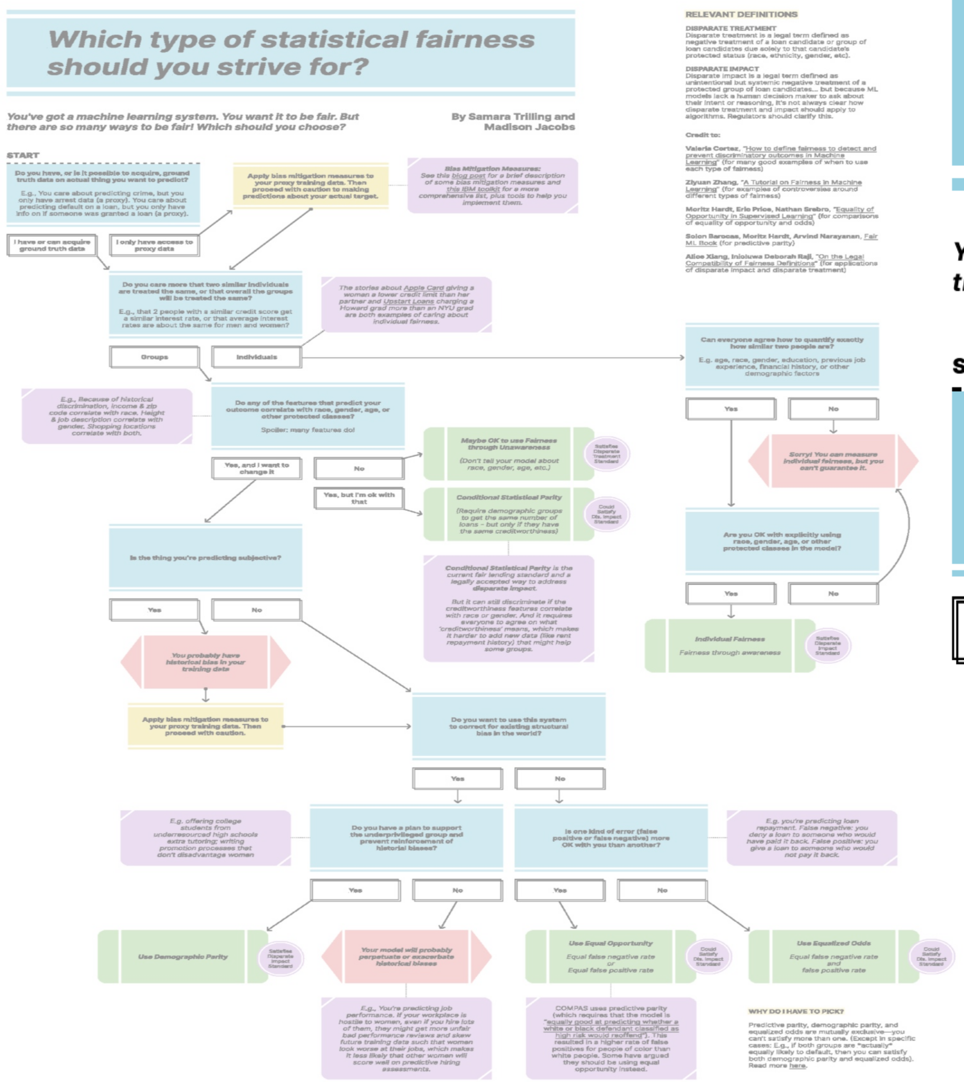
What factors must be considered during AI model-development and performance? How and where are they vulnerable to introducing bias? [1]

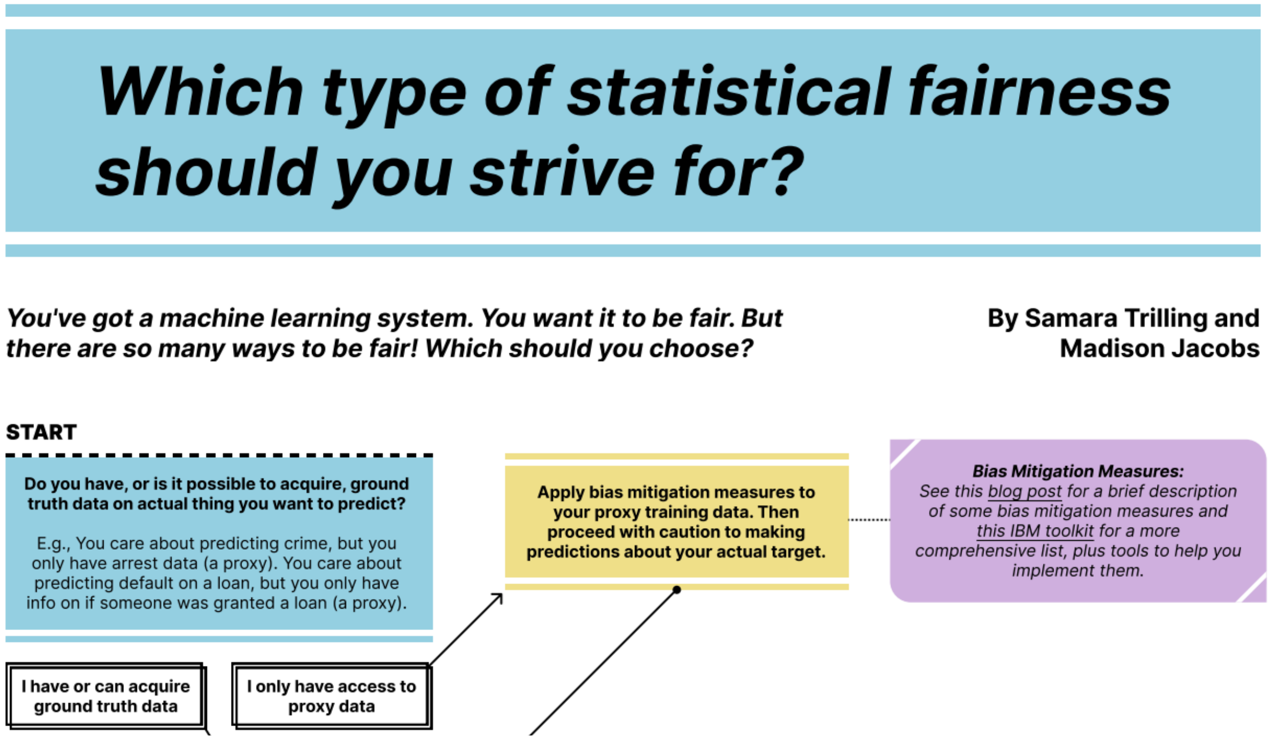
- Deployment Infrastructure [3]
- Bias detection [27]
- Feature Selection[28]
- Performance metrics [28,29, 31]
- Training, Testing, Validation [27,29,30]
- Model Transparency [30,31]

Post-deployment considerations

How and what should be measured to assess downstream impact of AI, and what factors should be used to audit for bias and clinical impact[1]

- Model Documentation[1]
- Regulatory Compliance[1,3]
- Continuous Model Monitoring[2]
- Social Determinants of Health[3]





[Quickly revisiting last week's demo…]

https://www.aspentechpolicyhub.org/wp-content/uploads/2020/07/FAHL-Tree.pdf

DISCUSSION QUESTIONS

- How can we determine what bias mitigation techniques to use?
- What should we do if we encounter unexpected results from our bias mitigation?
- Given our use case and data, what bias mitigation techniques seem reasonable? Unreasonable?

BIAS MITIGATION TECHNIQUES

Preprocessing

Used to mitigate bias in training data, before building model

- **Reweighing**
- **Optimized Preprocessing**
- **Learning Fair Representation**
- **Disparate Impact Remover**

In-processing

Used to mitigate bias in classifiers, while developing model

- **Adversarial Debiasing**
- **Prejudice Remover**

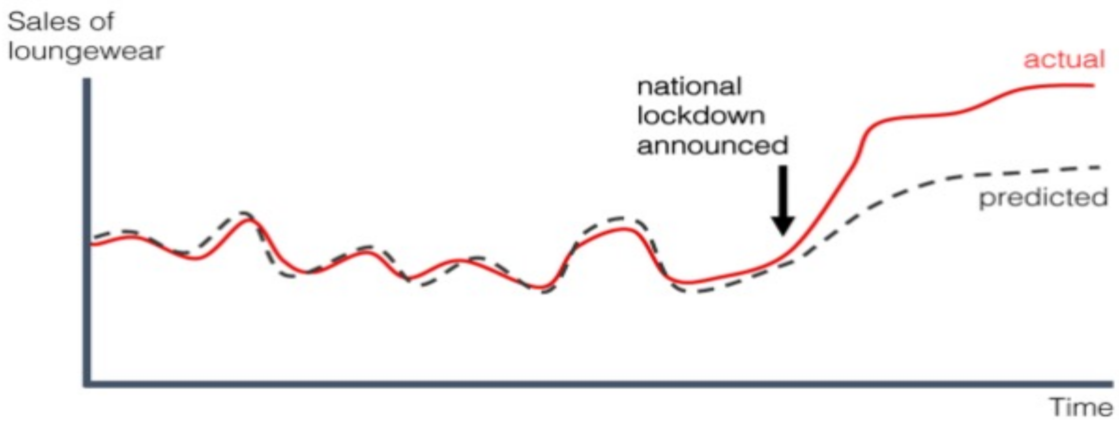
Post-processing

Used to mitigate bias in predictions/outcomes, after training the model

- **Equalized Odds Postprocessing**
- **Calibrated Equalized Odds**
- **Reject Option Classification**

RE-LEARNING AND RE-DEPLOYING MODELS

Data Drift



1. Deploy Model

Test model trained on 2014 (Panel 19) data on 2015 (Panel 20) data.

Does it exhibit fairness and maintain accuracy?

2. Re-Deploy Model

Test model trained on 2014 (Panel 19) data after **reweighing** on 2016 (Panel 21) data

Is there any drift?

3. Re-Learn Model

Re-learn model from 2015 (Panel 20) data.
Train and evaluate on **transformed** 2016 (Panel 21) data.

Is it relatively fair and accurate?

4. Re-Deploy Model

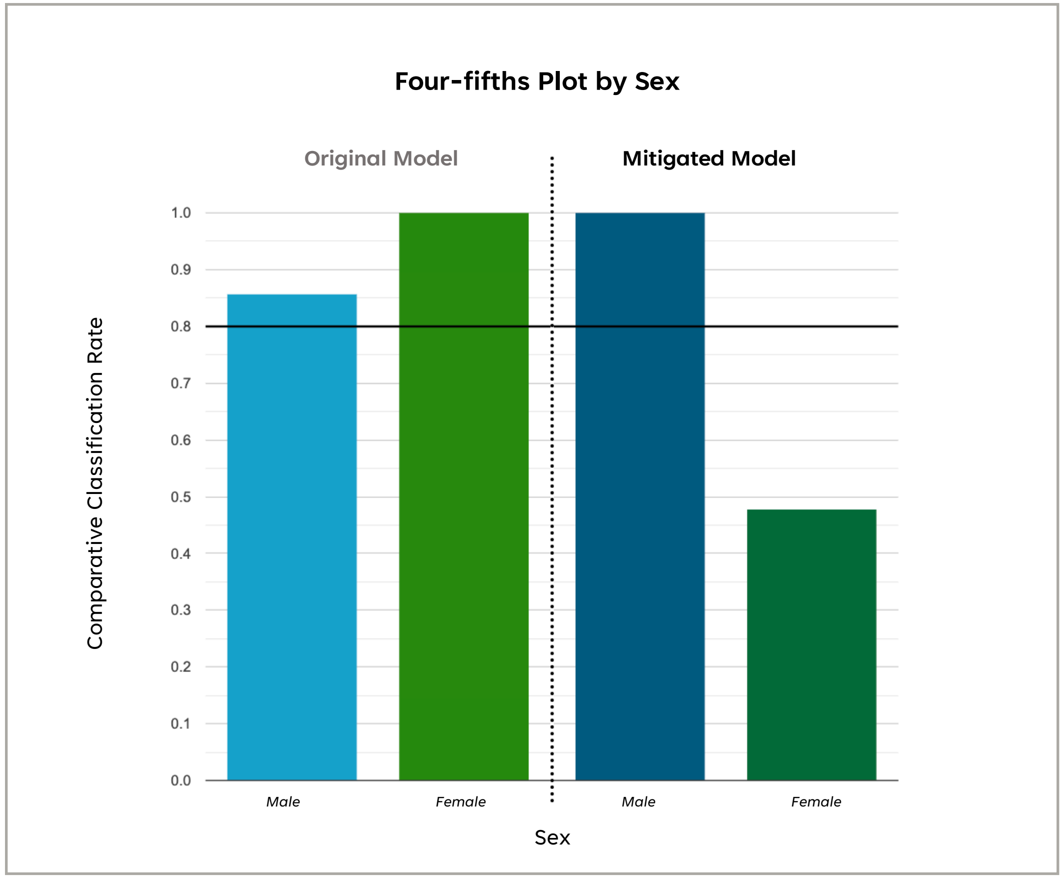
Test model trained on **transformed** 2015 (Panel 20) on 2016 (Panel 21) deployed data

Does it original fairness meet& accuracy specs?

FAIRNESS PLOT EXAMPLES WITH MITIGATED MODELS

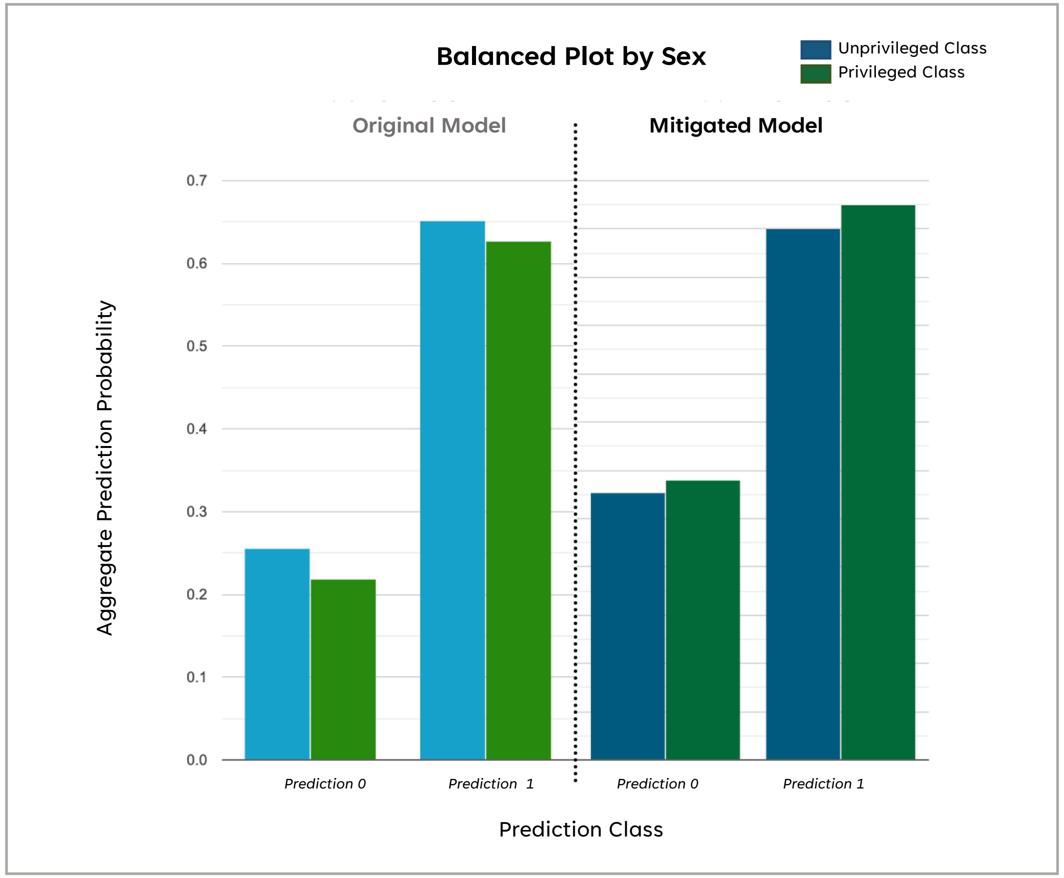
Four Fifths Plot

Identifies if there is adverse impact for unprivileged groups in comparison to the group with the highest selection rate.



Balance Plot

Examines whether average score received by individuals in positive and negative instances are similar regardless of sensitive attributes.

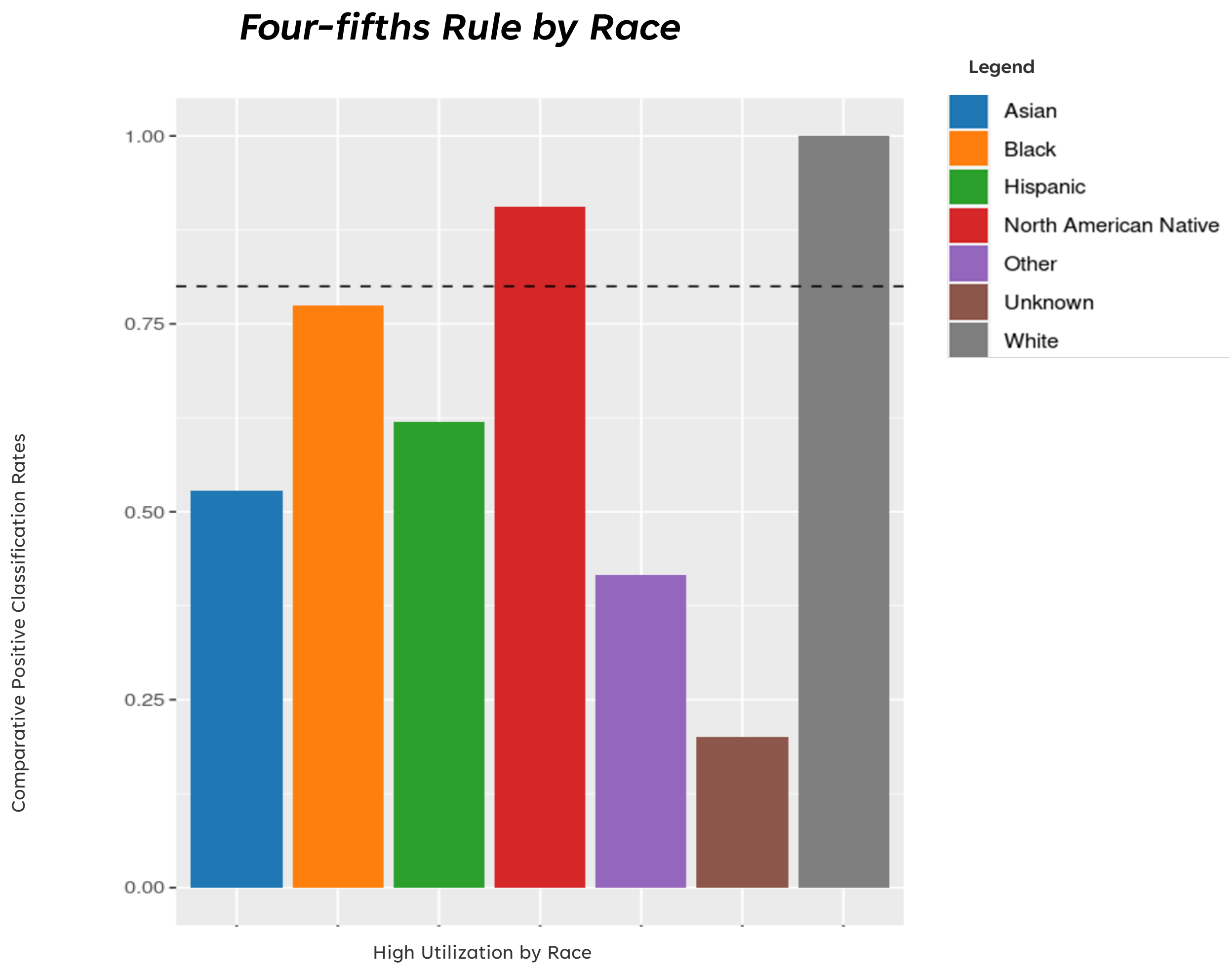


FAIRNESS PLOT EXAMPLES WITH RACE AND HEALTH EXPENDITURE

The four-fifths rule is a legal standard positing that if the selection rate for a certain group is less than 80 percent of that of the group with the highest selection rate, there is adverse impact on that group.

Whether White beneficiaries had the highest rates of 'high' utilization; could be largely reflective of different base rates in the data.

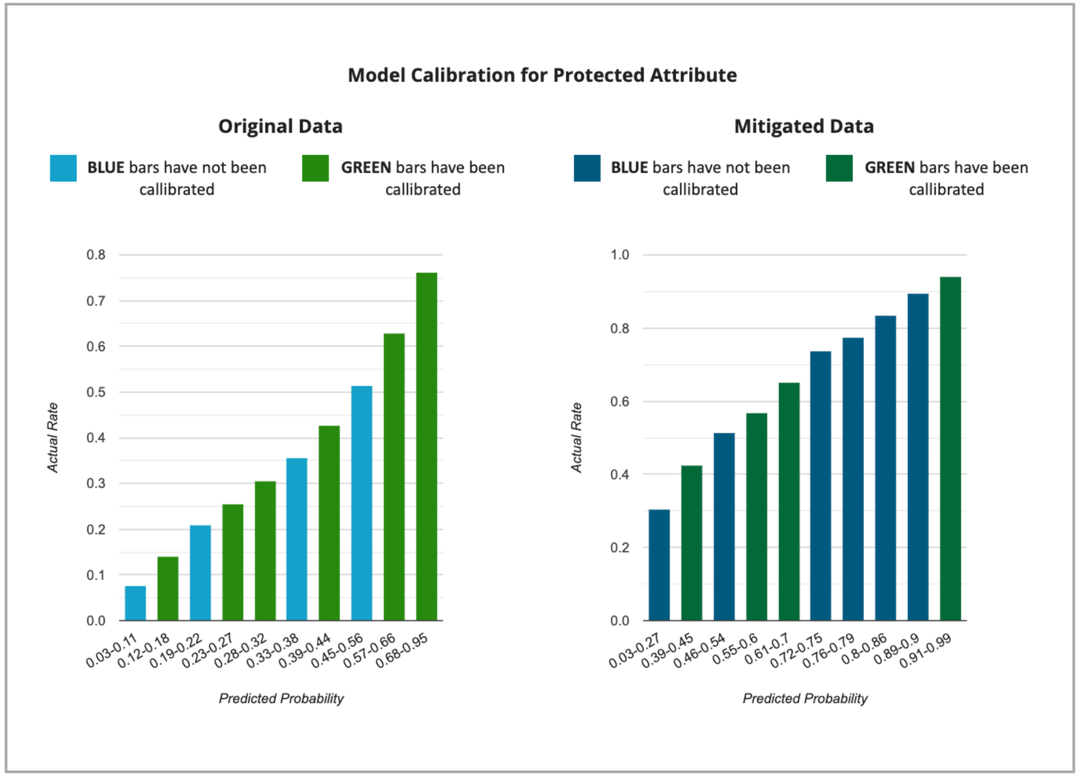
In the example on the right: the greatest discrepancy is for Other/Unknown groups – may be **indicative of missing data bias**.



OTHER FAIRNESS VISUALIZATIONS (*IMPUTED*) WITH MITIGATED DATA

Calibration Plots

Checks if model makes accurate predictions in aggregate for members of each class.



Fairness Ratios

A selection of metrics is designed to help data scientists detect and evaluate bias within AI models.

MIN — MAX

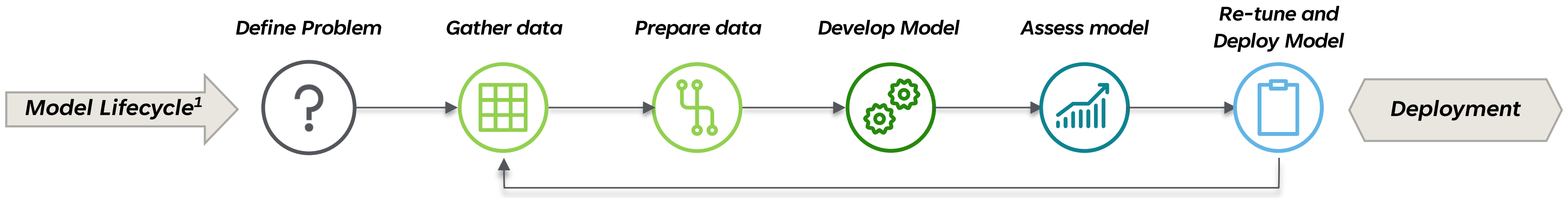
	Original	Mitigated
Statistical Party Difference	-0.1240	0.0036
Equal Opportunity Difference	-0.1985	0.0123
Average Odds Difference	-0.1414	0.0338
Theil Index	0.2718	0.1448
False Positive Rate	0.1309	0.5115
False Positive Rate Difference	-0.0844	0.1153
False Negative Rate	0.7215	0.0250
False Negative Rate Difference	0.1985	-0.0523

	Original	Mitigated
Disparate Impact	0.5289	0.1372
False Positive Rate Ratio	1.3403	0.5777
False Negative Rate Ratio	0.5612	1.2827

REPLICATION PROJECT PART 3

REPLICATION PROJECT: RESPONSIBLE AI IN ACTION

OPERATIONALIZING RESPONSIBLE AI ISN'T JUST USING CODE.
IT ALSO REQUIRES HUMANS IN THE LOOP FOR EACH STAGE OF THE MODEL LIFECYCLE



Start: Week 03

Part 0: Overview of Data, Scope, and Preparation
- Dataset
- Scope Definition

Week 04

Part 1: Exploratory Data Analysis
- Review of protected classes
- Analysis insights and explanation
- Data imbalances

Week 06

Part 2: Model Development and Fairness Evaluation (Bias Detection)
- Model Development
- Bias detection metrics
- Fairness metrics, assessment and visualizations

Week 07 (11/10/22)

Part 3: Bias Mitigation Techniques
- Pre, in, post-processing bias mitigation methods
- Discussion

End: Week 09 (12/01/23)

Part 4: Presentation, Conclusions and Explainability
- Local, Global Explainability Results
- Explainability integration into Model Lifecycles
- Future perspectives on explainability

FOR NEXT WEEK

- Complete next week's **reading**
  - If you signed up to present **Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models (Black et al.)** come prepared to present next week and submit your presentation to Gradescope by 10 AM PT, Thursday, November 17th

- Submit your answers to next week's participation questions on Gradescope by 10 AM PT, Thursday, November 23 (are there participations due on gradescope for thanksgiving break?

- **Replication Part #3:** Notebook and writeup on bias mitigation techniques
  - Primary contact for replication project: Emily Ramond & Parker Addison
  - Released on website.
  - Office hours: Tuesday 3:30-4:30pm (even on thanksgiving break!)
    - If you want extra OH, message us!

MODEL EXPLAINABILITY METHODS

MODEL EXPLAINABILITY: LIME

***L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations*

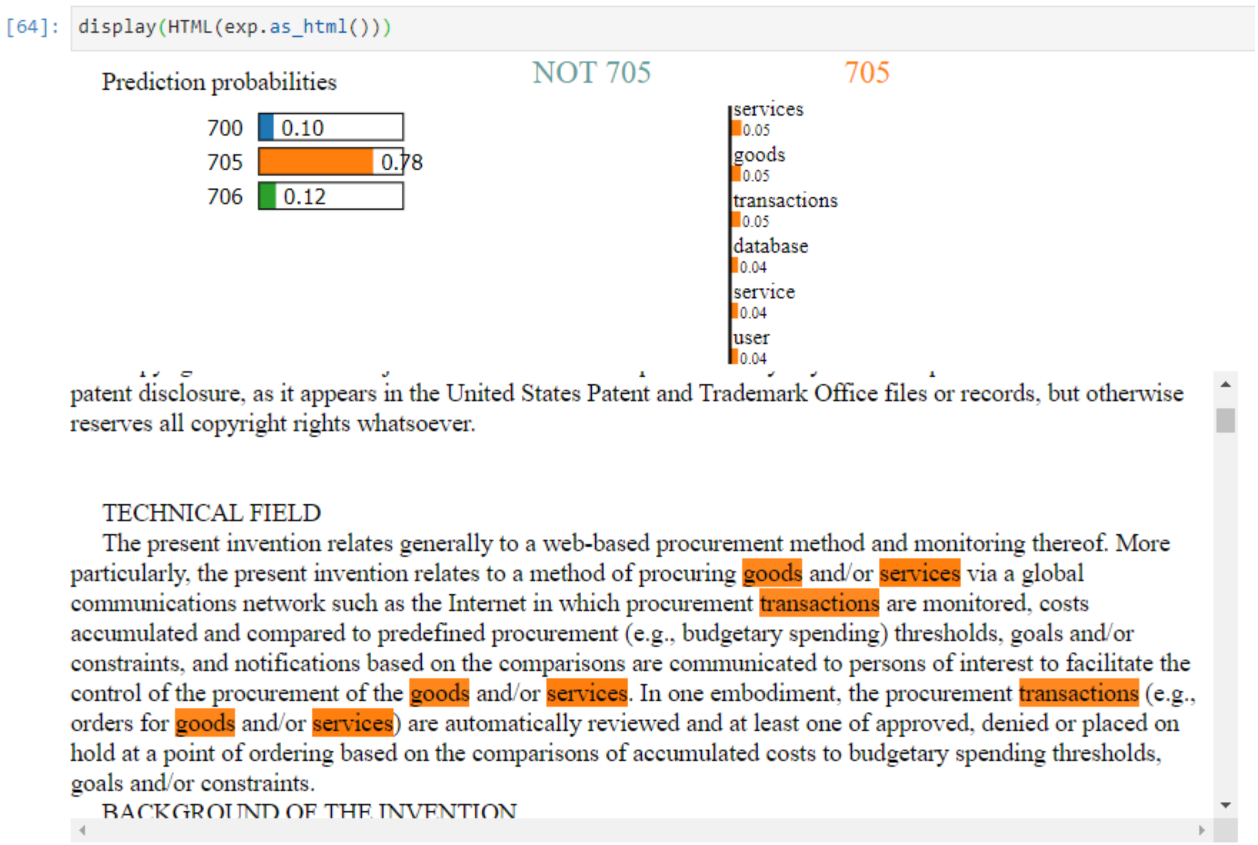
Using the LIME technique, a more complex model is interpreted by training less complex 'surrogate' models which provide explanations for individual data points.

LIME and SHAP are **attribution methods**, meaning that the prediction of a single instance is described as the **sum of feature effects.**

Patent Example:

This technique shows you which words were the most important. The numbers (700, 705, and 706) are different review groups. The NLP model works by connecting certain words to each of these groups.

1. Select the patent for which you'd like to interpret the explanation (e.g., why was patent X sorted into the review queue 705) .
2. LIME then creates fake patent data to train a new model, the ***surrogate model.*** This new model is ***local,*** meaning it can only be used to explain the one patent you selected.
3. The surrogate model uses a simpler algorithm (such as linear regression) and is more ***explainable***, allowing for the creation of bar charts like the one show to the right.
4. By analyzing the bar charts, you can explain which words or ***model features*** contributed the most to its predictions or determination of *class* (review queue number).

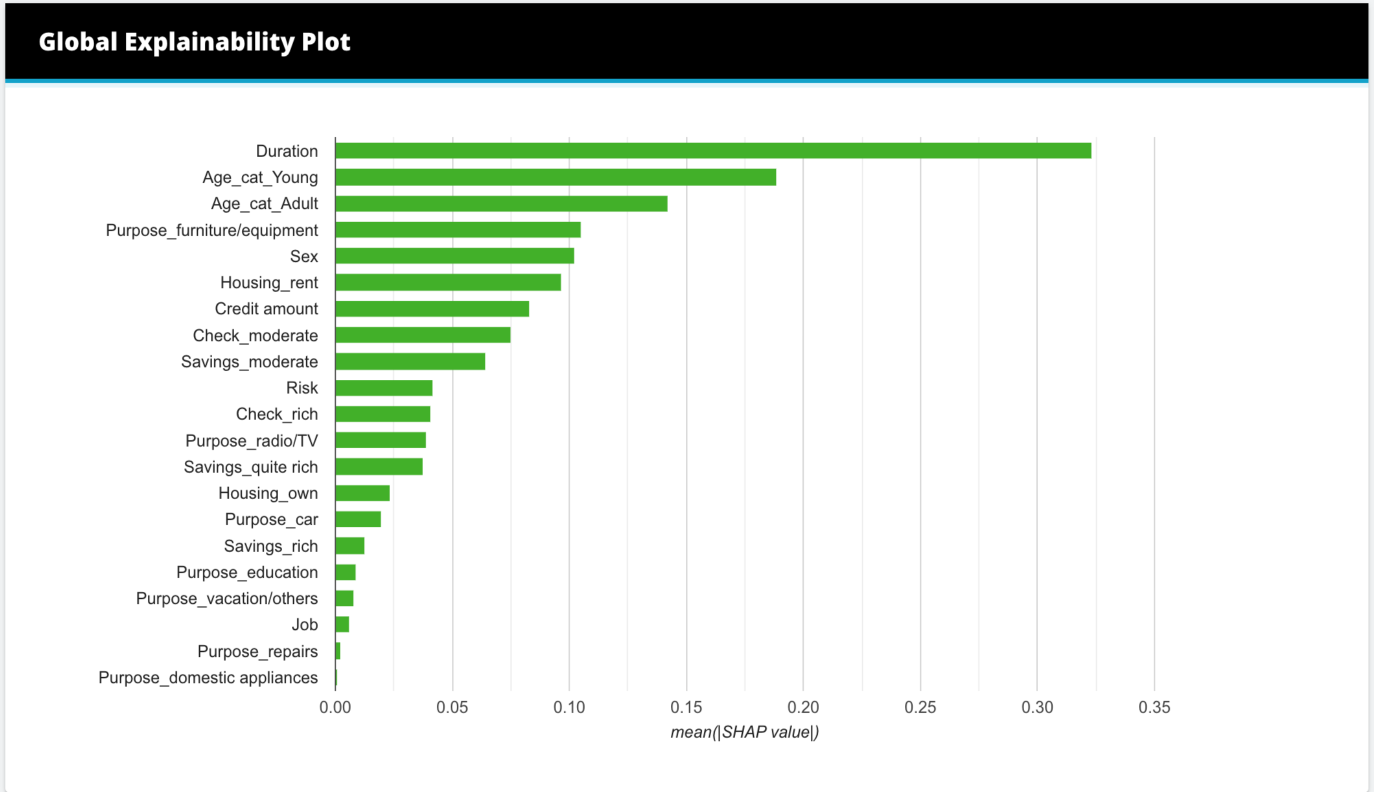


EXPLAINABILITY: SHAP

SHAP (**SH**apley **A**dditive ex**P**lanations)

Global Explainability

Global explainability ensures the entire process of decision making is completely transparent



In the bar plot, features are sorted by decreasing overall importance to a model's ability to predict an outcome.

Features list at the top of Y-axis have the most predictive power

However, these features should not be interpreted as having casual impact on the model's prediction

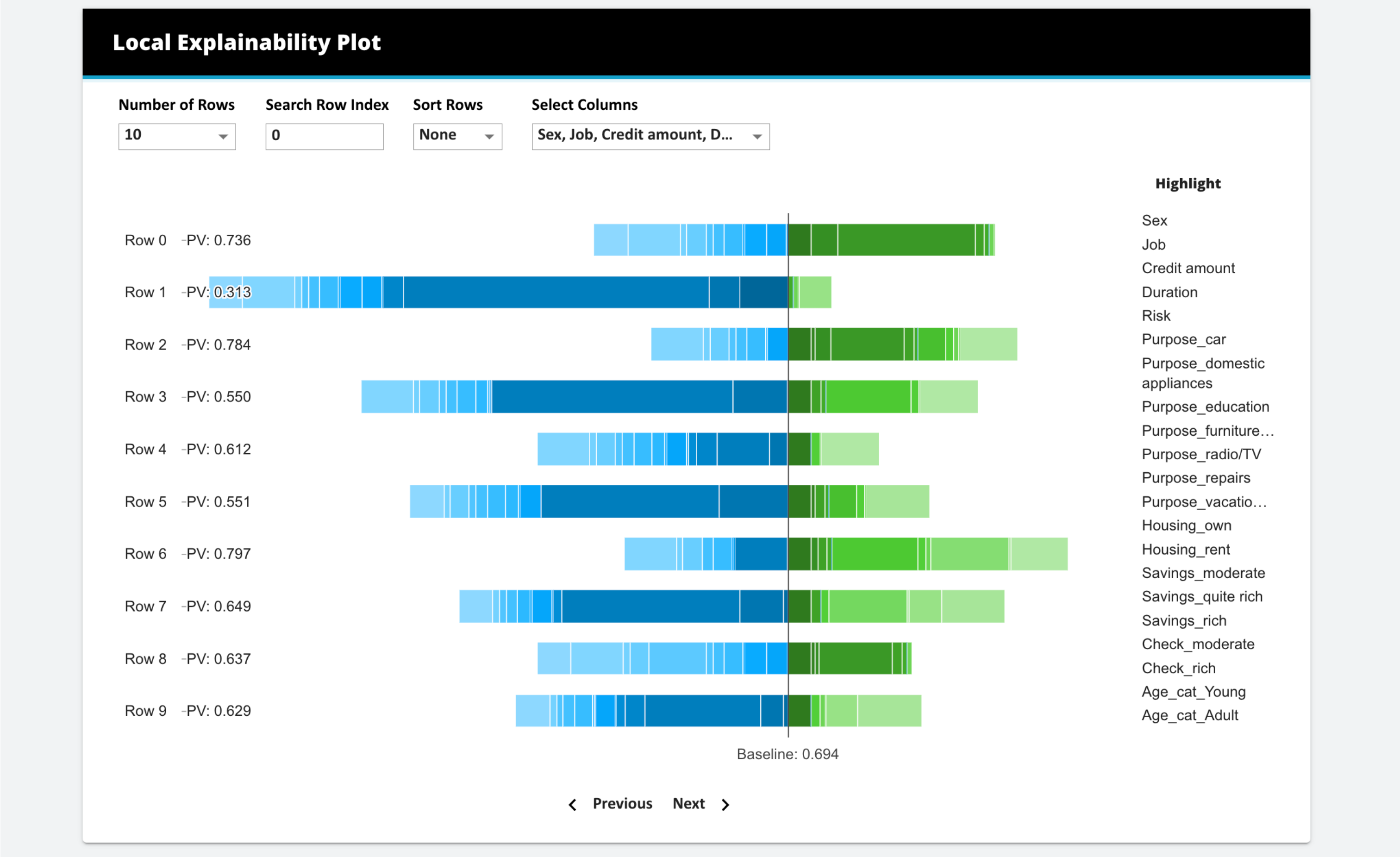
SHAP feature importance is measured as the "average Shapley value PER feature across ALL of the data."

Therefore, if a feature in this example plot has a mean Shapley value of 0.015, then that feature changes the model's predicted output on average by 1.5 percentage points

EXPLAINABILITY: SHAP

Local Explainability

- Local explainability provides explanations for each decision
- It takes a granular look at explaining how individual data points are used to make decisions



The **BASELINE** value represents the **average prediction across a dataset**

The **PREDICTION (PV)** value represents the **model's prediction** for that observation

The BLUE bars represent features that push the prediction value **higher**

The GREEN bars represent features that push the prediction value **lower**

What about other protected classes?

Potential things to consider when addressing bias for **Age**, **Race**, and **Sex**

Age

- Future work could consider identifying additional data sources to understand where proxies might pop-up, such as using age as a predictor – this would both increase accuracy and decrease the age-linked differential in error rates
- Youngest beneficiaries may be qualitatively different from older due to rates of disability.
- Consider separate models or hierarchical models which would explicitly learn the different dynamics of these types of subpopulations

Race

- Consider explicitly including race in models to correct for bias using pre-, in- and post- processing for bias mitigation
- Considerations of Impact (benefiting historically underserved groups by directing resources to group members at risk of mortality/hospitalization)
- Considerations for community stakeholders as “human values in the loop”[1], to suggest better data collection methods.
- “Unknown/other” missing data problem remediation requires investigating the source of upstream data collection

Sex

- Monitor for bias on an ongoing basis and examine intersectional bias for combinations of age and sex.
- Use visualizations such as the 4/5th plot to understand and communicate data remediation needs

Aggregation bias

Measurement bias

Historical bias

Evaluation bias

Deployment bias

Representation bias