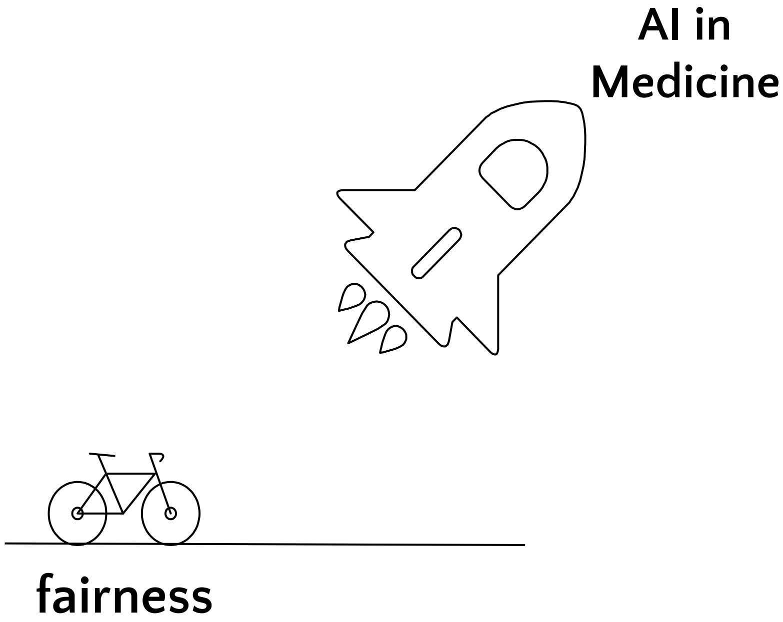
Non-Discrimination in AI: An Application to Fair Cardiovascular Disease Diagnosis

Presentation of Master Thesis Work by Analise Burko

The Big Picture



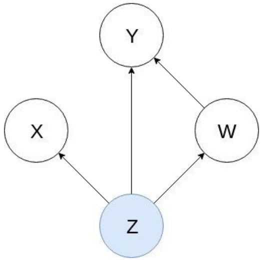
Problem

Little work has been done to evaluate model fairness, resulting in inequitable detection and therefore worse health outcomes for those that face bias

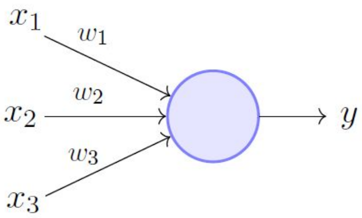
Origins of Bias

Historical Injustice

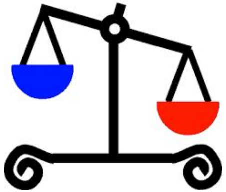
Proxy Variables



Algorithm Choice



Unbalanced Samples



Feedback Loops

Cardiovascular Disease

***Definition:** a group of disorders of the heart and blood vessels; normally associated with atherosclerosis and an increased risk of blood clots*

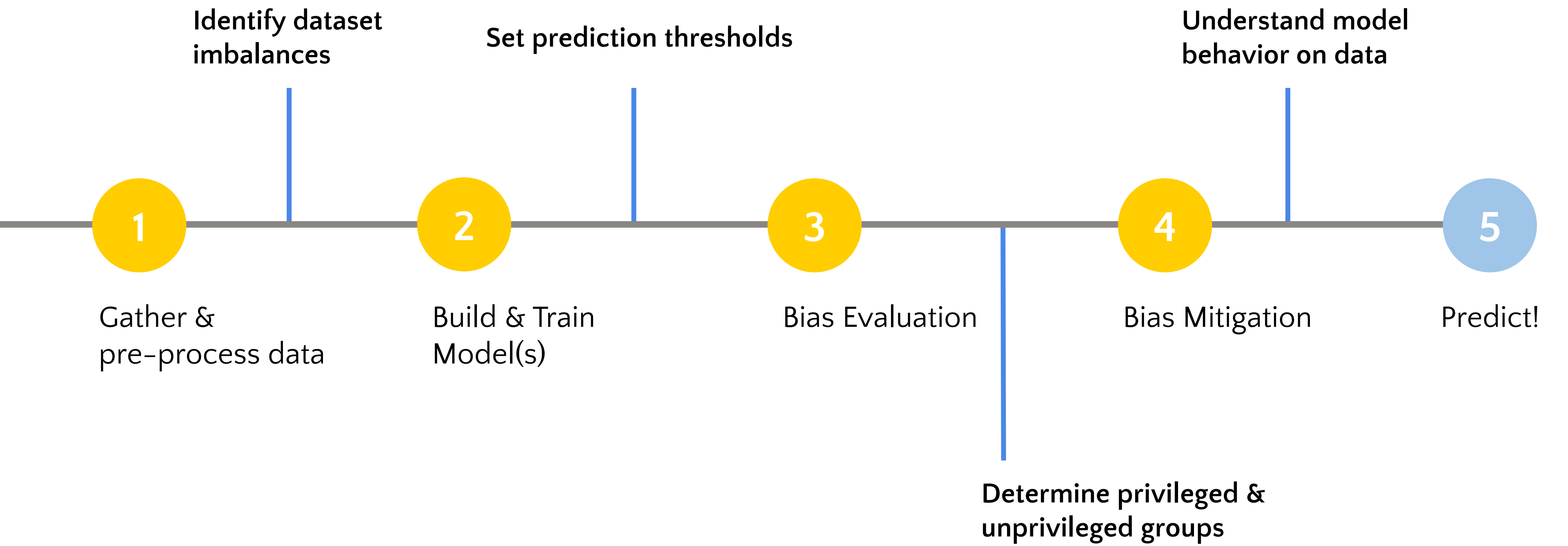
- Leading cause of global burden and mortality (Global Burden of Disease Study– 2019)
- Can most often be prevented through leading a healthy lifestyle
- Early diagnosis is critical in improving patient quality of life and allowing actionable measures to be taken sooner, ensuring the best outcomes possible

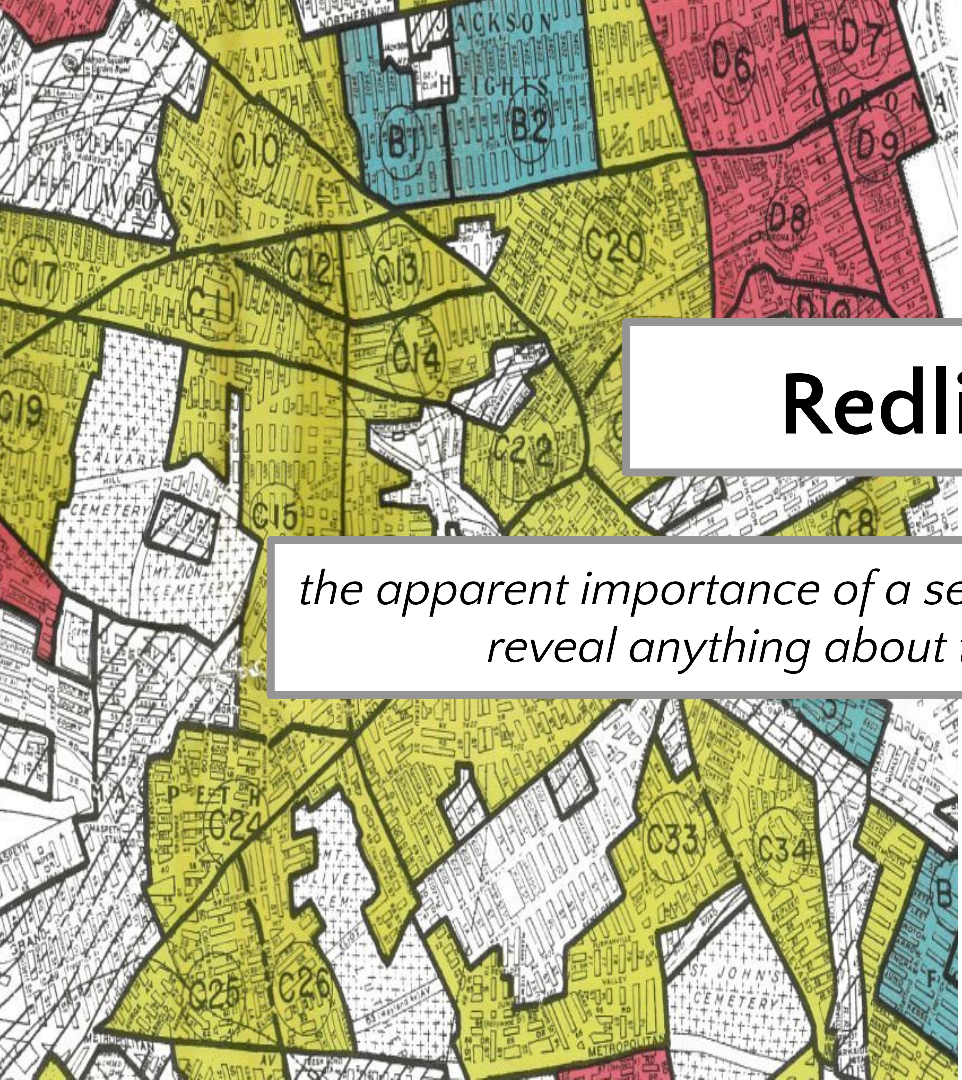
Goals

1. Design a framework of model building, bias evaluation, bias mitigation, and comparison to effectively diagnose the presence of CVD at baseline fairly against *sex, age, and race*

2. Background research on problem domain to provide insightful interpretation and understanding of fairness results

Project Roadmap





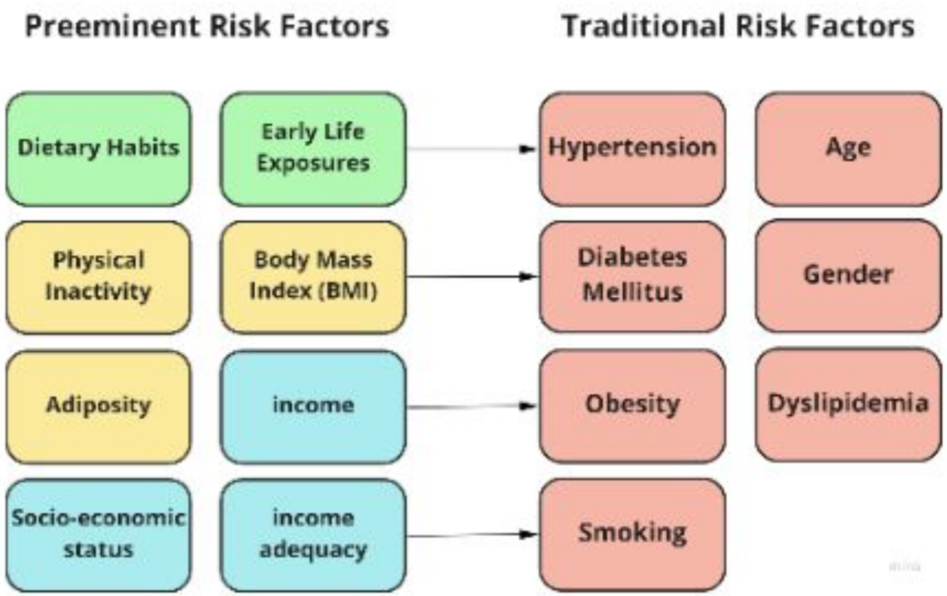
Redlining

the apparent importance of a sensitive feature does not reliably reveal anything about the fairness of a model

The Dataset

- Tabular Dataset of 502k Records
- Highly Unbalanced: 9,2% Positive for CVD
- Types of Features:
  - Physical
  - Sociodemographic
  - Lifestyle
  - Environmental
  - Health Outcomes...

CVD Risk Factors & Outcomes

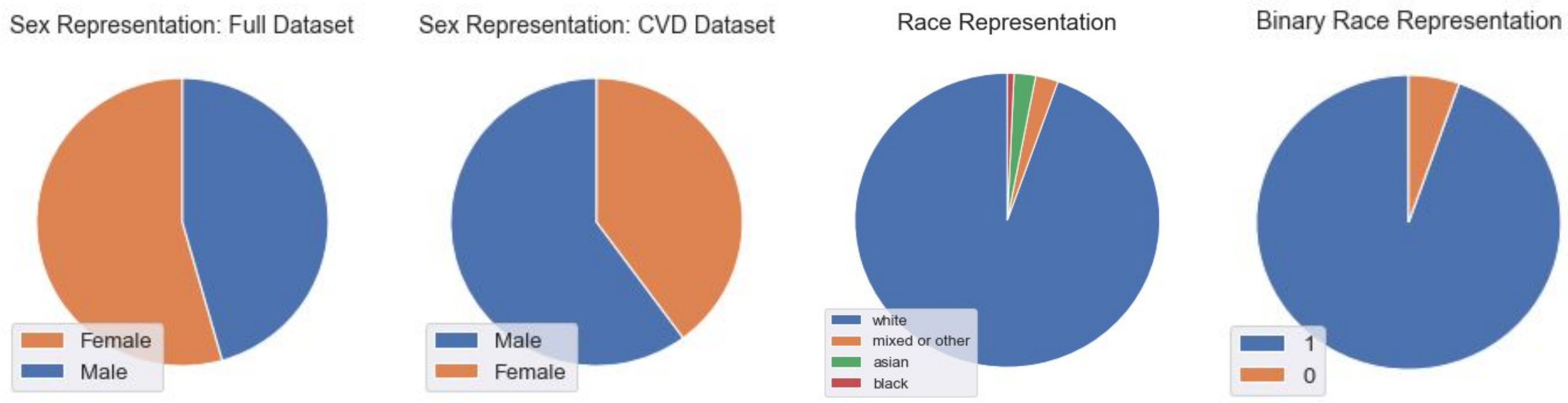


CVDs:

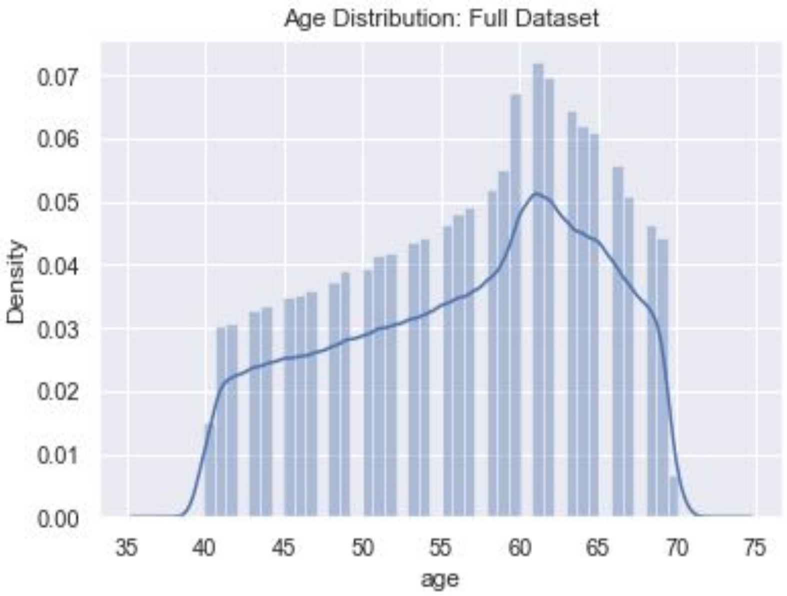
1. Cardiomyopathies
2. Ischemic Heart Disease
3. Heart Failure
4. Peripheral Vascular Disease
5. Cardiac Arrest
6. Cerebral Infarction
7. Arrhythmia
8. Myocardial Infarction

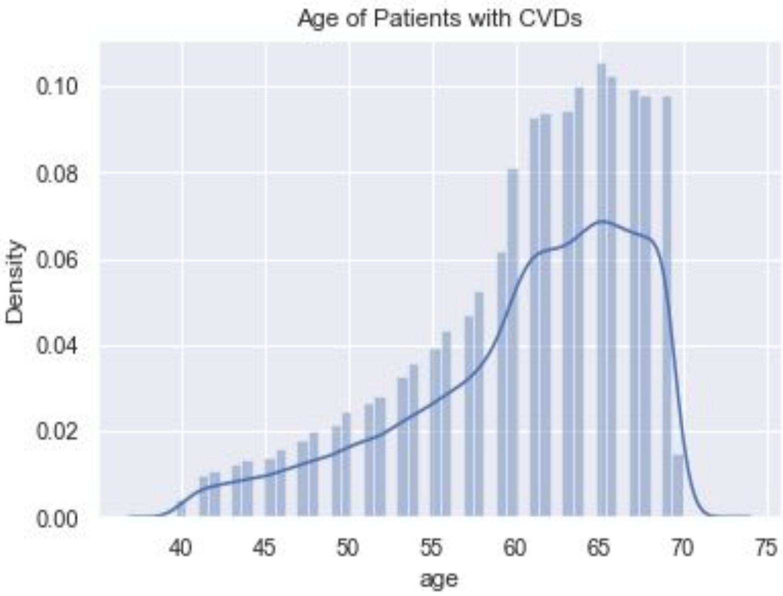
** 61 available UKBB inputs fall within the widely accepted preeminent and traditional risk factors for CVDs

Protected Attribute Distributions



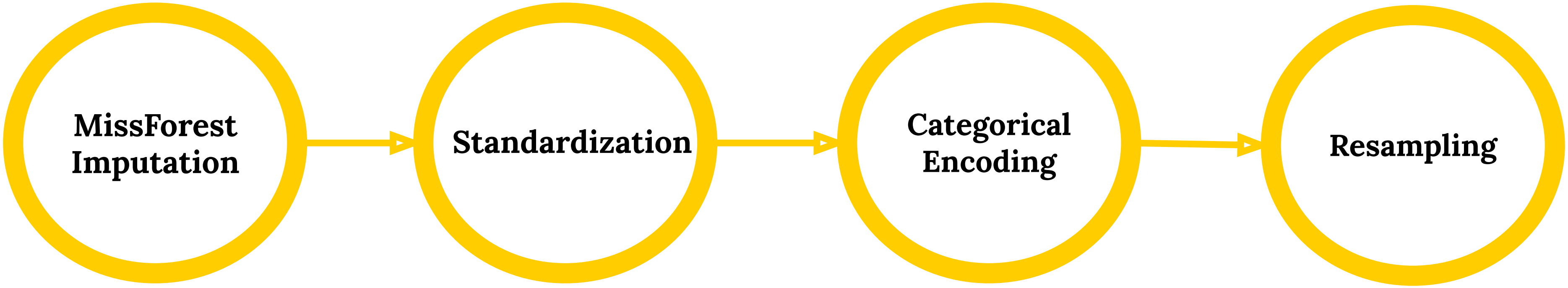
Protected Attribute Distributions





1 Model Development

Preprocessing & Feature Transformation



Model Architectures

XGBoost

- Leader for learning on tabular data

MLP

- Feedforward NN
- Batch normalization to address input-sensitivity common with tabular data

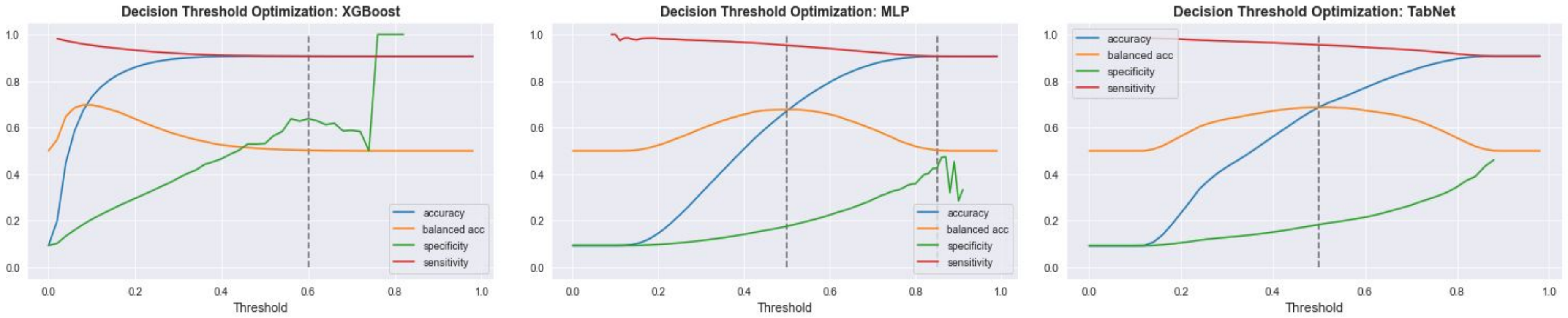
TabNet

- DL framework for tabular data
- Offers local explainability via feature masking

Experiments

- train/validate/test split (80/10/10)
- Binary classification– single target +/- for CVD in general

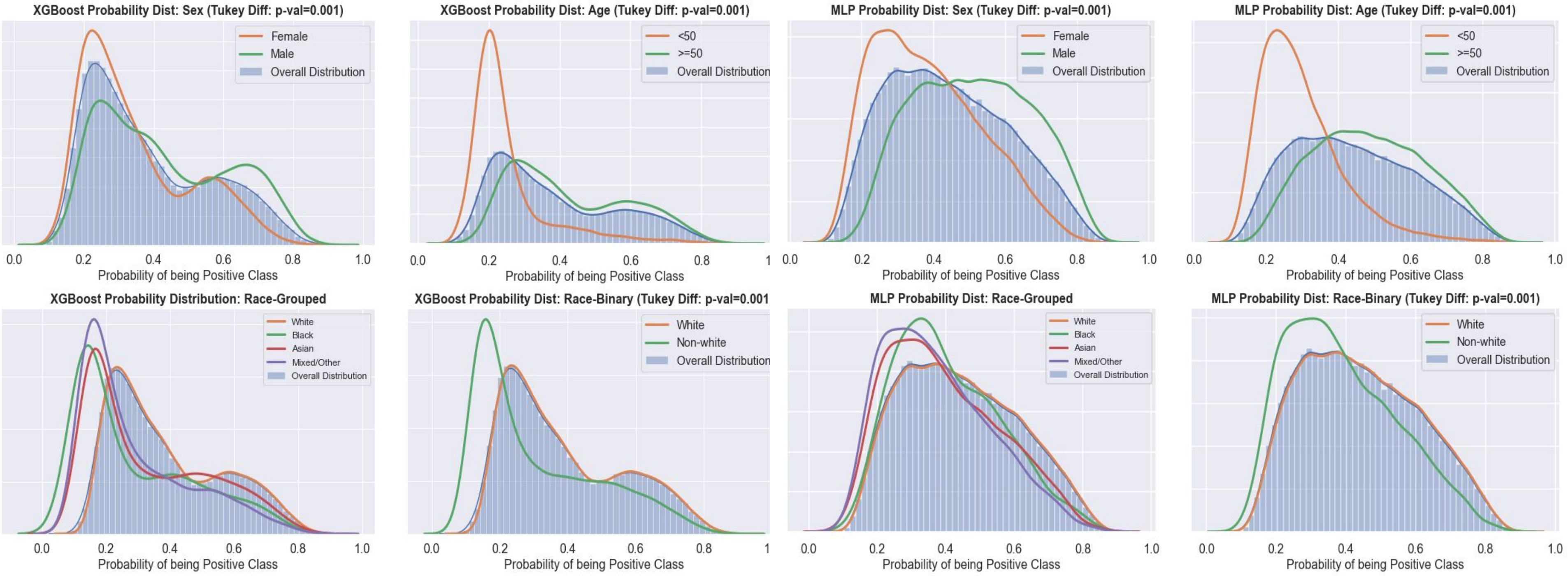
Model Performance



Model	Threshold	Accuracy	Bal Accuracy	Sensitivity
XGBoost	0.6	0.91	0.51	0.91
MLP	0.5	0.67	0.68	0.95
TabNet	0.5	0.69	0.69	0.96

2 Bias Evaluation

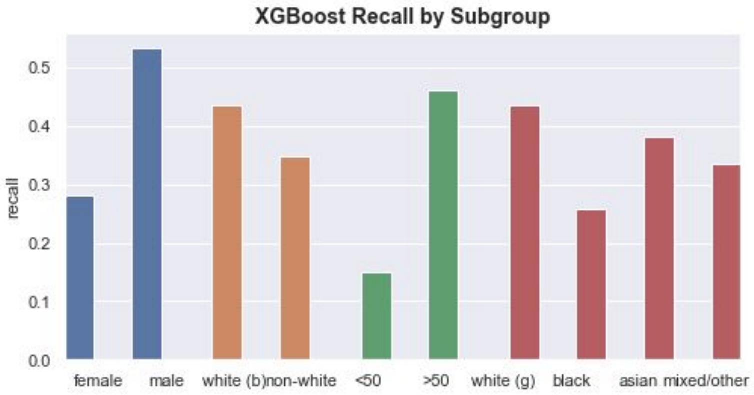
Determination of Privileged & Unprivileged Groups

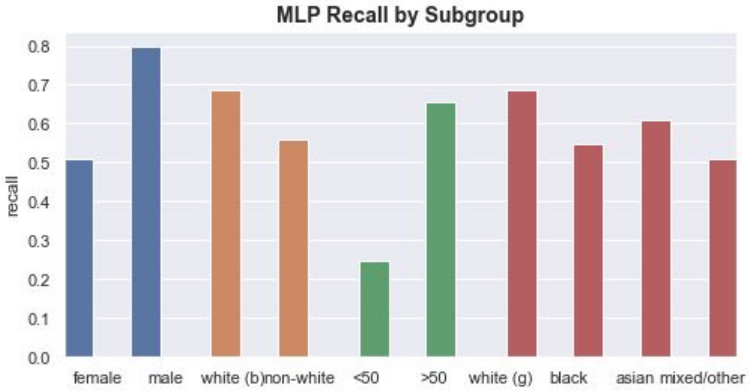


Apply Tukey's HSD test to determine significant difference in prediction probability means

Determination of Privileged & Unprivileged Groups

- **Sex**
  - Privileged: Male
  - Unprivileged: Female
- **Age**
  - Privileged: >50 years old
  - Unprivileged: <50 years old
- **Race**
  - Privileged: White
  - Unprivileged: Non-white





Fairness Metrics

- **Average Odds:** $\frac{1}{2}[(FPR_{D=\text{group 1}} - FPR_{D=\text{group 2}}) + (TPR_{D=\text{group 2}} - TPR_{D=\text{group 1}}))]$
- **Disparate Impact:** $DI = P(\hat{Y} = 1|A = minority)/P(\hat{Y} = 1|A = majority)$
- **Equal Opportunity:** $TPR = TP/P$,
- **Statistical Parity:** $P(\hat{Y} = 1|D = \text{group 1}) - P(\hat{Y} = 1|D = \text{group 2})$
- **Theil Index:** $\frac{1}{n}\Sigma_{i=1}^{n}\frac{b_i}{\mu}ln\frac{b_i}{\mu}, \quad where \;\; b_i = \hat{y}_i - y_i + 1$

Table 5.1: Fairness Metric Thresolds

Metric	**Optimal Value**	**Acceptable Range**	**Interpretation**
DI	1	0.8 to 1.0	<1 favors privileged group >1 favors unprivileged group
SPD	0	-0.1 to 0.1	<0 favors privileged group >0 favors unprivileged group
AOD	0	-0.1 to 0.1	<0 favors privileged group >0 favors unprivileged group
EOD	0	-0.1 to 0.1	<0 favors privileged group >0 favors unprivileged group
Theil Index	0	-	bias increases as score increases

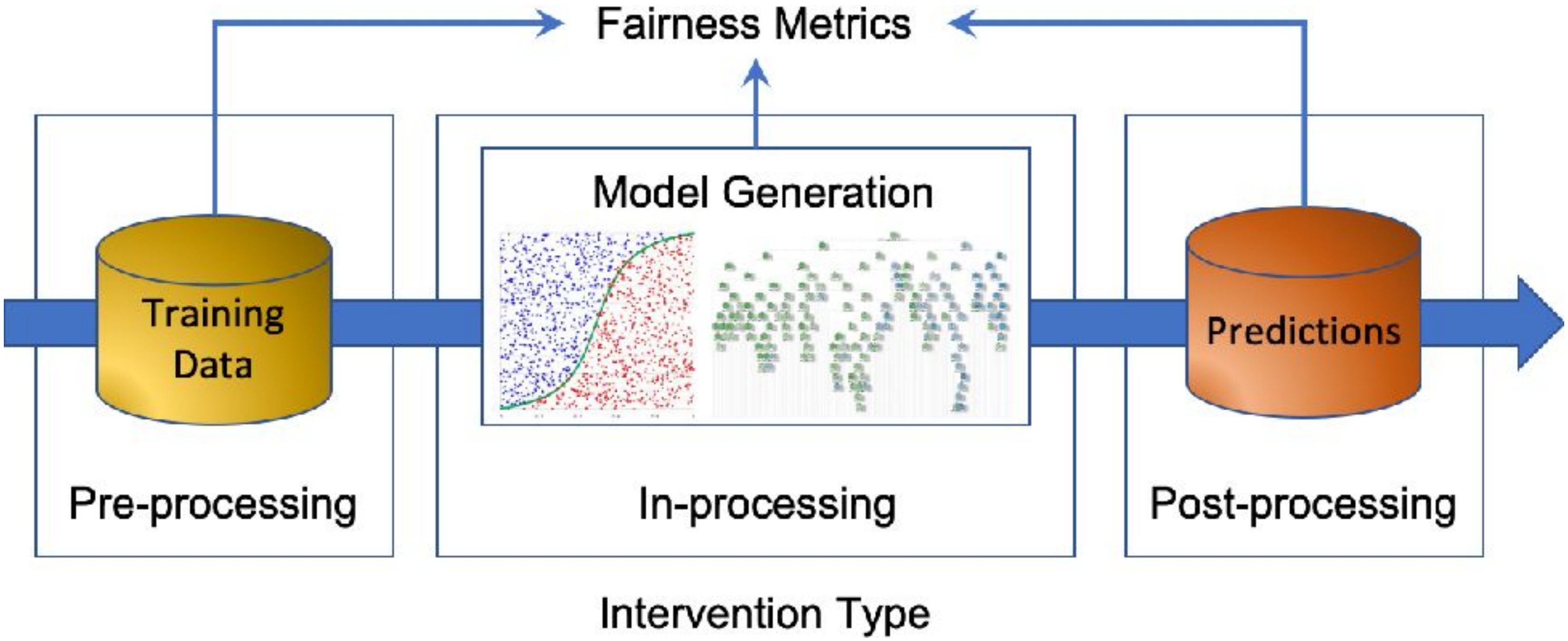
Bias of Original Models

Model	attribute	AOD	DI	SPD	EOD	Theil
XGBoost	sex	-0.27	0.22	-0.20	-0.38	0.09
	race	-0.10	0.58	-0.06	-0.15	0.09
	age	-0.17	0.22	-0.14	-0.22	0.09
MLP	sex	-0.27	0.35	-0.24	-0.32	0.09
	race	-0.08	0.61	-0.09	-0.10	0.09
	age	-0.29	0.08	-0.23	-0.39	0.09

3

Bias Mitigation

Possible Intervention Points



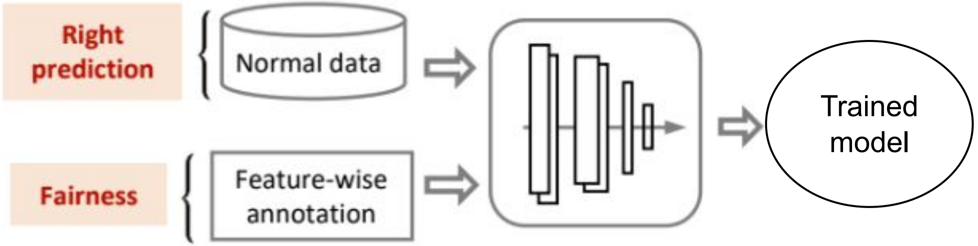
Pre-processing Interventions: XGBoost

Disparate Impact Remover

- removes the model's ability to distinguish between subgroups

Reweighing

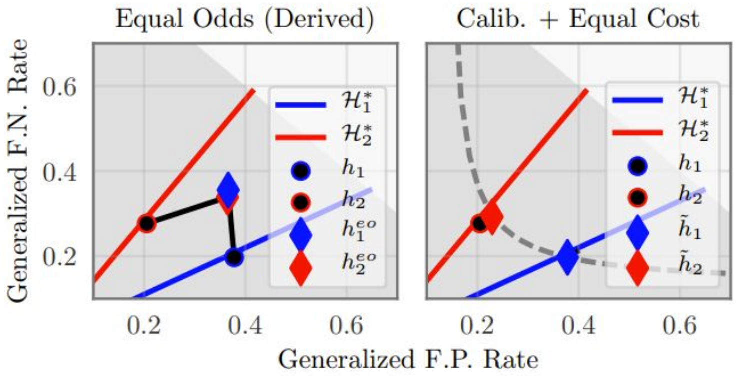
- weights samples to ensure fairness before classification



Post-processing Interventions: MLP

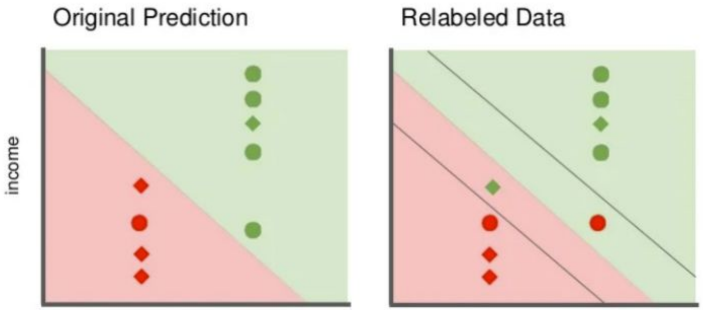
Calibrated Equalized Odds

- constrained by **FNR**, FPR or a weighted combination



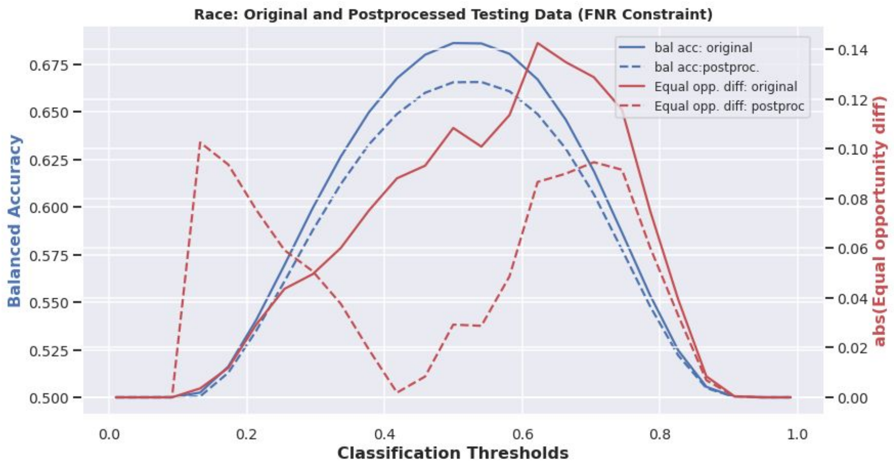
Reject Option Classification

- Positively discriminates in a confidence band around uncertainty
- Optimizes on AOD, **SPD,** or **EO**

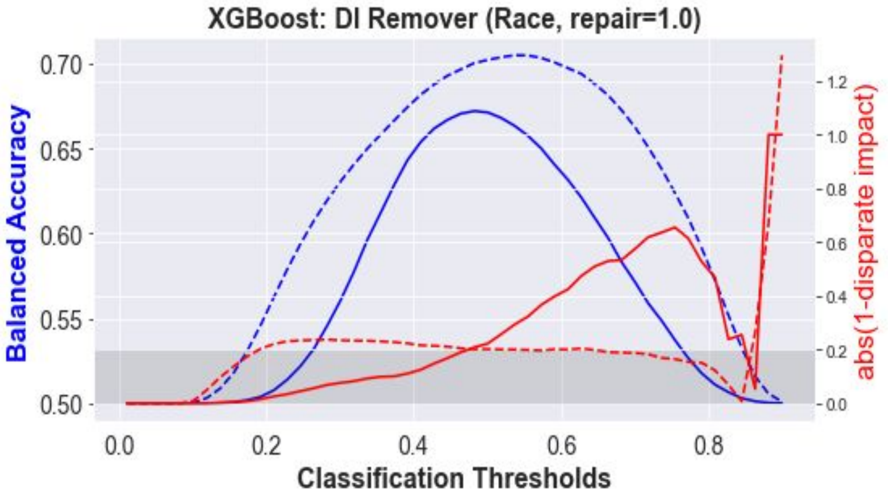


4 Results

Fairness-Utility Tradeoff

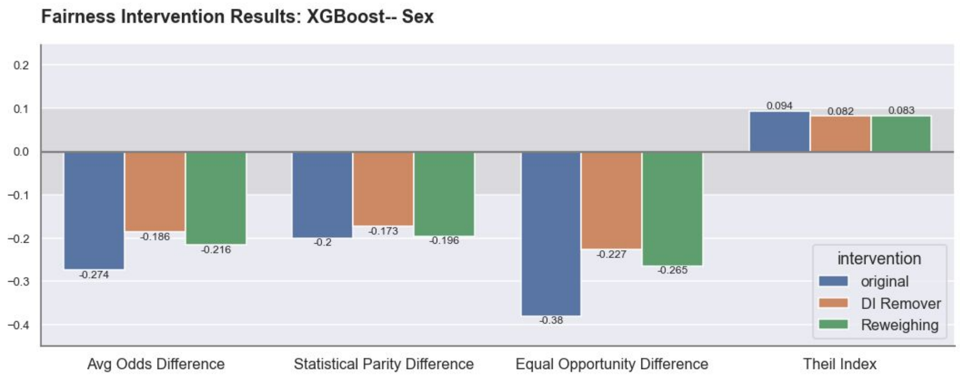


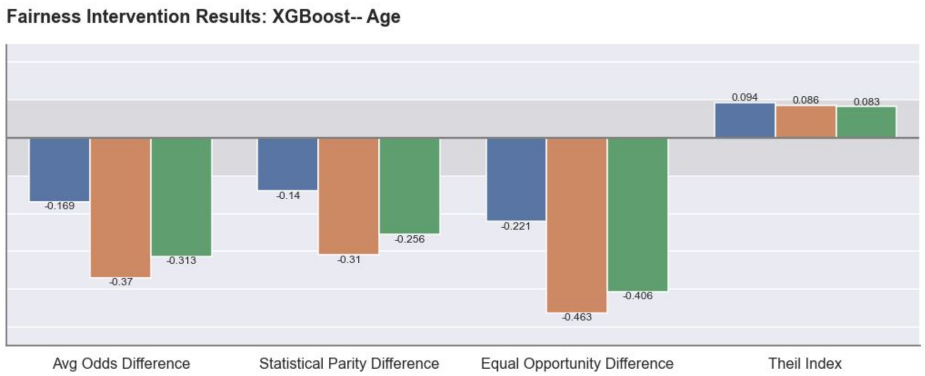
Calibrated Equalized Odds

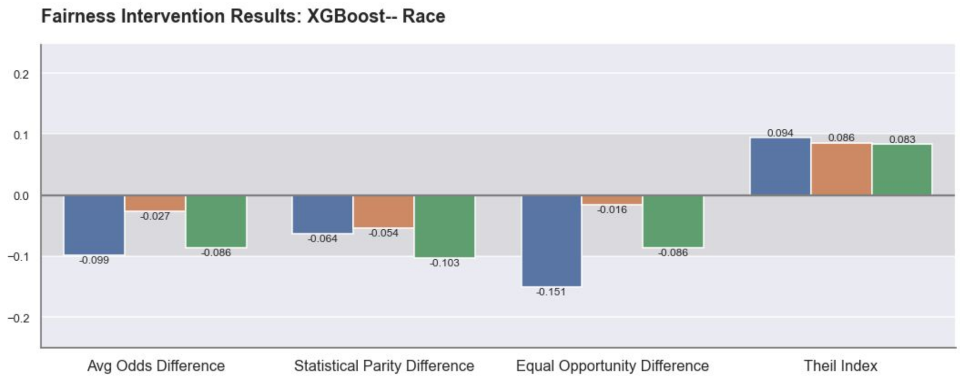


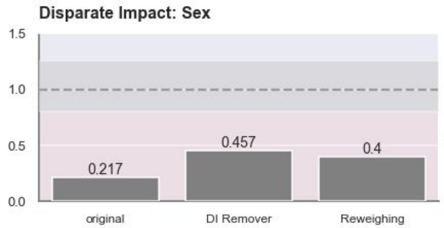
Disparate Impact Remover

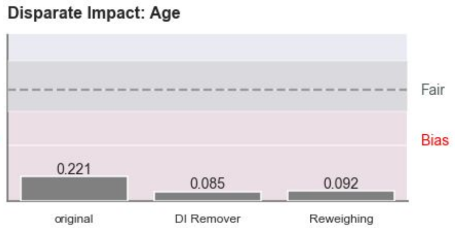
Pre-processing Fairness Results

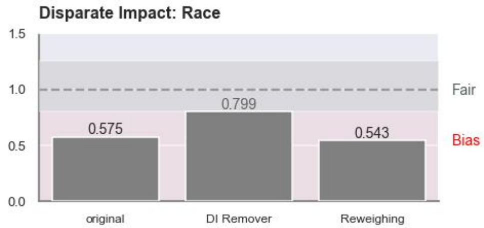




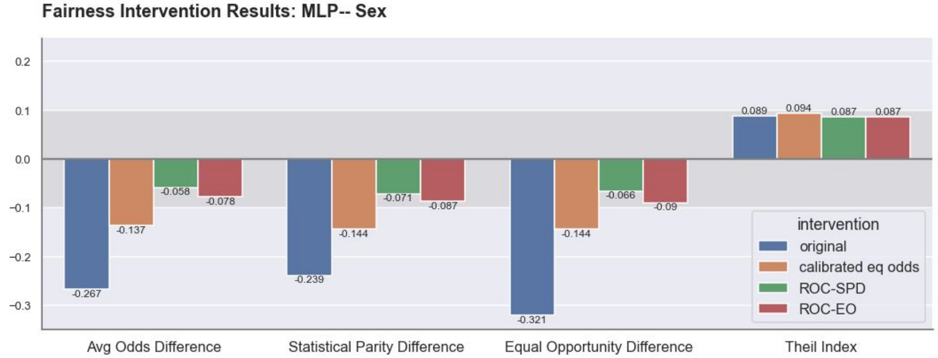


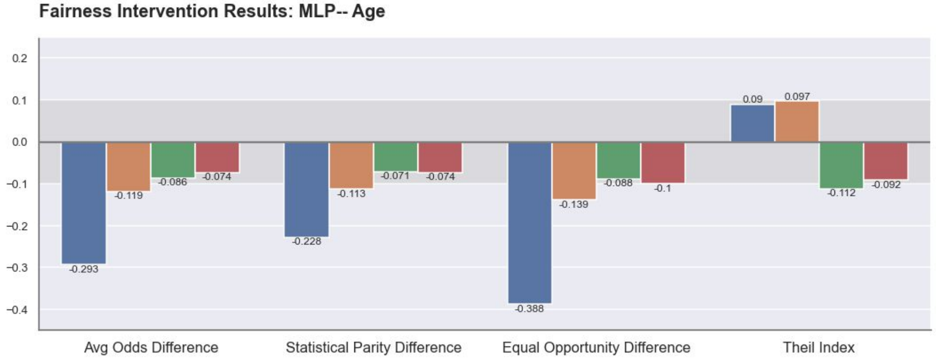


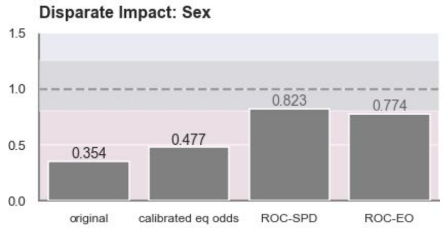


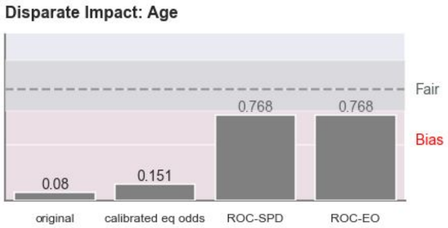


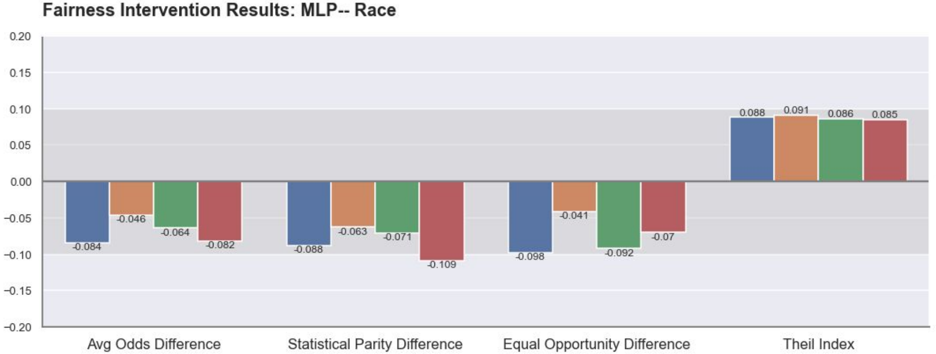
Post-processing Fairness Results

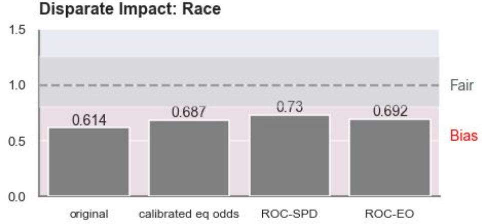












Effect on Utility: Pre-processing Interventions

Model	Intervention	Attribute	Balanced Accuracy	Sensitivity	Specificity
XGBoost	original	-	0.63	0.88	0.38
	DI Remover	sex	0.71	0.82	0.61
	DI Remover	race	0.70	0.77	0.62
	DI Remover	age	0.70	0.77	0.62
	Reweighing	-	0.82	0.82	0.60

Effect on Utility: Post-processing Interventions

Model	Intervention	Attribute	Balanced Accuracy	Sensitivity	Specificity
MLP	original	-	0.68	0.79	0.56
	Cal. odds	sex	0.65	0.83	0.45
	Cal. odds	race	0.66	0.83	0.49
	Cal. odds	age	0.59	0.91	0.27
	ROC - SPD	sex	0.68	0.67	0.69
	ROC - SPD	race	0.69	0.70	0.69
	ROC - SPD	age	0.69	0.67	0.67
	ROC - EO	sex	0.68	0.70	0.67
	ROC - EO	race	0.69	0.69	0.70
	ROC - EO	age	0.67	0.73	0.60

Key Takeaways

Pre-processing

- increase balanced accuracy **(+7–19%)**
- slight decrease in sensitivity **(–6–11%)**
- Only slightly improve fairness for sex and race, worsen fairness for age
  - fully resolved ⇒ Race: DI, EOD

Post-processing

- all techniques meet or almost reach acceptable fairness (cal. odds slightly less effective)
- Calibrated equalized odds slightly decreases balanced accuracy **(–2–9%)** & increases sensitivity **(+4–12%)**
- ROC interventions maintain balanced accuracy **(±1%)** and decrease sensitivity **(–6–12%)**

Future Work

- revisiting of the multi-label classification problem with a more robust dataset
- investigation of in-processing interventions and interventions in unison
- application of fairness framework to the TabNet model
- Address trustworthiness as a whole – *fairness, explainability, robustness*

Thank You!

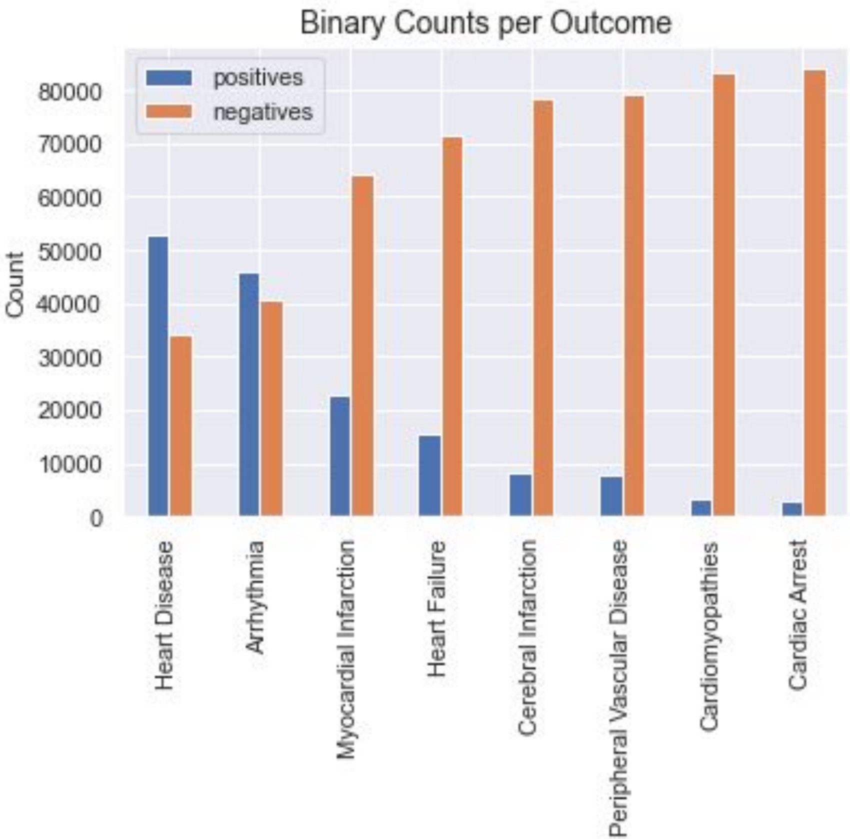
Supplementary Slides

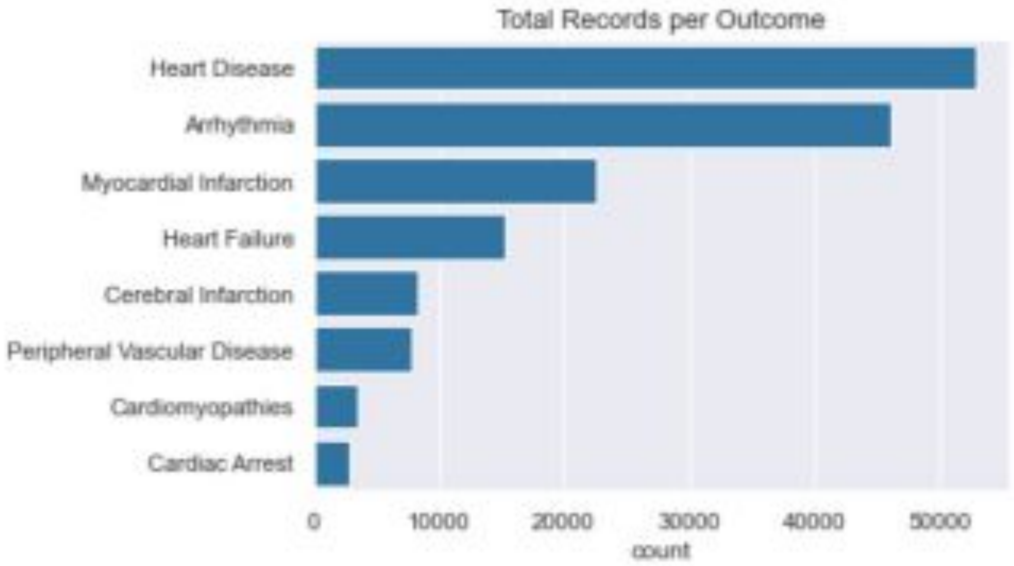
Risk Factor Category	**Feature**	
Physical Measures	Hypertension	Waist Circumference
	Hip Circumference	Diastolic Blood Pressure
	Systolic Blood Pressure	Body Mass Index (BMI)
	Body Fat Percentage	Whole body fat mass
	Whole body fat-free mass	Pulse Rate
	Impedence of whole body	
Sociodemographics	Sex	Qualifications
	Current Employment	Ethnic Background
	Age completed education	
Lifestyle/Environment	Sleep duration	Insomnia
	Current tobacco smoking	Past tobacco smoking
	Cooked vegetable intake	raw vegetable intake
	Fresh fruit intake	Dried fruit intake
	Oily fish intake	Non-oily fish intake
	Processed meat intake	Poultry intake
	Beef intake	lamb intake
	Pork intake	Cheese intake
	Coffee type	Alcohol status
	Variation in diet	Spread type
	Daytime dozing/sleeping	Water intake
	Major dietary changes	Non-butter spread
	Alcohol consumed	Daily alcohol consumption
Mental Health	Freq of Depressed Mood	Anxiety
	Seen a psychiatrist	
Blood Assays	Vascular	APOB
	Cholesterol	CRP
	Glucose	HDL
	LDL	LP-a
	Triglyceride	IGF-1
	Testosterone	HbA1c

CVD Distribution

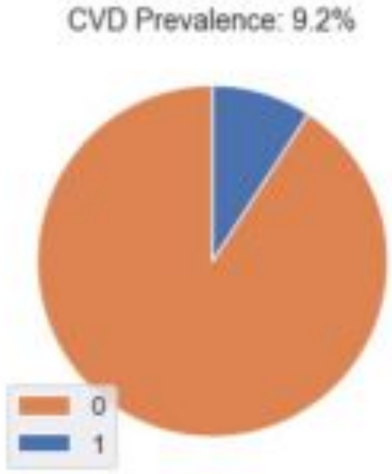
- Overall CVD prevalence in entire dataset (~500k records): **9.2%**

- Bar chart shows each CVD prevalence against all other CVDs
    - Cardiac Arrest, Cardiomyopathies, Peripheral Vascular Disease, and Cerebral Infarction are all greatly under-represented

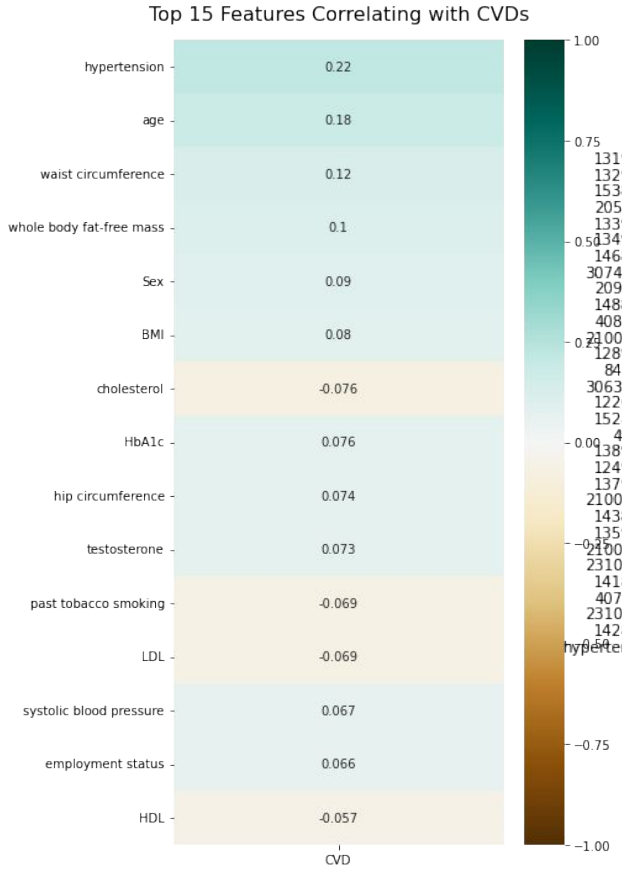


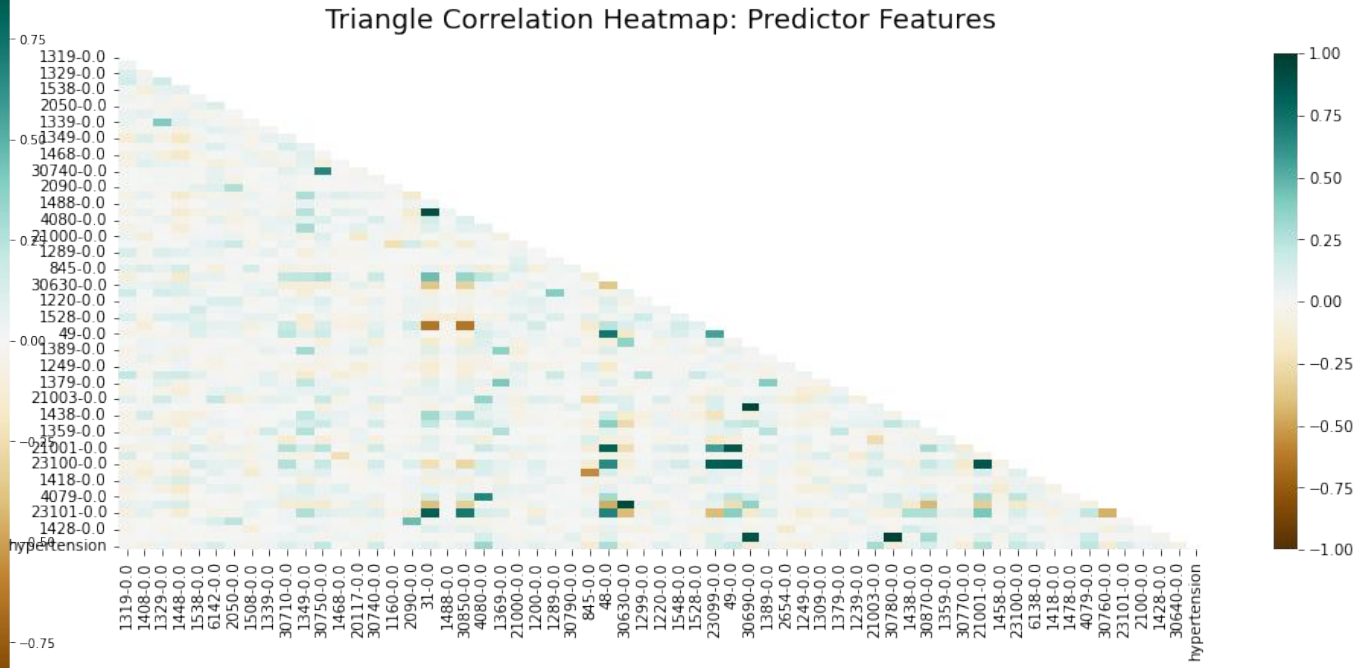


Total Records per Outcome
Heart Disease
Arrhythmia
Myocardial Infarction
Heart Failure
Cerebral Infarction
Peripheral Vascular Disease
Cardiomyopathies
Cardiac Arrest
0
10000
20000
30000
40000
50000
count



CVD Prevalence: 9.2%
0
1


Top 15 Features Correlating with CVDs
hypertension 0.22
age 0.18
waist circumference 0.12
whole body fat-free mass 0.1
Sex 0.09
BMI 0.08
cholesterol -0.076
HbA1c 0.076
hip circumference 0.074
testosterone 0.073
past tobacco smoking -0.069
LDL -0.069
systolic blood pressure 0.067
employment status 0.066
HDL -0.057
CVD
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Triangle Correlation Heatmap: Predictor Features
1319-0.0
1329-0.0
1538-0.0
2050-0.0
1339-0.0
1349-0.0
1468-0.0
30740-0.0
2090-0.0
1488-0.0
4080-0.0
21000-0.0
1289-0.0
845-0.0
30630-0.0
1220-0.0
1528-0.0
49-0.0
1389-0.0
1249-0.0
1379-0.0
21003-0.0
1438-0.0
1359-0.0
21001-0.0
23100-0.0
1418-0.0
4079-0.0
23101-0.0
1428-0.0
hypertension
1319-0.0
1408-0.0
1329-0.0
1448-0.0
1538-0.0
6142-0.0
2050-0.0
1508-0.0
1339-0.0
30710-0.0
1349-0.0
30750-0.0
1468-0.0
20117-0.0
30740-0.0
1160-0.0
2090-0.0
31-0.0
1488-0.0
30850-0.0
4080-0.0
1369-0.0
21000-0.0
1200-0.0
1289-0.0
30790-0.0
845-0.0
48-0.0
30630-0.0
1299-0.0
1220-0.0
1548-0.0
1528-0.0
23099-0.0
49-0.0
30690-0.0
1389-0.0
2654-0.0
1249-0.0
1309-0.0
1379-0.0
1239-0.0
21003-0.0
30780-0.0
1438-0.0
30870-0.0
1359-0.0
30770-0.0
21001-0.0
1458-0.0
23100-0.0
6138-0.0
1418-0.0
1478-0.0
4079-0.0
30760-0.0
23101-0.0
2100-0.0
1428-0.0
30640-0.0
hypertension
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

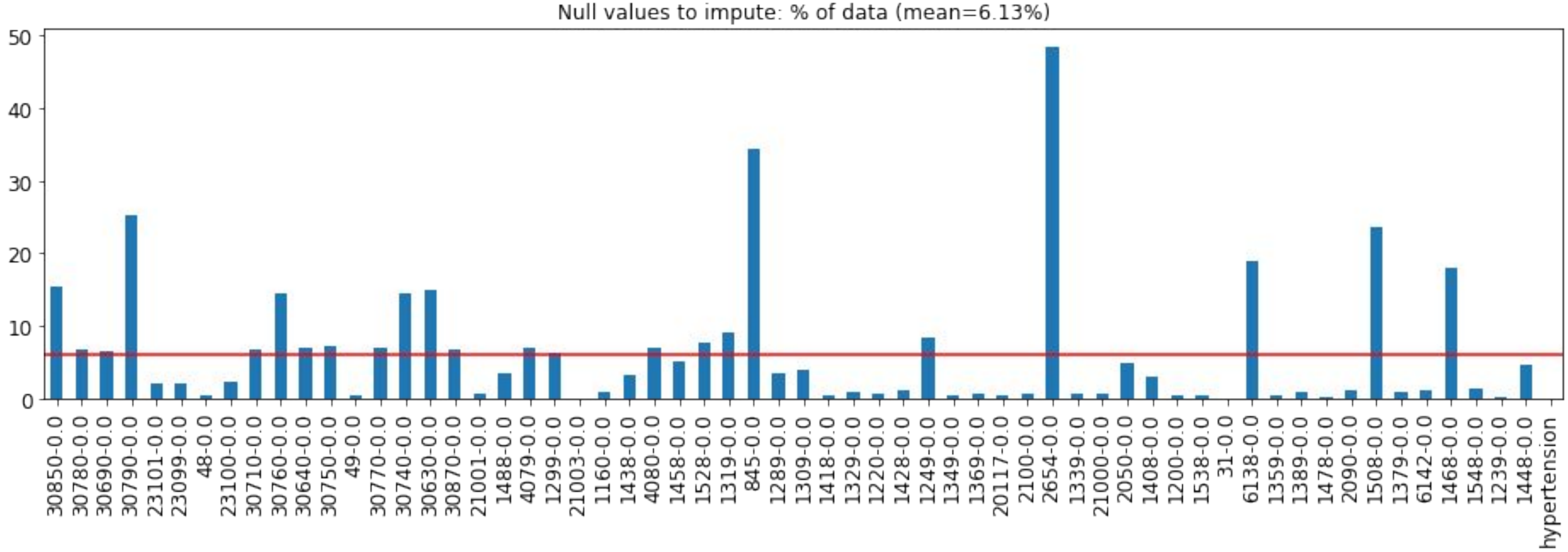
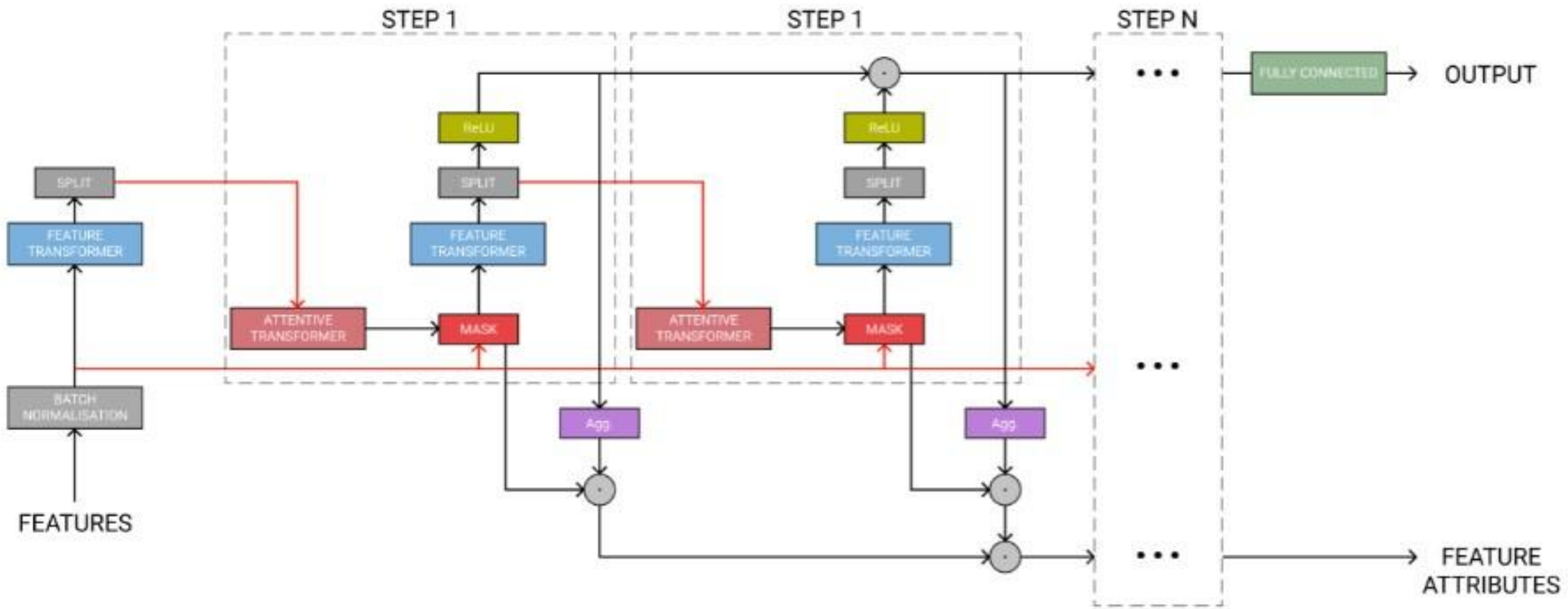


Figure X. *Data Preprocessing. Null value rates for each feature before implementation of MissForest imputation. All features with >=65% null values were discarded.*

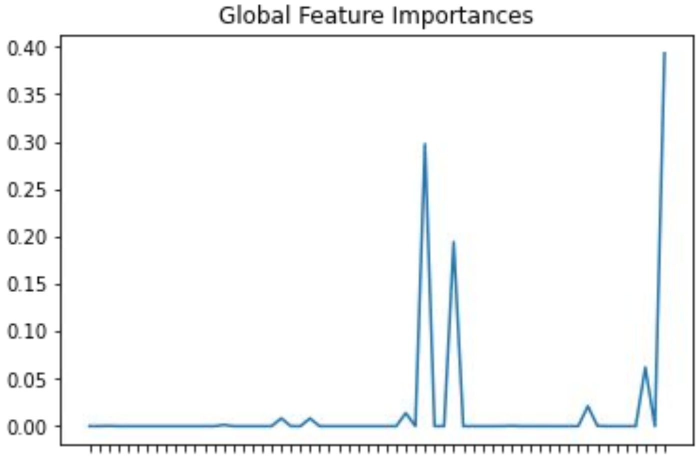
TabNet Architecture

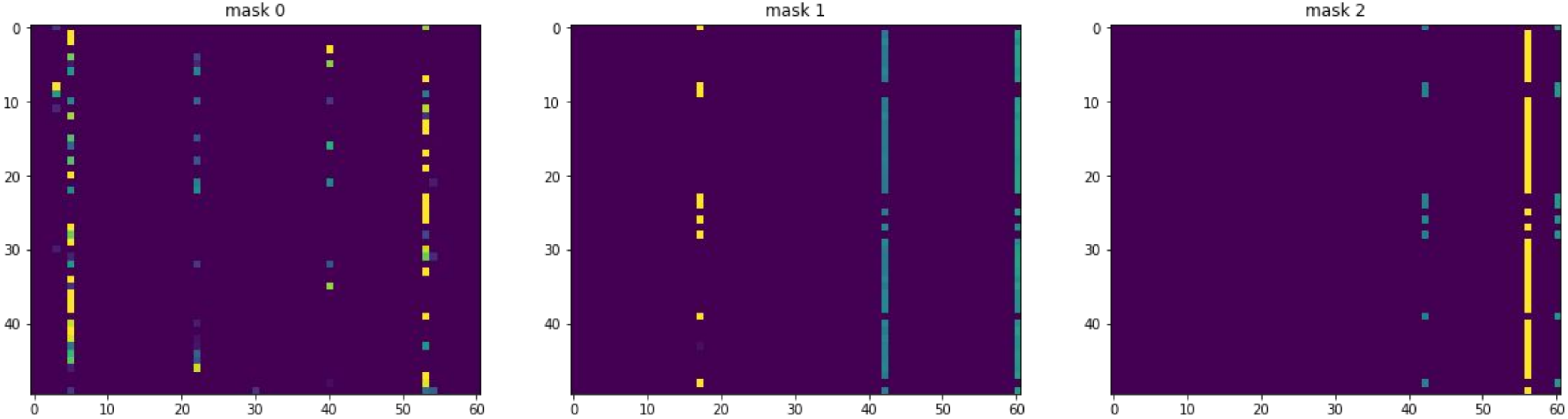
TabNet Model Architecture



Global & Local Feature Importance

- **Top contributing features:** hypertension, age, whole body fat-free mass, employment status, sex, ethnic background
- **31/61 features are not contributing at all**
- In general feature contribution is sparse over all inputs. Need to determine how well dataset captures problem-space





IBM AIF360

Adversarial Robustness 360

↳ (ART)

github.com/IBM/adversarial-robustness-toolbox

art-demo.mybluemix.net

AI Fairness 360

↳ (AIF360)

github.com/IBM/AIF360

aif360.mybluemix.net

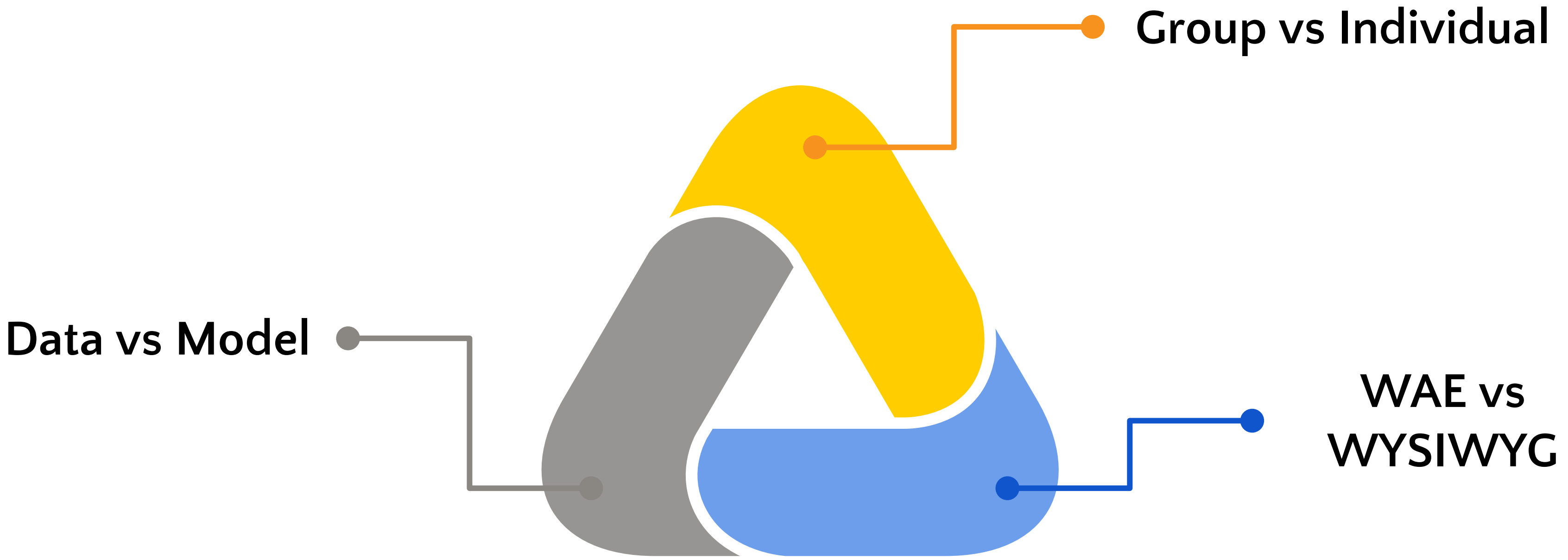
AI Explainability 360

↳ (AIX360)

github.com/IBM/AIX360

aix360.mybluemix.net

Perspectives of Fairness



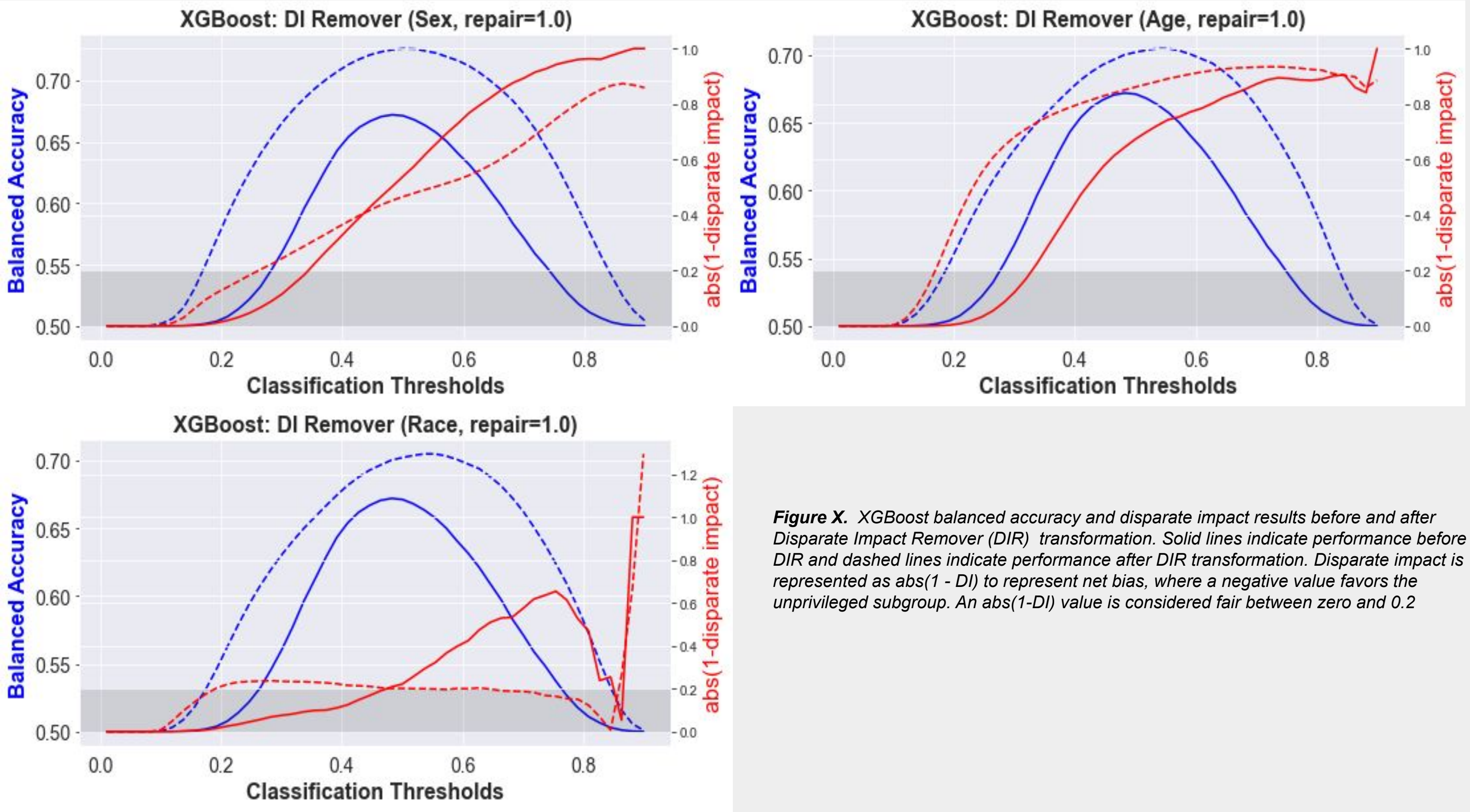


Figure X. *XGBoost balanced accuracy and disparate impact results before and after Disparate Impact Remover (DIR) transformation. Solid lines indicate performance before DIR and dashed lines indicate performance after DIR transformation. Disparate impact is represented as abs(1 - DI) to represent net bias, where a negative value favors the unprivileged subgroup. An abs(1-DI) value is considered fair between zero and 0.2*

XGBoost: Fairness after Reweighing

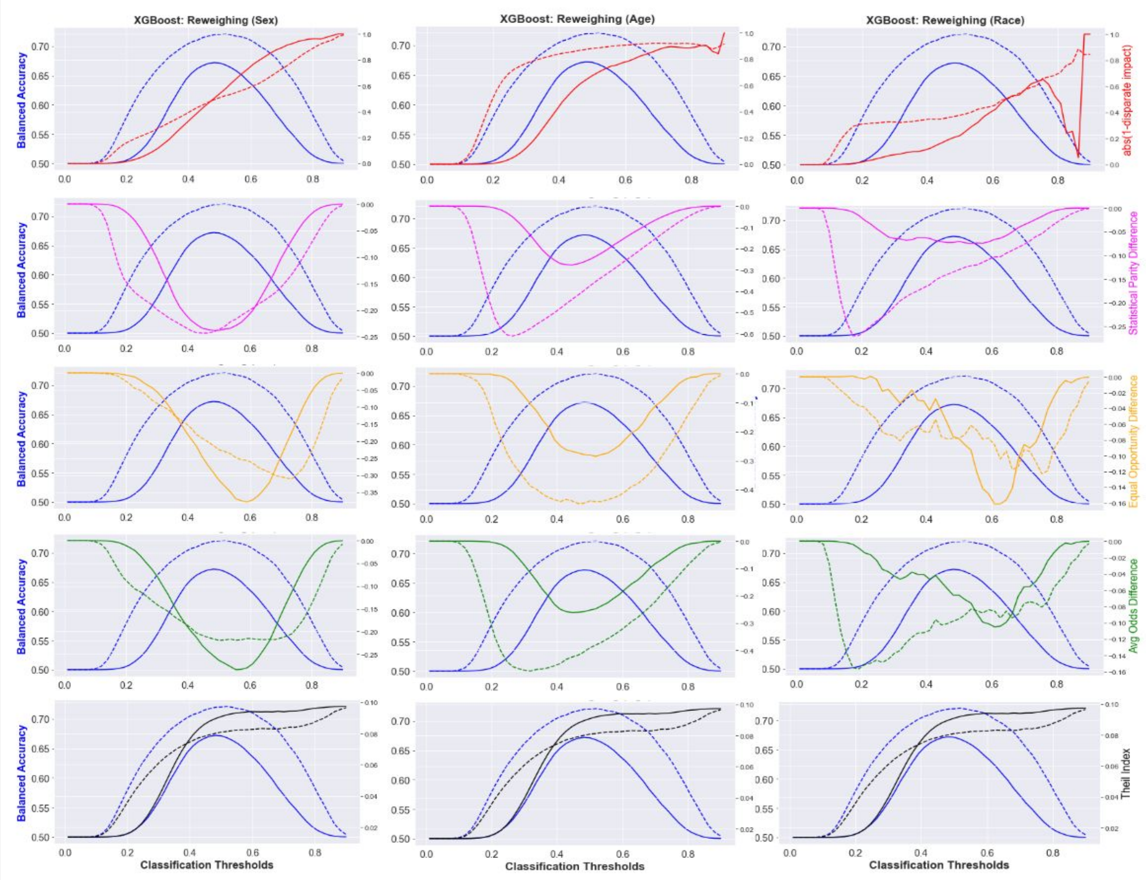
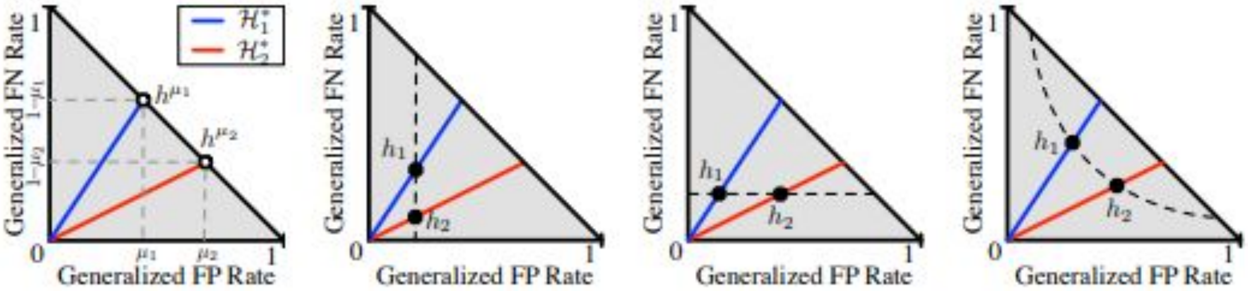


Figure X. *XGBoost fairness metrics & balanced accuracy vs classification decision thresholds before and after implementation of AIF360 Reweighting preprocessing*



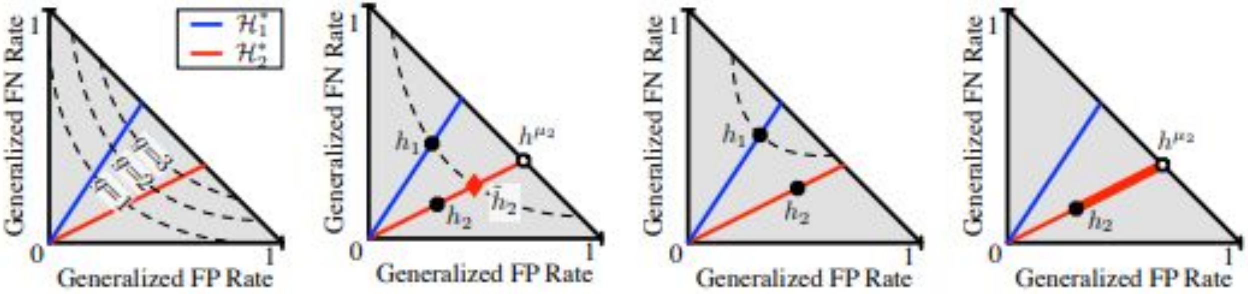
(a) Possible cal. classifiers $\mathcal{H}_1^*, \mathcal{H}_2^*$ (blue/red).

(b) Satisfying cal. and equal F.P. rates.

(c) Satisfying cal. and equal F.N. rates.

(d) Satisfying cal. and a general constraint.

Figure 1: Calibration, trivial classifiers, and equal-cost constraints – plotted in the false-pos./false-neg. plane. $\mathcal{H}_1^*, \mathcal{H}_2^*$ are the set of cal. classifiers for the two groups, and h^{μ_1}, h^{μ_2} are trivial classifiers.



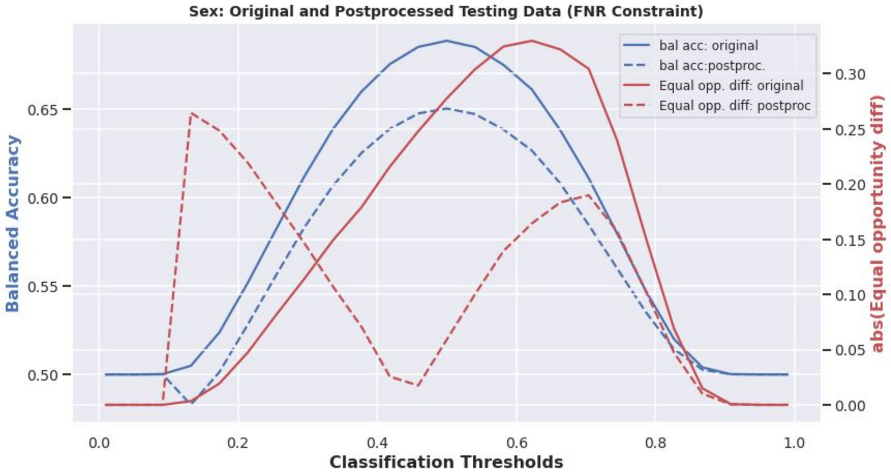
(a) Level-order curves of cost. Low cost implies low error rates.

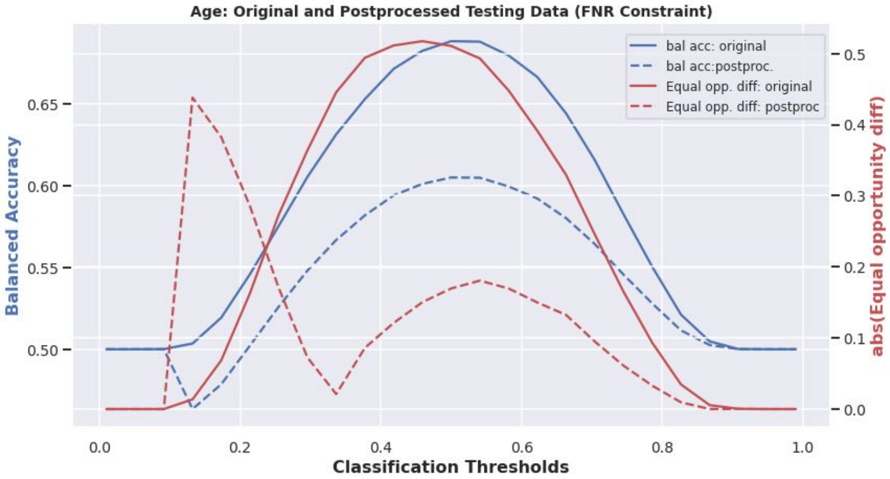
(b) Usually, there is a calibrated classifier $\tilde{h}_2$ with the same cost of h_1.

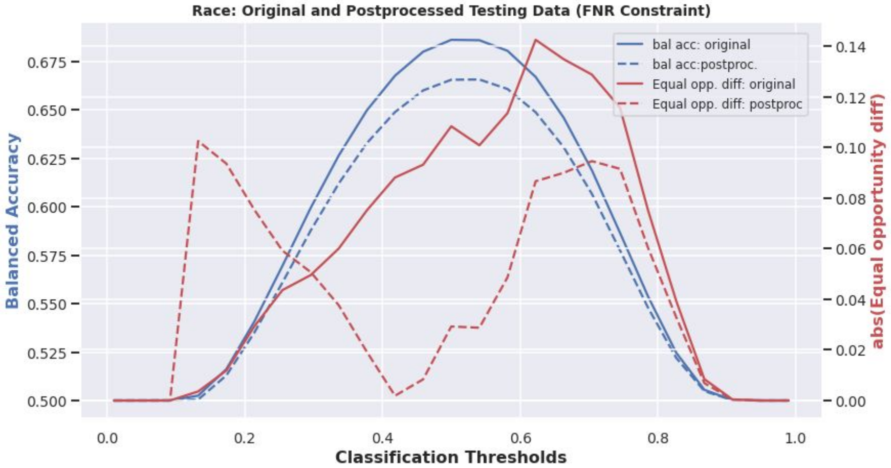
(c) Cal. and equal-cost are incompatible if h_1 has high error.

(d) Possible cal. classifiers for G_2 (bold red) by mixing h_2 and h^{μ_2}.

Figure 2: Calibration-Preserving Parity through interpolation.


Sex: Original and Postprocessed Testing Data (FNR Constraint)
bal acc: original
bal acc:postproc.
Equal opp. diff: original
Equal opp. diff: postproc
Balanced Accuracy
abs(Equal opportunity diff)
Classification Thresholds
0.50
0.55
0.60
0.65
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0


Age: Original and Postprocessed Testing Data (FNR Constraint)
bal acc: original
bal acc:postproc.
Equal opp. diff: original
Equal opp. diff: postproc
Balanced Accuracy
abs(Equal opportunity diff)
Classification Thresholds
0.50
0.55
0.60
0.65
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0


Race: Original and Postprocessed Testing Data (FNR Constraint)
bal acc: original
bal acc:postproc.
Equal opp. diff: original
Equal opp. diff: postproc
Balanced Accuracy
abs(Equal opportunity diff)
Classification Thresholds
0.500
0.525
0.550
0.575
0.600
0.625
0.650
0.675
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.0
0.2
0.4
0.6
0.8
1.0