BIAS MITIGATION & POLICY

Assignment 3: Tutorial Presentation
GRAD-E1394
Hannah, Carlo & Jorge

RELEVANCE OF THE TOPIC

For public policy

Analysis

UK risks scandal over 'bias' in AI tools in use across public sector

Kiran Stacey

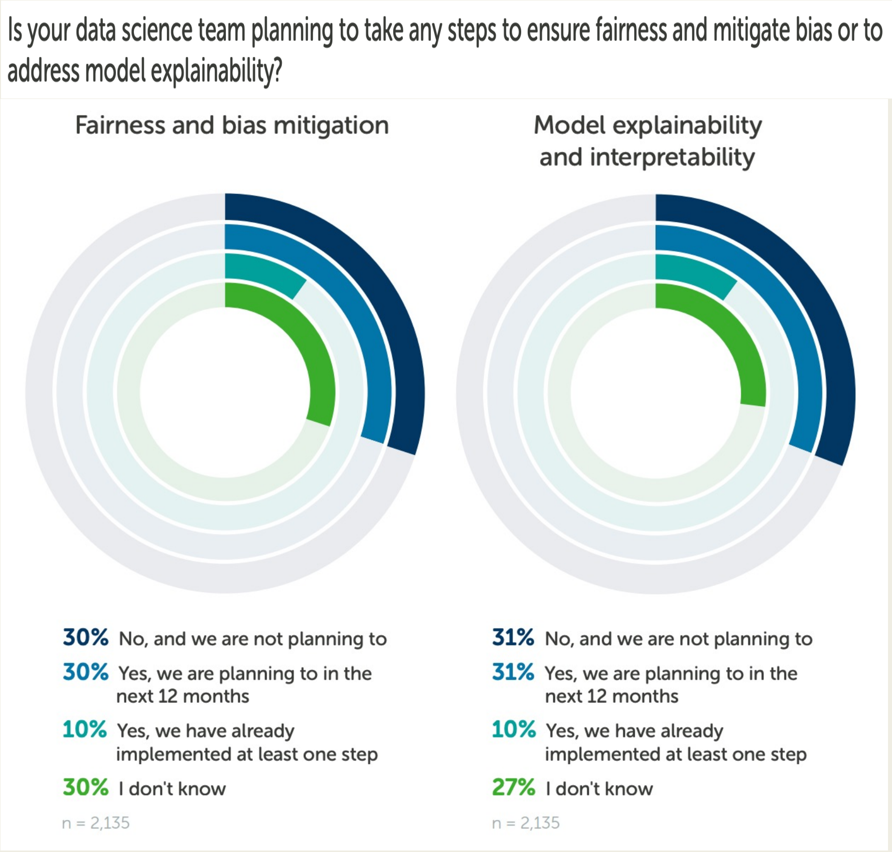
Systems operating across government departments and police forces raise concerns about accountability and discrimination

- UK officials use AI to decide on issues from benefits to marriage licences

https://www.theguardian.com/technology/2023/oct/23/uk-risks-scandal-over-bias-in-ai-tools-in-use-across-public-sector

Relevance of the topic

- Deep learning models trained on biased data will lead to biased results with harmful real-world consequences
  - Model learns from (biased) patterns in the data
  - Already existing historical/societal biases are perpetuated or even amplified
  - Examples show overpolicing and unfair credit scoring decisions for marginalised groups
- As a result, trust in DL applications erodes
- Although these harmful consequnces are widely known, bias mitigation is still no part of the established model pipeline



Source: 2021 Anaconda State of Data Science Report

Relevance of the topic

„Touch upon very private & sensitive areas of citizens‘ lives“ a.o.
- Identities
- demographic attributes
- preferences
- future behaviour

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, Nicol Turner Lee, Paul Resnick, and Genie Barton May 22, 2019

Algorithms

„Play a critical role in
1. identifying and
2. mitigating

biases, while ensuring that the technologies continue to make positive economic and societal benefits“

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, Nicol Turner Lee, Paul Resnick, and Genie Barton May 22, 2019

Policymakers

BACKGROUND

Task & Model

Bias & Fairness, two sides of the same coin

Bias:

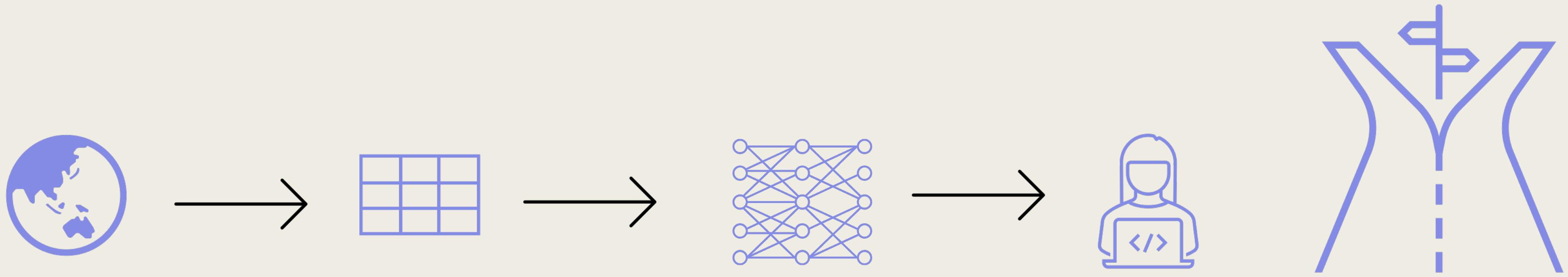
the presence of prejudice in the form of a systematic error in a model's predictions or decisions

Fairness:

the absence of prejudice or preference for an individual/group based on their characteristics

https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

Different kinds of biases



Bias in the world around us

Selection bias
Representation bias
Measurement bias
…

Algorithmic bias
Evaluation bias
Aggregation bias
…

Presentation bias
Interpretation bias
…

Harmful decisions

Characteristics of most common biases

Data bias

- **Representation:**
  Parts of the population have lower/higher probability of being selected

- **Measurement:**
  Available data does not realistically resemble the population

- **Exclusion:**
  Parts of the population are absent in the data

- **Historical**
  Data does not reflect current reality

Model/algorithmic bias

- **Model development:**
  Bias in training, model selection, and metric selection

- **Aggregation bias:**
  Single prediciton model for distinct populations

- **Model evaluation**
  Chosen metrics used for model evaluation are not representative for the general population

Three main approches for mitigation

Pre-processing

e.g.
reweighting

In-processing

e.g.
Adversarial Training

Post-processing

e.g.
calibration/debiasing

Metrics to assess bias

- huge variety of metrics
- challenge of measuring bias often lies in definition of fairness → many of the metrics cannot be balanced across subgroups at the same time

Binary Classification	Multi Classification	Regression	Clustering	Recommender System
Area Between ROC Curves	Multiclass Accuracy Matrix	Average Score Difference	Cluster Balance	Aggregate Diversity
Accuracy Difference	Confusion Matrix	Correlation difference	Minority Cluster Distribution Entropy	Average f1 ratio
Average Odds Difference	Confusion Tensor	Disparate Impact quantile	Cluster Distribution KL	Average precision ratio
Classification bias metrics batch computation	Frequency Matrix	MAE ratio	Cluster Distribution Total Variation	Average recall ratio
Cohen D	Multiclass Average Odds	Max absolute statistical parity	Clustering bias metrics batch computation	Average Recommendation Popularity
Disparate Impact	Multiclass bias metrics batch computation	No disparate impact level	Minimum Cluster Ratio	Exposure Entropy
Equality of opportunity difference	Multiclass Equality of Opportunity	Regression bias metrics batch computation	Silhouette Difference	Exposure KL Divergence
False negative rate difference	Multiclass statistical parity	RMSE ratio	Social Fairness Ratio	AExposure Total Variation
False positive rate difference	Multiclass True Rates	Statistical parity (AUC)		GINI index
Four Fifths	Multiclass Precision Matrix	Statistical Parity quantile		Mean Absolute Deviation
Statistical parity	Multiclass Recall Matrix	ZScore Difference		Recommender bias metrics batch computation
True negative rate difference				Recommender MAE ratio
Z Test (Difference)				Recommender RMSE ratio
Z Test (Ratio)				

https://www.holisticai.com/blog/measuring-and-mitigating-bias-using-holistic-ai-library

Deep dive: metrics

Disparate impact

- Goal: compare the proportion of individuals that receive a positive output across privileged vs. Unprivileged group
- Method: Calculate proportion of the unprivileged w. pos. outcome divided by privileged group w. pos. outcome

$$\frac{Pr(Y=1|D=\text{unprivileged})}{Pr(Y=1|D=\text{privileged})}$$

https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1

Equalized odds

- Goal: equalize the accuracy of prediction for all demographics
- Method: minimize the difference between the TPR and FPR of the privileged/unprivileged groups
- Define an acceptable error rate threshold

Men

	Qualified	Unqualified
Hired	56	9
Rejected	14	21
TOTAL	70	30

Women

	Qualified	Unqualified
Hired	24	21
Rejected	6	49
TOTAL	30	70

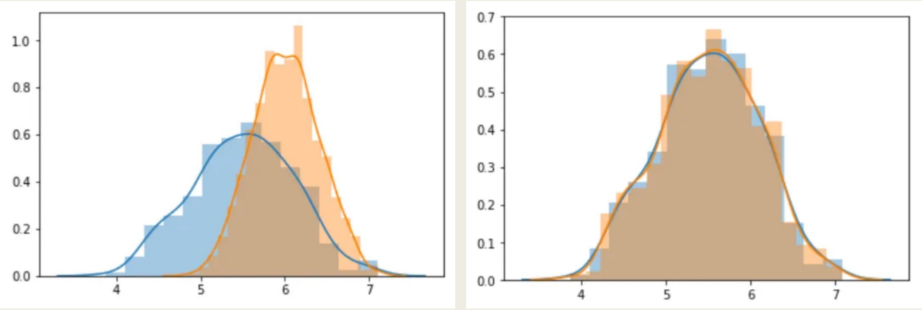
https://www.monitaur.ai/blog-posts/top-bias-metrics-and-how-they-work

Debiasing approaches

Disparative impact remover

→ pre-processing technique editing feature values

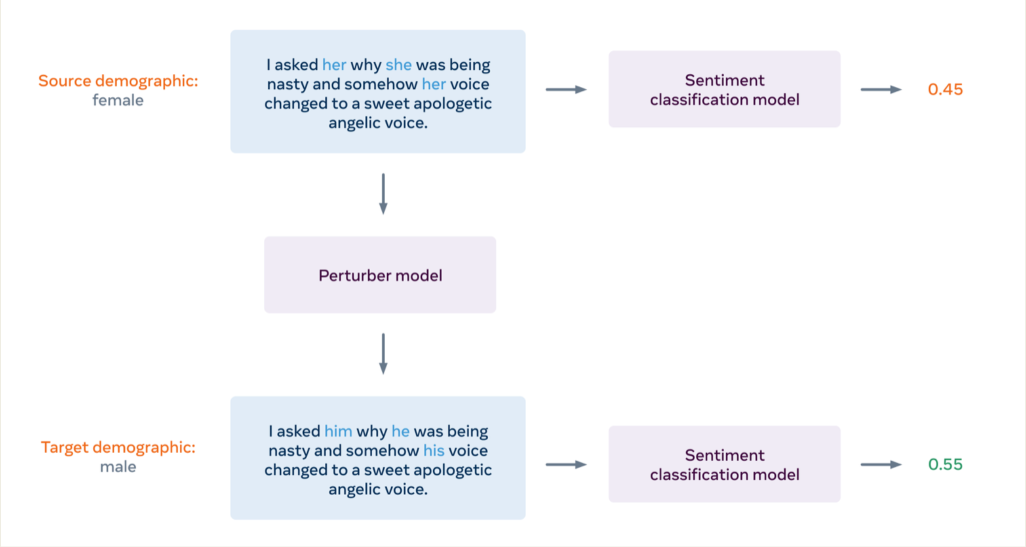
simply remove the protected variable from the data often doesn‘t help



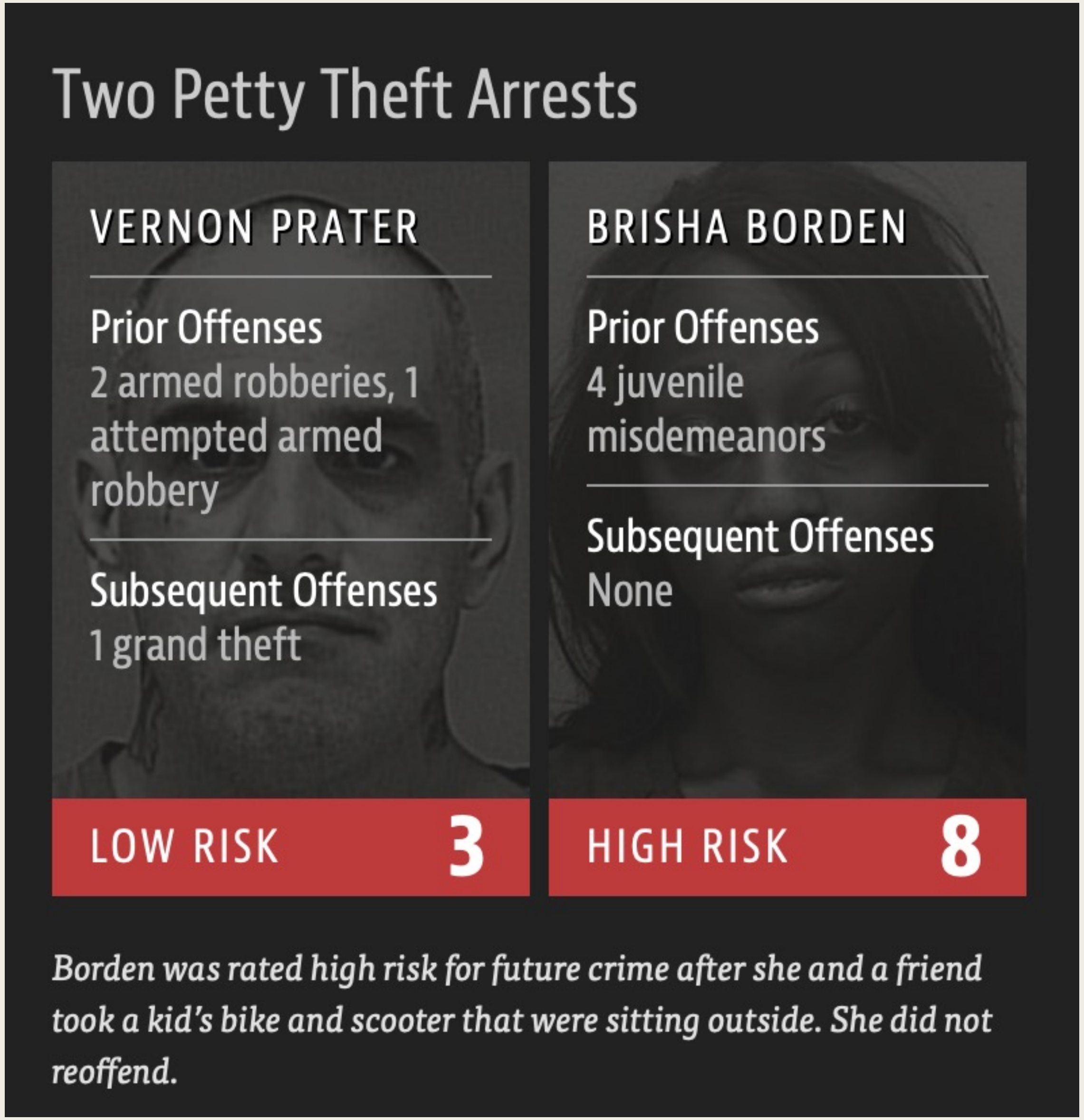
(Feldman et al. „Certifying and Removing Disparate Impact“. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2015, S. 259–68. *DOI.org (Crossref)*, https://doi.org/10.1145/2783258.2783311.)

Data augmentation

→ augment the dataset with modified demographic variations



Meta, Introducing two new datasets to help measure fairness and mitigate AI bias, May 23, 2022, https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias/



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Data

COMPAS Dataset:

- Historical entries related to defendants' demographics, criminal history, and the predicted recidivism risk assessment scores
- Used in US courtrooms
- likelihood of falsly categorizing black defendants as potential future offenders is notably higher, leading to mislabeling at nearly twice the rate observed for white defendants.

TUTORIAL

Sneak Peek

3 steps

1.- Identify bias through metrics

2.- Identify bias through a Feed Forward NN

3.- Fix bias to achieve more fairness

Aequitas library

open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias

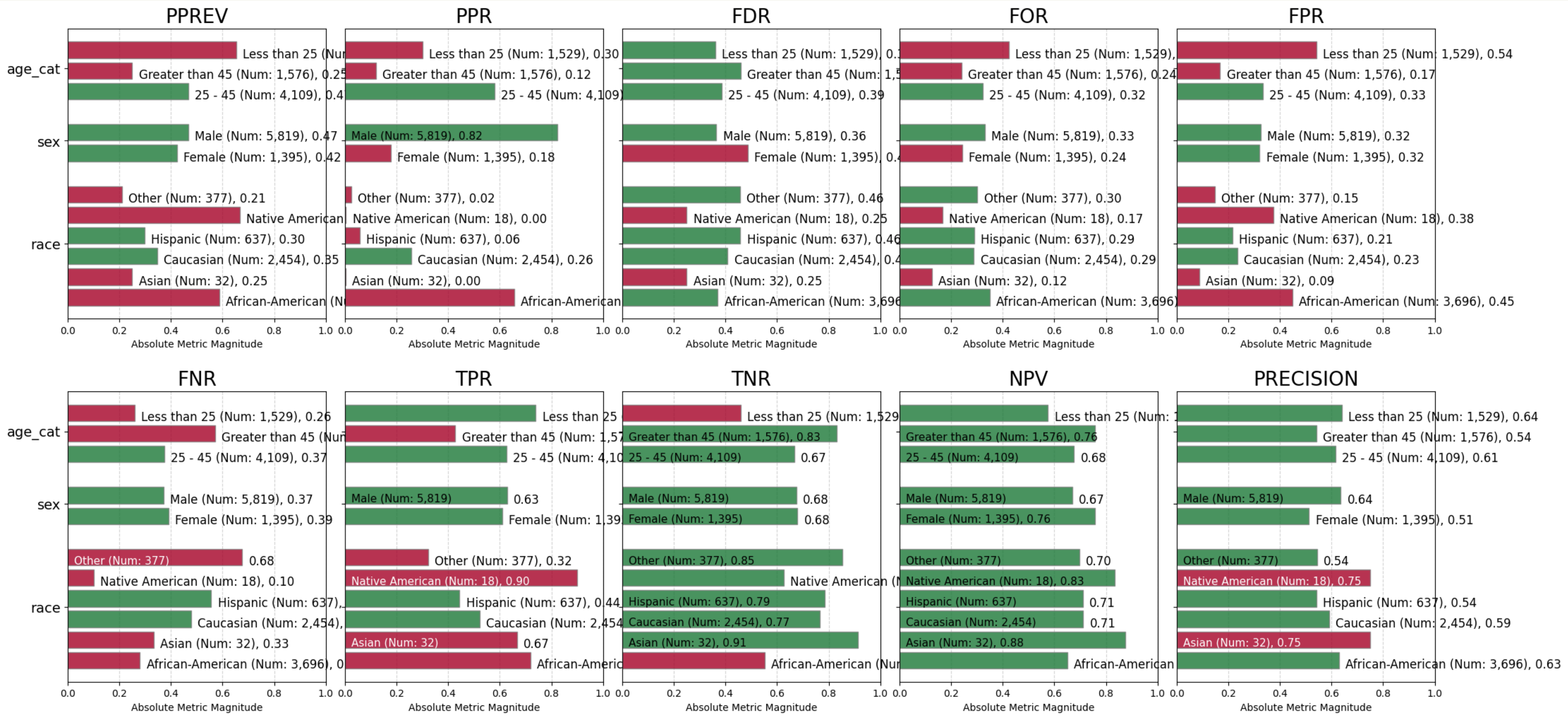
Replicate Compas Bias

Classify an individual depending on his/her reincidence risk.

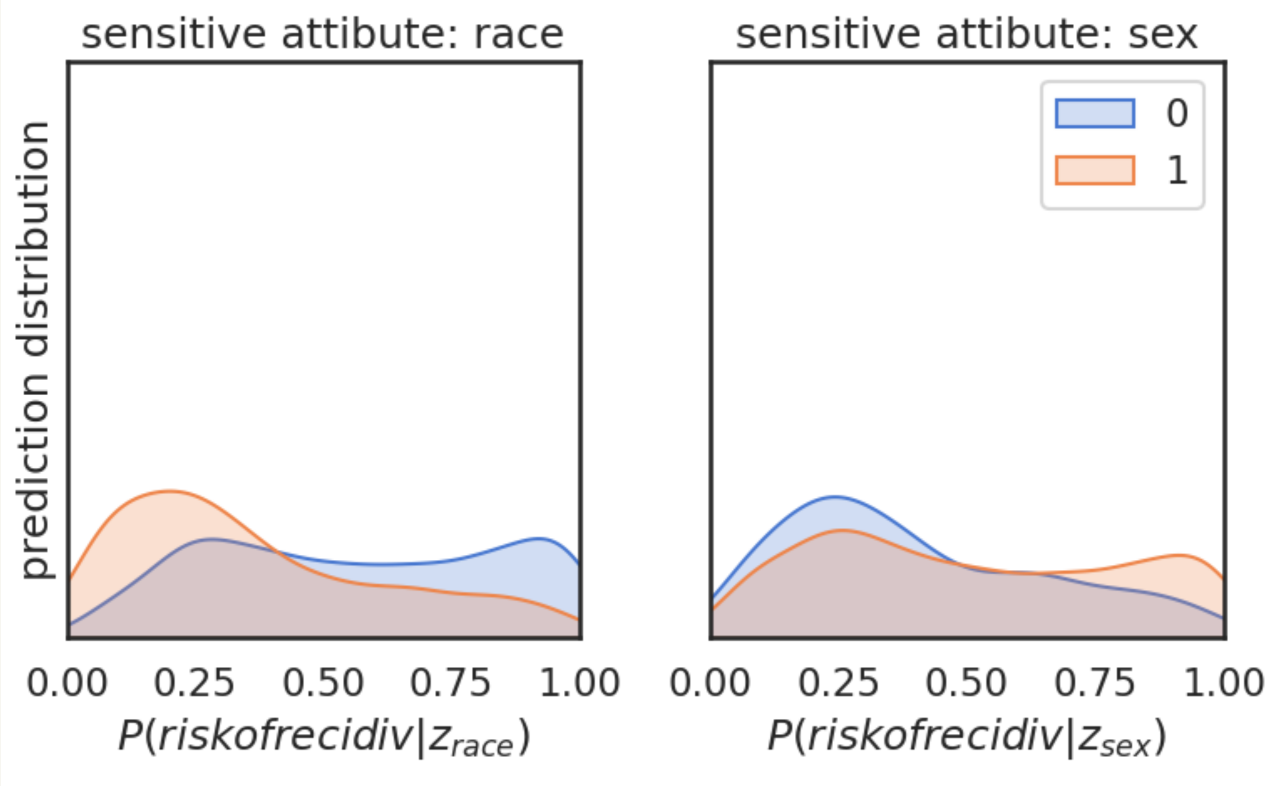
Disparate Impact Reparaing

Modify dataset to stop unfairly descrimination and use it to train a new machine learning model.

1.-Identify bias through metrics



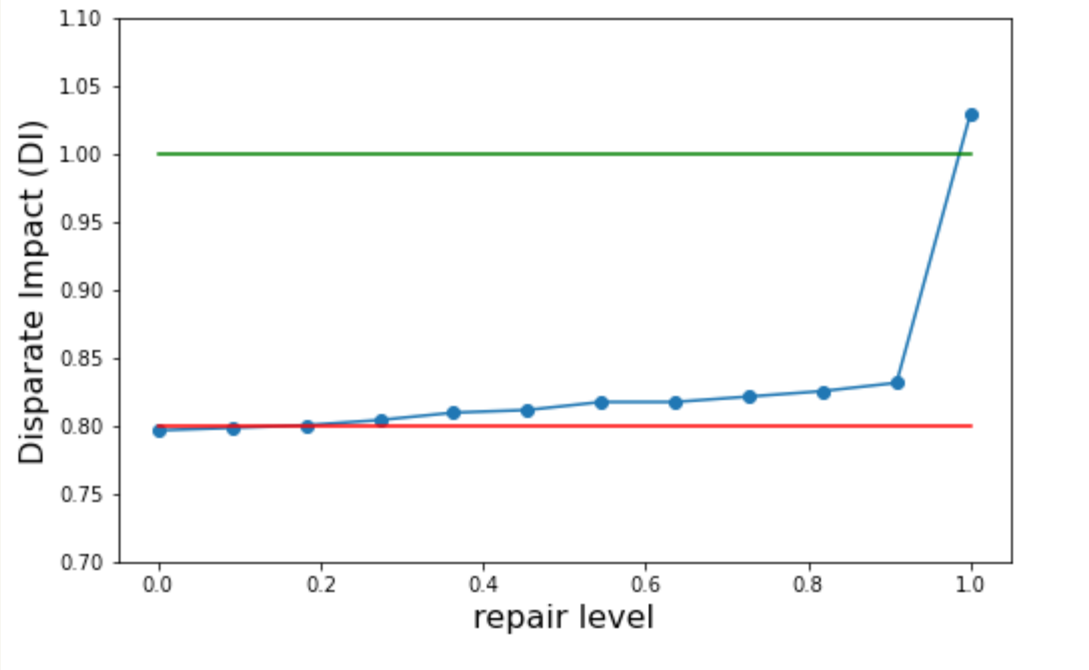
2.-Identify bias through a Feed Forward NN




```
def nn_classifier(n_features):
    inputs = Input(shape = (n_features,))
    dense1 = Dense(40, activation = 'relu')(inputs)
    dropout1 = Dropout(.4)(dense1)
    dense2 = Dense(40, activation = 'relu')(dropout1)
    dropout2 = Dropout(.3)(dense2)
    dense3 = Dense(32, activation = 'relu')(dropout2)
    dropout3 = Dropout(.3)(dense3)
    outputs = Dense(1, activation = 'sigmoid')(dropout3)
    model = Model(inputs = [inputs], outputs = [outputs])
    opt = Adam(learning_rate=0.001)
    model.compile(loss = 'binary_crossentropy', optimizer = opt, metrics = ['accuracy'])
    return model
```

- Predictions based on race are not fair. White defendants are more likely to be predicted as low-risk than black defendants, even when they have similar risk factors.
- Predictions based on sex are slightly fairer. Women are slightly more likely to be predicted as low-risk than men, but this difference is not as pronounced as the difference seen between race groups.

3.-Fix bias to achieve fairness




```
Balanced accuracy = 0.8963
Statistical parity difference = -0.0099
Disparate impact = 0.9777
Average odds difference = -0.0205
Equal opportunity difference = -0.0413
Theil index = 0.1155
```

Resources

- Angwin et al., 2023. "There’s software used across the country to predict future criminals. And it’s biased against blacks". https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Anaconda, 2021 State of Data Science Report. https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-2021-SODS-Report-Final.pdf
- Clark, Andrew. September 19, 2022. "Top bias metrics and how they work". Monitaur. https://www.monitaur.ai/blog-posts/top-bias-metrics-and-how-they-work
- Feldman et al. „Certifying and Removing Disparate Impact“. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, S. 259–68. DOI.org (Crossref), https://doi.org/10.1145/2783258.2783311.)
- Gichoya, Judy Wawira, et al. "AI pitfalls and what not to do: mitigating bias in AI." The British Journal of Radiology 96.1150 (2023): 20230023
- Lee, Nicol Turner, Paul Resnick, and Genie Barton. "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms." Brookings Institute: Washington, DC, USA 2 (2019)
- Meta, May 23, 2022. "Introducing two new datasets to help measure fairness and mitigate AI bias". https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias
- Muñoz et al., 2023. "Measuring and Mitigating Bias: Introducing Holistic AI's Open-Source Library". https://www.holisticai.com/blog/measuring-and-mitigating-bias-using-holistic-ai-library
- Regean, Mary. 2021. "Understanding bias and fairness in AI system". https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3