

# Investigating discrimination bias in predictive modelling

Sara Thiringer   Jonathan Rittmo

26/10/2020

## Background: The Problem

In recent years, many scandalous examples have shown that statistical models trained on large amounts of data can “act” discriminatory. Examples include:

- ▶ Adds of high-income jobs being shown less frequently to women, presumable because they've been predicted to be less interested or suitable<sup>1</sup>
- ▶ Black people's health status being underestimated, leading to inappropriate health care measures<sup>2</sup>
- ▶ Black people being predicted a higher risk for crime recidivism, leading to higher penalties<sup>3</sup>

---

<sup>1</sup>@datta\_automated\_2015

<sup>2</sup>@obermeyer\_dissecting\_2019

<sup>3</sup>ProPublica (2016)

# Project Aims

- ▶ How can we quantify fairness in order to be able to evaluate algorithmic fairness?
- ▶ What methods are available to increase algorithmic fairness?  
In what type of situations do they apply? (i.e. In what kind of situations can we expect them to be successful?)

## Background: Why Discrimination Bias?

- ▶ Correlation between outcome  $y$  and protected characteristic  $x_p$
- ▶ Correlation between important predictors  $x_i$  and protected characteristic  $x_p$
- ▶ Undersampling of groups with protected characteristic  $x_p$

## Possible Solutions

<b>Pre-Processing</b>	<b>Training</b>	<b>Prediction</b>
Resampling	Penalty	Threshold
Mapping	Model bias	adjustments
Altering labels	Tuning for fairness	Alter predictions

We've chosen to work with resampling and threshold adjustment.

# Possible Goals

Demographic parity

$$P(Y = 1|X = 1) = P(Y = 1|X = 0)$$

Equalized odds

$$P(G = 1|X = 0, Y = 1) = P(G = 1|X = 1, Y = 1)$$



# Models

Model	Tuning
Random Forest	Predictors at each split
Artificial neural net	Number of hidden nodes
Logistic ridge regression	Penalisation
K-nearest neighbour (left out)	Number of neighbours
AdaBoost	Predictors at each split



## Accuracy and Fairness for the Initial Models

# Disparate Impact Removal

# Preferential Resampling

# Uniform Resampling

# Comparison

# Final Model Performance

# Conclusions

## Slide with R Output

```
summary(cars)
```

##	speed	dist
##	Min. : 4.0	Min. : 2.00
##	1st Qu.:12.0	1st Qu.: 26.00
##	Median :15.0	Median : 36.00
##	Mean :15.4	Mean : 42.98
##	3rd Qu.:19.0	3rd Qu.: 56.00
##	Max. :25.0	Max. :120.00



# Slide with Plot

