Responsible AI: Group Fairness

1. Case study: Male fashion models

The model agency Top Models uses an application (i.e., ADS) that creates a list of potential male fashion models based on their social media profile pictures.

The classification algorithm is trained on images of their current male model pool: https://www.wilhelmina.com/new-york/men/main/

Desirable features: short hair, beard, muscular, tall, etc.

Sensitive/protected attribute: Race (For the sake of this example, groups: White and Asian).

2. Train set

Imbalanced --> Asian males are underrepresented in the data --> problematic because the classifier learns the bias; i.e., it is very unlikely that a male model is Asian/it is very likely that a male model is White.

Asian males: High TN and FN

White males: High TP and FP

Representative --> Train data is representative of the population; i.e., there are more White male fashion models than Asian male fashion models.

- Historical bias --> Equal base rates across sensitive/protected attribute groups --> Independence.

Allow for unequal base rates --> investigate if the difference in equal base rates is caused by the sensitive/protected attribute --> Separation or Sufficiency.

3. Metrics

Asian males = A
white males = W

Actual = y

predicted = $\hat{y}$	0	1

0	not on the list, does not have desirable features
	A = 12
	W = 2
	Tn
		Fn
		not on the list, does have desirable features
		A = 14
		W = 4

1	Fp
	on list, does not have desirable features
	A = 2
	W = 10
		Tp
		on list, does have desirable features
		A = 3
		W = 19

Calculate (positive) base rates for Asian and White males:

$$\text{Base rate} = \frac{Fn + Tp}{Fn + Tp + Fp + Tn}$$

$$\text{Asian males} = \frac{14 + 3}{14 + 3 + 2 + 12} = 0.54$$ → of all the outcomes 54% was positive

$$\text{white males} = \frac{4 + 17}{4 + 17 + 10 + 2} = 0.63$$

Separation: 'Gotta catch 'em all' --> recall --> positive outcome is favorable (e.g., hiring example).

What proportion is truly positive (i.e., correctly classified as positive); A male that is on the list, and has the desirable features.

$$TpR = \frac{Tp}{Tp + Fn}$$

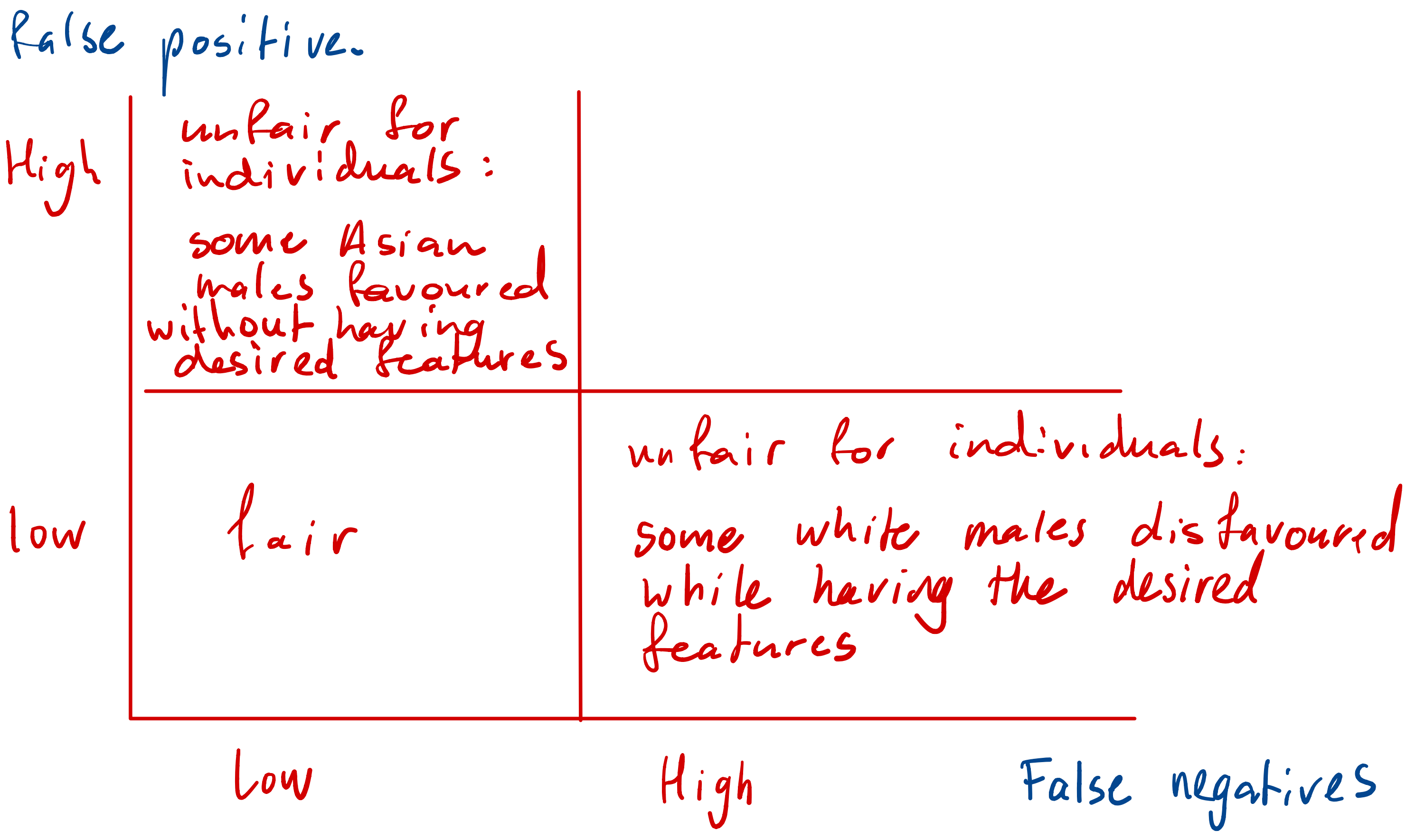
$$\text{Asian males} = \frac{3}{3 + 14} = 0.17$$ → of the positives only 17% was predicted correctly

↳ decrease Fn, increase Fp

$$\text{White males} = \frac{17}{17 + 4} = 0.80$$

↳ increase Fn, decrease Fp

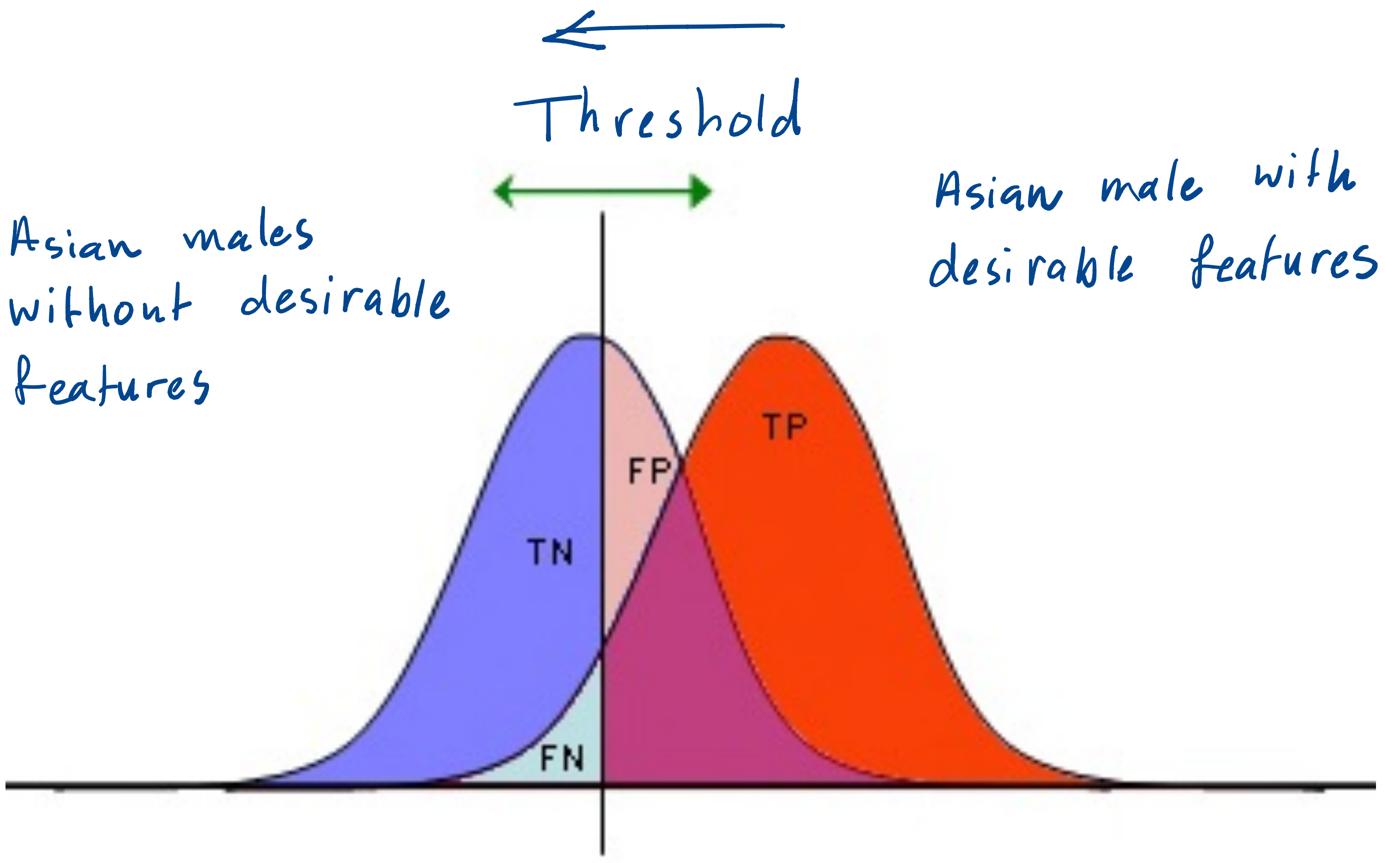
Equalized Opportunities: Equal chance to get a correct prediction, i.e., be on the list generated by the application, when they have the desirable features.



unfair for individuals: some Asian males favoured without having desired features

fair

unfair for individuals: some white males disfavoured while having the desired features



for white males the threshold goes the other way.