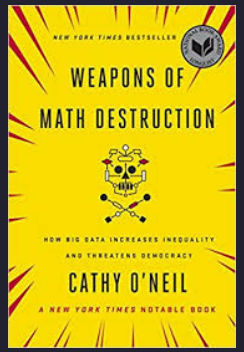
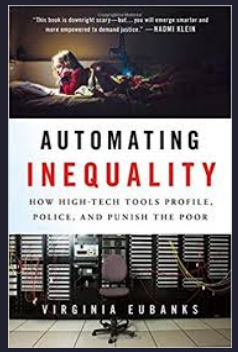
Big Data and AI

Algorithmic Fairness
Equitable Impact

Big Data Under Review



Weapons of Math Destruction



Automating Inequality

ProPublica

COMPASS

General Data Protection Regulation

European Union Regulatory Framework

- Curbs algorithmic decision making
- Defines “right to explanation”

Goodman, Flexman, 2016 - European Union regulation on algorithmic decision making and a “right to explanation”

Table of Content

1 Definitions of Fairness, “Catalog of Evils”

2 Framework for testing models for fairness

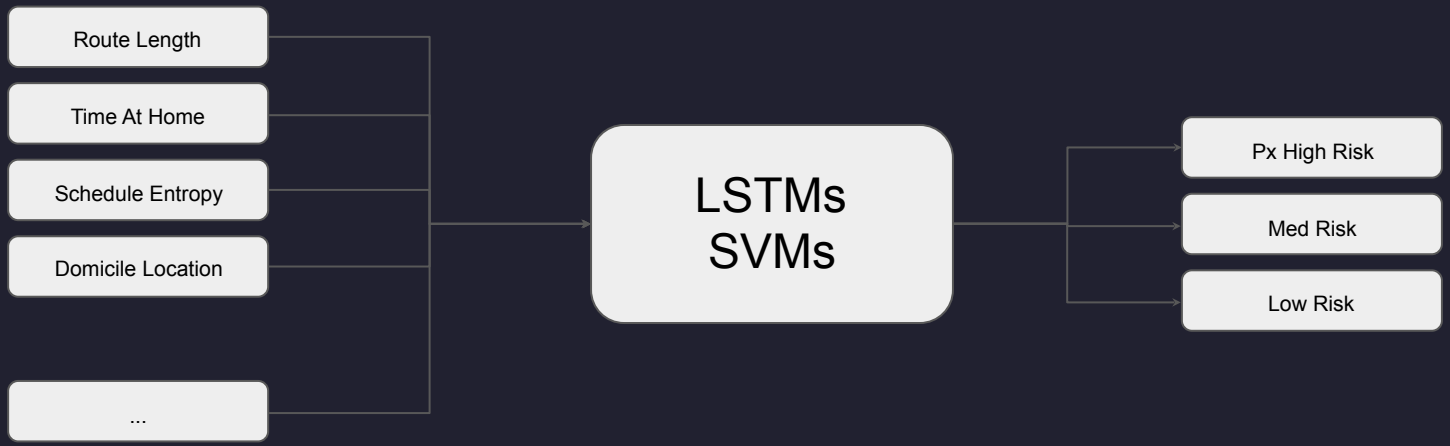
3 Case Study

4 Questions

Retention Model

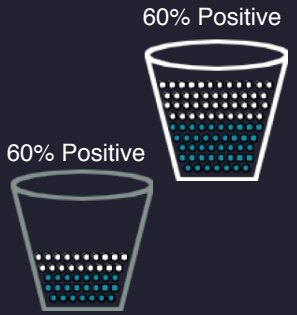
PROBLEM: Extreme driver turn-over in transportation industry, and high driver replacement cost.

SOLUTION: Retention model to predict drivers at high risk of leaving. Retain through intervention.



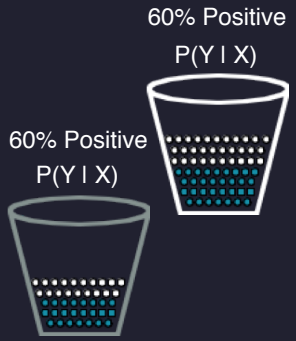
Model is deemed safe

Definitions of Fairness



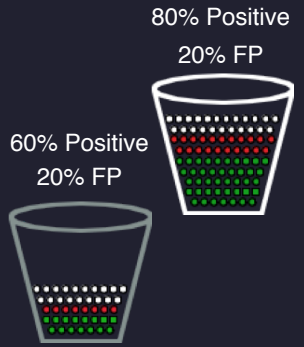
Statistical Parity

The rate of positives is the same between groups



Conditional SP

The rate of positives is the same between groups when controlled for legitimate factors



Predictive Equality

The rate False Positives is the same between groups.

Corbett-Davies et al, 2016 - Algorithmic decision making and the cost of fairness

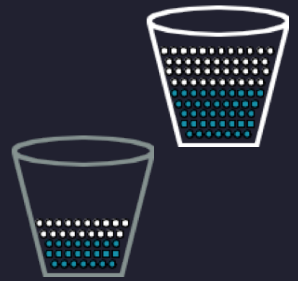
Definition of Fairness - 2

1. Calibration - for each group, the same fraction of people in each are positive

2. Balance for positive class - the average score for positive members in group A equals the average score of positive members in group B

3. Balance for negative class - the average score of negative members in group A equals the average score of negative members in group B

E(Na) = E(Nb)



E(Pa) = E(Pb)

Kleinberg et al, 2016 - Inherent Trade-Offs in the Fair Determination of Risk Scores

Catalog of Evils

1. Mis-calibration - systematic overestimation of group's risk
2. Redlining - ignoring relevant attributes
3. Sample bias and label bias - not representative population, wrong targets
4. Subgroup Validity - using differentially predictive features
5. Use of protected characteristics - race, gender, sex, or proxies
6. Disparate impact - adverse effects at different rates
7. Disparate error rates - unequal false positive rates

Dwork et al, 2011 Corbett-Davies, 2016

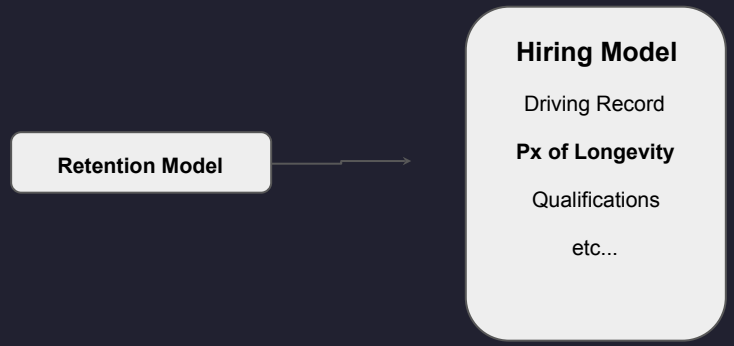
Retention for Hiring

Using retention model for maximizing the utility of each hire:

1. Can the retention model remain fair?
2. Is the hiring model non-discriminatory?

Goal - Predictive Equality:

1. Inversion Test
2. Identify sub-populations
3. Evaluate feature importance

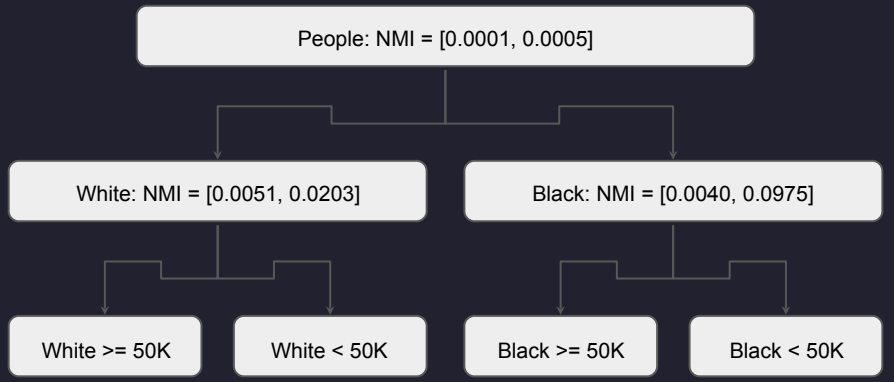


Existing Frameworks for Testing Fairness

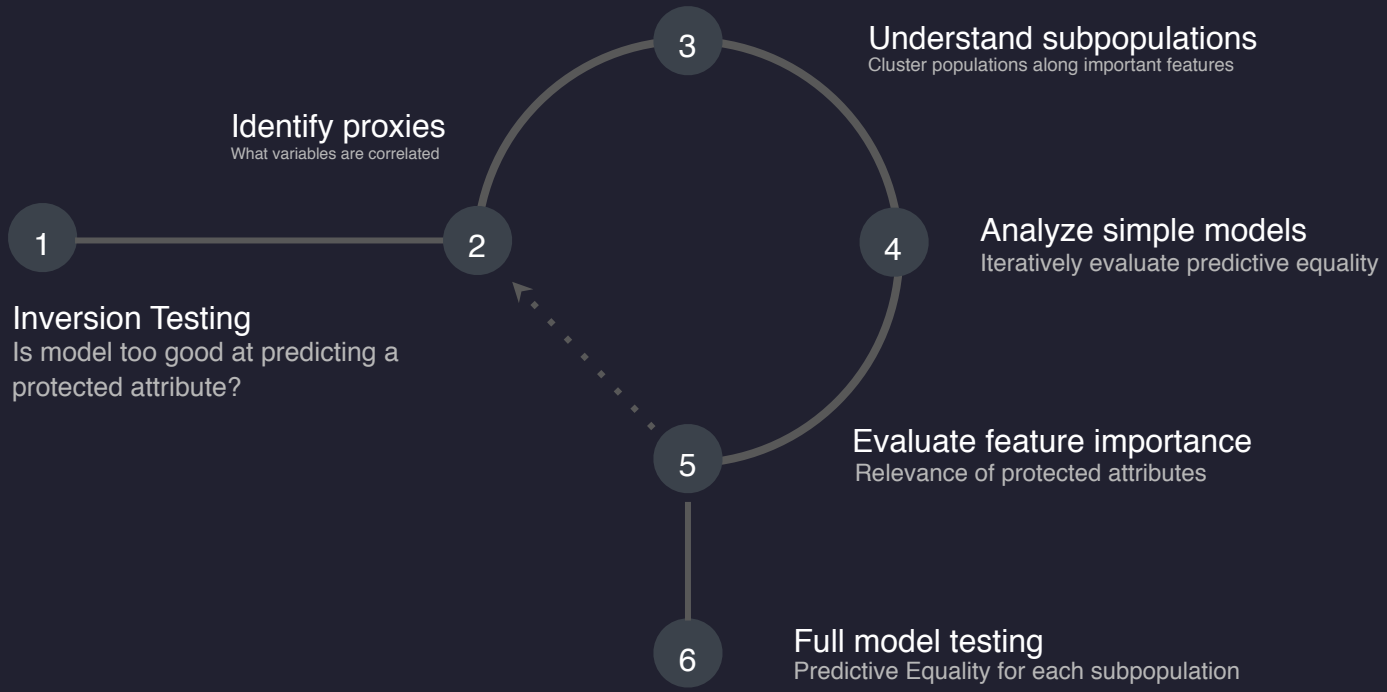
1. Enterprise tools:
    a. Microsoft - AI toolkit
    b. Facebook - FairFlow
    c. Accenture - Fairness Tool

1. Open Source tools:
    a. FairTest
    b. Fair ML

FairTest



Fairness Audit



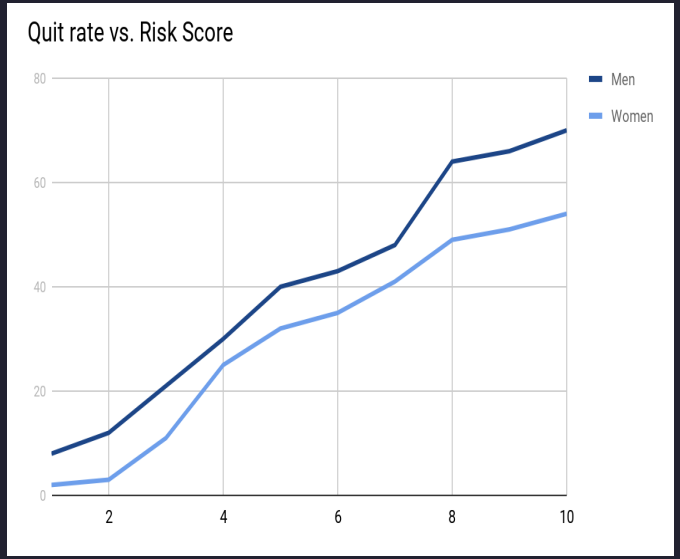
Fixing the models

Option 1: Thresholding scores

Option 2: ‘Repairing’ the data

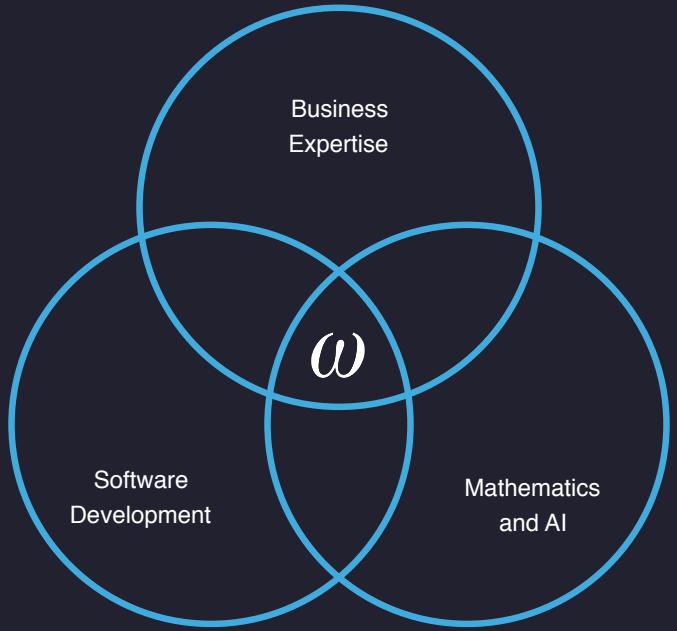
In cases of

- Different base rates
- Differentially predictive features
- Specific priorities for sub-population
- Correcting implicit bias in data
- etc.



(not actual for demonstration only)

Feldman et al, 2014 - Certifying and removing disparate impact



Confluence of science and business