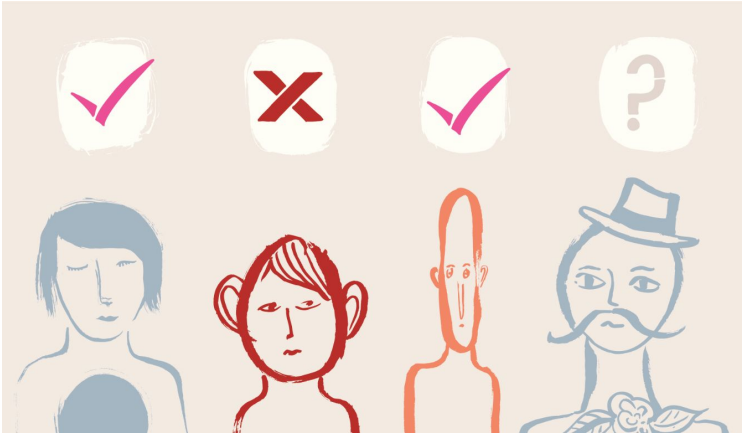
Algorithmic Discrimination and Equality of Opportunity

Saad Padela

The Humanity of Data

- Data is socially constructed, and inherits our human imperfections and biases with startling fidelity
- Instead of curbing the potential for systemic discrimination against disadvantaged groups, the use of data and algorithms may be expanding it



Trina Dalziel—Getty Images/Ikon Images

How Does Systemic Discrimination Occur?

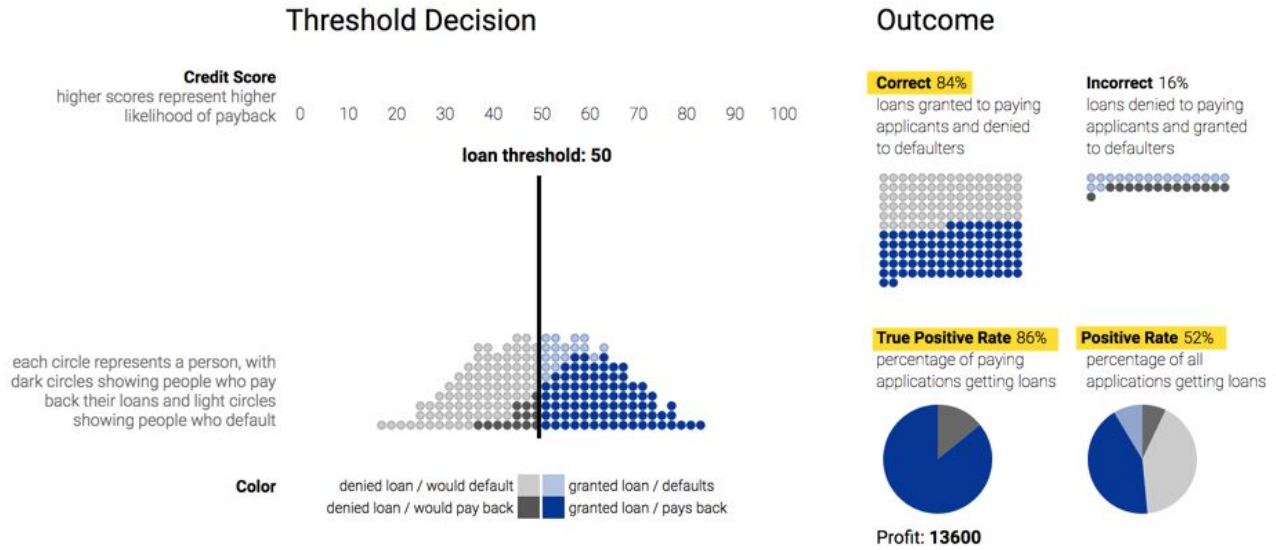
- Systemic discrimination occurs when members of a particular group are consistently at a disadvantage vis-a-vis decisions made by a predictive model
- Crucially, we consider this to be discrimination even if the outcome is correlated with membership in that particular group
- Three major reasons why the outcome may be correlated with the population segment:
    a. **Historical Reasons:** Historical imbalances in opportunity for various demographic segments
    b. **Data Quality:** Quality of the signal data coming in may be different for various groups
    c. **Model Quality:** Insensitivity/miscalibration of the model itself

Equality of Opportunity

- Concept proposed by Moritz Hardt (UC Berkeley), Eric Price, Nathan Srebro in 2016
- Well-suited for contexts where we can designate an “advantaged outcome” -- e.g., receiving a loan, acceptance to a college, etc.
- Equal opportunity has a precise mathematical definition, but informally it is best described with a few examples:
    - ***Loan Applications:*** Among applicants who will pay back the loan, probability of receiving it is equal among White/Black applicants
    - ***College Admissions:*** Among students who will successfully maintain a first-year GPA of 3.5, probability of admission to a school is equal among Male/Female applicants
- Shows that, given any learned predictor **Y**, we can adjust to **Y*** such that **Y*** satisfies the criteria of equal opportunity

Example: Simulating Loan Applications

- Google BigPicture visualization to simulate impact of loan thresholds



Example: Simulating Loan Applications

- Now suppose you had two populations (blue and orange)
- Scenario 2: Group Unaware
  - Applies a stricter standard to all members of group Orange (higher threshold)
- Scenario 3: Demographic Parity
  - Controls for acceptance rate, but virtuous applicants in group Blue are disadvantaged
- Scenario 4: Equal Opportunity
  - Harmonizes acceptance rate across demographic segments for individuals who would successfully pay back loan

Scenario 2: Group Unaware

Population	*Correct*	*Positive Rate*	*True Positive Rate*
Blue	79%	52%	81%
Orange	79%	30%	60%

Scenario 3: Demographic Parity

Population	*Correct*	*Positive Rate*	*True Positive Rate*
Blue	77%	37%	64%
Orange	84%	37%	71%

Scenario 4: Equal Opportunity

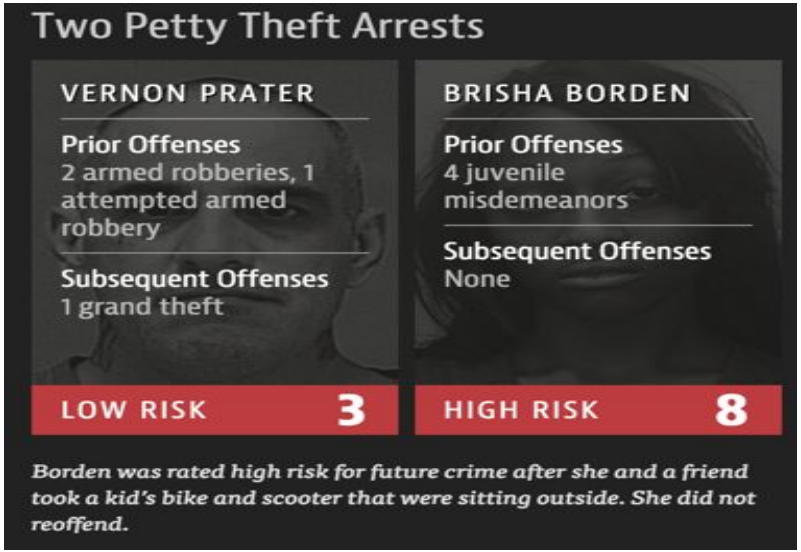
Population	*Correct*	*Positive Rate*	*True Positive Rate*
Blue	78%	40%	68%
Orange	83%	35%	68%

Algorithms for Criminal Justice

- Courts rely on third-party predictive algorithms to quantify the risk that a convicted criminal will commit a future crime (known as recidivism)
- Historically, judges have made these subjective determinations based on personal experience and professional expertise
- The introduction of an objective, data-driven algorithm into these settings seems like a sensible thing to do

Algorithms for Social Injustice (?)

- In 2016, ProPublica published an analysis of the COMPAS Recidivism Risk Score algorithm, which was being used by at least 9 state courts at the time
- ProPublica's analysis showed that Black defendants were almost twice as likely to be wrongly classified as "high-risk", whereas the opposite was true for White defendants



Source: ProPublica article

ProPublica's Methodology

- Link to full description of methodology
- Obtained two years worth of COMPAS scores (2013-2014) from Broward County Sheriff's Office via a public records request
- 11,757 pre-trial defendants were scored for "Risk of Recidivism" and "Risk of Violence"
- Scores range from 1-10 (highest), with 1-4 labeled as "Low Risk"
- Used public criminal records to build a profile of criminal history before/after assessment
- Determined race based on data from Broward County Sheriff's Office

A First Look At The Data

- Summary-level data shows that the algorithm is clearly picking up on race as a latent variable -- 56% of African-American defendants are labeled as “Low Risk”, vs. 78% of Caucasians
- Here’s the equal-opportunity test: A law-abiding African-American had a 31% chance of being incorrectly classified as Med/High Risk, versus 15% for Caucasians

	2-Year Recidivism					
Race:	*No*		*Yes*		*Total*	
African American	#	%	#	%	#	%
Low Risk	1,046	33%	743	23%	1,789	56% (1)
Med/High Risk	468	15% (3)	918	29%	1,386	44%
Total	1,514	48%	1,661	52%	3,175	100%

	2-Year Recidivism					
Race:	*No*		*Yes*		*Total*	
Caucasian	#	%	#	%	#	%
Low Risk	1,084	52%	564	27%	1,648	78% (2)
Med/High Risk	197	9% (4)	258	12%	455	22%
Total	1,281	61%	822	39%	2,103	100%

Operationalizing “Equal Opportunity”

- **Accuracy**: The ratio of correct predictions to the total number of defendants
- **Positive Rate:** The ratio of defendants who were classified as “Low Risk”
- **True Positive Rate:** The ratio of defendants *who subsequently did not recidivate* that were classified as “Low Risk”

race	total	accuracy	positive_rate	true_positive_rate
Asian	31	0.84	0.84	0.96
Other	343	0.67	0.78	0.85
Caucasian	2103	0.64	0.78	0.85
Native American	11	0.82	0.55	0.83
Hispanic	509	0.64	0.74	0.81
African-American	3175	0.62	0.56	0.69

Finding An Equal Opportunity Classifier

- In this case, we envision an equal opportunity classifier to be a decision rule that results in harmonized true positive rates across various values of race
- The original COMPAS algorithm generates scores 1-10, with 4 being the cutoff for “Low Risk”
- Since the score is not a continuous variable, we cannot simply pick varying thresholds by population
- Instead, we generate a random variable (picked from a different distribution depending on race) and use it to modify the raw score, then apply the original decision rule
- Under this formulation, finding “equal opportunity” becomes an optimization problem on the various parameters of the random variables

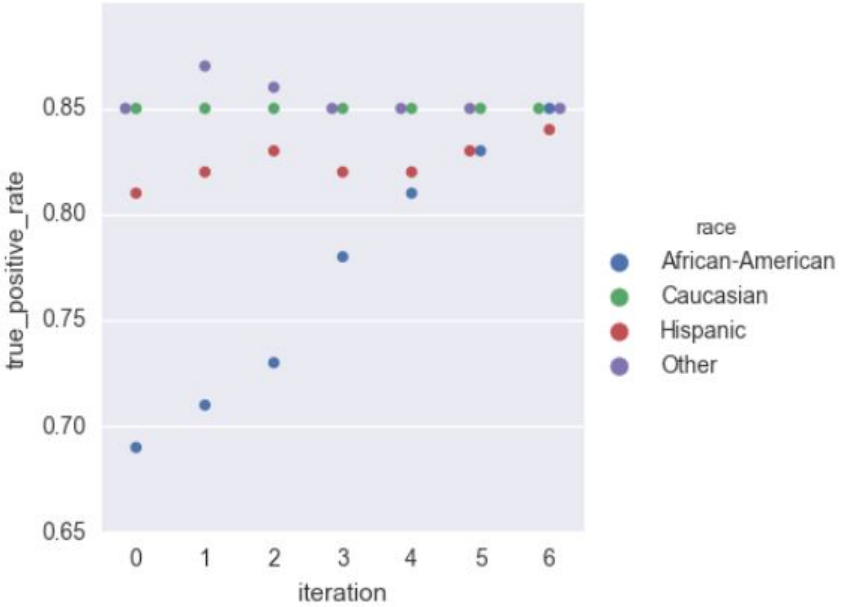
Final Result

- Optimization took 6 iterations to converge, but successfully harmonized True Positive Rates to 85%, granting “equal opportunity” among defendants who did not recidivate

Final Weights

Race	Mean	Std
African-American	2.43	2.46
Caucasian	-1.37*	0.18
Hispanic	0.53	1.32
Other	-1.53*	0.08

* Negative weights are ignored



Final Result (contd)

- By definition, an equal opportunity classifier entails a cost in accuracy; in this case, the accuracy of prediction dropped from 62% to 58% among African-Americans
- The interesting consequence here is that the incremental cost of equal opportunity (vs. optimal predictor) is borne by the organization making use of the algorithm, rather than the individual

	True Positive Rate			Accuracy		
Race	*Original*	*Final*	*% Δ*	*Original*	*Final*	*% Δ*
African-American	0.69	0.85	23%	0.62	0.58	-6%
Caucasian	0.85	0.85	0%	0.64	0.64	0%
Hispanic	0.81	0.84	4%	0.64	0.64	0%
Other	0.85	0.85	0%	0.67	0.67	0%