Fairness in Machine Learning

ECE324, Winter 2023

Michael Guerzhoy

Content from Moritz Hardt, Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel

COMPAS

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

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

https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

https://www.propublica.org/article/technical-response-to-northpointe

https://www.liebertpub.com/doi/pdf/10.1089/big.2016.0047

COMPAS

- "Correctional Offender Management Profiling for Alternative Sanctions"
  - Developed by Northpointe (currently Equivant)
  - Used by *a lot* of probation departments to assess the likelihood of a defendant becoming a recidivist
  - Defendants who are defined as medium or high risk are more likely to be detained before trial
    - (N.B., this is only suggestive of importance)
  - Race is not an input to the algorithm

COMPAS Probation Risk and Needs Assessment Questionnaire

OFFENDER NAME: NYSID: STATUS:
RACE: SEX: DOB:
DATE OF ASSESSMENT: MARITAL STATUS:
SCALE SET: Full COMPAS Assessment v2 AGENCY/COUNTY NAME:

PART ONE: CRIMINAL HISTORY / RISK ASSESSMENT

CURRENT CHARGES

What offenses are covered by the current charges (check all that apply)?

Homicide	Arson	Property/Larceny
Assault	Weapons	Fraud
Robbery	Drug Sales	DWI / DWAI
Sex Offense (with force)	Drug Possession	AUO
Sex Offense (without force)	Burglary	Other

1 Do any of the current offenses involve domestic violence?
Yes No

2 What offense category represents the most serious current charge?
Misdemeanor Non-Assault Felony Assaultive Felony

3 Was there any degree of physical injury to a victim in the current offense?
Yes No

4 Based on your judgment, after reviewing the history of the offender from all known sources of information (PSI, police reports, prior supervision, victim, etc.) does the defendant demonstrate a pattern of violent behavior against people resulting in physical injury?
Yes No

http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf

PART TWO: NEEDS ASSESSMENT

A. ASSOCIATES / PEERS

17 The offender has peers and associates who *(check all that apply)*:

Use illegal drugs	Lead law-abiding lifestyles
Have been arrested	Are gainfully employed
Have been incarcerated	Are involved in pro-social activities
None	

18 What is the gang affiliation status of the offender :

- Current gang membership
- Previous gang membership
- Not a member but associates with gang members
- None

19 Does the offender have a criminal alias, a gang-related or street name?

Yes No

20 Does unstructured idle time contribute to the opportunity for the offender to commit criminal offenses?

Yes Unsure No

21 Does offender report boredom as a contributing factor to his or her criminal behavior?

Yes Unsure No

B. FAMILY

22 Are the offender 's family or household members able and willing to support a law abiding lifestyle?

Yes Unsure No

23 Is the offender's current household characterized by *(check all that apply)*:

PART THREE: OFFENDER QUESTIONNAIRE

NYSID : Name : DOB :

Please look at the following areas and let us know which of them you think will present the greatest problems for you. *Please check one response for each question in the column provided* .

	Please answer questions as either No, Yes or Don't Know	**No**	**Yes**	**Don't Know**
48	Do you feel you need assistance with finding or maintaining a steady job?			
49	Do you feel you need assistance with finding or maintaining a place to live?			
50	Will money be a problem for you over the next several months?			

	How difficult will it be for you to...	**Not Difficult**	**Somewhat Difficult**	**Very Difficult**
51	manage your money?			
52	keep a job once you have found one or if you currently have one?			
53	find or keep a steady place to live?			
54	have enough money to get by?			
55	find or keep people that you can trust?			
56	find or keep friends who will be a good influence on you?			
57	avoid risky situations?			
58	learn to control your temper?			
59	find things that interest you?			
60	learn better skills to get or keep a job?			
61	find a safe place to live where you won't be hassled or threatened?			
62	get along with people?			

COMPAS Probation Risk Assessment

Offender: **Joe Sample** DOB: **2/2/1950** Gender: **Male**

Screening Date: **9/13/2007** Screener: **Hellem, Dan** Ethnicity: **Native A**

Scale Set: **DMB-PSI** Case: **009943** Marital Status: **Single**

Observational measures of fairness

- C – output of the classifier
- Y – ground truth (rearrested/was not rearrested)
- D – demographic
  - For simplicity 0 or 1
- X – features
- Demographic parity
  - $P(C = 1|D = 0) = P(C = 1|D = 1)$
- False positive parity (“equal opportunity”)
  - $P(C = 1|D = 0, Y = 0) = P(C = 1|D = 1, Y = 0)$

Observational measures of fairness

- Demographic parity
  - $P(C = 1|D = 0) = P(C = 1|D = 1)$
  - Everyone is predicted to re-offend at the same rate, regardless of demographic
  - A type of "classification parity"
- False positive parity ("equal opportunity")
  - $P(C = 1|D = 0, Y = 0) = P(C = 1|D = 1, Y = 0)$
  - People who did not reoffend predicted to reoffend at the same rate, regardless of demographics
  - A type of "classification parity"
- Predictive Value Parity
  - $P(Y = 1 \middle| C = 1, D = 0) = P(Y = 1 \middle| C = 1, D = 1)$ and $P(Y = 1 \middle| C = 0, D = 0) = P(Y = 1 \middle| C = 0, D = 1)$
  - (Positive predictive value (PPV) parity + Negative predictive value (NPV) parity)
  - People predicted to reoffend actually reoffend at the same rate, regardless of demographics

Calibration

- $P(Y = 1|s(X) = s, D = 0) = P(Y = 1|s(X) = s, D = 1)$
  - The probability of re-arrest for people who got the same risk scores is the same
  - N.B.: if the score is 0/1, this reduces to

$$P(Y = 1|C = 1, D = 0) = P(Y = 1|C = 1, D = 1)$$
$$P(Y = 1|C = 0, D = 0) = P(Y = 1|C = 0, D = 1)$$

Anti-classification

- Protected characteristics are not considered
- $P(C = 1|X) = P(C = 1|X')$ if X and X' only differ by protected demographic

Utility functions

- Can assign a cost to each of true positive/true negative/false positive/false negative, and then compute the expected utility for a rule for making decisions
- Optimal rules are of the form $P(Y = 1|X) \geq thr$
- Sketch of proof
  - An exchange argument: always better to predict C = 1 for riskier individuals

Generally, can’t satisfy two measures simultaneously

Accuracy parity vs. PPV Parity

Low-risk: 10% chance of re-arrest

High-risk: 80% chance of re-arrest

Group A	Group B
Low-risk: 40, High-risk: 60	Low-risk: 50, High-risk: 50

- Assume the system perfectly identifies low vs. high-risk
- Group A: Predict 60 will be arrested. 12/60 won't be.
- Group B: Predict 50 will be arrested. 10/50 won't be.
- Group A: error rate is $\frac{12+4}{100} = 16\%.$ False positive rate is $\frac{12}{48}$
- Group B: error rate is $\frac{10+5}{100} = 15\%.$ False positive rate is $\frac{10}{50}$
- Equalizing the error rates (perhaps by randomly erring when deciding about group B, if the user is acting in bad faith) will mess up the predictive value parity

Accuracy disparity when False Positive Parity holds

- The mix of False Positives is different for different populations
  - Mix of high-risk individuals and low-risk individuals who did not end up re-offending

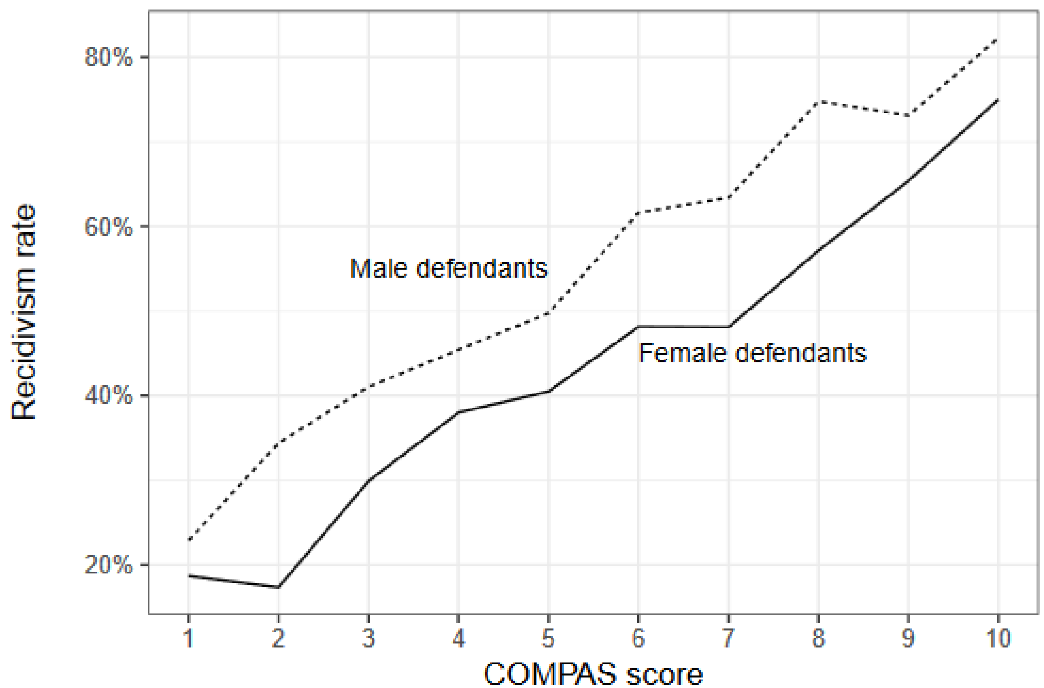
Discrimination before Fairness in ML

- Statistical discrimination
  - Charging male drivers more for insurance
  - Predicting younger people are more likely to reoffend
  - Predicting male defendants are more likely to reoffend
- “Taste-based discrimination”
  - Discrimination by the decision-maker that decrease an objective measure of the decision-maker’s utility (the decision-maker has a “taste for discrimination”) (Gary Becker 1957)

Discrimination before Fairness in ML

- Law usually focuses on the *intent* of the decision-maker to commit taste-based discrimination
  - If there is an observed disparity, that can trigger “strict scrutiny”: the decision-maker needs to justify their decision
- In the US, housing and employment, statistical disparities can be illegal unless they are justified
  - Griggs v Duke Power: the company could not require a high-school diploma for promotion since it was found there was no relation between job performance and having a diploma, because of racial disparity in promotion/having a diploma
  - “Unjustified disparate impact”: intent to discriminate *not* needed for the requirement to be illegal

Limitations of Anti-Classification



Sometimes need to consider demographics to get the best probability. COMPAS didn't, So there's no calibration wrt gender

Limitations of demographic parity/FP parity/etc

- Not necessarily compatible with each other
- Not compatible with calibration
  - (Again, calibration: scores mean the same thing regardless of demographic)

Limitations of calibration

Presence of discrimination despite calibration

- Redlining: the practice of not approving loan applications for predominantly black neighborhoods
- When predicting default rates just based on the zip code, calibration could be satisfied
  - If black neighborhoods are also generally poorer
  - There can be discriminatory intent in neglecting to use other features of the individuals

Label bias

- The y's (outcomes) in the training set might not be labelled correctly
  - In the COMPAS data, y = 1 if there was re-arrest
  - But we *want* to measure violent crime
    - Racial bias in the amount of policing in different neighborhoods
      - But could downweight e.g. drug arrests
    - Some arrests are not for violent crime
  - We don't have counterfactual information
    - We observe data that's conditioned on a judge's past decision
      - But can look at the two years after the release

Sample bias

- If the training set is not representative of new data, that is a problem

Simple and transparent models

- Advantages:
  - More likely to be adopted/trusted
  - Less sensitive to changes in data
- Disadvantages
  - Worse accuracy

Externalities + Equilibrium Effects

- Sometimes useful to think of decisions on a group level rather individual level
  - E.g. diversity is a measure of the group rather than individuals
- Predictive policing may create a feedback loop
  - More predicted crime => more policing => more detected crime => more predicted crime