Algorithmic Fairness Examples: Hiring

Technicalities

The raising hand functionality will be handled through a dedicated website

1) Go to: iraisemyhand.com

2) Enter channel name: **RML2023**

3) Enter your name, and join

Keep the website running in the background and simply press on the raise hand icon any time you have a question/reaction.

Algorithmic Fairness: Example of COMPAS

COMPAS men and COMPAS women – factual basis

- A risk assessment tool used to estimate recidivism risk
- Comprised of two different algorithms, COMPAS general and COMPAS Women, which is more tailored to the needs and unique characteristics of female
- COMPAS Women was created because females comprise a very small (statistically insignificant) portion of the criminal justice system
- COMPAS Women takes into account economic marginalization, trauma, victimization and abuse, mental health, dysfunctional intimate relationships, self-efficacy, and parental stress

COMPAS men and COMPAS women – contd.

- COMPAS developers claim that integrating gender sensitivity into the risk assessment tool will help agencies to achieve fairer results
- The notion of fairness applied by COMPAS is decoupling,
- Do you think that having two versions of COMPAS one for men and women for women is legal in this case?
- Do you think using an algorithmic risk assessment tool in the criminal justice system is desirable?
- What are the rights that might be jeopardized when using an algorithmic risk assessment tool in the criminal justice?
- Loomis v. Wisconsin, 881 N.W.2d 749 (Wis. 2016)

COMPAS general – facts

- COMPAS is a risk assessment tool developed by Northpoint now owned by Equivant.
- The tool covers different stages of the criminal justice process, and implemented in several jurisdictions in the U.S.
- COMPAS generates for each defendant a score on a scale of 1-10 that indicates how likely they are to reoffend if released during pretrial.
- The 1-10 score is divided into three categories, low risk, medium risk and high risk.

COMPAS general – continued

The facts that COMPAS considers are as follows:

1. Felony top charge
2. Pending case
3. Prior failure to appear
4. Prior arrest on bail
5. Prior jail sentence
6. Drug abuse history
7. Employment status
8. Length of residence
9. General questionnaire

Is there any problem with such factors?

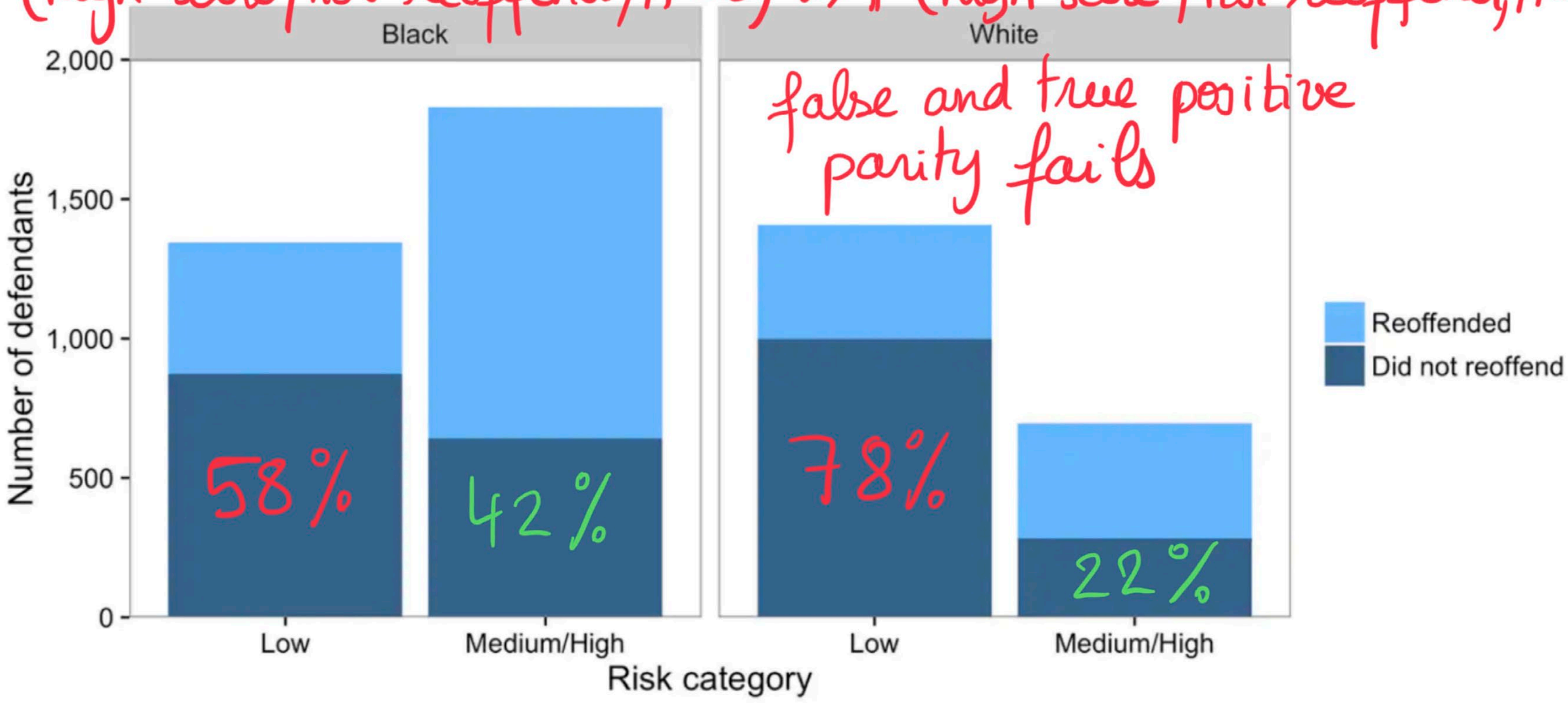
COMPAS general – continued

- The news outlet ProPublica obtained the risk score of more than 7,000 defendants from one of Florida's counties who have been released from jail in the pretrial phase.
- ProPublica examined whether after two years from the release, the defendants committed any additional crime, and this is in order to verify how good was the score given to them by COMPAS two years earlier.
- ProPublica concluded that COMPAS is biased against black.
- 42% of black defendants were wrongly classified by COMPAS as high risk, while among white defendants the mistake happened only in 22% of the cases.

ProPublica claimed that COMPAS is biased against black defendants

$\mathbb{P}(\text{high score} \mid \text{not reoffend}, A=b) \gg \mathbb{P}(\text{high score} \mid \text{not reoffend}, A=w)$

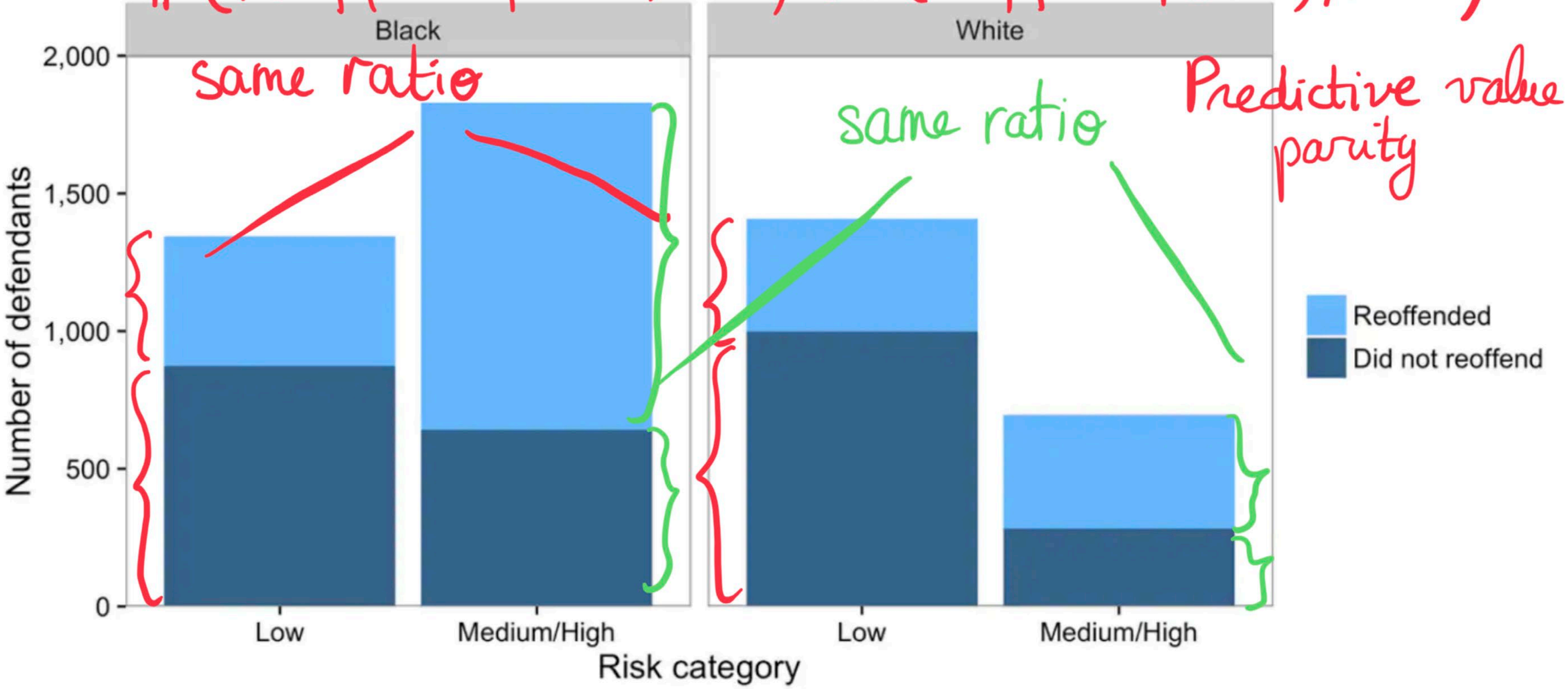
false and true positive parity fails



Distribution of defendants across risk categories by race. Black defendants reoffended at a higher rate than whites, and accordingly, a higher proportion of black defendants are deemed medium or high risk. As a result, blacks who do not reoffend are also more likely to be classified higher risk than whites who do not reoffend.

Nortpointe released its own report questionning the analysis

$$\mathbb{P}(\text{reoffend} \mid S=s, A=b) \approx \mathbb{P}(\text{reoffend} \mid S=s, A=w)$$



same ratio

same ratio

Predictive value parity

Distribution of defendants across risk categories by race. Black defendants reoffended at a higher rate than whites, and accordingly, a higher proportion of black defendants are deemed medium or high risk. As a result, blacks who do not reoffend are also more likely to be classified higher risk than whites who do not reoffend.

Maximizing accuracy

- The best case scenario yield 66% accuracy, and no matter where we place the threshold we always jail defendants who never get rearrested, and release defendants who do get rearrested.

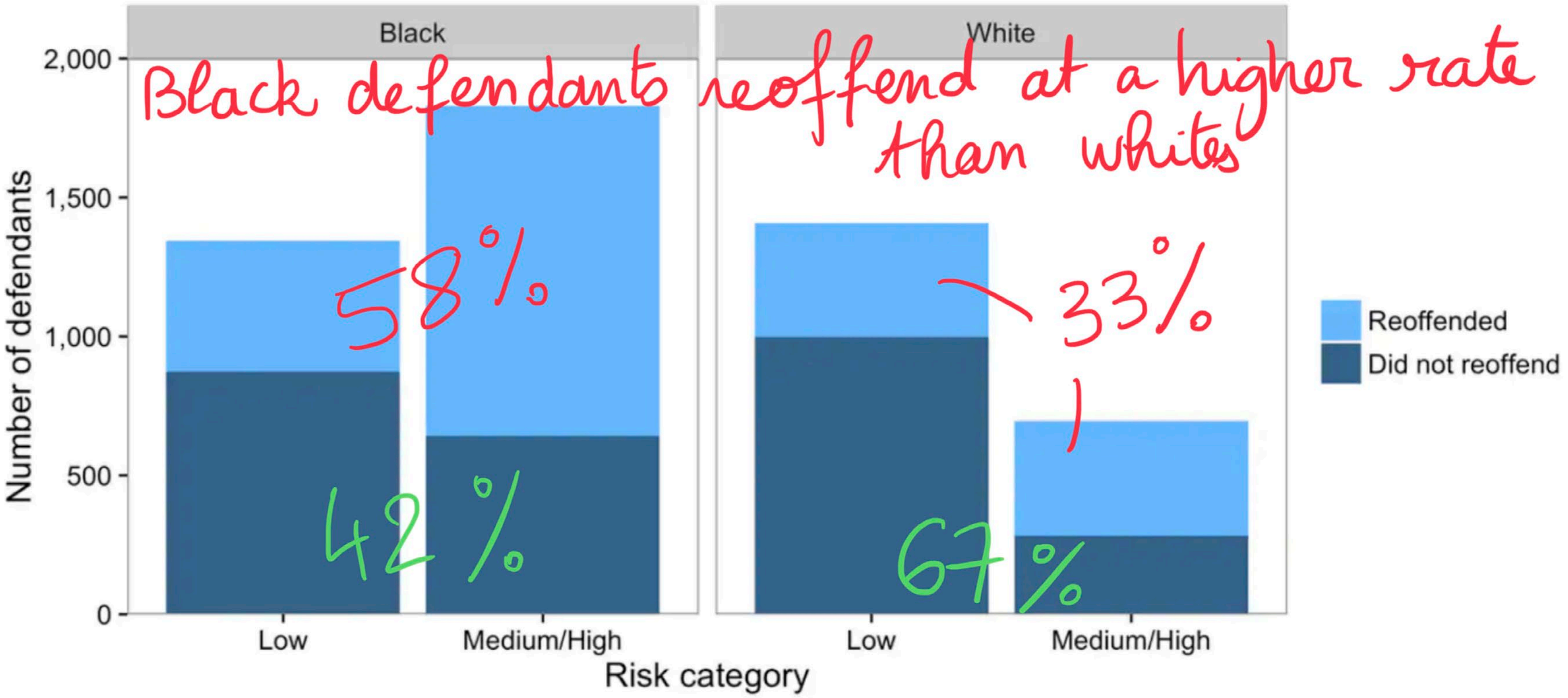
Check out interactive calculation:

https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

The Blackstone Ratio

“It is better that ten guilty persons escape than that one innocent suffer.”

What happens if we try to move the threshold so white and black defendants are needlessly jailed at roughly the same rate?



Distribution of defendants across risk categories by race. Black defendants reoffended at a higher rate than whites, and accordingly, a higher proportion of black defendants are deemed medium or high risk. As a result, blacks who do not reoffend are also more likely to be classified higher risk than whites who do not reoffend.

You cannot satisfy all notions of fairness simultaniously

- Tradeoffs are very hard to quantify
- What is the role of the judge?
- Actuarial tools have been used for decades in the criminal justice system

The COMPAS example- questions

- Which notion of fairness you think is fairer?
- Do you have other idea of how to adjust the values in order to achieve better results?
- In cases where the base rate for certain group is very different, what can be done?

Resources

on COMPAS:

- https://www.tml.cs.uni-tuebingen.de/teaching/2020_statistical_learning/downloads_free/luxburg_statistical_learning_slides.pdf ;
- https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/
- https://aif360.mybluemix.net/data

on loan application

- https://research.google.com/bigpicture/attacking-discrimination-in-ml/
- https://aif360.mybluemix.net/data

on adult dataset:

- https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/module-four-case-studies/case-study-mitigating-gender-bias/MITRES_EC001S19_video7.pdf
- https://aif360.mybluemix.net/data

on hiring:

- https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/12/Princeton-AI-Ethics-Case-Study-5.pdf

Credit scoring

Check out: https://scholarship.law.columbia.edu/faculty_scholarship/3312

Legal provisions prohibiting discrimination in lending

- The two laws that form the core of credit pricing discrimination are the Fair Housing Act (FHA) of 1968 and the Equal Credit Opportunity Act (ECOA) of 1974.
- The FHA protects renters and buyers from discrimination by sellers or landlords and covers a range of housing related conduct including the setting of credit terms.
- The FHA prohibits discrimination in the terms of credit based on race, color, religion, sex, disability, familial status, and national origins.
- In 1974, Congress passed the Equal Credit Opportunity Act, (ECOA), banning discrimination in all types of credit transactions.
- ECOA complements FHA by expanding discrimination provisions to other credit contexts beyond housing related credit.

Disparate treatment and disparate impact

- Disparate treatment involves the direct conditioning of the decision on a protected characteristic and therefore focuses on the causal connection between a protected characteristic and a credit decision.
- Disparate treatment can be triggered by directly considering a protected characteristic, such as race, in a specific credit decision or when a protected characteristic is used in setting general lending policy, such as in the case of "red-lining.
- Disparate impact covers cases in which a facially neutral rule has an impermissible disparate effect.
- A disparate impact case typically follows the burden-shifting framework that was developed primarily in the employment discrimination context.
- Establishing causality is a requirement in both cases

Personalization of credit decisions

- Credit contracts are often personalized, meaning that lenders will determine the specific terms of the contract based on the characteristics of the borrower and the specific loan.
- Pricing inputs could include borrower characteristics, such as the borrower's income or years of education, as well as the characteristics of the loan application, such as the loan amount.
- In traditional mortgage lending, a borrower's creditworthiness is assessed based on past credit behavior, or based on a borrower's FICO score.
- Lenders also use the specific characteristics of the loan, and the securitized property, to determine the terms of the loan.
- The exact terms of the loan vary greatly across borrowers, and so there is a degree of personalization of the prices paid by borrowers.

The impact of machine learning on lending

- Whereas traditional lending relied on relatively few defined characteristics, lenders are increasingly using new data and additional borrower characteristics to assess creditworthiness.
- Among them are data on payment and consumer behavior, social media behavior, and digital footprints, as well as information on education, such as the school attended and degree attained
- Using non traditional data could lead to more accuracy and expand lending to populations that have been traditionally excluded
- However, using non traditional data could also perpetuate discrimination

1. Achieving fairness by excluding protected attributes

- Prior to running the algorithm on the training set, a lender would exclude any protected characteristics from the inputs of the algorithm, even if they were available to the lender.
- An algorithm could use race for example because it correlates with other characteristics that the algorithm cannot observe directly, such as wealth or access to credit, which in turn affect default risk
- Algorithms will have a better ability to learn when using protected attributes
- When a characteristic should be interpreted differently for various racial groups, excluding "race" could increase disparities. This is because by excluding the race variable, we are imposing a similar interpretation of a characteristic for both white and non-white applicants a protected characteristic

2. Achieving fairness by prohibiting proxies

- Exclusion of protected characteristics is meaningless if the algorithm can use proxies
- Prohibiting the use of zip-codes for example
- The problem is that we do not have a good understanding of the "model" of default, nor of the variables that are causal of default
- A further difficulty is that many variables can be an indicator of a protected characteristic and also independently contain information relevant to the outcome of interest
- Intuition about which variable can serve as proxy could be misleading

3. Achieving fairness by restricting inputs only to pre-approved ones

- Instead of allowing use of any variable not barred, as in the traditional antidiscrimination model, actors can only use pre-approved variables.
- Focusing on variables that do not penalize protected groups
- The main challenge for the third approach is to define which variables are permissible. That definition depends on what the restriction is meant to achieve
- If we limit the variables to only inputs that predicts default, there will be no reason to exclude anything
- If we limit the algorithm to characteristics that are used in traditional credit pricing, such as FICO scores or a borrower's income. But this would undermine the benefits of big data and machine learning in extending access to credit

Correlation does not imply causation

- Because of the absence of identifiable causal relationships, input-based approaches are unsuitable for discrimination law in an algorithmic setting.
- This is true for both disparate treatment and disparate impact.
- For disparate treatment, we have no reliable way to detect proxies for protected characteristics.
- For disparate impact, we need new tools to evaluate the effects of algorithmic pricing.

Outcome based test

Three stages test:

1. The lender determines what inputs and which algorithm to use to predict default and price accordingly.
2. The regulator then takes that prediction or pricing rule and applies it to a dataset of people to see the distribution of prices the rule produces.
3. The regulator evaluates the outcome to determine whether the disparities created by the pricing rule amount to discriminatory conduct

Who are the similarly situated?

- In the algorithmic context, we can consider a set of characteristics which determines who is similarly situated. Any differences that are explained by this set of characteristics are not deemed to be impermissible discrimination.
- This set can intuitively be understood as adding control variables into a regression in that they explain differences between people.
- The size and scope of the similarly situated set are likely to have a significant effect on whether there is a finding of impermissible disparity.
- creating a test that relies on similarly situated characteristics makes the tradeoff between accuracy and other policy goals explicit, rather than rendering it opaque as input-based approaches do when they restrict inputs to those that seem intuitively relevant to default

Comparing to a pre-determined baseline

- Rather than considering the absolute levels of disparities created by a pricing rule, the focus is on how these disparities compare to traditional credit pricing rules.
- A regulator could compare the prices produced under the use of traditional lending variables with new data available to a lender, such as consumer and payment behavior
- Incremental approach

Thank you