Fairness, bias, and transparency in Machine Learning

Module 2: Intro to bias and fairness

This module

- 2.1: Introduction
- 2.2: Bias definitions
- 2.3: Fairness definitions
- 2.4: Impossibility results
- Case study: a detailed look at COMPAS

(approx. 1 week)

Intro to bias and fairness

2.1: Introduction

This section

- Fairness & Algorithmic Decision Making: https://afraenkel.github.io/fairness-book/
- A Survey on Bias and Fairness in Machine Learning: https://dl.acm.org/doi/pdf/10.1145/3457607
- Fairness in Machine Learning: https://arxiv.org/pdf/2010.04053
- A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle: https://arxiv.org/pdf/1901.10002
- (lots of others in slides)

What are bias and fairness?

- Based on https://afraenkel.github.io/fairness-book/content/02-frameworks.html
- and https://afraenkel.github.io/fairness-book/content/03-harms.html

- First let’s look at some of these concepts informally (or at least “non-mathematically”), before thinking about how we can formalize them

Motivating example: COMPAS Recidivism Algorithm

Example: COMPAS is a “decision support tool” used to measure recidivism risk

Although ostensibly an “advisory” tool, the tool’s risk assessment scores have been cited in rulings (i.e., judges base their decisions partly on the algorithm’s prediction)

COMPAS doesn’t use any “sensitive attributes” to predict outcomes (e.g. an individual’s race); however several of the features are *correlated* with race

https://en.wikipedia.org/wiki/COMPAS_(software)
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Motivating example: COMPAS Recidivism Algorithm

Roughly speaking a number of features are used to predict a number of outcomes:

- Pretrial release risk (failure to appear, or commit felonies while on release)
  - Features based on: current charges, pending charges, prior arrest history, previous pretrial failure, residential stability, employment status, community ties, and substance abuse
- General recidivism (new offenses upon release)
  - Features based on: individual's criminal history and associates, drug involvement, and indications of juvenile delinquency
- Violent recidivism (violent offenses following release)
  - Features based on: history of violence, history of non-compliance, vocational/educational problems, the person's age-at-intake and the person's age-at-first-arrest

https://en.wikipedia.org/wiki/COMPAS_(software)
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Motivating example: COMPAS Recidivism Algorithm

Some findings about COMPAS (see link):

- Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
- White defendants were often predicted to be less risky than they were. Analysis found that white defendants who reoffended within the next two years were mistakenly labeled low risk almost twice as often as black reoffenders (48 percent vs. 28 percent).
- The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
- Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.
- The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

https://en.wikipedia.org/wiki/COMPAS_(software)
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

This module

In this module:

- We’ll study how such biases and unfair outcomes can arise as a result of problems in datasets, algorithm design choices
- We’ll explore the relationship between biases in algorithms and unfair outcomes (roughly speaking, the effect that biases have on specific groups)
- We'll present and compare **many, many** potential definitions of bias and fairness
- Discuss fundamental limitations and incompatibilities between bias definitions

Afterwards, we’ll revisit the COMPAS algorithm in detail as a case study

Mostly, this module will be completely focused on *measurement* of bias and fairness, whereas Module 3 will focus on *intervention*

Intro to bias and fairness

2.2: Bias definitions

This section

- What are some of the common sources of bias in machine learning systems?
- Categorization of common sources of bias:
  - **Data-to-algorithm:** biases present in the data itself, causing algorithms to be biased
  - **Algorithm-to-user:** biases that result from algorithm design choices
  - **User-to-data:** biases that arise from models trained on user-generated data
- We will **not** spend much time discussing intervention strategies until the next module

Ways to categorize bias in ML

There are many possible ways to categorize the types of bias in ML systems, and too many possible sources of bias to cover in a few lectures; we'll mostly focus on those covered in the following papers:

A Survey on Bias and Fairness in Machine Learning. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. 2021 (https://dl.acm.org/doi/pdf/10.1145/3457607)

A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Harini Suresh, John Guttag. 2021 (https://arxiv.org/pdf/1901.10002)

Ways to categorize bias in ML

Three possible characterizations:

- **Data-to-algorithm** biases in a **dataset** cause algorithms trained on that data to have biased outcomes
- **Algorithm-to-user** biases are a result of algorithm **design choices**; these design choices then impact user behavior (possibly leading to further biases in user behavior)
- **User-to-data** biases arise from models trained on **user-generated data**; biases from users will appear in the data they generate

(exercise: how would we categorize the examples we saw in the last module?)

Data-to-algorithm bias – measurement bias

Measurement bias arises based on how we measure particular model feature. Model features are generally *proxies* for things we care about predicting.

1. Proxies can be oversimplifications. E.g. using “GPA” as an indicator of student success in a program
2. Method of measurement varies across groups

https://en.wikipedia.org/wiki/COMPAS_(software)
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Data-to-algorithm bias – measurement bias

E.g. in the case of COMPAS, the training data measures recidivism (i.e., whether people “reoffend”)

But, to be observed reoffending, the police must come into contact with, arrest, and charge a defendant; *communities that experience more police contact are more likely to be prosecuted*, *even if* actual *crime rates are the same*

(we’ll come back to this in a case study at the end of the module)

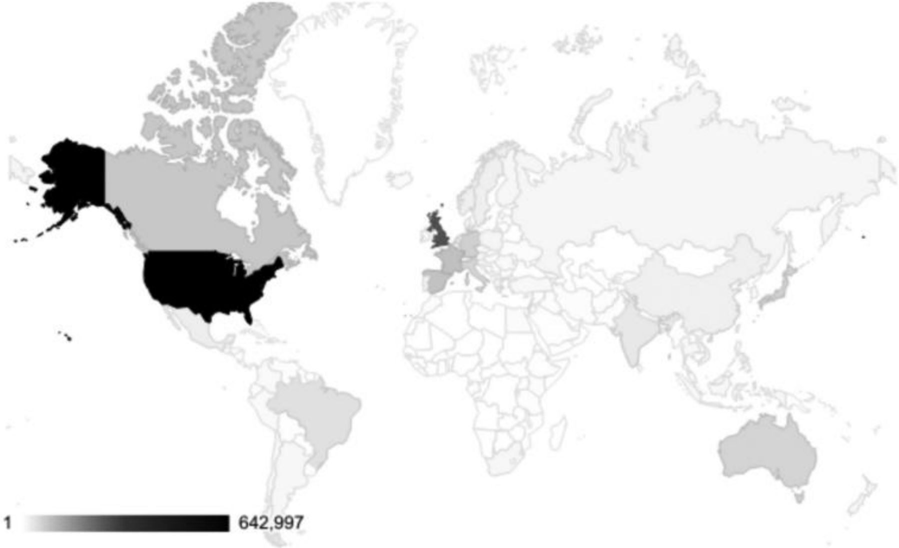
https://en.wikipedia.org/wiki/COMPAS_(software)
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Data-to-algorithm bias – sampling bias

Representation bias (or **“sampling bias”**) arises from how we sample data from a population. Non-representative samples may lack the diversity of the population, leading to poor performance for certain subgroups

E.g. image classifiers trained using data from Western cultures may have poor performance in non-Western contexts

Data-to-algorithm bias – sampling bias



Geographic distribution of countries in the Open Images dataset; from “No classification without representation: Assessing geodiversity issues in open data sets for the developing world”, Shankar *et al.*

Data-to-algorithm bias – sampling bias

Exercise: In the previous module, we looked at scenarios where image classifiers worked much better for male than for female users:

- Could these outcomes have been the result of sampling bias?
- Could these outcomes have been the result of something *other* than sampling bias?

Data-to-algorithm bias – omitted variable bias

Omitted variable bias arises when leaving a variable out of a model causes us to misattribute the influence of the missing variable to other variables in the model

Example:

- Students who have a long commute have higher GPA
- Students who have a long commute tend to be older; when we add an “age” variable, commute time is negatively associated with GPA

Data-to-algorithm bias – aggregation bias

Aggregation bias arises when we draw conclusions about populations that may not be true for individuals or subgroups

Data-to-algorithm bias – aggregation bias

Code example: Covid delta-variant outcomes by age group

workbook2.ipynb (see course webpage)

from: https://www.openintro.org/data/index.php?data=simpsons_paradox_covid

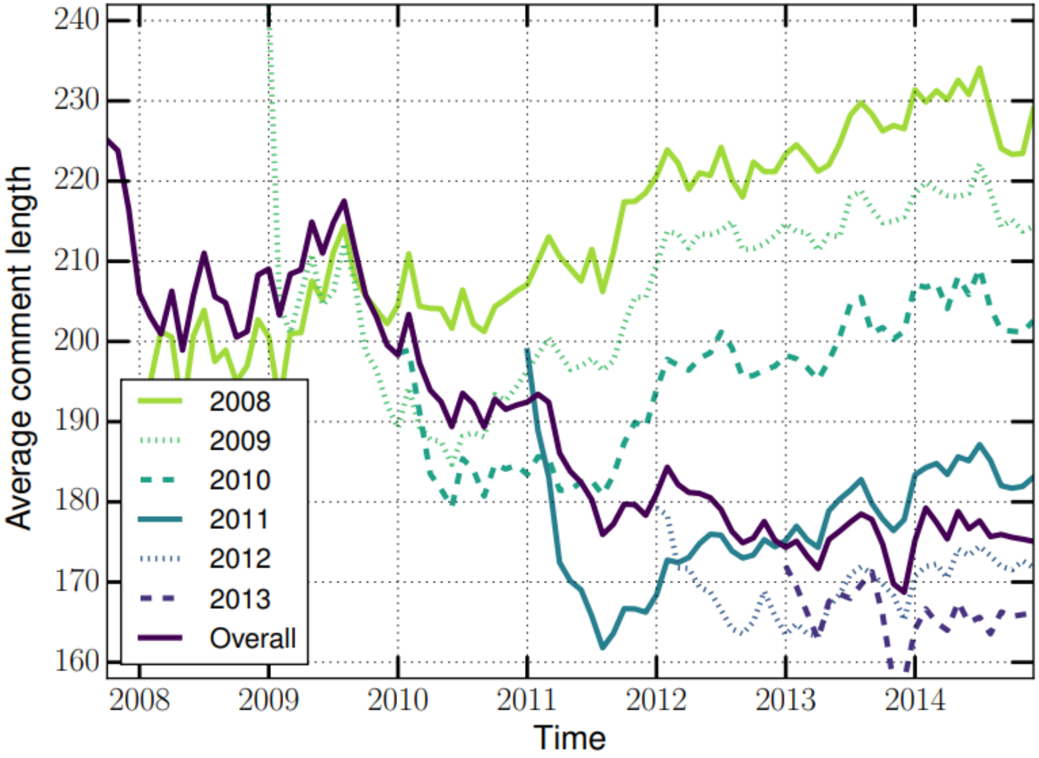
Data-to-algorithm bias – longitudinal data fallacy

The **longitudinal data fallacy** occurs when temporal data is aggregated in a way that mixes diverse cohorts

(**note:** really another form of aggregation bias, just with temporal data)

Data-to-algorithm bias – longitudinal data fallacy

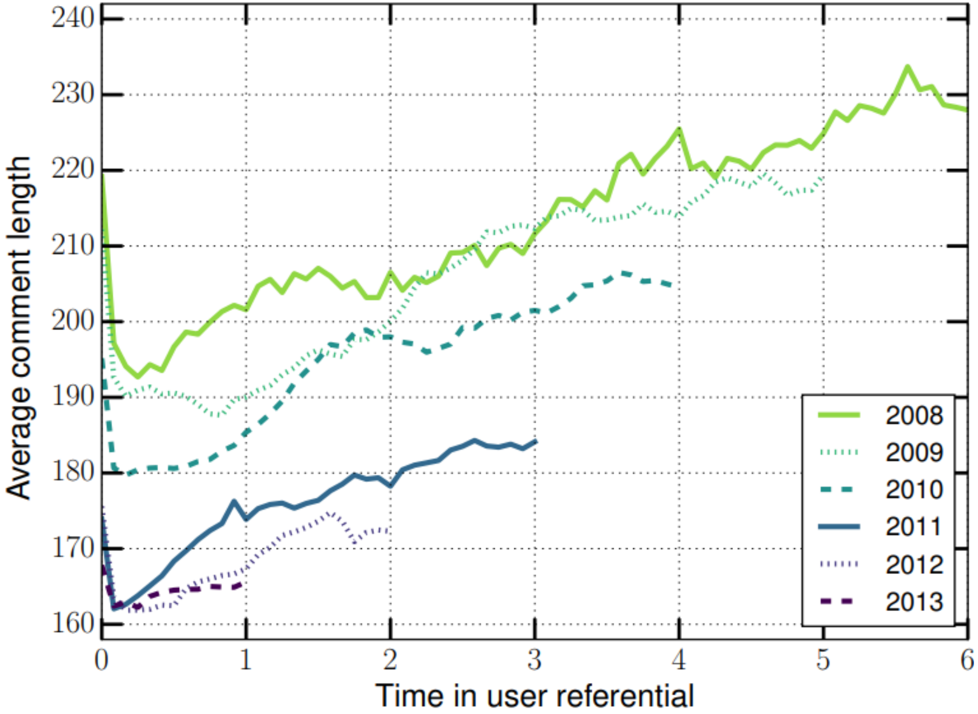
E.g. bulk reddit data combines users who joined reddit at different times; the data shows that comment length decreases over time



from: https://dl.acm.org/doi/pdf/10.1145/2872427.2883083

Data-to-algorithm bias – longitudinal data fallacy

But if we disaggregate the data by different cohorts, *comment length within each cohort increases* over time:



from: https://dl.acm.org/doi/pdf/10.1145/2872427.2883083

Algorithm-to-user bias

Algorithm-to-user bias occurs when algorithms (which might range from modeling decisions to user interface / presentation considerations) introduce new biases into the system that were not present in the data

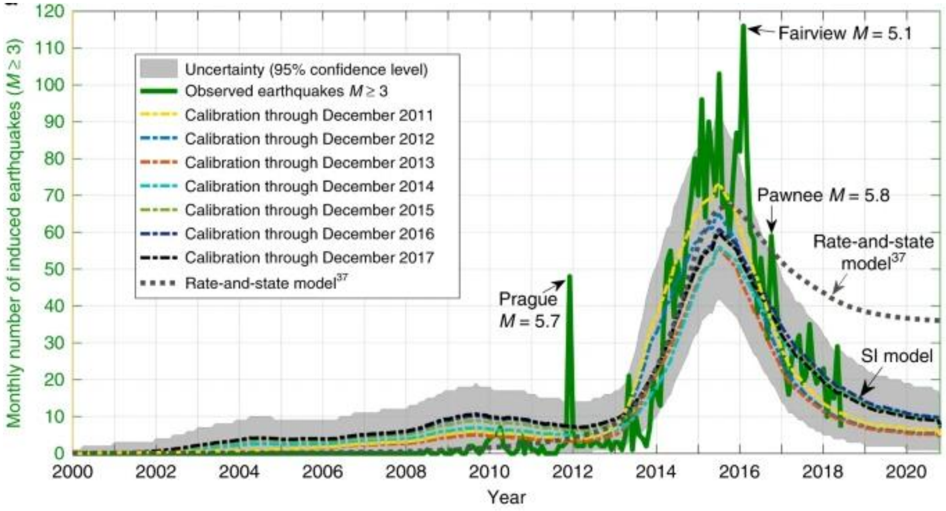
Alternately, the bias in the algorithm causes "downstream" biases in user behavior

Algorithm-to-user bias – algorithmic bias

Algorithmic bias refers to bias that is not present in the input data and is added purely by the algorithm (not necessarily any "user" in this case)

Algorithm-to-user bias – algorithmic bias

Example: we saw an example of this in the previous module: fitting a model using a mean-squared error can cause it to systematically overpredict most instances in a dataset containing large (positive) outliers, while underpredicting the outliers themselves



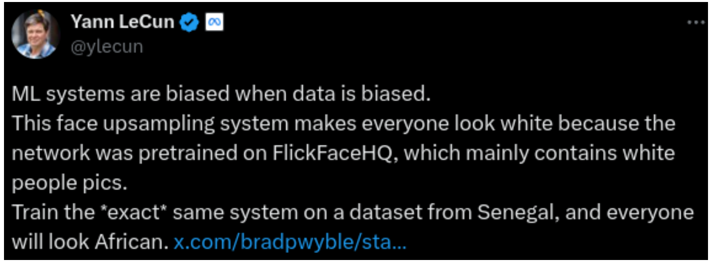
(seismicity prediction)

Algorithm-to-user bias – algorithmic bias

Note: in a different context, the “outliers” could be members of underrepresented groups!

Something as simple as a choice of objective (e.g. penalizing the *square* of the error rather than the *absolute value* of the error) could contribute to systematically bad performance for certain (underrepresented) types of user, while disproportionately favoring the majority group

Examples like this are often cited to point out that it’s *not* just the dataset that causes bias!



Yann LeCun
@ylecun

ML systems are biased when data is biased.
This face upsampling system makes everyone look white because the network was pretrained on FlickFaceHQ, which mainly contains white people pics.
Train the *exact* same system on a dataset from Senegal, and everyone will look African. x.com/bradpwyble/sta...

(tweet from June 2020)

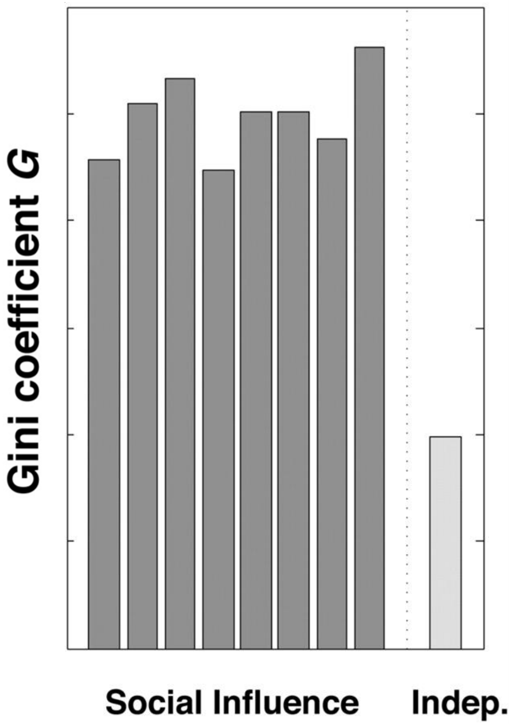
Algorithm-to-user bias – user interaction bias

Information is presented to the user in a certain way (e.g. the ordering in which items are ranked, or how past ratings are shown to a user) which might e.g. bias that user's opinion

Algorithm-to-user bias – user interaction bias

Example: users who can see whether a song is already popular will perceive it as better (and thus amplify its popularity) (basically, the "social influence" condition exposes users to existing download counts)

(paper also considers different ways of presenting social information via the UI)



from: https://www.science.org/doi/full/10.1126/science.1121066

Algorithm-to-user bias

Similarly, see:

Presentation bias: users interact with the content that is shown to them, while other content may not be seen;

Ranking bias: top-ranked items are perceived as being more relevant (and get more engagement), possibly feeding back into the algorithm that is used to rank

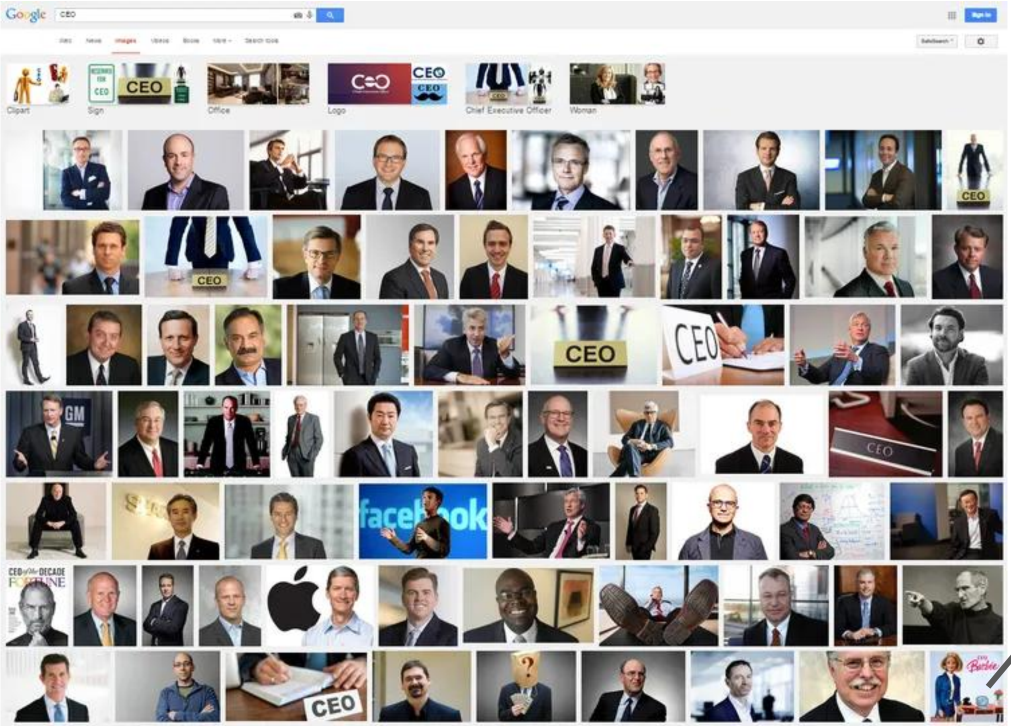
Popularity bias: items that are more popular get exposed more, and become more popular as a result (can also be subject to manipulation!)

User-to-data bias

Just as algorithms can expose data to users in a biased way (and that biased data can be fed back into the algorithm), users can inject their own biases into datasets, which are then propagated into the algorithm

This type of bias is called **user-to-data** bias

User-to-data bias – historical bias



Google image search fails to return female CEOs (from 2015); ~5% of Fortune 500 CEOs were female at the time

from: https://www.theverge.com/tldr/2015/4/9/8378745/i-see-white-people

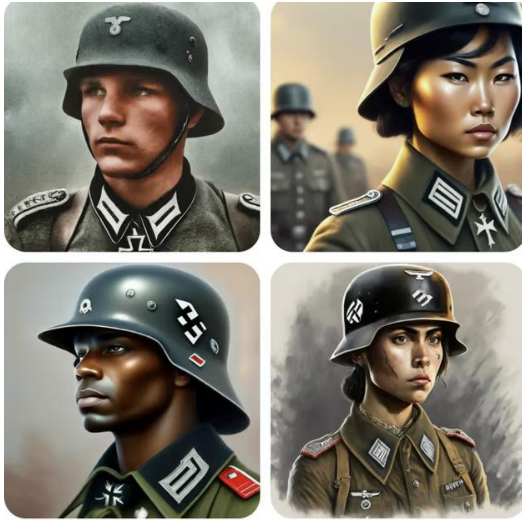
User-to-data bias – historical bias

Again, we should be careful not to fully attribute the above biases to users or to "data": certainly the data contains historical bias (women are not historically CEOs), but a different choice of algorithm might have surfaced results that better represented the tails of the distribution

User-to-data bias – historical bias

Related: recent high-profile diversity issues in image synthesis models (and attempts to mitigate them)

Food for thought: what *should* algorithms do in such cases?



User-to-data bias – more examples

Population bias arises when the characteristics of users on a platform are different from those of the target population (e.g. women are more likely to use Pinterest/Facebook/Instagram, men being more active on Reddit/Twitter)

Self-selection bias occurs when subjects "select themselves", e.g. by choosing to participate in an online poll

Social Bias happens when others' actions affect our judgment, e.g. we might be less likely to enter a negative review if we see that others' reviews are positive

see https://dl.acm.org/doi/pdf/10.1145/3457607

Plenty of other biases!

Plenty of other examples in https://dl.acm.org/doi/pdf/10.1145/3457607 (I won't go into all of them)!

Hopefully the point has been made that there are countless sources via which algorithms can become biased!

Study points & take-homes

- Understand the different sources of bias from users, data, and algorithms
- We've looked at these concepts at a high-level (i.e., non-technically) so far; before studying the next section (and next module), think critically about:
    - How you might measure some of these concepts formally
    - (next module) What you might do to fix some of these bias issues

Intro to bias and fairness

2.2: Fairness definitions

This section

- How is “fairness” different from and related to “bias”?
- *Many* different notions of fairness and discussion of their relative merits (see list at end of section)

Fairness vs bias

When studying bias so far, we’ve looked how issues of datasets and models can cause us to draw incorrect conclusions

We’ve already shown how some of these wrong conclusions can lead to unfair outcomes, so “bias” and “fairness” seem not all that different!

Formally speaking:

- “Bias” is a **mathematical** concept, describing the *tendency of a method to systematically mis-predict an outcome*
- “Fairness” is a **social** concept, concerned with differences in outcomes across different social groups (age, race, gender, etc.)

(so **biased** predictors lead to **unfair** outcomes)

Fairness vs bias

As such, in this section, we’ll introduce concepts like **protected characteristics** and **groups** (based around e.g. age, race, gender) with respect to which we’ll measure algorithmic outcomes

Desiderata of fair algorithms

We'll discuss some potential *desiderata* (desirable attributes) of a "fair" algorithm

These can be used in two ways:

1. As a *measurement* of whether an algorithm is fair (or how unfair it is)
2. As a *constraint*, i.e., we might want to find the most accurate possible algorithm that enforces a fairness requirement

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Some definitions (mostly recap)

- **Positive class:** *labeled* as positive
- **Negative class:** *labeled* as negative
- **Positive outcome:** *predicted* as positive by the model
- **Negative outcome:** *predicted* as negative by the model
- **Protected group:** a binary attribute (e.g. "is female") against which we want to measure a fairness outcome
- **Unprotected group:** the complement of the above
- **Classifier:** the function that decides the outcome

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Some definitions (mostly recap)

- Prevalence:

How common is the positive class?

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Some definitions – mathematical notation

- Positive class: $y_i = 1$
- Negative class: $y_i = 0$
- Positive outcome: $\hat{y}_i = 1$
- Negative outcome: $\hat{y}_i = 0$
- Protected group: $z_i = 1$
- Unprotected group: $z_i = 0$
- Classifier: $f(x_i)$ or $f(x_i, z_i)$

<u>lots</u> of different notation for these across papers!

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Some definitions – mathematical notation

- **Prevalence:** $\frac{1}{N} \sum_i \delta(y_i = 1)$

Sometimes care about prevalence per-group:

$$\frac{\sum_i \delta(y_i = 1 \mid z_i = 1)}{\sum_i \delta(z_i = 1)} \quad \text{and} \quad \frac{\sum_i \delta(y_i = 1 \mid z_i = 0)}{\sum_i \delta(z_i = 0)}$$

(prevalence for males & females)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Impact and treatment disparity

We’ll some times use terms like “disparate treatment” or “disparate impact”.

Algorithms exhibit **treatment disparity** if members of different subgroups are *explicitly* treated differently, e.g. the sensitive attribute might directly be used to make a decision

Algorithms exhibit **impact disparity** when *outcomes* differ across subgroups

definition from https://cseweb.ucsd.edu/~jmcauley/pdfs/nips18.pdf

Impact and treatment disparity

We'll explore the relationship between these notions more deeply in a case-study in the next module

For a preview, *affirmative action* policies aim to deliberately use **disparate treatment** in order to reduce **disparate impact,** which (to say the least) is legally disputed

definition from https://cseweb.ucsd.edu/~jmcauley/pdfs/nips18.pdf

Fairness through unawareness

An algorithm is fair as long as any protected attributes are not explicitly used in the decision making process

e.g. when predicting admissions outcomes
just don't use gender

Q: could you use it during training but not inference?

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Fairness through unawareness

An algorithm is fair as long as any protected attributes are not explicitly used in the decision making process

Is this sufficient to guarantee fairness?

not really defined yet...

No! Other features could be correlated

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Fairness through unawareness

An algorithm is fair as long as any protected attributes are not explicitly used in the decision making process

Is it *desirable*?

- Legally defensible (probably required)
- Not effective at ensuring fairness
- (Arguable) Maybe we could make things fairer by using the sensitive attribute (see: affirmative action)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Accuracy Parity

The accuracy of a classifier should be equal across two groups

$$P(\hat{Y} = y \mid z_i = 1) = P(\hat{y} = y \mid z_i = 0)$$

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Accuracy Parity

The accuracy of a classifier should be equal across two groups

Why do we want this?

- Algorithms should "work" about as well for either group
- ?

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Accuracy Parity

The accuracy of a classifier should be equal across two groups

Why do we *not* want this?

- Different error types could be more / less "costly"
- E.g. black defendants misclassified as "high-risk"
  white defendants misclassified as "low-risk"
  (but overall accuracy equal for both groups)

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Accuracy Parity

The accuracy of a classifier should be equal across two groups

Why do we *not* want this?

In the case of COMPAS: accuracy parity was approximately satisfied; the algorithm makes up for detaining releasable black defendants by wrongly releasing white defendants

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Demographic parity (statistical parity)

The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group

$$P(\hat{Y}=1 \mid z_i=1) = P(\hat{Y}=1 \mid z_i=0)$$

sometimes used as a ratio:

$$\frac{P(\hat{Y}=1 \mid z_i=0)}{P(\hat{Y}=1 \mid z_i=1)}$$

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Demographic parity (statistical parity)

The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group

Why do we want this?

- Different races should be assessed as "high-risk" at same rate (?)
- Women and men should be admitted with same probability (?)

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Demographic parity (statistical parity)

The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group

Why do we *not* want this?

- What if pevelance is different between groups?

e.g. unqualified women don't apply, but unqualified men do?

high prevalence

low pevalence

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Demographic parity (statistical parity)

The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group

Why do we *not* want this?

In this case, demographic parity would mean admitting unqualified males

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Demographic parity – legal definition

Some **food for thought:**

- The legal definition is one-sided
- In cases where demographic parity seems like a good idea, what are you assuming about the underlying distribution (**hint:** prevalence)

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Demographic parity – legal definition

The State of California Guidelines on Employee Selection Procedures (1972) defines a (fairly arbitrary) threshold of 80% (knows as “the 80% test”) to define **Disparate Impact,** that is:

$$\frac{P(\hat{y}=1 \mid z_i=0)}{P(\hat{y}=1 \mid z_i=1)} \overset{?}{<} 0.8$$

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Disparate impact: p-% rule

Impact disparity is sometimes measured using a quantity known as the **p-% rule** which measures *ratio between the probability of being assigned to the positive class for the advantaged versus disadvantaged group*

$$\frac{P(\hat{y}=1 \mid z_i=1)}{P(\hat{y}=1 \mid z_i=0)} > \frac{p}{100}$$

So, following the legal definition from the previous slide, we would say a classifier exhibits *disparate impact* if it fails to satisfy a “80-%” rule

Equalized odds

*The probability of a person in the **positive class** being (correctly) assigned a **positive outcome** and the probability of a person in the **negative class** being (incorrectly) assigned a **positive outcome** should be the same for both the protected and unprotected group members*

$$P(\hat{Y}=1 \mid z_i=0, \underbrace{Y=y}) = P(\hat{Y}=1 \mid z_i=1, Y=y)$$

$$y \in \{0,1\}$$

same as previous defn.
but adds actual qualification
as condition

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equalized odds

The probability of a person in the ***positive class*** *being (correctly) assigned a* ***positive outcome*** *and the probability of a person in the* ***negative class*** *being (incorrectly) assigned a* ***positive outcome*** *should be the same for both the protected and unprotected group members*

How is this different from demographic parity?

- Adds label (eg. "qualified") as condition
- Splits demographic parity into two goals
  -> qualified candidates should have parity
  -> unqualified candidates should have parity

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equalized odds

The probability of a person in the ***positive class*** *being (correctly) assigned a* ***positive outcome*** *and the probability of a person in the* ***negative class*** *being (incorrectly) assigned a* ***positive outcome*** *should be the same for both the protected and unprotected group members*

Why do we want this?

"Fixes" previous issue with demographic parity (different prevalence)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equalized odds

*The probability of a person in the **positive class** being (correctly) assigned a **positive outcome** and the probability of a person in the **negative class** being (incorrectly) assigned a **positive outcome** should be the same for both the protected and unprotected group members*

- Any specific objections?

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equalized odds

Exercise: Express equalized odds in terms of (per-group) True Positive (False Negative, etc.) rates

$$\underbrace{P(\hat{y}=1 \mid z_i=0, y_i=0)}_{\frac{\sum_i \delta(\hat{y}_i=1,\, z_i=0,\, y_i=0)}{\sum_i \delta(z_i=0,\, y_i=0)} = FPR_0}$$

$$\boxed{\begin{array}{c}\text{Equalized odds:} \\ FPR_0 = FPR_1 \\ TPR_0 = TPR_1\end{array}}$$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Code example

Code example: Equalized odds & demographic parity

workbook2.iypnb

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Predictive Value Parity

Predictive value parity states that the chance of a positive label should be equalized across groups given a positive prediction (for both classes):

PPV : $P(y=1 \mid \hat{y}_i=1, z_i=0) = P(y=1 \mid \hat{y}_i=1, z_i=1)$

↳ positive

NPV : $P(y=0 \mid \hat{y}_i=0, z_i=0) = P(y=0 \mid \hat{y}_i=0, z_i=1)$

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Predictive Value Parity

Predictive value parity states that the chance of a positive label should be equalized given a positive prediction (for both classes).

How is this different from equalized odds?

$$P(\underset{\uparrow \text{ prediction}}{\hat{Y}} \mid y, z) \quad \text{vs} \quad P(\underset{\uparrow \text{ label}}{Y} \mid \hat{y}, z)$$

Eq. odds PPV

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Predictive Value Parity

Predictive value parity states that the chance of a positive label should be equalized given a positive prediction (for both classes).

Why do we want this?

$P(\hat{y}|y,z)$: qualified people should be predicted as qualified

$P(y|\hat{y},z)$: people predicted as qualified should be qualified

see https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Equal opportunity

The probability of a person in a positive class being assigned a positive outcome should be equal for both the protected and unprotected groups

$$\underbrace{P(\hat{y}=1 \mid z_i=0, y_i=1)}_{TPR_0} = \underbrace{P(\hat{y}=1 \mid z_i=1, y_i=1)}_{TPR_1}$$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equal opportunity

The probability of a person in a positive class being assigned a positive outcome should be equal for both the protected and unprotected groups

Compare to equalized odds:

equalized odds

$y \in \{0, 1\}$

equal opp

$y = 1$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equal opportunity

The probability of a person in a positive class being assigned a positive outcome should be equal for both the protected and unprotected groups

Write in terms of rates:

$$\boxed{TPR_0 = TPR_1}$$

vs.

$$\boxed{\begin{array}{c} TPR_0 = TPR_1 \\ \underline{\text{and}} \\ FPR_0 = FPR_1 \end{array}}$$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equal opportunity

The probability of a person in a positive class being assigned a positive outcome should be equal for both the protected and unprotected groups

Why do we want this?

- Given that somebody is (e.g.) qualified, then p(qualified) should be equal for both groups
- Sometimes only care about positive outcomes

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Equal opportunity

The probability of a person in a positive class being assigned a positive outcome should be equal for both the protected and unprotected groups

Why do we *not* want this?

- For (e.g.) recidivism prediction, FP and FN could both be consequential

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Calibration

Probabilities output by a classifier should have semantic meaning, e.g. if 100 people in group ***g*** have ***f(x) = 0.6,*** then we would expect 60 of them to belong to the positive class.

A classifier ***f*** is perfectly calibrated if:

$$P(y=1 \mid f(x)=p) = p \quad \forall p$$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Calibration

Probabilities output by a classifier should have semantic meaning, e.g. if 100 people in group ***g*** have ***f(x) = 0.6,*** then we would expect 60 of them to belong to the positive class.

Why do we want this?

- Just another accuracy metric
- In some settings, users might assess confidence, or "risk scores" rather than predicted labels (so scores should be accurate)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Generalized equalized odds (and rates)

Above definitions apply to classifiers that output *decisions*, i.e., classifiers of the form *f(x) → {0,1}*. These ideas can be generalized to classifiers that output *probabilities*, i.e., *f(x) → [0,1]*.

Generalized false positive rate:

$$FPR: \frac{FP}{FP+TN} \sim p(\hat{y}=1 \mid y_i=0)$$

$$GFPR: E(f(x) \mid y_i=0) = \frac{1}{FP+TN} \sum_{y_i=0} f(x_i)$$

Generalized false negative rate:

$$GFNR: \frac{1}{FN+TP} \sum_{y_i=1} (1-f(x_i))$$

(can also index by group, eg. $GFPR_0$, $GFPR_1$, etc)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Generalized equalized odds (and rates)

The Generalized (or “Probabilistic”) Equalized Odds definition now simply states that these two quantities should be equal:

$$GFPR_0 = GFPR_1$$

$$GFNR_0 = GFNR_1$$

(again has the same intuitive definition: errors *of a certain type* should not be biased against any group)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Treatment equality

Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories:

$$\frac{FP_0}{FN_0} = \frac{FP_1}{FN_1}$$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Treatment equality

Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories.

Why do we want this?

- FP or FN can be more/less costly for certain groups

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Other fairness definitions

The remaining definitions are more technical, and generally from specific papers

I don’t want to go through these forever, but hopefully you get the idea that there are **lots** of potential criteria that one might want to satisfy depending on what “fair” means in a particular scenario

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Individual versus group fairness

Group fairness metrics define notions of statistical parity between members of different (e.g. male / female, protected / unprotected) groups (i.e., nearly all of the metrics we've covered)

Individual fairness instead suggest that "similar" individuals should be treated similarly, regardless of their group attributes, i.e., if task-relevant features are nearby, outcomes should be similar

Fairness through awareness

Two individuals who are similar (with respect to a distance function) should receive a “similar” outcome

Will see an example of this in the next module:

$$f_1(\mathbf{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\mathbf{x}_i, y_i)\in S_1\\ (\mathbf{x}_j, y_j)\in S_2}} \underbrace{d(y_i, y_j)}_{\text{dist. between outcomes}} \Big(\underbrace{\mathbf{w}\cdot\mathbf{x}_i - \overset{\text{weighting}}{\mathbf{w}}\cdot\mathbf{x}_j}_{\text{dist. between individuals}}\Big)^2$$

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Fairness through awareness

Two individuals who are similar (with respect to a distance function) should receive a “similar” outcome

Q: Why is this called “fairness through awareness”?

A: “Awareness” just refers to understanding that notions of similarity between individuals depends on the context of a specific task (see “Fairness Through Awareness”, Dwork *et al.*)

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Counterfactual fairness

*The outcome for any individual **would be** the same even if they belonged to a different demographic group*

Roughly speaking, z does not have a causal effect on $\hat{y}$ for any instance

definitions from survey https://dl.acm.org/doi/pdf/10.1145/3457607 and related refs

Study points & take-homes

- **Disparate treatment:** members of different groups are treated differently by an algorithm
- **Disparate impact:** members of different groups receive different outcomes
- **Fairness through unawareness:** just don’t use the protected attributes explicitly in decision-making
- **Accuracy parity:** classifier accuracy should be the same across both groups
- **Demographic parity:** the likelihood of a positive outcome should be the same across both groups
- **Predictive value parity:** chance of a positive label should be the same across groups given a positive prediction (sim. for negative label)
- **Equalized odds:** similar to predictive value parity, but condition on label rather than prediction

Study points & take-homes

- **Equal opportunity:** the probability of a person in a positive class being assigned a positive outcome should be equal across groups (like equalized odds, but only positive outcomes)
- **Calibration:** probabilities output by a model should match to actual proportions
- **Generalized equalized odds:** like equalized odds, but defined over probabilities, rather than labels
- **Treatment equality:** ratio of false negatives and false positives should be the same across groups
- **Lots more:** hopefully you get the idea of how “fraught” the topic of defining fairness can be!

Intro to bias and fairness

2.4: Impossibility results

This section

- Explore the relationship between different fairness objectives, and discuss whether they are mutually compatible (**spoiler:** given that the section is called “impossibility results,” they probably aren’t!)
- Manipulate fairness objectives to construct simple proofs demonstrating their incompatibility
- Discuss the consequences of fairness goals being fundamentally incompatible

Are fairness goals compatible?

So far we’ve seen lots of potential fairness definitions or “desiderata” of fair classifiers.

Is it possible for a classifier to satisfy all of these goals simultaneously?

(No!)

Is it surprising that it’s not possible?

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

Theorem: Given features *X*, labels *Y*, and group membership *Z*:

Fix a *non-perfect*, binary classifier C*(X,Z)* and outcome *Y*. If the **prevalence** of *Y* across *Z* is not equal, then:

1. If *Z* and *Y* are not independent, then Demographic Parity and Predictive Value Parity cannot simultaneously hold
2. If *Z* and C are not independent of *Y*, then Demographic Parity and Equalized Odds Parity cannot simultaneously hold
3. If *Z* and *Y* are not independent, then Equalized Odds and Predictive Value Parity cannot simultaneously hold

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

Consider that in real contexts (for features *X*, labels *Y*, and group membership *Z*):

- Classifiers are almost never perfect
- Base-rates of outcomes (prevalence) are rarely equal across groups
- *Z* and *Y* are usually not independent (where fairness issues are concerned)
- C and *Y* are usually associated, if the classifier is any good

↳ $\hat{y}$

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

Proof: start from PPV and NPV (positive predictive value, negative predictive value, see earlier)

$$PPV = \frac{TP}{TP + FP} = \frac{TP}{\#\text{predicted positive}}$$

$$NPV = \frac{TN}{TN + FN} = \frac{TN}{\#\text{predicted negative}}$$

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

These identities can be rewritten in terms of prevalence (not particularly hard to verify, this is shown on e.g. wikipedia: https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values):

$$PPV = \frac{TPR \cdot p}{TPR \cdot p + FPR(1-p)}$$

$$NPV = \frac{(1-FPR)(1-p)}{(1-TPR) \cdot p + (1-FPR)(1-p)}$$

(*p* is “prevalence”, i.e., the fraction of positive labels)

from https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

Note: these identities just measure the *proportion of individuals assigned to positive (PPV) or negative (NPV) outcomes*

Our previous definition of **predictive value parity** simply states that these two values must be equal across groups:

$$\boxed{\begin{aligned} PPV_0 &= PPV_1 \\ NPV_0 &= NPV_1 \end{aligned}}$$

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

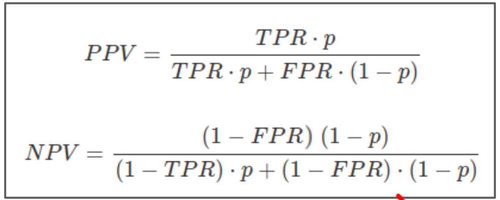
Let’s try to compute these values for different groups. First for **equalized odds.**

Recall: equalized odds requires that

$$\boxed{\begin{aligned} TPR_0 &= TPR_1 \\ FPR_0 &= FPR_1 \end{aligned}}$$

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?



$$PPV = \frac{TPR \cdot p}{TPR \cdot p + FPR \cdot (1-p)}$$

$$NPV = \frac{(1-FPR)\,(1-p)}{(1-TPR)\cdot p + (1-FPR)\cdot(1-p)}$$

(previous slide)

So, if TPRs and FPRs are equal, we should have:

Group 0:

$$PPV_0 = \frac{TPR \cdot p_0}{TPR \cdot p_0 + FPR \cdot (1-p_0)} = \frac{t \cdot p_0}{(t+f)p_0 + f}$$

$$\|$$

Group 1:

$$PPV_1 = \frac{TPR \cdot p_1}{TPR \cdot p_1 + FPR (1-p_1)} = \frac{t \cdot p_1}{(t+f)p_1 + f}$$

$NPV_0 =$ etc.

$\|$

$NPV_1 =$ etc.

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

But, **predictive value parity** requires that *PPV_g* and *NPV_g* be the same across both groups (*g=0* and *g=1*), which cannot be true if *p_0 != p_1*

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

Above was **Part 3** of the theorem (incompatibility of Equalized odds and Predictive Value Parity)

exercise: try it for the other two claims!

- Write the fairness constraints in terms of rates (*TPR* / *FPR* etc.)
- Compute *PPV_g* and *NPV_g* for both groups (*g=0* and *g=1*)
- Show that the two fairness constraints being compared lead to incompatible requirements in terms of *PPV* / *NPV* values

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Are fairness goals compatible?

The above is one of several “impossibility” results; see also e.g.

- *Inherent trade-offs in the fair determination of risk scores* (https://arxiv.org/pdf/1609.05807)

This paper proves the result for the specific cases of *calibration* and *equalized odds*

Are fairness goals compatible?

Recall:

Calibration: *Probabilities output by a classifier should have semantic meaning, e.g. if 100 people in group g have f(x) = 0.6, then we would expect 60 of them to belong to the positive class*

(Generalized) equalized odds: *The probability of a person in the **positive class** being (correctly) assigned a **positive outcome** and the probability of a person in the **negative class** being (incorrectly) assigned a **positive outcome** should be the same for both the protected and unprotected group members*

Are fairness goals compatible?

These two goals cannot be simultaneously satisfied (except for trivial edge-cases)

I won’t spend time going into this one (it’s not *that* hard but you wouldn’t learn that much from it, having already seen a different impossibility result); just want to make the point that many fairness definitions are, generally, at odds with each other

Are fairness goals compatible?

If fairness goals *aren't* compatible, what does this mean for fairness? Can a classifier ever be made "truly" fair?

No. (Imperfect) classifiers will require trade-offs. E.g. we might achieve equalized odds by increasing the chance of predicting a positive label for one group; but predictive value parity requires that the chance of predicting a positive label is preserved across groups

(put differently, the manipulations required to achieve some fairness goals violate others)

based on https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Food for thought

- If fairness goals are incompatible, what does this mean for “fairness” more generally?
- Should we give up on trying to make classifiers fair, or even measuring it?

Study points & take-homes

- Many fairness objectives are fundamentally incompatible!
- Although this may seem to undermine the objectives of fair ML, making trade-offs between such objectives is simple unavoidable when dealing with imperfect classifiers
- **Study:** try to write out different fairness objectives in terms of rates (TPR, FPR, etc.); this will give you a better sense of their meaning and differences

Intro to bias and fairness

Case study: A more detailed look at COMPAS

COMPAS recidivism algorithm

This section is based on

- Original article by propublica: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- A methodological explanation of the same article: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- Github repository of propublica's analysis: https://github.com/propublica/compas-analysis (and jupyter notebook: https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb)
- Detailed write-ups (on which these notes are directly based):
  https://afraenkel.github.io/fairness-book/content/04-compas.html
  https://afraenkel.github.io/fairness-book/content/08-compas-2.html

COMPAS recidivism algorithm

The original propublica article describes two separate incidents in Broward County, Florida:

- After school in 2014, two 18-year-old girls, Brisha Borden and Sade Jones, briefly grabbed an unlocked bicycle and scooter and rode them down the street; upon being confronted, they dropped the goods; police arrived and arrested the girls for burglary and theft of $80 worth of goods; one of the girls had previously had a minor run-in with the law, whereas the other had no record
- A 41-year-old man, Vernon Prater, was caught shoplifting $86 worth of goods from Home Depot; he had prior convictions for armed robbery and had previously served five years in prison

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

COMPAS recidivism algorithm

Both groups were booked into jail, where a judge decides how to set bail: should they be released from jail, with some amount of money as collateral, while they await trial? Broward County used COMPAS to assist the judge in making such decisions:

- A judge set bail for Brisha Borden and Sade Jones at $1000; COMPAS labeled both as high risk; a $0 bail would not be unusual in a case like this (considering age and circumstance); both spent the night in jail
- While the article doesn’t mention bail amount for Vernon Prater, it does mentione that COMPAS labeled him as low risk

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

COMPAS recidivism algorithm

Although anecdotal, COMPAS was specifically wrong in this case: the girls (labeled high-risk) never reoffended; the man was later arrested for grand-theft and is serving time in prison

Evidence points to COMPAS's risk scores having contributed to the judge's decisions for setting bail

It's worth thinking about whether COMPAS's risk scores are reasonable (beyond being wrong in this particular case), and why there's a discrepancy between the scores of these two groups

The propublica article hypothesized that the difference in scores may be due to race, so mostly we'll look at outcomes across racial groups

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

COMPAS recidivism algorithm

(if curious about the justice system generally, and the differences between “arrest”, “arriagnment”, “pretrial detention”, etc., see link)

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Digging into the data

Data from which risk scores are derived come from (a) a 137 survey (https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html) and the defendant's record. Variables include:

- Prior arrests and convictions
- Address of the defendant
- Whether the defendant a suspected gang member
- Whether the defendant ever violated parole
- If the defendant's parents separated
- If friends/acquaintances of the defendant were ever arrested
- Whether drugs are available in the defendants neighborhood
- How often the defendant has moved residences
- The defendants high school GPA
- How much money the defendant has
- How often the defendant feels bored or sad

Note: race is *not* included, but many of these variables are highly correlated with race

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

What is being modeled?

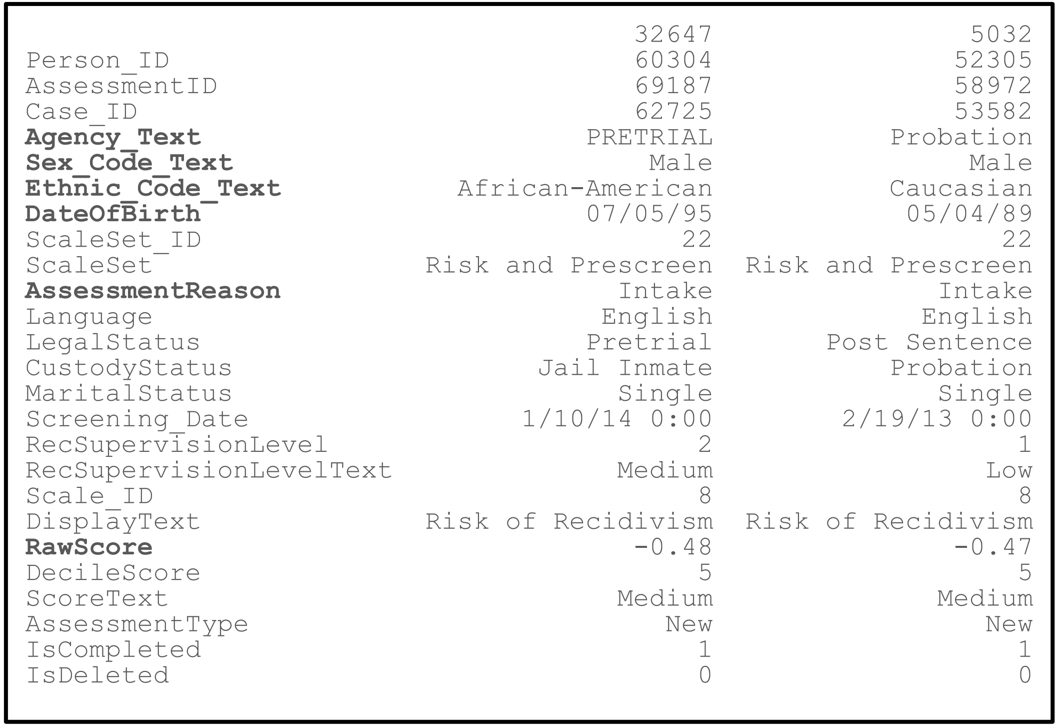
The specific outcome being modeled is *whether a defendant will commit another (felony) crime upon early release from custody*

Some thoughts:

- When considering releasing an individual charged with murder, the likelihood that they may be arrested for drug possession seems irrelevant
- Some time frame must be used for measurement (e.g. 2 years)
- To be observed reoffending, the police must come into contact with, arrest, and charge the defendant; communities that experience more police contact are more likely to be prosecuted, even if *actual* crime rates are the same

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Analysis



```
                                      32647                5032
Person_ID                             60304               52305
AssessmentID                          69187               58972
Case_ID                               62725               53582
Agency_Text                        PRETRIAL           Probation
Sex_Code_Text                          Male                Male
Ethnic_Code_Text           African-American           Caucasian
DateOfBirth                        07/05/95            05/04/89
ScaleSet_ID                              22                  22
ScaleSet                 Risk and Prescreen  Risk and Prescreen
AssessmentReason                     Intake              Intake
Language                            English             English
LegalStatus                        Pretrial       Post Sentence
CustodyStatus                   Jail Inmate           Probation
MaritalStatus                        Single              Single
Screening_Date                 1/10/14 0:00        2/19/13 0:00
RecSupervisionLevel                       2                   1
RecSupervisionLevelText              Medium                 Low
Scale_ID                                  8                   8
DisplayText              Risk of Recidivism  Risk of Recidivism
RawScore                              -0.48               -0.47
DecileScore                               5                   5
ScoreText                            Medium              Medium
AssessmentType                          New                 New
IsCompleted                               1                   1
IsDeleted                                 0                   0
```

Example samples (columns)

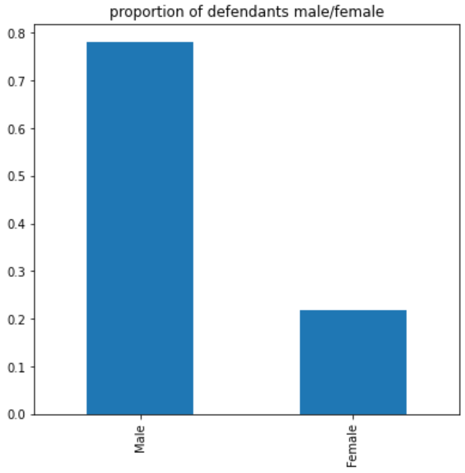
We'll look at sex/gender, race/ethnicity, age, purpose of assessment (e.g. pretrial release), type of assessment (recidivism, violent recidivismn), and the risk score itself

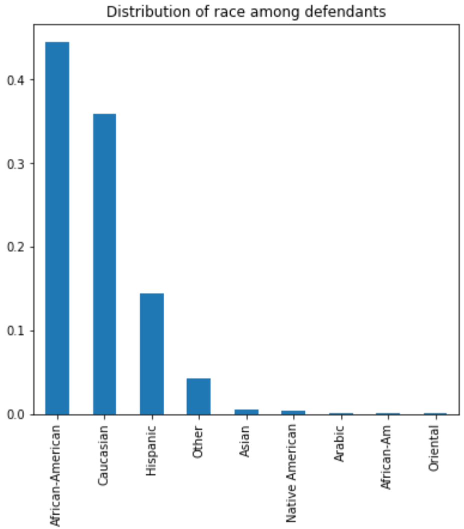
based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Analysis

Basic stats:

- Number of defendants: 20,281
- 80% male, 20% female
- Racial breakdown (right)



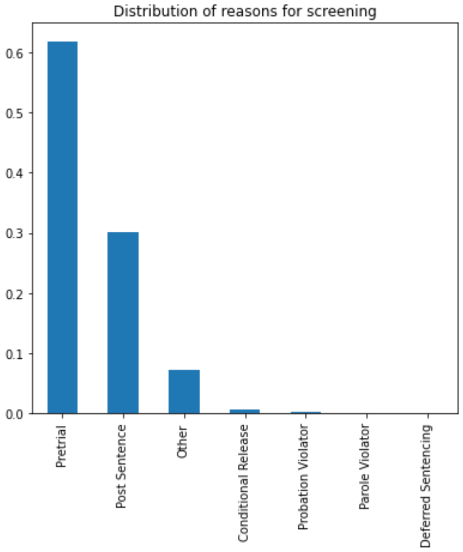


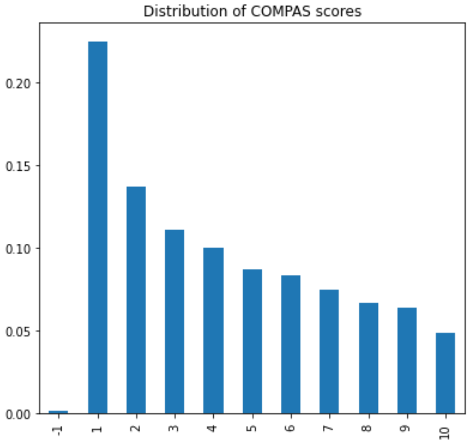
based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Analysis

Most uses of the risk assessment tool are for pretrial screening

After the lowest risk score, recidivism deciles taper off



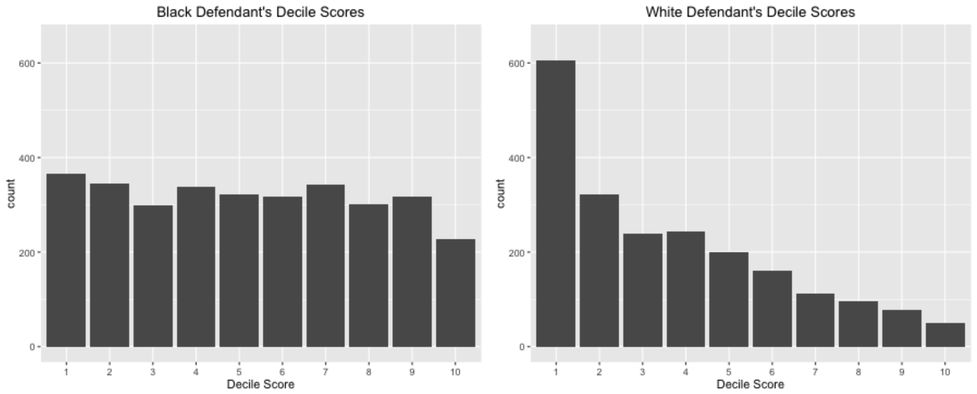


based on https://afraenkel.github.io/fairness-book/content/04-compas.html

How do deciles differ for black vs white defendants?

Decile scores are uniformly spread for black defendants, whereas for black defendants there's a downward trend in terms of decile scores

Q: is this evidence of a fairness issue? Why else might this happen?

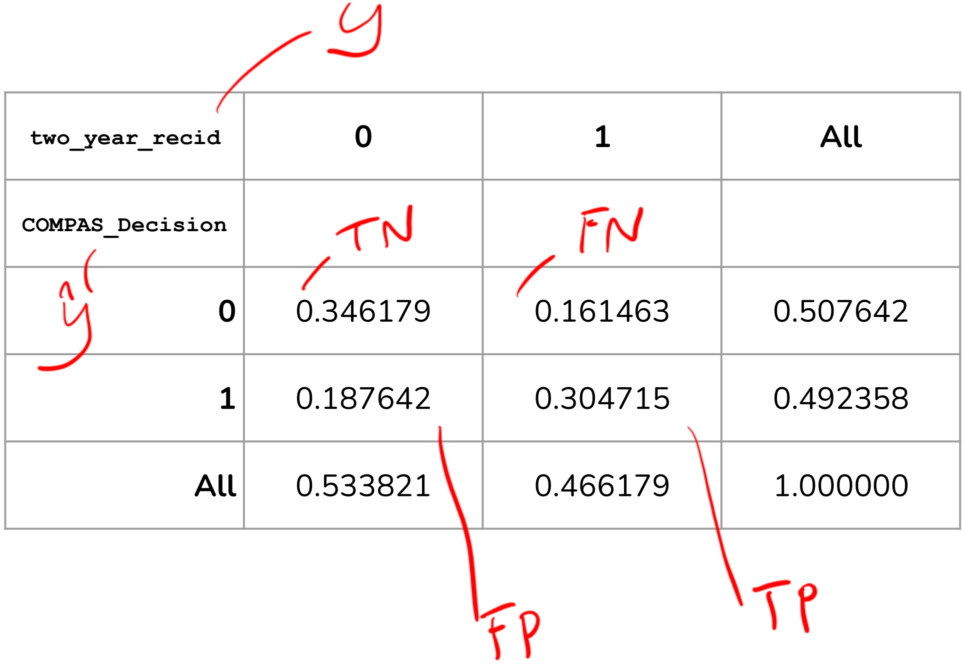


based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Did people predicted to reoffend *actually* reoffend

The two_year_recid field measures whether the defendant *actually* reoffended within two years of screening (though this variable is not always available)

- About half predicted to reoffend, slightly less than actual proportion of reoffenders
- About 35% (FP + FN) experienced an incorrect prediction



two_year_recid	0	1	All
COMPAS_Decision			
0	0.346179	0.161463	0.507642
1	0.187642	0.304715	0.492358
All	0.533821	0.466179	1.000000

y
TN
FN
$\hat{y}$
FP
TP

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Did people predicted to reoffend *actually* reoffend

When looking at the Black and white populations separately, a different picture emerges:

- A greater proportion of black defendants experience an incorrect (strict) “will reoffend” prediction than their white counterparts.
- A greater proportion of white defendants experience an incorrect (lenient) “won’t reoffend” prediction than their Black counterparts

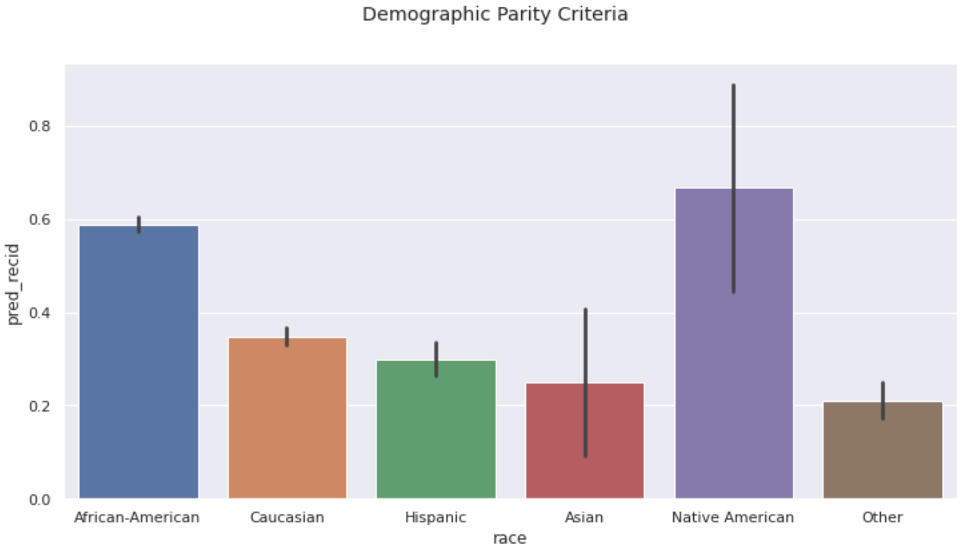
		Black			White	
two_year_recid	0	1	All	0	1	All
COMPAS_Decision						
0	0.267857	0.143939	0.411797	0.464140	0.187857	0.651997
1	0.217803	0.370400	0.588203	0.142217	0.205786	0.348003
All	0.485660	0.514340	1.000000	0.606357	0.393643	1.000000

based on https://afraenkel.github.io/fairness-book/content/04-compas.html

Parity measures – demographic parity

Q: Are the rates at which COMPAS predicts reoffending equal across groups?

A: No – but we *probably* wouldn't expect demographic parity to hold in this context (each group has different "true" rates in the data)

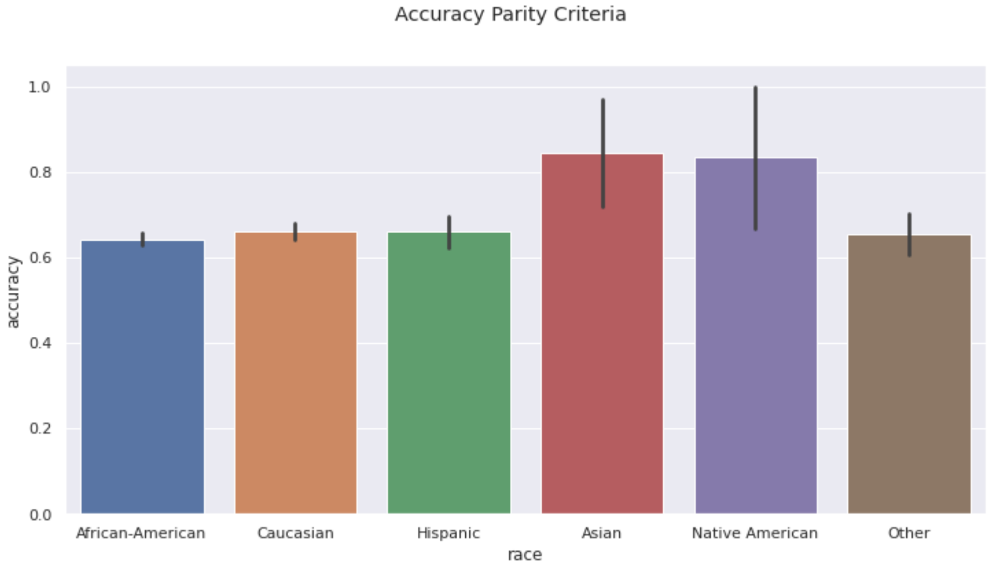


based on https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Parity measures – accuracy parity

Q: What about accuracy parity?

A: The rates of accurate predictions are roughly the same across groups



based on https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Parity measures – accuracy parity

E.g. looking at black versus white defendants specifically:

```
    accuracy (All):         0.650894
    accuracy (Black):       0.638258
    accuracy (White):       0.669927
```

There *is* a statistically significant difference at the 5% but not at the 1% level (p = 0.015)

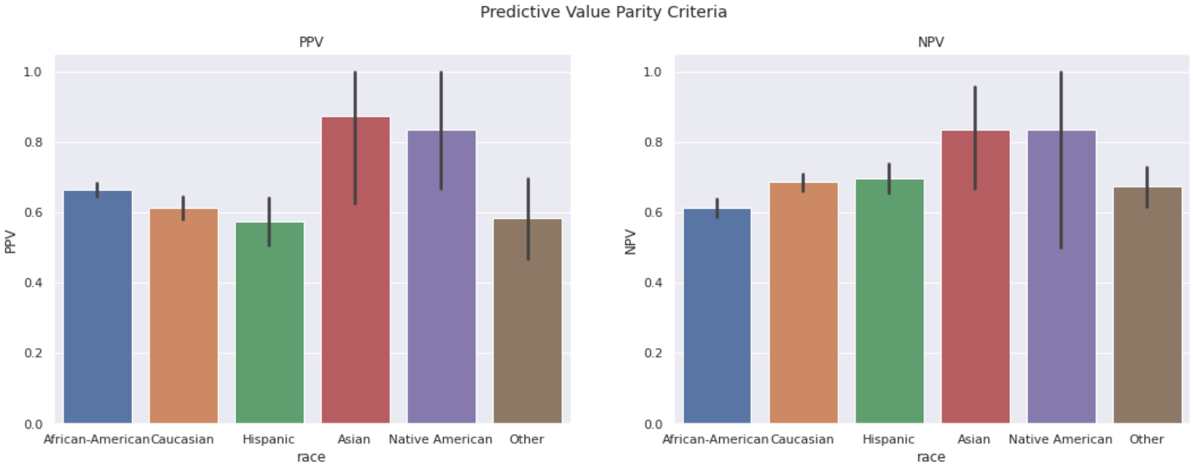
Note: *accuracy parity was used as a criterion when developing the COMPAS model*

based on https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Parity measures – predictive value parity

COMPAS largely maintains consistent rates of recidivism across groups:

- Among those labeled high risk, approximately 60% reoffended
- Among those labeled low risk, approximately 60% did not reoffend

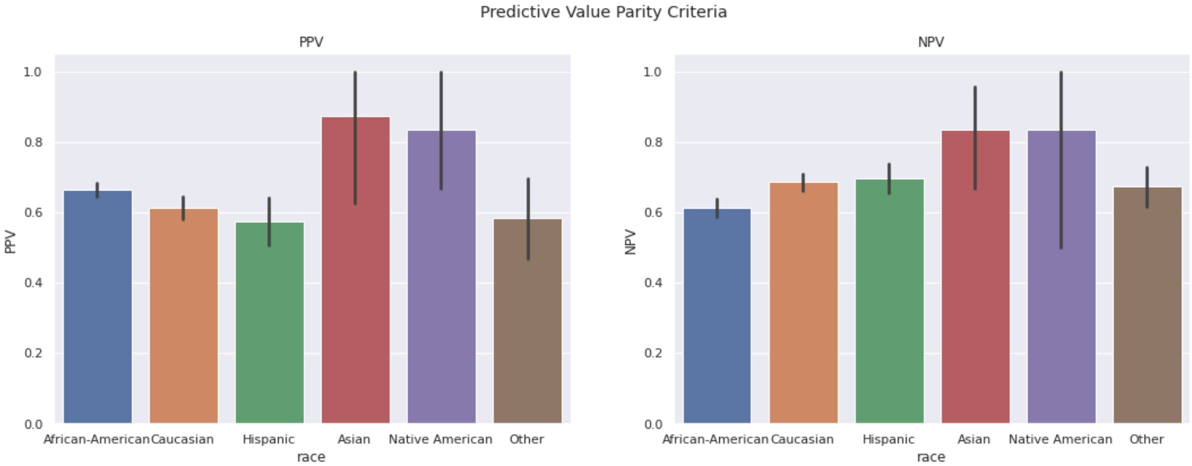


based on https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Parity measures – equalized odds

False positive and false negative rates vary significantly across groups:

- White defendants are improperly released 66% more often than black defendants
- Black defendants are improperly denied release twice as often as white defendants



based on https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Parity measures

Lots more in link:

- Calibration
- Balance of positive/negative class
- ROC Curves

Much of it is perfectly interesting but I think you get the idea…

based on https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Food for thought – what *should* COMPAS do?

- We’ve already seen some impossibility results, e.g. that accuracy parity is incompatible with other fairness definitions; so maybe we shouldn’t be surprised that COMPAS (which is tuned to achieve accuracy parity) is “unfair” according to other measures
- If it satisfied *other* fairness criteria, it mightn’t satisfy accuracy parity anymore; would that be any better?

Food for thought – what *should* COMPAS do?

- COMPAS has been argued to violate 14th Amendment Equal Protection rights, i.e., it has been argued to be racially discriminatory (https://scholarship.law.umn.edu/cgi/viewcontent.cgi?article=1680&context=lawineq)
- But with so many (incompatible) fairness definitions, couldn't *any* algorithm be argued to be racially discriminatory in the same way?

Food for thought – how should we react?

- How should we react when (e.g.) an article points out a fairness issue, given that some definitions *must* be violated (even if others are satisfied, as is the case with COMPAS)?

Food for thought – how should we react?

- COMPAS has systematic inaccuracies, but so do human judges (whether explicitly or implicitly)
- If COMPAS were *less biased than humans*, would that be sufficient justification to use it? Or should algorithms be held to a different (possibly higher) standard than people?

Food for thought

- So, algorithms might be “better” than humans, even if they are still flawed; imagine if algorithms were widely used for college admissions, replacing (more biased) human judges
- One possible concern is that algorithms might be biased in the *same* way (so an applicant might be rejected everywhere for the same reason), whereas humans are biased in *different ways* (so a candidate might get lucky if they submit enough applications)
- Does that make humans preferable to algorithms? Think back to our point from the very beginning of the module about whether adding randomness to decisions is a form of “fairness”

Study points & take-homes

- Go through the whole notebook: great place to get a sense of how all these measures are actually computed on a real dataset
- Try to think critically about the extent to which results from COMPAS are troubling, surprising, unavoidable, etc.

References for Module 2

- Fairness & Algorithmic Decision Making: https://afraenkel.github.io/fairness-book/
- A Survey on Bias and Fairness in Machine Learning: https://dl.acm.org/doi/pdf/10.1145/3457607
- A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle: https://arxiv.org/pdf/1901.10002
- Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior: https://dl.acm.org/doi/pdf/10.1145/2872427.2883083
- Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market: https://www.science.org/doi/full/10.1126/science.1121066