Fairness and algorithmic based decision making

Motivation

Issues

- Target and classes.
- Training data.
- Features.
- Looking ahead: Statistical bias and feedback loop.
- Modeling framework.
- Moral Hazard: Masking.
- A positive view.

A framework

- Formal fairness definitions.
- An impossibility theorem.
- Avoiding impossibility: equal opportunity.
- Fixing feedback loop.
- Fixing protected unobserved class bias estimation

Summing up and further work

Motivation

Motivation

White House Podesta Report (2014). Big Data: Seizing opportunities and preserving Values:

“big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.”

Barocas & Selbst (2016) suggest there are:

“areas that regulate legitimate economic discrimination, such as credit and insurance.

Motivation

- Protected classes defined by Fair Housing Act (FHA) and Equal Credit Opportunity Act (ECOA):

Table 1: Protected classes defined under fair lending laws.

Law	FHA[28]	ECOA[29]
age		X
color	X	X
disability	X	
exercised rights under CCPA		X
familial status (household composition)	X	
gender identity	X	
marital status (single or married)		X
national origin	X	X
race	X	X
recipient of public assistance		X
religion	X	X
sex	X	X

Issues

Target variable and class labels

- Target variable are usually defined based on the data available, not necessarily an objective definition.
  - Example: “creditworthiness” is an artifact of the problem definition itself. There is no way to directly measure creditworthiness because the very notion of creditworthiness is a function of the particular way the credit industry has constructed the credit issuing and repayment system (Barocas et al.).

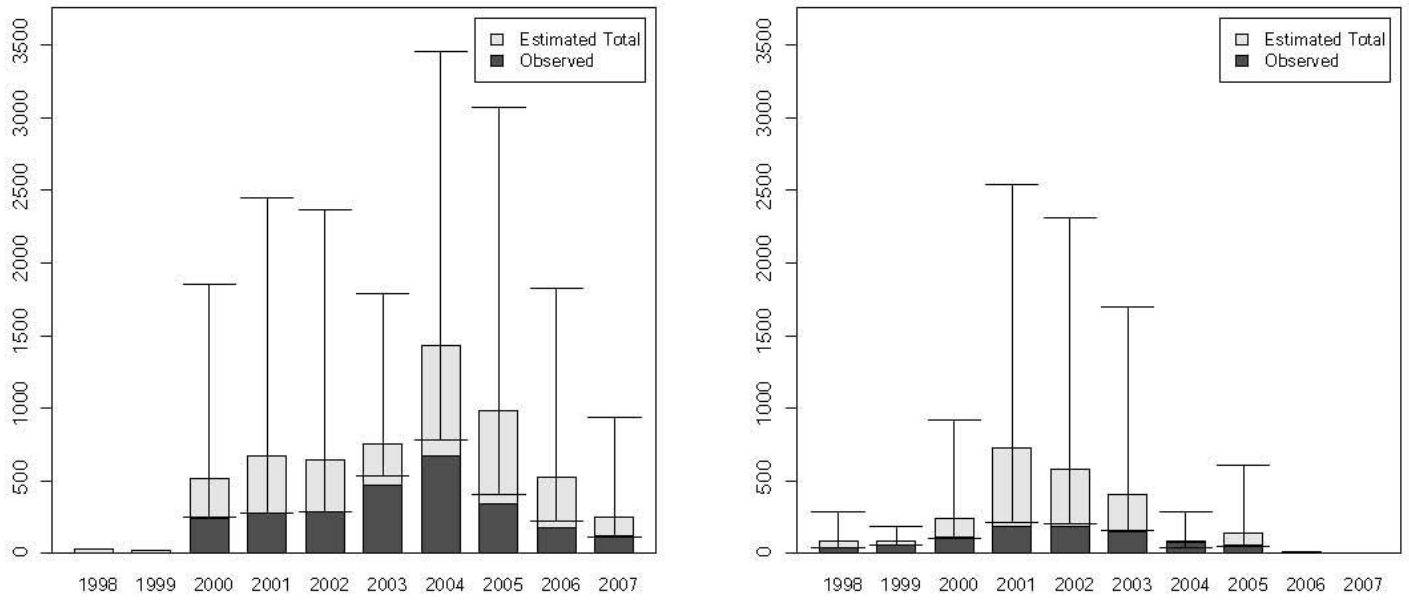
Training data

- Inferences from a biased sample of the population (missing data, sub reporting, etc.) are likely to be biased.
  - Street Bump, an application for Boston residents that takes advantage of accelerometers built into smart phones to detect when drivers ride over potholes. Potential biased by the uneven distribution of smartphones across populations in different parts of the city: underreporting of road problems in the poorer communities that lack the capability of using a smartphone (Barocas et al.).
  - Overrepresentation in a dataset can also lead to disproportionately high adverse outcomes for members of protected classes. Consider an example from the workplace: managers may devote disproportionate attention to monitoring the activities of employees who belong to a protected class and consequently observe mistakes and transgressions at systematically higher rates than others (Barocas et al., 2016).

Training data

– Sub reporting.

Statistics, Politics, and Policy, Vol. 1 [2010], Iss. 1, Art. 2



(a) Killings

(b) Disappearances

Training data

– Crime incidents are very likely to be sub reported.

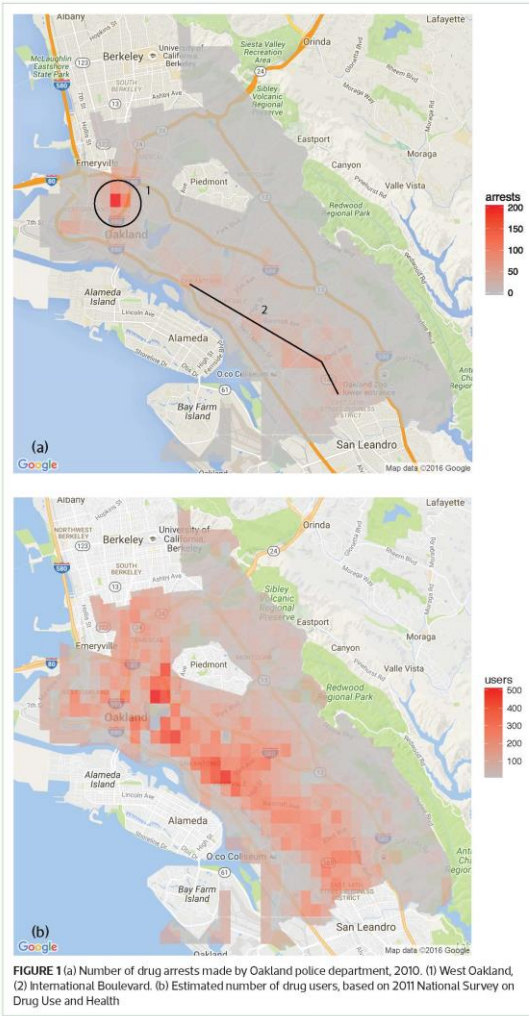


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

Training data

– … and biased.

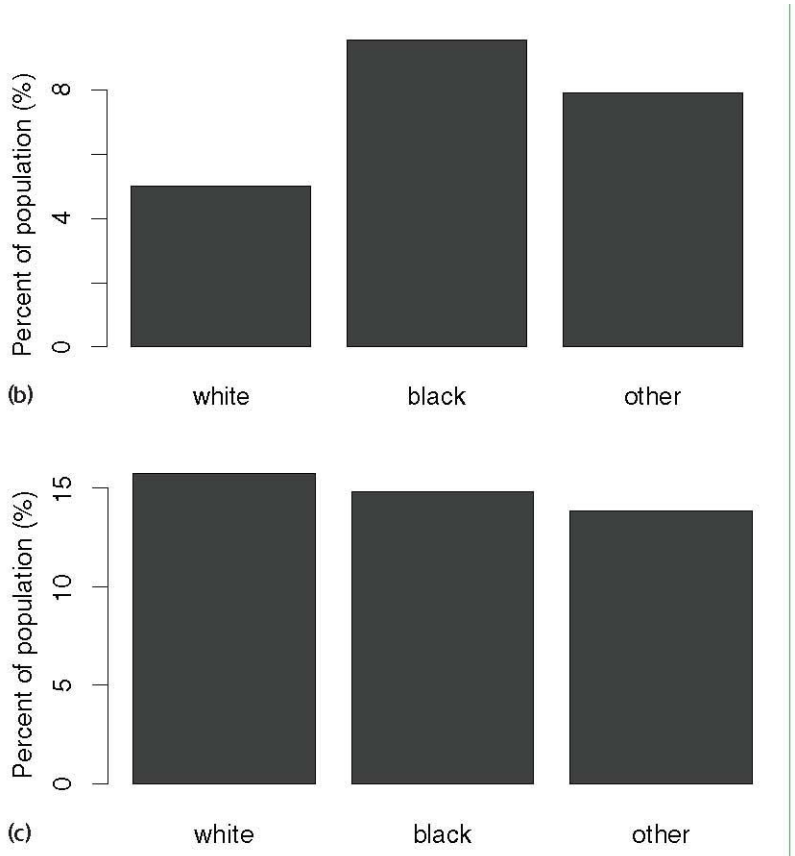


FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

Training data

- Prejudice has played some role as valid examples to learn from.
  - Google queries for black sounding names were more likely to return contextual (i.e., key-word triggered) advertisements for arrest records than those for white-sounding names. Reflecting differential delivery of advertisements that reflect the kinds of prejudice held by those exposed to the advertisements (Barocas et al.).

    https://dataprivacylab.org/projects/onlineads/

Features

- Datasets can never fully encompass the full complexity of the individuals they seek to represent, revealing the inherent inadequacy of such representations (Barocas et al., 2016).
  - Redlining: When financial institutions employ especially general criteria to draw distinctions between subpopulations (i.e., the neighborhood in which individuals happen to reside. Redlining is illegal because it can systematically discount entire areas composed primarily of members of a protected class, despite the presence of some qualified candidates.

Features

- Fairness through unawareness is ineffective due to redundant encodings (Hardt et al., 2016).
- Triangulation problem may uncover protected attributes (Fuster et al., 2018).
  - The essential insight of this paper is that a more sophisticated statistical technology (in the sense of reducing predictive mean squared error) will, by definition, produce predictions with greater variance (i.e., uneven impact in the population).
  - One possibility is that the additional flexibility available to the more sophisticated technology allows it to more easily recover the structural relationships connecting permissible variables to default outcomes.
  - Another possibility is that the structural relationship between permissible variables and default is perfectly estimated by the primitive technology, but the more sophisticated technology can more effectively triangulate the unobserved restricted variables using the observed permissible variables.

Features

- They find that Black and Hispanic borrowers are disproportionately less likely to gain from the introduction of the new technology, in terms of having lower predicted default probabilities.

- Most of the the predictive accuracy gains from the more sophisticated machine learning model are attributable to the increased flexibility of the model, with 8% or less attributable to pure triangulation.

- Simply prohibiting certain variables as predictors of default propensity will likely become increasingly ineffective as technology improves (i.e., unawareness is ineffective).

Features

- Kleinberg et al., 2019.
  - It is not always the case that the potential uneven impact hurts the protected group.
  - Kleinberg et al. examine pre-trial release decisions in New York and find that algorithms better distinguish low-risk from high-risk defendants.
  - By prioritizing the highest-risk people to detain, it becomes feasible in principle to jail 42% fewer people with no increase in crime.
  - The biggest benefits would accrue to the two groups that currently account for nine of every ten jail inmates: African-Americans and Hispanics.

Features

- Unobserved protected classes (Chen, et.al 2019)
- Probabilistic models (i.e., Bayesian Improved Surname Geocoding - BISG) for predicting the protected class based on observable proxies, such as surname and geolocation for race, are sometimes used to impute these missing labels for compliance assessments.
- Empirically, these methods are observed to exaggerate disparities.

Features

- Unobserved protected classes estimation bias (Chen, et.al 2019)

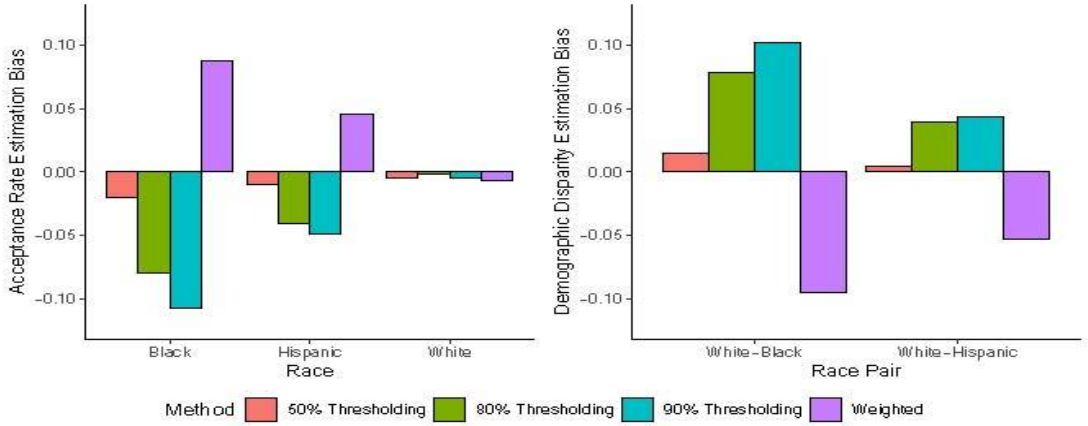
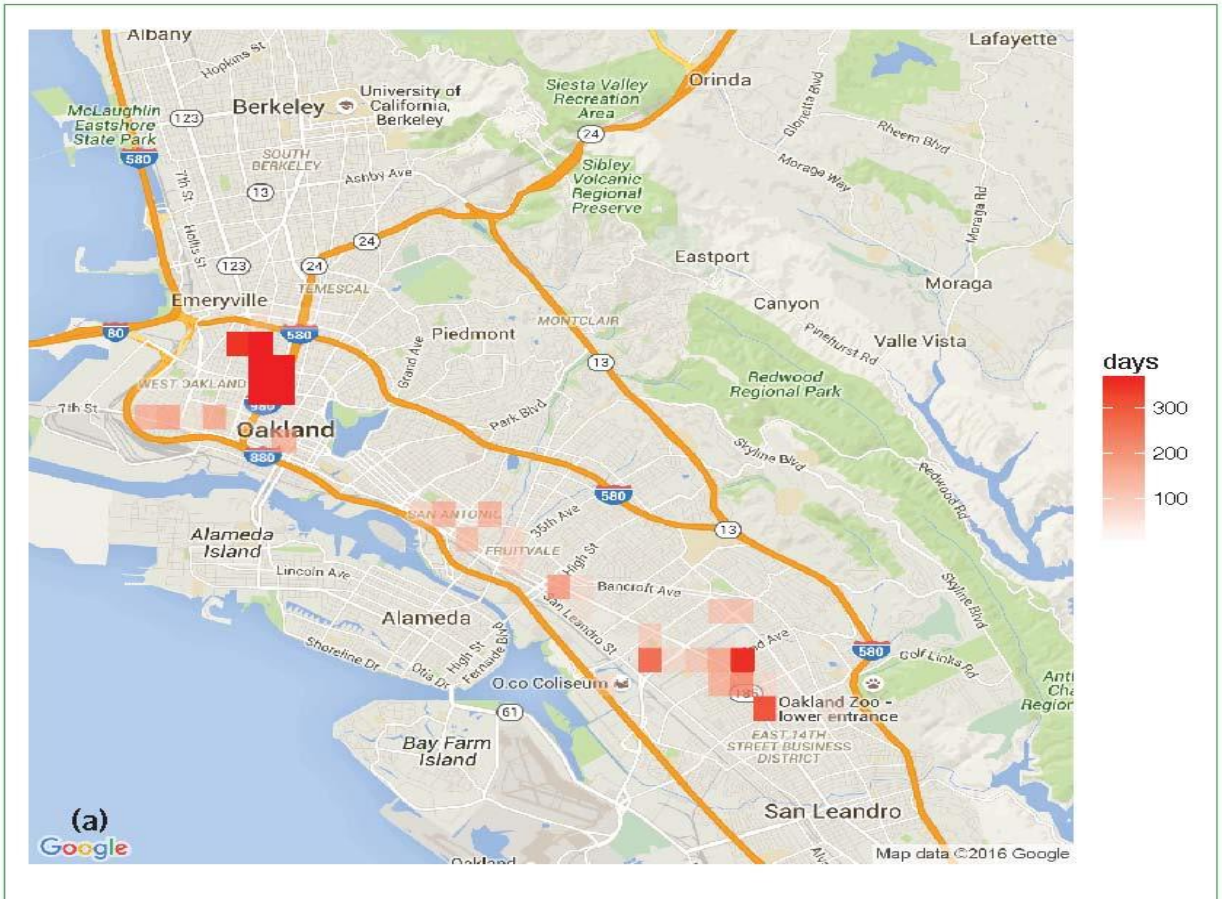


Figure 3: Left: estimation biases of loan acceptance rates μ for different races in the HMDA data set of Section 4.2, using the thresholded estimator for mean group outcome $\hat{\mu}_q$ $(q = 0.5, 0.8, 0.9)$ from Definition 2.4, as well as the weighted estimator $\hat{\mu}_W$ from Definition 2.5, relative to the true mean group outcome μ calculated using the actual race labels. Right: estimation biases of demographic disparity δ between pairs of races, using the thresholded estimator $\hat{\delta}_q$ from Definition 2.4 and weighted estimator $\hat{\delta}_W$ from Definition 2.5, relative to the true demographic disparity δ calculated using the actual race labels.

Feedback loop

- Self exciting poisson models at work (e.g., Predpol, Bogota crime prediction model, etc.)



Modeling framework

- Kleinberg et al., 2019.
  - Both screening and training algorithms should be analyzed.

Moral hazard

- Masking
  - Intentional exploitation of each one of the mechanisms enumerated above.
  - Any form of discrimination that happens unintentionally can also be orchestrated intentionally.

(Legal) External validation test

- Assessing these issues requires to think on (Barocas et al., 2016, Kleinberg et al., 2019):

  - Disparate treatment: The prohibition on disparate treatment reflects a commitment to a kind of neutrality. For example, public officials are not permitted to favor men over women or white people over black people. Civil rights statutes forbid disparate treatment along a variety of specified grounds, such as race, sex, national origin, religion, and age.

  - Disparate impact: The prohibition on disparate impact means, in brief, that if some requirement or practice has a disproportionate adverse effect on members of protected groups the defendant must show that the requirement or practice is adequately justified. Employers must show that the practices are justified by “business necessity.”

A positive view

- Kleinberg et al., 2019.
  - “The use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred.”
  - “By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central tradeoffs among competing values.”
  - “Algorithms are not only a threat to be regulated; with the right safeguards in place, they have the potential to be a positive force for equity”.
  - “Regulating the algorithmic building process will stand out for transparency and specificity. Formalizing, quantifying and posing well difficult questions.”
  - Traditional obstacles such as the difficulty in quantifying subjective intentional and unintentional human biases, may be addressed when using algorithms.

A framework

Fairness definitions

- Kleinberg et al., 2016. Defining a tractable measure of fairness is a challenge.

- Common approaches such as *demographic parity* are flawed: In a binary classification task it means that the algorithm has to predict the same proportion of positive outcomes across protected groups (i.e., prediction is independent of protected attribute).

  – Attempting to enforce equal impact between men and women in recidivism prediction systems, if men have higher reoffending rates, could result in women remaining in prison longer despite being less likely to reoffend (Bins, 2018).

Fairness definitions

- Kleinberg et al., 2016. Three principles (note that these principles are based on observed data = outcomes, covariates, protected attributes and forecasts):

  1. Probability estimates should be well calibrated. For all instances such that your algorithm forecasts a probability $p \in [0,1]$ of a positive class, observed frequencies should match p, and this should hold also conditional to each protected attribute (i.e., group).

Fairness definitions

Example:

If the model predicts a group of instances have probability 30% of being of positive class. Then approximately 30% of these instances are observed to be of positive class.

If the model predicts a group of women instances have probability 25% of being of positive class. Then approximately 25% of these instances are observed to be of positive class.

Fairness definitions

- Uncalibrated algorithm (Chouldechova, et.al. 2018)

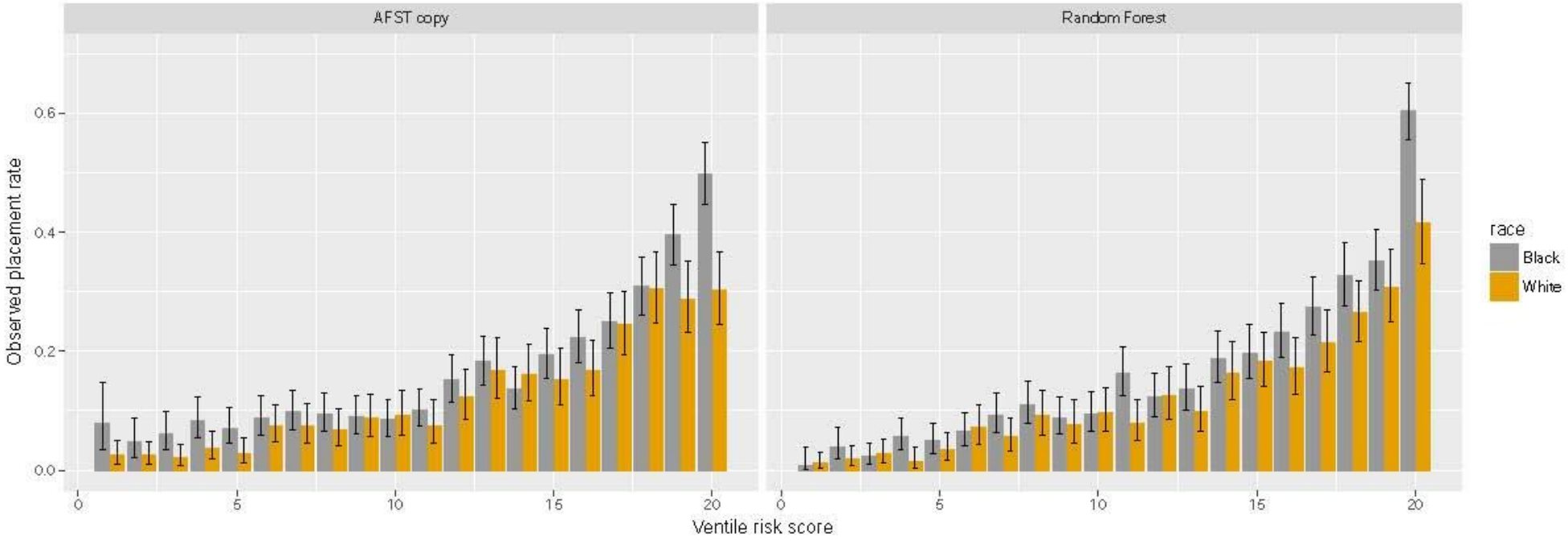


Figure 4: Observed placement rates by AFST model (left) and Random forest model (right) risk score ventile broken down by victim's race. Error bars correspond to 95% confidence intervals.

Fairness definitions

2. Balance for positive class: Conditional on being of positive class, the probability of being of this class is equal between protected attributes (*true positive rates are equal across protected attributes*).

Fairness definitions

3. Balance of negative class: Conditional on being of negative class, the probability of being of positive class is equal between protected attributes (*false positive rates are equal across protected attributes*).

Fairness definitions

- Unbalanceness (Chouldechova, et.al. 2018)

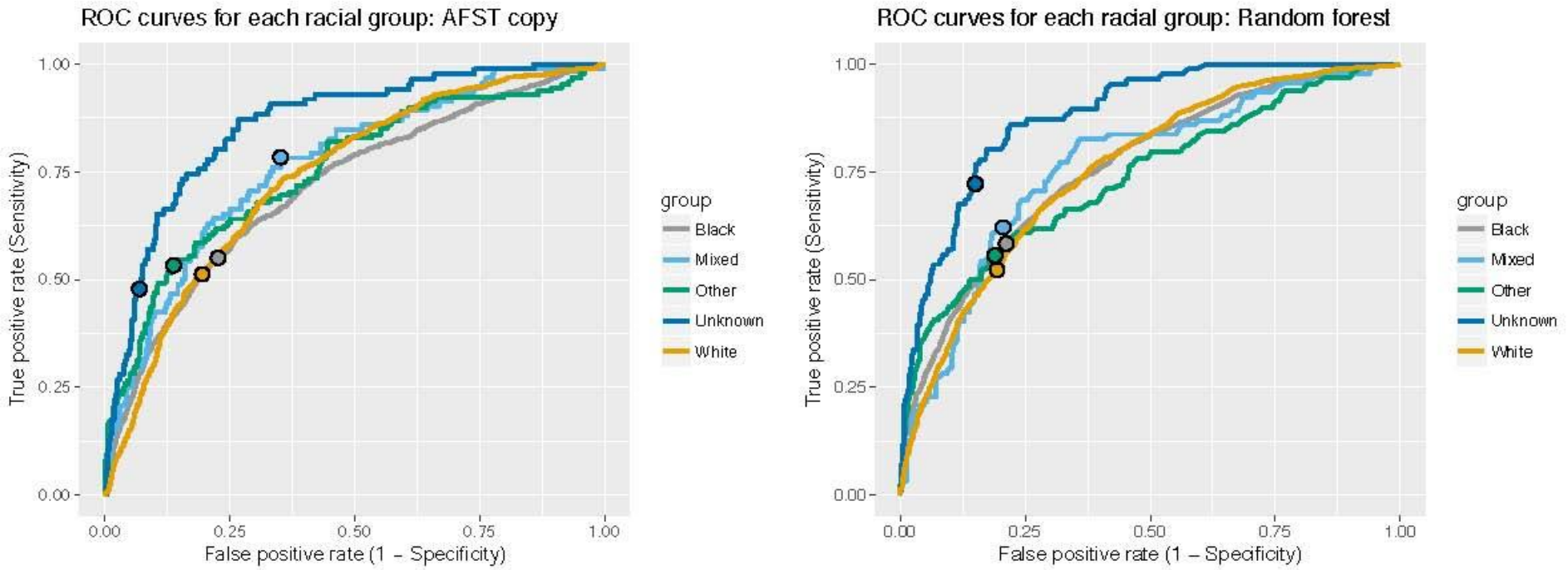


Figure 5: Race-specific ROC curves for the AFST copy (left) Random forest (right) models. Points overlaid on the curves correspond to the (FPR, 1 - FNR) values at the 25% highest risk cutoff delineating mandatory screen-ins (see Section 2).

Fairness definitions

- Fairness and causality (Khademi, et.al. 2019):
  - Instead of asking, is a decision (algorithmic) discriminating with respect to a protected attribute?
  - Ask: Does the protected attribute have a casual effect on the decision (causal effect)?

- The second version of the problem calls for a decision-making procedure that is not sensitive to a counterfactual instance of the problem: If the protected attribute changes, in a way that never seen before, the decision shouldn't change.

- Two definitions: Fair on average causal effect (FACE), fair on average causal effect of the treated (FACT)

An impossibility theorem

- Kleinberg et al., 2016.
- These three: calibration within groups, balance of positive classes and balance of negative classes are natural measures of fairness.
- Except for very special cases (perfect prediction or equal base rates): There is no algorithms that simultaneously (even approximately) satisfies the three properties.

Avoiding impossibility: equal opportunity

- Hardt et al., 2016.

- Assume the positive outcome is the preferred normative outcome (good client, non defaulter, safe trip, etc.)
- A weaker notion of fairness is: Balance of positive classes (i.e., *equal opportunity*).
- If you are good (of positive class), the algorithm should not discriminate against you.
- Even if the algorithm is not *fair* under this definition, it can be approximately fixed.

Avoiding impossibility: equal opportunity

- Hardt et al., 2016.

- Given a predictor (algorithm) a derived (expost) predictor can be constructed that optimizes a loss function (minimizes loss) subject to positively balance (equal opportunity).

Fixing feedback loop

- Ensign, D. et al. 2018.
- Consider modifying the training set as new incidents are reported.
- For a safety trip with a very high probability (p) if in fact it is observed that there was no incident drop the observation from your training data set with probability p.

Fixing protected unobserved class bias estimation

- Work in progress

Summing up and further work

Summing up and further work

- Identified issues provide a taxonomy and template to study fairness.
- The framework suggests testing experiments using the outcomes of current testing phases.
- Moreover, this framework also suggests some ways of mitigating some of the potential issues already identified.

References

- White House Podesta Report (2014). **Big Data: Seizing opportunities and preserving Values.**
- Barocas, S. and A. Selbst (2016). **Big Data's Disparate Impact.**
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T and A. Walther (2018) **Predictably Unequal? The Effects of Machine Learning on Credit Markets.**
- Hardt, M., Price, E. and N. Srebro (2016). **Equality of Opportunity in Supervised Learning.**
- Kleinberg, J., Ludwig, j., Mullainathan, S., and C. Sunstein. 2019 **Discrimination in the age of algorithms.**
- Lum, K. and W. Isaac (2016). **To predict and serve?**
- Kleinberg, J., Mullainathan, S. and M. Raghavan (2016). **Inherent Trade-Offs in the Fair Determination of Risk Scores**.
- Enslgn, D., Firedler, S.,Neville, S., Scheldegger, C. and S. Venkatasubramanian (2018**). Runaway Feedback Loops in Predictive Policing.**

References

- Binns, R. (2018). **Fairness in Machine Learning: Lessons from Political Philosophy.**
- Chouldechova, A., Putman-Hornstein, E., Benavides-Prado, D., Flako, O. and R. Vaithianatha (2018). **A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions.**
- Chen, J., Kallus, N., Mao, X., Svacha, G. and M., Udell (2018). **Fairness under unawerness: Assessing Disparity when Protected Class is Unobserved.**
- Khademi, A., Sanghack, Lee., Foley, D. and V. Honaver. **Fairness in Algorithmic Decision Making: An Excursion through the lens of causality.**