Building a Robot Judge: Data Science for Decision-Making

12. Algorithms and Decisions III: Fairness

Outline

Recap: Using Machine Learning to Guide Audit Policy

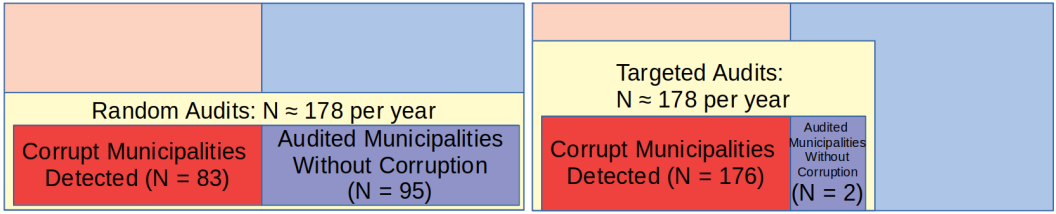
Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

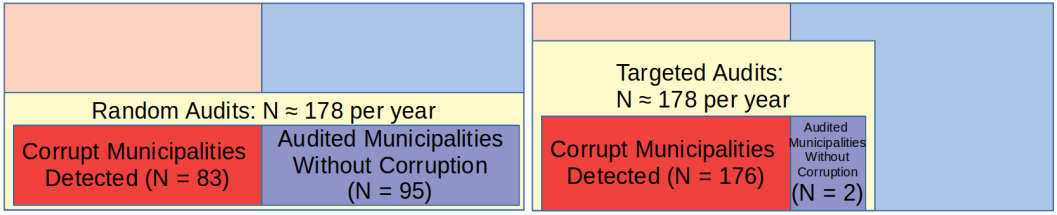
Problems with Algorithmic Fairness

Comparing the Policies



- Holding number of audits constant, targeting increases detections by 120%.
- Detection probability per corrupt municipality more than doubles – from 2.9% to 6.7%.

Comparing the Policies



- Holding number of audits constant, targeting increases detections by 120%.
- Detection probability per corrupt municipality more than doubles – from 2.9% to 6.7%.
- To achieve same number of detections as status quo (83 municipalities), only 84 targeted audits are needed.
  - Decrease of 94 audits per year (53%), a major reduction in audit resources.
- ***Why don't we need to use the contraction method a la Kleinberg et al 2018?***

Incentive Effects of Targeted Audits

- Remember that one of our criteria for ML-powered decision-making is that decision subjects don't respond to the algorithm.
- But **in the case of detecting corruption, this is exactly what we want**:
  - corruption makes audits more likely $\rightarrow$ reduces incentives and probability of corruption!

Mechanism Design Issues

- With repeated audits, there could be behavioral responses by local officials.
  - could produce significant errors favoring savvy mayors.
  - Would still deter corrupt fiscal actions that are not easily substitutable.

How much information to publicize about audit targeting?

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.
- mayors might learn how algorithm works over time.
- weights could be updated in response to behavioral responses

Mixing random and targeted audits

- Random audits could be maintained (along with targeted audits).
  - Preserves some deterrence incentive for all municipalities.
  - Results of random audits could be used to update algorithm parameters.

Outline

Recap: Using Machine Learning to Guide Audit Policy

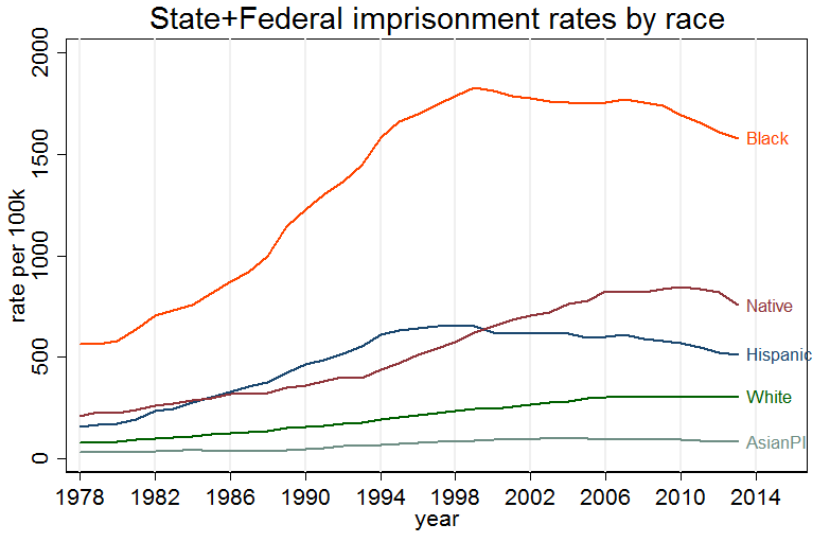
Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

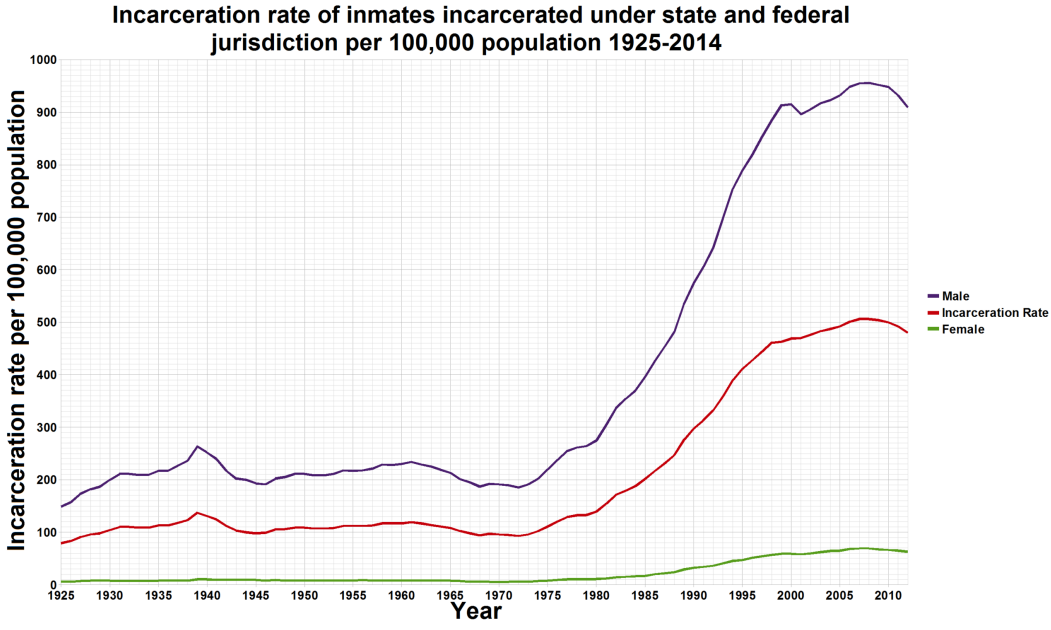
Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

Incarceration Rates by Race in U.S.A, 1978-2014



Incarceration Rates by Gender in U.S.A, 1925-2014



Homicide Offending Rates, by Race and Gender

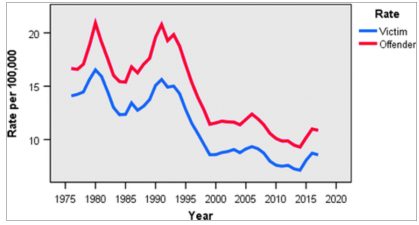
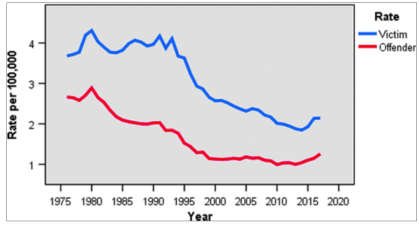


FIG. 1. Offending and victimization rates for men, 1976–2017.



Can group differences in preferences/ability explain variation in crime and incarceration?

- Is answer different for race and gender?

Can group differences in preferences/ability explain variation in crime and incarceration?

- Is answer different for race and gender?
- Preferences/ability could be the result of past discrimination/disadvantage.
  - disparities in health/education
  - prejudice leading to demotivation
  - etc.

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are
- Prejudice will reduce profits $\rightarrow$ in a competitive market, discriminating firms will be competed out.

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are
- Prejudice will reduce profits $\rightarrow$ in a competitive market, discriminating firms will be competed out.
  - could remain with other labor market frictions, e.g. imperfect competition

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are
- Prejudice will reduce profits $\rightarrow$ in a competitive market, discriminating firms will be competed out.
  - could remain with other labor market frictions, e.g. imperfect competition
  - could remain in public sector (e.g. judicial decisions)

In-Group Bias on Criminal Trial Juries (Anwar, Bayer, and Hjalmarsson 2012)

- Examine jury racial composition and trial outcomes in Florida, 2000-2010
- Exogenous treatment: day-to-day variation in composition of jury pool
  - Identification check: composition of jury pool uncorrelated with characteristics of the defendant and case.

In-Group Bias on Criminal Trial Juries (Anwar, Bayer, and Hjalmarsson 2012)

- Examine jury racial composition and trial outcomes in Florida, 2000-2010
- Exogenous treatment: day-to-day variation in composition of jury pool
  - Identification check: composition of jury pool uncorrelated with characteristics of the defendant and case.

TABLE IV

REDUCED-FORM BENCHMARK REGRESSIONS

	(1)	(2)	(3)	(4)
Dependent variable	Any guilty conviction		Proportion guilty convictions	
Black defendant	0.150***	0.164***	0.156***	0.160***
	[0.056]	[0.058]	[0.055]	[0.057]
Any black in pool	0.069	0.105**	0.063	0.090*
	[0.048]	[0.051]	[0.047]	[0.050]
Black defendant * any black in pool	−0.168**	−0.166**	−0.174**	−0.155**
	[0.070]	[0.074]	[0.069]	[0.072]
Constant	0.656***	0.627***	0.600***	0.576***
	[0.039]	[0.041]	[0.038]	[0.040]
Includes controls for:				
Gender/age of pool	No	Yes	No	Yes
County dummy	No	Yes	No	Yes
Year of filing dummies	No	Yes	No	Yes
Observations	712	712	712	712
R-squared	0.01	0.07	0.01	0.08

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality
- Different priors (stereotypes) about productivity/criminality.

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality
- Different priors (stereotypes) about productivity/criminality.
  - could be self-confirming: employer/judge doesn't give the stereotyped group a chance to prove themselves.

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality
- Different priors (stereotypes) about productivity/criminality.
  - could be self-confirming: employer/judge doesn't give the stereotyped group a chance to prove themselves.
  - another channel for self-confirmation: minority workers expect to be discriminated against, and therefore don't invest in education/skills.

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced
- Can formally test in this context:
  - statistical discrimination $\rightarrow$ contraband discovery (successful search) rates will be the same for both groups.

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced
- Can formally test in this context:
  - statistical discrimination $\rightarrow$ contraband discovery (successful search) rates will be the same for both groups.
  - prejudice $\rightarrow$ contraband discovery rates will be lower for black drivers, as threshold for search is lower.

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced
- Can formally test in this context:
  - statistical discrimination $\rightarrow$ contraband discovery (successful search) rates will be the same for both groups.
  - prejudice $\rightarrow$ contraband discovery rates will be lower for black drivers, as threshold for search is lower.
- Empirical test:
  - data on 1500 traffic searches in Maryland, 1995-1999
  - contraband discovery rates are the same across races, consistent with statistical discrimination, but not taste-based discrimination

Arnold, Dobbie, and Hull (AER 2022), Discrimination in Bail Decisions

- Judge bias in bail decisions could be due to differences in recidivism risk. Arnold et al train a regression model to predict recidivism and then condition on it.

Arnold, Dobbie, and Hull (AER 2022), Discrimination in Bail Decisions

- Judge bias in bail decisions could be due to differences in recidivism risk. Arnold et al train a regression model to predict recidivism and then condition on it.
- Judges still biased to release white defendants after adjustment:

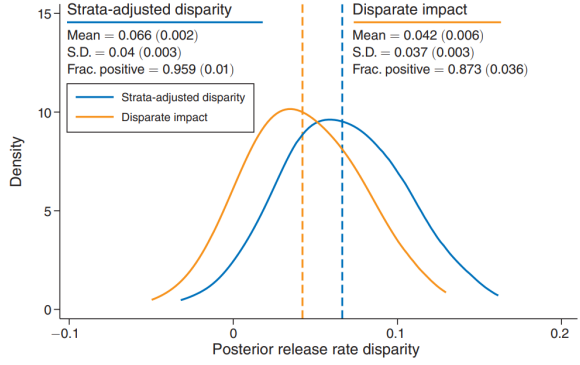


FIGURE 3. OBSERVATIONAL DISPARITIES AND DISPARATE IMPACT ESTIMATES

Arnold, Dobbie, and Hull (AER 2022), Discrimination in Bail Decisions

- Judge bias in bail decisions could be due to differences in recidivism risk. Arnold et al train a regression model to predict recidivism and then condition on it.
- Judges still biased to release white defendants after adjustment:

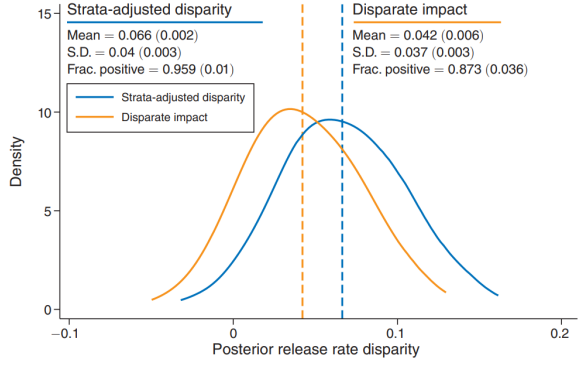


FIGURE 3. OBSERVATIONAL DISPARITIES AND DISPARATE IMPACT ESTIMATES

- Does not necessarily entail bias (Canay et al 2022)

Detecting discrimination by humans is difficult

(Kleinberg, Ludwig, Mullainathan, Sunstein 2020)

- A firm has fewer male than female employees, plaintiff alleges discrimination.
- Firm says that it is a pipeline problem.

Detecting discrimination by humans is difficult

(Kleinberg, Ludwig, Mullainathan, Sunstein 2020)

- A firm has fewer male than female employees, plaintiff alleges discrimination.
- Firm says that it is a pipeline problem.
- HR manager testifies they are not biased.
  - even if HR manager is biased, might not realize it.

Detecting discrimination by humans is difficult

(Kleinberg, Ludwig, Mullainathan, Sunstein 2020)

- A firm has fewer male than female employees, plaintiff alleges discrimination.
- Firm says that it is a pipeline problem.
- HR manager testifies they are not biased.
  - even if HR manager is biased, might not realize it.
- Statistics show that equally competent women to men were not hired
  - but what "competency metrics" to use?
  - many competency factors, and they are subjective

Detecting discrimination by robots is easier

(Kleinberg, Ludwig, Mullainathan, Sunstein 2020)

- Firm can submit the data and model for audit.

Detecting discrimination by robots is easier

(Kleinberg, Ludwig, Mullainathan, Sunstein 2020)

- Firm can submit the data and model for audit.
- change the gender in the dataset and see if predictions change.

Detecting discrimination by robots is easier

(Kleinberg, Ludwig, Mullainathan, Sunstein 2020)

- Firm can submit the data and model for audit.
- change the gender in the dataset and see if predictions change.
- can see if its possible to train a model with similar performance and less disparities.
  - different predictors
  - different algorithm
  - different outcome label

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

“Fair ML” / “AI Fairness”

- There is growing concern about social harms and disparities produced by AI decisions.

“Fair ML” / “AI Fairness”

- There is growing concern about social harms and disparities produced by AI decisions.
- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?

"Fair ML" / "AI Fairness"

- There is growing concern about social harms and disparities produced by AI decisions.
- "ML" or "AI" refer to statistical algorithms
  - can learning algorithms be fair or not?
- Rather: *fairness* is a property of *decisions*.
  - so *"AI Fairness"* should be understood as "*fairness of AI-supported decision-making*".

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
- Firms using ML to screen job applicants might wish to incorporate diversity objectives.

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
- Firms using ML to screen job applicants might wish to incorporate diversity objectives.
- Judges might want to reduce biases in legal decisions.

List of Protected Attributes Specified in US Fair Lending Laws

- Fair Housing Acts (FHA)
- Equal Credit Opportunity ACts (ECOA)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

- Machine learning researchers take these as given.

Data can be biased

- e.g. Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).

Data can be biased

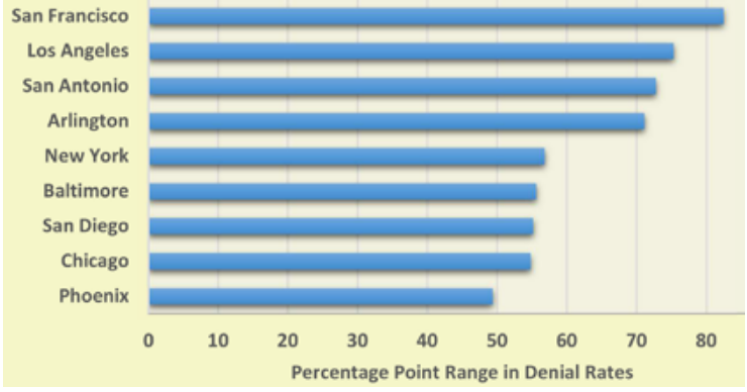
- e.g. Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.

Data can be biased

- e.g. Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
- If possible, reduce data bias!
  - the other fairness approaches are second-best.

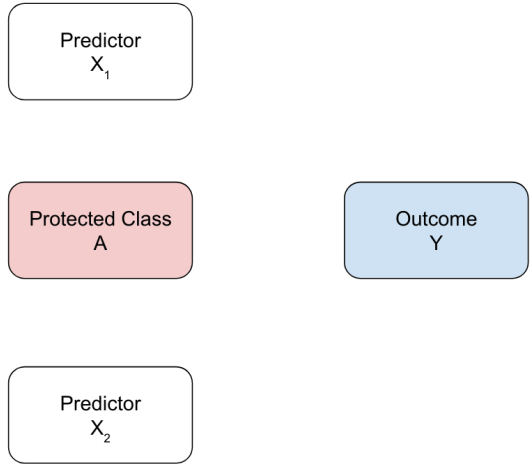
Humans are Biased

- Before getting into bias towards particular groups, it should be emphasized that humans are "biased" in the sense that some are more/less lenient:



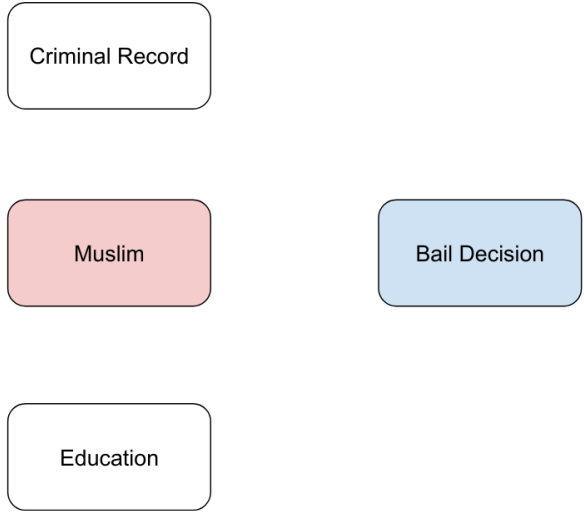
- A robot judge would generate consistent decisions for same evidence, correcting individual-level leniencies across judges.

Overview: Fairness in Decision-Making

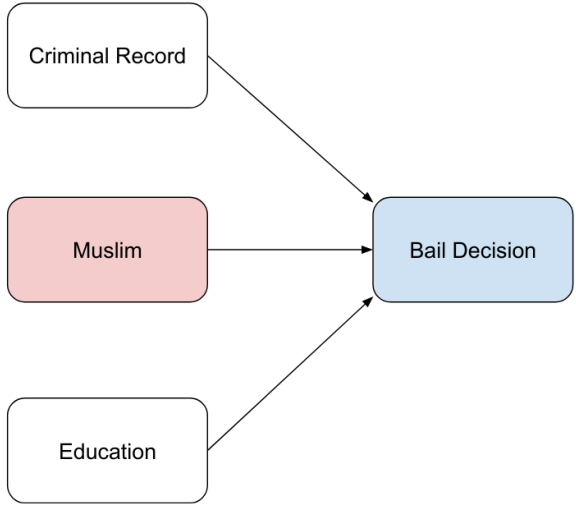


- $A \in \{0,1\}$= protected class, X= other predictors, Y= outcome.
- let $\hat{Y}(X,A)$ be our model predictions.

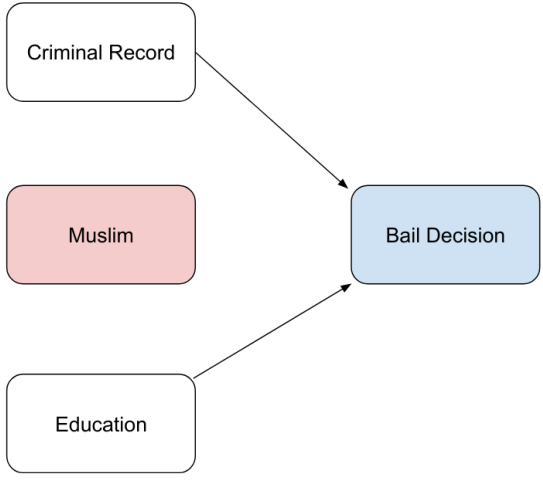
For example:



Standard Approach: Use All Data

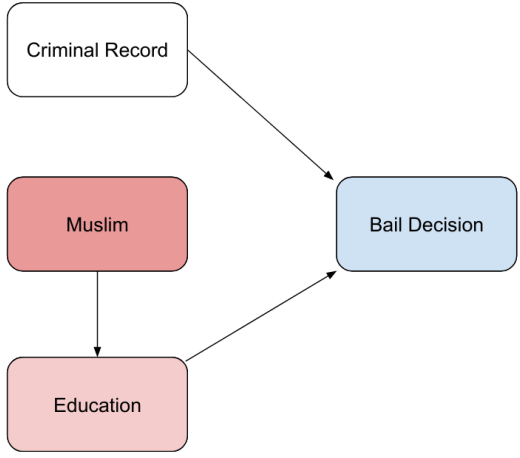


Fairness through Unawareness



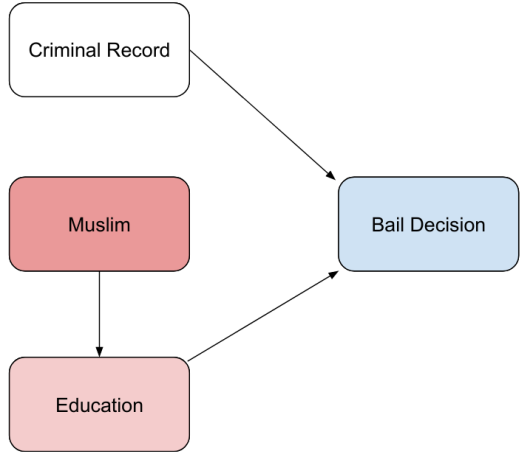
- **Fairness through unawareness:** protected attributes are not explicitly used in the prediction process.
  - that is, $\hat{Y}(X,0) = \hat{Y}(X,1)$, $\forall X$.

Problem: Indirect Discrimination



- sensitive factors are implicitly being used by the model, to the extent that they are correlated with included predictors.
  - e.g., muslims have lower education than rest of population.

Problem: Indirect Discrimination



- sensitive factors are implicitly being used by the model, to the extent that they are correlated with included predictors.
  - e.g., muslims have lower education than rest of population.
- in most datasets, if you drop the sensitive attribute and train a new classifier, the resulting predictions will be the same or very close.

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

Review: Classification Metrics

	Predicted Positive	Predicted negative
Actual Positive	$TP = \#$ true positives	$FN = \#$ false negatives
Actual Negative	$FP = \#$ false positives	$TN = \#$ true negatives

- Identify the correct sequence of labels for the following four metrics, separated by commas.
  1. $\frac{TP+TN}{TP+TN+FP+FN}$
  2. $\frac{TP}{TP+FP}$
  3. $\frac{TP}{TP+FN}$
  4. $\frac{FP}{FP+TN}$

Classification Metrics

Event	Condition	Associated metric: Pr{event | condition}	Formula: $\frac{\#\text{ event}}{\#\text{ condition}}$
$\hat{Y}=1$	$Y=1$	True positive rate [Recall for positive class]	$\frac{TP}{TP+FN}$
$\hat{Y}=0$	$Y=0$	True negative rate [Recall for negative class]	$\frac{TN}{TN+FP}$
$Y=1$	$\hat{Y}=1$	Positive predictive value [Precision for positive class]	$\frac{TP}{TP+FP}$
$Y=0$	$\hat{Y}=0$	Negative predictive value [Precision for negative class]	$\frac{TN}{TN+FN}$
$\hat{Y}=1$	$Y=0$	False positive rate	$\frac{FP}{TN+FP}$
$\hat{Y}=0$	$Y=1$	False negative rate	$\frac{FN}{TP+FN}$
$Y=1$	$\hat{Y}=0$	?	$\frac{TP}{TN+FN}$
$Y=0$	$\hat{Y}=1$	?	$\frac{TN}{TP+FP}$

- $Y \in \{0,1\}$= outcome label, e.g. reoffends or not; $\hat{Y} \in \{0,1\}$ = classifier output label
- TP = # true positives, FN = # false negatives, FP = # false positives, TN = # true negatives

Classifier Setup

- $Y \in \{0,1\}$ = outcome label, e.g. reoffends or not
- X = predictors, e.g. criminal history
- $A \in \{0,1\}$ = protected class, e.g. gender

Classifier Setup

- $Y \in \{0,1\}$ = outcome label, e.g. reoffends or not
- X = predictors, e.g. criminal history
- $A \in \{0,1\}$ = protected class, e.g. gender

Classifier output:

- $\hat{y}(X,A) \in [0,1]$ = the **score**, usually interpreted as a predicted probability

Classifier Setup

- $Y \in \{0,1\}$ = outcome label, e.g. reoffends or not
- X = predictors, e.g. criminal history
- $A \in \{0,1\}$ = protected class, e.g. gender

Classifier output:

- $\hat{y}(X,A) \in [0,1]$ = the **score**, usually interpreted as a predicted probability
- $\hat{Y}(X,A) \in \{0,1\}$ = the assigned class label
  - usually assigned by a threshold rule: $\hat{Y} = 1$ if $\hat{y} \geq \bar{y}$, $\hat{Y} = 0$ if $\hat{y} < \bar{y}$, for some $\bar{y} \in (0,1)$.
  - if $\hat{y}(\cdot)$ is well-calibrated, would typically set $\bar{y} = 0.5$.

Statistical Fairness Criteria

Based on Berk et al (2017) and Barocas et al (2021):

1. Equalizing outcomes across groups (statistical parity / independence)
2. Equalizing recall across groups (separation)
3. Equalizing precision across groups (calibration / sufficiency)

1. Equalizing Outcomes Across Groups

Statistical Parity

Average predicted outcome ($\frac{\#\text{ predicted positive}}{\text{sample size}}$) should be the same across groups.

$$\Pr(\hat{Y} = 1 | A = a) = \Pr(\hat{Y} = 1 | A = b)$$

- also called "demographic parity" or "disparate impact". This is probably the most commonly used fairness metric.

1. Equalizing Outcomes Across Groups

Statistical Parity

Average predicted outcome ($\frac{\#\text{ predicted positive}}{\text{sample size}}$) should be the same across groups.

$$\Pr(\hat{Y} = 1 | A = a) = \Pr(\hat{Y} = 1 | A = b)$$

- also called "demographic parity" or "disparate impact". This is probably the most commonly used fairness metric.
- Pros:
  - simple and intuitive
  - sometimes legally required (e.g. EEOC's four-fifths rule)
- Cons:
  - enforcing statistical parity tends to reduce accuracy, especially when the true label varies across groups (different base rates).
  - e.g.: if decision to grant bail is based on $\hat{Y}$, can lead to perhaps undesirable outcomes, such as imprisoning a lot more women who are not risky.

- In practice, achieving equal outcomes could be too restrictive.
- Instead, could impose a slack condition:

$$|\Pr(\hat{Y} = 1|A = a) - \Pr(\hat{Y} = 1|A = b)| \leq \epsilon$$

  - where, e.g. ϵ could be set to satisfy the "four-fifths rule" from disparate impact law.

1. Equalizing Outcomes Across Groups

Relaxed Statistical Parity and Independence

- In practice, achieving equal outcomes could be too restrictive.
- Instead, could impose a slack condition:

$$|\Pr(\hat{Y} = 1|A = a) - \Pr(\hat{Y} = 1|A = b)| \leq \epsilon$$

  - where, e.g. ϵ could be set to satisfy the "four-fifths rule" from disparate impact law.
- "Independence" (Barocas et al 2021)
  - a stronger criterion that implies statistical parity
  - requires independence of the **score** and the protected attribute: $\hat{y} \perp A$.
  - If $I(z,x)$ is mutual information between z and x, equivalent to requiring $I(A;\hat{y}) = 0$ or $I(A;\hat{y}) \leq \epsilon$.

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.
- Another set of fairness criteria penalize divergence in error metrics across groups, conditional on the true label.
  - e.g. accuracy, recall, FPR, FNR
  - allows for different treatment of groups if justified by variation in base rates

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.
- Another set of fairness criteria penalize divergence in error metrics across groups, conditional on the true label.
  - e.g. accuracy, recall, FPR, FNR
  - allows for different treatment of groups if justified by variation in base rates
  - e.g. men and women should have same model accuracy

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.
- Another set of fairness criteria penalize divergence in error metrics across groups, conditional on the true label.
  - e.g. accuracy, recall, FPR, FNR
  - allows for different treatment of groups if justified by variation in base rates
  - e.g. men and women should have same model accuracy
- Can also combine multiple criteria:
  - e.g., the ratio of false positives to false negatives should be the same for men and women.

2. Equalizing Recall Across Groups

Separation

Barocas et al (2021) discuss the more general criteria, "**separation**":

- requires $\hat{y} \perp A|Y$: that is, the score is independent of the sensitive attribute, conditional on the true label.
- In the binary case, equivalent to equalizing **both** true positive rates (recall for positive class) **and** false positive rates across groups.
- can also be achieved subject to a slack condition.

3. Equalizing Precision Across Groups

Definition

- A third set of metrics requires equalizing precision across groups
  - precision for both positive and negative outcomes
  - i.e. positive/negative predictive value
  - also called “predictive parity”

3. Equalizing Precision Across Groups

Definition

- A third set of metrics requires equalizing precision across groups
  - precision for both positive and negative outcomes
  - i.e. positive/negative predictive value
  - also called "predictive parity"
- Barocas et al (2021) call this "**sufficiency**" and formalize it as

$$\Pr(Y = 1 | \hat{Y}, A = a) = \Pr(Y = 1 | \hat{Y}, A = b)$$

  - that is, conditioning on the score, both groups get the same label.

3. Equalizing Precision Across Groups

Calibration

- An intuitive way to achieve sufficiency (equalizing precision across groups) is to require that the classifier is **well-calibrated** for each group.
- that is,

$$\hat{y}(X, A) = \Pr(Y = 1|X), \forall A$$

the scores provide the probability that the true label equals one, for all groups.

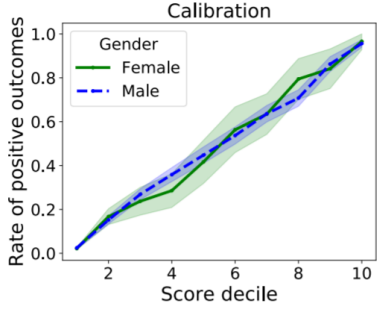
3. Equalizing Precision Across Groups

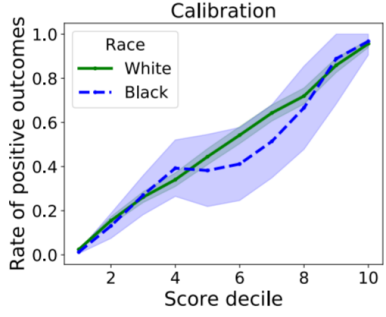
Calibration

- An intuitive way to achieve sufficiency (equalizing precision across groups) is to require that the classifier is **well-calibrated** for each group.
- that is,

$$\hat{y}(X, A) = \Pr(Y = 1|X), \forall A$$

the scores provide the probability that the true label equals one, for all groups.





What notions of fairness does this classifier satisfy?

Group A			
	$\hat{Y} = 1$	$\hat{Y} = 0$	
$Y = 1$	30	20	TPR=.6
$Y = 0$	20	20	TNR=.5
	PPV = .6	NPV = .5	
avg $\hat{Y}$ = .55, acc. = .55, FP/FN = 1			

Group B			
	$\hat{Y} = 1$	$\hat{Y} = 0$	
$Y = 1$	60	40	TPR=.6
$Y = 0$	60	60	TNR=.5
	PPV = .5	NPV = .4	
avg $\hat{Y}$ = .55, acc. = .55, FP/FN = 1.5			

1. Equality of outcomes (statistical parity / independece)
2. Equality of recall (separation)
3. Equality of precision (sufficiency)

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

1. Equality of outcomes (statistical parity / independece)
2. Equality of recall (separation)
3. Equality of precision (sufficiency)

Except in highly artificial datasets, Criteria (1), (2), and (3) are all mutually incompatible with each other!

(1) Statistical Parity and (2)/(3) Equal Recall/Precision are incompatible

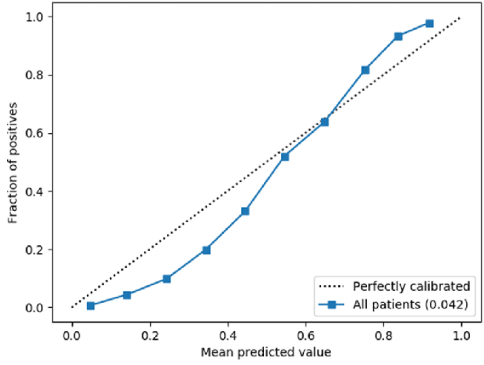
- If the outcome Y varies by group status A, a classifier achieving statistical parity means that average $\hat{Y}$ does not equal average Y for at least one of the groups.
  - that is, if statistical parity is imposed, there will be differences in both recall (error rates conditional on true label) and precision (error rates conditional on predicted label) across groups.
- Hence, satisfying (1) precludes satisfying (2) or (3) except in the special case of identical base rates across groups.

(2) Equal Recall and (3) Equal Precision are incompatible

- If base rates differ by group, these requirements cannot hold simultaneously:
  - error rate balance (equality of FPR/FNR across groups)
  - predictive parity (equality of PPV/NPV across groups)
- try to draw a confusion matrix that satisfies it.
- this is often called the precision-recall tradeoff.

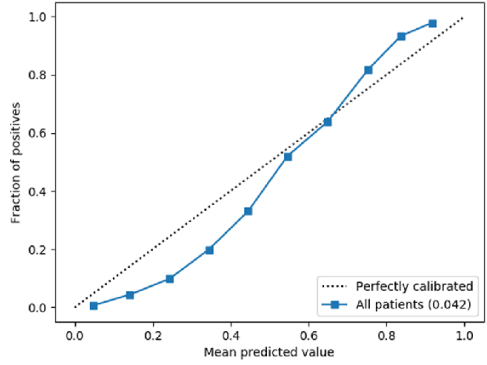
(2) Equal Recall and (3) Calibration are incompatible

- recall that in a well-calibrated model, we can bin observations by their predicted outcome probabilities, and the outcome rates should roughly match in those bins.
- good calibration requires equalizing false positive and false negative rates.



(2) Equal Recall and (3) Calibration are incompatible

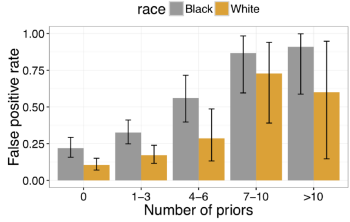
- recall that in a well-calibrated model, we can bin observations by their predicted outcome probabilities, and the outcome rates should roughly match in those bins.
- good calibration requires equalizing false positive and false negative rates.



Trade-off: If base rates differ by group, error rate balance (equality of FPR/FNR across groups) precludes calibration.

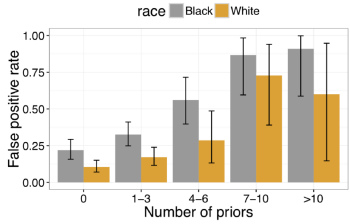
Example: COMPAS recidivism risk prediction

FPR is higher for black defendants! (Chouldechova'17):

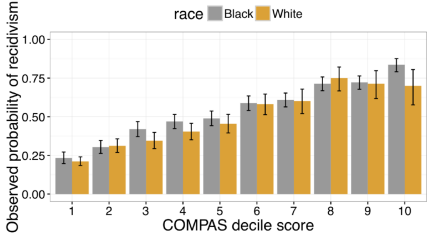


Example: COMPAS recidivism risk prediction

FPR is higher for black defendants! (Chouldechova'17):



But the scores are well-calibrated (or PPV similar across all groups)! (Chouldechova'17):



COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

But:

- also asked human annotators to produce recidivism predictions, and race info was not provided.
- humans were almost identically biased.

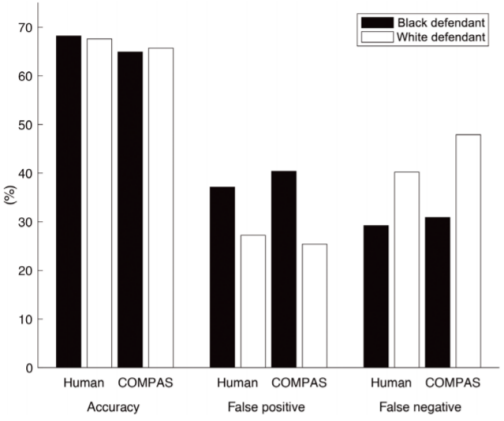


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

But:

- also asked human annotators to produce recidivism predictions, and race info was not provided.
- humans were almost identically biased.

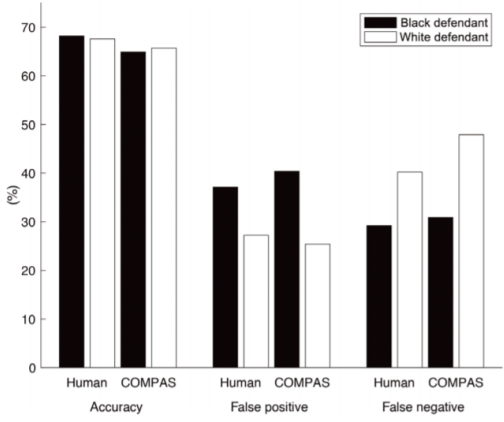


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

- giving the human annotators information on the race of the defendant made no difference.

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

How to make ML-Based Decisions fair?

- So far, our metrics can be used to assess the fairness of classifiers and the resulting decisions.
- What if our decisions is biased? What do we do?

How to make ML-Based Decisions fair?

- So far, our metrics can be used to assess the fairness of classifiers and the resulting decisions.
- What if our decisions is biased? What do we do?
- There are three groups of approaches:
  - **Pre-processing**: Adjust the feature space to be uncorrelated with the sensitive attribute.
  - **At training time**: Work the constraint into the optimization process that constructs a classifier from training data.
  - **Post-processing**: Adjust a learned classifier so as to be uncorrelated with the sensitive attribute.

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

Post-Processing with the Score

- Given a score function $\hat{y}(\cdot)$ and a cost for false negatives and false positives, find the derived classifier that minimizes the expected cost of false positive and false negatives subject to the fairness constraint at hand.
  - can depend on the sensitive attribute
  - can add randomness
- Advantages:
  - simple and transparent
  - works for any black-box classifier regardless of its inner workings.
  - no need for re-training models
- Disadvantage:
  - requires and uses the protected attribute.

Achieving Fairness with Post-Processing

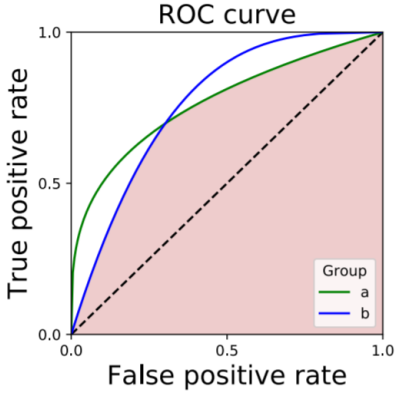
- Statistical Parity:
  - Just set the thresholds for each group k such that the average $\hat{Y}$ is the same.

Achieving Fairness with Post-Processing

- Statistical Parity:
  - Just set the thresholds for each group k such that the average $\hat{Y}$ is the same.
- Calibration:
  - just calibrate the classifier separately by group.

Achieving Fairness with Post-Processing

- Statistical Parity:
  - Just set the thresholds for each group k such that the average $\hat{Y}$ is the same.
- Calibration:
  - just calibrate the classifier separately by group.
- Separation (equality of true positive rates and false positive rates):



- In the binary case, a classifier satisfying separation is limited to the region in red.
- Set separate group thresholds and randomize across multiple classifiers to equalize the rates.

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.

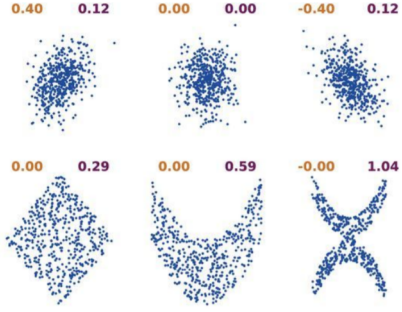
- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance
- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance
- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.
  - Then $\text{corr}(\tilde{X}_j, A) = 0$ by construction.

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance
- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.
  - Then $\text{corr}(\tilde{X}_j, A) = 0$ by construction.

Problem: Uncorrelated $\neq$ Independent (e.g. Ince et al 2016)



- relations could be non-linear
- could be interactions between predictors, $X_j X_k$, $j \neq k$, correlated with A.
- X_j and A could have an interaction effect on Y.

correlation $\neq$ mutual information

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

- See Zemel et al (2013), "Learning fair representations" and follow-up papers for sophisticated approach to this problem.

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

- See Zemel et al (2013), "Learning fair representations" and follow-up papers for sophisticated approach to this problem.
- Seemingly unrecognized problem: unobserved confounders relating A to X and Y.

Wang et al (adversarial de-biasing approach using gender and images)

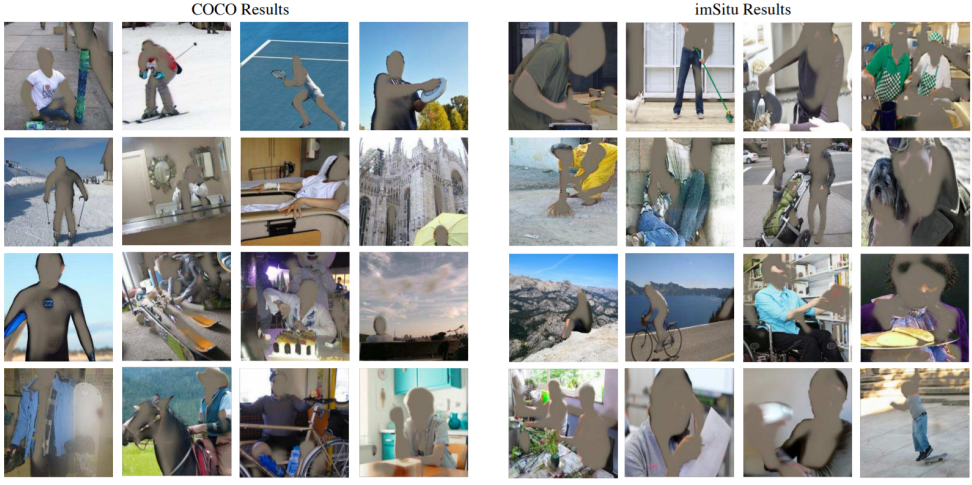


Figure 6. Images after adversarial removal of gender in image space by using a U-Net based autoencoder as inputs to the recognition model. While people are clearly being obscured from the image, the model selectively chooses to obscure only parts that would reveal gender such as faces but tries to keep information that is useful to recognize objects or verbs. 1st row: WWWM MMWW; 2nd row: MWWW WMWW; 3rd row: MMMW MMWM; 4th row: MMMW WWMM. W: woman; M: man.

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

Constrained optimization

Minimize model loss, subject to the expected outcome being similar across groups (beneath some threshold ϵ):

$$\min_{\theta} L(\theta)$$
$$\text{s.t.} |\mathbb{E}(\hat{Y}|\text{white}) - \mathbb{E}(\hat{Y}|\text{black})| \leq \epsilon$$

Constrained optimization

Minimize model loss, subject to the expected outcome being similar across groups (beneath some threshold ϵ):

$$\min_{\theta} L(\theta)$$
$$\text{s.t.} |\mathbb{E}(\hat{Y}|\text{white}) - \mathbb{E}(\hat{Y}|\text{black})| \leq \epsilon$$

- Reductions Approach (Agarwal 2018): solve a series of cost-sensitive classification problems using off-the-shelf methods.
  - also works for error rate balance (but not predictive parity)

Constrained optimization

Minimize model loss, subject to the expected outcome being similar across groups (beneath some threshold ϵ):

$$\min_{\theta} L(\theta)$$
$$\text{s.t.} |\mathbb{E}(\hat{Y}|\text{white}) - \mathbb{E}(\hat{Y}|\text{black})| \leq \epsilon$$

- Reductions Approach (Agarwal 2018): solve a series of cost-sensitive classification problems using off-the-shelf methods.
  - also works for error rate balance (but not predictive parity)
- Many options like this with neural nets – multi-task learning, adversarial models, gradient reversal, etc.
  - in general, there appear to be many approaches and no consensus on the best approach yet.

Outline

Recap: Using Machine Learning to Guide Audit Policy

Fairness, Bias, and Discrimination

Statistical Fairness
- Evaluating Classifier Fairness
- Fairness Criteria are Incompatible

Adjusting ML Decisions to Improve Fairness
- Post-Processing with the Score Function
- Pre-Processing the Data
- Constraining Classifiers at Training Time

Problems with Algorithmic Fairness

Fairness Metrics are Fragile (Li, Goel, Ash 2022)

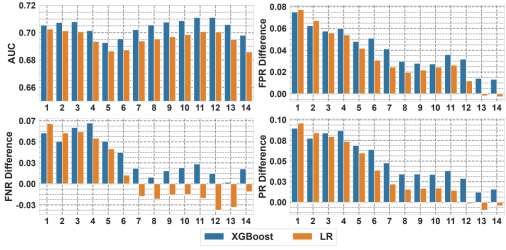


Figure 4: The Role of Time: X axis corresponds to training datasets from two consecutive years between 2000 and 2018 (e.g., "1" on x axis denotes the training data from the years 2000 & 2001, "2" denotes the training data from the years 2001 & 2002 and so on. The test data comes from the next two years after a two year gap (e.g. if training data is from 2000 & 2001, test comes from 2003 & 2004.)

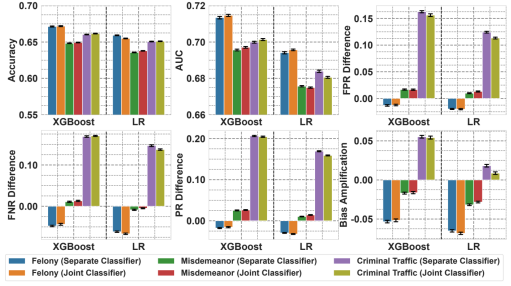


Figure 7: Offense Type Separated Classifiers: A separate classifier for each offense type is trained on data from that offense type. The performance of the classifiers are then observed on respective offense types. For comparison, the performance of a joint classifier, that is trained on all the data and uses offense type as a predictor, is also shown by offense type. 95% confidence intervals are constructed by multiple train and test splits.

Unobserved confounders

Unobserved confounders

- There is a deeper problem with any of these approaches:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.

Unobserved confounders

- There is a deeper problem with any of these approaches:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
- Counterfactual fairness (e.g. Kusner et al 2018):
  - "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."

Unobserved confounders

- There is a deeper problem with any of these approaches:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
- Counterfactual fairness (e.g. Kusner et al 2018):
  - "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."
  - e.g., had a defendant been from a different race, he would have had different education, different residence location, etc.

Unobserved confounders

- There is a deeper problem with any of these approaches:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
- Counterfactual fairness (e.g. Kusner et al 2018):
  - "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."
  - e.g., had a defendant been from a different race, he would have had different education, different residence location, etc.
  - Kusner et al (2018) and following papers:
    - build a structural causal model to predict education and other intermediate outcomes based on sensitive attribute
    - flip the attribute to find counterfactual
    - fairness requires no difference in outcome.

Unobserved confounders

- There is a deeper problem with any of these approaches:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
- Counterfactual fairness (e.g. Kusner et al 2018):
  - "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."
  - e.g., had a defendant been from a different race, he would have had different education, different residence location, etc.
  - Kusner et al (2018) and following papers:
    - build a structural causal model to predict education and other intermediate outcomes based on sensitive attribute
    - flip the attribute to find counterfactual
    - fairness requires no difference in outcome.
  - More of a proof of concept, because too many strong assumptions to be practically relevant.

Lightning Essay

For last minutes of class:

`https://bit.ly/BRJ-W9-Essay`