Machine Learning Bias

Nick Eubank

Outline

1. How Bias Sneaks In (Biased Factors)

Outline

1. How Bias Sneaks In (Biased Factors)
2. How Bias Sneaks In (Biased Measures)

Outline

1. How Bias Sneaks In (Biased Factors)
2. How Bias Sneaks In (Biased Measures)
3. Ethics of Type 1 versus Type 2 Errors

Outline

Supervised Machine Learning

Conceptual that because Supervised Machine Learning (SML) is built on math, it can't be biased.

Supervised Machine Learning

Conceptual that because Supervised Machine Learning (SML) is built on math, it can’t be biased.

When bias creeps in, it is assumed to be because of researcher negligence.

Supervised Machine Learning

Conceptual that because Supervised Machine Learning (SML) is built on math, it can’t be biased.

When bias creeps in, it is assumed to be because of researcher negligence.

By design, SML will always try and be biased.

Supervised Machine Learning

SML models are designed to find any patterns they can to help predict outcomes / classify records.

Because sexism, racism, xenophobia, homophobia, etc. shape outcomes in the world,

$\rightsquigarrow$ SMLs generally perform better when they are sexist/racist/xenophobic/homophobic!

Supervised Machine Learning

We're building SML model designed to predict performance reviews using resumes.

Supervised Machine Learning

We’re building SML model designed to predict performance reviews using resumes.
Train using resumes and performance evaluations of current employees.

Supervised Machine Learning

We're building SML model designed to predict performance reviews using resumes.
Train using resumes and performance evaluations of current employees.

- Help decide who to hire.

Supervised Machine Learning

We're building SML model designed to predict performance reviews using resumes.
Train using resumes and performance evaluations of current employees.

- Help decide who to hire.

If supervisors tend to discriminate against women, then our SML will look for signals that an applicant is a woman, since they can use this to give women lower reviews, better matching the training data.

Proxies

OK, but what if I don’t include data on gender, race, sexuality, etc. in my model?

Everything in society is correlated:

Proxies

OK, but what if I don't include data on gender, race, sexuality, etc. in my model?

Everything in society is correlated:

- Going to a women's college (Scripps College, Barnard College)

Proxies

OK, but what if I don't include data on gender, race, sexuality, etc. in my model?

Everything in society is correlated:

- Going to a women's college (Scripps College, Barnard College)
- Going to a Historical Black University (Howard University)

Proxies

OK, but what if I don't include data on gender, race, sexuality, etc. in my model?

Everything in society is correlated:

- Going to a women's college (Scripps College, Barnard College)
- Going to a Historical Black University (Howard University)
- Many activities are gender-correlated (Yoga, Football)

Proxies

OK, but what if I don't include data on gender, race, sexuality, etc. in my model?

Everything in society is correlated:

- Going to a women's college (Scripps College, Barnard College)
- Going to a Historical Black University (Howard University)
- Many activities are gender-correlated (Yoga, Football)
- Geography is *extremely* correlated with race and income (Princeton Review)

Proxies

OK, but what if I don't include data on gender, race, sexuality, etc. in my model?

Everything in society is correlated:

- Going to a women's college (Scripps College, Barnard College)
- Going to a Historical Black University (Howard University)
- Many activities are gender-correlated (Yoga, Football)
- Geography is *extremely* correlated with race and income (Princeton Review)

In COMPAS, race wasn't in the model.

Biased By Design

Bias in Machine Learning isn’t the result of negligence.

Biased By Design

Bias in Machine Learning isn't the result of negligence.

So long as society has biases, Machine Learning has an affirmative incentive to be biased too!

What Can I Do To Prevent Bias?

1. Target an unbiased outcome.

What Can I Do To Prevent Bias?

1. **Target an unbiased outcome.**
    - In hiring example, variables correlated with gender created bias because the target (performance evaluations) were biased!

What Can I Do To Prevent Bias?

1. **Target an unbiased outcome.**
    - In hiring example, variables correlated with gender created bias because the target (performance evaluations) were biased!

Less biased targets will reduce the incentive for your algorithm to be biased.

What Can I Do To Prevent Bias?

1. **Target an unbiased outcome.**
    - In hiring example, variables correlated with gender created bias because the target (performance evaluations) were biased!

Less biased targets will reduce the incentive for your algorithm to be biased.
Picking unbiased outcomes is not as easy as it seems...

What Can I Do To Prevent Bias?

COMPAS: Predicted probability of future arrest.

What Can I Do To Prevent Bias?

COMPAS: Predicted probability of future arrest.

Arrests are pretty objective, right?

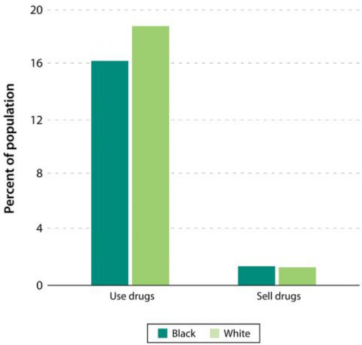
FIGURE 6A.

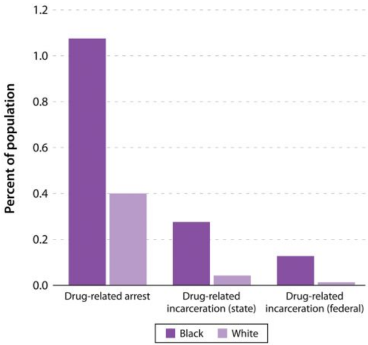
Rates of Drug Use and Sales, by Race

FIGURE 6B.

Rates of Drug-Related Criminal Justice Measures, by Race

At the state level, blacks are about 6.5 times as likely as whites to be incarcerated for drug-related crimes.





Source: BLS n.d.c; Carson 2015; Census Bureau n.d.; FBI 2015; authors' calculations.

THE HAMILTON PROJECT
BROOKINGS

Probability of arrest $\neq$ probability of committing a crime

What Can I Do To Prevent Bias?

2. Test your models *thoroughly.*

COMPAS:

- Same accuracy scores for Black and White suspects!

What Can I Do To Prevent Bias?

2. Test your models *thoroughly.*

COMPAS:

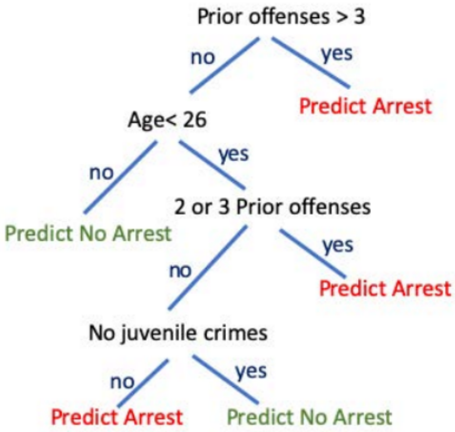
- Same accuracy scores for Black and White suspects!
- But... very different rates of false positives and false negatives.

What Can I Do To Prevent Bias?

3. Try to Use Interpretable Models

What Can I Do To Prevent Bias?

3. Try to Use Interpretable Models



An interpretable decision tree to predict whether an individual will be arrested in the future. Hu et al. NeurIPS 2019