Auditing Black Box Algorithms for Fairness and Bias

Daniel B. Neill, Ph.D.
Carnegie Mellon University (Heinz College)
and NYU (Center for Urban Science & Progress)

E-mail: neill@cs.cmu.edu / daniel.neill@nyu.edu

Joint work with Zhe Zhang (CMU Heinz College)

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals.
And it's biased against blacks.

Source:
Julia Angwin,
Jeff Larson,
Surya Mattu and
Lauren Kirchner, *ProPublica*

Motivating questions for this work

- Is the COMPAS algorithm for predicting re-offending risk **fair**, or is it **biased** against some subpopulation defined by observed characteristics?
  - **Black box** algorithm. All we observe is predictions vs. gold standard (re-offending) for a sample of individuals (ProPublica data from Broward County, FL).
  - Many possible biases: race, gender, age, past offenses…
  - Combinations of factors, e.g., “elderly white females”
- This led us to develop a **general approach** to auditing black box algorithms for fairness or bias.

Broward County data

- Source: ProPublica's data on criminal defendants in Broward County, FL, in 2013-2014
- Outcome: re-arrests (!) assessed through April 2016.
- Score: **COMPAS** score from 1 (low risk) to 10 (high risk)

Background	Black $(n = 3696)$		White $(n = 2454)$
Age	32.7 (10.9)	<	37.7 (12.8)
Male (%)	82.4	>	76.9
Number of Priors	4.44 (5.58)	>	2.59 (3.8)
Any priors? (%)	76.4	>	65.9
Felony (%)	68.9	>	60.3
COMPAS Score	5.37 (2.83)	>	3.74 (2.6)

Slide credit: Alexandra Chouldechova

What does it mean to be "fair"?

There are at least three possibilities (and probably more):

1) Group Fairness: The same proportion of each group should be classified as "high risk".

- Doesn't seem reasonable for COMPAS: observed reoffending rates are not constant across groups. For Broward County, 51% of black defendants and 39% of white defendants reoffended.
- Doesn't handle real-valued predictions (% chance of reoffending).
- Not easily generalizable to evaluating (un)biasedness across many different features and combinations of features.

2) **Disparate Impacts**: Comparing false positive and false negative rates across groups. (Good idea: see Alex's talk!)

- Impacts depend on how predictions are used (particularly if the prediction is a probability). Can we separate **fairness of prediction** from **fair decisions** using these predictions?

What does it mean to be “fair”?

There are at least th ... re):

1) G

sh

3) We focus on **unbiasedness** of probability estimates.

Individual risk probabilities should be predicted accurately, **without systematic biases** based on any observed attributes or combinations of attributes.

→ Are there any **statistically significant** biases?

→ Can we automatically **correct** these systematic biases, in order to improve fairness of prediction?

2) **Dis** ... d false negative ra ... s g ... dea: see Alex’s talk!)

– Impacts depend on how ... are used (particularly if the prediction is a probability). Can we separate **fairness of prediction** from **fair decisions** using these predictions?

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.

Assume a dataset with inputs x_i, binary labels $y_i \in \{0,1\}$, and the classifier's risk predictions $\widehat{p_i} = \Pr(y_i = 1)$.

Search space: subspaces defined by a subset of values for each attribute (e.g., "white and Asian males under 25")

Score function: a log-likelihood ratio statistic. H_0: $\widehat{p_i}$ correctly calibrated; $H_1(S)$: constant multiplicative increase or decrease in odds of $y_i = 1$ for subspace S.

$$F(S) = \max_q \log \prod_{s_i \in S} \frac{\Pr\left(y_i \sim Bernoulli\left(\frac{q\widehat{p_i}}{1 - \widehat{p_i} + q\widehat{p_i}}\right)\right)}{\Pr\left(y_i \sim Bernoulli(\widehat{p_i})\right)}$$

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.

1. Start with randomly chosen **subsets of values** V_j for each attribute. Let subspace S = Cartesian product of V_j.

2. Choose an attribute (randomly or sequentially) and find the highest scoring subset of values, conditioned on all other attributes. Update S.

3. Iterate step 2 until convergence to a local maximum of the score function F(S), and use multiple restarts to approach the global maximum.

<u>Key idea</u>: for a given optimization step, the **linear-time subset scanning** property (Neill, 2012) can be used to exactly identify the highest scoring subset of attribute values, evaluating $O(|V|)$ subsets instead of $O(2^{|V|})$.

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.

1. Start with randomly chosen **subsets of values** V_j for each attribute. Let subspace S = Cartesian product of V_j.

2. Choose an attribute (randomly or sequentially) and find the highest scoring subset of values, conditioned on all other attributes. Update S.

3. Iterate step 2 until convergence to a local maximum of the score function F(S), and use multiple restarts to approach the global maximum.

For interpretability, we maximize the penalized score $F(S) - \log \prod |S_j|$, where attributes with no excluded values are ignored. For each conditional optimization, we can use the simple penalty, $\log(|S_j|)\, 1\{|S_j| < \text{arity}(A_j)\}$.

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.

To determine whether the highest-scoring (most biased) subset is significant, we compare its score to the **maximum** subset scores of a large number of replica datasets generated under the null hypothesis of “no bias”.

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.

Real: $x_i \quad \hat{p}_i \quad y_i$

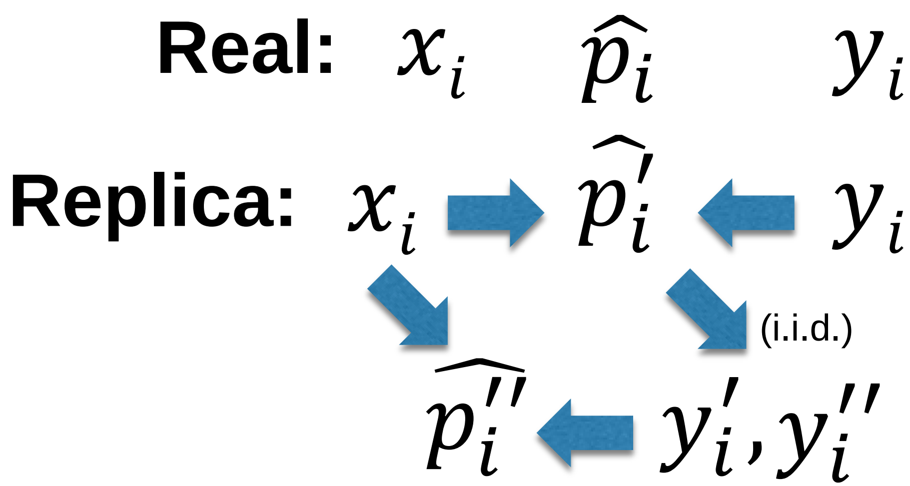
Replica: $x_i \quad \hat{p}_i \rightarrow y_i'$

The simplest randomization approach, the "sharp null", redraws all y_i from Bernoulli($\hat{p}_i$). But this null hypothesis does not account for the expected variance in $\hat{p}_i$ caused by learning model parameters from training data.

Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.

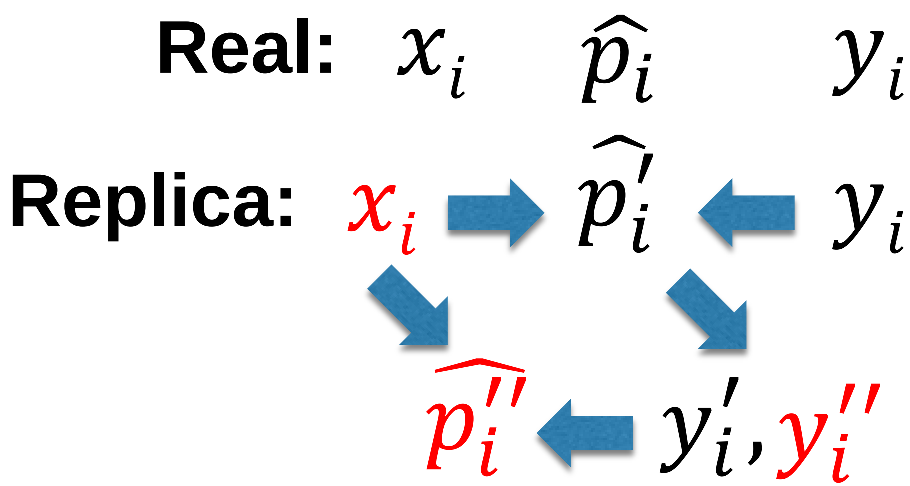


We have developed an alternative randomization approach, the “split null”, that compares the observed maximum bias to the bias one would expect from learning a correctly specified model, resulting in fewer false positives.

Bias scan (Zhang and Neill, 2016)

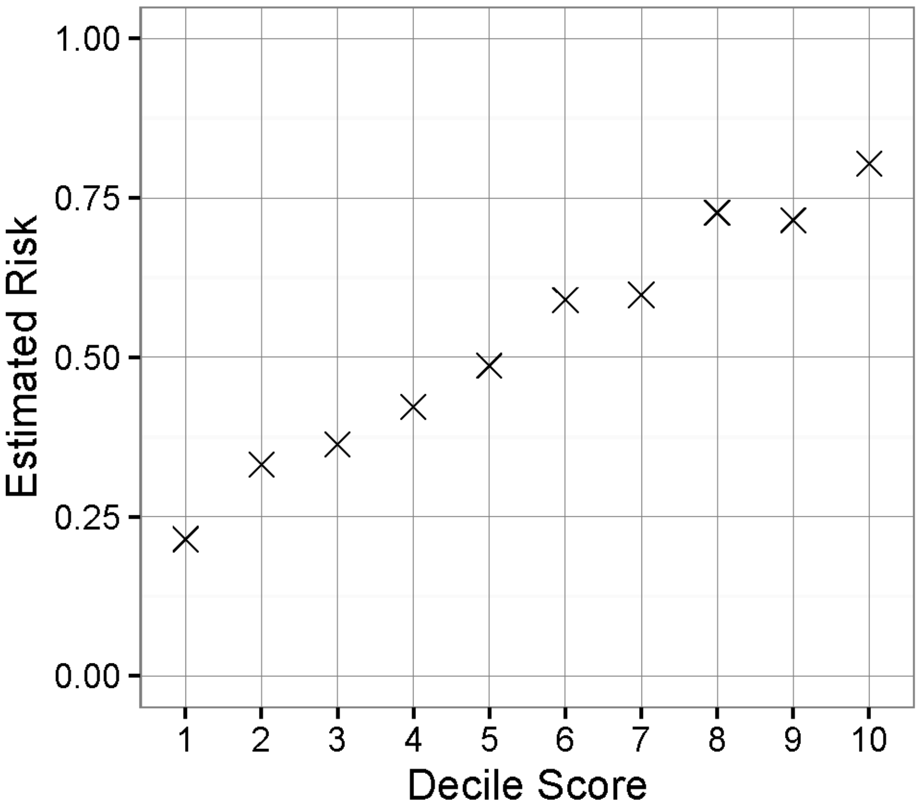
Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the **multidimensional subset scan** to identify subgroups where classifier predictions are significantly biased.



We have developed an alternative randomization approach, the “split null”, that compares the observed maximum bias to the bias one would expect from learning a correctly specified model, resulting in fewer false positives.

Results of bias scan on COMPAS

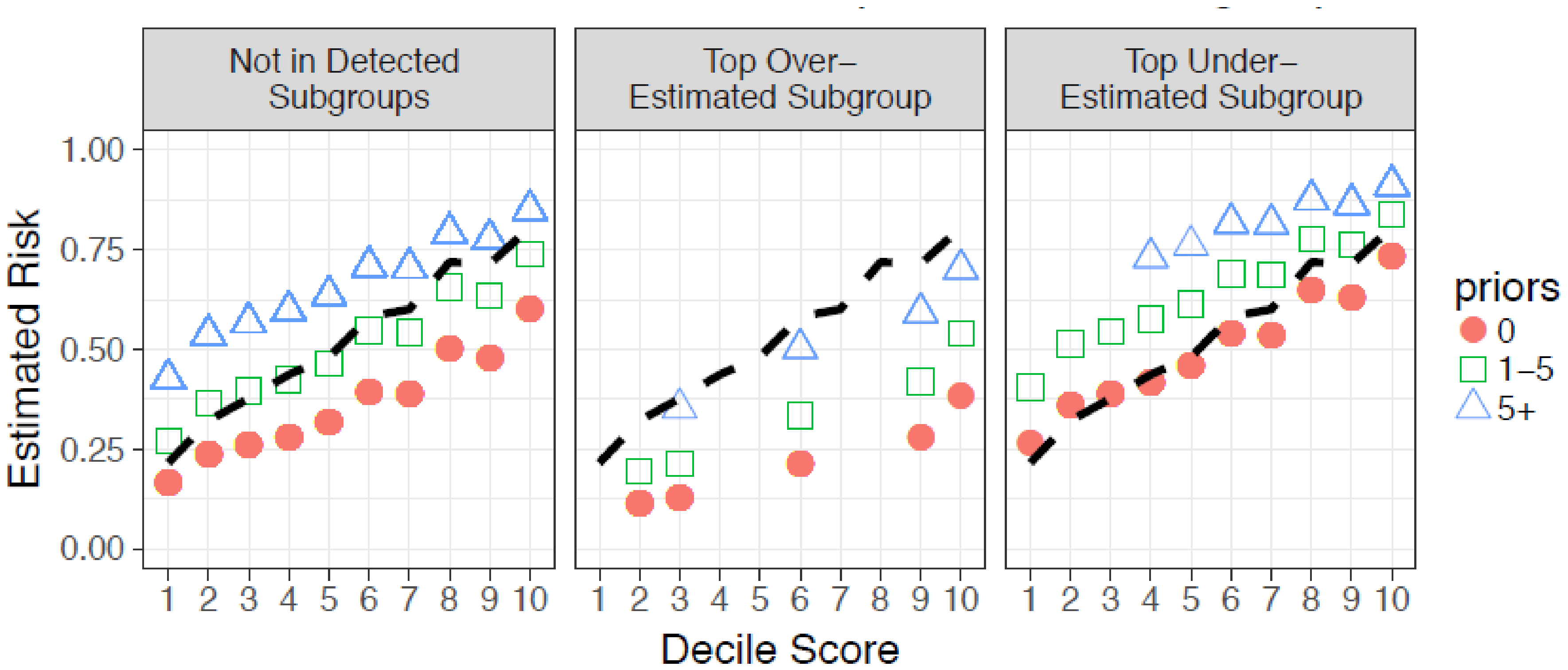


Start with maximum likelihood risk estimates for each COMPAS decile score.

Detection result 1: COMPAS underestimates the importance of prior offenses, overestimating risk for 0 priors, and underestimating risk for 5 or more priors.

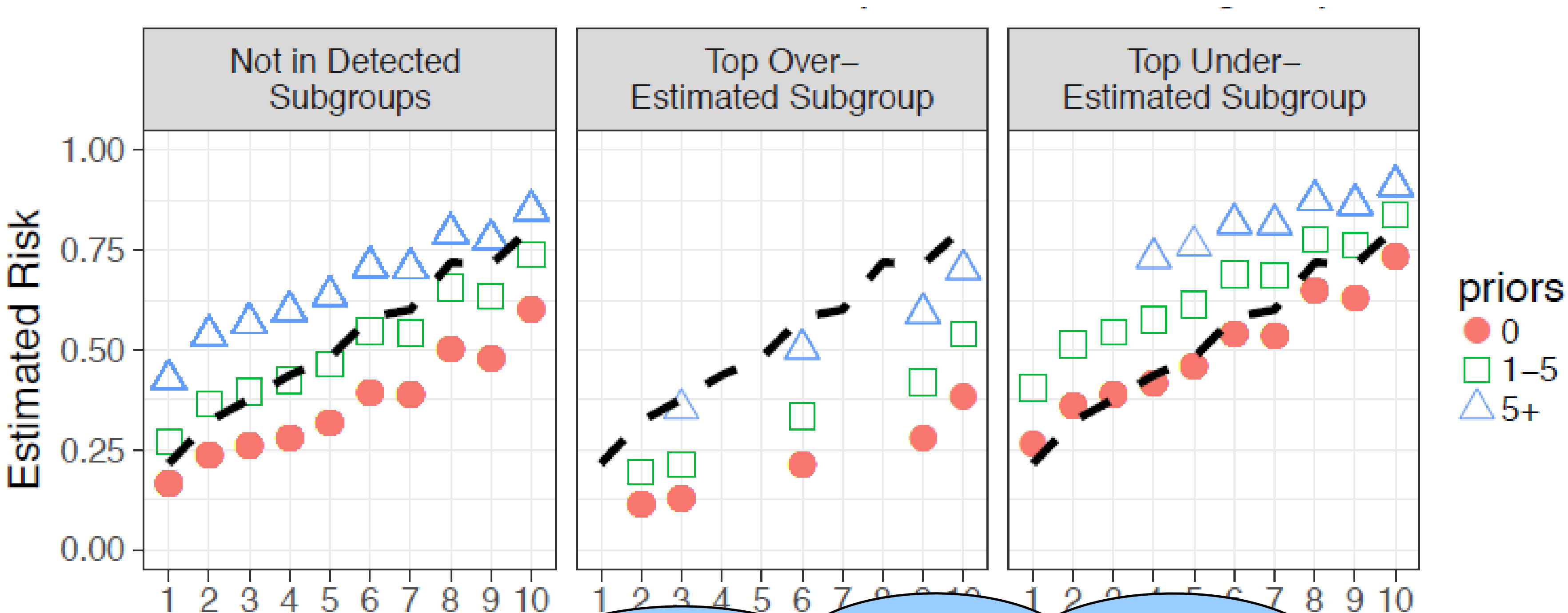
Detection result 2: Even controlling for prior offenses, COMPAS still underestimates risk for males under 25, and overestimates risk for females who committed misdemeanors.

Results of bias scan on COMPAS



After controlling for number of prior offenses and for membership in the two detected subgroups, there are no significant systematic biases in prediction.

Results of bias scan on COMPAS



The resulting probabilistic classifier has greater interpretability (though still based partially on a black box) and is less biased than the original COMPAS predictions… but does this mean it is "fair"?

Discussion: predictive fairness in context

- The method does not account for **target variable bias**: we predict re-offending risk but the gold standard is based on re-arrests not re-offenses.
  - Big problem with drug possession, weapon possession charges. Leads to feedback loops.
- How to avoid **disparate impacts** when making decisions based on even unbiased predictions? (60/40 example)
  - Integration with other data sources? Probability matching?

The resulting probabilistic classifier has greater interpretability (though still based partially on a black box) and is less biased than the original COMPAS predictions… but does this mean it is "fair"?

Another application example

- We also apply bias scan to a loan delinquency prediction dataset, “Give Me Some Credit”, provided by Kaggle.
- We start with a simple predictive model (lasso: L1-penalized logistic regression, penalty chosen by cross-validation) and compare the predictions to ground truth.
- Bias scan identifies an interesting group whose delinquency risk is significantly over-estimated ($p < 0.01$):
  - Consumers who are above the median in credit utilization and have at least 1 occurrence of a late payment of 30-59, 60-89, and 90+ days late → observed rate 79%, predicted rate 90%.
- This group is only 1.7% of the total dataset, but makes up 95% of the 496 consumers judged as the “riskiest 1%”.
- After adjusting the model to account for this bias, this group would only make up 58% of “riskiest” consumers.

References

- **Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. https://arxiv.org/abs/1611.08292.**
  - Version 1: NIPS 2016 Workshop on Interpretable Machine Learning.
  - Version 2: Presented at Fairness, Accountability, and Transparency (FAT/ML 2017).
- D.B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.

Thanks for listening!

More details on our web site:
http://epdlab.heinz.cmu.edu

Or e-mail me at:
neill@cs.cmu.edu