Week 9: *Can Criminal Risk Scores Balance Bias?*

MEMORANDUM

To: The Mayor
From: Terra R. Edenhart-Pepe
Subject: Recidivism

Primary question: Should we use decision-making algorithms to guide rehabilitation resource allocation decisions?

The first step in determining the primary question's answer, is to ask the right secondary questions, like is the algorithm both effective and fair?

Northpointe co-founder, Tim Brennan, says "the rate of accuracy for COMPAS scores… was the same for black and white defendants," or racially neutral (Angwin; Larson, 2016).

But some metrics don't represent the nuance implicit in social science research. Thus, it's essential to evaluate a given model with a wide variety of metrics and thorough definitions.

What is fairness?

The notion of "fairness" is an ethical concept with a diverse array of social impacts. While a predictive tool may have minimal bias, the use and application of the tool still could cause disproportionate adverse impact on distinct racial groups, an alarming idea when considering that, scores are increasingly combined with rehabilitation needs evaluations and courtroom decision-making (Angwin, et. al., 2016; Chouldechova, 2016).

Interpretation

The plot in figure 3 shows both models demonstrate "fair - good" predictive power and that the model without race is more only slightly more useful, but the difference between AUC values of .73 and .72 is not statistically significant. This metric is course; to get a closer look, predictions by race type (and filtering for 3 races) are plotted as an AUC curve (fig. 4 & 5). The model without race predicts well 73 out of 100 times, and there is minimal statistical difference between predictions for different race groups, which supports Northpointe's claim. An additional insight made apparent by the ROC curve and AUC values differentiating for race (fig. 4 & 5) is that there is negligible benefit to including race. The race variable could be removed without drastically impacting the model.

EXPLORATORY ANALYSIS

Frequency of Recivism by Race

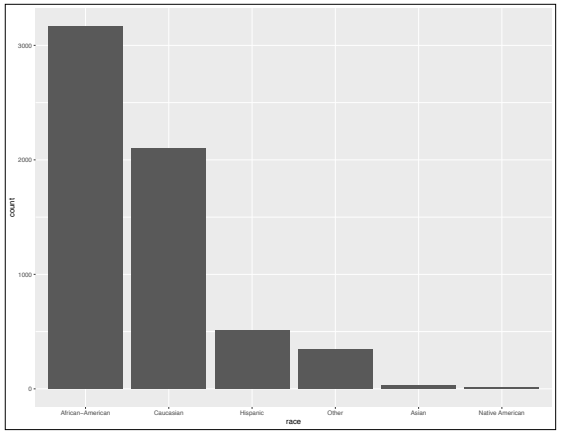

figure 1

This Frequency tables shows how often a given value occurs.

Rate of Recidivism Versus No-reoffense by Race

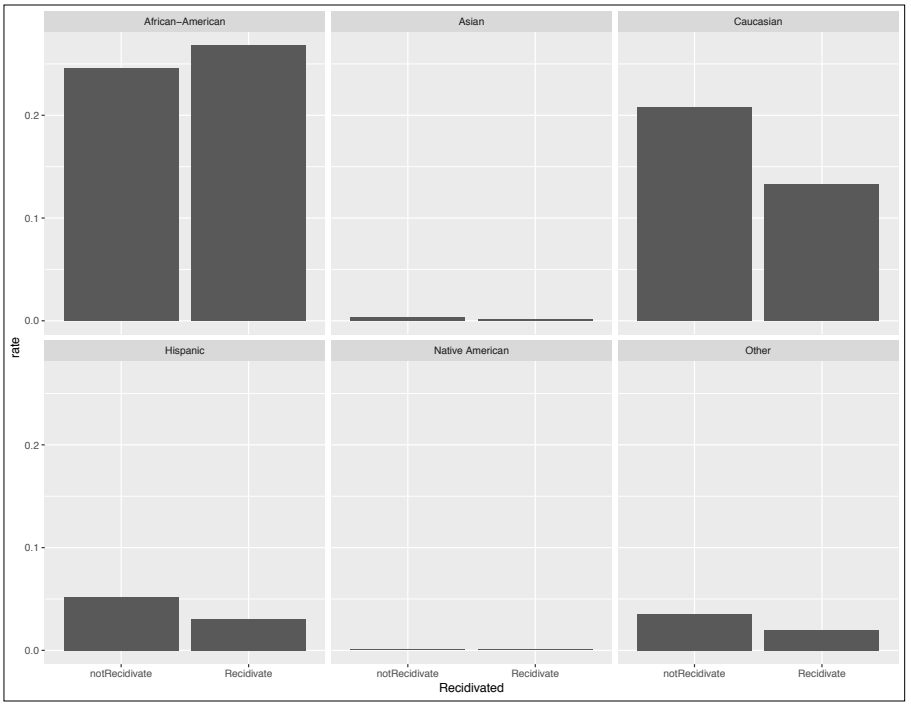

figure 2

This plot shows that rates of recidivism across different racial groups are not the same.
African Americans are most likely to recidivate.

ACCURACY

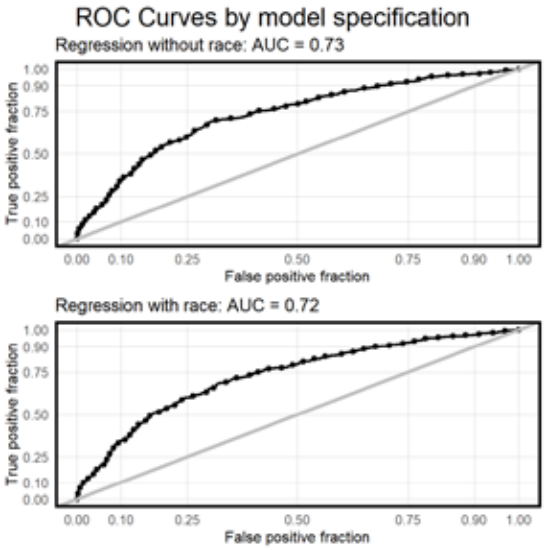

figure 3

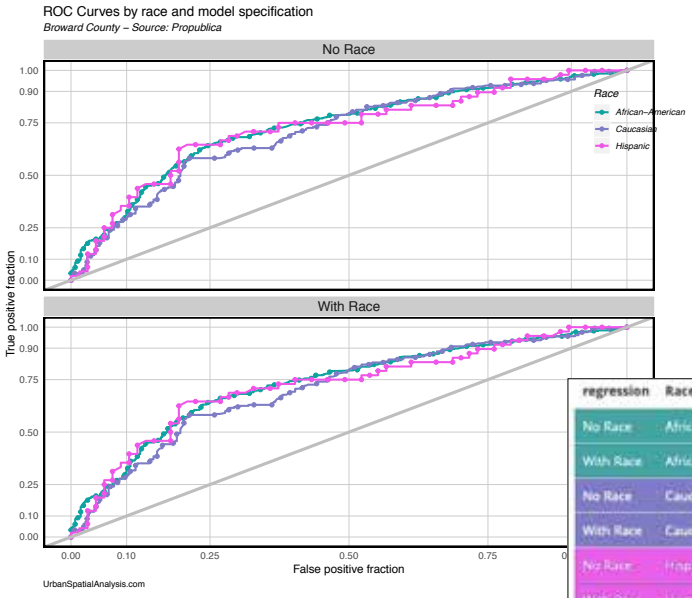

figure 4

regression	Race	AUC
No Race	African-American	0.7317832
With Race	African-American	0.7320427
No Race	Caucasian	0.7090416
With Race	Caucasian	0.7089330
No Race	Hispanic	0.7232799
With Race	Hispanic	0.7234142

figure 5

While the ROC curve (fig. 4) is a viable metric for assessing accuracy, it is relatively course analysis that may not fully capture the nuance of this challenge. An examination of observed and predicted recidivism rates for African American versus other races (fig. 6) shows a disparity in predictive power. While the COMPAS model is useful (as seen with the ROC curve), it may also be considered unfair (as evidenced by the values which inaccurately predict the likelihood of recidivism in Caucasians and Hispanics). Even though the model is fairly accurate on the whole, the chart of observed versus predicted rates show that the predictive power does not translate equally for every racial group.

Why is the model unfair?

A model is a representation of reality; naturally, it leaves part of reality out when selecting variables. If the model is better fit to certain race groups, than others then the data mined for the algorithms failed to integrate critical factors, like education outcomes and employment history. The model also omits the community gestalt. When the community is broken into pieces as variables, the model can't take into consideration the impact of the place.

Chicago's murder rates are off the charts right now, but only in black areas on the Southside of the city. It might be argued that the selection bias is due to more violent crime in certain neighborhoods. However, the COMPAS model has been "remarkably unreliable in forecasting violent crime" (Angwin, et. al., 2016), a fact that conclusively negates this argument. Instead, over-policing impacts base rates, which cause bias in the model as follows: the model "guarantees black defendants will be inaccurately identified as future criminals

GENERALIZABILITY

Race	regression	Observed.Recidivism.Rate	Predicted.Recidivism.Rate
African-American	No Race	0.4943396	0.5031447
African-American	With Race	0.4943396	0.5169811
Caucasian	No Race	0.4015444	0.2837838
Caucasian	With Race	0.4015444	0.2702703
Hispanic	No Race	0.4173913	0.3043478
Hispanic	With Race	0.4173913	0.2173913

figure 6

more often than their white counterparts" (Angwin; Larson, 2016). Said in another way, the model is overfit for a specific group and underfit for others, e.g. it doesn't generalize fairly.

The metric critical to generalizability analysis is the difference in false-positive and false-negative prediction rates across racial groups. Focusing on outcomes is the best definition of fairness (Angwin, 2016). Due to the unequal base rates (over-policing), a risk score cannot be both "equally predictive or equally wrong for all races," (Kleinburg, 2017). There are tradeoffs and the two metrics for measuring fairness are incompatible (Kleinburg, 2017). "If you have two populations that have unequal base rates, then you can't satisfy both definitions of fairness at the same time," (Angwin, 2016).

Uses

Although COMPAS does not achieve a balance of accuracy and generalizability that would warrant the distinction of "fair", that doesn't mean it should be trashed. Limitations of the model must be known and it should not be used as liberally. The model may not represent absolute truth but could be useful in the decision-making process for the allocation of rehabilitation funding. In this context, if mistakes are made, there may not be a loss. Figure 11 shows the possible positive influence of rehabilitation funding for any defendant.

Solutions

Ultimately, the model "would have to treat people differently" to create equal outcomes (Chouldechova, 2016).

Facetted barplot of observed recidivism rates : predicted recidivism rates by select race groups

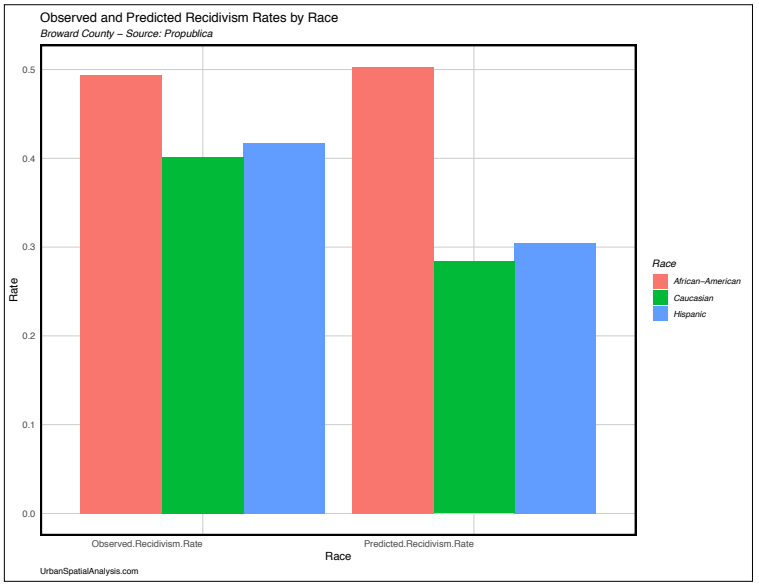

figure 7

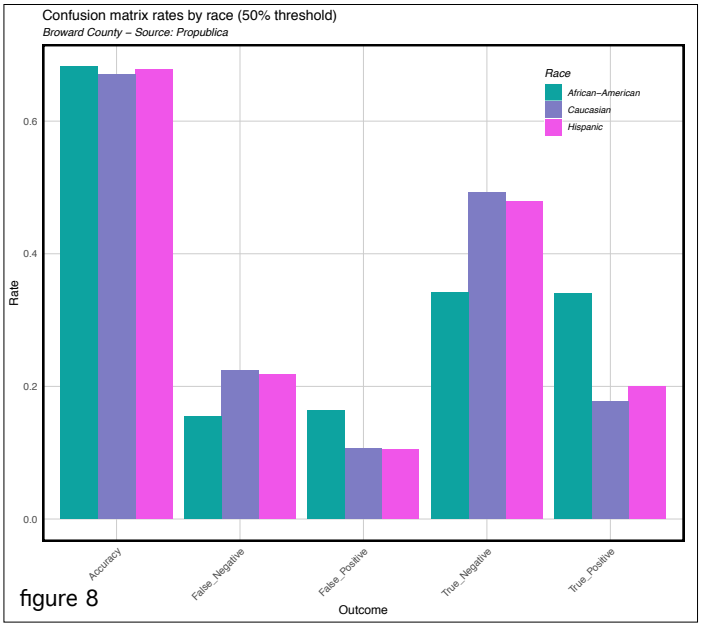

figure 8

The plot above shows that "black defendants were twice as likely to be incorrectly labeled as higher risk than white defendants. Conversely, white defendants labeled low risk were far more likely to end up being charged with new offenses than blacks with comparably low COMPAS risk scores" (Angwin; Larson, 2016). Black defendants are predicted to recidivate almost as much as they do but white and Hispanic predicted rates are very low, while the actual rate is much higher.

Race	True_Positive	True_Negative	False_Negative	False_Positive	Accuracy
African American	0.3396226	0.3421384	0.1547170	0.1635220	0.6817610
Caucasian	0.1776062	0.4922780	0.2239382	0.1061776	0.6698842
Hispanic	0.2000000	0.4782609	0.2173913	0.1043478	0.6782609

figure 9

Race	regression	Difference
African American	No Race	0.0088050
African American	With Race	0.0226415
Caucasian	No Race	0.1177606
Caucasian	With Race	0.1312741
Hispanic	No Race	0.1130435
Hispanic	With Race	0.2000000

figure 10

figure 11

Repeat offenders crime is more likely to be a lower degree. this is a positive trend and supports the idea that any public service or program for previous inmates could be beneficial.