A Diary Study on Algorithmic Auditing:Analyzing Racial Bias in the Algorithm of COMPAS

Walid Medani

05/04/21

As the overton window in public opinion shifts to more accountability, transparency and oversight in algorithms or black box models. The ethical dilemma of using models as the sole or major factor in deciding incarceration has permeated the discourse surrounding Risk Assessment Instruments (RAI). RAIs are algorithmic tools that assess a defendant's future risk of recidivism and court appearance based on a variety of factors and historical group tendencies. However, crime is an artificial concept and the way it's defined is subjective. The methods to measure it are imperfect and often display racial biases. It discounts and reinforces structural racial disparities in environmental factors, material conditions, education, and population density. In a 2013 study, (Olver, Stockdale, and Wormith (2014) found that ethnic minorities had higher risk scores in the U.S. in comparison to Canada and noted that "systemic bias within the justice system may distort recidivism." RAIs' predictive accuracy is also not supported by current evidence (Douglas et al. (2017),

> Existing data suggest that most risk assessment tools have poor to moderate accuracy in most applications. Typically, more than half of individuals judged by tools as high risk are incorrectly classified—they will not go on to offend [13]. These persons may be detained unnecessarily. False positives may be especially common in minority ethnic groups [14], [15].

The trade-secrets of these RAIs is currently justified under our current set of intellectual property laws, but it's unethical to justify private profiteering when it impacts incarceration. As algorithms models decide upon the balance of positive and negative impacts, the public requires transparency to understand how persons are being sorted and judged in order to have freedom from bias. The need for transparency is imperative due to the validity of RAIs being examined mostly by the same people who developed the instruments (Desmarais and Singh (2013).

In 2016, ProPublica questioned the validity of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in predicting recidivism and racial bias. COMPAS is an RAI that is widely used throughout the U.S. in pretrial and sentencing. It analyzes a questionnaire of 137 variables such as age, sex, criminal history and computes a scale for recidivism at the time of booking in jail. COMPAS does not use race in factoring its recidivist scores. To ensure equitable application, ProPublica wanted to determine if these risk scores were in fact accurate and void of racial biases despite it not accounting for race. To examine the collective capacity of auditing algorithms, with only knowledge of introductory statistics and programming in R, I will attempt to replicate ProPublica's findings to journal my experiences and the cognitive labor necessary for algorithmic literacy.

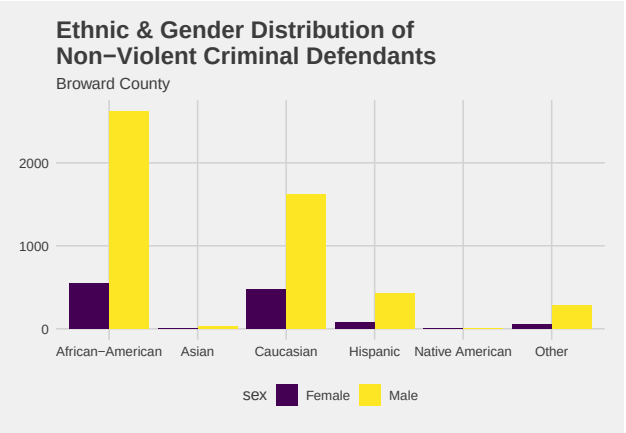
Data & Visualization

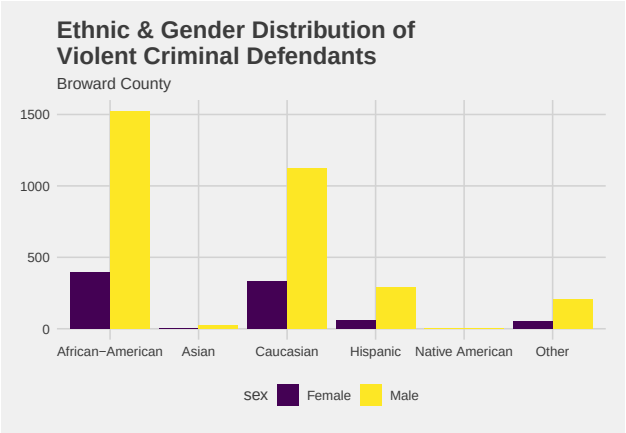
To analyze bias within COMPAS, I will be using the data ProPublica collected of 18,610 criminal defendants in Broward County. Broward County was specifically chosen since it utilizes COMPAS to oversee a large population size and benefits from Florida's transparent public record laws. Scores of criminal defendants that were assessed in stages outside of pre-trial were discarded, reducing the total criminal defendants down to 11,957 people.

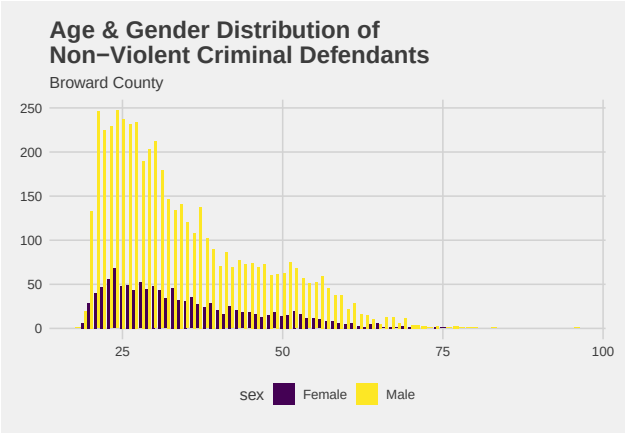
In order to clean up the raw data, defendants that were not screened by COMPAS within 30 days of the arrest were dismissed to ensure quality. Those who did not receive a COMPAS

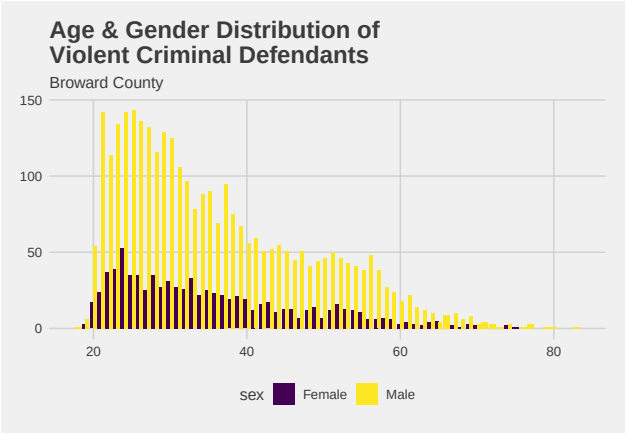
screening or committed traffic offenses with no jail time were removed.

Below is the demographic breakdown of the sample of non-violent and violent criminal defendants processed by COMPAS in Broward County.









COMPAS's decile scores are rankings of normative groups in ascending order. In Northpointe's practitioner guide for COMPAS, decile scores are interpreted as following:

- **1 − 4**: scale score is **low** relative to other offenders in norm group.
- **5 − 7**: scale score is **medium** relative to other offenders in norm group.
- **8 − 10**: scale score is **high** relative to other offenders in norm group.

We start to visually see bias in the data as shown in the figures below. The trend for Caucasian decile scores decreases as the normative scale score increases,

whereas for African-Americans it is much more distributed throughout the rankings. Although scores for Native Americans is high, we will later on find out it's not statistically significant due to a sample size of 11 criminal defendants.

Those who receive a decile score of 5 or more are classified as medium to high risk in relation to the normative group. By taking the mean of those who were predicted to recidivate (decile scores 5-10) and the actual recidivism rate, we can find the accuracy of

recidivism predictions. We find that COMPAS is 62% accurate in predicting non-violent recidivism and 35% accurate for predicting violent criminal defendants.

```
##   nonviolent_accuracy violent_accuracy
## 1            62.00583         35.14925
```

Since we have data on those who did recidivate within two years of receiving a decile score. We can run a confusion matrix to evaluate the performance of COMPAS's classification model in predicting recividism.

- True positives (TP): predicted to recidivate and did recidivate (correct classification)
- False positives (FP): predicted to recidivate but did NOT recidivate (incorrect classification)
- True negatives (TN): NOT predicted to recidivate and did NOT recidivate (correct classification)
- False negatives (FN): NOT predicted to recidivate but did recidivate (incorrect classification)

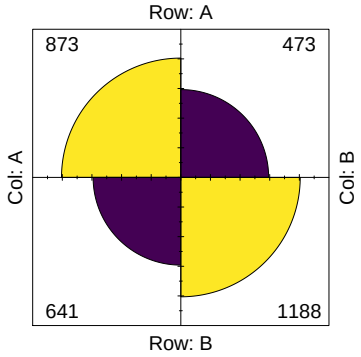
By using the formula $^{FP/TP+FP}$ we can find the false positive rate (incorrect classification). African American defendants had a 35% false positive rate in comparison to 28% in Caucasians.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  873  473
##          1  641 1188
##
```

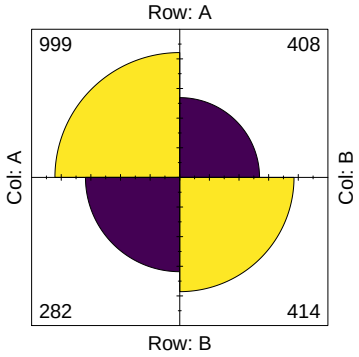
```
##                Accuracy : 0.6491
##                  95% CI : (0.6322, 0.6657)
##     No Information Rate : 0.5231
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2933
##
##  Mcnemar's Test P-Value : 5.63e-07
##
##             Sensitivity : 0.7152
##             Specificity : 0.5766
##          Pos Pred Value : 0.6495
##          Neg Pred Value : 0.6486
##              Prevalence : 0.5231
##          Detection Rate : 0.3742
##    Detection Prevalence : 0.5761
##       Balanced Accuracy : 0.6459
##
##        'Positive' Class : 1
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 999 408
##          1 282 414
##
```

```
##                Accuracy : 0.6719
##                  95% CI : (0.6514, 0.692)
##     No Information Rate : 0.6091
##     P-Value [Acc > NIR] : 1.417e-09
##
##                   Kappa : 0.2915
##
##  Mcnemar's Test P-Value : 1.949e-06
##
##             Sensitivity : 0.5036
##             Specificity : 0.7799
##          Pos Pred Value : 0.5948
##          Neg Pred Value : 0.7100
##              Prevalence : 0.3909
##          Detection Rate : 0.1969
##    Detection Prevalence : 0.3310
##       Balanced Accuracy : 0.6418
##
##        'Positive' Class : 1
##
```

African American Confusion Matrix



Caucasian Confusion Matrix



The distributions of decile scores might visually indicate bias but it does not take into consideration other factors that may be impacting it. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary and we want to explain the relationship between one dependent binary variable and one or more independent variables. To test racial bias in decile scores, the logistic regression model will control for race, age, criminal history, future recidivism, charge degree, gender and age.

```
##
## Call:
## glm(formula = is_med_or_high_risk ~ gender_factor + age_factor +
##     race_factor + priors_count + crime_factor + two_year_recid,
##     family = "binomial", data = compasglm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9966  -0.7919  -0.3303   0.8121   2.6024
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.52554    0.07851 -19.430  < 2e-16 ***
## gender_factorFemale         0.22127    0.07951   2.783 0.005388 **
## age_factorGreater than 45  -1.35563    0.09908 -13.682  < 2e-16 ***
## age_factorLess than 25      1.30839    0.07593  17.232  < 2e-16 ***
## race_factorAfrican-American 0.47721    0.06935   6.881 5.93e-12 ***
## race_factorAsian           -0.25441    0.47821  -0.532 0.594717
## race_factorHispanic        -0.42839    0.12813  -3.344 0.000827 ***
## race_factorNative American  1.39421    0.76612   1.820 0.068784 .
## race_factorOther           -0.82635    0.16208  -5.098 3.43e-07 ***
```

```
## priors_count                 0.26895    0.01110  24.221  < 2e-16 ***
## crime_factorM               -0.31124    0.06655  -4.677 2.91e-06 ***
## two_year_recid               0.68586    0.06402  10.713  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8483.3  on 6171  degrees of freedom
## Residual deviance: 6168.4  on 6160  degrees of freedom
## AIC: 6192.4
##
## Number of Fisher Scoring iterations: 5

##
## Call:
## glm(formula = score_factor ~ gender_factor + age_factor + race_factor +
##     priors_count + crime_factor + two_year_recid, family = "binomial",
##     data = compasglm_v)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9304  -0.5667  -0.3161   0.4192   2.8386
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.24274    0.11326 -19.802  < 2e-16 ***
## gender_factorFemale        -0.72890    0.12666  -5.755 8.66e-09 ***
```

```
## age_factorGreater than 45   -1.74208    0.18415  -9.460  < 2e-16 ***
## age_factorLess than 25       3.14591    0.11541  27.259  < 2e-16 ***
## race_factorAfrican-American  0.65893    0.10815   6.093 1.11e-09 ***
## race_factorAsian            -0.98521    0.70537  -1.397   0.1625
## race_factorHispanic         -0.06416    0.19133  -0.335   0.7374
## race_factorNative American   0.44793    1.03546   0.433   0.6653
## race_factorOther            -0.20543    0.22464  -0.914   0.3605
## priors_count                 0.13764    0.01161  11.854  < 2e-16 ***
## crime_factorM               -0.16367    0.09807  -1.669   0.0951 .
## two_year_recid               0.93448    0.11527   8.107 5.20e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4731.8  on 4019  degrees of freedom
## Residual deviance: 2998.8  on 4008  degrees of freedom
## AIC: 3022.8
##
## Number of Fisher Scoring iterations: 6
```

Based on the coefficients above, African American defendants were 47% more likely to receive a medium or high score. African American violent defendants were 65% more likely to receive a medium or high score. However these numbers aren't necessarily perfectly accurate due to the need of further calculating predicted probabilities.

Reflections

As a beginning student in the field of data science, the process of analyzing bias in datasets was a laborious task. Data preparation is a time-consuming but critical component in preparing data for exploratory analysis. But the ways to manipulate data are endless, and without a thorough background in statistics, it is easy to poison the data generating your analysis. Luckily there were a few case studies on COMPAS for me to learn from on how to manipulate datasets in order to begin my investigation. Once the data was cleaned, I was able to perform exploratory data analysis to summarize the characteristics of the dataset and formulate hypotheses without any modeling. This is the statistical method a beginner can perform to measure or visualize bias in data through discovering patterns, spotting anomalies, and checking assumptions with summary statistics.

The next step after conducting exploratory data analysis however is complex and requires the need of creating models and algorithms. I aspired to model the predictive accuracy of COMPAS's decile scores. I spent hours trying to learn the basics of machine learning and how to build classifiers to train my dataset. But ultimately I didn't have the foundation in multivariate statistics to apply models such as Cox Proportional Hazards or Survival Analysis to test for collider bias or casual influence. Open-source models to measure algorithmic fairness were readily available and comprehensively documented (e.g. IBM's Fairness 360, Fairness in R), with metrics such as demographic parity, proportional parity, ROC AUC comparison, etc. Yet it was still difficult for a beginner to implement these machine learning metrics without being cognizant of sensitive variables and which techniques to utilize in order to supplement these metrics. As (Kozodoi and Varga (2020) points out in relation to applying fairness metrics in COMPAS,

> First, excluding ethnicity from the features slightly increases precision for some defendants (Caucasian and African_American) but results in a lower precision for some other groups (Asian and Hispanic). This illustrates that improving a

model for one group may cost a fall in the predictive performance for the general population. Depending on the context, it is a task of a decision-maker to decide what is best.

Executing the models mentioned above with statistical computing was fairly easy, however interpreting the output from these models obstructed my exploration. This form of algorithmic opacity results from the fact that writing and reading code is a specialized activity (Burrell (2016). Even with publicly available open source models, the operation remains largely incomprehensible for those without specialized training.

Conclusion

ProPublica's analysis of COMPAS found a false positive rate (defendant is predicted as medium/high risk but does not re-offend) for African-American criminal defendants much higher than it is for caucasian ones, specifically for violent crimes. These findings sparked a debate between academics and RAI developers in what constitutes fairness. COMPAS's defense is that the proportion of recidivism is the same regardless of race, therefore making the algorithm fair. We would consider it unfair If the algorithm was to assign caucasians higher decile scores in order to mitigate racial bias.

However the assumption underlying this debate is that we blindly accept and rely on digital technology to accomplish ordinary goals. In a study to determine the need of such software, (Dressel and Farid (2018) recruited 400 non-expert participants from Amazon's Mechanical Turk to predict recidivism with only seven variables in comparison to COMPAS's 137; participants had a prediction accuracy of 63 percent in comparison to the 67 percent of COMPAS. In another study, (Angelino et al. (2017) developed an algorithmic model comprising of only a criminal defendant's sex, age, and prior convictions to replicate the same predictive accuracy as COMPAS. The appeal of algorithms and big data may influence a judge's decision but would they still consider a RAI to set bail and sentencing if it performs

the same as random surveyors? The problem is this reliance on algorithmic systems as oracles and without proper interpretation, the decision-making of algorithmic systems could devolve to perceive meaningful connections between seemingly unrelated things. And when these patterns are made, it could erode the dignity and autonomy of people and impose on their freedom from bias.

What ProPublica's analysis points to is most concerning possible sources of bias, which can come from the historical outcomes that an RAI learns to predict. Due to systemic racism engulfing the criminal justice system, models will only learn to replicate the outcomes of unjust practices. African Americans have historically been convicted at higher rates and racial disparity is exhibited in many forms from wage gaps, over-policing, education, sentencing, parole, and to bail. Datasets do not address the underlying social and structural hierarchies at play and are only modeled around the status quo. In order to circumvent racism in technology, (Benjamin (2019) profoundly pinpoints the concept of how "blackness can be both marginal and focal to tech development." That if we compare and contrast ostensibly different technologies we can better sort through what is consequential to racial inequality.

Bibliography

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2017. "Learning Certifiably Optimal Rule Lists for Categorical Data." *arXiv Preprint arXiv:1704.01701*. https://arxiv.org/abs/1704.01701.

Benjamin, Ruha. 2019. "Race After Technology: Abolitionist Tools for the New Jim Code." *Social Forces*.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 2053951715622512. https://journals.sagepub.com/doi/full/10.1177/2053951715622512.

Desmarais, Sarah L, and Jay P Singh. 2013. "Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States." *Lexington, KY: Council of State Governments*. https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf.

Douglas, Thomas, Jonathan Pugh, Ilina Singh, Julian Savulescu, and Seena Fazel. 2017. "Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data." *European Psychiatry* 42: 134–37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408162/.

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4 (1): eaao5580. https://advances.sciencemag.org/content/4/1/eaao5580.

Kozodoi, Nikita, and Tibor V. Varga. 2020. "Algorithmic Fairness in r." https://kozodoi.me/r/fairness/packages/2020/05/01/fairness-tutorial.html.

Olver, Mark E, Keira C Stockdale, and J Stephen Wormith. 2014. "Thirty Years of Research on the Level of Service Scales: A Meta-Analytic Examination of Predictive Accuracy and Sources of Variability." *Psychological Assessment* 26 (1): 156. https://pubmed.ncbi.nlm.nih.gov/24274046/.