Evaluating Racial Bias in COMPAS Assessments and Machine Learning Algorithms

Žan Mervič
zm3587@student.uni-lj.si

ABSTRACT

The goal of this project, as given by Prof. Dr. Ivan Bratko, was to investigate: Are COMPAS scores racially biased? Is machine learning using these data biased? Can we reproduce the results from Dressel and Farid (2018)[4] and Rudin (2019)[5] with Orange? Would excluding certain data, such as race, change the results? Additionally, can race be reliably predicted from other offender data?

1 INTRODUCTION

Artificial intelligence and machine learning are increasingly used in everyday life. This can be useful for applications such as autonomous driving and weather prediction, but what about making decisions that could significantly impact an individual's life, such as employment or criminal justice? In this project, we will focus on the latter. We will investigate if the scores produced by the risk assessment tool COMPAS, which tries to predict the likelihood of a defendant reoffending, are racially biased and try to replicate the results from different articles, most of which will be done using the Orange data mining software[1].

All the Orange workflows and code used in this project are available in our GitHub repository[2].

2 DATA

We will use the same data ProPublica used for their Machine Bias [1] article. The dataset contains information about 7214 defendants from Broward County, Florida, whom COMPAS scored. For most of the analysis, we will only use a subset of features from the data and only the defendants who were labeled as either "African-American" or "Caucasian," as it was done in most articles. The features we will use are the same as in the Dressel and Farid (2018) article [4]; these include age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior (nonjuvenile) crimes, crime degree, and crime charge, in most cases we will also include the race feature.

3 ARE COMPAS SCORES RACIALLY BIASED?

In the ProPublica article [1], the authors claim that the COMPAS scores are racially biased. They argue that the scores are biased mainly because the false positive rate for African-American defendants is higher than for Caucasian defendants; the opposite is true for the false negative rate, meaning that African-American defendants are more likely to be falsely labeled as high risk to reoffend, while Caucasian defendants are more likely to be falsely labeled as low risk. Using Orange, we successfully replicated all the results from the ProPublica article. So, does this mean that COMPAS scores are racially biased? It depends. The "problem" with fairness is that there are many different definitions and metrics for fairness, and very often, these definitions and metrics are mutually exclusive. That was also one of the main arguments in the response to the ProPublica article from Northpointe [3], the company that created COMPAS. So, are COMPAS scores racially biased? The answer is that it depends on how bias is defined.

4 IS MACHINE LEARNING USING THESE DATA BIASED?

Before we can answer this question, we need to set some ground rules. To avoid the "depends on how you define fairness" problem, we will use two popular metrics for fairness:

- Equalized/Average odds: The false positive rate and the false negative rate should be equal for all groups. Calculated as the average of the true positive rate difference and the false positive rate difference between the protected and non-protected groups. The ideal value is 0 (no bias).
- Disparate impact: The positive outcomes ratio between the groups should be equal. Calculated as the ratio of the positive outcomes between the protected and non-protected groups. The ideal value is 1 (no bias).

In both cases, values under the ideal value indicate a bias towards the protected group (African-American). In contrast, values above the ideal value indicate a bias towards the non-protected group (Caucasian).

Using Orange, we trained logistic regression and random forest models on the data and evaluated them using the two metrics. The results were as follows:

	AUC	Equalized odds	Disparate impact
Logistic Regression	0.722	-0.196	0.687
Random Forest	0.693	-0.163	0.712

The results show that the predictions made by both models contain a significant amount of bias towards the unprivileged (African-American) group. So, is machine learning using these data biased? Yes, it is, according to the two metrics we used.

5 CAN WE REPRODUCE THE RESULTS FROM DRESSEL AND FARID (2018) AND RUDIN (2019) WITH ORANGE?

Long story short, no, we could not. The main reason was that the authors did not provide information about either the models or the data they used (or both), making it hard to replicate the results. There are some interesting findings from the articles and attempts to replicate them.

In Dressel and Farid (2018) [4], they did an experiment where they gave COMPAS data (with and without the race attribute) to

[1] https://orangedatamining.com/
[2] https://github.com/ZanMervic/Evaluating-Racial-Bias-in-COMPAS

humans and asked them to predict the likelihood of a defendant to reoffend (the same task COMPAS does). The results showed that the predictions made by humans were very similar to the predictions made by COMPAS, both in terms of accuracy and bias. From this, we can conclude that, in this case, the bias is not due to the algorithms but possibly due to the data or simply the nature of the task itself.

In the Rudin (2019) [5] article, the authors advocate for the use of interpretable models instead of black box models and claim that when used on COMPAS data, interpretable models can achieve the same results (similar accuracy and bias) as the COMPAS algorithm, while being transparent. In our attempts, that was not the case; the interpretable models consistently scored lower than the COMPAS algorithm and other models (logistic regression, random forest). However, this could be due to the different data and hyperparameters used.

Interestingly, in one of the Cross-Validation folds, the interpretable CORELS model included the race attribute in its rules, where the race "African-American" increased the likelihood of being labeled as high risk to reoffend.

```
if [priors_count=< 10 && not sex=Male]:
prediction = False
else if [age=20 - 40 && race=African-American]:
prediction = True
else if [priors_count=< 10 && not age=< 20]:
prediction = False
else
prediction = True
```

Gender and age "bias" were also present in the rules, with males and younger defendants being more likely to be labeled as high risk to reoffend.

6 WOULD EXCLUDING CERTAIN DATA, SUCH AS RACE, CHANGE THE RESULTS?

To test this, we trained a logistic regression and a random forest model on the data with and without the race feature. The results were as follows:

	AUC	Equalized odds	Disparate impact
Models trained on data with the race attribute			
Logistic Regression	0.722	-0.196	0.687
Random Forest	0.693	-0.163	0.712
Models trained on data without the race attribute			
Logistic Regression	0.722	-0.167	0.719
Random Forest	0.692	-0.145	0.734

From the results, we can see that the accuracy of the models did not change; on the other hand, the bias did improve, but only slightly. So, would excluding the protected attribute (race) change the results? No, not by a meaningful amount. To improve the results further, we need to use more advanced techniques.

So why did removing the protected attribute not solve our problems? One of the reasons might be that the protected attribute is correlated with other attributes, meaning it can be predicted from other attributes. To test this, we trained a logistic regression and a random forest model on the data, where race is the target attribute. Logistic regression achieved an AUC of 0.694, and random forest achieved an AUC of 0.667. This shows that race can be predicted from other attributes to some extent.

The other reason might be that the race attribute is not very important for making predictions. We tested this by calculating feature importance for the models. We found that for both models, the race attribute was one of the least important features (the least important feature for logistic regression and the third least important feature for random forest).

7 ADVANCED BIAS MITIGATION TECHNIQUES

So far, we have learned that the data we are using is possibly biased; depending on the definition, the models trained on the data are biased, and removing the protected attribute does only a little to improve the results. So, what can we do to mitigate the bias? In Orange, the Fairness addon implements some of the AIF360 [2] algorithms. We tried using the Adversarial Debiasing and the Equalized Odds Postprocessing (EOP) algorithms, and here are the results:

	AUC	Equalized odds	Disparate impact
Logistic Regression	0.722	-0.196	0.687
Adversarial debiasing	0.687	-0.111	0.793
EOP	0.611	-0.001	0.958

As we can see, using bias mitigation techniques can significantly improve the fairness of the models (in the case of EOP, it almost completely mitigates bias), but at the cost of accuracy. This is expected because the models are tasked with not only accurately predicting the target attribute but also to be fair. The trade-off between accuracy and fairness is a common problem in AI fairness, and it is up to the user to decide which is more important.

Note: By removing the crime charge attribute, the result of the Adversarial Debiasing algorithm is much better, with an AUC of 0.713, equalized odds of 0.023, and disparate impact of 0.978. It often takes some domain knowledge and experimentation to get the best results.

8 CONCLUSION

Bias and fairness in AI and machine learning are complicated issues, especially in critical areas like criminal justice. Our analysis confirms that COMPAS scores and models trained on similar data show signs of racial bias, but it all depends on the definition of bias. Excluding race from the models only slightly reduced bias, indicating deeper systemic issues.

Replicating studies like Dressel and Farid (2018) and Rudin (2019) proved challenging due to a lack of detailed methodology, yet it highlighted that bias often persists regardless of the approach. Advanced bias mitigation techniques showed promise but at the expense of accuracy, emphasizing the need for careful, context-specific applications.

Addressing bias and fairness is not just a technical task but requires interdisciplinary collaboration. Bias mitigation should be applied judiciously, involving experts from various fields to ensure AI systems are both fair and effective.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[2] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[3] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 7, 4 (2016), 1–36.

[4] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[5] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.