AUDTING FOR BIAS

Vedant Venkatesh Yelsangikar, G01379948

Master of Science in Computer Science, George Mason University, vyelsang@gmu.edu

Harsha Masandrapalya Vanarajaiah, G01396731

Master of Science in Computer Science, George Mason University, hmasandr@gmu.edu

1. ABSTRACT

The ProPublica COMPAS investigation was a study conducted by ProPublica in 2016 that analyzed the performance of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm in predicting recidivism risk for individuals in the criminal justice system. The study found that black defendants were almost twice as likely as white defendants to be falsely labeled as having a high risk of recidivism, while white defendants were more likely to be falsely labeled as having a low risk of recidivism. The investigation also found that the algorithm was less accurate in predicting recidivism for black defendants than for white defendants.

Our objective is to reduce bias in your algorithms or decision-making processes, there are steps which are:

- Identify and measure bias.
- Collect diverse data.
- Train your models on diverse data
- Evaluate and test for bias.
- Test with different models for efficient fit and accuracy
- Note Observations and findings.

2. INTRODUCTION

The COMPAS dataset includes information on the likelihood of recidivism (reoffending) within two years for individuals who were arrested in Broward County, Florida, between 2013 and 2014. The risk score generated by the COMPAS algorithm predicts the probability of reoffending within that two-year period, based on a variety of factors such as criminal history, age, gender, and other demographic and personal characteristics.

It's crucial to remember that the COMPAS dataset only provides information for the first two years, and its projections are only valid for the population and region from which the data were gathered. Therefore, it should be utilized with caution when extrapolating its findings to other contexts as the dataset could not be representative of other populations or places. Overall, the COMPAS dataset has been utilized extensively in studies on recidivism prediction and criminal justice, and it has provoked significant discussions and arguments concerning the use of algorithmic tools in the criminal justice system.

Significant racial bias was found in the COMPAS algorithm, which was used to forecast recidivism risk for those involved in the criminal justice system, according to the ProPublica COMPAS investigation. We set out to evaluate the "compas-scores-two-years" edition of the ProPublica's COMPAS datasets in order to better understand the algorithm's performance and try to reduce the bias instilled in the algorithm. About 7214 defendants' incarceration and background data are included in this dataset, together with 53 features, such as whether they will reoffend within two years of the judgement.

3. DATA PREPROCESSING AND METHODS

We carried out several various data preprocessing techniques, mainly converting data to numerical values so that it is easy to analyze and predict on the machine learning model.

Before starting the classification task, we examined all 53 features in the COMPAS dataset to identify the most relevant ones for the task at hand. We Used domain knowledge and data exploration techniques to help me in this process. We found that not all 53 features were relevant for the classification task, and some of them were highly correlated with other features. The crime data collected includes name, COMPAS screening date, sex, date of birth, age, age category, race, juvenile felony count, decile score etc., and more. We concentrate on variables that could obviously affect recidivism rates for the sake of our analysis. This includes features like: sex, age, race, juv_fel_count, decile_score, juv_misd_count, juv_other_count, priors_count, c_charge_degree, is_violent_recid, two_year_recid, c_jail_in, c_jail_out, score_text, days_b_screening_arrest also created a few features like F, M, African-American, Asian, Caucasian, Hispanic, Native American, Other, Female, Male, High, Low, Medium, days_in_jail for better understanding of selected features.

The ‘days_in_jail’ feature was added by calculating from the features 'c_jail_in', and 'c_jail_out' to improve the accuracy of the model. We chose to eliminate rows where the offenses do not cause jail time and removed the cases in which charge date is more than 30 days (about 4 and a half weeks).

We want to transform our data set such that we have only binary values for each column (like race, sex, etc.,) also known as encoding, so we use a method called get_dummies(): Our Data Frame now contains a binary value in each column (including the individual race columns: African-American, Asian, Caucasian, Hispanic, Native American, Other) and then we standardized the features of a dataset. Standardization involves scaling the features so that they have zero mean and unit variance. This is often done to ensure that all features are on a similar scale, which can help to improve the performance of some machine learning algorithms. After cleaning the data, the dimensions of our dataset were (6172 rows × 23 columns).

4. METHOD IMPLEMENTATION

Table 1: Accuracy of Prediction, False positive rates and calibration values for different models

Model	Accuracy	False Positive Rate		Calibration Value	
		African American	Caucasian	African American	Caucasian
Logistic Regression	0.714	0.257	0.105	0.733	0.748
Adaboost	0.720'	0.275	0.122	0.727	0.74
Random Forest	0.702	0.321	0.176	0.704	0.665
KNN	0.663	0.326	0.234	0.681	0.561

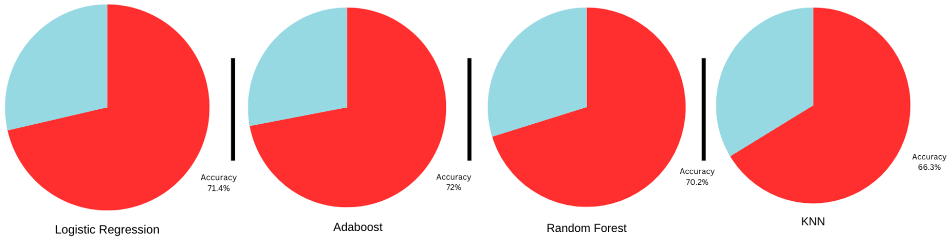

Figure 1: Accuracy depicted by different models

The accuracy values range from 0.663 to 0.720, indicating that all models perform reasonably well on this task. However, the false positive rate (the proportion of individuals who are predicted to recidivate but do not) and the calibration value (a measure of how well the model's predicted probabilities match the observed probabilities) vary across models and across racial groups. For example, the false positive rate for KNN is the highest among all models, and the calibration value for African Americans is the lowest for the Logistic Regression model.

However, based on the information from the table, the Adaboost and Logistic Regression model appears to be the best performer overall, with the highest accuracy and the lowest false positive rate among the four models.

5. OUR OBSERVATIONS

To check whether there is any bias in terms of opportunity cost, we compared false positive rates for each group. We can calculate the false positive rate for each group as follows:

- False positive rate for African Americans = FalsePositive_AfricanAmerican / (TrueNegative_AfricanAmerican + FalsePositive_AfricanAmerican)
- False positive rate for Caucasians = FalsePositive_Caucasian / (TrueNagative_Caucasian + FalsePositive _Caucasian)

Table below shows the false positives rates of different groups of people.

Table 2: False positive rates for African Americans and Caucasians

	African-American	Caucasian
False Positives	132	45
True Negatives	385	382
False Positive Rate	0.255	0.105

The false positive rate is the proportion of individuals who are predicted to be positive but are actually negative. In this case, the false positive rate is higher for African-Americans (0.255) than for Caucasians (0.105), indicating that the model has an approximately 2.5x higher rate of false positives for African-Americans defendant and falsely predicting that they will reoffend within two years when in fact they will not, compared to falsely predicting the same for a Caucasian defendant. This is a concerning finding, as it suggests that the model may be biased against African American defendants.

Next, we evaluated calibration, which measures the agreement between the predicted probability and the true probability of recidivism.

Table 3: Calibration values and rates for African Americans and Caucasians

	African American	Caucasian
True Positive	363	134
Predicted Positive	495	179
Calibration value	0.733	0.748

We compared the calibration for African American and Caucasian defendants who were predicted positive by our model. This suggests that our model is slightly biased towards Caucasian defendants in terms of calibration. In this case, the calibration value is slightly lower for African Americans (0.733) than for Caucasians (0.748), suggesting that the model's predicted probabilities may be less reliable for African American defendants. While the true positive rate may be higher for African Americans, the higher predicted positive rate and potential calibration issues should also be considered when evaluating model performance and potential biases.

We were interested in determining whether removing the race factor during model training would affect racial bias. However, after looking at the results below we can see that accuracy did not change and there was minute difference in the false positive rate and calibration value, we can infer that racial bias in the data set is not eliminated by not accounting for race; it is probable that one or more additional variables are a pseudo-race indicator. In other words, even if race is taken out of the equation, there may still be traces of the defendant's race in the data set due to a trait that has a significant connection with race. Indicating that race does not impact our prediction.

Table 4: False positive rates and calibration value for Protected race features

Logistic Regression (Protected race feature)		
Accuracy		0.714
False Positive Rate	**African American**	0.253
	Caucasian	0.103
Calibration Value	**African American**	0.733
	Caucasian	0.748

After trying out many different models and methods we can still see there is a significant bias towards the African Americans groups, so we used fair classifiers like Demographic Parity Classifier from Scikit-lego library and EqualOpportunityClassifier, Prior to this, we reweighted the training data to correct for imbalances in the dataset and to ensure demographic parity. This entails giving underrepresented groups heavier weights and overrepresented groups lighter weights. Below table shows our observations.

Table 5: Accuracy, False positive rates and calibration values for EqualOpportunity and Demographic Parity classifiers.

	Accuracy	**False positive**		**True Negative**		**False Positive Rates**		**True Positive**		**Predicted Positive**		**Calibration Value**	
Classifiers		**African American**	**Caucasian**	**African American**	**Caucasian**	**African American**	**Caucasian**	**African American**	**Caucasian**	**African American**	**Caucasian**	**African American**	**Caucasian**
Equal Opportunity	0.663	20	5	497	422	0.038	0.0117	172	68	192	73	0.895	0.801
Demographic Parity	0.671	41	20	476	407	0.799	0.046	200	81	241	101	0.932	0.801

We discovered that the Demographic parity classifier produced significantly better outcomes in terms of both calibration and opportunity cost. Nevertheless, the dataset's implicit bias continues to favor African American criminals over Caucasians. Although the frequency of false positives has decreased, American rates are about twice as high as those of Caucasians.

6. CONCLUSION

From this we can conclude that the models performed differently for different demographic groups, with African American defendants experiencing higher false positive rates and lower calibration values compared to Caucasian defendants. Our analysis highlights the importance of considering issues of fairness in machine learning models, especially in high-stakes domains such as criminal justice. The issue does not lie only with the COMPAS algorithm, but also depends maybe on data and the process by which we determine "risk of recidivism" which may be inherently flawed. By using appropriate techniques and approaches, we hoped that our models were fair, transparent, and accountable, and more contributive to a better justice and equitable society.

7. REFERENCES

- Sentence phrasing – www.wordtune.com
- Sklearn - https://scikitlearn.org/stable/modules
- https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb
- https://www.youtube.com/watch?v=pByOoO7UvgM&ab_channel=AbheeshKhanal
- https://towardsdatascience.com/5-tools-to-detect-and-eliminate-bias-in-your-machine-learning-models-fb6c7b28b4f1#:~:text=To%20check%20if%20your%20machine%20learning%20model%20is%20biased%20or,train%20or%20test%20the%20model
- https://brainnwave.ai/solutions/services/auditing-model-bias/

8. VIDEO LINK

Video Link