CS 584 - Final Project Report on 'Auditing for Bias'

Praveen Reddy Duligunti, (G01355892), pduligun@gmu.edu

Lakshmi Chetana Gomaram Bikshapathireddy, (G01353352), lgomaram@gmu.edu

Hrithikender Reddy Bathula, (G01354518), hbathula@gmu.edu

1 ABSTRACT

We have examined the ProPublica COMPAS dataset to see whether a defendant in criminal case would recidivate. Later, we concentrated on determining whether the given dataset is racially biased and utilized machine learning to improve the model's fairness. To predict the recidivism, we employed machine learning models such as Logistic Regression and Random Forest. We later discovered the bias in prediction by comparing different findings such as false positives, false negatives, and calibrations. As a result, we discovered that the supplied ML Model's prediction is biased against African-Americans(Blacks) and Caucasians(Whites), and we attempted to enhance the model's fairness. By using logistic regression, we have got an accuracy of 73.0 percent whereas it is 69.63 percent with random forest model. We got these accuracies after performing feature selection and making sure there is no complete effect or dominance of only one attribute which is highly correlated to the target attribute.

Additional Keywords and Phrases: Bias, False-positive Rate (FPR), False-Negative Rate (FNR), Fairness, Equal Opportunity, Calibration.

2 INTRODUCTION

Our main motivation is to predict the recidivism and to determine whether the given Dataset is biased, and in this case, if it is biased on race, age, or gender. Evaluation of the models with the 'race' feature is done to examine this bias option and to understand the impact on the models. Finding the bias and improving fairness of the model is one of the important objectives of our project.

The ProPublica COMPAS dataset contains the data of the crimes committed within the last two years. We have taken one of the two dataset variants provided, to examine and predict the possibility of recidivism i.e., the chance of going back to prison within the next two years. Our main question is examining potential bias based on several metrics on a task and evaluating how they change under different circumstances. Also to compare false positive rates, we will first determine whether there is bias in terms of opportunity cost. The second topic we'll address is if there is bias in a different sense, as measured by calibration.

Initially, we have performed the Exploratory Data Analysis to know the relation and comparisons between the attributes. While doing the exploratory analysis, we found the correlations between the attributes and noticed that there are two

attributes that are more correlated towards the target variable. Feature selection is done which is followed by applying the machine learning models with and without the highly correlated attribute and we have observed that there is a lot of difference between the accuracies of the models. 'One hot encoding' is done to convert categorical values into numerical values. Based on the training set, the models we have chosen to calculate the predictions are 'Logistic Regression' and 'Random Forest'. We have compared the model accuracies with and without the 'is_recid' attribute as it was highly correlated (~0.95) with the target variable('two_year_recid'). If we use this attribute as a feature, the model will result in possible overfitting. As a result, we noticed that the accuracy is more (~0.96) when the attribute 'is_recid' is considered and it becomes less (~0.73) when this attribute is not considered. So, we are not considering this attribute as a feature.

To find out the bias with respect to race which consists of various categories like 'African-American', 'Caucasian', 'Hispanic' and others, we have calculated the false positive and false negative rates. This clearly showed that the predictions are biased towards a particular race in different conditions. For providing equal opportunities to all the races, fairer classification model like equal opportunity classifier is used. This will ensure that the bias is reduced.

3 METHOD

The ProPublica COMPAS dataset contains a database of defendants' criminal histories, jail and prison time, demographics, and COMPAS risk assessments of 2 years. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a well-known commercial algorithm that courts and parole boards use to assess the likelihood of a criminal defendant reoffending (recidivism). The COMPAS score for each defendant ranged from one to ten, with ten being the most dangerous. Scores 1 to 4 were given a "Low" rating, while scores 5 to 7 were given a "Medium" rating, and scores 8 to 10 were given a "High" grade. The COMPAS algorithm is skewed in favor of Caucasian defendants and against African-American inmates, according to a two-year follow-up investigation. The error pattern, as measured by precision, is intriguing. Among the given variants of the dataset, we have chosen the one variant with 7214 data entries and 53 features. Based on our analysis, 'African-American', 'Caucasian' and 'Hispanic' are the races which have more than 500 observations.

Firstly, we preprocessed the data to make sure there are no data discrepancies. After that, Correlation is observed to understand how each attribute is related to the target attribute. To make an accurate analysis and for model construction, we perform the feature selection where the relevant subset of the given attributes are selected and used. This helps us in getting rid of the noise and unrelated attributes in data too. Hence, we have selected the features which include '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' for better results and analysis. In our case, handling the numerical data is easier than the categorical data and hence we applied the 'one-hot-encoder' and converted it to numerical data. The dataset has then been split up 70% for training and 30% for testing.

As we are only predicting the Output between 0(Individual will not Recidivate) and 1(Individual will Recidivate), we have selected the Classification model of Logistic Regression. The accuracy from the Logistic Regression model is 0.73, from the random forest model is 0.69. It is clearly observed how our current Machine Learning model's predictions are biased towards the 'African-American' and 'Caucasian' race. We have also analyzed the results of which 'age' and 'gender' are highly likely to go back to jail after two years of serving.

4 Result

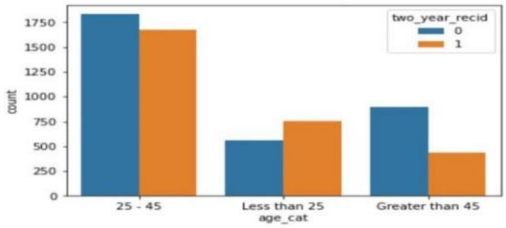

Figure 4.1: Age Category, Count and two_year_recid of individuals

From the above figure we can observe that young individuals below 25 years of age are more likely to Recidivate than elderly people above 45.

Table 1: Comparison with Risk and Race and getting their Counts Percentage

Race	Risk	Count Percentage
African-American	High	27.7
African-American	Low	41.1
African-American	Medium	31.1
Caucasian	High	11.2
Caucasian	Low	65.4
Caucasian	Medium	23.5

As we can observe from the above table that the African-American individuals who have been categorized as High Risk is double to Caucasian. Similarly, Caucasian Race have many Low-risk individuals. The risk factor is based on “Decile Score”. “Decile Score” feature is highly corelated (~0.35) with the target variable(“two_year_recid”).

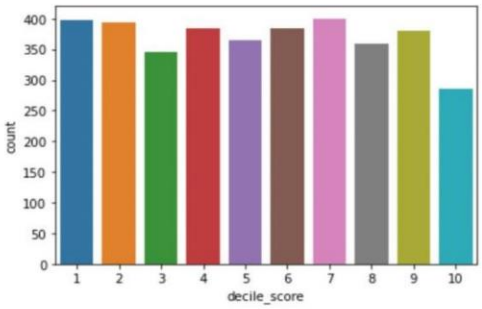

Figure 4.2: Decile Score and Count of African-American Individuals

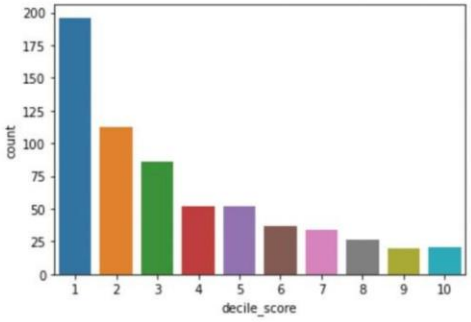

Figure 4.3: Decile Score and Count of Caucasian Individuals

Here, as we can observe in the above figure 4.3 that majority of Caucasian Individuals have very low Decile Scores (1-4) compared to the African-American individuals in figure 4.2. This will result in bias towards Caucasian and will in turn effect our predictions.

Table 2: Comparison with Cross Validation Rates and Race

Race	False Negative Rate	False Positive Rate	True Positive Rate	True Negative Rate
African-American	0.24	0.45	0.71	0.54

Race	False Negative Rate	False Positive Rate	True Positive Rate	True Negative Rate
Caucasian	0.47	0.26	0.53	0.75

False Positive Rate indicates that the individual is predicted positive by our algorithm, but they did not actually recidivate. False Negative Rate indicates that the individual is predicted innocent by our algorithm, but they did recidivate. Due to the bias, we can observe that the African-American individuals have nearly double False Positive Rates compared to Caucasian. This indicates that many African-American individuals did not recidivate after releasing from the jail but our model predicted that they would when compared to Caucasian. If decisions are made based on this prediction, the individuals are expected to serve more jail time even though they are innocent.
Similarly, Caucasian individuals have higher False Negative Rate compared to African-American. This indicates that a Caucasian individual is more likely to recidivate if he is released from jail based on our models prediction.
True Positive Rate indicates that the individual recidivates and is predicted positive by our algorithm. If we observe the above scores the model is biased towards predicting an African-American as an individual who recidivates and a Caucasian person as an individual who does not recidivate. The potential societal implications of using such an algorithm will result in more jail time for African-American individuals, while the Caucasian individuals are more likely to recidivate if they are released from jail based on the model prediction. **"Race" feature does not have much effect on the prediction. But due to the difference in "decile_score", we can observe bias in predictions.**

We can reduce this Bias in the prediction using Classifiers designed to be fairer like Equal Opportunity classifier. We are expecting to reduce the False Positive and False Negative rates using Race as a sensitive column. Using a fair Classifier our Accuracy score has reduced (~0.52) but the model is not biased towards a particular race now.

5 CONCLUSION AND LEARNING

If we look at the overall dataset, we cannot find bias in the predictions. To find the bias, we have to look at all the features individually and compare it with the results. In this case, by considering the role of the "Race" variable as a protected feature, there is no significant difference between the predicted Accuracy with and without "Race" variable. But when individually considering the False Positive and False Negative Rates of the predictions we can clearly observe the Bias in the Race variable between African American and Caucasian individuals. Sometimes, human biases and unfairness in society are evident in the data. Similarly, in recent findings Researchers have observed that Computer Vision algorithms pertained on ImageNet exhibit multiple biases on Race and Gender.

Judges and other Officers are using algorithms to decide which individuals to release on bail or parole (depending on whether they are awaiting trial or serving a sentence). The bias in the prediction towards certain type of race or category will not result in effective decision making and this in turn will have other societal effects. If we observe such bias, we must use classifiers which are designed to be fairer.

6 REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica. May 23rd, 2016. Machine Bias retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Anne L. Washington. 2018. 'How to argue with an algorithm: Lessons from the COMPAS PROPUBLICA debate' from https://ctlj.colorado.edu/wp-content/uploads/2019/03/4-Washington_3.18.19.pdf

[3] Matias Barenstein. August 11th, 2019. https://towardsdatascience.com/the-data-processing-error-in-one-of-the-most-prominent-fair-machine-learning-datasets-4fa205daa3c4

7 VIDEO LINK:

https://youtu.be/kfSefwbMGI8

1 Project Report 75 / 100

+ 100 pts Excellent project and report in all respects! Fantastic work.

+ 95 pts Very good project report with no shortcomings. All tasks performed competently, sound analysis, and good discussion.

+ 90 pts Good project and report with some room for improvement (see comments)

+ 85 pts Good project with scope for improvement in a couple of dimensions (see comments)

✓ **+ 80 pts** *Adequate project and report. Could be improved in several dimensions (see comments)*

+ 70 pts Some deficiencies (see comments)

- 5 pts Major formatting / length violation

✓ **- 5 pts** *Late submission*

(Late submission: Can't give very detailed comments, but happy to discuss if requested)

Your initial steps are reasonable, and the analysis of difference in false positive rates by race is also reasonable, but then I don't see the comparison of calibration or any in-depth analysis or discussion of this. The decile score is just another prediction, so I'm not sure what it is telling you that you shouldn't be able to replicate yourself with your own predictions. I also don't understand your conclusion that "we cannot find bias in the predictions" given the discussion of FPR above.

6 REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica. May 23rd, 2016. Machine Bias retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Anne L. Washington. 2018. 'How to argue with an algorithm: Lessons from the COMPAS PROPUBLICA debate' from https://ctlj.colorado.edu/wp-content/uploads/2019/03/4-Washington_3.18.19.pdf

[3] Matias Barenstein. August 11th, 2019. https://towardsdatascience.com/the-data-processing-error-in-one-of-the-most-prominent-fair-machine-learning-datasets-4fa205daa3c4

7 VIDEO LINK:

https://youtu.be/kfSefwbMGI8

2 Project Video **95 / 100**

+ 100 pts Check plus. Good video, no problems.

✓ **+ 95 pts** *Check. Adequate.*

+ 90 pts Check minus. Some issues (see comments)