Auditing for Bias

Sai Rohith Pasham, G01348426

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

Srikanth Reddy Dubba, G01353043

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

Ramaswamy Iyappan, G01348097

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

Across the world, judges and parole officers are increasingly using machine learning algorithms to assess a criminal defendant's likelihood to re-offend. It is essential to ensure that the models that are developed using machine learning are not discriminatory towards a particular group or population. In this experiment of Auditing for bias, we examined about the COMPAS dataset which was collected by ProPublica. The dataset contains features such as race, name, age, decile score, days of arrest, etc. and the target feature was two_year_recid, which tells us whether the defendant has recidivated in the next 2 years. By comparing different classifiers, we designed a random forest classifier suitable for this dataset that predicts if the defendant will recidivate in the next two years or not and got an accuracy of 90% on predicting a fixed validation set of the original dataset. The major task was to examine if there is any bias in terms of opportunity cost and also check whether there is any bias in a different sense, by comparing calibration. We analyzed about more than 7000 criminal defendants and observed that about 56.7% of the incorrectly predicted recidivism rate (false positive rate) by the Random Forest classifier were categorized as African-American by the race variable, but only 24.74% of the incorrect prediction belong to the Caucasian race category. Whereas, when comparing calibration, the probabilities for African-American and Caucasian are 90.00% and 92.18% respectively, indicating that using calibration as a metric for prediction in this domain is more fair to classify defendants irrespective of their race. Next, by considering "race" as a protected feature and carrying out the same experiment, we observed that there was no significant change in the probabilities of the false positives and calibration, i.e., the model did not discriminate on the basis of this race feature. Lastly, we used a classifier called Demographic parity classifier as a fair model which showed an accuracy of 91% in predicting the true recidivates of the validation set. The comparison of results given by this fair classifier and those given by our model, provided insights for both of them to be similar. But still, from our analysis of comparing opportunity cost in depth, we found that our model and the fair classifier was biasing against the African-Americans and is more likely to misclassify African American as high risk over Caucasians based on opportunity cost.

1 INTRODUCTION

There are dozens of these risk assessment algorithms in use. Our objective was to experiment with the COMPAS dataset, which is one of the most popular algorithms used by the judges in the United States in pretrial and sentencing, for scoring criminal defendant's likelihood of recidivism. This was meant to produce a fair prediction in a diverse population and not be biased. Overall, our main objective is to check if there is any bias associated against any particular demographic. We set out to assess the "compas-scores-two-years" variant of the available datasets under ProPublica's COMPAS, which includes jailing and background information of about 7214 defendants along with 53 features including their status of recidivism within the next two years of decision. The dataset was put together by ProPublica and comprises of about two years' worth of compas scores of criminals from the Broward County of Florida who were scored in 2013 and 2014.

We carried out a number of steps for preprocessing the dataset and converting them into numerical data which is easy to handle and be useful for predictions by machine learning models. Then, we split the dataset into train and test data where two_year_recid was the output column and the rest were taken as input. Next, we used the train data on different classification algorithms to find which model was appropriate to this domain. We then used cross validation and hyper parameter tuning using Grid search CV to improve the performance of the model. We found that the random forest classifier performed better among all the above as it produced the highest accuracy, precision, recall and f1 score.

Next, we move on to carry out our tasks of checking if there is any bias in terms of opportunity cost and calibration. We found that the model misclassified African Americans as recidivated almost twice as much as it incorrectly classified Caucasians. This shows us there is bias associated with the model in terms of opportunity cost.

We checked if there is any bias in different terms such as calibration and found that the model was fair in classifying defendants who recidivated irrespective of the race variable. Next, we considered the race attribute as protected and tried training the model with the rest. We applied the same steps and observed that even by removing the race variable there is no significant impact on the performance of the model as well as not much of a change in the predictions. This shows that the model does not bias against any criminal over the other in terms of race. There could be other reasons for the bias like decile score, sex etc. As a last experiment, we used a fair classifier called Demographic parity classifier for training the same dataset and observed that the accuracy of prediction improved to 91% and has performed slightly better than our previous model in terms of opportunity cost and calibration.

2 METHOD

Compas scores for each defendant was mentioned in the "decile_score" column ranging from 1 to 10 where, scores 1- 4 were labeled as "Low Risk", scores 5-7 were labeled as "Medium Risk" and scores 8-10 were labeled as "High Risk". The race of the defendants was identified as African American, Caucasian, Hispanic, Asian, Native American and others, by using the race classifications used by Broward County Sheriff's office. More information about the criminals was also included in the 53 features such as first & last name, screening date, sex, dob, juvenile score, number of days in jail, charge degree, etc.

Machine learning algorithms can process only numerical data. Since the dataset was mostly having categorical values (textual form of data) and did not include values for many fields (Null values), it is impossible for machine learning algorithms to process this dataset in its raw form. Hence some preprocessing is needed to clean up the raw data by removing irrelevant or Null fields and converting categorical values to numerical values. At first, we visualized the dataset in terms of sparsity using the matrix () function of missingno library and found that about 22 of the 53 features in the dataset was having null fields and was also irrelevant to this experiment. So, we had to drop those 22 features from the dataset. With the remaining attributes, we performed feature selection by comparing correlations between the columns as well as compared that result to that of applying a dimensionality reduction metric such as PCA. Finally, we arrived at a sweet spot of having 18 attributes in total including the target variable "two_year_recid".

Categorical data in the remaining features was then converted into numeric values using the label encoder and One Hot encoder. By now, having the clean form of dataset to be easily processed by machine learning algorithms, we split the dataset using train test split from Sci-kit Learn library, to holdout about 1/3 (33%) of the dataset as a fixed validation set (not at random by using random_state) for testing the learned model. So, from here, any model learnt will be based on a fixed proportion of 67% of the original dataset, which will be used to test on the remaining unseen 33%.

For selecting a suitable classifier for predicting whether a defendant recidivated in the next two years or not, we experimented learning the training dataset and predicting the validation set using a number of models such as Random Forest classifier, Logistic Regression, K Nearest Neighbors and XG Boost models.

Table 1: Accuracy of predicting the validation set trained using different models

Classifier	Accuracy (%)
Random Forest	90
Logistic Regression	90
KNN	74.50
XG Boost	89.92

After comparing the accuracy, precision and f1 score by these models, it was observed that the Random Forest Classifier produced the best results of predicting the validation set. Hence, we chose the Random Forest Classifier as our base model for carrying out this experiment.

3 RESULT

3.1 Task 1:

To check whether there is any bias in terms of opportunity cost, by comparing False positive rates.

Table 2: Analysis of False Positives by the Random Forest Classifier

Race	High Risk	Medium Risk	Low Risk	Total by Category
African-American	25	17	13	55
Caucasian	4	6	14	24
Total Number of False Positives by the model				97

From table 2, it is observed that 97 defendants in the validation set were incorrectly classified as positive (recidivated) while they didn't actually recidivate as per the true values in the unseen test set. Of those, 55 were categorized as African-American and 24 were categorized as Caucasian by the 'race' variable. In order to check for bias, we compare the probability that an individual is predicted positive by the model given that they did not actually recidivate and are categorized as 'African-American' by the race variable, to the probability of an individual predicted positive given that they did not actually recidivate and are categorized as 'Caucasian' by the race variable.

Probability [predicted +ve | did not actually recidivate, African-American] = 55/97 = 56.7%
Probability [predicted +ve | did not actually recidivate, Caucasian] = 24/97 = 24.74%

By comparing the probabilities, we could clearly see that the model is biased against the African-American defendants, since it misclassifies them as recidivated almost twice as much as it misclassifies the Caucasian defendants. Also, notice that majority (42 individuals) of the misclassified African-Americans are labeled either 'High Risk' or 'Medium Risk' while they did not commit any crime in two years, whereas only a small portion (10 individuals) of the misclassified Caucasians is labeled 'High Risk' or 'Medium Risk' and majority of them as 'Low Risk'. Therefore, it is evident that the model is highly biased against the Black (African-American) defendants since it is more likely to misclassify them as re-offended and mark them as high risk and acts in favor of the White (Caucasian) defendants by marking them as low risk even when it less likely misclassifies them.

Using this metric (opportunity cost) with such an algorithm in times of pretrial and sentencing might be critical and dangerous, since it is more likely to criticize individuals again and again even though they proved to be innocents, which can change the meaning of justice overall, and the defendants experiencing such kind of treatment may get frustrated of the false accusation which might lead them into committing new crimes. Hence scoring criminals using opportunity cost with this algorithm might just increase the crime rate rather than controlling it.

3.2 Task 2:

To check whether there is any bias in terms of calibration.

Table 3: Analysis of Calibration by the Random Forest Classifier

Race	African-American	Caucasian
Predicted positive	567 recidivated 63 did not recidivate	271 recidivated 23 did not recidivate
Total predicted +ve	630	294

From table 3, we can observe that out of 630 defendants that were predicted positive by the model categorized as African-American by the race variable, 567 individuals actually recidivated and 63 did not. Similarly, out of 294 defendants who were predicted positive by the model and categorized as Caucasian by the race variable, 271 individuals actually recidivated and 23 did not. Hence, by comparing the probabilities, we have:

Probability [individual recidivates | predicted +ve, African-American] = 567/630 = 90.00%
Probability [individual recidivates | predicted +ve, Caucasian] = 271/294 = 92.18%

Both the probabilities turned out to be so close, indicating that the model correctly predicts almost similar proportion of defendants as re-offended irrespective of their race. Hence, the algorithm is not biased towards one or the other group based on this measure (calibration) and proves to be fair enough based on the same. Using calibration as a metric with such an algorithm is very helpful and fair in predicting criminal defendants' likelihood of recidivism, where outcomes are independent of protected features such as race, color, etc., conditional on risk estimates.

When comparing results by the two metrics from task 1(opportunity cost) and task 2 (calibration), it can be clearly seen from the above statistics that considering calibration as a metric for predicting recidivism of the defendants proves to be more fair than using opportunity cost as a metric, since the model discriminates based on one's race while predicting using opportunity cost, whereas using calibration cuts down this influence of protected features and predicts in terms of risk factor rather than discrimination. Hence using calibration as a metric is more effective in this domain.

3.3 Task 3:

To consider the 'race' variable as a protected feature and observe whether it changes the results.

Table 4: Analysis of False Positives

Race	Total by Category
African-American	64
Caucasian	28
Total false positives	118

Table 5: Analysis of Calibration

Race	African-American	Caucasian
Predicted positive	569 recidivated 68 did not recidivate	271 recidivated 29 did not recidivate
Total predicted +ve	637	300

Probability [predicted +ve | did not actually recidivate, African-American] = 64/118 = 54.23%
Probability [predicted +ve | did not actually recidivate, Caucasian] = 28/118 = 23.72%
Probability [individual recidivates | predicted +ve, African-American] = 569/637 = 89.32%
Probability [individual recidivates | predicted +ve, Caucasian] = 271/300 = 90.33%

It was observed that there is no significant change in the opportunity cost or calibration, even after removing the 'race' variable. This means that the model does not discriminate, and the outcomes are fair with respect to the race feature. However, using False positive rates as a metric still seems to bias against the African-American defendants. This implicit bias in the dataset might be due to the influence of other protected factors such as gender, decile_score, is_recid, etc.

3.4 Task 4:

To carry out the whole experiment using a classifier designed to be more fair such as Demographic Parity Classifier from Scikit-lego library.

Table 6: Analysis of False Positives

Race	Total by Category
African-American	34
Caucasian	18
Total false positives	62

Table 7: Analysis of Calibration

Race	African-American	Caucasian
Predicted positive	553 recidivated 34 did not recidivate	265 recidivated 18 did not recidivate
Total predicted +ve	587	283

Probability [predicted +ve | did not actually recidivate, African-American] = 34/62 = 54.84%
Probability [predicted +ve | did not actually recidivate, Caucasian] = 18/62 = 29.03%
Probability [individual recidivates | predicted +ve, African-American] = 553/587 = 94.20%
Probability [individual recidivates | predicted +ve, Caucasian] = 265/283 = 93.63%

We found that the Demographic parity classifier provided much better results in terms of both opportunity cost and calibration. However still the presence of implicit bias in the dataset discriminates the African-American criminals over the Caucasians.

4 CONCLUSION

From our analysis, we learned that our model's prediction is biased in terms of opportunity cost as the false positive rates for African-Americans is more than that of the Caucasians. We found that the predictions were similar between both the races and fair enough in terms of calibration, which predicted the defendant's likelihood of recidivism in terms of risk rather than discriminating based on race. Also, the model's predictions were not significantly impacted when the race attribute was considered to be a protected feature.

REFERENCES

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[2] https://github.com/propublica/compas-analysis

[3] https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

VIDEO PRESENTATION

https://drive.google.com/file/d/1JgL-_79f6iW_rZdj6JGuglwpQMjqtfsO/view