Auditing for Bias

Shreenath Sivadas, G01269232

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

Sreeram Bangaru, G01406291

Master of Science in Computer Science, George Mason University, sbangar2@gmu.edu

Sai Sirichandana Kanuri, G01371765

Master of Science in Computer Science, George Mason University, skanuri4@gmu.edu

As the use of machine learning algorithms to predict the likelihood of criminal defendants to re-offend increases, it becomes crucial to ensure that these models are free from bias towards any particular group or population. In this study, we focused on the COMPAS dataset, which includes features such as race, name, age, decile score, days of arrest, and two_year_recid, indicating whether the defendant has recidivated within two years. We aimed to develop supervised learning models suitable for this dataset, we implemented multiple models like Logistic Regression, Decision Tree, Random Forest and Adaboost out of which Logistic Regression achieved an accuracy of 67% on a fixed validation set. To assess bias, we examined opportunity cost and calibration. We observed that the false positive rate for logistic regression classifier was higher for African-Americans than for Caucasians. However, when we compared calibration, we found that true positive rates for African-Americans and Caucasians were 66.7% and 70%, respectively, using logistic regression classifier indicating that using calibration as a metric for prediction would be fairer. We also conducted the same experiment with "race" as a protected feature and found no significant change in false positives, indicating that the model did not discriminate based on race. Finally, we compared our logistic regression model with a Demographic parity classifier, which showed comparable results. However, when we delved deeper into opportunity cost, we found that both models were biased against African-Americans and more likely to misclassify them as elevated risk compared to Caucasians. In conclusion, while our models and the fair classifier showed similar results, it is essential to consider opportunity cost when designing a fair model for predicting recidivism, especially in cases where there are significant differences in the proportions of races in the dataset.

1 INTRODUCTION

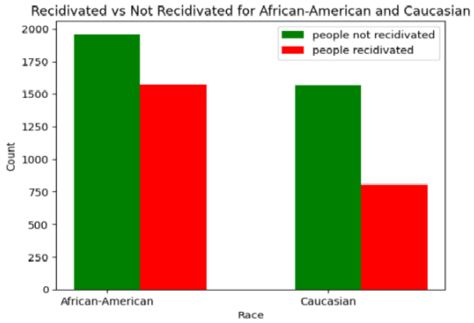
This dataset, compiled by ProPublica, contains jail and background information of over 7,000 defendants in Broward County, Florida, who were scored in 2013 and 2014, along with 53 features including their status of recidivism within the next two years of decision. To prepare the dataset for machine learning models, we preprocessed it and used various classification algorithms to find the best model for this domain. We found that the logistic regression model outperformed other models, achieving the highest accuracy. We then evaluated the model for bias in terms of opportunity cost and calibration. And then repeated the experiment with the race variable removed, which did not significantly impact the model's performance or prediction. Finally, we used a fair classifier called the Demographic Parity Classifier on the dataset.

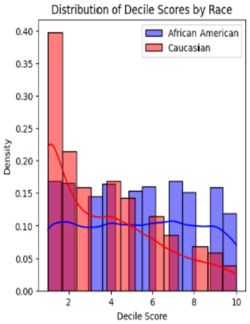
2 METHOD

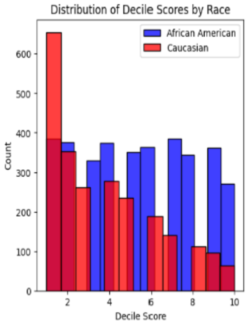
To prepare the dataset for training, we pre-processed it. The original dataset contained 53 columns. Columns that had irrelevant data / textual values were not considered for the final training dataset. Several rows containing null values for many fields were also removed as a machine learning model cannot work on data that holds a null field. By understanding the ProPublica's COMPAS analysis we have noticed that some features were referring to the future information such as 'is_recid' which indicates whether the defendant was re-arrested withing two years after their assessment hence is not included in the final training dataset as this may result in poor generalization on unseen data and might overfit the model. Several textual data columns were also encoded using label encoder and One hot encoder to integers for the model to make sense. The final columns we are considering for training the model are:

'age', 'days_b_screening_arrest', 'decile_score', 'two_year_recid', 'c_charge_degree_F', 'c_charge_degree_M', 'race_African_American', 'race_Asian', 'race_Caucasian', 'race_Hispanic', 'race_Native_American', 'race_Other', 'age_cat_25-45', 'age_cat_Greater than 45', 'age_cat_Less_ than 25', 'score_text_High', 'score_text_Low', 'score_text_Medium ,'sex_Female', 'sex_Male'

To get an understanding of the data and its distribution, we visualized the number of people who recidivated and who did not for both African American and Caucasian races. Also, to get a better understanding we plotted the decile score distribution for both the races. From the below figures it is evident that there is a higher number of Caucasians in the lower decile score bracket while the number of African Americans is higher in the higher decile score bracket. We can see that the dataset contains a higher number of African Americans in both the categories when compared to Caucasians.







Now we split the dataset into 1/3 and 2/3 for testing and training purposes, respectively. We implemented several supervised learning models namely Logistical regression, Decision tree classifier, Random Forest classifier and a Adaboost classifier (using Decision stump as weak classifier) then calculated their accuracies. From the below values we can see that Logistic regression is giving better accuracy results than other models. Hence we are going to consider this as the primary model in checking whether there the predictions are biased or not.

The below table shows the accuracies of different models that we implemented.

Classifier	**Accuracy**
Logistic Regression	67%
Decision Tree	64%
Random Forest	65%
Adaboost	61%

3 RESULTS

a) Task 1:

To check whether there is any bias in terms of opportunity cost, we compared the False positive rates.

Logistic Regression	Predicted recidivated	Predicted did not recidivate
Recidivated	598	351
Did not recidivate	306	782

False positive: African American – 201

False positive: Caucasian – 72

Probability [predicted +ve |, did not recidivate and is African American] – 201/306 = 65%

Probability [predicted +ve | did not recidivate and is Caucasian] – 72/306 = 23%

From the above tables and probabilities, we can see the false-positive values for the African American and Caucasian races for the Logistical regression classifier. It is evident that the majority of the false positive rates are categorized as African Americans when compared to Caucasians. This suggests that there is a bias towards African American race.

The use of opportunity cost as a metric in sentencing a person can pose significant risks. It can lead to the repeated condemnation of individuals even after they have been proven innocent, undermining the fundamental principles of justice.

b) Task 2:

To check whether there is any bias in terms of calibration. We compared the true positives for the African American and Caucasian races for the classifier and calculated the probabilities for the same.

Logistic Regression	African American	Caucasian
Predicted positive	403 recidivated	146 recidivated
True Label	201 did not recidivate	72 did not recidivate
Total	604	218

Probability [individual recidivates | predicted +ve, African American] – 403/ 604 = 66.7%

Probability [individual recidivates | predicted +ve, Caucasian] – 146/ 218 = 70%

The above results show that there is still bias towards African-American race, but when compared to the results obtained using the opportunity cost metric, calibration yields much better results. The calibration metric is particularly useful for predicting recidivism since it provides independent outcomes based on risk estimates and does not rely on protected features such as race or color. The opportunity cost metric can lead to discriminatory outcomes based on race, whereas calibration limits the impact of protected features and focuses on risk factors to produce fairer predictions. Therefore, utilizing calibration as a metric is a more effective approach in predicting the likelihood of recidivism.

c)Task 3:

We did the same analysis removing the race column to see how it affects the predictions and accuracy of the model. The below table summarizes the above said calculations.

Classifier	**Race**			**Accuracy**
		False Positives	True Positives	
Logistic regression	African American	200	403	67%
	Caucasian	71	142	

[Note: Total False positives = 305]

Probability [predicted +ve |, did not recidivate and is African American] **-** 200/305 = 65.57%

Probability [predicted +ve | did not recidivate and is Caucasian] **-** 71/305 = 23.28%

The above values show us that there is no significant change when the race feature is removed and the bias towards African American race still exists. Also, the feature analysis on the model showed us that the race feature has very little impact compared to other features in deciding the model performance.

d)Task 4:

Now we do the same analysis by implementing a fair classifier. We implemented a random forest classifier using Exponentiated Gradient optimizer to calculate the fairness measures: demographic parity and equalized odds and then calculated the false positive and true positive values for both the races.

Logistic Regression	African American	Caucasian
Predicted Positive	339 recidivated	185 recidivated
False Positive	180	140
Total Predicted Positive	519	325

[Note: Total False positives = 376]

Demographic parity difference: 0.193

Equalized odds difference: 0.047

Accuracy: 64%

Equalized odds ensure that the true positive rate and the false positive rate are equal between different groups defined by a sensitive attribute (race in this case) and hence it was used as a constraint for the fair model. The demographic parity difference and equalized odds difference values suggest that the classifier is fair but there is still a slight bias in the model. In pursuit of achieving fairness there is a tradeoff

Probability [predicted +ve |, did not recidivate and is African American] – 180/376 = 47.8%

Probability [predicted +ve | did not recidivate and is Caucasian] – 140/376 = 37.2%

Probability [individual recidivates | predicted +ve, African American] – 339/519 = 65.31%

Probability [individual recidivates | predicted +ve, Caucasian] – 185/519 = 56.92%

From the above probabilities, the Demographic parity classifier gives us better results but there is still a bias in the model which is due to the implicit bias in the dataset which is corroborated from our visualization and analysis done in the Method section of this report. Also, it is evident that considering the calibration metric yields much fairer results.

From the Demographic parity difference and the equalized odds difference values from above, we can infer that as the values are closer to 0 hence the fair classifier is able to predict with very less chances of bias. So, the fair classifier can reduce bias significantly with respect to the true positive rate and false positive rate when compared with logistic regression model. Finally, if we observe as we have focused mainly on making our classifier fair we have faced a backlash on the accuracy as it has been reduced from 67% to 64%. Hence there is a tradeoff between reducing bias and the accuracy

4 CONCLUSION

Our investigation revealed that our model's assessment of recidivism risk was biased in terms of opportunity cost as there was a higher false positive rate for African Americans compared to Caucasians. However, we discovered that in terms of calibration, although there is still bias, the model gave much better results. Furthermore, we found that the model's predictions remained unaffected when we designed the race attribute as a protected feature. This implies that the model did not unfairly discriminate against any defendant based on race.

REFERENCES

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

VIDEO PRESENTATION: **https://www.youtube.com/watch?v=Bp5i2nbcFzc**

SOURCE CODE: **https://github.com/shreenath96/Supervised-learning-models-on-COMPAS-Dataset**