Final Project of CS – 584

"Auditing for Bias on ProPublica COMPAS dataset"

Possibility of a recidivist criminal defendant, we are using Logistic Regression and Random Forest models to analyze the bias on COMPAS dataset.

Uday Kumar, UK, Kamalapuram

Computer science grad student at George Mason University, ukamalap@gmu.edu

Preeti, PB, Bhattacharya

Computer science grad student at George Mason University, pbhatta2@gmu.edu

Sandeep, SP, Potturi

Computer science grad student at George Mason University, vpotturi@gmu.edu

In our examination of Northpointe's COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) program indicated that black defendants were significantly more likely than white defendants to be mistakenly evaluated to be at a higher risk of recidivism, while white defendants were far more likely to be incorrectly classified as low risk. In this project, we are looking for evidence that the ProPublica COMPAS dataset favors white over black persons. We have examined more than 7000 criminal records and analyzed the various features from the given data. We filtered the various feature columns and implemented the Logistic Regression and Random Forest models on selected features to predict whether the defendant will recidivate in two years or not. From the results of our model accuracies and False Positive Rate, we discovered that black defendants were significantly more likely to misrepresent than white defendants to be wrongly rated as having a higher risk of recidivism, while white defendants were far more likely than black defendants to be incorrectly marked as having a low risk. So according to our model analysis and dataset, bias exists but is not solely dependent on race, but also on other COMPAS data set attributes because the accuracy of this model did not alter significantly once the race column was removed.

1 INTRODUCTION

Criminal recidivism occurs when a formerly imprisoned person returns to crime. The criminal justice system is continuously seeking ways to better anticipate which criminals are more likely to remain lawful after release and who are at risk of recidivism. Bias can lead to mistakes by erroneously weighing variables, and the criminal justice system is taking the help of machine learning to improve recidivism predictions. But there are chances of bias in the model trained by discrepancy data.

Risk assessments, or scores from machine learning model predictions, are used to guide choices about who can be released at every level of the criminal justice system, from determining bond amounts. The results of such assessments are presented to judges in different states of the USA during felony sentences. Based on its analysis, focusing on one set of predictive metrics, ProPublica concluded that the COMPAS risk score was biased against African Americans. The bias in the two-year datasets is clear, there are a disproportionate number of recidivists. In this project we are on the lookout for if there is a difference in prediction between African and Caucasian races, we are trying to figure out what the bias is (and which race the model is biased towards) and how that influences the final prediction results. What effect will there be on the model's performance if we do not consider race features?

In this project, we have chosen ProPublica COMPAS data set, consisting of two years' worth of COMPAS scores from the Broward County Sheriff's Office in Florida. We analyzed more than 7000 criminal defendants' data and their predicted recidivism rates with the rate that occurred over a two-year period. At least three COMPAS scores were assigned to each pretrial defendant: "Risk of Recidivism," "Risk of Violence," and "Risk of Failure to Appear." Each defendant's COMPAS score ranged from 1 to 10, with ten being the highest risk. COMPAS classified the scores from 1 to 4 as "Low," 5 to 7 as "Medium," and 8 to 10 as "High."

Firstly, our specific question is to find any discrepancy in the prediction of being recidivated between African-American and Caucasian races, and in particular, we are trying to identify the bias in misclassifying the predictions i.e., by calculating

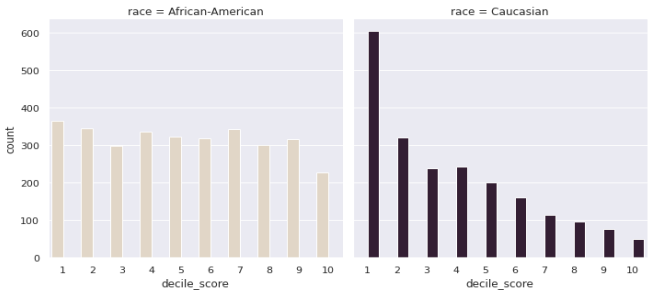
the false-positive rates. Secondly, we want to find out the impact using metrics like model accuracy if we do not consider the race feature.

Our approach to solving the above-stated question is, we intend to identify whether there is a bias predicting more black defendants to be recidivated by creating a model based on Logistic Regression and Random Forest techniques. We want to understand the behavior of the COMPAS data set on our models and document how the results turnout and check the accuracy for various scenarios (like considering race and not considering).

2 METHOD

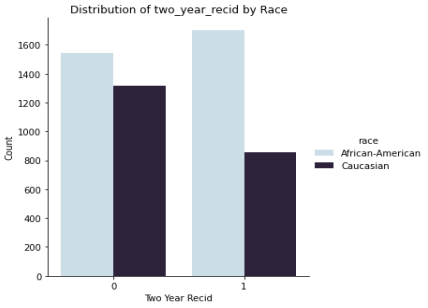
We have taken the ProPublica COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions) dataset. This includes 7214 records and 53 features that are used to predict recidivism over the course of two years. COMPAS is a key dataset for investigating algorithmic (un)fairness. This information was used to forecast recidivism (whether or not a criminal will re-offend) in the United States. The program was designed to overcome human biases by providing an automated, fair approach for predicting recidivism in a varied population.

Key columns in the dataset include age, race, c_charge_degree, score_text, sex, decile_score, two_year_recid, and so on.



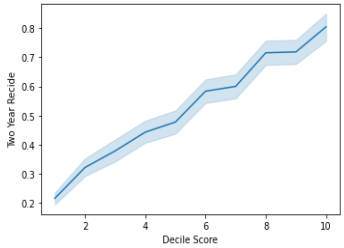
We started by looking at the risk of recidivism score. Our initial analysis is focused on the simple distribution of COMPAS decile scores among the African American race and Caucasian races. We plotted the distribution of decile scores vs the number of African American and Caucasian defendants. (Image in left).

The following representations describe that white defendants' scores tend towards lower-risk categories, whereas black defendants' scores were uniformly distributed across the board. The following image represent the distribution of two_year_recid for both African American and Caucasian races. (Image in right).



We preprocessed our dataset because it was not always evident which criminal case was linked to an individual's COMPAS score. We considered cases with arrest or charge dates within 45 days after a COMPAS evaluation to link COMPAS scores with related cases. In certain cases, we found some NaN values and we filtered the respective columns by using the dropna() function. There are some categorical values in our selected features i.e., ('score_text', 'age_cat', 'sex', 'race', 'c_charge_degree') and have converted those values into binary categorical values by using pandas.get_dummies() function and we have given new columns. We have taken two_year_recid as a target label ‘y’ for our model and dropped two_year_recid from our X_Train data. Further, we have split the train data using train_test_split from sklearn.model_selection i.e., ⅓ as test data and ⅔ as train data. We have standardized the x_train_data using standardscalar.fit_transform() function for implementing a logistic regression model.

When we also analyzed the relationship between the decile_score and two_year_recid column we got to know that the defendant chances of getting decile in two years is high, which is represented in the graph to the left. In this project, we have implemented two different models to predict fairness. The first model is a Logistic regression that overestimates odds ratios in studies with small to intermediate sample sizes.



On the other hand, Random Forest employs many techniques to reduce variance in predictions while retaining (to some extent) the low variance that was distinctive of the lone Decision Tree. It accomplishes this essentially by averaging a series of extremely weakly connected (if not totally uncorrelated) trees.

3 RESULT

When we run our model implemented based on Logistic Regression with a 45-day compass charge date from the test data, the accuracy is 95.31%. The FPR for the entire model is 0.0729.

When we consider African American test data, the accuracy is 94.31% and the FPR for the African American race is 0.109. Whereas when we filtered the test data for the Caucasian race, we got an accuracy of 96.3% and the FPR for the Caucasian race is 0.038. By comparing the FPRs, when we calculated the probability of an individual that is predicted positively by our model but actually he did not recidivate and categorized as African American by race is 0.109 to the probability of an individual to be predicted positively by our algorithm but who did not actually recidivate and categorized by Caucasian by race is 0.031. When we calculated these values taking the FPR of African Americans as the numerator and the FPR of the Caucasian race as the denominator, we got a ratio of 3.51. So, from these results, we have observed that there is 2.5 times (i.e., 250%) more chance of misclassifying African Americans as recidivate even though they did not recidivate compared to the Caucasian race. The interesting observation from the results is that the False Negative rate in the Caucasian race is 0.031 whereas in African Americans it is 0.0071. When we compare these values, we can see that there is a chance of 3.36 times higher representing Caucasians as non-recidivates even though they are recidivate compared to the African American race. This leads to a significant effect on the African American race which represents a bias in misleading the African race as more recidivates even though in the ground truth, they do not recidivate.

When we compared the FPR of African American and Caucasian races resulting from the random forest model, we got a similar bias as in Logistic Regression. There was a 1.64 times chance of misclassifying the African American defendants as recidivists compared to Caucasian defendants.

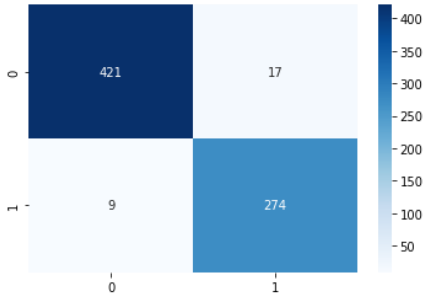
Model	**Logistic Regression**	**Random Forest**
Accuracy	95.31	96.41
False Positive Rate of whole model	0.0841	0.057
False Positive Rate of African Race	0.071	0.082
False Positive Rate of Caucasian Race	0.041	0.031
Accuracy without Race Feature	95.79	97.1

To examine if there is a bias in a different sense, by comparing calibration, we have calculated the probability of an individual getting recidivated when our model has predicted positive is 0.56 for African American race. Whereas for the Caucasian race, the probability of an individual getting recidivated when our model predicted positive is 0.403. From these results, we have learned that there is a 39% higher probability that an African American defendant will be predicted to recidivate compared to a Caucasian defendant, which indirectly represents the bias against the African American race. From the observations so far, there is a clear bias indicating toward the Caucasian race. It would affect the African Americans of getting recidivated when the lawmakers follow this prediction data. This is an example of the societal implication of using a prediction algorithm.

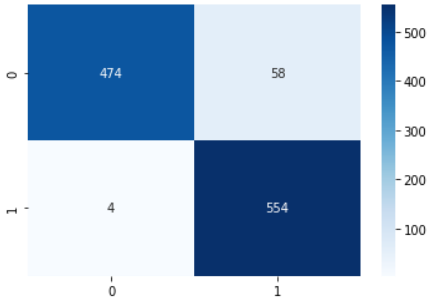
Similarly, when observing the results from the random forest model, and when tried examining the bias in a different sense by comparing calibration, the results indicated that there is a 34% higher probability that an African American defendant will be predicted to recidivate compared to a Caucasian defendant, the bias was present against African race.

The following are the confusion matrix images generated in the Logistic Regression Model:

African-Race

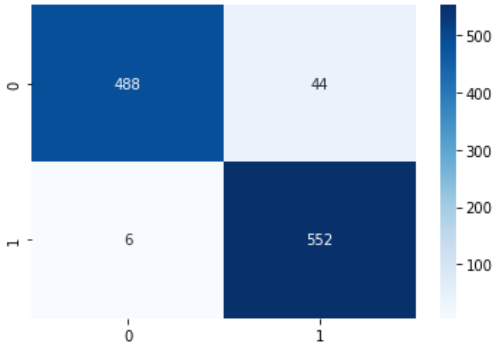


Caucasian- Race

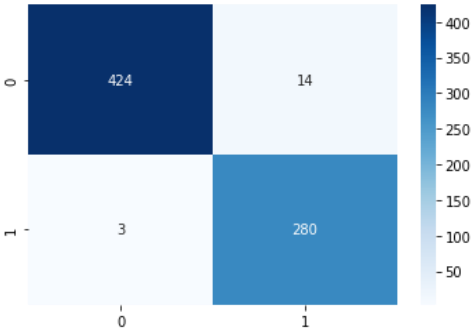


The following are the confusion matrix images generated in Random Forest Model:

African-Race



Caucasian-Race



The Classification Report of Logistic Regression Model with Race feature:

	precision	**recall**	**f1-score**	**support**
0	0.99	0.93	0.96	1110
1	0.92	0.99	0.95	927
accuracy			0.95	2037
macro avg	0.95	0.96	0.95	2037
weighted avg	0.96	0.95	0.95	2037

In further examination of our result, when we selected the Race column as a protected feature and removed the race column from the training data of a model, we got a similar results in terms of accuracy of a model.

In the Logistic Regression model, we got an accuracy of 95.79% which is almost similar to model with Race feature.

We got similar results in Random Forest model when we excluded Race feature there is similar accuracy as 97.1 without race feature and 96.41 with race.

4 CONCLUSION

What we discovered:

A model can be biased in terms of age, ethnicity, and gender if those characteristics are not included as input to the model. There are numerous metrics of fairness, and it may be impossible to meet all of them at the same time. Human biases and inequity in society penetrate the data used to train machine learning models. From the above result, we can say that the accuracy of this model did not alter significantly once the race column was removed. According to the model analysis and dataset, bias exists but is not just dependent on race, but also other variables in the COMPAS data set.

We would have used multiple models and characteristics to locate the bias, and I would have asked numerous questions and evaluated various scenarios within our data to find the bias in our dataset. For example, we would have examined whether the performance of our model changes if one data point changes, or if a different sample of data is used to train or test the model.

The interesting fact was that there is a bias toward the African race in our model selection. Initially, we thought that if we remove the race feature from our model the prediction might behave differently but when we make the race column a protected feature and train the model again with the new data set it gives a similar result as earlier. So, it was very interesting that the bias against African Americans was not solely based on the race feature, but it was through the entire dataset. Hence, we can conclude that these biased predictions through machine learning models in real-time scenarios are not good if they are biased or do not yield great accuracy. As a dangerous criminal could walk free if something goes wrong in one direction. If it is incorrect in one direction, it could lead to someone obtaining an unfairly harsher sentence or having to wait longer for parole than necessary.

VIDEO LINK : https://youtu.be/Dlr3_DJAe7U

REFERENCES

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

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

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] http://blog.kennylee.info/projects/python/data/machinelearning/bias/2020/11/01/analyze-Compas.html

[5] https://scikit-lego.readthedocs.io/en/latest/fairness.html