Group 35
Jae Hyeok Kwak
Sachin Tiwary
Manya Singh
CSE-574

Programming Assignment 3 Report
Fairness in Machine Learning Role*: Volunteer for Non-Governmental Organization*

1. **What is the motivation for creating a new model to replace COMPAS? What problem are you trying to address?**
   COMPAS is a Risk assessment instrument developed by Northpointe and is used to rate a defendant's risk to become a recidivist. This RAI is used by many states at nearly every step from sentencing to parole. Propublica, a non-profit news Organization reviewed COMPAS by taking Broward county's data for 7000 arrested people into account, and checked if these people went on to commit crimes in future as well. According to the Propublica's story, the predictions were different based on the races, showing the model was biased.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

   It proves that "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend." Likewise, incorrect information leads to incorrect conclusions. Therefore, for our new model, the main motivation is to avoid such unfair ethical bias that happens in the data. Also, we as the volunteer for NGO, we want to focus on the accuracy rather than the cost, and look at every ethnical groups.

2. **Who are the stakeholders in this situation?**
   Investors, NorthPointe, ProPublica and other NGOs, The US jurisdiction system(Judges, Attorneys, Defendants, Accused), Government and AI businesses could be identified as the stakeholders.

3. **What biases might exist in this situation? Are there biases present in the data? Are there biases present in the algorithms?**
   Since Nortpointe has refused to disclose the details of the algorithm, we cannot say if there was a bias in the algorithm or not, but there certainly was a bias in the data, be it the data Propublica tested the COMPAS with, or how the model was trained. The following bias in the data were identified:
   - To evaluate COMPAS, Propublica used Broward County's data which can be seen as picked up in a non-random manner, hence it can be said that there is a possibility of Sampling Bias to occur.
   - According to Propublica's report, "Northpointe's core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records." The questions like:
     - "Was one of your parents ever sent to jail or prison?" : may result in a detection bias, because there is a possibility of overestimation, that the kid would also attempt to commit a crime.
     - "How many of your friends/acquaintances are taking drugs illegally?": It's not certain if the defendant would correctly answer this question, hence a Response Bias can occur.

4. **What is the impact of your proposed solution?**

The proposed solution is Naive Bayes classifier with Equal Opportunity Post-Processing method. As a member of a Non Profit Organization, the sole purpose is to ensure social justice and improve accuracy. Seeing most of the matrices like TPR, FNR, TNR the disparity among races is pretty low.

	Threshold	FPR	FNR	TPR	TNR	Total Cost To State Government	Total Accuracy
African-American	0.11	0.512	0.250	0.749	0.487	$-752,448,0782	0.631
Caucasian	0.09	0.475	0.258	0.742	0.524		
Hispanic	0.05	0.513	0.245	0.754	0.486		
Other	0.03	0.479	0.248	0.751	0.502		

The other stakeholders like Judges and Police officers can rely more on this risk assessment tool, to make decisions on how harsh the sentence should be or if they need to imprison the accused for court trials.

5. **Why do you believe that your proposed solution is a better choice than the alternatives? Are there any metrics (TPR, FPR, PPV, etc?) where your model shows significant disparity across racial lines? How do you justify this?**

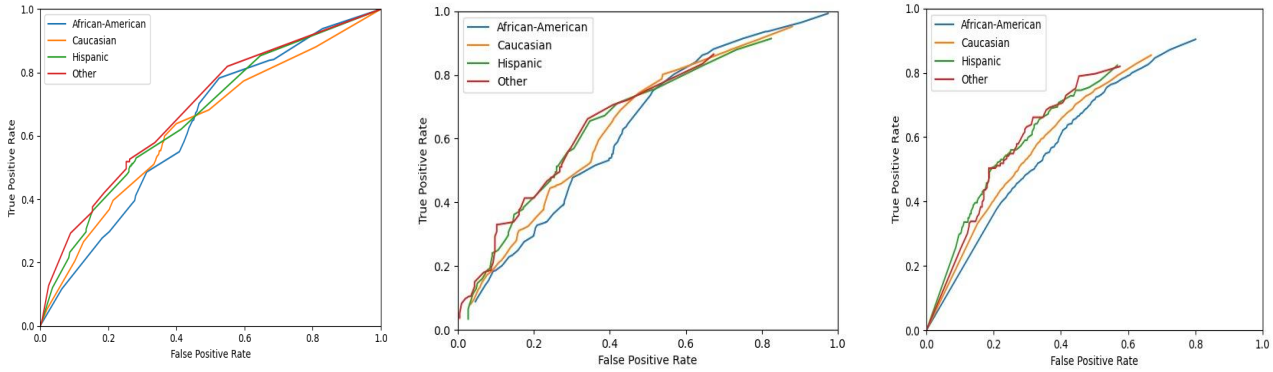
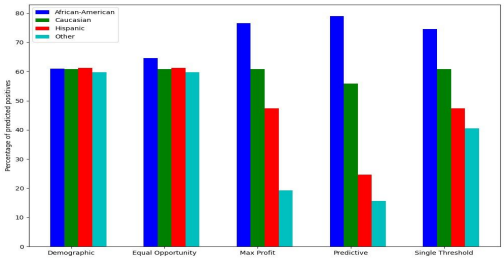


Figure shows the ROC curve for Neural Network , SVM and Naive Bayes respectively. So as we can see the graph AUC is more and the graph is closer towards the top left corner, hence this model was chosen. Demographic Parity and Equal Opportunity had the least amount of disparity among the sensitive attribute race.

The bars show the percentage of predicted positives for various races compared to post processing methods(demographic parity, equal opportunity, max profit, predictive parity, single threshold on X axis.



Equal Opportunity was chosen over Demographic Parity due to:

- Demographic parity would enforce false positives for minority groups in the dataset. So for example if we see COMPAS, just to enforce fairness we would wrongly rate a person in the minority group to equalize the probability for all groups to be at high risk.
- Demographic Parity does not take the true label into account, which would lead to lesser accuracy.
- Demographic Parity impairs an ideal predictor's utility if the tendency to recidivate is even slightly correlated to the sensitive attribute.

A small amount of disparity was noted for PPV, this happened because the dataset doesn't consider the base recidivism rate for different races in Broward County according to Northpointe's reply to ProPublica.

Market Competition Submission

The reasons for picking up Naive Bayes model with equal opportunity has already been enlisted, but here are some other observations as well which can be taken into account:

- Considering the metrics TPR, FPR, FNR, TNR we can see if our model is accurate or not, constituting the fairness as well. On the other hand we can see in the graph represented in figure 2 that the Predictive Parity, which takes PPV into account, shows large amounts of disparity as well.
- Data was only picked on Broward county, and recidivism rate according to various sensitive attributes hasn't been taken into consideration.
- The way assignment is presented, we are assuming that there is only one sensitive factor race, but there can be many like sex, age, c_charge_degree, priors_count,'c_charge_desc.
- Fairness and accuracy are not supposed to go hand in hand, if we would try to make the algorithm fairer we would be inflicting that white people who would supposedly not commit any crime in future according to previous data should be assigned a higher score to make the disparity somewhat equal. And on the other hand if we make our algorithm unfair to the other sensitive group that would also be a violation of a person's right to equality. Hence Equal opportunity is at a place where we can achieve fairness to an extent keeping accuracy in mind.
- The disparities for other sensitive groups were similar to race, for example if we see the data acc. to gender:

	Threshold	FPR	FNR	TPR	TNR	Total Cost To Government	Total Accuracy
Female	0.12	0.439	0.315	0.685	0.560	$-763,910,494	0.627
Male	0.23	0.448	0.302	0.698	0.551		

References:

- Report by ProPublica on COMPAS: Machine Bias — ProPublica
- Equality of Opportunity in Supervised Learning (arxiv.org)
- NorthPointe's response to ProPublica: ProPublica-Commentary-Final-070616 - DocumentCloud
- A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.: http://www.cs.yale.edu/homes/jf/Feller.pdf