Programming Assignment 3
CSE 574: Intro to Machine Learning

Bhavin Jawade, Akhil Singh Chauhan, Anuj Narayan Swami

1. Model Choice: **Naive Bayes**
2. Algorithm Choice: **Equal Opportunity**
3. Secondary Optimization Criteria: **Accuracy**
4. Overall Cost: **$-758,282,364**
5. Overall Accuracy: **0.6282586510808053**

1. What is the motivation for creating a new model to replace COMPAS? What problem are you trying to address? The COMPAS model is biased towards certain groups, especially against African-Americans. We are focused on making the model as fair as possible, by processing the predictions from the model so that they represent all groups equally. This is done based on 5 different post-processing methods, using a set of statistical metrics. Each method is an algorithm that determines the appropriate threshold values so that the data satisfies the requirements of the function. The requirement becomes our primary optimization criteria. As we are an NGO, we want to ensure that accuracy of our predictions are high as possible along with fairness treatment of all groups. Hence we have accuracy as our secondary optimization criteria.

2. Who are the stakeholders in this situation? Stakeholders in this situation are: **Criminals and their families**, Non-Government Organizations **(NGOs)** working on verifying claims of COMPAS, **Northpointe** the corporate that developed COMPAS, **Justice Department** and people who are part of it. And most importantly as this situation might impact the uplifting of a certain group and also impacts the distribution of **taxpayers** money, society as whole is a very important stakeholder.

3. What biases might exist in this situation? Are there biases present in the data? Are there biases present in the algorithms? As we can see from demographics or race-groups, each group has a disproportionate number of members classified as either recidivistic or not recidivistic. The African-American group have substantially more members classified as recidivistic, Caucasians an equal number, and Hispanic show a larger portion classified as non-recidivistic. Because the model has been trained over data that is skewed towards certain races with a history of showing greater signs of recidivism, the predictions also eventually tend to be biased towards the same result.

As per Barocas, Apart from the direct use of sensitive feature like race for training, bias might exist in data due to:

1. Human labelling or examples used as initial examples might be tainted or biased.
2. Bias in initial samples that got propagated and compounded.
3. Features might present less information for the minority class
4. Training data might have unequal number of data-points for both the classes
5. Even if sensitive feature is not explicitly present in the data, other features might act as proxy because of close correlation to sensitive feature.

Gender: **Females : 1395, Male : 5820**
Race: **African-American: 3696, Asians: 32, Caucasians: 2454, Hispanic: 637, Others: 439**
As we can see above their is unequal number of data-points for all classes of sensitive attributes, hence it contains **bias due to Sample size disparity**. As per the Pro-Publica report and Brennan, it is difficult to construct data-set that does not have features which are in someway correlated to race. Hence quite certainly, **the Compas dataset contains Proxy Bias**.

4. What is the impact of your proposed solution?

Our Proposed solution will enforce fairness in predicting recidivism across groups of society. Equal TPR rate will ensure that all groups have same rate of being correctly predicted as recidivist. This solution will help in limiting disparate impact [cite big data's disparate imapact]. Disparate impact arises when outcomes disproportionately benefit or hurt people with certain sensitive attribute values.

5. Why do you believe that your proposed solution 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? Answer: To justify our choice of algorithm, we will scrutnize all 5 algorithms based on TPR, FPR, PPV, FNR rates and their inherent pros and cons. In [1; 2], Prof. Moritz Hardt argues that demographic parity does not ensure fairness as this classifier will select qualified applicants in the demographic Class1=0, but unqualified individuals in Class1=1, as long as the ratios of

acceptance match (also referred as laziness behaviour of this method). His second argument states that this method cripples ML as whole because it loses the fundamental idea of gaining higher accuracy and that too for no good. Demographic parity will simply rule out the ideal classifier that might be possible hypothetically. This point is also quite evident from the accuracy we can see in our code, Demographic parity has got very less accuracy specifically for African-Americans and Caucasians along with overall less accuracy (0.625), when compared with Predictive Parity (0.631) or equal opportunity (0.628). Hence looking at the drawbacks of Demographic parity, we can certainly say that Predictive parity is better than Demographic parity. Now coming to comparison between **Equal Opportunity** and other methods. To compare these we will refer [3] which talks about equal opportunity as a fairness algorithm to be better than Max Profit simply because maximum profit has no notion of fairness, better than Race Blind or single threshold because it uses a single threshold for all groups which means it will put more African-American as recidivists. Now comes the most important comparison between Predictive parity and Demographic parity. It is quite evident from the results that we got that, equal opportunity not just imposed same TPR for all groups but also insured that FNR, FPR and TNR values are close enough for all groups. In case of predictive parity, FNR, FPR and TNR for African-Americans are unusually high or low with respect to other groups, showing distinctive unfairness. Incase of equal opportunity not just values of FNR, FPR are close enough they are also lower than that of Predictive parity which is good (having low false rate). Additionally, Equal opportunity had both PPV and TPR values high enough and decently close to each other for all groups. Hence, we selected **Equal Opportunity** algorithm[2] as the best algorithm here keeping in group fairness and disparate impact [4] in mind.

1 Additional Questions

1. How do you justify valuing one metric over the other as constituting "fairness"? Answer: To answer this we need to first define fairness (or unfairness for that matter). [4] define unfairness using 2 notion: disparate treatment, disparate impact. If decisions are made using the sensitive attribute then its disparate treatment and if the outcome of decision impacts different classes of sensitive attribute differently then its disparate impact. Now when we think about it, disparate treatment is something that more part of the decision making process which could be considered the human part of the system. Disparate impact is more implicit and could arise because of the data or algorithm. We want to limit disparate impact, and to do so we want to make sure that all groups must benefit (impact) equally from the outcome. Hence we compare the metrics based on how equally they impact different groups. TPR is metric which captures this picture very closely.

2. What assumptions are made in the way we have presented the assignment? Are certain answers presupposed by the way we have phrased the questions? Answer: Yes one of such assumption is regarding maximizing profit. Groups were divided into Corporate and NGO and were supposed to choose between accuracy or financial as secondary optimization criteria. Corporates will financials and NGO will select accuracy. If we read the Your role section of problem statement, it says societal considerations and positive reform are main concern of NGO.

3. In what ways do these simplifications not accurately reflect the real world? Answer: Although these simplifications try to reflect the real world as much as possible, a better simulation could be created by considering more stakeholder working simultaneously trying to optimize multiple features.

4. How do uncertainty and risk tolerance factor into your decision? Answer: Accuracy and uncertainty hold a inverse relationship. If accuracy increases uncertainty decreases.

To what extent should base rates of criminality / recidivism among different groups be factored into your decision? Answer: Base rate refers to the proportion of a population (group) showing particular characteristic, in this case recidivism. The whole point of fairness is to make treatment and impact independent of the sensitive attribute, in this case race. By considering base rates in the decision making process we are inherently bringing the sensitive attribute into the picture hence creating possibility of bias. This bias once entered into the pipeline in any form can then propagate in the whole process. Hence we strongly believe base rate should not be considered anywhere in our decision. Additionally, when we consider individual fairness instead of group fairness, point of base rate clearly represents a bias.

References

Moritz Hardt, Approaching Fairness in Machine Learning, *http://blog.mrtz.org/2016/09/06/approaching-fairness.html* 2016.

Dwork, Hardt, *Fairness through awareness*

M. Hardt, E. Price, N. Srebro, *Equality of Opportunity in Supervised Learning*, 2016.

Barocas, Selbst, *Big Data's Disparate Impact*