**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 also set accuracy of the model as a secondary optimization as our goal is to make the post-processed model free of bias.

**Who are the stakeholders in this situation?**

The stake holders are all the individuals within the US criminal justice system. This includes the criminals, their families, and as this issue is influenced by race, it affects the society as a whole. Wrongful imprisonment is unfair, and thus it also undermines the decision-making process of the criminal justice system.

**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.

Hence in this exercise, we are also taking race as a factor that can influence the outcome.

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

Our proposed solution aims to resolve two issues in the model:

1. Demographic disparity
2. Equal Opportunity

**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? (8pts)**

We have optimized the methods using certain metrics (TPR, Predictive Positives, PPV). These metrics give a better understanding of how the model is biased and shows that there is significant disparity across the race groups.

Method 1: Enforce ‘demographic parity’, by ensuring that each group has an equal number of members classified as recidivistic (label: 1) and non-recidivistic (label: 0). We have set an epsilon tolerance of ±2%.

Demographic parity can be logically calculated by the formula:

d.p. = Predicted Positives / (Predicted Positives + False\_Negatives + True\_Negatives)

Method 2: Enforce ‘equal opportunity’, by ensuring all groups show the same value of recall (probability of a prediction of ‘recidivistic’ given ‘crime committed’). In other words, we want to each group an equal opportunity to be classified as ‘recidivistic’. This also means that all groups should have the same true positive rate (TPR). We have set an epsilon tolerance of ±1%.

TPR (Recall) = True Positives / (True Positives + False Negatives)

**Reference:**

<https://arxiv.org/pdf/1908.09635.pdf>

**BONUS Questions**

**How do you justify valuing one metric over the other as constituting “fairness”?**

The metric that gives an equal representation over the ‘privileged’ outcome, in this case ‘recidivistic’, should be chosen. In other words, the metric should ensure that all individuals that are predicted to be recidivistic, get an equal opportunity at being labeled recidivistic. However, demographic parity is not as efficient at enforcing this equality. Demographic parity at most strives to ensure that each group has the same proportion of members labeled as recidivistic. This does not ensure that, in each group, the same proportion of members were given equal opportunity at being labeled recidivistic. Single threshold raises a lot of ambiguity in its methodology as only one threshold value is subjected to all groups. Maximum accuracy will find any means necessary to maximize the accuracy at the cost of unequal representation of groups, hence this is also not a fair metric. We can conclude that choosing the best metric is dependent on the role and situation. In our case, equal opportunity shows the best fairness.

**What assumptions are made in the way we have presented the assignment? Are certain answers presupposed by the way we have phrased the questions?**

Assumption 1: That it is sufficient to divide the model output into 5 groups, including ‘Other’ that encompasses all miscellaneous groups. The miscellaneous groups may be a collection of different race types, some which may have a greater disparity than the African-American group.

Assumption 2: The prediction-label pairs given are exhaustive, and values generated by metrics such as TPR, PPV, Predicted Positives, False Positives/Negatives, True Positives/Negatives serve as a good approximation for each of the post-processing methods.

Assumption 3: We have also assumed that considering at most 2 roles or perspectives, NGO and Corporate, and 2 secondary optimizations, accuracy and profit, is enough to analyze the issue. There can be other perspectives and secondary optimizations.

**In what ways do these simplifications not accurately reflect the real world?**

**How do uncertainty and risk tolerance factor into your decision?**

In each post-processing method, we apply a risk tolerance factor to line up the probabilities instead forcing them to match exactly. It ensures that our metrics work on data with a lot of variance. In case variance is high, the tolerance levels may have significantly larger values, and vice versa. Essentially, we would like to conduct our analysis over a neighborhood of data points which span over this tolerance level. It is also true that stretching or increasing this tolerance level will introduce more uncertainty and inaccuracy in our analysis.

**To what extent should base rates of criminality / recidivism among different groups be factored into your decision?**

Base rates of any feature serve as prior knowledge, and hence will introduce some bias. In small amounts, they can serve as a good benchmark, but if base rates are dominant then it will make skew the data towards the sensitive attribute.

**The tools we provide can split the predictions into different protected categories, such as by age or gender. What disparities arise in these groups? How do these disparities compare to those shown when the predictions are split by race?**

Age as a protected category will introduce new disparities such as the legal age limit that allow convictions, harshness of sentence determination and duration of term assigned. Gender will introduce disparities on male/female ratio of recidivism.