Kieren Singh Gill
Responsible Data Science
Professor George Wood
Homework #1

Problem 1 (20 points): Fairness from the point of view of different stakeholders

1a)

A. Accuracy

If the model's accuracy is optimized, most of the stakeholders involved will beneft. Defendants who were falsely assigned a higher risk score (predominantly black defendants) due as a result of false positives will benefit because they will receive a fairer and more accurate risk score. If their model has a higher accuracy rate, Northpointe will be able to bolster their reputation, and more companies and systems will be inclined to use their model, which can result in economic benefits for the firm.

B. Positive predictive value

If the model's positive predictive value (PPV) is optimized, black defendants stand to benefit the most. White defendants also benefit, and so do the judges and decision makers in the criminal justice system that utilize COMPAS. Positive predictive value can be calculated by $\frac{True\ Positive}{True\ Positive + False\ Positive}$. When this metric is optimized, the PPV should be as close to one as possible. This means that when PPV is optimized, the false positive rate would have decreased. Since black defendants were wrongly labelled as reoffenders at twice the rate as white defendants, they are severely impacted by the nature of the false positive rate, which is why they stand to benefit the most if it decreases. White defendants also benefit from this, and the decision makers in the criminal justice system benefit from this because they can use the model to make sentencing decisions that are more accurate.

C. False positive rate

If the model's false positive rate is optimized, black defendants stand to benefit the most. White defendants will also stand to benefit, but not at the same rate as black defendants. This is because as explained in part B, black defendants were wrongly labelled as reoffenders at twice the rate as white defendants (44.9% to 23.5%). Hence, if the false positive rate is optimized, they won't be subject to the same amount of mislabeling, and will be subject to fairer sentencing.

D. False negative rate

If the false negative rate is optimized, society at large stands from benefit from this. This is because a false negative results in a defendant being labeled low risk when they should not be low risk, and it is likely that these defendants will recidivate, which will negatively affect society. By optimizing the false negative rate, we reduce the chances of this happening, so society stands to benefit.

E. Statistical parity (demographic parity among the individuals receiving any prediction)

If statistical parity is optimized, black defendants stand to benefit from this. If a model is said to have statistical parity, it means that it treats the general population similarly to how it should treat a protected class within the general population. In this case, the protected class are black defendants because they are currently unfairly treated by the COMPAS algorithm. By making sure that black defendants are treated the same way as the rest of the general population, they stand to benefit.

1b)

Consider a hypothetical scenario in which TechCorp, a large technology company, is hiring for data scientist roles. Alex, a recruiter at TechCorp, uses a resume screening tool called Prophecy to help identify promising candidates. Prophecy takes applicant resumes as input and returns them in ranked (sorted) order, with the more promising applicants (according to the tool) appearing closer to the top of the ranked list. Alex takes the output of the Prophecy tool under advisement when deciding whom to invite for a job interview.

A. Pre-existing Bias
Because the tech industry is male-dominated, especially in technical roles, it is likely that in the data set that was used to train Prophecy, most of the applications would have come from men. This pre-existing bias in the data will reflect the lack of women in technical positions. In effect, the system could end up teaching itself that male candidates are preferable, and might penalize resumes that included the word "women's", as in "women's chess club captain", or downgrade applicants that graduate from all-women colleges. In this case, women would be adversely affected by the hiring algorithm through disparate impact. To mitigate this bias, Alex could use AI Fairness 360 (AIF360), an open-source toolkit used to check for unwanted biases and to mitigate them. With AIF360, he could set the protected attribute to "gender" so that the model will not take into account if the applicants are male or female.

B. Technical Bias
Resume screening tools like prophecy rely on Natural Language Processing methods to parse resumes in order to rank them. If more design-oriented applicants applied with creatively designed resumes to highlight their creative ability, Prophecy might encounter difficulty in parsing their resume since it would probably be formatted in a suboptimal manner for the screening tool to parse through it. These applicants would be at a disadvantage, simply because their resumes could not be "read" by Prophecy. Also, most resumes are one page long, and Prophecy might disregard data on multiple-page resumes. This could disadvantage participants who had strong and relevant experiences on the second page of their resume, as it simply would not be taken into consideration. This would also disadvantage TechCorp, as TechCorp could lose out on hiring qualified candidates simply because of Prophecy. To overcome this, the algorithm could be redesigned to accommodate multiple page resumes, and if there is a difficulty in parsing through a resume, the resume should be flagged for a human to go over instead of it being instantly disregarded.

C. Emergent Bias
In the tech industry, technical skills and side projects are valued highly, and is on par with or could even be preferred over other work and leadership experiences. If Prophecy is used across multiple industries, it would be trained with data from all these various industries which prioritize work and leadership experiences, and it might develop scoring metrics that cater to these characteristics. Hence, the model might undermine the value of technical skills and side projects, and score resumes with these skills as

lower than resumes with other work and leadership experiences. This would be a case of emergent bias, as the algorithm is “behaving” in this manner only because its feedback from other industries is making it biased. In this case, TechCorp would be losing out because they would be missing out on hiring potentially qualified candidates. The candidates also lose out because they would be undervalued based on the algorithm. To mitigate this, companies should constantly be mindful of the training data of their models, and they should provide their models with updated data to reflect candidates with the skills required that are looking for.

1c)

A.

Formal EOP says competition is fair when competitors are only evaluated on the basis of their relevant qualifications - in any contest, the most qualified person wins. In this scenario, following Formal EOP, the admissions officer would only evaluate candidates based on their GPA and SAT score. Formal EOP rejects hereditary privilege as a basis of winning positions. Hence, family income brackets should not be considered. In addition to this, since GPA varies from school to school, the admissions officer might prioritize SATs over GPA, since SATs are standardized across the nation.

B.

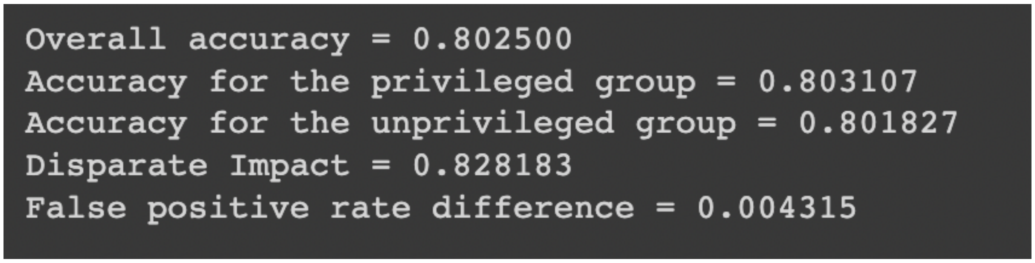
Rawlsian Fair EO/Substantive EO would be consistent with the goal of correcting such differences in the applicant pool. Substantive EO seeks to provide all individuals with realistic opportunities to develop qualifications, and hence a realistic shot at competing for a broad range of positions. In this case, the Substantive EO will consider the circumstances of the applicant, will take into consideration the fact that applicants are from lower income families, and will adjust the admission criteria to account for the applicants’ circumstances when evaluating their SAT scores.

C.

Luck-egalitarian is a selection procedure that believes no factors that an applicant did not choose for themselves (choice-luck) should affect the chances of admission. Factors like socioeconomic status are called brute luck, and luck-egalitarian controls for these factors to even out the playing field. The procedure for this intervention is to create brackets based on matters of brute luck and compare the candidates to others in their own brackets. In this case, admissions officers would divide the candidates up by their parent’s income (low, middle, high) and then evaluate the candidates by the relevant metrics against those in their own brackets. The candidates may also be evaluated on other factors of choice luck like their extracurricular activities and supplemental essays.

Problem 2 (40 points): Fairness-enhancing interventions in machine learning pipelines

a) I trained a baseline random forest model to predict income, using the hyperparameters in the provided notebook. Here are the results as follows:



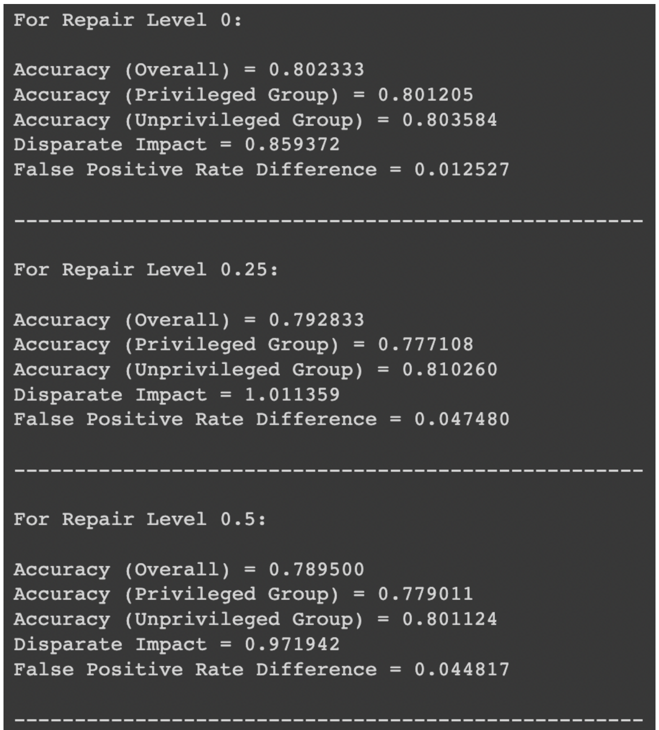
```
Overall accuracy = 0.802500
Accuracy for the privileged group = 0.803107
Accuracy for the unprivileged group = 0.801827
Disparate Impact = 0.828183
False positive rate difference = 0.004315
```

My overall accuracy was 0.8025, which is decent, but definitely far from excellent. Accuracy for the privileged group was higher than the accuracy for the unprivileged group, at 0.803107 and 0.801827 respectively. The model's disparate impact was at 0.828183, which means that it was not too unfair, but it still shows that there is room for improvement, because ideally, disparate impact should be closer to 1. The false positive rate difference was quite low at 0.004315, which is good because it is close to the ideal value of 0, and indicates that my model has mitigated some bias.

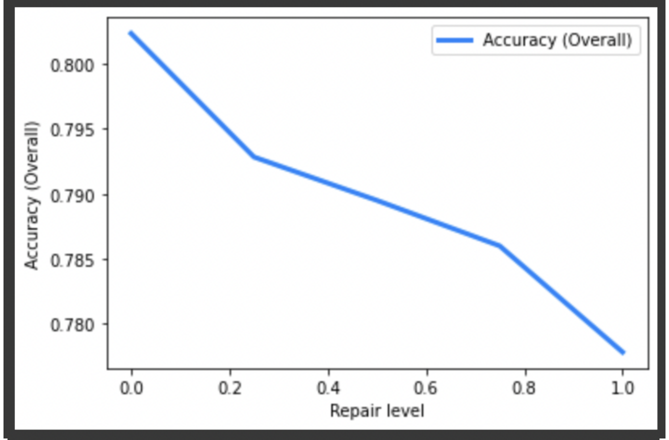
b) I used Consider Disparate Impact Remover (DI-Remover), a pre-processing fairness-enhancing intervention, to analyze the metrics at different metrics of repair. As repair level increases, overall accuracy decreases. However, disparate impact increases as repair level increases. This is expected because there is a tradeoff between accuracy and disparate impact - to reduce the unfairness between privileged and unprivileged groups, you should increase the repair levels, but in doing so, you sacrifice accuracy. It can also be observed that the accuracy of privileged groups decreases similarly to overall accuracy, but accuracy for underprivileged groups seems to increase, decrease, and increase again. Note that the magnitude of these changes for the accuracy for underprivileged groups is extremely small.

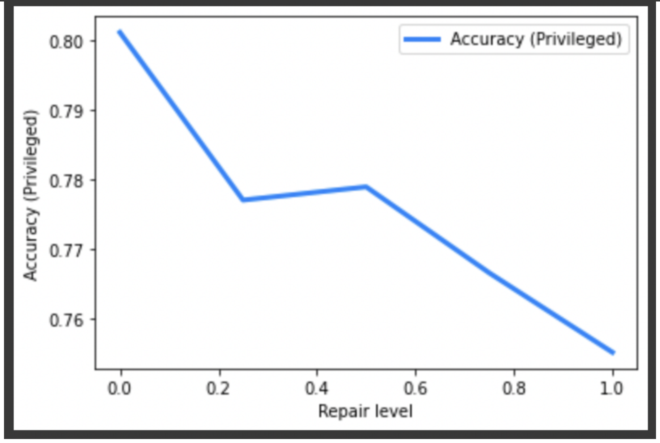
Also, it is interesting to note that by looking at the metrics below, the disparate impact is over 1 when the repair level is at 0.25, 0.75, and 1, which means that the disparate impact remover has overcorrected for the bias, and now the privileged group is disparately impacted. Given these metrics, I believe that the optimal disparate impact level would be at 0.25, because the disparate impact will be extremely close to 1, and not too much accuracy will be sacrificed from the model. Also note that the false positive rate difference increases as repair levels increase - this is because accuracy goes down, which means that there will be more false positives, which leads to an increase in the false positive rate difference.

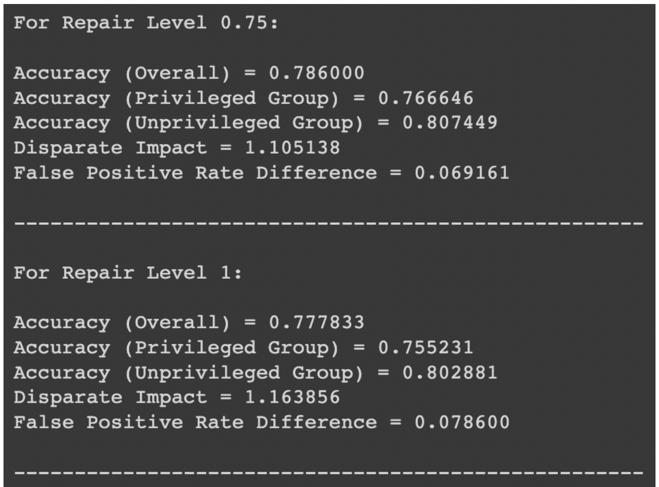
The graphs and metrics of my results are on the following page:



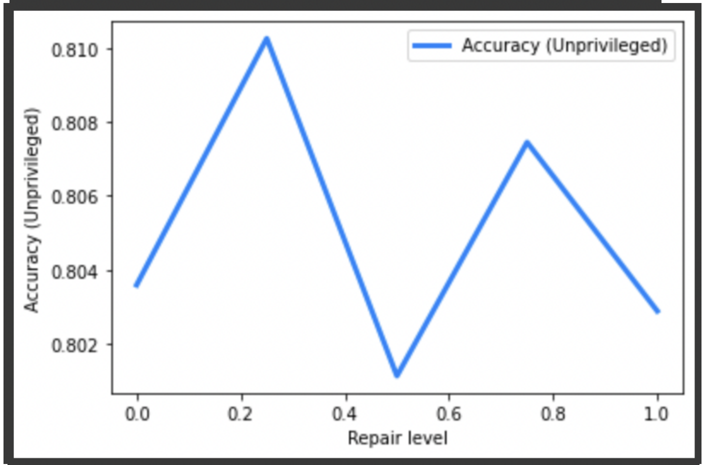
```
For Repair Level 0:

Accuracy (Overall) = 0.802333
Accuracy (Privileged Group) = 0.801205
Accuracy (Unprivileged Group) = 0.803584
Disparate Impact = 0.859372
False Positive Rate Difference = 0.012527

----------------------------------------------------

For Repair Level 0.25:

Accuracy (Overall) = 0.792833
Accuracy (Privileged Group) = 0.777108
Accuracy (Unprivileged Group) = 0.810260
Disparate Impact = 1.011359
False Positive Rate Difference = 0.047480

----------------------------------------------------

For Repair Level 0.5:

Accuracy (Overall) = 0.789500
Accuracy (Privileged Group) = 0.779011
Accuracy (Unprivileged Group) = 0.801124
Disparate Impact = 0.971942
False Positive Rate Difference = 0.044817

----------------------------------------------------
```

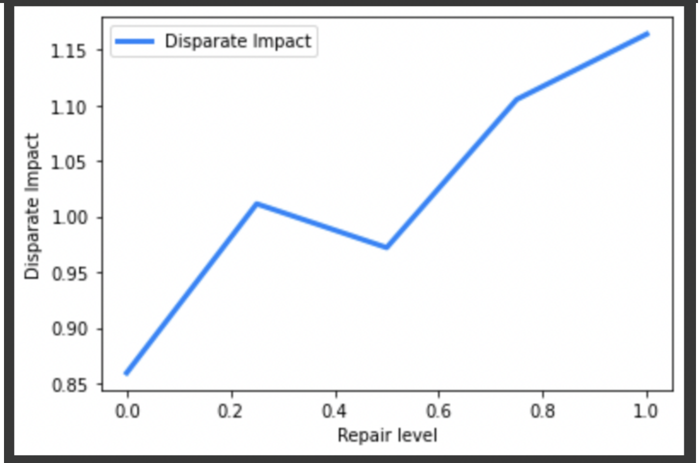


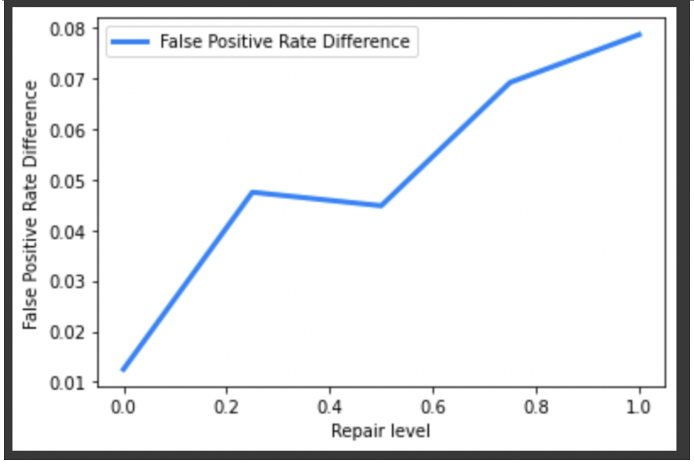




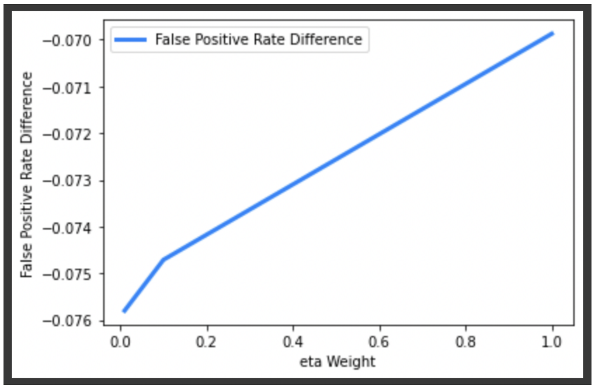
```
For Repair Level 0.75:

Accuracy (Overall) = 0.786000
Accuracy (Privileged Group) = 0.766646
Accuracy (Unprivileged Group) = 0.807449
Disparate Impact = 1.105138
False Positive Rate Difference = 0.069161

----------------------------------------------------

For Repair Level 1:

Accuracy (Overall) = 0.777833
Accuracy (Privileged Group) = 0.755231
Accuracy (Unprivileged Group) = 0.802881
Disparate Impact = 1.163856
False Positive Rate Difference = 0.078600

----------------------------------------------------
```

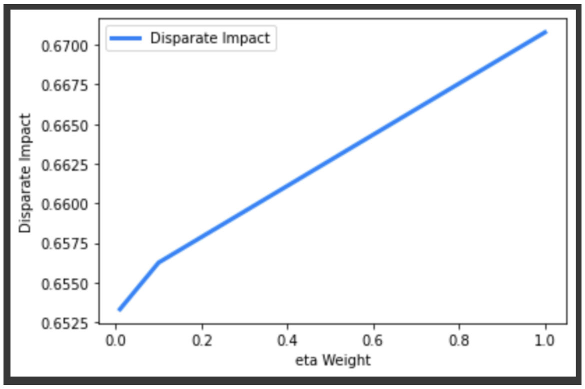


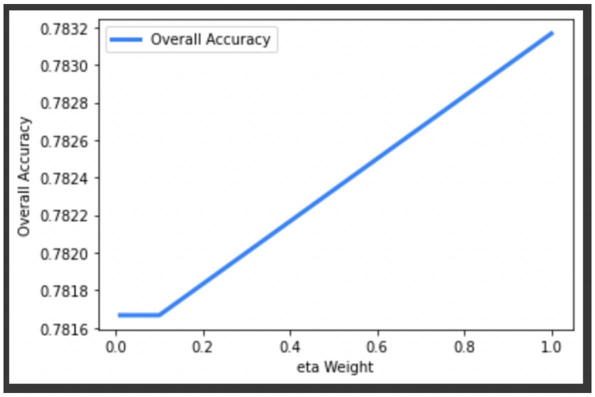




c) The Prejudice Remover in-processing technique in AIF360 is implemented as a regularizer, which means that it can be applied to a wide variety of prediction algorithms with probabilistic discriminative models. The Prejudice Remover focuses on classification, and the regularizers are built into logistic regression models.







As the eta parameter increases, overall accuracy and disparate impact increase as well. This positive correlation is good because ideally, we want disparate impact and overfall accuracy to be as close to 1 as possible, and is unlike the DI-Remover which appears to have an inverse relationship between disparate impact and overall accuracy. However, both accuracy and disparate impact levels are significantly lower than when the DI-Remover is used.

In conclusion, both the DI-Remover and Prejudice Remover are some of the many tools available to account for and adjust for biases in algorithms and data. These tools work to varying extents, and certain tools work better in certain scenarios, depending on what metric the researcher/data scientist is prioritizing. In this case, the DI-Remover presents a tradeoff between overall accuracy and disparate impact, whereas the Prejudice Remover has lower levels of both metrics instead.

Problem 3 (15 points)

3)

“AI for whom?”

In this lecture by Danya A Glabau, we explore the data science fairness issue that is the prevalence of discrimination in AI algorithms - how it benefits certain demographic groups and negatively impacts others. She discusses these issues through the lens of the hiring industry, specifically the role of AI in the screening and recruitment of employees.

Danya discussed a report on an Amazon system in which computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a ten-year period. Because the tech industry is male-dominated, especially in technical roles, most of the applications came from men. In effect, the system ended up teaching itself that male candidates were preferable, and consequently penalized resumes that included the word “women’s”, as in “women’s chess club captain”. It also downgraded the applicants that graduated from two all-women colleges. In this case, women were being adversely affected by the hiring algorithm through disparate impact, assuming that the algorithm was created without malicious or discriminatory intent. This was a result of pre-existing bias in the data because ten years' worth of data that reflected the lack of women in technical positions resulted in the algorithm being trained to penalize female applicants. While AI promises to correct human biases by having fewer humans in the hiring process, the biases of data used to train tools can perpetuate past and present biases, as indicated by this example.

Another example that was discussed was Predictim’s recruitment AI, a service that scrapes the social media accounts of babysitter applicants and scores them based on certain indicators of potential success. Interestingly, in a piece written by Brian Merchant, he decided to put Predictim to the test. He used Predictum on his child’s babysitter whom he had never had any issues with, and he used it on his child’s godfather, who was a comedian. Despite the comedian having multiple profanities littered throughout his Twitter feed, he scored close to perfect on Predictum, significantly higher than his child’s nanny. In this case, one key difference between the babysitter and the comedian is that the comedian was a white man and the babysitter was a black woman. In this case, black women in particular were being adversely affected by the hiring algorithm through disparate impact. This shows that while AI promises to evaluate people based on proven indicators of success when implemented incorrectly, this can result in AI hiring more people like the ones who are currently employed. This will adversely impact marginalized groups that aren’t in the industry, making life harder for them to break into the industry. This could be an example of technical bias because Predictim is a black-box algorithm that appears to be biased without offering any justification. Since Predictim promises to scrape data across so many different platforms for each babysitter, there could have been an oversimplification of certain parameters that leads to a misjudgment when generating a result.