Algorithmic Fairness

Preethi Narayan

March 1, 2022

Problem 1

Part A

A. Accuracy:

In this scenario, accuracy specifically refers to how well the risk scale is able to determine who will recidivate and who will not. The best case scenario would involve having an accuracy of 1. The stakeholders that would benefit from a model that optimizes accuracy include the decision maker and defendants who end up in the false positive group, which ProPublica's data has suggested are often those in minority races, will benefit from the optimization of accuracy. Improved accuracy will allow the decision maker to confidently make judgements and defendants who do not recidivate will be less likely to be classified in the recidivist category.

B. Positive predictive value:

This refers to the true positive rate of the model, or in other words, the probability that a person that is labeled as a recidivist will truly recidivate. The best case scenario would involve having a positive predictive value of 1. The optimization of this variable would result in a false negative rate of 0. As a result of a decreased false negative rate, white defendants would be harmed. It would not be reasonable to optimize this metric because, as we saw in one of the labs, optimizing this metric could lead the algorithm to make incorrect predictions, resulting in decreased accuracy of the model. This, in turn, would harm the decision makers.

C. False positive rate:

The best case scenario would be having a false positive rate of 0. Having a false positive rate of 0 would benefit those people that do not recidivate as they do not have to be concerned that they will falsely be labeled as high risk to recidivate. Optimizing this

metric to have a false positive rate of 0 would possibly negatively impact the overall accuracy of the model as the algorithm could start to make more incorrect predictions. This would result in harming the decision makers as well as harming Northpointe.

D. False negative rate:

   This refers to the rate of people that are marked as recidivists but end up not recidivating. The best case scenario would be having a false positive rate of 0. In the case that this is optimized, white defendants would be more negatively impacted than other groups because ProPublica's research has indicated that the false negative rate is higher than that of nonwhite defendants, leading those who would recidivate to be marked as non recidivists. Optimizing this metric could once again possibly affect the accuracy of the model, leading the algorithm to make more incorrect predictions, which would harm the decision makers.

E. Statistical parity:

   This refers to the demographic parity among those that are receiving any prediction from the algorithm. The best case scenario would be having an equal proportion of defendants in each group. While this would harm white defendants, it would not necessarily help minority groups. Rather, this might cause higher false positive and false negative rates across the board. Optimizing this metric would result in sacrificing accuracy for the sake of fairness.

Part B

A. As discussed in class, pre-existing bias is bias that is independent of an algorithm and has its origins in society. Pre-existing bias may arise in the scenario described above if the training data for the resume screening tool Prophecy had racial or gender bias. This would be considered pre-existing bias because racial and gender biases are present in society as a whole, which would result in them being present in training data that is based off of data points in the real world. Historically, women and non-white employees have been discriminated against in professional settings, and this would be reflected in the training data. This would then result in the resume screening tool Prophecy exhibiting those biases by choosing certain candidates over others that may not necessarily be more qualified, but rather fit the gender and racial biases that the tool has been trained on.

   Applicants may be harmed by pre-existing bias because, though they may be fully qualified for the job, the system may rank them lower than other candidates due to their race or gender as a result of the pre-existing biases that the algorithm is exhibiting. Specifically, applicants that are parts of minority groups that are often at a disadvantage during job applications (Women, POC, LGBTQ+, etc.) may be harmed by this.

   An intervention that might help to mitigate this tool is instrument calibration, so that the tool will perform equally amongst different groups. Another intervention that might mitigate this tool is for Prophecy to remove the variables that represent the protected

characteristics of minority groups. If this information was removed, then only information that is related to the qualifications of the candidates would be left, and the system wouldn't carry over pre-existing biases.

B. Technical bias is bias exhibited by systems as a result of technical constraints/considerations. This may result from a limitation of the system or computer software, a system or algorithm that fails to treat all groups fairly under all conditions, from inadequacies in random number generation, or from errors in translation when there is an attempt made to translate human constructs to something understandable for computers/algorithms. In this scenario, technical bias may arise in the situation that members of minority groups that are typically discriminated against may fill out a pre-interview survey with incomplete or incorrect information that may prevent the algorithm from giving accurate results. For example, if a person that is mixed-race does not have an accurate option to fill out when the survey asks them what race they are, they may incorrectly fill out a category that does not accurately represent them, or in the case that they system offers an "other" option, they may select that. In any of these scenarios, this person is not being accurately represented in the system's data, and this will result in there being bias in the tool's ranking.

   TechCorp could be harmed by technical biases because the biases in the algorithm could be costing them valuable qualified candidates by ranking the resumes of privileged but less qualified candidates over the resumes of other minority and non-privileged but more qualified candidates. This may happen because less privileged people may not choose options in surveys that accurately describe them, which may work against them as mentioned earlier. As a result of these biased rankings, TechCorp may choose candidates that are less qualified, and this may hurt the company in the long run.

   An intervention that may help to mitigate this bias is if Tech Corp devised a way for Prophecy to handle null values with these types of questions that does not negatively affect a candidate's ranking. To handle the null values, Prophecy could conduct statistical profiling on the columns that may contain null values. After calculating the number of percentages there are of null values for a certain variable, it could set a threshold, and then calculate whether the variable is providing fair representation for all the groups considered. If the variable does not provide fair representation for all the groups, then the variable can be dropped from the algorithm.

C. Emergent bias is that which arises from context of the use of a system. In this scenario, emergent bias could occur from Alex making use of Prophet's ranking systems to hire people for TechCorp. If Alex was to use Prophet's ranking systems to hire people for TechCorp, then whatever biases the system has will eventually pass on to Alex because he is trusting the system to make the correct decisions, and he is reinforcing those decisions by using those decisions to decide who will be hired. This will result in his idea of the ideal candidate to be influenced by Prophecy, and as a result Alex will carry the biases from Prophecy into his hiring decisions. This will then create an infinite loop of Alex reinforcing the biases from Prophecy, and Prophecy's biases influencing Alex's hiring decisions.

The people that might be harmed by this bias are minority groups and non-privileged people such as women and POC because Prophecy may rank these candidates lower than other candidates due to the biases it carries. These biases are then passed onto Alex, and then they are exhibited in his hiring decisions, which will further put minority groups at a disadvantage in the hiring process.

This bias could be mitigated by Alex preventing the system's bias to be imparted onto him by consciously making decisions that go against the system's bias. If Alex went out of his way to find qualified minority and non-privileged candidates that would be a good fit for the job, it would mitigate the emergent bias.

Part C

A. If an admissions officer were following the ideals of Formal EO, they would only care about the qualifications about an applicant and nothing else. In this case, the features that the admissions officer would use to evaluate applicants are SAT score and High School GPA. The admissions officer would only care about these because in formal EO, "a competition is fair when competitors are only evaluated on the basis of their relevant qualifications". Following formal EO, the admissions officer would not be concerned with family income bracket because they would only care about the pure qualifications of the applicant.

B. The EO doctrine that is consistent with the goal of correcting income-based differences in SAT scores would be Rawl's-Fair EO. The concept of Rawl's-Fair EO follows that "all people, regardless of how rich or poor they are born, should have opportunities to develop their talent". This doctrine believes in ensuring that being born into a privileged life doesn't compound into a lifetime of outperforming and outcompeting those that were not born into a privileged life whose disadvantages may have compounded over time. Therefore, correcting income-based differences that are observed in applicants' SAT scores would be consistent with the doctrine of Rawl's-Fair EO.

C. An applicant selection procedure that is fair according to Luck-Egalitarian EO would possibly include separating applicants into brackets for factors that provide unfair advantage or disadvantage. One example would be a system that assigns each applicant to a bracket based on their race, family's income bracket, sex, and disability status. Applicants would be competing amongst other applicants within their own brackets rather than competing with other applicants in more privileged brackets.

Problem 2

Part A

A.

```
Mean difference = -0.061586
Disparate Impact = 0.857283

Error rate difference (unprivileged error rate - privileged error rate)= -0

False negative rate for privileged groups = 0.215599
False negative rate for unprivileged groups = 0.189612
False negative rate ratio = 0.879463

False positive rate for privileged groups = 0.126050
False positive rate for unprivileged groups = 0.133216
False positive rate ratio = 1.056846
```

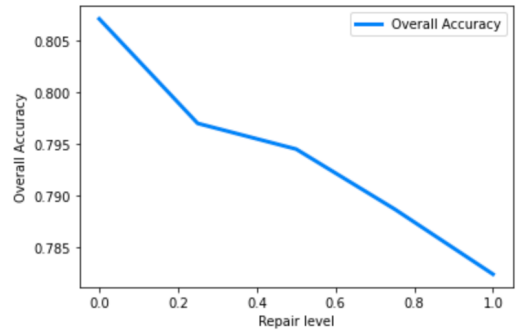
The false negative rate for privileged groups is higher than the false negative rate for unprivileged groups, and this is almost always true. For every privileged person that is marked as a false negative for recidivism, .88 non-privileged people are marked as a false negative for recidivism.

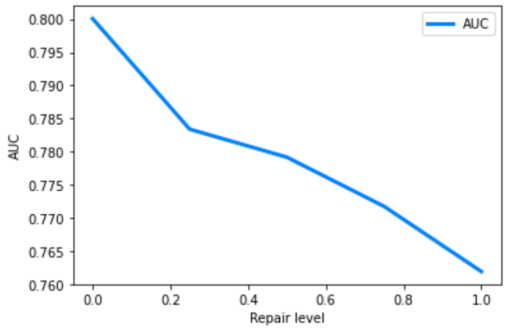
The false positive rate for non privileged groups is higher than the false positive rate for the privileged groups, which is also almost always true. For every 1 privileged person that is marked as a false positive for recidivism, 1.06 non-privileged people are marked as a false positive for recidivism.

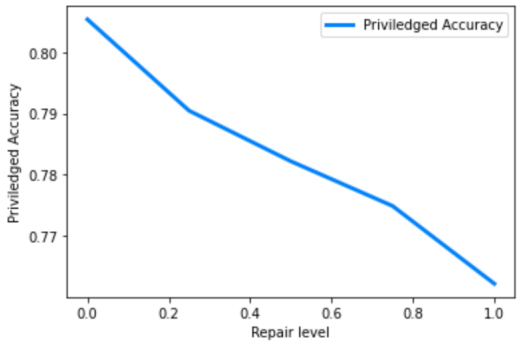
```
accuracy: 0.8393333333333334
privileged accuracy: 0.8323982615566969
unprivileged accuracy: 0.8470691934773028
disparate impact: 0.8572834544789694
false positive rate difference: 0.007165462742896311
```

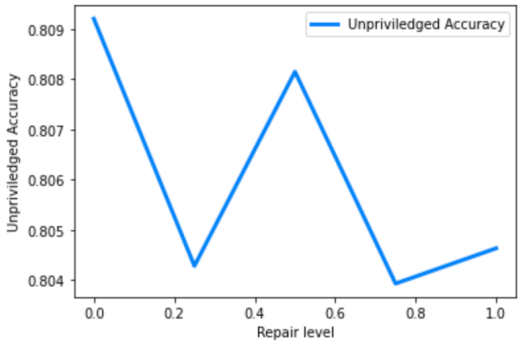
The accuracy for this model was relatively good. The accuracy for the privileged group and unprivileged group is daily close, but the accuracy for the unprivileged group is slightly higher. There is relatively high disparate impact before the dataset is altered with the disparate impact remover. The false positive rate difference is relatively small.

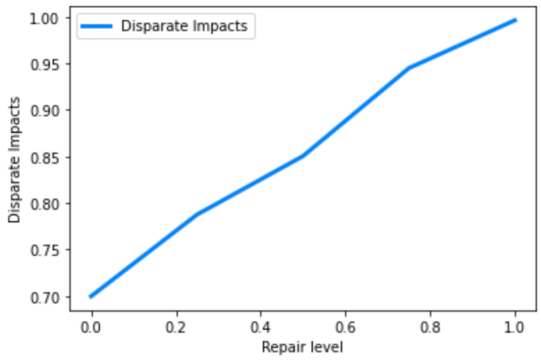
B.

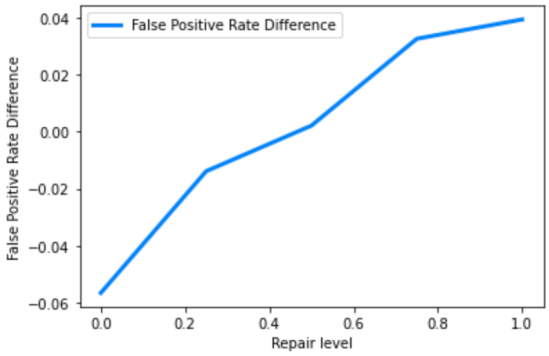






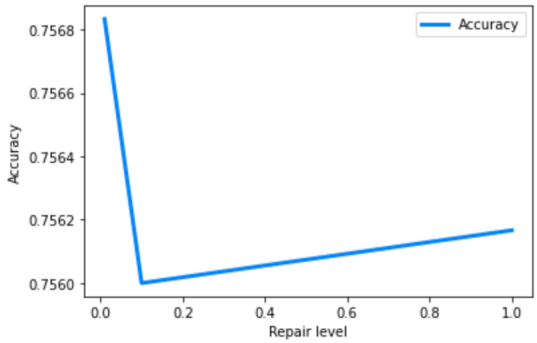


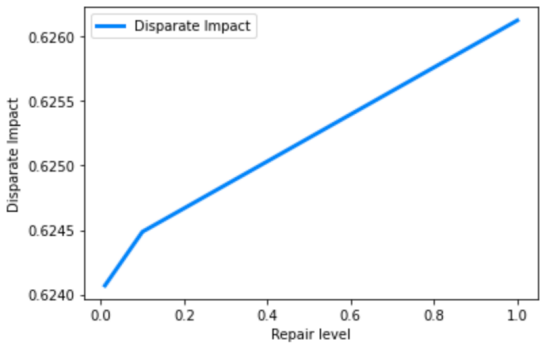




As repair level increased for the models with the DI-remover, overall accuracy decreased. So it seems that there is a tradeoff between decreasing disparate impact or decreasing accuracy. The same seems to be true for AUC and accuracy in the privileged group. The results are somewhat weird for unprivileged accuracy, as repair level increases, the accuracy switches from decreasing to increasing. As the repair level increases, the disparate impact increases proportionately, which makes a lot of sense. False positive rate difference also increases somewhat proportionately as repair level increases.

C.





The eta parameter has a much different effect on disparate impact than the earlier graphs that did not have the Disparate Impact Remover calculated in. The effect of repair level on accuracy is interesting. With the first increase in repair level, accuracy sharply decreases, and then over time as repair level increases, accuracy begins to increase as well.

Overall, accuracy is lower with the DI remover calibrated into the model. Disparate Impact is also much lower with hte DI remover calibrated into the model, as one would expect.

Problem 3

The data science application discussed in this lecture was AI, Artificial Intelligence. In this lecture there were many different stated purposes of AI, but to summarize, AI is often used by companies or organizations to improve efficiency, to remove human error from systems, to increase accessibility of different systems to people, to attempt to remove human bias from systems, and many others. Many organizations, industries, and populations can benefit from the application of Artificial Intelligence. For example, as we have discussed before, many recruiting systems will often use AI to analyze resumes. This benefits the companies that are hiring and recruiting because it reduces the amount of work that they must do to find qualified candidates, and thereby increases the efficiency of their recruiting process. The healthcare industry is another that is heavily benefiting from AI, often using algorithms to decide who needs treatment in a hospital, who needs organs on the organ transplant list, and there are algorithms being developed that will soon allow for prediction of illness and diagnosis in patients. The lecture also discussed how AI has also reached the restaurant/food delivery industry, as in this particular case, a company with a food delivery platform used AI to decide which employees should be assigned deliveries at what time.

Though AI has great potential and has had great benefits for countless industries and individuals, the bias existing in AI has harmed many people as well. The populations and groups that have been adversely affected and are most likely to continue to be adversely affected by AI are minority groups such as POC, women, those with disabilities, LGBTQ+, etc. In the recruiting example mentioned above, women and POC are often discriminated against because the AI that is used to analyze resumes often carries pre-existing biases that cause them to rank the resumes of white men above the resumes of POC and women. In the healthcare example, black patients have historically been ranked lower by organ donation algorithms as a result of either technical or pre-existing biases that result from misconceptions about the health of black people. In the specific food delivery example that was mentioned in the lecture, the AI used by the food delivery platform ended up discriminating against people who were not able to work during dinner time, which could have been biased against mothers who need to feed their children during dinner time (women), POC who cannot afford childcare, low-income people that need to be home during dinner time. This was a disparate impact that resulted from a technical bias because the system prioritized profit.

One example of disparate impact from Artificial Intelligence that was referenced in the lecture was when the US tried to use a model to decide whether people should be granted parole or not, and it ended up being heavily biased against black people. This resulted from pre-existing bias, and if it was put into practice, would have resulted in emergent bias.