Problems

1. Article: Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed

   Academics confirms that a major predictive policing algorithm is fundamentally flawed. The reason is the training data that this algorithm used is based on imperfect and cannot represent the whole crime data. The algorithm is inspired by statistical modeling method used to predict earthquakes but the key difference between earthquake and crime is that we basically can observe and obtain the earthquake data but not all crime activities are reported.

   Article: A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

   There's a debate about a computer program that used for bail and sentencing decision was considered biased against blacks as among the defendant who ultimately didn't reoffend, blacks were more than 2 as likely as white to be classified as medium or high risk. However, Northpointe stated that given a risk score, regardless the race, 60 percent of white defendants reoffended, 61 percent of black defendants reoffended. On this point of view, this tool is fair. We cannot be fair in both way.

   Article: Automating Bias

   Even using an online, automating algorithm, the process can be built based on designers assumptions. The first difficulty that this article mentioned was it's hard to make sure that all the families received all the benefits they were entitled to. An online application can be hard to apply for the poor families since they didn't have access to internet. Another problem is the data they relied on are only the people who reach out to public service for family support. It can skewed the algorithm too much because of oversurveillance.

2. We are surprised that in order to let the machine to learn a model, we should extremely carefully justify the data we have. If the way of collecting the training data cannot represent the population, then the model would have the issue the first article mentioned and even the issue that the article "Automating Bias" raised.

3. We're struggling that fact that we cannot find a better way to define fairness. As the second article mentioned it's impossible to satisfy two fairness standards simultaneously. And we had a debate about which standard raised by Northpointe and ProPublica is more robust than the other. It turned out that there's no definite conclusion for this answer.

4. We believe that the the most important ethical issue in the use of machine learning raised in these articles is can we be fair even if we only have imperfect training data. Our thought is inspired by the policing and defendant articles. Since we can only observe the data that was generated by human bias, is there a chance that we can use them to train an algorithm? Like the first article mentioned, we can only have the report records once they were been filed. And the same issue in the second article, people must be identified and reported to be a defendant. Under this circumstance, we still need to achieve the goal that, at least in one fairness standard, our model doesn't skew to a certain group's favor.

   This issue would have impact to any of us. If police officers are not going to patrolling our neighbor simply because a flawed algorithm told them this our area is safe but actually it's not, our neighbor might be in danger. And if an algorithm tends to use harsher treatment to the blacks, that can be interpreted as the algorithm treats white nicely and may have negative impact to our society.

5. In class, we mentioned three fairness standard.

   Anti-classification: The first one is the result should be the same for any two data points the unprotected features are the same.

Classification parity:

The second one is that the result should be the same no matter the protected characteristics are, and the false positive rate should also be them same regardless protected characteristics.

Calibration:

The third one is calibration, given the predicted result, the unprotected characteristics will not have any effect on the probability of the true value equals to a certain number.

We can focus on the issue we presented in `problem 4.` Even if with imperfect data, we still need to give a fair model. But the thing is there are three fairness definition, which one we would like to use? As stated in second article, `Northpointe` did not use `race` as input variable, so based on our opinion, it should satisfy the first fairness standard, that is, anti-classification. But then someone still points out that the people who ultimately didn't reoffend, blacks were more than twice as likely as white to be classified as medium or high risk. This represented the algorithm didn't satisfy the `Calibration`. But according to the designer, the model did meet the criteria of `classification parity`, that is, no matter defendant's race, the model labeled them fairly.

We conclude that there's no way to satisfy more than 1 notions in practice. Here's simple proof. Suppose we have 2 subgroups, `A` and `B`. `A` has higher reoffending rate than `B`. And there are two categories, `High` and `Low`. In order to let `High` category has higher reoffending rate, we need to put more `A` into `High` category, which means there's no way get equal for false positive rate in `High` and `Low`.

And even if we can satisfy more notions, there's must be some results can be identified as unfairness. So satisfying one or more than those notions cannot resolve these problems.

6. We would probably just follow the instructions we have. We'll not use protected features in our model, including race, gender, age, and etc.

7. The issue is a dilemma, depending on what we want to present then there's a different way to measure fairness. Our way is to not using protected features and let our algorithm learn a model. If that contains some result can be considered unfair, we'll reexamine our hypothesis.

The relevant decision makers is obviously our boss or supervisors XDD. After we assure that there's no protected features in our model, then the most important remaining issue would be whether our way to build the algorithm contains our assumptions that we are not realized. As the automating bias article mentioned during designing an algorithm, there's a chance that designers program our assumptions into the tool and hide consequential choices behind a math-washed facade.

We think our way to communicate with them is elaborating the design of the tool and carefully scrutinizing the fairness standard.

The technical part is can we find a way to have the dataset that can represent the population. We postulate it's the most important thing since our model would be based on the training dataset. The non-technical part is how can be convince people that our model can do accurate predictions without being unfair to a certain subgroups. That part needs some communications skills and persuasive stories.

Collaboration Statement

I collaborate this assignment with Jingru Hu, 466024, jingruhu@wustl.edu.