DS GA 1017 HW 1

Algorithmic Fairness

Joby George (jg6615)

Due 3/2/2022

Problem 1A Prompt:

Fairness from the point of view of different stakeholders

For each of the following metrics

A. Accuracy
B. Positive Predictive value
C. False Positive Rate
D. False Negative Rate
E. Statistical Parity

Provide a 1-2 sentence summary of which stakeholders benefit and are harmed by optimizing a model using that metric, and why. Add commentary on whether or not it would be reasonable within the context of the COMPAS (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing) investigation by ProPublica, and Northpointe's response (http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf)

Problem 1A Answer

Accuracy

Accuracy is defined as the number of correct predictions (TP and TN) divided by the total sample size (N). It gives a baseline estimate to the average performance of a model, but can often be a misleading evaluation criteria. For example, for cancer diagnosis, a simple yet highly accurate model would be predicting that no new patient has cancer, given the incidence rate of Cancer is approximately .22% globally.[4].

In the case of COMPAS, optimizing for accuracy benefits **Northepointe** and the direct interpreter's of the model's report (**the US judiciary system**) the most. By optimizing for accuracy, the model score can be interpreted in aggregate as a probability of recidivism.

Northepoint asserts there is an imbalance in the base rate between whites and blacks. Therefore, in optimizing accuracy, as later shown in Problem **2b**, the larger group will have more weight and receive more false negatives compared to the underprivileged groups, which will see more false positives.

While not perfect, optimizing for accuracy does have some advantages, namely the interpretability of the model in aggregate and a virtuous goal. If the model was 100% accurate, there would be fairness, however, this usually only happens in the physical sciences. The harms would be to the underprivileged groups, as they would be disparately impacted by the burden of false positive error. I still believe accuracy **is a reasonable metric to optimize for** in concert with other components.

Positive Predictive value

Positive Predictive value is defined as:

$$\frac{True\ Positive}{True\ Positive + False\ Positive}$$

.

Or more intuitively, if the COMPAS model scores someone as high risk, what is the **actual** probability of recidivism.

Just like in medicine, this is strongly determined by the base of the underlying phenomena (rate of recidivism and not recidivism). In optimizing for PPV, the model would try to be selective with positive predictions, this would hopefully alleviate false positive rate differences, however, it is not guaranteed.

I believe for this reason, PPV is a reasonable metric to optimize for.

False Positive Rate

False positive rate is defined as:

$$\frac{Labelled\ High\ risk\ and\ did\ not\ recidivate}{Didnot Recidivate}$$

If we optimize (minimize) false positive rate, we are incentivized to score people as less risky, and are more likely to make a false negative error. This would benefit defendants, as they would be less likely to have their rights infringed, however this would harm victims of recidivists. Given the high costs of a false negative, especially in the case of violent crime I would rank False Positive Rate as not reasonable to optimize for.

Intuitively, we want to make correct predictions, by optimizing for the minimal of false positive predictions, we could predict no one is likely to recidivate. If combined with another metric to balance metrics (accuracy, for example), it would be reasonable to optimize for, but by itself it can lead to bad models.

False Negative Rate

False negative rate is defined as:

$$\frac{Labelled\ low\ risk\ and\ did\ \ recidivate}{Recidivated}$$

Optimizing FNR would lead to a model that would be more likely to incarcerate people, harming defendants and the US tax payer. One could argue that crime would be lowered by the increased in incarcerated people, benefitting society as a whole, but it is not a guarantee.

However, I do think, with a carefully applied use of a COMPAS model optimized on False Negative Rate, that the tool could be used well. If the model is optimized towards false negative risk, then defendants that score low on the model's risk score would be unlikely to re-offend. If justices had that information it could help ease the burden on the judicial system and allow true negatives some relief. However, the model should not be used in a punitive way, such that high risk scores are used to penalize defendants. Meaning the justice system would still be overwhelmed by the number of cases it has to work.

The big risk would be whether the score distribution would still remain the same as it does by sub-groups, where Black Americans would have fewer in the least risky bucket after the algorithm is repurposed for a different type of prediction. With this in doubt, I would rank false negative rate as the second best metric to optimize for.

Statistical Parity

Statistical Parity is that the model's false positive and false negative error are balanced among sub groups. This would benefit defendants and the US justice system, as it does mean that the risk assessment could be employed without fear of racial bias. I would recommend optimizing for this metric, as while there may be more errors in general, no one group is bearing the majority of the consequences of the model's error.

While this would guarantee fairness, the broad usage and lower predictive power of the metric are downsides.

Sources:

4: CA: A Cancer Journal for Clinicians (https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21660)

Problem 1B Prompt: Tech Corp Recruiting System

Describe the:

1. Pre-existing Bias
2. Technical Bias
3. Emergent Bias

of an algorithmic hiring recommendation system, detailing

an example of how this type of bias may arise in the scenario described above
a stakeholder group that may be harmed by this type of bias
propose an intervention that may help mitigate this type of bias.

Problem 1B Answer:

Pre-existing Bias

Pre-existing Bias would exist in the data gathering process when training the AI. Tech as an industry is heavily dominated by men, meaning women resumes would be deprioritized by the algorithm. The primary stakeholders that would be hurt by this type of bias are those in small minority groups in the tech industry, whether racial (Blacks, Latinos), gender (women), or disability status, the algorithm would not prioritize these groups given these populations small prevalence in the training data.

An intervention that would help mitigate this type of bias would be over-sampling or synthetic data creation to ensure that the algorithm has enough sample of under-represented groups. This way the decision making system would not codify existing industry biases.

Technical Bias

Technical bias is a bias that is caused by limitations in technology or how it is used. In this case, the ADS sorts candidates by descending score. Candidates whose names fall outside the first screen's worth of top candidates, or are highly ranked but have a first and last name that starts with a Z would be chosen less frequently, given their worse placement in the sorting algorithm.

To remediate these separate biases, rather than a strict descending order that's displayed on a page that has to be iterated on for the next set fo candidates, the algorithm could present the top x candidates, in a shuffled order by rating and name. This way, someone with a last name starting with an A with the same score as someone with a last name starting with a Z would not be systemically favored. Additionally, top candidates would all be given a more equal chance by the recruiter, while the recruiter him or herself would be more engaged with each individual candidate as they have to apply their own heuristics to each candidate, rather than exclusively on the first page.

Emergent Bias

Emergent bias is a cause-and-effect feedback loop caused by ADS. In this case, groups that are already biased-against by the ADS are further discriminated by the ADS as the next iteration of training data is further skewed by the implementation of Prophecy.

For example, women of color might comprise 10% of ProphecyV1's training data. Once Prophecy is in use, the proportion of new hires that are women of color is now 5%, as the algorithm negatively penalizes candidates for these demographic attributes. Therefore, when re-versinong to ProphecyV2 and each future iteration, the model will more strongly penalize these demographic attributes as women of color are now further under-represented in the training data. Eventually, women of color will come to officially, or unofficially learn of this bias, and avoid applying to the company, completing the feed-back loop of emergent bias.

Remediating this bias can be done by the modeling or HR functions. This vicious cycle can be transformed into a virtuous cycle if the model developers and/or company correct for this bias by over-sampling, sample re-weighting, or modifying their loss function in each iteration of model training.

Alternatively, instituting explicit DEI organizational hiring policy will interrupt the emergent bias's feedback loop. If both practices are done simultaneously, the organizational will be able to cultivate a more tolerant culture that will create a virtuous cycle for talent.

Of important note in regards to emergent bias, is that it can be addressed through non-technical intervention. This highlights that when thinking of how to solve a problem data scientists should not limit themselves to only modeling based interventions

Problem 1C: A Prompt:

University Admissions at Best University are defined by:

1. SAT Score
2. GPA
3. Income Bracket (low/mid/high)

Describe the argument the Formal Equality of Opportunity (EOP) fairness doctrines would recommend to the admission office at Best University:

1. Formal Equality of Opportunity

1C: A Answer:

Formal EOP would require that the admissions criteria be the moment in time evaluation criteria (SAT/ GPA). The doctrine would be blind to Income Bracket as it deems it irrelevant to the qualification criteria, thus selecting the students with the best SAT and GPA would be a fair system that negates hereditary advantage.

Problem 1C: B Prompt:

Suppose that income-based differences are observed in applicants' SAT scores: the median score is lower for applicants from low-income families, as compared to those from medium- and high-income families. Which EO doctrine(s) is/are consistent with the goal of correcting such differences in the applicant pool? Briefly justify your answer

1C: B Answer:

The two doctrines that would attempt to correct for this are:

1. Substantive / Luck egalitarian
2. Substantive / Rawlsian.

Luck Egalitarians would posit that nothing you did not chose for yourself should impact your life prospects. Therefore, wealth, which is determined by birth as a child should not factor into the admissions criteria for Best University.

Rawlsian Egalitarians would posit a slightly different argument. Equally talented babies at birth must be given equal life prospects. Best university could either use equalized odds, such that the true positive rate and false positive rates are equal, or implement equality of opportunity which enforces true positive rate to be the same.

Equal Odds:

$$P(\hat{Y} = 1|income = Low, Y = 1) = P(\hat{Y} = 1|Income \neq Low, Y = 1)$$

and

$$P(\hat{Y} = 0|income = Low, Y = 0) = P(\hat{Y} = 0|Income \neq Low, Y = 0)$$

Equal Opportunity:

$$P(\hat{Y} = 1|income = (Low), Y = 1) = P(\hat{Y} = 1|Income \neq Low, Y = 1)$$

Problem 1C: C Prompt:

Describe an applicant selection procedure that is fair according to luck-egalitarian EO.

1C: C Answer:

Luck Egalitarians would posit that nothing you did not chose for yourself should impact your life prospects. Therefore, wealth, which is determined by birth as a child should not factor into the admissions criteria for Best University. Since this is impossible to enforce at the moment in time of admissions, we would compare members of the low income group to each other and members of the high income group to each other.

Problem 1D:

Prove that if the underlying base rate between two groups is different, we cannot simultaneoulsy achieve:

1. Equal Accuracy
2. Equal False Positive Rates
3. Equal False Negative Rates.

Definitions

Using the Compas example from Chouldechova's Fiar *Fair prediction with disparate impact* and the following definitions, listed below:

$$P_w \neq P_b$$

$$P = \frac{TP + FN}{N}$$

$$Accuracy = \frac{TP + TN}{N}$$

$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{FN + TP}$$

$$FP = FPR * N * (1 - p)$$

(from the [wikipedia (https://en.wikipedia.org/wiki/Sensitivity_and_specificity)](https://en.wikipedia.org/wiki/Sensitivity_and_specificity))

Solution

We know:

$$FNR_w = \frac{FN_w}{FN_w + TP_w}$$

Step 1

We can replace the denominator in FNR_w with the defintion:

$$N_w P_w = FN_w + TP_w$$

giving us

$$FNR_w = \frac{FN_w}{N_w P_w}$$

Step 2:

Next, we replace the numerator, using the definition:

$$FN_w = N_w P_w - TP_w$$

giving us

$$FNR_w = \frac{N_w P_w - TP_w}{N_w P_w}$$

Simplifying, we get:

$$FNR_w = 1 - \frac{TP_w}{N_w P_w}$$

Step 3:

Replacing TP_w using the following definition:

$$N_w Accuracy_w - TN_w = TP_w$$

we get:

$$FNR_w = 1 - \frac{N_w Accuracy_w - TN_w}{N_w P_w}$$

Simplifying this we get:

$$FNR_w = 1 - \frac{Accuracy_w}{P_w} - \frac{TN_w}{N_w P_w}$$

Step 4:

Now, we can substitue TN_w with FPR_w with the following formula:

$$TN_w = \frac{FP_w(1 - FPR_w)}{FPR_w}$$

$$FNR_w = 1 - \frac{Accuracy_w}{P_w} - \frac{\frac{FP_w(1-FPR_w)}{FPR_w}}{N_w P_w}$$

Step 5:

Now, substituting the definition of FP_w:

$$FP_w = N_w * (1 - p_w)FPR_w$$

we get:

$$FNR_w = 1 - \frac{Accuracy_w}{P_w} - \frac{\frac{N_w(1-p_w)(FPR_w)(1-FPR_w)}{FPR_w}}{N_w P_w}$$

Step 6:

Simplifying this we get

$$FNR_w = 1 - \frac{Accuracy_w}{P_w} - \frac{(1 - p_w)(1 - FPR_w)}{P_w}$$

Now, if our $P_w \neq P_b$ with $Accuracy_w = Accuracy_b$ and $FPR_w = FPR_b$, then it is impossible for FNR_w to equal FNR_b as FNR_w and FNR_b is a function of the three variables

We can show this mathematically:

$$FNR_w = FNR_b$$

Giving us:

$$FNR_w = 1 - \frac{Accuracy_b}{P_b} - \frac{(1 - P_b)(1 - FPR_b)}{P_b}$$

and

$$FNR_w = 1 - \frac{Accuracy_b}{P_w} - \frac{(1 - P_w)(1 - FPR_b)}{P_w}$$

Since $P_W \neq P_b$, but all other variables are equal we have proved our the contradiction, as FNR_{w} cannot equal those two things

Q.E.D

Problem 2

Problem 2a

Discuss your results of a baseline Random Forests's (defined as max_depth = 1, and n_estimators = 1) performance on the five metrics of interest:

Answer

We observe:

1. Accuracy .6146
2. Privileged group accuracy .6066
3. unprivileged groups accuracy is 0.6237
4. disparate impact is 0.8238
5. false positive difference rate is −0.1475

The model accuracy is not amazing with a max depth of 1 and one estimator, however, it is interesting to note that it performs better in regards to false positives for the unprivileged group, with a negative false positive difference rate.

Disparate impact of .823 implies that women receive the favorable outcome (income > 50k) at a ratio of .823 to 1, which shows some bias in the predicted labels. It is worth noting that the training dataset had a disparate impact ratio of .75, meaning the model was slightly more fair in its assignment of positive outcomes than the training dataset.

Prior to any modeling processes, it is important to note the MinMax Preprocessing that occurred for our numerical variables, which was chosen intentionally as the numerical variables (Hours worked per week) would not be normally distributed, therefore using the StandardScalar would have been inappropriate.

Problem 2b

Discuss the impact of hyper-parameter tuning of both fairness and accuracy for both models, and hypothesize about any differences between the models.

Answer

Hyper parameter tuning created a much more stable model, accuracy, for the privileged and unprivileged group went up and had a much smaller variance compared to the baseline model, which was much more reliant on the training data it saw.

However, disparate impact became noticeably worse, hovering in the .7-.77 range, while our untuned model had ranges from .8 to 1.2. False Positive difference rate was stable in a window between -.05 and -.03, meaning we were not incorrectly predicting women to have a high income at a rate noticeably different from men.

The main takeaways from hyperparameter tuning a model is that it can substantially improve performance on unseen data, and can do so consistently (lower variance in metrics) through the Random Search Cross Validation process.

The reason behind the difference in the models is that the untuned, shallow decision stump can only create a simple decision boundary, heavily dependent on the training data it sees. Whereas the tuned random forest, which had a max depth of 10 with 20 estimators can create a true ensemble model which has the impact of reducing variance due to training data and creating more a complex decision boundary.

Problem 2c

Discuss in your report how these results compare with the metrics from the baseline random forest model from (b), paying particular attention to the impact of repair level

Answer

When we vary the repair level from 0 to 1, we see an extreme degradation in accuracy, for all groups (privileged and unprivileged). At repair level = 0, the model is exactly the same as our model from **b**, and accuracy started at approximately .8. By with the first increment of repair level to .1, accuracy drops down to .55. From there, accuracy does not improve meaningfully, either in aggregate, or for the subgroups.

However, Disparate Impact which started at .88 for our optimal model from **b**, quickly rose and started converging near .98 by repair level =.2 indicating men and women were receiving favorable outcomes at almost the same rate.

Lastly, when looking at false positive difference rate, the tuned model with a repair level = 0, was close to -.01, when we modified repair level, false positive rate decreased further to -.015. After a repair level of .4, we see that false positive difference rate actually goes up all the way to .02, but quickly shoots back down to approximately the -.01 value. It's interesting that repair level has a nonlinear relationship with false-positive rate difference, making it tough to choose.

Problem 2d

Discuss in your report how the effect of the eta parameter compares to what you observed for DI-Remover. (Remember: Prejudice Remover is not a pre-processing method that is combined with an existing Random Forest model. It's a different model altogether, which fits a Logistic Regression under the hood.)

Answer

As we increase η from 0 to 1, we notice accuracy dips slowly from .76 to .75, until η = .8, where accuracy peaks at .775 , driven by a increase in unprivileged false accuracy.

Above .8, η accuracy metrics decrease slightly for all groups. The model's performance on these metrics across the η was consistent, with the large perturbations in the graph being differences no more than .03 on the y axis.

Compared to our tuned model, accuracy across the board is lower. This is likely due to the different model being used to predict the target label, logistic regression versus a Random Forest. However, when compared to our DI-removed model, accuracy was regularly better.

The Prejudice remover worsened disparate impact compared to our the tuned model and our DI-Removed model, with the value starting at .66 and peaking at .68 when η was .5, compared to a DI ratio of approximately 1 for our DI-removed model. Additionally, performance on the disparate impact ratio is also noticeably worse than our tuned random forest, which had a disparate impact in the .7 range.

Lastly, false positive difference rates were more skewed towards men, meaning men had a higher false positive rate across the board compared to our tuned model. Previously we were seeing false positive difference rates of approximately -.01, compared to a range of -.1 to -.07 it's most fair.

Problem 2e

Discuss in your report how these results compare with the metrics from the baseline random forest model from (b), paying particular attention to the impact of repair level, and how they compare with the results in (c).

Answer

Reject Option Post-Processing significantly improved performance across the board. Accuracy was on average, greater than 90%, far better than the tuned model, and tightly concentrated in this range. Additionally, accuracy was almost equivalent between our the privileged and unprivileged groups and tightly concentrated. Disparate impact had a wider

range, from .85 to 1.1, however, these values are considerably better than our tuned model. Interestingly False positive difference rate spiked, varying from .05 to .15, with an average of .1.

This does mean that our Post-Processed model was more likely to predict a false positive for our unprivileged group, compared to the tuned and baseline model.

When comparing this model to our DI Removed model, we notice again substantially better accuracy across all groups. Disparate impact is in a slightly wider range for our post processed model, however, the minimum (.85 and the maximum 1.1) contain the range of Disparate impacts throughout the repair level, which ranged from .88 to 1.02.

Lastly, False positive difference rate in comparing the Post-Processed model to the DI Removed model was substantially higher across all values of repair level.

The high False Positive Difference rate could be a function of the extremely high accuracy, and the difference in base rates of the labels in the training data itself.

Conclusion Prompt

Conclude your report with any general observations about the trends and trade-offs you observed in the performance of the fairness enhancing interventions with respect to the accuracy and fairness metrics

Conclusion write-up

This example project showed many interesting facets of modeling as it relates to fairness and performance.

A poorly tuned model, will be less performant for the privileged and under-privileged group compared to a tuned model. The fact that our untuned model was more variable in performance is a cause for concern, in addition to the fact that the range was generally lower than our tuned model, which had a very tight window on all it's performance and fairness metrics.

The trade-off we observed here was that Disparate Impact did concentrate at a more unfair level. Thinking about real models deployed that are finely tuned, we would tend to expect that these models are having a disparate impact on the unprivileged group as a result of the hyper-parameter tuning. Without explicitly correcting for this bias, we can see how emergent bias is created.

In analyzing the ways we can remediate the accuracy-fairness trade-off, the Disparate Impact Remover had a substantial impact on assigning equal amounts of labels to our two groups, however accuracy was substantially reduced.

Prejudice Remover was again less performant than our tuned RandomForest, most likely due to the difference in model choice. Additionally, the model prejudiced-removed model was more variable in performance, indicating that different datasets could result in significantly different performance across privileged and unprivileged groups in fairness metrics and accuracy.

Lastly, the post-processing showed the most amount of promise. Not only did it improve accuracy across groups, it increased Disparate Impact to a higher level compared to the tuned model. However, the variation in the Disparate Impact and False Positive difference rate is a cause for concern, as a model that is relatively fair and accurate one day, could become less fair.

Problem 3

Facial Recognition technology in Police Investigations

Police facial recognition to identify criminal suspects and potentially incarcerate them. based ADS.

The use case from a Data Scientists perspective is facial matching, using camera data to come up with the known person that matches that facial data. This general technology can be used for a number of purposes (FaceID, for example), but in this case the Police were using it to identify criminal suspects based on camera data.

beneficiaries

The primary stakeholders are the police investigations unit, and the corporations selling their technology.

Police Investigations

Facial recognition, if it was as infallible as DNA matching, would be a utopian technology for law enforcement as they could ID the true suspect with any picture of the criminal. However, our current technology has systemic performance problems with minorities, especially those with darker skin, leading to incorrect ID of suspects.

Tech Companies

The companies selling this facial recognition software to Police Departments are financially rewarded for their development of the facial recognition software, even if it comes with considerable flaws.

Caucasians Males

White men have been shown to have the most representation in the training data, and comprise a large portion of the technology professionals managing facial recognition software. In studies, it has been shown that the White Males have an extremely low error rate on commercial facial recognition software, this prevents them from being falsely incarcerated, and creates an illusion of accuracy in the model

penalized stakeholders

Women

Studies have found that women in general, and especially women PoC, facial recognition error rates are magnitudes higher compared to other demographic groups. This means that women, and women PoC, are more likely to be wrongly identified and incorrectly prosecuted as a result of bias.

PoC (and the intersection)

Similarly to women, PoC, and those with darker complexions are found have demonstrably lower accuracy. As mentioned before, the intersection of these groups see magnifying error rates and are thus exponentially more likely to be misidentified by the police.

Disparate Impact for PoC in facial recognition

Dr. Timnit Gebru and Joy Buolamwini's paper (http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf) contains the main known ethical issues with facial recognition today.

In the paper, the authors recap existing biases:

```
    1. Pre-existing bias of imbalanced datasets for intersecti
ons of PoC and gender
    2. Pre-existing bias, the companies that are responsible f
or fixing them are not incentivized to fix them or proclaim th
e model's errors to their customers.
    3.Emergent bias -- incorrectly incarcerating and violating
civil liberties of PoC as more localities and countries adopt
this law enforcement practice
```

In this case, a few large corporations backed out of agreements with law enforcement (https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/), however, Google, the company which notoriously fired Dr. Gebru (https://www.bbc.com/news/technology-55164324) is now claiming to have made significant progress in this space. (https://www.youtube.com/watch?v=gGZx0RrL-ow)

Word count = 464

Taking a Philosophical View on COMPAS

not a question assigned in the hw assignment

Before I started answering 1A, i tried to examine the full-extent of implications of the COMPAS ADS., applying some useful frameworks learned in class. **None of this was asked for in any questions**, but it helped me fully utilize core concepts of the course for better understanding.

When thinking about algorithmic fairness there are several important questions:

```
    1. Who are we building this model for?
    2. How do we define and measure whether the automated deci
sion system improves upon our current system.
    3. Algorithmic Impact Assessment (AIA) criteria
```

The stakeholders involved in the scope the COMPAS model are numerous and of extreme consequence:

```
1. The consumer: the United States Federal and State Governmen
ts and it's judiciary system
2. The producer: Northepoint
3. The subjects: Defendants on trial
4. Long term impacted stakeholders: Future United States citiz
ens
```

Who are We building this Model For

In the Mirror Mirror comic, a core question about Automated Decision Systems was: who are building models for?

These stakeholders impacted by COMPAS all have various different needs and priorities, touched upon below. Balancing the need of all stakeholders becomes more complex the bigger the picture becomes.

The United States itself

To create a more functioning society for everyone, the government of the United States would prefer to minimize crime, especially violent crime as it destabilizes communities and undermines the authority of the United States government itself. Examples of unfettered crime being disastrous to the average citizen include many narco-states that corrode the integrity of democracy and enforce the fairness motto of *might is right.*

However, minimizing crime through incarceration comes with a cost. The United States is the world leader in incarcerated population per 100,000 people[1]. Unsurprisingly when it comes to crime, and most importantly violent crime, the US is not the world's leader. According to

the World Bank, the United States has an intentional homicide rate of around 5 people per 100,000, closer to countries like Afghanistan (7/100k) and the Philippines (6/100k) than other Western Democracies France and Britain (both with 1/100K)[2]

It is evident from a performance, and financial perspective the United States government is eager to find a scalable, more performant and cheaper solution to criminal justice, where less people are in jail and less crime occurs, especially violent crime.

Northpointe

Northpointe has a simple goal compared to the consumer. Provide a model that is legally compliant and more performant than the existing system when it comes to reducing crime rates and incarceration years of non-recidivist defendants. If they can provide this solution, they have a very willing and large customer poised to buy their model.

Defendants on trial

Defendants are people who have been apprehended by police for committing some crime (whether or not they have actually committed the crime in question). The defendant is primarily interested in him or herself, maximizing their own utility. In the majority of cases, we could assume this would mean that the defendant pays a minimal bail or serves a minimum sentencing.

Future United States citizens

Future United Citizens are stakeholders that are impacted by emergent bias. If the COMPAS model is widely used and has a bias, certain communities can bear the cost of disproportionate incarceration which undoubtedly impacts their families and children.

America has seen parallels with this in the War on Drugs, where low level dealers of marijuana and crack cocaine were severely punished through the use of mandatory minimums compared to white drug users and dealers.

HumanRightsWatch.org states this succinctly: "In the poor urban minority communities from which most black drug offenders are taken, the high percentage of men and, increasingly, women sent to prison may also undermine their communities' moral and social cohesion. By damaging the human and social capital of already disadvantaged neighborhoods, the "war on drugs" may well be counterproductive, diminishing opportunities for social and economic mobility and even contributing to an increase in crime rates."[3]

Defining a better society

In class we've discussed the different viewpoints of fairness and what a more fair or better society would look like.

The three most commonly discussed viewpoints were:

```
    1. Utilitarianism, which boils down to creating the most g
ood for the most people,
    2. Raising the minimum, not compromising the well being of
an underprivileged minority
        to raise the average for a larger body.
    3. Rawl's theory of Justice, positing that if no one knew
who they would in a hypothetical society,
        the designers would design a system where any subgroup
would not be
        harmed to benefit another subgroup.
```

In thinking about what an improved society would look like, the main goals to optimize would be (in no particular order):

1. Crime Prevention

Minimizing the ability and frequency of crime, especially violent crimes which create loss of life for innocent victims.

2. Criminal Justice

Preventing someone who did not commit a crime to wrongly have their liberties infringed (inappropriate bail, inordinate sentencing).

3. Social Justice

Alleviating the burden of citizens for exploding financial costs from high incarceration rates and preventing any reformed system to propagate negative societal externalities that lead to worse societal outcomes.

Ideally, a system would co-optimize among all three of these priorities, but there exists an inherent tension, especially when considering the framework of fairness to be applied (utilitarianism vs raising the minimum vs Rawlsian).

Algorithmic Impact Assessment

The best framework we have now to think about how to balance risks in ADS is the AIA which posits:

```
1. The likelihood of harm caused by an ADS and
2. The severity of harm caused by an ADS
3. Determine the level of oversight on an ADS
```

Taking all of these things into account, I will now proceed to answer the primary question for Problem 1.

Sources:

1: BBC, prison stats (http://news.bbc.co.uk/2/shared/spl/hi/uk/06/prisons/html/nn2page1.stm)
2: World Bank Data (https://data.worldbank.org/indicator/VC.IHR.PSRC.P5?most_recent_value_desc=true)
3. humrightswatch.org (https://www.hrw.org/legacy/reports/2000/usa/Rcedrg00.htm)