Homework 1

Yunjeon Lee(yl7143)

March 5, 2022

1 Problem 1

Fairness from the point of view of different stakeholders

1.1 (a)

Consider the COMPAS investigation by ProPublica and Notrthpointe's response. (You may also wish to consult Northpointes's report.) For each metric A-E below, explain in 1-2 sentences which stakeholders would benefit from a model that optimizes that metric, and why. If you believe that it would not be reasonable to optimize that metric in this case, state so and explain why.

A. Accuracy
: African American groups are stakeholders and benefited from a model if the model optimize Accuracy. This is because currently the accuracy for violent crime is only 20% and many of defendants who got unfairly high risk scores are African Americans. After optimizing the accuracy, African Americans who got unfairly high scores can get their proper score.

B. Positive predictive value
: Neighborhoods are stakeholders and benefited from a model if the model optimize Positive predictive value. This is because if a defendant who got the higher score recidivate, that means this model is reliable and we can predict recidivism rate with this model in the future and be careful based on the predicted results.

C. False positive rate
: African American groups are stakeholders and benefited from a model if the model optimize False positive rate. This is because currently 44.9% of African American groups are predicted to be higher risk, but they did not recidivate in fact. In the process of optimizing the false positive rate, African American defendants who are classified as higher risk are reclassified and corrected as lower risk, and this will mitigate unfairness given to African American groups.

D. False negative rate
: Neighborhoods are stakeholders and benefited from a model if the model optimize False negative rate. This is because currently 47.7% of White defendants are predicted to be lower risk, but they did recidivate in fact. In the process of optimizing the false negative rate, White defendants who are classified as lower risk is reclassified and corrected as high risk.

E. Statistical parity (demographic parity among the individuals receiving any prediction)
: It would not be reasonable to optimize Statistical parity in this case. Considering that risk scores can be different based on attributes(especially with protected attributes), and there is a possibility that defendants who should get high risk score can get lower risk score than what they really should if we optimize with statistical parity.

1.2 (b)

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.
In their 1996 paper "Bias in computer systems", Friedman & Nissenbaum discussed three types of bias: A. pre-existing, B.technical, and C.emergent. We also discussed these types of bias in class and in the "All about that Bias" comic.

A. pre-existing
: For example, if Prophecy was trained on historical data about past employees which were predominantly male, the Prophecy will give results that male candidates are in a high rank. In other words, female candidates can be harmed by gender bias. In order to mitigate this bias, "gender" feature should be removed when training because feature of "gender" is not relevant with the job data scientist.

B. technical
: For example, if one of the candidates used special resume format to emphasize his or her skills, which is not traditional resume format and Prophecy is not trained to parsing this new resume format. In this case, because of the special formatting itself, the candidate has low probability of getting an interview opportunity and cannot pass the resume screening. Furthermore, if one of the candidates is good at new library and framework which is useful to job of data scientist, but if Prophecy cannot acknowledge that word because of lack of training of skillsets, then TechCorp loses the proficient candidate. In above examples, both the candidate and company are harmed by the bias. This is because the candidate cannot get interview opportunity and the company can lose proper and proficient candidates for the data scientist position.
In order to mitigate this bias, we should make a procedure to check the performance of Prophecy in the design process.

C. emergent
: For example, if Prophecy is trained and make a decision that prefer male candidates over time(by making decisions), then the system will penalize the resume of female applicants and resume that contains words related to the female such as women's club, women's in tech. In this case, female candidates can be harmed by the bias from the system that has cumulative decision making experience. In order to mitigate this bias, the system should not trained on the data whose decision is made by the machine itself.

1.3 (c)

Consider a hypothetical scenario in which an admissions officer at Best University is evaluating applicants based on 3 features: SAT score, high school GPA, and family income bracket(low, medium, high). We discussed several equality of opportunity (EO) doctrines in class and in the "Fairness and Friends" comic: formal, substantive / luck egalitarian, and substantive / Rawisian.

A. In a selection procedure that is fair according to formal EO, which of these features would the admissions officer use? Briefly justify your answer.
: The admissions officer would use SAT score and high school GPA for evaluating applicants based on formal EO, which appeals relevant skills in and irrelevant characteristics out. In the perspective of formal EO, applicants should only be evaluated by what they have(their ability - studying ability in this example), not by their background or special treatments. Among 3 features, SAT score and high school GAP show how much this applicant is prepared for studying at University. While family income bracket(low, medium, high) is not directly related to applicants' ability to study at Best University.

B. 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.
: Substantive EO doctrine is consistent with the goal of correcting such differences in the applicant pool. Applicants from medium- and high-income families might have more opportunity to prepare for the

exams(SAT, high school GPA), because parents can support students by registering private institutions and tutoring (because they are economically stable). From the perspective of students(applicants), living in medium- or high-income families is not their choice and it is "brute luck". In other words, the applicants(students) cannot choose families when they are born, so Subtantive EO whose motto is "Nothing that you did not choose for yourself should not affect life." is suitable for correcting this difference.

C. Describe an applicant selection procedure that is fair according to luck-egalitarian EO.
: Basic motto for luck-egalitarian EO is that nothing that you did not choose for yourself should not affect your life. In other words, outcomes should only be affected by "choice luck", not by "brute luck". Within this perspective, all of the 3 features can be influenced by brute luck and should be considered in the process. This is because, students who have grown up in rich family might have more time and materials(private institution or tutoring) to study for SAT score and high school GPA. While students who have grown up in low-income family might have less time to study(do part-time job for living) and less materials(relatively hard to have private institution or tutoring). Accordingly, students should compete with other students who are in the same family income bracket. Firstly, the headcount should be divided in to 3 (low, medium, high) groups. And for this 3 groups of headcount, it should be divided based on the ratio of low-income family, medium-income family and high-income family. The reason why I use ratio here is that if 3 groups are allocated the same number of headcounts, and if there are more headcount of low-income family students than low-income family students (for example), then all the students from low-income family get admission, even though the student do not study and get 0 point on the exams which is unfair for other students who studied hard regardless of their backgrounds. And then, within the headcount in each group, students should compete with others in the same family income bracket.

1.4 (d)

Consider a binary classification problem where the population consists of two groups. The "Fair prediction with disparate impact" by Chouldechova paper showed that if the base rate for the outcome of interest is different across groups – that is, if fraction of each group with a positive outcome is different – then no classifier can simultaneously achieve (i) equal positive predictive value, (ii) equal false positive rates, and (iii) equal false negative rates across groups.

Assumption 1: $P_A \neq P_B$
$P_A = \frac{TP_A+FN_A}{TP_A+FP_A+FN_A+TN_A}$
$P_B = \frac{TP_B+FN_B}{TP_B+FP_B+FN_B+TN_B}$

Assumption 2: $ACC_A = ACC_B$
$\Rightarrow \frac{TP_A+TN_A}{TP_A+FP_A+FN_A+TN_A} = \frac{TP_B+TN_B}{TP_B+FP_B+FN_B+TN_B}$

1 - ACC_A = 1 - ACC_B
$\Rightarrow \frac{FP_A+FN_A}{TP_A+FP_A+FN_A+TN_A} = \frac{FP_B+FN_B}{TP_B+FP_B+FN_B+TN_B}$

Assumption 3: $FPR_A = FPR_B$
$\Rightarrow \frac{FP_A}{FP_A+TN_A} = \frac{FP_B}{FP_B+TN_B}$

1 - $ACC_A = \frac{FP_A+FN_A}{TP_A+FP_A+FN_A+TN_A} = \frac{FP_A}{TP_A+FP_A+FN_A+TN_A} + \frac{FN_A}{TP_A+FP_A+FN_A+TN_A}$
$\Rightarrow (TP_A + FP_A + FN_A + TN_A)(1 - ACC_A) = FP_A + FN_A$
$\Rightarrow \frac{(TP_A+FP_A+FN_A+TN_A)(1-ACC_A)}{FP_A+TN_A} = \frac{FP_A}{FP_A+TN_A} + \frac{FN_A}{TP_A+FN_A} * \frac{TP_A+FN_A}{FP_A+TN_A}$
$= FPR_A + FNR_A * \frac{P_A*(TP_A+FP_A+FN_A+TN_A)}{FP_+TN_A}$
1 - $ACC_A = \frac{(FP_A+TN_A)FPR_A+(P_A*(TP_A+FP_A+FN_A+TN_A))*FNR_A}{TP_A+FP_A+FN_A+TN_A}$

1 - ACC_A = 1 - ACC_B
$\Rightarrow \frac{(FP_A+TN_A)*FPR_A+P_A*(TP_A+FP_A+FN_A+TN_A)*FNR_A}{TP_A+FP_A+FN_A+TN_A} = \frac{(FP_B+TN_B)*FPR_B+P_B*(TP_A+FP_A+FN_A+TN_A)*FNR_B}{TP_B+FP_B+FN_B+TN_B}$
$\Rightarrow$ (1-P_A)FPR_A + $\frac{P_A(TP_A+FP_A+FN_A+TN_A)FNR_A}{TP_A+FP_A+FN_A+TN_A}$ = (1-P_B)FPR_B + $\frac{P_B(TP_B+FP_B+FN_B+TN_B)FNR_B}{TP_B+FP_B+FN_B+TN_B}$

$\Rightarrow FPR_A$ - P_A FPR_A + P_A FNR_A = FPR_B - P_B FPR_B + P_B FNR_B
$\Rightarrow P_A$ (FNR_A - FPR_A) = P_B (FNR_B - FPR_B)

Since $FPR_A = FPR_B$ and $FNR_A = FNR_B$, $P_A = P_B$ from above equation.
However, we assumed that $P_A \neq P_B$ at first, the assumption contradicts. In other words, if $P_A \neq P_B$, equal accuracy, equal false positive rates, and equal false negative rates cannot be achieved.

2 Problem 2

Fairness-enhancing interventions in machine learning pipelines

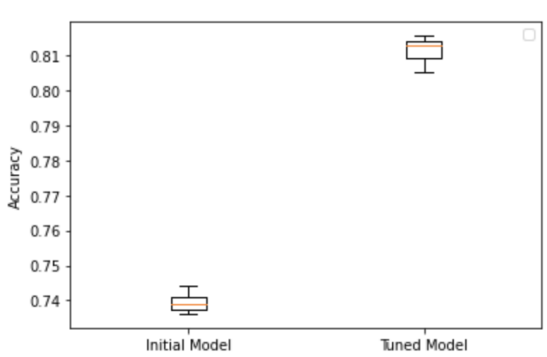
2.1 (a)

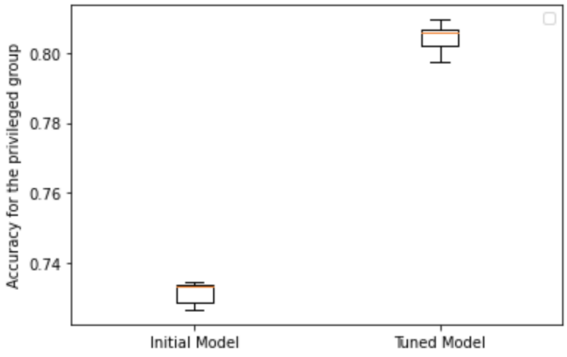
Overall accuracy:	0.740571
Accuracy for the privileged group:	0.732752
Accuracy for the unprivileged group	0.749175
Disparate Impact:	0.761670
False positive rate difference:	-0.053378

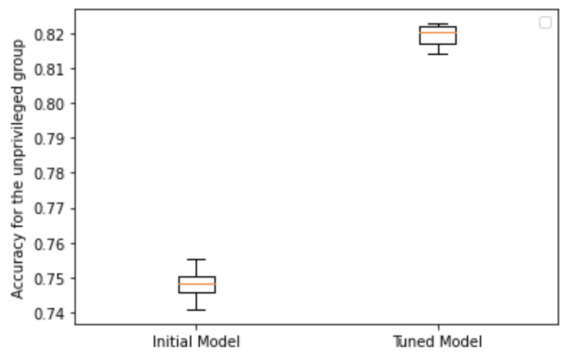
The baseline model has accuracy of 0.740571. And the accuracy for privileged and unprivileged groups are similar to overall accuracy which are 0.732752 and 0.749175 respectively. Since the Disparate Impact is less than 1, this is favorable to privileged group(male). For the False positive rate, which is $\frac{FalsePositivesofunprivileged}{N}$ - $\frac{FalsePositivesofprivileged}{N}$, is smaller than 0, and this means privileged group(male) has more people who are counted as higher income than their real income and this also shows the data is favorable to privileged group.

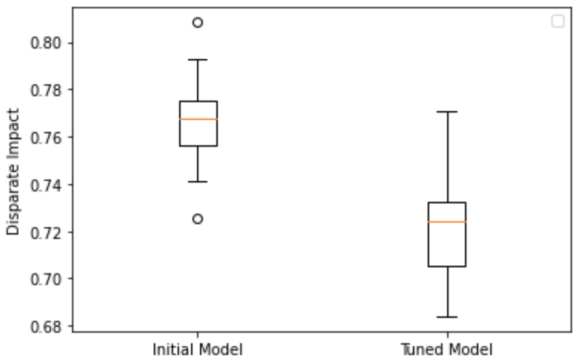
2.2 (b)

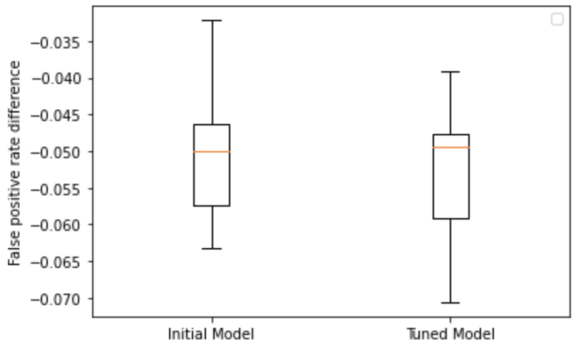
For the result of hyperparameter tuning, the accuracy of model is the highest when max_depth is 10 and n_estimators is 20.





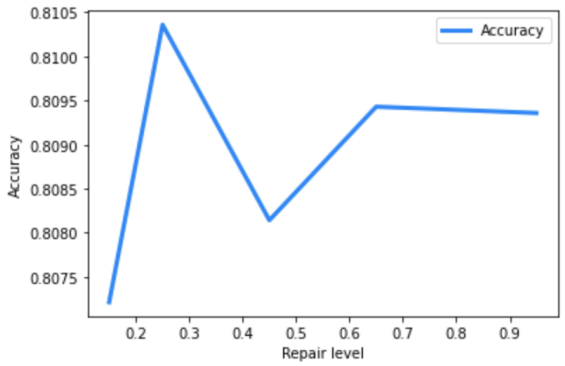


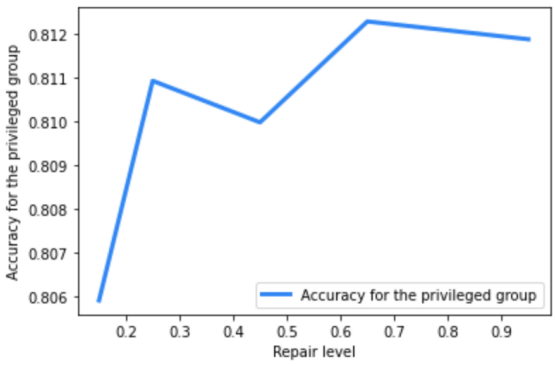


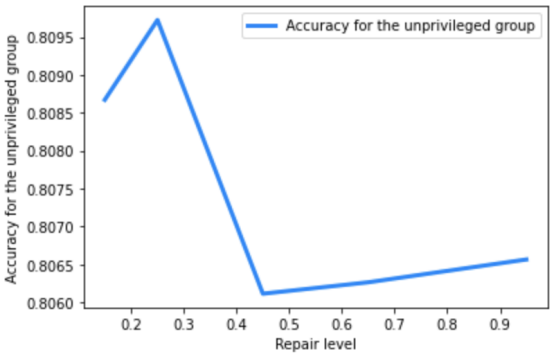


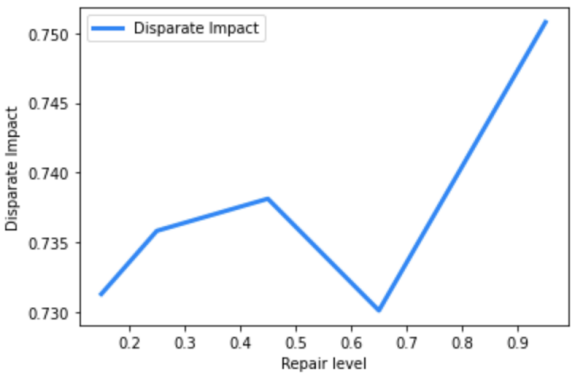
For the accuracy, it is improved to around 0.81 which was previously around 0.74. The main reason is that I select the hyperparameters based on the accuracy, which means I did hyperparameter tuning for the accuracy.
For the Accuracy for the privileged group it is improved to around 0.80 which was previously 0.72. And for the Accuracy for the unprivileged group, it is improved to around 0.82 which was previously around 0.74. The reason why accuracy is improved here is that I tuned the hyperparameters based on the accuracy.
However, in the perspective of fairness, it becomes more favorable to privileged group considering that Disparate Impact became around 0.73 which was around 0.77 before. In addition, False positive rate difference become more smaller which means more favorable to privileged group.

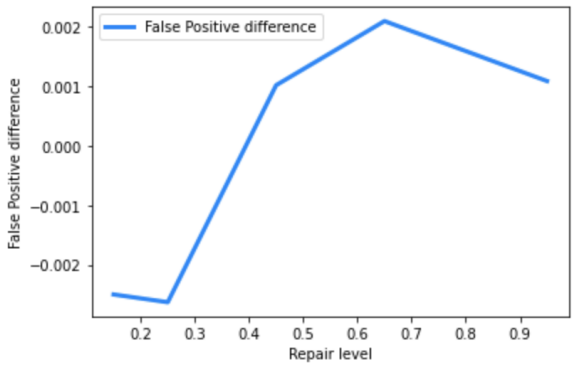
2.3 (c)





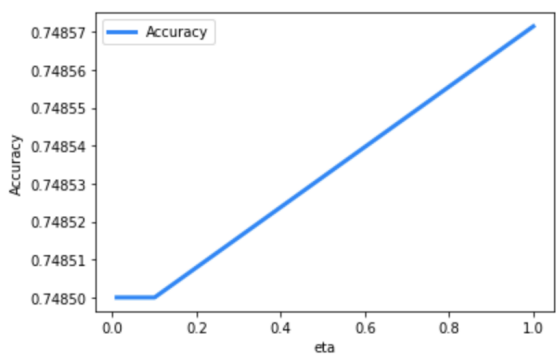


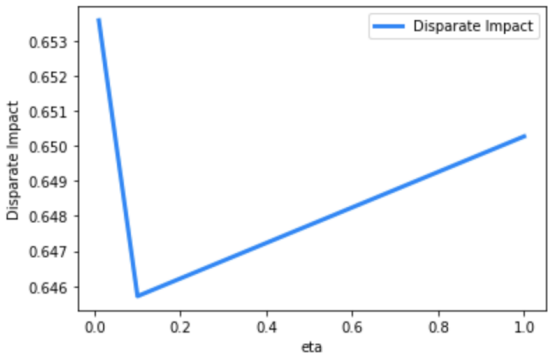




The overall accuracy is increased from 0.74 to 0.80 compared to baseline model. This is because of hyperparameter tuning for Random Forest model(max_depth:10, n_estimators:20). If I compare the performance with hyperparameter tuned model, then the accuracy is similar. For accuracy for privileged group, it is also increased to around 0.81 which was previously 0.7265189615332337. For accuracy for unprivileged group, it is increased to around 0.80 which was previously 0.7472527472527473. For the disparate impact, it increases as repair level increases. This is because from the DI-Remover, as the repair level increases it gives more effects of removing disparate impact, which means the model is closing to fairness. As repair level increases, it is closer to 0.75 which was previously 0.7409586541510732 in baseline model. This is also can be seen in the False Positive rate difference. As repair level increases, the value of False Positive rate difference becomes larger and becomes positive. This means the model becomes favorable to unprivileged group. Compared to baseline model which was previously -0.06325765966015812, False Positive rate difference becomes larger(from around -0.002 to 0.002). This also shows that it becomes favorable to the unprivileged group.

2.4 (d)





The overall accuracy becomes around 0.74 previously around 0.80. One of the reason is that we used Random Forest model with best hyperparameters in above problems while in this problem we used Prejudice Remover which is based on Logistic Regression model so the baseline can be different. And for the accuracy, if the value of eta increases, then the value of overall accuracy also increases.
For the Disparate Impact, the value increases when the eta increases. That means as the eta increases, it removes the effect of favorable to privileged groups and becomes favorable to unprivileged group compared to before.
Even though, the accuracy increases, it is almost the same as before when I just check the number itself. However, Disparate Impact itself increases in bigger differences, and it is because of Prejudice Remover in-processing technique used through the PrejudiceRemover class.

2.5 (e)

Accuracy	0.7958571428571428	0.8074285714285714	0.8081428571428572
privileged	0.7990961380443714	0.8060043080236942	0.8093873652612907
unprivileged	0.7923260674828307	0.8090383444917834	0.8067755958626892
Disparate Impact	0.908187	0.900053	0.914475
False positive rate difference	0.023066	0.005507	0.027320

Accuracy	0.8067142857142857	0.7994285714285714	0.8011428571428572
privileged	0.8076455491017155	0.8018139975632869	0.804920913884007
unprivileged	0.8056692435955738	0.7967639497958566	0.7969104952294411
Disparate Impact	0.914170	0.911531	0.918862
False positive rate difference	0.020907	0.021689	0.024174

Accuracy	0.8057857142857143	0.8013571428571429	0.8020714285714285
privileged	0.8061034115138592	0.8034993270524899	0.8032293377120964
unprivileged	0.8054187192118226	0.7989345509893455	0.800806933652122
Disparate Impact	0.906350	0.912120	0.901290
False positive rate difference	0.026616	0.018085	0.012766

Accuracy	0.7991428571428572
privileged	0.80209324452902
unprivileged	0.7958753575191931
Disparate Impact	0.918293
False positive rate difference	0.032997

The overall accuracy becomes around 0.80 which is lower than hyper parameter tuned model 0.81 and higher than baseline model from (b) and similar to model in (c). In other words, the overall accuracy is almost the same from (b), (c) and (e) (for the hyperparameter tuned model). For the accuracy of privileged group and the unprivileged group, it follows the values of accuracy. For the accuracy for the privileged group and accuracy for the unprivileged group, it is similar with the accuracy.
For the disparate impact, it becomes around 0.90 which is the best value in this Problem 2. In other words, Reject Option Classification shows the best performance for improving the Disparate Impact and gives fairness. In the initial model from (b), the disparate impact was around 0.77 and from the model in (c), the highest disparate impact was around 0.75. In (c), even though it removes with repair level, the highest disparate impact was around 0.75.
For the false positive rate difference, it becomes positive value which means the ratio of False Positive of unprivileged is larger than that of privileged. Compared to base line model from (b) which was around -0.05 it becomes more favorable to the unprivileged. In addition, when compared to the False Positive rate difference in (c), it is more favorable to the unprivileged group. In (c), as the repair level increases, the False Positive difference becomes larger and positive. The False positive rate difference from (e) which is from Reject Option Classification is larger and shows more favorable to the unprivileged group.

3 Problem 3

3.1 AI in recruiting

(a) Amazon recruiting
: The purpose of Amazon's AI recruiting system was to get rid of human bias with the help of AI and evaluate applicants based on applicants' ability(solely on indicators of success). In addition, the system can help to save time and money in recruiting and protect the safety and well-being of current employees and clients. However, because of pre-existing bias in tech industry, this system lead to gender discrimination in recruiting. Since there were much more male engineers in tech industry, the system is trained based on the data of male engineers, and this leads to make deduction in the keywords that relates to female. In other words, male applicants are benefited(stakeholders) of this

system, while female applicants are adversely affected(undervalued). In this system, disparate treatment based on the gender leads to disparate impact which means female applicants are undervalued compared to male applicants.

(b) Predictim
: The purpose of Predictim is to help finding suitable caregivers such as babysitters and evaluate potential employees based their SNS services. However, because of pre-existing bias in race and gender, this system leads to race and gender discrimination in the process of hiring caregivers. For example, this system high valued Nick who uses bad words in SNS, while undervalued Kianah who is kind and nice person. Considering that Nick is a White male and Kianah is a Black female, the system has bias on race and gender. In other words, because of pre-existing bias in race and gender White male are benefited(stakeholders), and Black female are adversely affected(undervalued). In this system, disparate treatment based on the race and and gender leads to disparate impact which means Black females who are nice and kind are undervalued compared to White male who are saying bad words and risky.

3.2 AI in healthcare

(a) AI for benefits allocation
: The purpose of AI system in healthcare is to reduce in benefits fraud, to reduce in biased decision making by caseworkers, efficiency in scheduling, and automated detection catches health indicators providers might overlook. In the example, Sophie need $6000 per month to treat her illness. However, her mother did not submit the application form for health subsidiary because she could not get the form, so Sophie could not get the subsidiary. The system concluded that the Sophie's case as intentional lack of compliance. In this case, the stake holders are people working in the related departments because they can work efficiently. However, in this case, one of the patients were adversely affected because of omission.

(b) AI for healthcare access
: The purpose of AI system for healthcare access was to efficiently allocating the time of patients by booking less or more time for each patients. However, this system turned out to be overbooking the time for Black patients. In addition, in the high-risk management access system, it makes a decision based on the cost of care, not by the health needs. In the system, the risk score of Black patients are more severe than it really is, because the system undervalued the risk of the Black patients. In these case, the stakeholder is the hospital who wants to maximize their profit, and Black patients are adversely affected. In these systems, disparate treatment based on the race leads to disparate impact which means Black patients are undervalued their risk and overbooked by the system.

References