paper2

June 9, 2022

1 The Inputs, Outputs, Decisions Being Made, and Relationship Between Decisions and True Outcome

The New York City Civilian Complaint Review Board (CCRB), the system in this context, is an agency independent from the New York Police Department that investigates complaints against NYPD officers regarding allegations of excessive or unnecessary force, abuse of authority, discourtesy, or offensive language. Civilians can file complaints to the CCRB online, in-person, over the phone, or via mail. In the online form, the inputs include information regarding the complainant, details regarding the incident, any information on the victim(s)/witness(es) if the complainant is not sole victim, and any information regarding the accused officer. Complainants can also submit any supporting documents. Throughout the investigation process, the review board may obtain further evidence from the NYPD (e.g., body camera footage) and conduct interviews of any complainants, victims, witnesses, and officers involved. Specific complainant/victim attributes collected include age, gender, and ethnicity. The complainant may also provide details on the officer's sex, race, and physical appearance.

The dataset only contains information regarding the complainant, the accused officer, and the nature of the complaint/allegation. It also contains general information regarding the complaint itself, like the dates it was submitted and closed and the precinct it took place in. The dataset does not include any additional evidence obtained after the complaint was filed (e.g., interviews).

The output of the decision-making system is the same as the decision being made – the disposition. Put simply, the CCRB chooses to "substantiate," meaning "there was a preponderance of evidence that the acts alleged occurred and constituted misconduct," a complaint or not (Title 38-A: Civilian Complaint Review Board - New York City 2018). Since the decision-making system (the CCRB) is deciding whether the allegations occurred, the true outcome is whether the CCRB is actually correct. In other words, the true outcome is whether the alleged actions in question (one complaint can have multiple allegations) actually occurred, whereas the output/decision made by the CCRB is only the result of their investigations and taken as a proxy for the ground truth.

2 Model Card

2.1 Description of the Model

To replicate this classification of complaints into substantiated or unsubstantiated, a logistic regression modle with L2 regularization was used. To train the model, a simple 75%-25% train-test split was used. One parameter used is the `class_weight` parameter, which assigns a weight to each class that the model uses for penalizing. The class weights calculated using `n_samples / n_classes * np.bincount(substantiated)` help combat the class imbalance in the dataset (only about 25%

of the observations are substantiated) and make substantiated observations more important for the model. The features used in the model are `contact_reason` (or text indicating why the officer approached the civilian), `mos_ethinicity` (officer's ethnicity), `rank_incident` (officer's rank at time of incident), `mos_gender` (officer's gender), `complainant_gender` (complainant's gender), `mos_age_incident` (officer's age at time of incident), `complainant_age_incident` (complainant's age at time of incident). `borough` (the borough in which the incident took place), `black` (whether the complainant is Black), `allegation` (brief description of the allegation), `fado_type` (type of complaint), and time/date related features (`month_received`, `year_received`, `month_closed`, and `year_closed`). All categorical features except `allegation` and `fado_type` were one-hot encoded while the exceptions were ordinal encoded. The numerical features were scaled.

2.2 Intended Use

Primary intended uses: This model would be used to help determine whether a given allegation should be substantiated or not. An allegation is substantiated if there is enough evidence to determine that the alleged action(s) happened and violated NYPD rules. Presumably, the model would not be the sole decision maker in this context; investigators would use the outcome of the model as supplementary information.

Intended users: The intended users of this model is the New York City Civilian Complaints Review Board and its Board members. Since the CCRB is independent of the NYPD, people affiliated with the NYPD should not use this model to, for example, try and predict if behavior against a given civilian would result in punishment.

Out of-scope use cases: Any use cases related to the law enforcement system (e.g., use by the police departments or criminal courts) would be out of scope. Since this model is designed specifically for civilian complaints, it should not be used in relation to criminal activity. Furthermore, the CCRB only investigates complaints in the four specific categories aforementioned, so any complaints that fall outside of the CCRB's jurisdiction would be out-of-scope for this model.

2.3 Factors

2.3.1 Relevant factors

One of the most relevant groups of this dataset and the model is complainant ethnicity. As described more in depth in Paper 1, policing in the United States has a racially charged background that continues to affect how Americans are policed. Given disparities in policing across different neighborhoods and different groups' perceptions of the police, we can assume that people of different ethnicities interact with the police in different ways. Additionally, despite only making up around 20% of NYC's population, Black people made up most complainants in the dataset the model is built on (New York City Department of City Planning | Population Division, 2022). Thus, the model's performance will be looked at for Black complainants vs. non-Black complainants.

2.3.2 Evaluation factors:

Complainant ethnicity is the factor reported because of its apparent relevance in the distribution of the dataset and dispositions of complaints (whether a complaint is substantiated or not). The historical significance of ethnicity in the context of policing make this factor important to consider.

2.4 Evaluation Data

Datasets: The dataset used for training and testing comes from ProPublica and consists of 31,686 unique complaints to the New York Police Department submitted by 11,312 complainants for the Civilian Complaint Review Board's review from January 2000 to January 2020 (*The NYPD Files*, 2020). Each complaint comes with information regarding the complainant's demographics, the police officer's information and demographics, the nature of the complaint and the alleged action in question, and whether the CCRB deemed the allegation as substantiated.

Motivation: This dataset was introduced to me through DSC 80, so having done previous analysis with this dataset, I chose to continue analyzing the dataset. Its context is highly relevant to today's society and discussion on policing in the United States.

Preprocessing: Before creating the model, the allegations column was simplified so that it can be binned. All categorical variables were one hot or ordinal encoded and all quantitative variables were scaled. The disposition column (target variable) was also simplified to simply include substantiated and not substantiated. The complainant ethnicity column was binned into Black and non-Black.

2.5 Metrics

2.5.1 Model performance measures

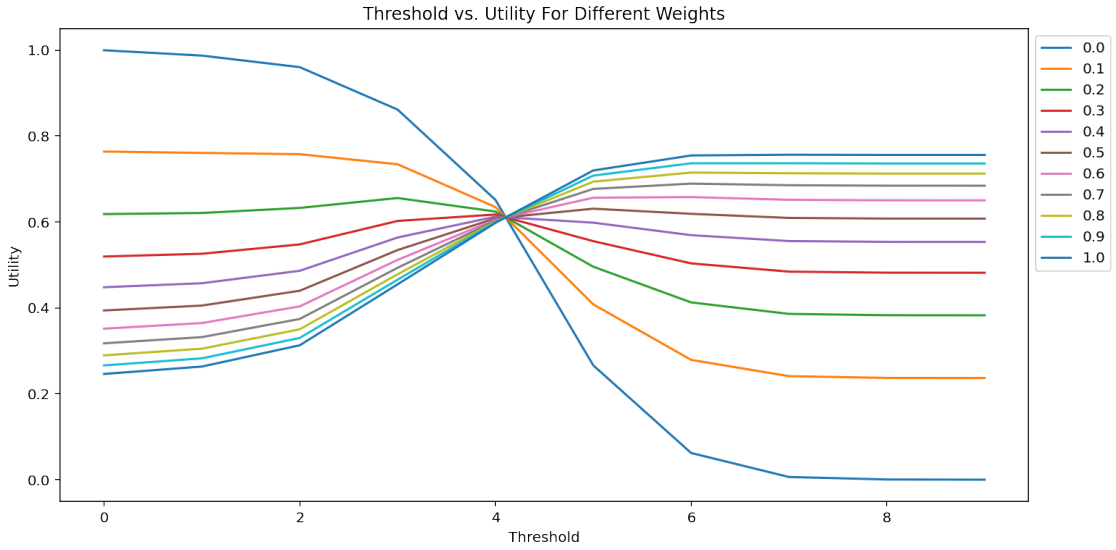
Accuracy is not the focus of this model because of the class imbalance in the data. Since 75% of allegations were unsubstantiated, the model could theoretically predict the negative class all the time and achieve a decent accuracy of 75%. Instead, recall is one of the primary model performance metrics. Since the CCRB is independent of the NYPD and wants to encourage civilians to file complaints if they feel they've been victims of police misconduct, the model should prioritize substantiating all complaints that should be substantiated. The CCRB rarely substantiates any allegations, so a model that attempts to find all complaints that should be substantiated would be a large improvement from the previous decision-making system. However, it is also important that the CCRB doesn't recklessly substantiate too many underserving complaints because the board has a reputation to uphold. If the board makes too many mistakes, then neither the NYPD nor the civilians would be well served. This is why precision is also important for this model. So since recall and precision are important, the F1-score, or the harmonic mean of recall and precision, is the primary metric.

2.5.2 Decision thresholds

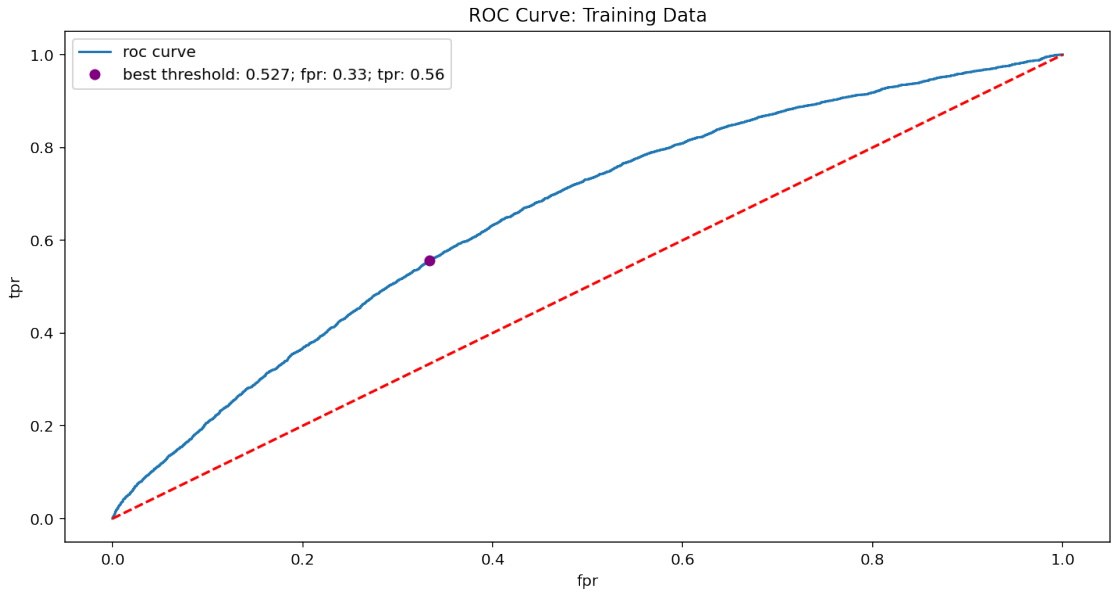
In order to choose a suitable threshold for all complaints, different utility functions were compared, as demonstrated in *Fairness & Algorithmic Decision Making* (Fraenkel). The utility function calculates weighted accuracy as a means to compare different thresholds and weights. In this model, a positive prediction corresponds to substantiation, which yields the positive utility (sense of justice) for the complainant. Thus, positive predictions should be weighed at least as much as negative predictions (unsubstantiated). Since we do not have access to the ground truth (whether the alleged action actually occured/if the CCRB is correct), the CCRB's decision is used as a proxy for it and is used to calculated utility.

At weights 0.0-0.3, the utility is higher for thresholds lower than 0.5, while for weights 0.4-1.0, utility is higher for thresholds above 0.5. This means that when substantiation's weight isn't much higher than unsubstantiation's, the threshold is at least 0.5. When substantiation is weighted over 2x as more important the unsubstantiation, the threshold is below 0.5. Since the CCRB was designed

to deal with police misconduct, substantiating complaints and ensuring officers are reprimanded for misconduct is a higher priority than ensuring complainants cannot recklessly submit false complaints. With this in mind and how utility fluctuates as described above and in the graphs, a weight of 0.4 seems reasonable. This weight indicates that substantiation is over 2x more important than unsubstantiation.



With a weight of 0.4, the threshold that maximizes utility is 0.527. As shown in the ROC curve below, for the training set, a threshold of 0.527 yields a false positive rate of 0.33 and a true positive rate of 0.56.



3 Performance Metrics

Unitary Results: The performance of the model for Black complainants was slightly better than non-Black complainants in terms of accuracy, but the test recall and precision were much better for the non-Black complainants than the Black complainants. This may mean that the model does a better job at detecting negative cases in general. Since there are more negative cases in the original dataset for Black complainants, accuracy for Black complainants is stronger but the metrics that look at the positive classes (precision and recall) are lower. There is a higher proportion of positive classes for non-Black complainants, so this may be why the model's precision and recall is stronger for non-Black complainants.

```
Test Performance Metrics For All Groups:
Accuracy Score: 0.623882406425216
Recall score: 0.526413921690491
Precision score: 0.3299571484222828
F1 score: 0.4056513409961685

Test Performance Metrics For Black Complainants:
Accuracy Score: 0.6321701720841301
Recall score: 0.49948400412796695
Precision score: 0.31469440832249673
F1 score: 0.3861188671719187

Test Performance Metrics for non-Black Complainants
Accuracy Score: 0.6095238095238096
Recall score: 0.5671875
Precision score: 0.35276967930029157
F1 score: 0.4349910125823847
```

3.1 Ethical Considerations

The data contains three protected classes – complainant/officer ethnicity, complainant/officer gender, and complainant/officer age. While the model's intended use doesn't involve anything that's central to human life, one may argue that civilians should have the right to critique and make complaints against the policing body. For the sake of this exercise, no risks or harms were explicitly considered during model building; this may be addressed in the audit. This model could theoretically be used to punish or not punish an officer that has broken the rules or give/deny civilians who were victims of police misconduct a sense of justice, so its use case can have some considerable ramifications on one's life. Furthermore, since the CCRB is a government body, a model such as this one that uses protected classes may not be apporpriate; the CCRB may be legally obliged to omit these proected classes.

3.2 Caveats and Recommendations

If this model were to be used by the CCRB, it should not replace the investigation and input of the board, but may be used as an ancillary factor in the ultimate disposition. As for caveats of the dataset, the limited view of complainant ethnicity (e.g., doesn't consider Afro-Latino people) could mean that a significant group in NYC's population wasn't explicitly considered. Since about 75% of complaints fall in the negative class (unsubstantiated), it would be ideal that we use a dataset with more balanced classes for both training and testing.

4 Parity Measures, Utility, and Fairness

In this context, utility refers to the sense of justice a complainant, who feels that they're a victim of police misconduct, receives when their complaint(s) get substantiated. Actual utility would be the result the model yields. While not observable, effort-based utility can refer to the ground truth; if a reported incident actually happened, then the complaint should be substantiated and the complainant should receive that sense of justice.

Accuracy: In Paper 1, it was mentioned that accuracy parity isn't a strong parity metric to enforce because a model that either substantiates all complaints or unsubstantiates all complaints would satisfy accuracy parity, but not be useful in reality. Similarly, accuracy itself is not a strong performance metric because of the class imbalance; an accuracy of 75% may sound strong, but the model could just be predicting unsubstantiated every time an achieve this accuracy. The model above achieved a test accuracy of about 63%. For comparison, the COMPAS algorithm which involves the criminal justice system had an accuracy of about 68% (Julia Angwin, 2016). A machine learning algorithm predicting Supreme Court behavior had about 70% accuracy (Katz et al.). Considering the limited data and minimal feature engineering, an accuracy of 63% doesn't seem wholly unreasonable.

Recall: Maximizing recall would maximize utility (sense of justice) for the complainants. However, if recall is maximized and there are some false positives, innocent police officers may be unjustly penalized. The model yielded a test recall score of about 52%, meaning the model correctly classified 52% of all substantiated complaints in the test set. 52% of these complainants would rightfully receive a sense of justice (the utility) from this model, while the remaining 48% of missed out true substantiated cases wouldn't.

Precision: Since there is a tradeoff between recall and precision, the test precision is only 33%. This means that the model correctly predicted the positive class for 33% of complaints. The false positive complaints would mean that the alleged police officers would lose utility because of the punishment they would receive and potential damage to their reputation. A low precision score isn't necessarily detrimental to complainants, so long as complaints that should in reality be substantiated are correctly classified, regardless of the false positive rate. On the other hand, a low precision may mean a lot of false positives, which would mean a lot of police officers are wrongfully reprimanded. However, one may argue that since the CCRB is independent of the NYPD that their priority should be civilian utility.

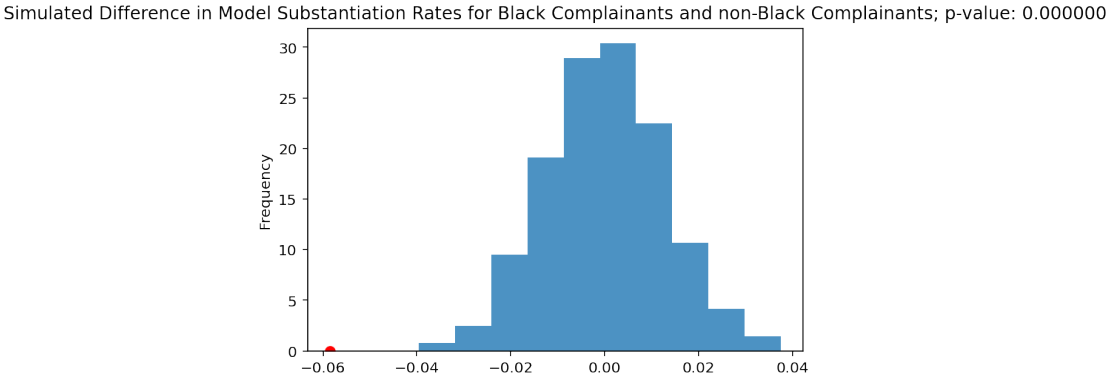
4.1 Demographic Parity

For demographic parity, we expect the model rates of substantiation to be equal across complainant ethnicity and borough. Enforcing demographic parity implies that effort-based utility is actually constant across all groups. In this context, that would mean that all complainants, whether or not their complaint is true, have an intrinsic right to be substantiated by the CCRB. Assuming that no group is more likely to submit a false complaint, this notion of constant effort-based utility can hold.

In Paper 1, the data showed that complaints submitted by Black complainants were substantiated about 2% less of the time than complaints submitted by non-Black complainants. The model yielded an even stronger violation of demographic parity by ethnicity. The rate of positive classification (substantiation) for complaints submitted by Black complainants is 6% less than that for complaints submitted by non-Black complainants, a difference that is significantly different from

0 by permutation test at significance level 0.05. The model amplified the disparity seen in the original dataset.

```
Model Substantiation Rates for Black Complainants vs. non-Black Complainants

 classification  Substantiated  Overall Proportion
 black
 False                0.426087            0.365965
 True                 0.367591            0.634035

Observed Difference in Model Substantiation Rate for Black vs. non-Black
Complainants: -0.05849613434200679
```



4.2 Equality of Odds

With the original dataset analyzed in Paper 1, there was no attribute reflecting the ground truth, which was stated to be whether the CCRB correctly deemed a complaint as (un)substantiated. While the ground truth is still unobservable, we can use the actual CCRB disposition as a *proxy* for Y (the true binary label). Using the CCRB's decision as a proxy, we can calculate false positive and true positive rates, meaning we can calculate equality of odds. Furthermore, if we assume that the proxy for true label (whether the CCRB substantiates a given complaint) reflects a complainant's effort-based utility (i.e., a complaint that should be substantiated is actually substantiated by the CCRB), then Fair Equality of Opportunity translates to equality of odds. As discussed in Paper 1, assuming that no particular group is more likely to submit a false complaint, all complainants across groups are equally qualified for substantiation. So, since we expect Fair Equality of Opportunity to hold, we would also expect the model to satisfy Equality of Odds.

$$P(C = 1|Y = 1, \text{complainant ethnicity} = \text{Black}) = P(C = 1|Y = 1, \text{complainant ethnicity} \neq \text{Black})$$

and

$$P(C = 1|Y = 0, \text{complainant ethnicity} = \text{Black}) = P(C = 1|Y = 0, \text{complainant ethnicity} \neq \text{Black})$$

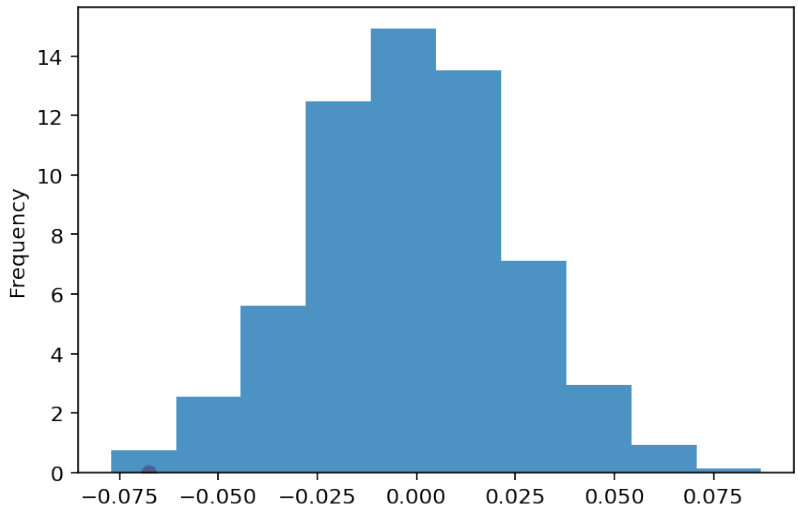
The true positive rate for Black complainants is about 7% lower than that for non-Black complainants, meaning that the model missed about 7% more substantiated complaints for Black

complainants. The false positive rate for Black complainants is about 5% lower than for non-Black complainants. The model gave more complaints submitted by non-Black complainants the "benefit of the doubt" and classified more CCRB-unsubstantiated complaints as substantiated than for Black complainants. This violation of equality of odds is verified via permutation test in which the differences in TPR and FPR both yielded significant p-values.

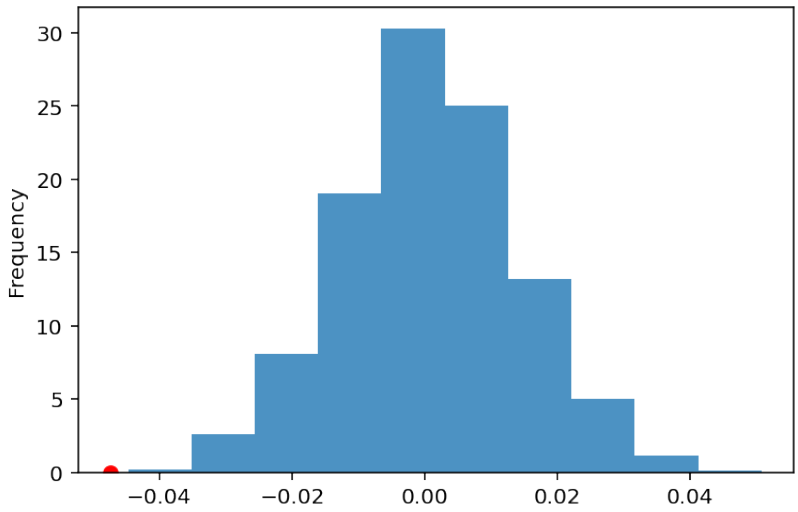
`Observed Difference in True Positive Rate for Black complainants vs. non-Black complainants: -0.067703495872033`

`Observed Difference in False Positive Rate for Black complainants vs. non-Black complainants: -0.047373009440781544`

Simulated Difference in TPR for Black vs. non-Black Complainants; p-value: 0.008000



Simulated Difference in FPR for Black vs. non-Black Complainants; p-value: 0.001000

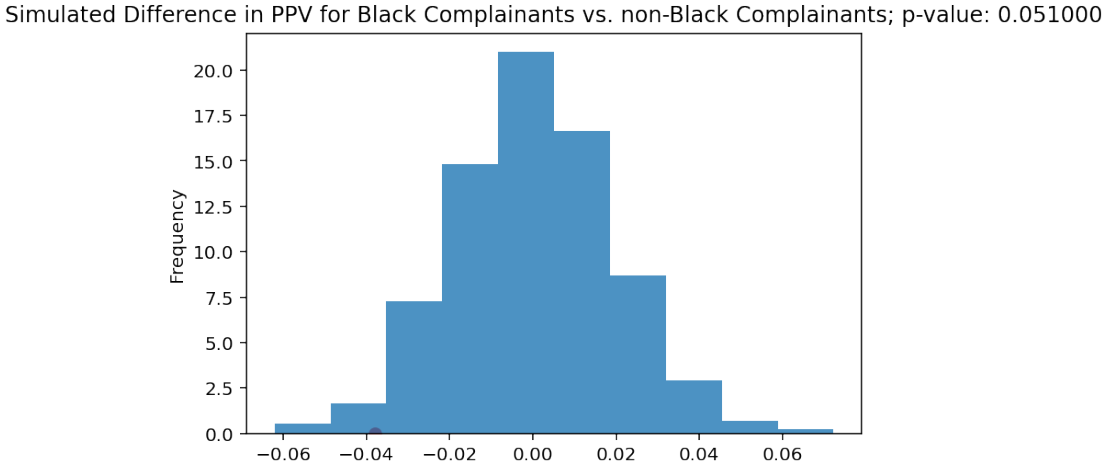


4.3 Predictive Value Parity

This parity measure requries that the true substantiation rates are the same across groups (or circumstances). Since we assume that complainant ethnicity (S) and the model (C) are not independent of the CCRB's decision (Y), then predictive value parity cannot hold if demographic parity or equality of odds hold. But, since neither of those parity metrics hold, we might expect PPV-Parity and NPV-Parity to hold. Satisfying predictive value parity would translate to luck egalitarianism if we assume effort-based utility is equal to the model prediction (i.e., a complaint that should get substantiated is substantiated by the model) and actual utility (what the model classifies) is the same as what the CCRB rules. Since the ground truth is unobservable, the CCRB's decision is again used as a *proxy*; Y in this case represents a proxy for the ground truth, not the actual ground truth.

PPV-Parity: $P(Y = 1|C = 1, \text{complainant ethnicity} = \text{Black}) = P(Y = 1|C = 1, \text{complainant ethnicity} \neq \text{Black})$. The observed difference between positive predictive value for Black complainants vs. non-Black complainants is about -0.038. While this isn't 0, a permutation test with a significance level of 0.05 yields a p-value of 0.051, indicating that this difference in PPV is not significantly different from 0. Thus, PPV-parity is satisfied.

```
Observed Difference in Positive Predictive Value for Black complainants vs. non-
Black complainants: -0.03807527097779484
```



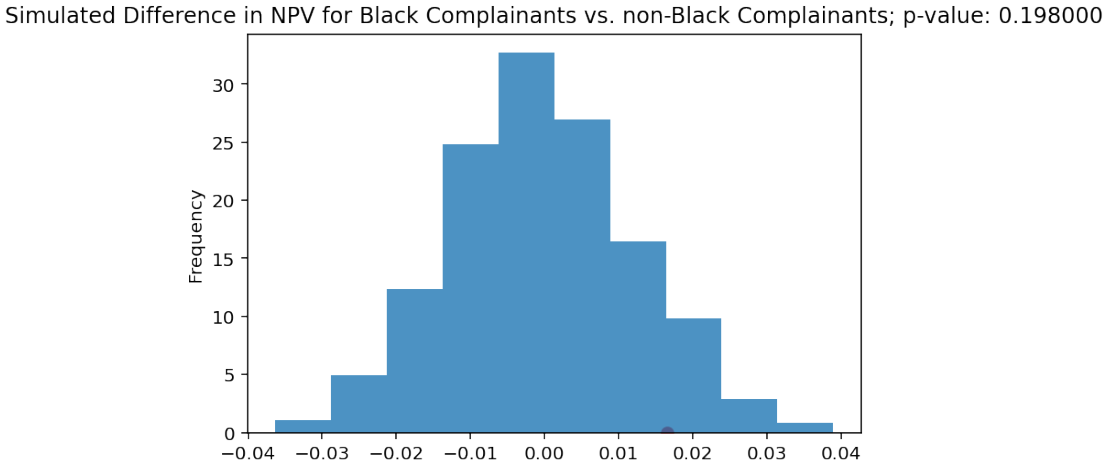
NPV-Parity: $P(Y = 0|C = 0, \text{complainant ethnicity} = \text{Black}) = P(Y = 0|C = 0, \text{complainant ethnicity} \neq \text{Black})$

The negative predictive value parity for Black complainants is about 0.016 higher than that for non-Black complainants. A permutiation test with significance level of 0.05 yielded a p-value of 0.198, meaning this difference is not statistically different from 0. Thus NPV-parity is satisfied.

Since both PPV-parity and NPV-parity between Black and non-Black complainants are satisfied, predictive value parity is satisfied. This means that the model does an equally "good" job of ensuring complaints that should be substantiated (as determined by the proxy - CCRB's original decision) are correctly classified and that complaints that should be unsubstantiated are predicted the negative class. Since demographic parity and equalized odds parity cannot simulataneously hold

(assuming the proxy for the ground truth Y, the CCRB's decision, is dependent on complainant ethnicity and the model C), it makes sense that predictive value parity does.

`Observed Difference in Negative Predictive Value for Black complainants vs. non-Black complainants: 0.016560159417302267`



5 Threshold Tests

5.1 Group Aware Thresholds For Maximizing Utility

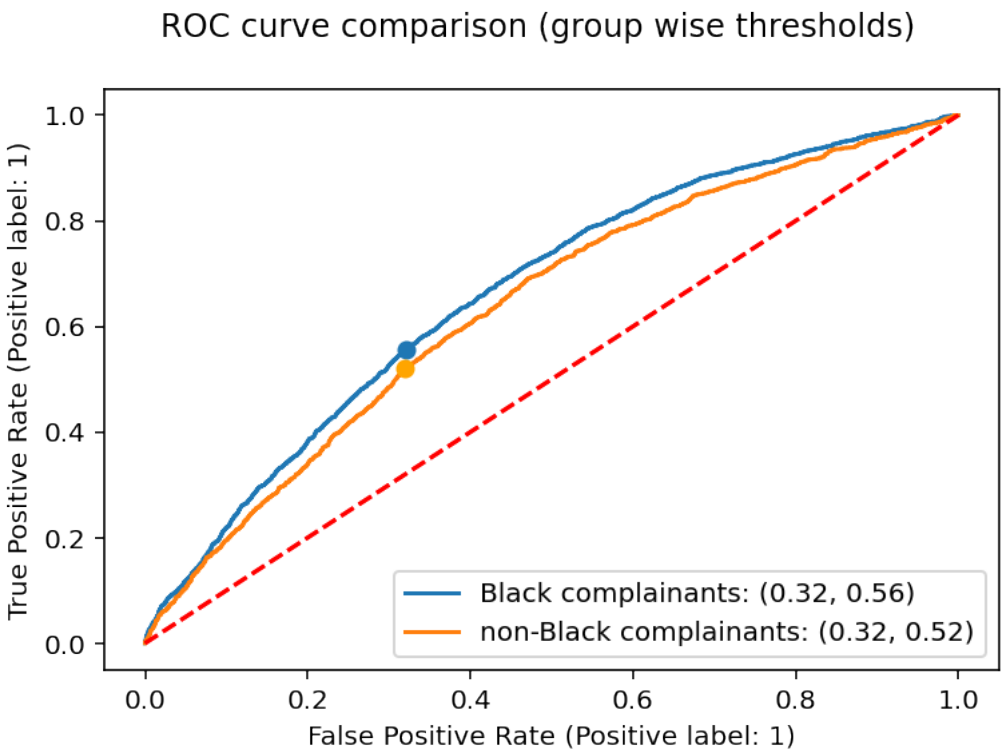
When creating the original model, the entire dataset was used to determine an optimal threshold. Alternatively, using the same logic, we can find the optimal threshold for a given *group* by simply providing only the samples within that group. The same function/methodology of maximizing utility and the same weight of 0.4 will be used to find the group-wise best thresholds. Once the group-wise best thresholds are found with the training data, the parity measures can be recalculated on the test data to see if the new thresholds yield any changes or improvements.

`Overall Best Threshold: 0.527`

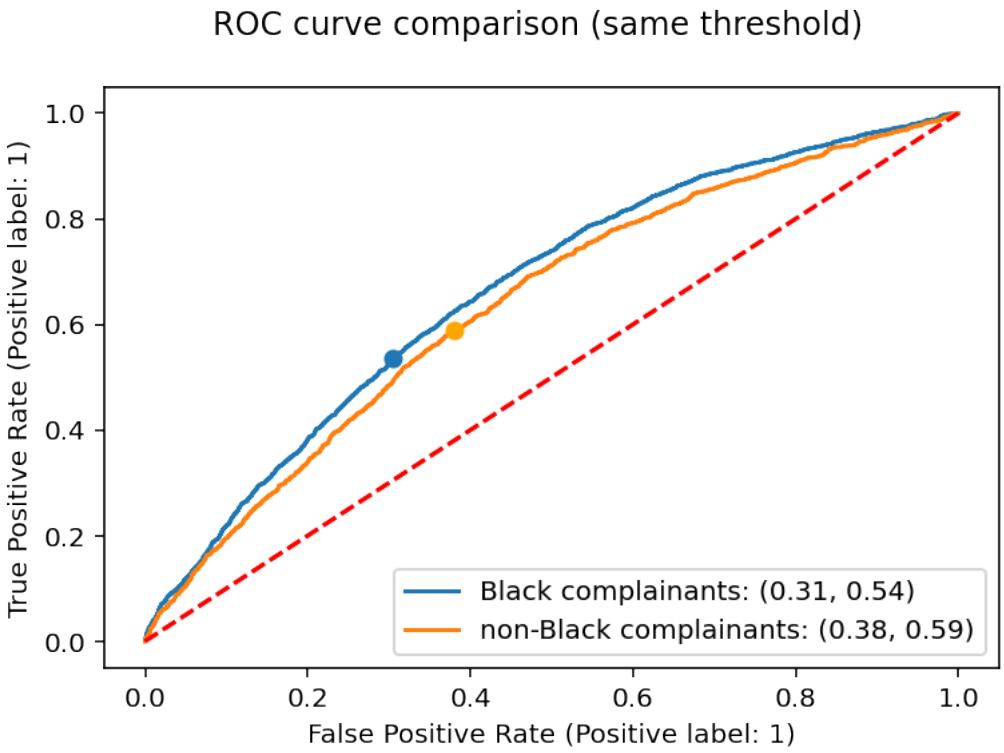
For Black complainants, the best threshold is 0.522, corresponding to a false positive rate of about 0.32 and a true positive rate of 0.56. The best threshold for non-Black complainants is 0.546, associated with a FPR of 0.32 and TPR of 0.52. At these groupwise thresholds, the model performs fairly similar, with the model performing a bit better for Black complainants with the higher true positive rate and fairly similar false positive rate.

`Best Threshold for Black Complainants: 0.522`

`Best Threshold for non-Black Complainants: 0.546`



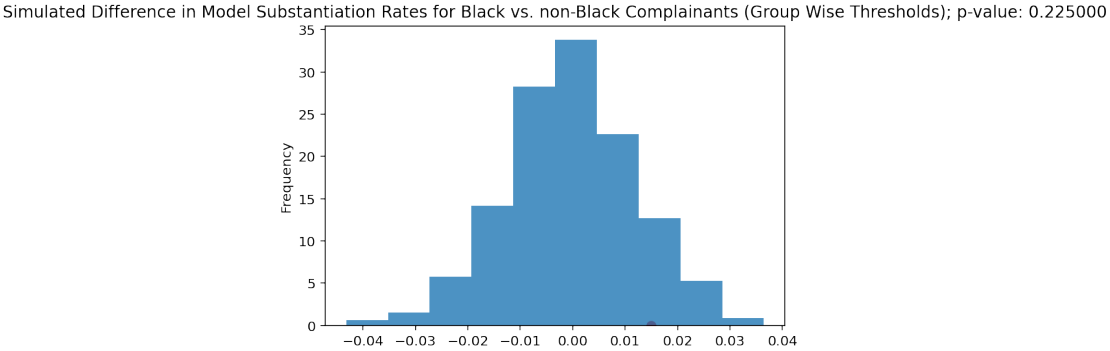
When comparing this to how the performance overall best threshold (0.527), the false positive rate for Black complainants is a bit lower, but the true positive rate is also lower. For non-Black complainants, both the false positive and true positive rates increased. Additionally, the FPR and TPR of each group under the single threshold are further apart. It seems that choosing a threshold for each group made the model perform more similarly per group.



5.1.1 Demographic Parity

The group wise thresholds that maximize utility also happen to satsify demographic parity. With an observed difference in substantiation rate between Black and non-Black complainants of +0.015 and a permutation test revealing this difference to not be significantly different from 0, the thresholds of 0.522 for Black complainants and 0.54 for non-Black complainants satisfy demographic parity. When compared to the single threshold results, the group wise thresholds substantiate a bit more complaints from Black civilians and a bit less complainants from non-Black civilians.

```
Model Substantiation Rates for Black Complainants vs. non-Black Complainants

 classification_gw  Substantiated  Overall Proportion
 black
 False                  0.370600            0.365965
 True                   0.385516            0.634035

Observed Difference in Model Substantiation Rate for Black vs. non-Black
Complainants: 0.014915838311382434
```

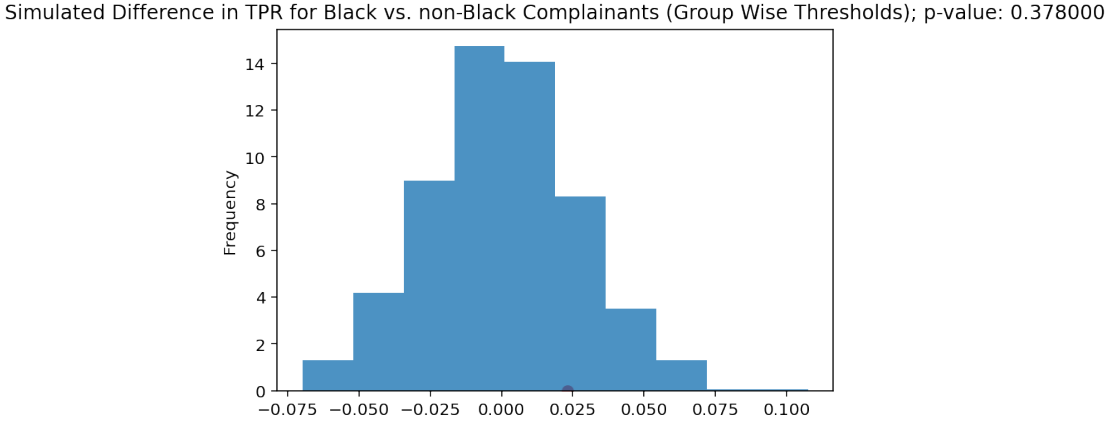


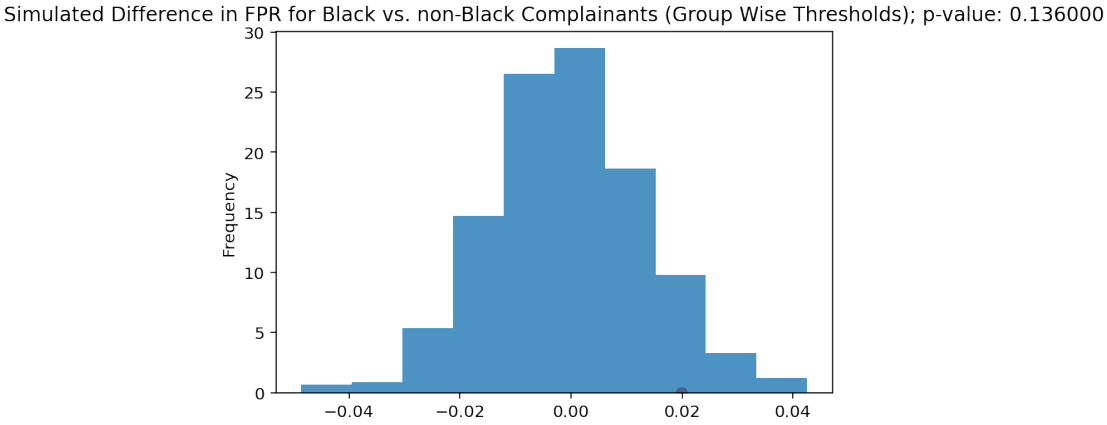
5.1.2 Equality of Odds

Additionally, the group wise thresholds that maximize utility also satisfy equality of odds. The observed difference in true positive parity between Black and non-Black complainants is +0.02, and the difference in false positive parity is -0.047. With the permutation test yielding high p-values for both true positive and false positive parity, equality of odds is satisfied.

```
Observed Difference in True Positive Rate for Black complainants vs. non-Black
complainants: 0.023248839009287925
```

```
Observed Difference in False Positive Rate for Black complainants vs. non-Black
complainants: -0.047373009440781544
```

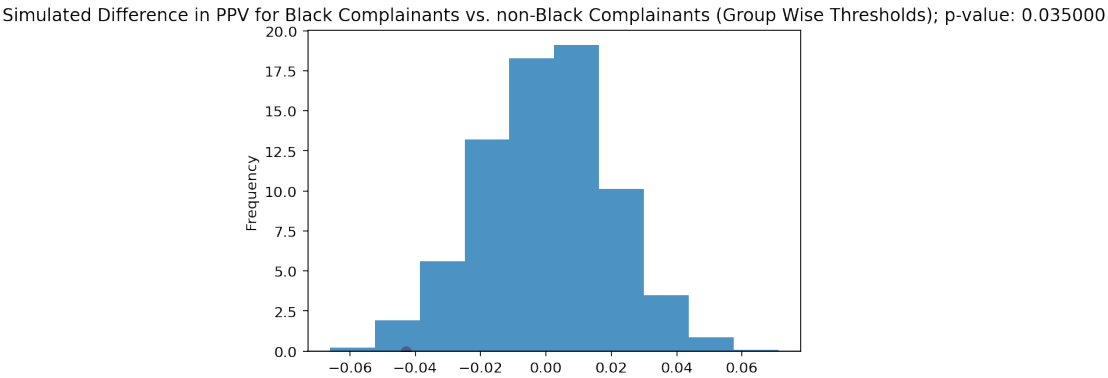




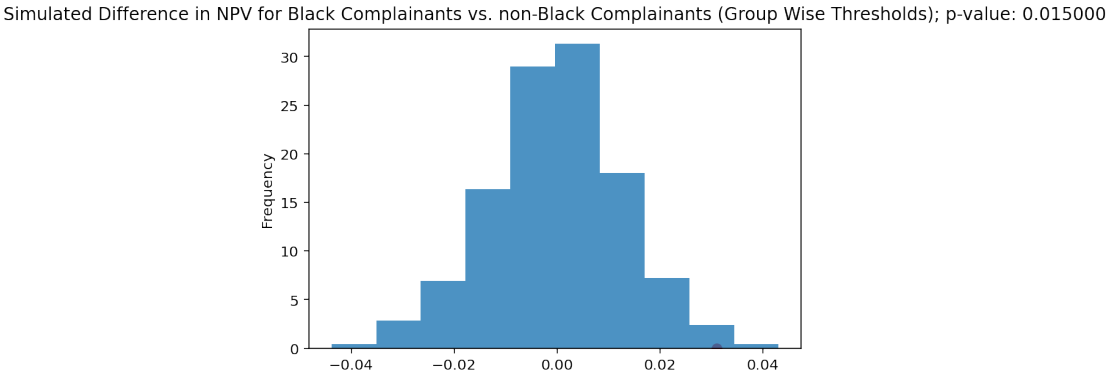
5.1.3 Predictive Value Parity

With the single threshold of 0.527 that maximizes utility overall, predicitve value parity was already satisfied. Under a significance level of 0.05, the two-sided permuation test yielded p-values smaller than 0.05, meaning the observed differences in positive value parity (-0.035) and negative value parity (0.015) resulting from the group wise thresholds are significantly different from 0. Predictive value parity doesn't hold when thresholds of 0.522 for Black complainants and 0.54 for non-Black complainants are used. This is expected if we assume that complainant ethnicity and the *proxy* (CCRB decision) for the ground truth Y (whether the CCRB is correct) are not independent.

`Observed Difference in Positive Predictive Value for Black complainants vs. non-Black complainants: -0.04284601024497192`



`Observed Difference in Negative Predictive Value for Black complainants vs. non-Black complainants: 0.030978628016950283`



5.2 Group Thresholds for Parity

Another way to decide on group-wise thresholds is to find thresholds for each group that satisfy parity. To do so, we'll first use the training set to determine a set of thresholds for which parity holds. Then, using these thresholds and the test data, the parity measures will be recalculated to see if they hold with "real world"/"unseen" data.

5.2.1 Demographic Parity

Since the dataset is fairly small, the approach was to simply loop through possible threshold combibnations and find the ones for which demographic parity is satisfied. The 2 thresholds for which demographic parity is satisfied (on the training set), ended up being the same thresholds found that maximize utility per group: 0.522 for Black complainants and 0.546 for non-Black complainants. These thresholds already proved to satisfy demographic parity on the test set, so the permuation tests weren't re-performed.

`Thresholds That Satisfy Demographic Parity and Yield Highest Utility`

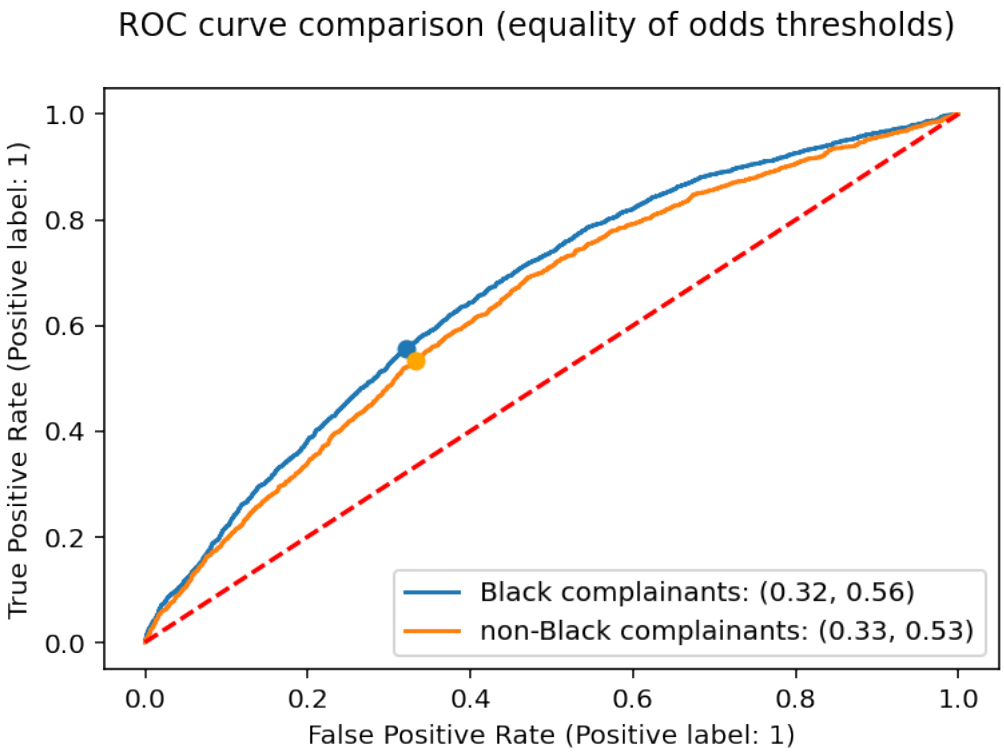
	black_threshold	nb_threshold	black_utility	nb_utility	mean_utility
160	0.522	0.546	0.625187	0.606897	0.616042

With the utility-maximizing singular threshold, only predictive value parity was satisfied. Since demographic parity and equalized odds parity cannot simulataneously hold (assuming the proxy for the ground truth Y, the CCRB's decision, is dependent on complainant ethnicity and the model C), there are likely different thresholds where equalized odds parity is satisfied, but not demographic parity. Similarly, since demographic parity and predictive value parity cannot simulataneously hold (assuming the proxy for the ground truth Y, the CCRB's decision, is dependent on complainant ethnicity), there are likely different thresholds where predictive value parity is satisfied, but not demographic parity. In this context, while it is assumed that all groups should have an equal chance at being substantiated (and thus demographic parity should be met), it may be more valuable to equalize false and true positive rates or predicitve value across groups. Thus, these other parity measures will also be explored with group wise thresholds.

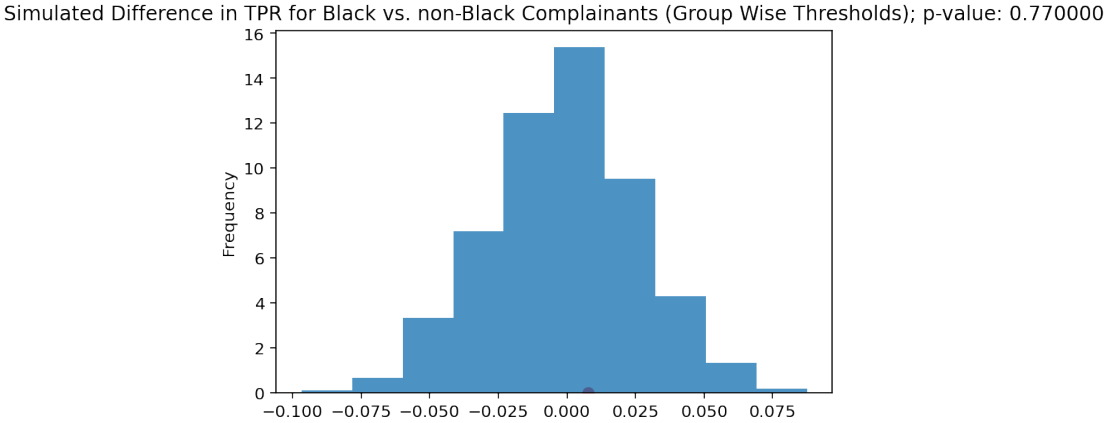
5.2.2 Equality of Odds

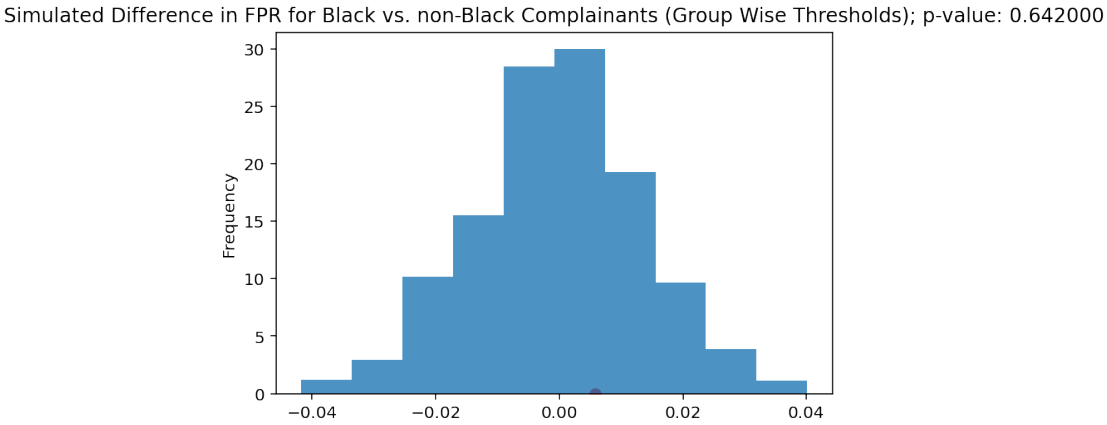
Taking a similar approach as with demographic parity, the thresholds that satisfy equality of odds on the training set are 0.522 for Black complainants and 0.541 for non-Black complainants. This is fairly similar to the results found earlier when thresholds were found so that utility is maximized for each group. When calculating equality of odds on the test set and running a permutation test, it was found that these thresholds yield differences in false positive and true positive parity that aren't significantly different from 0, meaning equality of odds is satisfied. While both these results and the previous results satisfied equality of odds, these new thresholds more strictly enforce equality of odds and only allow for differences in TPR/FPR of less than 2%; the observed differences with these thresholds are under 1%.

```
Thresholds That Satisfy Equalized Odds Parity and Yield Highest Utility

   black_threshold  nb_threshold  black_utility  nb_utility  mean_utility
6            0.522         0.541       0.625187    0.606414      0.615801
```




```
Observed Difference in True Positive Rate for Black complainants vs. non-Black
complainants: 0.00762383900928798

Observed Difference in False Positive Rate for Black complainants vs. non-Black
complainants: 0.005790638074168186
```

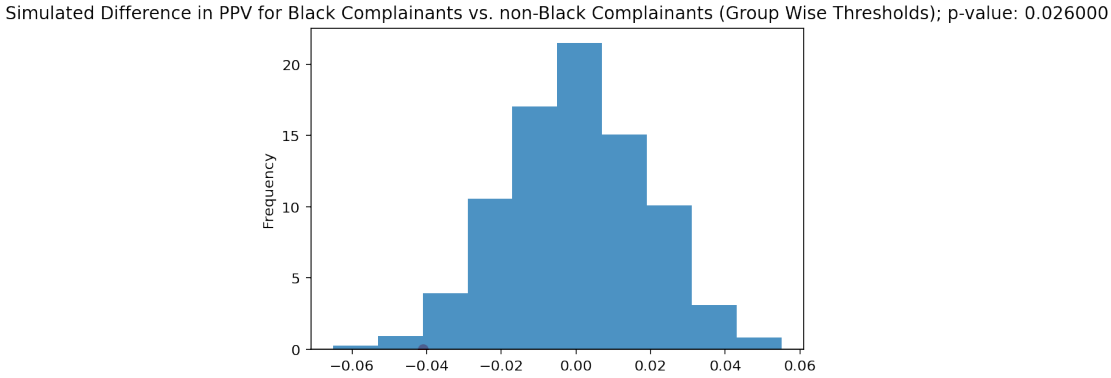




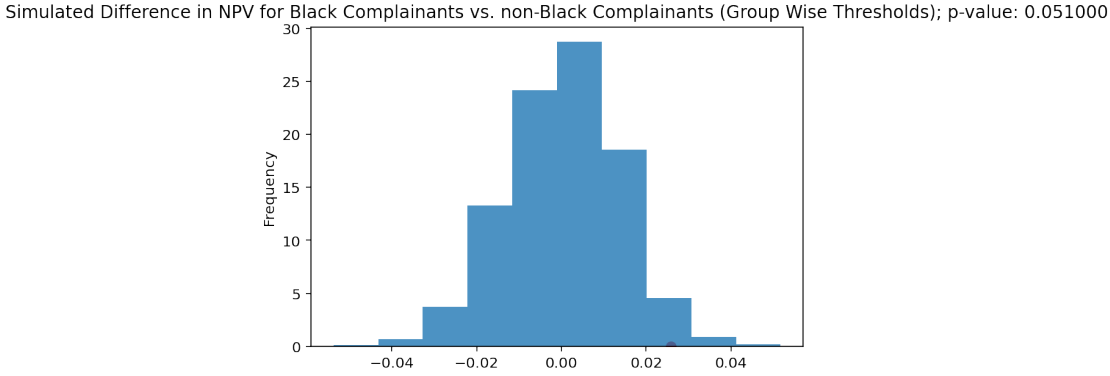
5.2.3 Predictive Value Parity

Since the group wise thresholds that maximize utility did not satisfy predictive value parity, there should be different thresholds that enforce predictive value parity. By iterating through possible thresholds and calculating the differences in PPV and NPV, the best thresholds found for holding predictive value parity were 0.522 for Black complainants and 0.535 for non-Black complainants. However, these thresholds do not hold on the test set, and the observed difference in PPV was found to be significantly different from 0 via permutation test at significance level 0.05.

```
Thresholds That Satisfy Predictive Value Parity and Yield Highest Utility

     black_threshold  nb_threshold  black_utility  nb_utility  mean_utility
157            0.522         0.535       0.625187    0.606752       0.61597

Observed Difference in Positive Predictive Value for Black complainants vs. non-
Black complainants: -0.04103196978150031
```




```
Observed Difference in Negative Predictive Value for Black complainants vs. non-
Black complainants: 0.025867291984561613
```



5.3 Thresholds Summary

Depending on what is most important to the decision-making body (CCRB), different threshold(s) would be chosen. Maximizing utility overall enforces equality for all complainants, while maximizing utility for each group is more equitable. Satisfying demographic parity ensures that substantiation doesn't depend on complainant ethnicity. Equality of odds would focus on ensuring that the chance of obtaining the benefit of substantiation is equal across groups, while predictive value parity focuses on ensuring deserving complainants receive substantiation and un-derserving complainants don't. While I believe that predictive value parity should be the priority, the CCRB may think otherwise and would thus choose the thresholds accordingly. See below for a table summarizing all the thresholds tested and what parity measure(s) they satisfy (if any).

	Threshold for Black Complainants	Threshold for non-Black Complainants	\
0	0.527	0.527	
1	0.522	0.546	
2	0.522	0.541	
3	0.522	0.535	

```
                              Test Demographic Parity      Equality of Odds  \
0              Max utility overall      Not satisfied         Not satisfied
1            Max utility per group          Satisfied             Satisfied
2         Enforce equality of odds                N/A  *Strictly* satisfied
3  Enforce predictive value parity                N/A                   N/A

  Predictive Value Parity
0               Satisfied
1           Not satisfied
2                     N/A
3           Not satisfied
```

6 Luck Egalitarian Preprocessing: Decorrelating The Features and Sensitive Attribute Complainant Ethnicity

The threshold tests performed above are an example of using post-processing to create a fairer classifier. This may not be applicable in all cases because it requires access to the sensitive attribute (complainant ethnicity) at score time. In this context, using the complainant's ethnicity when determining if their complaint should be substantiated may not be appropriate.

With preprocessing, the sensitive attribute is not needed when training or testing the model. This would mean that the CCRB would not need to know the complainant's ethnicity when making their decision. This assumption could be more applicable for the CCRB because of potential legal issues that would arise when directly using complainant ethnicity in the decision.

Luck egalitarian preprocessing seeks to "level the playing field" by decorrelating the features from the sensitive attribute. This would make the overall distribution of features the same across sensitive groups while also maintaining the relative ranking/order within groups. In Paper 1, it was discussed that Black New Yorkers more often interact with the police and that Black neighborhoods have different interactions with the police in general. This would imply that the playing field is not level across ethnicity. Preprocessing would attempt to level the playing field in the data so that the model evaluates complainants on their relative circumstances.

Example: The data shows that Black complainants are more likely to make allegations related to abuse of authority (a more severe allegation) than non-Black complainants, and non-Black complainants make more allegations related to discourtesy (a less severe allegation). With preprocessing, the data will be manipulated so that the general distribution of severity of allegations is the same across complainants while maintaining that distribution within the groups. A complaint regarding abuse of authority would send the same "signal" to the model despite ethnicity, but within groups, the same relative severity will hold.

To implement this decorrelation, I followed the process presented in *Certifying and removing disparate impact* (Feldman et al., 2015):

1. Remove the sensitive attribute (complainant ethnicity) and all other protected attributes (officer ethnicity, officer/complainant age, and officer/complainant gender).
2. Remove unordered categorical variables (borough) and convert ordered categories (allegation, type of complaint, and officer rank) to integers.

- Contact reason was dropped because most of complaints fall under "other" or simple violation, making it difficult to determine any order from the limited variance.
- Order of allegation and type of complainant was based on the NYPD *Disciplinary System Penalty Guidelines* (New York City Police Department, 2021).
- Order of officer rank was based "Duties of the NYPD ranks" (Zander, 2018).
  - Note that officer rank was treated as unordered originally, but due to lack of remaining features, it was retained for this model.

3. Scale all features so that minimum is 0 and maximum is 1.
4. Calculate the group-wise quantiles of each feature and take the median for each feature's quantiles.
5. Convert the original data so that the original value is transformed to the median of its percentile. For example, if feature x_1 for a Black complainant falls in the 20th percentile for all Black complainants, look at the 20th percentile in the median distribution (the median of the 20th percentile values for both Black and non-Black complainants).
6. Train the model on this decorrelated data.
   - The threshold that maximized utility for this new model was 0.546, compared to 0.527 in the original model.

Ideally, this decorrelation process will reduce any infra-marginality seen between the two groups while maintaining any relevant comparisons within groups. Since Black complainants in the 20th percentile of their group and non-Black complainants in the 20th percentile of their group are the same in the eyes of the model, we would assume the model's prediction to be the same for both of them, despite complainant ethnicity. This similarity across groups is what would make the model fair.

Model performance in terms of precision and recall both overall and across groups remained generally the same compared to the original model and the model trained on the decorrelated data. Since fairness and parity measures are the emphasis of this exercise, these will be looked at more closely. With making the features, and in turn the model, independent of the sensitive attribute complainant ethnicity, we assume demographic parity to hold.

```
Test Performance For All Groups:
Accuracy Score: 0.6129716623730869
Recall score: 0.5638957816377171
Precision score: 0.3293478260869565
F1 score: 0.41582799634034767

Test Performance For Black Complainants:
Accuracy Score: 0.6006221584111031
Recall score: 0.5908629441624366
Precision score: 0.31493506493506496
F1 score: 0.4108718672785034

Test Performance For Non-Black Complainants:
Accuracy Score: 0.6342975206611571
Recall score: 0.5215311004784688
Precision score: 0.35855263157894735
F1 score: 0.42495126705653014
```

6.1 Demographic Parity

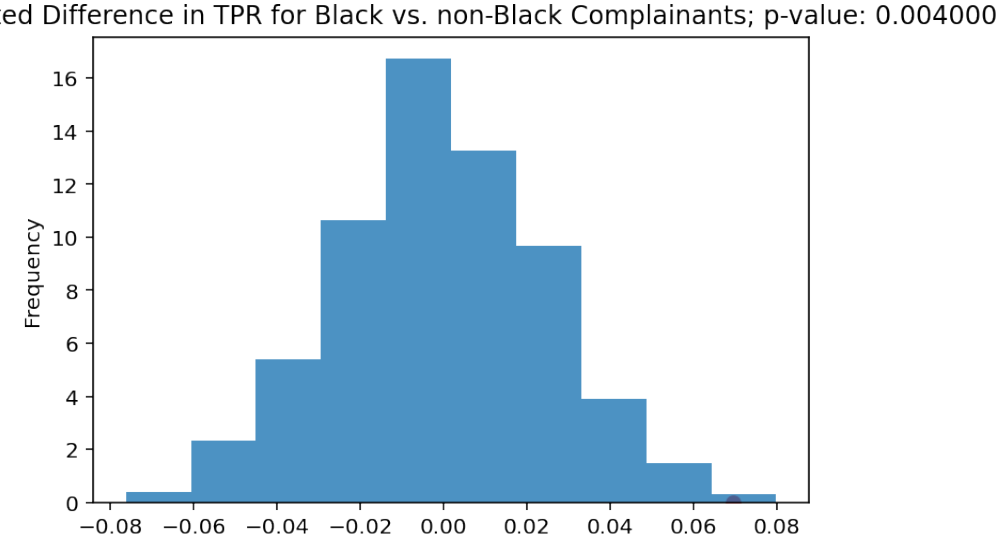
In the original model (prior to threshold tests), the model substantiated Black civilians' complaints about 6% less often than it did for non-Black complaints. Rather than moving this difference to 0, the new model switched the direction of the violation of demographic parity. Instead of a difference in substantiation rate of 0, the new model substantiated complaints from Black civilians 6.5% *more* often than it did complaints from non-Black civilians. It seems that this decorrelation may be an overcorrection of sorts when compared to the original model.

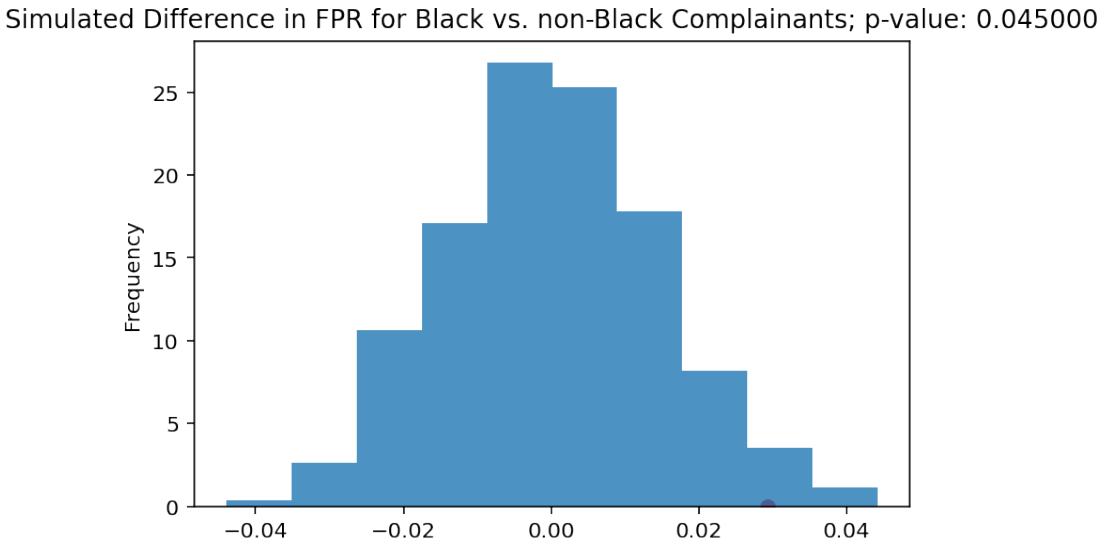
```
Model Substantiation Rates for Black Complainants vs. non-Black Complainants

 classification  Substantiated  Overall Proportion
 black
 False                0.376860            0.366722
 True                 0.442211            0.633278

Observed Difference in Model Substantiation Rate for Black vs. non-Black
Complainants: 0.06535155114415053
```

6.2 Equality of Odds

Since the model should be independent of complainant ethnicity, we could expect equality of odds to also be satisfied. However, the new model seemed to overcorrect once again. The true positive rate for Black complainants went from being 6.8% lower in the original model to 6.9% greater than that for non-Black complainants in the new model. The difference in false positive rate between Black and non-Black complainants went from -4.7% to +2.9%. At the 0.05 significance level, permutation tests revealed that both true positive and false positive parity are not met by the new model.

```
Observed Difference in True Positive Rate for Black complainants vs. non-Black
complainants: 0.06933184368396772

Observed Difference in False Positive Rate for Black complainants vs. non-Black
complainants: 0.029373116465293614
```

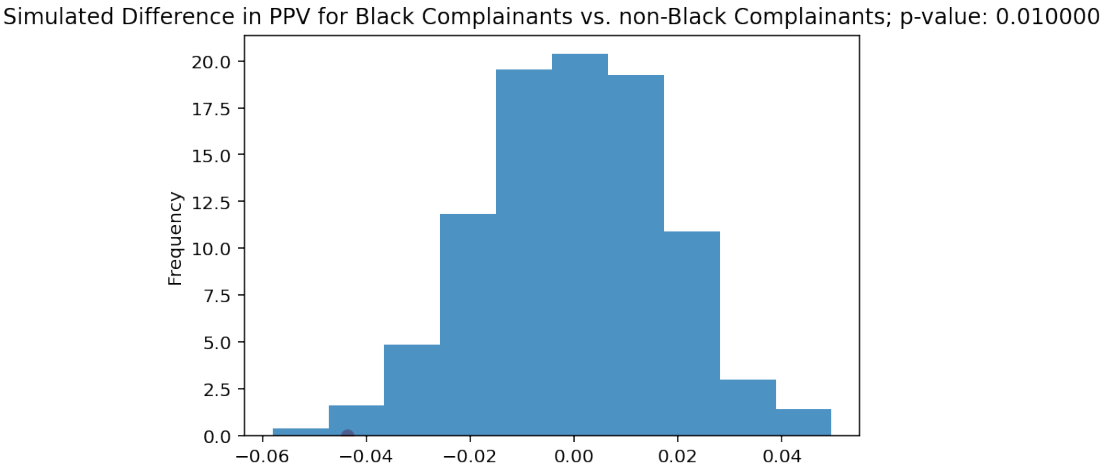




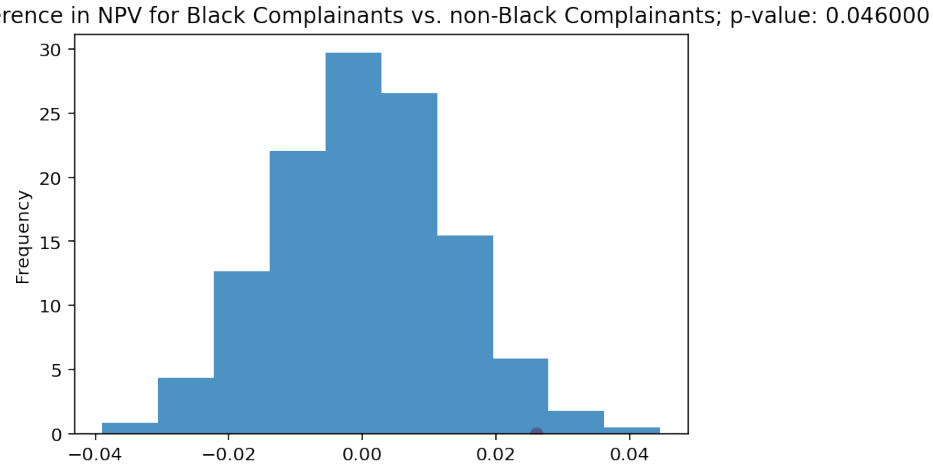
6.3 Predictive Value Parity

The new model also doesn't satisfy predictive value parity, despite the original model satisfying it. The differences in both negative and positive predictive value parity between Black and non-Black complaints slightly widened (less than a percent difference), making these differences significant from 0 when conducting a permutation test with 0.05 significance level.

`Observed Difference in Positive Predictive Value for Black complainants vs. non-Black complainants: -0.04361756664388239`



```
Observed Difference in Negative Predictive Value for Black complainants vs. non-
Black complainants: 0.02605181915526744
```



6.4 Explanation

It is clear that the decorrelation was unsuccessful and seemed to benefit Black complainants more than non-Black complainants. Below are some speculations as to why either this preprocessing is not suitable for this context or dataset or why the methodology didn't yield the expected results:

- The distributions of the remaining features (allegation, type of complaint, rank of officer, and time elapsed between opening and closing of complaint) were too similar across groups. Thus, the decorrelation process would not have changed the data enough to make the expected impact.
- Another factor could be that many of the dropped attributes, including the protected classes and the unordered categorical variables, could have high predictive power. When simply looking at the logistic regression coefficients of the original model, some of the higher coefficients are from the unordered categorical variables. These variables could have helped the model somehow differentiate between Black and non-Black complainants, leading to the original violations in parity. Since these variables are excluded in the new dataset/model, the new model doesn't fix the unfairness seen in the original model. Instead, the new model reverses the unfairness in favor of Black complainants.
- The attributes that may have put Black complainants at a disadvantage in the eyes of the model could be the same attributes that had to be excluded from the new model. Additionally, the distributions of features that were retained could have been too similar across complainant ethnicity. So, when decorrelating these retained features, Black complainants could have benefitted more.
- The features that led to an "unlevel playing field" across complainant ethnicity could have been the excluded features, so "leveling the playing field" with the retained features could have led to the overcorrection in differences in substantiation rate.
- The initial belief that the decision is dependent on complainant ethnicity could have been grossly incorrect. While demographic parity in the dataset is clearly violated, this may not be because of the dataset, but because of some unseen factors not present in the data. If this

were true, then this entire exercise of analyzing the dataset, creating a model, and auditing the model could have been useless. The information that most greatly influences the CCRB's decision making could be excluded in the dataset.

- Performing decorrelation is likely more successful with more groups and more features. Since there are only two groups to create the median distribution from, and the dataset had a limited amount of attributes appropriate for this process, the statistical degradation could have been too severe.

7 Works Cited

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2783258.2783311

Fraenkel, A. (n.d.). Fairness & Algorithmic Decision making. 8. COMPAS Analysis II - Fairness & Algorithmic Decision Making. Retrieved June 9, 2022, from https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Julia Angwin, J. L. (2016, May 23). Machine bias. ProPublica. Retrieved June 9, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Katz, D. M., II, M. J. B., & Blackman, J. (n.d.). A general approach for predicting the behavior of the Supreme Court of the United States. PLOS ONE. Retrieved June 9, 2022, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0174698

New York City Department of City Planning | Population Division. (2022, February 9). Dynamics of racial/Hispanic composition in NYC neighborhoods. ArcGIS StoryMaps. Retrieved June 9, 2022, from https://storymaps.arcgis.com/stories/46a91a58447d4024afd00771eec1dd23

New York City Police Department. (2021, January 15). Disciplinary System Penalty Guidelines. Retrieved June 9, 2022, from https://www1.nyc.gov/assets/nypd/downloads/pdf/public_information/disciplinary-system-penalty-guidelines-effective-01-15-2021-compete-.pdf

The NYPD Files. (2020, July 7). ProPublica Data Store. https://www.propublica.org/datastore/dataset/civilian-complaints-against-new-york-city-police-officers

Title 38-A: Civilian Complaint Review Board - New York City. (2018, January 1). Retrieved June 9, 2022, from https://www1.nyc.gov/assets/ccrb/downloads/pdf/about_pdf/Title38-A_20210526.pdf

Zander, J. A. (2018, June 29). Duties of the NYPD ranks. Work. Retrieved June 9, 2022, from https://work.chron.com/duties-nypd-ranks-21851.html