NYPD Stop And Frisk:

Nutritional Labels for Automated Decision Systems

Simran Arora(sa5476)
Yash Jajoo(yj1499)

DS-GA 1017
Responsible Data Science
Spring 2021

NEW YORK UNIVERSITY

Contents

1 NYPD Stop and Frisk: Nutritional Labels for Automated Decision Systems

1.1 Background

NYPD's "stop, question, and frisk" program is a controversial policy that allows police officers to, solely on suspicion, stop, question and search people for a weapon. It has been highly criticized for being a form of racial profiling. Consider the following: In 2011, 685,724 NYPD stops were recorded [1].

- 605,328 were innocent (88 percent).
- 350,743 were Black (53 percent).
- 223,740 were Latinx (34 percent).
- 61,805 were white (9 percent).
- 341,581 were aged 14-24 (51 percent).

More than half of those who were stopped and frisked were African-Americans while less than 10% were Caucasians.

This ADS aims to predict whether someone will be frisked (output feature: frisked) by the New York Police Department based on the circumstances of the stop and the features of the suspect. The ADS used the NYPD Stop And Frisk Data[2] which contains records of this policy dating from 2003 to 2019. It includes a record for each stop made in NYC, with features such as the time of day, the location of the

[1] NYCLU ACLU of New York. "Annual Stop-and-Frisk Numbers." In: *nyclu.org* ()

[2] NYPD. "NYPD Stop, Question and Frisk Data." In: *New York City Police Department* ()

stop, officer's details, the suspect's information such as race, sex, height, weight, etc. However, the ADS only uses 2011's data since the highest number of stops were recorded in said year.

1.2 Input and Output

NYPD Stop And Frisk Data[3]

- The initial dataset shape is (685724, 112) which is reduced to (685724, 29) before starting the preprocessing stage. Following are the 29 features and their corresponding data types:

Table 1.1: Features and Corresponding Data types

Features	Data Type
sex	String
race	String
age	Integer
height	Integer
weight	Integer
haircolr	String
eyecolor	String
build	String
city	String
pct	Integer
timestop	Integer
inout	String
trhsloc	String
typeofid	String
othpers	String
explnstp	String
offunif	String
officrid	String
offverb	String
offshld	String

[3]NYPD, "NYPD Stop, Question and Frisk Data"

Continuation of Table 1.1	
Features	**Data Type**
ac_rept	String
ac_proxm	String
ac_evasv	String
ac_assoc	String
ac_cgdir	String
ac_incid	String
ac_time	String
ac_stsnd	String
frisked (output feature)	String

- The shape of the final dataset, that is, after preprocessing, feature selection and down-sampling is (100000,11). Final feature set includes: timestop, age, weight, height, offunif, ac_incid, ac_time, others, ac_proxm, ac_evasv, frisked (output variable).

- A selection of said graphs from the ADS are shown below:

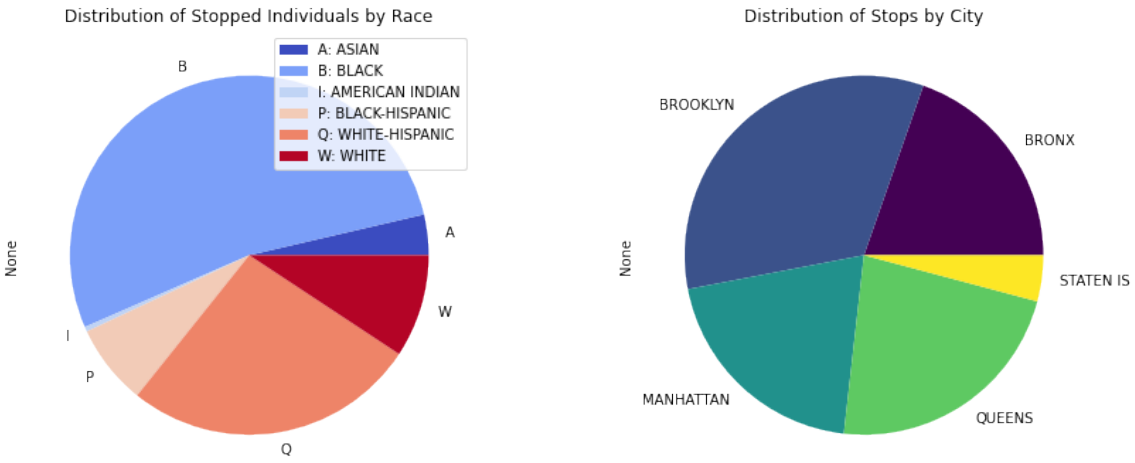

Figure 1.1: Distribution of Stopped Individuals by Race and by City

From Figures 1.1 and 1.2, we can see that in the dataset, there are many more records of non-white people who were frisked than their white coun-

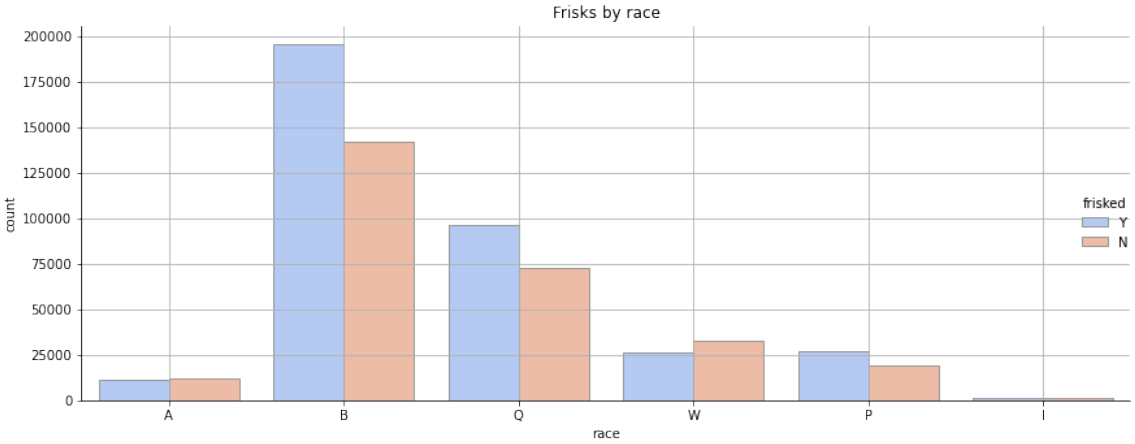

Figure 1.2: Distribution of Frisks by Race

terparts. This indicates that the non-Caucasians might be over-represented in the dataset.

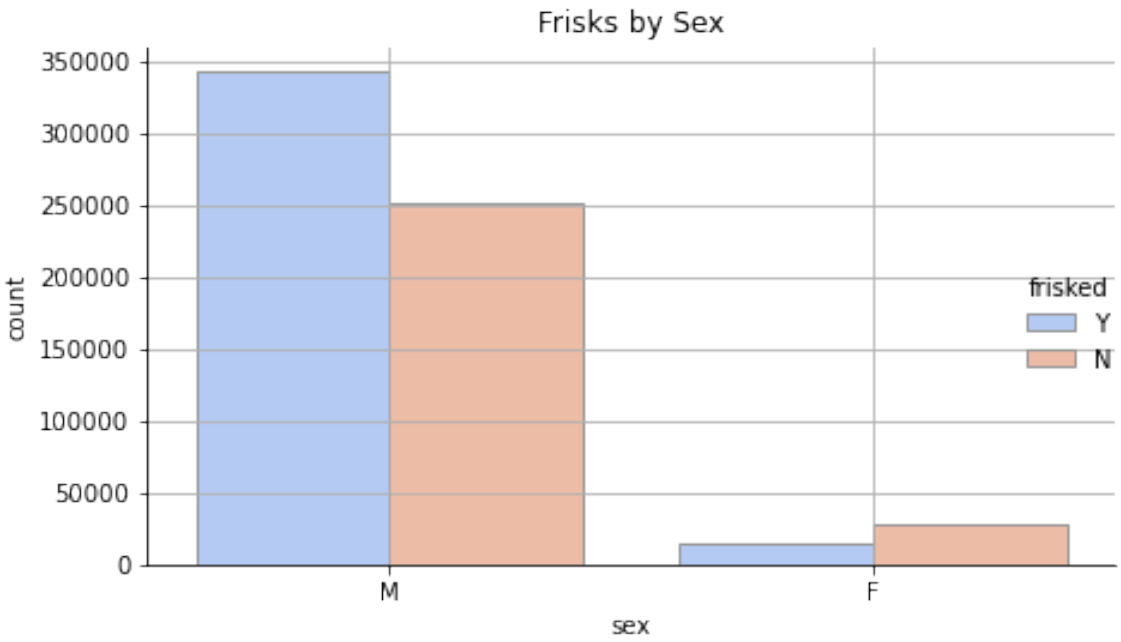

Figure 1.3: Distribution of Frisks by Sex

- Furthermore, the dataset also includes features that might as proxies for race, for ex, “city”, since a vast majority of African-Americans in NYC reside in Brooklyn as well as Bronx.
- Each of the models used in the ADS, i.e., SVM, Naive-Bayes and KNN is evaluated based on accuracy. Furthermore, **the ADS does not perform inference**, that is, it does not use any of the trained models for classification on unseen data. However, with some additions, the classifier can easily be used to predict the classification label based on the probability computed for each class.

1.3 Implementation and Validation

1. **Preprocessing and feature engineering:**

   (a) Proceeding with data preprocessing, out of the above mentioned features, officrid, offverb, offshld and city have missing data.

Feature	Missing Value
officrid	677711
offverb	548624
offshld	512650
city	38

Table 1.2: Feature and Missing Value

   The first 3 features were dropped since there were too many missing values. As for “city”, the values were filled using another feature “pct” (Data[4]) by checking which precinct number corresponds to which borough and afterwards, pct was dropped.

   (b) For simplicity, the ADS combines “ht_feet” and “ht_inches” freatures together into a single feature “height”.

   (c) Unknown and inconsistent values were handled as follows:

      i. **haircolr** - Rows with values “XX” (unknown) and “ZZ” (other) are dropped and the value “RA” was changed to “RD” for red.

      ii. **sex** – Rows with value “Z” (unknown) are dropped.

[4]New York City Police Department. “Precinct Numbers.” In: *nyc.gov* ()

iii. **race** - Rows with values “U” (unknown) and “Z” (other) are dropped.

iv. **eyecolor** - Rows with values “XX” (unknown) and “ZZ” (other) are dropped. Values “MC” and “P” was changed to “MA” for maroon and “PK” for pink respectively.

v. **build** - Rows with value “Z” (unknown) are dropped. Here “U” means muscular, hence rows with said value were retained.

vi. **typeofid** - Rows with value “O” (other) are dropped.

(d) The ADS checks for erroneous data for features age, height as well as weight by computing the upper and lower bounds, then

i. **age** – Since average lifespan of an American is 76 years and infants are not likely to be frisked, **values above 100 and below 5 are mostly erroneous and therefore removed**.

ii. height – Since the tallest person in America at the time was 92 inches tall and any value less than 40 inches is likely erroneous, **values above 90 and below 40 were removed.**

iii. weight – As an average American weighs between 159 to 191 lbs, **values below 50 and above 300 were removed.**

(e) Binarization and One-hot encoding was performed as follows:

i. For **binary** features -

A. sex – M: 1 and F: 0

B. inout – O: 0 and I: 1

All features with values Y/N: Y: 1 and N: 0

ii. For **non-binary** features ’race’, ’haircolr’, ’eyecolor’, ’build’, ’city’, ’trhsloc’ and ’typeofid’, one-hot encoding is performed using pandas get dummies function.

After all of the above, the dataset shape turns out to be (638447, 58).

2. **Feature selection:** This is performed using Random Forest Classifier which is fit on the preprocessed dataset and the ADS takes the top 10 most important features (based on feature importance) into the final feature set.

Furthermore, the ADS also performs data sampling and uses only 100000 rows [dataset’s shape: (100000,11)] as opposed to post one-hot encoding (600000+, 58). This is done by code authors primarily to ensure that the models don’t take “too much time”.

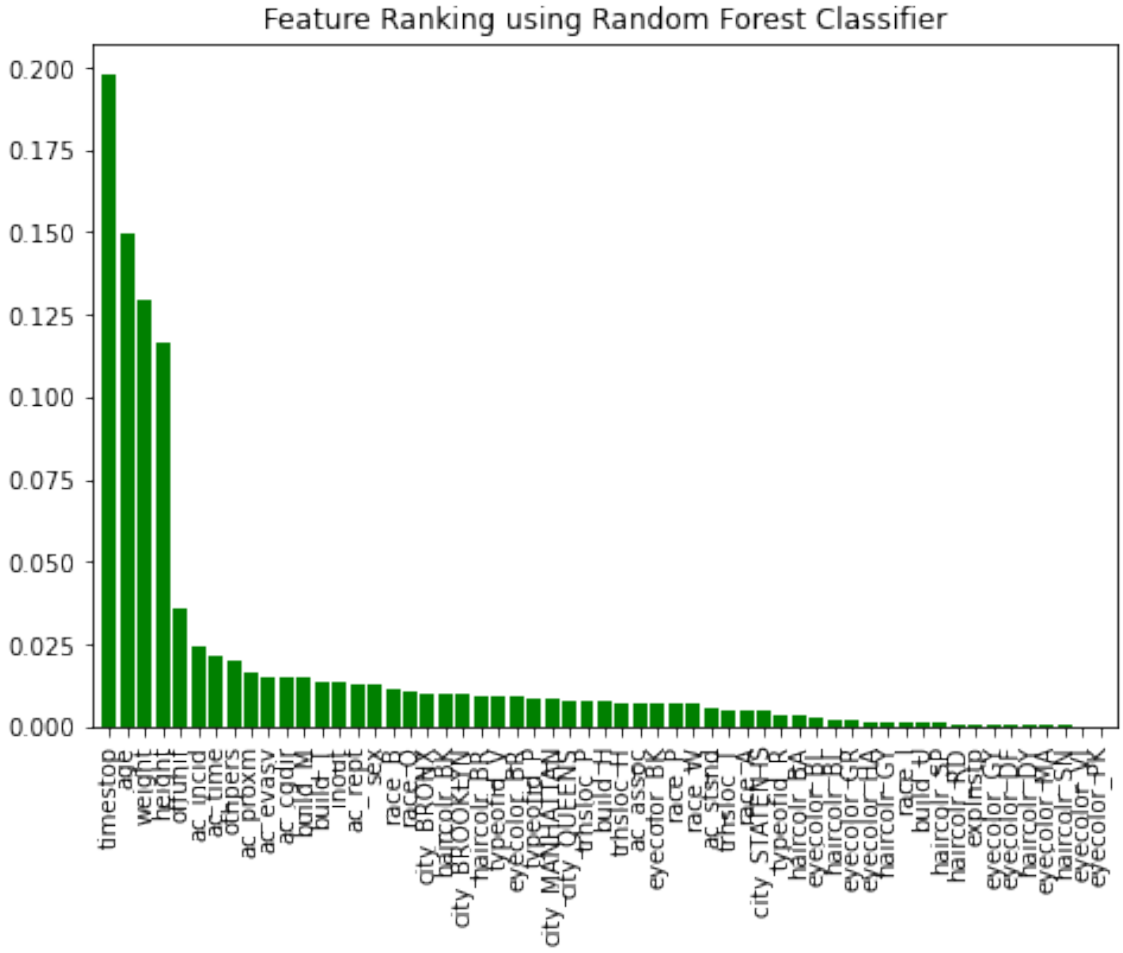

Figure 1.4: Feature Selection

3. **Model training and testing:**

   (a) First model used is a Support Vector Machine which after 5-fold cross validation and hyperparameter tuning using grid search, gives 61% accuracy with tuned hyperparameter kernel: Gaussian radial basis function.

   (b) Next is a Naïve-Bayes classifier for which the same approach is followed and the accuracy reported is 60%.

   (c) Lastly for KNN, the accuracy is 60% with best hyperparameters 29 neighbors, uniform weights and 9 pca components

1.4 Outcomes

Consider the ADS' methodology for feature selection again. We believe that the authors' approach is insufficient in that it fails to account for several factors that an officer might actually consider while deciding whether to frisk a suspect or not. Also, consider the following correlation matrix,

From the plot of correlation heat map(Figure 1.5), features such as haircolr_BK (hair colour = Black) seem to have a relatively greater correlation (>=0.25) with frisked_Y (frisked = Yes).

Notice that such features were actually not included in the final feature set. Thus given the nature of our project, we decided to retain all 58 features. Also, we think that the sample size of only 100000 rows is too less and hence, we also used all rows from the preprocessed dataset for training, testing and validation. Moreover since we wished to evaluate the model's fairness with respect to race which is a multi-value (non-binary) feature, we created a new feature "binary_race" which has values 1 for White and 0 for all other races. Thus, the shape of the dataset we used was (638447, 59).

Notice that none of these features were actually included in the final feature set. Thus given the nature of our project, **we decided to retain all 57 features**. Also, we think that the sample size of only 100000 rows is too less and hence, **we also used all rows from the preprocessed dataset** for training, testing and validation. Moreover since we wished to evaluate the model's fairness with respect to race which is a multi-value (non-binary) feature, **we created a new feature "binary_race"** which has values 1 for White and 0 for all other races. Thus, the shape of the dataset we used was **(638447, 59)**.

The training process using the entire preprocessed dataset was taking a lot of time and so, we decided to focus on one out of the three models that the ADS uses, the hyperparameter tuned kNN. We decided to focus on disparate impact as a measure of fairness. The reason for that is since we are evaluating the model's fairness with respect to a race, i.e., a group of people, we are actually dealing with group fairness and disparate impact solves for group fairness.

Now we evaluated the model's fairness with respect to race; baseline fairness and accuracy metrics are as follows:

1. Accuracy = 63.8%

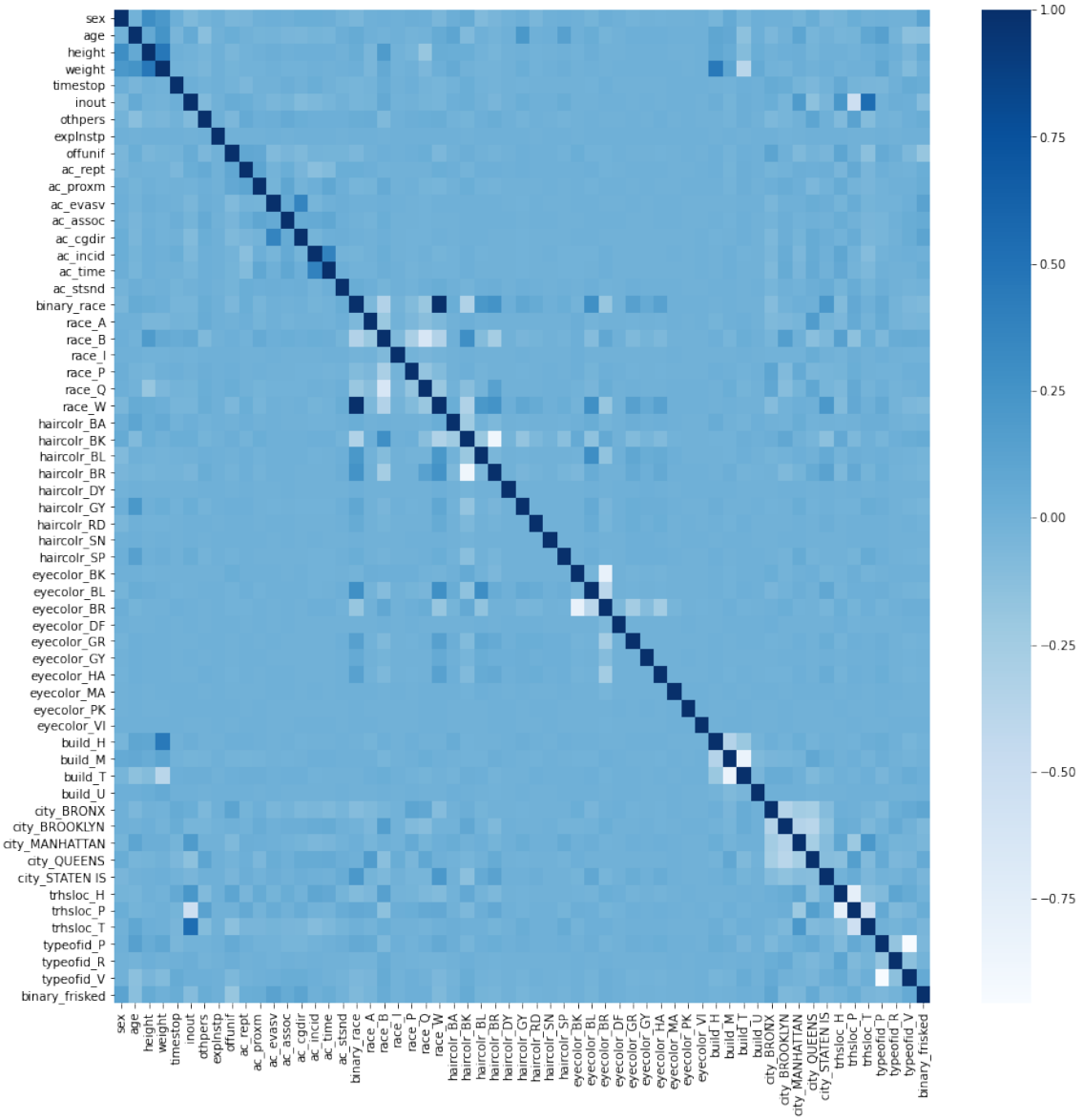

Figure 1.5: Correlation Heat Map

2. Accuracy for privileged group (White: binary_race = 1) = 0.64
3. Accuracy for unprivileged group (Non-White: binary_race = 0) = 0.63
4. **Disparate Impact = 0.58**

Clearly, the baseline model isn't very fair.

According to AIF 360, following are the ten state-of-the-art bias mitigation algorithms(Figure 1.6) that can address bias throughout AI systems

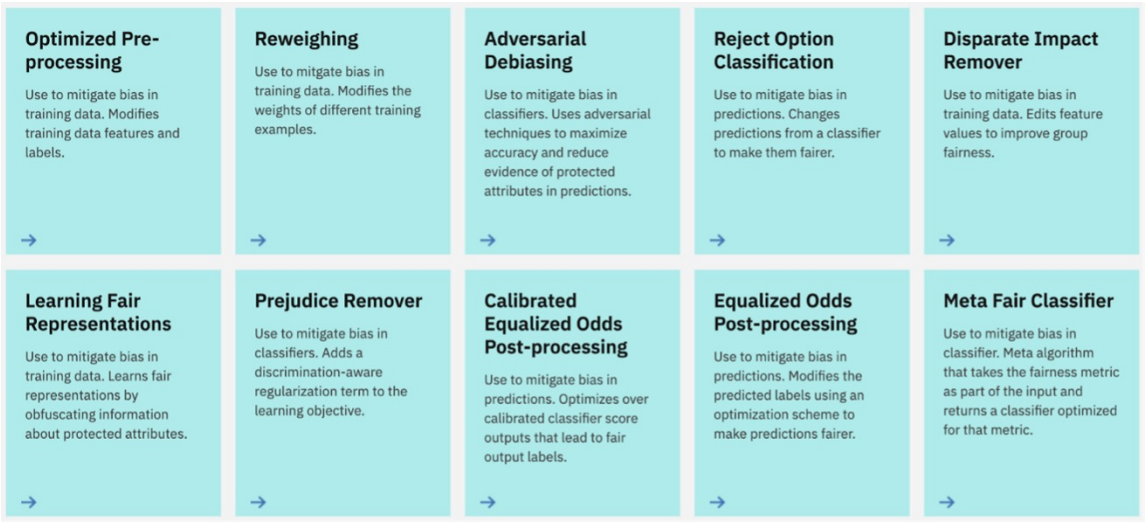

Figure 1.6: Bias Mitigation Algorithms

We decided to use two methodologies: **Disparate Impact Remover (DIR) and Reject Option Classification (ROC)**. The reason for choosing these two strategies was to apply bias removal directly on data (pre-processing) as well as on predictions (post-processing).

Disparate Impact Remover (DIR): DIR is a preprocessing technique that edits feature values in order to hide which group a given data point belongs to and thereby increase fairness. The distributions of the groups involved are made to overlap and the extent of said overlap is governed by level" where a repair level of 1 indicates complete overlap and 0 indicates no overlap. The 5 repair levels we considered are 0.2, 0.4, 0.6, 0.8 and 1. The expected behaviour is that with increasing values of repair level, from 0 to 1, disparate impact improves (that is moves closer towards 1). Also oftentimes with an increase in fairness, accuracy suffers.

Reject Option Classification (ROC): ROC is based on the assumption that most discrimination occur when models are least certain around their classifica-

tion thresholds. For instance, consider a classification threshold of 0.6 and if the 25 prediction is 0.1 or 0.91, the model is certain of its prediction but it is not for 0.59 or 0.6. The algorithm gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups. Thus, Bias can be reduced by taking advantage of a classifier's low confidence region and rejecting its predictions. Equal opportunity difference was set as the optimization metric. Minimum classification threshold was set to 0.01 and maximum classification threshold was set to 0.99 with 100 classification thresholds between them for optimization search.

Metrics	**Baseline**	**DIR**					**ROC**
		Repair Lvl 0.2	Repair Lvl 0.4	Repair Lvl 0.6	Repair Lvl 0.8	Repair Lvl 1	
Accuracy	**0.6383**	0.5582	0.5579	0.5603	0.5599	0.5600	**0.4681**
Accuracy privileged class	0.6440	0.6610	0.6604	0.6681	0.6673	0.6715	0.5514
Accuracy unprivileged class	0.6377	0.6522	0.6560	0.6654	0.6638	0.6643	0.4595
Disparate Impact	**0.5850**	0.5951	0.5907	0.6314	0.6343	0.6365	**1.0091**

Table 1.3: Tabulating the results (taking seed value 32 for DIR and ROC)

Disparate impact is least in baseline's case while it is the greatest after performing ROC. The opposite is the case for accuracy; **baseline model's accuracy is the highest while that of ROC is the lowest**.

DIR wasn't successful in making the model more fair in any significant manner and thus, it seems like post-processing bias mitigation is more effective for this dataset and model.

Analyzing using SHAP The features shown in the plot contribute the most towards the classification. Thus, we can see that city-BROOKLYN, city-BRONX, race-B are contributing heavily and surprisingly, the feature "timestop" that the ADS' authors deemed to be most important affects the model's prediction far less than other features.

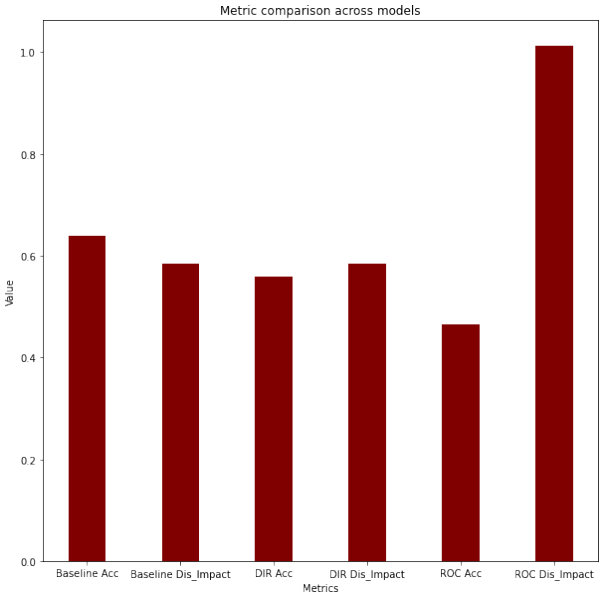

Figure 1.7: Metric Comparison Across Models

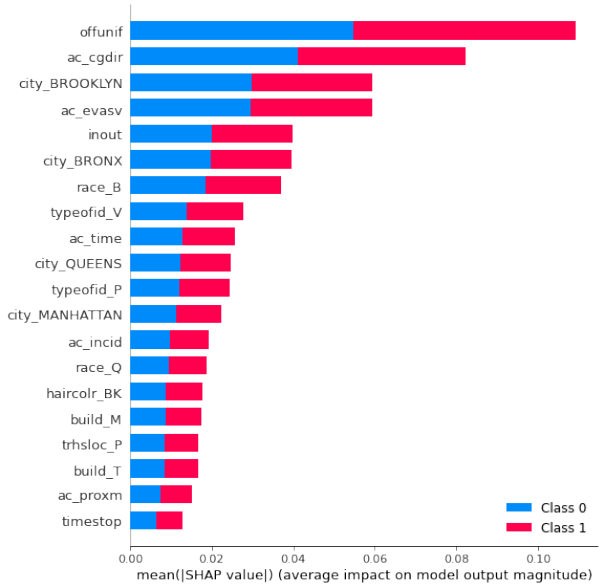

Figure 1.8: SHAP Summary Plot

Considering a record from the test dataset where the SHAP correctly classified

the outcome as Frisked where the true label was also Frisked. Here, we can see that the suspect is black(race_B=1)(Figure 1.9).

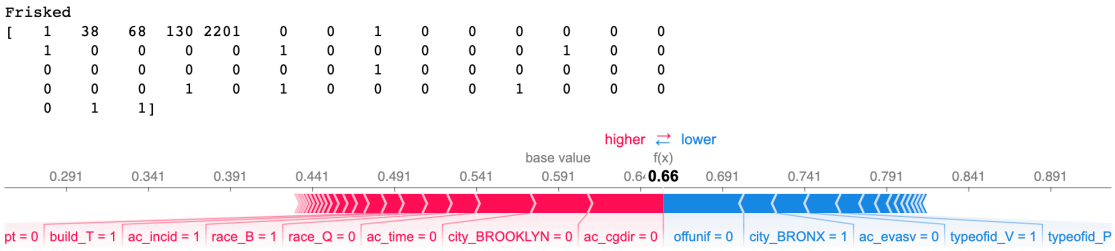

Figure 1.9: Classified Frisked

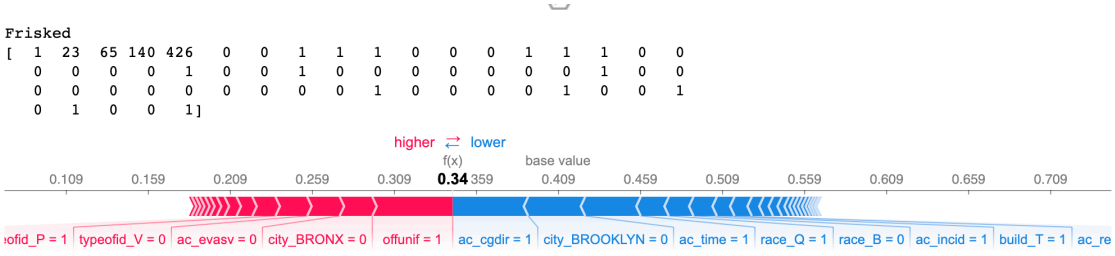

Figure 1.10: Classified Not Frisked

Now consider a record where the SHAP incorrectly classified the outcome Not Frisked where as the true label was Frisked. Here, we can see that the suspect is not black(race_B=0) and hence the SHAP classified that the suspect wasn't frisked(Figure 1.10).

Hence, there is clear indication that bias has indeed been embedded into the model.

1.5 Summary

Overall, we have serious doubts as to whether the data is appropriate for the ADS. The dataset, as mentioned previously, is **over-represented by non-Caucasians with frisked = Y** and hence the predictions shall be biased as well. Some features

like city, as stated earlier, may act as proxies as well. Building on that, we can see from the from the correlation matrix that the features city-BRONX and city-QUEENS (both are boroughs where predominantly non-Caucasians reside) indeed have a comparatively higher correlation with frisked=Y. Using this ADS' predictions for decision making will therefore produce a potentially dangerous positive feedback loop which reinforces the NYPD's' notion that certain races are more prone to commit crimes, a classic case of **confirmation bias**.

The original implementation of the ADS was lacking in that they didn't use enough of the dataset and also left out several features while training their models, as we know, exclusion of features alone doesn't help in bias mitigation and hurts accuracy as well. The changes that we made, i.e., using the entire dataset with 58 features instead of 10 lead to a **slightly higher accuracy but the model was highly biased (disparate impact = 0.585)**. Even with disparate impact removal, there wasn't any significant improvement in disparate impact. In case of reject option classification, even though disparate impact greatly improved (disparate impact = 1.009), the accuracy fell drastically. A higher disparate impact would be beneficial to the citizens, particularly non-Caucasians, as they will be treated fairly. It will also be beneficial to the NYPD since it will enable them to be non-discriminatory and thus protect their reputation from harm. An higher accuracy will obviously help the NYPD but also the citizens as they will then be able to enjoy better security. Thus, **optimizing this ADS' accuracy and disparate impact is in the best interests of both the citizens and the NYPD**.

There are several improvements which could be made to the ADS pipeline to improve it. **The problem of over-representation can be dealt by generating a synthetic dataset** that is more representative of the ground truth determined by using other data sources such as surveys. A **differentially private synthetic dataset can also reduce the impact of proxy features**. Bias mitigation strategies such as the one performed here can then be used to mitigate bias that may still exist post data generation.

NYPD possesses a great deal more information than the ADS has access to, including previous criminal records. In this context, the ADS does not appear to offer any tangible benefit to stakeholders; we conclude that **the project is not viable as an ADS**. However, it is possible that the project could be converted from a decision system to a tool for studying structural bias in law enforcement.

Bibliography

Department, New York City Police.
“Precinct Numbers.”
In: *nyc.gov* ().
New York, NYCLU ACLU of.
“Annual Stop-and-Frisk Numbers.”
In: *nyclu.org* ().
NYPD.
“NYPD Stop, Question and Frisk Data.”
In: *New York City Police Department* ().

NEW YORK UNIVERSITY