Group Monarchy: Maya Blumovitz, Rundi Zhou, Eddie Woo, Di Mei, Austin Porras

CPD Project - Final Report

Introduction

In recent years, the relationship between crime and race has been increasingly scrutinized by the general public. Police departments across the U.S. face criticism, accusing departments of discriminatory practices towards people of color. These criticisms come after decades of police brutality incidents, wrongful convictions, and a disproportionately large number of Black Americans in the prison system. Even though Black people represent only 13% of the US population, 38% of the US prison population is Black.[1] In response to these recent events, our group wanted to study race and crime data in the Chicago area and discover whether there is a statistically significant relationship between these two data points. In addition, we wanted to check if people of color are disproportionately targeted in arrests and felony charges compared to other races. We were also interested in discovering general trends in the CPD database, such as checking how the number of arrests changed over time and during what periods of time there are more arrests. This was our motivation to look into the Chicago Police Department's database.

Dataset

The dataset we chose to analyze for our final project was the Chicago Police Department's arrest dataset.[2] At the time when we downloaded the final dataset for our project, the file contained records from Jan 1st, 2014 to September 30th, 2022. This dataset had over 500,000 rows of data with 24 columns per row. This data includes only adult arrests, and excludes expunged records. The data includes the arrestee's race, case number, arrest date, and up to four charges and their associated information. The fields we are most interested in (race and charge-types) are all stored as plain-text except the arrest-date which is stored in a date-time format. We chose to convert these to boolean dummies using pandas.

To supplement the initial dataset, we also used the Chicago Police Department crimes dataset.[3] This dataset has over seven millions rows of data from January 1st, 2001 until September 30th, 2022 when we downloaded the data. This dataset contained valuable information relevant to our research questions such as the location of arrests and the ward in

[1] https://tinyurl.com/prisonpolicyresearch
[2] https://data.cityofchicago.org/Public-Safety/Arrests/dpt3-jri9
[3] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

which they occurred. The ward is stored as a number while the district is stored as plain text. The district was implicitly converted to a number during analysis.

Due to CPU limitations, our ability to join the data from such a massive dataset was limited. We searched through 720,000 entries in the crimes dataset and matched the data from those entries to rows in the arrest dataset using the case number. This is the suggested way of linking the datasets per the documentation associated with the datasets. We were able to find 106,042 rows of data that had case numbers matching our arrest data. This was our final dataset used for analysis after cleaning, which we then split randomly into eighty percent train data and twenty percent test data.

One clear limitation of our data gathering process is the reduction from half a million potential rows to less than a hundred thousand. Given unlimited time and better computers, we would have preferred to search the entire crimes dataset and match as many rows as possible to better train and test our models. However, given that our final notebook took several hours to run, this was a necessary choice.

Cleaning

The final dataset we used had 106,042 rows, which we assumed would be enough to get statistically significant results. This final dataset had information from two different ones. Our arrest dataset had twenty four columns, and we analyzed each one individually to determine what kind of data cleaning they require. Before cleaning the dataset, we had 580,663 rows of data, and after cleaning we ended up with 571,146 rows of data. We describe the process here:

1. CB_No
   This is a unique ID (numerical) for each row of data, and since we didn't find any missing values, we didn't have to do anything for this column.
2. Case number
   The case number appears (as a string) in rows where there is a case associated with the arrest - this number corresponds to case numbers in other datasets. We didn't find any errors or bad values in this column other than missing values, which we didn't use.
3. Arrest date
   The arrest date data was a floating timestamp of the form DD/MM/YYYY HH:MM:SS, but we wanted to be able to run some statistics on different periods during the year, days of the week, and time of day. Thus, we decided to separate this

column into four different columns: day, month, and year as numbers, and time as a datetime. There were no empty values.

4. Race
   We found no blank values or errors in this column, so we left it untouched. Values were categorical.
5. Charge 1 statute
   We found no blank values or errors in this column, so we left it untouched. Values were strings.
6. Charge 1 description
   We found no blank values or errors in this column, so we left it untouched. Values were strings.
7. Charge 1 type
   The charge type is broken down into three categories (strings), where "Other" was just left as empty. We decided to turn all the empty values into "O" (for Other) so we could use them to run analyses. We had to assume, since the CPD states that empty values are "Other", that there was no missing data.
8. Charge 1 class
   We didn't find errors in this column but there were 4717 empty values, and since it was a very small fraction of our total data, we decided to remove those rows. Values were strings.
9. Charge 2 statute
   Columns 9-20 only contained data if the arrests included more than one charge. Hence, empty values don't signify missing data but rather that the column wasn't relevant. Therefore, we decided to leave them as is (with no errors) and only use them to make analyses that are based on how many charges there were per arrest. Value types correspond to charge_1 columns of the same name.
10. Charge 2 description
    See no. 9.
11. Charge 2 type
12. See no. 9.
13. Charge 2 class
    See no. 9.
14. Charge 3 statute
    See no. 9.
15. Charge 3 description
    See no. 9.
16. Charge 3 type
    See no. 9.

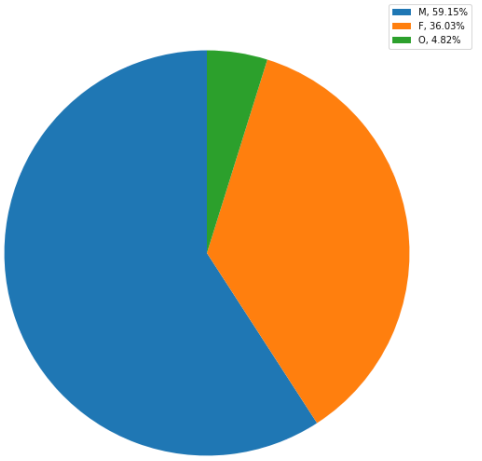
17. Charge 3 class
    See no. 9.
18. Charge 4 statute
    See no. 9.
19. Charge 4 description
    See no. 9.
20. Charge 4 type
    See no. 9.
21. Charge 4 class
    See no. 9.
22. Charges statute
    Columns 21-24 contained summaries (as strings) of the data provided in columns 5-21. We decided that we most likely will not need the same information twice, and will delete these columns.
23. Charges description
    See no. 21.
24. Charges type
    See no. 21.
25. Charges class
    See no. 21.

We also used the CPD's crimes dataset. The original dataset had 7,686,323 rows, but after dropping the rows with missing data, it shrunk down to 7,071,425 rows. Since this dataset had such massive amounts of data, and simply loading the entire amount could take at least a couple hours, we limited our API call to 720,000 rows (80,000 per year). We only needed the following three columns:

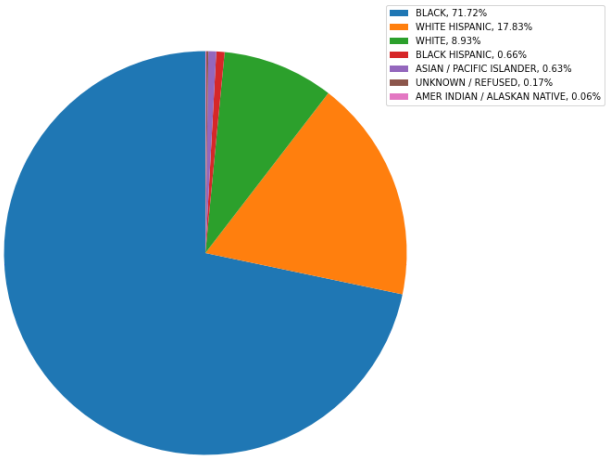
1. Case number
   The case number appears (as a string) in most data where there is a case assigned to the crime. We didn't find any errors or bad values in this column other than missing values, which we couldn't use since our merge was based on case number.
2. Ward
   The ward column has numerical data, and we didn't find any errors or bad values in this column. We did, however, encounter some missing values, so we removed those rows.
3. District
   The ward column has string data, and we didn't find any errors or bad values in this column. We did, however, encounter some missing values, so we removed those rows.

EDA

For the EDA, we first checked the percentage of different charge 1 types in the dataset (figure 1). We found there was an imbalance in that type M and F charges have significantly larger quantities than type O. The number of arrests for each race and its distribution with different charge 1 types (F, M, O) are shown in figures 2 and 3.

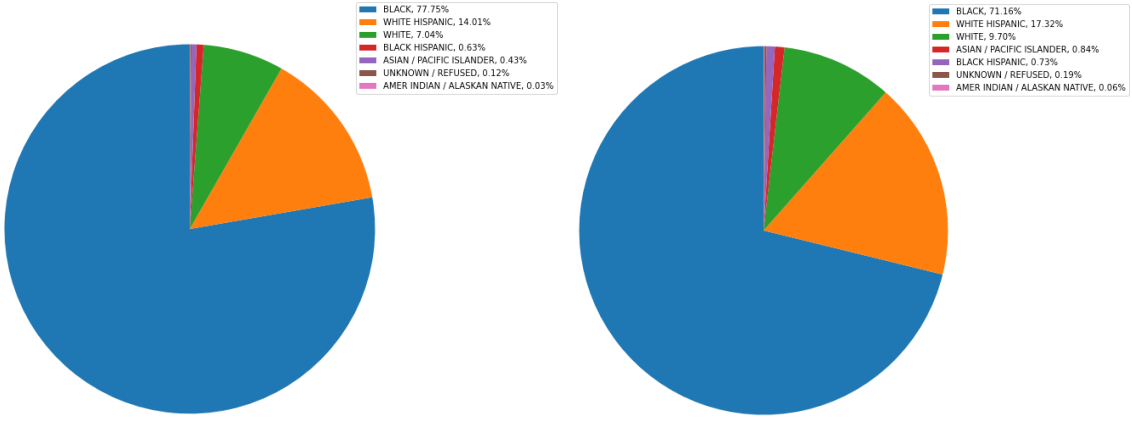


fig(1): Pie chart of charge 1 type

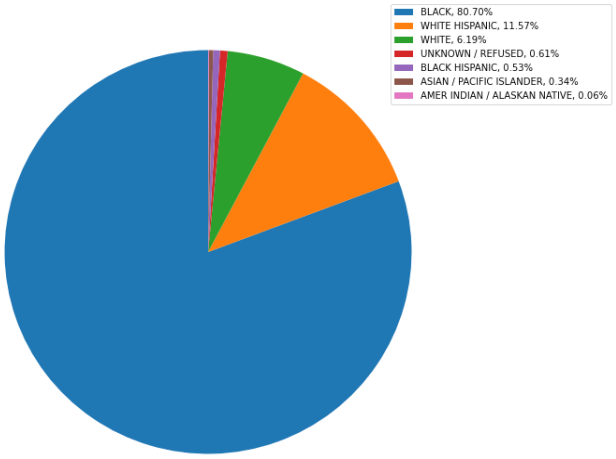


fig(2): Pie chart of the number of arrests per race

In figures 2 and 3, we saw that the racial distribution for different charge types doesn't vary significantly, but there was a concerning imbalance of races. However, this was a reflection of the reality of arrests in Chicago, so skewing the data to get better accuracy would have given us false models that don't reflect the true nature of arrests.

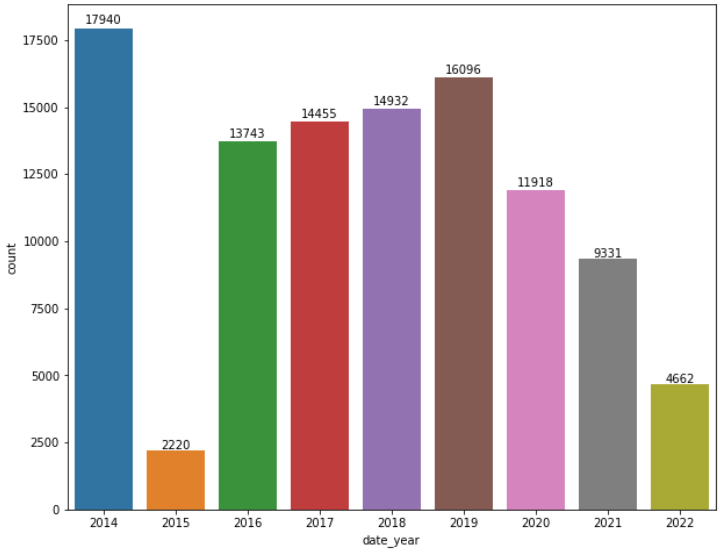


fig(3): Pie chart of the number of arrests per charge 1 type per race (F left, M right)

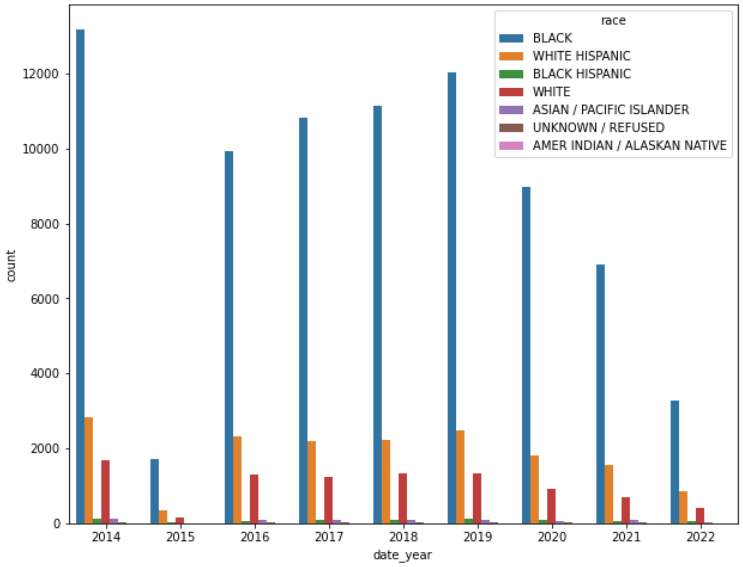


fig(3): cont. Pie chart of the number of arrests per charge 1 type per race (O)

We created plots of the number of arrests per year and per race (figures 4 and 5), from which we concluded that there was a significant decrease in the total number of arrests starting in 2020 but the distribution of races did not change noticeably. We also saw a major drop in arrests during 2015, but since we were taking 80,000 rows from each year in the crime database and merging the two databases on case number, the likeliest explanation is that there were a low amount of cases that year.

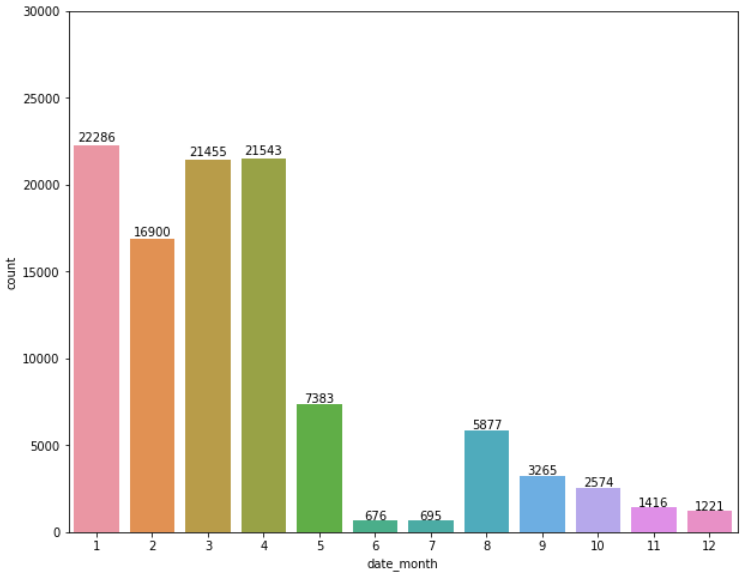


fig(4): The number of arrests per year



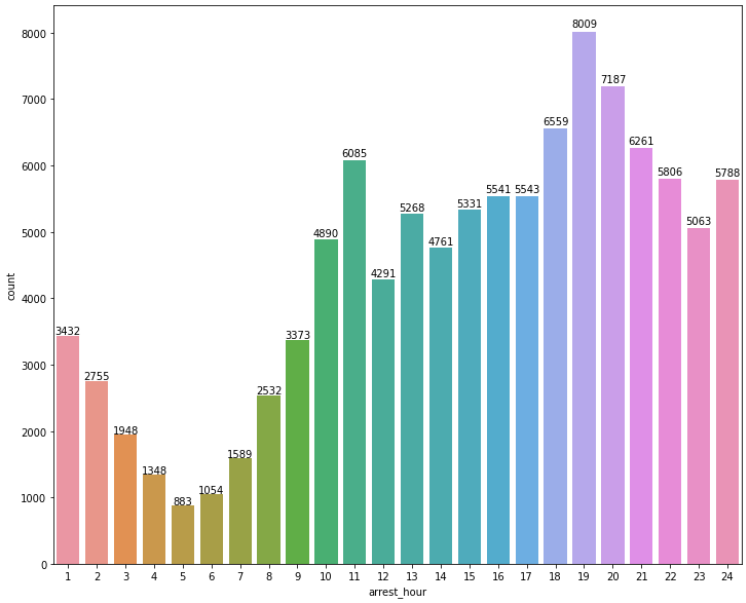
fig(5): The number of arrests per arrest type per year

Figures 6 and 7 are the plots of the total number of arrests per month and per hour of the day. Figure 6 shows a clear trend where the beginning of each year has significantly more arrests than the end. However, after checking the same statistic for each year separately, we believe that this is a distortion in the data associated with the way the crimes API works. When we first ran the API, we realized that we were getting results only from 2020 onwards, so to fix the issue we had to make separate calls for each year and merge the results. That, however, generated results that included more information about January-April. It was an imbalance that we preferred over having data from only 3 years.



fig(6): Total number of arrests per month

Figure 7 shows clearly that there are fewer arrests between 2am to 7am during the day, which we saw consistently throughout our EDA process.



fig(7): Total number of arrests per hour of a day

Machine Learning

In addition to the EDA, we tried to explore the CPD arrests dataset in a predictive manner. Namely, we used machine learning models to predict a crime case's charge-I type given a set of selected features including the criminal's race (turned into one-hot dummies), crime's time (day, year, etc) and location-related information which included ward and district numbers. There are 11 features in total and for our prediction target, the charge type is multi-class, involving felony, misdemeanor and others (which represent local ordinance, traffic arrests, etc). We joined multiple tables together to obtain the final dataset for this prediction and it contains 106,042 data points, which should be more than enough for us to conduct a persuasive machine learning analysis. Since this is a multi-class classification problem, we used the percentage accuracy as our scoring metric for the evaluation of machine learning models. Another scoring metric we tried is F1 score, which gives a more comprehensive evaluation of the machine learning models. This metric was computed for evaluating the prediction upon every single class first, and then we averaged them to obtain a "macro" F1 over multiple classes. We also looked at cross validation scoring for each model.

We developed different machine learning models to do the prediction of charge types and the training-testing set ratio is 4:1. The first model we tried is k's nearest neighbor (KNN), which is a commonly used classification technique. We assumed that crime cases with similar

charge types should be close to each other in LP space, and for each testing case, we computed the correlation between it and every crime case in the training set. Another model we used is a deep net with the multi-layer perceptron (MLP) architecture. Though it comes with very high time complexity, we wanted to observe whether our constructed feature vector is structurally meaningful (such as texts and images) or not. If so, its performance can be surprisingly impressive. In addition, we applied tree-based models like the random forest and gradient boosting (Histogram Gradient Descent based on LightGBM). Tree-based models can be readily interpreted and well compatible with both numerical and categorical data. Based on the built-in decision tree model provided by sklearn, we also built different types of ensemble models such as bagging, voting and stacking so as to compare their performances in a comprehensive manner.

KNN	**Neural Network**	**Random Forest**
leaf_size=35, n_neighbors=9, weights='uniform'	learning_rate_init=0.0009, random_state=200	max_depth=12, max_features=4, min_samples_leaf=9, min_samples_split=38, random_state=0
Gradient Boosting - XGB	**Gradient Boosting - Hist**	**Ensemble (Bagging)**
max_depth=6, min_child_weight=1	l2_regularization=0, max_depth=9, max_iter=100, max_leaf_nodes=31	max_depth=10, max_leaf_nodes=105
Ensemble (Stacking)		
passthrough=False, stack_method= 'auto'		

Table 1: Final parameters for each model

We fit and evaluated our machine learning models by using the 5-fold cross-validation to find out the set of optimal hyperparameters for each model. After the cross-validation, some hyperparameters' values were adjusted to the optimal ones (see table 1) while others kept the default ones. We also conducted different feature engineering techniques, which involves scaling (standard scaling & min-max scaling) and dimension reduction (principal component analysis), upon the CPD dataset. The testing performance of each model (under the optimal hyperparameters) is shown in table 2.

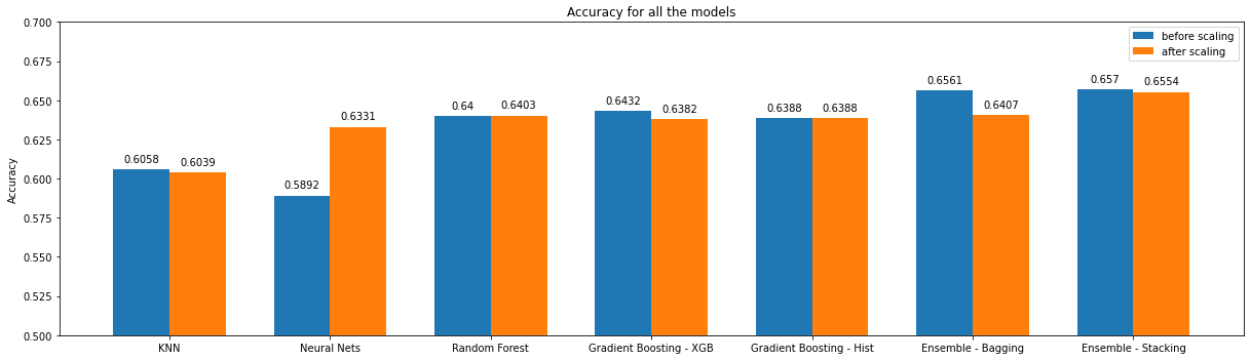
Machine Learning Testing Accuracy Results			
Model	**Accuracy Score**	**F1 Score**	**Average CV Score**
KNN	60.58%	44.08%	58.61%
Neural Network	63.11%	32.09%	52.55%
Random Forest	64.03%	44.34%	64.23%
Gradient Boosting - XGB	64.32%	44.17%	63.9%
Gradient Boosting - Hist	63.92%	44.54%	63.64%
Ensemble (Bagging)	65.61%	60.55%	64.77%
Ensemble (Stacking)	65.7%	37.07%	65.31%

Table 2: Final results

In Table 2, the ensemble model with the stacking method has the highest accuracy while the KNN has the lowest. The neural network does not show any advantage over other methods, and this implies that our constructed feature vector is not structurally meaningful. The performances of the random forest and gradient boosting classifiers are almost identical. The tradeoff between the variance and bias is actually balanced here regarding the task to predict a crime case's charge-I type. Another interesting observation is that most models' F1 scores are much worse than their percentage accuracy. One critical reason is that our CPD dataset is skewed towards the classes of misdemeanor and felony, and their amounts are significantly higher than those of others (local ordinance or traffic arrests). We got relatively good F1 scores (~70%) on predicting types like misdemeanor and felony, but for other types, the F1 score is only about ~15% since their proportion in the dataset may not be large enough to let the models be effectively trained. Consequently, the averaged F1 score for different classes should be lower than our expectation.

In addition, we found that the dimension reduction and scaling do not really improve the performance of machine learning models (see Figure 8). Scaling is only helpful for the neural network model (~4% increase of accuracy) while for others, its impact is not only negligible but also slightly decreases their performances since some of the features like ward or district are not numerically significant. This implies that a neural network is often sensitive to the

scale of data and data standardization can greatly improve its performance. We only have 11 features and thus the ineffectiveness of the dimension reduction is not quite surprising for us.



fig(8): Accuracy before and after scaling, comparison

Findings

Our conclusions indicated that there was some relationship between race and charge type. Using race, location, and time parameters, we were able to predict the charge type with a maximum accuracy of 65.7%. While this seems to indicate that race, location, and time have some predictive power in determining the charge type, they weren't the sole factors in influencing the severity of charge types subjects face. In other words, there are other factors not included in this model that would have improved the accuracy in determining the subject's crime charge type.

The biggest challenge we faced in this project was solving the above statement, which is to find more parameters that would improve the accuracy of the model. Initially, we collected data from a single dataset from the Chicago Police Department (CPD's arrests) that listed out the charge type, race, and arrest time of subjects. While the data was sufficiently large with over 500,000 entries, our initial models' performance was very weak. In order to solve this problem, we reconciled our original data with more data from the CPD that included information such as location of arrest, in the form of ward and district. Our model performance improved significantly thereafter reaching 65.7% as mentioned earlier. However, this came at the cost of removing a significant portion of our original dataset with only 106,042 entries left. While this was still a statistically significant amount of data, there could've been a lot to learn if we had larger computing power and more information about the arrests.

Although our models did not provide the highest performance, we acknowledge that the relationship between race and crime is not as clear cut. There are a lot of factors that go into

the arrest of a subject that cannot be assessed easily, such as existing prejudice of arresting officers or contextual information surrounding arrests. Despite this, data that could have been helpful, such as an officer's race, subject's past arrest history, or officer identification, were not available in the CPD data website. With race and crime being one of the most important issues impacting cities today, maintaining transparency and open access to data surrounding the topic can be an important way in reconciling the ramifications of potentially harmful urban crime policies.