Crime in Boston (CS109A-Project Group #14)

## Data Description

Our primary source of data is the [Crime Incident Reports](https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system) provided by the Boston Police Department. The data includes initial details of an incident to which police officers in Boston have responded. The data include details such as the type of incident, time and location of the incident, and the police district that the incident was assigned to. The dataset includes data from years 2015-2021.

Additionally, we have used [Boston neighborhood](https://data.boston.gov/dataset/neighborhood-demographics) and streetlight location data from City of Boston. The neighborhood data provides socio-economic information such as income, education, race, etc. for each Boston neighborhood. The [streetlight](https://data.boston.gov/dataset/streetlight-locations) data has the location of all streetlights within the city of Boston. We have also used Boston weather data from NOAA (https://www.ncdc.noaa.gov/cdo-web/) that has precipitation and temperature information for the time period of our study. With this data, we could potentially identify the impact of weather on our response variable.

## Initial Exploration, data cleaning, and reconciliation

During this phase we first merged all the crime data for 2015-2021. This resulted in more than 500k observations. The following cleaning and modifications were performed in order to combine this data with other datasets described in the previous section, the following were performed:

* The crime data until 2018 had column called UCR\_PART that categorized the crimes using FBI’s Uniform Crime Reporting Handbook. The UCR has more violent PART\_1 category that includes murder, rape, theft etc. and PART\_2 category that includes less violent crimes such as fraud, forgery, drug related crimes etc. The pre-2018 Boston data had also PART\_3 and ‘other’ for incidents that did not fall under those two categories (e.g., Sick calls). We created a list of keywords and new column called crime\_type and assigned all the incidents a type of 1, 2, or 3 based on FBI definition and Boston crime data.
* The Boston neighborhood data had information for each neighborhood while the crime data information was based on police districts. While some districts like C11 only include one neighborhood, there are many districts that cover more than neighborhood for example A1 that includes Downtown, North End, West End, and Beacon Hill. Combining the neighborhoods data to create district data was not a good option since we would lose lots of information and probably degrade the model as some very different neighborhoods could be within one district e.g., South Boston and South Boston Waterfront (Seaport). In order to assign neighborhoods to the crime incidents, we created latitude/longitude boundaries for each neighborhood using Google Map and used the already available latitude/longitude of the crime data and the district info to add a new neighborhood column to our primary data. Once this was done, we merged the crime data and neighborhood data using the neighborhood columns of the two datasets. In this process, we removed datapoints that were missing latitude/longitude information and were assigned to districts that covered more than one neighborhood.
* Weather data was merged with the dataset from previous step using the DATE variable so for each incident, we added weather information for that particular incident date. There were some days in the 2015-2021 period where no incident was reported and those weather information were removed.
* In order to include the information from the streetlight dataset, we calculated the distance if each incident using its latitude/longitude to the closest streetlight using the latitude/longitude information from the streetlight dataset. These distances will be used as a predictor for our model.

## Exploratory Data Analysis

### Neighborhood and crime

Our initial exploration showed that crime rates vary based on the neighborhood. For example, Mattapan has the highest crime rate, while an affluent neighborhood like Beacon Hill has comparatively low crime rate. As shown in the graph below, we can see the crime rate defined as number of incidents divided by population of the neighborhood varies significantly between different neighborhoods. Similar trend can be seen for shooting rate as well. The shooting rate is the number of shooting incidents divided by the population. Shooting rate is a good indicator of more violent crimes and it clearly shows why some neighborhoods were historically been referred to as dangerous neighborhoods.

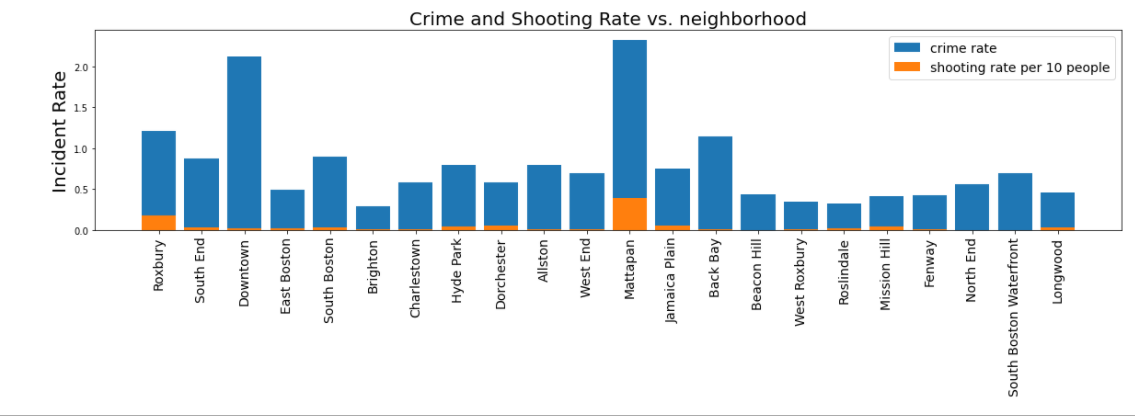


Figure 1-Neighborhood and crime

### Minority population (%non-white) and crime rate

It appears from the graph that the neighborhood with more minority race in percentage have higher crime rate in most cases. There could be other socio-economic factors which we will explore next.

Chart, histogram

Description automatically generated

### Income and crime rate

Here we look at the impact of income level on the crime rate. The below graph shows the percentage of households with high incomes (i.e., above 100k) for each neighborhood. We can see that the neighborhoods with less high earners have higher crime rate compared to more richer neighborhoods. As an example, Mattapan has fewer high earners and has higher crime rate compared to Beacon Hill which has lower crime rate and more high earners.

Chart, histogram

Description automatically generated

### Education and crime rate

As seen in the plot below neighborhoods which has lower education level have higher crime rate compared to neighborhood with higher education level. As an example, Mattapan has lower percentage of people with higher education, and has higher crime rate compared to Beacon Hill which has lower crime rate and higher education level.

Chart, histogram

Description automatically generated

### Weather and Crime type (Based on uniform crime report FBI)

Here we look at the impact of weather temperature on the crime rate defined here as number of incidents divided by number of days (with different temperature categories). For instance, the graph shows that on a given cold day where the average temperature was below freezing, on average around 40 type 1, 80 type 2, and 100 type 3 crimes happened. Based on the plot we can see that crime rate is higher for hot days compared to milder or cold days.

Chart, bar chart

Description automatically generated

## Research Question

What factors both socio-economic and non-socio-economic, can help us best categorize the severity of crimes as defined by Uniform Crime Report (UCR)?

Uniform Crime Report definitions are as follows:

* Part 1: For violent crime such as homicide, rape, murder etc.
* Part 2: For less severe crimes such as assault without aggravation, forgery, fraud etc.
* Part 3: For all other crimes that do not fall under Part 1 or Part 2.

## Feature Engineering

We performed feature engineering in our features such as street light distance and converted it to a binary predictor with 1 indicating crime incident happened within 20 meters, 0 otherwise. Similar feature engineering was performed on hour of the day. Population, and education, related features were scaled for each neighborhood. Holiday feature was added based on US holidays. We performed one hot encoding for all our categorical variables such as month and hour.

## Total Features for our Model

After feature engineering, and one hot encoding we have a total of 72 features. Our response is “crime\_type”.

## Splitting our data for training and test

We used a split of 80% for training, and 20% for test stratified by crime\_type.

## Naïve Model

The most common class in our response crime type was 3 or part 3 crimes as defined by UCR. Our dataset comprised of 51.72% of the crime incident that were categorized as 3. Our naïve model accuracy was 51.72%.

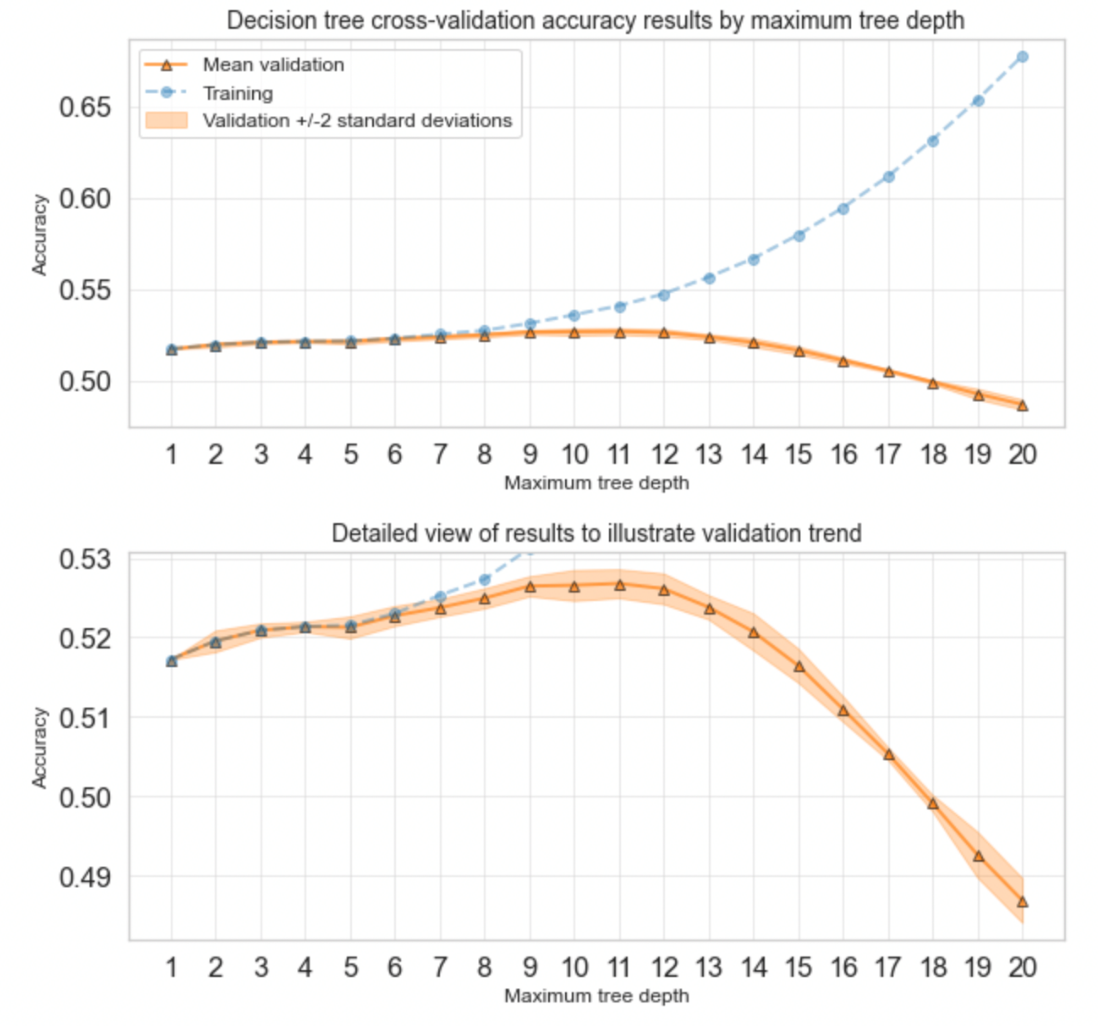
## Base Model

We selected Logistic Regression without any hyperparameter tuning as our base model. The accuracy of the model on test was 52.12%, just a slight improvement from our naïve model.

## Choice of models, hyperparameter training, model selection

**Logistic Regression with Lasso Regularization:** We trained this model with regularization with values between 1e-4 to 1e4 and cross validation of 10 folds. The model gave us accuracy of 51.72% on test. The model does not overfit but has a similar accuracy as our naïve model, and underperformed the base model.

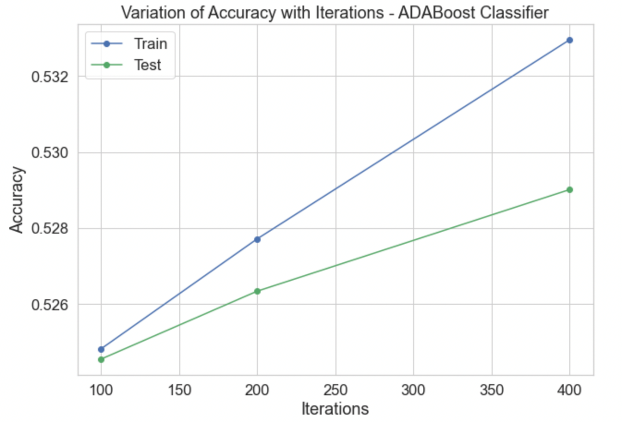
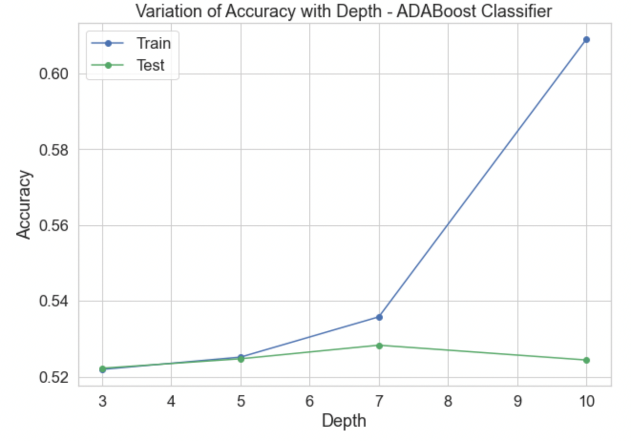
**Single Decision Tree with CV**: We trained single decision tree with depth hyperparameter using 5 folds cross validation. We did not achieve higher accuracy score using the best depth obtained from our hyperparameter training.



**Bagging:** Using the decision tree as a base, we trained bagging on depth above 7 with numbers of trees as our hyperparameter. However, we do not include this in our notebook, as because of the time limitation, we were not able to train more than 50 aggregated trees. And the results were similar to our single decision tree with best depth of 7. Instead, we decided to focus on Random Forest, and Boosting models.

**Random Forest Classifier:** As seen previously in our dataset response has imbalance classes with class 3 more than 50% of the total data; so, we have used “balanced random forest” classifier after up sampling the data using SMOTE; and the hyperparameter tree depth is tuned and 10 is chosen as optimal value for the model. Accuracy result has degraded compared to the base model, while the prediction is done for all the classes and performance metrics for individual classes have improved.

**Boosting:** We trained AdaboostClassifier using 2 hyperparameters, depth and number of estimators. We have used limited set of parameters as part of tuning exercise as it was slower to train on. The best tree depth as 5, and number of estimators as 100 were selected as part of this tuning exercise.



**Our own ensemble model**: After looking at the performance of all the above models, we decided to try an ensembled model of our own to see if accuracy gets improved. We ensembled three separate classification models i.e., Logistic Regression, Random Forest, and Adaboost and used the voting mechanism to choose the majority class, and final accuracy was better than our base model.

**Summary of Model Performance**

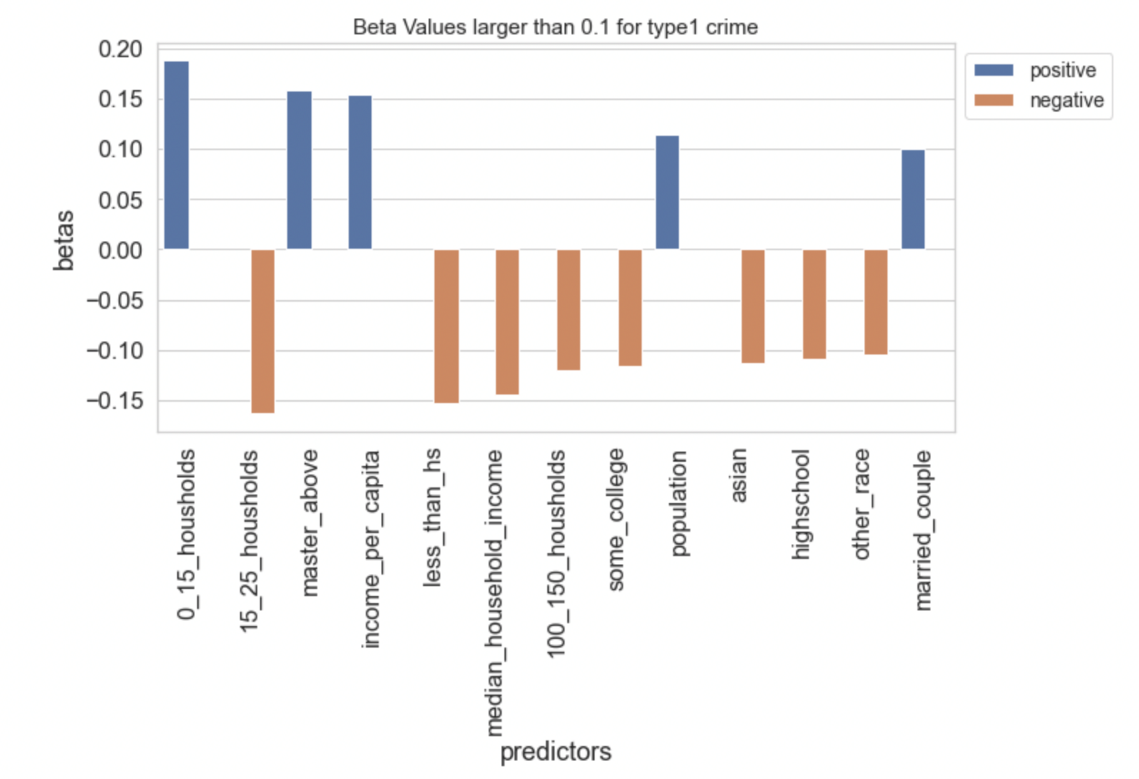
|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy %** | **AUC Score** |
| Naïve | 51.72 | - |
| Base Logistic Regression | 52.12 | 0.5716 |
| Logistic Regression with Lasso | 51.72 | 0.5591 |
| Single Decision Tree | 52.47 | 0.5577 |
| Balanced Random Forest | 47.04 | 0.5862 |
| AdaBoost | 52.60 | 0.5951 |
| Our own Ensemble Model | 52.60 | - |

## Model interpretation

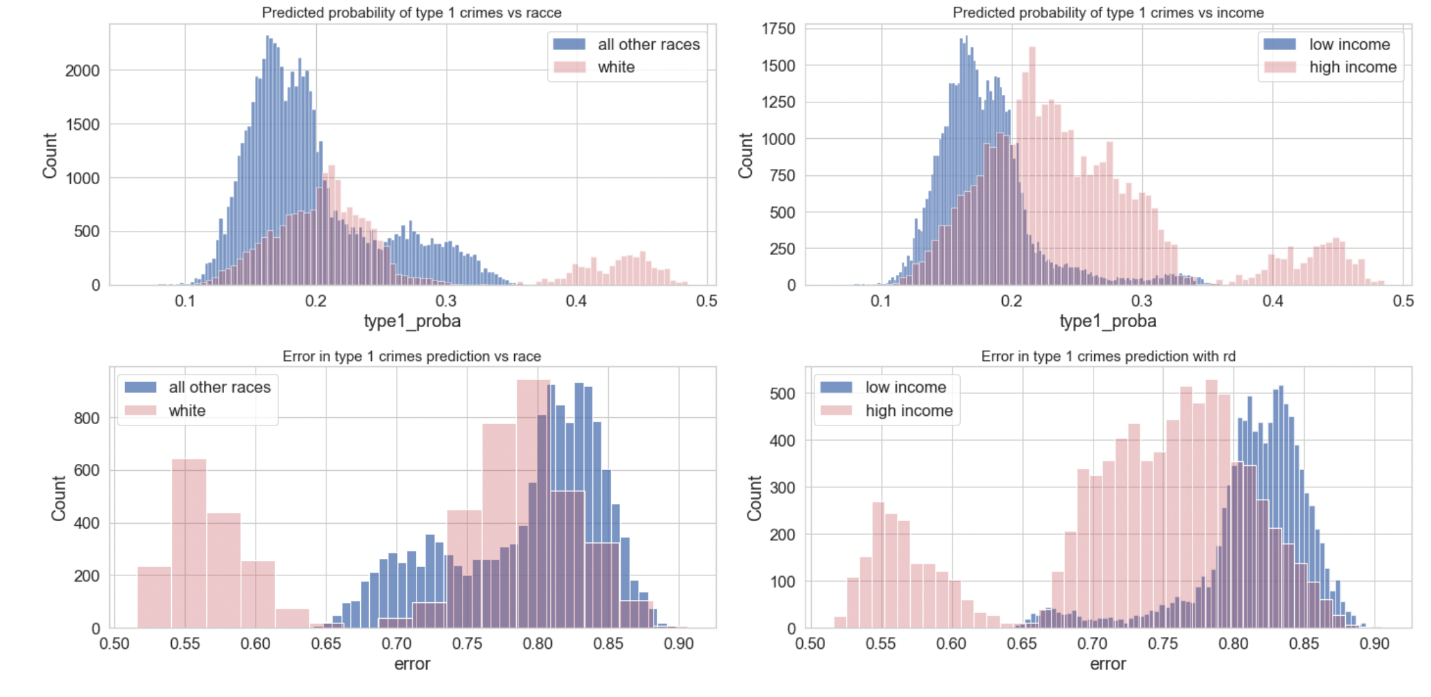
We have used below techniques to interpret the model outcomes.

1. Logistic Regression and Logistic Regression with Lasso Model explanation
2. Feature importance
3. Permutation importance

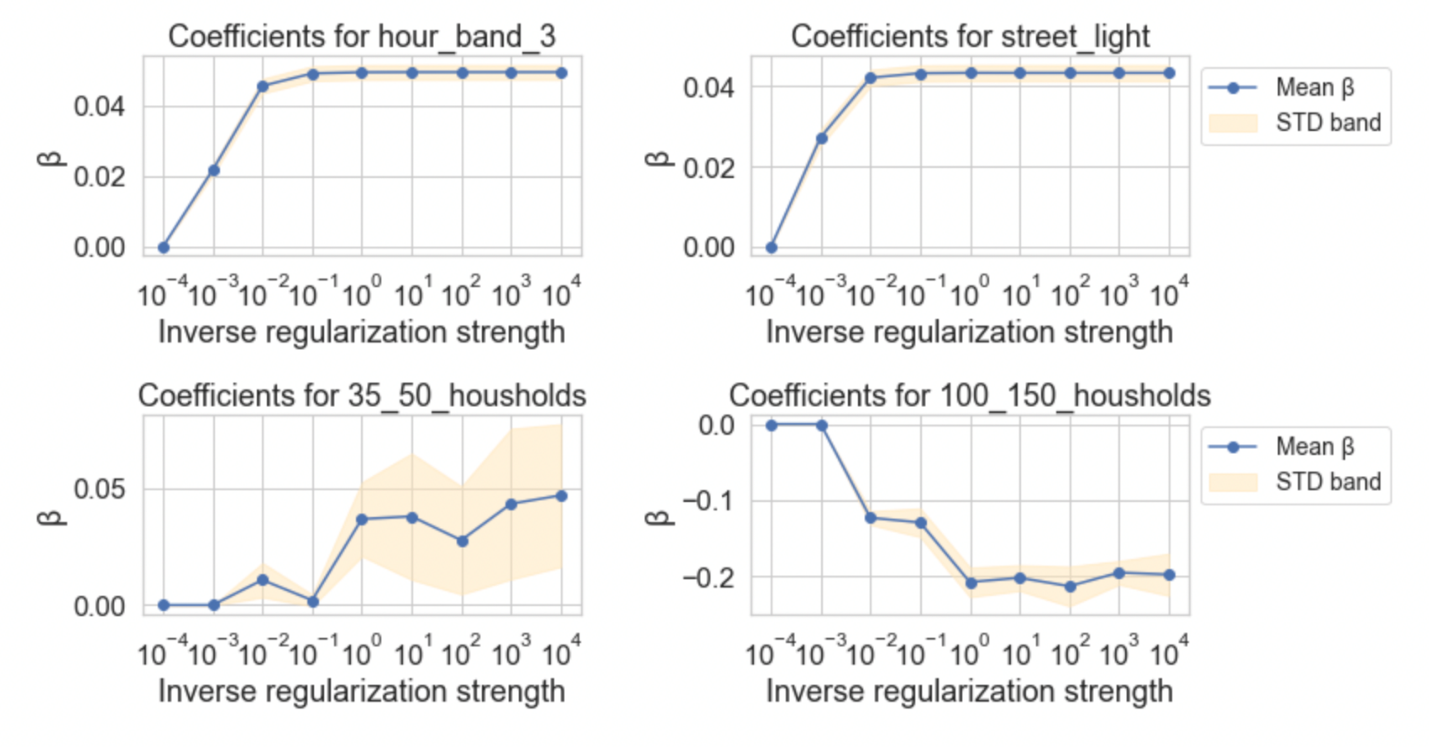
The results from the logistic regression model coefficients shows that the factors related to income, education and race have the largest impact on the prediction of type 1 crimes. The interpretation from these beta values should however be performed carefully because of correlation between different predictors. The high correlation between almost all socio-economic predictors is as a result of the nature of the data. We tried excluding some predictors however to remove the collinearity, we will lose all the socio-economic information.



The logistic regression model performs slightly better than the naïve model by predicting some observations as type 1. We looked at the impact of some factors such as race or income on the prediction capability of the model and as can be seen below, it appears that both type 1 prediction capability and accuracy is different between those different groups.



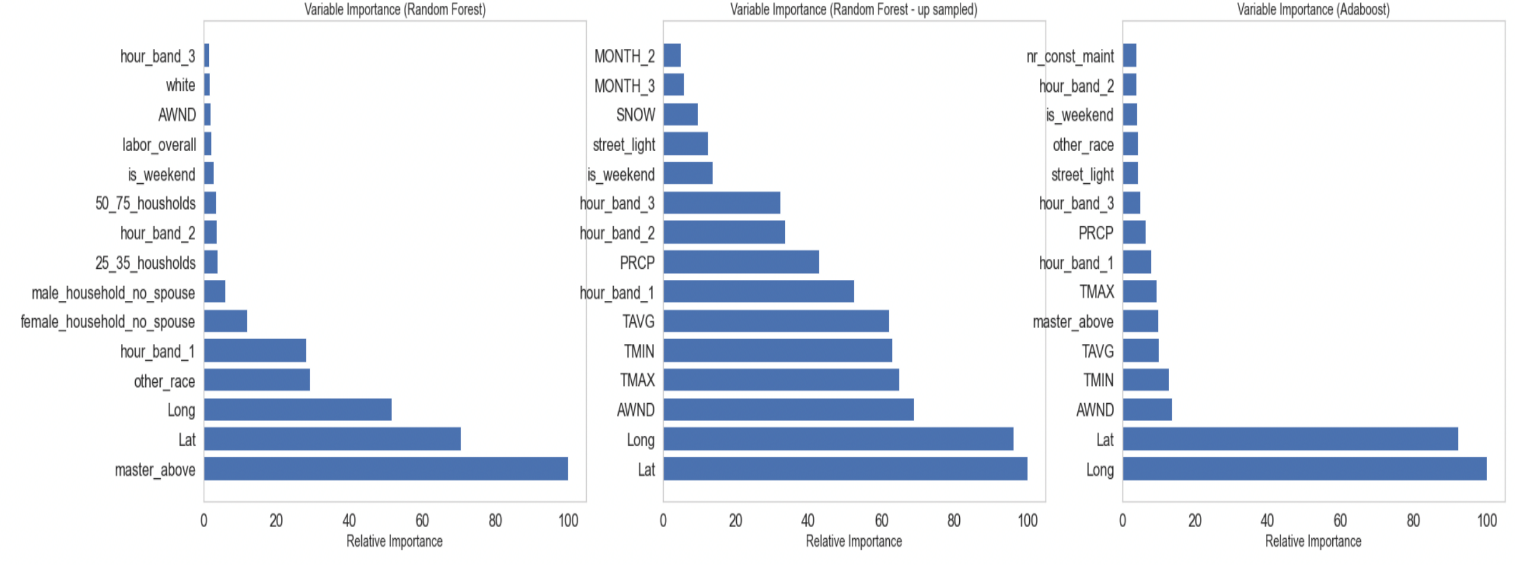
We then tried to use a Lasso-like logistic regression with different regularization strengths to see what the model will keep as significant predictors, and as it is shown in the notebook, most of the betas for type1 crime are zero for the selected model. We also show the estimation of significant betas with their confidence intervals for different cross validation sets. The results show that some betas such as street light or the hour band for the incident time has low standard deviation showing generally better estimations. The graph below shows the confidence interval for a subset of the coefficients.



## 

### Feature Importance

We have used tree-based models to predict the severity of crimes and tried to identify socio-economic and other factors those have played important role in predicting the severity of crime.



Based on our analysis of the features derived from tree models our findings shows that:

* Time of crime is one of the most important predictors as seen by many of the models.
* Another important predictor is weather information that too is associated with time, and all the tree models are conveying similar information.
* The distance of streetlight too has played an important role of predictor in the model.
* Another interesting predictor is the commute time to work in each neighborhood.
* One of the socio-economic factors, the household type, especially "household owned by male with no spouse" has also been one of the important predictors identified by these models.
* Another important socio-economic predictor "education" is also highlighted by few of the models as important predictors.
* "Race" too has come up as one of the important predictors from Adaboost, but the significant ones are non-white and non-black races.
* The type of industry in the neighborhood, especially "construction and maintenance" is identified as one of the important predictors.

So, we can see that predictors that have more variance with respect to crime, such as weather information, streetlight and time of crime have played a more important role, while the other socio-economic factors that are derived based on neighborhood data have played minor roles in predicting the severity of crime.  And intuitively we can say that this makes sense for this data set as we do not have many socio-economic predictors those can be directly associated with the crime incidents, all we have is for a given neighborhood and that is at very broad level; so, these factors have not provided major role in prediction; while the other items those are more granular with crime incidents (time, weather, streetlight) have more predictive power

### Permutation Importance

We have used model agnostic method to interpret our model to find the predictive power of the different socio-economic and non-socio-economic factors on prediction.

|  |  |
| --- | --- |
| **Adaboost** | **Random Forest** |
|  |  |
| **Logistic Regression** |  |
|  |  |

Based on our permutation importance analysis we see that tree models considers time and weather as the top predictors and some of the socio-economic factors; but logistic regression has picked up the median and average income as top predictors. When we look at "Adaboost" it has picked Socio economic indicators such as education and race as well as time and street light predictors. This is a very interesting finding and indicates that we need both types of predictors and might need to create ensembled model to correctly predict the severity of crime; and it might be important to see if more granular socio-economic information can help in better predicting a crime in Boston

## Model performance & Limitations

Our majority class for response is type 3 crime, and our naïve model accuracy based on predicting the majority class gave us an accuracy of 52%. We tried to improve upon the naïve model using different modeling algorithms, and training them using the optimal hyperparameters. The results are shown below:

**Summary of Model Performance**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy %** | **AUC Score** |
| Naïve | 51.72 | - |
| Base Logistic Regression | 52.12 | 0.5716 |
| Logistic Regression with Lasso | 51.72 | 0.5591 |
| Single Decision Tree | 52.47 | 0.5577 |
| Balanced Random Forest | 47.04 | 0.5862 |
| AdaBoost | 52.60 | 0.5951 |
| Our own Ensemble Model | 52.60 | - |

Based on the results shown we can see that each of the model only slightly improved upon the naïve model.

The main limitation of our models is because of the type of available predictors. Despite having 72 predictors, and 500K+ observation, there is not much variability between the predictors in each of the observation. For example, there are 22 neighborhoods in Boston, and the socio-economic data is similar across all the observations for a given neighborhood. Even though our EDA showed that socio-economic factor influence crime, the models did not identify the relationship between those socio-economic predictors and the crime type. We can see from our above analysis on feature importance that predictors that have more variance with respect to crime, such as weather information, streetlight and time of crime have played a more important role, while the other socio-economic factors that are derived based on neighborhood data have played minor roles in predicting the severity of crime.

## Model for city of Cambridge

In order to compare our findings with another geographic location, we looked at the crime in the city of Cambridge. The primary dataset provided by the City of Cambridge included the description of crimes as well as the time and location the crimes happened including the neighborhood. Additionally, we used the Boston weather data proximity of the two cities and also some socio-economic information for each neighborhood in Cambridge from [niche.com](http://niche.com/). We also created a similar response variable “crime\_type” using the description of crime and the keywords we created for UCR crime types.

 The final dataset for Cambridge had many similar types of information compared to Boston but the predictors were not exactly the same. Therefore, we could not use our Boston models to make predictions for Cambridge. We built a new logistic regression model for Cambridge and compared it to the Naïve model of predicting all Cambridge crimes as type 1. The results show that we have a similar accuracy issue with the Cambridge model. This was expected as we had the same issue of observation variance for all socio-economic data as they are similar at the neighborhood level. Additionally, Cambridge is a smaller city and unlike Boston, many of its neighborhoods are very similar in terms of socio-economic factors which makes it even harder to build a model that can accurately make predictions.

## Conclusion

Based on our EDA it appeared that socio-economic factors have influence on crime. However, our availability of socio-economic data was at neighborhood level. We only had 22 unique datapoints available for our whole dataset of 500K+ observations. Most of the predictors are correlated, and provided limited information to the model to predict crime type. The crime incident report provided us with only location, and time information for each crime incident reported. Based on these facts, model performance could not be improved upon by using complex models.

Even though the models were not able to identify the relationship between socio-economic factors and crime type, our EDA has confirmed that education, income and race of a neighborhood influences crime. This tells us that in order to truly model the socio-economic impact on crime we need more research in obtaining data that can be associated to individual crime incidents.

We have also seen that the non-socio-economic factors such as weather and time of the day has impact on crime. Our EDA showed there were more violent crime incidents when weather was warmer. Summer months have more crime incidents compared to winter months. Friday has overall higher crime incidents compared to other days. Our feature importance analysis on the model prediction show that these were the important features used.

## Next steps

For us to build a model to identify crime types, we will require predictors that are specific to each crime incident. We will also need to categorize the crime types consulting with subject matter expert, and filter out non-crime data points. We will also need to identify other indicators of crime incident such as historical crime trend associated with bias in society.