# Evaluation of Factors Affecting Violence, Property-related, and Other Trivial Crimes from 2020 to 2021 in Boston

Chenwei Wu, Joslyn Fu, Katherine Hu, Lu Yu

## Abstract

Horrified by the recent violent crime cases in universities across the U.S., our team hopes to better understand crime cases in Boston, a well-recognized city clustered with higher education institutions. Through evaluating various factors affecting violence, including shooting, property-related, and other more trivial crimes, we found out that all crimes are more likely to happen during night and less busy hours, away from streetlight and traffic, and in neighborhoods less educated and affluent. Distinctly, whereas violent crimes are related to weather factors, property crimes are more dependent on distances from streetlights and traffic, and also differentiate from the rest with respect to local residential structures.

# Introduction

Recent tragedies at higher education institutions have called our team's attention to crime cases. The graduate student at Columbia University, Davide Giri, was fatally stabbed in December 2021[1]. An 18-year-old man has been arrested and charged with the murder of a UChicago graduate student shot and killed during a robbery near the school campus in November 2021[2]. In Boston, the chance of becoming a victim of violent crime is 1 in 151, and the chance of becoming a victim of property crime is 1 in 51[3]. Due to our proximity to Boston, we intend to understand what factors contribute to different types of crime in the Boston area from 2020 to 2021.

We divided all types of crimes into 3 major categories. **Violence crimes** include simple assault, etc., **property-related crimes** include larceny, vandalism, investigate property, larceny from motor vehicle, etc., and **other more trivial crimes** include motor vehicle accident response, investigate person, medical assistance, fraud, towed, etc. For violence crime, we specifically sift out **shooting crimes** to see how shooting is different from other violent crimes. We build classification models to predict the type of crime with predictors, composed of various variables extracted from weather, location, MBTA ridership, neighborhood, streetlight, economic, and property assessment data.

Based on the above classification of crimes, our research directions are as below:

1) Shooting: we want to see how shooting is different from other violent crimes. We suspect that there are some factors that are more correlated with shooting crimes versus other violent crimes like robbery and assault.
2) Violence crimes: We are interested in how violence crimes are different from property crimes and other trivial crimes in terms of when and where it occurred. We are interested in exploring the major contributing factors to violence crimes.
3) Property crimes: Similar to violence crimes, we are interested in exploring the major contributing factors to property-related crimes.
4) Other more trivial crimes: Similar to violence crimes, we are interested in exploring the major contributing factors to other more trivial crimes.

# Data

## Data Source

We retrieve data from 9 data sources. They are Boston Crime data, Neighborhood data, Weather Data, Streetlight Data, MBTA Ridership and Stations Data, Police Districts Data, Zip Codes Data, Economic Indicators Data, and Property Assessment Data.

[1] https://www.nytimes.com/2021/12/03/nyregion/columbia-student-stabbed.html
[2]https://www.fox32chicago.com/news/man-18-charged-with-murder-in-shooting-death-of-university-of-chicago-graduate
[3] https://www.neighborhoodscout.com/ma/boston/crime

## Exploratory Data Analysis

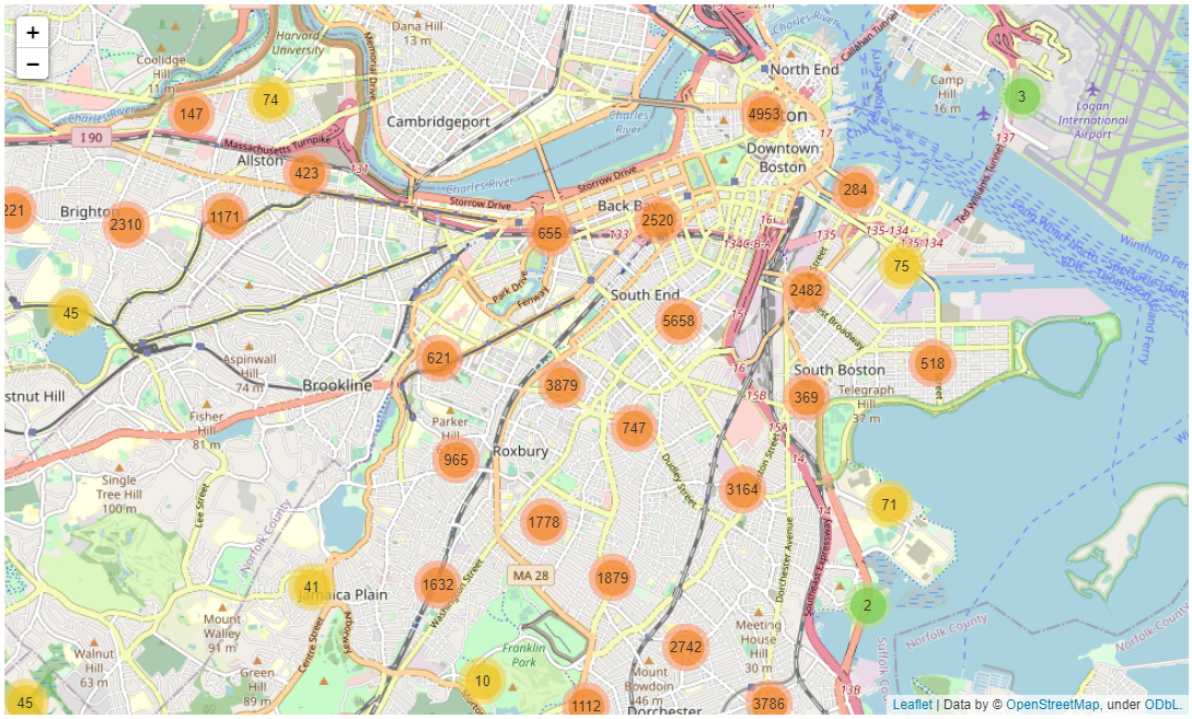


**Fig 1.** Geographical Distribution of Crimes Number across Boston.

This map is generated from the Boston crime data. From this geomap we could see that most crimes take place around several central clusters (Downtown Boston and Roxbury etc.), indicating that geographical location might be a strong indicator for crimes. We decided not to calculate the percentage of the crimes in each district because we are not setting boundaries for crimes and using the district population as a direct predictor for the target values. Instead, we use ridership of the nearby MBTA stations/stops to give a more accurate estimation of the crowdedness near the crime location in different time periods of a day.

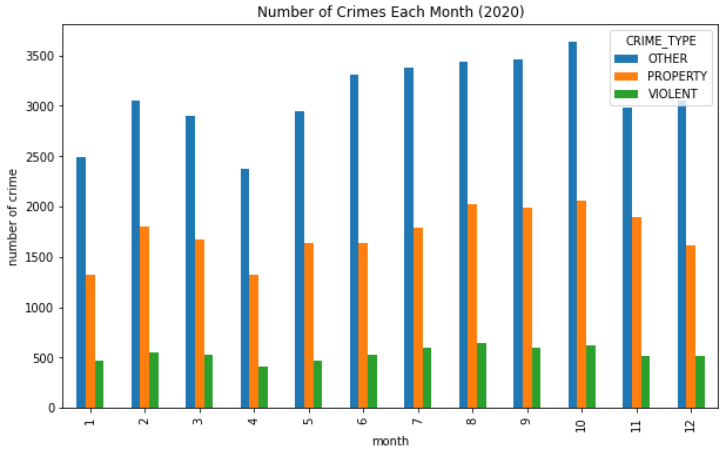


**Fig 2.** Monthly crime number in 3 types.

The number of crimes in all three types peaks around October but gets to the valley around January and April.

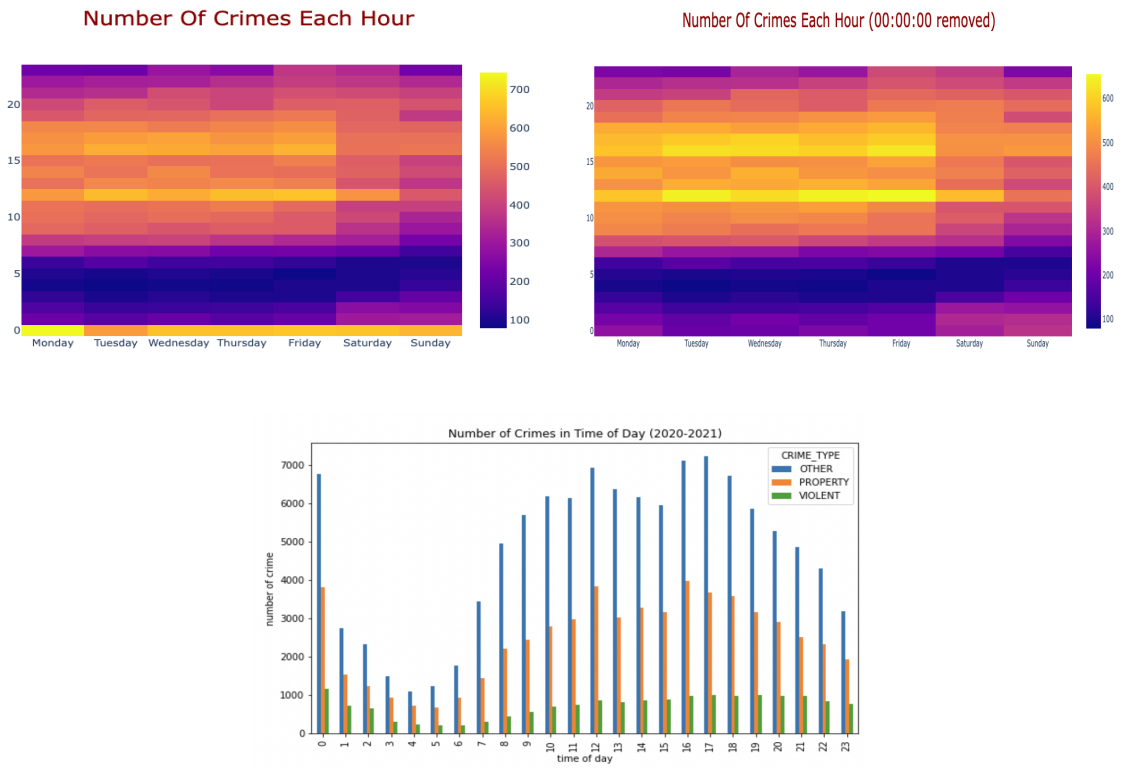


**Fig 3, 4, 5.** Number of Crimes in Time of Day.
**Upper Left:** # of crimes each hour for each day of week.
**Upper Right:** # of crimes each hour for each day of week removing the 00:00am for each day.
**Bottom:** # of crimes each hour.

The left upper plot shows the number of crimes in each hour of a day for each day of week. However, the number of crimes is exceptionally high for each day from 00:00 am-1:00 am. We can see such a phenomenon more obvious in our bottom plot. We found that a large number of crimes occured time is at 00:00:00 am. As TF mentioned in our milestone 1, it is possible for the record to default all crime incidents without a specific or known time-of-day to 00:00 am for that day and default those without a known time-of-week to 00:00 am next Monday for the previous week. We plot the upper right plot by removing the crime incidents recorded to occur at 00:00am where we can see a sharp decrease in crime number at 00:00 am of each day. However, because we could not find official documentation to support our hypothesis, we decided to leave it as it is for our model training part. This will be a limitation to our model.

## Cleaning Process

The crime incident reports provided by Boston Police Department (BPD) are our primary interest, and we decided to use crime type as the response variable. The raw data ranges from 2015 to 2021, however, due to the impact of COVID-19 on crimes, we decided to only focus on post-COVID-19 data from January 2020 to November 2021. Since the data after 2019 only has numerical offense codes and their group missing, we used past data from 2017 to 2018 to create a dictionary that maps from offense codes to offense groups. In this way we could guarantee that more than 95% of the crime incidents from 2019 to 2021 can be assigned an offense group code. Then we dropped incidents that are duplicates or could not be mapped to an offense code group. We also removed data with missing/zero latitude and longitude. And we also remove the data that occurred outside of Boston by filtering out those data with the district being ‘EXTERNAL’ or the longitude and latitude being located outside of the Boston zip codes. To make more thorough use of the data, we also did feature engineering to generate more features (refer to **Appendix A** for more details). After cleaning, there are in total 66973 data points available. Since so many predictors were available, it would be unwise to include them all since that would cause

multicollinearity and ruin predictions. Thus we conducted correlation analysis by generating Pearson correlation matrices and eliminated many variables that have too high of a correlation with each other. Lastly we performed one-hot encoding for the categorical variables such as months and days of the week and dropped one to avoid multicollinearity.

### Sample Data

There are in total 66822 rows of data with 4 columns of target values and 108 selected features as predictors after all the preprocessing. Each row is a reported crime in Boston and the corresponding predictors in regards to the occurred time, location-related factors, economic indicators of that month, weather during that time, neighborhood of the location, and surrounding property value. Please refer to the above feature engineering section or our data dictionary in **Appendix D** for detailed description of the features.

## Model Selection and Results

### Baseline/Imbalance Assessment

| Target Value | True | False | Total | Percentage (%) | Accuracy Rate |
|---|---|---|---|---|---|
| **Shooting** | 759 | 5257 | 6016 | 12.6 | 87.4% |
| **Violence** | 6016 | 60806 | 66822 | 9 | 91% |
| **Property** | 21093 | 45729 | 66822 | 31.6 | 68.4% |
| **Other** | 39713 | 27109 | 66822 | 59.4 | 59.4% |

**Table 1.** Percentage of Positive Class for Each Target Value

As shown in Table 1, for target value shooting and violence, imbalance class is a big problem. Only 12.6% of violence crimes are reported as shootings which are in total 760 cases. And only 9% of all crimes are recorded as a type of violence crime. Property-related crimes and other crimes are relatively more balanced than shootings and violence but the imbalance class issue is still obvious.

Our baseline models would be just the naive percentage of the larger class. For shooting, the baseline model would predict all cases as not shooting and achieve an accuracy of 87.4%. For violence, the baseline model would predict everything as non-violence crimes and have an accuracy rate of 91%. For property, the baseline model would assign every crime as non-property related and achieve a 68.4% accuracy rate. To predict other trivial crimes, the baseline model would predict all occurrences as other trivial and have an accuracy rate of 59.4%.

### Solutions to Imbalance Classification

Except for Adaboost that we do not have the corresponding parameter, we use class weighting to all other models we tried. For a logistic regression, it is basically weighting the loss of sample by its class weight to compute the loss function. For a tree, what it does is to assign a weight to each class (positive or negative) when evaluating the “impurity” of a split point. The weight is calculated by the inverse of the portion of class frequency in the input. For example, the weight of the positive class, like shooting, is simply n_samples / n_shooting whereas the weight of the negative class, like not shooting, is n_samples / (n_samples - n_shooting).

There are also other options to deal with the imbalanced class problem such as downsampling and upsampling. We decided not to use downsampling because we don't want to lose information by randomly dropping a large number of cases with non shooting. The reason not to use upsampling is that there are more than 100 features which makes the upsampling very difficult and not reliable.

Our evaluation metrics also take the imbalance class into account. Instead of only looking at accuracy scores, we took ROC-AUC scores, F1 scores, Recall scores and confusion matrix into consideration. For more details, please check the **Appendix E Evaluation Metrics.**

## Model Selection Process

First, we split the data into a training set and a test set with test dataset size 20%. Then, we find the best set of hyperparameters using GridSearch with 5-fold cross validation on the training set. Lastly, we train each model with the best combinations of hyperparameters and select the best model based on the 5 metrics. We prioritize ROC-AUC scores and F1 scores. If there's a discrepancy between the two scores, we will use the recall score as well.

## Experiments

We conducted experiments on Regularized Logistic Regression, Random Forests, AdaBoost, XGBoost and LightGBM. For detailed explanation of these models, please refer to **Appendix E Models Used**.
Grid search over sets of hyperparameters are used to fine tune our models and get best estimators for each of our outcome variables.

### Shooting

| model name | roc_auc_score | accuracy_score | f1_score | recall_score | tn | fp | fn | tp | Best Hyperparameters | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | C | lr | n_estimators | max_depth | min_sample_leaves | min_sample_split |
| XGBoost | 0.799 | 0.724 | 0.231 | 0.694 | 822 | 310 | 22 | 50 | - | 0.5 | 200 | 1 | default | default |
| LightGBM | 0.793 | 0.716 | 0.223 | 0.6805 | 813 | 319 | 23 | 49 | - | 0.1 | 60 | default | 2 | default |
| Logistic | 0.788 | 0.735 | 0.242 | 0.708 | 834 | 298 | 21 | 51 | 0.001 | - | - | - | - | - |
| **RandomForest** | **0.8** | **0.803** | **0.28** | **0.639** | **921** | **211** | **26** | **46** | **-** | **-** | **600** | **6** | **7** | **2** |
| AdaBoost | 0.746 | 0.936 | 0.115 | 0.069 | 1122 | 10 | 67 | 5 | - | 0.3 | 160 | 2 | default | default |

**Table 2.** Shooting - Best Hyperparameters for Each Model and Evaluation Metrics
We selected Random Forest as our best model to identify Shooting from other violence crimes because it has the highest ROC-AUC score and F-1 score.

### Violence

| model name | roc_auc_score | accuracy_score | f1_score | recall_score | tn | fp | fn | tp | Best Hyperparameters | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | C | lr | n_estim | max_de | min_sampl | min_samp |

| | | | | | | | | | | | ators | pth | e_leaves | le_split |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **XGBoost** | **0.598** | **0.607** | **0.193** | **0.525** | **7492** | **4680** | **567** | **626** | **-** | **0.5** | **900** | **1** | **default** | **default** |
| LightGBM | 0.585 | 0.569 | 0.189 | 0.563 | 6935 | 5237 | 521 | 672 | - | 0.1 | 340 | default | 2 | default |
| Logistic | 0.573 | 0.628 | 0.178 | 0.451 | 7853 | 4319 | 655 | 538 | 0.0001 | - | - | - | - | - |
| RandomForest | 0.6 | 0.764 | 0.187 | 0.305 | 9841 | 2331 | 829 | 364 | - | - | 200 | 10 | 4 | 8 |
| AdaBoost | 0.588 | 0.91 | 0 | 0 | 12160 | 12 | 1109 | 0 | - | 0.1 | 160 | 3 | default | default |

**Table 3**. Violence - Best Hyperparameters for Each Model and Evaluation Metrics
XGBoost is our best model to predict violence versus non violence crimes with the highest ROC-AUC score and F-1 score.

## Property

| model name | roc_auc_score | accuracy_score | f1_score | recall_score | tn | fp | fn | tp | Best Hyperparameters | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | C | lr | n_estimators | max_depth | min_sample_leaves | min_sample_split |
| **XGBoost** | **0.644** | **0.618** | **0.487** | **0.575** | **5838** | **3311** | **1791** | **2425** | **-** | **1** | **2000** | **1** | **default** | **default** |
| LightGBM | 0.645 | 0.621 | 0.485 | 0.565 | 5921 | 3228 | 1832 | 2384 | - | 0.01 | 340 | 20 | 100 | default |
| Logistic | 0.62 | 0.596 | 0.459 | 0.545 | 5663 | 3486 | 1920 | 2296 | 100 | - | - | - | - | - |
| RandomForest | 0.66 | 0.669 | 0.474 | 0.473 | 6945 | 2204 | 2222 | 1994 | - | - | 200 | 15 | 4 | 6 |
| AdaBoost | 0.647 | 0.71 | 0.257 | 0.162 | 8753 | 396 | 3535 | 681 | - | 0.1 | 180 | 3 | default | default |

**Table 4.** Property - Best Hyperparameters for Each Model and Evaluation Metrics
When predicting violence versus non violence crimes, XGBoost has the highest F-1 score. However, it has a ROC-AUC score lower than Random Forest, so we also take into account the recall score where XGBoost is higher than Random Forest.

## Other

| model name | roc_auc_score | accuracy_score | f1_score | recall_score | tn | fp | fn | tp | Best Hyperparameters | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | C | lr | n_estimators | max_depth | min_sample_leaves | |
| XGBoost | 0.624 | 0.592 | 0.642 | 0.615 | 3025 | 2384 | 3064 | 4892 | - | 1 | 3000 | 1 | default | default |
| LightGBM | 0.632 | 0.610 | 0.675 | 0.679 | 2757 | 2652 | 2554 | 5402 | - | 0.05 | 100 | 20 | 100 | default |
| Logistic | 0.614 | 0.589 | 0.65 | 0.641 | 2771 | 2638 | 2855 | 5101 | 100 | - | - | - | - | - |
| **RandomForest** | **0.65** | **0.642** | **0.749** | **0.898** | **1440** | **3969** | **813** | **7143** | **-** | **-** | **1000** | **default** | **40** | **6** |
| AdaBoost | 0.588 | 0.62 | 0.743 | 0.62 | 894 | 4515 | 578 | 7378 | - | 0.05 | 140 | 1 | default | default |

**Table 5**. Other Trivial - Best Hyperparameters for Each Model and Evaluation Metrics
We choose Random Forest as our best model to predict other trivial versus non other trivial crimes because it has the highest ROC-AUC score and F-1 score.

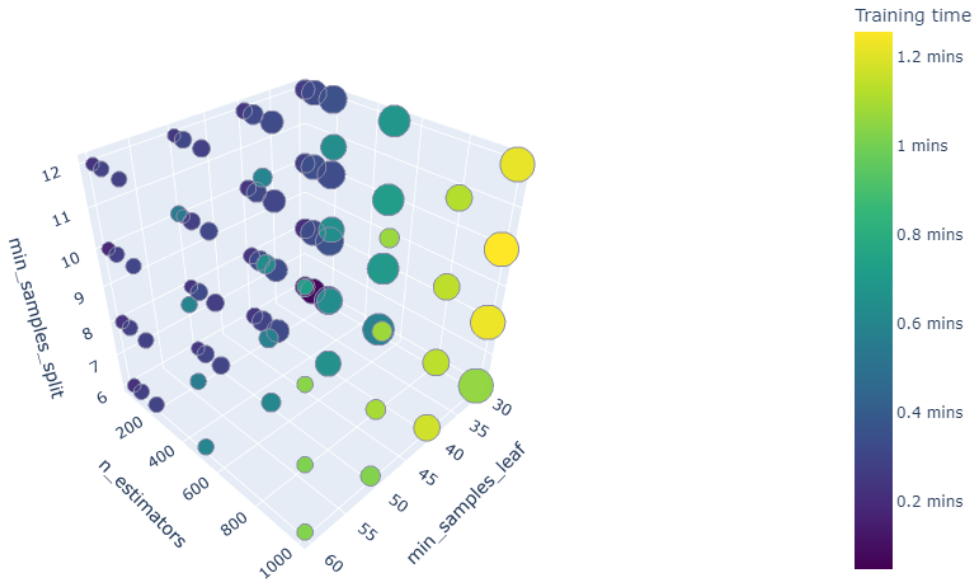


**Fig 6**. Example of grid search results for Random Forest,Other Trivial crimes

**Fig 6** is a visualization of the grid search hyperparameter tuning we conducted for Random Forest on Other Trivial Crimes prediction. We tuned the model on the number of estimators, min samples per split and min samples per leaf. The size of each data point represents the mean test AUC for the 5-fold cross validation conducted on a certain set of hyperparameters and the color represents the training time. We could see that the mean test AUC increases as the min sample per leaf decreases, the number of estimators increases and min samples per leaf increases.

# Feature importance & Assessment

The evaluation of the impacts of model predictors on class predictions are based on both model scale (global) importances including coefficients, gain, weight, cover, and SHAP values, as well as instance wise (local) importances including LIME.

## Shooting (Logistic Model & Random Forest)

For shooting we would like to investigate and compare Logistic Regression which has the highest recall as the prediction of the positive class (shooting) here is of quintessential interest to us, and also the overall best performing Random Forest which has highest F1 score.

For the logistic regression model selected for shooting prediction, feature coefficients in **Fig 7** are helpful to understand their importances. Here we could see that time features (night, hour_1), neighborhood affluence features(1850-1900,1950-2000,poverty_rate, bachelor's_degree), traffic(stop_dist, within_dist_stops_9) are all affecting the occurrences of shooting. The high positive coefficient of night and hour_1 indicate that shooting incidents are highly likely to happen at night, especially midnight. The positive coefficients for poverty rate and number of old houses, together with negative coefficients for education and number of more recently built houses illustrate that shooting is more likely to happen in neighborhoods that are less affluent. Further distance from a stop and fewer within distance stops are also factors for shooting.

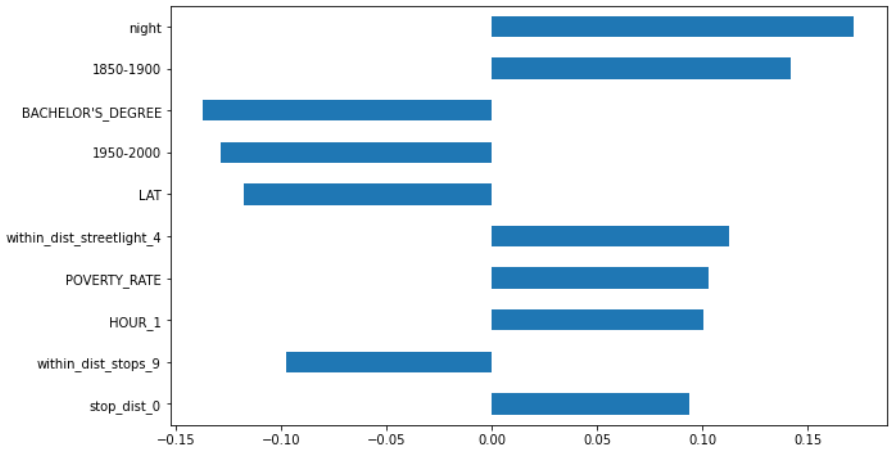


**Fig 7**. Logistic regression coefficients

For Random forest we are looking at its mean decrease in impurity in **Fig 8**. For each feature we can collect how on average it decreases the impurity (and gains information). We could tell that apart from the time and nearby traffic/streetlight factors indicated by logistic regression, Random Forest also places great importance on the household types, and educational factors.

| Weight | Feature |
|---|---|
| 0.0624 ± 0.1866 | LAT |
| 0.0419 ± 0.1804 | FEMALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT |
| 0.0361 ± 0.1744 | MALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT |
| 0.0348 ± 0.0923 | LONG |
| 0.0344 ± 0.1631 | TOTAL_VALUE_MEAN |
| 0.0320 ± 0.0761 | within_avgons_13 |
| 0.0319 ± 0.1520 | _MASTER'S_DEGREE_OR_MORE |
| 0.0316 ± 0.0904 | within_avgons_11 |
| 0.0284 ± 0.1293 | MARRIED-COUPLE_FAMILY |
| 0.0270 ± 0.0840 | within_avgons_9 |
| 0.0257 ± 0.0754 | within_avgload_11 |
| 0.0255 ± 0.0824 | night |
| 0.0230 ± 0.1197 | BACHELOR'S_DEGREE |
| 0.0226 ± 0.0641 | stop_dist_0 |
| 0.0214 ± 0.0606 | within_dist_streetlight_8 |
| 0.0210 ± 0.0533 | within_avgload_13 |
| 0.0199 ± 0.0549 | within_avgoffs_13 |
| 0.0198 ± 0.0620 | within_dist_streetlight_9 |
| 0.0179 ± 0.1119 | GED_OR_ALTERNATIVE_CREDENTIAL |
| 0.0175 ± 0.1153 | ASSOCIATE'S_DEGREE |

**Fig 8**. Mean decrease in GINI for RF (shooting)

*LIME and SHAP values analysis are also conducted but we will introduce them in the next section when we discuss violent crimes. See* ***Appendix C - Fig 28,29****.*

In short, shooting incidents have higher chances of occurrence at nights, in neighborhoods where education level is lower, householders are living by themselves, and poverty rate is high, and also in places that are further away from streetlight and public transport.

*Multiple sets of metrics are used to evaluate the tree models we selected for violence, property, and other crimes. For conciseness, visualizations for violence would be fully showcased here and rest of the graphs are included in the* ***Appendix C****.*

## Violence (XGBoost)

1) Three tree-model specific importance measures are generated for XGBoost (**Fig 9-11**):

| Weight | Feature |
|---|---|
| 0.1944 | night |
| 0.1234 | 1850-1900 |
| 0.0344 | HOUR_7 |
| 0.0341 | HOUR_8 |
| 0.0292 | HOUR_9 |
| 0.0256 | within_dist_streetlight_5 |
| 0.0245 | cpiu |
| 0.0233 | MONTH_8 |
| 0.0184 | edu_health_wage |
| 0.0174 | HOUR_21 |
| 0.0170 | finance_wage |
| 0.0159 | within_dist_stops_9 |
| 0.0149 | HOUR_15 |
| 0.0147 | within_dist_streetlight_8 |
| 0.0144 | within_avgoffs_11 |
| 0.0141 | NEWER_THAN_2000 |

| Weight | Feature |
|---|---|
| 0.0145 | 1850-1900 |
| 0.0144 | night |
| 0.0143 | MONTH_8 |
| 0.0142 | HOUR_8 |
| 0.0142 | edu_health_wage |
| 0.0142 | HOUR_9 |
| 0.0142 | HOUR_7 |
| 0.0142 | cpiu |
| 0.0142 | within_dist_stops_9 |
| 0.0142 | within_dist_stops_7 |
| 0.0142 | HOUR_1 |
| 0.0142 | NEWER_THAN_2000 |
| 0.0142 | info_wage |
| 0.0142 | TOTAL_VALUE_SUM |
| 0.0141 | HOUR_2 |
| 0.0141 | HOUR_18 |

| Weight | Feature |
|---|---|
| 0.1067 | stop_dist_0 |
| 0.1022 | LAT |
| 0.0922 | streetlight_dist_0 |
| 0.0878 | LONG |
| 0.0544 | within_avgoffs_13 |
| 0.0478 | within_avgons_13 |
| 0.0444 | within_avgload_13 |
| 0.0389 | FeelsLikeC |
| 0.0344 | within_avgons_11 |
| 0.0289 | within_avgload_11 |
| 0.0278 | within_dist_stops_13 |
| 0.0267 | within_avgons_7 |
| 0.0256 | within_avgoffs_11 |
| 0.0233 | within_avgoffs_9 |
| 0.0222 | within_avgons_9 |
| 0.0178 | windspeedKmph |

**Fig 9, 10, 11**. Gain, XGBoost, Violence; Cover, XGBoost, Violencel Weight, XGBoost, Violence

The gain metric averages each features' contribution for each tree in the model and calculates the relative contribution to prediction, while the cover metric indicates the relative quantity or support of a feature and weight metric counts the occurrences of a feature in the splits across the model. The three graph together indicates that time (night, hour, month), distance from streetlight (within_dist_streetlight_8) and nearby traffic(within_avgoffs) , neighborhood affluence (1850-1900, cpiu, edu_health_wage) , and weather(FeelsLikeC, windspeed) are having the most influence on occurrences of violent crimes.

Despite the useful information for overall strength of features these metrics may provide, other metrics are needed to understand if the predictors have a positive/negative impact on the predictions. It would be also of practical interest to look at individual instances via model agnostic importance metrics. Thus the SHAP value and LIME metrics are also deployed here.

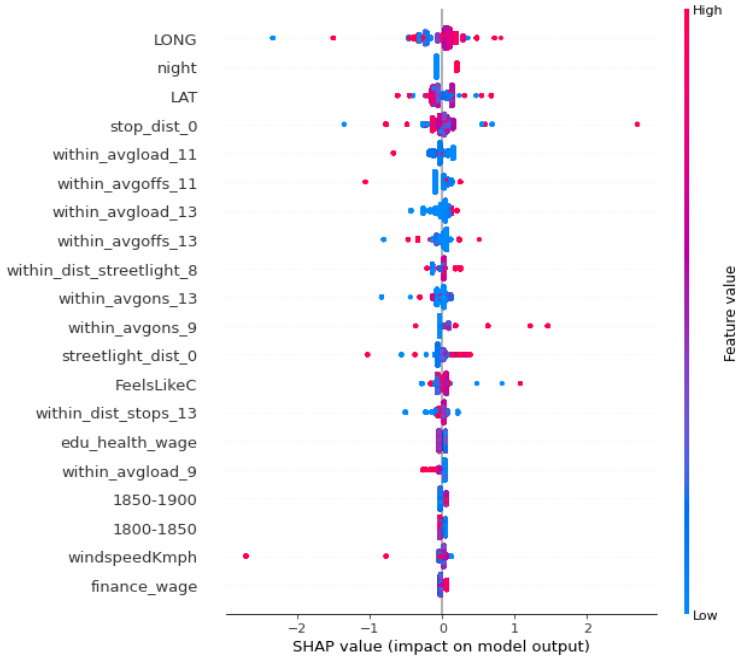


**Fig 12.** SHAP Value Swarm Plot, Violence

The SHAP values approach trains distinct models on subset coalitions in the feature space and explains the marginal change in probability of violent crime incidence. It not only tells which features are most important, and also their range of effects over the entire test set. The color palette allows matching of value changes with the impact on model output. Positive Y values means that it is contributing positively to the model output and increase in predicted probability of crime incidence and vice versa for negative values. Our SHAP value swarm chart in **Fig 12** indicates similar results as the tree model importances, with night, distance from streetlight (within_dist_streetlight_8) and nearby traffic(within_avgoffs), neighborhood affluence and weather dominating the predictions, except for that the geographical location itself plays much more importance in SHAP evaluations.

While a swarm plot provides model-scale understanding of each feature, plotting the SHAP dependence against each feature shows how the output is influenced across its value range and captures data trends. Examples for each of these feature groups are given below. It is clear that empirically, higher apparent temperature (FeelsLikeC) and closer distance to a stop and more traffic nearby generally contributes to higher probability of violent crimes.

For weather, we could see in **Fig 13-14** that higher temperatures are positively correlated with the violent crimes and higher wind speed are negatively correlated with violent crimes and both their effects increase drastically as it reaches more extreme weather conditions (>40C />40 Kmph).

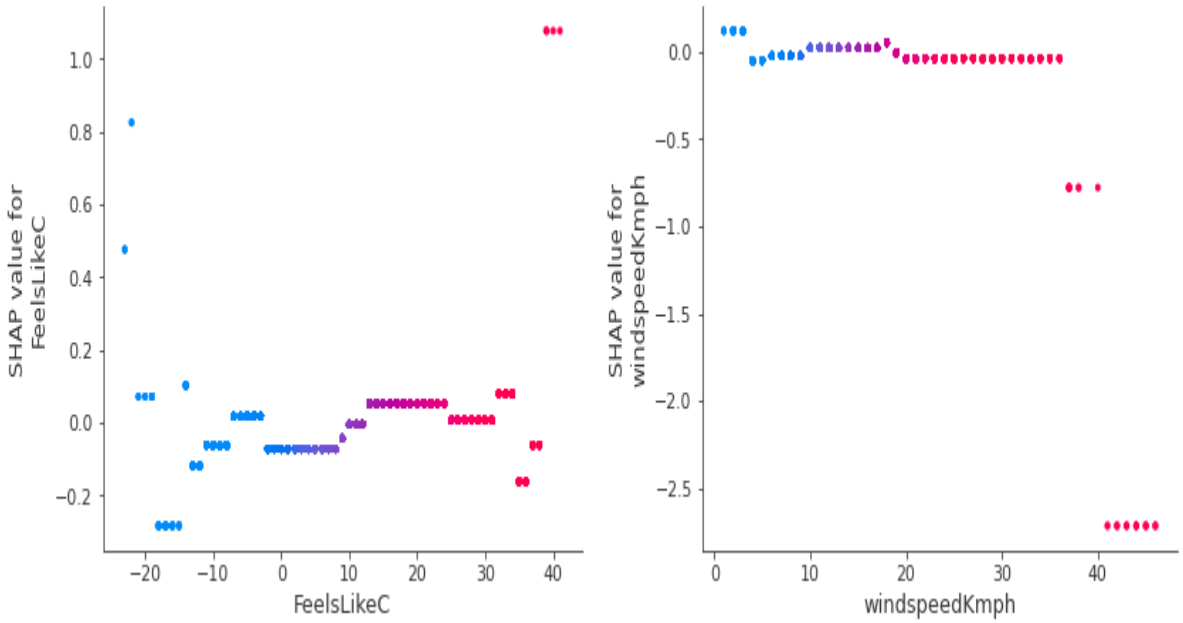


**Fig 13, 14.** Weather SHAP dependence plots

Juxtaposing the three dependence graphs **Fig 15-17** for streetlight distance, stop distance and nearby public transport loads, we could see that the further away from the streetlight, the higher probability that a violence crime is to occur. Also as the distance from a stop and generally average load and off of the passengers increases, the probability of a violence crime occuring tends to decrease, indicating that violent crime criminals might avoid places with higher traffic.

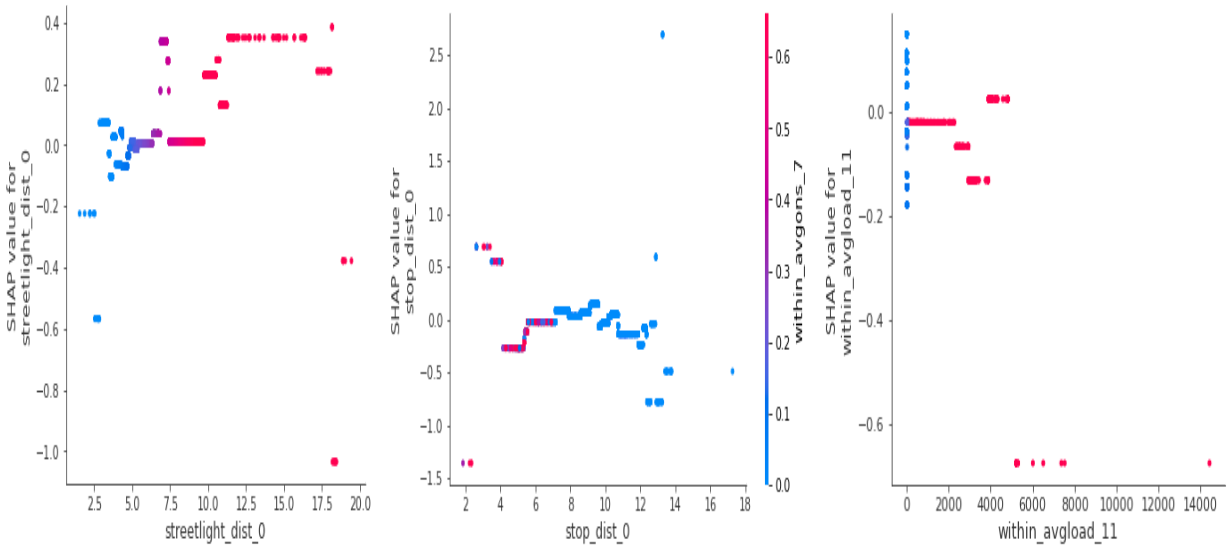


**Fig 15, 16, 17.** Traffic and Street Light SHAP dependence plots

We will also take advantage of Local Interpretable Model-agnostic Explanations (LIME), which perturbs the original data points, feeds them into the black box model, and then observes the corresponding outputs. The method then weighs those new data points as a function of their proximity to the original point and fits a surrogate model such as linear regression on the dataset with variations using those sample weights. Each original data point can then be explained with the newly trained explanation model. By looking at individual instances explained by LIME we could interpret better why a specific value for a data point is contributing to the final prediction. Take the below data point as an example:

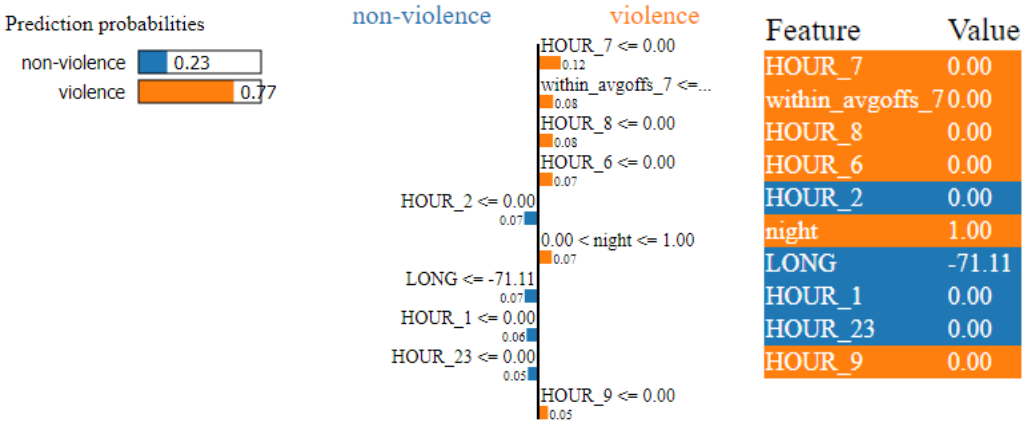


**Fig 18.** LIME explanation of why a incident is predicted as violent

Here in **Fig 18.** our model makes a correct prediction (violence crime=1), and we could see that the prediction again highly depends on time (non-morning traffic hours, night) and the traffic nearby (no passengers loaded off nearby), which agrees with our previous findings that violent crimes usually follow the pattern of taking place at night and with lower traffics (meaning fewer witnesses nearby).

In short, violent crimes are more likely to happen during the night and non-busy hours (non-morning rush hours), in less affluent neighborhoods with older houses and lower cpi, around places further away from street lights and traffic, and in weather conditions that are hot and less windy.

## Property(XGBoost)

For property crimes, we selected XGBoost as the best estimator like we did for violent crimes. Thus similar global and local feature analysis (3 types of tree importances, SHAP, LIME) were conducted (Check **Fig 31-36**) and we found that factors that contribute to property crimes overlap with violent and shooting crimes in that they all are more likely to occur during nights, in less affluent neighborhoods, around places far away from street lights and traffic. One factor that makes property crime occurrences stand out from other crime types is Householder_living_alone. From **Fig 19** we found that districts where more householders are living alone have a much higher chance of having property crimes, which may indicate that lone residents could be easier targets for criminals.

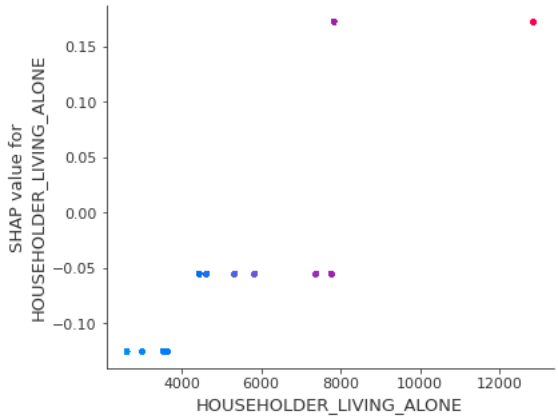


**Fig 19.** Householder_living_alone dependence plot vs shap values

## Other (Random Forest)

For other trivial crimes, we selected RF as the best estimator like we did for shooting. Similar global and local feature analysis (mean decrease in impurity, SHAP, LIME) were conducted and we found that factors that contribute to other crimes widely overlap with the 3 other types we have covered. Higher chances of other type crimes occurring are related to similar features like nights, lower educational level and lower affluence level of neighborhoods, and further distance from street lights and traffic.

**Fig 20 and 21** show that the difference is that top features in other trivial crimes are the most dominated by the distance from street lights and traffic. The LIME analysis example for other trivial variables is in **Fig 37**.

| Weight | Feature |
|---|---|
| 0.0505 ± 0.0683 | LAT |
| 0.0387 ± 0.0277 | streetlight_dist_0 |
| 0.0382 ± 0.0271 | stop_dist_0 |
| 0.0344 ± 0.0249 | within_avgons_13 |
| 0.0340 ± 0.0237 | within_avgload_13 |
| 0.0322 ± 0.0236 | LONG |
| 0.0318 ± 0.0224 | within_avgoffs_13 |
| 0.0303 ± 0.0373 | within_dist_streetlight_9 |
| 0.0270 ± 0.0245 | within_dist_stops_13 |
| 0.0262 ± 0.0164 | within_avgload_11 |
| 0.0256 ± 0.0163 | within_avgoffs_11 |
| 0.0251 ± 0.0263 | within_dist_streetlight_8 |
| 0.0248 ± 0.0163 | within_avgons_11 |
| 0.0248 ± 0.0120 | humidity |
| 0.0214 ± 0.0114 | FeelsLikeC |
| 0.0187 ± 0.0101 | windspeedKmph |
| 0.0175 ± 0.0703 | MALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT |
| 0.0173 ± 0.0549 | TOTAL_VALUE_STD |
| 0.0168 ± 0.0190 | within_dist_streetlight_7 |
| 0.0160 ± 0.0111 | edu_health_wage |

**Fig 20.** Mean decrease in GINI for RF (Other)

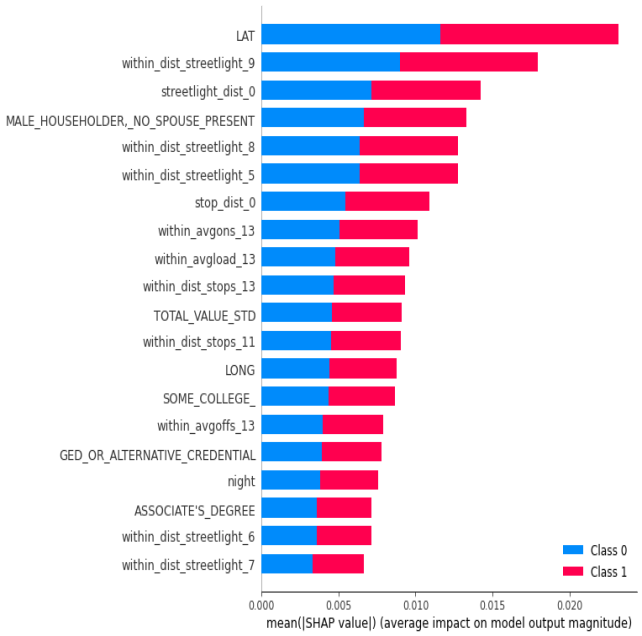


**Fig 21.** SHAP summary plot (Other)

# Conclusions

To sum up, after thorough investigation, we have found that albeit all having interesting distinctions, top contributing factors to different types of crimes in Boston overlap widely.
All the crimes are more likely to happen during the night and less busy hours, away from streetlight and traffic and in neighborhoods less educated and affluent.

For each crime specifically:

- Shooting crimes are more dependent on nearby traffic and street lights, education and household types. Higher education populations may generally have less motivation to commit crimes and places near school also benefit from the higher security level around campus, while households that are single may be more vulnerable to shooting incidents.
- Violent crimes are related a lot to time factors, and have a distinct relationship with weather. Higher temperatures may have bad effects on peoples' moods and trigger violent crimes while extreme weather like windy days with wind speed greater than 20mph may prevent people from going out and reduce chances of violent crimes.
- Property crimes are non-surprisingly more dependent on time and distances from streetlights and traffic, and also differentiate from the rest with respect to local residential structures. More residents living alone and renters may be identified as easier targets for property crimes.
- Other crimes are the most dependent on nearby crowd level (street lights and traffic).

Some features that we expected to have a relationship with crimes like snow level didn't show up amongst the top features at all. We suspect that the data points that have snow level>0cm only constitute very little part of the dataset and thus aren't able to explain most of the crime incidences.

# Limitations & Future Work

We are aware of several limitations of our research.

1. Our dataset is largely based on the crime reporting practices of the citizenry at Boston. We are also aware that

manual or system error could have been introduced in the data collection process, which would affect the data accuracy.
2. Selection bias is introduced by the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved. We admit that the selection of the dataset and variables unintentionally brings bais into our analysis.
3. Our analysis is constrained by the types of model we choose and the metrics we evaluate. We admit that there are more models and more metrics to take into consideration.
4. Our hyperparameter tuning process is constrained by the hyperparameter we choose and the selected values we choose to tune on. Due to the volume of our dataset, we were unable to apply more tuning values for various models, and we admit this puts further limitations on our modeling result.

Our analysis has a wide range of future directions that we encourage all scholars to explore.

1. Inclusion of more datasets and data points would strengthen our argument. For instance, cultural factors, recreational, and religious characteristics would be beneficial to better understand Boston's population. Data on the strength and the aggressiveness of a jurisdiction's law enforcement agency would help us better understand citizen's attitude towards crime, which is a huge contributor to crime rate.
2. Comparison of different cities, such as New York and San Francisco, would help us better understand the different crime composition in major cities, thus making people more informed of variables affecting crimes across geographic locations.
3. Applying causal and counterfactual analysis would help strengthen causal argument and enable evaluators to attribute cause and effect between interventions and outcomes.
4. Predictive analysis would help us identify future crime outcomes to reduce risks associated with crime. For instance, predictive policing involves using algorithms to analyze massive amounts of information in order to predict and help prevent potential future crimes. Place-based predictive policing, the most widely practiced method, typically uses pre-existing crime data to identify places and times that have a high risk of crime.

# Appendix A - Feature Engineering

## Neighborhood Data

The demographic data for Boston's neighborhoods ranges from 1950 to 2019, and we took the data from 2015 to 2019 only for time range consistency and used selected tables that are relevant to crime incidents: age, household type, race, group quarters population, nativity, geographic mobility, educational attainment, housing tenure, and poverty rates.

The crime incident data has a property called district code, which can be mapped to the neighborhoods using BPDNews, so we dropped crime incidents that do not have a valid district code and got the neighborhood name for the rest.

## Weather Data

The hourly weather data gathered using WorldWeatherOnline API are joined with the crime incident data on the timestamp. The sunrise and sunset time is a good indicator of when it is dark or not, so we created another binary predictor called night and assigned 1 to crime incidents if the timestamp is before sunrise or after sunset time. We used correlation matrix and pairwise correlation and selected the following predictors: total snow in cm, feels like temperature in Celsius, humidity, precipitation in mm, and wind speed in km per hour.

## Streetlight Data

To gain insight into the relationship between the streetlights and the crime location, we calculated the distance between streetlights and crime location in the same district. The distance is calculated by taking the log of the euclidean distance twice between the projected longitudes and latitudes of the crime location and streetlight location. The projection is converting from longitude, latitude to native map projection x,y coordinates. The Universal Transverse Mercator (UTM) is a map projection system for assigning coordinates to locations on the surface of the Earth.
Below are the additional features we added after feature engineering:

- **Streetlight_dist_0**: the normalized distance between the crime location and the closest streetlight
- **Within_dist_streetlight_{i}** where $i \in [4,9]$: the number of streetlights that are within i-unit of the normalized distance from the crime location. We chose the i's after looking at the distribution of street lights in different distances as shown in **Fig 22**. Very few crimes are within 3 unit normalized distance to the closest street lights. when the distance is below 3, fewer than 1000 cases on that. Starting from 4, the number of crimes increases sharply as the distance increases. Therefore, we choose the distance thresholds to be 4,5,6,7,8 and 9 as these thresholds all have different proportions of the distance layers.

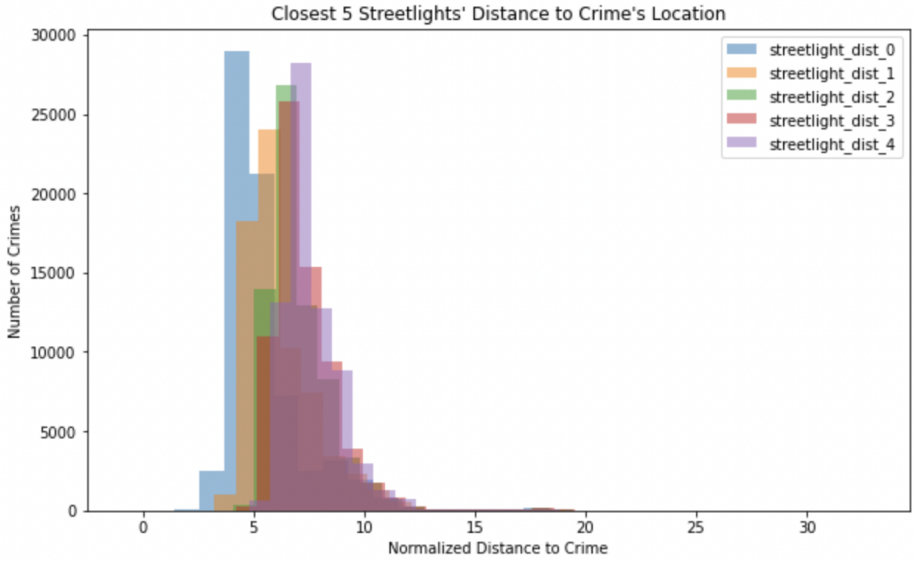


**Fig 22**. Closest 5 Streetlights' Distance to Crime's Location

## MBTA Station/Ridership Data

We did similar things as with street lights in terms of calculating distance. However, we chose different thresholds after looking at the distribution of station numbers in different distances.

According to **Fig 23**, we can see that very few crimes are within 5 units of normalized distance to the closest stations/stops. When the distance is below 5, fewer than 3000 cases on that. Starting from about 7, the number of crimes changes sharply as the distance increases by every 2 units of normalized distance. Based on this information, we choose the threshold of 7,9,11, and 13.

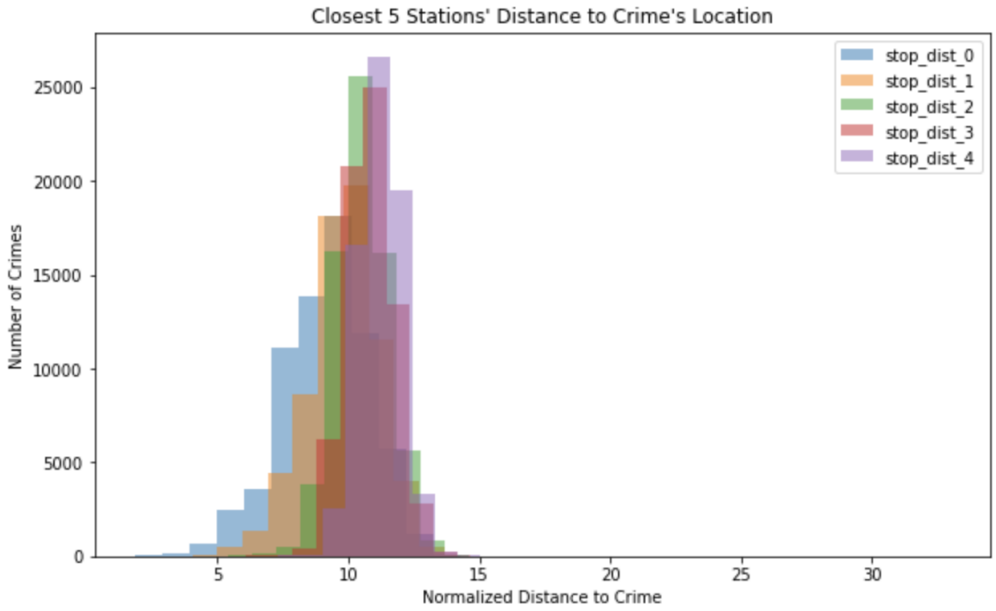


**Fig 23**. Closest 5 MBTA Stations' Distance to Crime's Location

Based on this information, we generated 5 kinds of features:

- **Stop_dist_0**: The normalized distance from the crime location to the closest MBTA stop/station.

- **Within_dist_stops_{i}** where i = {7,9,11,13}: the number of MBTA stops/stations that are within the normalized distance of 7.
- **Within_avgload_{i}** , **Within_avgons_{i}, and Within_avgoffs_{i}** where i = {7,9,11,13}: the average loadings, ons, and offs of the MBTA stops/stations that are within the normalized distance of 7 during the crime occured time period.

We also plotted the correlation heatmap (see **Appendix B Fig 26**).and did not find pairs of features to have correlation greater than 0.9.

### Economic Data

Based on heatmap in **Appendix B Fig 24**, we found several high correlation (>0.9) between pairs of economic indicators. To reduce the multicollinearity problem, we dropped features 'professional_business_wage', 'other_wage','mining_logging_construction_wage', 'total_nonfarm_wage', and 'cpiw' because they have been highly correlated with other data.

### Property Data

Due to a lack of data dictionary from the data source, we extracted the useful quantitative variables as much as possible. Specifically, 'ZIPCODE', 'LIVING_AREA', 'LAND_VALUE', 'BLDG_VALUE', 'TOTAL_VALUE', 'GROSS_TAX', 'YR_BUILT', 'YR_REMODEL'. After analyzing the correlations, zip-code, total value of a specific property, and the last time the property was built/remodeled are left. We further categorize the last time the property was built/remodeled into different timeframes: before 1800, 1800-1850, 1850-1900, 1900-1950, 1950-2000, and after 2000. For each of the zipcodes, we calculated how much percentage of the properties falls into the timeframes, and included those into the later modeling analysis. In addition, we calculated the mean total property value and sum of total property value for each zip-code. Finally, based on the latitude and longitude of the crime data, we mapped the zip-code, and merged the two datasets based on zip-code. For 151 rows, we did not map out the zip-code because the matching process failed. This could be due to cyber crimes that have some false or missing latitude and longitude. Later in the analysis, these 151 rows are deleted.

## Appendix B - Correlation Heatmap

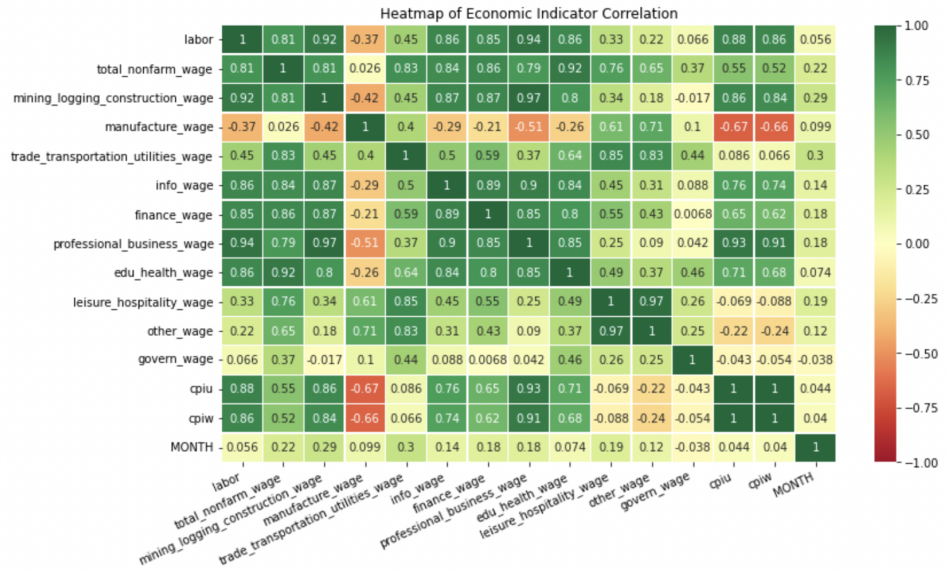


**Fig 24**. Heatmap of Economic Indicators before Dropping Features

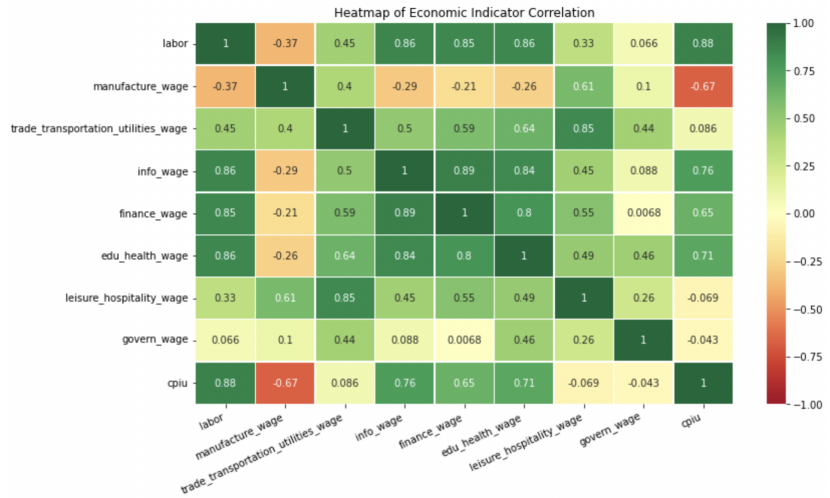


**Fig 25**. Heatmap of Economic Indicators after Dropping Highly Correlated Features

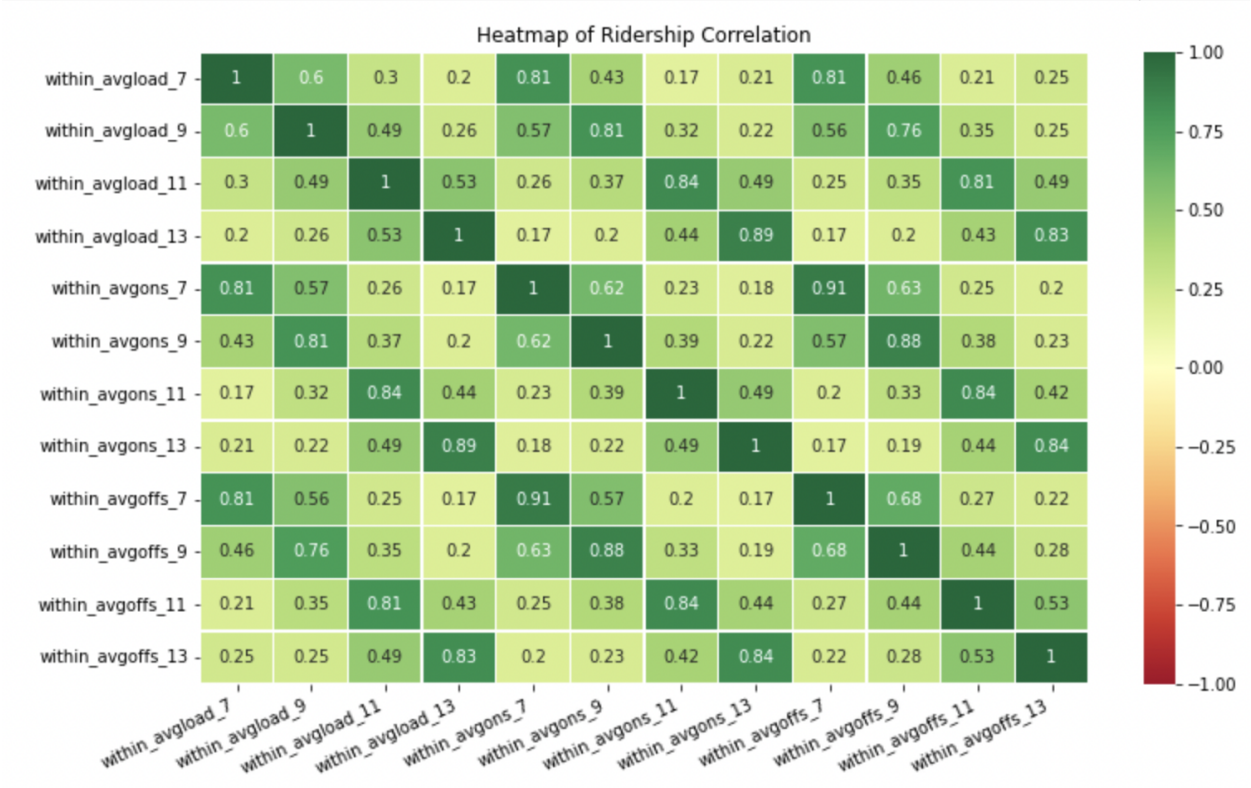


**Fig 26**. Heatmap of Average Ridership Loading, Ons, and Offs Correlation

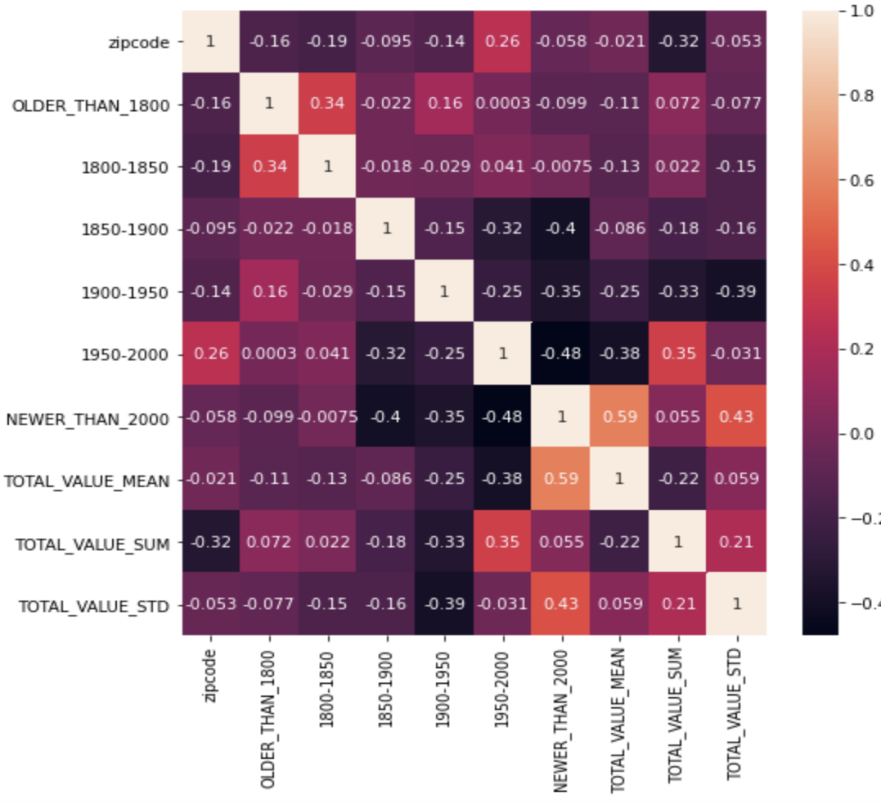


**Fig 27**. Heatmap of Property Assessment Data Correlation

# Appendix C - Feature Analysis

## Shooting

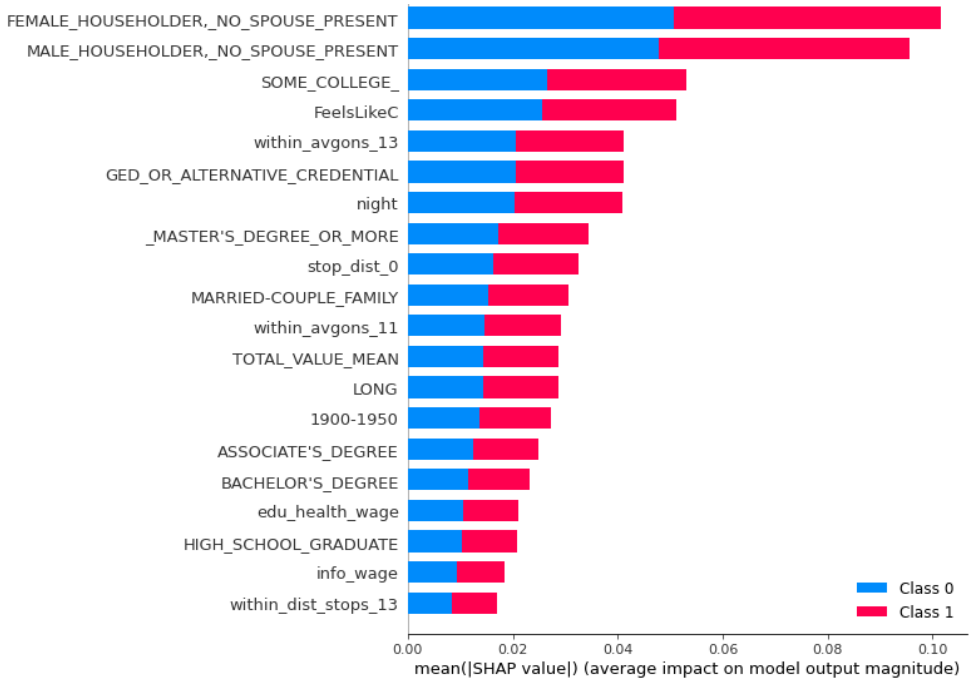


**Fig 28.** Summary barplot of the SHAP values for shooting

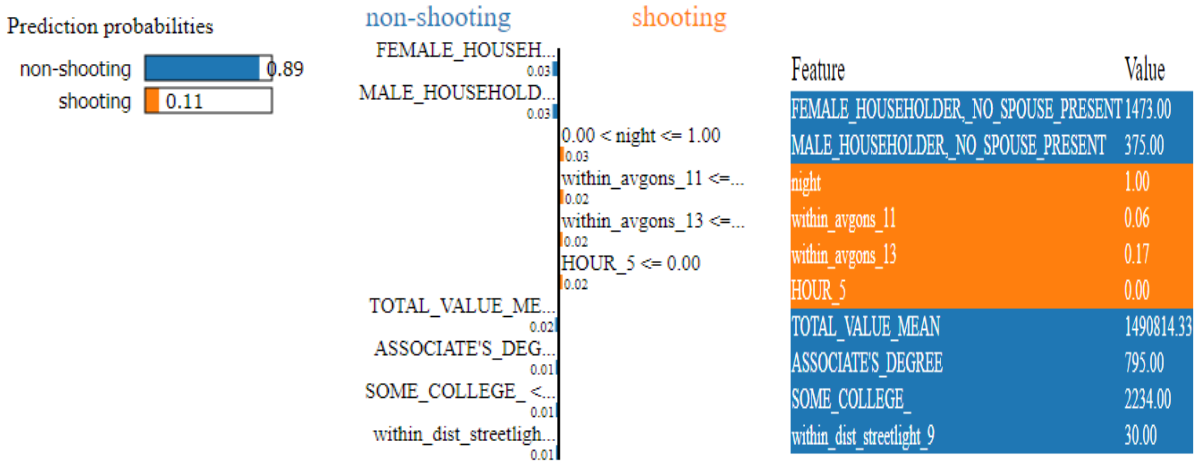


**Fig 29. LIME Example for Shooting**

## Property Crimes

| Weight | Feature |
|---|---|
| 0.1181 | within_dist_stops_7 |
| 0.1181 | TOTAL_VALUE_MEAN |
| 0.1162 | within_dist_streetlight_5 |
| 0.1004 | 1800-1850 |
| 0.0712 | HOUSEHOLDER_LIVING_ALONE |
| 0.0366 | MEDIAN_AGE |
| 0.0334 | DAY_OF_WEEK_Saturday |
| 0.0333 | 1900-1950 |
| 0.0312 | manufacture_wage |
| 0.0305 | labor |
| 0.0188 | edu_health_wage |
| 0.0172 | TOTAL_POPULATION_x |
| 0.0153 | DAY_OF_WEEK_Sunday |
| 0.0114 | 1950-2000 |
| 0.0100 | within_dist_streetlight_8 |
| 0.0096 | LAT |
| 0.0095 | POVERTY_RATE |
| 0.0093 | HOUR_23 |
| 0.0091 | HOUR_9 |
| 0.0087 | HOUR_10 |

| Weight | Feature |
|---|---|
| 0.0624 ± 0.1866 | LAT |
| 0.0419 ± 0.1804 | FEMALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT |
| 0.0361 ± 0.1744 | MALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT |
| 0.0348 ± 0.0923 | LONG |
| 0.0344 ± 0.1631 | TOTAL_VALUE_MEAN |
| 0.0320 ± 0.0761 | within_avgons_13 |
| 0.0319 ± 0.1520 | _MASTER'S_DEGREE_OR_MORE |
| 0.0316 ± 0.0904 | within_avgons_11 |
| 0.0284 ± 0.1293 | MARRIED-COUPLE_FAMILY |
| 0.0270 ± 0.0840 | within_avgons_9 |
| 0.0257 ± 0.0754 | within_avgload_11 |
| 0.0255 ± 0.0824 | night |
| 0.0230 ± 0.1197 | BACHELOR'S_DEGREE |
| 0.0226 ± 0.0641 | stop_dist_0 |
| 0.0214 ± 0.0606 | within_dist_streetlight_8 |
| 0.0210 ± 0.0533 | within_avgload_13 |
| 0.0199 ± 0.0549 | within_avgoffs_13 |
| 0.0198 ± 0.0620 | within_dist_streetlight_9 |
| 0.0179 ± 0.1119 | GED_OR_ALTERNATIVE_CREDENTIAL |
| 0.0175 ± 0.1153 | ASSOCIATE'S_DEGREE |

| Weight | Feature |
|---|---|
| 0.0134 | TOTAL_VALUE_MEAN |
| 0.0133 | 1800-1850 |
| 0.0132 | MEDIAN_AGE |
| 0.0132 | DAY_OF_WEEK_Saturday |
| 0.0132 | manufacture_wage |
| 0.0132 | 1900-1950 |
| 0.0132 | HOUSEHOLDER_LIVING_ALONE |
| 0.0132 | within_dist_stops_7 |
| 0.0131 | edu_health_wage |
| 0.0131 | TOTAL_POPULATION_x |
| 0.0131 | within_dist_streetlight_5 |
| 0.0131 | POVERTY_RATE |
| 0.0131 | HOUR_9 |
| 0.0131 | HOUR_10 |
| 0.0131 | HOUR_4 |
| 0.0131 | HOUR_19 |
| 0.0131 | HOUR_23 |
| 0.0131 | HOUR_13 |
| 0.0131 | DAY_OF_WEEK_Thursday |
| 0.0131 | HOUR_7 |

**Fig 30,31,32.** From left to right: XGBoost Gain Cover, Weight for property crimes

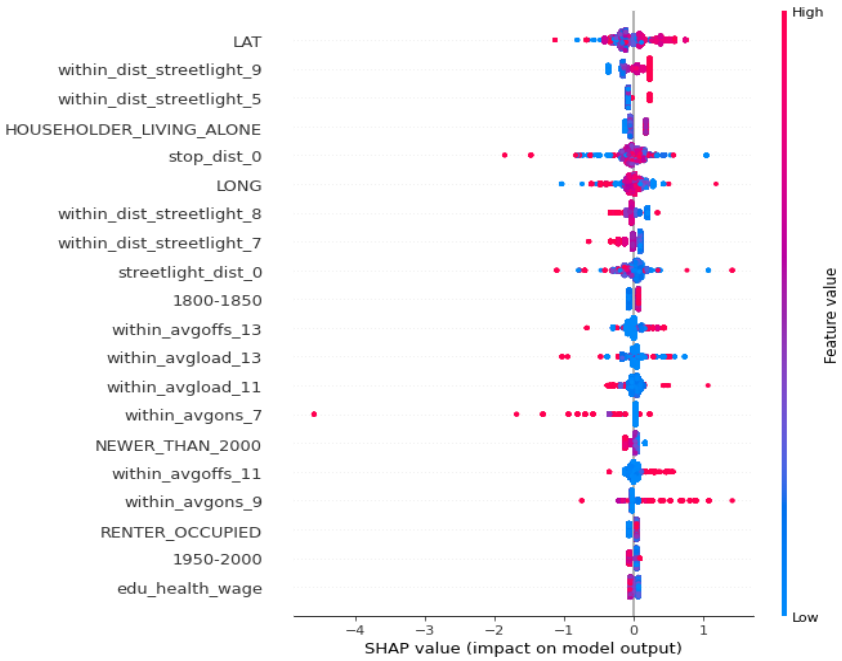


**Fig 33.** Shap summary plots for property crimes

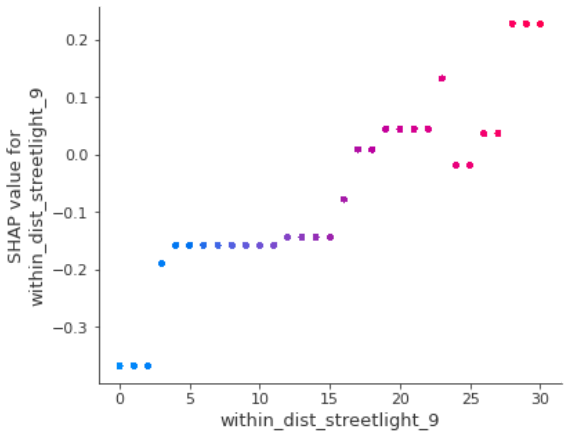


**Fig 34.** Shap dependence plots for within_dist_streetlight_9

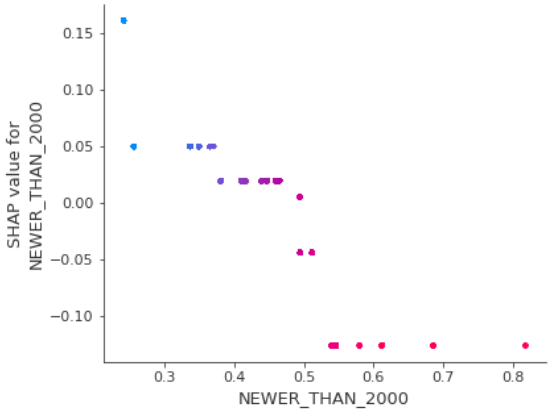


**Fig 35.** Shap dependence plots for NEWER_THAN_2000

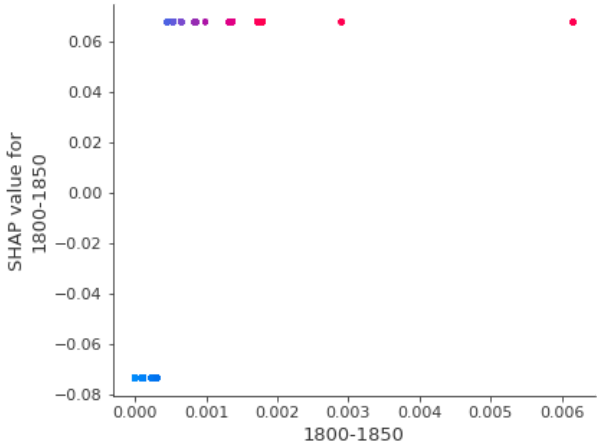


**Fig 36.** Shap dependence plots for 1800-1850

## Other Crimes

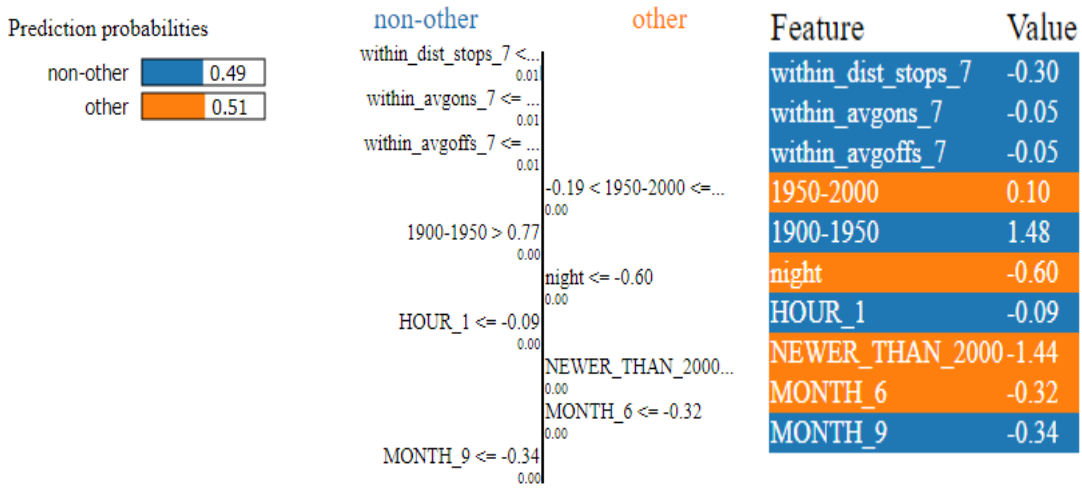


**Fig 37.** LIME Explanation of Why a instance is classified as Other Trivial Crimes

# Appendix D - Data Dictionary

Please refer to this link for a detailed data source for each variable.

| data type | variable name | description |
|---|---|---|
| **crime type (Target Value)** | SHOOTING | whether the crime is shooting or not |
| | CRIME_TYPE_VIOLENT | whether the crime belongs to violence crime or not |
| | CRIME_TYPE_PROPERTY | whether the crime belongs to property crime or not |
| | CRIME_TYPE_OTHER | whether the crime belongs to other non trivial crime or not (if it is not violence or property, then the crime falls into this category) |
| **property** | OLDER_THAN_1800 | The percentage of properties that was last remodeled/built before 1800 in this zipcode. |
| | 1800-1850 | The percentage of properties that was last remodeled/built between 1800 and 1850 in this zipcode. |
| | 1850-1900 | The percentage of properties that was last remodeled/built between 1850 and 1900 in this zipcode. |
| | 1900-1950 | The percentage of properties that was last remodeled/built between 1900 and 1950 in this zipcode. |
| | 1950-2000 | The percentage of properties that was last remodeled/built between 1950 and 2000 in this zipcode. |
| | NEWER_THAN_2000 | The precentage of properties that was last remodeled/built after 2000 in this zipcode. |
| | TOTAL_VALUE_MEAN | The mean value of the total value of each of the property in this zipcode |
| | TOTAL_VALUE_SUM | The total value of the total value of each of the property in this zipcode |
| | TOTAL_VALUE_STD | The standard deviation of the total value of each of the property in this zipcode |
| **streetlight** | streetlight_dist_0 | The distance from the crime location to the closest streetlight. The distance is calculated by taking the log of the euclidean distance between the projected longitudes and latitudes of the crime location and streetlight location. The projection is converting from longitude, latitude to native map projection x,y coordinates. The Universal Transverse Mercator (UTM) is a map projection system for assigning coordinates to locations on the surface of the Earth. |
| | within_dist_streetlight_4 | the number of streetlight that is within the distance of 4. Refer to project report for the distance definition |
| | within_dist_streetlight_5 | the number of streetlight that is within the distance of 5 |
| | within_dist_streetlight_6 | the number of streetlight that is within the distance of 6. Refer to project report for the distance definition |

| | | |
|---|---|---|
| | within_dist_streetlight_7 | the number of streetlight that is within the distance of 7. Refer to project report for the distance definition |
| | within_dist_streetlight_8 | the number of streetlight that is within the distance of 8. Refer to project report for the distance definition |
| | within_dist_streetlight_9 | the number of streetlight that is within the distance of 9. Refer to project report for the distance definition |
| **MBTA** | stop_dist_0 | The distance from the crime location to the closest MBTA stop/station. |
| | within_dist_stops_7 | the number of MBTA stops/stations that is within the distance of 7. Refer to project report for the distance definition |
| | within_dist_stops_9 | the number of MBTA stops/stations that is within the distance of 9. Refer to project report for the distance definition |
| | within_dist_stops_11 | the number of MBTA stops/stations that is within the distance of 11. Refer to project report for the distance definition |
| | within_dist_stops_13 | the number of MBTA stops/stations that is within the distance of 13. Refer to project report for the distance definition |
| | within_avgload_7 | the average loading of MBTA stops/stations that is within the distance of 7 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgload_9 | the average loading of MBTA stops/stations that is within the distance of 9 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgload_11 | the average loading of MBTA stops/stations that is within the distance of 11 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgload_13 | the average loading of MBTA stops/stations that is within the distance of 13 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgons_7 | the average number of passenger-ons of MBTA stops/stations that is within the distance of 7 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgons_9 | the average number of passenger-ons of MBTA stops/stations that is within the distance of 9 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgons_11 | the average number of passenger-ons of MBTA stops/stations that is within the distance of 11 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgons_13 | the average number of passenger-ons of MBTA stops/stations that is within the distance of 13 during the corresponding time period. Refer to project report for the distance definition |

| | | |
|---|---|---|
| | within_avgoffs_7 | the average number of passenger-offs of MBTA stops/stations that is within the distance of 7 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgoffs_9 | the average number of passenger-offs of MBTA stops/stations that is within the distance of 9 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgoffs_11 | the average number of passenger-offs of MBTA stops/stations that is within the distance of 11 during the corresponding time period. Refer to project report for the distance definition |
| | within_avgoffs_13 | the average number of passenger-offs of MBTA stops/stations that is within the distance of 13 during the corresponding time period. Refer to project report for the distance definition |
| **economics** | labor | the number of labor force in Boston-Cambridge-Quincy area |
| | manufacture_wage | manufacturing wage |
| | trade_transportation_utilities_wage | trade, transportation and utilities wage |
| | info_wage | information wage |
| | finance_wage | financial activities wage |
| | edu_health_wage | education and health services wage |
| | leisure_hospitality_wage | leisure and hospitality wage |
| | govern_wage | government wage |
| | cpiu | urban consumer price index |
| **location** | zipcode | zipcode of the crime location |
| | police district | police district number of the crime location |
| | LAT | latitude of crime location |
| | LONG | longitude of crime location |
| **time of day** | night | 0 if crime occured date is at day else night |
| | MONTH_2 | Calendar month Feburary |
| | MONTH_3 | Calendar month March |
| | MONTH_4 | Calendar month April |
| | MONTH_5 | Calendar month May |
| | MONTH_6 | Calendar month June |
| | MONTH_7 | Calendar month July |
| | MONTH_8 | Calendar month August |
| | MONTH_9 | Calendar month September |
| | MONTH_10 | Calendar month October |

| | | |
|---|---|---|
| | MONTH_11 | Calendar month November |
| | MONTH_12 | Calendar month December |
| | DAY_OF_WEEK_Monday | Calendar week Monday |
| | DAY_OF_WEEK_Saturday | Calendar week Saturday |
| | DAY_OF_WEEK_Sunday | Calendar week Sunday |
| | DAY_OF_WEEK_Thursday | Calendar week Thursday |
| | DAY_OF_WEEK_Tuesday | Calendar week Tuesday |
| | DAY_OF_WEEK_Wednesday | Calendar week Wednesday |
| | HOUR_1 | crime happened between 01:00 and 02:00 |
| | HOUR_2 | crime happened between 02:00 and 03:00 |
| | HOUR_3 | crime happened between 03:00 and 04:00 |
| | HOUR_4 | crime happened between 04:00 and 05:00 |
| | HOUR_5 | crime happened between 05:00 and 06:00 |
| | HOUR_6 | crime happened between 06:00 and 07:00 |
| | HOUR_7 | crime happened between 07:00 and 08:00 |
| | HOUR_8 | crime happened between 08:00 and 09:00 |
| | HOUR_9 | crime happened between 09:00 and 10:00 |
| | HOUR_10 | crime happened between 10:00 and 11:00 |
| | HOUR_11 | crime happened between 11:00 and 12:00 |
| | HOUR_12 | crime happened between 12:00 and 13:00 |
| | HOUR_13 | crime happened between 13:00 and 14:00 |
| | HOUR_14 | crime happened between 14:00 and 15:00 |
| | HOUR_15 | crime happened between 15:00 and 16:00 |
| | HOUR_16 | crime happened between 16:00 and 17:00 |
| | HOUR_17 | crime happened between 17:00 and 18:00 |
| | HOUR_18 | crime happened between 18:00 and 19:00 |
| | HOUR_19 | crime happened between 19:00 and 20:00 |
| | HOUR_20 | crime happened between 20:00 and 21:00 |
| | HOUR_21 | crime happened between 21:00 and 22:00 |
| | HOUR_22 | crime happened between 22:00 and 23:00 |

| | | |
|---|---|---|
| | HOUR_23 | crime happened between 23:00 and 00:00 |
| **crime** | CRIME_TYPE_OTHER | crime incident categorized as "other" |
| | CRIME_TYPE_PROPERTY | crime incident categorized as "property" |
| | CRIME_TYPE_VIOLENT | crime incident categorized as "violent" |
| | SHOOTING | whether shooting happened in the crim incident |
| **neighborhood** | TOTAL_POPULATION_x | total population in the neighborhood |
| | MEDIAN_AGE | median age in the neighborhood |
| | MARRIED-COUPLE_FAMILY | number of married couple family in the neighborhood |
| | MALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT | number of male householders in the neighborhood |
| | FEMALE_HOUSEHOLDER,_NO_SPOUSE_PRESENT | number of female householders in the neighborhood |
| | HOUSEHOLDER_LIVING_ALONE | number of householders that are living alone in the neighborhood |
| | POPULATION_IN_GROUP_QUARTERS | number of people live in group quarters rather than in households in the neighborhood |
| | LESS_THAN_HIGH_SCHOOL_ | number of people holding less than high school degree in the neighborhood |
| | HIGH_SCHOOL_GRADUATE | number of people holding high school degree in the neighborhood |
| | GED_OR_ALTERNATIVE_CREDENTIAL | number of people holding GED or alternative degree in the neighborhood |
| | SOME_COLLEGE_ | number of people holding some college degree in the neighborhood |
| | ASSOCIATE'S_DEGREE | number of people holding associate's degree in the neighborhood |
| | BACHELOR'S_DEGREE | number of people holding bachelor's degree in the neighborhood |
| | _MASTER'S_DEGREE_OR_MORE | number of people holding master's degree or higher in the neighborhood |
| | OWNER_OCCUPIED | number of houses occupied by owners in the neighborhood |
| | RENTER_OCCUPIED | number of houses occupied by renters in the neighborhood |

| | | |
|---|---|---|
| | POVERTY_RATE | poverty rate in the neighborhood |
| **weather** | totalSnow_cm | total amount of snow in centimeter |
| | FeelsLikeC | "feels like" temperature |
| | humidity | humidity in a day |
| | precipMM | precipitation in millimeter |
| | windspeedKmph | windspeed in kilometer per hour |

# Appendix E - Evaluation Metrics and Models Explained

## Evaluation Metrics

There are in total 5 metrics we measure to select the best model. But we prioritize F-1 score and ROC-AUC score because both metrics can deal with imbalance class evaluation. However, if there is any discrepancy between two scores, we also look at the recall score because we value the accuracy of the model in predicting positive cases.

1. ROC-AUC score: its full name is Area Under the Receiver Operating Characteristic Curve. It ranges from 0 to 1. The closer to 1 the better the model is at distinguishing crimes. If it is 0.5, it means the model is no different from the naive baseline model we built.
2. F1 score: it is the harmonic mean of precision and recall that ranges from 0 to 1. This balances precision and recall. The closer it is to 0 means either the recall or the precision is too small. We usually use this for imbalance class evaluation.
3. Recall score: It equals true positive / (true positive + false negative). It is basically evaluating how well the model can get all the positive cases.
4. Confusion matrix: we store the true positive (tp) , false positive (fp), true negative (tn), and false negative (fn).
5. Accuracy Score: (tp + tn) / (tp+tn+fp+fn). The percentage of data being correctly classified.

## Models Used

### Logistic Regression

Logistic regression models the probabilities of classification problems with two possible outcomes. Logistic classification is relatively easy to implement, interpret and easier to train, while its major

limitation is the assumption of the linearity between the dependent variables and independent variables. We mainly focused on tuning the hyperparameter C, which is the inverse of regularization strength. Lowering C would strengthen the Lambda regulator and, and a high value of C tells the model to give high weight to the training data.

## Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Random Forests implicitly perform feature selection and generate uncorrelated decision trees, but can be computationally intensive for large datasets.

## AdaBoost

Adaptive Boosting (AdaBoost) we used in the project is an algorithm used with a group of decision trees as weak learners and assigns higher weights to data that are wrongly classified. It is less likely to overfit since the model can gradually become a strong learner as long as each weak learner performs better than random guess. The aggregate performance generated from weak learners also brings some disadvantages: if data is too noisy or there are lots of outliers, the performance of weak learners would only reflect the data used to train the model, which may lead to overfitting.

## XGBoost

XGBoost, named eXtreme Gradient Boosting, is an implementation of gradient boosting decision trees that is popularly used in Kaggle. It is designed with parallel trees to have good performance with higher speed.

## LightGBM

LightGBM stands for Light Gradient Boosting Machine. It takes many of XGBoost's advantages like parallel training, bagging and regularization but constructs trees differently. Instead of growing trees level-wise it grows leaf-wise and picks leaves that yield the largest decrease in loss.

# Reference - Data Source Link

We have drawn data from 9 data sources. Our data features include:

1. **Boston Crime data**: the crime data are directly taken from boston.gov and used to link between different outside data sources. We get the unique crime offense code dictionaries from data before 2018 then fill in offense code types for the rest of the new data starting from 2020. Most importantly, we divide all types of crimes into 3 major categories as mentioned in the introduction: **violence crimes**, **property-related crimes**, and **other more trivial crimes**.

2. **Neighborhood data**: the neighborhood data we are using are also from boston.gov. We map the demographic information to all police districts in the crime dataset and investigate whether there exists a relationship between demographic factors including Age, Household Type, Race, Group Quarters Population, Nativity, Geographic Mobility, Educational Attainment, School Enrollment, Means of Commuting, Per Capita Income, Household Income, Family Income, Housing Tenure, Bedrooms, Vacancy Rates, Vehicles per Household, Poverty Rates and crime incidence.

3. **Weather data**: given the fickle nature of Boston's weather, its relationship with crimes is of particular interest to us. Past studies have largely found that crime levels tend to increase with temperature and obtained mixed results regarding how precipitation affects crime results on crime. Using short time periods of weather data may make it difficult to distinguish the effects of ambient weather conditions from those caused by short-term weather fluctuations. We plan to use hourly data first and then compare results using different time intervals. We also want to lump crimes together into violent and property categories and study the effects of weather on both broad and individual crime types. The wwo-hist package is used here to retrieve and parse hourly historical weather data from World Weather Online into pandas DataFrame and CSV files.

4. **Streetlight data**: we propose that streetlight data can be helpful in understanding the relationship between crime and distance between the crime occurrence location and the streetlights. We get the streetlight data from boston.gov which contains the longitude and latitude data for 40,000 streetlights, 2,800 gas lights, and 1,500 fire alarm lights in Boston.

5. **MBTA ridership and stops data**: we collect the ridership and stops/stations data from the MBTA website. The ridership data contains fall season average bus ridership, rapid-transit ridership, and commuter rail ridership data. The latest data is 2019 but we assume the flow in each station/stop would be relatively the same throughout these years. We use the police_districts data to find the corresponding district using the longitude and latitude data of the stations/stops.

1. Data Source for stops/stations:
    a. Bus stops locations: https://mbta-massdot.opendata.arcgis.com/datasets/bus-stops/explore?location=42.186700%2C-71.042550%2C9.17&showTable=true
    b. Rapid Transit stations/stops: https://mbta-massdot.opendata.arcgis.com/datasets/rapid-transit-stops/explore?location=42.186700%2C-71.042550%2C9.17
    c. Commuter rails stations: https://mbta-massdot.opendata.arcgis.com/datasets/commuter-rail-stations/explore?location=42.186700%2C-71.042550%2C9.17
2. Data Source for Ridership:

a. Bus stops ridership: https://mbta-massdot.opendata.arcgis.com/datasets/mbta-bus-ridership-by-time-period-season-route-line-and-stop/explore
b. Rapid Transit stations/stops ridership: https://mbta-massdot.opendata.arcgis.com/datasets/mbta-rail-ridership-by-time-period-season-route-line-and-stop/explore
c. Commuter rails stations ridership: https://mbta-massdot.opendata.arcgis.com/datasets/mbta-commuter-rail-ridership-by-trip-season-route-line-and-stop/explore

6. **Police districts in Boston**: there are in total 12 districts in Boston specified in this dataset. The district data in Geojson format was collected from boston.gov. This is the same district code that is used in the crime data district column. We retrieve this information to match the streetlight location in longitude and latitude to district in order to reduce computation in distance calculations.

7. **Zip codes in Boston**: The zip code data in Geojson format was collected from boston.gov. This is the same district code that is used in the crime data district column. We retrieve this information to match the crime location in longitude and latitude to zip codes with the property assessment data which only has zip codes quantified location data of each property.

8. **Economic data**: we collected the economic data from January 2011 to September 2021 in the U.S. Bureau of Labor Statistics. It contains the labor force data, consumer price index, and nonfarm wage and salary employment in the Boston-Cambridge-Nashua, MA-NH metropolitan area.

9. **Property assessment data**: we are also considering deriving property assessment information from the Boston government page. From this dataset, we can retrieve property-related information, such as the assessed value for property and the age of the property.