Analysis of Boston Crime Reports

Alex Moore
College of Behavioral, Health, and Social Sciences
Clemson University
Clemson, South Carolina
afm2@g.clemson.edu

Reetyan Das
College of Engineering, Computing, and Applied Sciences
Clemson University
Clemson, South Carolina
reetayd@g.clemson.edu

Abubeker Abdullahi
College of Engineering, Computing, and Applied Sciences
Clemson University
Clemson, South Carolina
abubeka@g.clemson.edu

Abstract—This Report contains our Analysis for the Boston city Crime report which was provided by the police Department in Boston City from year 2015 to 2018. The crime Dataset contains crimes along with their offense codes which is reported to the officers with occurrence of time and date. We have thought about measure and predict the level of violation along with the crime occurrence. After Finding the primary analysis our goal was to make a prediction model to predict the level of violation. For better prediction results we also worked on Feature Engineering and mapped them with Census Data. Most of our work progress is done with a start of EDA (Exploratory Data Analysis) and in a story telling way.

Dataset and Preprocessing

We have worked on the revised Data set which has less number of columns than the original Crime Report. The Crime report ranges from June 14th, 2015 to Sept 3rd, 2018 incidents. The revised Dataset contains 17 columns with an entry of 319073 variables. We have first preprocessed the Data to remove the null values with the mean values of the columns. We wanted to measure the level of violation. As there were too many dimensions (67 types) for the "Offense Code", it will be difficult if we would predict the "Offense Code" by building classification model. In this case, for a lower dimension we wanted to break the Offense code occurred in terms of 'Frequency' and measure it between some ranges. There are four level of violations. Those offense code which occurred more than 5000 times they are referred as Strong level of violation ('sv'), those who are less than 500 are referred as non-violation ('nv').

Var1	Freq	violation_level
Aggravated Assault	7807	strong
Aircraft	36	noviolation
Arson	94	noviolation
Assembly or Gathering Violations	955	low
Auto Theft Recovery	1051	medium
Auto Theft	4851	medium

After that we merged that Violation-level column with out main DataFrame and marked it as our label of the Dataset.

STREET	Lat	Long	Location	violation_level
LINCOLN ST	42.35779	-71.13937	(42.35779134, -71.13937053)	strong
HECLA ST	42.30682	-71.06030	(42.30682138, -71.06030035)	strong
CAZENOVE ST	42.34659	-71.07243	(42.34658879, -71.07242943)	strong
NEWCOMB ST	42.33418	-71.07866	(42.33418175, -71.07866441)	strong
DELHI ST	42.27537	-71.09036	(42.27536542, -71.09036101)	strong
TALBOT AVE	42.29020	-71.07159	(42.29019621, -71.07159012)	strong
NORMANDY ST	42.30607	-71.08273	(42.30607218, -71.08273260)	medium
LAWN ST	42.32702	-71.10555	(42.32701648, -71.10555088)	strong
MASSACHUSETTS AVE	42.33152	-71.07085	(42.33152148, -71.07085307)	medium

Influences

(1) Which type of crime is the most common?

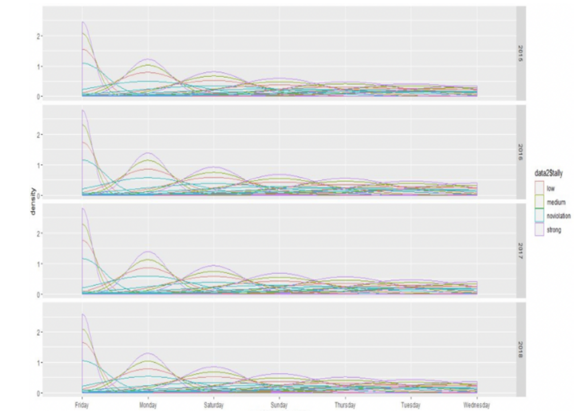
(2) When is the most freqent time of day strong violation occurences?

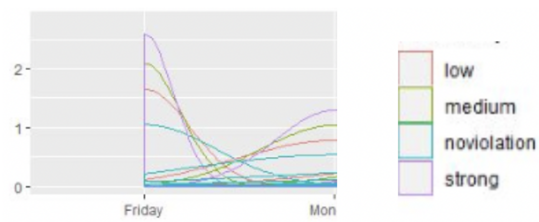
(3) What areas of Boston are most heavily impacted?

(4) What socioeconomic factors are related with the most common type of crime?

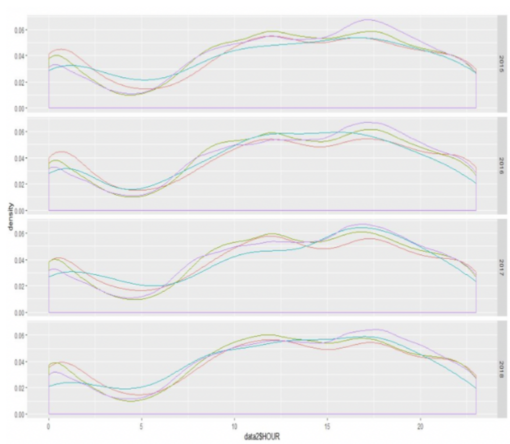
Primary Investigation

We wanted to see In 4 years which day has impact with high level of frequency of all types of Crimes. From the below ggplot it is clear that the strong level of violation start with a very high peak in Friday. Even, surprisingly the other non-strong crimes also started with the very high peak

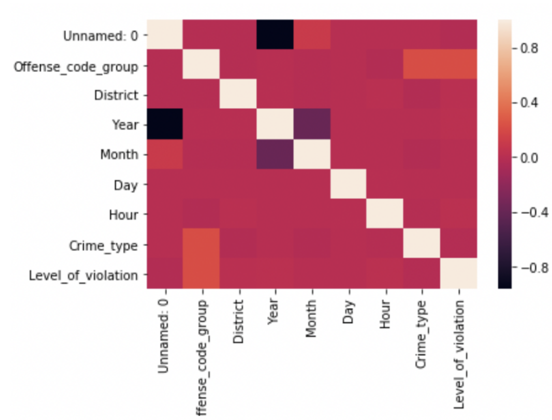




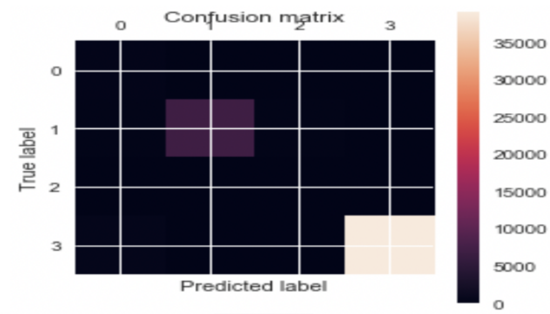
So When looked closely at the Friday Data, Its clearly visible that for all past 4 years the occurrence of all types of violation is very high. So we Seperated the Friday Data from the Data Frame and wanted to know the particular time of occurrence in this case. The Behaviour surprisingly also more or less same now, where we can see the occurrence of strong violation here happened at late after noon which is 17:30 PM which gave us a very important explanation



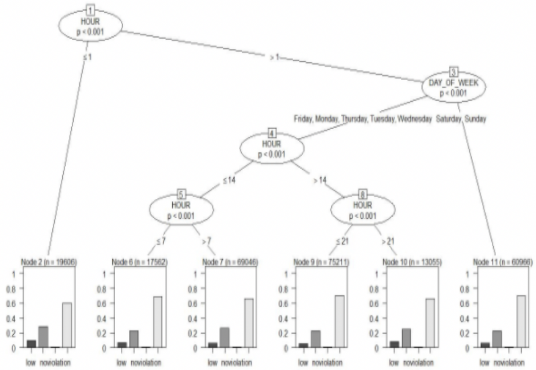
So, we wanted to know which are or could be the good predictors in this case. So we plotted a correlation matrix across our label. Below the matrix shows that “Offense Code group”, ”District” have the high correlation value from the color bar.



Next Our step was to make a prediction on our label “violation_level”. We chose “Offense_code_group”, ”District”, “Crime_type” as our predictor ans “level_of_violation” as our label. We made a decision tree which is balanced and with a high accuracy of 98.2%. “District” and “Offense_code” needed to be encoded to 0 and 1 by encoding process.



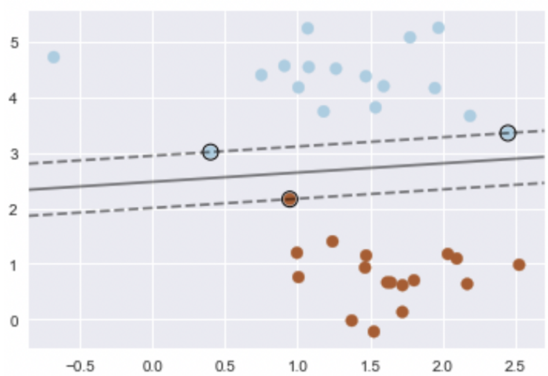
We made a good decision tree model but we wanted to see a more generalized Decision rule which explains more about the weekdays and weekends Decision Rule. Below the picture shows that there is a clear decision rule which explains clearly that even though Friday has the most occurrence of strong violation of crimes, it belongs to weekdays group which has a lower probability than the weekends crime rule. In the Week-Days crime rule the probability changes with the crime time before 2:00 (pm) and after (2:00) pm.

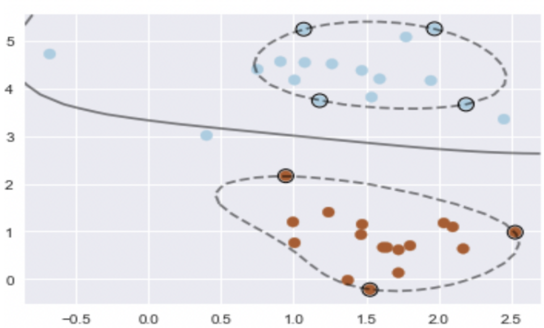


The accuracy was also decent then ($R^2 = 18.1\%$).

Feature Clustering

After the Prediction model we knew that those used features are very important an interesting and wanted to make them clustered in the basis of the level of violation. We applied support vector machine to find out the widest distance between two clusters. We applied linear kernel when we introduced "District" as a feature.

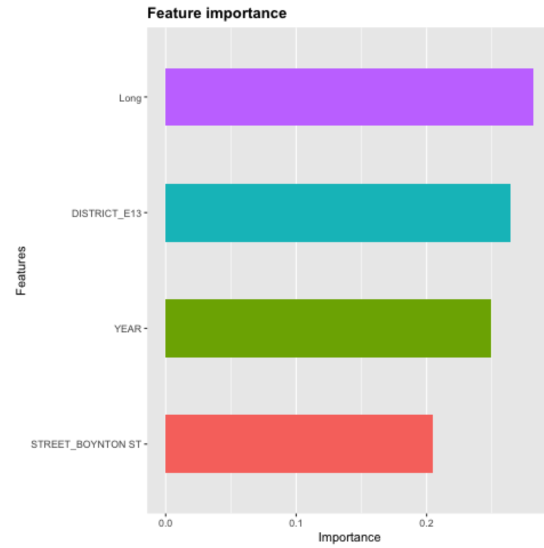




Analyses of Social Covariates

The Importance of Spatial Features

An Extreme Gradient Boosted Regression Tree model was used to predict whether a given police report was for the most frequent crime, Larceny. The model was trained on a randomly sampled subset of the data (n = 32489) and had a prediction accuracy of 91.7% when predicting occurrences of Larceny on the test data (n = 13924).



The above features above reflect their overall importance (i.e., Gain) on model predictions. Evidently, location-related variables like Latitude, specific districts and streets were most informative, with Time being the lone temporal variable related to Larceny.

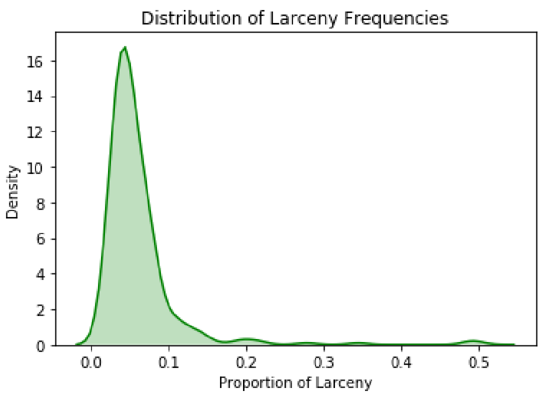
Initial Results suggest that Spatial features (i.e., Latitude, Longitude districts, streets) are more related to occurrences of Larceny than most Temporal features (i.e., Day of Week, Month, Year). Thus, the next phase of exploration will be to assess other forms of crime and whether the trends regarding Larceny differ between different forms of crime.

The present study sought to consider why certain parts of town might differ from one another in terms of Larceny frequency, as it may offer useful insights for members of the Boston community. Specifically, we considered what the distinguishing socioeconomic differences there might be between neighborhoods by using publicly available Census Data. Spatial location was used to make requests for corresponding Census tracts from the Census API. Then, variables like population density, resident demographics (i.e., age, gender, race) as well as median income were collected from Census Response data from the 2015 American Community Survey (ACS) specifically. The 2015 ACS was used because it provided a more granular level of measurement (encompassing blocks rather than districts) as with the 2018 ACS data. The sacrifice of temporal recency for spatial granularity seems fitting given the primary results and the relative effects of spatial features compared to temporal features.

Merging ACS Data

The geographic span of census tracts vary by population density; each district represents approximately 2,500 to 8,000 residents (Wiessies 2019). This particular dataset was chosen because it offered a compromise on spatial granularity with recency (the most granular survey would be from the 2010 Census, which is nine years old at the time of writing).

The ACS survey data contains population-weighted totals and percentages of Age, Gender, Race of residents, aggregated to the tract level. Beyond demographic statistics, and aggregate statistics of economic status (i.e.., educational attainment, median income, access to transportation) are reported for each tract as well. Planned analyses of these variables included the demographic information listed above, as well as median income and educational attainment. However, there were a large number of missing data for median income amongst the census tracts. Because the number of areas omitted from analyses of income would meaningfulness of results this critical measure of socioeconomic status was omitted from analyses.

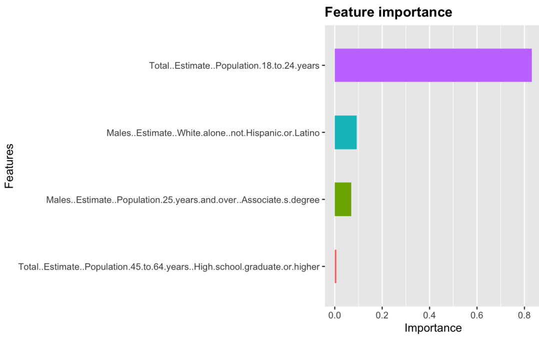


To merge the ACS data with the Crime Reports, Geolocation Data from the latter were used. The longitude and latitude of each respective report were used in requests to the Census API to collect Census information for coordinates. However, coordinates were rounded down to three decimal places to reduce the number of requests to the Census API and shorten the duration of data collection. Rather than submitting nearly 300,000 requests, this simplified approach contained approximately 7,650 unique requests. While this reduction in unique coordinates came at the cost of accuracy, it is not to a degree that impacts the veracity of results. Rounding our location from 8 decimal places to 3 meant that provides estimates within a range of approximately 80 meters at the 45th parallel (three degrees north of Boston), which is arguably a sufficient level of precision for identifying census tracts.

Analysis of Larceny with ACS Data

To assess the relative differences between tracts in terms of Larceny occurrences, a new feature was created to represent the proportion of crime reports within each tract that were for Larceny. Using this strategy, bivariate correlations for the relative frequency of Larceny were calculated with all other Census variables using Spearman's rho, a non-parametric rank-based measure of covariance with more lenient assumptions than Pearson's r coefficient (Kutner, Nachtsheim, Neter, & Li, 2004). This method was selected because the proportion of Larceny between different tracts were positively skewed.

A review of the rho coefficients for each numeric feature revealed a few noteworthy insights. First, coefficient with the largest magnitude was a cross-sectional variable that reflected specificities in Gender, Race, and Education. Specifically, Larceny was less common in neighborhoods that had a larger population of Black Females with a Bachelor's degree or greater (ρ = -.494, p < .01). Interestingly, the next largest coefficient was for Females with an Associate's degree or greater, irrespective of race (ρ =.08, p < .01). That is, when race is neglected and educational attainment is slightly lower, the population of educated females is negligible, with a magnitude that is much smaller than the former relationship, but notably in the opposite direction. More analyses need to be done to understand this trend, but what can be gathered from these relationships, however, is that the educational attainment and race of residents are in some way related with the frequency of Larceny within Tracts. Before progressing, it is worth re-iterating the analytics axiom that correlation does not imply causation. For example, there is evidence of racial profiling and discriminatory practices in the Boston area's police force (c.f., Antonovics & Knight, 2009). Such practices have an adverse effect on communities of racial minorities, meaning the trends in our data may have more to do with policing practices themselves, than group-level differences in criminal behavior. Additional analyses of bivariate relationships revealed more insightful trends: Age and gender appear related with the relative frequency of Larceny. The census estimated the population by gender using five age brackets: 18 to 24, 25 to 34, 35 to 44, 45 to 65, as well as 65 and over. Bivariate correlations with the Frequency of Larceny were greatest for ages 25 to 34, irrespective of gender (ρ_M = .33, p < .01; ρ_F = .34, p < .01). However, this relationship effectively dropped to 0 for Males aged 35-44 and 44-65, while the relationship became moderately negative for women in the same age respective brackets; (ρ_F = -.29, p < .01) and (ρ_F = -.32, p < .01). Coefficient Magnitudes were lowest for the youngest and oldest age brackets, which nonetheless show a pattern of Male population size being more related with Larceny Frequency. For the 18 to 25 year old population, the correlation by gender was (ρ_M = .132, p < .01; ρ_F = .086, p < .01), while the 65 and over population was (ρ_M = .172, p < .01; ρ_F = .100, p < .01). When another Gradient Boosted Regression Tree model was trained to predict Larceny exclusively with using features from from the ACS dataset. Validity estimates were comparable to the prior model (R^2 = 91.4%). The general pattern that population density matters most is apparent when assessing the realtive feature importance indices. A plot of the information gain each feature individually provided towards the model's overall accuracy is presented below.



It appears that far and away, the relative population of younger adults within a tract was particularly related with the frequency of Larceny within tracts. However, there was additionally evidence of additional social covariates, including demographic characteristics (i.e., non-white population) and educational attainment.

Conclusion

1) Calculating the violation level based on the frequency of the offense code group have helped us classify the different types of crimes as "non-violation","low","medium" and "high" opened room for further generalized analysis.

2) Friday has the most occurrence of "Strong" level of violation at 17:30.

3) Even though, Friday has the most occurrence of "Strong" level of violation, it belongs to the weekday crime decision Rule.

4) The frequency of the most common crime, larceny, is most common in areas with a larger population of younger Adults (ages 18-24).

References

Wiessies, Kathleen. 2019. https://libguides.lib.msu.edu/tracts.