Preventing Urban Crime

Jonathan Posada, Calvin Ying

ORIE 4741 Midterm Report

This is a final project for the course ORIE 4741: Learning with Big Messy Data offered at Cornell University and taught by Professor Madeleine Udell. The students seek to use machine learning to prevent urban crime. Using an open dataset for Philadelphia crime, we set out to find factors that are correlated with crime.

Exploratory Data Analysis

The first step in an ML project is to explore the data by way of viewing the raw data or use of summary statistics. By examining the source data we see that it contains a total of 14 features for each data point, i.e. for each crime incident. However, some of these features are redundant as date_time, date, time and hour are all features that encode the time of the incident. Furthermore, these features can be generated or parsed from the single feature of date_time. Similarly there are the features location_block and longitude and latitude coordinates, but the longitude and latitude coordinates can generate the location block of the incident.

The next step in exploring the data is cleaning the data. Luckily there was only one feature (Police_Districts) which was missing for some of the data samples. Since these data points with incomplete features only accounted for a fraction of all samples, roughly 7%, we chose to discard these samples to simplify analysis.

Analysis

After cleaning the data and making sure there were no missing values in our dataset, we divided the data into two parts – a training set with 80% of the data and a test set with the remaining 20%. We fit our model on the training set and observed our model's performance using data in the test set.

Before fitting any models, we created some plots to visualize our data.

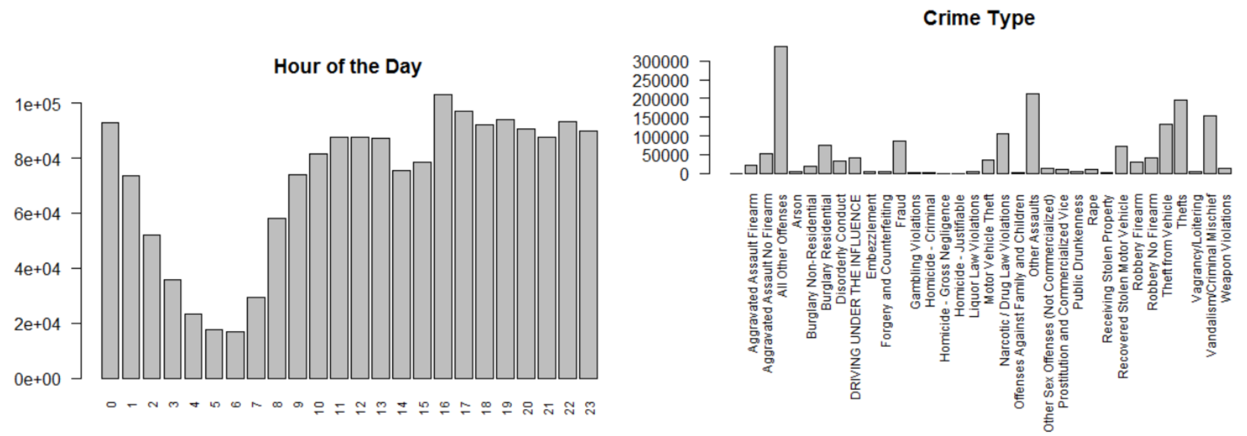

Figure 1.1: Number of Crimes by Hour of the Day (top left) and Number of Crimes by Crime Type (top right)

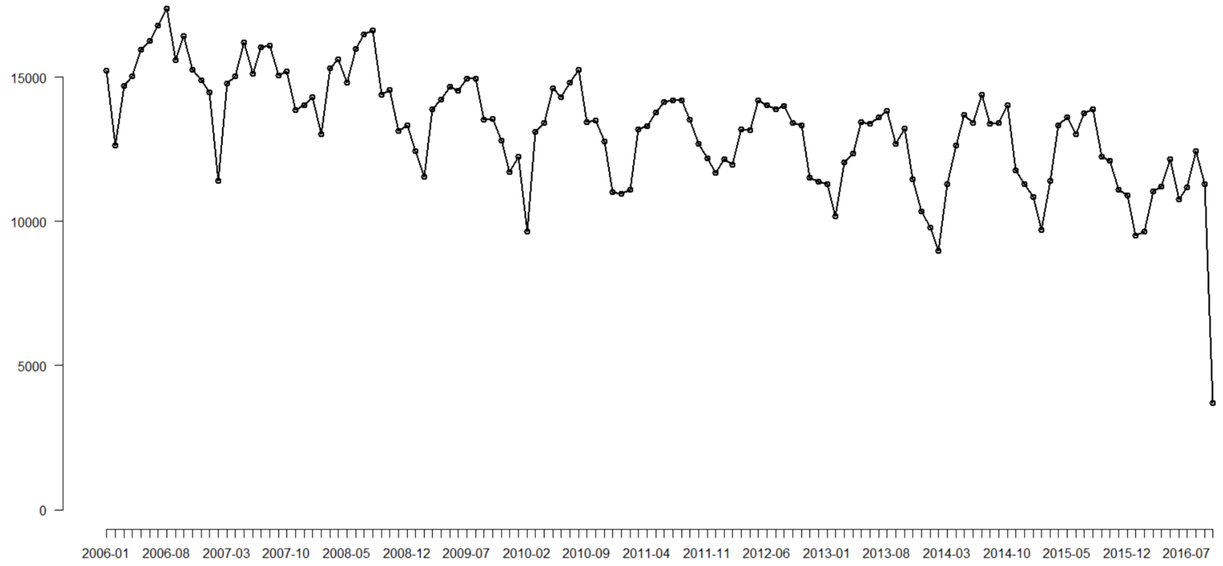

Figure 1.2: Number of Crimes by Month

We observed from these plots that more crimes happened in the evening than morning, and that one of the most popular crimes in Philadelphia is theft. We also noticed that the crime rate in Philadelphia has gradually decreased since 2006.

After eyeballing the data, we decided to fit linear regression to predict Crime Type using predictors such as Hour and Dc_Dist. Solving this problem helps us predict which crime type (i.e. rape, murder, etc.) will most likely occur given a time and place. The ”Hour” variable represents the hour of the day that a crime takes place, and the ”Dc_Dist” variable represents the Philadelphia district that includes the crime scene. Furthermore, the Crime Type variable contains all text data. In order to use this variable in our regression model, we converted the text data to numbers by assigning each crime type to a unique integer.

Our first order linear model shows that the coefficient for Hour is 0.0639 and the coefficient for Dc_Dist is -0.0364. Both predictors' coefficients have p-values close to 0, and the small p-values signal that both predictors are relevant and significant to our model. However, the adjusted R-squared in our model is 0.00271, which indicates that our model explains little variability of the response variable around its mean. Therefore, first order linear regression may not be the best model to fit our data. To prevent our model from overfitting the data, we will perform ridge regression and lasso to shrink the coefficients of our predictors. During the second half of the semester, we will fit these regularized models to obtain λ that will return the smallest mean squared errors and use that λ to find the desirable value for the coefficients.

Moving Forward

After taking a closer look at the dataset it is not as rich as we thought. At first glance, 14 features seemed like alot but there is a decent about of useless redundancy. It is also more difficult than anticipated to link housing data since Zillow does not allow for their data to be stored locally, it must be user-facing on the web. This prevents us from including this useful source of data in our project. We now plan to either find additional datasets to incorporate in our analysis in finding some essential factors that lead to crime.