Crime Time Midterm Report

Brendan Kirsh, Sudeshna Pontula, Dylan Tsai

November 8, 2019

Introduction

Our team is examining crime data from multiple major cities to determine trends in serious crimes over time. The dataset consists of complaints and police reports from Boston, Chicago, and San Francisco spanning the years 2015-2018. Through thorough analysis and the development of a machine learning model, we are able to predict higher levels of crime activity based on time, which will continue to yield similar trends across multiple cities. We are choosing not to analyze based on location to eliminate any biases that might be associated with different neighborhoods.

Data Specifications

Boston

For our Boston dataset, the following is a table of the columns we are given:

Column Name	Data Type	Description
offense_code	Categorical Integer	An index for the type of offense that occurred
offense_code_group	Categorical String	Generalized description of the offense, 67 categories
offense_description	Categorical String	Specific description of the offense, 242 distinct values
district	Categorical String	District number of offense
reporting_area	Categorical Integer	More specific location category
shooting	Boolean	True if there was a shooting, False if not
occurred_on_date	Date	Date and time the offense occurred
year	Integer	Value 2015-2018, year of offense
month	Integer	Value 1-12, month of offense
day_of_week	String	Day of week offense occurred
hour	Integer	Value 0-23, hour of offense
ucr_part	Ordinal String	"Part ...." One, Two, or Three, referring to severity of offense
street	String	Street name where offense occurred
lat	Decimal	Latitude of offense
long	Decimal	Longitude of offense
location	String	Ordered pair of (lat, long)

Using these fields, we were able to modify our dataset to eliminate location bias and allow for more complete analysis of offense severity. For the ordinal *ucr_part* column, the values given are "Part One", "Part Two", "Part Three", and a few scattered null or other values, with One being the most severe and others decreasing in severity respectively. Using one-hot encoding, we assigned matching number values to the three different part types, and the number 4 to all other values, as they are still less severe than Part 3 offenses. Using this, we then added three new columns, shown in the table below:

Column Name	Data Type	Description
ucr_lte_one	Boolean	True if offense is less severe than Part One, False otherwise
ucr_lte_two	Boolean	True if offense is less severe than Part Two, False otherwise
ucr_lte_three	Boolean	True if offense is less severe than Part Three, False otherwise

In addition to adding those columns, we have added a few other fields that will allow new views of our data. For example, we might want to examine whether a crime took place on a weekend, so a new Boolean field makes that distinction.

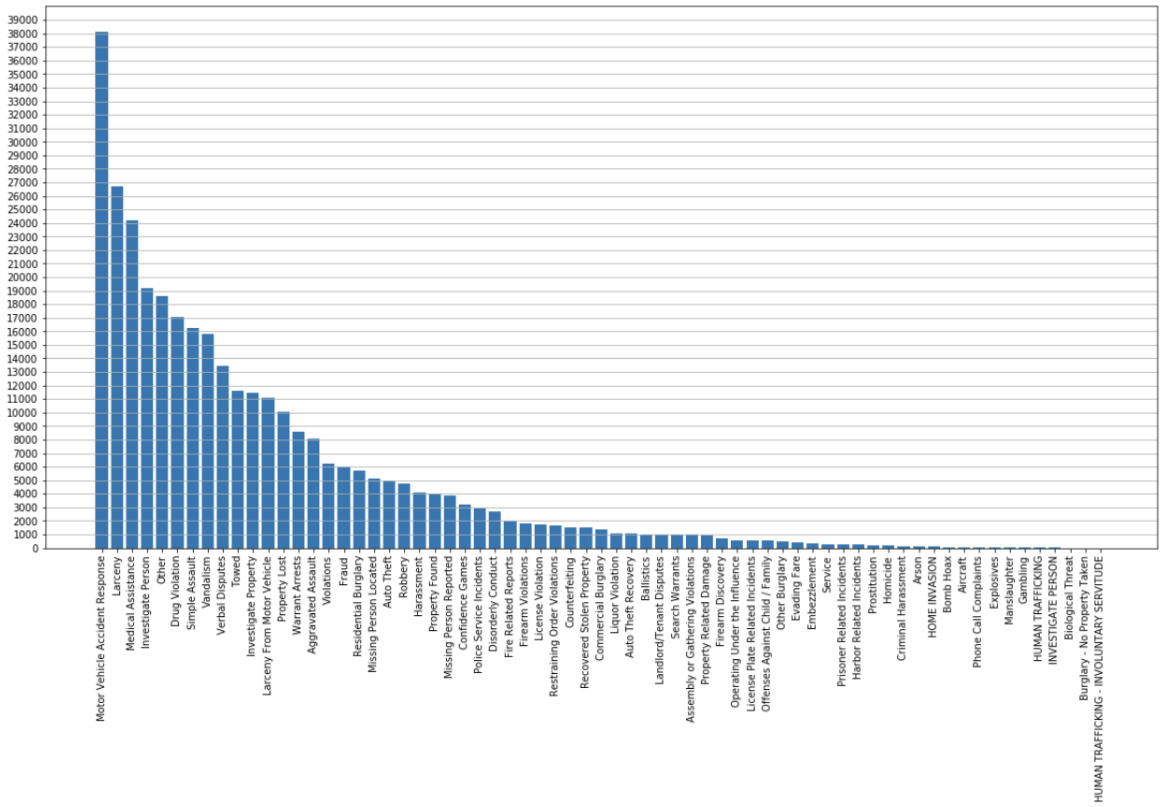
Because we are working to eliminate location bias in our data, we have removed all columns that are related to location. These columns are district, reporting_area, lat, long, location, and street. However, we found that it is useful to make a distinction between street types (i.e. highway, street, avenue, etc.), so we added a field to replace street that pulls the last word from the street column. For example, from "Maple St", we would get "St". Here is an explanation of the additional columns added:

Column Name	Data Type	Description
is_weekend	Boolean	True if offense occurred on a weekend, False otherwise
street_type	Categorical String	Type of street

Exploratory Data Analysis

First, we conducted a naive distribution of several features, to see what kind of biases might arise from data availability.

We determined that looking at the offense code does not prove useful at this point, as the frequency or severity of any given code does not correlate with the integer value. Instead, viewing the frequency of values in column offense_code_group is much more clear, as there are far fewer distinct values, and this view becomes useful when we get to the more granular analysis of crimes, based on type.



We can also view the frequency of offenses occurring on each day of the week, which shows us that while most days are similar, Friday is the most common and Sunday is the least common. Later analysis will reveal if the small differences are in fact significant. Viewing frequency by year is less useful to us, as our dataset only includes part of 2015 and 2018.

Next Steps

For our further analysis, we will be fitting a model to our Boston data as a training set, and using the Chicago and San Francisco data as a test set to determine if our crime predictions are accurate. We will be focusing on the features referring to crime severity as well as the time of day and day of the week features. We also plan to incorporate the shooting column into our model to predict the times with the highest volume of dangerous crimes.