Capstone Milestone Report

San Francisco Crime Classification

Mentor: Srdjan Santic

By Anushree Srinivas

Problem to be solved and motivation:

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay.

From Sunset to SOMA, and Marina to Excelsior, this competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. The objective of this project is to predict the category of crime that occurred at a given time and location.

Client:

City and County of San Francisco is trying to make the city and the neighborhoods more safe. The City authorities will be the client for whom the conducted data analysis as part of the capstone project will be beneficial.

Data Cleaning and Wrangling:

The data will be obtained from:
https://www.kaggle.com/c/sf-crime/data

There are **two** data sets in this project which I worked on:

1) **Train**.csv: The data set on which we'll be working on and deriving insights. This data set has **9** columns: Dates, Category, Descript, DayOfWeek, PdDistrict, Resolution, Address, X(latitude) and Y(Longitude).

2) **Test**.csv: The data set for which predictions of the crime category are to be done.

The first thing I did was to check for missing values in all the data sets. I used the **isnull()** function to check for null values in the data set and also tried to get the percentage of the null values in train data set.

To find the missing values I first found the **tota**l for each data set using the. **isnull() and .sum()** functions. This gave me the number of null values for each column in the train data set. There were no null values in this data set so therefore, there was no cleaning required to be done. I then proceeded to do the exploratory data analysis.

Exploratory Data Analysis:

I first started by looking at the distribution of the number of crimes occurred vs the day of the week. Using, **value_counts()** and **sns.countplot()**, I plotted the graph. Then since the data has dates, I converted them into hour, month and year using the **to_datetime()** function. Now using this, I decided to check the trend of crime activities with respect to the hour, month and year (2003-2015). The crime rate is maximum on Friday and Saturday and least on Sundays. The months of May and October recorded a high number of crimes. In terms of the time, the maximum crime activity is around 12pm and in the 5-7pm window the maximum being at 6pm.

I then went to explore how the crime activity affects the pdDistrict. So using **value_counts()** I plotted the graph of the crime activities vs regions(pdDristrict). The maximum crime rate was recorded for the Southern region followed by the mission region. Then I moved on, to find the top 5 crime categories.
Since in the top 5 categories, there was other offenses and non-criminal which doesn't give us much insight, I decided to drop using the **drop ()** those two rows to get a better idea about the top 5 categories. The 5 categories being Larceny/Theft, Assault, Vandalism, Drug/Narcotic, Vehicle/Theft.

Since the crime rate was maximum for Friday, Saturday and Wednesday and lowest on Sundays, I decided to look at the hourly distribution of the crime activities specifically on these days.

Then to get a better idea about the district specific crime category, I used a stacked bar graph to plot the top 5 crimes across the pd Districts.

Finally, I proceeded to see the frequency of the number of crimes for the top 5 categories in terms of day of the week and hour of the day.

Inferential Statistics:

For inferential statistics, I performed two different normality tests for each of the following variables: hour, year, month, day_num, X,Y. Shapiro-Wilk test couldn't be conducted since the sample size is > 5000). So I proceeded to do the Anderson-Darling, and the D'Agostino tests. To further check if any of the variables are correlated, I tried a correlation plot for the train dataset.I used **sns.heatmap()** and the **corr()** functions to get my corresponding correlation plot. There was a high positive correlation between X and Y of 0.56. The correlations between other variables were not that significant.

Future Work:

Use machine learning algorithms to predict the category of crime given the time and the area based on the previous crime history.