Capstone 1 - Milestone Report

22nd April 2018

Problem and Client

The city of Phoenix, AZ is the 5th largest city in the United States. Over the last 6 years the population of Phoenix has increased. With the increase in population there is the potential for an increase crime. This poses the question: 1) ***Does it make sense for the city of Phoenix to hire more police officers to support the population increase?; 2) Can we use historical data crime data to determine where crime hot spots will be to allocate resources more efficiently and not have to hire new officers?*** The clients for this project is the Phoenix Police department and City of Phoenix.. They can use the findings to allocate resources(officers, patrols routes, etc.), make budgetary decisions and develop community outreach programs to prevent crime.

Data

The city of Phoenix publishes an updated crime incident list at 11am daily (https://phoenixopendata.com/dataset/crime-data). The list contains crime incident data from November 1st, 2015 forward through 7 days prior to today's posting date. The data provided includes; Date/Time of Incident, UCR crime type, Block Address, Zip code and Premise Type. In addition, I will append demographic data for Phoenix to help answer our question which is also provided by the city of Phoenix.

Approach and Deliverables

The approach will be to use the historical crime data in combination with the demographic data to predict “hot spots” for crime. The type of machine learning technique has yet to be determined. The “hot spot” prediction will likely be presented as the Block Address or Zip code and the type of crime listed out to show here is *where* to put officers and the *type of crime* they should be looking out for. The deliverables for this project will be all the code, a report detailing the work completed and a slide deck providing the story and answers.

Data Wrangling

The dataset being used for this project is from the City of Phoenix Open data portal. The dataset in question is the Crime Data for Phoenix. The data provided is from November 1st, 2015 forward through 7 days prior to the newest posting date. The data used for the capstone project is from 11/1/15 to 2/20/18 (the csv file was downloaded on 2/27/18). This dataset has 148468 rows, 7 columns and 2 data types: object(6) and float64(1).

Reviewing missing data showed 4 columns have missing data:

1. Date_occurred_on = 336
2. Date_occurred_to = 43719
3. Zip code = 3
4. Presmise_type = 805

To clean the missing values for the date columns they will be filled in with the date/time from the populated date column of that incident. A check was run to display the rows that were NaN for date_occurred_on and date_occurred to and it returned 0 rows where that condition is true. Since we had missing values of 336 and 43719 we cannot omit those rows. The 43719 missing values are due to the victims knowing the exact date/time of the incident therefore the date_occurred_to field is left blank. The 3 zip code rows will be removed since 3 is a very insignificant number that will not affect the project when removed. The premise_type missing values will be classified as unknown, there is already an unknown variable entry. The 805 could not be removed since it is a large number of rows that could affect the project.

Overall the data is pretty clean aside from the large number of missing values. The clean data set is called *crime_clean* and has 148465 rows and 7 columns.

Exploratory Data Analysis

The crime data set is categorical type data set. The graphs created show counts of the groups within each variable based on different filters. To draw insights from the data using visualizations some initial questions were asked:

1. What are the top trending crimes?
2. When do crimes occur: Month, Day, Year, Season?
3. Do more crimes occur in certain zip codes?
4. What is the most common place a crime occurs e.g. house, business, etc.?

The answers to the questions above:

1. Top trending crimes are Larceny-Theft, Burglary, Motor Vehicle Theft, Drug Offenses
2. The most crimes occurred in the Winter months(Dec, Jan, Feb), Fridays and between the hours of 4-7pm.
3. Two zip codes stood out with the most crimes 85015 and 85008.
4. Crimes typically took place at Single Family Homes, Apartments, and Parking Lots.

These answers provide us with good information to look at how these categories are related. A technique to help identify relationships or associations is Market Basket Analysis (MBA). MBA takes the data and determines what items are commonly associated e.g. people who purchase Milk & Bread also purchase cheese. To accomplish this we use the Apriori function from the Apyori library. The main outputs of the Apriori function are the **Frequent itemsets** and **Support metric**. **Frequent itemsets** are the item or sets of items most frequently occurring together and the **Support metric** is the fraction of transactions where item(s) occur divided by the total number transactions. The higher the support the more frequently the item(s) occur. Along with frequent itemsets and the support metric there are 2 additional key metrics **Confidence** and **Lift** which are known as Association rules. **Confidence** is the probability of seeing an itemset that contains an item of the itemset. **Lift** is used to measure how more often an itemset occurs than we would expect if the items in the itemset were statistically independent. A Lift score of of 1 indicates independence.

Using the Apriori function with the following minimums: min_support=0.003, min_confidence=0.2, min_lift=4, min_length=4, a list of itemsets with support, confidence and lift were returned. The low Support and Confidence minimums were used to return more itemsets. The large Lift score was used to find itemsets that were far from independent. Reviewing the results showed the more items in the itemset the smaller the support and lower confidence which is expected. Single item and 2 item itemsets returned larger support and confidence values. The itemsets returned showed some patterns of items containing: Year, Month, Crime, Premise type, Time range and Season. What was not returned was itemsets containing zip codes, crimes, premise type, month, season, time range. The MBA was useful in returning frequent itemsets confirming suspected combinations.

Inferential Statistics

Using the information learned from the visualizations created and Market Basket Analysis, we created hypotheses to test on the population data. The data set is comprised of categorical variables therefore the Chi-Square test statistic was used. The Chi-Square test tests the strength of association between variables. 5 different hypotheses were tested and the results are listed below.

1. Compare Crime Category with Zip Code

- **Null Hypothesis:** There is no association between the crime type and zip code.
- **Alternate Hypothesis:** There is an association between the crime type and zip code.
- **Answer:** P-value = 0.0, null hypothesis is rejected

2. Compare Crime Category with Time Crime Occurred

- **Null Hypothesis:** There is no association between the crime type and time of day crime occurred.
- **Alternate Hypothesis:** There is an association between crime type and time of day crime occurred.
- **Answer:** P-value = 0.0, null hypothesis is rejected

3. Compare Crime Category with Season of the year

- **Null Hypothesis:** There is no association between crime type and season of the year.
- **Alternate Hypothesis** There is an association between crime type and season of the year.
- **Answer:** P-value = 3.4184825183899065e-07, null hypothesis is rejected

4. Compare Crime Category with Month

- **Null Hypothesis:** There is no association between crime type and premise type.
- **Alternate Hypothesis** There is an association between crime type and premise type.
- **Answer:** P-value = 1.6102836646351823e-18, null hypothesis is rejected

5. Compare Crime Category with Premise Type

- **Null Hypothesis:** There is no association between crime type and premise type.
- **Alternate Hypothesis** There is an association between crime type and premise type.
- **Answer:** P-value = 0.0, null hypothesis is rejected

From the hypothesis testing we see that Crime Type has significant associations with Zip Code, Time Crime Occurred, Premise Type, Season of the year or Months.