ZUHA AHMAD

DS-670: Capstone: Big Data & Business Analytics

Dr. Jaume

06 May 2017

**TELEPROMPTER**

Crime will remain to be erratic and irregular in many societies. However, with the power of analytics, we can hope to project the estimated outcome of future crimes and implement strategies to decrease all crimes, if not, some. The objective of the project is to conduct crime forecasting by performing a regression analysis, specifically a K-means clustering, based in the city of Portland, Oregon in the United States of America. The notion is to create a methodology to visualize the crimes in Portland and then further analyze the areas that are impactful to the community. By doing so, we can detect which areas are significant and how it can improve under law enforcement. The strategy is to use programming tools such as Zeppelin notebook to import and organize the comma separated values (CSV). Afterwards, R, will be used to generate a superlative algorithm and it will showcase findings through data visualization with an application called Shiny from *RStudio*. The dataset will be utilized is derived from the United States National Institute of Justice (USNIJ). Although, there are contributing variables, the main variables that will be considered over in this dataset are accidents, burglary, and shootings, will be taken into consideration. This would help to facilitate practical crime prevention solutions that correspond to specific times and places in Oregon.

According to the Oxford Dictionary, a crime is *an action or omission which constitutes an offence and is punishable by law*. A crime can occur at any place at any time which can be done discretely, unless the individual is caught in the act. Since there is always a possibility that crime will occur in the future by a series of chances, it would be practical and sensible to collect and review numerous amounts of information with data (i.e. qualitative, quantitative). Steps can be taken further as to create data visualization such as plotting, marking, and navigating “hotspots” of crime locations in maps and dashboards. The important aspect, above all else, is to predict the outcome and likelihood of the crime occurring in a different time, place, region or interestingly, repeated crime in the same time, place, and region. This leads to the emergence of crime analytics where overall, law enforcements analyze and effectively respond to crime patterns, series and trends by enabling data sharing, pattern analysis, predictive analytics, crime mapping and reporting. Consequently, this will establish law enforcement and police departments into action by constructing crime prevention strategies, raising public safety awareness, and feasibly modifying laws to governmental (city, state, federal) policies.

Oregon is a [state](https://en.wikipedia.org/wiki/U.S._state) located in the [Pacific Northwest](https://en.wikipedia.org/wiki/Pacific_Northwest) region on the [West coast](https://en.wikipedia.org/wiki/West_Coast_of_the_United_States) of the [United States](https://en.wikipedia.org/wiki/United_States). It is one of the most geographically diverse states in the United States, marked by volcanoes, abundant bodies of water, dense evergreen and mixed forests, as well as high deserts and semi-arid shrub lands (Jewell & McRae, 2014). Oregon's economy is mainly powered by various forms of agriculture, fishing, and hydroelectric power due to its landscapes and waterways. Portland is a port and the largest city in the state of Oregon, USA. The city covers 145 square miles and had an estimated population of 632,309 in 2015, making it the 26th most populous city in the United States (U.S. Census Bureau). According to the U.S. Census Bureau, Oregon's population as of 2015 is 4,028,977 with the Metropolitan Statistical Area (MSA) being 2,389,228. This leaves roughly 60% of Oregon's population residing within the metropolitan area. A dense population such as Portland would more than likely have high crime rates. Portland crime statistics indicate that crime is overall decreasing such as violent crime and property. Although improvements are being made, high crime rate indexes that occurred in 2012 for instance happened recently and no matter the predictions, future crimes remain to be random and changeable.

The competitor’s article that I have chosen is called ‘*Crime Forecasting Using Data Mining Techniques’*. I find the article to be compelling and applicable towards my research of crime forecasting. The competitor states that crime is ‘neither systemic nor entirely random’, but it can be classically ‘unpredictable’ which happens to be the case when combating with crime. There are several aims that the competitors had wanted to accomplish. The goal of their project was to explore a methodology for reliably predicting the location, time, and/or likelihood of future residential burglary. The dataset was derived from the United States National Institute for Justice (NIJ). Although the crimes were located in the Northeast, due to the sensitivity of the data; the name of their city was not specifically mentioned. First, they discussed how to generate architected data sets from original crime records. The architected data sets contain the aggregated counts of different types of crimes and related events as categorized by the city’s police department. Second, they had an ensemble of data mining classification techniques (i.e. SVM, J48, 1NN, Neural Networks, Naïve Bayes) that were chosen to perform the crime forecasting. Finally, they analyzed which classification approach is potentially the best method for predicting whether residential burglary will happen. In their results, they had listed out three findings having to do with spatial knowledge. First, there was success of the simple 1NN classifier modified with location constraint. It turned out that finding the most similar circumstance within the same neighborhood proved more effective than finding it within the entire city. Second, even though Neural Network is an ideal classifier, the Naive Bayes classifier had yielded better results in the F1 graph which is the total weight of precision and recall. Finally, the 24-by-20 grid data showed success measures that were consistently higher when using lower resolution data set which is due to each grid cell exhibiting a broader spatial knowledge.

The datasets are based on the locations listed in calls-for-service (CFS) records provided by the Portland Police Bureau (PPB) for the period of March 1, 2012 and had updated through February 28, 2017. However, I had collected the dataset from March 1, 2012 to 2016. The data is from the United States National Institute of Justice and it is based on a challenge called ‘Real-Time Crime Forecasting Challenge’ In this table, they have included crime categories, code (which had been re-categorized as “other”), and translation. They also had excluded data due to sensitive information and focus. The principal meaning of clustering refers to the process of making a group of abstract objects into classes of similar objects. The important aspects of clustering are that a cluster of data objects can be treated as one group. By performing cluster analysis, we would have to partition the set of data into groups based on data similarity and then assign the labels to the groups.

Then, I would talk about applications that were used in this project. After the Zeppelin notebook was applied, the following software that was used was R. R is a programming language, which is an open source platform and a software environment for statistical computing and graphics. It is a popular programming language which is widely used among researchers, statisticians, and even data miners for developing statistical software and data analysis. R has its own libraries that can be used for installation and implements a variety of statistical and graphical techniques. The RStudio environment was utilized and three sets of R scripts (i.e. ui.r, server.r, global.r) were being executed to run an application called R Shiny. The statistical codes and techniques that we applied was clustering. Some statistical examples include classical statistical tests, linear and nonlinear modeling, time-series analysis, classification, clustering, and many more. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. The packages that were part of the process were libraries of (dplyr), (DT), (ggmap), (ggplot2), (googleVis), (leaflet), (RColorBrewer), (readr), (shiny), (shinydashboard), (shinythemes), and (tidyr).

The R Shiny is a useful web application where you can show data visualization. This is used with a combination of a web browser and *RStudio*. Since, this is an interactive shiny application, it allows us to choose the crime variables that are needed to be analyzed and forecasted. The R shiny was created to visualize the city map of Portland, Oregon. It has a customized widget which helps to navigate the crime with latitude and longitude coordinates provided by the dataset. Overall, to forecast this application, we can change the date, months, even years in accordance to several crime variables. Since we are utilizing clustering algorithm, we can look at the time, trends, and the quantity of crimes in different spots of the city.

There are the two approaches that are used to improve the quality of hierarchical clustering. It is important to perform careful analysis of object linkages at each hierarchical partitioning. In addition, we have to integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters. This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points. For now, we will focus on the K-means algorithm. The K-means is one of the easiest unsupervised learning calculations that take care of the notable grouping issue. The method takes after a straightforward and simple approach to group a given informational collection through a specific number of bunches settled from the earlier. The primary thought is to characterize k centroids, one for each bunch. These centroids ought to be put shrewdly in view of various area causes distinctive outcome. Along these lines, the better decision is to place them however much as could reasonably be expected far from each other. The following stride is to take each direct having a place toward a given informational index and partner it to the closest centroid. At the point when no point is pending, the initial step is finished and an early gathering age is finished. Now we have to re-ascertain k new centroids as barycenters of the groups coming about because of the past stride. The R Shiny is a useful web application where you can show data visualization. This is used with a combination of a web browser and *RStudio*. The way that we are going to display our findings and results is through our visual application. This is an exceptional way to make crime forecasting more quantifiable and resourceful. There are multiple analyses that we can make, however, we are only going to concentrate on these variables such as accidents, burglary, and shooting. With each variable, we are going to count the number of cluster with the number of crime incidents from the year 2012. The date will start from March 1, 2012 since this is the first record to the last date, December 31, 2012. Then, we are going to count the number of cluster with the number of crime incidents from the year 2016. The date will start from January 1, 2012 since this is the first record to the last date December 31, 2016. In the R Shiny Map, there is a color distinction between the group number of clusters. If the cluster is blue, then it is a single crime incident or possibly an outlier. If the cluster is green, then the number of crime incidents range from 2-9. If the cluster is yellow, then the number of crime incidents range from 10-99. Finally, if the cluster is orange, then the number of crime incidents are 100+. Overall, there were Identified crime clusters over a period of time. The crimes in 2016 overall had increased from 2012 for all three variables. The majority of crime activities occur near the rivers especially in the Lloyd district and Pearl district is heavily impacted. We had achieved results by updating data that is supplied into the application. There were are clusters that are moving up north towards Washington State, particularly in Vancouver.