CHICAGO SCHOOL SAFETY ANALYSIS, CLUSTERING AND PREDICTION

Jan 2020

Author: Sujita Chatterjee

1.0 Introduction:

Safety in schools is a very important requirement and the neighborhood of the school has a big influence on the school safety. This project will use Chicago school and crime data from the Course 5 of this certification. School data will be used with Four Square API to find the available security ( number of Police stations). Depending on the crime counts in the neighborhood of existing schools, it can be decided if more police stations are required. For new schools, the crime data analysis and developing a model using machine learning can help with predicting the crime in the neighborhood in future. The project can be used by any location with the location, school and crime rates data available. This will help the school administration and government bodies to make accurate decisions on which locations are appropriate for new schools as well as which existing schools need more security arrangements.

2.0 Data:

This project will use the Chicago school and crime data from the Course 5 of the IBM Data Science Professional certification. Only certain columns of interest are retrieved from the datasets. The records with missing data in the datasets are dropped. Crime data prior to 2013 is used to develop a linear regression model to be used for predicting crime counts for future years. Chicago location data is obtained from the school data. The location co-ordinates for Chicago is obtained using the geopy.geocoders library Nominatim. Maps for the Chicago school data is displayed. FourSquare API will be used to find the police stations near schools to determine the available security and if additional security is required based on the current crime counts. Kmeans clustering is used to cluster the schools into clusters for further analysis.

3.0 Methodology:

First the python and machine learning packages and libraries to be used for the project are imported and installed.

Below is a list of the packages and libraries:

Lxml # Processing XML and HTML

Geopy # to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders

ProgressBar # to provide visual (yet text based) progress to long running operations

BeautifulSoup # library for web scraping

numpy # library to handle data in a vectorized manner

pandas # library for data analysis

json\_normalize # for flattening semi-structured JSON objects

matplotlib.cm # Colormap reference for a list of built in colormaps

matplotlib.colors # list of the named colors supported in matplotlib

requests # library to handle requests

geopy.geocoders Nominatim # convert an address into latitude and longitude values

matplotlib # library for visualization

sklearn.cluster KMeans # import k-means from clustering stage

folium # map rendering library

re # specifies a set of strings that matches it

Next, the Chicago crime data set is retrieved and cleaned.

Source of crime data:

<https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv>

The data is read into a python dataframe from the csv file. Only columns of interest are saved and other columns are dropped. The records with missing data are also dropped. Data prior to the year 2013 is saved and is used to develop a Linear Regression model to predict the crime count for a year for a certain location.

Next, the Chicago school data set is retrieved and cleaned.

Source of school data:

<https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv>

The data is read into a python dataframe from the csv file. Only columns of interest are saved and other columns are dropped. The records with missing data are also dropped.

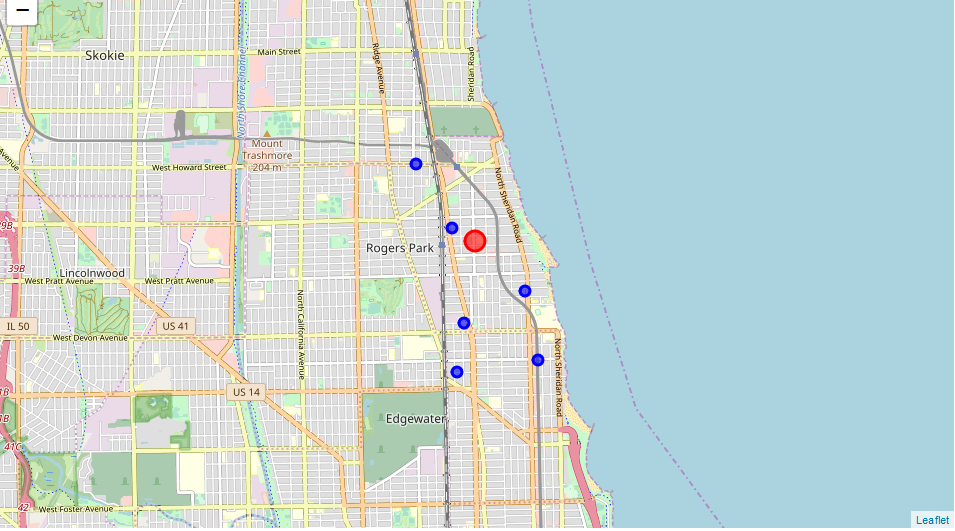
The location co-ordinates for Chicago is obtained using the Nominatim geocode and used to develop a Folium map to visualize the Chicago schools.

Foursquare API is used to find the police stations within 2000 meters of schools. A map is developed using Folium to visualize a sample school along with the police stations in its neighborhood.

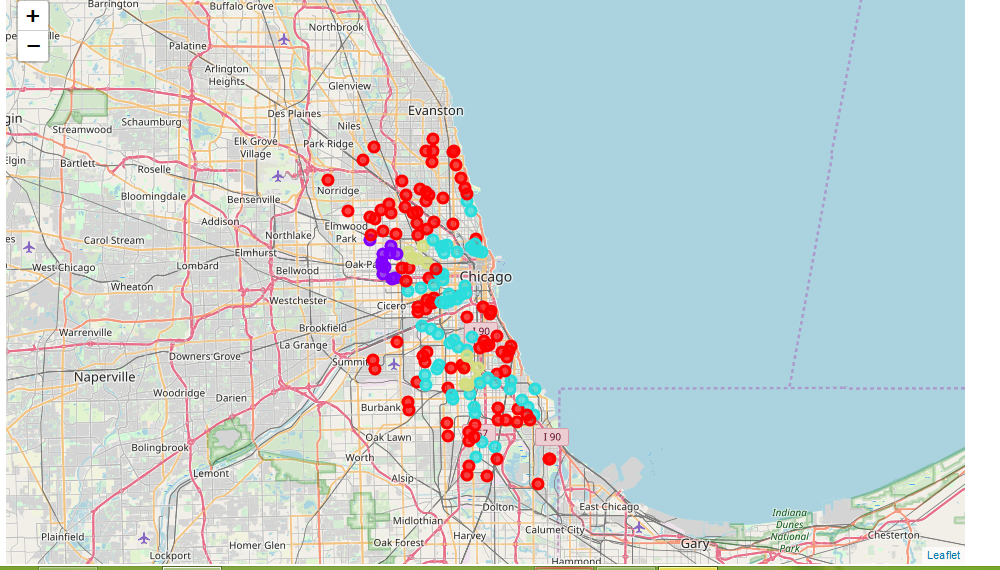
Sample map of Eugene Field Elementary School along with the police stations in its vicinity is provided in the next page.

The crime counts for each area is calculated and then joined with the school data. The combined school and crime data is used for clustering the schools using KMeans based on the crime count. The clusters are displayed in a map using Folium. The clusters are also analyzed to obtain the average, maximum and minimum crime counts in each cluster. The map with the clusters is displayed in a following page.

Map of Eugene Field Elementary School along with the police stations in its vicinity:



Map of KMeans clusters based on school and crime data:



4. Results:

The model developed using Linear Regression predicts the crime count for an area for a future year very accurately as seen by the R2 score.

The results of the FourSquare API query provide the police stations within a certain radius of a school.

The cluster maps and data analysis show that some clusters are spread across many areas while some are in one or two areas.

5. Discussion:

Machine learning models using Linear Regression are very effective tools for prediction and should be used to develop the model for predicting the crime count in an area for a future time.

FourSquare API also provides accurate information about existing venues. This is very helpful with locating the existing security as well as any other facilities of interest.

Both the crime count and available security in an area is very helpful for officials to determine if additional security is required for existing schools or if the areas are safe for new schools.

Visualizing the school and crime data in clusters in a map also helps to understand the areas with similar crime counts and which areas need further attention.

6. Conclusion:

The sample of Chicago schools used for clustering into four clusters based on crime counts produced one cluster across many areas, one with few areas, one with two areas and another in only one area. The Folium map helps in visualizing the clusters.

The model using Linear Regression for crime counts predicts the crime count for an area for a future time accurately.

FourSquare API helps provide the different venues for any location and can be used effectively to find the available security (police stations) in a schools vicinity.

As seen in the clusters map, more schools are in areas with less crime count. And the areas with high crime count have fewer schools. Further effort should be taken to add additional security in the areas with high crime count if their existing security is determined to be not adequate.