Coursera - IBM Data Science Certification

Course - Applied Data Science Capstone

Capstone Project Report - The Battle of Neighborhoods

Presented by – Anand Joshi

Date: 21-Aug-2019

Version 1.0

Table of Contents

Executive Summary

This report provides an analysis of the higher claim rates, higher insurance premiums and customer dissatisfaction of Toronto Insurance Ltd. Data Science Methods used include exploratory analysis of various public data sources, Geopy library and Foursquare API. Other tools used are Python and various packages available for statistical analysis. All calculations can be found in the appendices. Results of data analysed show that there is a strong correlation between higher crime rates in a neighbourhood and location of a police station.

Recommendations discussed include:

- Adoption of proven Data Science Methodology and Machine Learning Tools to gain further insight into the challenges faced by the Toronto Insurance Ltd
- Use and integration of reliable and publicly available data sources
- Integration of Location Technology / API in determining insurance premiums

Introduction

Business Problem

Toronto Insurance Ltd (a Personal Home and Auto Insurance company) based in Canada with a significant market share in Toronto city is lately facing a higher claim rates by its customers. This has resulted into lower profits, higher insurance premiums and customer dissatisfaction. The CEO of company would like to carry out a Poof of Concept (PoC) using readily available Data Science & Machine Learning tools, a popular location technology API and most importantly using the publicly available crime data. The key success factors of this PoC are:

- to identify a minimum of 6 neighborhoods in Toronto city having highest and lowest Crime rates
- successful integration of publicly available data sources on Crime Rates and location mapping technology
- Identify unknown clusters or data patterns of crimes which visually may not be identifiable
- enable the Insurance company to offer a targeted insurance premium based on the neighborhood in which customer lives or does the business.

The key success factors of this PoC are:

- to identify a minimum of 6 neighborhoods in Toronto city having highest and lowest Crime rates
- successful integration of publicly available data sources on Crime Rates and location mapping technology
- identify unknown clusters or data patterns of crimes which visually may not be identifiable
- enable the Insurance company to offer a targeted insurance premium based on the neighborhood in which customer lives or does the business.

Data Sources

The focus of this PoC is on Toronto city so the key data sources are explored locally. The below datasets considered for analysing Toronto city:

Data 1: a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. This dataset exists for free on the web. The key features used are Postal Code, Borough and Neighborhood

Data 2: Second data source is the Geospatial to get the latitude and the longitude coordinates of each neighborhood in a CSV format

Data 3: Third data source is Toronto Neighborhood Crime Rates related details from 2014 to 2018 available at Toronto Police Service: Public Safety Data Portal. The key features used are Neighborhood and various types of crimes that includes Assault, Auto theft and Break-ins etc.

Data 4 : Fourth data source is Foursquare to explore the neighborhoods and segment them. Foursquare is a Location Technology Company which provide an API and Data to further drill down the neighborhoods of Toronto

Methodology

In this exercise we will focus our efforts on visualizing the Toronto neighbourhoods with highest and lowest crime rates and overlay with locations of Police Stations.

In first step we have collected the required data: Postal Codes of Toronto starting with letter M.

Second step in our analysis will be enriching the Toronto neighbourhood data with Latitude and Longitude using the Geospatial API.

In third step we will focus on integrating external and publicly available crime data at ***Toronto Police Service: Public Safety Data Portal*** from 2014 to 2018

And in the fourth step, we will use ***Foursquare API*** to query Government category to get the venue for Police Stations and overlay on top of Toronto neighbourhoods with crime rates

Throughout various steps, we apply various exploratory data analysis methods to assess the data quality and apply appropriate remediation steps.

Results

This picture shows the distribution of Toronto neighbourhood with top six areas with highest and lowest crime rates. The red circles are with high crime rates and blue ones with lowest, and yellow circles indicates in between.

This picture shows the distribution of Toronto neighbourhood with top six areas with highest and lowest crime rates along with nearby location of Police Stations.

Conclusion

This concludes our analysis. We have highlighted Toronto neighborhood with the highest to lowest crimes and nearby Police Stations. It is interesting to observe high crime neighborhoods in peripheral areas on Toronto city and downtown core is having moderate crime rates. The Insurance company can factor in the distribution of the crime rates and location of the police stations for determining the premium rates.

All key success factors of this PoC are achieved i.e.:

1. identify a minimum of 6 neighborhoods in Toronto city having highest and lowest Crime rates
2. successful integration of publicly available data sources on Crime Rates and location mapping technology
3. identify unknown clusters or data patterns of crimes which visually may not be identifiable
4. enable the Insurance company to offer a targeted insurance premium based on the neighborhood in which customer lives or does the business.

Final decision of adopting Data Science and ML tools, and use of publicly available data sources will be made by stakeholders based on specific characteristics of neighborhoods and crime rates, taking into consideration additional factors like police station location etc.

References & Appendix

Data 1 - a list of postal codes in Canada: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data 2 - Geospatial to get the latitude and the longitude coordinates of each neighborhood in a CSV format: http://cocl.us/Geospatial_data

Data 3 : Toronto Neighborhood Crime Rates related details from 2014 to 2018 available at Toronto Police Service: Public Safety Data Portal: http://data.torontopolice.on.ca/datasets/neighbourhood-crime-ratesboundary-file-/data?geometry=-80.686%2C43.542%2C-78.346%2C43.89

Data 4 – Foursquare API: https://foursquare.com/

Special thanks Coursera Instructors for presenting a quality course and peer assignment reviewers for their timely reviews and encouraging comments.