The Battle of the Neighbourhoods - Week 1

Part 1 - Introduction

I am a Data Scientist working in Toronto's Financial Hub where top Canadian Banks and Insurance companies' headquarters are located. A few months back on my trip to Singapore, I met with a fellow traveller who happened to be the CIO of a leading Insurance Company in Canada and he brought up an interesting business challenge his company was facing and wondering how Data Science and Machine Learning tools could be used to solve the business challenge. In the following section, I am going to articulate the business problem and how Data Science / ML tools can be used for solving this challenge.

Business Problem

A leading Personal 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 Proof of Concept (PoC) using readily available Data Science & Machine Learning tools, a popular location technology API and most importantly leveraging the publicly available crime data.

The key success factors of this PoC are:

1. to identify a minimum of 6 neighbourhoods 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 neighbourhood in which customer lives or does the business.

High-level approach

1. use open source and freely available data science and machine learning tools like Python, Juypter Notebook, Sci-kit learn ML library and Github *(a code hosting platform for version control and collaboration.)*
2. use publicly available Crime data using Toronto Police Service - Public Safety Data Portal
3. identify top Toronto neighbourhood using ForeSquare API (a popular location Technology provider)
4. from this list of top neighbourhoods, the list is augmented with additional geographical data
5. present the historical crimes within a predetermined distance of all neighbourhoods are obtained
6. a map is presented to the to the CIO showing the selected neighbourhoods and crime statistics of the area.
7. future probability of a crime happening near or around the selected top sites is also presented to the user

I am given the opportunity to lead this PoC, present the findings and business rationale to the CIO of the Insurance company.

Target Audience

This solution is targeted at ***the CIO*** of the Insurance company and the ***Customers*** to explains the reasons how premiums are calculated and why it varies based on the neighbourhood. This approach may of interest to ***other Insurance Companies*** encountering similar business challenge.

Part 2 - Data

Description of the data and its sources

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

Data 1 This is 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. Link to the dataset is: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

To create the below (df) dataframe: The dataframe will consist of three columns: PostalCode, Borough, and Neighbourhood Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned. More than one neighbourhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighbourhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighbourhoods separated with a comma as shown in row 11 in the above table. If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough. So, for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighbourhood columns will be Queen's Park.

The following screen shot lists top 10 entries after the link is scrapped into a DataFrame.

	PostalCode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford Heights

There are a total 103 unique postal coded with one or more Boroughs and Neighbourhoods.

Data 2 : Second data source is the Geospatial to get the latitude and the longitude coordinates of each neighbourhood in a CSV format from http://cocl.us/Geospatial_data. The following screen shot lists top 10 entries after the file is loaded into a DataFrame.

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford Heights	43.757410	-79.273304

Data 3 : The third data source is Toronto Neighbourhood 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-rates-boundary-file-/data?geometry=-80.686%2C43.542%2C-78.346%2C43.89

The below is a screen shot of a sample list:

OBJECTID_1	Neighbourhood	Hood_ID	Hood_ID	Neighbourhood	Assault_2014	Assault_2015	Ass
1	Yonge-St.Clair	097	97	Yonge-St.Clair	58	38	51
2	York University Heights	027	27	York University Heights	78	101	111
3	Lansing-Westgate	038	38	Lansing-Westgate	216	203	223
4	Yorkdale-Glen Park	031	31	Yorkdale-Glen Park	121	141	136
5	Stonegate-Queensway	016	16	Stonegate-Queensway	109	140	124
6	Tam O'Shanter-Sullivan	118	118	Tam O'Shanter-Sullivan	63	58	50
7	The Beaches	063	63	The Beaches	349	392	380
8	Thistletown-Beaumond Heights	003	3	Thistletown-Beaumond Heights	45	47	39
9	Thorncliffe Park	055	55	Thorncliffe Park	111	124	157
10	Danforth East York	059	59	Danforth East York	214	203	214

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

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude
0	Rouge, Malvern	43.806686	-79.194353
1	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	Guildwood, Morningside, West Hill	43.763573	-79.188711
4	Guildwood, Morningside, West Hill	43.763573	-79.188711

Once all data sources are normalized, they will be graphically overlay using Foursquare API to show the neighbourhoods with Crime rates and to pick top six neighbourhoods with (3) highest and (3) lowest crimes.