Coursera : Capstone project

IBM datascience professional certificate

Title: Opening a big chain grocery store in new neighbourhoods

March 2019

Introduction

The Globe is rapidly urbanizing. More than half of the world's population lives in cities, and by 2050 – if current trends continue – the urban population in developing countries is expected to nearly double to 6.7 billion, according to the U.N. Hence it only seems natural that cities are expanding increasingly. And once a deserted and empty area of a city is now rising with population. In my city, Casablanca, not rising areas have a big grocery chain market where they can have legions of choices though they are highly populated. And are left with small retailers that impose super high prices and not many brands to choose from. Consequently, the majority of people living in those areas have to either drive or take public transportation and go to whatever nearest market hence large crowds are shopping at the same time which makes the whole experience excrutiating and unecessarly long. It is evident that opening a new super market is more complicated than one might think as with any business decision. It takes serious work and consideration and one of the most crucial decisions is where to locate the project and where exactly it would be more benificial to build it.

Business problem

The objective of this project is to deploy the tools acquired along the whole course such as machine learning's techniques to find a solution to a business problem. And the one we will be solving in this project is how to find the best location for super markets in all the rising neighbourhoods if a property developper is looking to open a new supermarket.

Target audience

This project is targeted at any inverstors or property developesr that would like to open a new super market in the capital city of Morocco, Casablanca. This project is timely as the city is currently suffering from oversupply of super markets.

Data

To solve the problem, we will need the following data:

→ List of neighbourhoods in Casablanca. This defines the scope of this project which is confined to the city of Casablanca, the capital city of the country of Morocco in North Africa.

→ Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

→ Venue data, particularly data related to super Markets. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methodolgy to go about this problem

The following wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Casablanca) contains a list of neighbourhoods in Casablanca, with a total of 24 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Casablanca. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Casablanca). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Casablanca. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering.

Since we are analysing the “Supermarkets” data, we will filter the “Supermarket” as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Supermarket”.

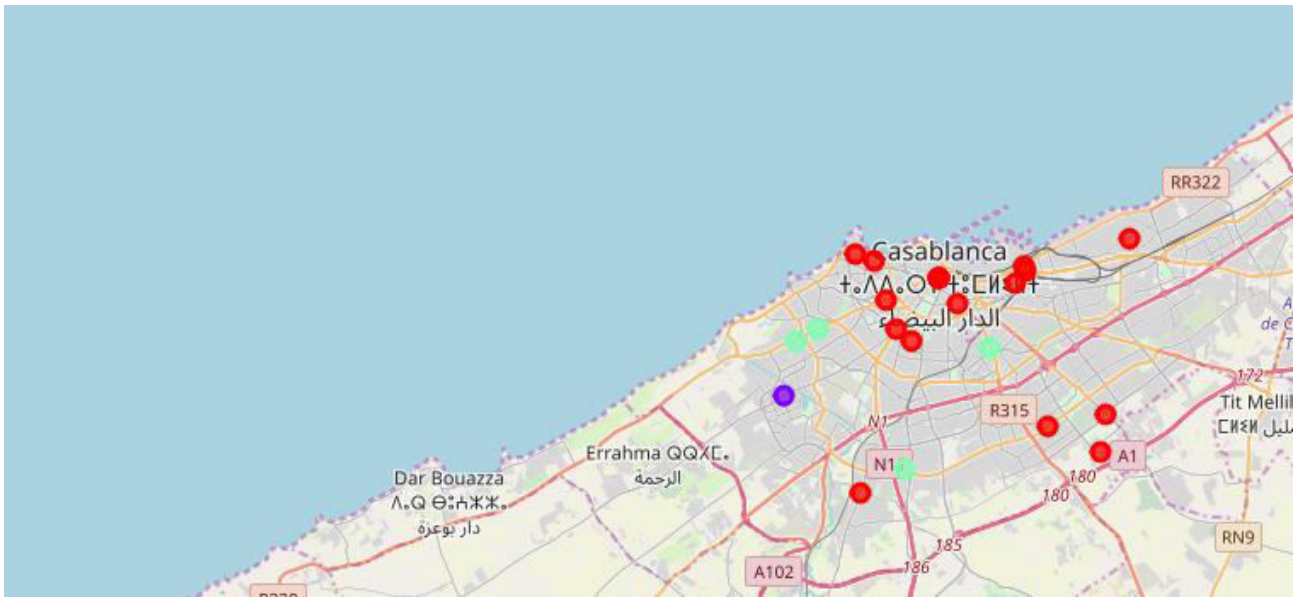
The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new supermarkets

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Supermarket":

→ **Cluster 0:** Neighbourhoods low number to no existence of supermarkets
→ **Cluster 1:** Neighbourhoods with moderate number of supermarkets
→ **Cluster 2:** Neighbourhoods with high concentration of supermarkets

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour



Discussion

As observations noted from the map in the Results section, most of the supermarkets are concentrated in the outer area of Casablanca, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to totally no supermarkets in the neighborhoods. This represents a great opportunity and high potential areas to open new supermarkets as there is very little to no competition from existing ones. Meanwhile, supermarkets in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of supermarkets. From another perspective, this also shows that the oversupply of supermarkets mostly happened in the outer area of the city, with the central area still having very few supermrkets. Therefore, this project recommends property developers to capitalize on these findings to open new supermarkets in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new supermarkets in neighborhoods in cluster 1 with moderate competition.

Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of supermarkets and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of supermarkets, there are other factors such as population and income of residents that could influence the location decision of a new supermarket. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new supermarket. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new supermarket. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new supermarket.