IBM Applied Data Science Capstone

Leveraging Data Science to Predict New Shopping Mall Location in Rome, Italy

Kenneth Zong

Introduction/Business Problem:

Shopping malls are a place of happiness for the ordinary consumer to actively shop, eat, relax, and spend time with loved ones. The mall experience is unique such that there is always something for someone to do. Not only that, but for the average tourist, a shopping mall can be a great way to learn about a country's culture without having to travel many miles. Many surrounding clubs, schools, and venues have the flexibility to market themselves more especially at a centralized hub such as a shopping mall. However, what makes the shopping mall experience so great is not just the extracurricular events that are held there or the restaurants and consumer stores that reside within it, but also the location. If the location is within a cluster radius of too many malls, it will be difficult to earn profit as the new mall will be competing with other ones within their vicinity.

It is important to look not only where there are little to no malls but also where there are many malls. So a common business problem can be asked: If a contractor is told to suggest a few of the places that a shopping mall would flourish, where would it be located, based on the data of Rome, Italy?

Data:

The data needed would be the neighborhood surroundings and location names within Rome. Latitude and Longitude coordinates of these neighborhood locations would allow us to plot the locations on a map and accurately generate the cluster model. Also, it would be interesting to know what other malls and venues are within those neighborhoods already. It is important to understand this because it may be meaningless to open a new mall within the vicinity of another one.

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Subdivisions_of_Rome) contains a list of neighborhoods in Rome , with a total of 46 neighborhoods. Web scraping techniques can be utilized to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Geographical coordinates of the neighborhoods can be extracted using the Python Geocoder package which will allow us to plot the latitude and longitude coordinates of the neighborhoods. Another package used will be the Foursquare API to get the venue/mall data for those neighborhoods. Foursquare API (https://foursquare.com/))

can provide copious amounts of data; however, we are only interested in the shopping malls within the vicinity.

Methodology:

We need the list of neighborhoods in the city of Rome. This data is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Subdivisions_of_Rome). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods . 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 addresses 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 neighborhoods in a map using the 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 Rome . 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 neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighborhood by grouping the rows by neighborhood 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 “Shopping Mall” data, we will filter the “Shopping Mall” as a venue category for the neighborhoods. 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 neighborhoods into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighborhoods have higher concentrations of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the

occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results :

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:
Cluster 0: neighborhoods with low number to no existence of shopping malls
Cluster 1: neighborhoods with high concentration of shopping
Cluster 2: neighborhoods with moderate number of shopping malls
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 shopping malls are concentrated in Infernetto, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has a very low number of no shopping malls in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls 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 shopping malls in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of shopping malls and suffering from intense competition.

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 shopping mall. To answer the business question that was raised in the introduction

section, the answer proposed by this project is: The neighborhoods 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 in high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

References :

Category: Subdivisions of Rome in Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:Subdivisions_of_Rome

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs