Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Pune, India

Simran Kaur Ahluwalia
March 2021

Introduction

A shopping centre is defined as an aggregate of trade enterprises and enterprises providing services located in a certain territory, planned, built and managed as a whole and providing parking for the vehicles within its territory. For many shoppers, visiting malls are a great way for relaxation. They can do grocery shopping,dine at resturants, shop at various fashion outlets, watch movies and perform many more activities. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market tgeir products and services. As a result, there are many shopping malls in Pune and many more are being built. In particular, the location of shopping mall is one of the most important decisions that will determine whether the mall will be success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in Pune, India to open a new shopping mall.Placement of shopping centers is the most important aspect of developing the concept of designing a modern shopping center.Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Pune, India, if a property developer is looking to open a new shopping mall, where would you recommend that they will open it?

Target Audience of this project

This project would particularly be useful for investors & property developers looking out for investment and developing opportunities in Pune, India. This project is timely as the city is currently suffering from oversupply of shopping malls.

Data

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

- List of neighbourhoods in Pune. This defines the scope of this project which is confined to the Pune city.
- Latitude & Longitude coordinates of those neighbourhoods. This required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page(https://en.wikipedia.org/wiki/Category:Villages_in_Pune_district) contains a list of neighbourhoods in pune, with a total 137 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhood using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare API has one of the largest database of 105+ million places and used by over 125,000 developers. Foursquare API will provide many categories of the venue data; we are particularly interested in the shopping mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping(Wikipedia), working with API(Foursquare),data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization(Folium). In the next section, we will present the methodology section where we will discuss the steps taken in this project, the data analysis that we did the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Pune. Fortunately the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Villages_in_Pune_district) . 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 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 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

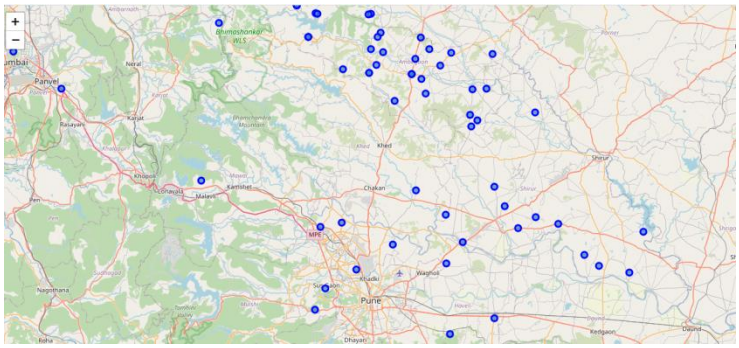
plotted in the city of Pune. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and secret key. Then we make API calls to Foursquare passing the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venues data in JSON format and we will extract the venue name, category, 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 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 venues category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-meansclustering. 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 algorithm and is particularly suited to solve the problem for this project. We will cluste the neighbourhoods into 3 cluster based on their frequency of occurrence for "Shopping Mall'

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 shopping malls.

Results

Most of the shopping malls are concentrated in the Pune city, with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has only two shopping mall in the neighbourhoods. 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.



Discussion

As observation noted from map in the result section, most of the shopping malls are concentrated in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has

only two shopping mall in the neighbourhoods. 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. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 1 which already have higher concentration of shopping malls and suffering from intense competition.

Limitation & suggestions

In this project, we only consider one factor i.e frequency of occurrence of shopping malls ,there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. 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 shopping mall. 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 shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open 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 shopping mall.