Coursera Capstone

IBM APPLIED DATA SCIENCE CAPSTONE

Opening a New Residential Project in Patna, Bihar.

By: Aditya

July, 2019

Introduction

Real State is property consisting of land and the building on it, along with its natural resources such as Crops, Minerals or Waters.

Residential real estate may contain either a single family or multifamily structure that is available for occupation or for non-business purposes.

Residences can be classified by and how they are connected to neighbouring residences and land. Different types of housing tenure can be used for the same physical type. For example, connected residences might be owned by a single entity and leased out, or owned separately with an agreement covering the relationship between units and common areas and concerns.

Of course, as with any business decision, opening a Real State requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the project is one of the most important decision that will determine whether the project will be a success or a failure.

Problem

The Objective of this capstone project is to analyse and select the best location in the city of Patna, Bihar, India to open a new Real State Project. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business questions: In the city of Patna, Bihar, if a property developer is looking to open a new Real State Project, where would you recommend that they open it ?

Target Audience of Project

This project is useful for property developers and investors who are looking to invest in new Real State Project in the city of Patna, Bihar. As according to 2011 stats Patna had an estimated city population of 1.68 million in 2011, making it the 18th largest city in India. With over 2 million people, its urban agglomeration is the 18th largest in India.

Data

To solve the problem, we will need the following data

- List of neighbourhoods in Patna. This defines the scope of this project which is confined to the city of Patna, the capital city of Bihar, a state in India.
- 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 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:Neighbourhoods_in_Patna) contains a list of neighbourhood in Patna, Bihar. 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 neighbourhoods 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 has one of the largest database of 105+ millions places and is used by over 125,000 developers.

Foursquare API will provide many categories of the venue data, we are particularly interested in the Real State category in order to help us to solve the business problem put forward. This is a project that will make us use of many data science skills, form web scraping, to working with APIs, data cleaning, data wrangling, to Machine learning (K-Means Clustering) and finally Map visualization (using Folium).

In the next documentation section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysed and the machine learning techniques that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Patna. Fortunately, the list was present on Wikipedia (link already mentioned above)

We will do web scraping using Python requests and beautifulSoup packages to extract the list of neighbourhoods data. However, this was just the 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 on map of Patna.

Next we will use Foursquare API to get the top 100 venues, but unfortunately here the data was not present for all the location of city Patna. So the scope of clustering the data was not possible in this case.

Limitation and Suggestions for Future Research

In this project, we are bounded by the lack of data on Foursquare API

Future research could devise a methodology to estimate such data to be used in thr clustering algorithm to determine the preferred location to open a new Real Estate Project. 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, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a New Real Estate Project and to answer the business questions that was raised in the introduction section. 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.