Coursera Capstone

Set up Café in Bangalore

IBM DATA SCIENCE CERTIFICATION

Amitabha Kanjilal
September 2019

Introduction

In the modern days that we live in, cafés have become very popular, and is a great place to hang out with friends or family. Any major city would usually have a bunch of cafés around, and many investors still look out for business opportunities for opening brand-new cafés in town. However, the location is a vital factor for getting profits out of it.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Bangalore, India to open a new café. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Bangalore, if an entrepreneur is looking to set up a new café, where would you recommend that they open it?

Data Collection

We need the following data to solve the problem –

- List of neighbourhoods in Bangalore
- Latitude and Longitude for those neighbourhoods
- Venues, especially cafés in 10 km radius of these co-ordinates

We are extracting the neighbourhood data by web-scraping the Wikipedia link for localities in Bangalore. After we get the neighbourhood names, we are searching for the co-ordinates of those neighbourhood in google and again scraping for the latitude and longitude values.

Once we have the co-ordinates, we are using the foursquare API to fetch top 100 venues within a 10 km radius of each neighbourhood.

Methodology

Once we have collected the data as explained above, we will transform the data into another data to 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 Café data, we will filter the Café as venue category for the neighbourhoods.

Finally, 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 4 clusters based on their frequency of occurrence for Café. The results will allow us to identify which neighbourhoods have higher concentration of cafés while which neighbourhoods have fewer number of cafés. Based on the occurrence of cafés in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new cafés.