IBM

Data Science Project Week 4th

INTRODUCTION

This is a capstone project for IBM Data Science Professional Certificate. I am developing a hypothetical scenario in this project for a concept that does not have enough Indian Restaurants in the Toronto area. So this might be a perfect opportunity for a Canadian-based entrepreneur. Since Indian food is popular among the Asian community, this entrepreneur may think about opening up his business in areas where the Asian community resides. With the intention in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I'm designing this project to help him find the most suitable place.

BUSINESS PROBLEM

The goal of this capstone project is to find the most appropriate place for the entrepreneur to open a new Indian Restaurant in Toronto , Canada. This project aims to provide answers to the business problem by using data science methods and techniques along with machine learning algorithms such as clustering: in Toronto, if an entrepreneur decides to open an Indian Restaurant, where will they consider opening it?

TARGET AUDIENCE

The entrepreneur who wants to find the location to open authentic Indian restaurant.

DATA

We'll need below data to solve this problem:
● List of Toronto neighbourhoods, Canada
● The latitude and longitude of those districts
● Venue data concerning restaurants in India. This will help us find the neighborhoods better suited for opening an Indian Restaurant.

EXTRACTING THE DATA

- Scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

  In the next section we will discuss the techniques used in data analysis and machine learning.

METHODOLOGY

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

I scraped the site using HTML table scraping approach for pandas as Pulling tabular data directly from the web is simpler and more convenient Page in frame of data.

It's just a list of the community names and postal codes, though. I need to go to get their co-ordinates to use Foursquare to draw a map of locations Near to those barrios. I tried using Geocoder to get the coordinates package but it didn't work so I used the CSV file IBM provided unit to suit community co-ordinates in Toronto. Upon reunion I imagine the Toronto map with these coordinates using the Folium kit to verify that these

coordinates are right. Afterwards, I use the Foursquare API to draw a list of the top 100 venues within a radius of 500 metres.

To get account ID and API key to pull the info, I built a Foursquare Developer account. I can pull the names, categories, latitude, and longitude of the venues from Foursquare. With this info, I can also test how many specific categories I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of each venue category occurring. This is to plan the clustering for later completion.

Here, I made a justification to specifically look for "Indian restaurants. "Finally, I performed the clustering method 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. This is one of the simplest and most common unsupervised machine learning algorithms, and is therefore highly appropriate for this project. I have grouped the Toronto neighbourhoods into 3 clusters based on their incidence frequency for "Indian cuisine." I'll be able to recommend the best place to open the restaurant based on the findings (cluster concentration).

RESULT

CLUSTERS

The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with the less number of Indian restaurants.
- Cluster 1: Neighborhoods with no Indian restaurants.
- Cluster 2: Neighborhoods with a more number of Indian restaurants .

The results are visualized in the above map with Cluster 0 in green, Cluster 1 in blue, Cluster 2 in red.

RECOMMENDATIONS

Most of the Indian restaurants are in cluster 2 which is around Central Bay Street, Church and Wellesley, Berczy Park, Union Station, Richmond,lowest in Cluster 1 areas which are in North Toronto West and Parkadeareas. Also, there are good opportunities to open near St James Town,Cabbagetownlooking at nearby venues it seems cluster 0 might be a good location asthere are not a lot of Indian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Indian restaurant in these locations.