IBM Applied Data Science Capstone Project

Opening a new Chinese food restaurant in the Los Angeles metropolitan area

Lin Zhu

10/28/2020

Introduction

The Los Angeles metropolitan area is the second-largest metropolitan area in the United States and is known for its ethnic diversity. Ethnic diversity also brings food diversity, especially for Chinese food. No matter American Chinese cuisine, or fast Chinese food like Panda Express, or different types of Chinese cuisine, Chinese restaurants have sprung up in this city. Chinese restaurants are mainly concentrated in areas where Chinese people gather. They are very popular with locals and tourists. Investors are also taking advantage of this trend to open more Chinese restaurants to cater to the demand.

Business problem

Now seems like the market is almost saturated. Consider the population is scattered in different parts of the Los Angeles metropolitan area. The main objective is to find a suitable location for opening a Chinese restaurant in the Los Angeles metropolitan area. There are many varieties of Chinese cuisine, such as Cantonese cuisine is sweeter, Sichuan cuisine is spicy. People who live in different areas will have different preferences. So we can take advantage of this to find possible opportunities. This project will take advantage of Foursquare data, to answer the question that, can we find an ideal spot and also recommend types of Chinese cuisine that could maximize profits to Chinese restaurant invertor?

Data

The data I will use in this project include:

- List of cities in the Los Angeles metropolitan area. This define the scope of this project which is confined to the Los Angeles metropolitan area. The link to this information is: https://en.wikipedia.org/wiki/Los_Angeles_metropolitan_area#Urban_areas_of_the_region
- Geographic information of those city (latitude and longitude)

- Venue data from Foursquare. Those data will be used to perform clustering and find the ideal spot for the new Chinese restaurant

Methodology

In this section, I will discuss and describe how I prepared and analyzed data, machine learning that I used, to solve this problem.

For data, firstly, I applied beautifulsoup python library to extract information about cities within the Los Angeles metropolitan area from Wikipedia page (https://en.wikipedia.org/wiki/Los_Angeles_metropolitan_area#Urban_areas_of_the_region).

However, the information I got from this page is just a list of the cities' names. In order to take advantage of Foursquare API, I need to get geographic information for those cities. To do so, I used the Geocoder python package that allows me to convert the address into geographical coordinates in the form of latitude and longitude. After gathering this information and populate them into a pandas Dataframe format, I visualized the neighborhood in a map using the Folium package. In this way, I can check if the geographic information is correct and make sure everything looks good.

After geographic data are collected, I moved to pull venue information from the Foursquare API process. I used the API the get the top 150 venues that are within a radius of 1500 meters. I passed the geographical coordinated in the form of latitude and longitude to call Foursquare API. The returns are venue data in JSON format. The information I need is the venue name, venue category, venue latitude, and venue longitude. Then I calculated the mean of occurrence of the frequency of each venue category for each neighborhood.

After all the needed information is collected, the next step would be to create a clustering model for the final analysis. Our goal in this project is to find an ideal location for a Chinese restaurant. Since Chinese cuisine has lots of types, I manually selected all types of Chinese restaurants based on the venue category which include Chinese Restaurant, Cantonese Restaurant, Dim Sum Restaurant, and Szechuan Restaurant. Consider it is an unsupervised

machine learning, I applied the K-means cluster model to group the data into four clusters based on the frequency of occurrence for all Chinese restaurants. In this way, we can see which neighbor has a high concentrate on Chinese restaurants and which types of Chinese restaurants are concentrate. That information could bring us an idea about the competition condition and where would be a good spot for opening a new Chinese restaurant.

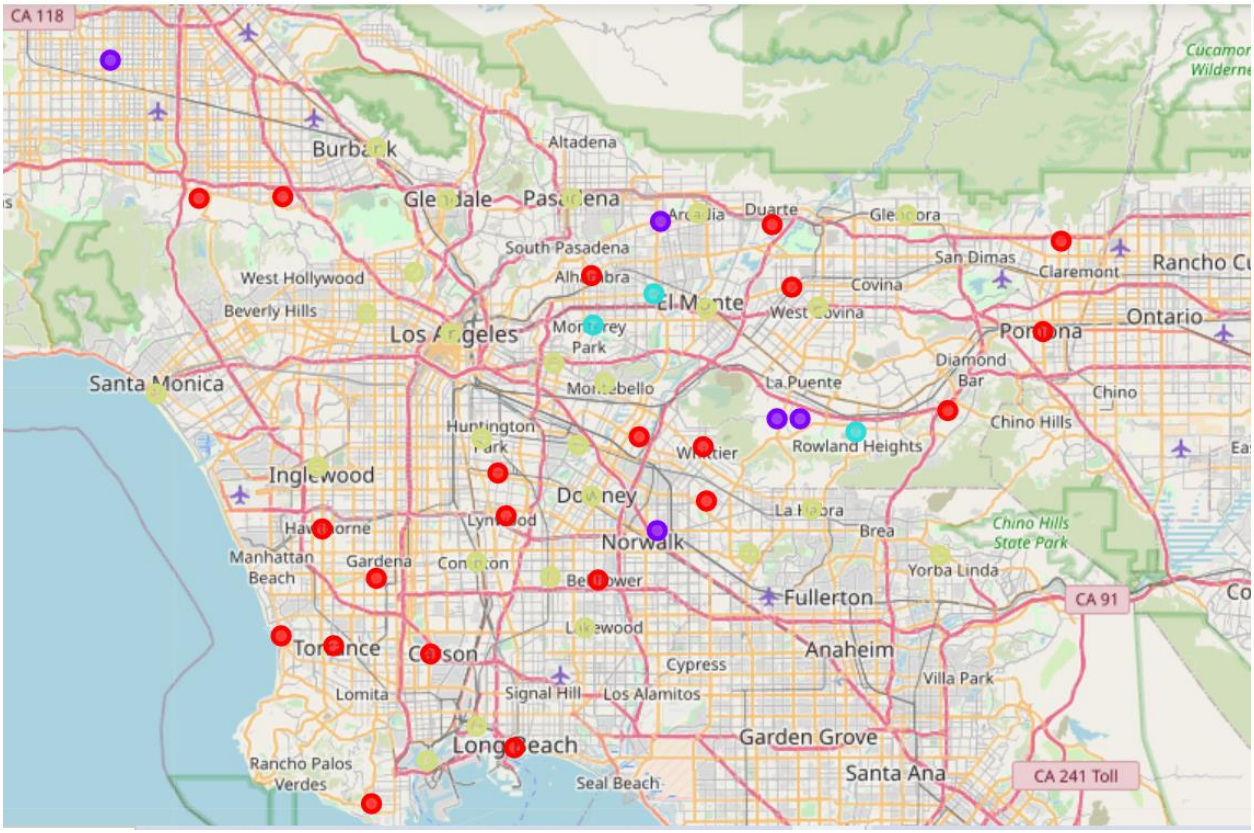
Results

After applied K-means Clustering, the neighborhoods are categorized into 4 cluster based on the frequency of occurrence for Chinese restaurant. From the clustering result, we can see that neighborhoods in cluster 0 has low number of Chinese restaurant, neighborhood in cluster 1 has moderate number of Chinese restaurant, neighborhood in cluster 2 has high concentrate of Chinese restaurant, and neighborhood in cluster 3 has very less or no Chinese restaurant.

Discussion

From the visualization attached below, most of the neighborhoods with the high and moderate concentrate of Chinese restaurants (cluster 1 and 2) are gathered on the east side of Los Angeles which means those area has more demands but also face higher competition. Let take a close look at each cluster. Neighborhoods in cluster 2 have a high concentrate of Chinese restaurants and I could find all four types of Chinese restaurants (Chinese Restaurant, Cantonese Restaurant, Dim Sum Restaurant, Szechuan Restaurant) here. So cluster 2 is likely suffering from intense competition due to the high concentrations of Chinese restaurants. Comparing neighborhoods in cluster 1 which has a moderate number of Chinese restaurants, I found only one place has more than one type of Chinese restaurant. Other neighbors in cluster 1 only have Chinese restaurants. We can say that those places have a high demand for Chinese food. Although still have intense competition, because of the single type of Chinese restaurants in those areas, investors can use the advantages of diversified restaurants to open other types of Chinese restaurants such as Szechuan Restaurant. Although neighbors in cluster 3 have very

low or no number of a Chinese restaurant and seem like a good opportunity, I need to consider if that area has demanding which means we need to do more investigations. In terms of cluster 0 where has a low number of Chinese restaurants, I think those areas could be a potentially ideal place as well since they have demand but low competition. Especially those neighbors located on the east side of Los Angeles.



This project has limitations for sure. For future research, I would suggest that firstly, consider the competition from other restaurant such as Japanese restaurants and Korean restaurants since they are also very popular types of Asian food. Secondly, I only used the free plan in Foursquare API which has limitation on number of API calls and number of results returned. So future research could try the paid account which could provide more information about venue in the target areas.

Conclusion

In this project, we identified the business problem, specified the data required, extracted and prepared data, and performed machine learning by clustering the data into 4 clusters based on their similarities. In the end, we could provide valuable suggestions to investors who want to open a Chinese restaurant in the Los Angeles metropolitan area and also discussed the limitation of this project. After analysis, investors could find an ideal spot in neighborhoods in both clusters 0 and 1. Cluster 0 has low competition and is good for opening a general Chinese restaurant, on the other hand, cluster 1 has more demands but could take advantage of the diversified restaurants to open specific types of Chinese restaurants such as Szechuan Restaurant or Dim Sum restaurants.