Capstone Project - The Battle of Neighborhoods (Week 2)

Prediction of the Japanese Food Restaurant in New York City

Date: August 15, 2020

Author: Anson Lo Kwong Shing

1. INTRODUCTION AND BUSINESS PROBLEM

In this Capstone Project, the assessment of the Japanese food restaurant will be implemented before launching a new Japanese food restaurant in New York City first. Due to the competition in this city, the comprehensive data analysis shall be implemented in order to understand the capability, price, categories (Japanese food restaurant or sushi restaurant) and rating for the Japanese Restaurant in New York City. The assessment can help understand what **business strategy** of Japanese restaurant shall be made (e.g. pricing strategy, decision for the acquisition of higher rating, location, etc.) before its new launch.

In order to investigate the market of Japanese restaurants in New York City, the number of Japanese restaurant for the assessment shall be as large as possible to have comprehensive information, particularly for Location, Rating and Price, the complete data set will be used for the data analysis of Japanese Restaurant. There are several parts to be performed in the analysis, including:

- Exploratory Data Analysis
- Clustering by K-Means
- Classification using Random Forest and XG Boost

The conclusion will be made after the subsequent works aforesaid and provide the findings and suggestions for the investor of Japanese Restaurant to provide a good competitive edge for its launch.

2. ACQUISITION OF DATA

The data will be collected from Foursquare API to obtain 50 numbers of Japanese food restaurant at most. The original features of the Japanese Restaurant in New York City are gathered by the Foursquare API as follows:

1. Name
2. Team
3. Categories
4. Address
5. Crossstreet
6. Lat
7. Lng
8. Labeledlatlngs
9. Distance
10. Postalcode
11. Cc
12. City
13. State
14. Country
15. Formattedaddress
16. Neighborhood
17. Id
18. Rating
19. Price

The search query of Japanese restaurant is adopted and the searching radius of 40000 will be used. The data will be cleaned and then some significant features will be extracted as well, covering Name, Categories, Distance, City, Rating and Price.

Prior to the acquisition of data from Foursquare API, the client ID and client secret shall be obtained in the Foursquare's website. Moreover, the version and limit shall also be provided for the connection of API.

3. METHODOLOGY

3.1 Exploratory Data Analysis

In regards to exploratory data analysis, the bar charts have been prepared for the analysis and listed below.

1. Rating versus Categories
2. Rating versus Price

The selection of bar charts can clearly show how the relationship between rating and categories as well as rating versus price which can provide obvious observation by graphic representation.

3.2 Folium Map

Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.

The use of Foilium Map can clearly show the locations of selected Japanese food restaurant acquired in Foursquare API. The latitude, longitude of the individual restaurant could be delivered to Folium Map and generate the spots where the selected restaurants are located.

3.3 Data Cleaning and Encoding for data analysis

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results.

As some rating and price cannot be identified in the Foursquare API, the vacant cell will be left because of the unavailability of rating and/price for Japanese food restaurants. Some data cleaning shall be made before the data analysis. The removal of the row will be made for the vacant cell in the data column of rating and price.

Moreover, the word cannot be handled for the input of data modelling. The encoding shall be implemented before training the model and predicting the result. Some features are applying to encode, covering:

1. Categories: Sushi Restaurant and Japanese Restaurant
2. City: New York, Forest Hills, Brooklyn, Staten Island, etc.
3. Price: Very Expensive, Expensive, Moderate, Cheap

3.4 Clustering by K Mean and Application of Elbow Method

Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is a collection of objects based on similarity and dissimilarity between them.

Two sets of clustering implemented by K Mean, including

1. Rating and Price
2. Categories, Rating and Price

For Rating and Price, 2-D clustering is performed and 3-D clustering is performed for Categories, Rating and Price. The cluster centres for two sets of clustering are projected at the same time.

In this analysis, Elbow Method has been used before implementing K-Means for cluster analysis, the elbow method is a heuristic used in helping determination of the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

3.5 Classification for the Problem

After the consideration of the problem, four features will be selected to predict the rating for the Japanese food restaurant, namely:

1. Categories
2. City
3. Distance
4. Price

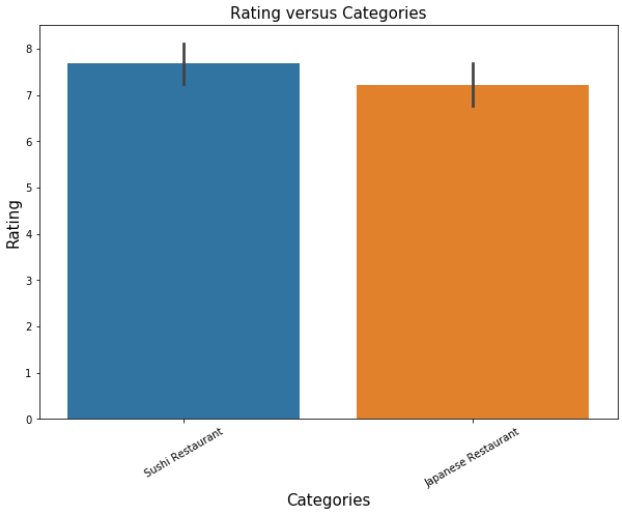
Before the modelling, the train test split shall be used to split the data set into training set and testing set. The training set is used for the training of the model and the remaining is the testing data to test the model. The test size of 0.2 is used for the train test split.

In regards to Classification problems, the algorithm of Random Forest and XG Boost will be used for the modelling and prediction. Tree-based algorithms are considered to be one of the best and mostly used supervised learning methods. Tree-based algorithms empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

4. RESULT

4.1 Exploratory Data Analysis – Bar Chart

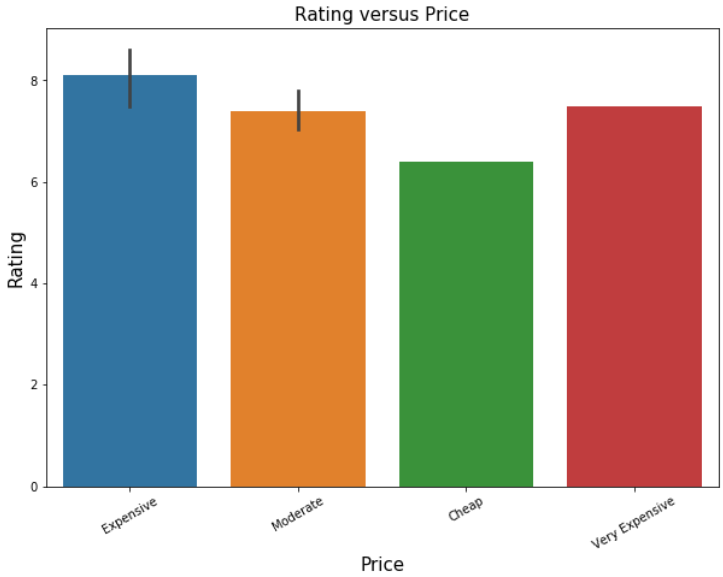
It is identified categories of sushi restaurants has a higher rating than Japanese restaurants. If the new restaurant would like to have a higher rating, it is better to strategically set the restaurant as a sushi restaurant.



The second Exploratory Data Analysis is to identify the relationship between rating and price. In fact, there are four types of price arrangement for the assessment, including very expensive, expensive, moderate and cheap.

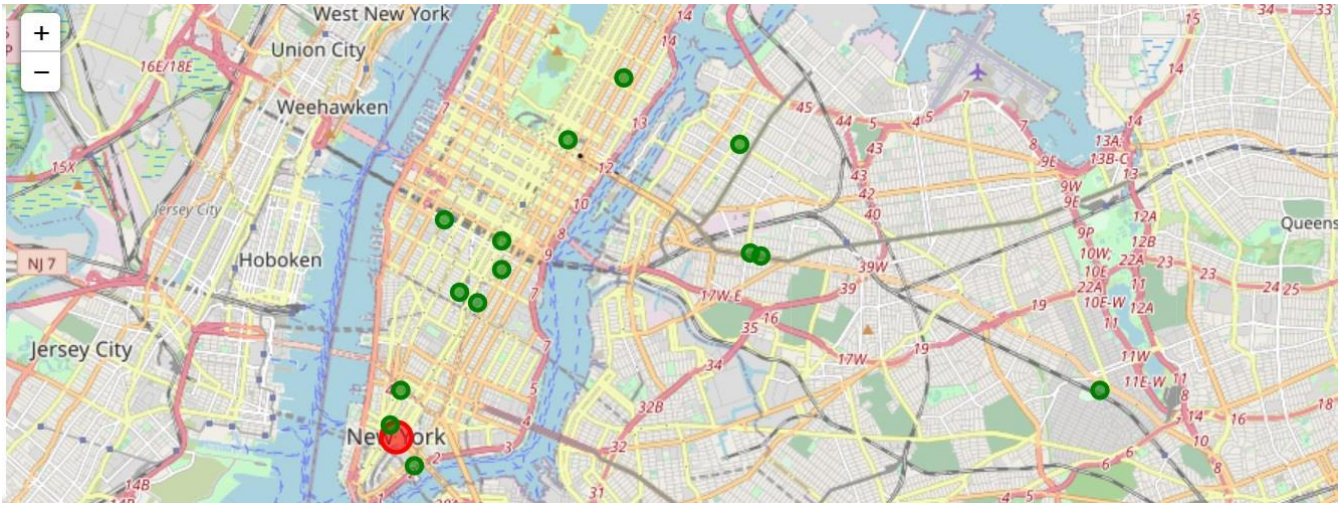
The result is listed from the high to low in terms of rating

1. Expensive
2. Very Expensive
3. Moderate
4. Cheap



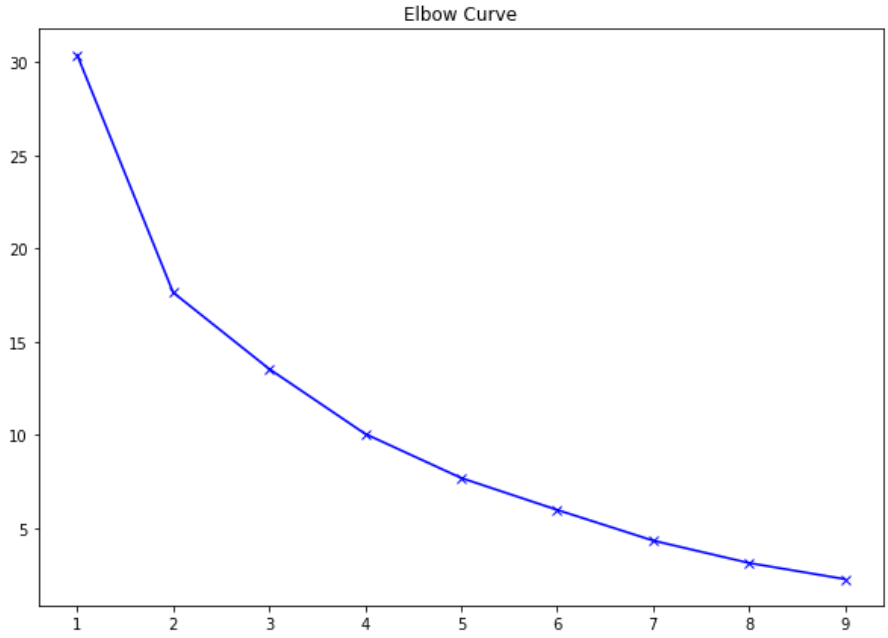
4.1 Exploratory Data Analysis – Folium Map

The Folium Map has been used for the location of Japanese food restaurant from Foursquare API. The diagram is provided below for information.

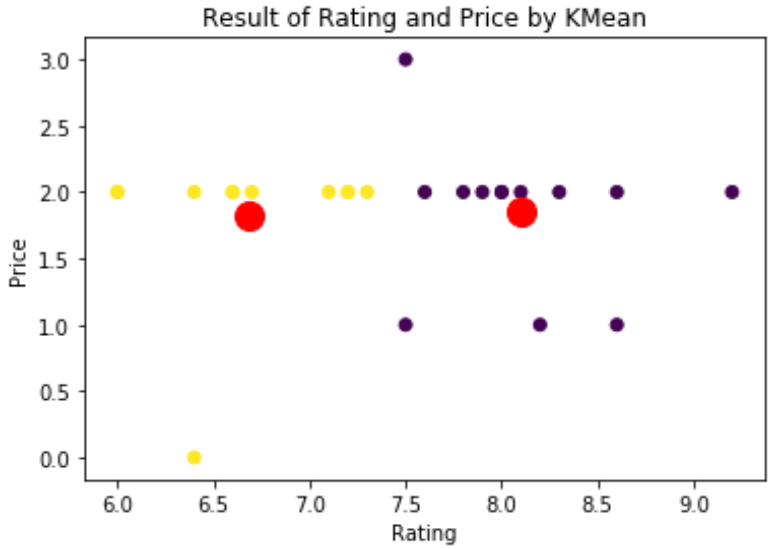


4.2 Clustering by K Mean and Application of Elbow Method

After the processing in Elbow Method, two numbers of clusters shall be preferred to have for clustering. Please refer the graph below for reference.



For the 2-D Clustering, the rating between 6 - 7.5 and cheap and moderate price are grouped and marked as yellow. In the meantime, the rating between 7.5 - 9.5 and moderate, expensive and very expensive price are grouped and marked as purple.



For the 3-D Clustering, the rating between 6 - 7.5 and cheap, moderate and expensive price are grouped and marked as yellow for both sushi restaurant and Japanese restaurant. In the meantime, the rating between 7.5 - 9.5 and moderate, expensive and very expensive price are grouped and marked as purple for both sushi restaurant and Japanese restaurant.

4.3 Classification by Random Forest and XG Boost

i. Result of Classification by Random Forest:

```
Scoring - RandomForest Classification: 0.8
Mean Absolute Error: 0.3
Mean Squared Error: 0.5
Root Mean Squared Error: 0.7071067811865476
Prediction of the rating for new Restaurant: 2.0
```

ii. Result of Classification by XG Boost:

```
Scoring - XG Boost Classification 0.7
Mean Absolute Error: 0.4
Mean Squared Error: 0.6
Root Mean Squared Error: 0.7745966692414834
Prediction of the rating for new Restaurant: 2.0
```

5. DISCUSSION OF THE OBSERVATION IN DATA ANALYSIS

5.1 Exploratory Data Analysis

It is interesting to identify the higher rating for sushi restaurant compared to Japanese restaurant. If the owner would like to have positive feedback and rating for the restaurant providing Japanese Meal, it is better to name its restaurant with sushi.

Moreover, the most expensive Japanese food restaurant does not mean to obtain the highest rating in Foursquare. The category of expensive Japanese food restaurant can acquire the highest rating. The sequence of the rating according to the category of Japanese food restaurant is listed below (from high to low).

1. Expensive
2. Very Expensive
3. Moderate
4. Cheap

If the owner wants to have a high rating in Foursquare, it is recommended not to set the price of food too expensive and suggest to set the price in the category of expensive. However, the quality of the food shall be good enough and customers can feel worthwhile to pay for it.

5.2 Clustering by K Mean

It is identified there are two groups for the customer segmentation and listed below.

First Cluster:

- Rating between 6-7.5
- Cheap and Moderate price

Second Cluster:

- Rating between 7.5-9.5 and
- Moderate, Expensive and Very Expensive price

The owner can now clearly know the positioning of the Japanese Food Restaurant. If the owner would like to apply high pricing strategy for the Japanese Food the restaurant provides, the high rating shall be achieved.

5.3 Classification by Random Forest and XG Boost

For the models I have set up, it can achieve a moderately high accuracy by Random Forest and XG Boost (0.8 and 0.7). However, due to the limited access of Foursquare API, it is not possible to acquire more than 50 rows of information for Japanese restaurants in New York City which would be the sad news to know it.

Besides, the tentative location is identified with the following features, namely

1. Categories: Sushi Restaurant
2. City: New York City
3. Distance: 1000m
4. Rating: 7

The prediction made by Random Forest and XG Boost provides the result that the price shall be set in Expensive (2) for the rating of 7. The arrangement shall be established for the pricing and marketing strategy. Meanwhile, the scoring of the model by Random Forest and XG Boost is acceptable which are 0.8 and 0.7, even the size of the dataset is not big enough.

6. CONCLUSION

In this data analysis, several techniques have been adopted including Exploratory Data Analysis, Clustering and Classification by Random Forest and XG Boost. Before the data analysis, the preprocessing of data has been implemented before fitting the model, data cleaning and encoding have been performed before data analysis.

Clustering shows the grouping of the data for the rating, price as well as category of the Japanese food restaurant. It is identified the restaurant shall have the rating over 7.5 for the expensive and very price in the product.

Classification by Random Forest and XG Boost can provide a good price prediction for the input of categories, city, distance, rating. The prediction of price based on the specific categories, city, distance and rating could be successfully performed for the consideration of establishing a new Japanese Food Restaurant.