Submitted by: Eric Ng Date: May 26, 2019

Applied Data Science Capstone Project – Week 1

Finding "Michelin Quality" Restaurants in Hong Kong

I. Introduction and Background:

Michelin is probably one of the most influential companies in the restaurant industry in the world. Originally a company that sells tires, it started assigning ratings and awards to restaurants since 1926 in France, while putting together a guide which helps drivers find good food while they travel. The ratings are conducted by Michelin's own full-time food critics, who are passionate about food, have good eye for detail, and have a great taste memory to recall and compare types of foods. For many years now, the Michelin awards are most coveted by restaurant owners around the world, since being on the Michelin list will certainly bring prestige as well as more customers. The award system has changed over time, and currently the ratings are given as follows:

- **Bib gourmand**: A place that high-quality food and services and good value for the money. Bib is short for Bibendum the character of Michelin's logo.
- **One star:** A good place to stop on your journey, indicating a very good restaurant in its category, offering cuisine prepared to a consistently high standard.
- **Two stars:** A restaurant worth a detour, indicating excellent cuisine and skillfully and carefully crafted dishes of outstanding quality
- **Three stars:** A restaurant worth a special journey, indicating exceptional cuisine where diners eat extremely well, often superbly. Distinctive dishes are precisely executed, using superlative ingredients.

With the advance of mobile computing in the past 10 years, ratings to any restaurant can now be conducted by anyone conveniently through Apps such as Foursquare. Now that we have aggregated ratings and stats offered by these Apps, one may ask how the data would compare to those conducted by the Michelin judges. Is there opportunity to have the regular user data and the professional Michelin data leverage each other so it can help customers pick good restaurants?

II. Issue and Opportunity:

In every metropolitan city, there are thousands of restaurants. Being just one single company, Michelin is limited by its coverage. There are only so many restaurants a Michelin judge can go visit every year in a particular city, and as a result, many great restaurants may not ever get discovered or graded by Michelin. Now, data provided the mass and aggregated by Foursquare may help. Foursquare covers thousands of restaurants in town and possess many stats. Based on these stats, we can try to discover the special characteristics of the restaurants that had earned a Michelin award. After learning their special characteristics, we can then predict whether a non-Michelin restaurant possess "Michelin quality" too.

III. Objective – Predict "Michelin Quality" Restaurants

We would like to train a classification model that is able to identify bewteen a "Michelin quality" and a "Not Michelin quality" restaurant. Using this trained model, we will then generate a list of restaurants that is currently not awarded by Michelin; yet has a high probability that their overall quality is comparable to those that have already earned the Michelin award.

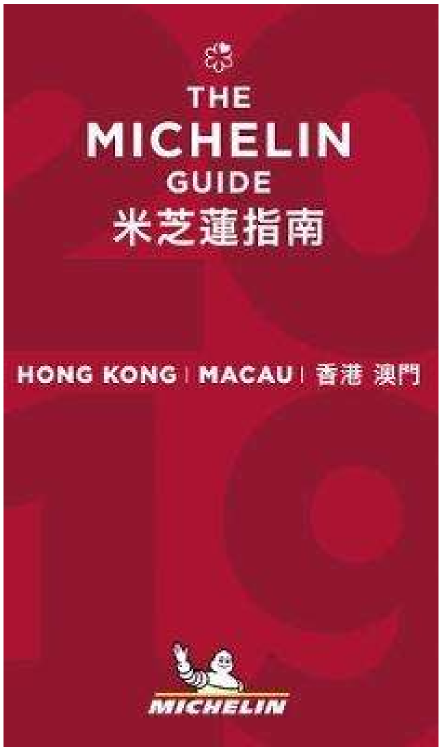
We can then advise customers about this list, so they have more choices beyond the Michelin list. Alternatively, we can also provide this list to Michelin so it may sends its food critics to these restaurants for a visit, just in case they missed them in their radar.

IV. City of choice and defining "Michelin Restaurants"

Being a Hong Kong citizen, I choose to use Hong Kong as the city of study, since I am more familiar with the restaurants and food being offered. I may also be able to make sense of the data and the result better. Hong Kong has a decent number of Michelin winners close to that of New York, so I shall have a workable sample size that helps me train the model.

Quantity of the Top 10 Most Michelin Starred Cities (including 1, 2, and 3 stars):

1. Tokyo, 230
2. Paris, 123
3. Kyoto, 103
4. Osaka, 97
5. New York, 76
6. London, 69
7. **Hong Kong, 63**
8. Singapore, 39
9. San Francisco, 38
10. Barcelona, 25



Besides the starred restaurants, we will also add the 71 Bib gourment restaurants to the list, giving us a total of 134. The model will be trained irrespective of whether the restaurants is 1 star, 2 star, 3 stars, or Bib Gourment. They are all grouped into the "Michelin Restaurants" category. These restaurants will be assigned with a value "1", while restaurants that did not earn the Michelin award will be assigned "0". The 1 and 0 will be the value of our "y", the dependent variable in our classifciation model.

V. Data sourcing and Pre-processing

The data sourcing is straight forward, but the data cleaning is challenging especially when the Chinese language is involved.

1. Michelin winners data
   Starred restaurant data from Michelin website –
   https://guide.michelin.com/hk/en/hong-kong-region/hong-kong/restaurants/3-stars-michelin/2-stars-michelin/1-star-michelin
   Bib gourment restaurant data from Michelin website –
   https://guide.michelin.com/hk/en/hong-kong-region/hong-kong/restaurants/bib-gourmand

   The data we will obtain include:
   a. Name of the winning restaurants
   b. Coordinates of the restaurants

   With these 2 pieces of data, we try to find them in Foursquare database, then learn about their venue ID's. Only when we have the venue ID's can we pull the ratings and stats from Foursquare.

2. Find Venue ID of Michelin Restaurants in the Foursqure database
   The coordinates of the restaurants used by Michelin and Foursquare do not match. There are slight differences, and therefore, we will use the explore function in the Foursquare API to find the nearby venues using Michelin's coordinates, and hopefully find the matching restaurants.

3. Name matching – manual work required due to language issues
   In the Foursquare database, the names of the restaurant are also mostly inconsistent with that of Michelin. Michelin use English only for names, but in Foursquare, the language used is in consistent. The names can be:
   a. English only
   b. English + Chinese
   c. Chinese only
   d. Simplified Chinese and Traditional Chinese are also used interchangeably

   Because of this, the data matching will have to be done by human manually on a spreadsheet instead of via python programming. Luckily there is only just over 100 names to go through.

4. Find Venue ID of Non-Michelin restaurants

We also need data from non-Michelin restaurants. To obtain their venue ID's, we will leverage on the same API calls which we used to explore the coordinates provided by Michelin. We will screen out all restaurants nearby, from which we will obtain a total of about 400 to 500 non-Michelin restaurants which we can use for model building.

5. <u>Find ratings and stats of Michelin and Non-Michelin restaurants</u>
   Now that all Venue ID's are available, we will use the venue_ID API call to acquire details for each restaurant. Since we are using a free package, we only have visibility to limited type of data. For example, we are not able to get the no. of dislikes and no. of total check-ins. However, there are still a number of useful data we may use as the independent variables (x) for our model:
   a. Ratings
   b. Number of likes
   c. Number of tips
   d. Number of photos
   e. Number listed (something like my favorite)
   f. Price level (1, 2, 3 to choose from, where 3 being most expensive)

   Area and type of cuisines of the restaurants are discrete data which cannot be used in our model building.

VI. Model Selection and Building

There two approaches being considered for this classification problem.

1. K-means clustering
   We try to cluster the restaurants into a few groups, and see which group includes the most Michelin Restaurants in it. If the cluster has a high share of Michelin Restaurants, then all the restaurants included in the cluster, with or without a Michelin award, are likely to be of the same Michelin quality due to their closeness and similarity.
2. Logistics regression
   This model provides a probability value, and such value will be realistic and useful when one decides whether a particular restaurant is highly likely to be of Michelin quality. However, its usefulness would depend on the soundness of the model after the analysis.

In any case, a list of "Michelin quality" restaurants that have not yet earned a Michelin award will be generated using the selected model.

VII. Key Assumption and Consideration

Here, we make a major assumption that the Michelin judges give the most reliable and accurate score to the restaurants. Some may also argue that the Michelin professionals have a "noble tongue" that is not representative of the laymen. But since the Michelin guide targets the regular consumers rather than a small group of food professionals, we should be confident that Michelin ratings are also well adjusted for the audience, and therefore not so far away from the tastes of the Foursquare users.