California Restaurant "Likes" Prediction Using Foursquare API and Machine Learning

By : Khalid.Meguenni
Capstone Project
IBM Data Science Professional Certificate

1. Introduction

California boasts an incredibly diverse collection of restaurants catering to different palettes and appetites. A large part of marketing for a modern restaurant (or any company) is social media, where the number of "likes" that the company can receive will dictate its brand and image to the general public.

For a new business owner (or existing company) to open a new restaurant in California, knowing ahead of time the potential social media image they can have would provide an excellent solution to the ever present business problem of uncertainty. In this case the uncertainty is regarding performance of social media presence.

We can mitigate this uncertainty through leveraging data gathered from FourSquare's API, specifically, we are able to scrape "likes" data of different restaurants directly from the API as well as their location and category of cuisine. The question we will try to address is, how accurately can we predict the amount of "likes" a new restaurant opening in this region can expect to have based on the type of cuisine it will serve and which city in California it will open in. (For the purposes of this analysis, we will contain the geographical scope of analysis to three heavily populated cities in California, namely San Francisco, Los Angeles, and San Diego).

Leveraging this data will solve the problem as it allows the new business owner (or existing company) to make preemptive business decisions regarding opening the restaurant in terms of whether it is feasible to open one in this region and expect good social media presence, what type of cuisine and which city of three would be the best. This project will analyze and model the data via machine learning through comparing both linear and logistic regressions to see which method will yield better predictive capabilities after training and testing.

2. Data

2.1 Data Scraping and Cleaning

In this section we will first retrieve the geographical coordinates of the three cities (San Francisco, Los Angeles, and San Diego). Then, we will leverage the FourSquare API to obtain URLs that lead to the raw data in JSON form. We will separately scrape the raw data in these URLs in order to retrieve the following columns: "name", "categories", "latitude", "longitude". and "id" for each city. We can also provide another column ("city") to indicate which city the restaurants are from.

It is important to note that the extracts are not of every restaurant in those cities but rather all of the restaurants within a 1000KM range of the geographical coordinates that geolocator was able to provide. However, the extraction from the FourSquare API actually obtains venue data so it will include venues other than restaurants such as concert halls, stores, libraries etc. As such, this means that the data will need to be further cleaned somewhat manually by removing all of the non-restaurant rows. Once this is complete, we have a shortened by cleaned list to pull "likes" data. The reason the cleaning takes precedence is mainly that pulling the "likes" data is the computing process which takes the longest time in this project so we want to make sure we are not pulling information that will end up being dropped anyways.

The "id" is an important column as it will allow us to further pull the "likes" from the API. We can retrieve the "likes" based on the restaurant "id" and then append it to the data frame. Once this is complete, we finally name the dataframe 'raw_dataset' as it is the most complete compiled form before needing any processing for analysis via machine learning.

2.2 Data Preparation

The data still needs some more processing before it is suitable for model training and testing. Mainly, the "categories" column contains too many different types of cuisines to allow a model to yield any meaningful results. However, the different types of natural cuisines have natural groupings based on conventionally accepted cultural groupings of cuisine. Broadly speaking, all of the different types of cuisine could be reclassified as European, Latin American, Asian, North American, drinking establishments (bars), or casual establishments such as coffee shops or ice cream parlours. We can implement manual classification as there really aren't that many different types of cuisines.

As this project will compare both linear and logistic regression, it makes sense to have "likes" as both a continuous and categorical (but ordinal) variable. In the case of turning into a categorical variable, we can bin the data based on percentiles and classify them into these ordinal percentile categories. I tried different ways of binning but in the end, splitting the sample into three different bins proved to yield the best classification results from a prediction standpoint.

As the last stage of data preparation, it is important to note that the regresses are categorical variables (3 different cities and 6 different categories of cuisines). Hence, they require dummy variable encoding for meaningful analysis. We can accomplish this via one-hot encoding.