Using Data Science to solve Business Challenges in Restaurant Industry

Sheetal Joshi

6th April 2020

Contents

[Introduction 2](#_Toc37158382)

[Data Collection, Cleansing and Preparation 2](#_Toc37158383)

[Methodology 3](#_Toc37158384)

[Results 3](#_Toc37158385)

[Discussion 4](#_Toc37158386)

[Conclusion 5](#_Toc37158387)

# Introduction

There is stiff competition in the restaurant industry. In larger cities, people dine out very often and they have a lot of food options to choose from. Right from the choice of world cuisine to the type of restaurant ambience to the cost, there are multiple factors that can make a restaurant popular.

#### Business Problem Description

It has become increasingly difficult for new Restaurant owners to select the ‘type’ of restaurant cuisine or location to ensure business success. They cannot rely merely on intuition and need solid analysis of facts to make their decisions.

Restaurant goers have started depending heavily on Social media recommendations. The likes, reviews or ratings of a restaurant are an important success factor for any restaurant.

Hence, can Business owners leverage Social Media to solve practical business questions like –

1. **Given a choice of Multiple Cities, which city may be the best for a new restaurant?**
2. **Within the city selected above, which type of cuisine will be a successful proposition?**

#### Possible Solution

Foursquare has immense data on restaurants, their locations and performance.

We can leverage the Foursquare API to compare performance (ratings/reviews) across cities and cuisines. 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.

# Data Collection, Cleansing and Preparation

We will collect the Foursquare Restaurant data from the 3 cities – San Francisco, Los Angeles and San Diego.

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 the restaurants within a 1000KM range of the geographical coordinates that geolocator was able to provide. However, the extraction from the Foursquare API 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 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.

# Methodology

This project will utilize both linear and logistic regression machine learning methods to train and test the data.

Linear regression will be used to predict the number of "likes" a new restaurant in this region will have. The Sci-Kit Learn Package will be used to run the model.

We can also utilize logistic regression as a classification method rather than direct prediction of the number of likes. Since the number of "likes" can be binned into different categories based on different percentile bins, it is also potentially possible to see which range of "likes" a new restaurant in this region will have.

Since the "likes" are binned into multiple (more than 2) categories, the type of logistic regression will be multinomial. Additionally, although the ranges are indeed discrete categories, they are also ordinal in nature. Therefore, the logistic regression will need to be specified as being both multinomial and ordinal. This can be done through the Sci-Kit Learn Package as well.

# Results

**Linear Regression**

A linear regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%. To see if this is a reasonable model the residual sum of squares score, and variance score were both calculated (45206.57 and 0.20 respectively).

Given the low variance score, this is probably not a valid/good way of modelling the data. Therefore, we move on to logistic regression.

**Logistic Regression**

A multinomial ordinal logistic regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%. To see if this is a reasonable model, its Jaccard similarity score and log-loss were calculated (48.48% and 1.06 respectively). Although this is not a perfect prediction, a similarity of 48% between the training set and test set is an acceptable result.

Given the modestly accurate ability of this model, we can also run the model on the full dataset. The coefficients show that following business propositions may run into Negative Likes

1. Restaurants with an American, Asian or Latino Cuisine
2. A new bar
3. Restaurant in San Francisco

# Discussion

The first thing to note is that given the data, logistic regression presents a better fit for the data over linear regression. Using logistic regression, we were able to obtain a Jaccard Similarity Score of 48.48%, which although not perfect, is more reasonable than the low variance score obtained from the linear regression. As stated before, please note that for the purposes of this project, we are assuming that likes are a good proxy for how well a new restaurant will do in terms of brand, image and by extension how well the restaurant will perform business-wise. Whether or not these assumptions hold up in a real-life scenario is up for discussion, but this project does contain limitations in scope due to the amount of data that can be fetched from the Foursquare API.

As such, to obtain insights into this data, we can proceed with breaking down the results of the logistic regression model. The results showed that the precision score for classifying whether the new restaurant would fall into classes 1, 2, or 3 (lowest, medium, or highest percentile of likes) were **50%, 25%, and 62%**. Therefore, the model is better at predicting if a restaurant will fall into the best or worst percentile of likes.

Additionally, not only are we attempting to predict the general business performance but also pull insights to inform on business strategy. In this case strategy insight can be gleamed from the coefficient values from running the logistic regression on the full dataset. As such, we can see that opening a restaurant with Restaurants with an American, Asian or Latino Cuisine, or a bar or a restaurant in San Francisco are associated with Negative Likes.

This suggests that the business opportunity should be opening a restaurant in either Los Angeles or San Diego, with a cuisine that is European, or casual in nature would be the best approach for maximizing likes.

# Conclusion

In conclusion, after analyzing restaurant "likes" in California from 300 restaurants, we can conclude that the approach to best take in regards to maximizing business performance (as measured by "likes") is to open a restaurant that is either European, Latino, or casual and that opening the venue in either Los Angeles or San Diego rather than San Francisco would be the best approach.