Additional data analysis for choosing the location of a new Chinese restaurant in Los Angeles, Ca.

Arshak Mkhitaryan
Data Scientist

January 2020

The business problem

In no more than 48 hours the CEO of a Chinese restaurant business is going to pitch a project to the board of stakeholders. The agenda of the project is related to expanding the business by opening another restaurant at a new location in Los Angeles.

But how to choose a location for a new venue?

She approaches me as the data scientist of the company with a request for additional insight and visualization on the matter. Specifically, she is interested in the mapping and clustering of potential regions within the city.

The final problem can be summarised as follows: cluster regions in Los Angeles based on information regarding Chinese restaurants and present relevant insight and visualizations that can be incorporated into the final presentation for the board of executives in. The results should be presented to the CEO in 24 hours so that she has an adequate amount of time to incorporate them into her presentation.

Background

How does one find the best location for a venue?

The problem of finding the best location for a new venue might seem trivial at a glance but it's not. There are many criteria to consider to support the decision. Some of them are as follows:

1. The number of similar restaurants in the neighborhood.
2. The house prices in the neighborhood.
3. The average income levels of the residents.
4. The average percentage of people eating out in the area.
5. The trends and migration patterns in the area that might be taken as a proxy for the demand in a particular cuisine.
6. The number of business centers in the area.
7. The history of restaurants that closed or prospered in the past and the reasons behind their failure or success.
8. etc.

And the list goes on and on. As we can see, it is obvious that a comprehensive prediction of success of a venue based on the location and subsequently an effective choice of such locations requires a large amount of data and complex data analysis as well as a deep understanding of underlying business processes.

That is why it is very important to define the limits of the given problem and possible solutions right away.

Nonetheless, it's worth undertaking quick research providing additional insights into the best practices and the state of the art. During my initial research, I found two main approaches frequently applied to the problem in question as well as combinations of those.

The first approach heavily relies on domain knowledge and business expertise. The ability of the business team to "sense" best locations relying on their experience.

The second approach that is becoming more and more popular with these sorts of problems is using Geolocational data of individuals together with locational data of venues to extrapolate information such as clustering customers into meaningful groups by their preferences and then matching those clusters with clusters of venues.

It's worth mentioning that both approaches are very similar to each other in the fact that they require an enormous amount of information. In one case that information is gained by humans through years of hands-on experience of opening new locations and running businesses. In the other, it's the amount of data collected in time and extrapolated using unsupervised machine learning techniques such as the Travel Time Factorization Model (TTFM)[1].

What makes it so hard?

My current understanding is that the problem is complicated not only by the amount of information needed to choose the location but also because of the Dynamically changing nature of the data.

Let's, for example, try to answer these types of questions. How our venue will be affected by the economic changes in the region? How would it be affected by another restaurant closing or opening nearby? How would it be affected by migration patterns within the area, such as the number of people that might be interested in visiting our new Chinese restaurant?

Having all of the complexity of the problem and the limitations of time and resources we first must define the scope and depth of our analysis.

In this particular case, the business decision would be mostly based on the expertise of the business team and sparingly supplemented by our analysis and visualization.

---

[1] For more information check out this study https://arxiv.org/abs/1801.07826 or this blogpost https://blog.safegraph.com/opening-a-new-restaurant-ditch-the-guesswork-and-turn-to-machine-learning-5276f44dd408.

Interest

Business owners, as well as upcoming entrepreneurs, will find the reports helpful in deciding on the best location for a Chinese restaurant in Los Angeles. Although this analysis is not providing the final answer it is a good informational supplement.

Data

The data used in the analysis are as follows:

1. The list of notable districts and neighborhoods of the city of Los Angeles, California.
2. Source: https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles

3. This list will be scraped and cleaned in Python using the Beautiful Soup library and then stored in a Pandas Dataframe.

4. The latitude and longitude coordinates of each neighborhood will be obtained using Geocoder and then stored in a Pandas DataFrame.

   Location data from Foursquare.
   Source: https://foursquare.com/
   Data will be acquired through a developer's account using Places API and then stored in a Pandas Dataframe.

All of the Data above will be merged, processed and used to fit the K-means clustering machine learning algorithm.

The clusters then will be visualized using the Folium library.