The Battle of Neighborhoods

Applied Data Science Capstone

New restaurant location

Introduction

The problem

Choosing a location for a new restaurant (or other any other business) is crucial for the success of the business

To anyone planning to open a new business it is crucial to make a data based, strategic decision on a new business location

However, many different neighborhoods have a lot of different factors differentiating them from each other, therefore we need a systematic data science approach

The background

A few factors to consider are:

- *Foot traffic- what areas are generally heavily frequent area*
- *Feeder traffic- area there other businesses in the area that attract people that are also in need of your service (food)*
- *Proximity to areas where people live or work*
- *Parking or public transportation*
- *Competition- are there a lot of other businesses offering the same service in a given area*
- *Many others….*

Data

Description of data

I will use data from Foursquare:

Queries of businesses around existing restaurants

I will use an area in the east of the San Francisco bay area. While less densely populated area than the city of San Francisco, it is still highly populated and therefore represents a good market for a restaurant. However, because of the decreased density car travel is most common and the selection of a location becomes more crucial for an emerging business to gain popularity.

Usage of data

To determine which factors are relevant to locate a successful new restaurant I will analyze what businesses are most prevalent in proximity to existing restaurants

Proximity to other restaurant of the same kind will likely be a penalizing factor as that might represent unnecessary competition (e.g.: when opening a Thai restaurant right next to another Thai restaurant). On the other hand other types of restaurants might be a positive factor in moderation. Other businesses that drive entertainment traffic (e.g. movie theater, Mini-golf, etc.) will most likely factor in positively.

I will use data from existing as a neighborhood conducive to restaurant business and use k-means to identify actual neighborhoods that are most similar to this profile to determine which neighborhood is most similar to the environment restaurants are usually found in.

Methodology

First I will find out which venue categories are most commonly present in proximity to restaurants.

Start with a query by using the GPS coordinates with a large radius and restaurants.

Then start another query with each of the identified restaurants to get close by venues

Then analyze the neighborhoods and make a data frame with relative frequencies and transform

Add a line with the restaurant's most common venue neighbors and use k-means to find out which neighborhood it clusters with.

This should be the most favorable neighborhood based on the data I fed in.

Results

The most common neighbor or restaurants are other restaurants, followed by bars, grocery stores, pharmacy and coffee shops.

The neighborhoods that are most similar to the environment existing restaurants are in are Farrelly Pond, Floresta Gardens and Mulford Gardens

Discussion

There are a few limitations with this approach, some related to the amount of data available with the free developer account. Here are a few that I came across and possible ways to address them. Most of the issues mentioned would require a time investment that goes well beyond this course:

- The methodology relies largely on the assumption that existing restaurant owners have made a good location choice. There's no data included on whether the restaurants are actually doing well financially, so that might not be the case. Furthermore, established restaurants might be doing well because they are already well known and clients are willing to travel there. The same might not be true for a new business.
- Since every analyzed restaurant requires a new query I have limited my data. This results in less accurate results, however since this is a learning experience and can be fixed by just changing the radius parameter I prefer to keep the functionality of several queries (things tend to go wrong in the beginning)

- The amount of businesses of one category ignore the size and draw of each business. For example: a big movie theater (which does not appear on the list of most frequent neighbor) will likely draw a lot more potential clients that several small restaurants (top frequent neighbor). This one is a little harder to fix, Google has some data on how busy a business is (google maps will display real time data) but I don't believe it's publicly available. Foursquare has the "trending" function but that doesn't give any absolute numbers I believe. So it would be technically possible to improve on this point but the data might not be readily available (for free- this seems like valuable data to have)
- Since the Foursquare interface just works of GPS coordinates and a fixed radius there's bound to be some overlap in the neighborhoods. With some research it would be possible to adjust the radius for each neighborhood as a first approximation. A more precise way would be to make a grid within the neighborhood and iterate through the GPS coordinates with smaller radii.
- Not all factors that come up high on the list might be relevant for the success of a business. With some data of financial success it might again be possible to tease out which are more likely to be causational or just correlational.
- There's no data integrated into this model speaking to the cost of doing business in a certain area. Some real estate data might improve this model, since rent is likely one of the more variable costs for areas in close proximities like neighborhoods within the same city.
- The density of businesses is overall lower than expected. An expansion of the analyzed area will probably lead to more accurate results. This can be easily accomplished with more Foursquare queries.

Conclusion

Chefs interested in opening a restaurant should consider the Farrelly Pond, Floresta Gardens and Mulford Gardens areas