Finding the best location for a vegan restaurant in Los Angeles

Introduction/Business Problem

A vegan diet is being adopted by more and more people in the USA, and the demand for vegan restaurants is correspondingly increasing. This is particularly true in large multicultural cities such as Los Angeles. Hence, this capstone project is tailored to stakeholders who are interested in opening a vegan restaurant in Los Angeles. We want to use data to answer the following question: What is the best location to open a vegan restaurant in LA?

Data sources

Using the Foursquare API (https://developer.foursquare.com/places) we can retrieve information about restaurants in Los Angeles, including location, neighbourhood characteristics (e.g. other venues that are within walking distance), rating, number of tips and type of cuisine. In addition, we will use the Yelp API (https://www.yelp.com/fusion) to find "vegan" venues, because the Foursquare category system only returns a few venues categorized as "vegetarian / vegan", even though there are hundreds according to Yelp or the HappyCow App.

To map out districts in Los Angeles we can scrape borough names from the LA-times website (https://maps.latimes.com/neighborhoods/neighborhood/list/) and use the geocoder python package to find corresponding geospatial coordinates. Using these coordinates, we can search on Foursquare and Yelp to find surrounding venues, including vegan restaurants.

For example, we can search for venues in Hollywood in Los Angeles using the explore Foursquare querry and find the surrounding venues and their characteristics. Based on the surrounding area characteristics (e.g. lots of food venues or close proximity to metro stations), we can predict vegan restaurant success as measured by both user ratings and number of tips (suggesting that the restaurant is popular).

Finally, we can predict the best locations to open a new vegan restaurant based on those area characteristics.

Methodology

I started by scraping Los Angeles neighborhood names from the LA-times website and obtained geospatial coordinates using the geopy package. Most neighbourhoods' geospatial coordinates were returned correctly, but a few were not found or returned incorrectly and had to be adjusted or removed. Considering that the list of neighborhoods from the LA-times website was somewhat arbitrary anyway, and more than 200 neighborhoods with correctly returned coordinates remained, I chose to simply exclude neighborhoods with missing or erroneous coordinates (the final neighborhoods can be seen in Figure 1).

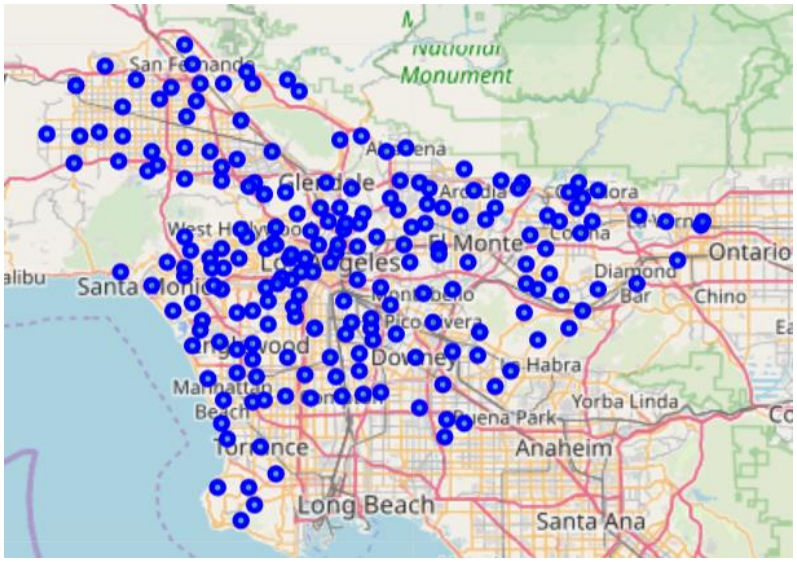

Figure 1. Neighborhoods in LA as specified by the LA-times.

Next, I tried to find all vegan restaurants in LA by using the explore command with the Foursquare API. However, Foursquare returned less than 50 restaurants in all of LA with the category label “vegetarian / vegan”, which is i. a small number of restaurants and ii. not even conclusively vegan. Instead, therefore, I used the Yelp API to find “vegan” restaurants for each neighbourhood, obtaining a few hundred venues (Figure 2).

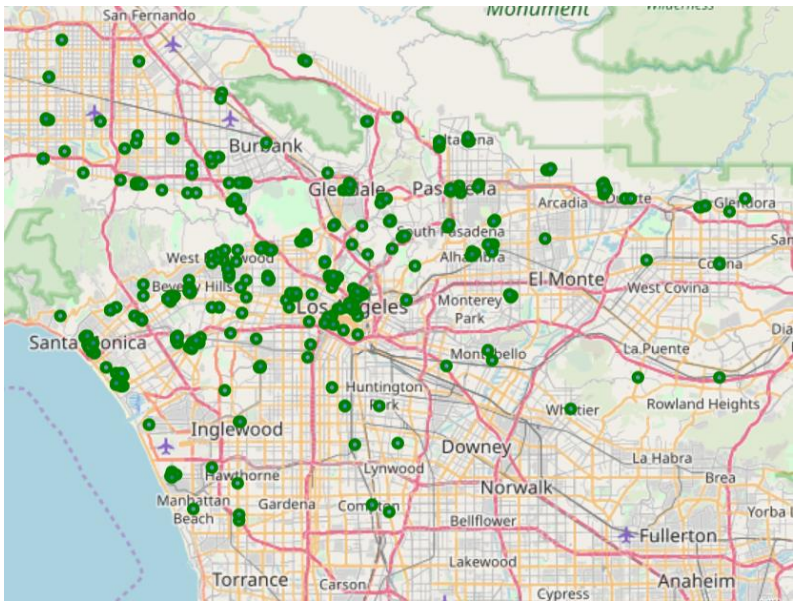

Figure 2. Vegan venues found using the Yelp API.

Next, I searched for the same venues on Foursquare to obtain Foursquare venue ID, category, rating, pricing, tipping, photo and likes information. As the Foursquare category system contains few “vegetarian / vegan” labels, it might be interesting to see what labels it assigns to the Yelp “vegan” labelled venues. It turns out that most venues are coffee shops. Moreover, some venues such as "7-Eleven" or "Aldi" should probably not be included in our analysis, as they hardly compare to a vegan restaurant. After cleaning and combining certain labels (e.g. café and coffee shop) 394 venues remained (Figure 3).

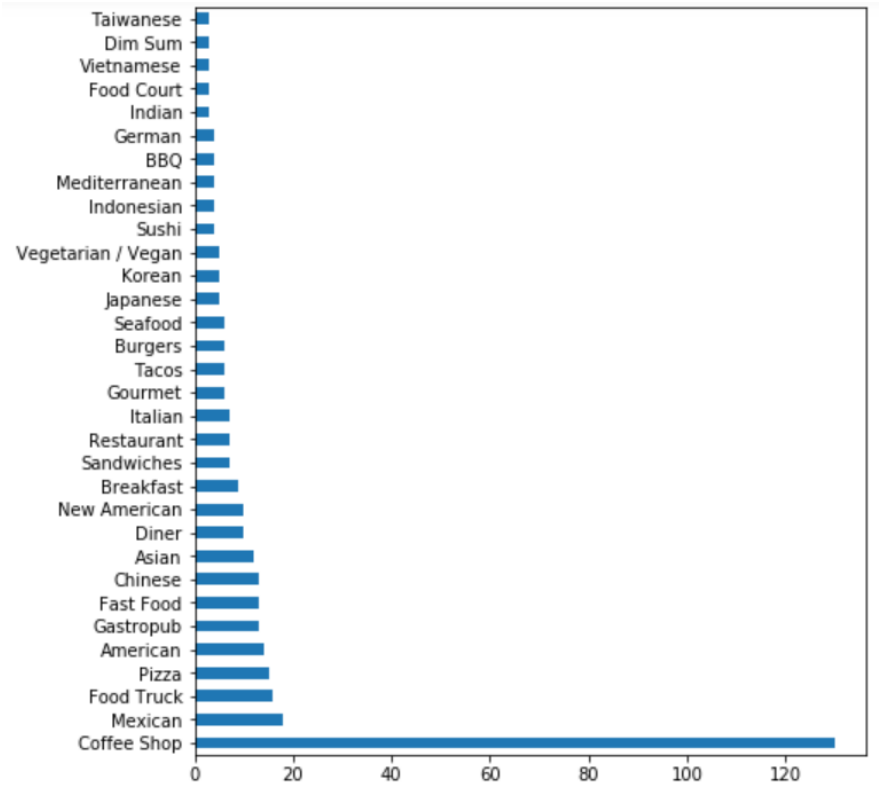

Figure 3. Number of vegan venues in LA by Foursquare venue label (min. number of venues = 3).

To assess what user related features (e.g. likes or rating) are useful to keep for further analysis and combine into a useful "popularity and success" dependent variable we performed pairwise Spearman correlations (Table 1). This revealed, surprisingly, that Yelp and Foursquare ratings were almost completely uncorrelated. While this is worrying, we chose to continue with only the Foursquare rating information, as it correlated to a sensible degree (around 0.4) with other measures of restaurant popularity such as likes or number of photos, whereas the Yelp rating did not. Moreover, Foursquare photos, likes and tips were highly correlated (close to 1), indicating that only one of these measures is sufficient to use for further analysis.

Table 1. Spearman correlation matrix between different variables putatively quantifying restaurant success.

	Rating	Rating FS	Price FS	Likes FS	Tips FS	Photos FS
Rating	1	-0.0444916	0.0118588	-0.157255	-0.132114	-0.174267
Rating FS	-0.0444916	1	0.237514	0.465174	0.356473	0.419765
Price FS	0.0118588	0.237514	1	0.406132	0.422209	0.418841
Likes FS	-0.157255	0.465174	0.406132	1	0.947115	0.961489
Tips FS	-0.132114	0.356473	0.422209	0.947115	1	0.917688
Photos FS	-0.174267	0.419765	0.418841	0.961489	0.917688	1

To obtain a single measure of restaurant success I combined restaurant ratings (from Foursquare) and number of likes in the following way: First, Likes FS was log-transformed to obtain a more “normal” distribution. Next, both Likes FS and Rating FS were normalized to range between 0 and 1. Finally, both measures were added and the result divided by 2 to yield the new measure “Success”, ranging between 0 and 1 (see Figure 4).

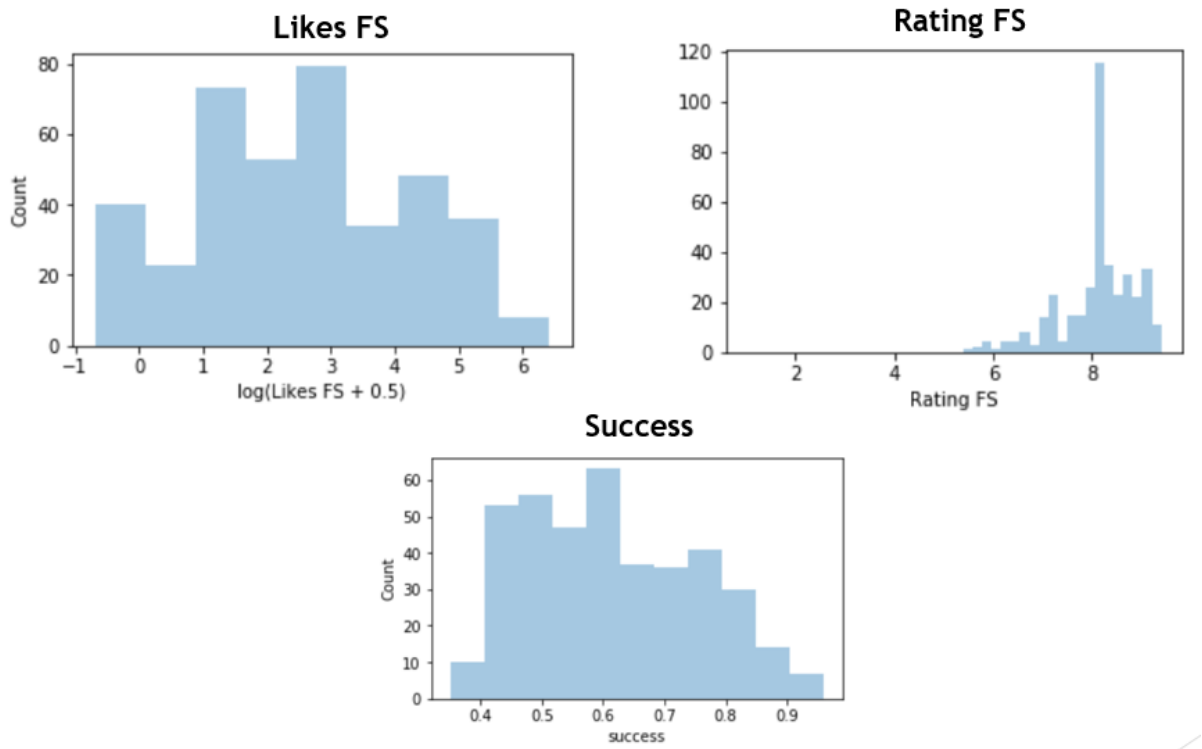

Figure 4. Histograms of variables useful for quantifying restaurant successfulness. "Likes FS" and "Rating FS" are combined to yield the measure "Success" (see Text for details).

Next, I searched for surrounding venues within walking distance (200m) from a given vegan restaurant on Foursquare focusing on the venue groupings "Food", "Shop & Service", "Bus Stop", "Metro Station", "Nightlife Spot" and "Arts & Entertainment". Using this new Information I created independent variables (or features) that quantified the ratio of how frequent a given venue category (e.g. Italian restaurant) was around a particular vegan venue relative to all other venues surrounding it. In total I obtained just under 400 Features (the data is illustrated in Figure 5).

Vegan Venue	ATM	Academic Building	Accessories	Acupuncturist	Administrative Building	Adult Boutique	African	American	Amphitheater	Antiques	Apparel	Arcade	...
1802 Roasters	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	
7 Leaves Cafe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.012658	0.0	0.0	0.025316	0.012658	
85°C Bakery Cafe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.007752	
A Divine H2O	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.034091	0.000000	
ACASA Food Truck	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	

Figure 5. Illustration of some of the features (columns ATM, Academic Building etc.) used for predicting restaurant success.

Having established a measure of " success" (dependent variable) and features of a restaurant's neighbourhood (independent variables) I decided to apply a K-nearest neighbour regression. To this end I divided the data into training (66%) and test (33%) sets, so I could assess model performance on the test set (to avoid overfitting). To assess how many neighbours K should be used I plotted the root mean squared error (RMSE) for a range of Ks and chose a reasonable one based on visual inspection (Figure 6).

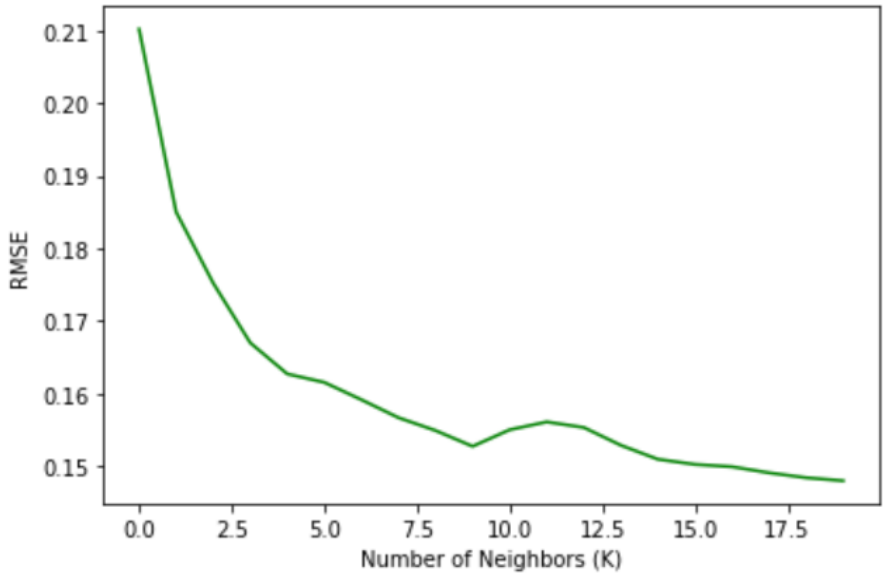

Figure 6. Change in root mean squared error (RMSE) with increasing number of neighbors (K). K=9 is a reasonable choice, as increasing K further hardly improves performance.

Results

Finally, we can use the trained model to predict the "success" of new venues based on a given location's surroundings. We can, for example, do this for every LA neighbourhood center (Figure 7). Although it should be noted that this is a very arbitrary choice, and only very loosely resembles a given neighborhood's suitableness for a new vegan venue (as it is based on a very narrowly defined location around a neighborhood's center).

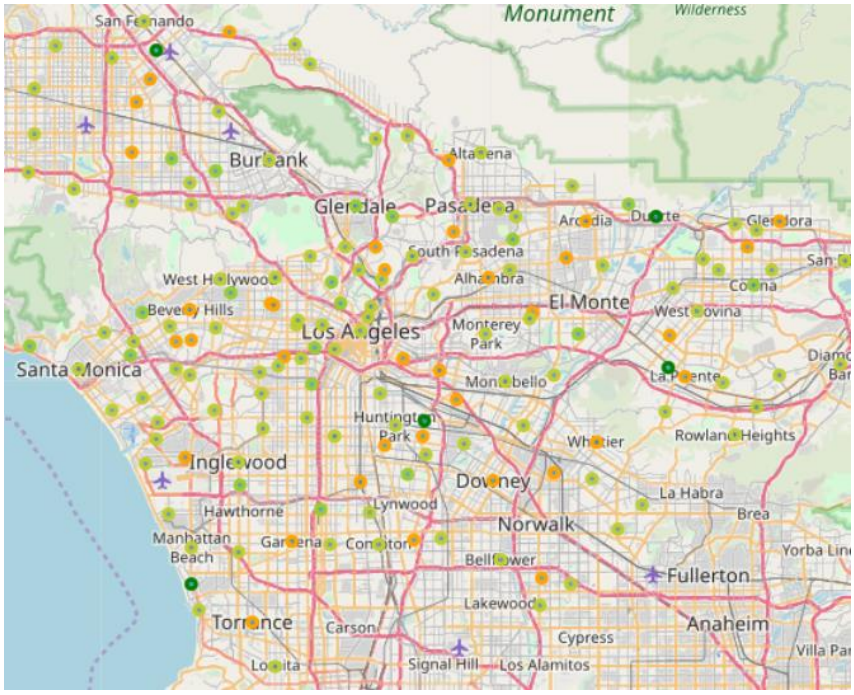

Figure 7. Predicted success of a new vegan restaurant for the coordinates associated with LA neighborhoods, color coded from green (successful) to red (not successful). The five restaurants most likely to succeed are shaded dark green.

A more sensible approach might be to search a given area in 300 m steps vertically and horizontally to narrow down more specific locations where a restaurant is predicted to be successful. An example of this is shown in Figure 8. The distribution of predicted success scores is shown in Figure 9.

Figure 8. Predicted success of a new vegan restaurant for the coordinates associated with LA neighborhoods, color coded from green (successful) to red (not successful). The five restaurants most likely to succeed are shaded dark green.

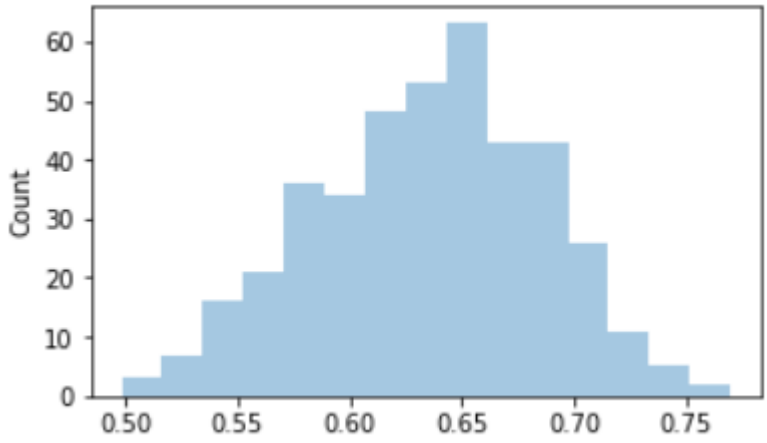

Figure 9. Distribution of predicted success scores for a grid of geospatial coordinates in LA.

Discussion

We leveraged user ratings and likes, as well as information about surrounding venue distributions from the Foursquare API to predict the success of vegan restaurants in Los Angeles using a K-nearest Neighbour regressor. While this approach is in principle sensible, the interpretation of the results is not straightforward.

First of all, having approximately 400 features is probably overkill. At the same time, simply grouping them into more broad categories (like the six parent categories we looked for on Foursquare – e.g. “food”), is probably too simplistic. Instead, we could try to create more derived features, such as “distance to public transport”.

Secondly, the venues included in the analysis are very diverse, and it might be fruitful to look at vegan venues over multiple cities to narrow down the best surroundings for a specific type of cuisine or restaurant (I doubt that a coffee shop benefits from the same surroundings as Lebanese restaurant).

If a stakeholder wants to invest in this project it will also be possible to explore restaurant menu information on a large scale (again over multiple cities, as many restaurants do not provide it), which would allow a quantification of how many vegan dishes are actually offered in a given restaurant.

Altogether, this analysis can only be seen as a first glimpse for what the Foursquare, Yelp and possibly other online data sources about urban venues can provide a data scientist in their quest to establish the best location for opening a new vegan restaurant.

Conclusion

Vegan restaurants are on the rise and more and more people are adopting a vegan lifestyle. Yet, vegan cuisine is still relatively sparse, which offers great opportunities for aspiring new restaurant owners. Using data from Yelp and Foursquare can provide crucial insights into where such a new venue should be build.