Prospective Coffee Shop Locations in Los Angeles

*R. Velunta*

**Introduction**

The coffee business has seen a recent influx of participants as consumers begin to expect more out of their caffeine experience than older formats such as Starbucks, Coffee Bean and Tea Leaf, or your standard cafe can offer. This “third-wave coffee” trend is quickly becoming a crowded space to occupy, especially in denser cities. In addition to the constant search for novelty in this market is the race to bring this new coffee-style to every uninitiated neighborhood across the globe. To aid in this effort, this analysis will attempt to predict new locations where these businesses can thrive.

**Data**

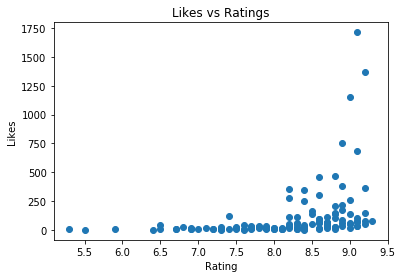
The metrics to be used for this analysis will primarily be sourced from Foursquare, a location-data and technology platform that serves as the basis for popular apps such as Uber and Apple Maps. In particular, this model will utilize Foursquare’s social aspects, such as ratings and likes, as a measure of a business’ success. The surrounding venues will reveal several key considerations such as: diversity of other activities, competition, and complimentary businesses. In order to gather a comprehensive list of coffee shops within the city we focus on, a set of postal codes are used as anchors for our search calls. The implementation of the model will also rely on coordinates of the most frequent venue neighbors as centers for new location considerations.

**Methodology**

Feature Selection

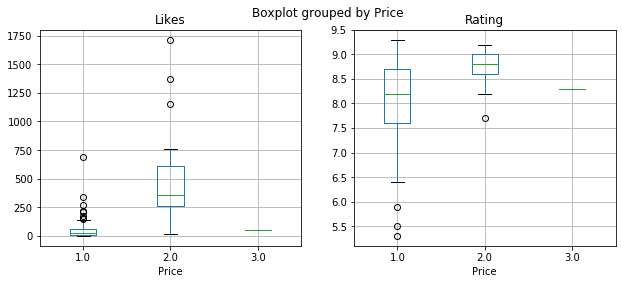
The structure of this analysis requires a restriction to the feature set that is based on the number of API calls to be made, with some coffee shops yielding the maximum (50) nearby venus in a walkable (200m) radius. A tertiary premium call would allow the aggregation of ratings, likes, and check-ins for these stores. Instead, we focus on the categories of nearby places and transform those into diversity, competition, and total; these are the number of unique categories, the number of coffee shops/cafes, and the total number of places respectively. In addition, a comprehensive one-hot matrix of all observed categories in our data set is used to track the five nearest venues of each coffee shop in order to provide a sense of location context.

The analysis scope is limited to the city of Los Angeles for this iteration to both demonstrate the elementary methodology and also in respect of cultural/geographical considerations that could affect the training of the model.

Labels

A few options exist for determining whether a business is successful: likes, rating, and price tier. Broadly, these metrics should positively correlate with one another, which is certainly what is seen when plotting them against each other.

A notable fact is the distribution of likes below 100 across the entire rating spectrum. This is interpreted as a count against ratings as a determination of a successful business since it does not always track with the other metrics. In contrast, a high amount of like will surely imply a high rating as well as an indication of pricing capability as shown in the box plot below (annotation)..



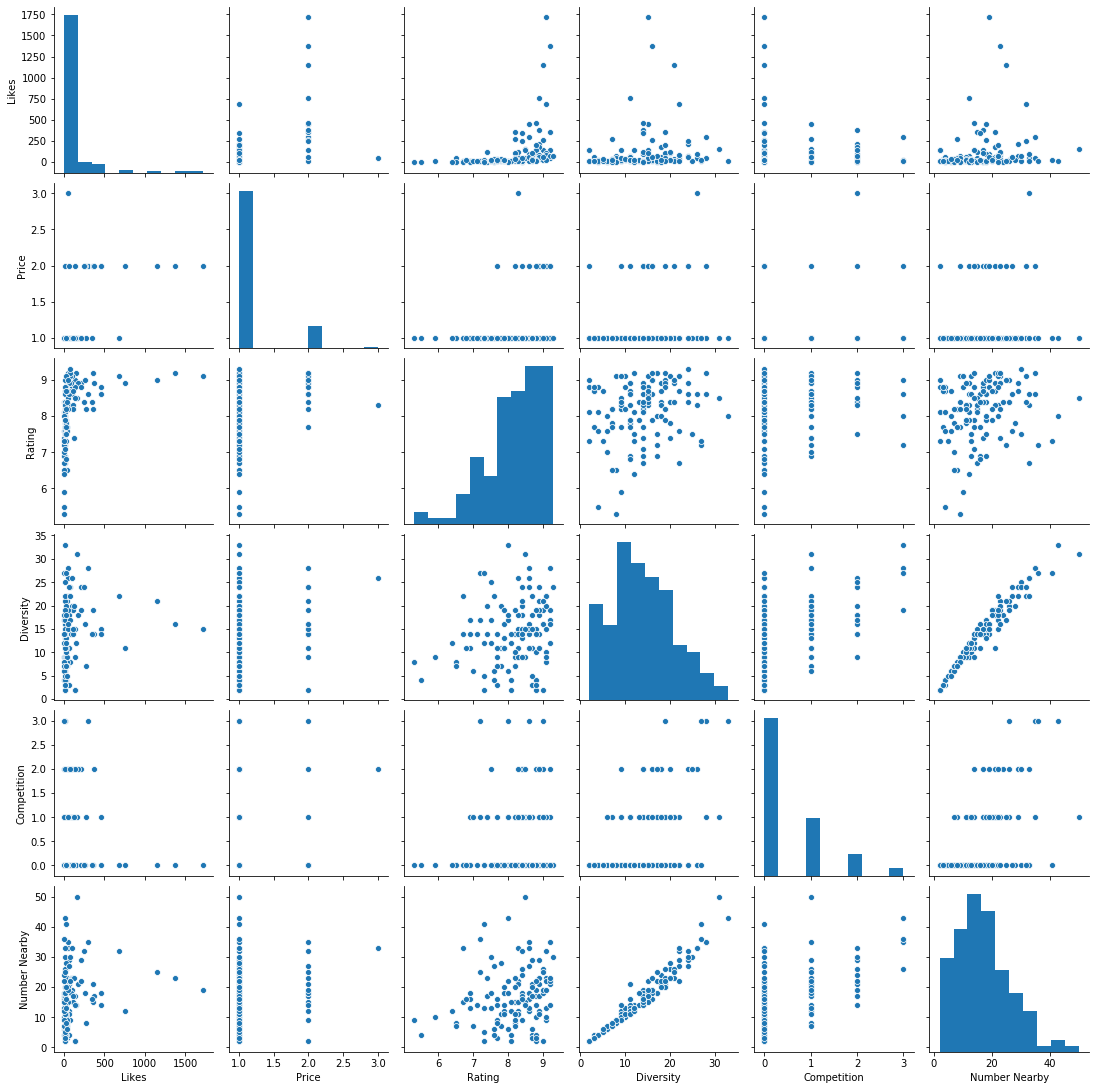
The number of likes also encapsulates weight by popularity despite our exclusion of check-in data which makes it a more robust indicator overall.

Cleaning and Retrieval

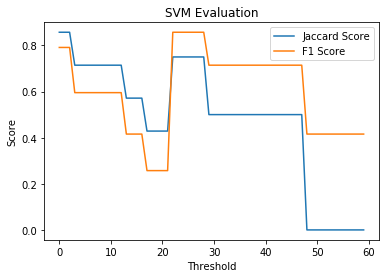
As described previously, gathering a list of coffee shops in Los Angeles was achieved through Foursquare API calls to each Los Angeles postal code coordinate with a generous 10 mile radius. Repetitions were removed by taking unique venue IDs. Since this was originally designed with specialty coffee shops in mind, large chains like Starbucks and The Coffee Bean and Tea Leaf were removed from the data set. Unfortunately, this reduced our sample size dramatically from 686 to 123. The last bit of cleansing required a removal of venues that did not have ratings or likes from Foursquare users.

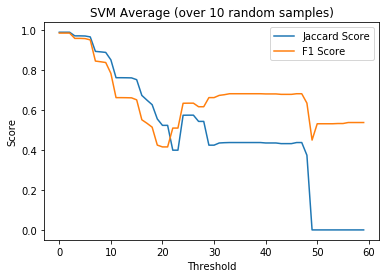
Exploration

Before building a model, the three context metrics described above (diversity, competition, and number of neighbors) were plotted against the success metrics in consideration (likes,ratings,price tier) to evaluate any standout trends among the data.

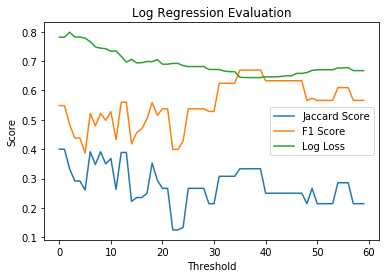


Slight normal distributions can be inferred among the ‘Number Nearby’ and ‘Diversity’ scatters when plotted against ‘Likes’. Additionally, we see a distinct feature of venues possessing 500 likes or more having a ‘Competition’ score of 0. These patterns do appear weakly in our data set, however, and may highlight the requirement of more data. Despite this, these relationships strengthen our choice of likes as a delineator of success.

Model Evaluation

Two machine learning algorithms were applied over a range of thresholds on the ‘Like’ column to optimize over our definition of success. This first model chosen was an SVM in consideration of the large categorical component of the feature vector. So this decision is primarily motivated by an attempt to locate a grouping in neighboring venue category space that was indicative of success. 

Initial testing of this method shows optimum predictive power in the 25 - 45 ‘Likes’ range for success threshold. However, this result is highly sensitive to changes in the training and test subsets chosen. An average over 10 randomized sets shows a more tempered result along the same range.

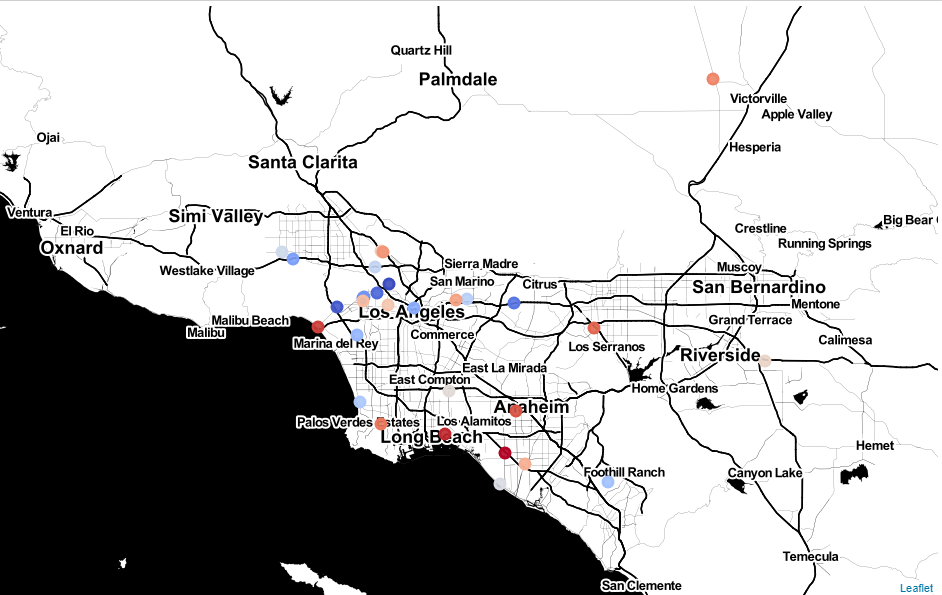
The second model utilized logistic regression in an attempt to soften our expectations for data separation and opts for a probability for success. The binary results do seem to suffer here with Jaccard and F1 scores only seeing a modest increase in the range we observed for the SVM. The log-loss also shows stability across the threshold range as well with a regularization factor of 0.01 chosen. Increasing this hyperparameter only affects the lower end of the threshold scale as underfitting becomes more impactful when more data points are included here. 

Implementation

The linear regression model provides a more useful output to the potential consumers of this data. The binary classification of the SVM is more difficult to act upon with the scores it displayed on the test data. A probability remains a useful consideration despite imperfect predictive power.

To implement the model, we consider a small set of venues consisting of categories that were both frequent and infrequently found near our data set locations to provide a mix of environments to predict. This set included: Sandwich Places, Pizza Places, and Clothing Shops as frequent nearby venues and Drug Stores and Dive Bars as infrequent nearby venues. Thirty venues were obtained and processed in an identical manner as the original data set to provide feature vectors compatible with the model.

**Results**





The resulting coordinates are plotted on a Los Angeles map with the color scale of each point corresponding to the probability of success. These locations should be interpreted as an analysis of the 200 meter area around the coordinate plotted.

**Discussion**

First Impressions

The values of our implementation are promising considering general intuition about the areas the points are in. Outlier points such as the one in the upper right near Victorville are on the outskirts of high density zones and probably would not garner the amount of likes we have used to label success. Additionally, locations along the Hollywood Hills and West LA are surrounded by other popular venues and have higher relative probabilities.

Data Size

A notable concern at early stages of this analysis was the incredibly low number of usable data points after cleaning, which was 123 coffee shops/cafes. This manifested most noticeably during our training step where changes to our training/test set composition and size had large effects on the predictive scores calculated for each method. Expanding the scope of consideration would help this, but would need to consider how much trends differ by region.

Information Limitations

Another limitation that should be addressed is the reliance of user participation in Foursquare ratings. Undoubtedly the distribution of active users is not uniform across any geographical location. Perhaps an outside metric of success should be considered in this case.

Furthermore, a unique characteristic of this problem is identifying characteristics that represent a small percentage of the total examples. The best cases will always represent the outlying data points and will thus be underrepresented in any training set.

Additional Considerations

Finally, one could also aggregate further characteristics from the Foursquare API such as the number of visits, the type of users who visit, and how popular the surrounding venues are. A more comprehensive feature vector like this would be possible with further API resources.

A more complete implementation would also consider enough venues to map probabilities across the entire geographical space, providing a more continuous function of success probability over the city.

**Conclusion**

This analysis shows that a predictive model is capable of being constructed using contextual details of existing businesses to map success probability across a geography. It also highlights some limitations of data size and quality that are unique to the business category in question. This was a major inhibitor for finding locations for Coffee Shops in Los Angeles, but did not prevent the development of models at least as good as a native’s intuition. The demonstration of this restricted implementation is promising for further augmentation.