Location evaluation for Gastronomy Businesses based on Geospatial Data

by Christopher Husmann

Introduction

This capstone projected aims to identify promising location for business like restaurants, bars and snack shops. For the success of such a business the quality of food and servers is essential success. However, the location of such business is also significant for their long-term success. For example, it is unlikely that a restaurant flourishes in an industrial industrial area even if it serves exceptional food. In contrast, a mediocre restaurant can be well visited if it is located next to famous tourist attraction. Hence a tool that could evaluate a location on its suitability for a certain establishment could be very valuable for people who contemplate the idea to become self-employed as well as for big changes who want to extent by opening new brunches.

As discussed above the envisioned tool is in general useful, however due to the current COVID-19 pandemic its benefit is quadrupled. The drastic emergency acts, which were ratified to enforce social distancing, hitting gastronomy especially hard. Hence it is expected that many independent establishments have to close permanently. For example, according to the New York Times as many as 75% of the independent restaurants of New York may not survive the COVID-19 pandemic. While such a cataclysmic event is extremely tragic for the millions of entrepreneurs who may lose the fruits of decades of work, it also sets the stage for new success stories. In other words, at the end of the COVID-19 crisis the shortage of gastronomy businesses of all kinds will present tremendous business opportunities. Furthermore, large numbers of prime location will be available. This unprecedented situation makes tools capable of reliably evaluate the suitability of a location for a particular type of gastronomy business more valuable than ever.

Dataset

The target of this project is to develop a tool which calculates a score for any location in the US, specified by longitude and latitude, and a particular type of business, where the score represents who suitable the location is for opening a new business of the given type. For this purpose, we will leverage geospatial data gather via the Foursquare API. In particular we query the recorded venues in Foursquare for each postcode of the US. To enrich our database, we use census data provided by the US government. In particular, we utilize the number of employees in a postal code, the total income of these employees, and the number of registered business to further characterize economic strength of the particular post code area.

In addition, we query the details of all business of interest. For example, if a client is interested in opening an ice cream shop, we query the details off all US-based ice cream shops in the Foursquare API. To build the training and test sets we rank the business based on the number of likes they received on Foursquare. Then, we label the post-codes of the top-30% business as promising locations. Postal codes without a business of interest (e.g., ice-cream store) are labelled as not promising locations.

This rich database, will then be used to train a logistic regression model to categorize locations into promising and non-promising location. Logistic regression is employed since it provides not only a label ('promising' or 'not promising') but also an estimate how reliable the classification was. Hence, the tool cannot only be used to determine if a given location is promising, but we can also compare different location against each other.

Methodology

First, we need to get a list of all post codes of the US and the corresponding coordinates, the data is available online as csv file. (https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/). As a next step, we collected the county business patterns of 2017. This official report of the US government allows to analyse economic activities of small areas and includes the number of establishments, employment during the week of March 12, first quarter payroll, and annual payroll. The dataset is again available online as csv file (https://www.census.gov/data/datasets/2017/econ/cbp/2017-cbp.html).

Next, we use the Foursquare API to get the top 100 venues that are within a radius of 500 meters around the centre of a given postal code area. After gathering the data, a panda dataframe is populated representing the geospatial data, the census data, and the Foursquare data for all postal codes of the US.

Depending on the client interest we query the Foursquare API for details for all US based business of interest. In this we report we use the example of an ice cream shop, but other venues such as restaurants or coffee shops would work as well. However, in this report we use ice cream shops as example.

In total we found 1510 ice cream store in Foursquare's database. Among others details, the number of likes for a given venue can be queried on Foursquare without requiring a business account. Therefore, we used this metric to indirectly measure the popularity and success of a given business. To enable supervised machine learning approaches we label the postal codes that correspond to one of the top-30% ice cream stores in the country as promising locations whereas all postal codes without an ice coffee shop registered at Foursquare are labelled as not promising. Given that we found a total of 1510 ice cream store, we labelled only 453 location as promising. To avoid that the test set is heavily skewed towards non-promising location we randomly selected 453 postal codes out of the set of as non-promising labelled locations. Hence, we use a set with a total of 906 labelled locations to train the classifier.

By exploring the vicinity of the on Foursquare registered ice cream stores, we found a total of 638 venue types. Theoretically, the all of the venue types could be used as features and therefore as input to the ML algorithm. However, since the number of features would then be of the same order then the number of observations in the labelled dataset, overfitting might be a problem when generalizing the model to new locations. Therefore, we further process the available dataset to extract the venues that are characterizing promising locations. To identify such venue type we use two indicators. As first indicator we utilize the who frequent a particular venue is found near a promising location. As second indicator we calculate the ratio of the frequency near promising location and the frequency near non-promising locations. Since in average promising locations have in total more venues in their vicinity, we normalized the frequencies by the average frequency over all venues before calculating the ratio. To train the ML algorithm we only utilize venue types, whose two indicators are above certain imperially found thresholds. The calculation of these thresholds is described in the result section.

Results

Feature selection

As a first step we divided the labelled dataset into training set (70% of the observations) and a test set (30% of the observations).

As discussed, we only utilized venues types as features which indicators, frequency of occurrence and frequency ratios, are above given thresholds. To find these thresholds we conducted a grid search over the two parameters. In particular, we calculated for each indicator pair the average F1 score using 3-folded cross-validation on the training set. To account for venues that are positive correlated with ice cream stores (frequency ratio is >>1) and venues and venues that are negatively correlated with ice cream stores (frequency ratio is << 1), we use the venue types as features with fulfil the following conditions:

$$\alpha < f_{+}(x),$$

$$\beta < \left|\frac{\overline{f_{+}}\,(x)}{\overline{f_{-}}\,(x)} - 1\right|.$$

Where α and β are the two threshold values, $f_{+}(x)$ is the frequency of occurrence for venue x near promising locations ($f_{+}(x = coffe\ shop) = 1.2$ means in average 1.2 coffee shops are found near promising locations) and $\overline{f_{+}}\,(x)$ and $\overline{f_{-}}\,(x)$ nominate the normalized (by average) frequency of occurrence for promising and non-promising locations, respectively.

Figures xx and xx shows the results of the grid search. In particular each square represents one α, β pair and the depicted values represent the corresponding F1 score and the number of venues selected as feature for the Figures xx and Figures xx, respectively.

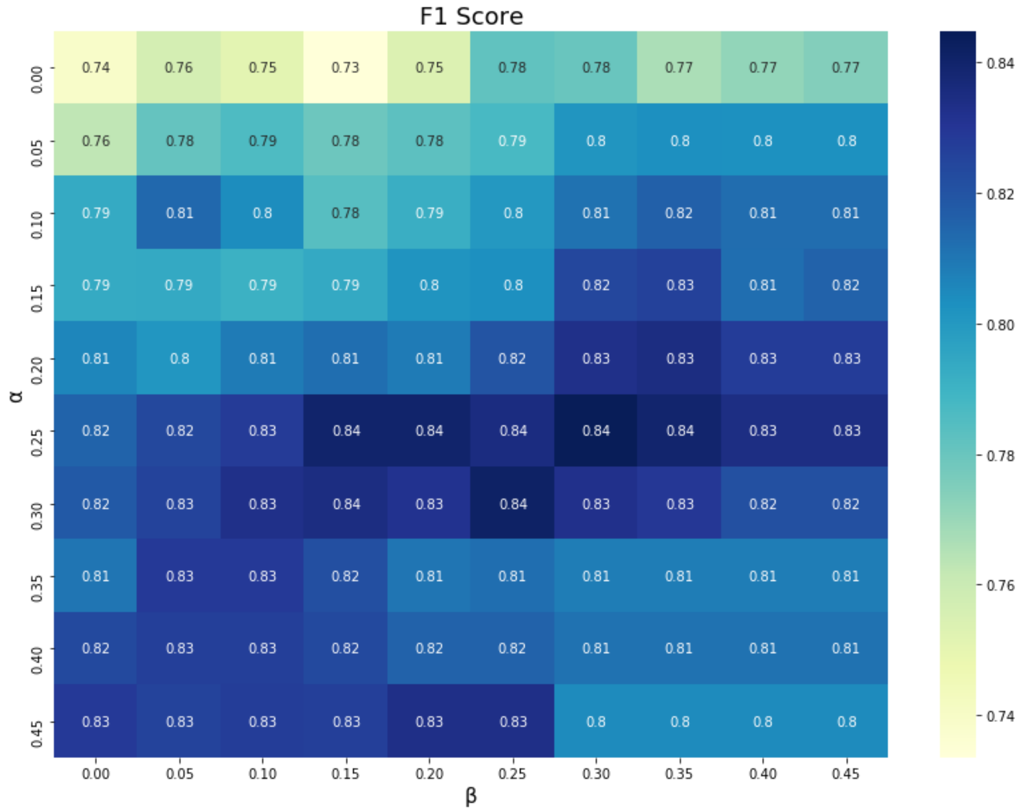


Figure 1: F1 score for given α, β pair

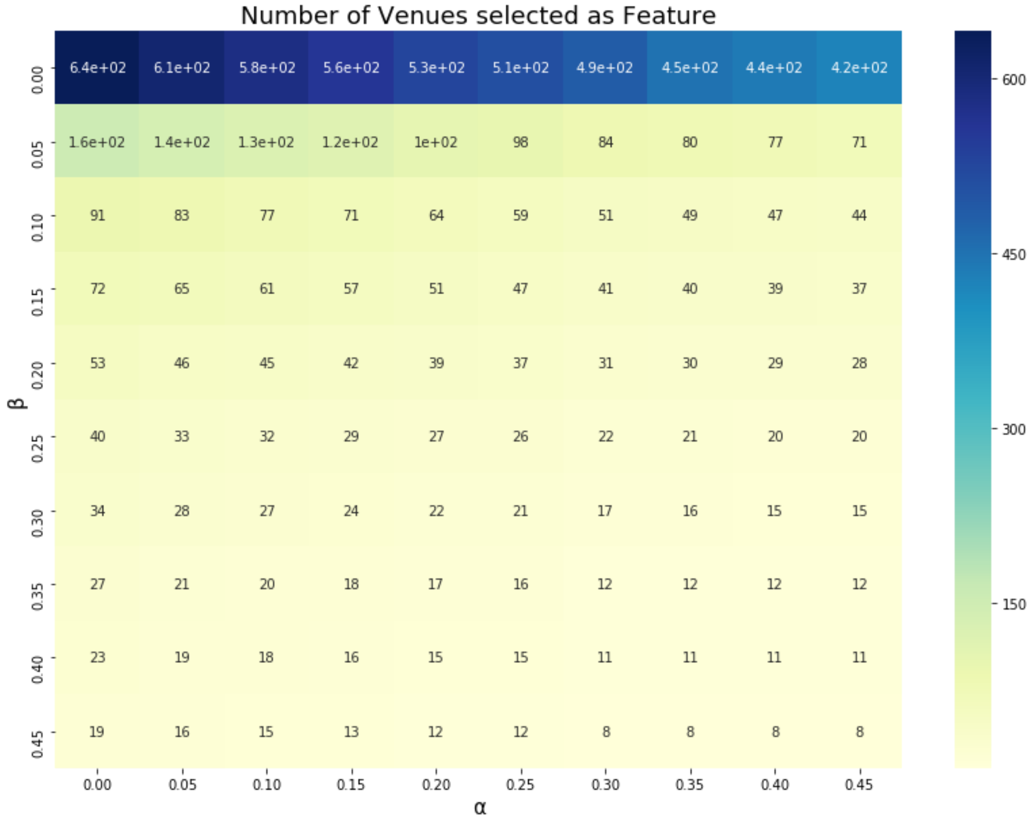


Figure 2: Number of selected features for given α, β pairs

In figure xx it can be seen that the F1 score is maximized for $\alpha = 0.25$ and $\beta = 0.30$. Here a total of only 20 venues are selected as features. Interestingly, the F1 score is notably better compared to the case when all venues are utilized as feature.

Figure ii depicts the frequency of occurrence and the frequency ratio for all venues which have been selected as features. We note the final set of features has venues which are positively correlated with ice cream stores (ratio larger than one) and negatively correlated ice cream stores (ratio smaller than one).

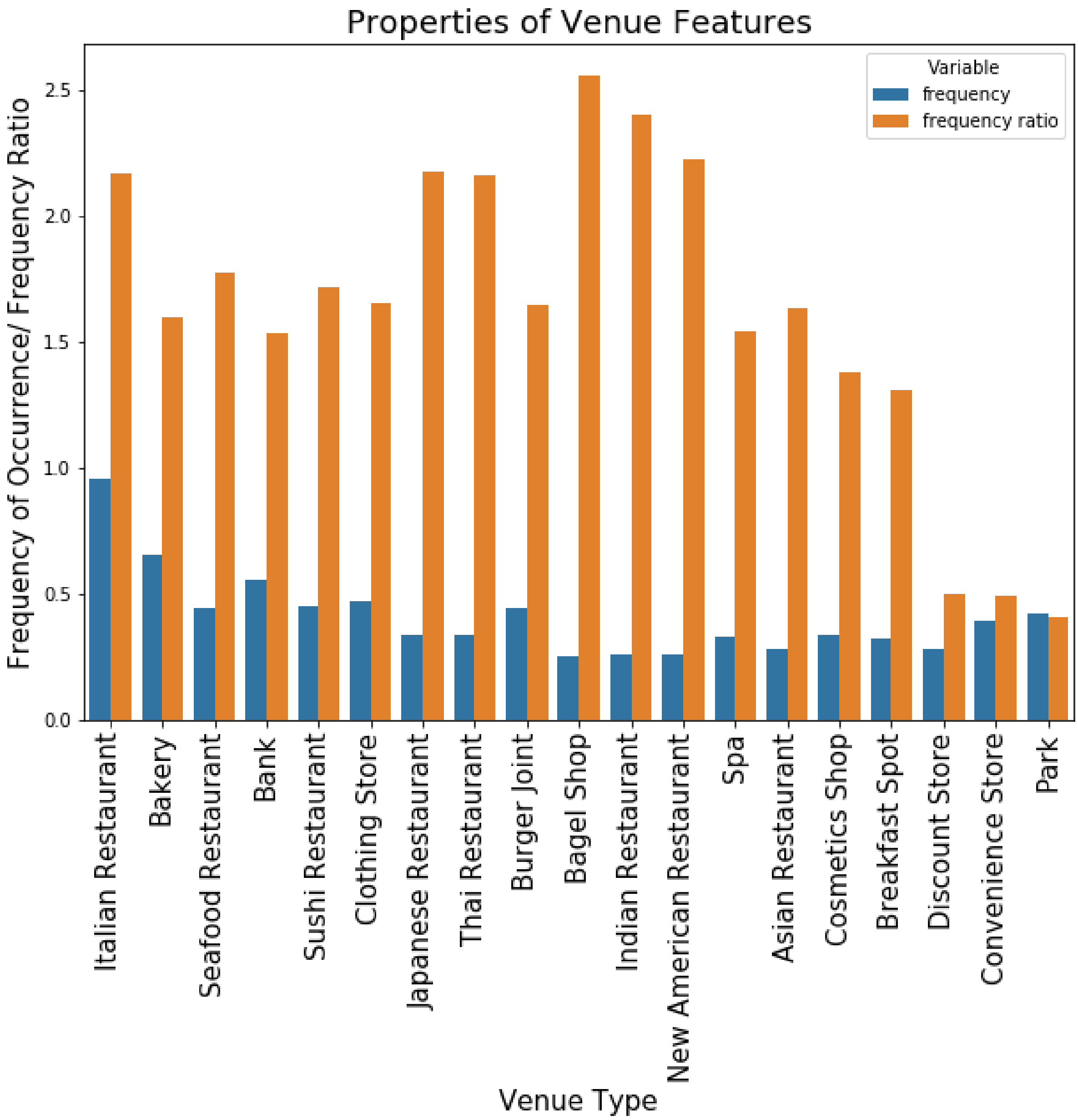


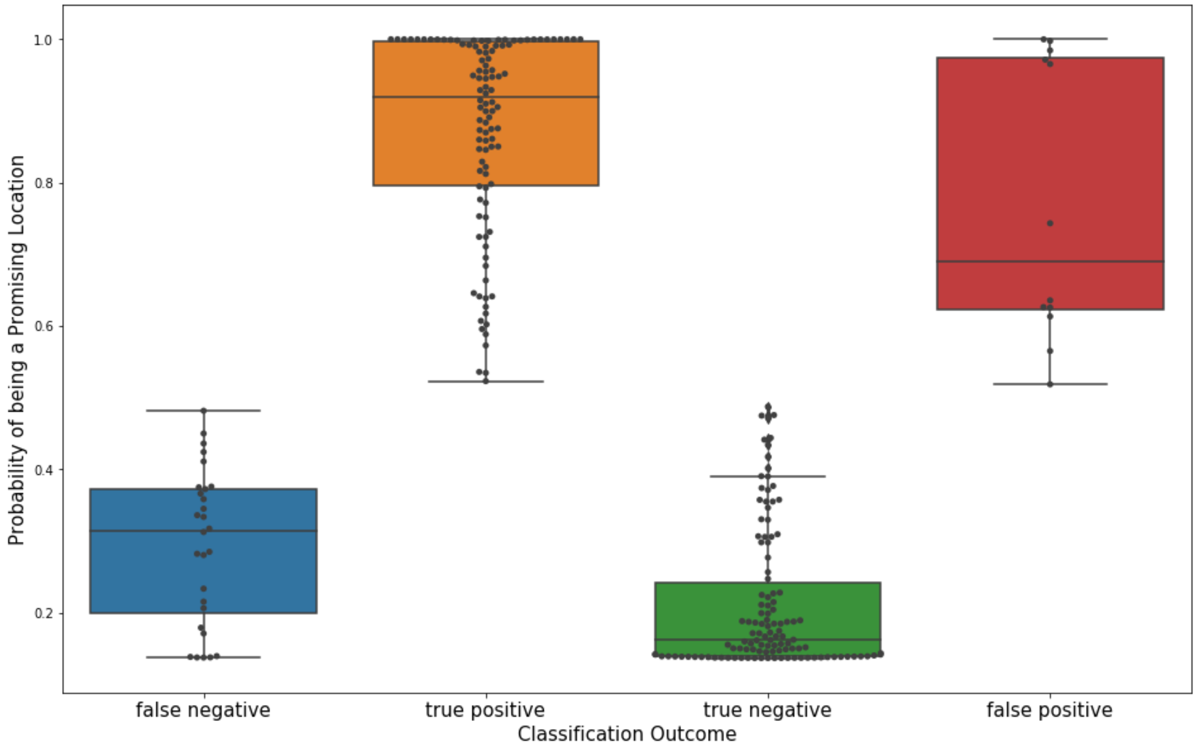
Figure 3: Frequency of occurrence and the frequency ratio for all venues which have been selected as features

Classification results

To fit a logistic regression model, we used the sklear library, where we used the liblinear solver and the Euclidean norm for penalization. The classifier reached an F1 score of 0.82 on the training set and 0.84 on the test set. The similarity between the F1 scores of training and test set indicate that the model generalizes well to new observations. Table xx depicts the confusion matrix:

	Labelled as non-promising	Labelled as promising
Non-promising	45 % (true negative)	4.5 % (false positive)
Promising	10.5 % (false negative)	40 % (true positive)

Figure SS shows the depicts the calculated probabilities of the logistic regression modelled for the four possible classification outcomes in the form of a boxplot.



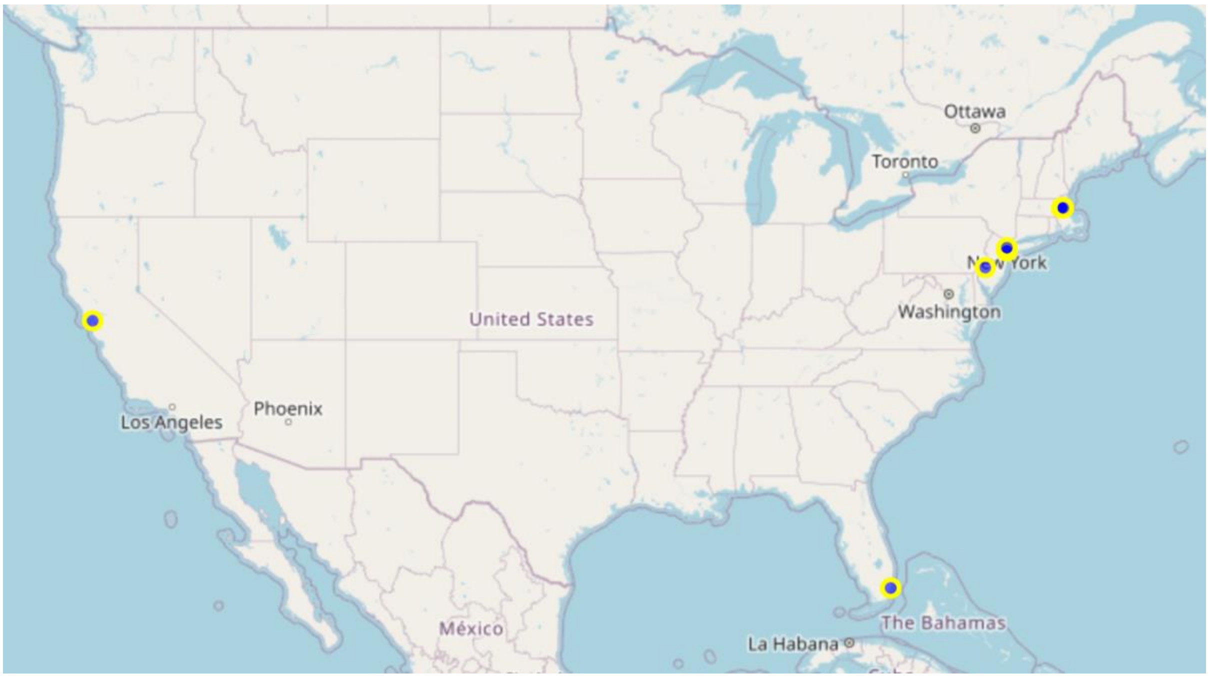
It can be seen that the majority of true positives have a score close to one. Similar, the majority of true negatives have a score close to zero.

Figure 4 US map with the ten most promising spots to open an ice cream store

Discussion

In the results sections we have shown that that we were able to reliability classify locations as promising or not promising for a certain type of venue. Interestingly, we were able to show that only a few key venue types have to be tracked to make accurate predictions. These key venues have a significantly higher probability of occurrence in the vicinity of successful ice cream stores than their unconditional probability of occurrence. While some of these venue types are related business types such as bakeries and bagel shops, we also unveiled less obvious connections. For example, the density of cosmetic shops is about 50% higher than average in the vicinity of ice cream stores. Another, surprising discovery is that density of parks is actually 60% smaller in the vicinity of ice cream stores.

In the result sections we showed that the majority of true positivises and true negative have scores close to one and zero, respectively. This indicates that score calculated by the logistic regression model is a good indicator the suitability of a location to open an ice cream shop. The indicative power of the score is further supported when plotting the postal codes most suitable to open new store. As shown in figure xx, the most promising are, as can be expected, in the heart of the biggest metropoles of the US.



Conclusion

In this report we explored the possibility to evaluated the suitability of location for a specific given gastronomy business based on geospatial data. For this we utilized localized economic data from the us government and location data form the Foursquare database. We presented to method to reduce the location data of Foursquare to few key features (in the order of ten) at the example of ice cream store. Still the presented method can be uses for any arbitrary gastronomy business. To classify the locations into the categories promising and not promising we used logistic regressions. For the ice cream shop example, we reached an F1 score of 0.84. Hence the classification results and classification score of the proposed method can be a valuable input for entrepreneurs who are looking to open a new gastronomy business.