Data-based recommendation on location for Houston restaurant business owners

Linhai Percy Zhao

Table of Contents

Introduction: Business Problem

When it comes to brick-and-mortal businesses, what always come on top of business owners' mind is location. In this project, we will be specifically focusing on restaurant businesses in the city of Houston, where a wide variety of foods and cuisines can thrive. An ideal location for a restaurant should be a great combination of reasonable cost (land/rent/utilities/personnel etc.), abundant customer traffic, nice neighborhood and relatively low competition.

However, in reality, the decisions on location are usually not done well-informed, with the limit of time and data. Besides, several important factors may not appear as apparent as others and can be easily neglected during the real selection process. Therefore, it is valuable to provide a data-driven recommendation for location pick to help business owners make an evidence-based decision.

Here, I will leverage the Foursquare API and other sources of data to depict a detailed profile of each neighborhood in Houston. Based on the profile data from other similar metropolitan cities, I will model this problem into a supervised learning model using most widely used classification models. The best prediction model will then be used to generate predictions on the neighborhoods in Houston about whether they are good choices for restaurant businesses. The whole process is driven by objective data and the final recommendation will therefore be completely data-driven. In more detail, I will also interpret the predicting model to generate the most important features for selecting the right neighborhood for the restaurant business, which can be used for future considerations.

Data Description

We will use data from four cities: Houston, Los Angeles, Dallas, and New York City (NYC). All of them are big cities with variety, one in the north, the other three in the south. With these similarities, we hope to capture some core features that are critical to restaurant businesses, from which we can build a reliable prediction model. We will be primarily using venue category data from Foursquare API. The venue category data will be used as important features describing the neighborhoods. Besides, we will also include other features such as housing rate, foot traffic, demographics and so on.

More specifically, for a total of four cities (Houston, Los Angeles, Dallas, and NYC), the data we will use include:

1. Full neighborhood list and corresponding geographic coordinates. (e.g. (40.586314, -74.190737))
2. Venue category data from Foursquare API. (e.g. "Spa", "Supermarket" etc.)
3. I was also able to find some social-economic data for each neighborhood in geocod.io, include: household type, housing prices, family composition, demographics and income. (e.g. Household type: Family, Non-family; Family composition: Married couple, Singles etc. For different types, there will be absolute counts and percentage data for a neighborhood.)

Together, we would be able to depict a rather detailed picture about the fitness to open a restaurant in a neighborhood.

Methodology

Because the decision of selecting a location can be modeled as a **classification problem**, we need to assign labels to neighborhoods of Los Angeles, Dallas, and NYC indicating whether the area is a good place to have a restaurant business. For simplicity, we used the percentage of total venues that are restaurant as the indicator. Here, we set the cut-off value as 30%, meaning that if a neighborhood has more than 30% of venues being restaurant, we will say it is a good place to have a restaurant business and the label "1" will be assigned. Otherwise, label "0" will be assigned.

Columns and rows with high missing rate (>50%) are removed. Any apparent outliers are removed. Categorical data is one-hot encoded. Besides, for social-economic data, we used percentage data instead of count data as a way of standardization. The data from all four cities is then aligned to make sure the features are consistent across different cities. And the data from the three cities is then split into training (80%) and test (20%).

We adopted some most well-known models: **Random Forest**, **Logistic Regression** and **Support Vector Machines** to model this problem. Hyperparameter tuning is also performed to reach better accuracy for each model.

We picked out the model with the highest accuracy on test data and apply it to the Houston dataset. Moreover, we dug deeper to interpret the prediction results to extract key insights in picking locations for restaurants.

Analysis and Results

1. Get neighborhood list for all cities

For all four cities, we used Wikipedia page for the neighborhood list, and used *BeautifulSoup* and *requests* to parse for neighborhood names. (Code details in notebook) Together we found 195 neighborhoods in Los Angeles, 216 neighborhoods in Dallas, 306 neighborhoods in New York City and 88 neighborhoods in Houston.

2. Geocode neighborhoods

Next, instead of using the geocoder from Google, I used an online geocode tool (geocod.io) to get geographic coordinates for the neighborhoods. The online geocode tool is also able to annotate additional social-economic data including: demographics (population size, age, gender, ethnicity), economic income, family construction, housing and education etc. And that's what we got for the neighborhood list.

3. Use geocodes to get venue category information from Foursquare API

After getting the geocodes, we can use the information to get nearby venues from Foursquare API, and we used a radius of 500 meters. Remember that venue categories are categorical data and in order to be used in a machine learning model it's best to one-hot encoding the venue categories into a list of binary indicators. Moreover, we calculated the mean value for each venue category by aggregating each neighborhood. So before our annotation steps the data frames of neighborhoods look like:

	Neighborhood	City
0	Angelino Heights	Los Angeles

Now the data frames look much more enriched (part of dataframe):

	Neighborhood	ATM	Accessories Store	Airport	American Restaurant	Arcade
0	Angelino Heights	0.0	0.0	0.0	0.0	0.0

The next thing is to preprocess the data to make it suitable for feeding into a machine learning model. Here, we did several processing steps: 1) remove columns and rows with high proportion of missing data 2) remove redundant columns 3) use percentage data instead of absolute value as the standardized data.

Additionally, using the method described in the methodology section we assigned target values in the "Label" columns for data from Los Angeles, Dallas and New York. Lastly, because the data frames from different cities may contain different number of columns after the preprocessing steps, we aligned the columns of data frames from 4 cities including Houston, to make sure the model analyzes same features.

4. Apply machine learning modeling

Here, we used three common classification models: Random Forest, Logistic Regression and Support Vector Machines. To evaluate the models, we split the data from Los Angeles, Dallas and New York into training (80%) and test data (20%). Hyperparameter tuning is performed by random search/grid search to find the optimal selection of hyperparameters for each model. The best model is compared with the default settings to compare differences.

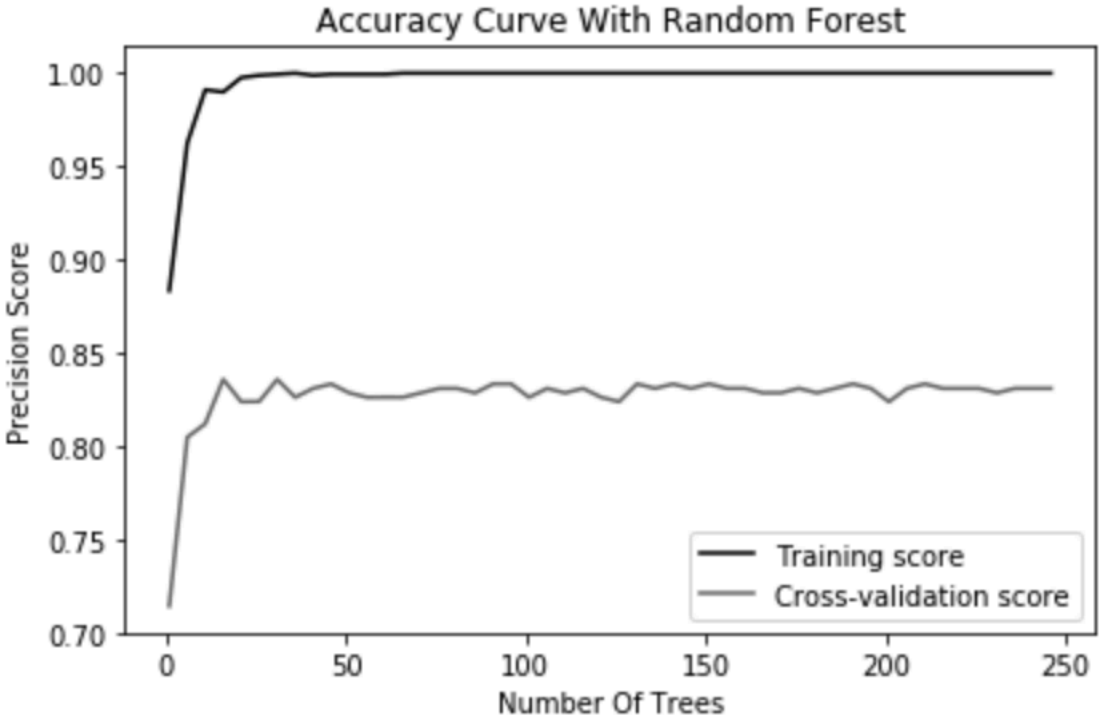
4.1 Random Forest model

We first applied Random Forest model to the training data. Considering the random nature of this model, cross-validation was used to get the averaged accuracy score as the measurement for performance.

Three hyperparameters were first tested individually: 1) class_weight 2) n_estimators 3) max_depth. For class_weight, we compared whether we should indicate balanced weighting to

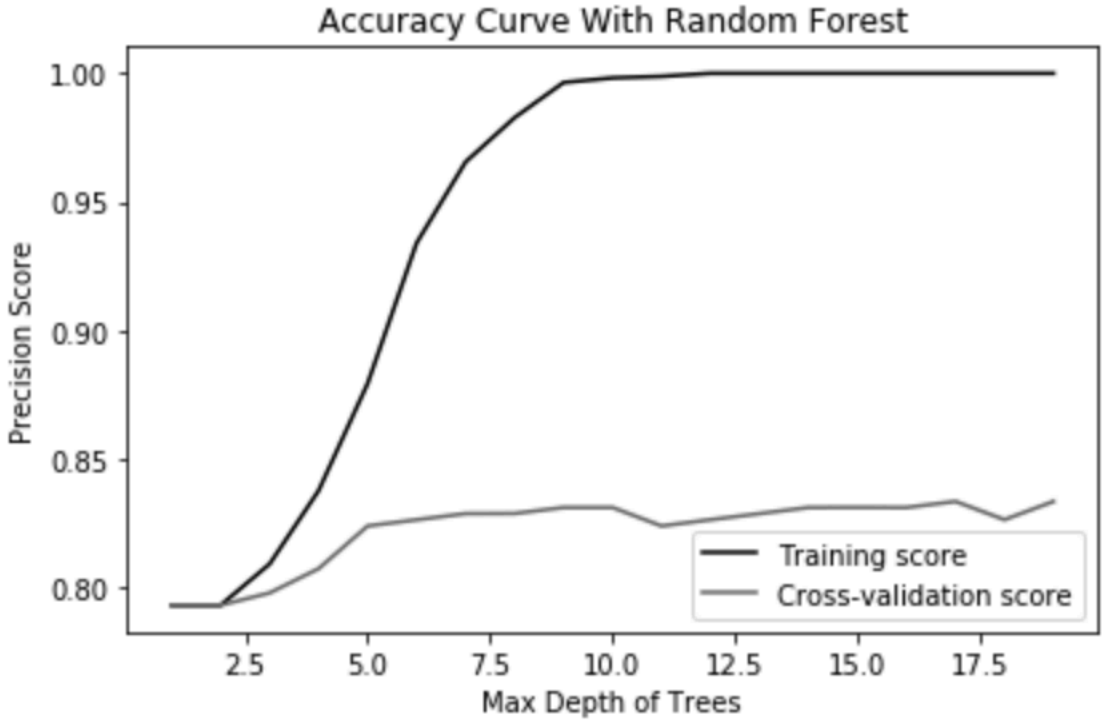
adjust for occurrences of different labels, but found out that the default setting provides better accuracy.

For n_esimators, we searched from 1 to 250 as the candidate values for number of trees, and the results are shown below:



It can be seen that while training score reaches close to 1.0 very quickly, the cross-validation score plateaued at around 30 trees. Therefore, we will use 30 as the optimal value for n_estimators.

For max_depth, we searched from 1 to 20 as the maximum depth for trees, and the results are shown below:



From the plot, it can be inferred that the cross-validation score goes up with maximum depth till around 10. Therefore, the selection for max_depth is 10 to reduce over-fitting tendencies.

Till now, we have built a calibrated model by looking at 3 hyperparameters individually. By comparing the performance of this model with the default Random Forest model on test data, we found that the calibrated model has an accuracy score of 0.821 compared to 0.811 in default model. The test data confirms the improvement.

Moreover, instead of looking at individual hyperparameters, we also searched for optimal combinations of hyperparameters. We used RandomSearchCV() and GridSearchCV() to search for hyperparameters: max_depth, max_features, min_samples_split, bootstrap and criterion. Through the searches, we identified the best Random Forest model: RandomForestClassifier(n_estimators=30,random_state=0,bootstrap=True,criterion='gini', max_features=100). Comparing with the default model, the best Random Forest model gives an accuracy score of 0.858 compared to 0.811 in default model, which shows a decent improvement.

4.2 Logistic Regression model

For logistic regression, we also performed grid search for combination of two hyperparameters: inverse of regularization strength C and penalty method ‘l1’ or ‘l2’. The best calibrated logistic regression model gives an accuracy score of 0.830.

4.3 Support Vector Machines model

For support vector machines, we performed grid search for combination of two hyperparameters: inverse of regularization strength C and kernel function ('linear','poly', 'rbf', or 'sigmoid'). And the calibrated model gives an accuracy score of 0.802.

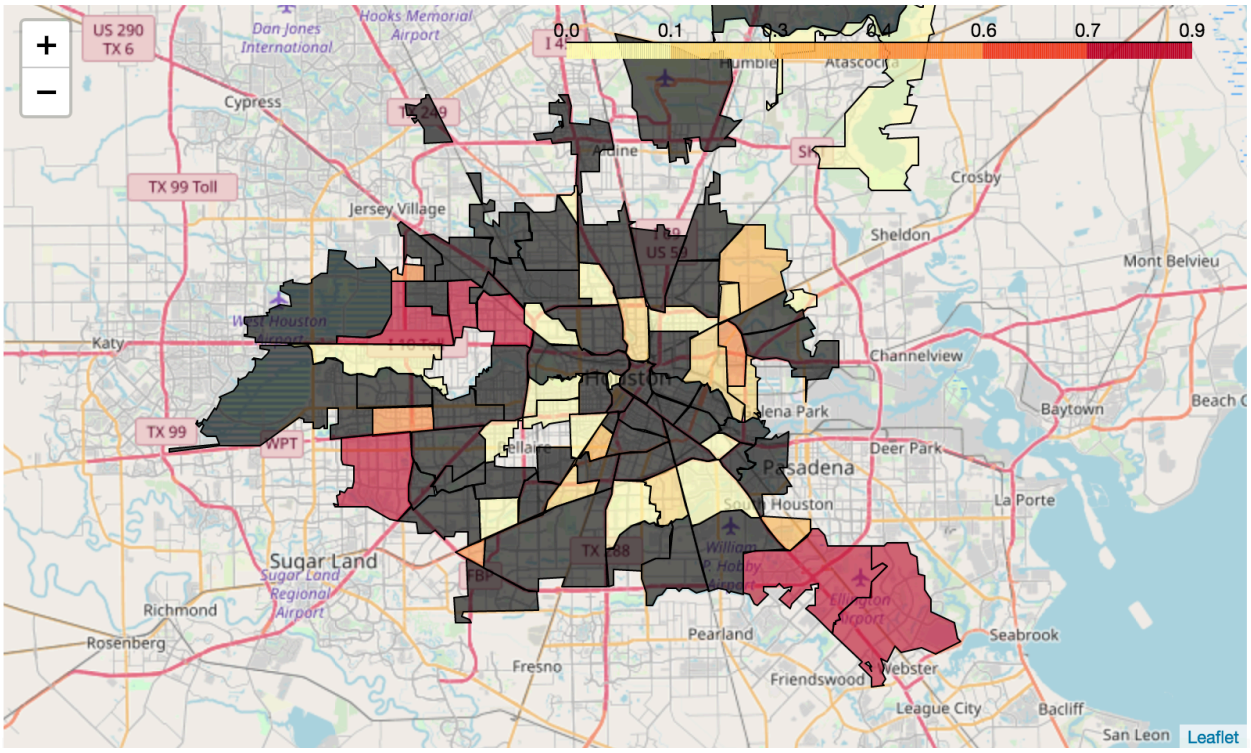
5. Use the best model to make predictions

Based on the accuracy scores from three different models on the test data, we identified Random Forest as the best performing model for this problem. Here are the top neighborhoods:
('South Belt / Ellington', 0.87), ('Alief', 0.87), ('Clear Lake', 0.83), ('Spring Branch East', 0.8), ('Spring Branch Central', 0.8), ('Spring Branch West', 0.8)

Additionally, we also used Logistic Regression model and the top neighborhoods:
('Spring Branch East', 0.99), ('Spring Branch Central', 0.99), ('Spring Branch West', 0.99), ('Clear Lake', 0.98), ('Alief', 0.74), ('South Belt / Ellington', 0.72)

We can see that the two models identified the same top neighborhoods, further suggesting that both models learned similar information from the data and the predictions are reliable.

A choropleth map is generated to visualize our findings:



The top neighborhoods are shown in red, and from the map, we can see three areas pop out: the northwest, southwest and southeast. Those are satellite cities around central Houston and are all considered good place to live. Besides, many areas do not have a prediction (showed in black), that's because those areas don't have social-economic data available therefore are not included in the prediction.

Moreover, we took the top neighborhoods and the bottom neighborhoods in terms of prediction probabilities, and we calculated the mean value for the top features in the Random Forest model to see if any differences can be found to potentially help understand the prediction.

For the top neighborhoods:

Mexican Restaurant	0.440476
Vietnamese Restaurant	0.007937
Chinese Restaurant	0.000000
Italian Restaurant	0.000000
Asian Restaurant	0.000000
Seafood Restaurant	0.000000
American Restaurant	0.000000
ACS Demographics/Race and ethnicity/Not Hispanic or Latino: Asian alone/Percentage	0.153000
ACS Housing/Value of owner-occupied housing units/$300,000 to $399,999/Percentage	0.051500
ACS Demographics/Population by age range/Female: 30 to 34 years/Percentage	0.076333

For the bottom neighborhoods:

```
Mexican Restaurant                                                                0.0000
Vietnamese Restaurant                                                             0.0000
Chinese Restaurant                                                                0.0000
Italian Restaurant                                                                0.0000
Asian Restaurant                                                                  0.0000
Seafood Restaurant                                                                0.0000
American Restaurant                                                               0.0000
ACS Demographics/Race and ethnicity/Not Hispanic or Latino: Asian alone/Percentage    0.1104
ACS Housing/Value of owner-occupied housing units/$300,000 to $399,999/Percentage     0.0466
ACS Demographics/Population by age range/Female: 30 to 34 years/Percentage            0.0718
```

We can see that compared to bottom neighborhoods, top neighborhoods have higher percentage of Mexican and Vietnamese restaurants in the area, higher proportion of Asian residents, higher percentage of housing prices ranging from $300k to $400k, and more females aged 30-34 years old.

To interpret these further, here are some key insights:

1) Regions with popular Mexican restaurants can be considered as nice neighborhoods for restaurant businesses. Probably because popular Mexican restaurants are supported by high foot traffic and people would stay around the area before/after the meal.

2) The demographics in a neighborhood also provides some information: If a neighborhood has higher percentage of Asian, it indicates a better business for restaurants. Probably Asians love dining out more?

3) Neighborhood housing. If a neighborhood has more housing prices ranging 300k-400k dollars, it's a good indicator that this neighborhood is a great option for dining.

4) Females in 30-34 years old. A higher percentage of this demographic group is also a positive sign for restaurant business. Looks like ladies in their 30s make great impact on restaurant businesses.

Conclusion

Based on the predictions from the best model: Random Forest model. We identified 3 areas as the top candidates for opening a restaurant: northwest (Spring Branch), southwest (Alief) and southeast (Clear Lake, South Belt/Ellington). These areas stand out because of already thriving restaurant business environment, high concentrate of Asian, healthy neighborhood housing and high concentration of young female group.

A good example is the Spring Branch district. This neighborhood harbors many headquarters of oil and gas companies as well as good quality residential real estate. Besides, this area also has the biggest Korean town in Texas and has rich resources that are appealing to customers especially female customers.

Discussion

Here, we used 3 common classification methods to model the problem of selecting best neighborhood for a restaurant. There are several discussion points.

First, we were using a lot of information to depict neighborhoods in a metropolitan area. Eventually, in the best model, we identified a few important features covering demographic, economic and industry specific information. While some features are more straightforward, some are more subtle for example the young female group. These findings suggest that using data-driven models can help uncovering some predictors that may not be apparent.

Second, we did not use the absolute values such as count as potential features because different cities have different scales and using absolute values may introduce bias, instead, here we used percentage values as a standardized value.

Third, for the missing data, we did not choose to impute them. The reason is that some neighborhoods miss all of their social-economic data, and one possible reason for missing data for such neighborhoods may be that these neighborhoods are less established, and this might indicate that there is a pattern for the missing data.

Finally, the model we built here did not contain more specific considerations such as: the type of the restaurant, the budget, the crime rate, the competitive landscape etc. I believe a more sophisticated model can be built with these further considerations.