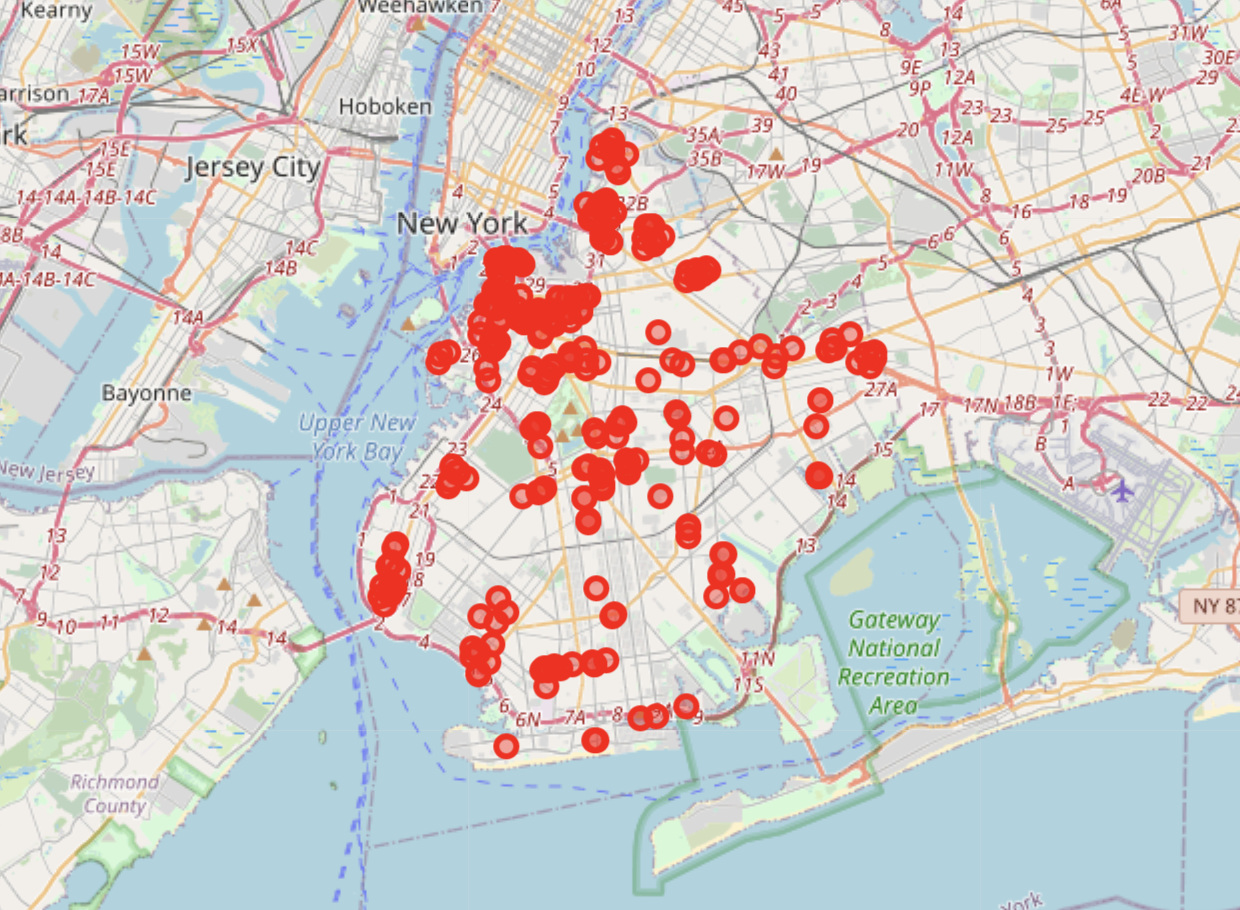
IBM Data Science Capstone Project

Predicting a Venue’s Price Range in Brooklyn Using Foursquare Location Data

Michelle Sun

# Introduction

* The business problem

The goal of the project is to predict the price range of a food venue. This can be potentially helpful to a new food business trying to set a competitive price compared to similar venues.

* Background

The price range for a restaurant can depend a range of factors like the cost of rent, labor and raw ingredient. These types of data are hard to gauge from your potential competitors in the area, but they are correlated to the location of the venue and the type of food it serves. Before opening the business, one would naturally want to know the profitability of the new venture by knowing what kind of prices it can charge the customers.

## Data

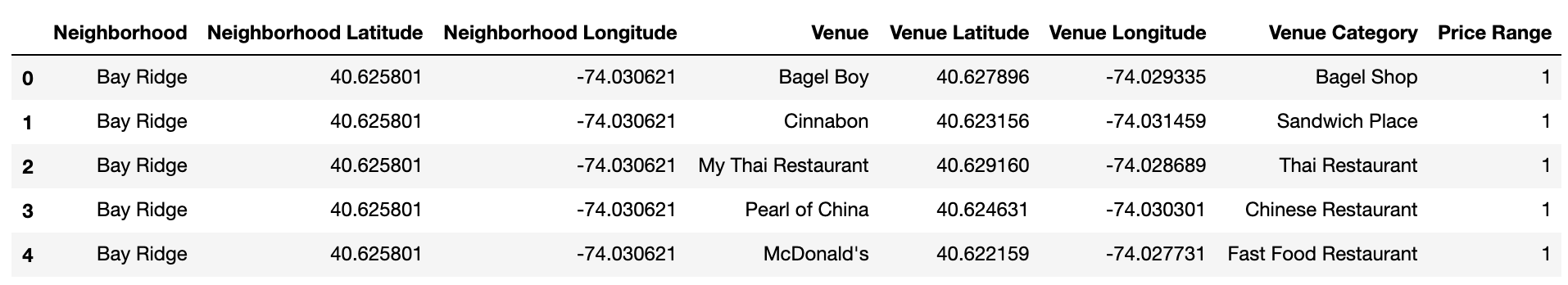
The Foursquare API free personal account has a limit on how many calls and premium calls can be made. Therefore, we chose to get the price information through the ‘explore’ endpoint over the ‘details’ endpoint, which is a premium call. There are 746 food venues in the Brooklyn area but only 500 premium calls can be made per day.

The ‘explore’ endpoint can retrieve price category specific results, category 1 being the cheapest and 4 being the most expensive. Hence, we will compile a data frame with the venue location, its neighborhood, category and price range.

## Methodology

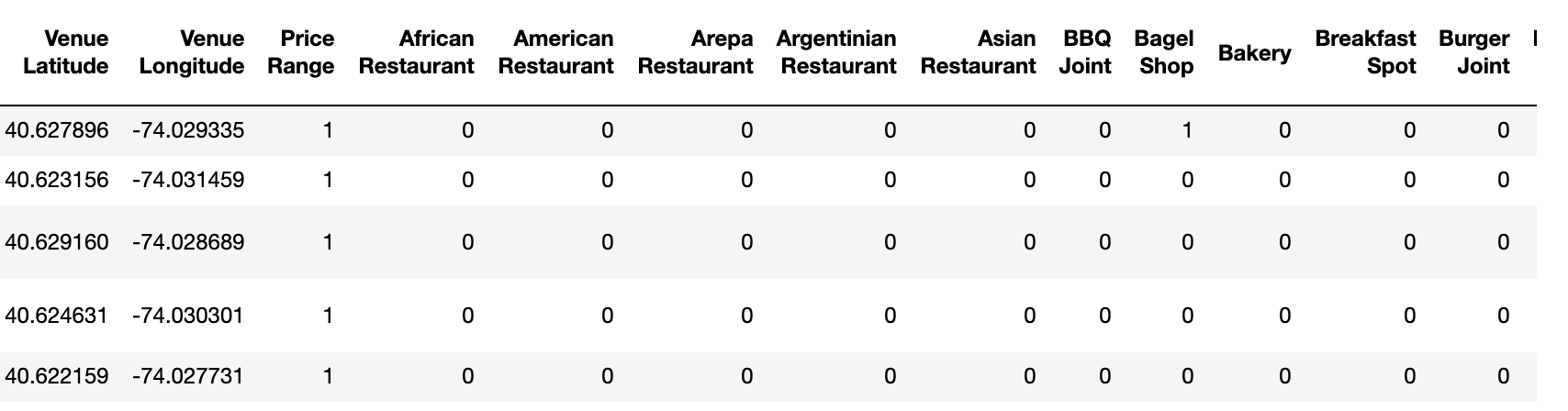
* Data Cleaning

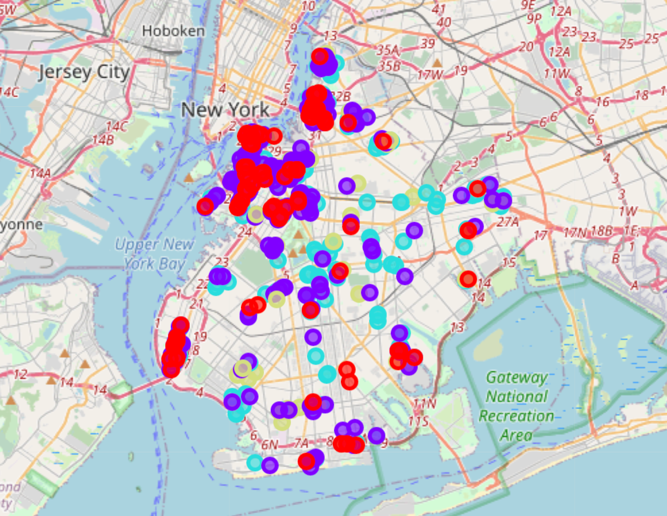
The data we get from Foursquare looked like this:



We can plot the venues with different price ranges on a single map. We can see clusters of regions where restaurants are cheaper (purple) and where they are more expensive (red). This indicates that the location of the venue has an effect on the price range. However, we would also like to know if the category and the neighbourhood matters when deciding a price range.

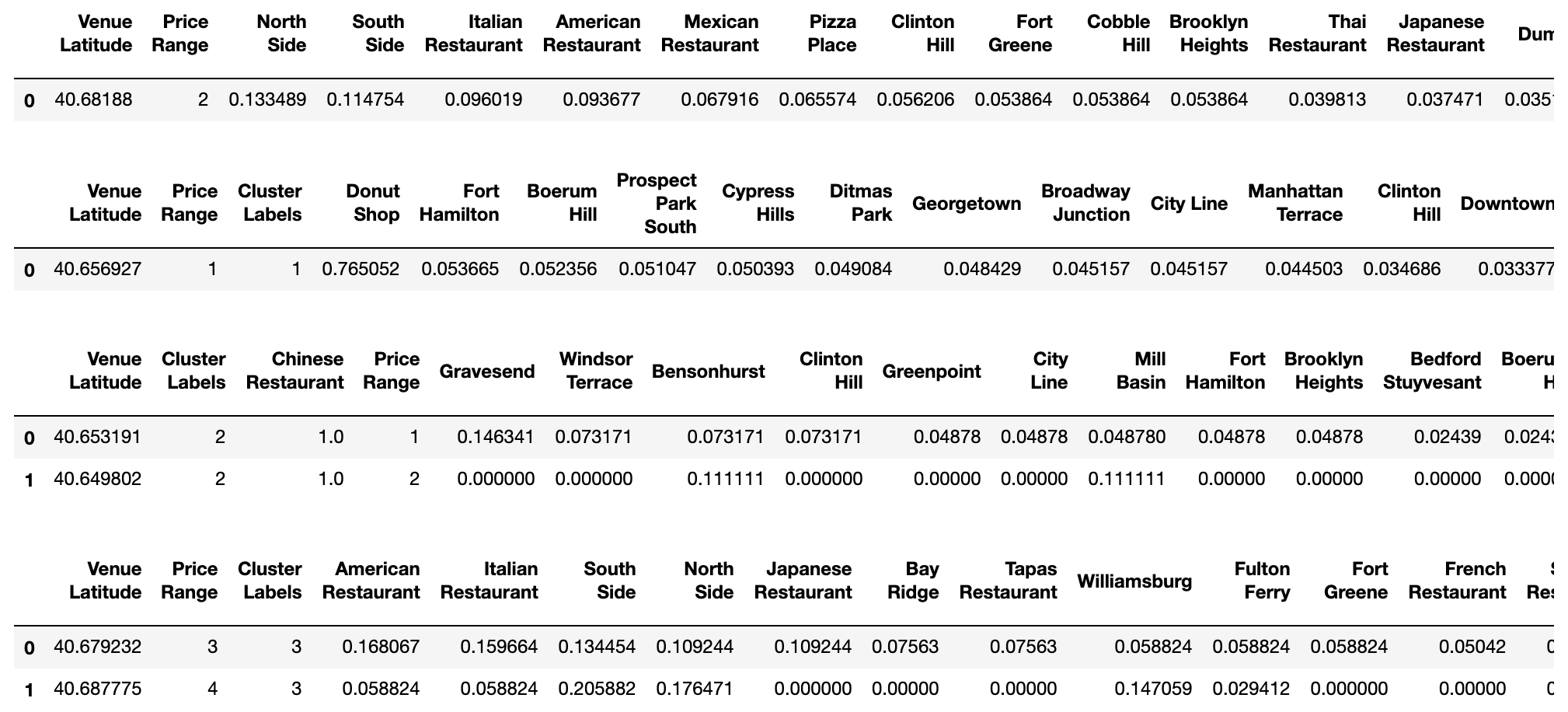
Before proceeding, the data need to be modified. The location of the neighbourhood is no longer needed. The name of the venue is also irrelevant. And the ‘Neighbourhood’ and ‘Venue Category’ cannot be string data. Eventually, we arrive at a data frame that looks like this:



* Unsupervised Learning (Clustering)

Now that the data is in an acceptable form, we will try to cluster the data together to investigate how separable the data is for classification. We first will use the K Means algorithm and display the clusters on a map:

The clusters are not entirely separable by location as we can see from the map. Therefore, we further examine the frequency of the categories and the neighbourhood in each cluster:



Some interesting observations can be made: first, clusters are related to price ranges: cluster 0 is lower mid-range, cluster 1 is mid-range, cluster 2 is low-range and cluster 3 is high; second, cluster 1 contains only Italian restaurants; and finally, cluster 2 has many donut shops.

We have also tried other clustering algorithms. However, they have yielded results that are less separable in terms of price ranges.

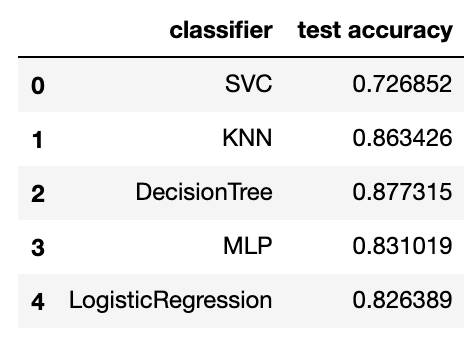
* Supervised Learning (Classification)

First, we need to separate the data into a training dataset and a test dataset by an 80:20 split. The training strategy will be to use a cross validation grid search on a variety of classification algorithms including K Nearest Neighbours, Neural Networks, Support Vector Machines, Decision Tree and Logistic Regression. These algorithms are chosen based on their popularity. In addition, they cover a wide range of non-linear and linear classification algorithms.

We also pre-processed the data to normalise it since the magnitude of the location data is significantly larger than the rest.

The grid search yielded the optimal parameters for each classification and we used them to build the final classifiers and apply them on the test dataset to obtain the final result.

Results

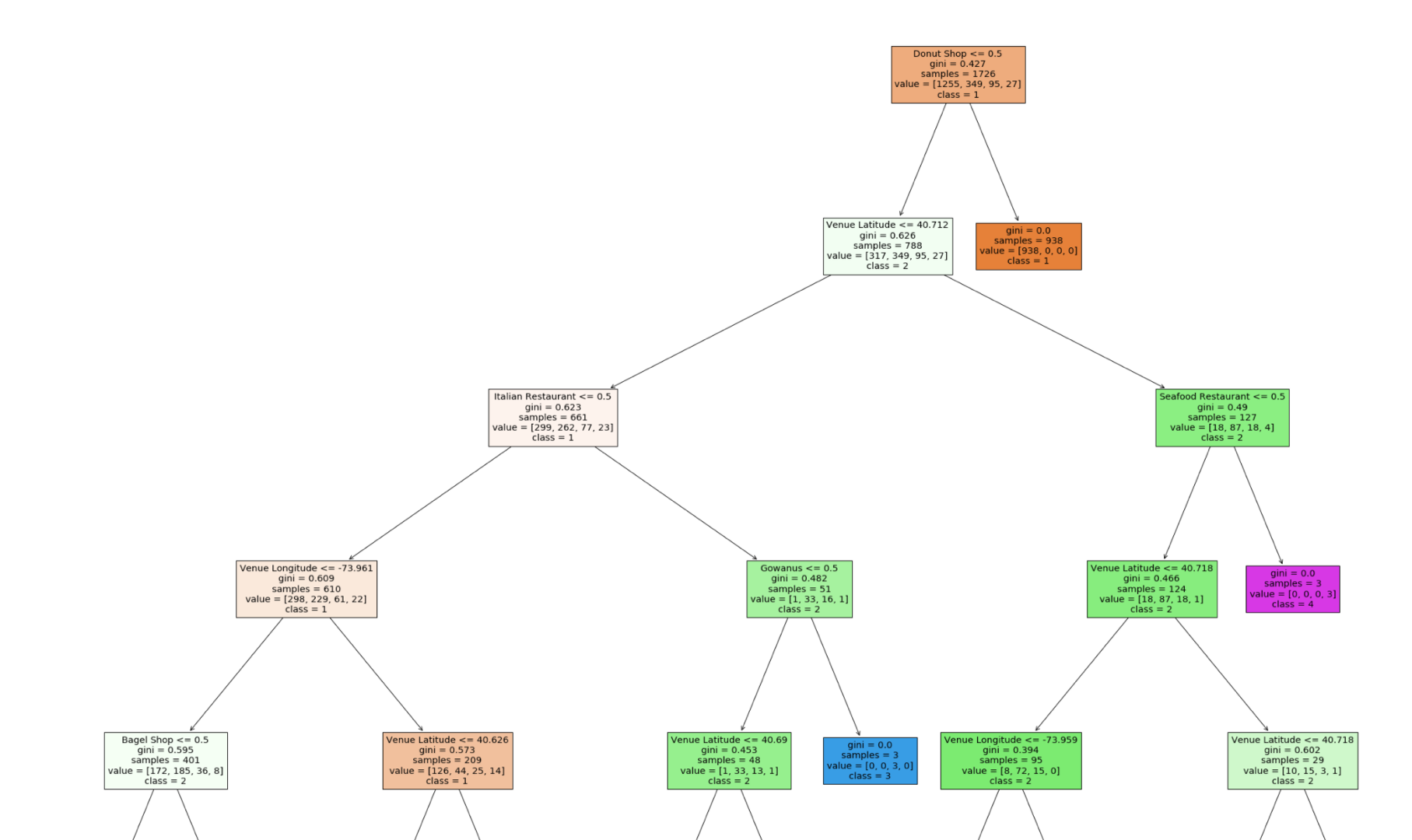
The results from each classifier compiled into a data frame as shown:

KNN and Decision Tree performed especially well in this case.

## Discussion

The relatively good performance of the decision tree and KNN means that the data is not linearly separable. KNN makes more intuitive sense as we can examine the ‘neighbours’ of a particular test case. For example, for a burger joint correctly predicted in the price range of 1 in the Downtown neighbourhood, the closest neighbour is a donut shop in the same area. Decision Tree is a bit harder to interpret as the depth of the tree is quite large. But we can print the feature importance of the classifier and potentially reduce the dimensionality of the data by removing some of the features with zero importance:



We can also print out the first couple of layers in the decision tree:

Some rules can be spotted immediate: like donut shops are always in the price range of 1 regardless of the venue location. Or that seafood restaurants above a certain venue latitude are always in price range 4 (closer to Manhattan perhaps?).

## Conclusion

The Foursquare data alone can provide a fairly good prediction on the possible price range of a new food venue. However, for us to perform more accurate predictions, we need more precise price data. This requires access to the menu and price for each item. In the future, these data can be retrieved from Foursquare with a paid account that has a higher limit on the API’s premium calls. A regression model in that case will be more suitable for predicting a precise pricing strategy.