Predicting the best neighbourhood for Opening a Restaurant in Scarborough, Toronto

Bashar Jaan Khan

July 23, 2019

1. Introduction

1.1 Background
Scarborough is an administrative division in Toronto, Ontario, Canada. Situated atop the eastern the Scarborough Bluffs, it occupies the eastern part of the city. Scarborough is a popular destination for new immigrants in Canada to reside. As a result, this increases the potential in revenue for recreational activities as more people from different cities with different tastes settle there. Therefore it is advantageous to have information about the existing restaurants to gain a competitive edge over other such businesses.

1.2 Problem
Data that might contribute to determining success of a restaurant might include the locality, popularity of food taste, etc. may define the success of that restaurant in the future. This project aims to predict the best locality of the restaurant based on such factors.

1.3 Interest
Restaurant joints and would be primarily interested in such information as it would help them target their customers in a more effective manner.

2. Data acquisition and cleaning

2.1 Data sources
The information about the Neighborhoods come from Wikipedia from here. This dataset lacks latitude and longitude information which is included from the file created by Cognitive Class from here. Also the location and popularity of the venues in these neighborhoods is found out using the FourSquare API.

2.2 Data cleaning
Data download or scraped from Wikipedia and Foursquare was combined into one pandas Dataframe. Scraping was done using BeautifulSoup. One Neighborhood had some NULL in latitude and longitude (postal code M1X) so it was dropped from the analysis.

The categories of the venues in that area were converted to one hot vectors to apply k means on them.

Foursquare provided the popularity of the various venues in the form of number of likes.

2.3 Feature Selection
After data cleaning, there were 16 data samples and 64 features. The most common 10 venue categories were selected to analyze. Then the popularity of each venue category (for eg: Italian Restaurant, Coffee Shop, etc) for each neighborhood was found by taking the mean for all such instances. This was then added to the cleaned dataset.

Furthermore, the total popularity of the area was calculated by summing the likes for all the 10 most common venue categories for each locality.

3. Exploratory Data Analysis

3.1 Relationship between restaurant popularity and neighborhood popularity

It was seen that as the more popular the Restaurants in the area was the more popular was the Neighborhood it belonged in. This can be due to a higher density of people using that area. Moreover, this trend was found in less popular areas too. This suggests that the new restaurants in the area are expected to generally well compared to other areas.

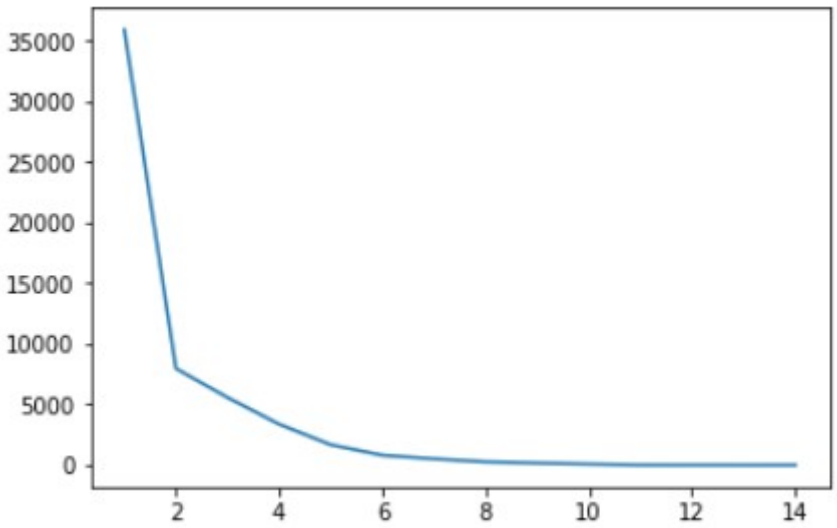
3.2 Applying standard K Means and its problems

Kmeans tends to not work very effectively on high dimensional datasets. Fortunately our input space had 64 vectors which is not a problem (compared to large values like 256x256x3 or more in case of images). Also different values for the number of clusters need to experimented before using one.

4. Modelling

4.1 Clustering with different number of clusters

The number of clusters were set from 1 to 11. There seems a clear elbow at around n = 6 after which the model does not perform any better. The error in this case is 806.3395916052206.



A measure of error vs number of clusters

5. Conclusions

In this study, the optimal location to open a new restaurant was found by clustering the localities by taking in account their similarites and dissimilarities in popularity. This model can be useful for stakeholders as a starting point for opening a new restaurant in Scarborough, Toronto.

6. Future Directions

Data from more social media outlets could be collected to perform a better evaluation of the popularity of the venues. Also some other algorithms for clustering could be performed which work on less data and higher dimensional datasets.