Untitled

January 22, 2019

1 Capstone Project

1.1 Introduction

I am going to be trying to help both businesses and individuals who are interested in social, outdoor activities. People in this category may want to know how many parks are close to a given neighborhood since this will give more opportunities for outdoor social activities. Businesses may want to know what neighborhoods are lacking these amenities since it may indicate areas of the market that are under-served, and they could offer alternatives to parks (such as golf courses). If, for example, a business is considering building a golf course, they will need to know if there are other golf courses in the area they will need to compete against, and they will need to know how many parks are available which may be (economically speaking) a substitute for golf courses. On the other hand, parks may be a compliment to golf courses, either way, it is vitally impr=ortant for businesses to know information about parks and golf courses in the area.

1.2 Data

I will be leveraging four-square data to group neighborhoods in Toronto based on the number of parks within a given radius, number of golf courses within the same radius, and the average distance of those parks and golf courses (this will indicate whether the parks and golf courses are near the center or concentrated on the edges of the circle).

1.3 Methodology

For my query to Four Square I will be using the query 'Parks Golf'. This query has shown to be very effective in finding all golf courses and parks in the supplied search radius. Once I have queried Four Square for each neighborhood in my data set, I will engineer features by counting the number of golf courses and parks in a radius around each neighborhood. Additionaly, I will find the average distance for the parks and golf courses. I can tell which results are parks and which are golf courses by the category column returned from the search query. Last of all, once the features are created, I normalize each feature.

Once the results were obtained, and the features created and normalized, I used the Kmeans function supplied by sklearn, to cluster based on these features. You can see the visualization of these clusters in the provided notebook. Upon initial examination, it appeared that the clusters were determined almost entirely by golf courses, so I weighted the parks features (both distance and count) so that the clusters are determined by both parks and golf courses. The weighting was done by simply doubling the values in the park count and average distance features. When

I rewighted in this manner, I found the clusters to be much more dependent on all features. I further found that there were some very logical splits to create clearly defined clusters.

1.4 Discussion

Based on the results of our clustering, we can clearly see that there are groups which are underserved. The purple and orange groups are far away from golf courses, so these are good candidates for building a golf course. In particular, the purple has very few golf courses, and very few parks, thus these neighborhoods can take some serious consideration. If an individual is using this data to find what neighborhoods are desirable to live in, the green group seems to have a good combination of parks and golf courses.

1.5 Conclusion

The results seem to indicate the possibility of investment opportunities for businesses, but the results are only indicators. More analysis may be necessary before a choice is made. For example, some of the neighborhoods may on average attract individuals who are not looking for outdoor social activities. If this is the case, the lack of outdoor ammenities in certain areas may be indicative of a lack of demand for these services, and not a lack of service. In any case, these results provide a valuable starting point for further research.