**Coursera and IBM**

**Data Science Capstone Final Report**

**Michael Ding**

**Applied Data Science Capstone**

**June 14, 2020**

Data Science Capstone Final Report

**Introduction/Business Problem**

It is no doubt that cultural diversity serves an integral role in the success of many Western nations in an increasingly globalized world. Multiculturalism helps countries build long-lasting products that capture a broad global audience. But when immigrants come to countries like Canada, it is critical that they find a supportive community for their shift and integration as part of Canadian society. Elsewise, immigrants feel secluded and become unwilling to contribute the cultural melting pot of diverse cities like Toronto, NYC, and Sydney.

Our problem today will focus on determining whether we are in Greek town, given any pair of geocoordinates in Toronto. Our scenario is to help a Greek family determine if the house locations they are looking to purchase are part of Greek town. To solve this problem, we will fit a decision tree model and use venue frequencies that are classified by the Foursquare API. Ultimately, with future development, we can determine the cultural community of any home for homebuyers from foreign nations.

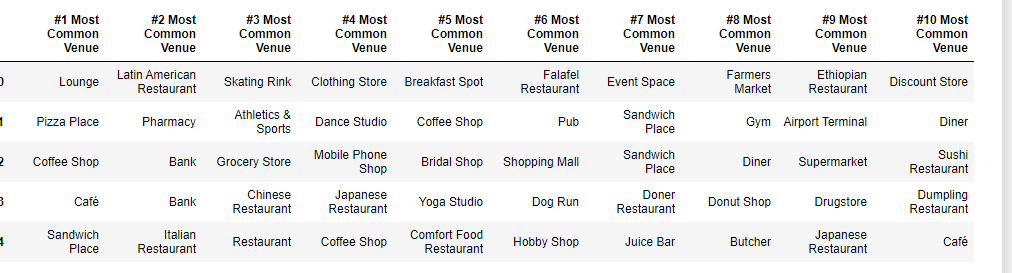
**Data**

As mentioned, the method we will use to solve our problem is to examine places data from the Foursquare API. For defining individual neighbourhoods, we use a Wikipedia link of all postal codes and neighbourhoods within the Toronto area ([Link](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)). We will find the 10 most common venues within neighbourhoods and build a decision tree to determine if any given geocoordinates are within Toronto’s Greek town or not.

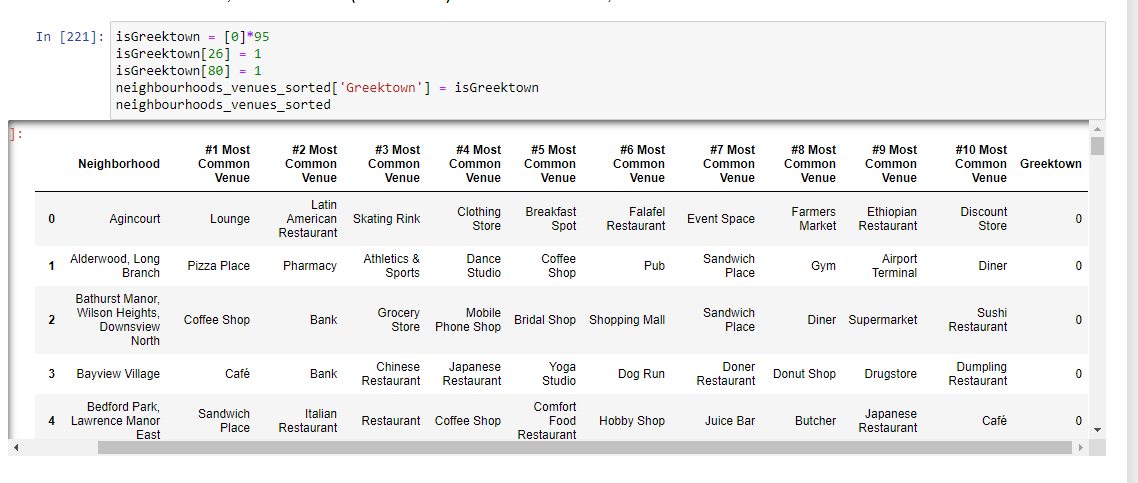
**Methodology**

In terms of exploratory data analysis, upon loading all the individual neighborhood latitudes and longitudes and searching for local venues using the Foursquare API, we compiled all the different types of venues that belong to each neighbourhood. This ultimately allowed us to classify the frequency of different types of venues within each neighbourhood and was integral to our development of the decision tree model.

A decision tree model was used to determine whether, given the top 10 most common venues, if we were in Greek town or not. We use the decision tree model because we want to ultimately build a classification model, and not a clustering model like the K-clusters algorithm. We classified each neighborhood as Greek town (1) or not (0). We casted all strings to distinct ints in order to train our model.

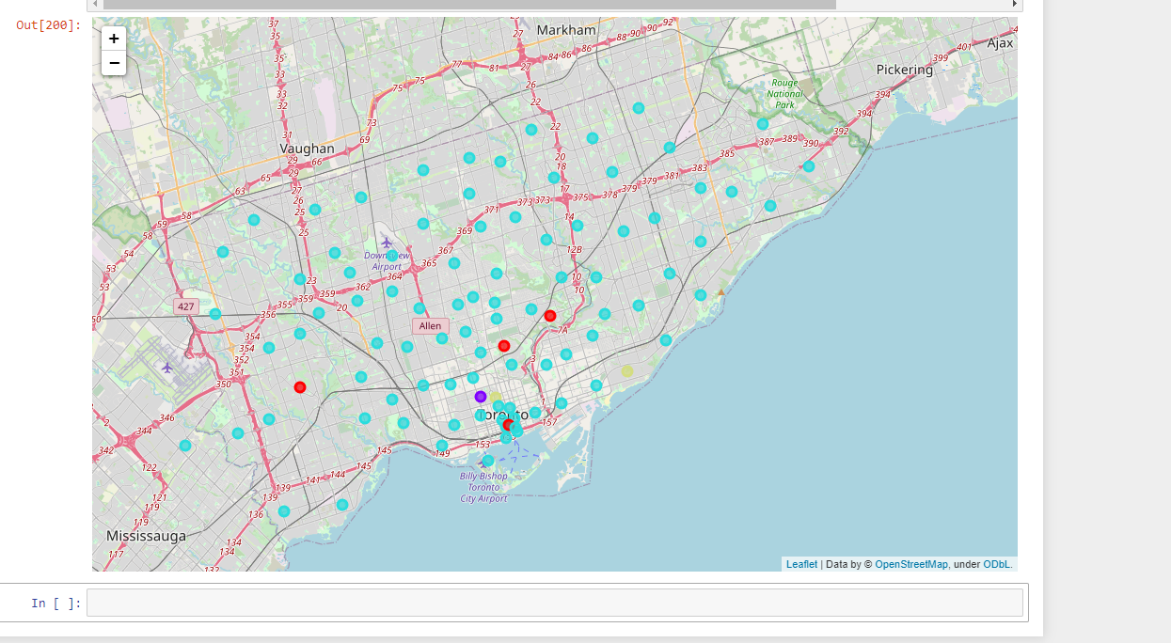


**Figure 1**

****

**Figure 2**

**Results**



When put through our decision tree model, the following results were found. All blue dots represent cells that were correctly classified as not greektown by our decision tree while the purple dots are false negatives and green dots are true positives. The red dots were false positives. The accuracy of our model was 93%.

**Discussion**

We should in the future use more datasets and train our model using more attributes about local venues in which the classification for greektown were “1” so that we can build a more accurate decision tree. Furthermore, we could perhaps have increased the depth of our tree, but it is possible that this could have generated an overfitting model.

**Conclusion**

Based on our model, we were able to correctly predict whether a given location was in a Greek town or not with 93% accuracy.

Overall, this course has given a lot of insight as to how data scientists structure their work and the process they undergo to solve problems. I hope to build more models like this in real world applications such as data science challenges in the future.