IBM Data Science Certificate Capstone Project

Introduction

Where’s the best neighborhood to build a coffee shop in Toronto?

Say we want to start our own coffee shop, because we are passionate about coffee, but Toronto is already flooded with coffee shops. We want to give our coffee shop a chance at succeeding in such a market, and location is key.

The first step in choosing a location is choosing a neighborhood. Every neighborhood has coffee shops, but some have more than others. We could just count the number of coffee shops in every neighborhood, but this raw approach is not nuanced enough. For example, some tourist areas may already have many coffee shops; but because of the number of people coming though these neighborhoods, they may be able to support more coffee shops.

I propose an approach where we sample the venues around each neighborhood and determine the proportion of each type of venue. We use these proportions as features for a linear regression to predict the proportion of coffee shops in a neighborhood. This should provide our analysis with the nuance we need.

With the regression, we can evaluate several potential neighborhoods. For each neighborhood, we compare the output of the regression with the actual proportion. Those neighborhoods in which we predict a larger proportion of coffee shops than exist hold the most promise of supporting additional coffee shops.

Data

The data for this project closely parallels the data used in the Toronto lab assignment. In particular, we start by using postal code information from Wikipedia to identify neighborhoods. We get the location information from each neighborhood using the .csv file provided for the lab.

For each neighborhood, we query FourSquare to get a sample of venues from each neighborhood. For each sample, we can calculate the proportion of each type of venue.

We can combine “coffee shop” and “café” since they are, for our purposes, identical. We extract this combined field’s proportion as our target and the remaining venue proportions as our features.

Methodology

TODO

Results

TODO

Discussion

TODO

Conclusion

TODO