Prediction of Ideal Neighborhoods to Open New Coffee Shops

Introductiuon

Toronto is the biggest city and financial center of Canada with a population of 6,197,000 (as of year 2020). The city is divided into 103 postal code neighborhoods, which are highly heterogeneous in function and population. The residence in each neighborhood also highly diverse in income, education, ethnicity origin et al.. Thus the demand of coffee shops is neighborhood specific.

An ideal neighborhood to open new coffee shops is determined by the law of supply and demand. If the predict number of total coffee shops in a neighborhood is larger than the existing number of coffee shops, there is demand over supply. Thus, the neighborhoods would support a new coffee shop.

Data Acquisition and Cleaning

The Toronto city postal code and neighborhood were scrapped from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) . I parse the table of postal code to a panda dataframe with BeautifulSoup.

The latitude and longtitude data of each postal code were obtained from geocoder. However, due to some technical difficulties with geocoder server, I use the alternative approach, a csv table provided by capstone project. The postal code and coordinate tables were merged into a single dataframe.

I explored the neighborhood with FourSquare API with the latitude and longtitude of each neighborhood. The venues were limited by radius of 500. All venues names and category were added into a new dataframe, which then transform into one-hot format. This new one-hot table has 2123 venue entries and 269 categories.

I tallied the number of venues in each categories in each neighborhood. Since some categories have very small number of venues, I only keep the top 20 categories (at least 22 venues in each categories). This new table was used for training linear regression model to predict the number of coffee shop in each neighborhood.