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

In my initial exploration of the data, I noticed that both "Coffee Shop" and "Cafe" were categories. The differences between the two are subtle enough that I chose to relabel all instances of "Cafe" as "Coffee Shop".

When I sampled the venues from the neighborhoods from FourSquare, I limited my search to 100 items per neighborhood. After transforming the venues into a one-hot, and grouping by neighborhoods, I tried

summing up each neighborhood. The vast majority, but not all, of the neighborhoods summed to 1.0 as expected. The ones that did not were because the FourSquare query returned less than 100 venues. Thus, I renormalized my neighborhoods so that the sum of each row would be 1.0. Thus, the frequency measure reflects the probability of a venue in a neighborhood being a particular type.

For the purposes of this activity, I randomly chose 10% of the neighborhoods to be in the test set. If I truly were looking to open a coffee shop, and I had a set of neighborhoods in mind, I would have saved those for the test set.

I chose a RandomForestRegression, as ensemble methods usually produce better results. I purposely let the decision trees go deeper than usual in the forest since I have over 200 features. I used mean-squared-error as the metric to evaluate the regression.

For each value in the test set, I subtracted the actual value from the predicted value and I chose the maximum difference. This showed the neighborhood most underserved by coffee. (On the other hand, the minimum value would have shown the most over-saturated coffee market in the set of neighborhoods.)

As a final step, I ran the regression on all the data, and again identified the biggest difference. This would show the best neighborhood in the entire city.

Results

The mean squared error return a value of 0.0481, which gives me some confidence in the test.

Downsview Northwest is the neighborhood with the largest difference.

When I ran the analysis across the entire set, Downsview Northwest remained the best location.

Discussion

Based on the analysis, Downsview Northwest is the best location for a coffee shop.

Although Google Maps labels it as Jane and Finch, the postal service considers it to be "Downsview Northwest." You can read more about Jane and Finch here: https://en.wikipedia.org/wiki/Jane_and_Finch

Although it is not reflected in the Notebook file, I looked at the other extreme – the neighborhood with the smallest difference (or rather, the largest negative number.) Scarborough's "Woburn" neighborhood was the most over-saturated coffee market.

Conclusion

This exercise was theoretical, but it shows how data can be used to make important decisions. Of course, identifying the least-saturated market is not enough to make a final decision about where to open a coffee shop. Several other factors would be involved in making this decision.