**Finding the Best Location to Open a Pizza Store in Toronto using Foursquare API.**  
Author: Nicholas Boldt

**Introduction:**

Opening a new business is a risky venture. If the proper location is not chosen, then the business may fail, and the investors will have lost their money for nothing.

Assume a theoretical client wants to open up a pizza store in Toronto. Using Foursquare, we can gather information about nearby venues centered around postal code locations. Census data is also available from the Canadian government. Using this data, we can predict where the client should open their new pizza business through means of a dissimilarity metric, as well as a logistic regression model.

**Data:**

Foursquare API can be used to find venues around a given location. Using it, we queried up to 100 venues within 500 meters around each postal code. This would allow use to create a profile for each postal code location.

The Canadian government’s 2016 census data is publicly available, and was used to find the population at each postal code location (https://www12.statcan.gc.ca).

**Methods:**

The Foursquare API was used to collect information about local venues at each Toronto postal code. We queried up to 100 venues within 500 meters around each postal code. Using this query result, we created a one-shot matrix listing the count of each possible type of venue at each postal code.

There were 276 unique venues, whose counts were averaged across postal code, such that we created a profile of the types of businesses at each postal code. Hence, each postal code’s profile had 276 dimensions. We standardized the values for each dimension across all postal codes.

There were 34 postal code locations that had a nearby pizza place, and 66 locations that did not have a pizza store nearby. We filtered out the locations without pizza stores that were within 1 km of a postal code with an existing store, leaving 59 potential pizza store locations.

The postal codes with pizza places were used to find the average profile of a postal code with a pizza store. The profiles of each potential location were compared to this average profile (Euclidian distance) to create a dissimilarity metric.

Using this dissimilarity approach essentially applies an equal weight to all measures, which can be suboptimal. Hence, we also tried logistic regression, which would learn the optimal weights to apply to all provided inputs. As before, we created the profiles for each postal code, but did not filter out the locations near existing pizza stores, as these could be valuable inputs to feed to the logistic model. We also included the population information at this time.

Standardizing the profiles with population information, we first used in-sample prediction with a logistic model, where the model created predictions for data it was trained on. The problem with this approach is that we already assigned a zero label (no pizza store) for our potential locations, which will influence the model’s weights.

Therefore, we performed out-of-sample prediction using all the data to train the model, minus the data for a potential location. We repeated this for all possible locations, to find probabilities that there should be a pizza store at that given postal code location.

**Results:**

The dissimilarity metric, in-sample probabilities, and out-of-sample probabilities are plotted separately on maps of Toronto.

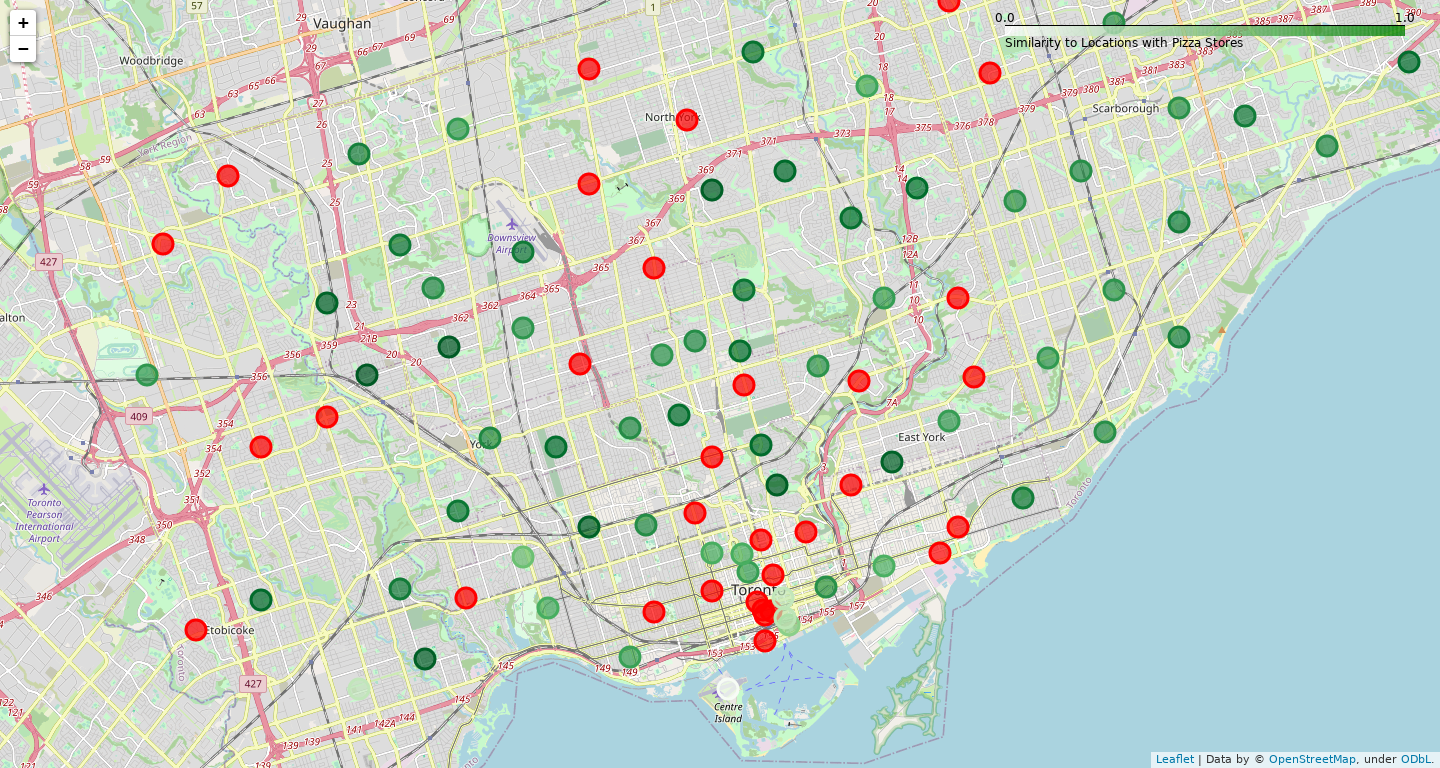


Figure 1: Dissimilarity metric results. Each dot represents a postal code location, where the red dots are locations with existing stores. The shade of the green dots represent similarity of locations without pizza stores to locations with existing pizza stores.

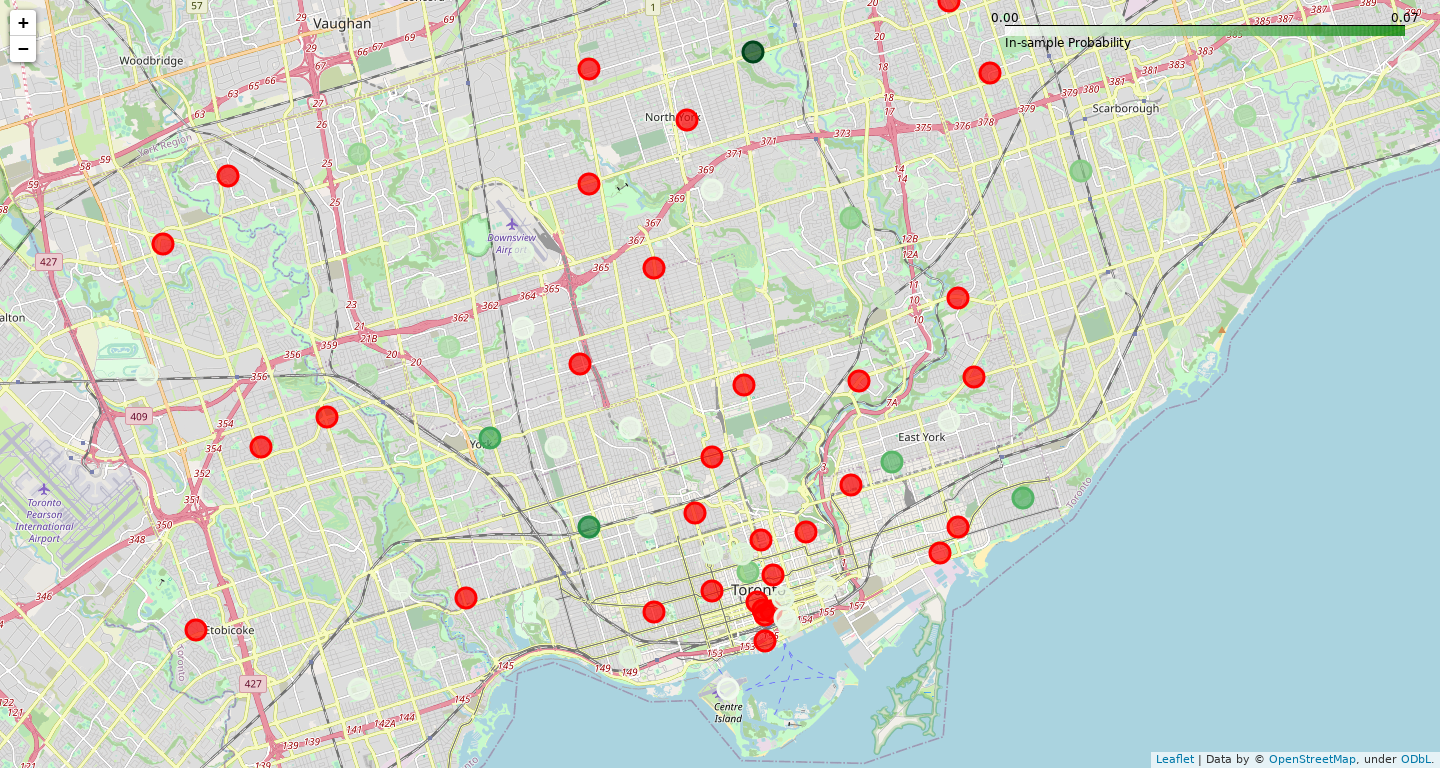


Figure 2: In-sample logistic regression prediction results. Each dot represents a location, where the red dots are locations with existing stores. The shade of the green dots represents the probability that a given postal code location without a pizza store ‘should’ have a pizza store.

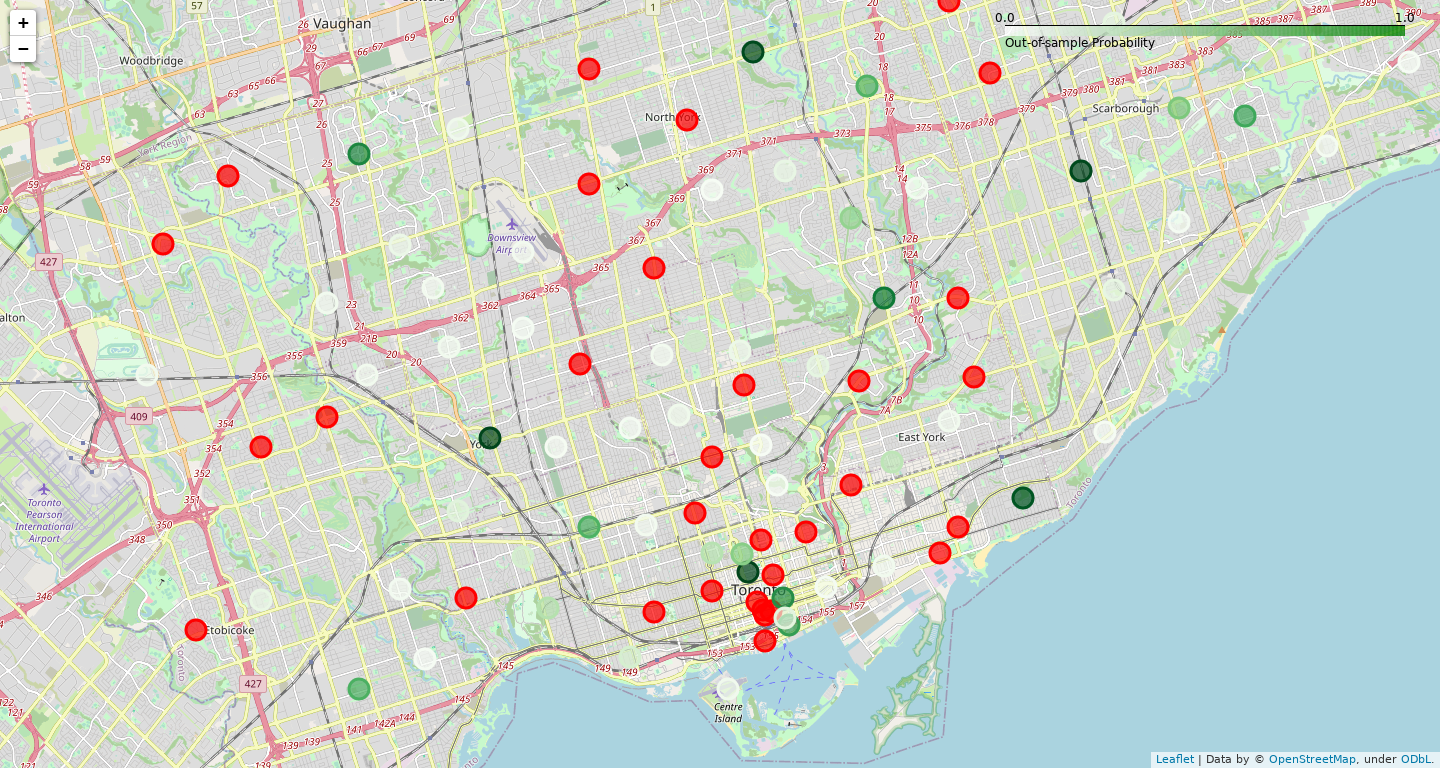


Figure 2: Out-of-sample logistic regression prediction results. Each dot represents a location, where the red dots are locations with existing stores. The shade of the green dots represents the probability that a given postal code location without a pizza store ‘should’ have a pizza store.

|  |  |
| --- | --- |
| **Metric** | **Weight** |
| Sandwich Place | 0.795417 |
| Pharmacy | 0.629754 |
| Pub | 0.583500 |
| Furniture / Home Store | 0.527276 |
| Turkish Restaurant | -0.439646 |
| Coffee Shop | 0.429255 |
| Gym | 0.390059 |
| Health Food Store | -0.383757 |
| Cosmetics Shop | 0.379041 |
| Fried Chicken Joint | 0.345571 |

Table 1: Greatest weights of the logistic regression model (out-of-sample).



Table 2: Top ten potential locations, sorted by out-of-sample probability determined with logistic regression. Note that some locations are close to a postal code location with an existing pizza store.

**Discussion:**

First, we tried using Euclidian distance to measure the dissimilarity between the venue profiles of potential pizza store locations and an average profile of those with existing pizza stores. The results in Figure 1 show that many locations are considered good choices, when all input features are weighted equally. However, not all features are equally important, so this approach is not optimal.

Using logistic regression, we first calculated the probability that a postal code location should have a pizza store using all the data, including the possible locations that we wanted to predict. The results of this in-sample prediction are shown in Figure 2.

This is a huge improvement over using Euclidian distance, as seen by the much narrower selection of good choices (darker green dots show locations with a higher probability). We decided to try in-sample prediction due to the low number of data samples (n=100), but this may have negatively impacted the results, since it assigns zero labels to potential locations when training the logistic regression model.

Hence, we trained many logistic regression models, using the entire data set minus a given potential location. As a result, we calculated the probability that a potential postal code location should have a pizza store, as shown in Figure 3.

As we pointed out earlier, the dissimilarity metric was suboptimal as it weighted all features equally. For out-of-sample prediction with logistic regression, a different model was trained for each potential location. The weights for all models were averaged, and the weights with the greatest magnitude are shown in Table 1. For example, postal codes with pizza stores tend to also have sandwich places and pharmacies. Conversely, postal codes with a nearby Turkish restaurant or health food store are unlikely to also have a pizza store. Notably, the weights for the census population features were quite low, at -.014 and -.007 for (Population, 2016) and (Total Private Dwellings, 2016) respectively.

The top ten locations, sorted by out-of-sample probability, was reproduced in Table 2. We did not exclude locations based on their proximity to nearby postal codes with existing pizza stores when training our logistic regression models, as these data samples could reveal information about why nearby stores did not have a pizza store. This is seen in Table 2, where some of the top locations are within a kilometer of postal codes with existing pizza stores.

A safe recommendation would be to exclude those locations that are too close to pre-existing stores. Therefore, we would recommend our client start their business in one of the other top locations, namely at the postal codes ‘M6M’, ‘M1P’, ‘M2K’, or ‘M4E.’ Our model predicts that these locations ‘should’ have a pizza store, with a probability of 94.9% or higher. In other words, these locations have similar environments and thus similar underlying consumer populations, as those locations with existing pizza stores. Hence, our client should be able to open a successful pizza store business at one of these recommended locations.

**Conclusion:**

Our goal was to determine where our client should start a pizza business, using venue and population information collected about each postal code location in Toronto. Using logistic regression, our model indicates that our client should open their store at the following postal codes: ‘M6M’, ‘M1P’, ‘M2K’, or ‘M4E.’