Predicting Cluster Labels of Edmonton Banks based on Neighborhood Characteristics

Sean Hong

April 06, 2020

Introduction: Business Problem

This capstone project analyzes the characteristics of the most common establishments that are found closest to banks in Edmonton. Hypothetically, banks would be found in areas where they can provide banking services to a good number of people enough to stay profitable. Factors for choosing a location would depend on each individual bank company's ability to scale, their number of clients, and even the presence of other banks, in what is known as game theory in economics. Less-known, however, is the connection of the bank branches' physical location with those of nearby establishments. Do these establishments tend to cluster in the retail industry? In coffee or other refreshment services? How about other banks themselves? And given information on these types of establishments, can they be used to predict which cluster a particular bank branch belongs to? This report will examine such questions.

I have chosen the city of Edmonton in the Canadian province of Alberta as a case study. Edmonton is Canada's sixth largest city with a population of over 1.3 million. Its urban landscape is distinctly suburban and has a decent-sized downtown. Edmonton is home to a fairly robust banking industry and houses the Canadian Western Bank, the only publicly traded Schedule I chartered bank headquarters west of Toronto.

I examined specifically what the most common venues situated near banks are, as well as their corresponding venue category. The primary stakeholders of this capstone project are multitude: it will benefit Edmonton's urban planning department, business investors, and even the banking companies by providing useful insight into the businesses that cluster around banks, including but not limited to:

- Which banks are clustered around each other
- The types of establishments that are located near banks
- Relation between a specified bank cluster and the most common establishments that operate within said cluster

Data

To retrieve the banks' addresses in Edmonton, I referred to BankChart.ca for information from the site: https://bankchart.ca/catalogue/branches/3. Two variables are of particular interest:

- Bank name

- Addresses of the bank branches

Once the relevant data was stored in a dataframe, the following were performed:

- The approximate locations of each branch's address were retrieved using Google Maps API reverse geocoding.
- Information on the venue category of the establishments were collected using Foursquare API.

These steps provided the necessary data ripe for answering the business problem, upon which the results were reviewed for drawing useful conclusions.

Methodology

To introduce latitudes and longitude coordinates to our data set, I made use of Google Maps API reverse geocoding.

I called in a prepared csv file containing the names of banks in Edmonton and their corresponding addresses. Then, a loop was performed to retrieve the respective coordinates and match them with the addresses.

	Bank Name	Address	long	lat
0	Bank of Montreal	208 KINGSWAY MALL, EDMONTON, AB T5G3A6: Edmonton	-113.505	53.5634
1	Bank of Montreal	2 HEBERT RD. UNIT 200, ST.ALBERT, AB T8N5T8: E...	-113.607	53.6223
2	Bank of Montreal	2304-24TH STREET NW, EDMONTON, AB T6T0G9: Edmo...	-113.378	53.4544
3	Bank of Montreal	17504 - 100 AVE NW, EDMONTON, AB T5S2S2: Edmonton	-113.623	53.5392
4	Bank of Montreal	300 222 BASELINE RD, SHERWOOD PARK, AB T8H1S8:...	-113.317	53.5421

Having compiled the location data of our banks in Edmonton, I then used the Foursquare API to get info on each branches' nearby establishments.

First, I defined and made use of the getNearbyVenues to retrieve the venues that are closest to the bank coordinates, and then store the information into a dataframe. After that, further data wrangling was performed to ready the dataset for analysis.

The establishments' venues were grouped by each bank location. I proceeded by defining a function that returns the most common venues located for each given coordinate. Then I performed a clustering analysis to segment the banks and their respective locations. Five clusters were formed to segment the banks.

For this capstone project, I limited the number of independent variables to just the top three most common venues, and then predicted the corresponding cluster label that we got from the previous clustering machine learning analysis that I had completed, resulting in this dataframe.

	Bank Name	Address	long	lat	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Bank of Montreal	208 KINGSWAY MALL, EDMONTON, AB T5G3A6: Edmonton	-113.505	53.5634	2	Fast Food Restaurant	Clothing Store	Coffee Shop
1	Bank of Montreal	2 HEBERT RD. UNIT 200, ST.ALBERT, AB T8N5T8: E...	-113.607	53.6223	2	Fast Food Restaurant	Grocery Store	Coffee Shop
2	Bank of Montreal	2304-24TH STREET NW, EDMONTON, AB T6T0G9: Edmo...	-113.378	53.4544	2	Liquor Store	Grocery Store	Pharmacy
3	Bank of Montreal	17504 - 100 AVE NW, EDMONTON, AB T5S2S2: Edmonton	-113.623	53.5392	2	Fast Food Restaurant	Coffee Shop	Hotel
4	Bank of Montreal	300 222 BASELINE RD, SHERWOOD PARK, AB T8H1S8:...	-113.317	53.5421	2	Fast Food Restaurant	Pizza Place	Gas Station

For further analysis, I was interested in finding out whether the type of most common establishments located near banks have any relation to the cluster of their respective banks. Is there any observable pattern in these neighborhoods?

The dataframe below is a screenshot of the relevant data that I used for the classification analysis. Models were trained to predict the five 'Cluster Labels' using the information we had on the most common establishments that were situated closest to banks.

I used two classification techniques, Logistic Regression and K-Nearest Neighbours Classifier to predict the cluster label of the banks in the dataset given their corresponding top three most common establishments.

Overall Results and Discussion

Clustering

From the results, it was observed that 'Coffee Shop' is the most common type of establishment that is located close to Edmonton banks, at around 15%. This was followed by 'Fast Food Restaurant' and 'Pizza Place', which came at 10% and 8% respectively. In fact, the top three venue categories involved sit-down dining establishments and collectively account for over a third of establishments that are closest to banks.

After performing the clustering analysis, it was found that the sizes of the clusters were very uneven, with the cluster label '1' being almost as large as the other clusters combined. In fact, the smallest cluster, '4', had a population of one, that is, only one bank location was found to belong to this cluster, which suggests that this data point is an outlier. The clustering analysis showed that in Edmonton, there was a main node (representing cluster 1) for which most banks were located. As this specific cluster accounted for almost half of the locations in our database, it can be said that this node plays an outsized role in banking services in Edmonton or at least, that this node is the most representative of the venue categories that are situated near banks. This was followed by a smaller but still significant node (population of 40) and then to a yet smaller one (population of 23), after which only distant outliers remained.

Classification

Two classification methods, Logistic Regression and K-Nearest Neighbours, were performed to predict the cluster labels given the top three most common venues as the independent variables that form the framework of the hypothesized analysis.

The logistic regression analysis that was performed resulted in a Jaccard Similarity Score of 0.533 and an F-1 score of 0.371. Whereas the K-Nearest Neighbours resulted in a Jaccard Similarity Score of 0.633 and an F-1 score of

0.550. This suggests that the K-Nearest Neighbours is better suited than logistic regression to predict the cluster labels given the data we have on the most common venues as provided by Foursquare.

Using K-Nearest Neighbours as our algorithm in classifying venue categories and bank clusters, we can expect to get our predictions right more than half of the time.

The results of this project provides further insight into the methodology of assigning clusters following the theory of 'neighborhood characteristics'. Narrowing down the criteria for selecting which venue category to include (and exclude), and performing other classification techniques, are possible avenues for further refinement of the methodology.

Conclusion

Broadly-speaking, the purposes of this project was multifold:

- To segment the banks into clusters by location
- To observe the most common types of businesses that tend to be established nearby banks and,
- To determine whether specific characteristic of a bank's neighborhood can accurately predict the bank's cluster

From retrieving the coordinates of the banks' addresses, we were able to generate a 'neighborhood picture' of each bank index using Foursquare data. Then, using clustering analysis, assigned the banks a cluster label from 0 to 4, representing five clusters in total. Using the first index as an example, we could say that:

Bank of Montreal, located at *208 KINGSWAY MALL, EDMONTON, AB T5G3A6* with the neighbourhood description: "Fast Food Restaurant", Coffee Shop", and "Clothing Store", belongs to Cluster 1.

Finally, we were able to build a decent classification model using the K-Nearest Neighbours method to predict the cluster labels given the generated 'neighbourhood picture'.

As mentioned earlier, the main stakeholders that will benefit from this project are Edmonton's urban planning department, investors, and even the bank companies themselves. Inasmuch as they learn what types of establishments are located near banks, decisions on how land in Edmonton should be zoned, what businesses (e.g. coffee shops, clothing stores) are suited to be located nearby banks, and where the banks ideally should expand their footprint are all decisions that this study will help in facilitating.