Ammar Baig

IBM - Capstone Project Report

May 28, 2020

Open an Indian Restaurant in Calgary, AB

1. Introduction/Business Problem

The primary object of this Capstone project is to propose a neighbourhood in Calgary for an individual or a business franchise owner, who is considering opening an Indian Restaurant. This vibrant multi-cultural city is quit diverse in its population attracting people from a wide variety of ethnic backgrounds. Being Canada's primary oil and gas economic engine, Calgary attracts a diverse population of talent including South Asians often working in engineering or information technology. The city is however still relatively small and underdeveloped compared to metropolitan cities like Toronto, Montreal and Vancouver. Nevertheless, Calgary's population is expected to rise as the city matures and diversifies its economy and demographics. In 2011, Immigration Canada, 80% of the population who reported speaking a language other than English, French or Aboriginal language lived in one of Canada's largest census metropolitan areas, and Calgary is one of them.

2. Data & Methodology

In first part of this project, we explored and developed insight on Toronto and its neighborhoods. Toronto postal codes and respective neighborhoods were extracted using Python's Beautifulsoup package on Toronto's Wikipedia page. Neighbourhood names were then passed into Python's geocoder package to obtain location latitude and longitude coordinates. We then explored Toronto's neighborhoods such as Thorncliffe Park in East York, which is predominantly known as the immigrant hub of the greater Toronto area. Neighborhood venues data were then acquired through API connection to well-established location based service, Foursquaures. Similar neighborhoods were clustered using K-means clustering algorithm.

To apply machine-learning model effectively, we took mean frequency of occurrence for each venue category for their respective neighborhoods. We then applied a regressor that will handle data with over 200 features in Toronto's data set. We trained and fitted support vector machines (SVM) regressor on the mean frequency dataset to help us predict neighborhoods with high frequency occurrence for an Indian Restaurant in any city.

The second part of this project was to effectively apply the model to Calgary. We again extract neighborhoods using Calgary's Wikipedia page through Python's beautifulsoup package. Unlike the Toronto dataset, the location data for each neighborhood was directly extracted from Wikipedia and stored as csv file for further analysis.

3. Results & Discussion:

Results from first part of the project which focused on Toronto, identified two neighbourhoods where Indian Restaurant was the most common venue. These neighbourhoods correlated directly with high density of immigrants in these areas. When comparing their venues, we noted other Asian restaurants as top five venues. The clustering confirmed that these areas are similar as they fell under the same cluster. The clustering however did not produce

anything more meaningful, which could be effectively applied to Calgary neighbourhoods, which is where the SVM regressor performed as expected.

Modeling

In order for the SVM regressor to apply effectively, we ensured both Toronto and Calgary data frames contain the same model features (or venue types) on which the original model was trained. Therefore, venues that were not common between the two cities were dropped off from the Calgary data set.

The SVM regressor was successfully trained and fitted on Toronto dataset. The SVM regressor parameters were tuned using Sckhit-learn's GridSearchCV optimization package. The resulting model mean square error was only 0.009.

The machine-learning model from Toronto was then applied to Calgary data set to determine possible neighbourhoods where an Indian Restaurant is likely to succeed. Five neighbourhoods were recommended (Southview, Shaganappi, Shawnessy, Aspen Woods and Coach Hill) with Southview being the most recommended for opening an Indian Restaurant. Further review of the recommended neighbourhoods in Calgary revealed that there is presence of Asian restaurants in the area. This is the same observation we made on the Toronto neighbourhoods, where top 5 venues in neighbourhood with high Indian restaurants density also had variety of other Asian Restaurants.

4. CONCLUSION

The primary objective of this project was to help business owners and investors determine possible neighbourhoods in Calgary to open an Indian. The modeling aspect of this project could be further enhanced as reasonable amount of venues on which the model was trained in the Toronto dataset where not found in Calgary's foursquare data. This may change as city of Calgary grows economically and demographically, attracting a more diverse group of ethnicities in years ahead. Last but not least, as with any data science project in real life, data modeling results need to be combined with other factors before taking any actions. The neighbourhoods recommended, will therefore require further analysis before the potential investor make their final decision. Factors like overall Alberta Economy, performance of existing Asian venues in the neighbourhoods, and projected immigrant population will be few to consider. For now we recommend Southview in Calgary as the top neighbourhood for such consideration.