IBM Capstone Project

An analysis of optimal retail store location based in Waterfront station, Vancouver

Last Update: 6th May, 2020
Author: Hua-Wei Huang

Table of Contents

1. Introduction

Finding an optimal location for a new retail store has been critical in the success of a business. A well-selected retail location can not only increase the revenue of a company but also increase its brand awareness to the public. The determinants of a popular retail store vary across different types of business. Customers react differently to an Italian restaurant and a supermarket. With the growth of location-based software and online platforms, detailed data recording customers' preferences and behaviours has become available. As a result, this report focuses on a single type of retail store, coffee shops, and apply various machine learning models in predicting the popularity of the retail store.

1.1 Problem description

This project aims at, given a set of potential locations, predicting the customer rating of a coffee shop in downtown Vancouver, Canada. A customer's preference can be associated with various external factors such as the distance to nearby subway stations, price of products, income level, branding, and coffee shop density of the area, etc. Customer rating would be considered as an indication of the popularity of a store. Furthermore, we assume an investor is interesting in opening a high rating store to increase his/her brand exposure. These variables will be considered when approaching this problem.

2. Data description

The data is extracted from Google API, Foursquare API, Statistic Canada, and censusmapper.ca. The data contains information of a coffee shop which are the location, price tier, type of coffee shops (chain or individual), number of coffee shops in the designated area, customer rating, distance to subway stations in the area, population, average household income and population density of the dissemination area (DA).

The distance to subway stations is calculated after retrieving the latitude and longitude through Google API using Python. Foursquare contains information related to venues. The type of coffee shops is categorised as

1. Waves Coffee House
2. Starbucks
3. Tim Hortons
4. Blenz Coffee
5. Other

These datasets are merged according to the postal code of each coffee shop. Information retrieved through Google API and Foursquare API contains latitude and longitude. Information from censusmapper.ca only contains ID of each DA which data from Statistic Canada contains postal code, ID of DAs, latitude and longitude. Therefore, dataset from Statistic Canada is used to merge all the variables from the sources.

176 coffee shops are retrieved from Foursquare which 48 of them are removed due to missing data for customer rating.

3. Methodology

Our approaches follow the steps

1.	Define neighbourhoods in the targeted area, 1.5km within Waterfront Station, see figure 1.
2.	Retrieve venue information from Foursquare
3.	Explore coffee shops through clustering
4.	Implementing machine learning techniques to predict customer rating
5.	Perform prediction on the selected potential locations based on the model with the lowest Root Mean Square Prediction Error (RMSE)

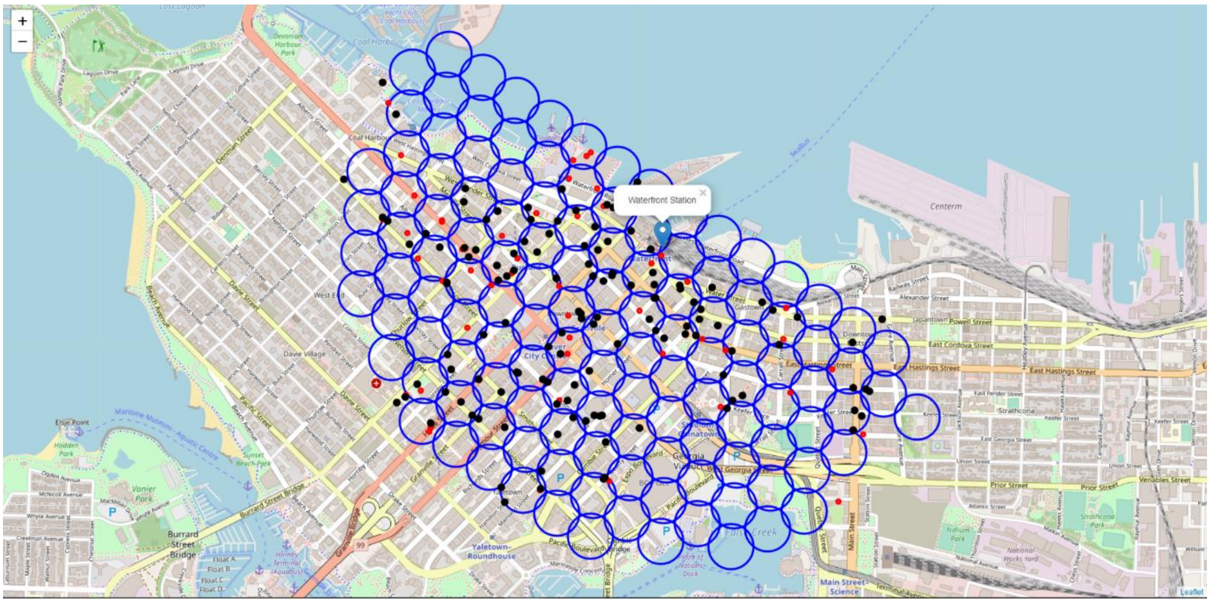

Figure 1. Defined neighbourhoods and retrieved coffee shops

The red dots are coffee shops without customer rating. Each blue circle is the self-defined neighbourhood.

4. Analysis

4.1 Exploratory analysis

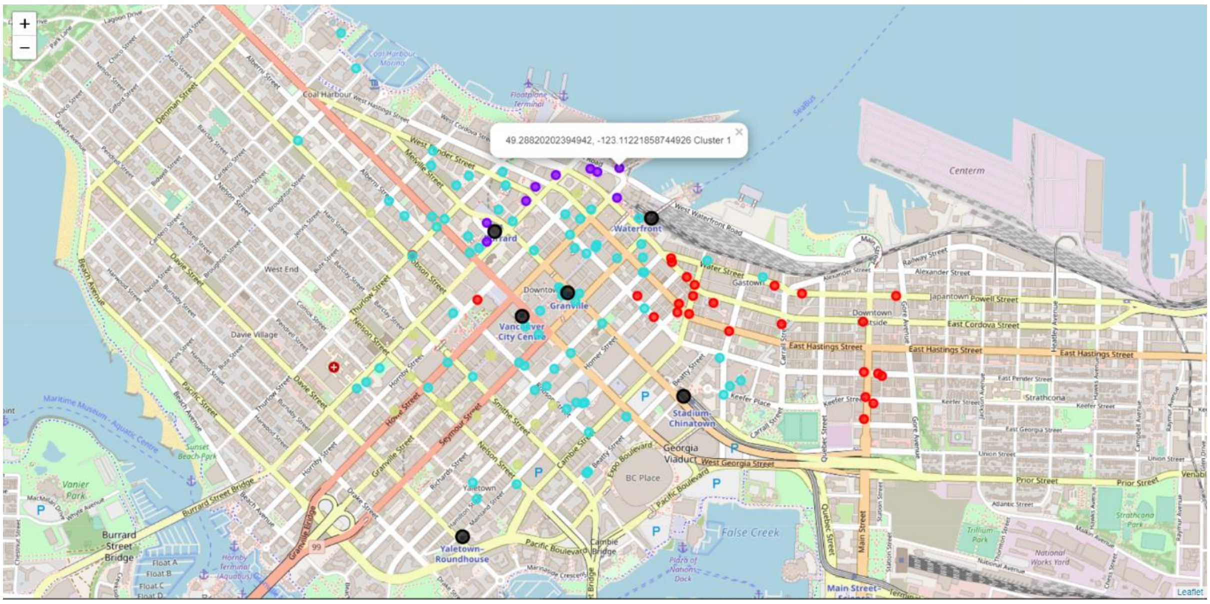

Figure 2. Clustering of all coffee shops

Figure 2 is generated using K-means clustering methods for coffee shops while the black dots are subway stations within the area. As seen from figure 2, there are three major areas where coffee shops have similar characteristics as suggested by K-means clustering. Coffee shops in the red cluster locate in the east of waterfront station. Those in the purple cluster are near the coast and extend along on one of the streets. Coffee shops in the bright green/blue cluster locate at the centre.

Through clustering, we observe that coffee shops in certain regions may have similar characteristics and hence attract different groups of customers. These differences are the potential factors that may influence customer rating when opening a coffee shop.

Section 4.1 provides a brief overview of possible correlations between the characteristics and location of coffee shops. We will apply six different models in estimating the customer rating on the coffee shop data to bring insights in their relationships.

4.2 Machine learning models

Model 4.2.1 Linear Regression

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon_i \tag{1}$$

Our dependent variable, Y, measures customer rating. i indicates index of the number of variables. X_i indicates independent variables. ε_i is the error term. In ordinary least squares (OLS), the parameter of equation (2) are typically estimated using $\beta = (X'X)^{-1}(X'Y)$ (Bajari *et al.*, 2015).

Model 4.2.2 Stepwise Regression

A forward stagewise regression is considered by Bajari *et al.* (2015) with the following algorithm:

1. Find the predictor with the highest correlation with the error term
2. Update the coefficient with the covariance of the variable from the first step
3. Repeat step 1 and 2 until no predictor has any correlation with the error term

Model 4.2.3 LASSO

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda(t - \sum_{j=1}^{p}|\beta_j|) \tag{2}$$

where t is the tuning parameter governing how strictly additional regressors are penalised (Bajari *et al.*, 2015). LASSO is similar to the ridge regression but overcomes the disadvantage that the coefficients are never exactly shrunk to zero. LASSO typically results in some predictors being given zero weights.

Model 4.2.4 Bagging, Boosting, Random Forest

Bagging, random forest, and boosting are methods using the idea of trees in decision trees.

Bagging obtains a predictor through resampling and model combination. Bagging creates multiple copies of data sets and fits a separate decision tree to each copy. Then, it averages the prediction of decision tree under each copy. Boosting holds a similar idea as that of bagging; however, instead of multiple copies, boosting uses information from previously grown trees. Random forest is similar to bagging. The difference is that random forest considers smaller number of exploratory variables by introducing randomness into the set of variables when splitting.

We split the dataset into training data and test data where training data contains 70% of the original data. We compute root mean squared prediction error across the training and test dataset, see table 1.

Table 1. Model comparison: prediction error

	Test RMSE	Training RMSE
Linear	0.896	0.703
Stepwise	1.209	1.063
LASSO	0.975	0.745
Bagging	0.883	0.339
Boosting	0.895	0.303
Random forest	0.895	0.303

Table 1 report the RMSE of both test and training datasets. Based on training prediction error, the best two models are random forest and boosting. When looking at the test prediction error, the best model is bagging; however, its value is not much different from boosting and random forest. The main reason of similar values is that these three methods are similar to each other. Nonetheless, we will utilise the method with the lowest test RMSE to predict the selected potential store locations in next section.

4.3 Prediction on potential locations

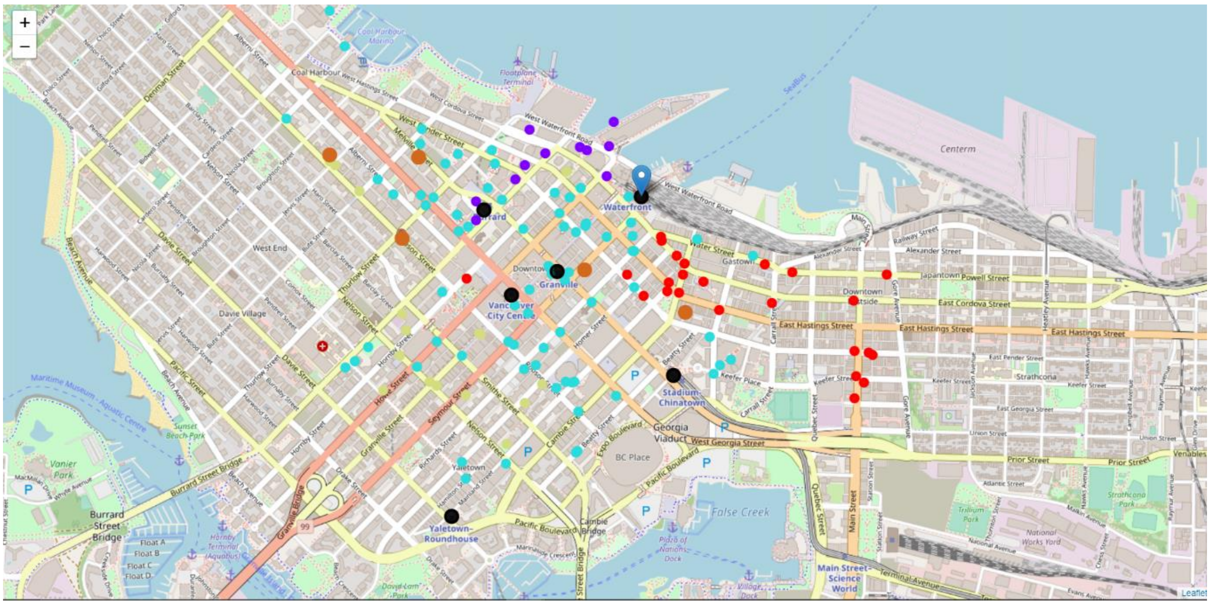

Figure 3. Potential location and clusters, potential location in brown

Table 2. Customer rating prediction on potential locations

	Address	Postal Code	Customer Rating
1	789 Jervis St, Vancouver, BC	V6E 2B1	7.34
2	560 Seymour St, Vancouver, BC	V6B 3H7	6.62
3	1160 Melville St, Vancouver, BC	V6E 2S8	7.07
4	150 W Hastings St, Vancouver, BC	V6B 1R3	7.22
5	1098-1008 Robson St, Vancouver,	V6E 1A7	7.19

Figure 3 displays the clusters of existing coffee shops, subway stations, and selected potential locations. As seen from table 2, the highest predicted customer rating is location one while the lowest is location 2. According to this result, we suggest an investor to open a new coffee shop at location 1 given his choice of these five locations.

5. Conclusion and discussion

In this study, customer rating has been estimated and applied on five potential locations. The highest prediction customer rating by using bagging model is 7.34. Overall, our study provides a brief insight into retail location analysis that could be carried out using the growing online platform, Foursquare. Nonetheless, there are certain limitations in this study. From a business perspective, an investor may want to know the expected profit or revenue he could earn from opening a new retail store. Unfortunately, we cannot retrieve any revenue related datasets. An alternative way of measuring revenue is using user mobility or traffic flow. The time and amount of people staying a an area could potentially transform into spending in stores in that area.

Reference

Patrick Bajari *et al.* (2015) 'Machine Learning Methods for Demand Estimation', The American Economic Review, 105(5), p. 481. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.43821932&site=eds-live (Accessed: 18 March 2020).