DATA FINDS THE BEST SHOPPING MALL

Applied Data Science Capstone Project

How machine learning improves human learning

Ahmed Sabit Faisal

Financial and Business Data Analyst
BEXIMCO Holdings Limited
Dhaka, Bangladesh
Email: asabitfaisal@gmail.com

Data of Submission: August 25, 2019

Table of Contents

Executive Summary

The newly emerged techniques of machine learning are enabling people to identify patterns in data which were not recognized previously. This pattern can be quite insightful in solving age old problems of many industries. The optimization of resources as well as of business processes are some of the usefulness that is being derived using the data which is now available in different formats and from different sources.

This project tries to use this data by applying a specific kind of machine learning methods named as clustering to direct a user (in this case, an explorer in a new city) to the trendy shopping malls as perceived as popular in the city. From a list of shopping malls present in Dhaka, the capital of Bangladesh, one can group these malls in different categories such as below average, average and above average quality. The data analyzed is extracted from the location provider service, Foursquare.

The final outcome of this project portrayed how clustering methods of machine learning discipline bundled the shopping malls in three different clusters and help the user choose best and trendiest shopping malls (also which shopping malls to avoid) of Dhaka city in Bangladesh.

1. Introduction

1.1 Background

Exploring a new city can be challenging at times. If that city is a busy megacity, then the exploring can become a daunting task. It is easy for someone to get lost in the convoluted maze of streets in the city. A city may offer lots of interesting places for a new visitor. But it would not only be exhaustive for the visitor, but also expensive in terms of time and money. So, to choose the most popular spots for the purpose of checking out, it only makes sense if someone mines the data of all places and somehow find the best place to explore. To assist the tourist in this endeavor, Foursquare data can come handy and by analyzing the data in proper manner, a person can easily identify the popular spots (of certain category). This will be just another way where data science comes to rescue by sheding light in myriad data and answers the question posited.

1.2 Problem

This project aims to guide a tourist to select the most trending shopping malls in a predetermined radius (as selected by the user) from the current location of the tourist by clustering the shopping malls of the said region based on the “Likes” as given by other users to those venues. There will be three (3) clusters which will divide the shopping malls in three classes namely – Above Average, Average and Below Average.

1.3 Interest

Anyone who would like to explore a given area for a venue of a specific categories can find this application useful as it will arrange the venues with similar popularities in respective clusters, making it easier for user to understand the quality of the venues. The location provider can also use the features to present the list of venues in an easy-to-understand way for the users.

2. Data Processing

2.1 Data Source and Collection

The location data will be collected using the Foursquare, which is a location service provider. The Places API of Foursquare Developer will be used for sourcing the data. In this case, the test data will be selected from Dhaka, the capital city of Bangladesh. As a particular area of Dhaka, Dhanmondi is chosen. Dhanmondi is a popular zone in Dhaka city for many recreational activities as well as for shopping. So, the data of shopping malls of Dhanmondi will be obtained using the Foursquare places API and will be processed accordingly.

2.2 Data Preparation

Using the API of Foursquare, within the radius of 3000 meters, keeping Dhanmondi as the center, a list of nearby shopping malls is generated. The raw data is converted to a JSON dataframe for better understanding. A total of 30 shopping malls with 18 columns (features) have populated the dataframe.

Then, only columns that include venue name and anything that is associated with location were kept and column names were appropriately edited. A venue ID list is generated for the purpose of extracting Likes from each venue. After getting like count of each venue, a list of likes for each venue is produced.

This list was appended to the main dataframe and then the entire table was sorted in descending order based on likes.

2.3 Feature Selection

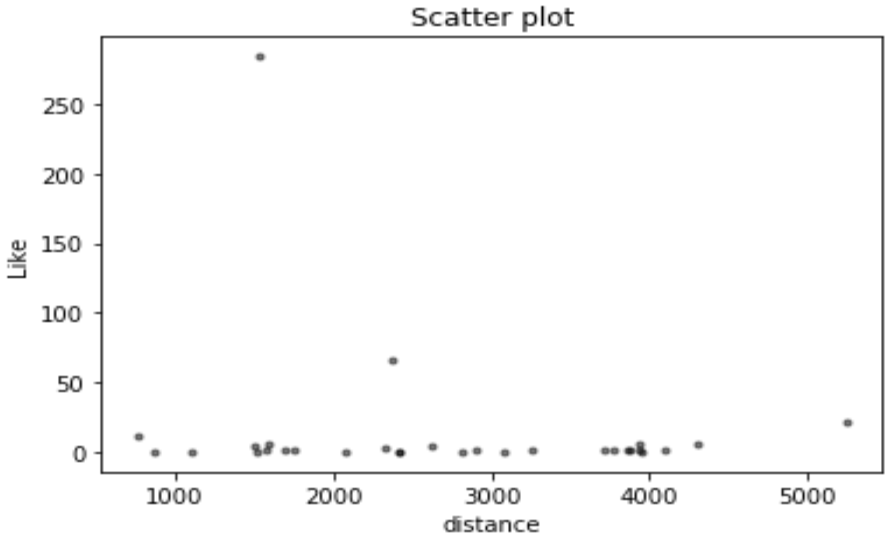
Finally, only venue name, the distance from the tourist and the corresponding likes of the venue are kept as the selected features from the aforementioned dataframe. This is the final dataset which is used for feeding into the machine learning algorithm. All other columns/features are deemed unnecessary at this point for the purpose of this project, so they were dropped.

3. Methodology

To find out the best shopping mall based on the count of Likes, three (3) clustering methods are used which are as follows:

1. K-means clustering
2. Agglomerative hierarchical clustering
3. Density-based spatial clustering of applications with noise (DBSCAN)

These three (3) clustering approach should ensure whether the data is producing same result or not. Before processing with the clustering methods, scatter plot is produced to have a visual on the data distribution.



As it is seen, there are clearly three outliers in the given dataset and rest of the venues almost have the same number of Likes.

3.1 K-means clustering

To apply K-means clustering on the dataset, first we normalize the dataset and then initialize the K-means clustering method where the number of clusters is set at 3 because 3 clusters will represent 3 types of shopping malls which is given below:

Label	Description
0	Below Average
1	Average
2	Above Average

After generating the Labels and assigning the labels to the corresponding each row, the centroid values by averaging the features in each cluster:

Labels	distance	Like
0	3765.714286	3.214286
1	1770.133333	6.466667
2	1530.000000	284.000000

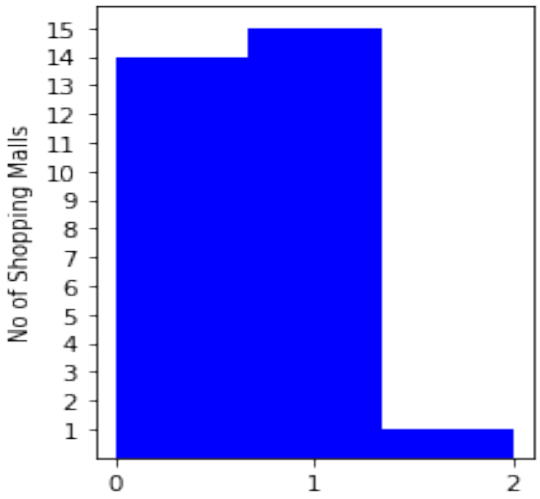
K-means labels came as below:

array ([2, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0]

And the cluster centers are:

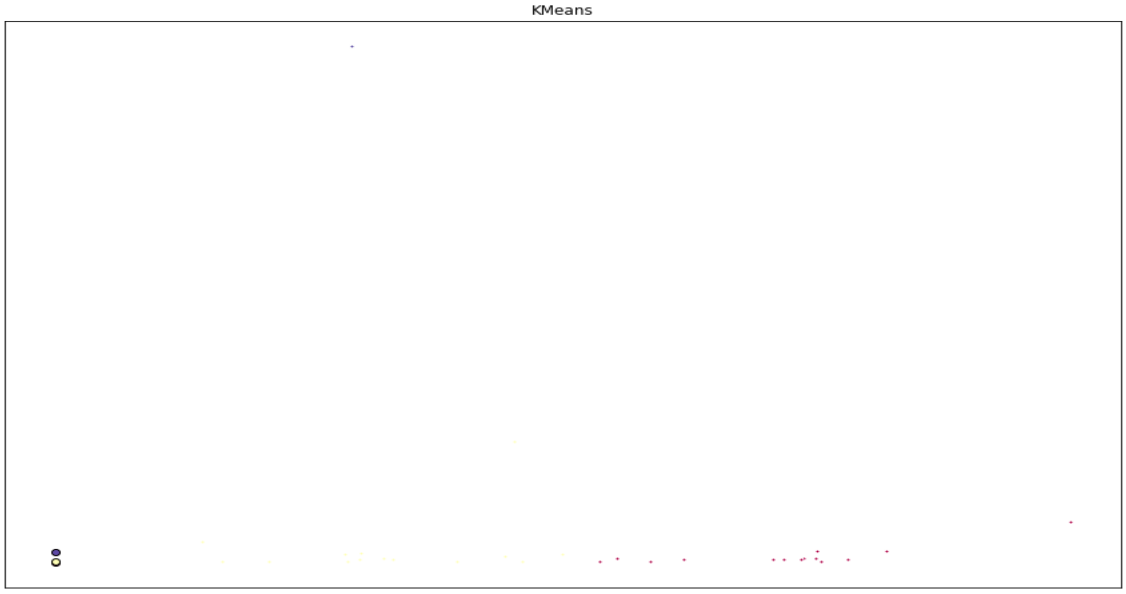
	0	1
0	0.926618	-0.213099
1	-0.797822	-0.150010
2	-1.005328	5.233533

The Histogram generated for the number of shopping malls in every label is shown below:



The Histogram clearly shows Label 2 contains only one shopping mall, Label 1 contains 15 shopping malls and Label 0 contains 14 shopping malls.

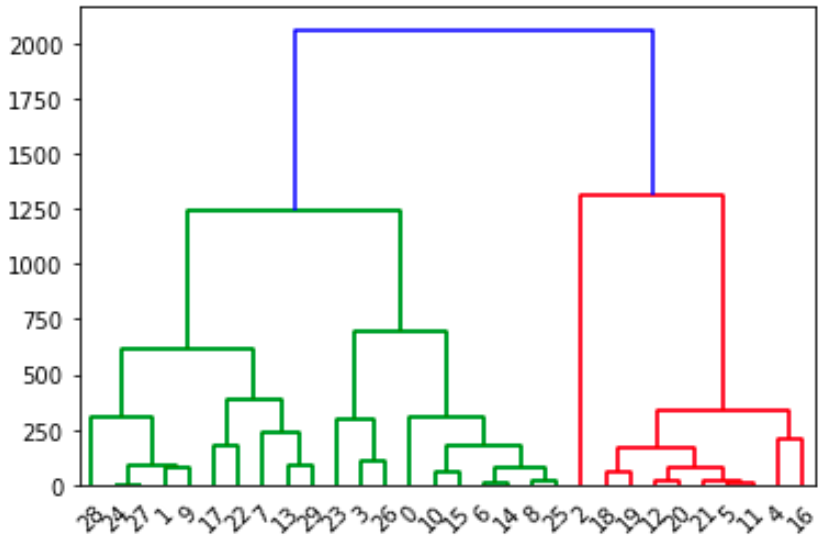
The K-means clusters are plotted below:



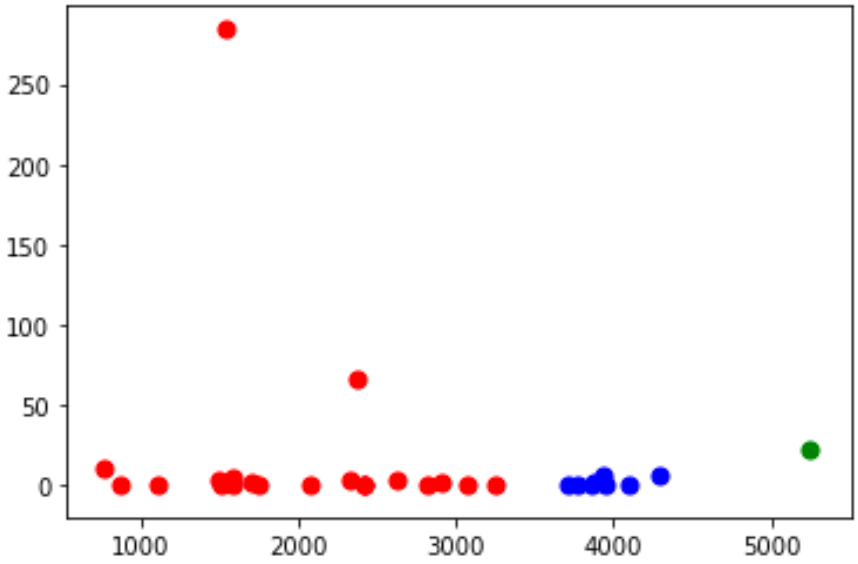
3.2 Agglomerative hierarchical clustering

For Agglomerative hierarchical clustering, the distance and Like data for each shopping mall is taken and fit the model for agglomerative clustering with the Euclidean affinity and average linking method thus produced the labels.

The Dendogram can be found below:

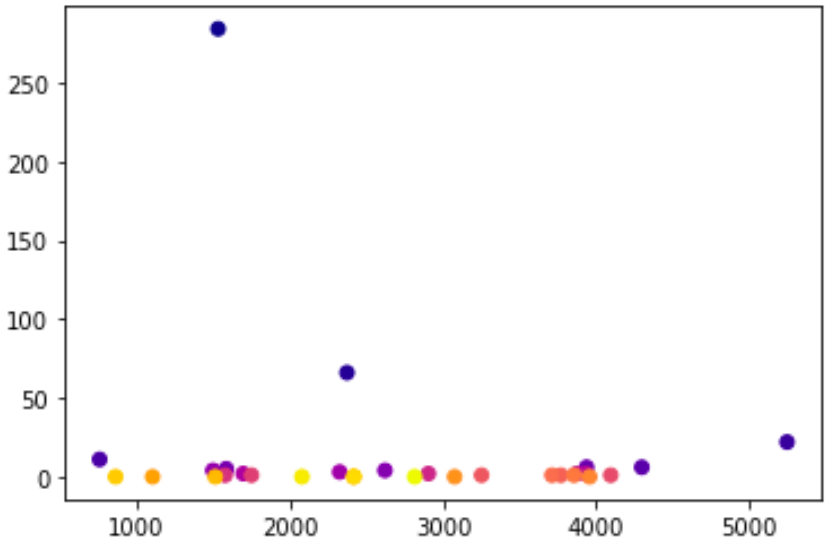


The scatter plot of the labeled data is given below:



3.3 Density-based spatial clustering of applications with noise (DBSCAN)

For DBSCAN approach, the minimum sample requirement set at 1 and Epsilon at 0.123 on the data. The DBSCAN plot is given below:



4. Results

Three clustering machine learning approach is applied to find out the best shopping mall for a tourist. Three approaches brought almost similar results. The results are documented below accordingly:

4.1 K-means clustering

The only above average shopping mall is Bashundhara City Shopping Complex. The others are given below:

Below Average

	name	distance	Like	Labels
27	Twin Tower Shopping Complex	3937	6	0
3	Eastern Plus Shopping Mall	4298	6	0
20	Balaka & Chadni Chawk Shopping Complex	2904	2	0
29	Tropical Razia Shopping Complex	3934	2	0
2	Karnafuly City Garden Shopping Complex	3873	2	0
22	Fortune Shopping Mall	3711	1	0
26	A.R Bhaban. Ayesha Shopping Complex	3855	1	0
25	Police Plaza Concord Shopping Mall	4095	1	0
24	Easy Shopping BD	3251	1	0
23	Century Arcade Shopping Center	3769	1	0
28	Destiny Shopping Center Ltd.	3901	0	0
21	The Grand Plaza Shopping Mall	3075	0	0
19	Eastern Mollika Shopping Complex	2814	0	0
16	Hosaf Shopping Center	3957	0	0

Average

	name	distance	Like	Labels
1	Shimanto Square Shopping Mall	2372	66	1
4	Metro Shopping Mall	756	11	1
10	Aarong, Bashundhara City Shopping Mall	1583	5	1
18	Eastern Mollika Shopping Complex	2620	4	1
7	Gallery Apex, Bashundhara City Shopping Mall	1496	4	1
14	BAFWA Shopping Complex	2324	3	1
11	Bata Mega Store, Bashundhara City Shopping Mall	1694	2	1
12	Agora Shopping mal.hatirpol	1748	1	1
8	A.R.A Center Shopping Mall	1576	1	1
13	Globe shopping center	2078	0	1
5	Globe Shopping Centre	860	0	1
6	lion shopping complex	1102	0	1
17	Priyangon Shopping Mall	2417	0	1
9	American Burger, Bashundhara City Shopping Mall	1511	0	1
15	Noor Mansion Shopping Cente	2415	0	1

Undoubtedly, Bashundhara City Shopping Complex appeared as the best shopping mall of the city, which indeed it is. The Shimamoto Square Shopping Mall, which is also true. K-means clustering puts the Shimanto Square Shopping Mall as an average shopping mall, although the intra-cluster difference with other shopping malls are high.

4.2 Agglomerative hierarchical clustering

Agglomerative hierarchical clustering identified Bashundhara City Shopping Complex as the best shopping malls as well. The difference of Agglomerative hierarchical clustering with K-means clustering is Agglomerative hierarchical clustering correctly identified Shimanto Square Shopping Mall as an outlier. One interesting aspect of this clustering method is it categorized the data based on the distance from the user. So, Agglomerative hierarchical clustering is more decisive compared to K- means clustering.

4.3 DBSCAN

DBSCAN result also yielded almost similar result like Agglomerative hierarchical clustering but with more distinct features. The outliers are the first three best shopping malls of city. These three shopping malls along with the next best shopping mall are:

1. Bashundhara City Shopping Complex
2. Shimanto Square Shopping Mall
3. Pink City Shopping Complex
4. Metro Shopping Mall

DBSCAN method visualized the result based on distance and number of likes.

In all cases, Bashundhara City Shopping Complex and Shimanto Square Shopping Mall emerged as the best shopping malls of Dhaka. The difference between these two clearly indicates that Bashundhara City Shopping Complex is much more popular than Shimanto Square Shopping Mall because Bashundhara City Shopping Complex is much bigger with more shops than Shimanto Square Shopping Mall, which can be further confirmed by external approach. Therefore, the machine learning approach has picked the best shopping mall from the available data.

5. Discussions

In this particular case, the best shopping mall was easy to find out from the given dataset because of the stark difference in the Likes count of the shopping malls. But the clustering approach can be useful if the dataset contains more shopping malls with closer Like counts with other features as well. Due to sparsity in the data and being a small city with small number of shopping malls, only distance and like counts are used as features and it served the purpose. The machine learning approach for clustering can deal with higher number of features and would be able to reach at solutions demanded by the user.

The Foursquare free account only gives access to the Likes count of the venues. Ratings and the number of users rated the venue (which is accessible through Foursquare premium accounts) can be insightful features to combine with the above features. A rating of 10 or 0 may not correctly reflects the true quality of a venue if it is only rated by a single user. Machine learning can distinguish this fact and provide user better understanding of the scenario by filtering/separating the dubious venues.

6. Conclusion

In conclusion, using the clustering method, best (or at least the better) shopping malls were identified. Different kinds of clustering method were applied, and, in this case, all produced the same result, although this may not hold true for a larger dataset with more features. These models should be a helpful tool for someone who would like to find out a shopping mall worth visiting. The same model can be extended for different kinds of venue too such as restaurant etc. Further development on more varied dataset will be required in the future for more assuring results.

----------xxx----------