Concordia University

Gina Cody School of Engineering and Computer Science

Department of Building, Civil and Environmental Engineering

CIVI 691K: Big Data Analytics for Smart City Infrastructure

Winter 2019

Instructor

MAZDAK NIK-BAKHT

Homework assignment #3
Bixi_station_code- 6184; year- 2016

Submitted by,

Vivekkumar Chatrola- **40059267**

1. **Classifier's Performance-** The performance of the decision tree is evaluated based on the following 5 metrics value (average +/- SD) obtained with 10-fold cross validation.
    - Precision: 74.52% +/- 6.75%
    - Recall: 2.87% +/- 0.67%
    - Selectivity: 99.81% +/- 0.07%
    - Accuracy: 83.77% +/- 0.12%
    - AUC: 0.513 +/- 0.003

kNN model accuracy obtained from performance (classification) operator is given in the table 1:

Table 1 Accuracy of kNN classifiers based on classification performance

k	1	2	3	4	5	6	7	8	9
Accuracy	99.78%	99.78%	97.07%	87.72%	80.9%	74.43%	68.35%	62.41%	57%

- Accuracy of kNN for k=1 based on the 10-fold cross validation is only 14.87%. So, without any doubt, kNN is unacceptable. AUC for the binary classifier DT is only 51.30% which is nearly 50%. So, we cannot accept the decision tree classifier as well.

2. **Clustering-** Table 2 shows the list of DBI for different k-means clustering:

Table 2 Davies-Bouldin Index (DBI) for k-means clustering

k	2	3	4	5	6	7
DBI	0.639	0.420	0.268	0.597	0.621	0.670

In my case, k=4 means clustering is the best cluster since it has the smallest Davies–Bouldin index. Table 3 and Figure 1 to figure 5 shows information related to 4-means clustering:

Table 3 Information related to the clusters of the 4-means clustering

Information	***cluster_0***	***cluster_1***	***cluster_2***	***cluster_3***
Size of clusters	25002	2806	8002	3901
Average within centroid distance	-0.016	-0.017	-0.035	-0.038
Attributes of the centroid				
is_member = True	1	0	1	0
is_member = False	0	1	0	1
end_station_longitude	-73.578427	-73.579484	-73.579378	-73.5794093
end_station_latitude	45.525345	45.523079	45.525850	45.523395
duration_sec	594.30	1124.997	611.875	1041.56
day	0	1	1	0
Most frequent station code	6154	6216	6154	6015
Most frequent station longitude	-73.57544	-73.58503	-73.57544	-73.5613
Distance between my station and the centroid	0.3298 km	0.2974 km	0.2796 km	0.2829 km
Distance between my station and the most frequent station	1.012 km	1.097 km	1.012 km	1.956 km

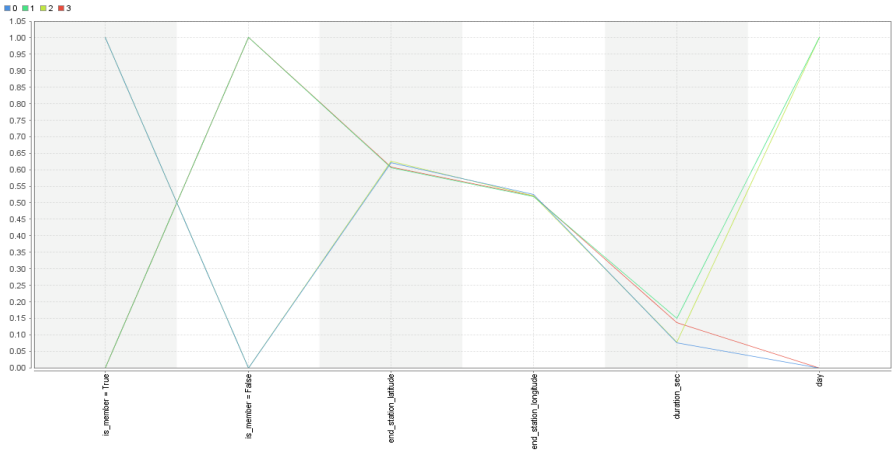

Figure 1 Parallel diagram of the clusters

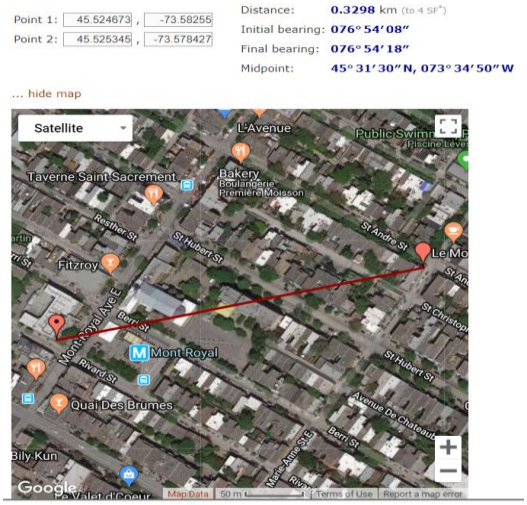

Figure 2 Snapshots of direct path between my station and centroid of cluster_0(acquired from Move Type Scripts Tools)

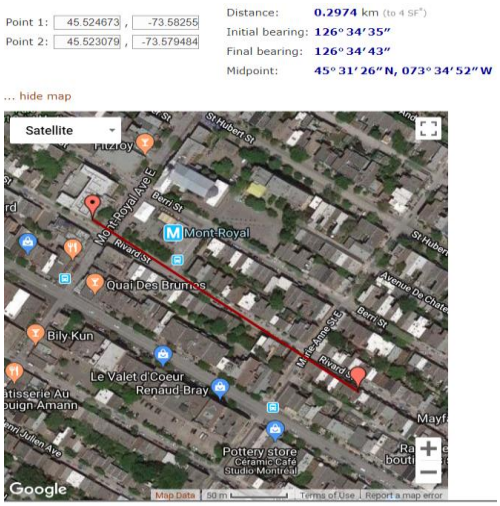

Figure 3 Snapshots of direct path between my station and centroid of cluster_1

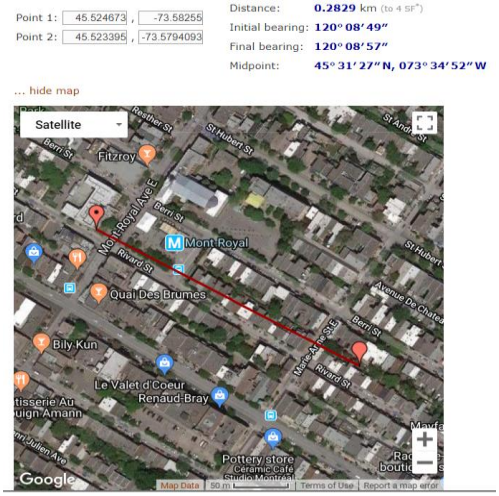

Figure 4 Snapshots of direct path between my station and centroid of cluster_3(acquired from Move Type Scripts Tools)

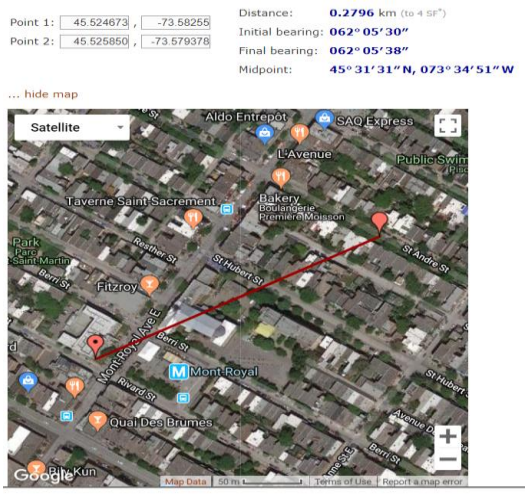

Figure 5 Snapshots of direct path between my station and centroid of cluster_2(acquired from Move Type Scripts Tools)

Following Insights can be obtained from the review of the clusters:

- Direct paths between cluster_0 & cluster_2 and cluster_1 & cluster_3 are same and their distance are almost similar as well.
- Surprisingly, cluster centroid without bixi members tend to travel almost double than centroid with bixi members.
- Users' (member or non-member) average trip length from my station is around 300m.
- Except cluster_3, most frequent stations in all the clusters are at the distance of 1km.
- Members who starts trip from my station are more likely to end trip at station 6154.

3. Regression

I used 'write csv" operator at the end of clustering process to get the cluster id. In excel, I added that attribute to my original dataset of trip started from my station and removed unnecessary attributes as well. I selected duraction_sec attribute although it was not mentioned in the instructions. Then, I labelled duration, mapped days into 0 and 1, transferred nominal to numerical, used split validation with shuffled sampling and keeping local random seed marked with nested process of linear regression using greedy feature selection and regression performance. Regression was time consuming for k-separate models since I had to filter each cluster and modify the process related to it.

Table 4 Linear Regression performance metrics for different models

Model	Equation	R^2	RMSE
Model 1	25.410 * day = 1 + 463.805 * is_member = FALSE - 57.360 * start_weather_type = light rain - 70.894 * start_weather_type = light intensity shower rain - 105.574 * start_weather_type = moderate rain - 64.066 * start_weather_type = mist - 97.813 * start_weather_type = heavy intensity rain + 73.747 * start_weather_type = fog - 77.011 * start_weather_type = proximity shower rain - 4.805 * start_time + 0.677 * start_humidity - 4.458 * start_wind_speed + 2.348 * start_temperature - 0.626 * start_air_pressure + 1239.142	0.099	534.049
Model 2	79.161 * day = 1 - 437.802 * is_member = TRUE - 57.844 * start_weather_type = light rain - 71.986 * start_weather_type = light intensity shower rain - 106.072 * start_weather_type = moderate rain - 64.171 * start_weather_type = mist - 97.794 * start_weather_type = heavy intensity rain + 74.295 * start_weather_type = fog - 76.919 * start_weather_type = proximity shower rain - 68.229 * cluster = cluster_2 - 4.822 * start_time + 0.682 * start_humidity	0.099	534.026

	- 4.382 * start_wind_speed + 2.340 * start_temperature - 0.674 * start_air_pressure + 1729.104		
Model 3 - Cluster 0	61.876 * start_weather_type = few clouds + 55.208 * start_weather_type = sky is clear + 52.113 * start_weather_type = scattered clouds + 43.310 * start_weather_type = broken clouds + 21.263 * start_weather_type = overcast clouds + 120.491 * start_weather_type = fog + 124.379 * start_weather_type = haze + 210.212 * start_weather_type = thunderstorm with light rain - 7.131 * start_time + 0.321 * start_humidity - 4.621 * start_wind_speed + 1.227 * start_temperature - 2.064 * start_air_pressure + 2727.062	0.014	430.938
Model 3 - Cluster 1	- 88.837 * start_weather_type = sky is clear + 130.237 * start_weather_type = broken clouds - 231.412 * start_weather_type = light rain - 262.419 * start_weather_type = moderate rain - 326.272 * start_weather_type = haze + 3.537 * start_humidity - 27.530 * start_wind_speed- 8.414 * start_temperature + 3.926 * start_air_pressure - 2831.262	0.026	952.362
Model 3 - Cluster 2	37.643 * start_weather_type = few clouds + 46.272 * start_weather_type = sky is clear + 79.934 * start_weather_type = scattered clouds + 50.152 * start_weather_type = broken clouds + 53.819 * start_weather_type = overcast clouds + 237.504 * start_weather_type = thunderstorm with light rain - 1.881 * start_time - 3.652 * start_wind_speed + 5.639 * start_temperature + 1.613 * start_air_pressure - 1127.978	0.014	445.961
Model 3 - Cluster 4	253.887 * start_weather_type = few clouds + 221.794 * start_weather_type = sky is clear + 235.477 * start_weather_type = scattered clouds + 157.870 * start_weather_type = broken clouds + 271.456 * start_weather_type = overcast clouds + 124.882 * start_weather_type = light rain + 357.142 * start_weather_type = fog + 337.621 * start_weather_type = haze + 775.003 * start_weather_type = proximity thunderstorm + 2.441 * start_humidity + 15.221 * start_wind_speed - 3.479 * start_air_pressure + 4161.058	0.001	828.466

Table 5 List of 4-star features & top 3 features (in descending order of their significance)

Model	***List of four-star features (****)***	***List of top 3 features (in descending order of their significance)***
Model 1	day = 1; is_member = FALSE; start_weather_type = light rain; start_weather_type = light intensity shower rain; start_weather_type = moderate rain; start_weather_type = mist; start_time; start_humidity; start_wind_speed; start_temperature; (Intercept)	is_member = FALSE start_time start_temperature
Model 2	day = 1; is_member = TRUE; start_weather_type = light rain; start_weather_type = light intensity shower rain; start_weather_type = moderate rain; start_weather_type = mist; cluster = cluster_2; start_time; start humidity; start wind speed; start_temperature; (Intercept)	is_member = TRUE start_time day = 1
Model 3 - Cluster 0	start_weather_type = few clouds; start_weather_type = sky is clear; start_weather_type = scattered clouds; start_weather_type = broken clouds; start_weather_type = haze; start_time; start_wind_speed; start_air_pressure; (Intercept)	start_time (Intercept) start_weather_type = few clouds start_weather_type = sky is clear
Model 3 - Cluster 1	start_humidity	start_humidity start_weather_type = broken clouds start_weather_type = light rain
Model 3 - Cluster 2	start_weather_type = scattered clouds start_temperature	start_temperature start_weather_type = scattered clouds start_weather_type = broken clouds
Model 3 - Cluster 4	start_weather_type = few clouds start_weather_type = sky is clear start_weather_type = scattered clouds start_weather_type = overcast clouds	start_weather_type = few clouds start_weather_type = overcast clouds start_weather_type = sky is clear

- None of the model has R^2 more than 10% which makes them all unacceptable. So, I would not recommend any model to use as a forecasting tool for bixi travel time.