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 #2
Bixi_station_code- 6184; year- 2016

Submitted by,

Vivekkumar Chatrola- **40059267**

1. Association Analysis

Table 1 shows the rules that can be generated by setting minsup of the FP-Growth algorithm at 20% and minconf of the rule generation at 16.50%.

Table 1 Rules generated from association analysis (at 20% minsup and 16.50% minconf)

#	Antecedent	Travel length (conclusion)	Confidence
1	end_humidity: low	Medium	48.10 %
2	end_weather: few clouds	Medium	47.60%
3	end_wind_speed: medium	Medium	47.40%
4	end_humidity: medium & end_temperature: high	Long	33.30%
5	end_temperature: high	Long	33.20%
6	end_wind_speed: low & end_temperature: high	Long	33.10%
7	end_wind_speed: low & end_humidity: high	Short	26.90%
8	end_humidity: high	Short	26.90%
9	end_wind_speed: low & end_temperature: medium	Short	24.60%

Important points:

1. Temperature and Humidity have high influence on travel duration.
2. High temperature is the most important factor for long travels.
3. Trips are shorter in very high humidity and medium in very low humidity.
4. Wind speed is the 3rd important attributes. Although it is very hard to predict anything solely based on the wind speed except for the medium travel (rule 3) because in rule 6 and 9 wind speed is the same but there is a difference in temperature which I think is more important.
5. Weather type doesn't have much impact on short and long travel length.
6. None of the rules have more than 50% of confidence.
7. All the rules are meaningful in association with weather condition.

2. Classification with Decision Tree

Table 2 List of the attributes available in DT model with depth of 5 (in descending order of their weights)

No.	Attribute name	Weight
1	duration_second	0.710
2	end_temperature	0.132
3	start_humidity	0.095
4	start_wind_speed	0.064

Table 2 shows the list of attributes available in DT model in a descending order of their weights based on maximum depth of 5 and information gain ratio. Only 4 classifying attributes are achieved by using information gain ratio. Duration is the most important attribute for the DT model with 71% weightage which clearly overshadows other 3 attributes. It is also the main reason for getting only 4 attributes. This high weight is logical as users with the bixi membership tend to travel longer than non members because of the cheap fares. Other 3 factors with very less weights are not that much important since weather condition has logically no relation with bixi membership. There are no surprises in the list or the order.

(Note: To get top 5 attributes, depth had to be changed to 6. In that case, 5th attribute obtained was end_humidity with different weights for all the attributes than shown in the table 2.)

Decision tree explanation:

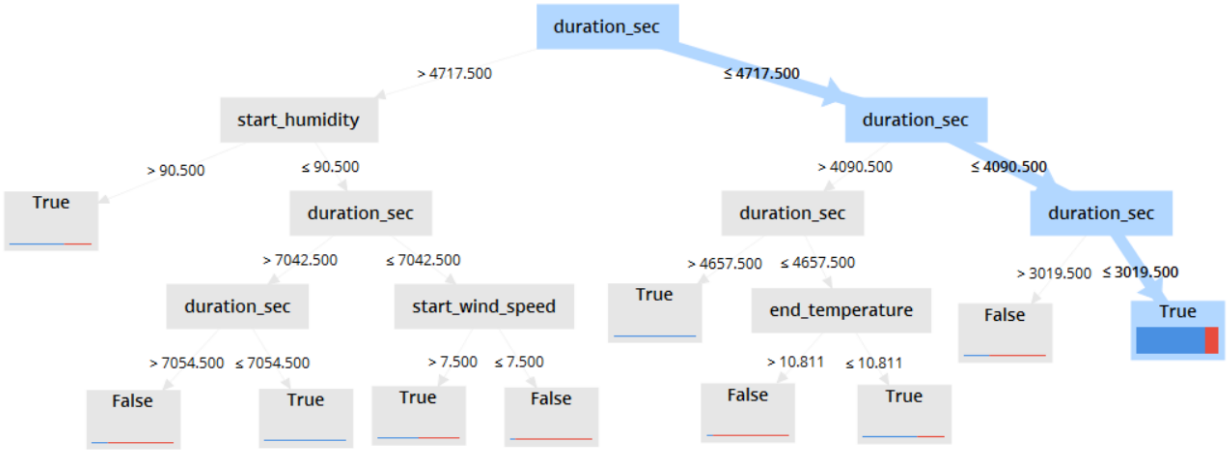

Figure 1 Screenshot of Decision Tree Graph (by using max depth=5 & gain ratio)

1. If duration_sec > 4717.500 & start_humidity > 90.500 then user is a bixi member with a confidence level of = 66.67%
2. If duration_sec > 7054.500 & start_humidity ≤ 90.500 then user is not a bixi member with a confidence level of = 80%
3. If 7042.500 < duration_sec ≤ 7054.500 & start_humidity ≤ 90.500 then user is a bixi member with a confidence level of = 100%
4. If 4717.500 < duration_sec ≤ 7042.500 & start_humidity ≤ 90.500 & start_wind_speed > 7.500 then user is a bixi member with a confidence level of = 50%
5. If 4717.500 < duration_sec ≤ 7042.500 & start_humidity ≤ 90.500 & start_wind_speed ≤ 7.500 then user is not a bixi member with a confidence level of = 93.65%
6. If 4657.500 < duration_sec ≤ 4717.500 then user is a bixi member with a confidence level of = 100%
7. If 4090.500 < duration_sec ≤ 4657.500 & end_temperature > 10.811 then user is not a bixi member with a confidence level of = 92.45%
8. If 4090.500 < duration_sec ≤ 4657.500 & end_temperature ≤ 10.811 then user is a bixi member with a confidence level of = 66.67%
9. If 3019.500 < duration_sec ≤ 4090.500 then user is not a bixi member with a confidence level of = 68.89%
10. If duration_sec≤ 3019.500 then user is a bixi member with a confidence level of =83.80%

DT Accuracy: Total 62772 bixi members and 12447 non-members are there in the training set. But Prediction on the same dataset based on DT model shows only 409 non-members and 74810 members. So, Total 12038 false predictions are provided by our trained DT model. Just to understand, percentage of the true predictions is 1- [12038/ (62772+12447)] = 84%. So, just to get an idea, DT can be considered as 84% accurate.

3. Classification with k-NN

Table 3 Predictions based on k-NN model for the test cases with different values of k

Case #	k = 1		k = 5		k = 9	
	end_station	Confidence	end_station	Confidence	end_station	Confidence
1	6158	100%	6158	20%	6158	11.10%
2	6154	100%	6154	39.90%	6154	22.40%
3	6750	100%	6750	21.30%	6204	21.50%
4	6398	100%	6398	20.50%	6252	22.10%
5	6266	100%	6234	39.70%	6234	22.40%
6	6152	100%	6152	20.10%	6913	22.20%
7	6346	100%	6346	20.50%	6354	22.10%
8	6165	100%	6165	20.50%	6158	22.10%
9	6070	100%	6070	22.20%	6070	11.80%
10	6178	100%	6178	20.10%	6178	11.20%
11	6902	100%	6902	20.50%	6221	22.20%
12	6221	100%	6156	36.60%	6221	23.40%
13	6182	100%	6182	20.80%	6155	22.00%
14	6225	100%	6225	20.30%	6225	11.30%
15	6018	100%	6018	20.40%	6018	11.30%
16	6023	100%	6727	40.20%	6727	22.60%
17	6272	100%	6268	39.10%	6268	22.20%
18	6154	100%	6154	20.00%	6154	33.30%
19	6750	100%	6145	40.10%	6145	22.30%
20	6750	100%	6750	20.20%	6015	22.10%