

BIS2216 Data Mining and Knowledge Discovery Fundamentals

Semester of August 2019

No.	Content to Report																																																																		
1.	1.1. Dataset Name: London Bike Sharing Dataset																																																																		
	1.2 Website Link & Source of Dataset																																																																		
	URL: https://www.kaggle.com/hmavrodiiev/london-bike-sharing-dataset Data Source: Kaggle																																																																		
	1.3. Background of the Dataset																																																																		
	This dataset is the historical data for bike-sharing in London licensed by TFL Open Government Data, containing 17,414 rows of data with 10 variables from January 2015 to January 2017. The data is acquired from 3 different sources; https://cycling.data.tfl.gov.uk/ which shows the cycling bike shares count, freemeteo.com which provides the weather and season data, and https://www.gov.uk/bank-holidays which determines on which days are holidays. The data from cycling data is ground by “Start time”, this represents the count of new bike shares grouped by hour, and the long-duration shares are not taken in the count for this dataset.																																																																		
	1.4. Metadata Information:																																																																		
	<table><tr><th>Attribute</th><th>Description</th><th>Data Type</th><th>Levels</th><th>Value Range</th><th>Role</th></tr><tr><td>timestamp</td><td>Timestamp of bike was shared</td><td>TimeID</td><td>dd/mm/yyyy H:M</td><td>-</td><td>Rejected</td></tr><tr><td>cnt</td><td>Count of new bike shares</td><td>Interval</td><td>Min = 0, Max = 7860</td><td>-</td><td>Rejected</td></tr><tr><td>t1</td><td>Real temperature in Celsius</td><td>Interval</td><td>Min = -1.5, Max = 34</td><td>-</td><td>Input</td></tr><tr><td>t2</td><td>Temperature in Celsius “feels like”</td><td>Interval</td><td>Min = -6, Max = 34</td><td>-</td><td>Rejected</td></tr><tr><td>hum</td><td>Humidity in percentage</td><td>Interval</td><td>Min = 20.5, Max = 100</td><td>-</td><td>Input</td></tr><tr><td>wind_speed</td><td>Wind speed in km/h</td><td>Interval</td><td>Min = 0, Max = 56.5</td><td>-</td><td>Input</td></tr><tr><td>weather_code</td><td>Category of the weather (1 = Clear with fog patches, 2 = Few clouds, 3 = Broken clouds, 4 = Cloudy, 7 = Light rain, 10 = Heavy Rain, 26 = Snowfall)</td><td>Interval</td><td>Min = 1, Max = 26</td><td>-</td><td>Input</td></tr><tr><td>is_holiday</td><td>Holiday or not holiday.</td><td>Interval</td><td>2</td><td>0, 1</td><td>Input</td></tr><tr><td>is_weekend</td><td>Weekend or not weekend</td><td>Interval</td><td>2</td><td>0, 1</td><td>Input</td></tr><tr><td>season</td><td>Category field meteorological seasons (0 = Spring, 1 = Summer, 2 = Fall, 3 = Winter)</td><td>Interval</td><td>Min = 0, Max = 3</td><td>-</td><td>Input</td></tr></table>	Attribute	Description	Data Type	Levels	Value Range	Role	timestamp	Timestamp of bike was shared	TimeID	dd/mm/yyyy H:M	-	Rejected	cnt	Count of new bike shares	Interval	Min = 0, Max = 7860	-	Rejected	t1	Real temperature in Celsius	Interval	Min = -1.5, Max = 34	-	Input	t2	Temperature in Celsius “feels like”	Interval	Min = -6, Max = 34	-	Rejected	hum	Humidity in percentage	Interval	Min = 20.5, Max = 100	-	Input	wind_speed	Wind speed in km/h	Interval	Min = 0, Max = 56.5	-	Input	weather_code	Category of the weather (1 = Clear with fog patches, 2 = Few clouds, 3 = Broken clouds, 4 = Cloudy, 7 = Light rain, 10 = Heavy Rain, 26 = Snowfall)	Interval	Min = 1, Max = 26	-	Input	is_holiday	Holiday or not holiday.	Interval	2	0, 1	Input	is_weekend	Weekend or not weekend	Interval	2	0, 1	Input	season	Category field meteorological seasons (0 = Spring, 1 = Summer, 2 = Fall, 3 = Winter)	Interval	Min = 0, Max = 3	-	Input
	Attribute	Description	Data Type	Levels	Value Range	Role																																																													
	timestamp	Timestamp of bike was shared	TimeID	dd/mm/yyyy H:M	-	Rejected																																																													
	cnt	Count of new bike shares	Interval	Min = 0, Max = 7860	-	Rejected																																																													
t1	Real temperature in Celsius	Interval	Min = -1.5, Max = 34	-	Input																																																														
t2	Temperature in Celsius “feels like”	Interval	Min = -6, Max = 34	-	Rejected																																																														
hum	Humidity in percentage	Interval	Min = 20.5, Max = 100	-	Input																																																														
wind_speed	Wind speed in km/h	Interval	Min = 0, Max = 56.5	-	Input																																																														
weather_code	Category of the weather (1 = Clear with fog patches, 2 = Few clouds, 3 = Broken clouds, 4 = Cloudy, 7 = Light rain, 10 = Heavy Rain, 26 = Snowfall)	Interval	Min = 1, Max = 26	-	Input																																																														
is_holiday	Holiday or not holiday.	Interval	2	0, 1	Input																																																														
is_weekend	Weekend or not weekend	Interval	2	0, 1	Input																																																														
season	Category field meteorological seasons (0 = Spring, 1 = Summer, 2 = Fall, 3 = Winter)	Interval	Min = 0, Max = 3	-	Input																																																														
2.	Problem formulation: To predict the count of future bikes shares by examining the natural environmental factors and government-related factor. By knowing whether these factors will affect the usage of public bike sharing in London, this will provide some insights to organizations to predict changes and further allows for future strategies for public bike shares.																																																																		

3.

Summary Statistics of the Dataset

Figure 1: Class Variable Summary Statistics

Class Variable Summary Statistics

(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	is_holiday	INPUT	2	0	0	97.79	1	2.21
TRAIN	is_weekend	INPUT	2	0	0	71.46	1	28.54
TRAIN	season	INPUT	4	0	0	25.23	1	25.19
TRAIN	weather_code	INPUT	7	0	1	35.32	2	23.17

Figure 2 Interval Variable Summary Statistics

Interval Variable Summary Statistics

(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
hum	INPUT	72.32495	14.31319	17414	0	20.5	74.5	100	-0.57278	-0.25577
t1	INPUT	12.46809	5.571818	17414	0	-1.5	12.5	34	0.203862	-0.26168
wind_speed	INPUT	15.91306	7.89457	17414	0	0	15	56.5	0.669011	0.449236
future_cnt	TARGET	0.28701	0.452379	17414	0	0	0	1	0.941749	-1.11324

4.

Data Preparation

4.1. Rejected Attribute & Rationale

Attribute Rejected	Rationale for rejection
timestamp	This analysis focuses on whether the natural environment and factors will affect future bike shares. Thus, the timestamp variable served no purpose and was rejected.
t2	The t2 variable was rejected because the temperature is based on the feeling of the bike users, unlike t1 as it uses the real recorded temperature of a timestamp.
cnt	Since the analysis is an estimation problem, the current cnt variable will be rejected and a transformed cnt called 'future_cnt' will be used as the target for prediction as seen in figure 2.

4.2 & 4.3. Pre-processing Attributes and Methods Used

Attribute Processed	Rationale for processing	Methods for processing
cnt	This analysis would like to predict future counts of bike shares above 1500. This formula provides a good range of skewness and kurtosis.	The cnt variable was changed to rejected and using the Transform Variable Node to create a new variable called future_cnt to be used as the target, where the formula is cnt > 1500.
weather_code	A nominal data type is used to name or label a series of values, given that there are only 7 different values.	Changing the data type of the variable from interval (default) to nominal when assigning the data levels.
is_holiday	There are only 2 possible outcomes for this variable.	Changing the data type of the variable from interval (default) to binary when assigning the data levels.

is_weekend	There are only 2 possible outcomes for this variable.	Changing the data type of the variable from interval (default) to binary when assigning the data levels.
season	A nominal data type is used to name or label a series of values, given that there are only 4 different values.	Changing the data type of the variable from interval (default) to nominal when assigning the data levels.
hum	In hopes to improve model performance.	Grouping the variable using the Interactive Binning Node into groups of 4.
t1	In hopes to improve model performance.	Grouping the variable using the Interactive Binning Node into groups of 4.
wind_speed	In hopes to improve model performance.	Grouping the variable using the Interactive Binning Node into groups of 4.

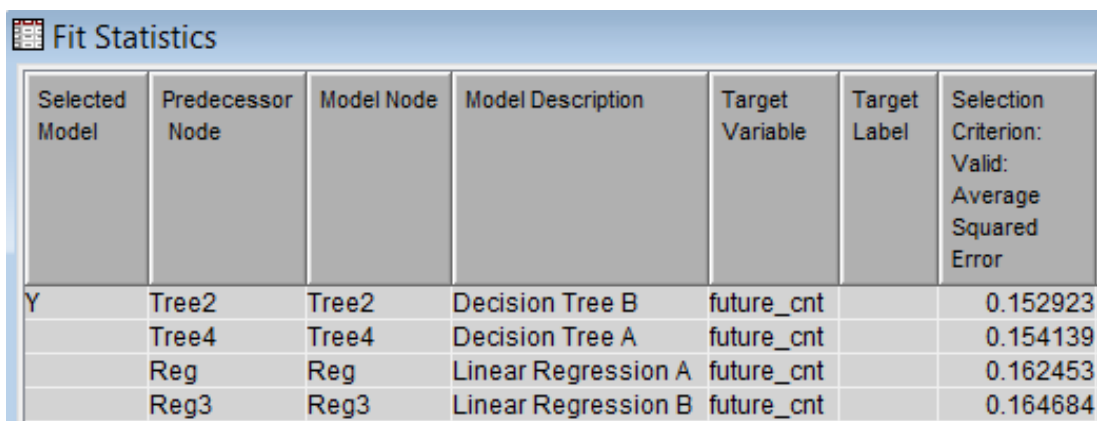
4.4 Unused Methods

Unused Nodes	Rationale
Impute	The dataset does not contain any missing value. Therefore, imputation was not required.
Replacement	The dataset does not contain any outlier values. Therefore, there is nothing to replace with the replacement Node.

5. 5.1. Modeling Methods and Model Performance

No.	Modeling technique and Naming	Partition ratio	Partition Method	Other preparation methods applied	Model Performance
1.	Decision Tree A	Train: 60% Validation: 40%	Default	Connect after the Data Partition Node, the property of Decision Tree A is default.	Valid Average Squared Error: 0.15414
2.	Decision Tree B	Train: 60% Validation: 40%	Default	Connect after the Data Partition Node, the edited properties of Decision Tree B are as follows: <ul style="list-style-type: none"> Maximum Depth = 8 Leaf Size = 6 	Valid Average Squared Error: 0.15292
4.	Linear Regression A	Train: 60% Validation: 40%	Default	Connect after the Data Partition Node, the property of Linear Regression A is default.	Valid Average Squared Error: 0.16245
5.	Linear Regression B	Train: 60% Validation: 40%	Default	Connect after the Interactive Binning Node, the property of Linear Regression B is default.	Valid Average Squared Error: 0.164684

5.2. Screenshot of the Model Comparison Node's Fit Statistics. (Figure 3)



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Y	Tree2	Tree2	Decision Tree B	future_cnt		0.152923
	Tree4	Tree4	Decision Tree A	future_cnt		0.154139
	Reg	Reg	Linear Regression A	future_cnt		0.162453
	Reg3	Reg3	Linear Regression B	future_cnt		0.164684

6. Interpretation of the Best Selected Mode.

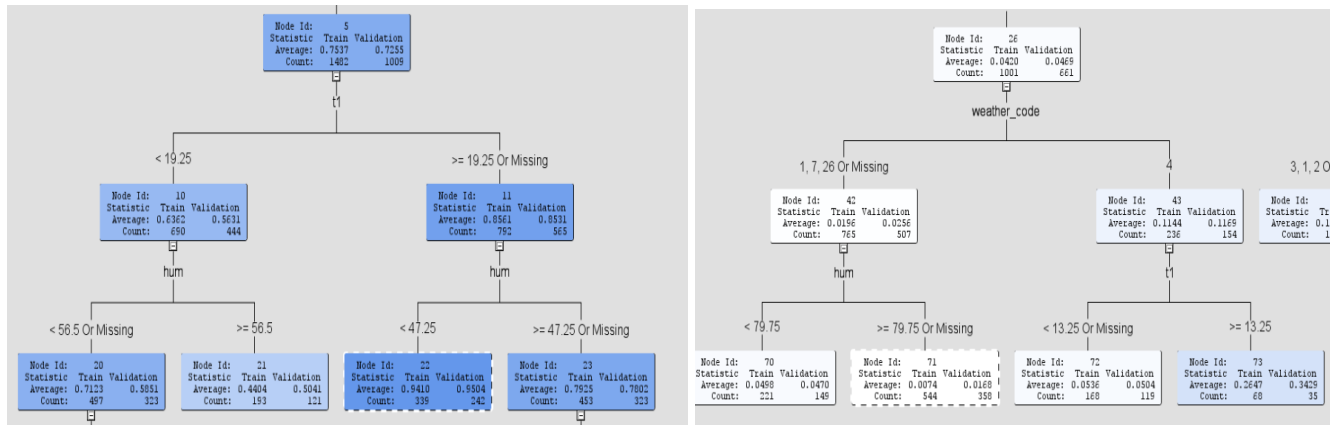


Figure 4: Decision Tree Interpretation

The result of this analysis offer meaningful insights into the counts with more than 1500 future predicted bike shares. Considering the counts of more than 1500 future predicted bike shares, there are several approaches interpretation to take note:

- Temperature and humidity affecting bike shared: A person may first look at the temperature and humidity of the day before they decide that if they want to share a bike or not. For example, Node 22 and Node 23 identify the best performing Nodes. While Node 20 identify reasonably high counts of future bike shares. This shows that people are most likely to go share a bike on a hotter and humid day.
- Weather affecting bike shared: A person may first look at the weather of the day before they decide that if they want to share a bike or not. For example, based on the known demographics of the worst-performing counts of future bike shares, Node 71 represents that fewer counts of bike-sharing was shown. While in Node 73 identified as having better performance but fewer counts were associated with it. This shows that people are more likely to avoid using a bike on a bad weather day.

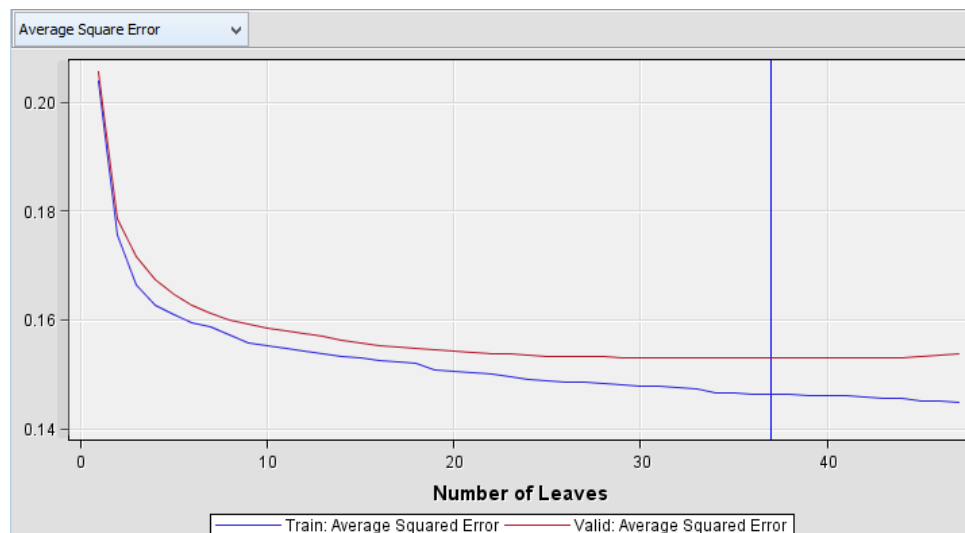


Figure 5: Subtree Assessment Plot for Decision Tree B

In figure 5, the final Nodes show a wide range of percentages for counts having more than 1500 future predicted bike shares and the optimal leaf for the decision tree is 37. Moreover, Node 71 is the lightest background and has the lowest percentage of counts with more than 1500 future predicted bike shares (7%). Nodes 22 is the darkest, with the highest percentages of counts with more than 1500 future predicted bike shares (94%).

```

57  *-----*
58  Node = 22
59  *-----*
60  if t1 >= 19.25 or MISSING
61  AND hum < 47.25
62  then
63  Tree Node Identifier    = 22
64  Number of Observations = 339
65  Predicted: future_cnt = 0.9410029499

```

Figure 6: View Node Rules - Highest Performing Node for Decision Tree B

Node 22 counts of bike shares, with a 94% rate of predicting counts more than 1500, can be described as having a temperature of more than 19.25 Degrees Celsius and humidity of less than 47.25%. It can be assumed that this group of people seems to most likely share a bike when the temperature is above 19.25 degree celsius while humidity is below 47.25%. Using these very specific details we are able to predict the future bike shares to an accuracy of 94%.

```

327  *-----*
328  Node = 71
329  *-----*
330  if weather_code IS ONE OF: 1, 7, 26 or MISSING
331  AND is_weekend IS ONE OF: 1
332  AND hum >= 79.75 or MISSING
333  then
334  Tree Node Identifier    = 71
335  Number of Observations = 544
336  Predicted: future_cnt = 0.0073529412

```

Figure 7: View Node Rules - Lowest Performing Node for Decision Tree B

Node 71 is the lowest-performing group, with the count of future bike shares rate more than 1500 of only 7%. These counts are characterized as having clear with fog patches, light rain, or snowfall weather with humidity more or equal than 79.75% on a weekend. This shows that people are less likely to share a bike during a bad weather day on a weekend.

7. Screenshot of the Modelling Process Diagram.

