

Seoul Bike Sharing





Table of Contents

|01 Problem Statement

|02 Data Selection

|03 Data Preprocessing &
Visualization

|04 Model Selection

|05 Model Limitations



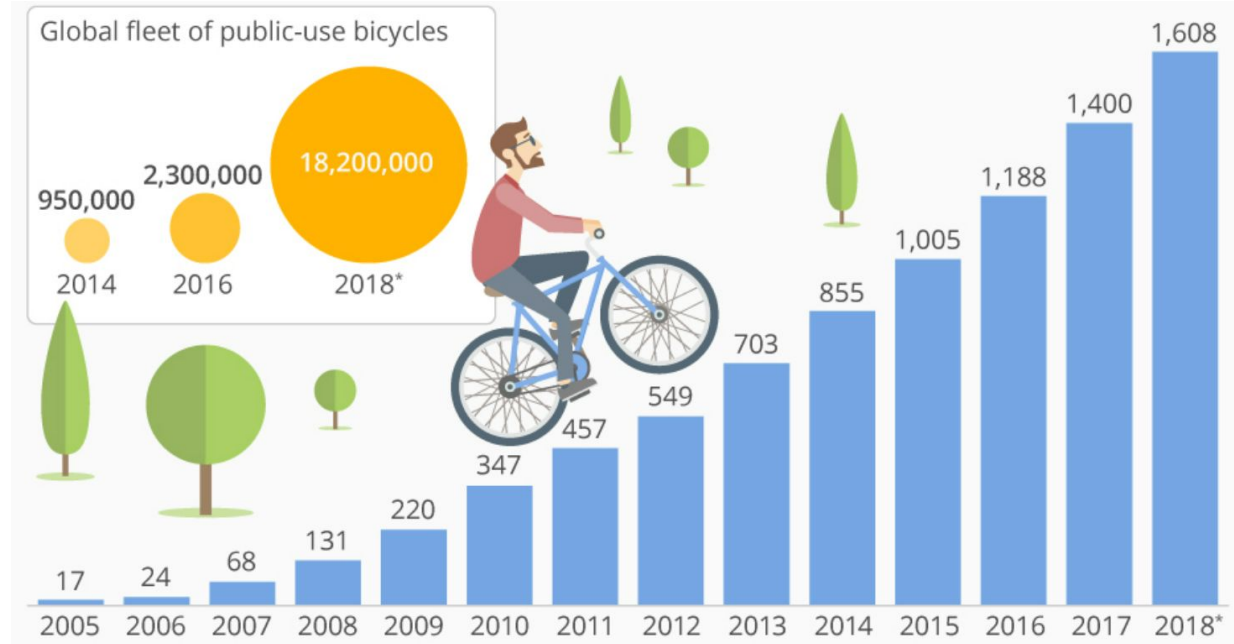
01

Problem Statement

“The rental bike business is thriving worldwide especially after 2012”

Several Potential Reasons:

- Environment friendly
- Healthy Lifestyle
- Less traffic jams
- Convenient on short distance traveling



Goal

Build a model to predict the demand of rental bikes in Seoul (Capital of South Korea) . So that city government and rental bike companies may be able to allocate their bikes more wisely by time.



02

Data Selection

Seoul-Bike Demand Dataset

Variable Name	Explanation
Date	Date in Year-Month-Day format
Rented Bike Count	Demand of bike within certain hour
Hour	Hour dummies of 24 hours
Temperature(°C)	Temperature in Degree Celsius
Humidity(%)	Humidity
Wind speed (m/s)	The speed of wind.
Visibility (10m)	Visibility
Dew point temperature(°C)	Temperature air needed to saturate and condense into liquid water
Solar Radiation (MJ/m2)	Amount of solar radiation
Rainfall(mm)	Amounts of rainfall in millimeters
Snowfall (cm)	Amounts of snowfall in centimeters
Seasons	Season dummies of 4 seasons
Holiday	Dummies variable whether certain date is a holiday or not
Functioning Day	Dummies variable if the software is functioning on certain day

- 14 attributes
- 8760 instances
- No missing values
- 445 outliers

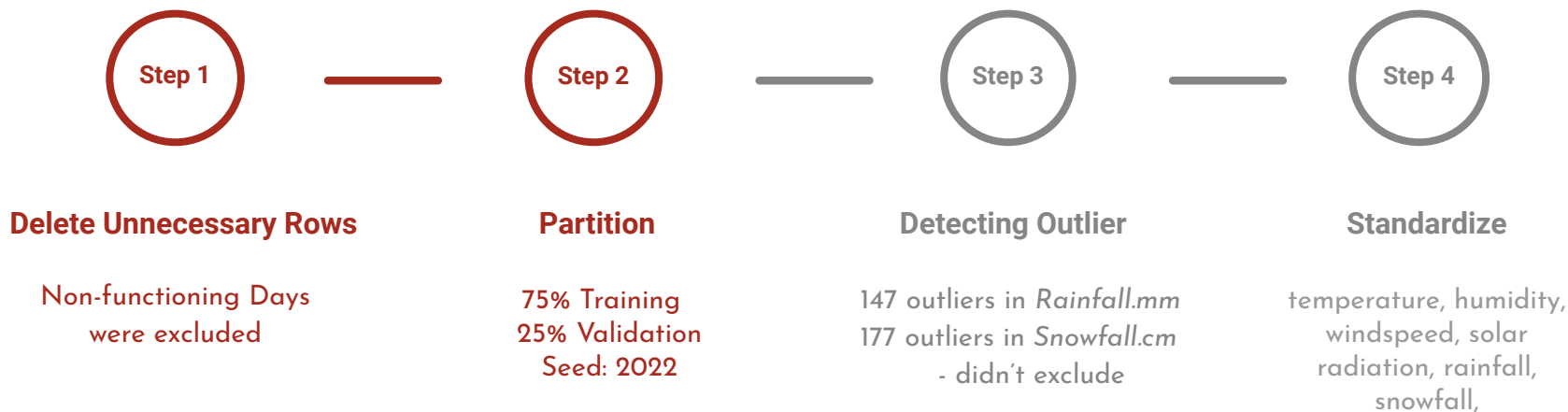
Column	Lower Prob	Upper Prob	Lower Quantile	Upper Quantile	Low Threshold	High Threshold	Number of Outliers
Date	0.1	0.9	3.6e+9	3.62e+9	3.52e+9	3.7e+9	0
Rented Bike Count	0.1	0.9	64	1671.9	-4759.7	6495.6	0
Hour	0.1	0.9	2	21	-55	78	0
Temperature(°C)	0.1	0.9	-3.7	28	-98.8	123.1	0
Humidity(%)	0.1	0.9	32	86	-130	248	0
Wind speed (m/s)	0.1	0.9	0.6	3.2	-7.2	11	0
Visibility (10m)	0.1	0.9	436.1	2000	-4255.6	6691.7	0
Dew point temperature(°C)	0.1	0.9	-15.3	21	-124.2	129.9	0
Solar Radiation (MJ/m2)	0.1	0.9	0	2.059	-6.177	8.236	0
Rainfall(mm)	0.1	0.95	0	0.4	-1.2	1.6	184
Snowfall (cm)	0.1	0.95	0	0.2	-0.6	0.8	261

03

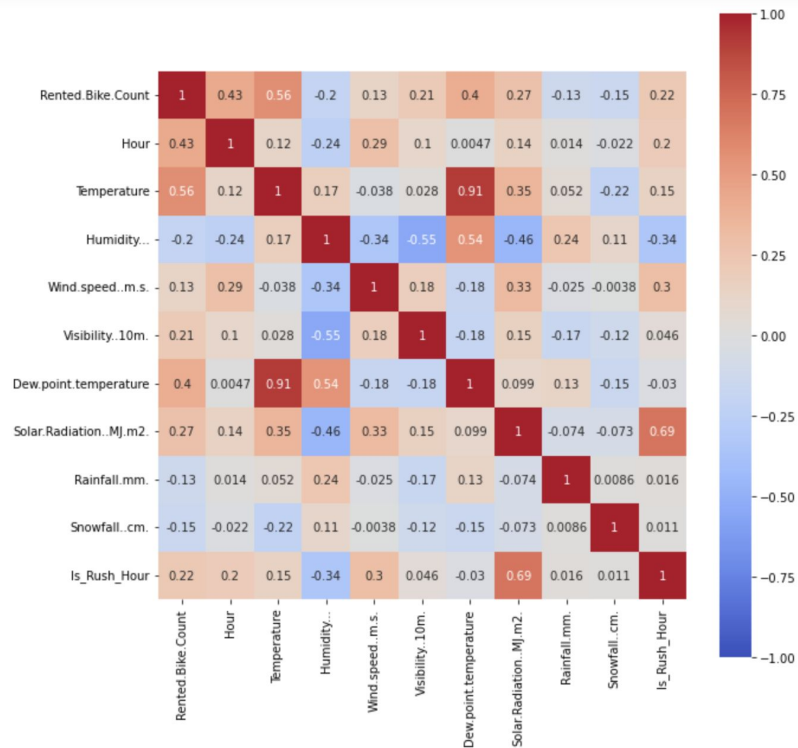
Data Pre-processing & Visualization



Data Pre-processing Procedure



Correlation Matrix



Shows the correlation coefficients between several variables related to rented bike count:

- **Finding 1:** the correlation between temperature and the dew point temperature is 0.91 (strongly positively correlated)
- **Finding 2:** the correlation between solar radiation and rush hour is 0.69 (relatively positively correlated)

*note: since there is a hyper correlation between columns temperature(°C) and Dew point temperature(°C) so we can drop the column Dew point temperature(°C).

Inspect Multicollinearity

❖ Finding:

VIF on Dew.point.temperature=10.8658

	GVIF	Df	$GVIF^{1/(2*Df)}$
Hour	1.206624	1	1.098464
Temperature	90.014794	1	9.487613
Humidity...	20.351053	1	4.511214
wind.speed..m.s.	1.299927	1	1.140143
visibility..10m.	1.702329	1	1.304733
Dew.point.temperature	118.066770	1	10.865853
Solar.Radiation..MJ.m2.	2.023486	1	1.422493
Rainfall.mm.	1.085400	1	1.041825
snowfall..cm.	1.120201	1	1.058396
Seasons	5.235148	3	1.317715
Holiday	1.023907	1	1.011883

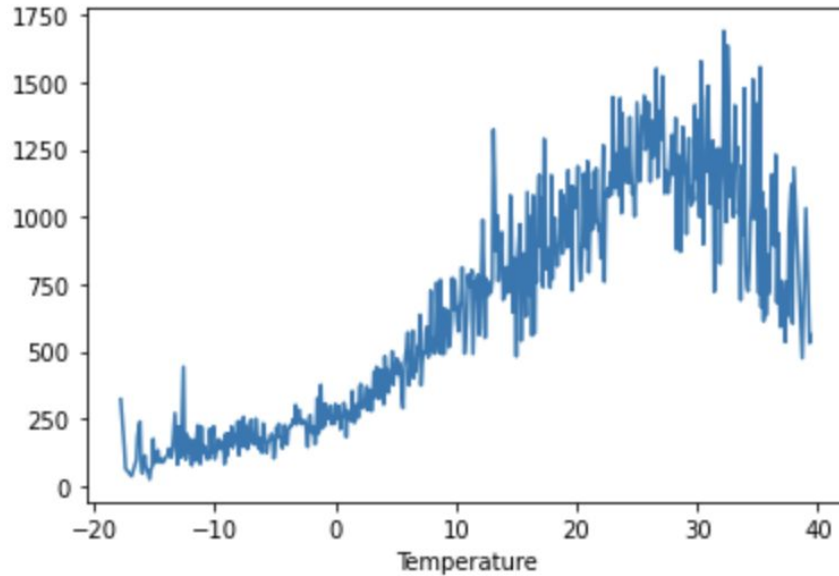
Correlation Matrix Cont'd

Right: list of sorted correlation of variables to the rented bike count from largest to smallest

out[26]:

	Correlation to the target
Rented.Bike.Count	1.000000
Temperature	0.562740
Hour	0.425256
Solar.Radiation..MJ.m2.	0.273862
Is_Rush_Hour	0.219067
Visibility..10m.	0.212323
Wind.speed..m.s.	0.125022
Rainfall.mm.	-0.128626
Snowfall..cm.	-0.151611
Humidity...	-0.201973

Visualization

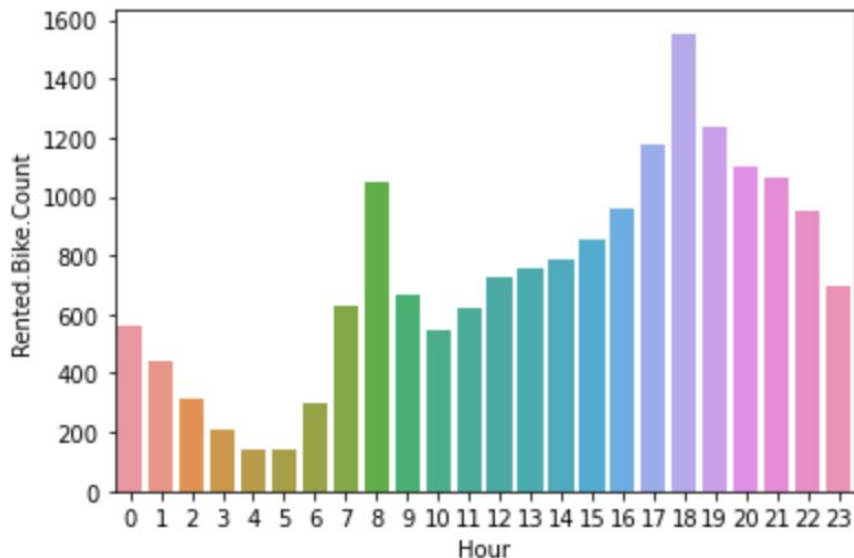


Rented Bike Count v.s. Temperature

- There is an increasing trend between temperature and rented bike count, which reaches peak at around 30°C .

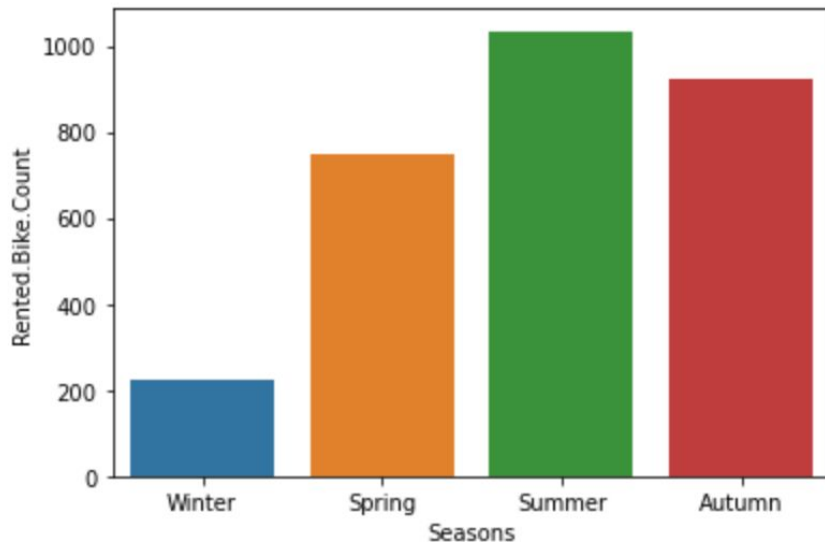
Visualization Cont'd_(python)

Rented Bike Count v.s. Hour



- peak is at 8.am and 6p.m., which makes good sense since there are the rush hours

Rented Bike Count v.s. Seasons



- Citizens in Seoul rented bike more frequently in Summer and Autumn, which matches the weather condition of the city (Summer: 60.8°F ~ 82.4°F / 16° C ~ 28°C)



04

Model Selection

Linear Regression

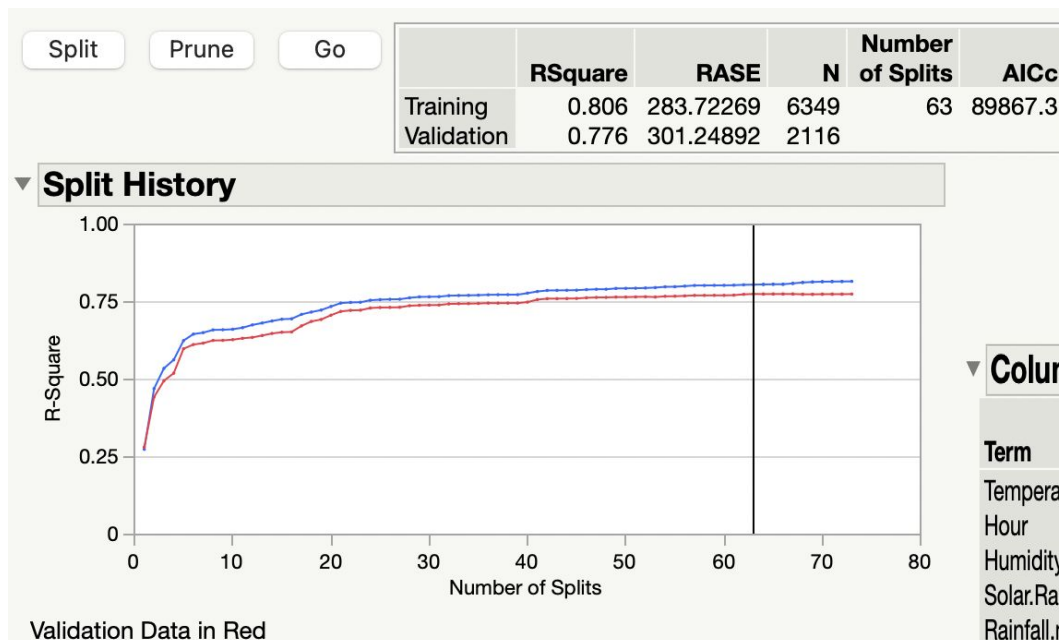
- **X Variable:** hour, temperature, humidity, windspeed, solar radiation, rainfall, holiday
- **Y variable:** rented bike count (numerical)
- **Stepwise regression:** Forward & backward using minimum BICs
- ❖ **Finding:**
 - R-Square on Validation: 0.50
 - RASE on Validation: 448.14

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	424.66605	28.11818	15.10	<.0001*
Hour	29.375405	0.856059	34.31	<.0001*
Temperature	32.223208	0.549487	58.64	<.0001*
Humidity...	-7.836527	0.35514	-22.07	<.0001*
Solar.Radiation..MJ.m2.	-79.29898	8.332543	-9.52	<.0001*
Rainfall.mm.	-65.39598	5.274667	-12.40	<.0001*
Holiday[Holiday]	-78.3773	13.42051	-5.84	<.0001*

Effect Tests	

Crossvalidation			
Source	RSquare	RASE	Freq
Training Set	0.5167	447.86	6349
Validation Set	0.5040	448.14	2116

Decision Tree



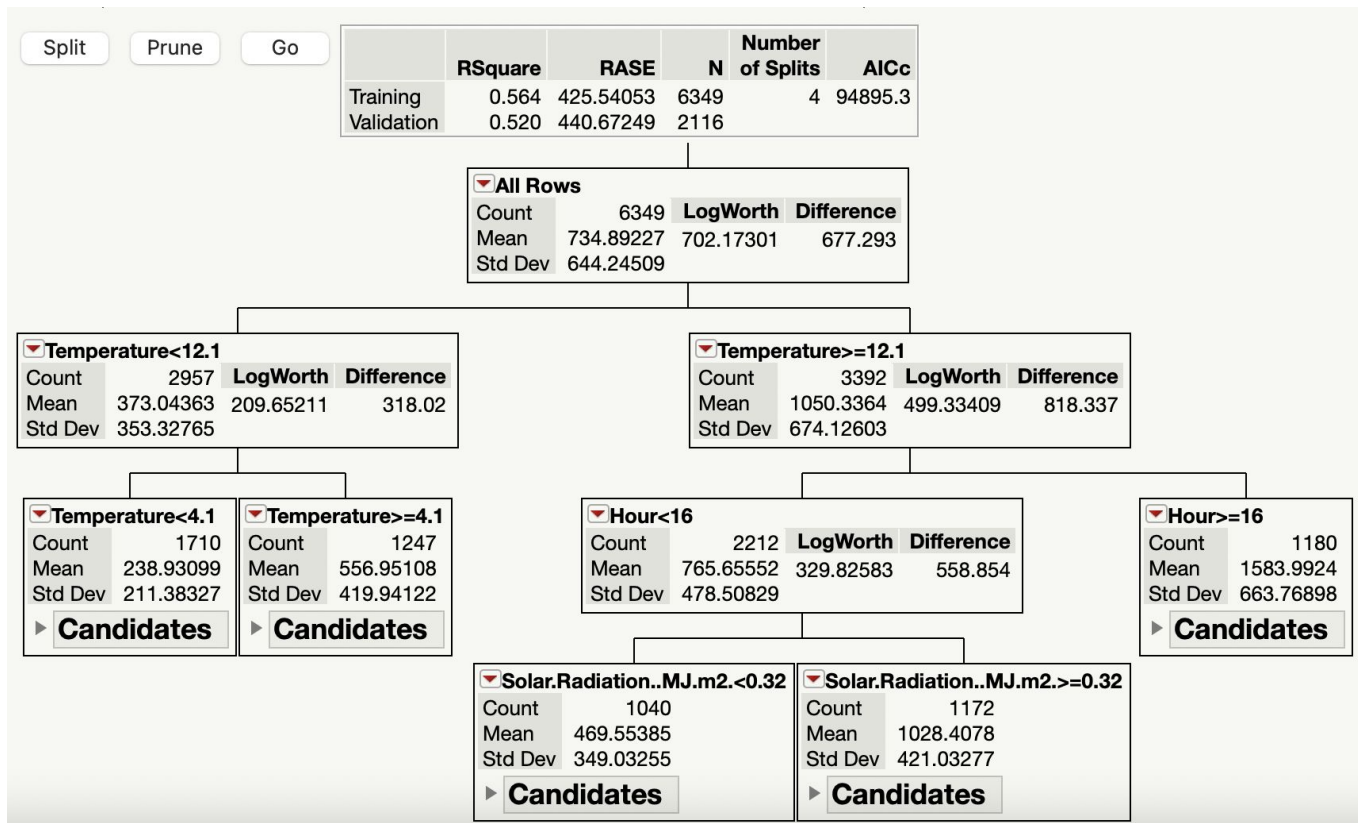
- **Best number of splits: 63**
- **X Variable:** hour, temperature, humidity, windspeed, solar radiation, rainfall, holiday, snowfall (removed season due to multicollinearity)
- **Y variable:** rented bike count (numerical)

▼ Column Contributions

Term	Number of Splits	SS	Portion
Temperature	17	898665393	0.4232
Hour	24	763200389	0.3594
Humidity...	11	222389732	0.1047
Solar.Radiation...MJ.m2.	4	205192655	0.0966
Rainfall.mm.	5	33313368.9	0.0157
Holiday	1	719799.196	0.0003
Wind.speed..m.s.	1	181677.485	0.0001
Snowfall..cm.	0	0	0.0000

- ◆ **Finding:**
- R-Square on Validation: 0.776
 - RASE on validation: 301.25

Decision Tree



Bootstrap Forest

- **X Variable:** hour, temperature, humidity, windspeed, solar radiation, rainfall, holiday, snowfall (removed season due to multicollinearity)
- **Y variable:** rented bike count (numerical)
- ❖ **Finding:**
 - R-Square on Validation: 0.85
 - RASE on validation: 248.86

▼

Specifications

Target

Rented.Bike.Count

Training Rows:

6349

Validation Column:

Validation

Validation Rows:

2116

Number of Trees in the Forest:

1000

Test Rows:

0

Number of Terms Sampled per Split:

3

Number of Terms:

8

Bootstrap Samples:

6349

Minimum Splits per Tree:

10

Minimum Size Split:

8

▼

Overall Statistics

Individual Trees

RASE

In Bag

196.3160

Out of Bag

322.1751

RSquare

RASE

N

Training

0.891

212.59498

6349

Validation

0.847

248.85973

2116

►

Cumulative Validation

►

Per-Tree Summaries

▼

Column Contributions

Term

Number of Splits

SS

Portion

Temperature

22845

571197807

0.4027

Hour

19447

485422725

0.3422

Solar.Radiation..MJ.m2.

11882

131860996

0.0930

Humidity...

18335

127651922

0.0900

Rainfall.mm.

2186

60975701.7

0.0430

Wind.speed..m.s.

15658

26351148.8

0.0186

Snowfall..cm.

1792

12072452.9

0.0085

Holiday

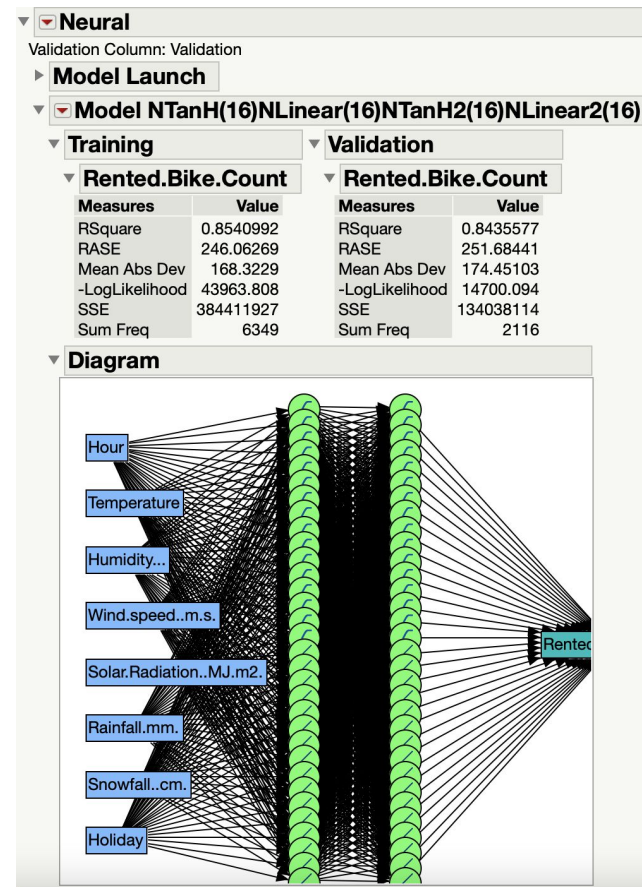
1375

2899634.49

0.0020

Neural Networks

- **X Variable:** hour, temperature, humidity, windspeed, solar radiation, rainfall, holiday, snowfall
- **Y variable:** rented bike count (numerical)
- 2 hidden layers and 32 hidden neurons (2 architectures and 16 each)
- ❖ **Finding:**
 - R-Square on Validation: 0.84
 - RASE on Validation: 251.68



Summary

- We select Bootstrap Forest as our final model(highest R-square,lowest RASE,free from overfitting)
- The 'Temperature' and 'Hour' has significantly higher positive contributions to the variance of demand of rental bike than other variables.
- The major negative contributor is the 'Solar radiation'.
- Seoul citizens use rental bike more frequently when the temperature is higher especially when the rush hours come.

General suggestions based on prediction:

1. Increase the supply of shared bikes during 8 a.m. and 17~19 p.m., especially in summer and autumn.
2. Winter may be a good time to provide bike maintenance.



05

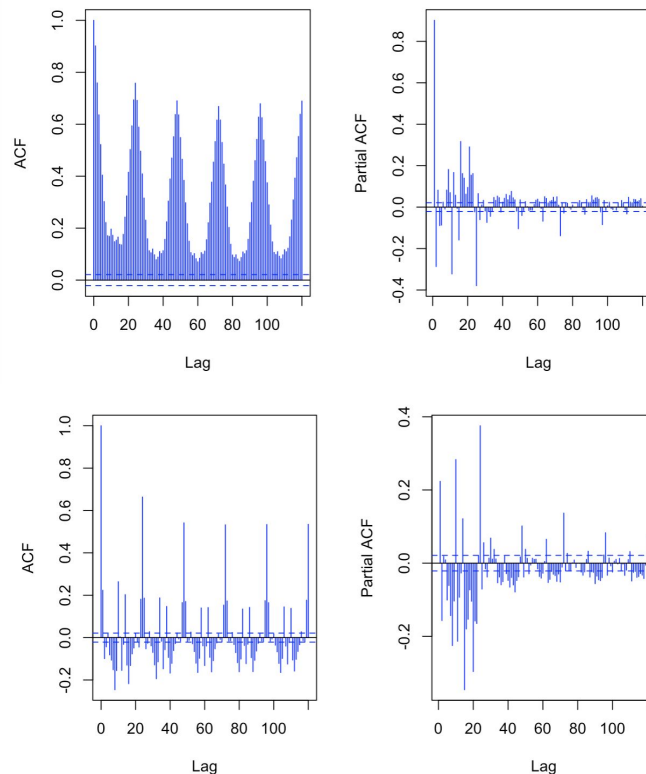
Limitations

Limitation 1

- **Time series dataset:** our data is a time series data. We observe our data by using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).
- **Stationarity:** the ACF and PACF of origin dataset shown above, and that of the first difference of dataset shown below. The plot shows that the dataset is non-stationary and contains both seasonal and non-seasonal AC. A multiplicative SARIMA may be appropriate.

❖ **Decision:**

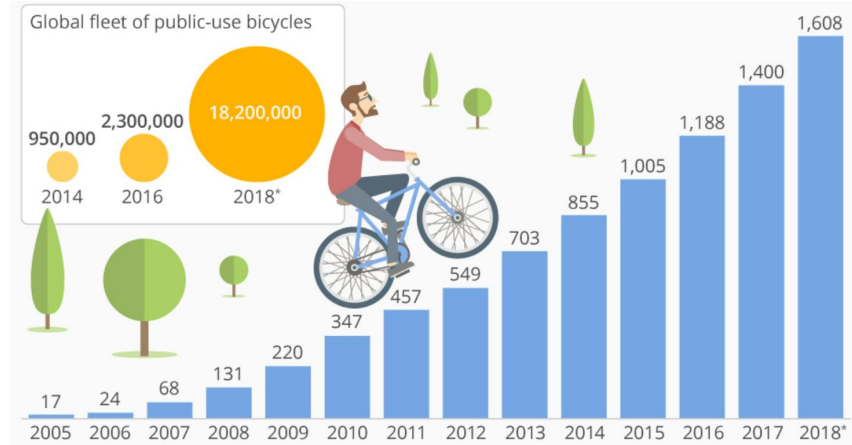
- Given Bootstrap Forest gave us a good fit and the difficulty of using this model, we decide not to fit a multiplicative SARIMA model.



Limitation 2

- **Industry growth:** bike rental is a fast growing industry and our data used to fit model and validate is data from 2017 to 2018, which did not account for the rapid growth of the bike rental industry and the impact of COVID-19.

- ◆ **Finding:**
 - Thus, our model may be not as accurate to predict the real life bike rental demand today in Seoul.





Thank You!

Any questions?

Reference

- (Seoul, South Korea - Average Annual Weather - Holiday Weather (holiday-weather.com))
- Chart: Bike-Sharing Clicks Into Higher Gear | Statista
- [SeoulBikeDemand](#)