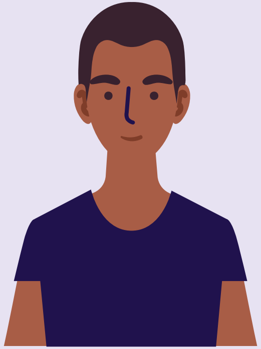
Capital Bikeshare Consulting Report

OUR TEAM



Feng Yi

ZHANG Huiyan

PANZHIYONG

Lai ziwei

TABLE OF CONTENTS

01 Company Introduction

02 Business Understanding

03 Data Understanding

04 Data Preparation

05 Modeling

06 Evaluation

07 Conclusion

08 Business Recommendations

01

Company Introduction

About Capital Bikeshare

5000+ Bikes

600+ Stations

7 Jurisdictions

365 Days a Year

Ref: https://ride.capitalbikeshare.com/about

About Capital Bikeshare

Ref: https://ride.capitalbikeshare.com/about

Subscription Plans

Single Trip

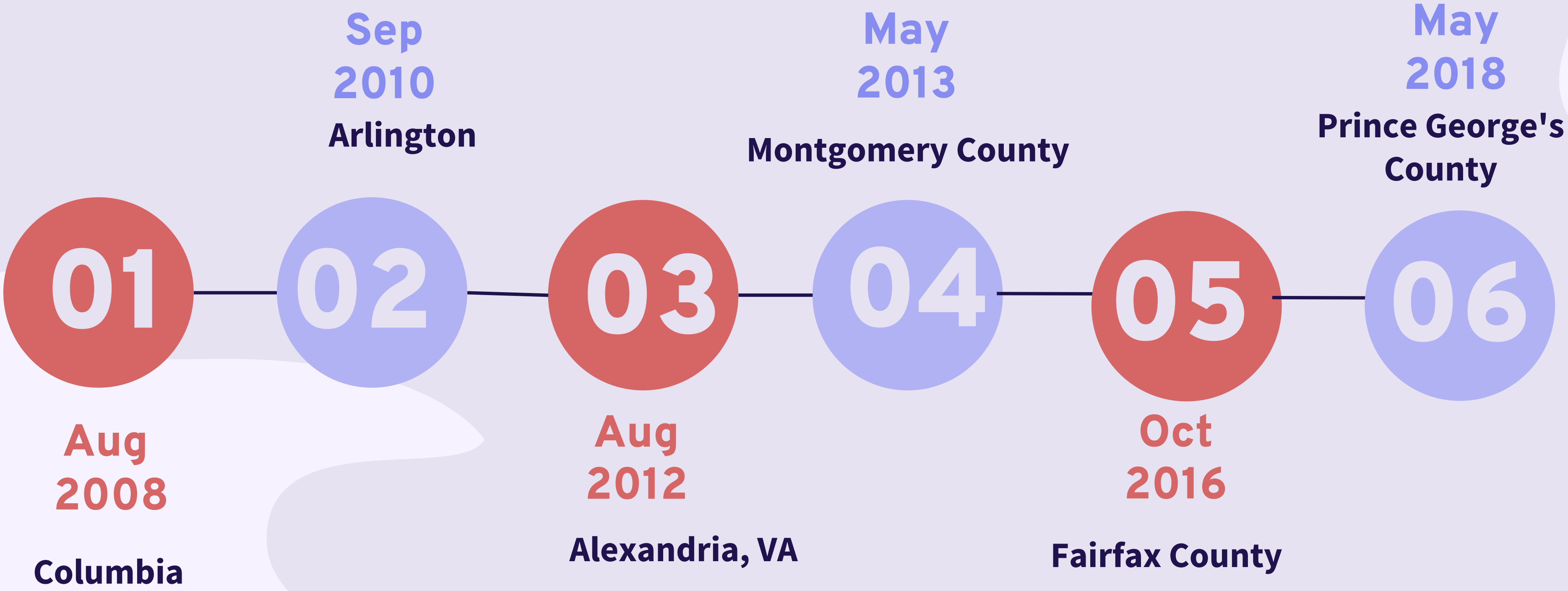
US$1/unlock+US$0.05/min

Annual Membership

US$7.92/month

Ref: https://ride.capitalbikeshare.com/pricing

Capital Bikeshare History



01

Aug 2008

Columbia

02

Sep 2010

Arlington

03

Aug 2012

Alexandria, VA

04

May 2013

Montgomery County

05

Oct 2016

Fairfax County

06

May 2018

Prince George's County

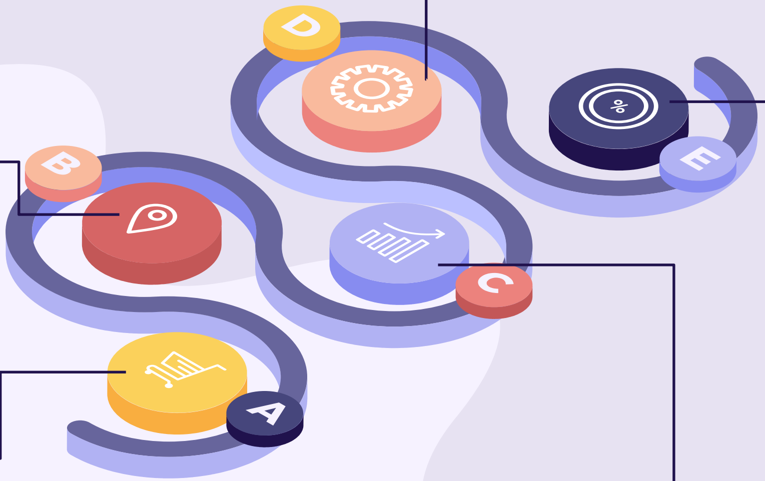
Ref: https://ride.capitalbikeshare.com/about

02

Business Understanding

Business Understanding

GOALS !



GOAL 4

Attract and retain users

GOAL 2

Specify different products demands

GOAL 5

Maximize Profits

GOAL 1

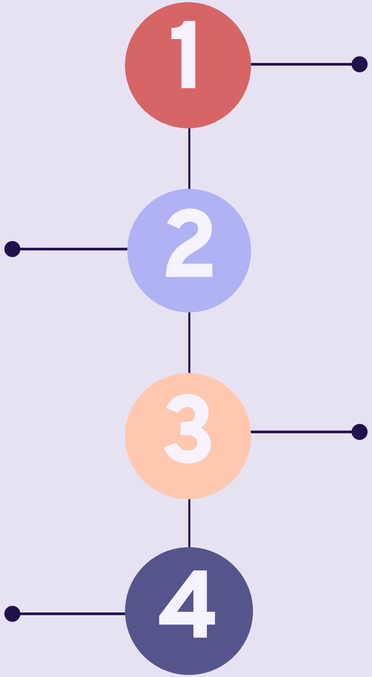
Better understand Users' needs

GOAL 3

Identify Seasonality and Weather Trends

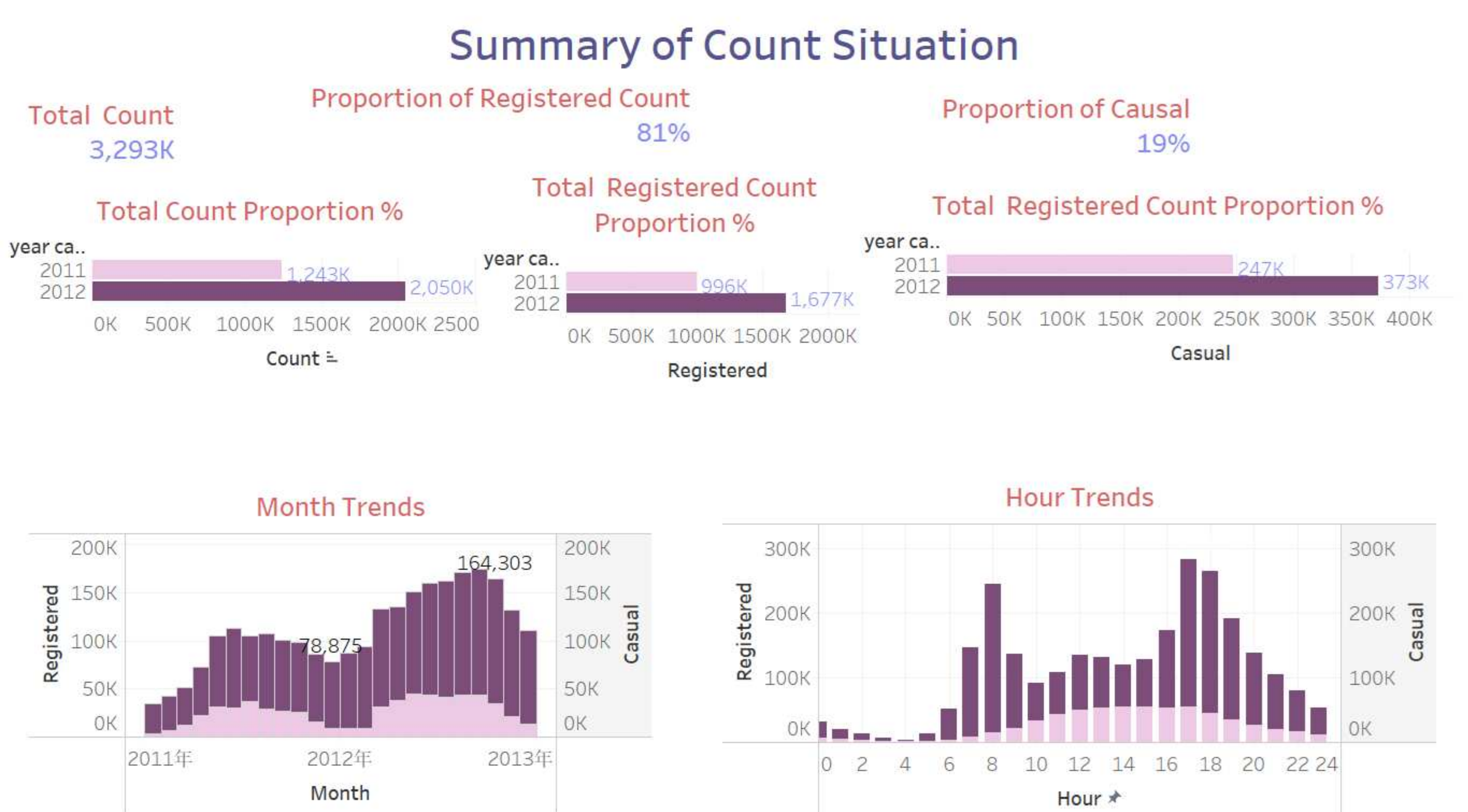
Business Understanding

CHALLENGES!



1. Modelling Overfitting
2. Models Sensitive to Outliers
3. Correlation does not imply causality meaning
4. Related business information is hard to obtain

Business Understanding



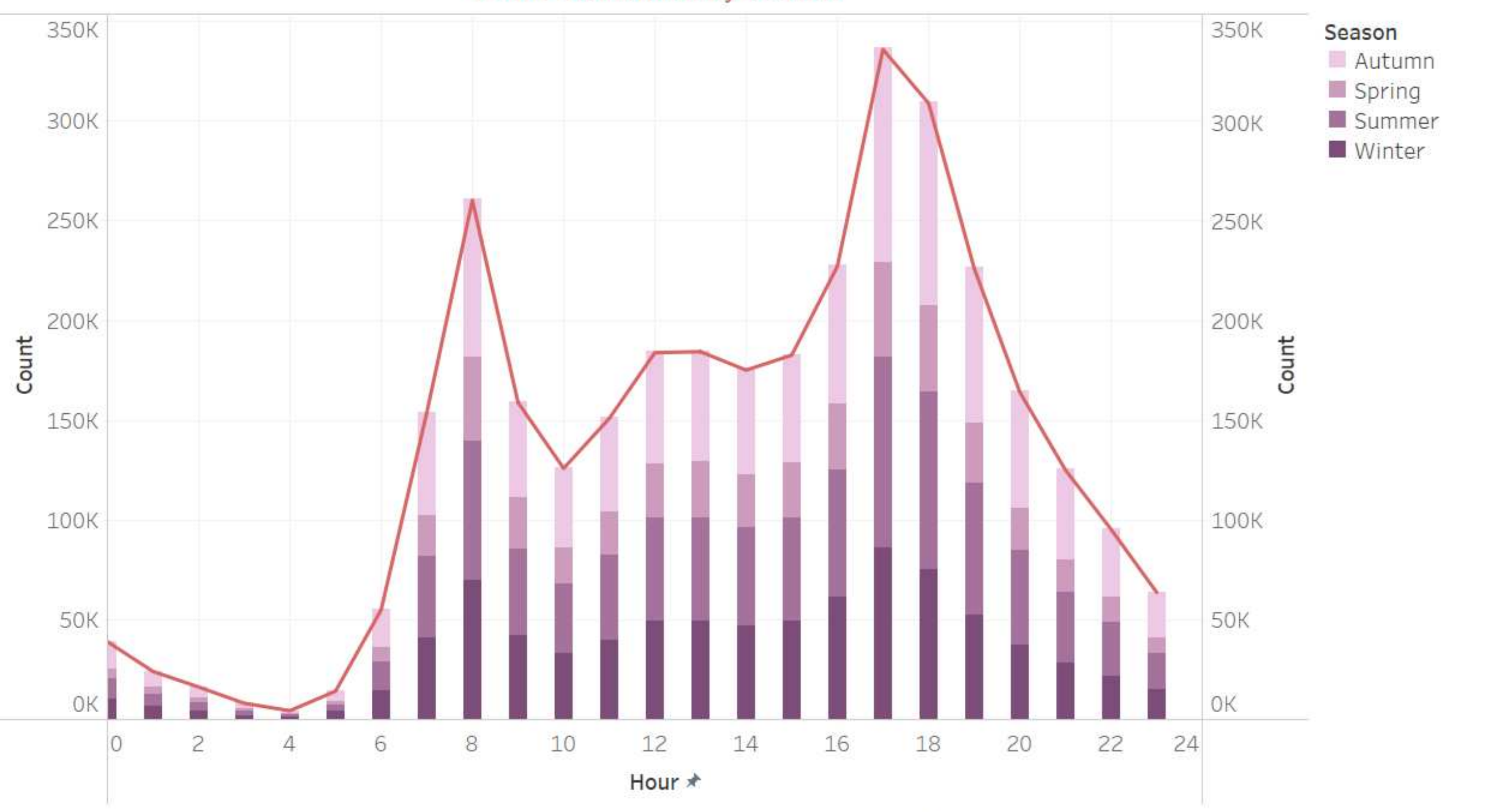
Summary of Count Situation

Total Count
3,293K

Proportion of Registered Count
81%

Proportion of Causal
19%

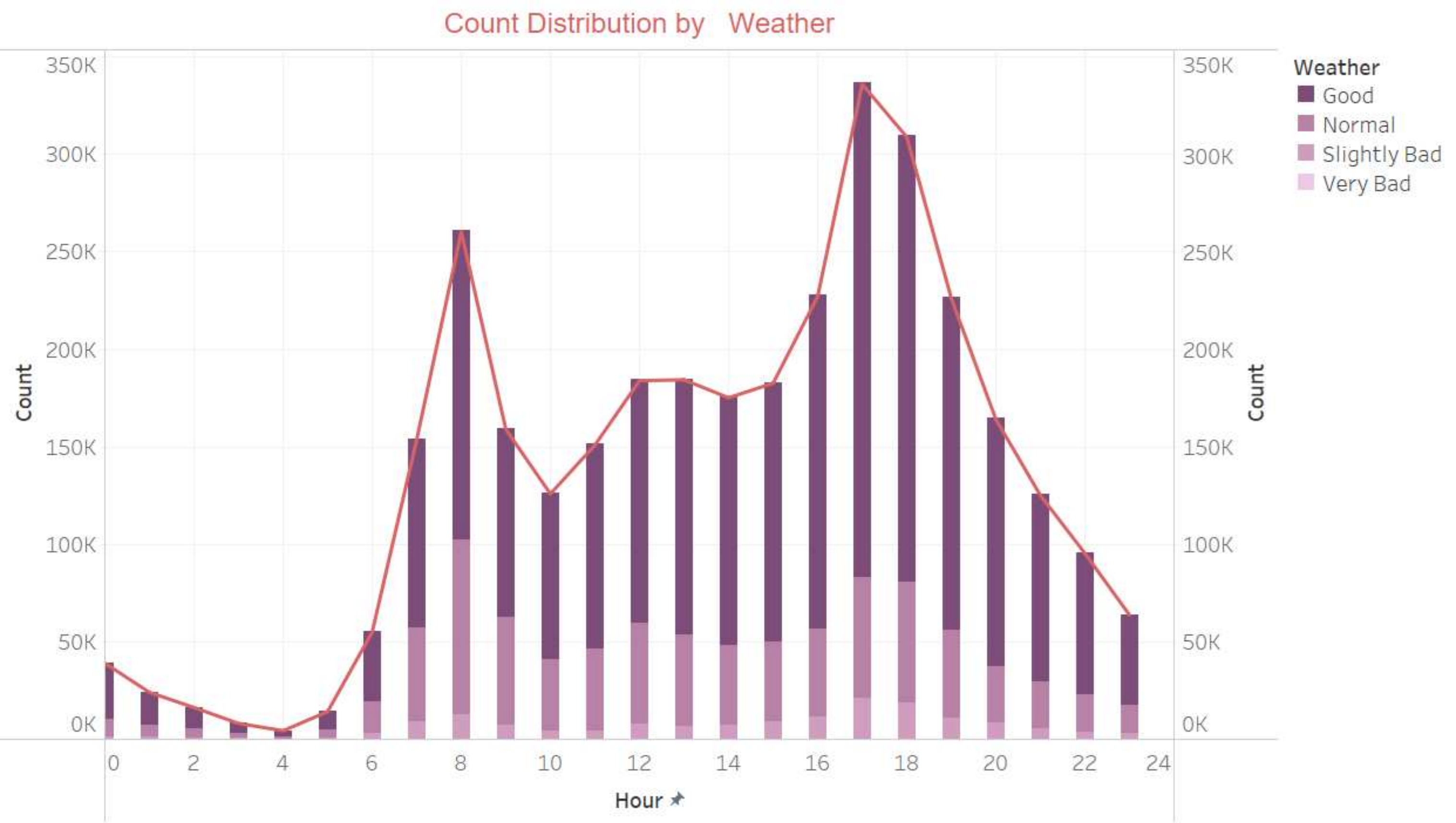
Business Understanding



Count Distribution by season

The trends of sum of Count and sum of Count for Hour. For pane Sum of Count: Color shows details about Season.

Business Understanding



The trends of sum of Count and sum of Count for Hour. For pane Sum of Count: Color shows details about Weather.

03

Data Understanding

Our Datasets

Hourly

Rows: 17,380

Cols: 17

Daily

Rows: 731

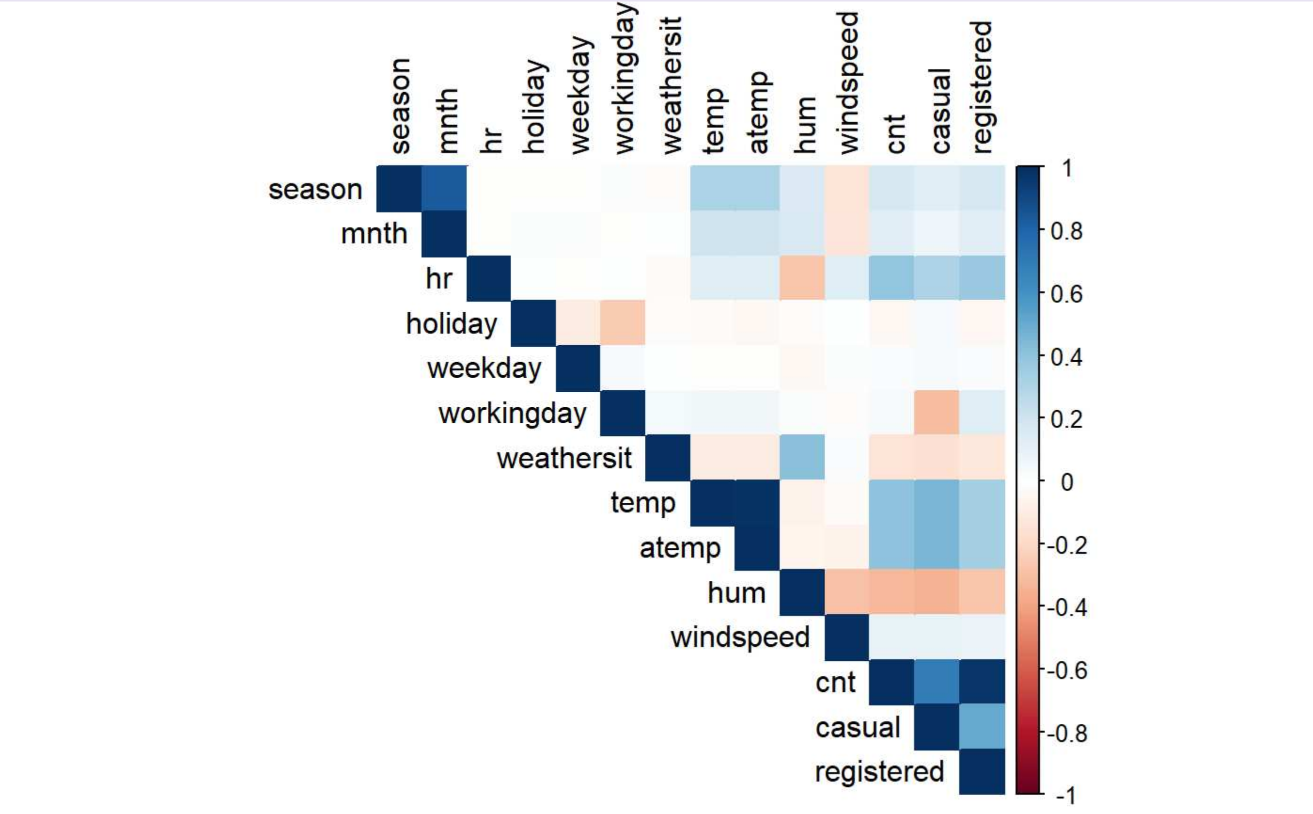
Cols: 16

About the key variables

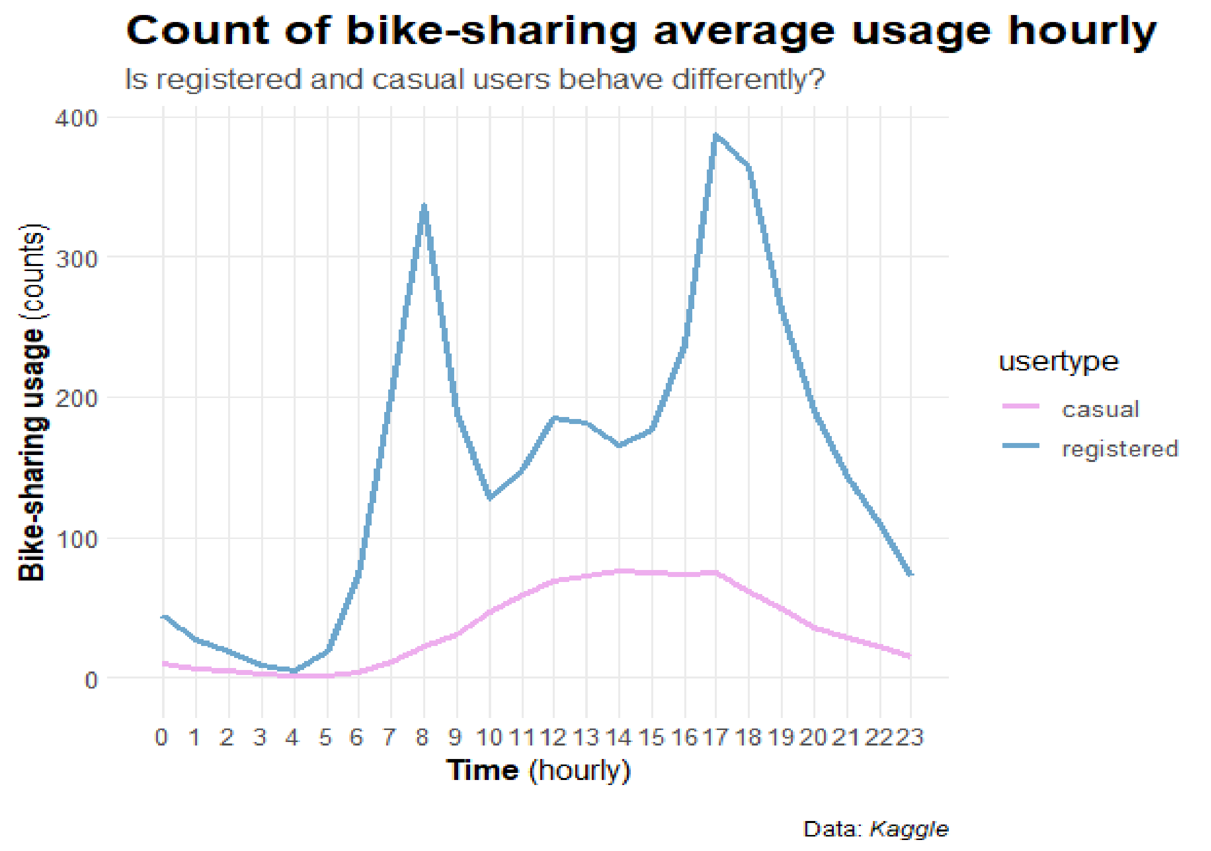
Name	Definition
Season	1:winter 2:spring 3:summer 4:fall
Holiday/working day	Is the day a holiday/workingday or not.(1 or 0)
Weather	Weather conditions(4 degree)
Temp/Atemp	The standardized temperature/feeling temperature.
Humidity	The standardized Humidity.
Windspeed	The standardized windspeed.
Registered	Registered users' ridership
Casual	Casual users' ridership
CNT	Total users

Data Understanding

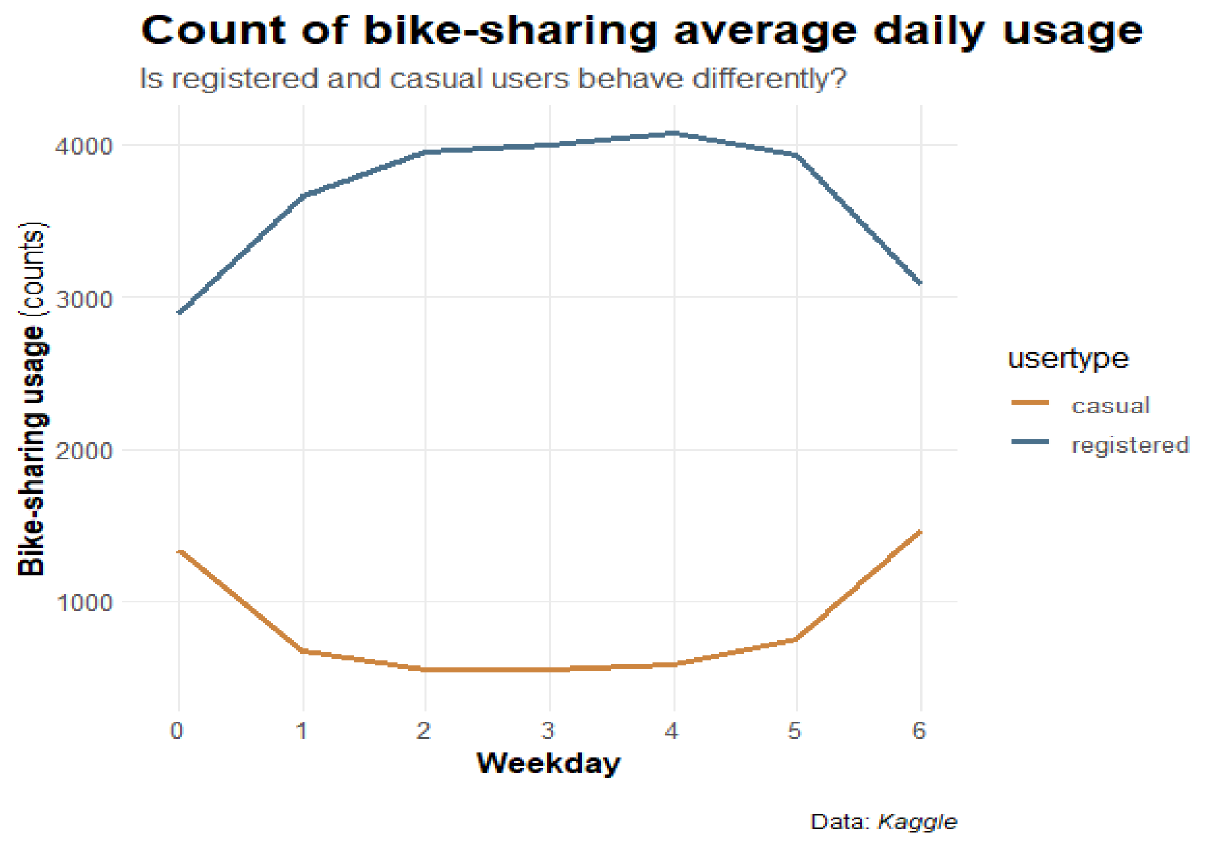
Correlations between Variables



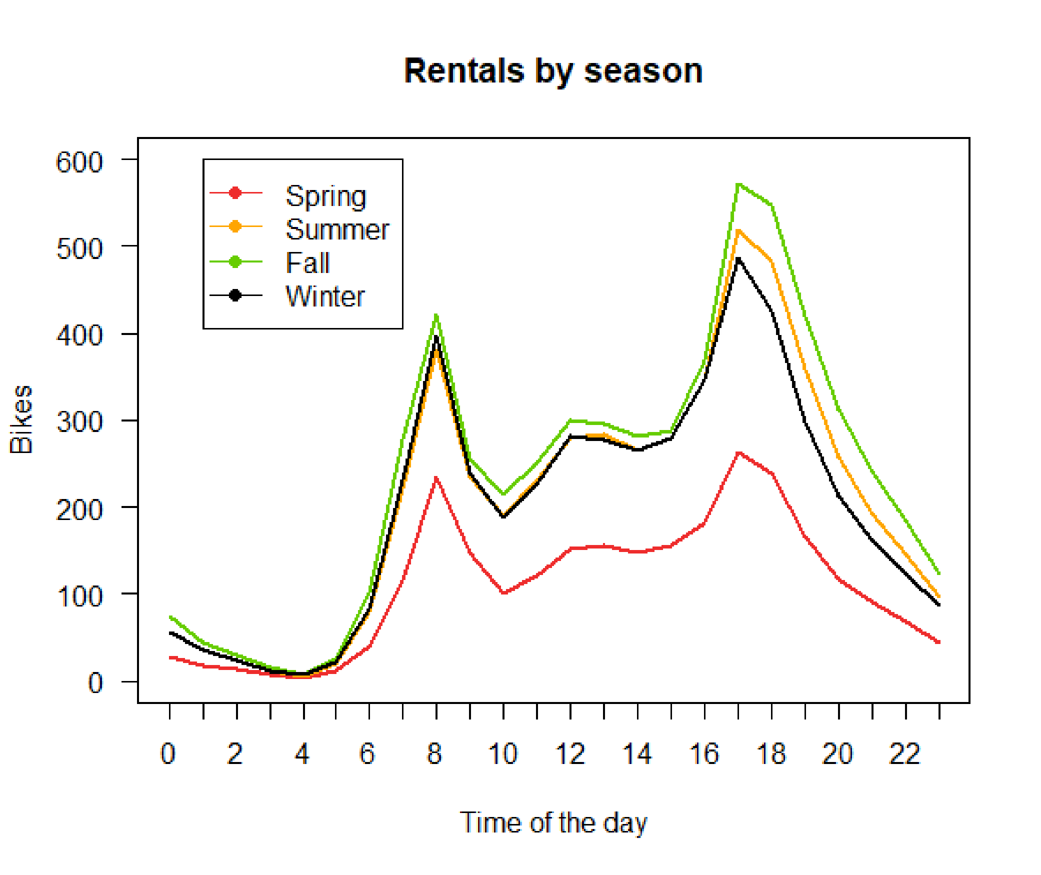
Some data explorations



Some data explorations



Some data explorations

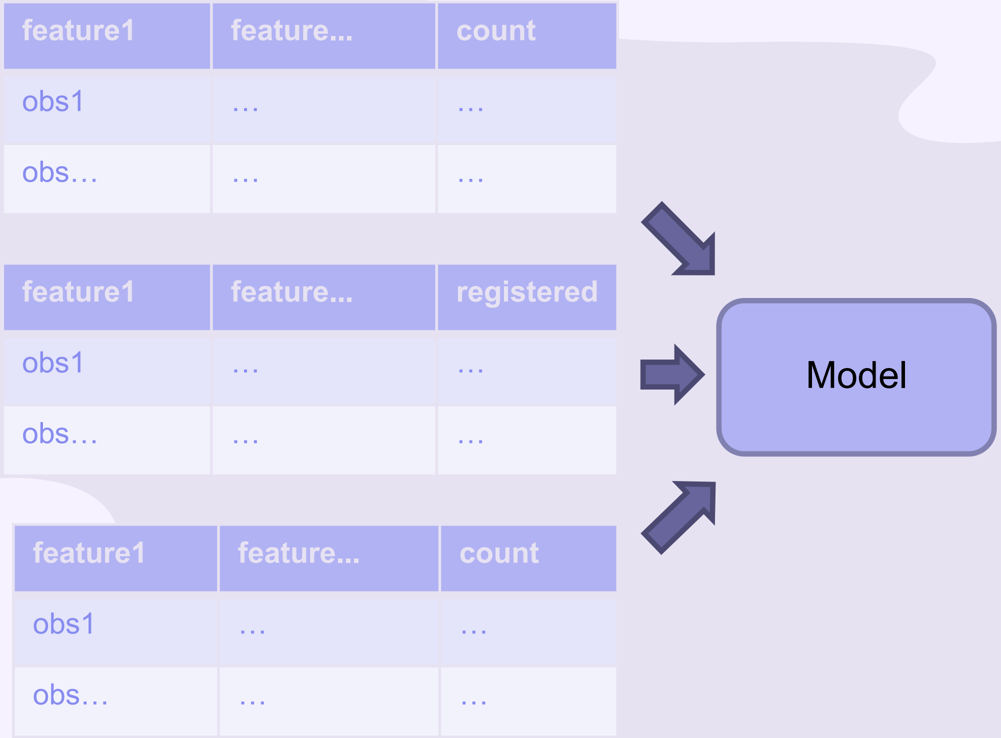


04

Data Preparation

3 potential target variable

- We decided to use all of them as target variable, but not together.



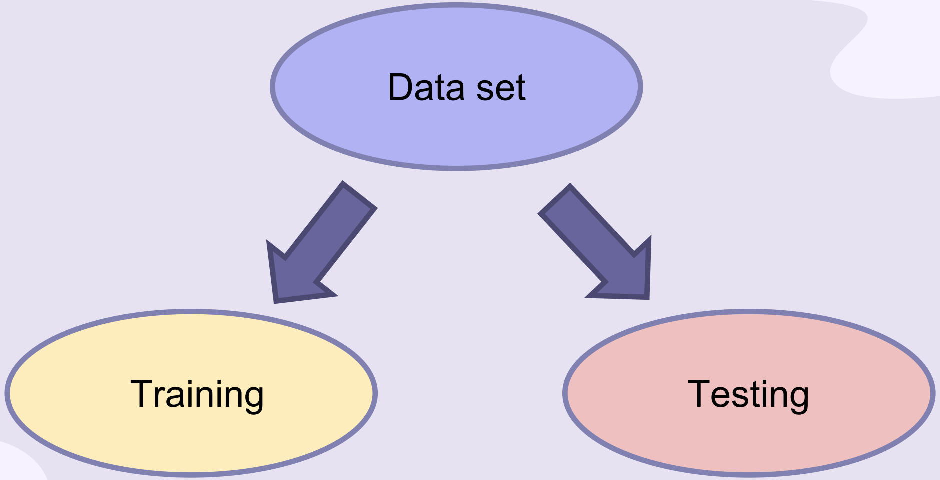
feature1	feature...	count
obs1	…	…
obs…	…	…

feature1	feature...	registered
obs1	…	…
obs…	…	…

feature1	feature...	count
obs1	…	…
obs…	…	…

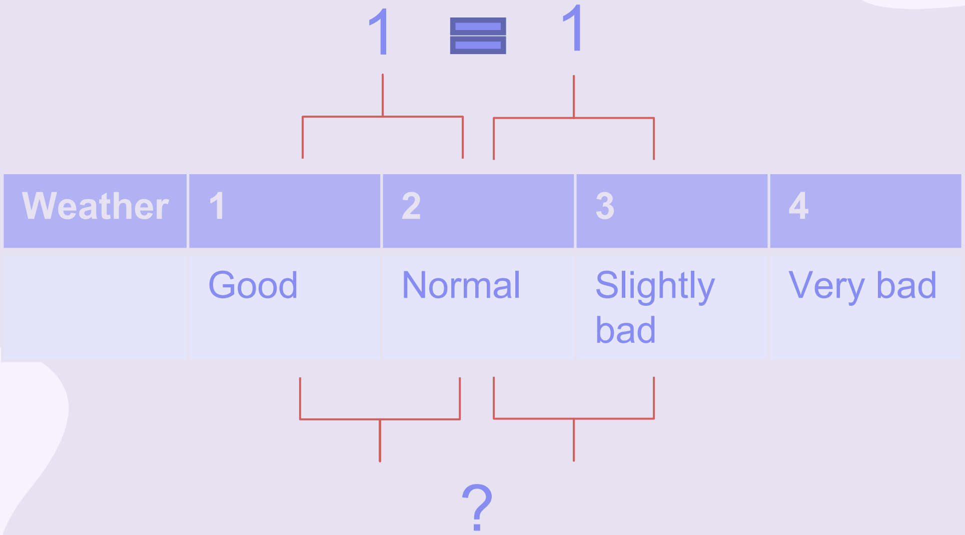
Evaluate the models

- We randomly divide the data set into training set and test set.



Categorical variables data type

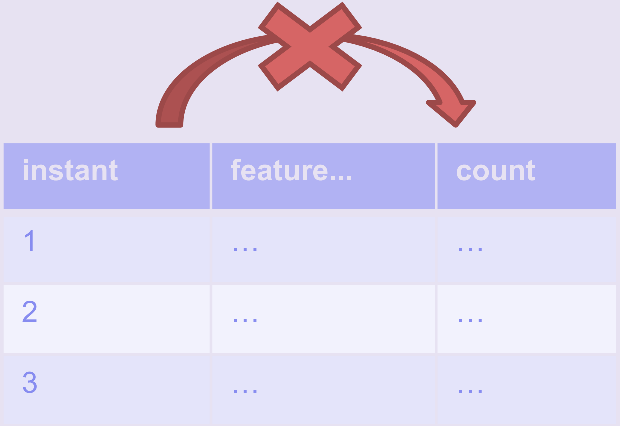
- We set the categorical variables as factor.



Weather	1	2	3	4
	Good	Normal	Slightly bad	Very bad

Irrelevant variables

- We delete the irrelevant variables.

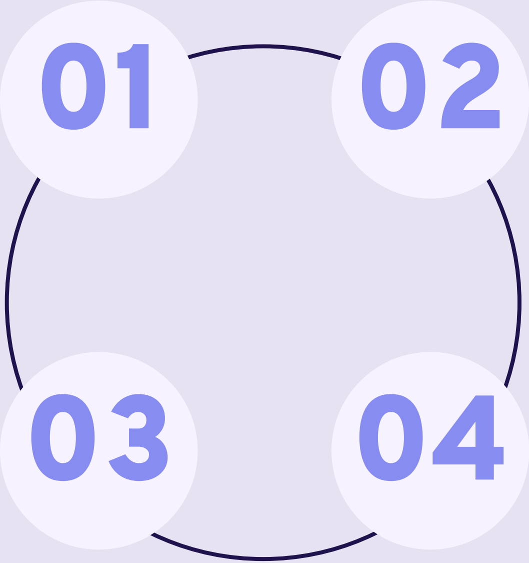


instant	feature...	count
1	…	…
2	…	…
3	…	…

05

Modeling

Modeling



Linear Regression

Baseline Model

Best Subsets Regression

Model Selections
Regression Models

Random Forest

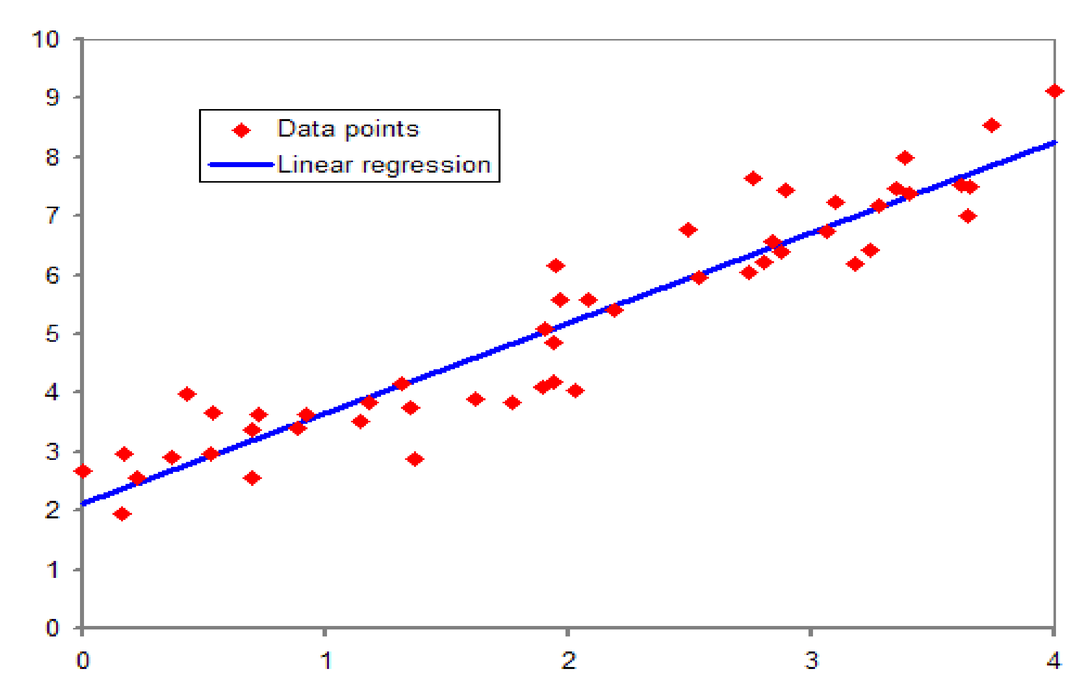
Bagging+Decision Tree

XGboost

Gradient
Boosting
Regression Tree

Modeling

Linear Regression



Ref:https://en.wikipedia.org/wiki/Regression_analysis

Pros:

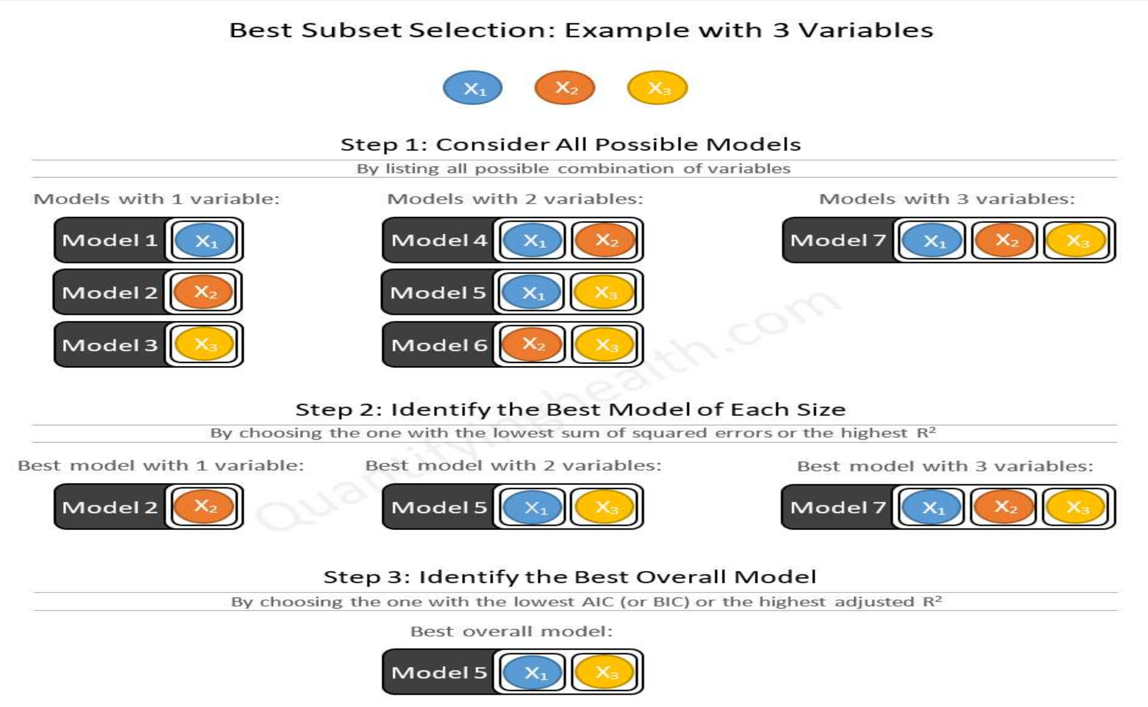
- Simple and very easy to interpret the result
- Handle overfitting very in dimension reductions and cross-validation
- Perform exceptionally well for linearly separable data

Cons:

- Prone to noise and overfitting
- Sensitive to the Outliers
- Prone to multicollinearity

Modeling

Best Subsets Regression


Best Subset Selection: Example with 3 Variables

Step 1: Consider All Possible Models

By listing all possible combination of variables

Step 2: Identify the Best Model of Each Size

By choosing the one with the lowest sum of squared errors or the highest R^2

Step 3: Identify the Best Overall Model

By choosing the one with the lowest AIC (or BIC) or the highest adjusted R^2

Pros:

- Improves generalizability by eliminating unnecessary predictors
- Simple and very easy to interpret the result
- Reproducible and Objective

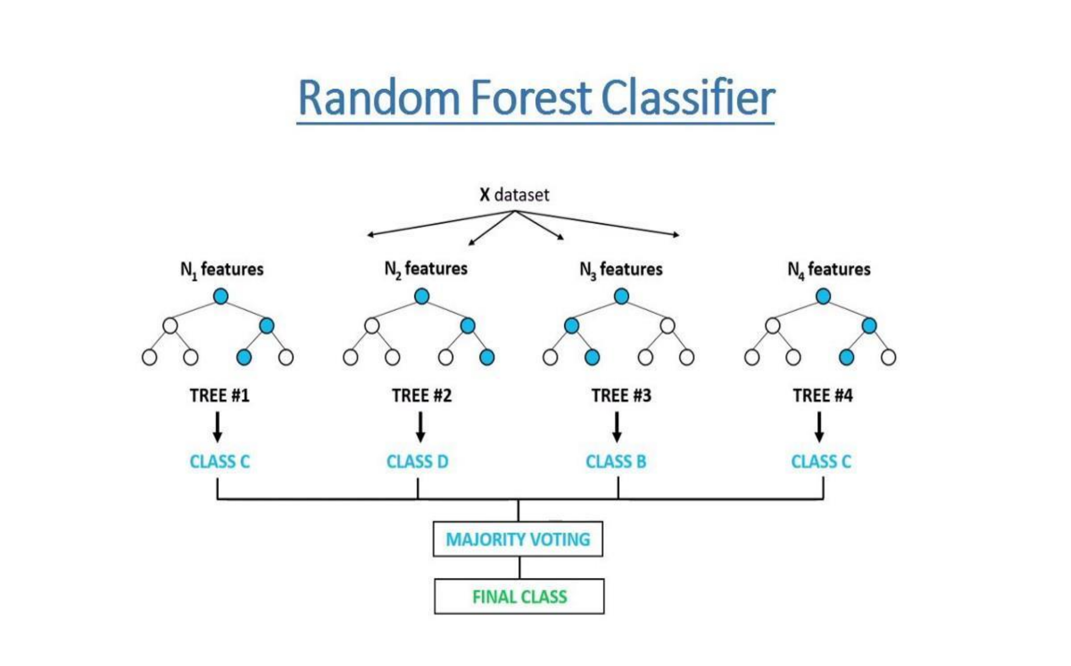
Cons:

- Computation Limitation
- Theoretical limitation

Ref:https://quantifyinghealth.com/best-subset-selection/

Modeling

Random Forest



Pros:

- Overcome overfitting by averaging the results of different tree models
- Reduce Variance
- Flexible and High Accuracy

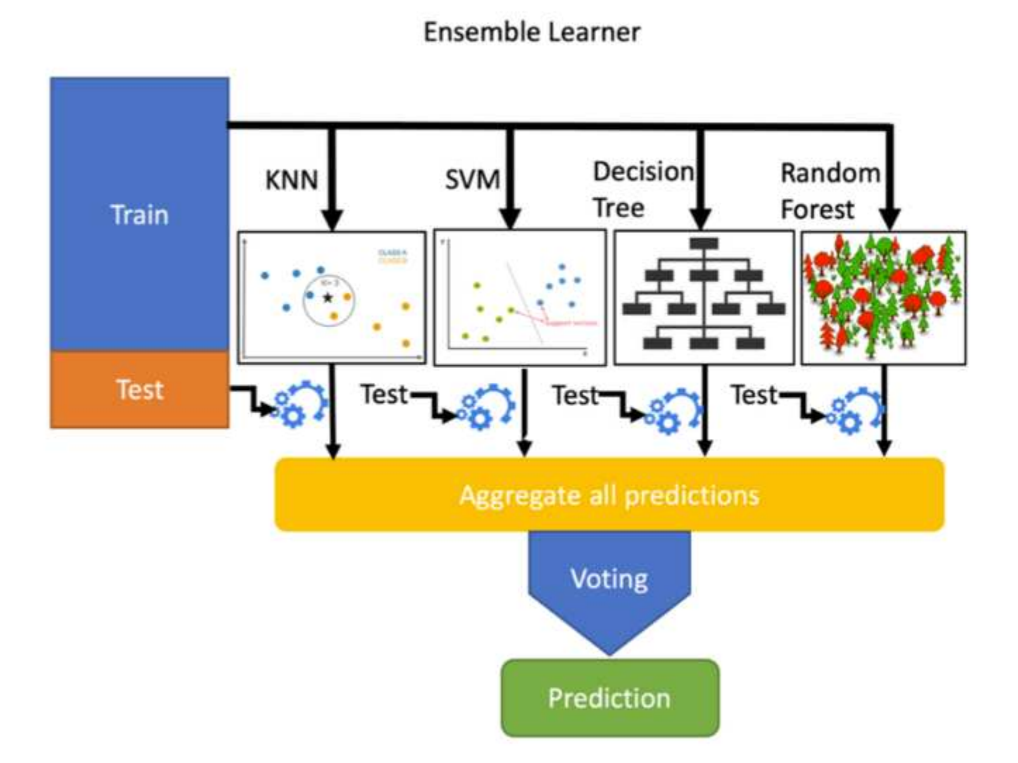
Cons:

- High Complexity and less intuitive compared to other tree models
- Harder and time-confusing to construct
- Computation Limitation

Ref:https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

Modeling

Xgboost



Pros:

- Reduce bias and increase accuracy
- Less prone to outliers and overfitting
- Regularization and missing values handling

Cons:

- Difficult interpretation , visualization tough
- Harder to tune as there are too many hyperparameters
- More training time is needed

Ref: https://medium.com/sfu-cspmp/xgboost-a-deep-dive-into-boosting-f06c9c41349
https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6

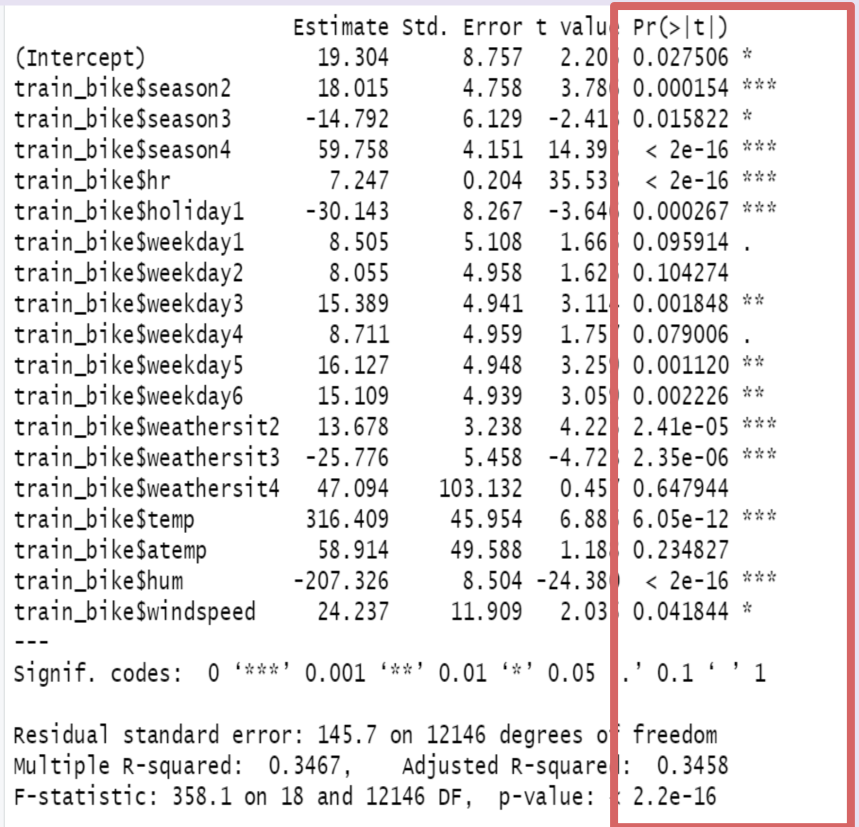
Modeling-Linear Regression

3 Target Variables

- CNT
- Registered CNT
- Causal CNT

Modeling-Linear Regression

CNT



```
                          Estimate Std. Error t valu Pr(>|t|)
(Intercept)                 19.304      8.757   2.20  0.027506 *
train_bike$season2          18.015      4.758   3.78  0.000154 ***
train_bike$season3         -14.792      6.129  -2.41  0.015822 *
train_bike$season4          59.758      4.151  14.39   < 2e-16 ***
train_bike$hr                7.247      0.204  35.53   < 2e-16 ***
train_bike$holiday1        -30.143      8.267  -3.64  0.000267 ***
train_bike$weekday1          8.505      5.108   1.66  0.095914 .
train_bike$weekday2          8.055      4.958   1.62  0.104274
train_bike$weekday3         15.389      4.941   3.11  0.001848 **
train_bike$weekday4          8.711      4.959   1.75  0.079006 .
train_bike$weekday5         16.127      4.948   3.25  0.001120 **
train_bike$weekday6         15.109      4.939   3.05  0.002226 **
train_bike$weathersit2      13.678      3.238   4.22  2.41e-05 ***
train_bike$weathersit3     -25.776      5.458  -4.72  2.35e-06 ***
train_bike$weathersit4      47.094    103.132   0.45  0.647944
train_bike$temp            316.409     45.954   6.88  6.05e-12 ***
train_bike$atemp            58.914     49.588   1.18  0.234827
train_bike$hum            -207.326      8.504 -24.38   < 2e-16 ***
train_bike$windspeed        24.237     11.909   2.03  0.041844 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05  .’ 0.1 ‘ ’ 1

Residual standard error: 145.7 on 12146 degrees o  freedom
Multiple R-squared:  0.3467,    Adjusted R-square :  0.3458
F-statistic: 358.1 on 18 and 12146 DF,  p-value:  2.2e-16
```

Key Figures

R Square: 0.3467

Adjusted R Square: 0.3458

Residual Standard Error: 145.7

MSE: 43,807.35

F Statistics:358.1

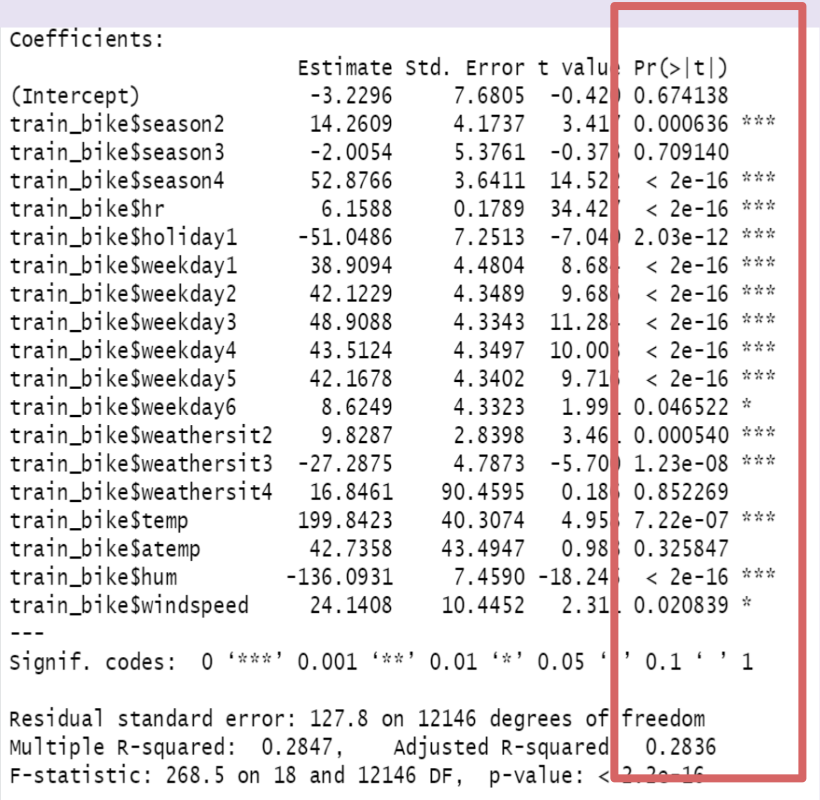
P-Value: <2.2e-16

AIC: 155,739.2

BIC: 155,887.3

Modeling-Linear Regression

Registered CNT



```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.2296     7.6805  -0.42  0.674138
train_bike$season2      14.2609     4.1737   3.41  0.000636 ***
train_bike$season3      -2.0054     5.3761  -0.37  0.709140
train_bike$season4      52.8766     3.6411  14.52   < 2e-16 ***
train_bike$hr            6.1588     0.1789  34.42   < 2e-16 ***
train_bike$holiday1    -51.0486     7.2513  -7.04  2.03e-12 ***
train_bike$weekday1     38.9094     4.4804   8.68   < 2e-16 ***
train_bike$weekday2     42.1229     4.3489   9.68   < 2e-16 ***
train_bike$weekday3     48.9088     4.3343  11.28   < 2e-16 ***
train_bike$weekday4     43.5124     4.3497  10.00   < 2e-16 ***
train_bike$weekday5     42.1678     4.3402   9.71   < 2e-16 ***
train_bike$weekday6      8.6249     4.3323   1.99  0.046522 *
train_bike$weathersit2   9.8287     2.8398   3.46  0.000540 ***
train_bike$weathersit3 -27.2875     4.7873  -5.70  1.23e-08 ***
train_bike$weathersit4  16.8461    90.4595   0.18  0.852269
train_bike$temp        199.8423    40.3074   4.95  7.22e-07 ***
train_bike$atemp        42.7358    43.4947   0.98  0.325847
train_bike$hum        -136.0931     7.4590 -18.24   < 2e-16 ***
train_bike$windspeed    24.1408    10.4452   2.31  0.020839 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127.8 on 12146 degrees of freedom
Multiple R-squared:  0.2847,	Adjusted R-squared:  0.2836 
F-statistic: 268.5 on 18 and 12146 DF,  p-value: < 2.2e-16
```

Key Figures

R Square: 0.2847

Adjusted R Square: 0.2836

Residual Standard Error: 127.8.

MSE: 28,719.25

F Statistics:268.5

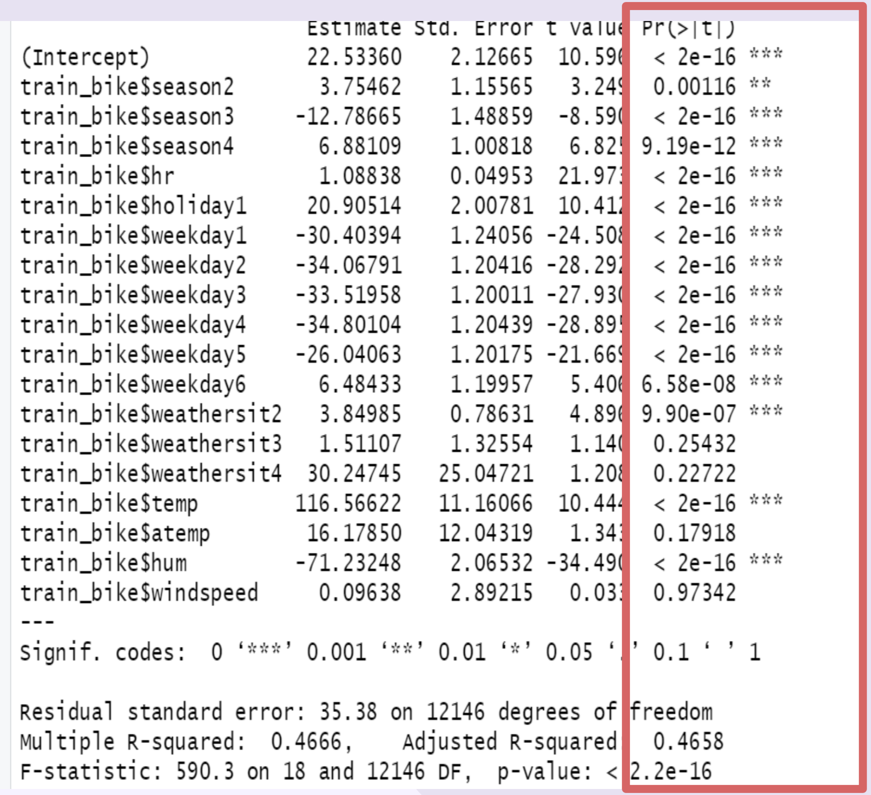
P-Value: <2.2e-16

AIC: 15,2549.2

BIC: 152,697.3

Modeling-Linear Regression

Causal CNT



```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             22.53360    2.12665  10.59  < 2e-16 ***
train_bike$season2       3.75462    1.15565   3.24  0.00116 **
train_bike$season3     -12.78665    1.48859  -8.59  < 2e-16 ***
train_bike$season4       6.88109    1.00818   6.82 9.19e-12 ***
train_bike$hr            1.08838    0.04953  21.97  < 2e-16 ***
train_bike$holiday1     20.90514    2.00781  10.41  < 2e-16 ***
train_bike$weekday1    -30.40394    1.24056 -24.50  < 2e-16 ***
train_bike$weekday2    -34.06791    1.20416 -28.29  < 2e-16 ***
train_bike$weekday3    -33.51958    1.20011 -27.93  < 2e-16 ***
train_bike$weekday4    -34.80104    1.20439 -28.89  < 2e-16 ***
train_bike$weekday5    -26.04063    1.20175 -21.66  < 2e-16 ***
train_bike$weekday6      6.48433    1.19957   5.40 6.58e-08 ***
train_bike$weathersit2   3.84985    0.78631   4.89 9.90e-07 ***
train_bike$weathersit3   1.51107    1.32554   1.14  0.25432
train_bike$weathersit4  30.24745   25.04721   1.20  0.22722
train_bike$temp        116.56622   11.16066  10.44  < 2e-16 ***
train_bike$atemp        16.17850   12.04319   1.34  0.17918
train_bike$hum         -71.23248    2.06532 -34.49  < 2e-16 ***
train_bike$windspeed     0.09638    2.89215   0.03  0.97342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.38 on 12146 degrees of freedom
Multiple R-squared:  0.4666,    Adjusted R-squared:  0.4658
F-statistic: 590.3 on 18 and 12146 DF,  p-value: < 2.2e-16
```

Key Figures

R Square: 0.4666

Adjusted R Square: 0.4658

Residual Standard Error: 35.38

MSE: 3,565.914

F Statistics:590.3

P-Value: <2.2e-16

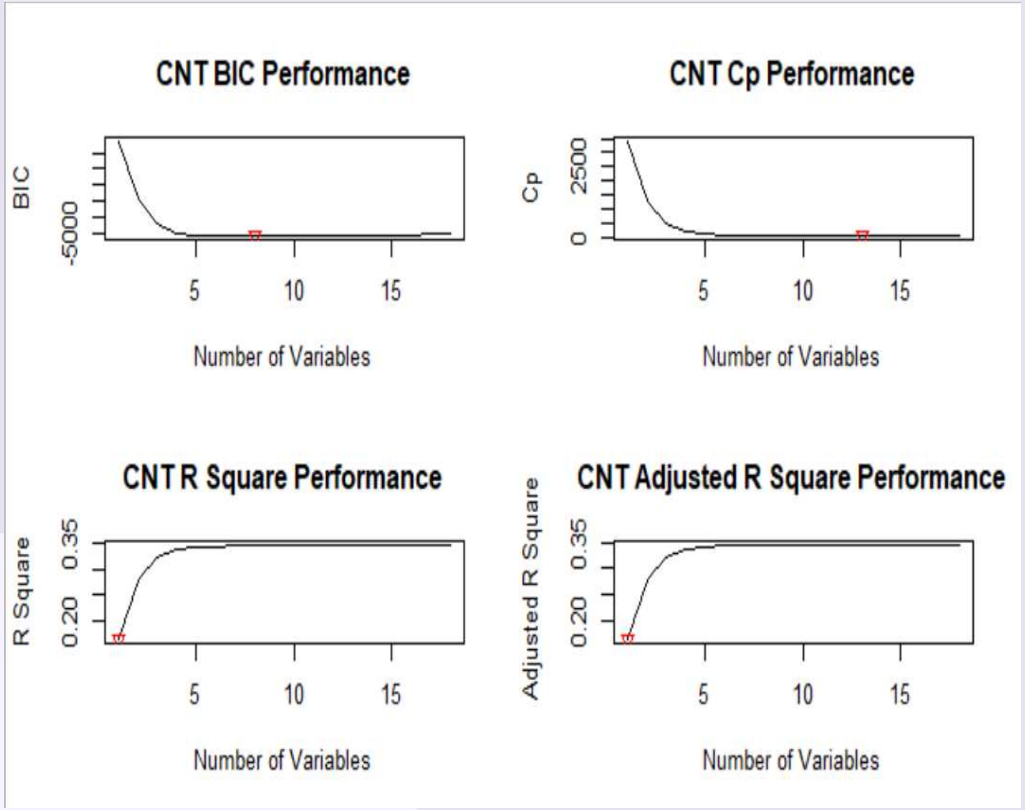
AIC: 12,1306.1

BIC: 12,1454.2

Modeling-Best Subsets Regression

Method

CNT-Training Set



Best 3 Model:
Temperature;humidity,Hour

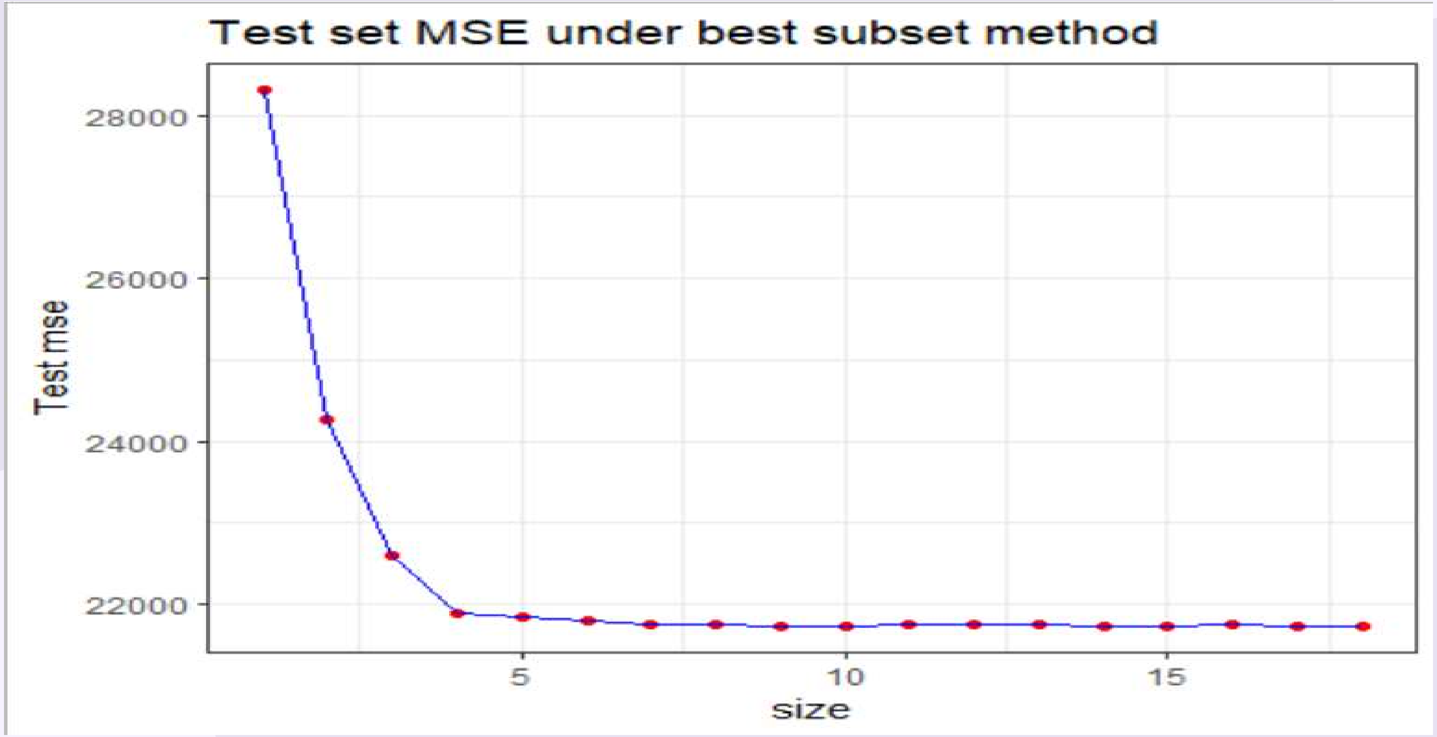
Best 4 Model:
Winter;Temperature,humidity,
Hour

Best 5 Model: Summer,Winter,Humidity,
Hour,Temperature

Best 8 Model: Summer,Winter,Hour
Holiday,Weather,Temperature,Humidity

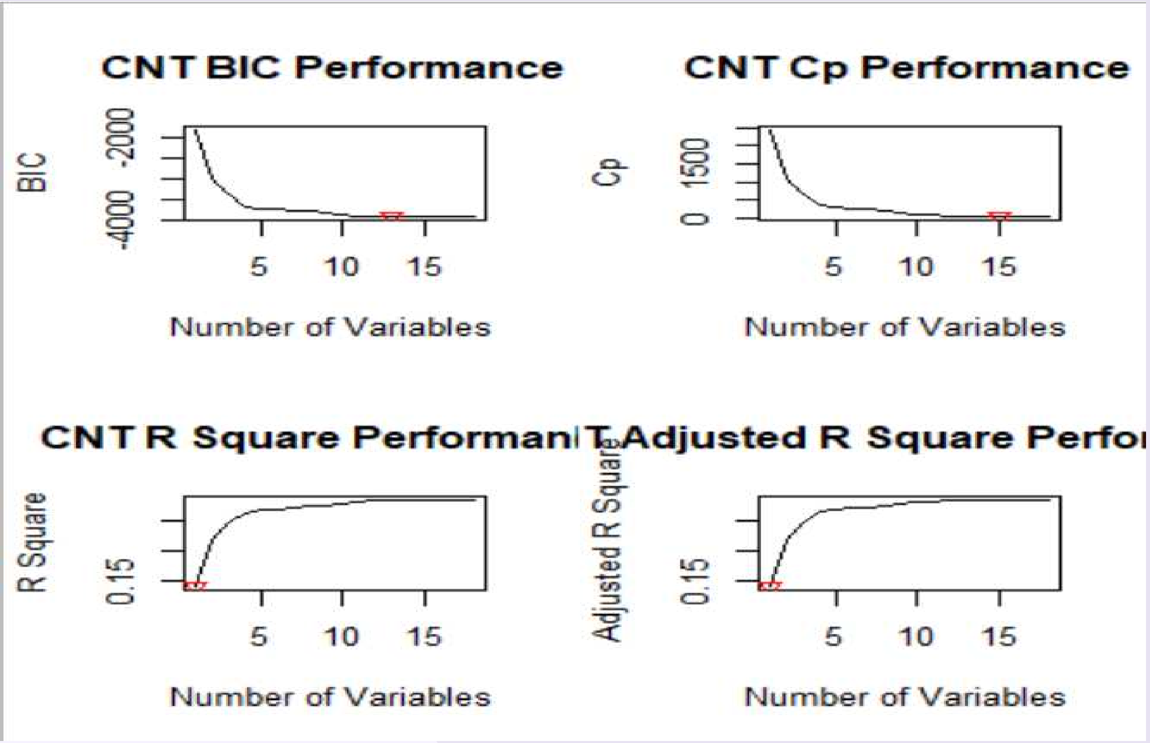
Modeling-Best Subsets Regression

CNT-Test Set



Modeling-Best Subsets Regression

Registered CNT -Training Set



Method

Best 3 Model: Hour,Temperature and Humidity

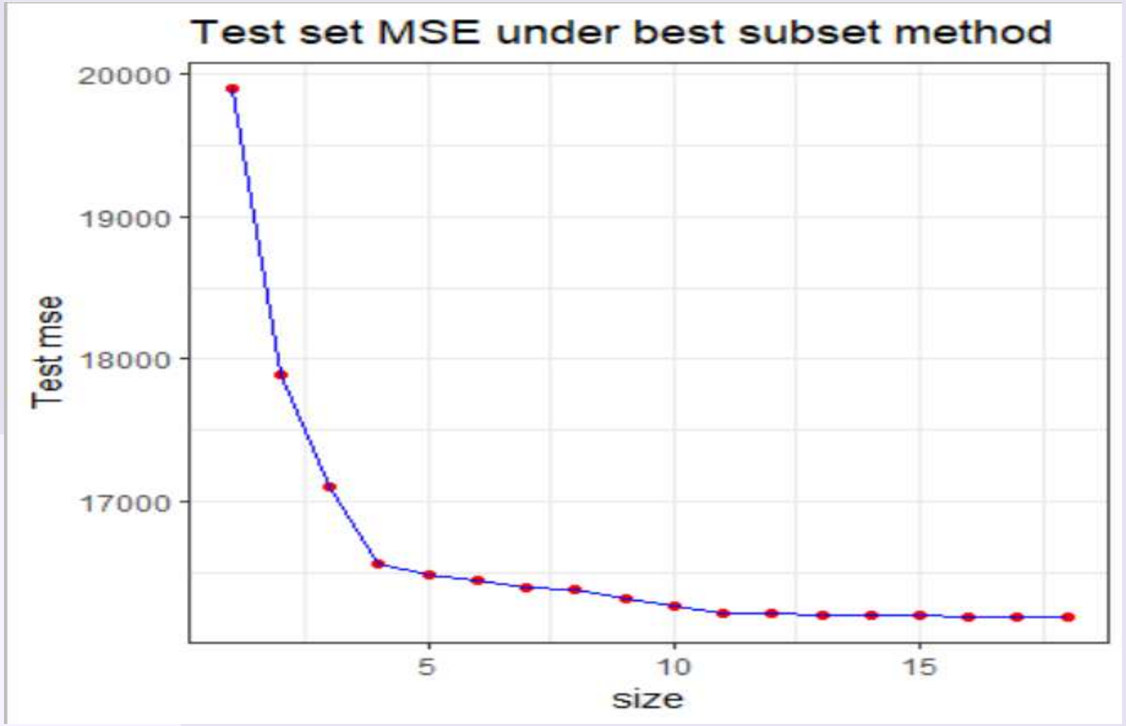
Best 4 Model: Winter;Temperature,humidity, Hour

Best 5 Model: Winter;Temperature,humidity, Hour,Saturday

Best 8 Model: Summer, Winter,Hour,Holiday,Saturday, Weather,Temperature,Humidity

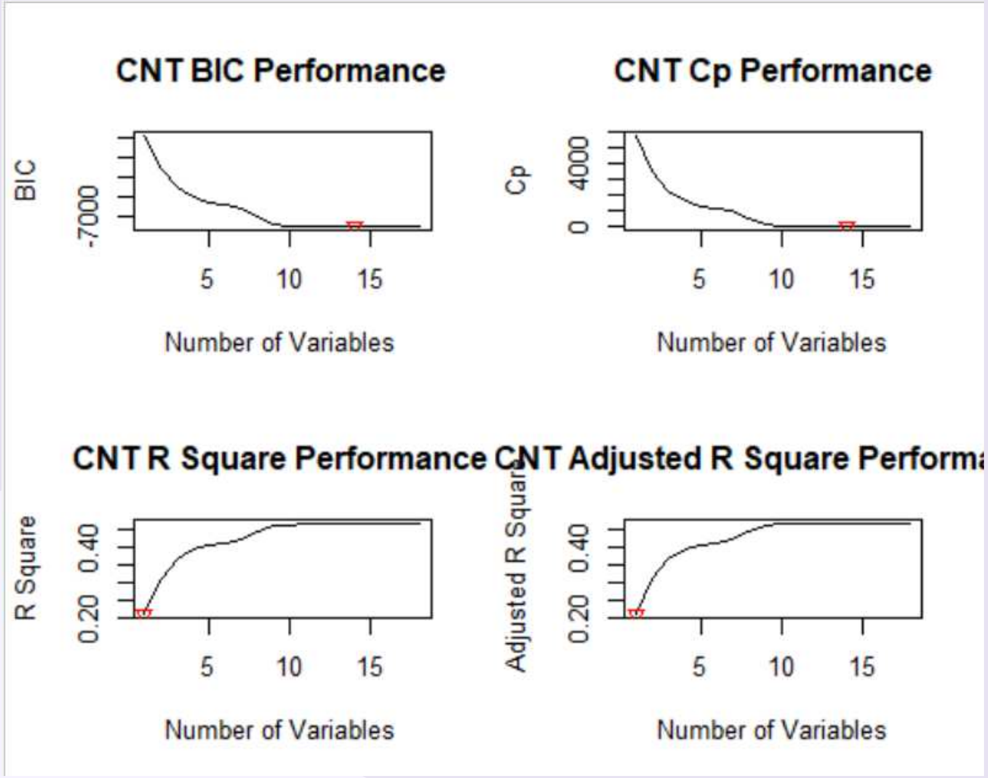
Modeling-Best Subsets Regression

Registered CNT -Test Set



Modeling-Best Subsets Regression

Causal CNT -Training Set



Method

Best 3 Model:
Saturday,Temperature,Humidity

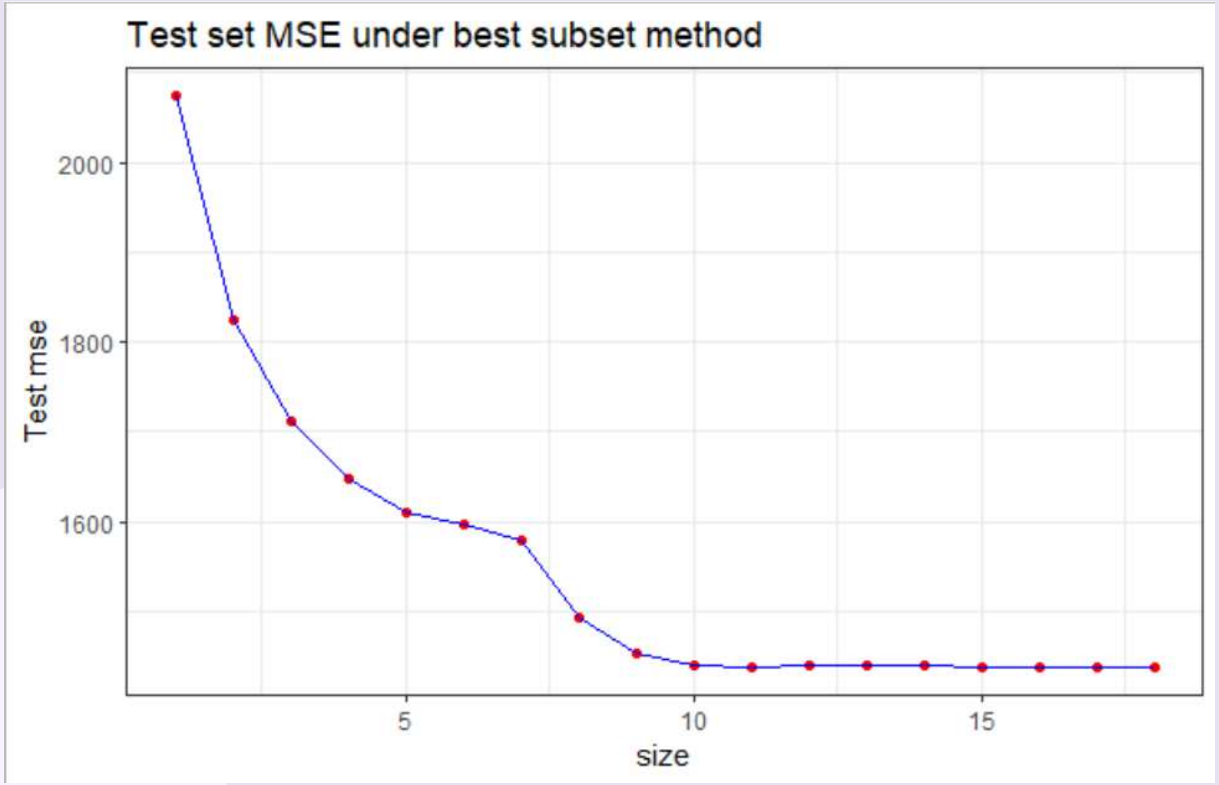
Best 4 Model:
Hour,Saturday,Temperature,Humidity

Best 5 Model:
Hour,Saturday,Temperature,Humidity,Autumn

Best 8 Model:
Hour,Temperature,Humidity,Monday,Tuesday,Wednesday,Thursday,Friday

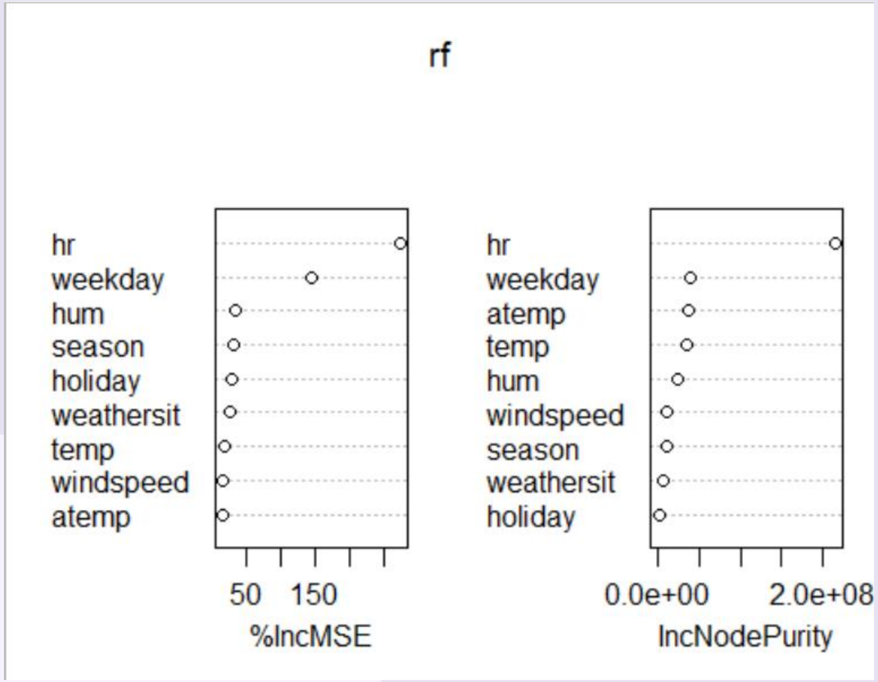
Modeling-Best Subsets Regression

Causal CNT
-Test Set

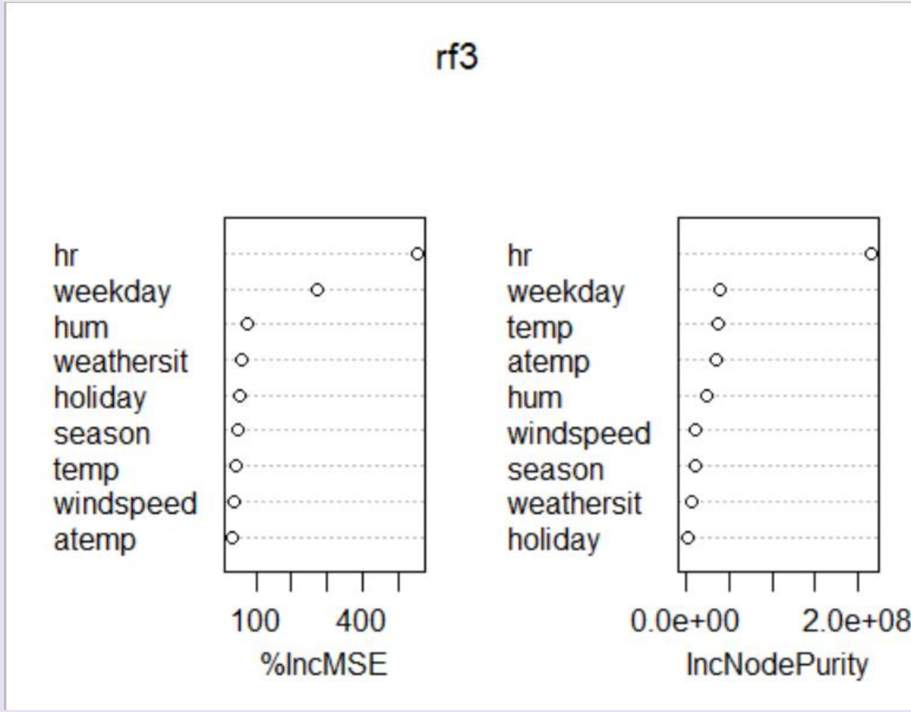


Modeling-Random Forest

CNT-Training Set
Tree=100 85.02% Var explained

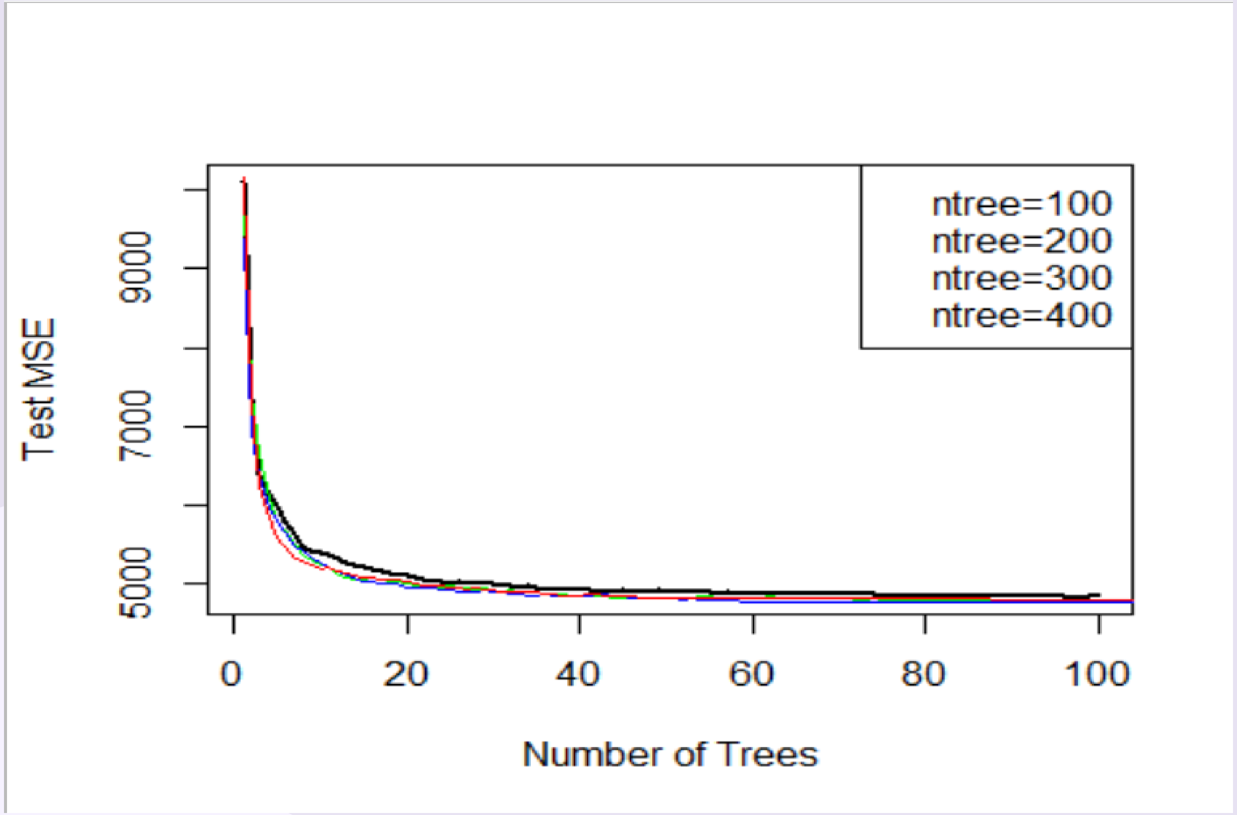


Tree=400 86.03% Var explained



Modeling-Random Forest

CNT-Test Set



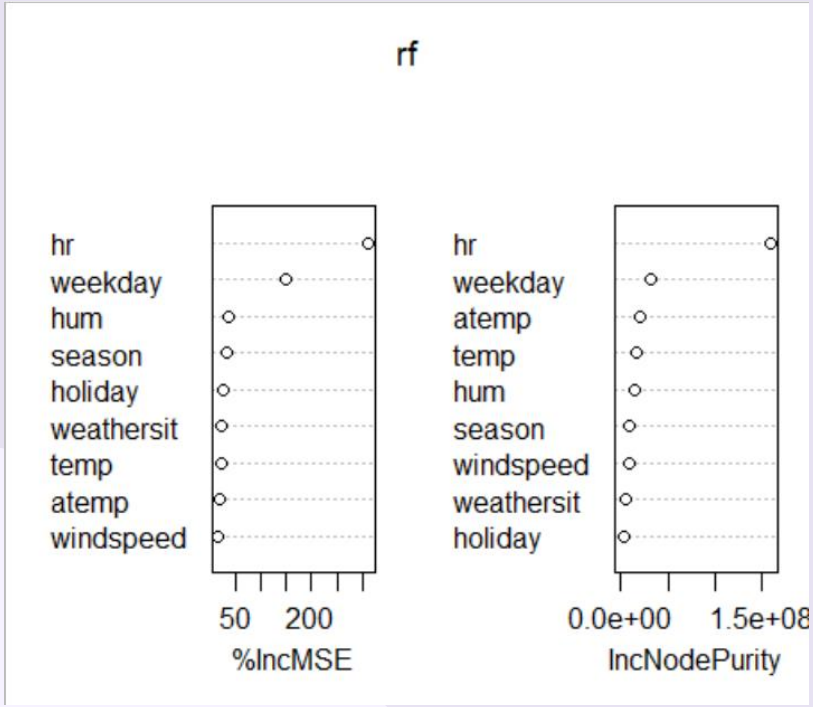
Tree=100
MSE:
4,834.429

Tree=400
MSE:
4,748.079

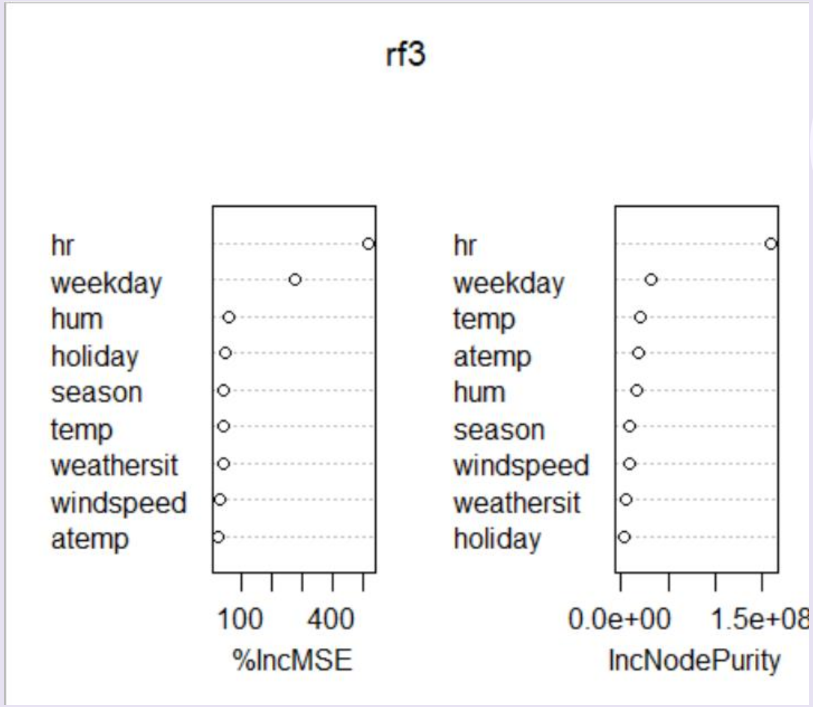
Modeling-Random Forest

Registered CNT-Training Set

Tree=100 84.75% Var explained

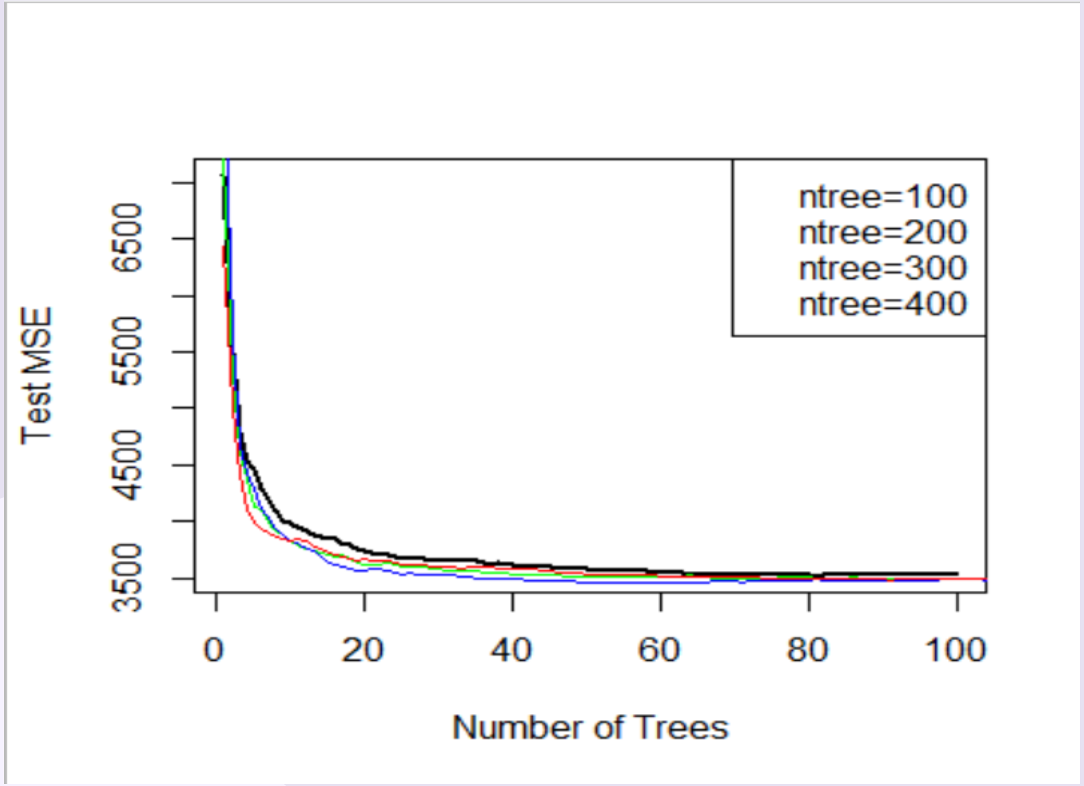


Tree=400 85.16% Var explained



Modeling-Random Forest

Registered CNT-Test Set



Tree=100 Minimized MSE: 3,521.261

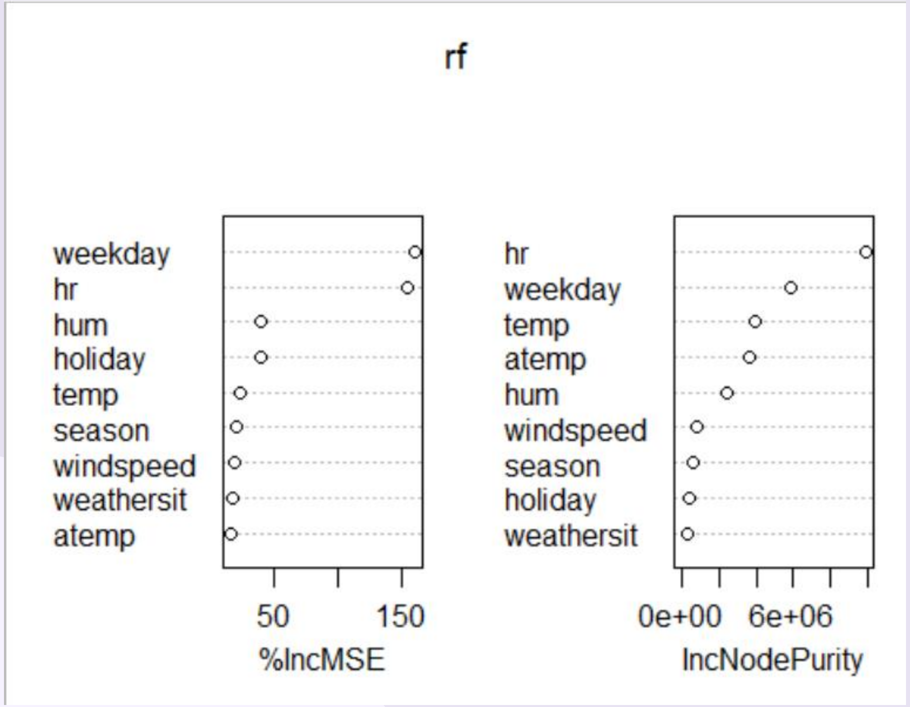
Tree=400 Minimized MSE: 3,439.367

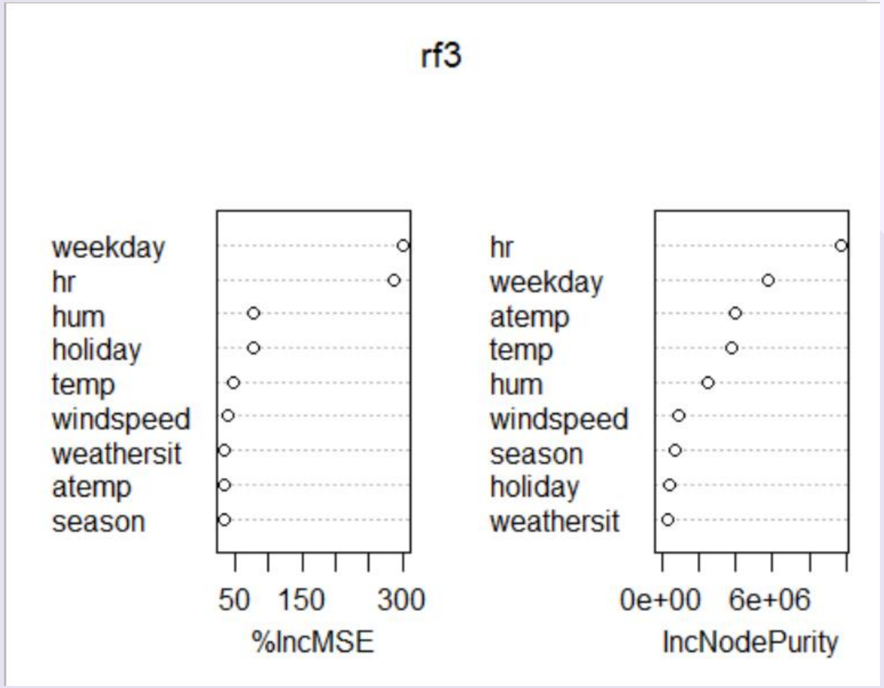
Modeling-Random Forest

Casual CNT-Training Set

Tree=100 88.13% Var explained

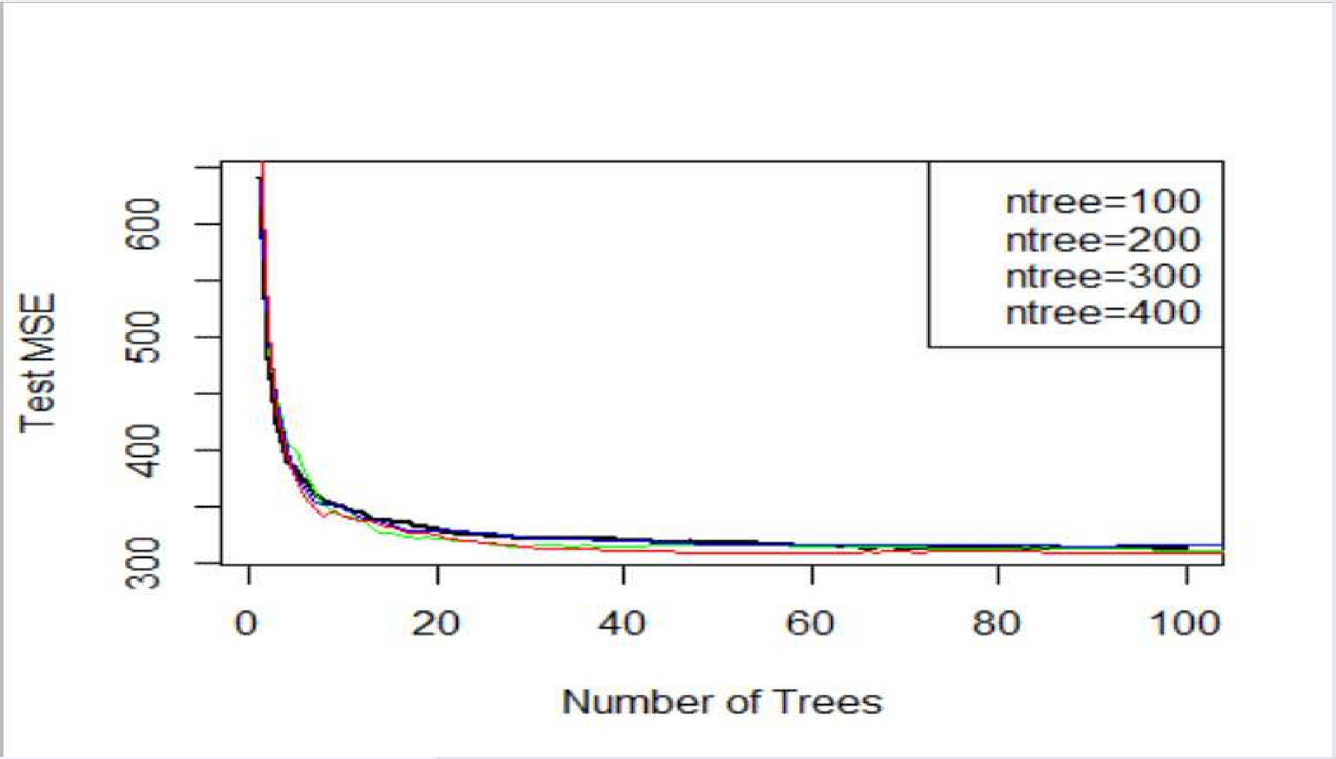
Tree=400 88.31% Var explained





Modeling-Random Forest

Casual CNT-Test Set

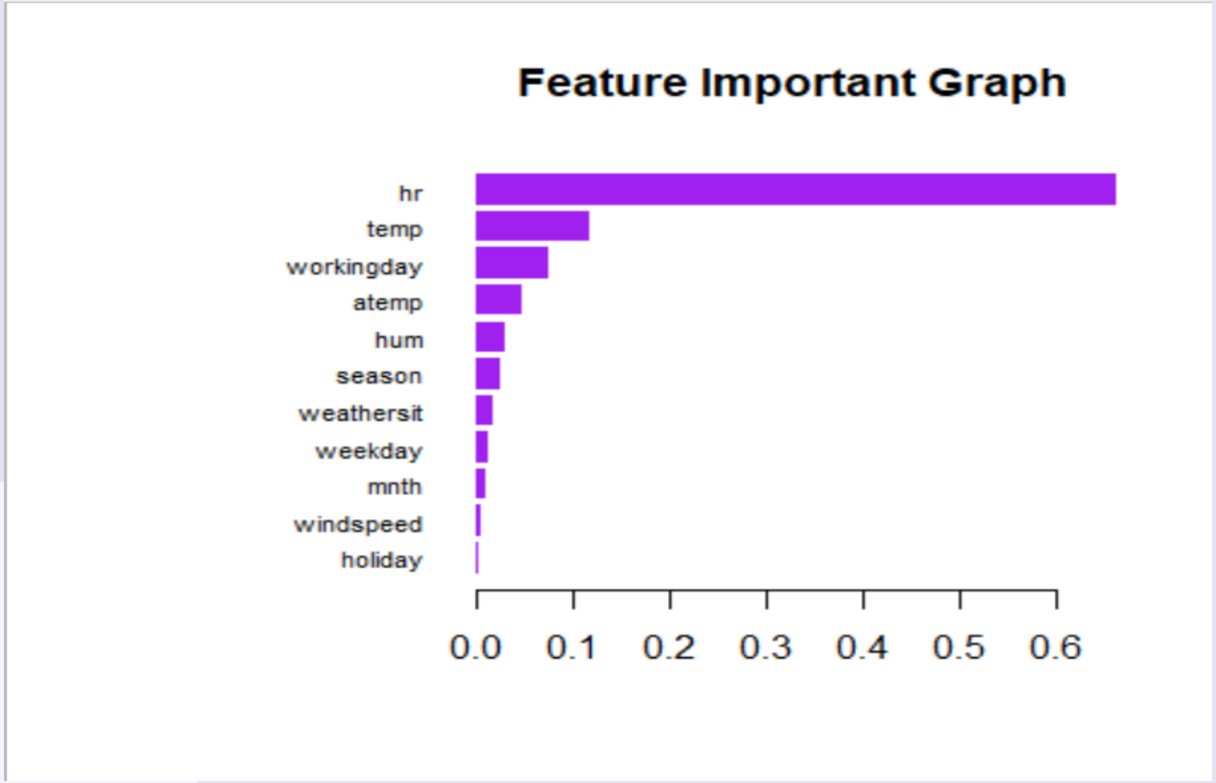


Tree=100 Minimized MSE: 312.5301

Tree=400 Minimized MSE: 307.3897

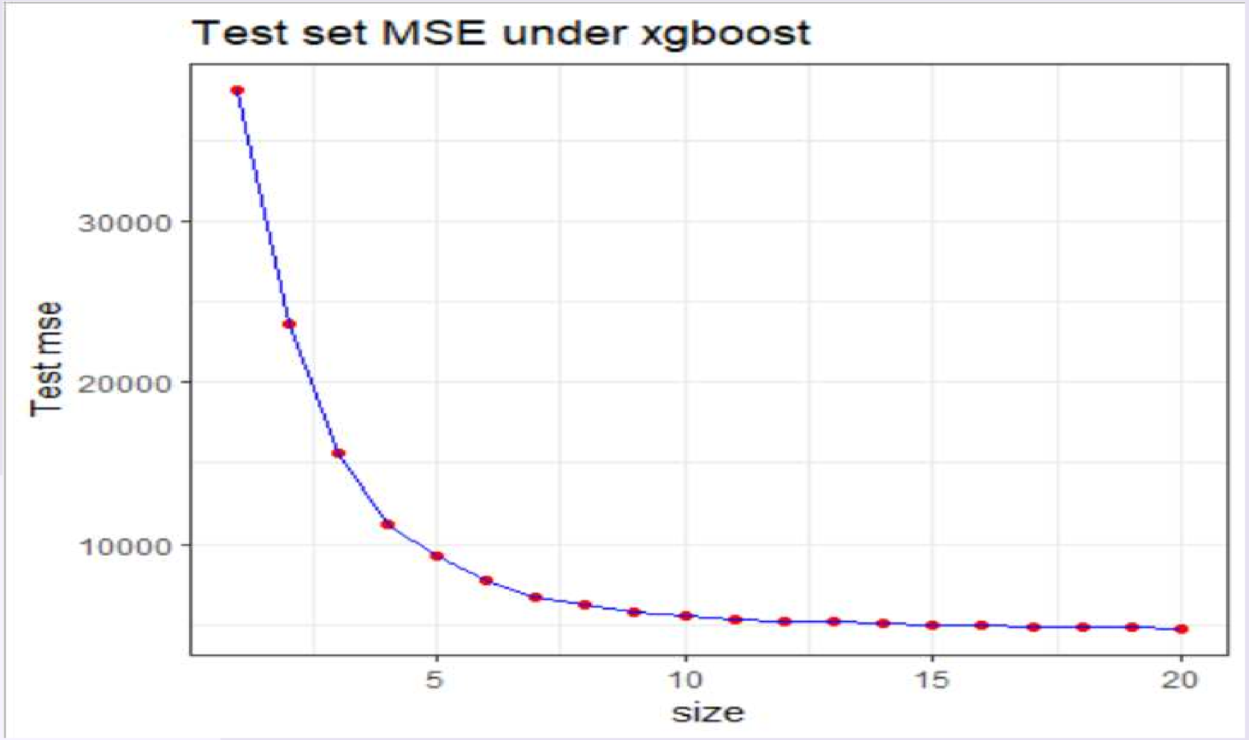
Modeling-XGboost

CNT-Training Set



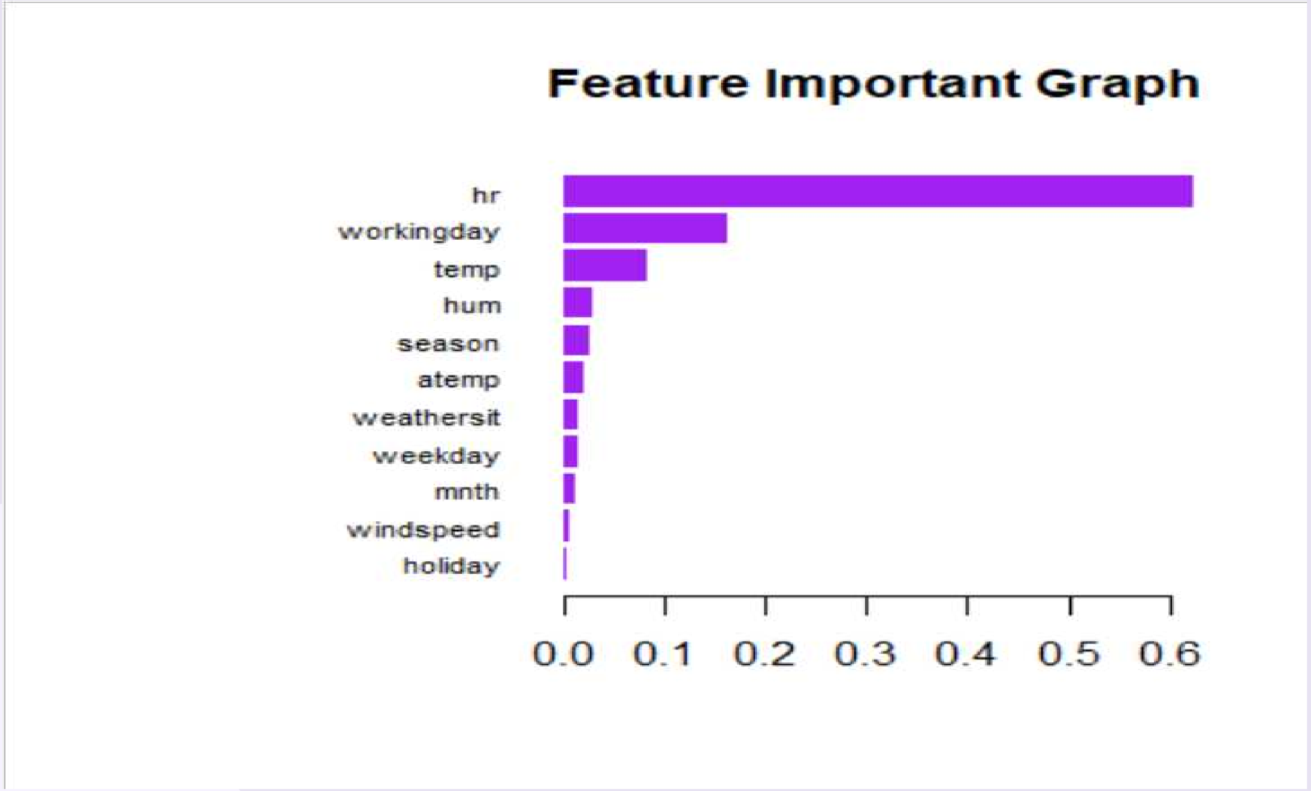
Modeling-XGboost

CNT-Test Set



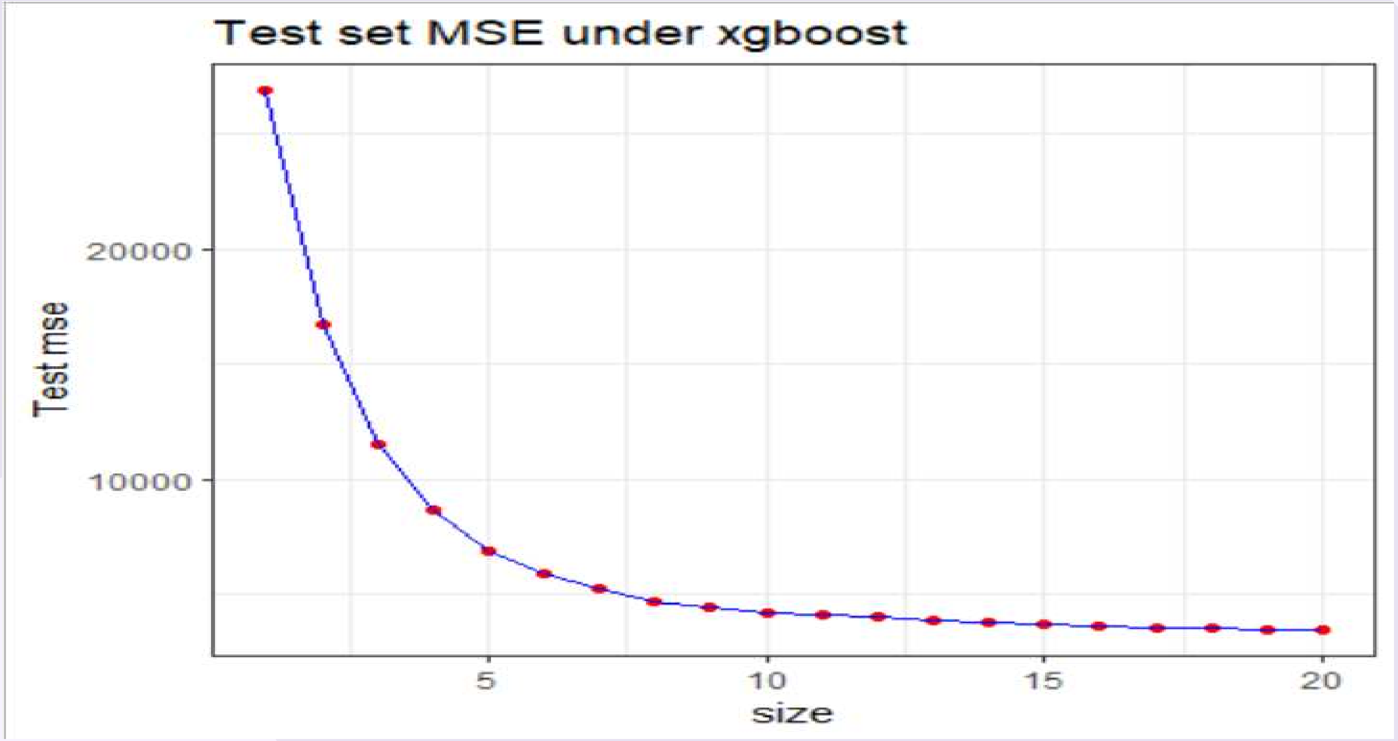
Modeling-XGboost

CNT-Registered-Training Set



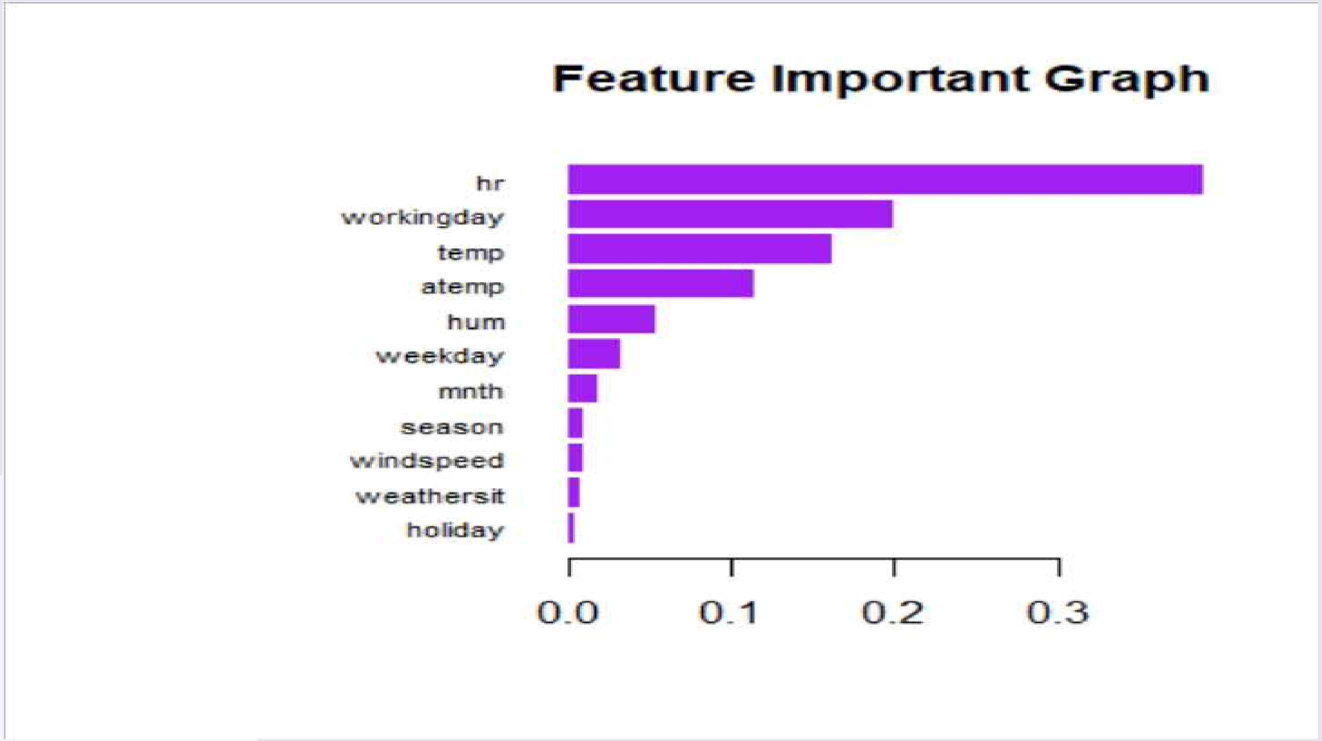
Modeling-XGboost

CNT-Registered-Test Set



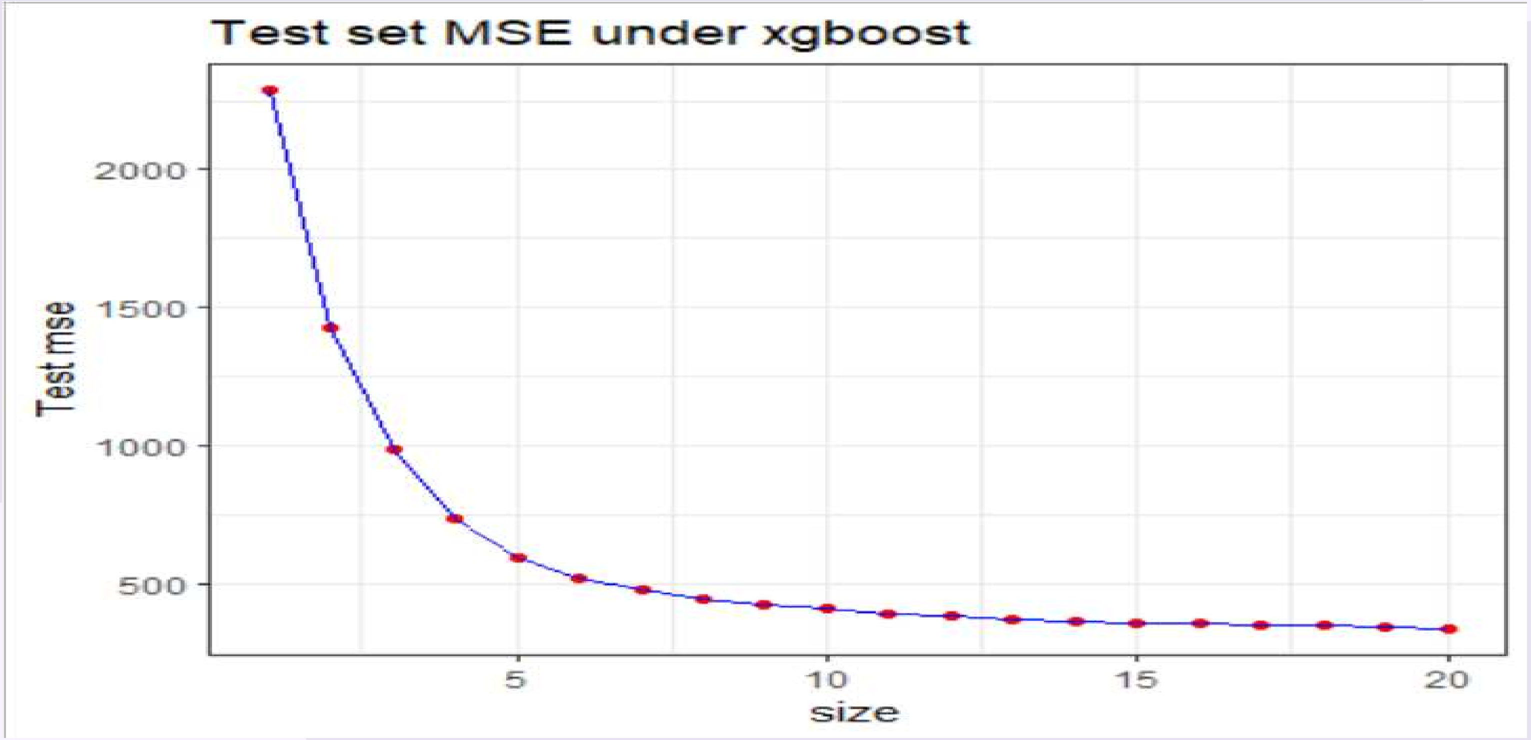
Modeling-XGboost

CNT-Casual-Training Set



Modeling-XGboost

CNT-Casual-Test Set



06

Evaluation

Evaluation

CNT

Model	Minimized MSE	Maximized Adjusted R Square	Minimized RMSE
Linear Regression	43,807	35%	145
Best Subsets Regression	21,730	16%	147
Random Forest	4,748	86%	69
XGboost	4724	85%	68

Evaluation

Registered-CNT

Model	Minimized MSE	Maximized Adjusted R Square	Minimized RMSE
Linear Regression	28,719	28%	128
Best Subsets Regression	16,186	14%	127
Random Forest	3,439	85%	58
XGboost	3478	85%	59

Evaluation

Casual-CNT

Model	Minimized MSE	Maximized Adjusted R Square	Minimized RMSE
Linear Regression	3,565	46%	35
Best Subsets Regression	1438	21%	38
Random Forest	307	88%	17
XGboost	339	87%	18

07

Conclusion

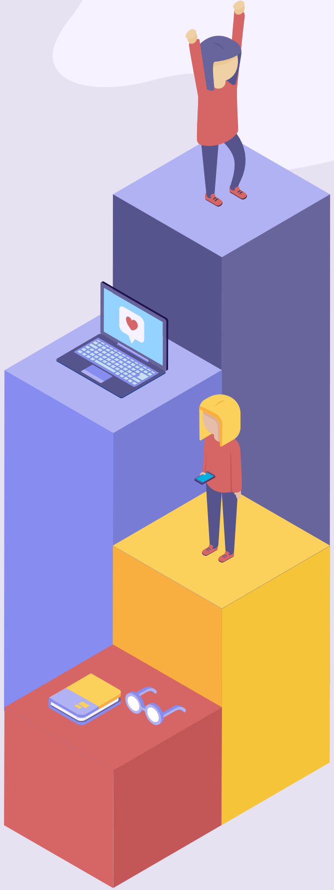
Conclusion

Best Model

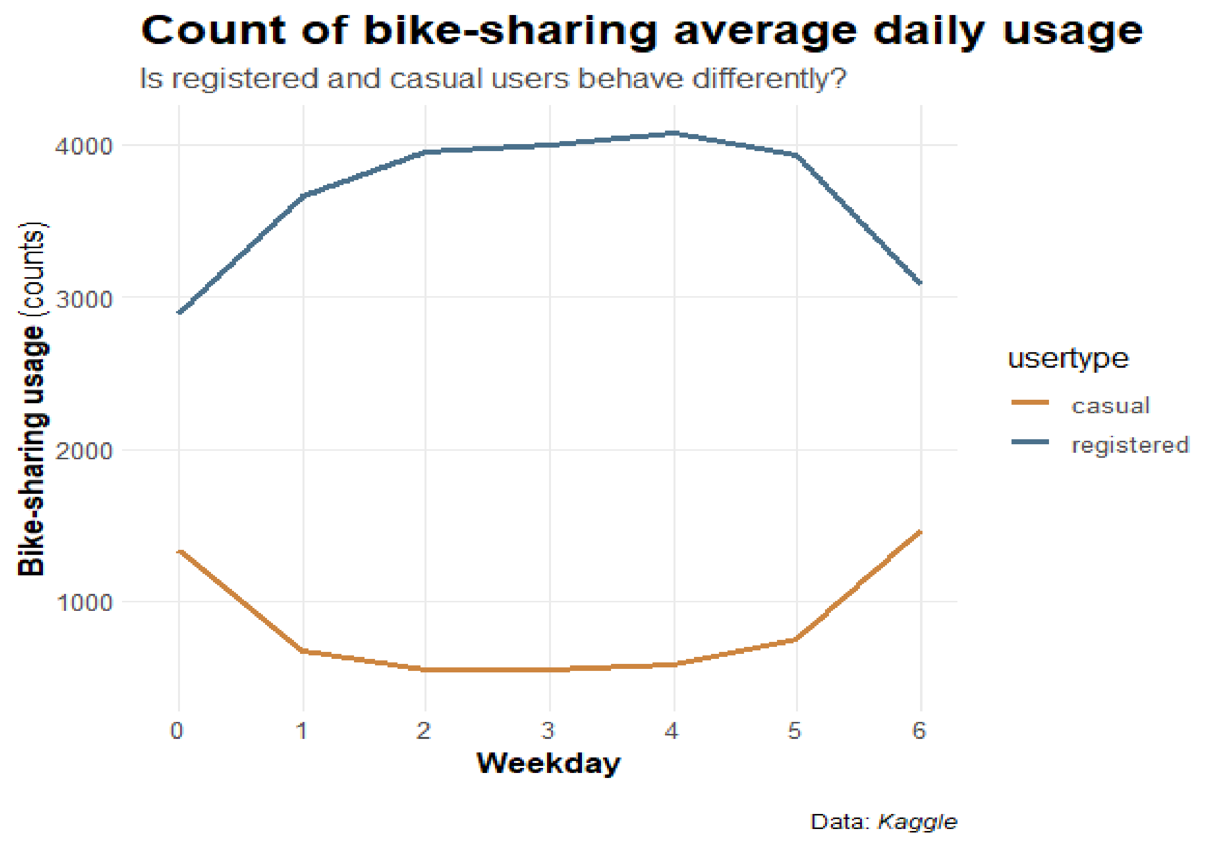
Xgboost!

Key Factors

Hour,Working day,Temperature,Humidity,Season



Some data explorations



08

Business Recommendati ons

Business Recommendations

Provide more discount packages for causal on Weekdays and registered at Weekends

Put the bike near the working place and dwelling place for the registered users

Carefully choose bike type and build SEO and Google Marketing Analytics to attract customers

Build a large AI Platform and Database to improve Customer experience

THANKS!

Do you have any questions?

Add your email at contact_us@capitalbikeshare.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution