KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics

Seoul Bike Rental – Urban Mobility

Wednesday Section

Team 2

Team Members: Silvy Saint-Jean, YangYang Li, Yen Chun Lin

Last updated: 11th December 2024

Deliverable Number: 4

1. Business case:

Urban mobility has become critical as cities seek sustainable and accessible transportation options. According to Statista, the growing demand for city transportation has steadily increased. In 2021, the daily rental count of Seoul bikes surged to around 87.8 thousand, up from 65 thousand the previous year; since 2015, this rate has doubled yearly. The website reports that Seoul bike rentals reached almost 41 million in 2022. Bike rental programs have become popular for residents and tourists, offering a cleaner and healthier way to navigate urban environments.

Since the COVID-19 pandemic, public transportation methods like buses and trains have declined by about 33 percent due to fears of illness. This gap in the market led many individuals in Seoul to opt for shared mobility options like bike rentals or scooters (Statista). However, maximizing the usage and profitability of bike rental services requires a comprehensive understanding of the factors that influence demand, particularly under varying weather conditions.

In addition to concerns related to COVID-19, weather conditions can significantly impact customer preferences for bike rentals. Bike rental companies should analyze rental patterns across different weather scenarios to maximize their service benefits. By understanding which conditions encourage higher bike usage and which deter it, companies can optimize bike availability—ensuring more bikes are available when demand is likely high and avoiding unnecessary fleet deployment when demand is expected to be low. This data-driven approach will help rental companies enhance customer satisfaction, improve operational efficiency, and boost profitability.

2. Business and Analytics Questions

2.1 Business question: What are the key factors that influence the rental bike count?

Our goal is to identify the primary drivers of bike rentals, with a particular focus on weather conditions. Understanding these drivers will help determine which weather conditions correlate with higher rental counts and guide promotional strategies to increase bike usage during favorable conditions.

2.2 Analysis question: Which factors have the most significant impact on rental bike count?

The analytics question aims to determine which features significantly affect the rental bike count.

3. Data Set Description

This dataset is sourced from the UC Irvine Machine Learning Repository. It contains the hourly count of public bicycles rented in the Seoul Bike Sharing System and corresponding weather data and holiday information. The dataset includes 8,760 observations across 14 variables, comprising six integer variables, one categorical variable, four continuous variables, two binary variables, and one date variable. This dataset provides a comprehensive basis for exploring the effects of weather, seasonality, and special calendar events on bike rental demand, making it well-suited for predictive modeling.

4. Descriptive Analytics

4.1-4.2 Visual and Descriptive Statistics:

This section explores and visualizes key variables influencing bike rental patterns, aiming to uncover trends and establish a foundation for predictive modeling and strategic decision-making. Our initial analysis explored key predictors that might influence bike rental patterns. We focused on three main analytical approaches: Normality and Residual Analysis, correlation analysis, and boxplot analysis.

Normality and Residual Analysis: The outcome variable, "Rented.Bike.Count," reflects daily rentals, showing a right-skewed distribution with most counts at lower levels and occasional peaks, likely influenced by seasons or weather. "Temperature" and "Humidity" have near-normal distributions, suggesting rentals occur in moderate weather. In contrast, "Wind Speed" and "Visibility" are right-skewed, indicating more rentals on calm, clear days. "Dew Point Temperature" also follows a normal pattern, aligning with mild weather conditions. "Solar Radiation" is right-skewed, showing rentals occur under moderate sunlight. "Rainfall" and "Snowfall" are skewed toward zero, confirming fewer rentals in adverse weather. These trends highlight the preference for bike rentals during mild and comfortable conditions.

Correlation Analysis: The correlation analysis focuses on identifying high-value relationships among continuous predictors. This analysis revealed a strong correlation (0.91) between temperature and dew point temperature and a moderate negative correlation (-0.46) between temperature and solar radiation, suggesting a climatic effect on rental rates. Correlations among other variables were within normal ranges, providing additional insights into variable interdependencies.

Boxplot Analysis: For categorical predictors, we used boxplot analysis to assess the impact of variables such as "Seasons," "Holiday," and "Functioning_Day" on bike rentals. The study showed that both "Seasons" and "Holiday" significantly impact bike rentals, while "Functioning_Day" displays substantial variation, as expected, since it reflects days when rentals were actively available. The boxplots reveal that rentals tend to be higher in spring and summer compared to winter, with lower counts on holidays, highlighting the seasonal and holiday influences on bike rental patterns.

4.3 Data Pre-Processing and Transformations:

After importing and cleaning the data, we refined the dataset by removing unnecessary variables, such as "Date," which held no predictive value. To address the non-normal distribution of "Wind_speed," we applied a log transformation to improve normality and manage zero values. "Dew Point Temperature" was excluded due to its high correlation with "Temperature" and a Variance Inflation Factor above 10, indicating potential multicollinearity issues. Additionally, categorical variables, including "Seasons," "Holiday," and "Functioning_Day," were set as categorical types, enabling ANOVA and other relevant analyses. These steps improved the dataset's suitability for modeling by reducing multicollinearity and aligning predictor distributions with normality assumptions where possible.

5. Modeling Methods and Model Specifications

5.1 Initial Set of Predictors

We selected continuous and categorical predictors to identify the factors influencing bike rentals. Continuous predictors included temperature, humidity, log-transformed wind speed, visibility, dew point temperature, solar radiation, rainfall, and snowfall to capture weather conditions affecting bike rentals. Categorical predictors included "Seasons," "Holiday," and "Functioning_Day."

5.2 Initial OLS Modeling

We performed two OLS regressions with these predictors: the first (OLS_1) was a model including only weather variables, and the second, OLS_2, refined the selection to include Hour, Temperature, Humidity, log_wind_speed, Solar_Radiation, Rainfall, Snowfall, Seasons, Holiday, and Functioning_Day. An ANOVA test confirmed that OLS_2 is a significantly better model, highlighting the selected predictors' more substantial explanatory power in capturing bike rental patterns.

5.3 OLS Assumptions Tested

We tested the assumptions of the refined model (OLS_2). The residuals vs. fitted values plot confirmed that the linearity assumption (LI) was satisfied, showing no significant non-linear patterns. The normality of residuals (EN) was approximately met, with the QQ plot and histogram indicating an essentially normal distribution except for minor deviations in the tails. The multicollinearity levels (XI) were acceptable, as the condition index was below the threshold of 30, and the highest VIF value was 5.05 for "Seasons," which is within tolerable limits. However, the homoskedasticity assumption (EV) was violated, as shown by a significant Breusch-Pagan test (p-value < 2.2e-16), indicating heteroskedasticity. Additionally, the independence of residuals (OI & EI) was violated, with the Durbin-Watson test (DW = 0.50738, p-value < 2.2e-16) confirming positive autocorrelation. Despite these issues, the mean of the residuals was effectively zero (7.34e-14), satisfying the error average assumption (EA). A Weighted Least Squares (WLS) model and the inclusion of lagged variables for autocorrelation correction will be necessary to address the violated assumptions.

5.4 Model Specifications Evaluated (and Variable Selection)

The first model specification in this exercise was OLS_2, which included the initial set of 10 predictors mentioned in 5.1. The second model specification applied a Weighted Least Squares (WLS) regression with a lagged outcome variable to address heteroskedasticity and autocorrelation issues identified in OLS_2. Variable selection for WLS was refined through statistical significance and model fit, resulting in 11 impactful predictors: Hour, Temperature, Humidity, log-transformed Wind Speed, Solar Radiation, Rainfall, Snowfall, Seasons, Holiday, Functioning_Day, and the lagged rental bike count. All predictors in this specification were found to be significant at the 0.05 level, highlighting their critical role in explaining bike rental demand.

5.5 Methods Evaluated

First, we examined a Weighted Least Squares (WLS) model with a lagged outcome variable using the specified predictors to address heteroskedasticity and autocorrelation issues. Next, we applied ridge and LASSO regression to OLS_1 and OLS_2 to shrink the models and reduce dimensionality. Finally, for both OLS_1 and OLS_2, we implemented a random forest model to evaluate improvements in predictive accuracy. All above models add the lagged rental bike count as a predictor. By comparing the MSE values among these models, we aim to determine the most optimal model and identify the factors that significantly influence the number of rented bikes.

5.6 Cross-Validation Testing:

We used 10-fold cross-validation to evaluate various combinations of models, ensuring the results were reliable and consistent. For the WLS models, after addressing assumption violations, the MSE was 221,467.6 for the model with only weather variables (OLS_1) and 189,182.8 for the model after variable selection and adding holiday variables (OLS_2). Ridge regression produced MSEs of 74,349.4 for OLS_1 and 72,284.6 for OLS_2. Random forest performed significantly better, with MSEs of 8,872.3 for OLS_1 and 5,472.3 for OLS_2. The LASSO model, ranking second overall, had MSEs of 71,770 for OLS_1 and 69,870.9 for OLS_2. These results guided our decision to balance accuracy and interpretability when selecting the final model.

5.7 Final Method/Specification Selected:

Although the random forest model with the small specification achieved higher predictive accuracy, we prioritized interpretability to meet better the needs of our primary audience—bike rental company executives and customers. As a result, we selected the LASSO model with the OLS_2, which also had the lowest MSE (69870.9).

6. Analysis of Results:

For the best model, OLS_2 with LASSO shrinkage, variables with high absolute coefficients are considered the most influential. These variables have the most significant impact on the outcome. Functioning_Day, Season, and solar radiation significantly influence the rented bike count.

Quantitative predictors: The quantitative variables that positively impact the rented bike count are Hour, Temperature, log wind speed and snowfall. Although the absolute value of the coefficient is not high, it is still significant.

On average and holding everything else constant, when log wind speed goes up by one unit, the expected rented bike count will increase by 27.049.

On average and holding everything else constant, when temperature goes up by one unit, the expected rented bike count will increase by 4.861.

On average and holding everything else constant, when snowfall goes up by one unit, the expected rented bike count will increase by 3.704.

On average and holding everything else constant, when solar radiation goes up by one unit, the

expected rented bike count will increase by 32.112.

The quantitative variables negatively impacting the rented bike count are humidity and rainfall. The coefficients for humidity and rainfall are –1.122 and –9.261, respectively, which are low absolute values, suggesting that humidity and rainfall have a week negative influence on the rented bike count, but it is still significant.

On average and holding everything else constant, when humidity goes up by one unit, the expected rented bike count will decrease by 1.122.

On average and holding everything else constant, when rainfall goes up by one unit, the expected rented bike count will decrease by 9.261.

Categorical predictors: The categorical variable Functioning Day significantly positively impacts the rented bike count. The coefficient for Functioning day(Yes) is 217.367, which is a high absolute value, suggesting that people are more likely to rent a bike during functioning days.

On average and holding everything else constant, the predicted rented bike count on functioning days would be 217.367 units higher than not functioning days.

The categorical variable Seasons significantly negatively impact the rented bike count. The spring, summer, and winter coefficients are –36.233, -36.755, and –68.949, respectively. This means that, on average, the rented bike count decreases by approximately 36.233, 36.755, and 68.949 bikes during spring, summer, and winter compared to the baseline season (autumn), holding all other variables constant.

The negative coefficients for all Seasons variables suggest that autumn likely has the highest rented bike count of the year. Winter has the most significant negative coefficient, indicating the most substantial drop in rented bike counts compared to autumn—spring and summer show smaller but notable decreases.

On average, and holding everything else constant, the coefficient for spring, summer, and winter indicates that the rented bike count is predicted to decrease by 36.233, 36.755, and 68.949 units, respectively, compared to the baseline season (Autumn).

7. Conclusions and Lessons Learned

7.1 Conclusions from the Analysis.

According to the best model, OLS_2 with shrinkage by LASSO, the important factors that might influence the rented bike count include functioning day, and seasons. Functioning day is the most influential factor affecting rented bike count, suggesting that customers are more likely to rent a bike which is a functioning day compared to not functioning day. Season is the second most influential factor affecting rented bike count, suggesting that seasonal variations significantly impact bike rentals. The Season’s coefficients represent each season's impact compared to the baseline (autumn). The rented bike count decreases in spring, summer, and winter compared to autumn. This result aligns with the assumption that seasonal factors, such as weather conditions, may reduce bike rentals in winter

(due to cold and snow) and possibly in summer (due to extreme heat).

These analysis results can help managers develop effective business strategies. Since [functioning day and autumn is the peak for bike rentals, additional bikes can be made available to meet customer demand during this time.

7.2 Project Issues, Challenges, and Lessons Learned

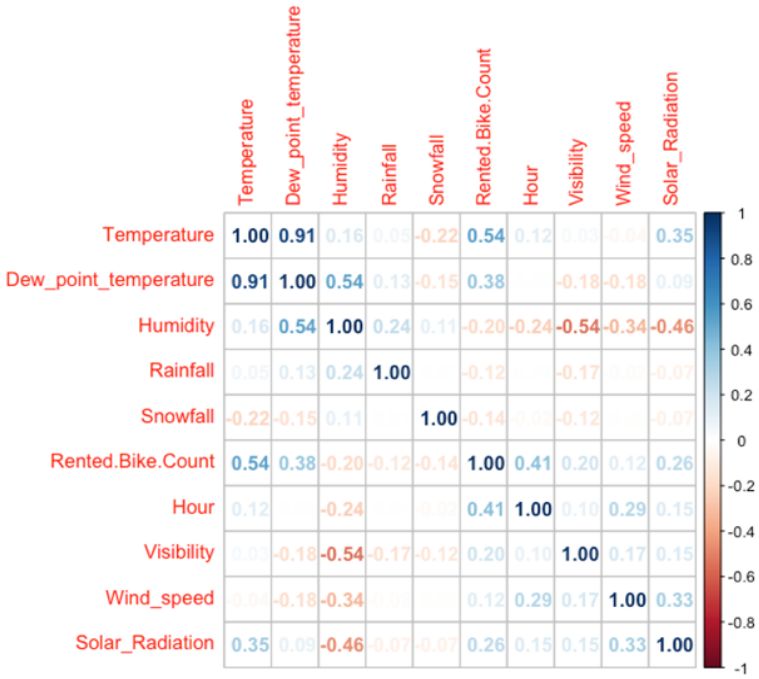
One challenge was variable selection due to the dataset's complexity and the inclusion of numerous weather-related predictors. Some variables, such as "Dew Point Temperature" and "Temperature," had similar meanings and showed overlapping explanatory power, leading to redundancy in the model. Determining how to balance accuracy with interpretability during variable selection became critical. By analyzing correlations and multicollinearity metrics, we decided to exclude "Dew Point Temperature," as it added little unique explanatory value compared to "Temperature." This approach streamlined the model and improved its clarity for stakeholders.

A key lesson learned was the need to balance model accuracy with interpretability. While the random forest model achieved the highest predictive accuracy, its complexity made it less suitable for communicating insights to stakeholders. Choosing the LASSO model as the final specification reflected the value of simplicity and interpretability for practical applications. Moreover, the project underscored the critical role of cross-validation in ensuring model reliability and the necessity of tailoring analyses to align with the intended audience's needs.

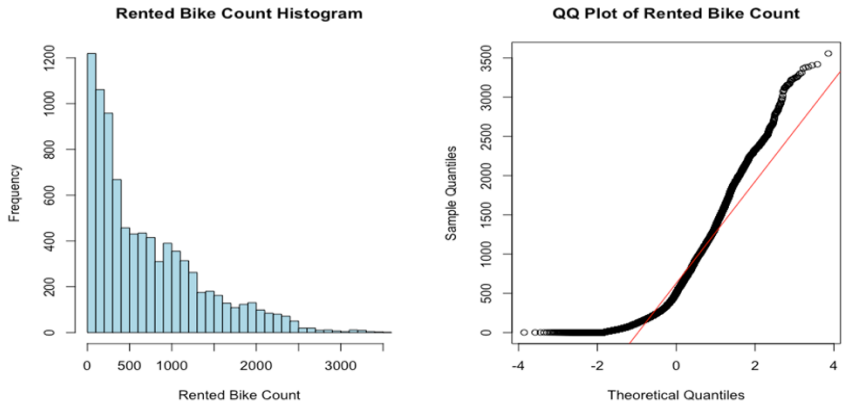
Appendix Contents

1. Descriptive Statistic:

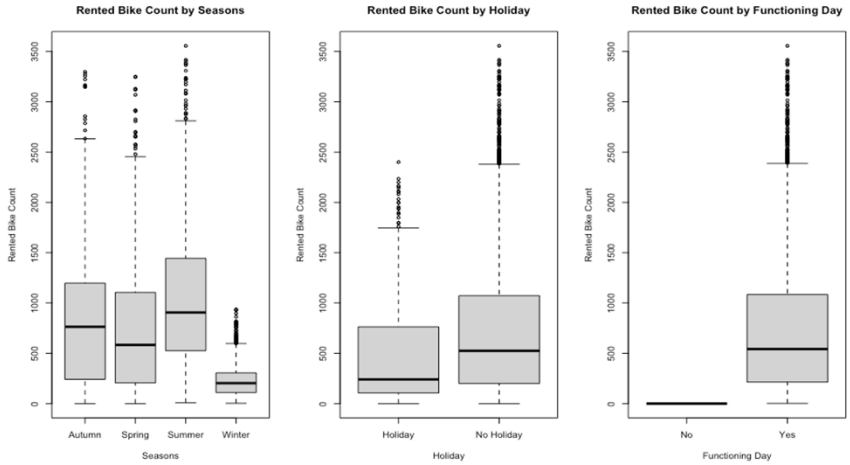
Correlation: a corrplot illustrates the correlations in our initial data set:



Normality and Residual Analysis: the outcome variable shows a right-skewed distribution

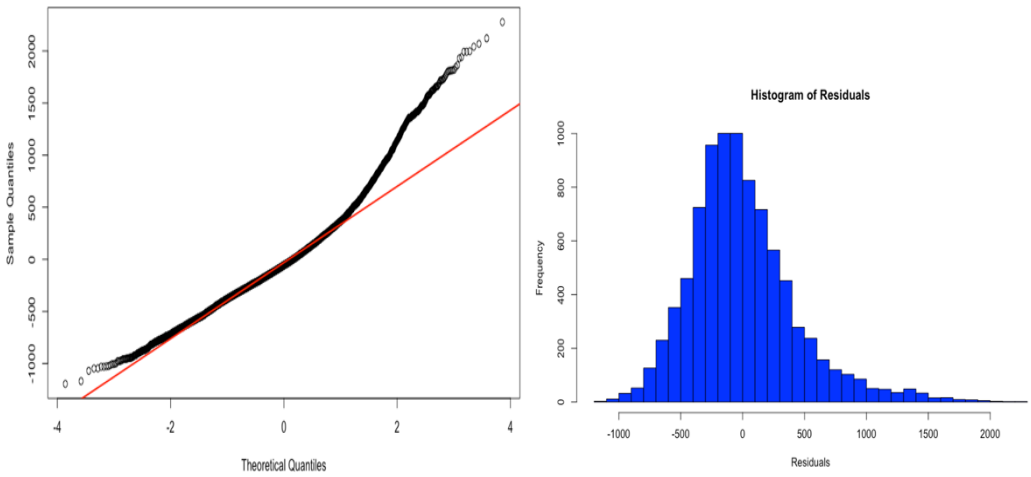


Boxplot Analysis: categorical predictors. Mention in 4.1-4.2



2. OLS Assumption Tests

Normality of Residuals: Mentioned in 5.3



Breusch-Pagan test for Homoskedasticity Check: Mentioned in 5.3

```
# Breusch-Pagan test for constant variance
library(lmtest)
bp_test <- bptest(ols_2)
print(bp_test)
```

```
    studentized Breusch-Pagan test

data:  ols_2
BP = 813.08, df = 12, p-value < 2.2e-16
```

Multicollinearity Diagnostics: Mentioned in 5.3

```
    Eigenvalue Condition Index
1  7.267938458        1.000000
2  1.516308817        2.189333
3  1.051549004        2.629002
4  1.005549416        2.688462
5  0.796226399        3.021255
6  0.636723457        3.378549
7  0.263658098        5.250311
8  0.198726092        6.047532
9  0.099513001        8.546055
10 0.068314634       10.314513
11 0.049373424       12.132743
12 0.036768745       14.059382
13 0.009350455       27.879774
```

```
# Variance Inflation Factor (VIF)
vif_ols_2 <- vif(ols_2)
print(vif_ols_2)
```

```
                   GVIF Df GVIF^(1/(2*Df))
Hour           1.208327  1        1.099239
Temperature    5.031924  1        2.243195
Humidity       1.811552  1        1.345939
log_wind_speed 1.310297  1        1.144682
Solar_Radiation 1.854782  1        1.361904
Rainfall       1.068913  1        1.033882
Snowfall       1.112992  1        1.054985
Seasons        5.054033  3        1.310005
Holiday        1.022780  1        1.011326
Functioning_Day 1.079337  1        1.038912
```

Durbin-Watson test for autocorrelation: Mentioned in 5.3

```
    Durbin-Watson test

data:  ols_2
DW = 0.50738, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Mean of Residuals Test: Mentioned in 5.3

```
mean_residuals <- mean(residuals(ols_2))
print(mean_residuals)
```

```
[1] 7.341626e-14
```

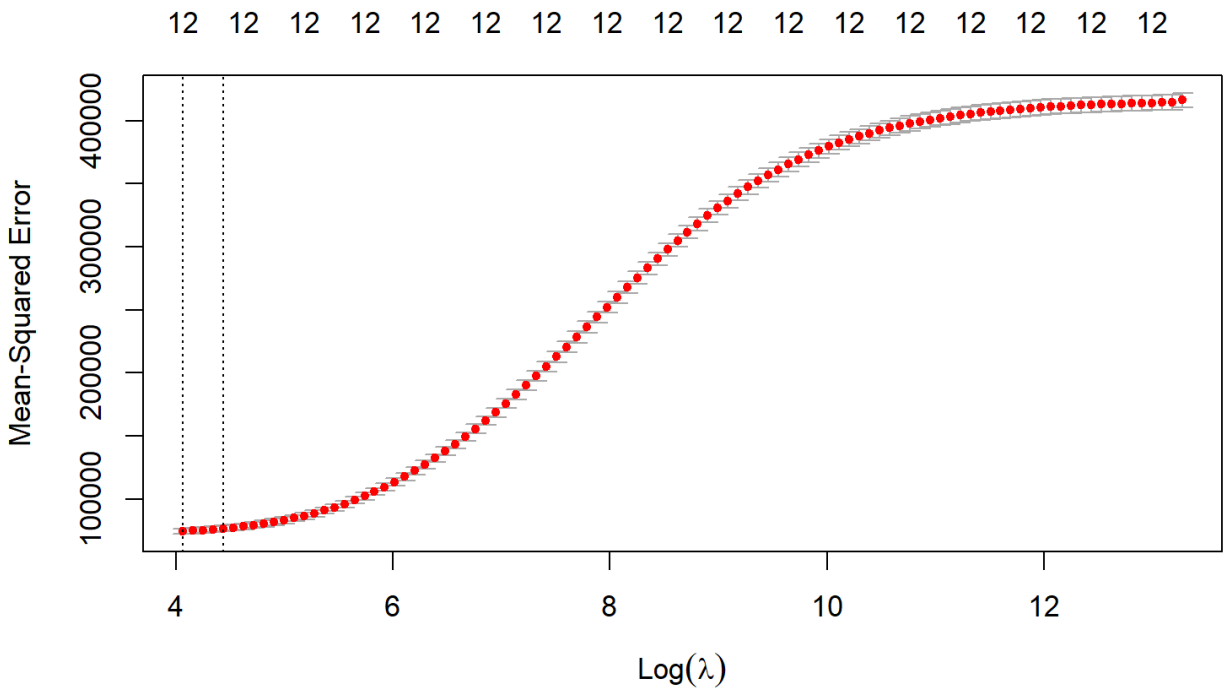
3. Variable Selection for WLS lagged Model

```
Call:
lm(formula = Rented.Bike.Count ~ Lagged_Rented_Bike_Count + Hour +
    Temperature + Humidity + log_wind_speed + Solar_Radiation +
    Rainfall + Snowfall + Seasons + Holiday + Functioning_Day,
    data = data_clean, weights = weights_clean)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.0137 -0.9980 -0.9936  0.9966  3.4580

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)               -3.252e+02  4.529e+00  -71.802  < 2e-16 ***
Lagged_Rented_Bike_Count   2.512e-03  5.623e-04    4.467 8.02e-06 ***
Hour                       2.701e+01  2.821e-02  957.476  < 2e-16 ***
Temperature                2.638e+01  3.496e-02  754.728  < 2e-16 ***
Humidity                  -8.166e+00  9.974e-03 -818.784  < 2e-16 ***
log_wind_speed             7.269e+01  4.644e-01  156.526  < 2e-16 ***
Solar_Radiation           -8.549e+01  2.623e-01 -325.964  < 2e-16 ***
Rainfall                  -5.997e+01  1.016e-01 -590.331  < 2e-16 ***
Snowfall                   2.881e+01  3.412e-01   84.429  < 2e-16 ***
SeasonsSpring             -1.426e+02  3.017e-01 -472.738  < 2e-16 ***
SeasonsSummer             -1.503e+02  5.202e-01 -288.886  < 2e-16 ***
SeasonsWinter             -3.720e+02  4.286e-01 -867.868  < 2e-16 ***
HolidayNo Holiday          1.166e+02  1.056e+00  110.478  < 2e-16 ***
Functioning_DayYes         9.264e+02  4.396e+00  210.734  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.999 on 8745 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.088e+06 on 13 and 8745 DF,  p-value: < 2.2e-16
```

4. 10F CV for 8 models

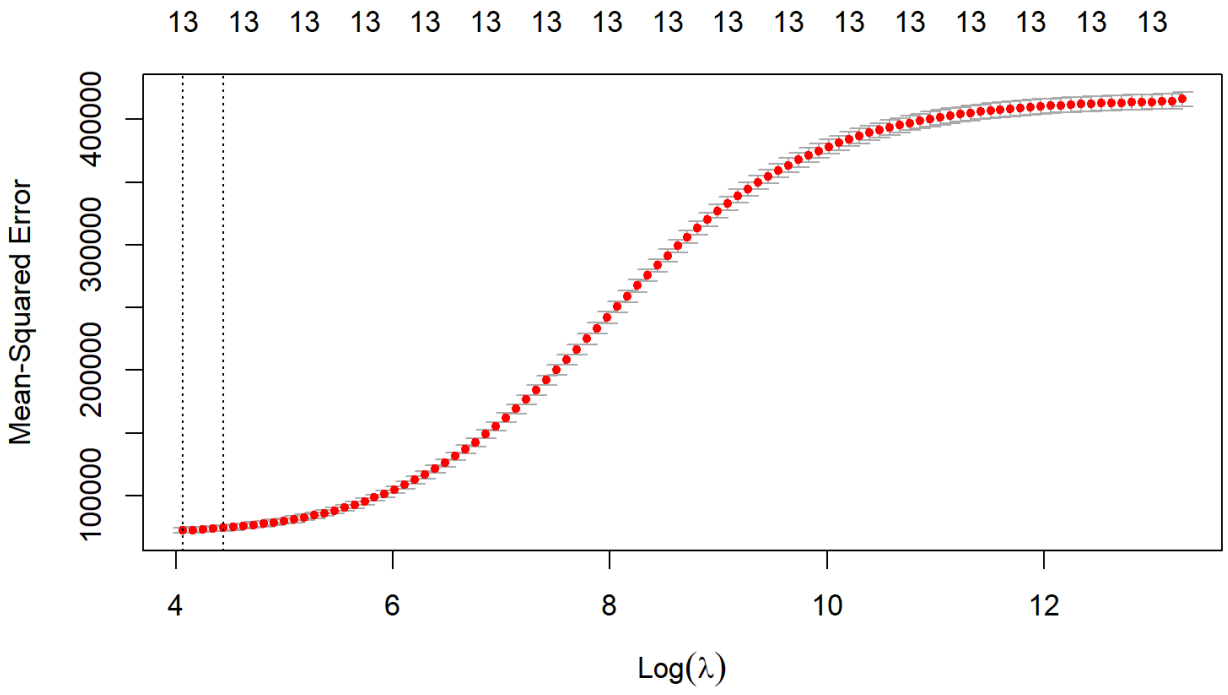
10F CV Ridge for OLS_1

```
      Best Lambda Best Log Lambda Best 10FCV
[1,]      58.271           4.065   74349.45
```



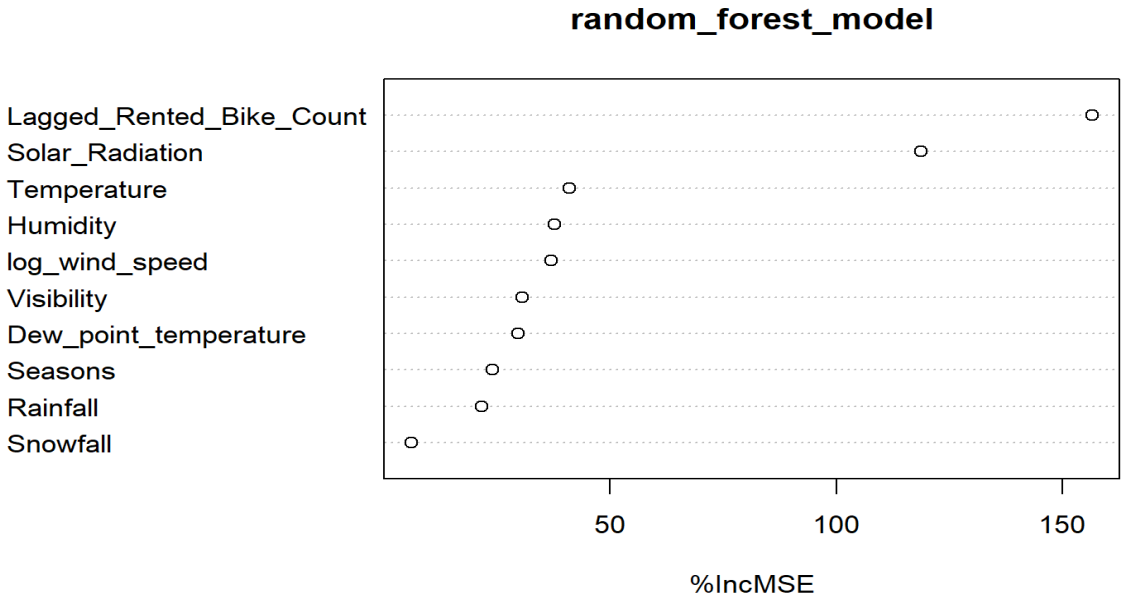
10F CV Ridge for OLS_2

	Best Lambda	Best Log Lambda	Best 10FCV
[1,]	58.271	4.065	72284.64



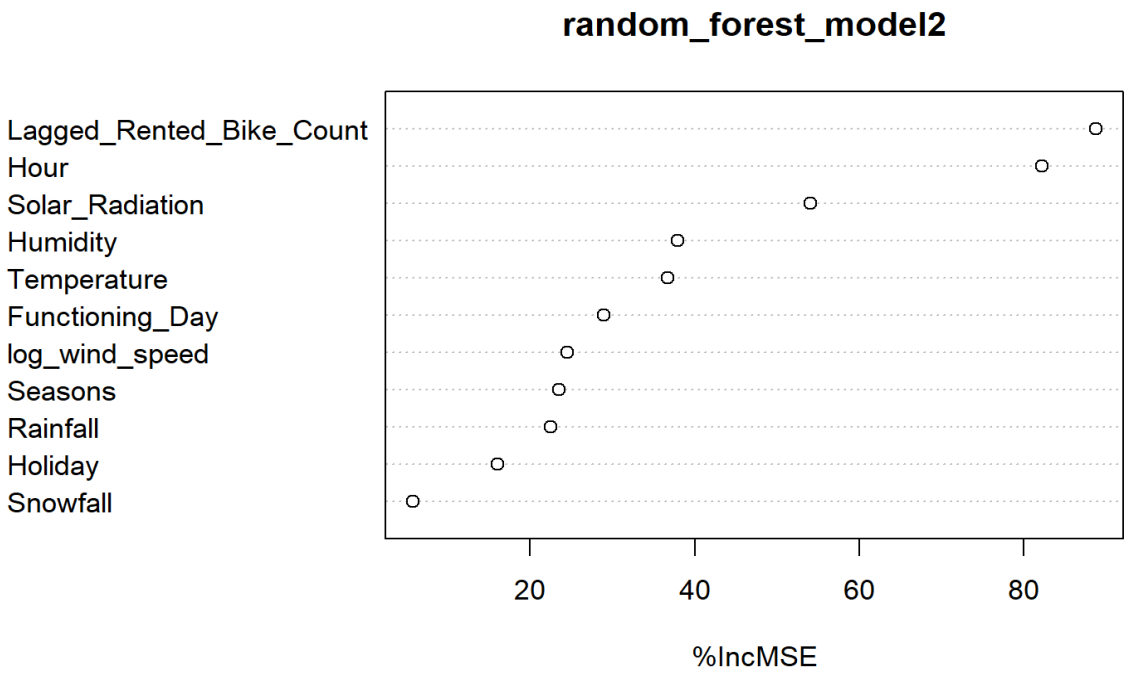
Random Forest for OLS_1

```
Mean Squared Error: 8872.253
RMSE: 94.19264
                           %IncMSE
Lagged_Rented_Bike_Count 156.566709
Temperature               41.048565
Humidity                  37.798442
log_wind_speed            37.059735
Visibility                30.581955
Dew_point_temperature     29.778334
Solar_Radiation          118.734197
Rainfall                  21.598157
Snowfall                   6.153791
Seasons                   23.951185
```



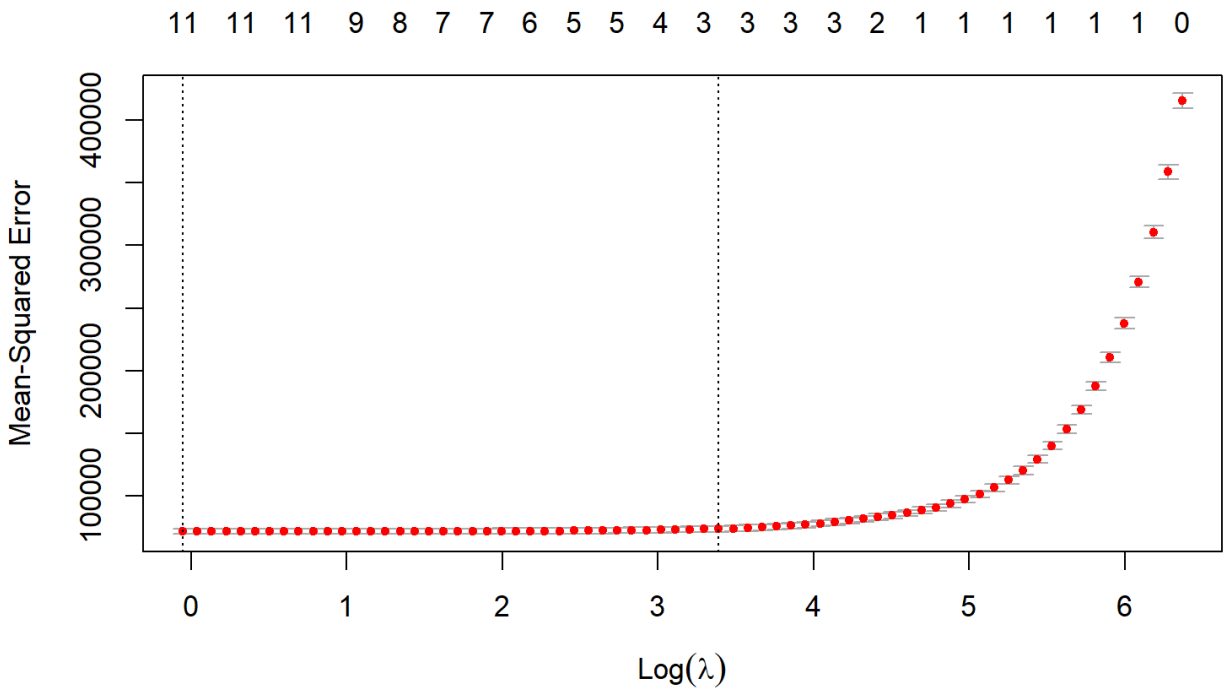
Random Forest for OLS_2

```
Mean Squared Error: 5472.322
RMSE: 73.97514
                           %IncMSE
Lagged_Rented_Bike_Count 88.770842
Hour                     82.214638
Temperature              36.646137
Humidity                 37.881662
log_wind_speed           24.528919
Solar_Radiation          54.010147
Rainfall                 22.521819
Snowfall                  5.749098
Seasons                  23.453574
Holiday                  16.064918
Functioning_Day          28.968129
```



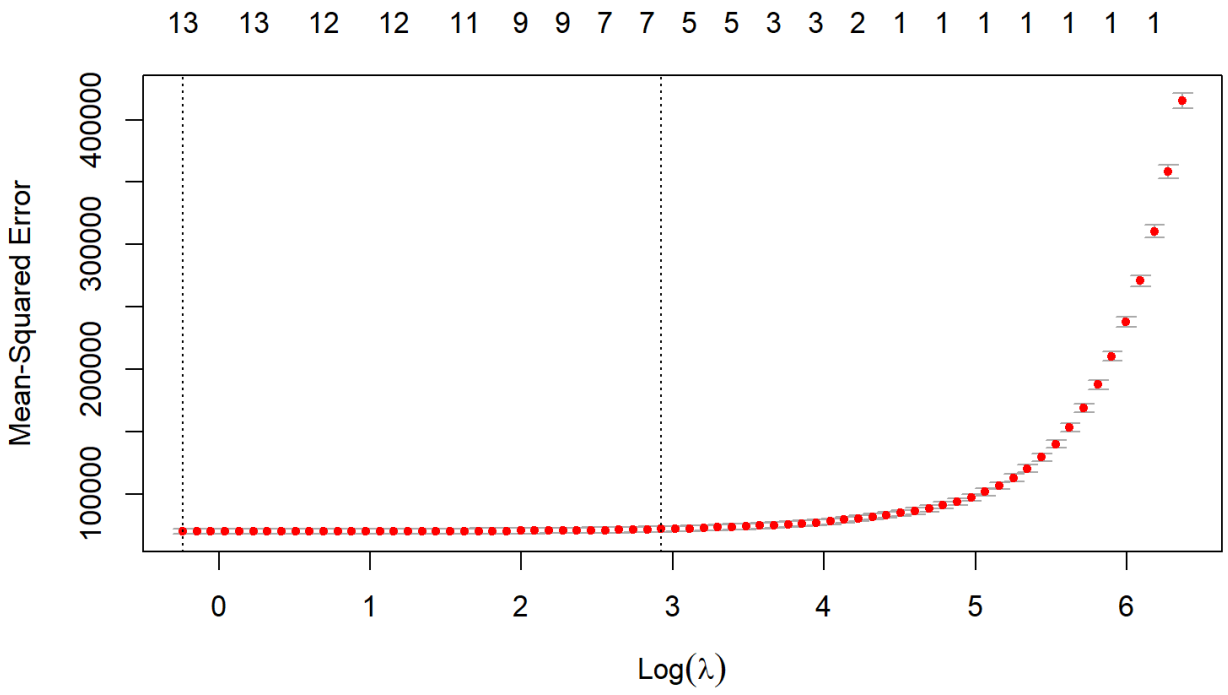
10F CV LASSO for OLS_1

```
     Best Lambda Best Log Lambda Best 10FCV
[1,]    0.949671     -0.05163966   71769.95
13 x 4 Matrix of class "dgeMatrix"
                          Best LASSO         Odds 0-Lambda LASSO         Odds
(Intercept)                   80.296 7.446504e+34         80.296 7.446504e+34
Lagged_Rented_Bike_Count       0.839 2.313000e+00          0.839 2.313000e+00
Temperature                    4.020 5.570300e+01          4.020 5.570300e+01
Humidity                      -1.009 3.650000e-01         -1.009 3.650000e-01
log_wind_speed                36.431 6.630987e+15         36.431 6.630987e+15
Visibility                     0.000 1.000000e+00          0.000 1.000000e+00
Dew_point_temperature          0.000 1.000000e+00          0.000 1.000000e+00
Solar_Radiation               35.357 2.266129e+15         35.357 2.266129e+15
Rainfall                      -5.011 7.000000e-03         -5.011 7.000000e-03
Snowfall                       5.639 2.811200e+02          5.639 2.811200e+02
SeasonsSpring                -13.339 0.000000e+00        -13.339 0.000000e+00
SeasonsSummer                -11.884 0.000000e+00        -11.884 0.000000e+00
SeasonsWinter                -30.463 0.000000e+00        -30.463 0.000000e+00
```

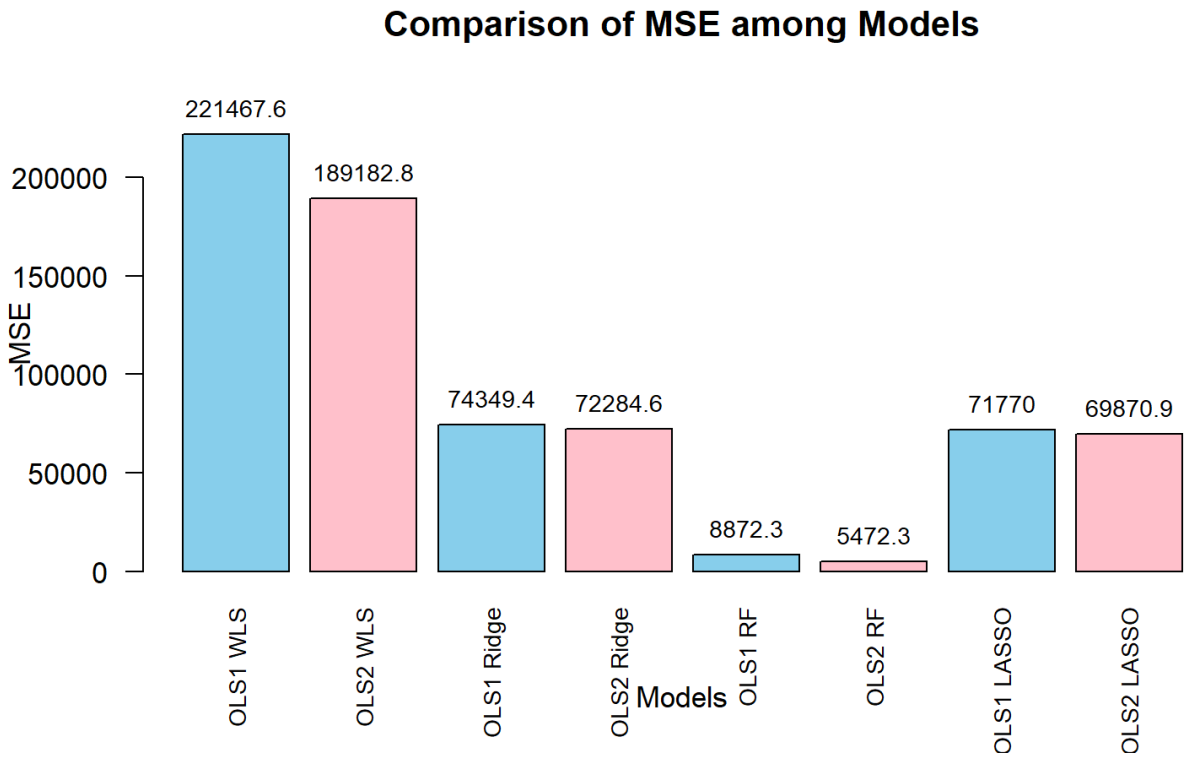


10 F CV LASSO for OLS_2

```
     Best Lambda Best Log Lambda Best 10FCV
[1,]   0.7884336      -0.2377071   69870.89
14 x 4 Matrix of class "dgeMatrix"
                          Best LASSO         Odds 0-Lambda LASSO         Odds
(Intercept)                 -142.728 0.000000e+00       -142.728 0.000000e+00
Lagged_Rented_Bike_Count       0.793 2.210000e+00          0.793 2.210000e+00
Hour                           4.375 7.941500e+01          4.375 7.941500e+01
Temperature                    4.861 1.292120e+02          4.861 1.292120e+02
Humidity                      -1.122 3.260000e-01         -1.122 3.260000e-01
log_wind_speed                27.049 5.587174e+11         27.049 5.587174e+11
Solar_Radiation               32.112 8.830377e+13         32.112 8.830377e+13
Rainfall                      -9.261 0.000000e+00         -9.261 0.000000e+00
Snowfall                       3.704 4.061600e+01          3.704 4.061600e+01
SeasonsSpring                -36.233 0.000000e+00        -36.233 0.000000e+00
SeasonsSummer                -36.755 0.000000e+00        -36.755 0.000000e+00
SeasonsWinter                -68.949 0.000000e+00        -68.949 0.000000e+00
HolidayNo Holiday             25.313 9.845242e+10         25.313 9.845242e+10
Functioning_DayYes           217.367 2.518655e+94        217.367 2.518655e+94
```



5. 10FCV MSE comparison between models



6. Coefficient of best Model- LASSO