Algorithm Comparison

Dataset

The dataset selected for above model is ‘Seoul Bike Sharing Demand Data Set’. The dataset and descriptions can be found at: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

Overview

This document contains two sections :

1. **Textbook segment**: It contains the description of 5 models that can be used to solve regression and/or classification problems.
2. **Practical segment**: It contains the process and results of using these 5 models to predict Bike demand every hour for the public in Seoul.

Textbook Segment

Below mentioned are the models selected for Analysis:-

1. Multiple Linear Regression
2. Polynomial Model
3. Penalised Regression - Lasso
4. K Nearest Neighbour
5. Generalised Linear Model - Poisson

Multiple Linear Regression

Description

Linear regression is a machine learning technique where independent[1] variable(s) are used for the prediction of a dependent[2] variable through a linear relation between them. When multiple variables are used for prediction of the response variable, it is known as multiple linear regression. The relationship between these variables can be explained by the following mathematical equation:-

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p ,$$

Where,

$\hat{Y}$ is the predicted or expected value of the dependent or response variable,

X1 through Xp are p distinct independent or predictor variables,

b0 is the value of Y when all the independent variables (X1 through Xp) are equal to zero,

and b1 through bp are the estimated regression coefficients.

Assumptions and Conditions

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. Independence: The residuals are independent. There is no correlation between consecutive residuals in time series data.
3. Homoscedasticity: The residuals have constant variance at all values of x.
4. Normality: The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

Advantages

1. One advantage is that it helps to measure the relative influence of one or more predictors on the response variable.
2. It also helps to find outliers or anomalies in the data.

Disadvantages

1. Limiting the analysis to only linear relationships is a setback as most relationships in the universe are not linear.
2. R and least square errors are not resistant to outliers.

Polynomial Model with transformation

Description

Polynomial & transformed models are an addition to linear regression models. In polynomial models, there exists an n degree relationship with the explanatory variables. i.e., where the explanatory variables are linear, but the model is complex which could be quadratic, cubic etc. A quadratic relationship will be represented like:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

where,

Y is predicted value for the response variable,

X represents explanatory variable which can be multiple,

B0 is the value of Y when all the independent variables (X1 through Xn) are equal to zero,

and B1,B2 are the estimated regression coefficients.

Transformation

It refers to the change of scale of axis for either both the explanatory & response variables or for one of them. A useful transformation is the natural log transformation to the base e. Some other transformations are square root or exponentials. We could try out several possible transformations and see which one gives the better plots.

Advantages

Most mathematical functions that satisfy reasonable conditions can be approximated by a Taylor series which is a polynomial. Therefore, it is quite reasonable to approximate an unknown function by a polynomial. A non linear transformation can help to bring out a linear relationship between the variables.

Disadvantages

The disadvantage is that the formula provides no insight and can become a rote technique. Transformations can have their own consequences. Interpretations about the model parameters and the units they are measured in are changed by transformations. In such cases when we apply one transformation to build the model, the results need to be retransformed to interpret the results in the original format. For e.g. Under root to square, and log to exponential.

Penalised Regression (Lasso Model Building)

Description

Penalised models are implemented when there is a high multicollinearity[3] Linear models can be penalised by adding a constraint in the equation for a regression model. Imposing a penalty to the coefficients is to shrink the effect of those variables that do not contribute to prediction of response variable. One such model is Lasso (Least Absolute Shrinkage and Selection Operator). Lasso shrinks the coefficients of undesired features equal to zero. Lasso Regression uses L1 regularization technique.

What is L1 regularisation?

L1 regularization adds a penalty that is equal to the absolute value of the magnitude of the coefficient. Some coefficients might become zero and get eliminated from the model. Larger penalties result in coefficient values that are closer to zero. Mathematical Representation:-

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Out of the two terms present in these equations, first is representing Residual Sum of squares. Lambda in the second term is a constant and can be used to fine-tune the amount of penalty.

Advantages

It produces simpler and more interpretable models that incorporate only a reduced set of predictors. It can avoid overfitting. It can be applied even when number of features is larger than amount of data.

Disadvantages

Lasso cannot work for a group of corelated variables, if there are two or more highly collinear variables then Lasso Regression will select one of them randomly which is not a good technique in Data Interpretation. This regularised regression can automate things and decrease the contribution of common sense and thought process of an analyst.

K Nearest Neighbour

Description

KNN is a supervised machine learning algorithm which can be used for regression as well as classification. It predicts the outcome of a new observation by comparing it to the most similar cases in the training dataset. The training dataset is plotted based on x variables. After training, test observations are also plotted on the graph based on attribute values. After plotting that test observation, it is compared to its 'K' neighbours as per its position. For regression, the value can be predicted by taking the mean value of neighbours. For classification, the class can be predicted by taking the mode of neighbours.

Now two questions arise?

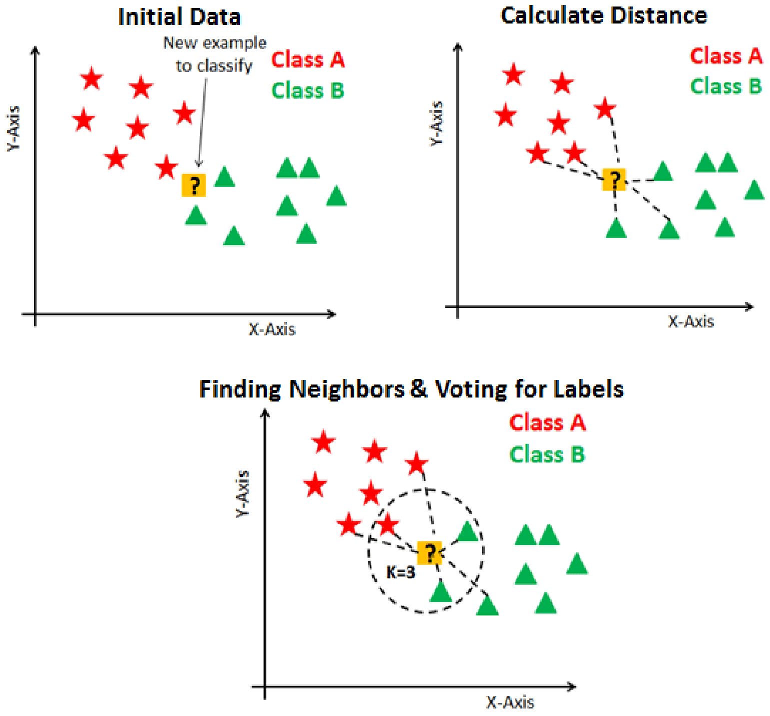
1. What should be the value of k?
2. How to calculate which observations are neighbours?

Most of the times k is taken as an odd number ranging from 1 to number of observations. Low value of K is more prone to errors. Bigger value will cost high computational cost. So, value of k shouldn't cost much and minimize the RMSE[4]. Less the value of RMSE, better is the value of K.

To find out the neighbours from the training observations some similarity measure can be used. Similarity measures such as distance metrics can be used. To calculate the distance(e.g., Euclidean distance) between the new observation and other observations and find the k observations with least distance.

Key Notes

- It is not necessary that more the value of K better will be the prediction.
- If there is a huge gap in attribute scales, then attributes should be standardised before the distance is calculated.



Generalised Linear Model - Poisson

Description

A Poisson Regression model is a Generalized Linear Model (GLM) that is used to model count data and contingency tables. The output Y (count) is a value that follows the Poisson distribution. It assumes the logarithm of expected values (mean) that can be modelled into a linear form by some unknown parameters. To transform the non-linear relationship to linear form, a link function is used which is the log for Poisson Regression. The general mathematical form of Poisson Regression model is:

$$\log(y)=\alpha + \beta1x1 + \beta2x2 + \ldots.+\beta pxp$$

where,

y: Is the response variable

α and β: are numeric coefficients, α being the intercept, sometimes α also is represented by $\beta0$, it's the same

x is the predictor/explanatory variable

Advantages

First, the Poisson model has a minimum value of 0. It will not predict negative values. This makes it ideal for a distribution in which the mean or the most typical value is close to 0. Also, Poisson model is estimated by the maximum likelihood method, the estimates are adapted to the actual data. In practice, this means that the sum of the predicted values is virtually identical to the sum of the input values, except for a very slight rounding off error.

Disadvantages

It makes strong assumptions about the underlying distribution of the data. Poisson GLM is very restricted to only one set of problems where counts is used to find probability and cannot be used vastly.

Practical Segment

The main object of this project was to predict ‘Rented Bike Count’ on a single day based on attributes available. The dataset was taken from UCI (Machine Learning Repository) which contains 14 attributes in total and around 8700 observations from the past. The dataset contains data from 1 December 2017 to 30th November 2018. The analysis & modelling is done using R programming. Most of the supporting features are related to weather such as temperature, humidity, season etc. Five models were created to predict the response variable based on desired explanatory variables. All the models were evaluated from qualitative & quantitative point of view by considering RMSE, execution time, R square value etc.

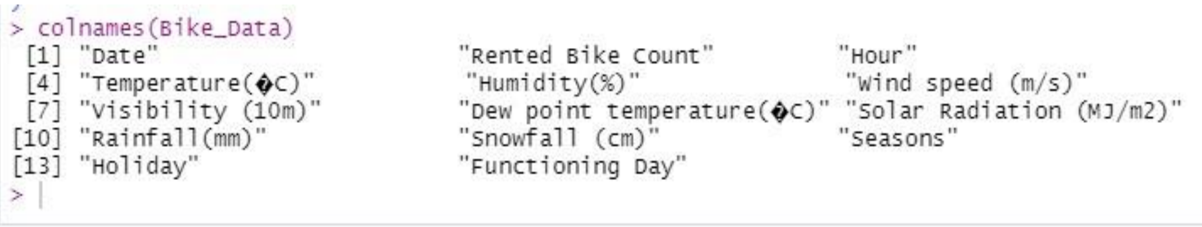
Data Preparation

Data Loading

- Using ‘read_csv’ from the package ‘tidyverse’

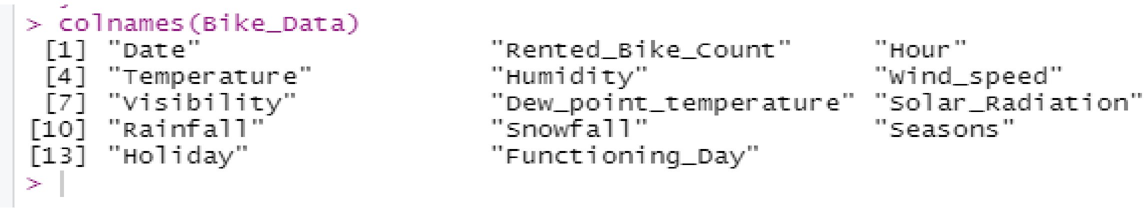
Data Cleaning

1. **Formatting column names:**
   Before:



```
> colnames(Bike_Data)
 [1] "Date"                   "Rented Bike Count"          "Hour"
 [4] "Temperature(�C)"         "Humidity(%)"                "Wind speed (m/s)"
 [7] "Visibility (10m)"       "Dew point temperature(�C)" "Solar Radiation (MJ/m2)"
[10] "Rainfall(mm)"           "Snowfall (cm)"              "Seasons"
[13] "Holiday"                "Functioning Day"
> |
```

After:



```
> colnames(Bike_Data)
 [1] "Date"                 "Rented_Bike_Count"     "Hour"
 [4] "Temperature"          "Humidity"              "Wind_speed"
 [7] "Visibility"           "Dew_point_temperature" "Solar_Radiation"
[10] "Rainfall"             "Snowfall"              "Seasons"
[13] "Holiday"              "Functioning_Day"
> |
```

2. Check for NA values:

```
+ }
[1] "There are no NA values in the dataset."
> 
```

3. Mutating Categorical variables:

Change some factor variables to numerical such as Holiday and Functional day by adding dummy values. For e.g., ‘Holiday’ assigned as 1 and ‘No Holiday’ assigned as 0.

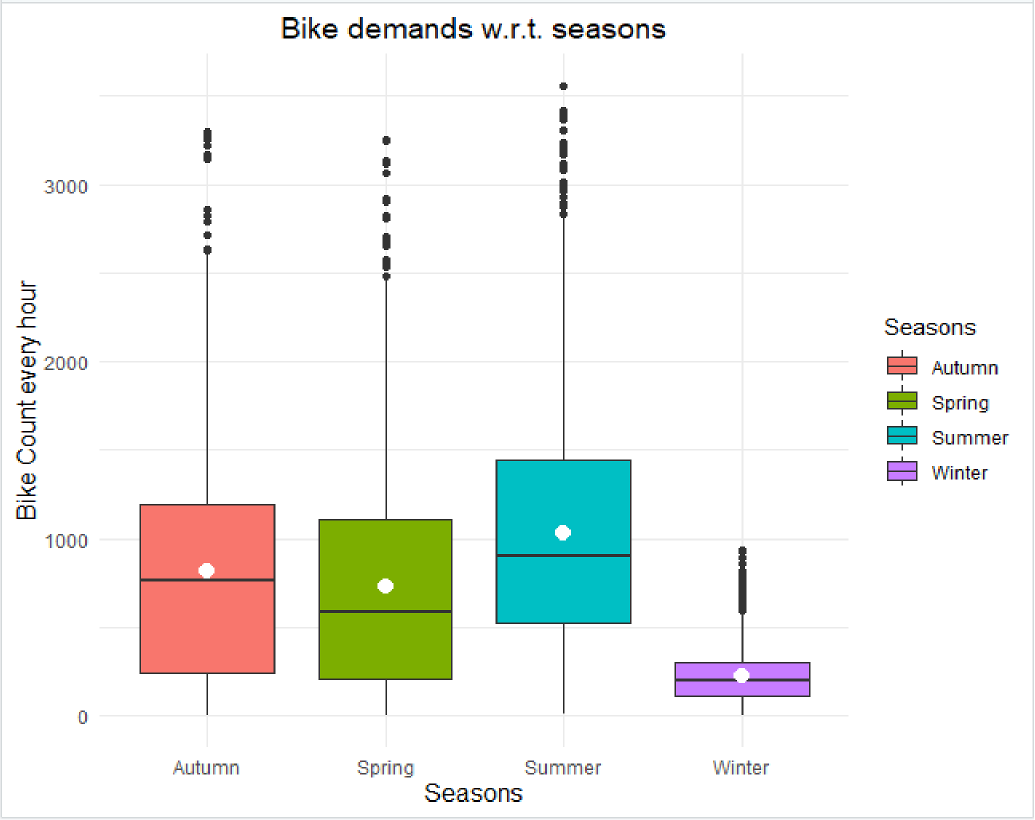
4. Removing Outliers:

Outliers are removed after analysis in the EDA part.

Exploratory Data Analysis

Effect of Seasons on the Bike Count:

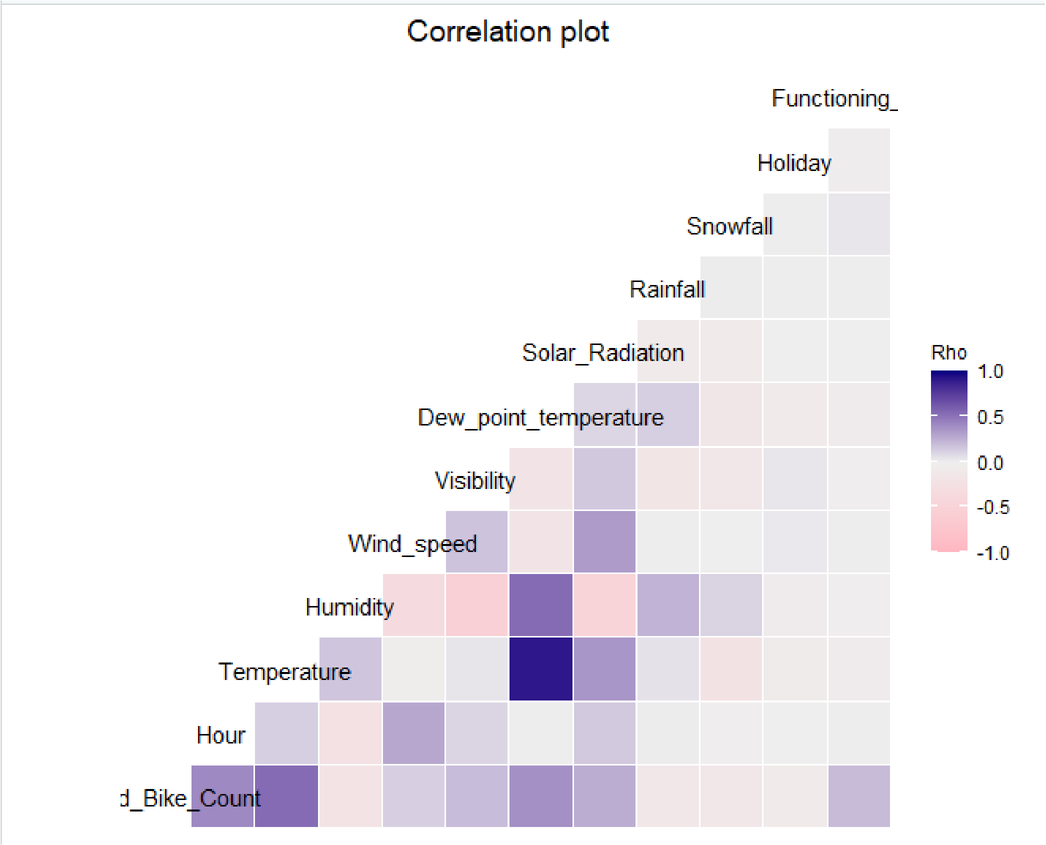
This is a Boxplot displaying the distribution of Bike count across all the seasons.



Season seems to have some major impact on predicting Bike demand for every hour. The demand is least in winter season and maximum in summer season. High number of outliers in all the seasons should be removed before modelling. Due to the outliers the mean has been deviated toward the skew in the right direction.

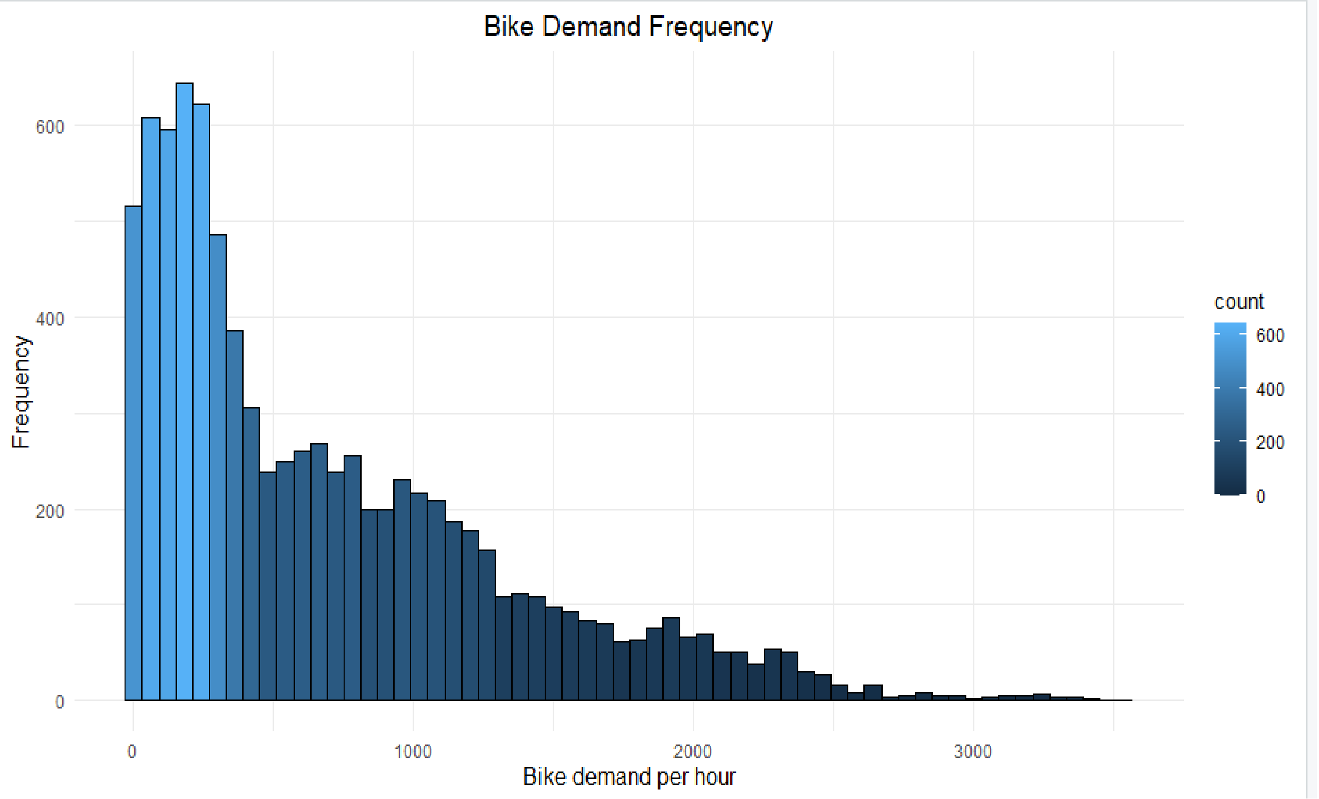
Correlation plot for Numerical variables:

This is the correlation graph displaying relations between all the features. Bike Count is in the last row displaying the relation with all other variables based on intensity of the colour in the box. Bike Count has strong relation with Hour, Temperature, Visibility, Dew point Temperature, Solar Radiation and Functioning Day.



Response variable:

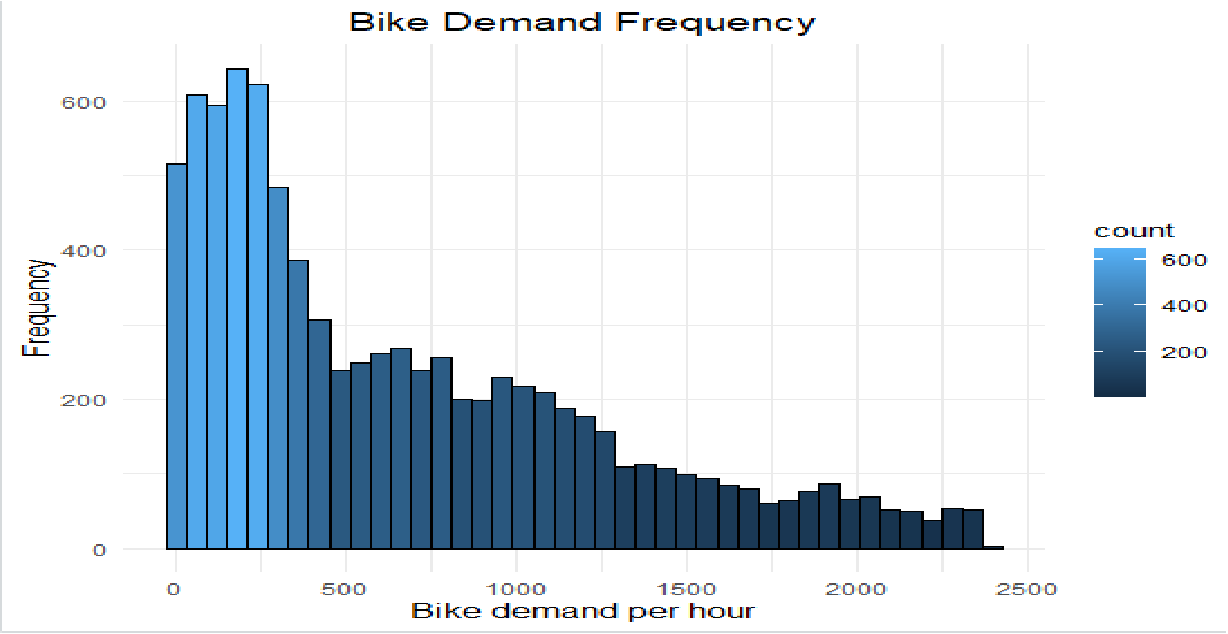
This is a histogram showing the frequent trends for response variable i.e. Rented Bike Count.



The response variable is right skewed and shows the possibility of outliers. The values range from 0 to around 3700 which is unrealistic, that the demand touched to 3700/hr. Maximum values lie between 0 to 500, so that is the demand most of the times.

Removing the Outliers

All the outliers present in the dataset based on ‘Rented_Bike_count’ are removed. Plot after removal:-



Modelling

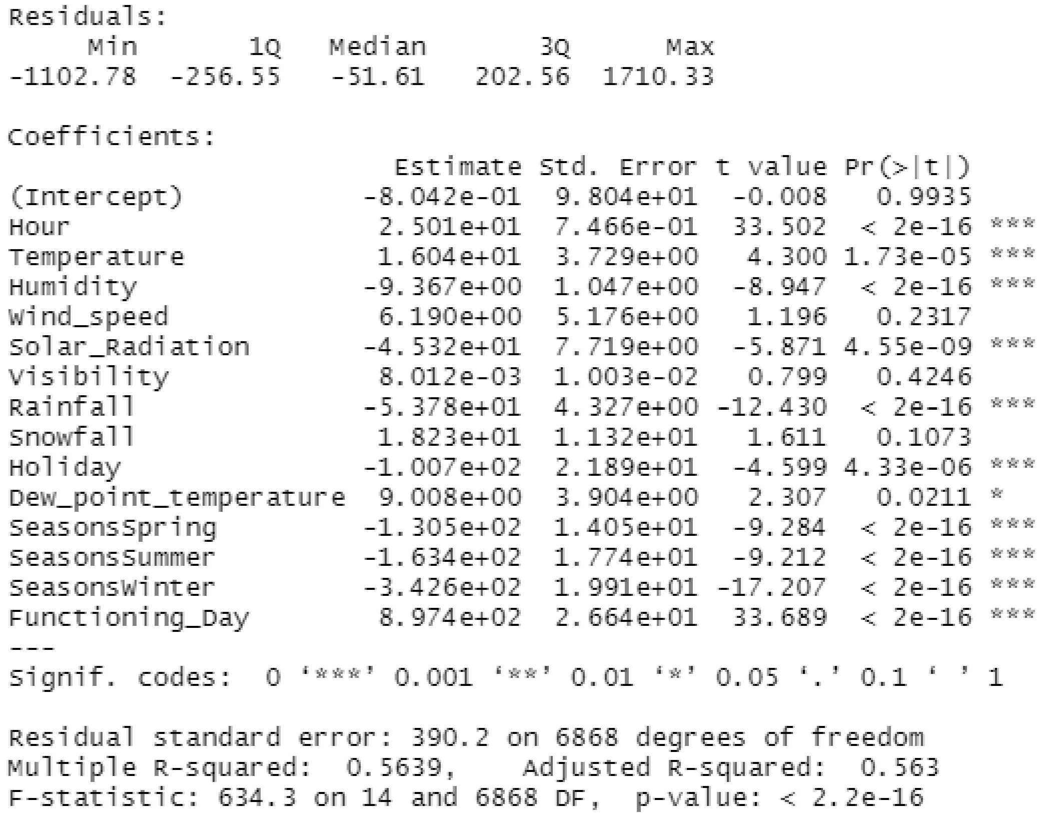
Splitting the Data

The complete dataset has been split into two smaller datasets. One dataset contains 80% of the total data being used for training. And another contains 20% of the total data being used for testing.

Multiple Linear Model

In the MLM model, only the Date variable is removed from the analysis. Rest all the variables are considered for prediction.

Results:



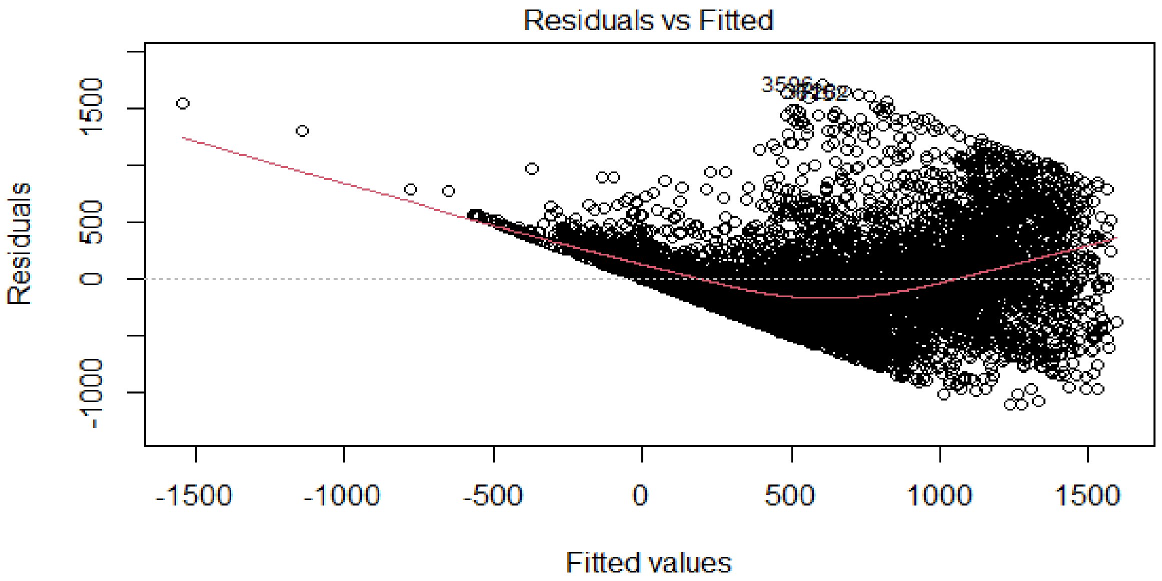
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1102.78  -256.55   -51.61   202.56  1710.33 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -8.042e-01  9.804e+01  -0.008   0.9935    
Hour                   2.501e+01  7.466e-01  33.502  < 2e-16 ***
Temperature            1.604e+01  3.729e+00   4.300 1.73e-05 ***
Humidity              -9.367e+00  1.047e+00  -8.947  < 2e-16 ***
Wind_speed             6.190e+00  5.176e+00   1.196   0.2317    
Solar_Radiation       -4.532e+01  7.719e+00  -5.871 4.55e-09 ***
Visibility             8.012e-03  1.003e-02   0.799   0.4246    
Rainfall              -5.378e+01  4.327e+00 -12.430  < 2e-16 ***
Snowfall               1.823e+01  1.132e+01   1.611   0.1073    
Holiday               -1.007e+02  2.189e+01  -4.599 4.33e-06 ***
Dew_point_temperature  9.008e+00  3.904e+00   2.307   0.0211 *  
SeasonsSpring         -1.305e+02  1.405e+01  -9.284  < 2e-16 ***
SeasonsSummer         -1.634e+02  1.774e+01  -9.212  < 2e-16 ***
SeasonsWinter         -3.426e+02  1.991e+01 -17.207  < 2e-16 ***
Functioning_Day        8.974e+02  2.664e+01  33.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 390.2 on 6868 degrees of freedom
Multiple R-squared:  0.5639,	Adjusted R-squared:  0.563 
F-statistic: 634.3 on 14 and 6868 DF,  p-value: < 2.2e-16
```

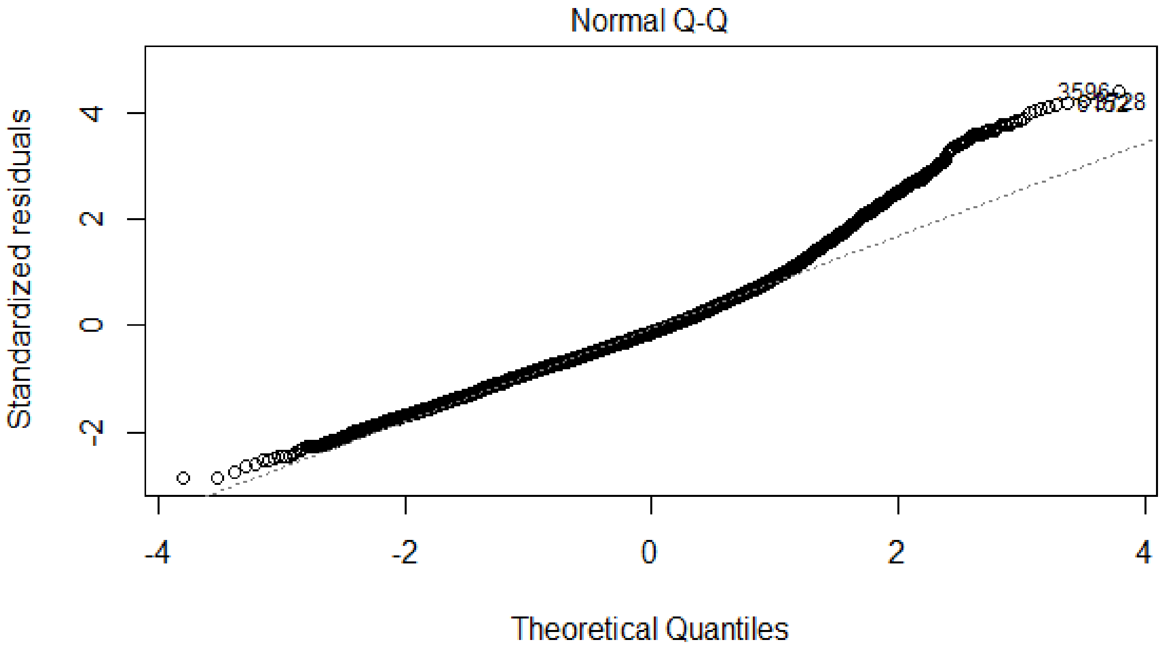
Most of the variables are showing significant results except ‘Wind sped’, ‘Visibility’, ‘Snowfall’ & ‘Dew point temperature’ as they have high significant value of being the contribution(slope) equal to zero in the model. So, to improve the model these variables can be removed. Adjusted R square is coming out as 0.563 which means that 56.3% of the prediction of rented bike count can be explained by all these variables.

The data equation formed can be displayed as:-

Y(Rented bike count) = (-0.804) + (25.01)*Hour + (16.04)*Temperature + (-9.367)*Humidity +(6.19)*Wind_speed + (-45.32)*Solar_radiation + (0.008)*Visibility + (-53.78)*Rainfall +(18.23)*Snowfall + (-100.7)*Holiday + (9)*Dew_point_temperature + (897.4)*FunctioningDay

These coefficients in the equation also tell whether the relation is negative or positive depending on the sign before the number.





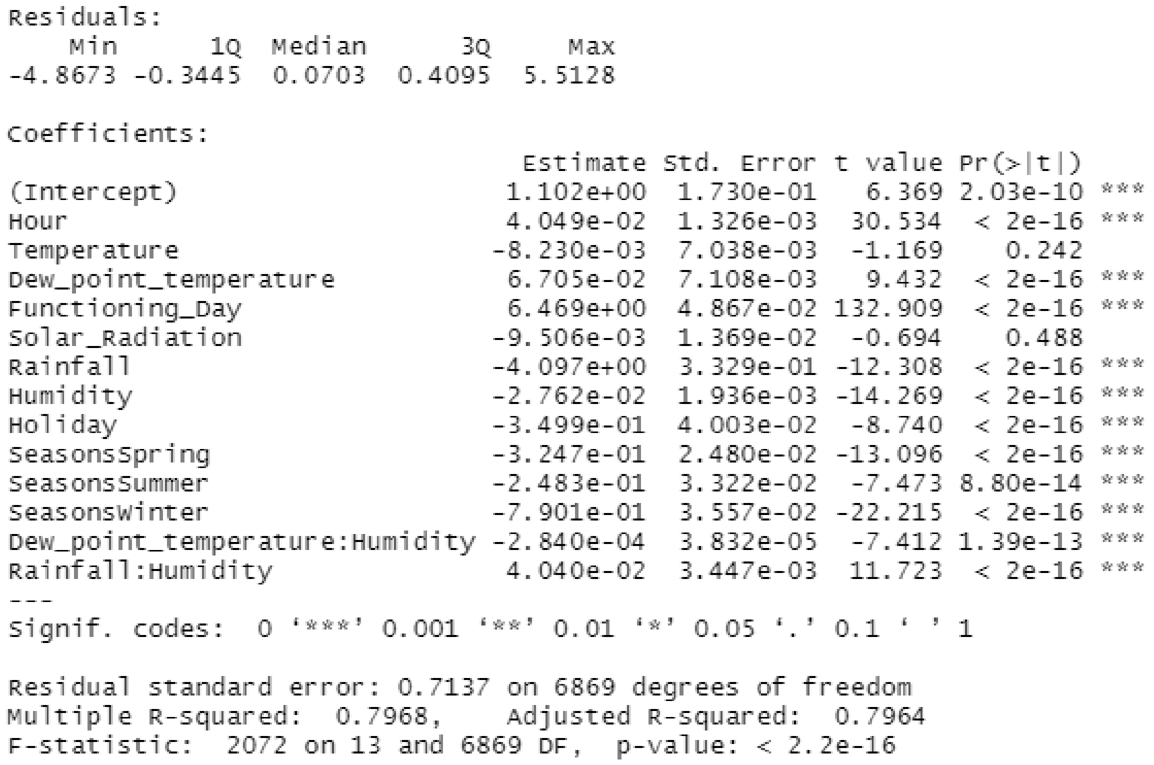
The results for these plots are not satisfying. The linear model assumptions are not satisfied here. In the first plot, the residuals are not scattered, and the trend line shows a somewhat quadratic bend. The second figure is a Normal QQ plot, which is not following a linear trend and has few outliers. The trend line in the quadratic plots is showing quadratic trend.

Polynomial Model with Transformation

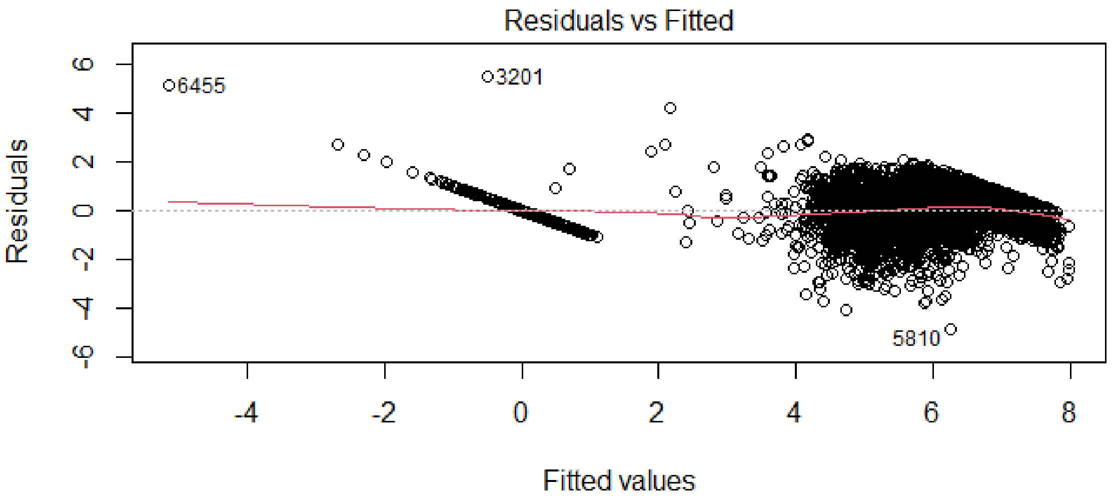
Based on the values found in MLM, the features that don’t contribute any significant relation have been removed from model. Along with that, to improve the basic MLM, transformations and interactions are introduced. Two changes are made in the model. One transformation of response variable by taking log(y). Another by including the interactions between two pairs of variables – (Dew_point_temperature & Humidity) and (Rainfall & Humidity) which seems to give decent contribution in prediction. These pairs have average correlation among themselves.

Results after transformation

The graphs are a little better than before. The trend line in residual plot is much more straightened as to what it was earlier. These two transformations have increased the R squared metric by 60%. The response variable was right skewed, so the transformation has improved the result significantly. In the summary table shown below, almost all the probabilities of t-values are not very significant depicting that each contribute to prediction. Although the graph still needs improvement as there is high collinearity between the variables. In this model, the variable temperature and solar radiation are failing to give any contribution. One possibility could have been square root instead of log transformation.



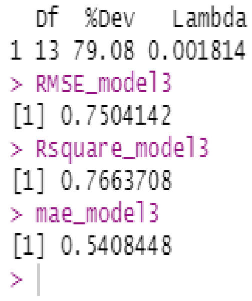
```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.8673 -0.3445  0.0703  0.4095  5.5128 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  1.102e+00  1.730e-01   6.369 2.03e-10 ***
Hour                         4.049e-02  1.326e-03  30.534  < 2e-16 ***
Temperature                 -8.230e-03  7.038e-03  -1.169    0.242    
Dew_point_temperature        6.705e-02  7.108e-03   9.432  < 2e-16 ***
Functioning_Day              6.469e+00  4.867e-02 132.909  < 2e-16 ***
Solar_Radiation             -9.506e-03  1.369e-02  -0.694    0.488    
Rainfall                    -4.097e+00  3.329e-01 -12.308  < 2e-16 ***
Humidity                    -2.762e-02  1.936e-03 -14.269  < 2e-16 ***
Holiday                     -3.499e-01  4.003e-02  -8.740  < 2e-16 ***
SeasonsSpring               -3.247e-01  2.480e-02 -13.096  < 2e-16 ***
SeasonsSummer               -2.483e-01  3.322e-02  -7.473 8.80e-14 ***
SeasonsWinter               -7.901e-01  3.557e-02 -22.215  < 2e-16 ***
Dew_point_temperature:Humidity -2.840e-04  3.832e-05  -7.412 1.39e-13 ***
Rainfall:Humidity            4.040e-02  3.447e-03  11.723  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7137 on 6869 degrees of freedom
Multiple R-squared:  0.7968,	Adjusted R-squared:  0.7964 
F-statistic:  2072 on 13 and 6869 DF,  p-value: < 2.2e-16
```



Penalised Regression (Lasso Model)

Due to high multicollinearity, a penalised model is needed to remove the unwanted variables. Lasso Regression is used as a 3rd model for predicting bike count. Library ‘glmnet’ from R was used for all the automatic calculations. As all these functions use only numeric variables, model.matrix() was introduced which would automatically convert all the categorical variables into numerical by assigning them some dummy values. The best lambda found is 0.0018 which is selected by the function cv.glmnet() itself. In the modelling function, ‘glmnet()’, alpha is passed as a parameter which is set equal to 1 for Lasso regression. All the variables are considered for analysis except the variable Date, and response variable is taken in the log format.

Results



```
  Df  %Dev   Lambda
1 13 79.08 0.001814
> RMSE_model3
[1] 0.7504142
> Rsquare_model3
[1] 0.7663708
> mae_model3
[1] 0.5408448
> 
```

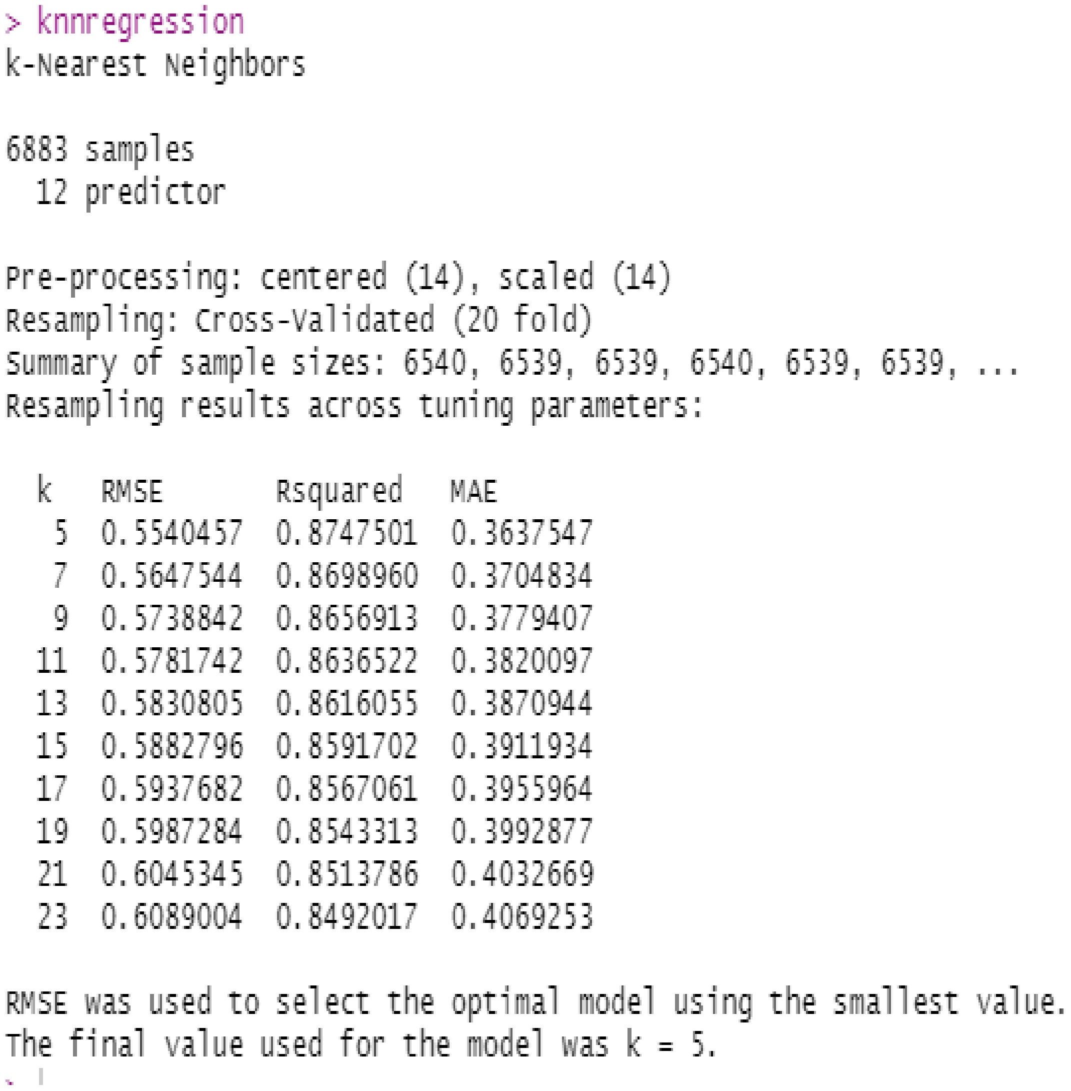
K Nearest Neighbour

KNN algorithm is used as a model for predicting bike count. For this algorithm also, response variable was selected as log(Rented bike Count) because of the skew in the original variable. In explanatory variables, all the variables were considered as original except the variable Date was removed. The scales were pre-processed and standardised. One additional thing included is cross-validation. Cross validation is a process that tries different sets of training and test data for some n number of times so that all the dataset gets equal weightage to be a part of both training and test data and finding the best pair of training and test data through this algorithm.

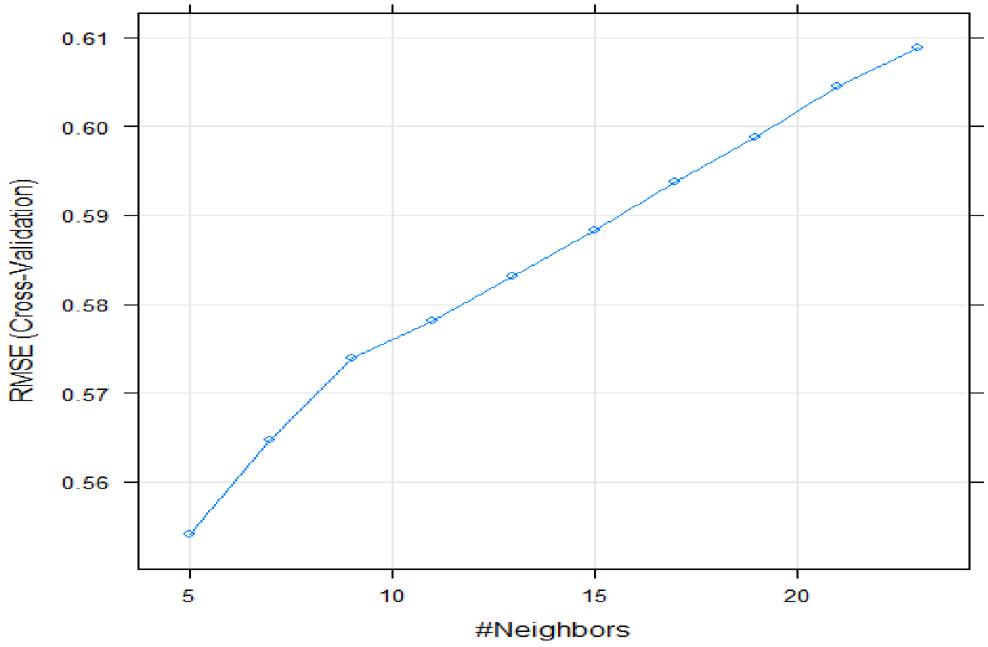
Also, for choosing the value of k, tune length feature was used to test some 20 values for k before picking the best value.

Results

The summary show below contains different values of RMSE, R square and MAE for all values of k that are tested. The graph below shows the tuning results for different values of k. Best value is the one that gives the minimal value for RMSE. Here best value of k came out to be 5. So, 5 nearest neighbours using Euclidean distance gave the predictions for test dataset.



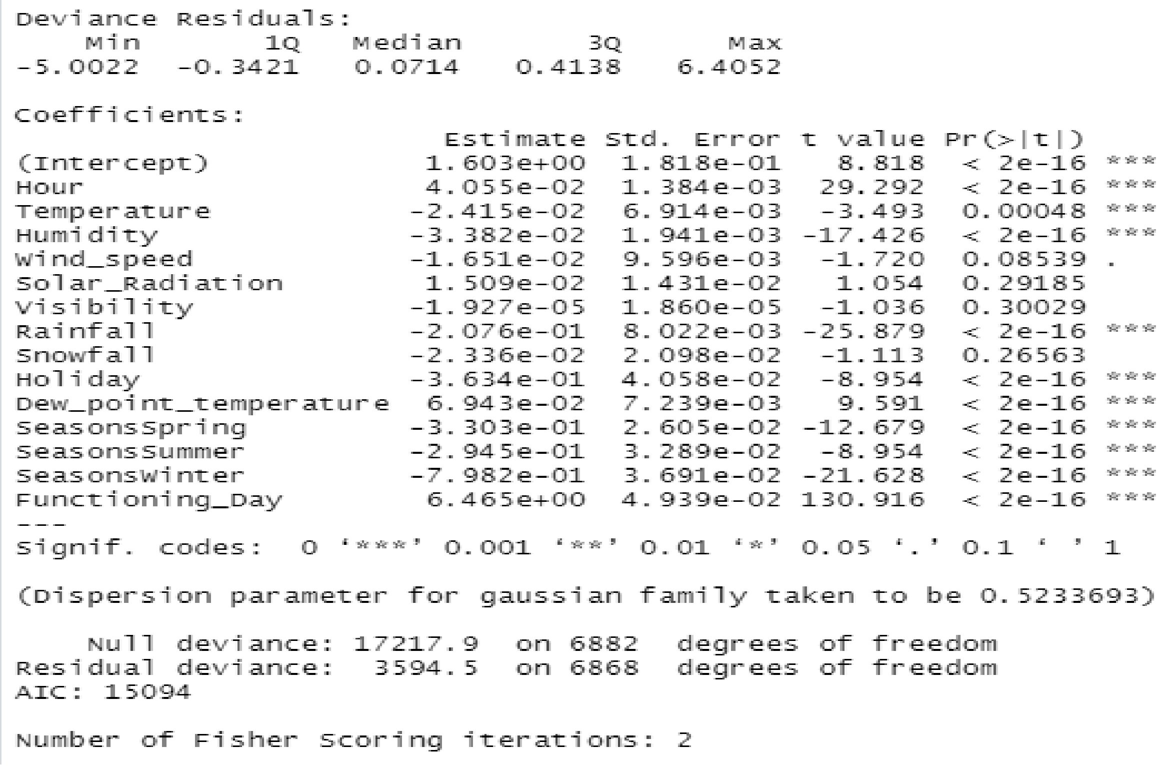
```
> knnregression
k-Nearest Neighbors

6883 samples
  12 predictor

Pre-processing: centered (14), scaled (14)
Resampling: Cross-Validated (20 fold)
Summary of sample sizes: 6540, 6539, 6539, 6540, 6539, 6539, ...
Resampling results across tuning parameters:

  k   RMSE       Rsquared   MAE
   5  0.5540457  0.8747501  0.3637547
   7  0.5647544  0.8698960  0.3704834
   9  0.5738842  0.8656913  0.3779407
  11  0.5781742  0.8636522  0.3820097
  13  0.5830805  0.8616055  0.3870944
  15  0.5882796  0.8591702  0.3911934
  17  0.5937682  0.8567061  0.3955964
  19  0.5987284  0.8543313  0.3992877
  21  0.6045345  0.8513786  0.4032669
  23  0.6089004  0.8492017  0.4069253

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 5.
```



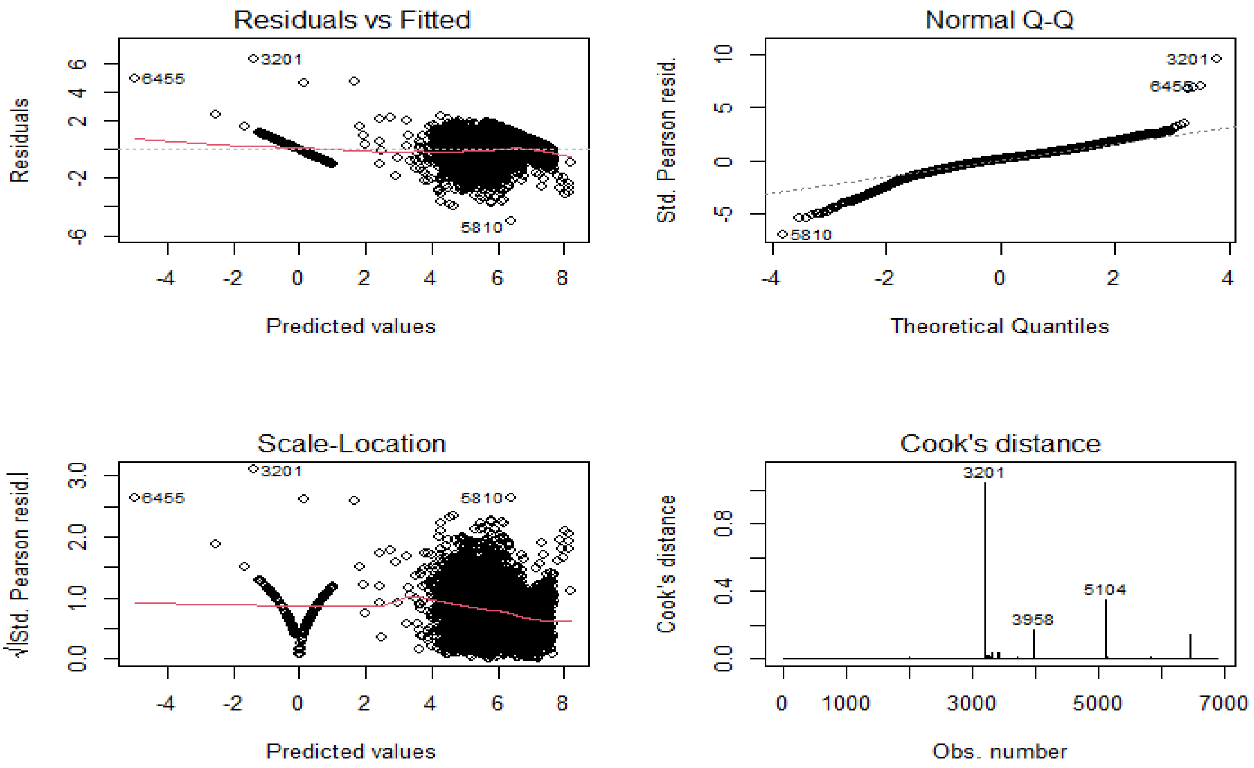
Model 4- Generalised Linear Model (Poison)

As the prediction variable is count. Poisson Linear model is implemented. All the variables have been taken except Date. In the R studio itself, there is a function glm() which can be used in where family attribute can be set to ‘Poisson’ for Poisson models.

Results



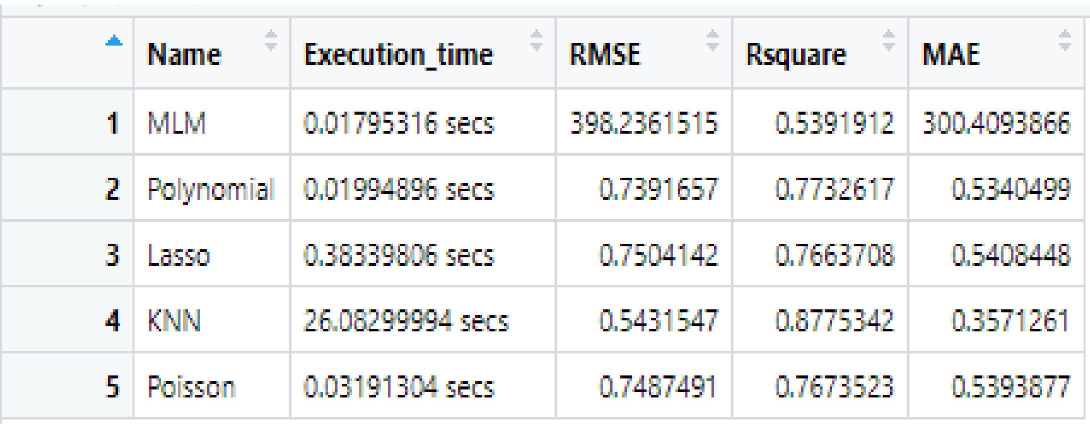
```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.0022  -0.3421   0.0714   0.4138   6.4052  

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.603e+00  1.818e-01   8.818  < 2e-16 ***
Hour                   4.055e-02  1.384e-03  29.292  < 2e-16 ***
Temperature           -2.415e-02  6.914e-03  -3.493  0.00048 ***
Humidity              -3.382e-02  1.941e-03 -17.426  < 2e-16 ***
Wind_speed            -1.651e-02  9.596e-03  -1.720  0.08539 .  
Solar_Radiation        1.509e-02  1.431e-02   1.054  0.29185    
Visibility            -1.927e-05  1.860e-05  -1.036  0.30029    
Rainfall              -2.076e-01  8.022e-03 -25.879  < 2e-16 ***
Snowfall              -2.336e-02  2.098e-02  -1.113  0.26563    
Holiday               -3.634e-01  4.058e-02  -8.954  < 2e-16 ***
Dew_point_temperature  6.943e-02  7.239e-03   9.591  < 2e-16 ***
SeasonsSpring         -3.303e-01  2.605e-02 -12.679  < 2e-16 ***
SeasonsSummer         -2.945e-01  3.289e-02  -8.954  < 2e-16 ***
SeasonsWinter         -7.982e-01  3.691e-02 -21.628  < 2e-16 ***
Functioning_Day        6.465e+00  4.939e-02 130.916  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.5233693)

    Null deviance: 17217.9  on 6882  degrees of freedom
Residual deviance:  3594.5  on 6868  degrees of freedom
AIC: 15094

Number of Fisher Scoring iterations: 2
```



The first plot shows there is not constant variance like there is an issue of Heteroscedasticity is there. We can solve this issue with taking log(x), sqrt(x).

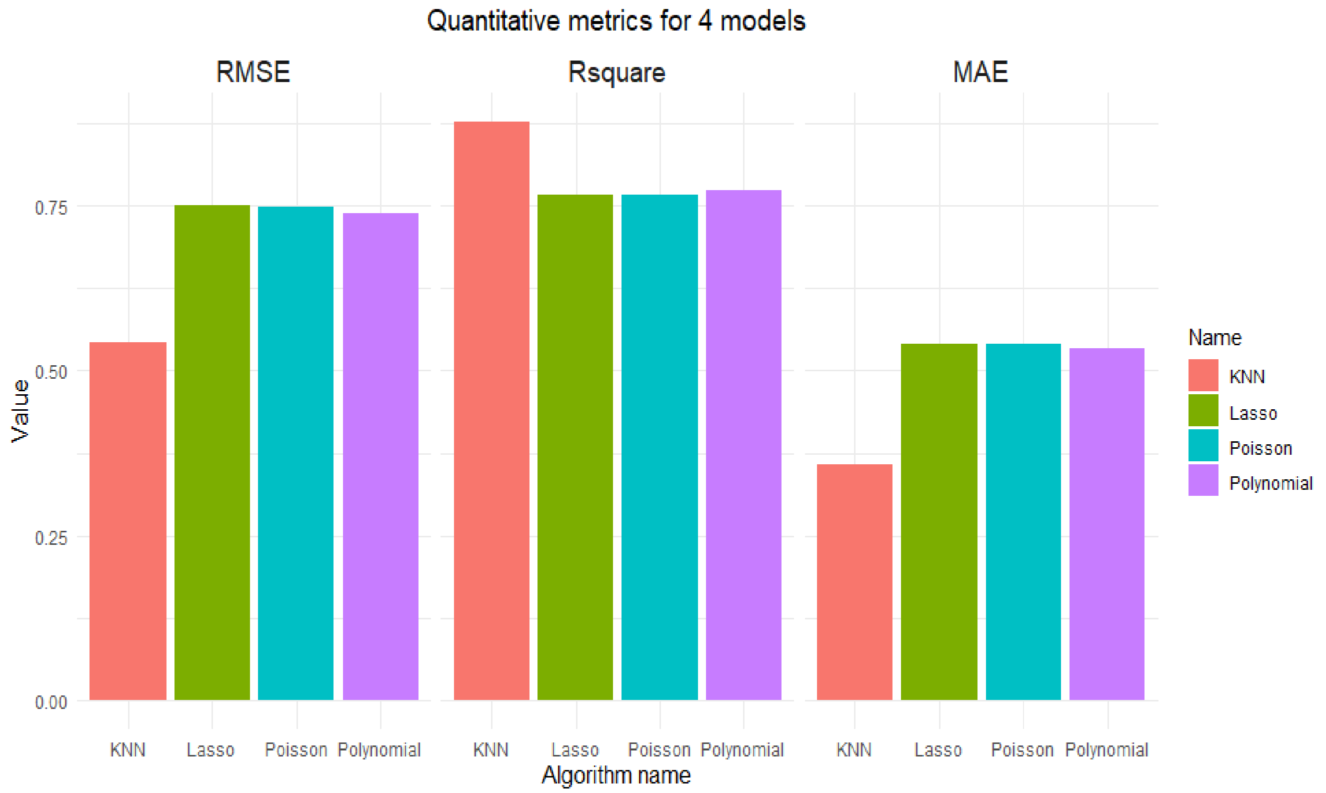
Evaluation

Qualitative Comparison of Models



	Name	Execution_time	RMSE	Rsquare	MAE
1	MLM	0.01795316 secs	398.2361515	0.5391912	300.4093866
2	Polynomial	0.01994896 secs	0.7391657	0.7732617	0.5340499
3	Lasso	0.38339806 secs	0.7504142	0.7663708	0.5408448
4	KNN	26.08299994 secs	0.5431547	0.8775342	0.3571261
5	Poisson	0.03191304 secs	0.7487491	0.7673523	0.5393877

The MLM results have a big difference because for all the other models the prediction was taken for log of response variable and for MLM model the prediction was done for the original count variable. From the statistics it looks like KNN model is the best when it comes to prediction. It has the biggest R squared(coefficient of determination) measure and least RMSE(Root Mean squared Error) and MAE(Mean Absolute Error). But the KNN model has taken the maximum time for execution as it is a big dataset and required a lot of time with 10-fold cross validation included.



References

1. https://www.statology.org/
2. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn
3. https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
4. https://www.ibm.com/cloud/learn/overfitting
5. https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/
6. https://www.sciencedirect.com/topics/mathematics/generalized-linear-model
7. https://www.dataquest.io/blog/tutorial-poisson-regression-in-r