PREDICATION OF BIKE RENTAL COUNT

RAHUL BHATTACHARYYA

5TH AUGUST 2018

Contents

1 Chapter 1
Introduction

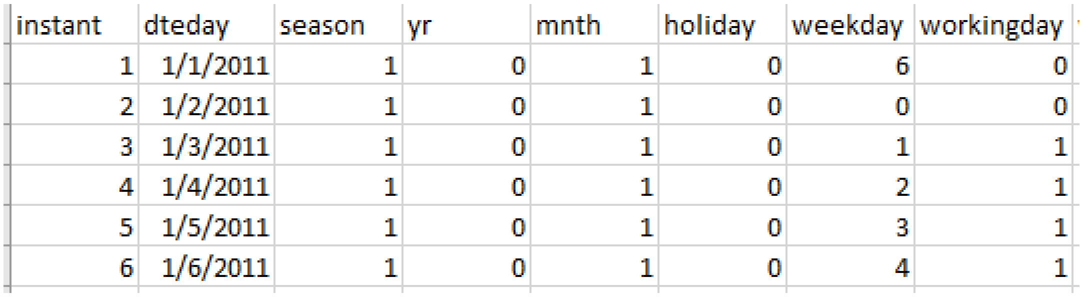
1.1 Problem Statement

The problem statement is regarding an organization who lends rental bike. The demand of bikes for a particular day depends upon several factors like weather situation, season, holiday etc. It is important for an organization to know the demand of a particular day beforehand, so that they can meet the demand smoothly. This problem can be solved using several regression techniques as discussed in detail in this report.

1.2 Data

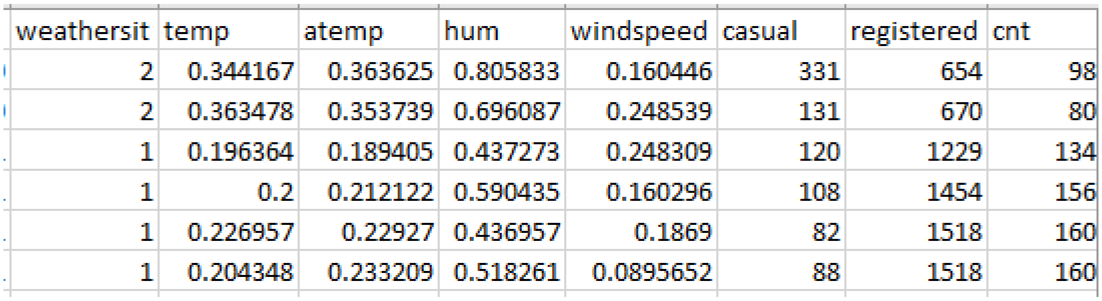
The main task is to build a regression model which will tell what will be the demand of cycles for a particular day. Below is the glimpse of the data

Column (1-8)



instant	dteday	season	yr	mnth	holiday	weekday	workingday
1	1/1/2011	1	0	1	0	6	0
2	1/2/2011	1	0	1	0	0	0
3	1/3/2011	1	0	1	0	1	1
4	1/4/2011	1	0	1	0	2	1
5	1/5/2011	1	0	1	0	3	1
6	1/6/2011	1	0	1	0	4	1

Column (9-16)



weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
2	0.344167	0.363625	0.805833	0.160446	331	654	98
2	0.363478	0.353739	0.696087	0.248539	131	670	80
1	0.196364	0.189405	0.437273	0.248309	120	1229	134
1	0.2	0.212122	0.590435	0.160296	108	1454	156
1	0.226957	0.22927	0.436957	0.1869	82	1518	160
1	0.204348	0.233209	0.518261	0.0895652	88	1518	160

So, in total we have 16 variables from which we would like to predict churn variable. The predictor variables as per the problems are.

instant: Record index

dteday: Date

season: Season (1: spring, 2: summer, 3: fall, 4: winter)

yr: Year (0: 2011, 1:2012)

mnth: Month (1 to 12)

holiday: weather day is holiday or not (extracted fromHoliday Schedule)

weekday: Day of the week

workingday: If day is neither weekend nor holiday is 1, otherwise is 0.

weathersit: (extracted fromFreemeteo)

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp: Normalized temperature in Celsius. The values are derived via

(t-t_min)/(t_max-t_min),

t_min=-8, t_max=+39 (only in hourly scale)

atemp: Normalized feeling temperature in Celsius. The values are derived via

(t-t_min)/(t_maxt_min),

t_min=-16, t_max=+50 (only in hourly scale)

hum: Normalized humidity. The values are divided to 100 (max)

windspeed: Normalized wind speed. The values are divided to 67 (max)

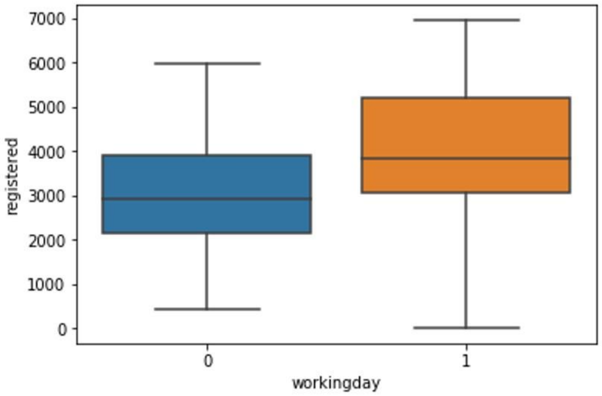
2 Chapter 2
Methodology

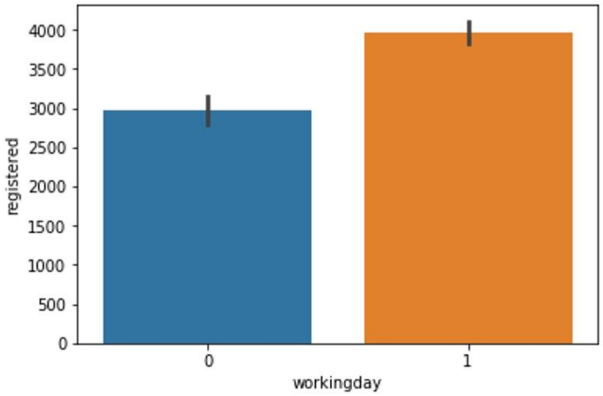
2.1 Pre Processing

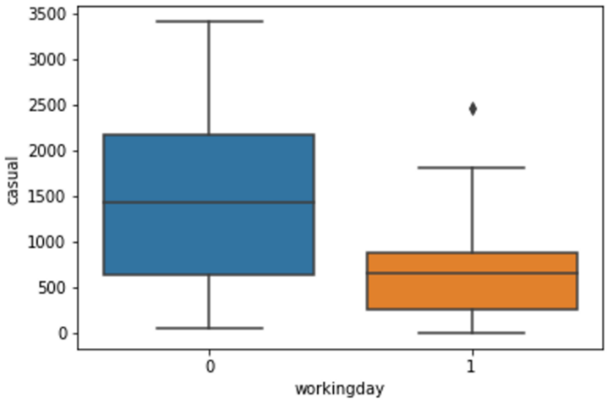
Any machine learning algorithms needs preprocessing techniques to prepare the data efficiently to fit the algorithms. In this problem several preprocessing methods like exploratory data analysis, chi-square test techniques are applied to select and prepare the actual data. It was also noticed there was no missing value in the entire data. The outlier analysis is being avoided for this particular dataset as an outlier in the numerical variables can have valuable information which can be really useful from the model perspective (natural outliers). Example: It is possible that on a certain day excessive bikes are rented based on the weather condition or holiday. Also before proceeding with exploratory analysis “instant”, “dteday” have been dropped.

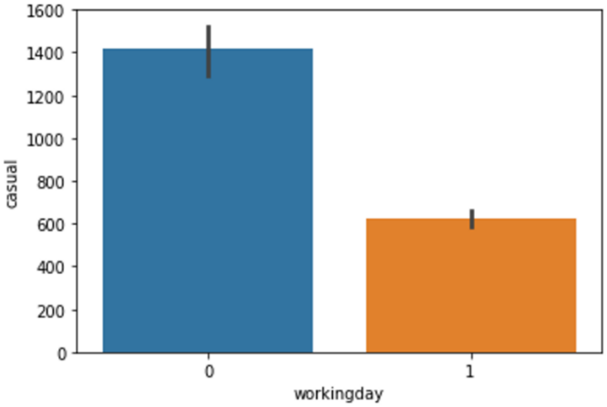
2.1.1 Exploratory analysis

Exploratory analysis is one of the most important aspect of any analysis as we try to get an early insight from the data. Here also we tried to see various distribution for each of the variable which helped us in feature selection process. We also selected “registered” and “casual” as our target variable as “cnt” is just the sum of these two variables. So there is a chance of losing critical data pattern if we choose “cnt” as our only target variable. First, the distribution of registered vs working day and casual vs working day were plotted and it has been found that on a working day the number of registered bike users is significantly higher than on a non-working day and the number of casual bike users are higher in non-working day.



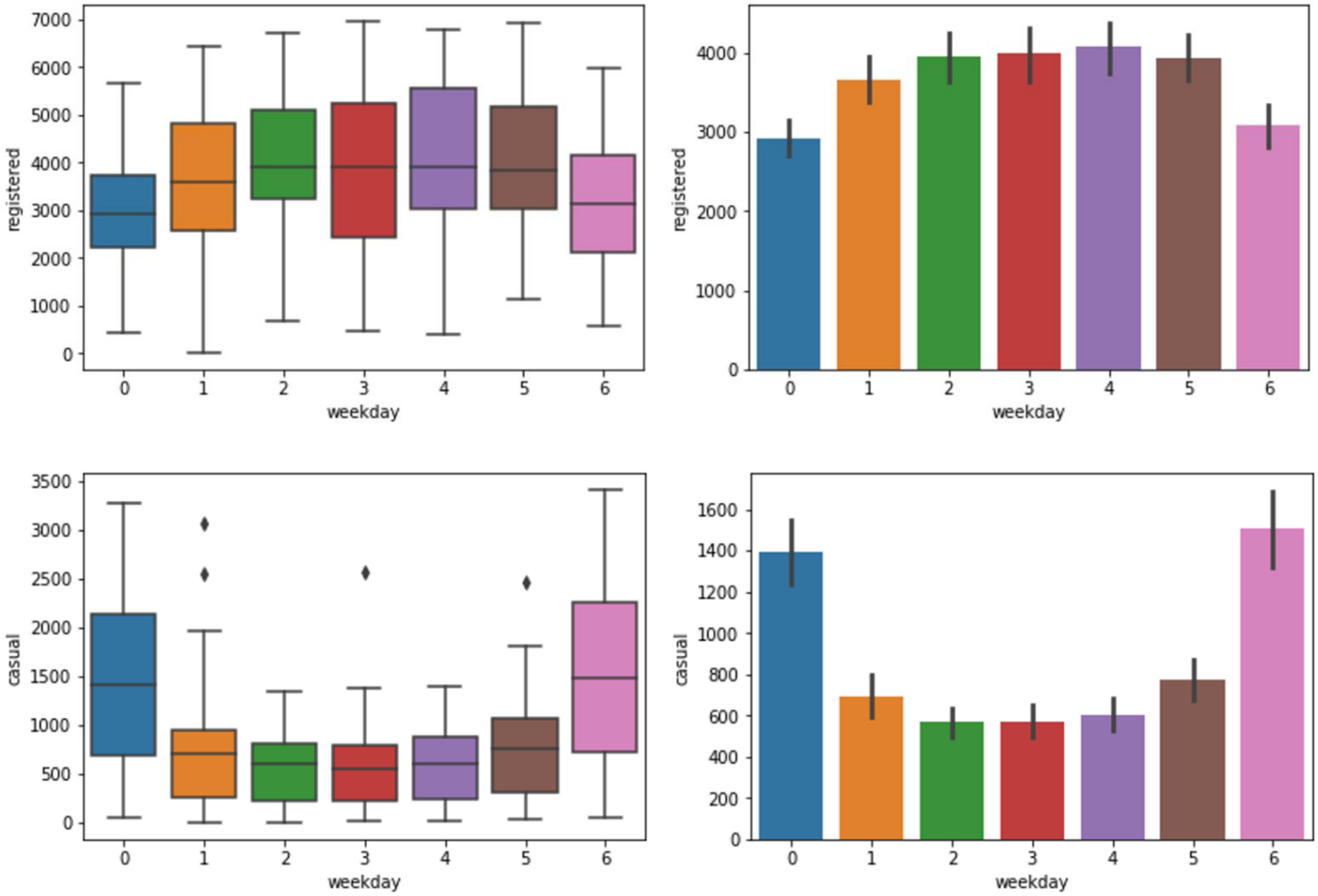






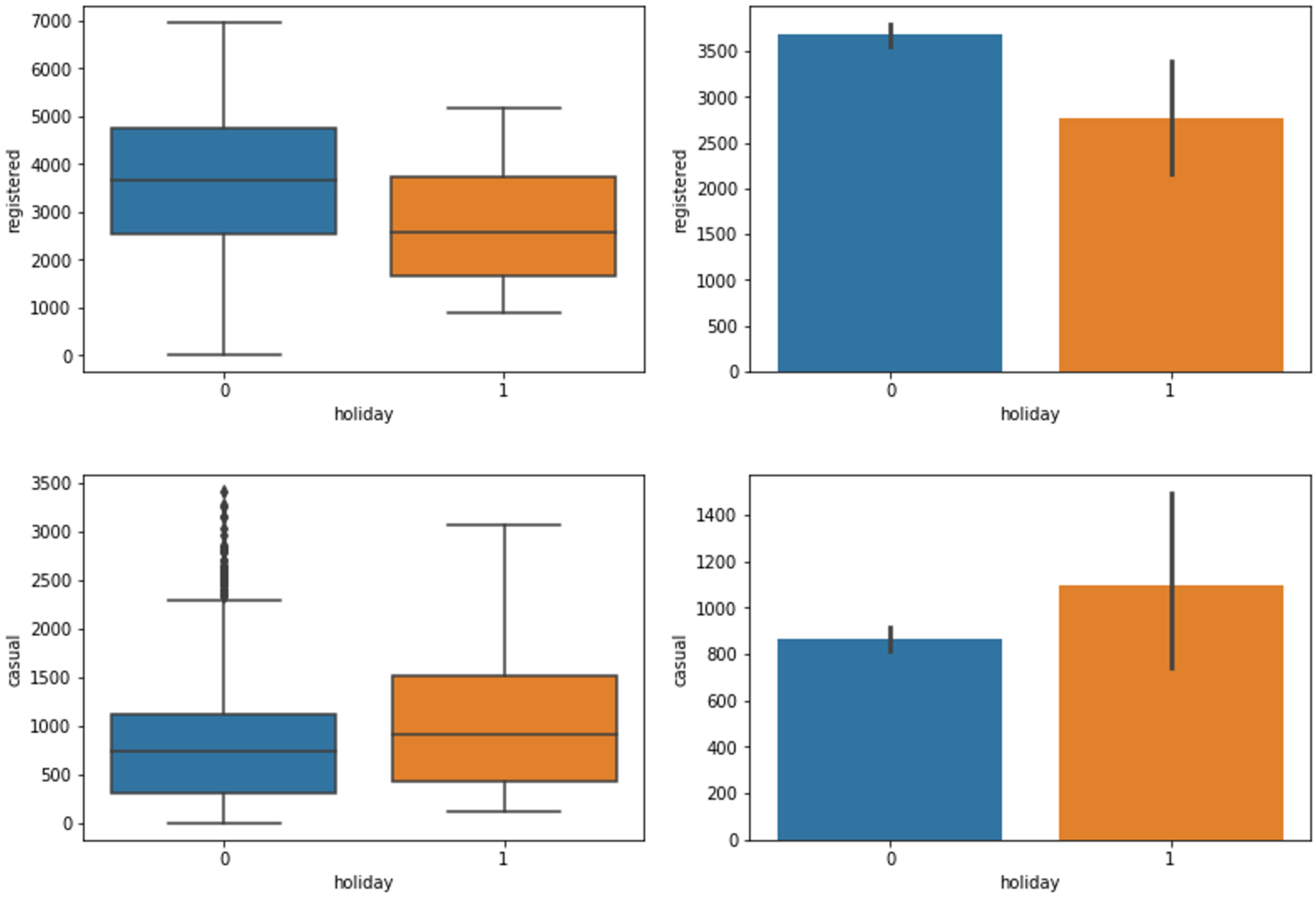
Here the box plot shows the range of the variables and the barplot shows the mean of each variable.

Further the registered vs weekday and casual vs weekday was plotted. The 0-6 variables denote day of the week which starts from Sunday and end on Saturday. It can be clearly seen that on weekend (Saturday and Sunday number of registered bike users are less than on weekdays. This is a similar trend shown by previous graphs as weekends are non-working days. Similarly, number of casual users are much higher on weekends than on any other days.



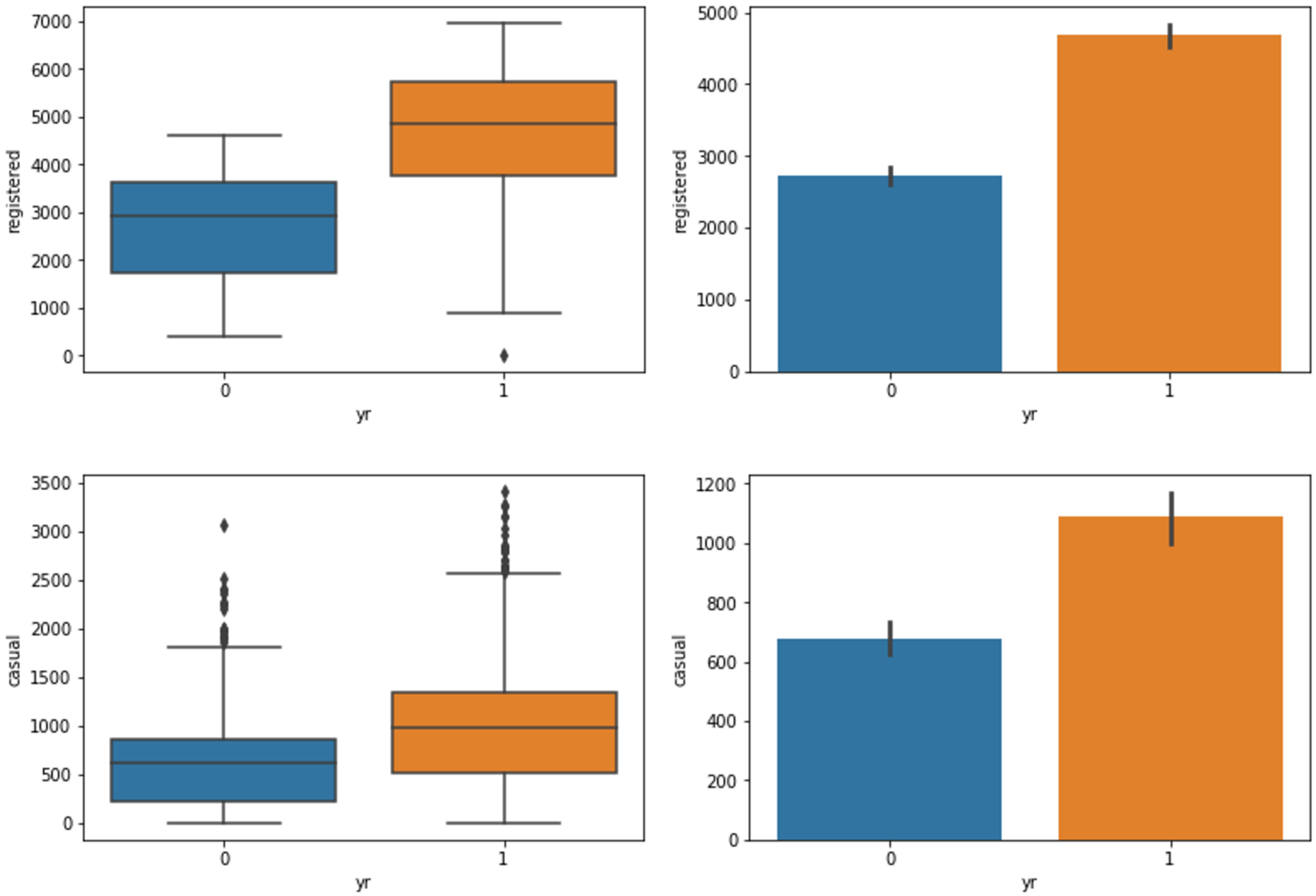
As these are the similar trends shown by the workingday and non-working day plots it was decided to check if the variability in the target variables explained by the weekday variable can be full explained by the working day variable.

The following graph explains the number of registered and casual users with whether a particular day is holiday or not. It also shows a similar trend as previously shown by the above graphs as a holiday is always a non-working day.

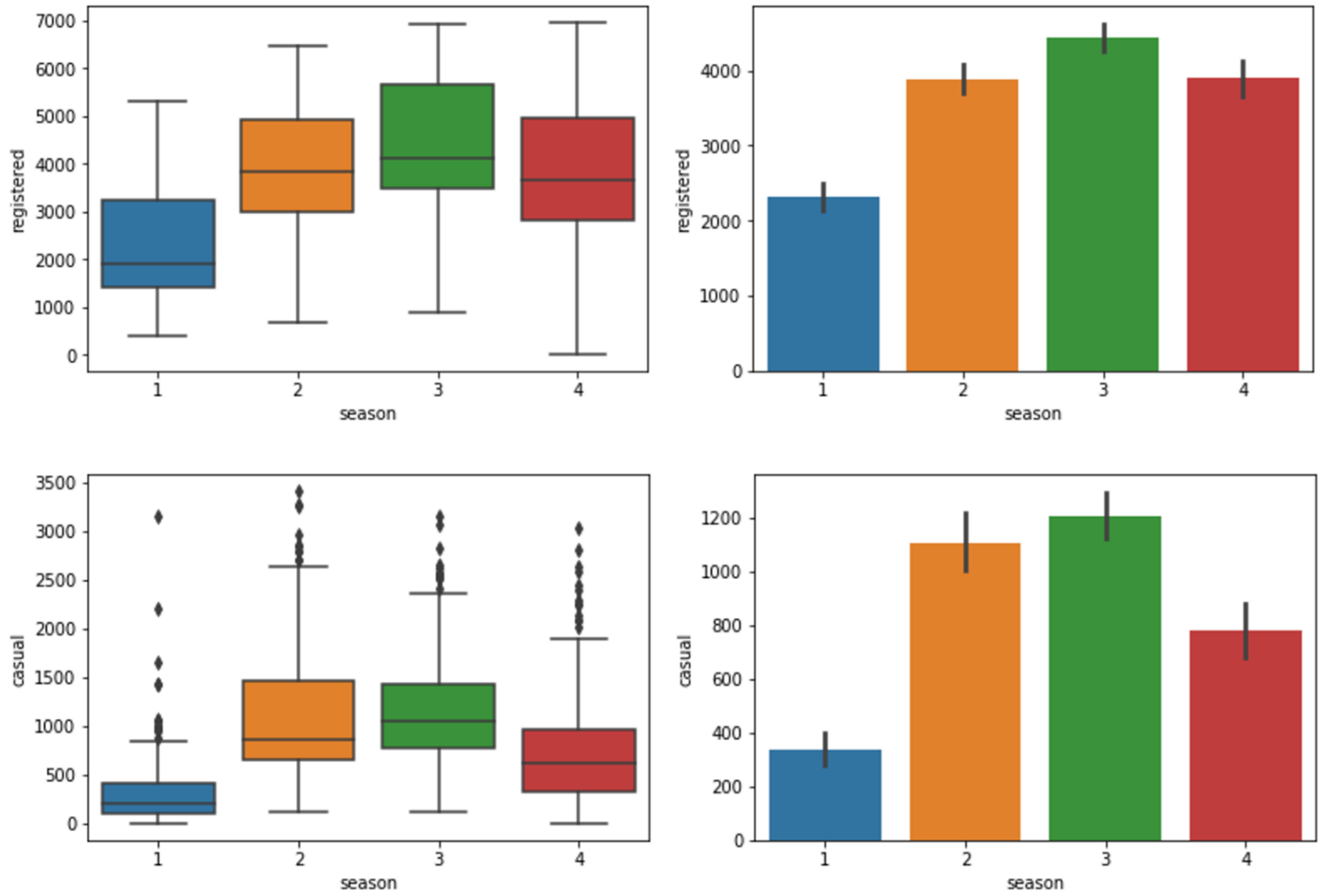


As these are the similar trends shown by the workingday and non-working day plots it was decided to check if the variability in the target variables explained by the holiday variable can be full explained by the working day variable.

The following graphs showed how the numbers of registered and casual users are behaving along the two years. It can be seen that for both the cases there is an increase. So it can be concluded the total number of users are increasing over the years.

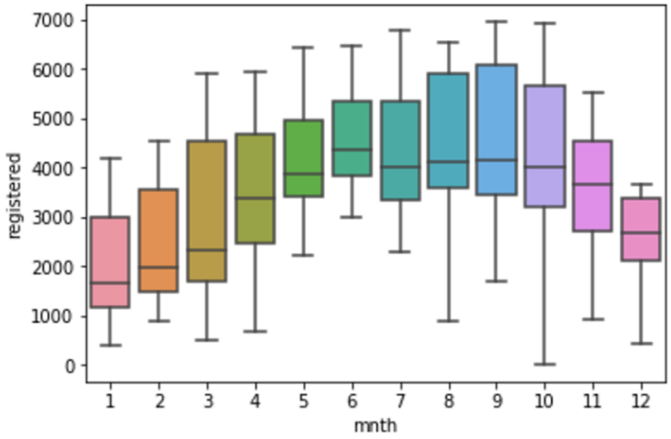


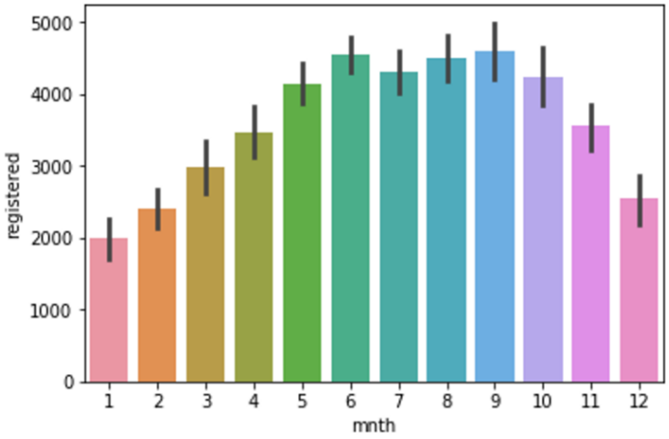
The next plots show how the number of registered and casual users varied through the season. It can be seen that the highest number of cycle rented in Fall and lowest number of cycles are rented in the Spring season. Both the registered and casual users showed the similar trend.

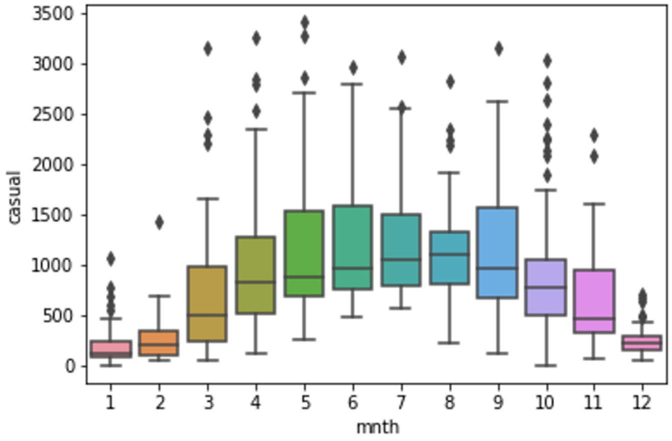


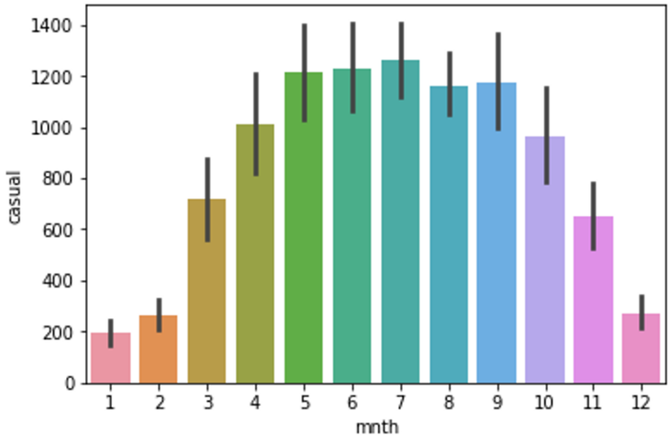
However, an interesting trend is that the number of casual users in winter is significantly less than number of registered users in winter. It can be a result of bad weather where casual users chose not opt for paddling.

The following graphs shows the number of registered and casual users over the month. The (1-12) codes showing months from January to December respectively. As a particular seasons consists of several months So it can be expected similar trend would be displayed by the graphs as shown by the season variable, and as expected it is showing a similar trend.



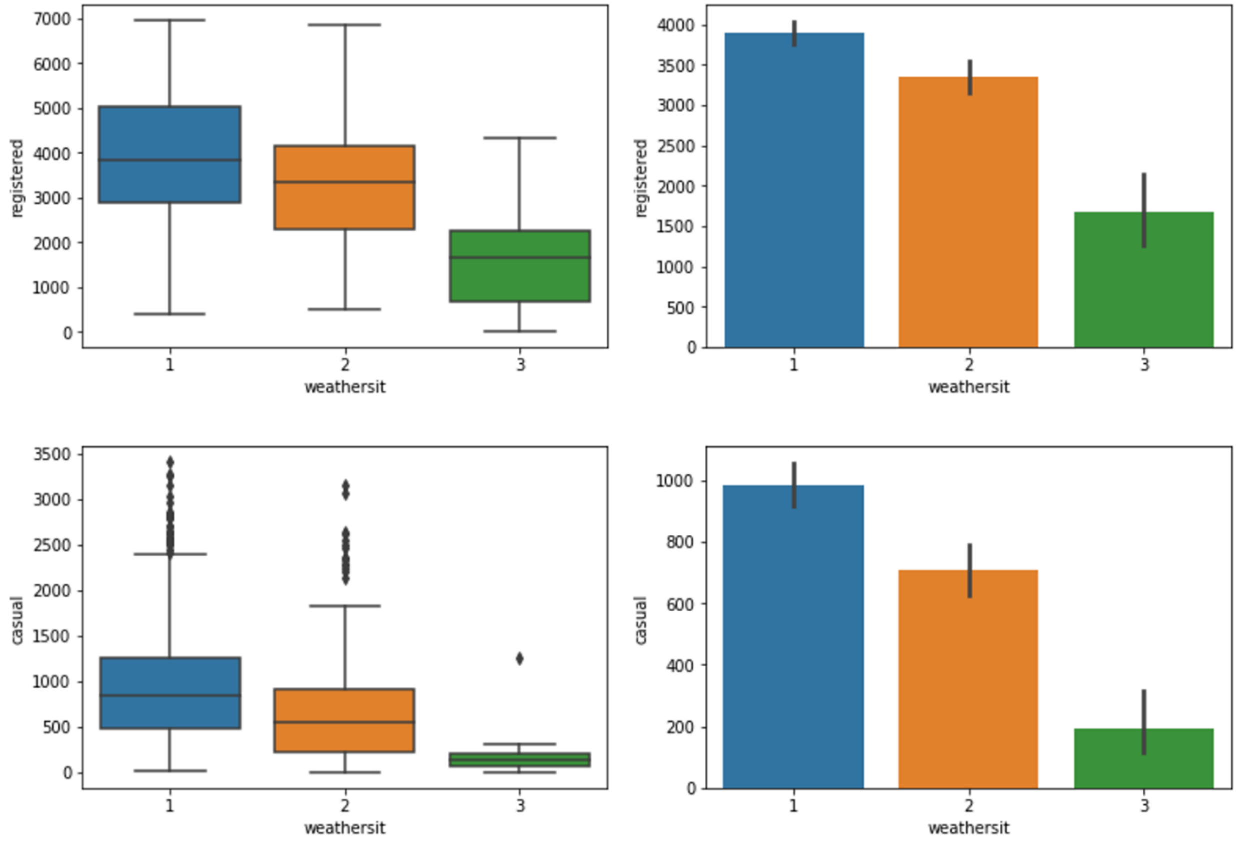




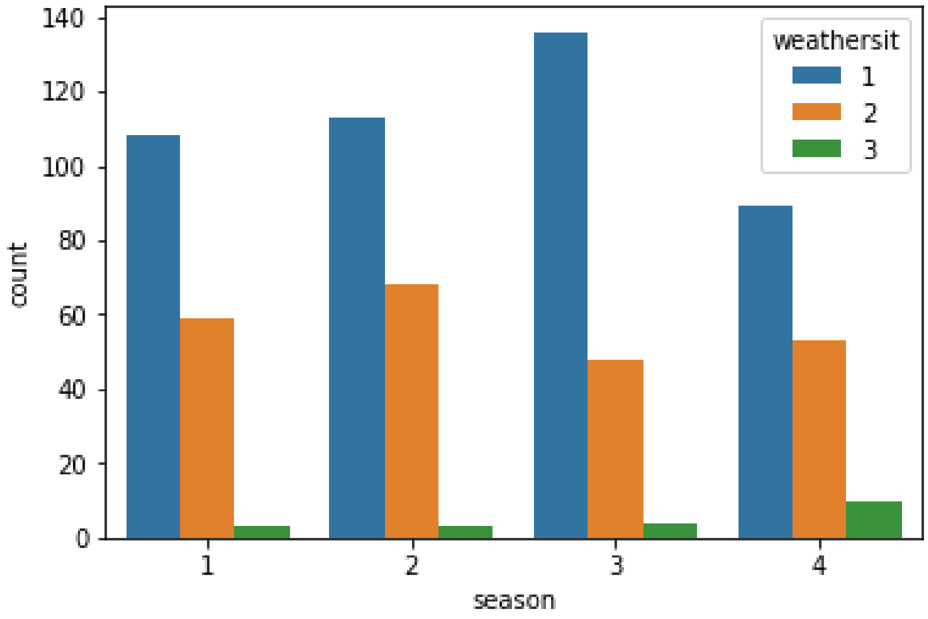


As these are the similar trends shown by the season variable it was decided to check if the variability in the target variables explained by the mnth variable can be full explained by the season variable.

The next graphs show the number of registered and casual users with the weather situation. The number of users have gone down as the weather situation has gone down from good to worse which is fairly intuitive.



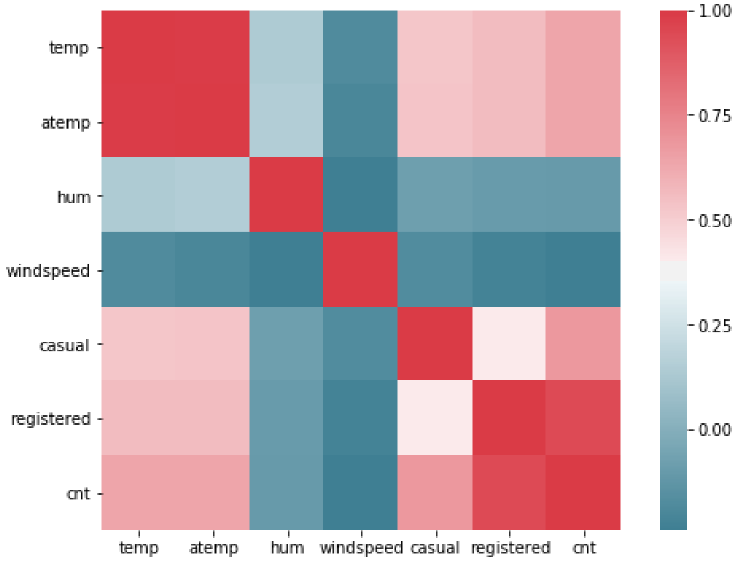
It can be postulated that weather situation is related to season which was not demonstrated by the following graph. So it was decided not to check the dependency between these two variables.

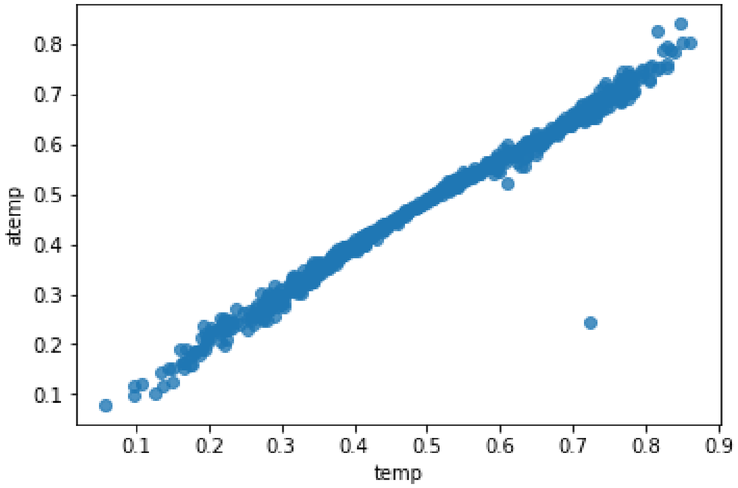


2.1.2 Feature selection

Feature selection is one of the most important aspect of any data analysis project as unnecessary features in any machine learning project can lead to erroneous results. So, before further analysis we decided to select the necessary features based on the insight gained from the exploratory analysis. As shown by the above mentioned graphs it can postulated that the variability explained in the target variable by the predictor variables "weekday" and "holiday" can actually explained by the "workingday" variable. So a chi-square test is done between these variables and it has been found out for both the cases p<0.5. So we discard both the variables. Similarly the "mnth" variable can be fully explained by the "season" variable and hence we discard it also after doing a chi-square test.

As this a regression problem, so multicolinearity is one of the aspect which we need to take care before proceeding with any model development. So it is decided to do a correlation analysis on the numerical variables. It can be seen from the following figure "temp" and "atemp" has a very strong correlation which is supported by the scatter plot as well.





As the correlation between these two variables are very strong we decided to do a VIF test and as expected the result is greater than 10, so the variable "atemp" is discarded due to higher VIF value.

2.1.3 Feature transformation

As several regression models depends upon the least square method so the categorical variables represented by numbers needs to be transformed into dummy variable to eliminate the effect of order. OneHotEncoding is applied on “season” and “weathersit” variable as these two have more than two categories. As the data is already scaled we decided not employ any scaling technique.

2.2 Modelling

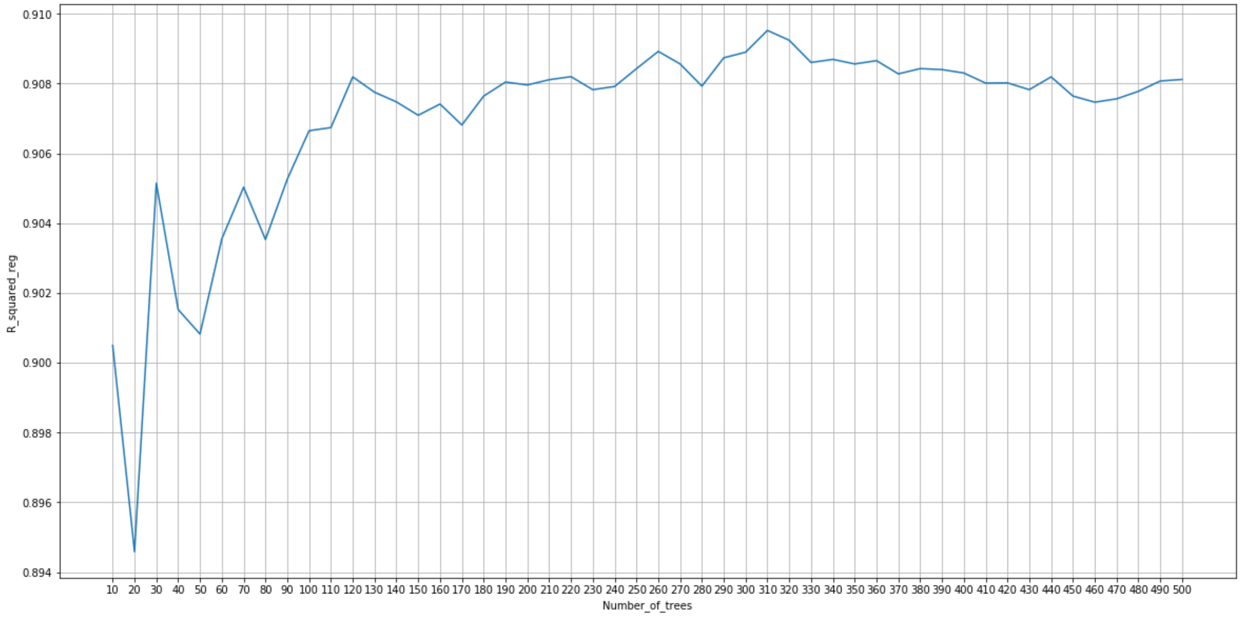
As this problem is a regression problem so we have employed several algorithms to predict our target variable. The models are:

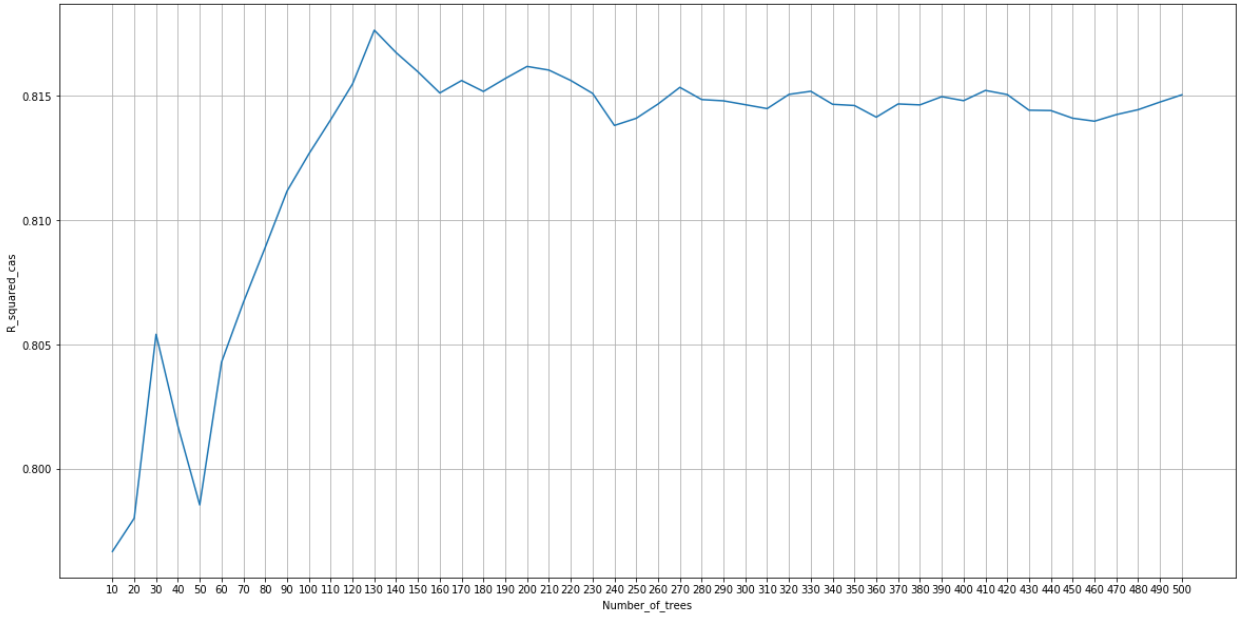
a) Random forest algorithm
b) Decision tree regression
c) Multiple linear regression
d) Polynomial regression

In this context MAPE is chosen as error metric along with the Rsquared value. The data provided by the “day.csv” is divided randomly into three parts training, testing and validation. The validation part is used to provide the sample output.

The training and testing data is further divided into two dataset. One containing all the predictor variables and the “registered” variable. The other one contains all the predictor variables and the “casual” variable. Separate models are built for each of the dataset from each of the algorithm.

In case of random forest algorithm, we employed an iterative algorithm to choose correct number of decision trees. The algorithm iteratively calculates Rsquared value for both the dataset and plot it in a graph to choose the number of trees in an optimized way.





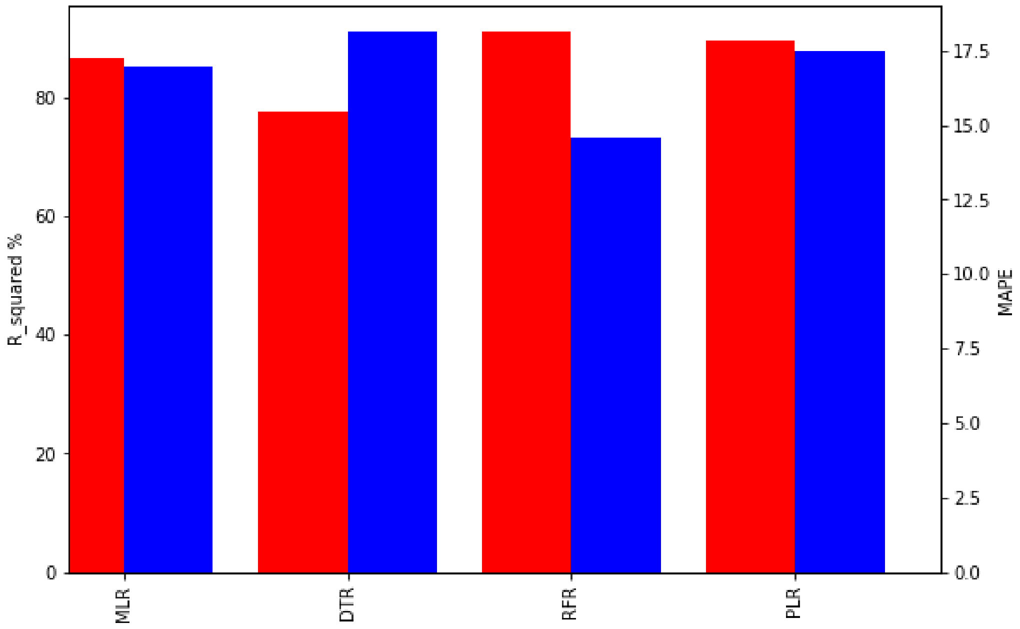
It can be noticed 310 and 130 are the most optimized no of trees to proceed with as it gives highest Rsquared value for registered and casual user's data respectively. The number of trees chosen in R is 60 and 10 respectively for registered and casual users. The supporting figures can be found in the Appendix. We also employed other algorithms as mentioned for both the dataset and in every cases Rsquared value and MAPE value is recorded.

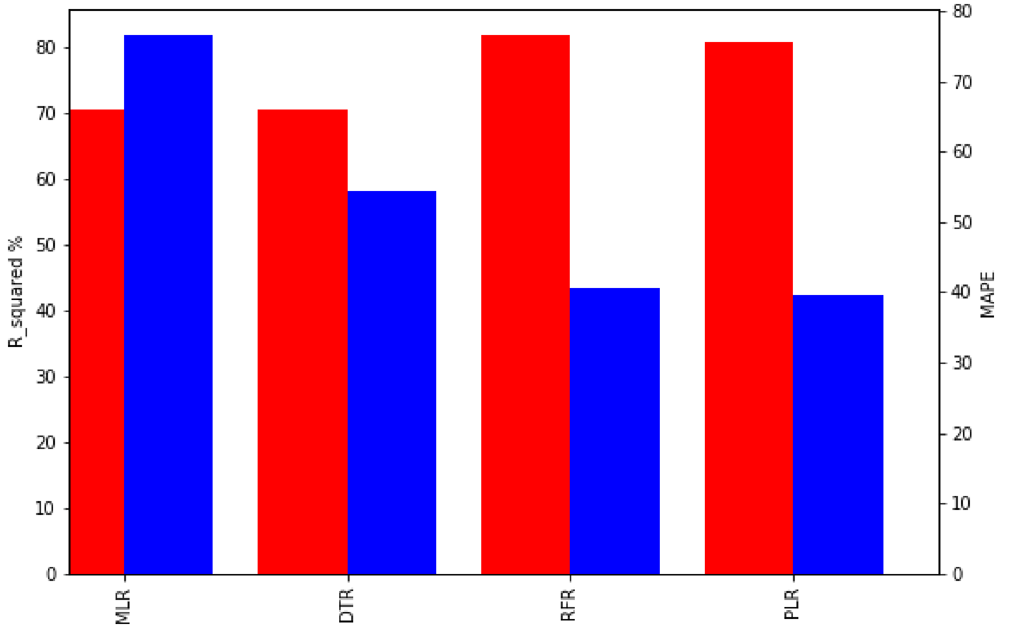
3 Chapter 3

Conclusion

3.1 Model selection

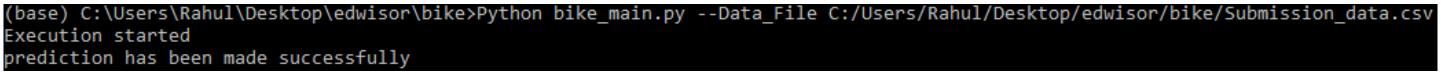
The final model is selected based on the both error metrices, namely RSquared value as well as the MAPE value. It can be seen clearly Random forest algorithm showed the most optimized result in both the dataset. In case of R also same trend is being showed. Though in case of R polynomial regression showed a better MAPE value it has not been chosen due to negative output which is not meaningful in this particular scenario. The representation can be found in appendix.





3.2 Model Deployment

The model is such that the end user only need to run it from the command prompt. The unnecessary training should not be done over and over again just for the prediction purpose. The bike_main.py file takes care of this. In case of R bike_main.R takes care of this particular job.The final model is saved into disk and loaded at the time of prediction. The “Submission_data.csv” file used for this purpose and “Predic_submission.csv” is our final outcome. In case of R "predict_r_submission.csv" is the submission data and "predict_r_submission_final.csv" is the final outcome.

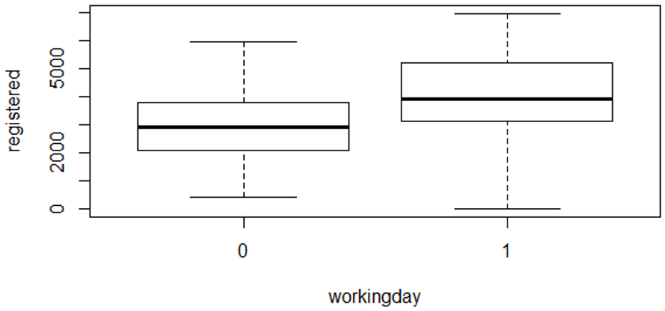


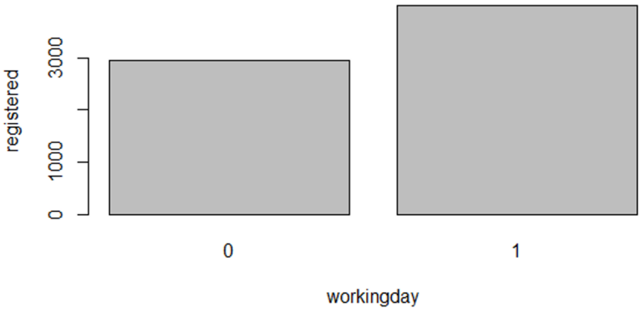
```
(base) C:\Users\Rahul\Desktop\edwisor\bike>Python bike_main.py --Data_File C:/Users/Rahul/Desktop/edwisor/bike/Submission_data.csv
Execution started
prediction has been made successfully
```

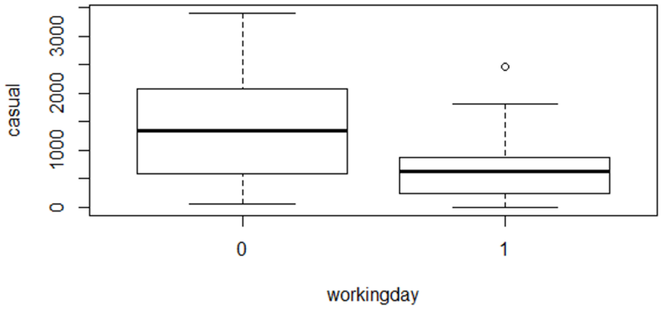
4 Chapter 4

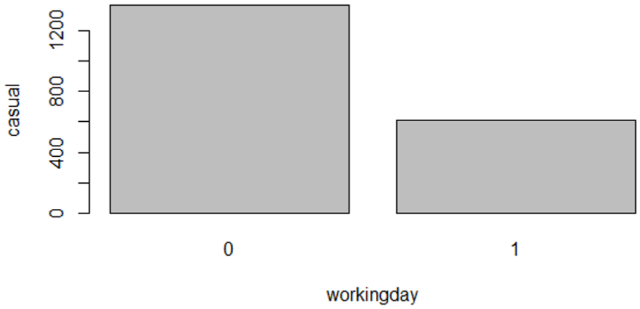
Appendix

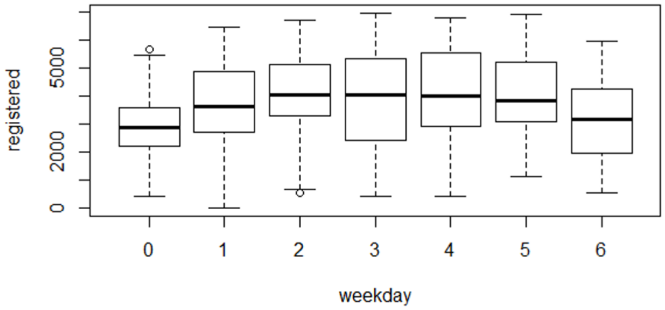
Visualization with R

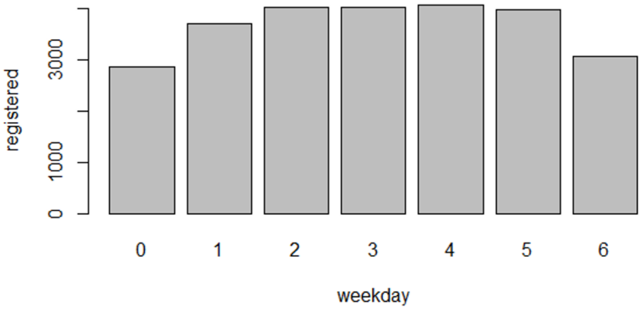


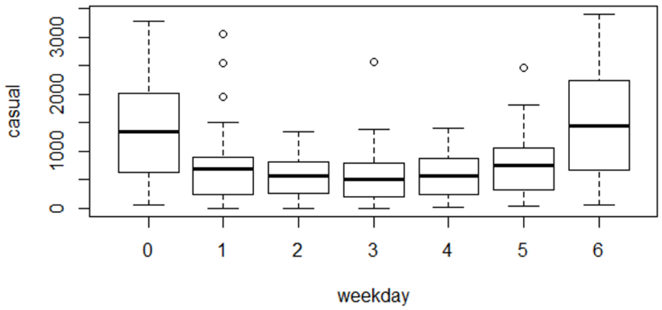


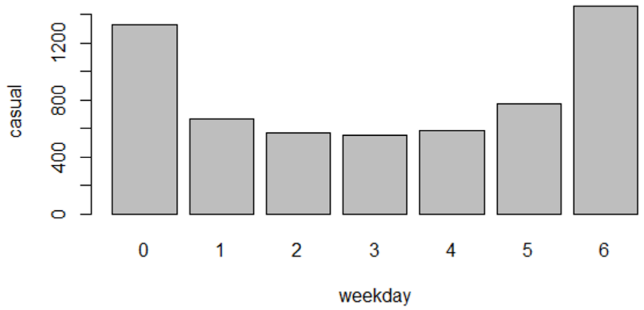


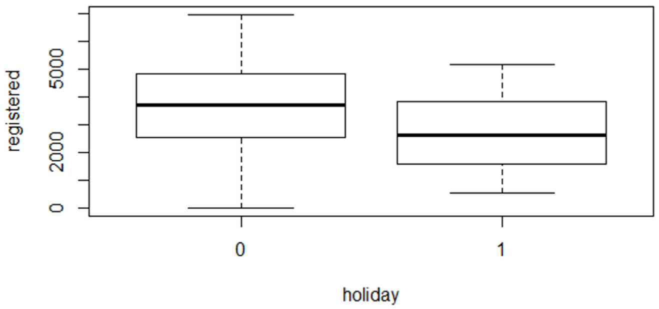


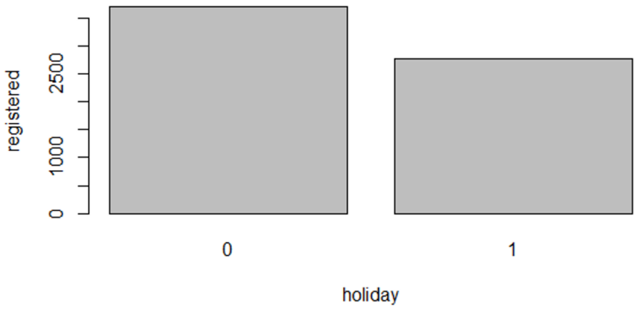


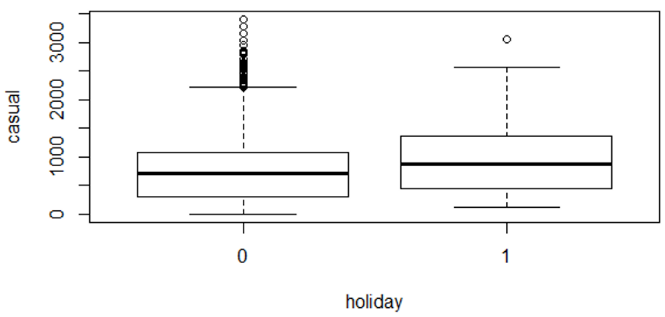


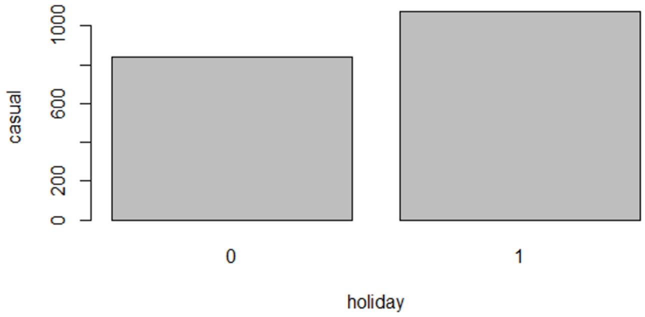

casual
0
1000
2000
3000
0
1
2
3
4
5
6
weekday

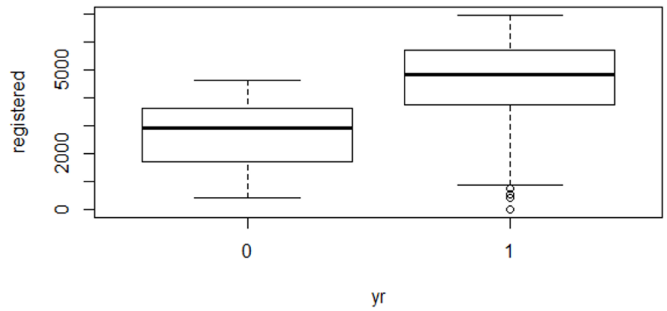

casual
0
400
800
1200
0
1
2
3
4
5
6
weekday

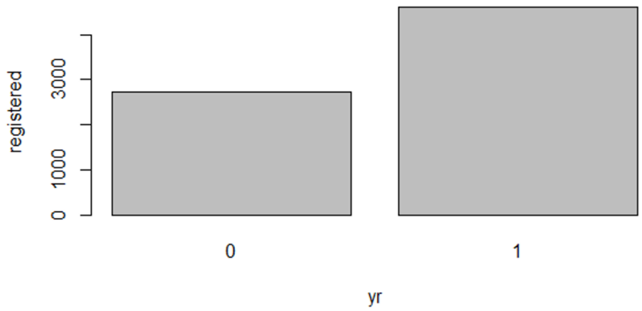

registered
0
2000
5000
0
1
holiday

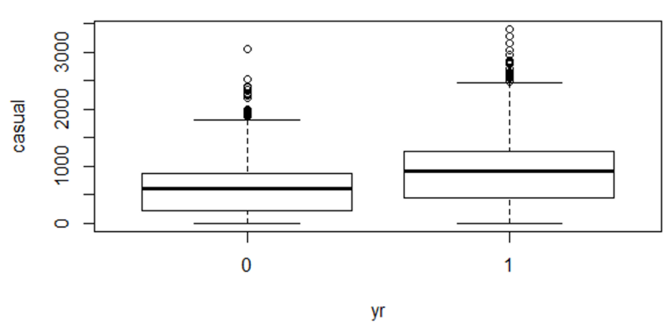

registered
0
1000
2500
0
1
holiday

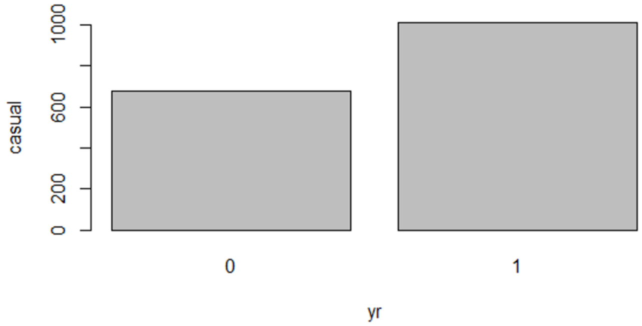

casual
0
1000
2000
3000
0
1
holiday

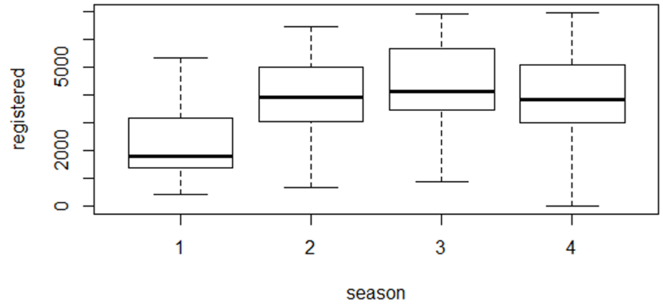

casual
0
200
600
1000
0
1
holiday

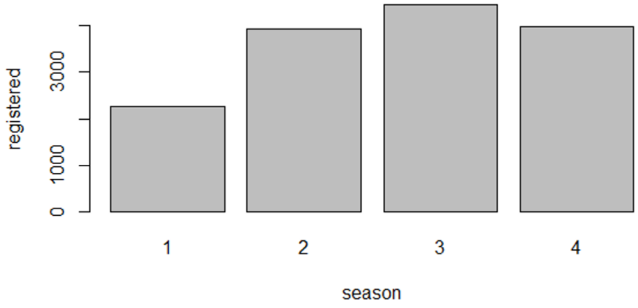

registered
0
2000
5000
0
1
yr

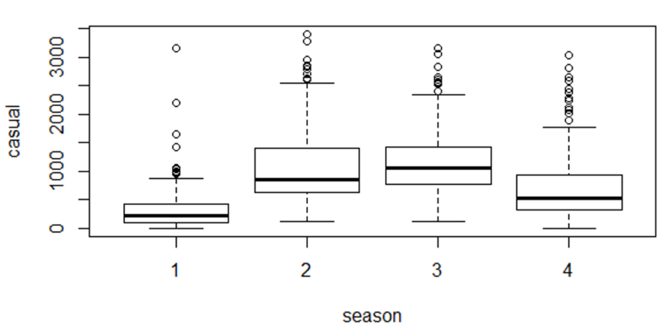

registered
0
1000
3000
0
1
yr

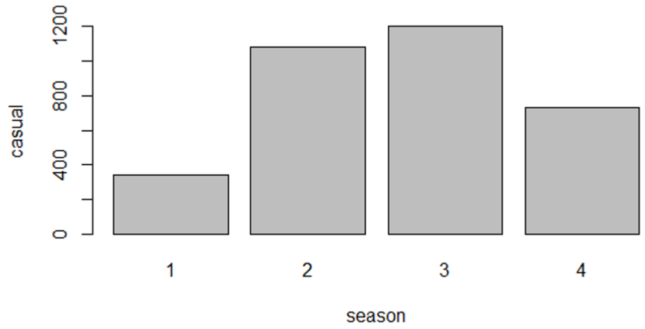

casual
0
1000
2000
3000
0
1
yr

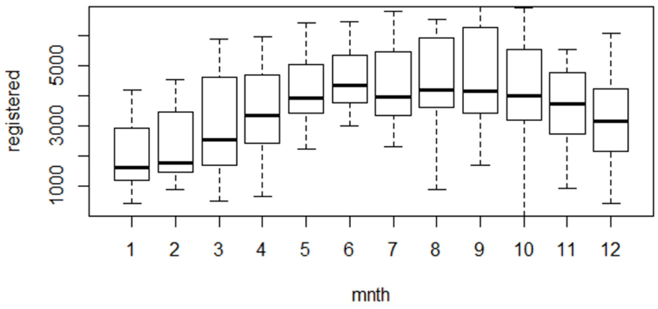

casual
0
200
600
1000
0
1
yr

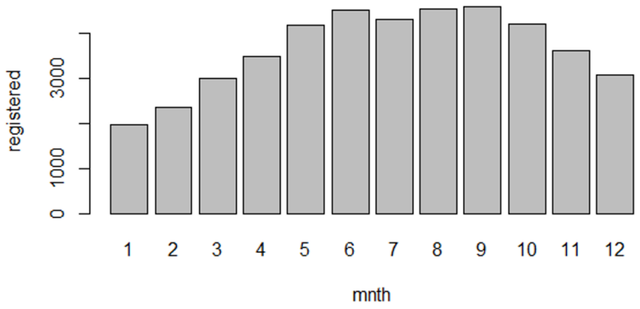

registered
0
2000
5000
1
2
3
4
season

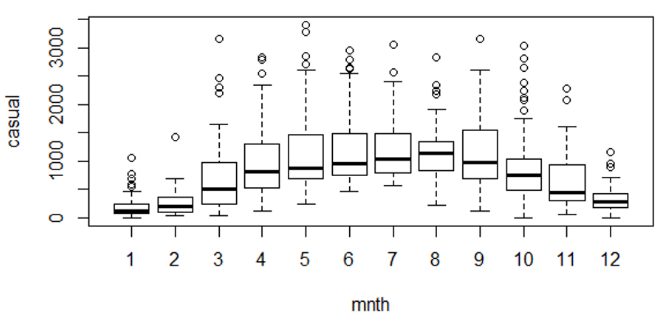

registered
0
1000
3000
1
2
3
4
season

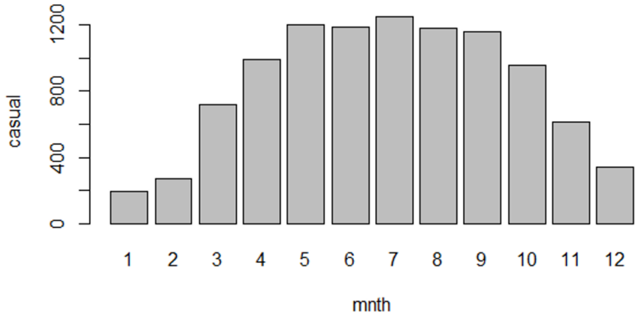

casual
0
1000
2000
3000
1
2
3
4
season

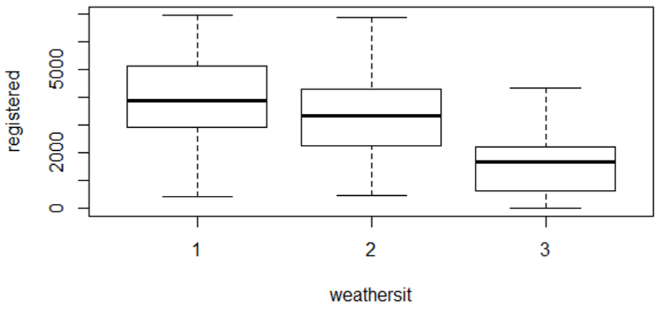

casual
0
400
800
1200
1
2
3
4
season

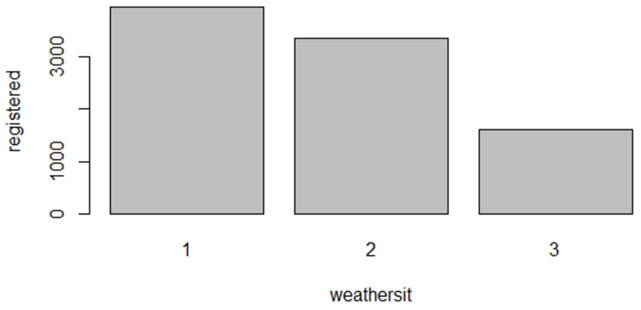

registered
1000
3000
5000
1
2
3
4
5
6
7
8
9
10
11
12
mnth

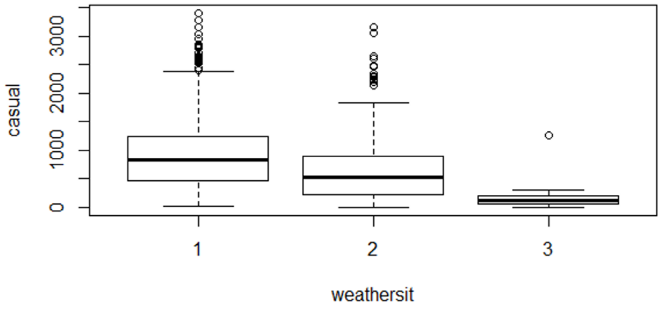

registered
0
1000
3000
1
2
3
4
5
6
7
8
9
10
11
12
mnth

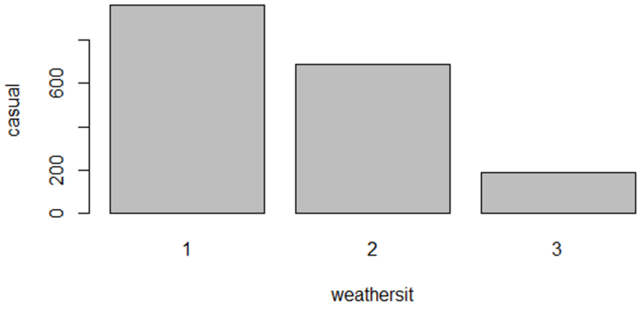

casual
0
1000
2000
3000
1
2
3
4
5
6
7
8
9
10
11
12
mnth

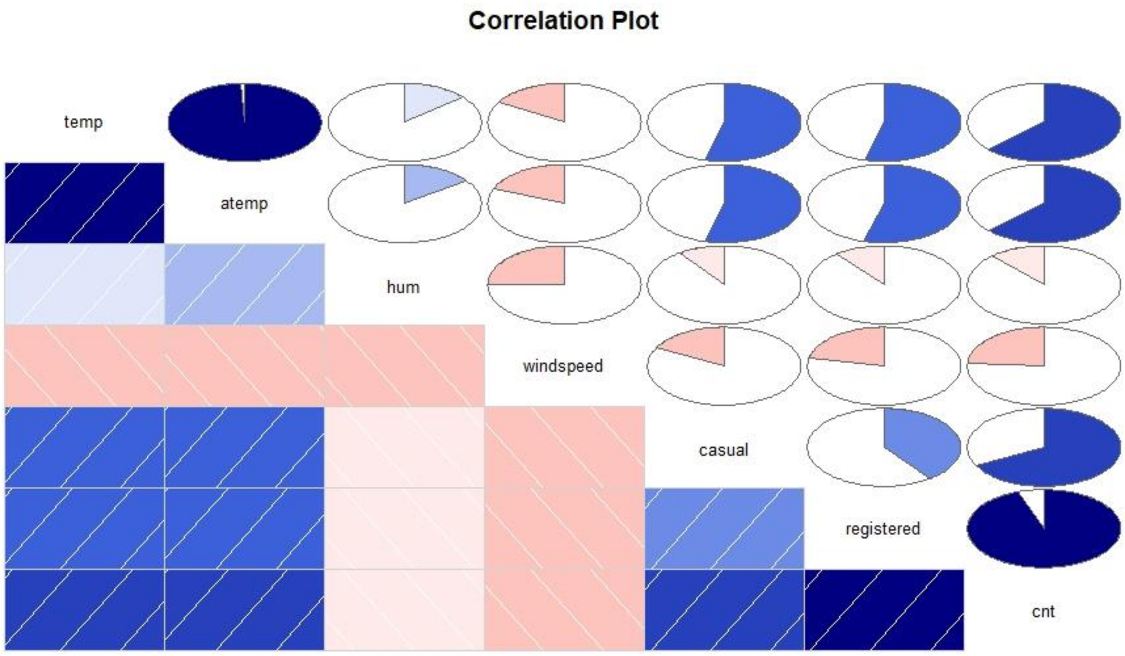

casual
0
400
800
1200
1
2
3
4
5
6
7
8
9
10
11
12
mnth

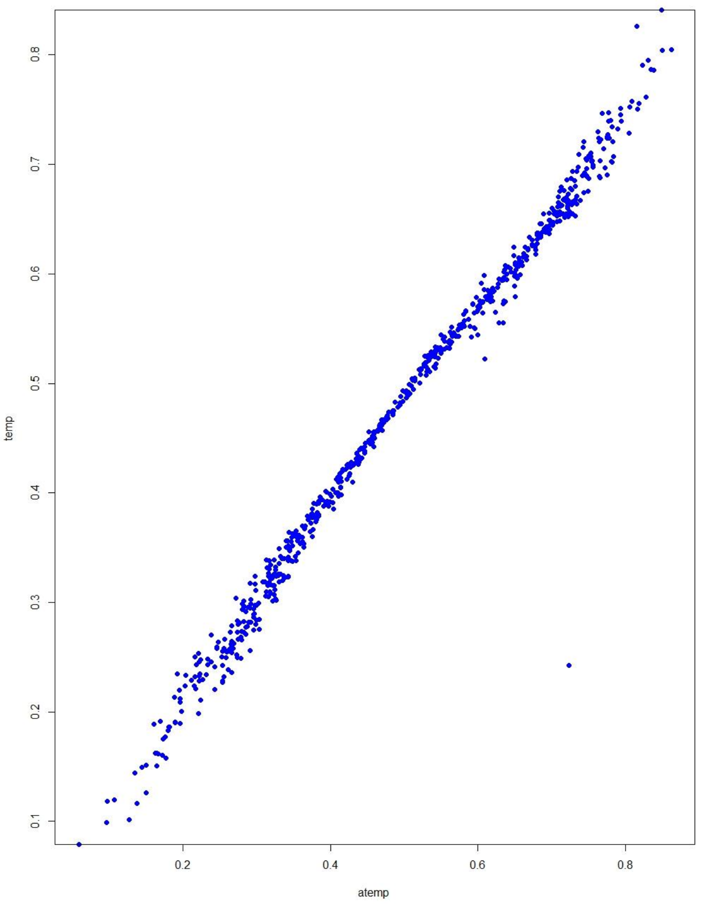

registered
0
2000
5000
1
2
3
weathersit

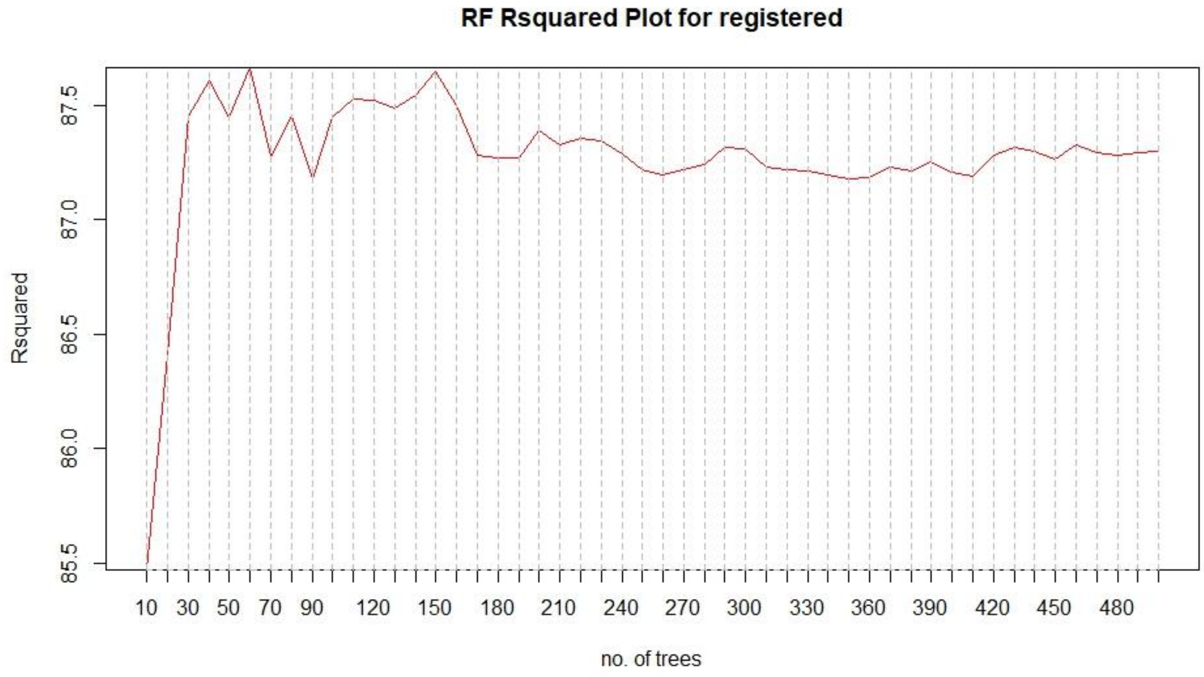

registered
0
1000
3000
1
2
3
weathersit

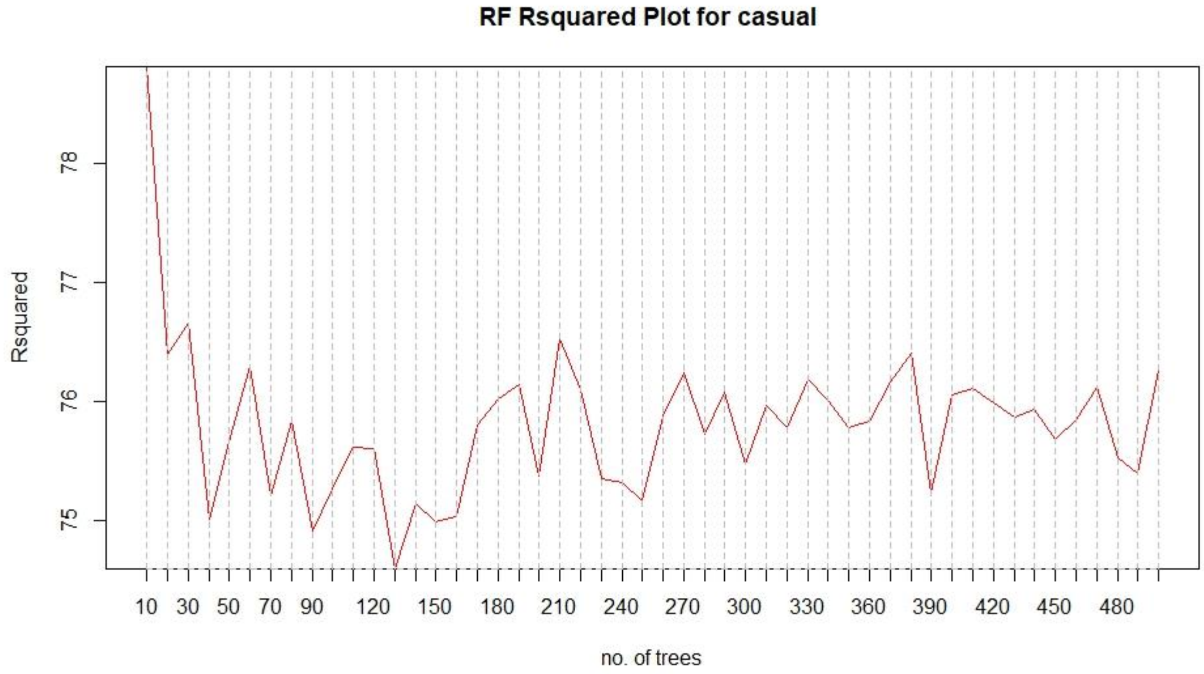

casual
0
1000
2000
3000
1
2
3
weathersit

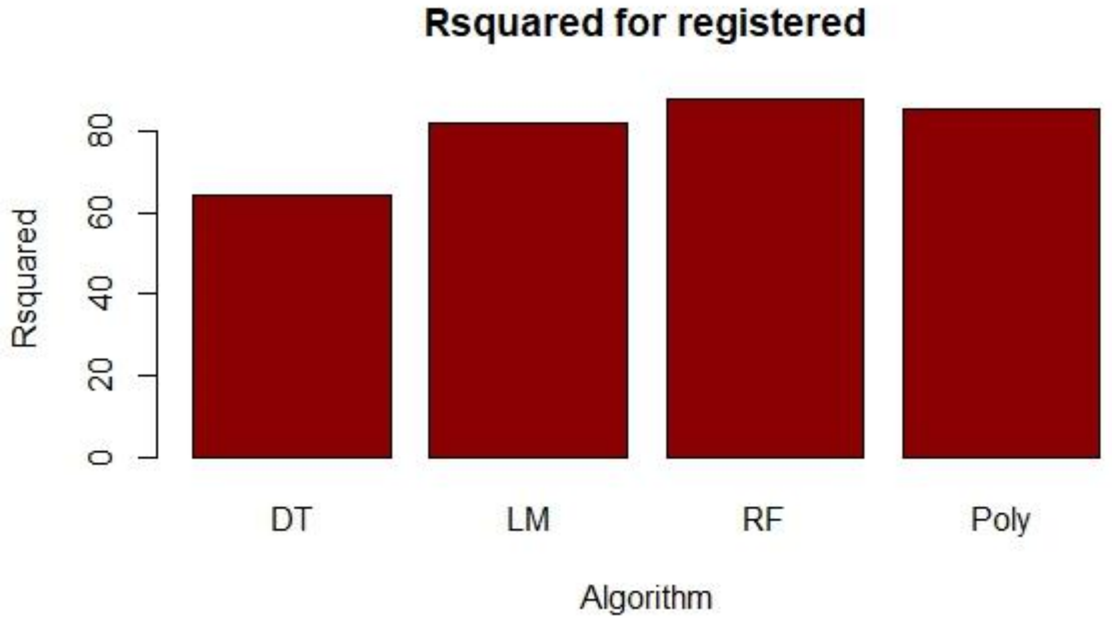

casual
0
200
600
1
2
3
weathersit

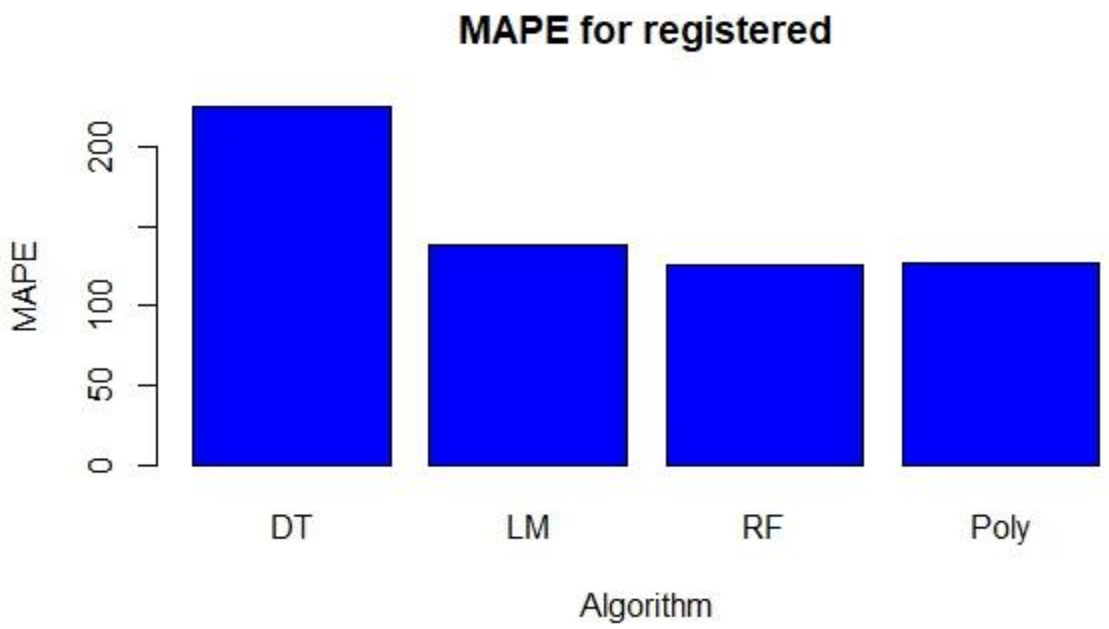

Correlation Plot
temp
atemp
hum
windspeed
casual
registered
cnt

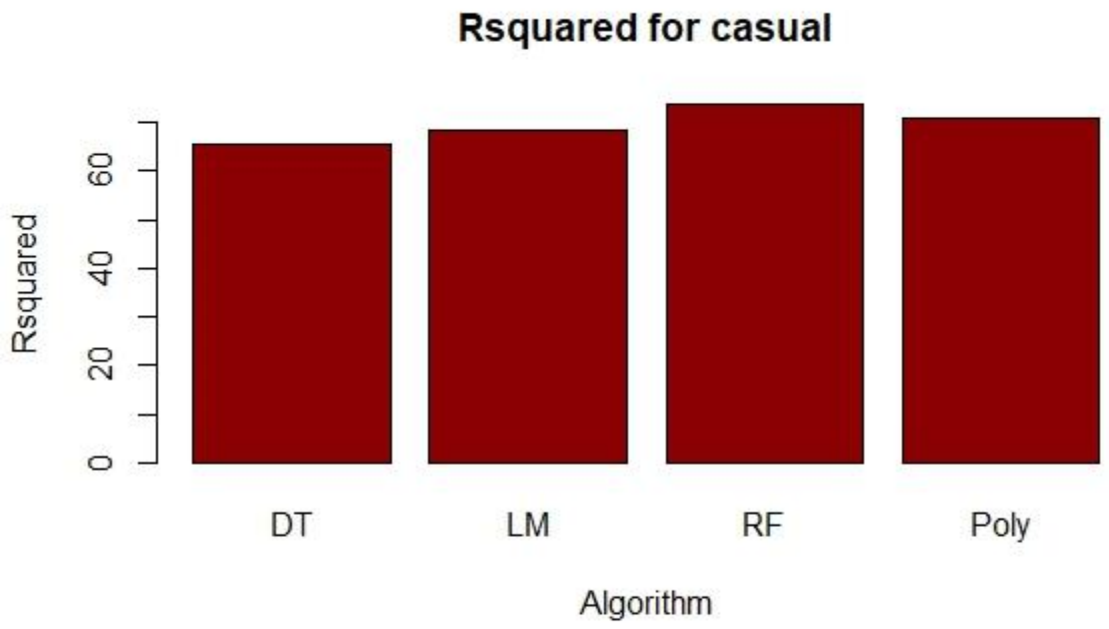

temp
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.2
0.4
0.6
0.8
atemp

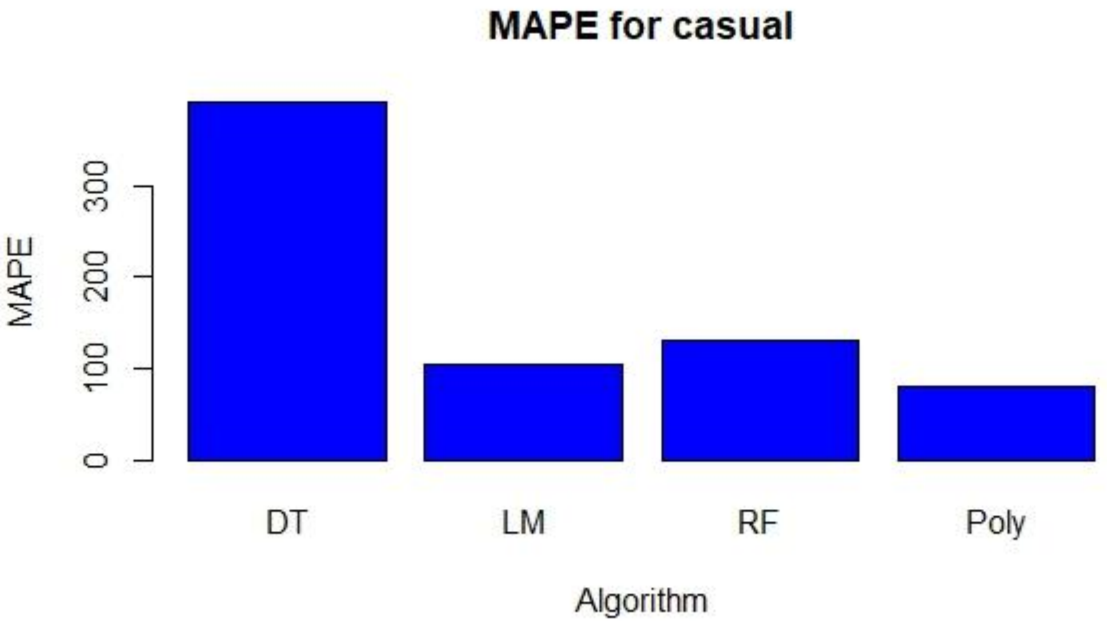

RF Rsquared Plot for registered
Rsquared
85.5
86.0
86.5
87.0
87.5
10
30
50
70
90
120
150
180
210
240
270
300
330
360
390
420
450
480
no. of trees

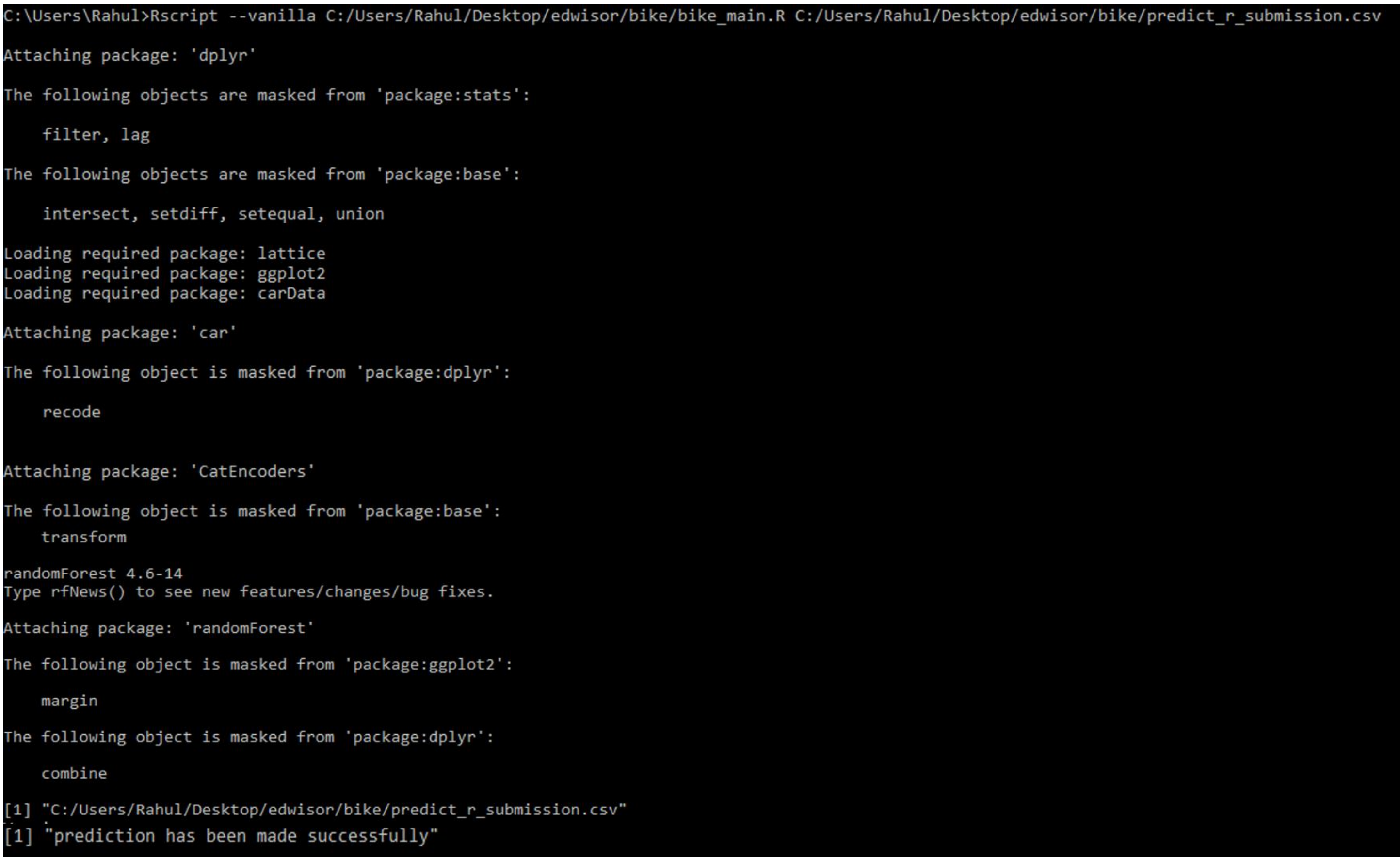

RF Rsquared Plot for casual
Rsquared
75
76
77
78
10
30
50
70
90
120
150
180
210
240
270
300
330
360
390
420
450
480
no. of trees


Rsquared for registered
Rsquared
0
20
40
60
80
DT
LM
RF
Poly
Algorithm


MAPE for registered
MAPE
0
50
100
200
DT
LM
RF
Poly
Algorithm


Rsquared for casual
Rsquared
0
20
40
60
DT
LM
RF
Poly
Algorithm





```
C:\Users\Rahul>Rscript --vanilla C:/Users/Rahul/Desktop/edwisor/bike/bike_main.R C:/Users/Rahul/Desktop/edwisor/bike/predict_r_submission.csv

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Loading required package: lattice
Loading required package: ggplot2
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode


Attaching package: 'CatEncoders'

The following object is masked from 'package:base':
    transform

randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:ggplot2':

    margin

The following object is masked from 'package:dplyr':

    combine

[1] "C:/Users/Rahul/Desktop/edwisor/bike/predict_r_submission.csv"
[1] "prediction has been made successfully"
```