Predicting Bike Rental Count

Srijith Rajeev

21st Oct 2018

Contents

Chapter 1

Introduction

1.1 Problem Statement

The objective of this project is to predict the daily count of bike rental based on environment and seasonal settings.

1.2 Data

A sample of the data set is shown below which is used to predict the count of bike rental which depends upon environmental and seasonal factors.

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	1/1/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	1/2/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	1/3/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	1/4/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	1/5/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	1/6/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
7	1/7/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	1/8/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	1/9/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
10	1/10/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321

Here we are given 16 variables. Of these, 13 variables are predictor variables and 3 variables are dependent variables. The dependent variables are ‘casual’, ‘registered’ and ‘cnt’. The relation between ‘casual’ , ‘registered’ and ‘cnt’ is

$$cnt = casual + registered$$

The list of predictor and dependent variables is shown below

Sl No	Dependent Variables
1	casual
2	registered
3	cnt

Sl No	Predictor Variables
1	instant
2	dteday
3	season
4	yr
5	mnth
6	holiday
7	weekday
8	workingday
9	weathersit
10	temp
11	atemp
12	hum
13	windspeed

The details of the data attributes in the dataset are:

1. instant – Record index
2. dteday – Date
3. season – Season (1:Springer, 2:Summer, 3:Fall, 4:Winter)
4. yr –Year (0:2011, 1:2012)
5. mnth – Month (1 to 12)
6. holiday – whether day is holiday or not: extracted from Holiday schedule
7. weekday – Day of the week
8. workingday – If day is neither weekend nor holiday:1 otherwise 0
9. weathersit – Weather (1: Clear, Few clouds, Partly cloudy; 2: Mist+cloudy, Mist+Broken clouds, Mist+few clouds; 3:Light snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain+ Scattered clouds; 4:Heavy Rain + Ice Pellets + Thunderstorm +Mist, Snow+Fog)
10. temp – Normalized temperature in Celsius
11. atemp – Normalized feeling temperature in Celsius
12. hum – Normalized humidity
13. windspeed – Normalized wind speed
14. casual – count of casual users
15. registered – count of registered users
16. cnt - count of total rental bikes including casual and registered.

The data is having numerical and categorical variables. The numeric variables are

- instant
- temp
- atemp
- hum
- windspeed

Chapter 2

Methodology

2.1 Pre Processing

The data has to be explored, cleaned, and visualized before doing predictive modeling, which is often, termed Exploratory data analysis.

2.1.1 Data **Summary**

The summary of the data is shown below:

```
    instant              dteday       season       yr          mnth
 Min.   :  1.0   2011-01-01:  1   Spring:181   2011:365   Jan    : 62
 1st Qu.:183.5   2011-01-02:  1   Summer:184   2012:366   Mar    : 62
 Median :366.0   2011-01-03:  1   Fall  :188              May    : 62
 Mean   :366.0   2011-01-04:  1   Winter:178              Jul    : 62
 3rd Qu.:548.5   2011-01-05:  1                           Aug    : 62
 Max.   :731.0   2011-01-06:  1                           Oct    : 62
                 (Other)   :725                           (Other):359
 holiday     weekday      workingday     weathersit        temp
 No :710   Sun:105   Min.   :0.000   Clear:463   Min.   :0.05913
 Yes: 21   Mon:105   1st Qu.:0.000   Mist :247   1st Qu.:0.33708
           Tue:104   Median :1.000   Light: 21   Median :0.49833
           Wed:104   Mean   :0.684   Heavy:  0   Mean   :0.49538
           Thu:104   3rd Qu.:1.000               3rd Qu.:0.65542
           Fri:104   Max.   :1.000               Max.   :0.86167
           Sat:105
     atemp              hum           windspeed          casual
 Min.   :0.07907   Min.   :0.0000   Min.   :0.02239   Min.   :   2.0
 1st Qu.:0.33784   1st Qu.:0.5200   1st Qu.:0.13495   1st Qu.: 315.5
 Median :0.48673   Median :0.6267   Median :0.18097   Median : 713.0
 Mean   :0.47435   Mean   :0.6279   Mean   :0.19049   Mean   : 848.2
 3rd Qu.:0.60860   3rd Qu.:0.7302   3rd Qu.:0.23321   3rd Qu.:1096.0
 Max.   :0.84090   Max.   :0.9725   Max.   :0.50746   Max.   :3410.0

   registered       cnt
 Min.   :  20   Min.   :  22
 1st Qu.:2497   1st Qu.:3152
 Median :3662   Median :4548
 Mean   :3656   Mean   :4504
 3rd Qu.:4776   3rd Qu.:5956
 Max.   :6946   Max.   :8714
```

The mean, median, minimum, maximum and quartiles of the numeric variables are shown in the summary. For categorical variables, the number of observations in each category is shown.

2.1.2 Missing Value Analysis

In the given data, there are no missing values. So there is no necessity for missing value analysis at this stage.

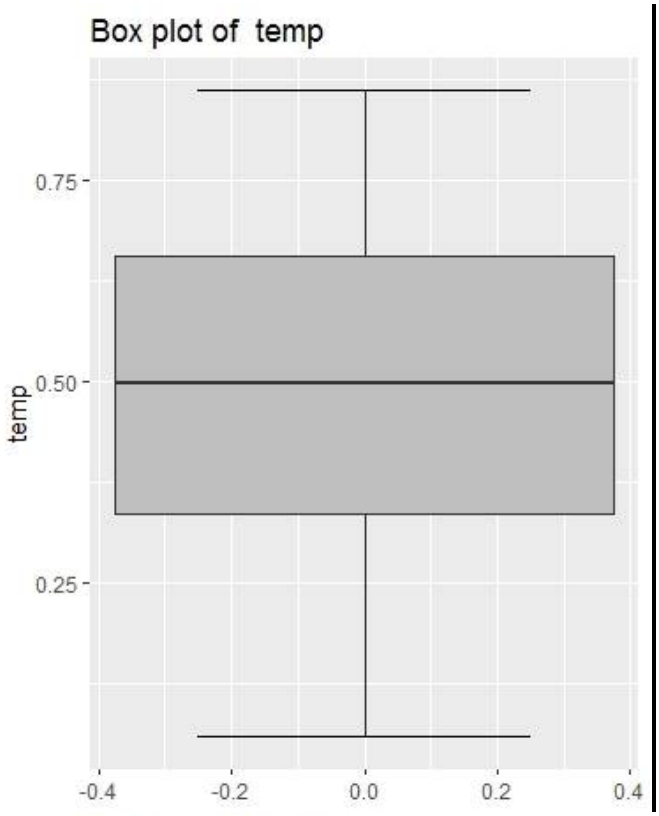
2.1.3 Redundant variable Removal

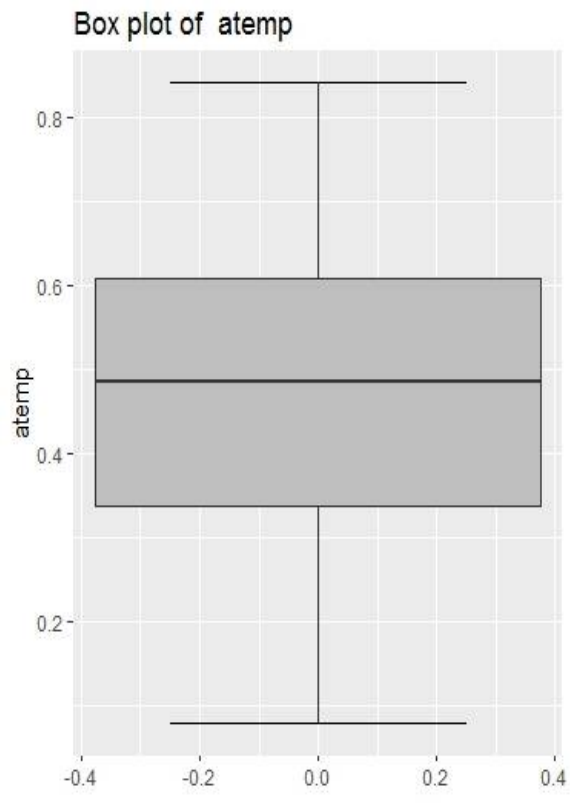
The variable ‘instant’ refers to the number of days starting from 1st Jan 2011. Therefore the information in the variable ‘dteday’ is available in ‘instant’ So for the analysis, the variable ‘dteday’ is redundant.

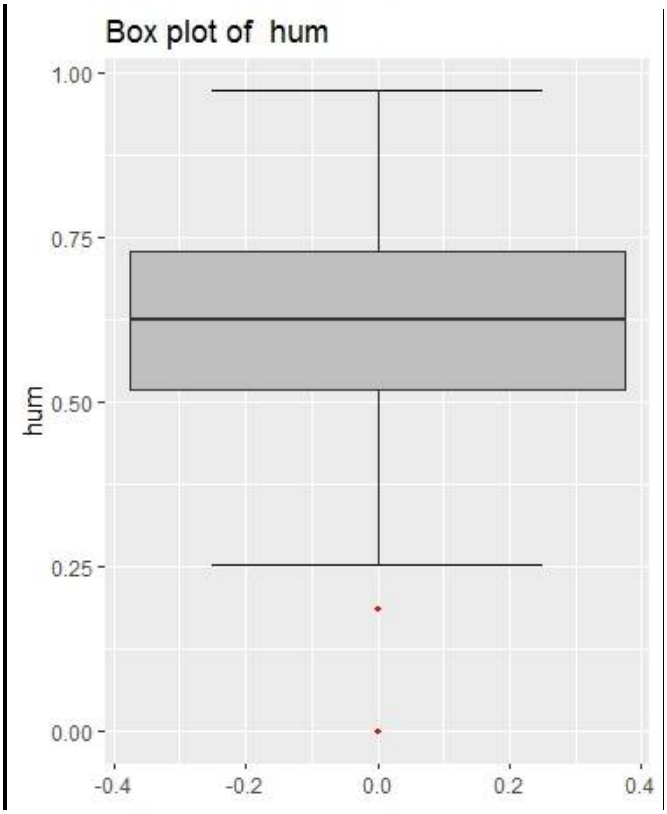
The variable ‘workingday’ is defined in terms of the variables ‘weekday’ and ‘holiday’. Therefore these three variables are perfectly multi collinear. Therefore one of these variables is redundant and can be removed. Here variable ‘workingday’ is removed.

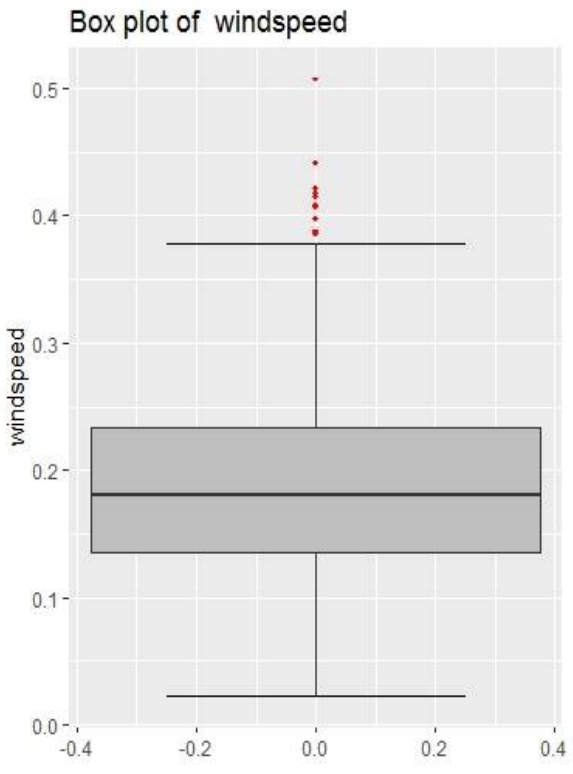
2.1.4 Outlier Analysis

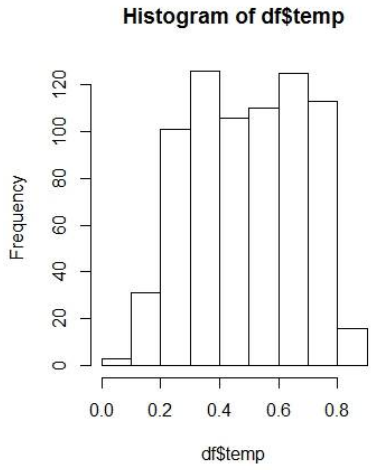
We visualize the outliers using boxplot. The boxplot of the numeric predictor variables are shown below. Here the variables ‘hum’ and ‘windspeed’ are having outliers. Two outliers are present in the variable ‘hum’, while ‘windspeed’ is having 13 outliers. The outliers are replaced with ‘NA’ and imputed by KNN imputation. The histogram of all numeric variables is plotted.

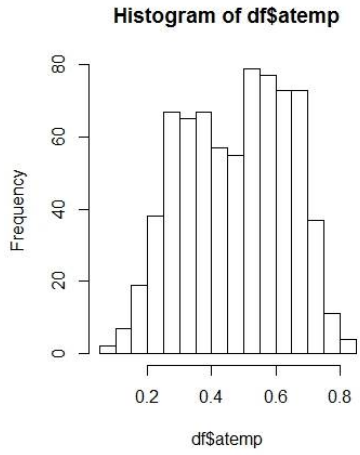

Box plot of temp
temp
0.75
0.50
0.25
-0.4
-0.2
0.0
0.2
0.4

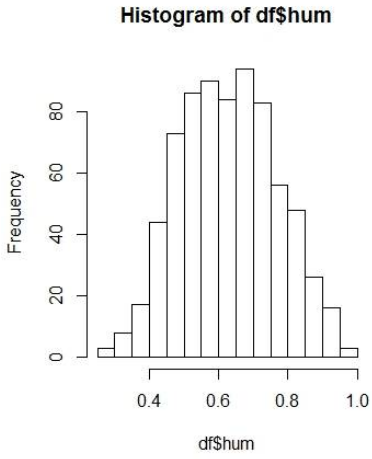

Box plot of atemp
atemp
0.8
0.6
0.4
0.2
-0.4
-0.2
0.0
0.2
0.4

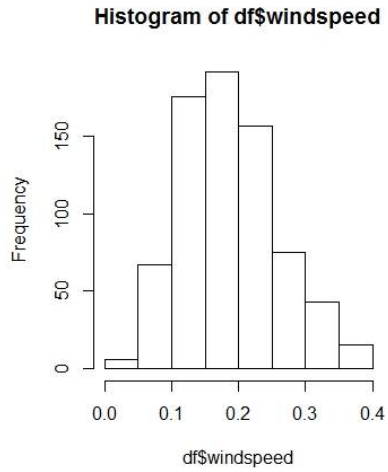

Box plot of hum
hum
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4

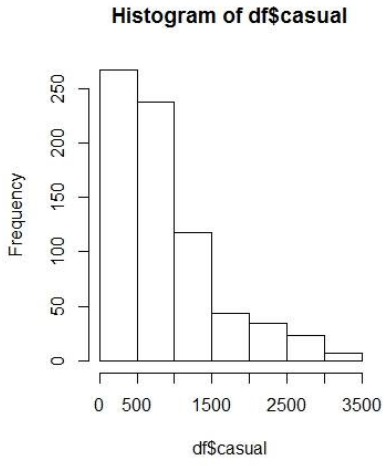

Box plot of windspeed
windspeed
0.5
0.4
0.3
0.2
0.1
0.0
-0.4
-0.2
0.0
0.2
0.4

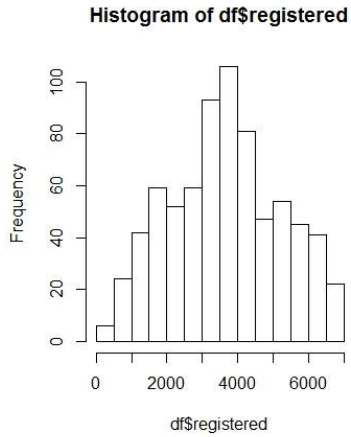

Histogram of df$temp
Frequency
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
df$temp

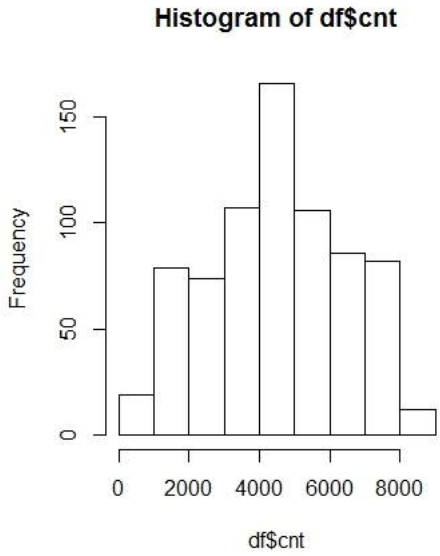

Histogram of df$atemp
Frequency
0
20
40
60
80
0.2
0.4
0.6
0.8
df$atemp


Histogram of df$hum
Frequency
0
20
40
60
80
0.4
0.6
0.8
1.0
df$hum


Histogram of df$windspeed
Frequency
0
50
100
150
0.0
0.1
0.2
0.3
0.4
df$windspeed


Histogram of df$casual
Frequency
0
50
100
150
200
250
0
500
1500
2500
3500
df$casual


Histogram of df$registered
Frequency
0
20
40
60
80
100
0
2000
4000
6000
df$registered

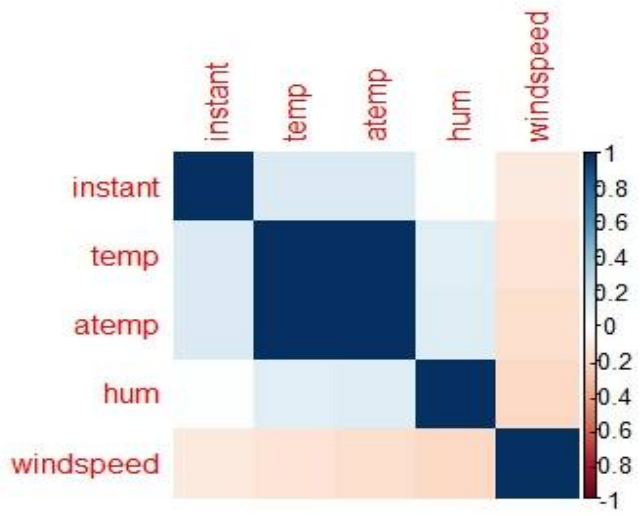

Histogram of df$cnt
Frequency
0
50
100
150
0
2000
4000
6000
8000
df$cnt

2.1.5 Correlation Analysis

The correlation between the numeric variables is studied. The correlation matrix is

	instant	temp	atemp	hum	windspeed
instant	1.0000000000	0.1505803	0.1526382	0.0001455301	-0.1150960
temp	0.1505803019	1.0000000	0.9917016	0.1229063822	-0.1479808
atemp	0.1526382379	0.9917016	1.0000000	0.1365301930	-0.1744354
hum	0.0001455301	0.1229064	0.1365302	1.0000000000	-0.2098121
windspeed	-0.1150959539	-0.1479808	-0.1744354	-0.2098121190	1.0000000

Also the correlation plot is also plotted.



From the correlation analysis, it's evident that the variable 'temp' and 'atemp' are having a high degree of correlation. So for further analysis, the variable 'temp' is neglected.

2.1.6 Train – Test Data

For further analysis after model fitting, the data is divided into train data and test data. The model is trained on train data and its performance is evaluated on test data.

2.2 Modelling

2.2.1 Model Selection

The objective is to predict the count of bike renting on any particular day given environmental and seasonal settings. This is a case of Regression Problem. The models to be fitted on this dataset are

- Multiple Linear Regression
- Decision Tree
- Random Forest

Train data is inputted to the regression model and two separate analyses are done to predict the variable 'cnt'.

- One method is to predict the variables 'casual' and 'registered' and add them to get the prediction of 'cnt'.
- Another method is to directly predict the variable 'cnt'.

For decision tree and random forest, if predicting for the future dates, the variable 'instance' doesn't seem to make sense, as one branch of the tree will be unused for higher 'instance' values. So separate analysis is also done removing the independent variable 'instant'

2.2.2 Multiple Linear Regression

- Linear Regression to predict 'casual'. The anova table is

```
Analysis of Variance Table

Response: casual
            Df    Sum Sq  Mean Sq  F value    Pr(>F)
instant      1  26075342 26075342 204.2210 < 2.2e-16 ***
season       3  78955616 26318539 206.1257 < 2.2e-16 ***
yr           1   2503191  2503191  19.6049 1.104e-05 ***
mnth        11  22843917  2076720  16.2648 < 2.2e-16 ***
holiday      1   2280113  2280113  17.8578 2.695e-05 ***
weekday      6 100121773 16686962 130.6916 < 2.2e-16 ***
weathersit   2  10156560  5078280  39.7729 < 2.2e-16 ***
atemp        1   8353612  8353612  65.4251 2.652e-15 ***
hum          1    852433   852433   6.6762  0.009972 **
windspeed    1   2383505  2383505  18.6675 1.781e-05 ***
Residuals  702  89632760   127682
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Linear Regression to predict ‘registered’. The anova table is

```
Analysis of Variance Table

Response: registered
             Df    Sum Sq   Mean Sq   F value    Pr(>F)
instant       1 604306092 604306092 1684.1338 < 2.2e-16 ***
season        3 236061779  78687260  219.2926 < 2.2e-16 ***
yr            1  64322315  64322315  179.2591 < 2.2e-16 ***
mnth         11  76899828   6990893   19.4828 < 2.2e-16 ***
holiday       1  10538597  10538597   29.3699 8.928e-08 ***
weekday       6 122597734  20432956   56.9444 < 2.2e-16 ***
weathersit    2  90731608  45365804  126.4294 < 2.2e-16 ***
atemp         1  11719750  11719750   32.6616 1.792e-08 ***
hum           1   2114564   2114564    5.8931   0.01552 *
windspeed     1   6063408   6063408   16.8980 4.540e-05 ***
Residuals   555 199146820    358823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Linear Regression to predict ‘cnt’. The anova table is

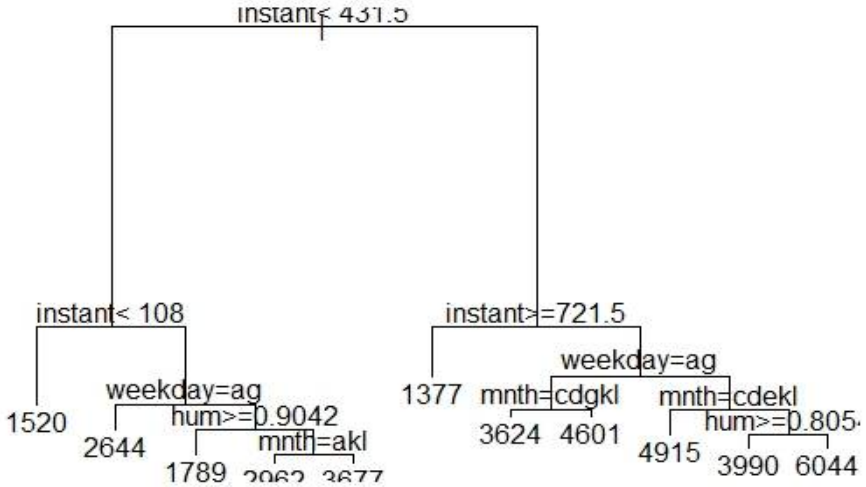
```
Analysis of Variance Table

Response: cnt
             Df    Sum Sq   Mean Sq   F value    Pr(>F)
instant       1 809269601 809269601 1369.5140 < 2.2e-16 ***
season        3 521390337 173796779  294.1135 < 2.2e-16 ***
yr            1  87810289  87810289  148.5999 < 2.2e-16 ***
mnth         11 162011645  14728331   24.9245 < 2.2e-16 ***
holiday       1   3786804   3786804    6.4083 0.0116334 *
weekday       6  14555720   2425953    4.1054 0.0004823 ***
weathersit    2 160770691  80385346  136.0348 < 2.2e-16 ***
atemp         1  32814469  32814469   55.5314 3.548e-13 ***
hum           1   5829619   5829619    9.8654 0.0017739 **
windspeed     1  15784609  15784609   26.7120 3.300e-07 ***
Residuals   555 327959146    590917
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

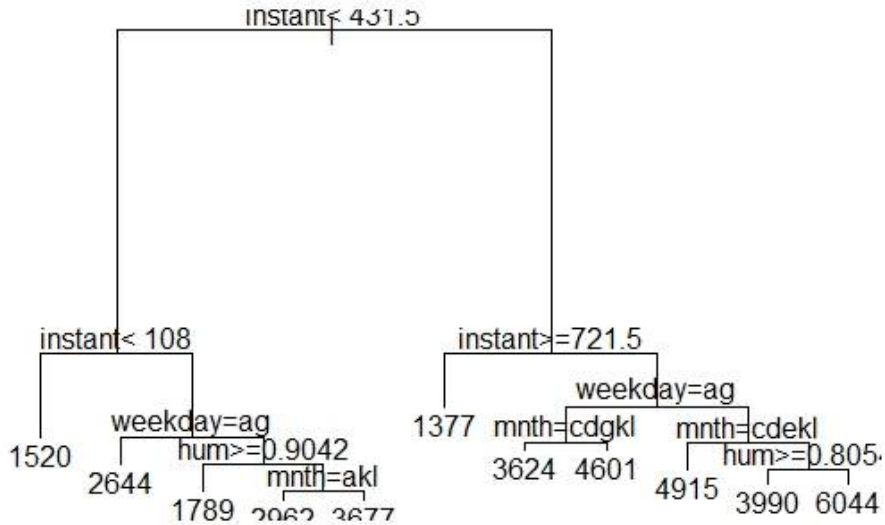
From the above anova tables, it’s evident that all variables are significant as p values are less than 0.05

2.2.3 Decision Tree

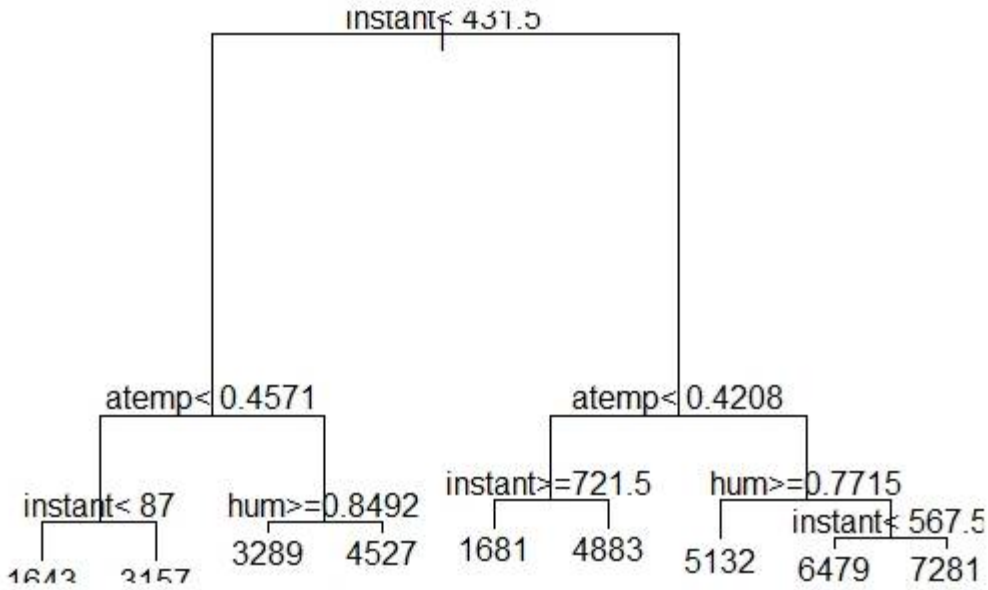
- Decision tree to predict ‘casual’



- Decision tree to predict ‘registered’



- Decision tree to predict ‘cnt’



2.2.4 Random Forest

Similar analysis is done in Random Forest also, with max number of trees limited to 200.

Chapter 3

Conclusion

3.1 Model Evaluation

Model evaluation is done by predicting the test data values, using the model which is trained in train data. MAPE (Mean Absolute Percentage Error) and MSE (Mean Square Error) are the error matrices used for the model evaluation.

Train data and test data are randomly generated in R and Python. So the results slightly vary. When entire data was used for training, the linear regression results were the same.

3.1.1 Results using R

The table having the MAPE and MSE of the three models namely Linear Regression, Decision Tree and Random forest is shown below. Two separate cases with and without the variable 'instant' is analyzed.

	Linear Regression			
	casual	**registered**	**combined**	**cnt**
MAPE	148	21	20	20
MAPE - without 'instant'	146	20	20	20
MSE	1095	2993	4644	4644
MSE - without 'instant'	1085	3012	4640	4640

	Decision Tree			
	casual	**registered**	**combined**	**cnt**
MAPE	81	23	23	25
MAPE - without 'instant'	81	26	26	22
MSE	1326	3965	5954	5924
MSE - without 'instant'	1327	4396	6068	5407

	Random Forest			
	casual	**registered**	**combined**	**cnt**
MAPE	57	17	17	18
MAPE - without 'instant'	22	17	15	17
MSE	748	2190	3373	3312
MSE - without 'instant'	128	1815	2034	2599

MSE is better error matrix than MAPE for our case. Random Forest evolved to be the best model, in comparison with other models.

3.1.1 Results using Python

The table having the MAPE and MSE of the three models namely Linear Regression, Decision Tree and Random forest is shown below. Two separate cases with and without the variable 'instant' is analyzed

	Linear Regression			
	casual	**registered**	**combined**	**cnt**
MAPE	90	17	17	17
MAPE - without 'instant'	90	17	17	17
MSE	959	2475	4051	4051
MSE - without 'instant'	958	2489	4067	4067

	Decision Tree			
	casual	**registered**	**combined**	**cnt**
MAPE	101	37	32	35
MAPE - without 'instant'	101	38	34	36
MSE	1858	7595	8490	9365
MSE - without 'instant'	1858	6977	8343	9499

	Random Forest			
	casual	**registered**	**combined**	**cnt**
MAPE	61	22	18	18
MAPE - without 'instant'	58	25	21	22
MSE	1152	3434	3432	2826
MSE - without 'instant'	1034	3941	4360	4382

MSE is better error matrix than MAPE for our case. Random Forest evolved to be the best model, in comparison with other models.