Predicting Bike Rental Count

Priyanka Thakare.

Contents:

Chapter 1.

Introduction

1.1 Problem statement-

The objective of this Case is to Predict bike rental count on daily based on the environmental and seasonal settings. The details of data attributes in the dataset are as follows

1.2 Data.

Our task is to build regressor model which will predict the count based on different factors.

Table 1.1 sample of Day.csv file.

Columns 1: 8

Instant	dteday	Season	yr	Mnth	holiday	weekday	workingday
1	01-01-2011	1	0	1	0	6	0
2	02-01-2011	1	0	1	0	0	0
3	03-01-2011	1	0	1	0	1	1
4	04-01-2011	1	0	1	0	2	1
5	05-01-2011	1	0	1	0	3	1
6	06-01-2011	1	0	1	0	4	1
7	07-01-2011	1	0	1	0	5	1

Columns 9:16

weathersit	temp	Atemp	Hum	windspeed	casual	registered	cnt
2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	0.363478	0.353739	0.696087	0.248539	131	670	801
1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
1	0.2	0.212122	0.590435	0.160296	108	1454	1562
1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
2	0.196522	0.208839	0.498696	0.168726	148	1362	1510

From the above table we can see there are total 16 variables and 731 rows. Out of which only first 7 are shown in the above table. As we can see from the problem statement , we

have to predict the count of the bike , so ‘cnt’ is our dependent variable. Rest are predictor variables.

Table : Predictor variables

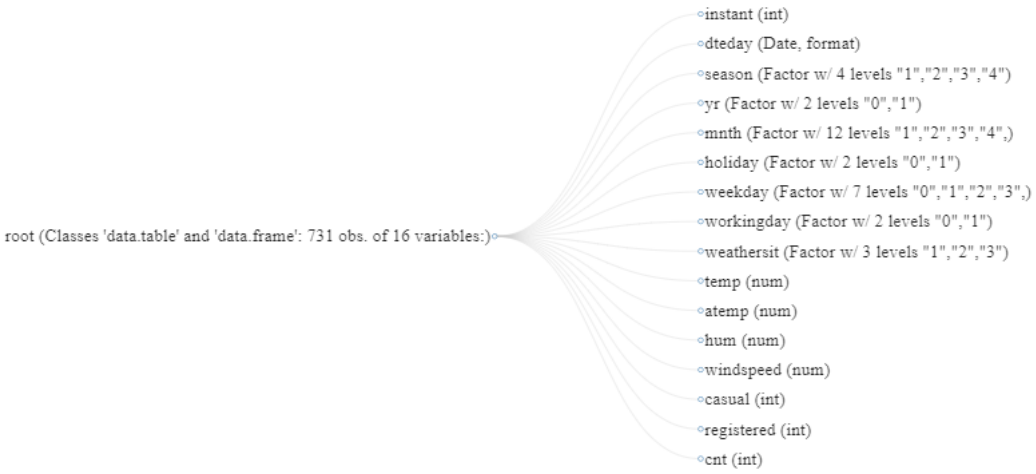
Instant	Temp
Dteday	Atemp
Season	Hum
Yr	Weahersit
Mnth	Windspeed
Holiday	Casual
Weekday	Registered
workingday	

Chapter 2 Methodology:

2.1 Data Preprocessing:

2.1.1 Exploratory Data Analysis:

Using the Data explorer library in r we can look at the structure of the data . After looking at the data we can differentiate data in below categories. Continuous variables and categorical variables.



Continuous	Categorical
Temp	Holiday
Atemp	Season
Hum	Weathersit
Windspeed	Workingday
	Yr
	Mnth
	Weekday

Let us plot the histogram for continuous variables and bar chart for categorical variables.
For detail data report refer to the Data profiling report generated by r studio.

Fig 2.1 Histogram plots for continuous variables.

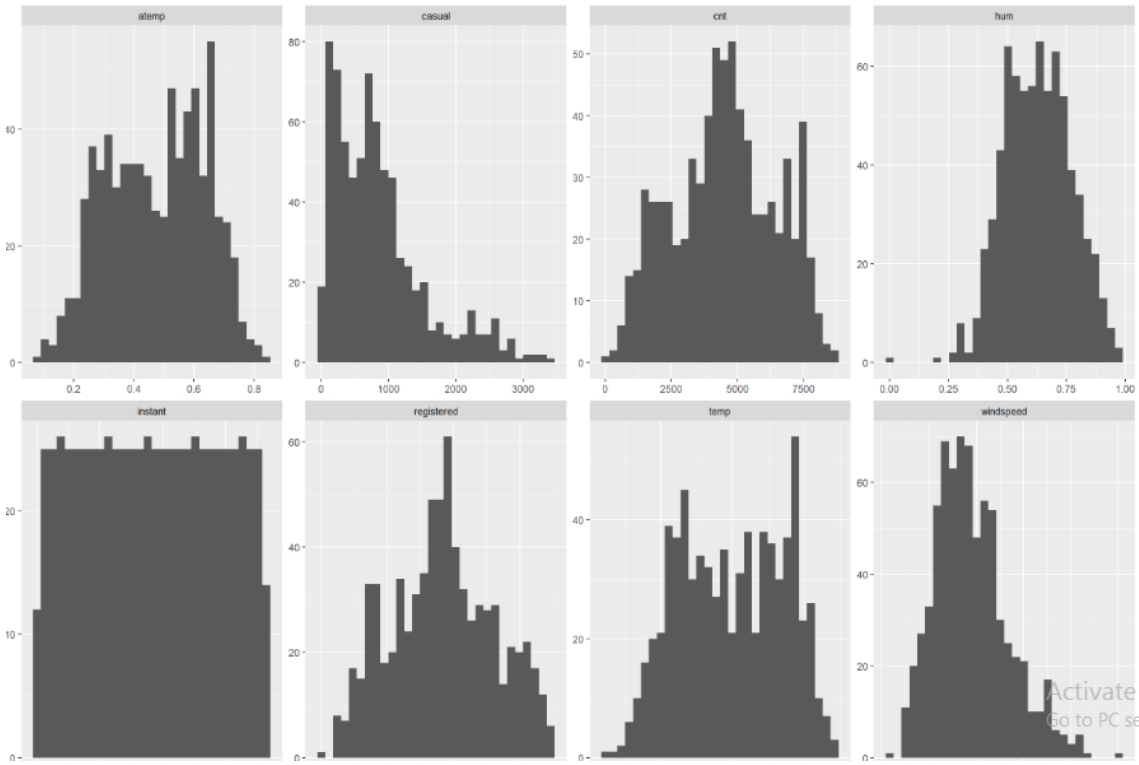
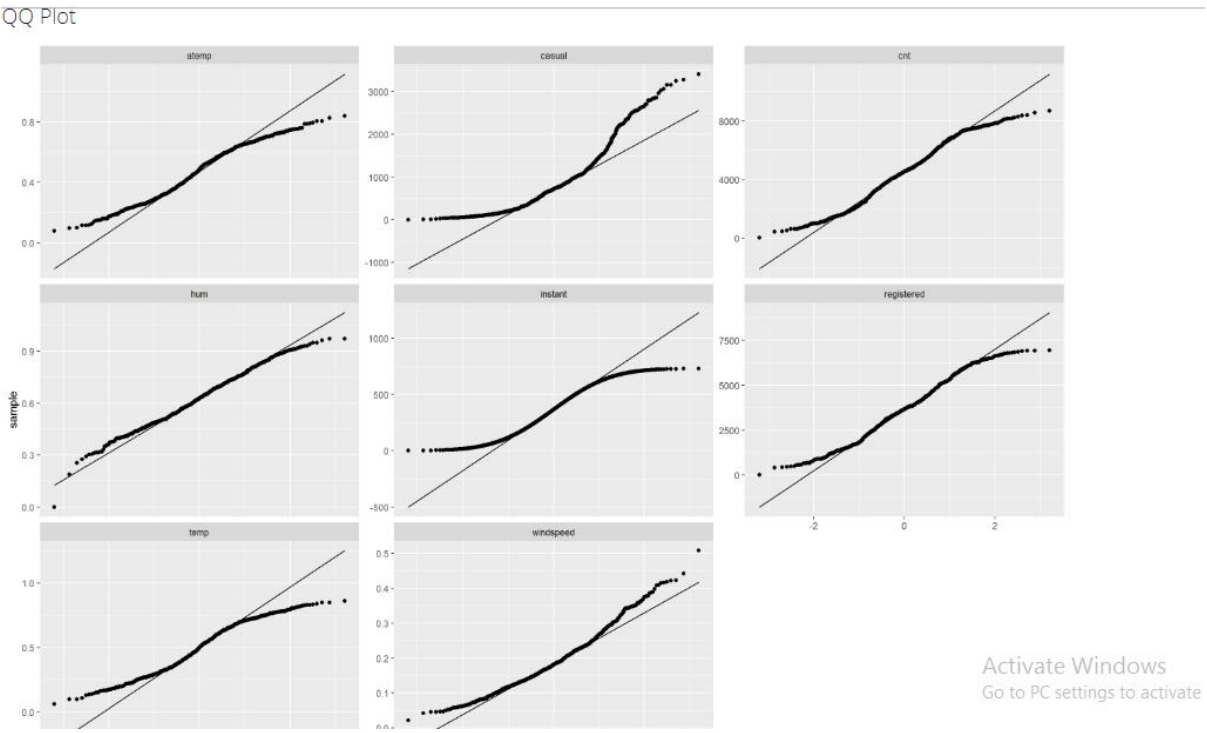


Fig 2.2 : Bar plots for categorical variables.

Below is the QQ plot to understand how values of continuous variables are varying.

Fig 2.3: QQ plots for continuous variables:



2.1.2 Missing value analysis:

No missing values are present .

2.1.3: Outlier analysis:

Let us plot box plots to understand if there is presence of outliers.

Fig 2.4 outlier plots for tem and atemp

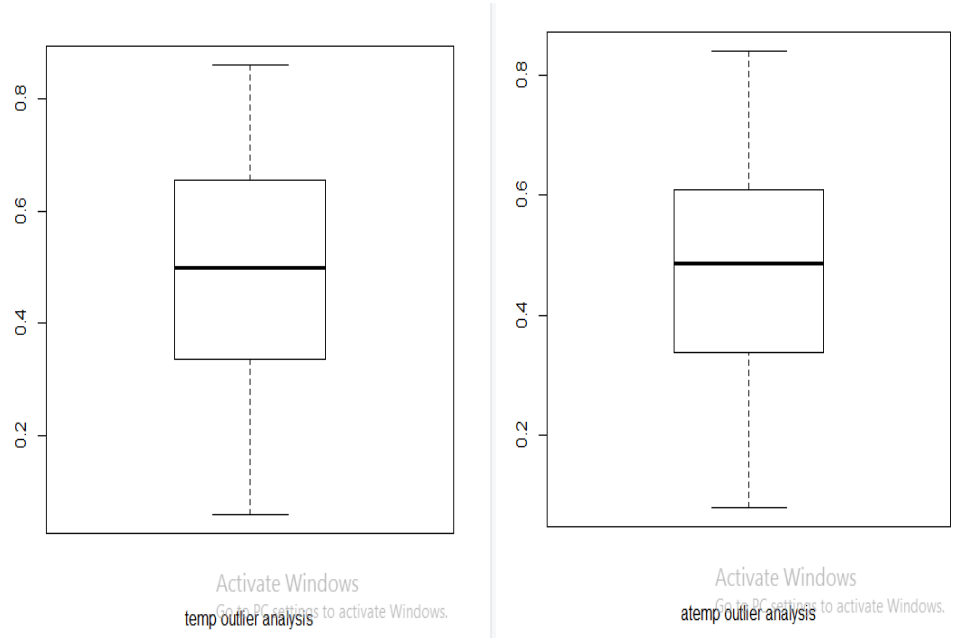
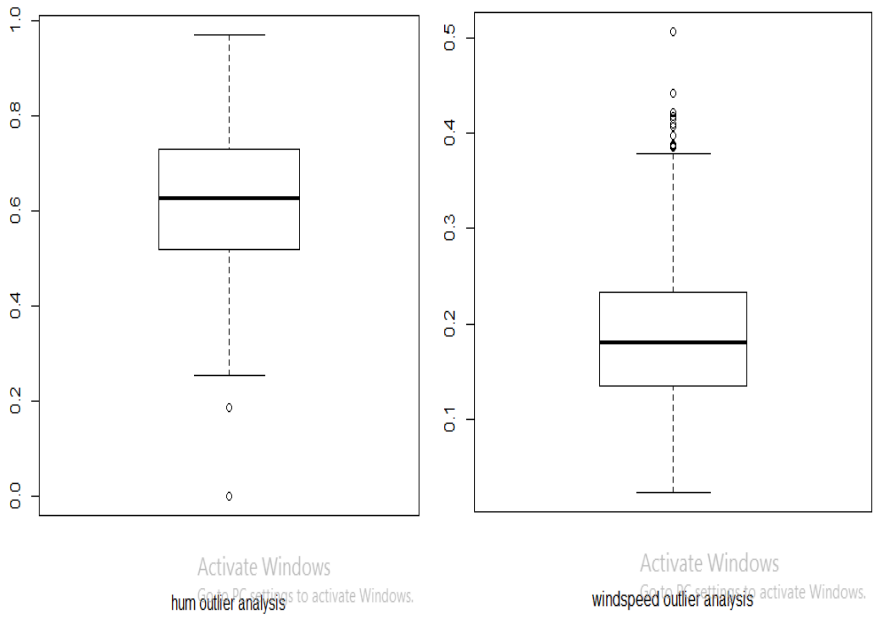
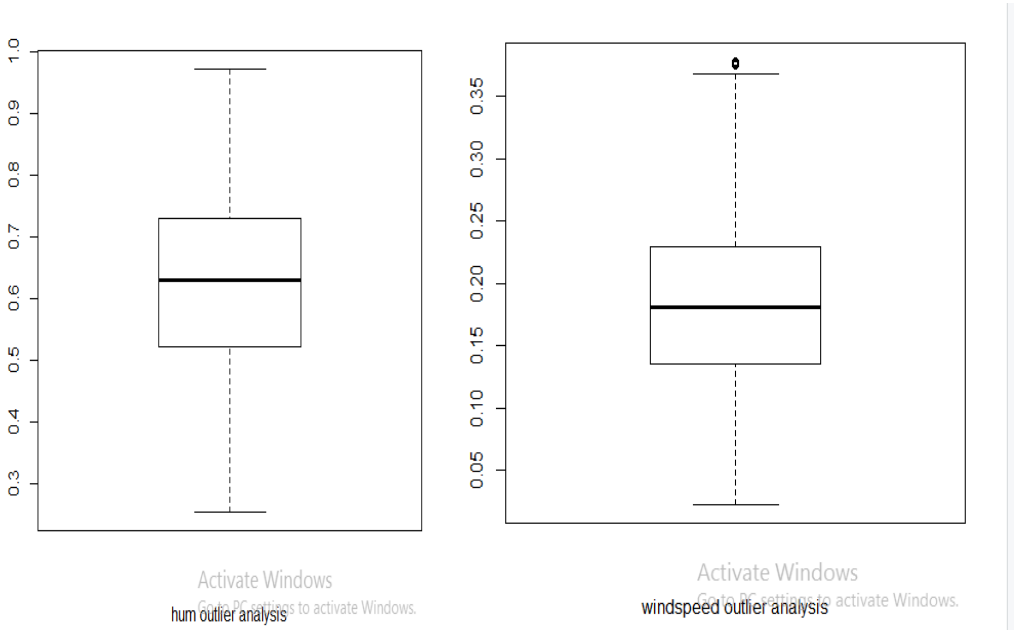


Fig 2.5 Outlier plot for hum and windspeed.



As we can see outliers are present in only hum and windspeed variables. Rest of the variab
outliers.
Mean method is used to remove outliers.

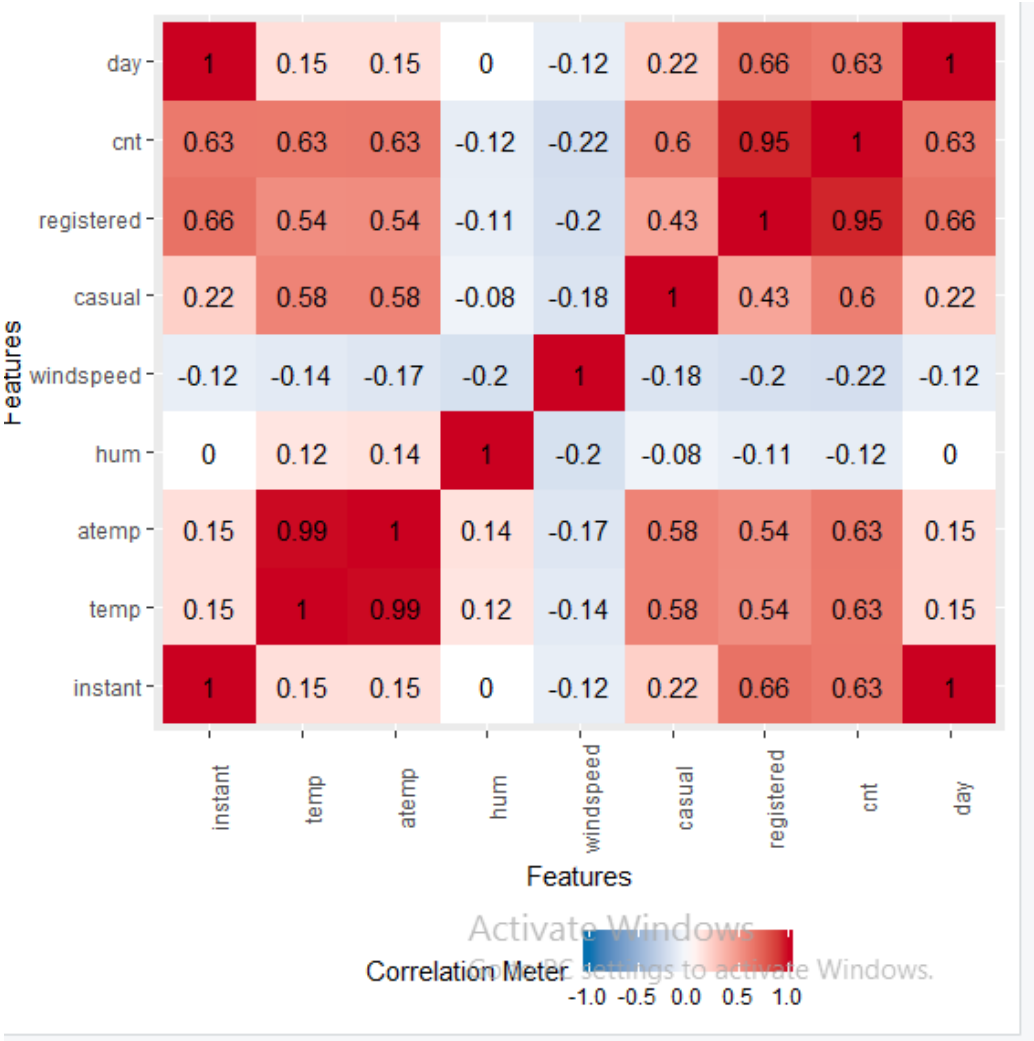
Fig 2.6 plots for hum and windspeed after removing outliers.



2.1.4. Feature selection and extraction:

For correlation analysis let us plot correlation plot of cnt .From the plot we can see that temp , atemp are highly positively correlated and registered and cnt are highly positively correlated. So we can drop atemp, registered and we can also drop casual because cnt is nothing but sum of casual and registered anywayswe are going to predict cnt we don’t need these two variables.Also we can drop insted od dteday we can only extract day from whole date and and can drop full date.Holiday and working day are nothing but opposite to each other so we don’t need both we will keep only workingDay and drop holiday and will also drop instant.
So after feature selection and extraction we have below 11 Variables.

```
colnames(data_bike)
 [1] "day"        "season"     "yr"         "mnth"       "weekday"    "workingday"
 [7] "weathersit" "temp"       "hum"        "windspeed"  "cnt"
```



2.2. Model selection and evaluation :

We need regression model we can go for linear regression, random forest etc.

2.2.1 Linear Regression:

mlr=lm(cnt ~.,data=train_bike)
pred_lm=predict(mlr,test_bike[,-11])

Summary of LR model :

```
Residual standard error: 770.5 on 555 degrees of freedom
Multiple R-squared:  0.8469,    Adjusted R-squared:  0.8392
F-statistic: 109.7 on 28 and 555 DF,  p-value: < 2.2e-16

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 146085.687  55519.818   2.631 0.008744 **
```

We can see Lr model is able to explain 83% variance from data. And p values is less than 0.05 . Intercept is 146085.687

Below are the predicted values by LR model:

```
> data.frame(test_bike$cnt,pred_lm)
    test_bike.cnt      pred_lm
10           1321  1043.46764
12           1162   885.11064
14           1421  1409.03841
23            986   425.74923
24           1416   816.91888
26            506 -1636.89432
38           1712  1945.68130
```

MAPE calculation/Performance:

```
731           2729  2743.86705
> regr.eval(test_bike$cnt,pred_lm)
          mae           mse          rmse          mape
6.252610e+02 6.827493e+05 8.262864e+02 2.213205e-01
>
```

As we can see the MAPE is 22.13 % which means accuracy is 77.87 %

2.2.2. Random Forest with 500 trees:

mrf = randomForest(cnt ~ ., train_bike, importance = TRUE, ntree = 500)
pred_mrf=predict(mrf,test_bike[,-11])
First model is built with 500 tress. Lets see its performance.

```
> regr.eval(test_bike$cnt,pred_mrf)
          mae           mse          rmse          mape
4.581142e+02 4.304811e+05 6.561106e+02 1.698359e-01
>
```

So the error rate is 16.98 % which means accuracy is 83.02 %

2.2.3. Random forest with 800 trees:

mrf2 = randomForest(cnt ~ ., train_bike, importance = TRUE, ntree = 800)
pred_mrf2=predict(mrf2,test_bike[,-11])

```
> regr.eval(test_bike$cnt,pred_mrf2)
          mae           mse          rmse          mape
4.509989e+02 4.133768e+05 6.429439e+02 1.656237e-01
>
```

Error rate: 16.5623 % accuracy: 83.4377

2.2.4. Random Forest with 1000 trees:

mrf3 = randomForest(cnt ~ ., train_bike, importance = TRUE, ntree = 1000)
pred_mrf3=predict(mrf2,test_bike[,-11])

```
> regr.eval(test_bike$cnt,pred_mrf3)
         mae          mse         rmse         mape
4.509989e+02 4.133768e+05 6.429439e+02 1.656237e-01
>
```

Error rate: 16.5623 % accuracy: 83.4377

Chapter 3 Conclusion :

From model performance we can see LR is giving r-sq value 83% but accuracy is only 77%On the other hand random forest is giving accuracy 83% . So random forest can be Selected out of two. Again in case of random forest accuracy can further be increased by increasing No of trees . model is giving same accuracy with 800 and 1000 trees so . Random forest model with 800 trees can be selected. Final Output predicted by RF model with 800 trees.

Output:

```
> output=data.frame(test_bike$cnt,pred_mrf2)
> output
    test_bike.cnt pred_mrf2
10           1321  1377.001
12           1162  1487.593
14           1421  1478.427
23            986  1160.652
24           1416  1415.517
26            506  1293.019
38           1712  1705.699
```

Appendix

R code: refer to attached r file
Python code : refer to attached python file