**Bike Renting**

*Diksha P. Jagre*

**Contents**

**1** **Introduction**   **2**

1.1 Problem Statement …............................................................................................................. 3

1.2 Data ….................................................................................................................................... 3

**2** **Methodology**  **4**

2.1 Pre -Processing ….................................................................................................................. 4

2.1.1 Outlier Analysis -------------------------------------------------------------------------- 4

2.1.2 Feature Selection ------------------------------------------------------------------------- 9

2.2 Modeling ----------------------------------------------------------------------------------------------- 16

2.2.1 Model Selection -------------------------------------------------------------------------- 16

2.2.2 Logistic Regression----------------------------------------------------------------------- 16

2.2.3 Random Forest---------------------------------------------------------------------------- 17

**3**  **Conclusion** 19

3.1 Model Evaluation -------------------------------------------------------------------------------------- 19

3.1.1 ROC Curve-------------------------------------------------------------------------------- 19

3.1.2 Concordance & Discordance ----------------------------------------------------------- 24

3.1.3 ks-plot ---------------------------------------------------------------------------------------22

3.1.4 ks-stat --------------------------------------------------------------------------------------- 23

3.1.5 Confusion Matrix --------------------------------------------------------------------------24

3.2 Model Selection --------------------------------------------------------------------------------------- 24

**Appendix A - Extra Figures 25**

**Appendix B - R Code 32**

Complete R File .----------------------------------------------------------------------------------------- - 39

**Chapter 1**

**Introduction**

**1.1 Problem Statement**

The Objective of this Case is to Predication of bike rental on daily based on the environmental and seasonal settings.

**1.2 Data**

The dataset will be given:

1) day.csv

The details of data attributes in the dataset are as follows –

|  |  |
| --- | --- |
| instant | Record index |
| dteday | Date |
| season | Season (1:springer, 2:summer, 3:fall, 4:winter) |
| yr | Year (0: 2011, 1:2012) |
| mnth | Month (1 to 12) |
| hr | Hour (0 to 23) |
| 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) |
| casual | count of casual users |
| registered | count of registered users |
| cnt | count of total rental bikes including both casual and registered |

There are total 16 variables in given dataset.

**Chapter 2**

**Methodology**

**2.1 Pre-Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. But Before exploring data, you should spend some time thinking about the business problem, gaining the domain knowledge and may be gaining first-hand experience of the problem.

The dataset shows hourly rental data for two years (2011 and 2012).The training data set is for the first 19 days of each month. The test dataset is from 20th day to month’s end. We are required to predict the total count of bikes rented during each hour covered by the test set.

In training data set, they have separately given bike demand by registered, casual users and sum of both is given as count. Training data set has 12 variables and Test has 9 (excluding registered, casual and count)

**2.1.1 Outlier Analysis**

Outlier analysis is one of the technique to understand, clean and prepare data for building a predictive model. Here we used multivariate analysis for testing the dataset.

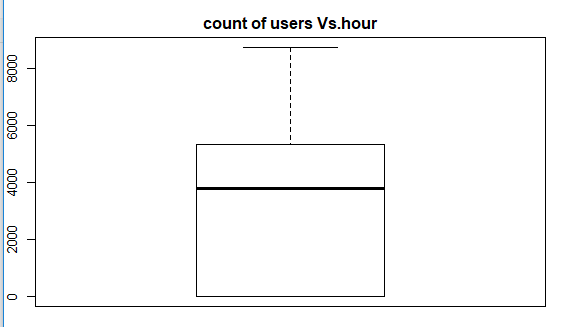
Till now, we have got a fair understanding of the data set. Now, let’s test the hypothesis which we had generated earlier. Here I have added some additional hypothesis from the dataset. Let’s test them one by one:

* **Hourly trend**: We don’t have the variable ‘hour’ with us right now. But we can extract it using the datetime (dteday) column.

****

Let’s plot the hourly trend of count over hours and check if our hypothesis is correct or not. We will separate train and test data set from combined one.

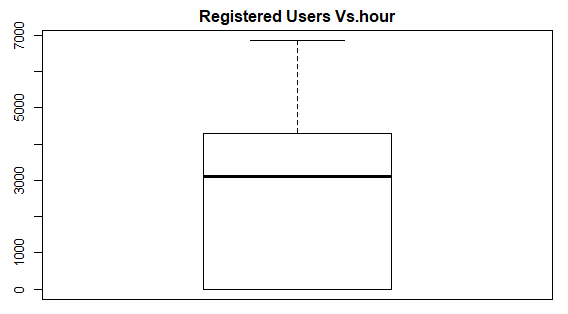
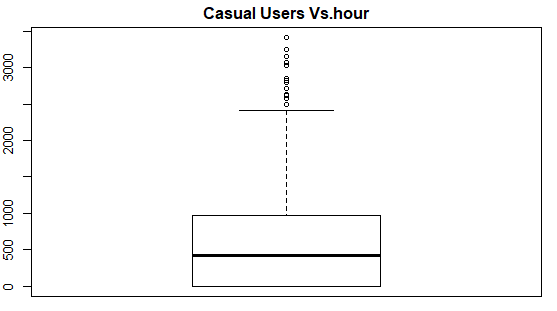


Above, you can see the trend of bike demand over hours.

Here I have analyzed the distribution of total bike demand. Let’s look at the distribution of registered and casual users separately.



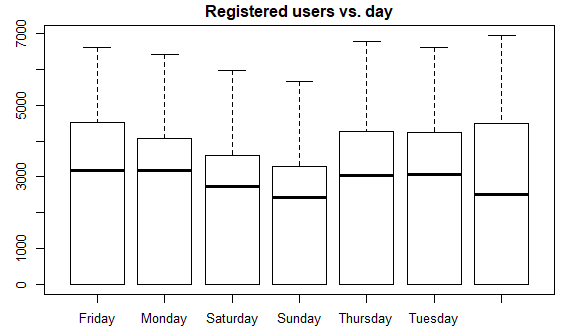
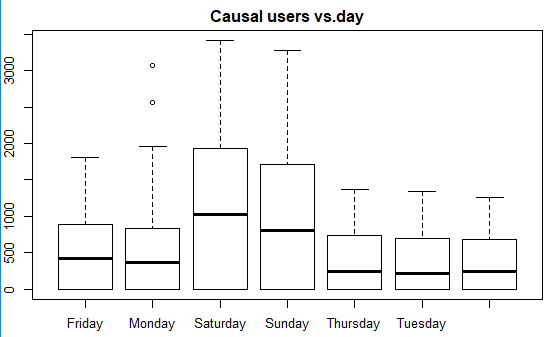
 

Above you can see that registered users have similar trend as count. Whereas, casual users have different trend. Thus, we can say that ‘hour’ is significant variable and our hypothesis is ‘true’.

* **Daily Trend:**Like Hour, we will generate a variable for day from datetime variable and after that we’ll plot it.

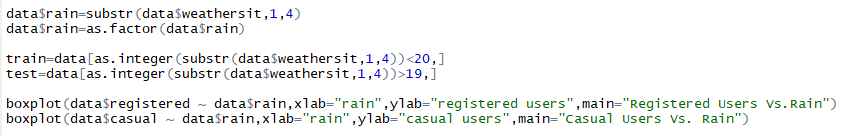
****

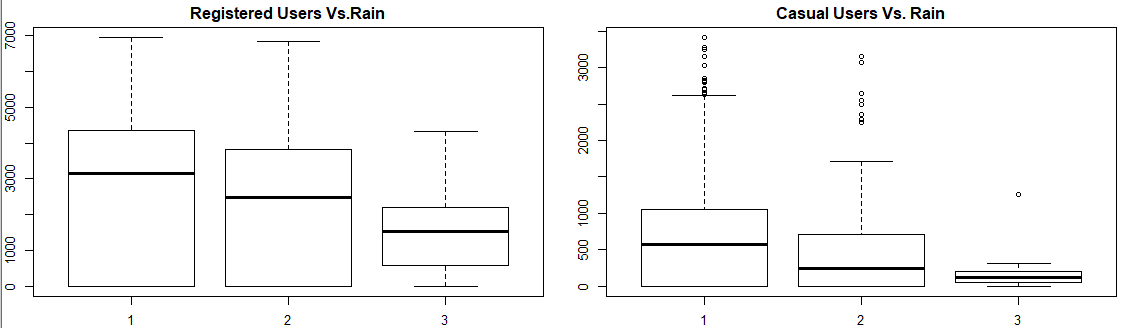
Plot shows registered and casual users’ demand over days.

** **

While looking at the plot, I can say that the demand of causal users increases over weekend.

* **Rain:**We don’t have the ‘rain’ variable with us but have ‘weathersit’ which is sufficient to test our hypothesis. As per variable description, weather 3 represents light rain and weather 4 represents heavy rain

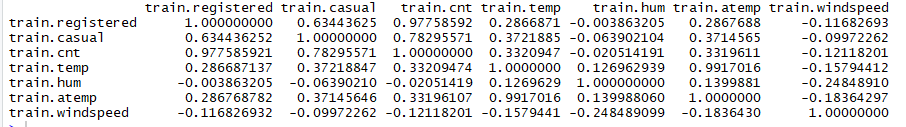
****Take a look at the plot:

****

It is clearly satisfying our hypothesis.

* **Temperature, Windspeed and Humidity:**These are continuous variables so we can look at the correlation factor to validate hypothesis.

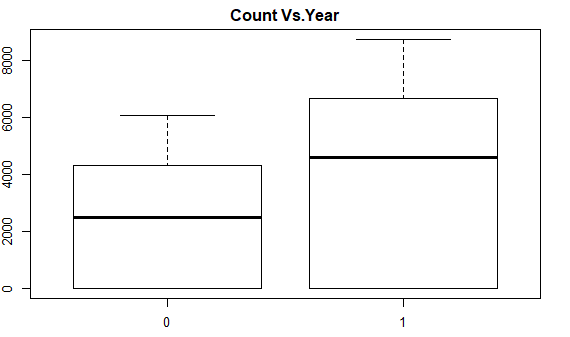
****

****

Here are a few inferences you can draw by looking at the above histograms:

* Variable temp is positively correlated with dependent variables (casual is more compare to registered)
* Variable atemp is highly correlated with temp.
* Windspeed has lower correlation as compared to temp and humidity
* **Time:**Let’s extract year of each observation from the datetime column and see the trend of bike demand over year.

****

****

Here 0 represent 2011 and 1 represent 2012.You can see that 2012 has higher bike demand as compared to 2011.

* **Pollution & Traffic:**We don’t have the variable related with these metrics in our data set so we cannot test this hypothesis.

**2.1.2 Feature Selection**

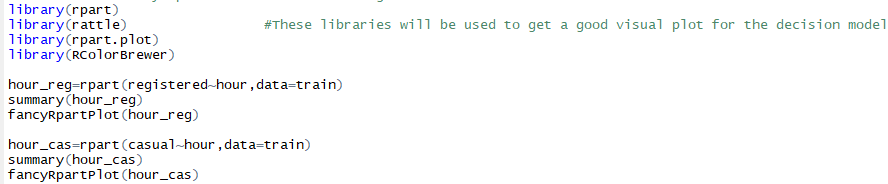
In addition to existing independent variables, we will create new variables to improve the prediction power of model. Initially, you must have noticed that we generated new variables like hour & day.

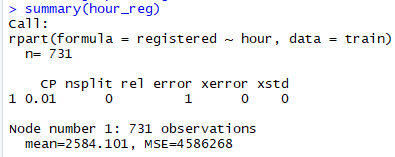
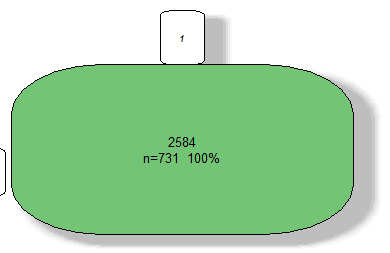
Here we will create more variables, let’s look at the some of these:

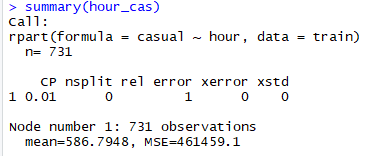
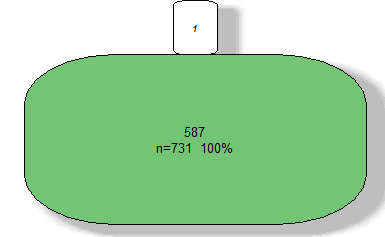
1. **Hour Bins:**Initially, we have broadly categorized the hour into three categories. Let’s create bins for the hour variable separately for casual and registered users. Here we will use decision tree to find the accurate bins.

****

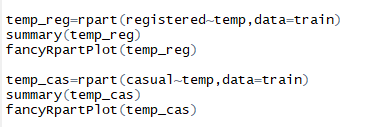
We use the library rpart for decision tree algorithm.

****

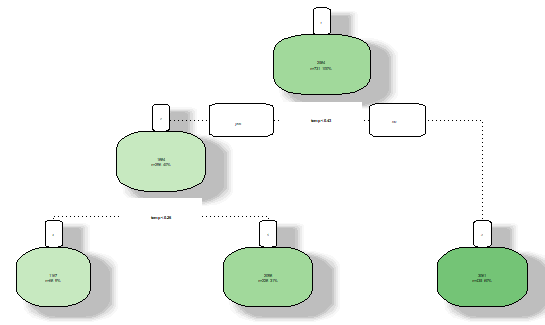
****

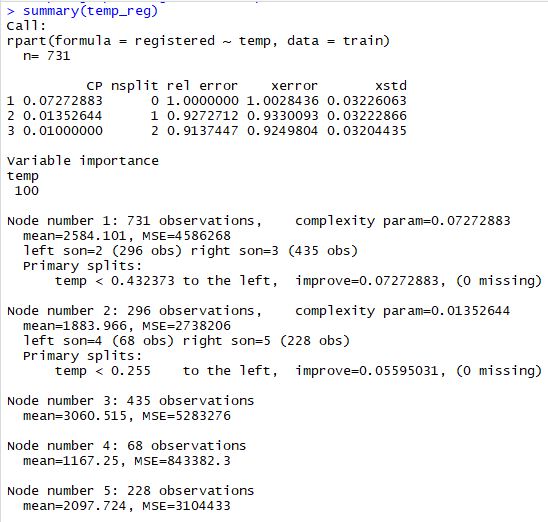
****

1. **Temp Bins:**Using similar methods, we have created bins for temperature for both registered and casuals users. Variables created are (temp\_reg and temp\_cas).

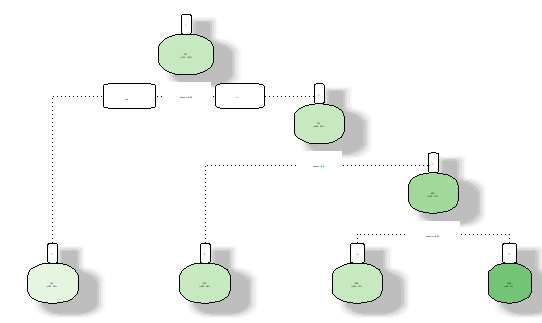


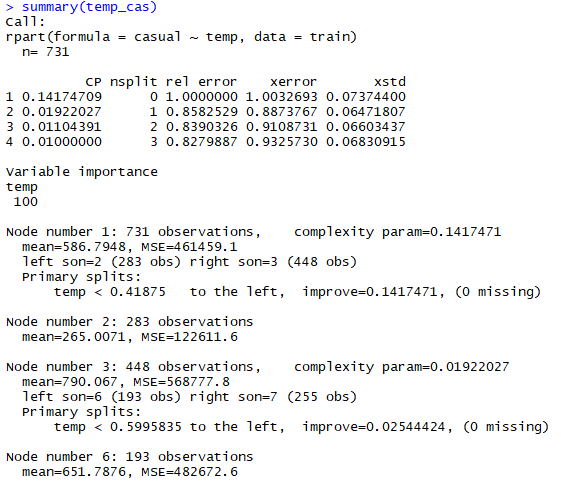
**Temp\_Reg :**

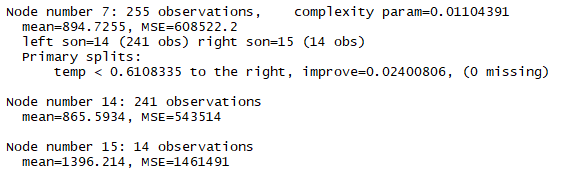




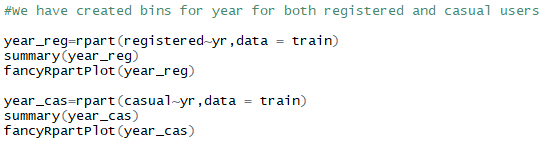
**Temp\_cas:**



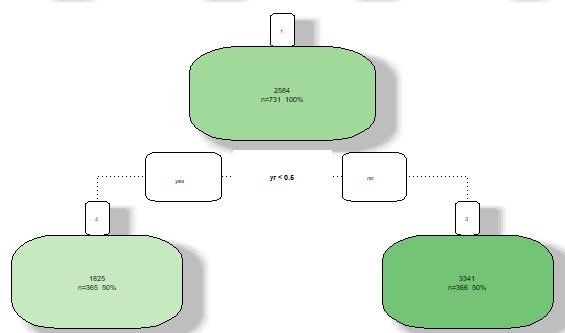


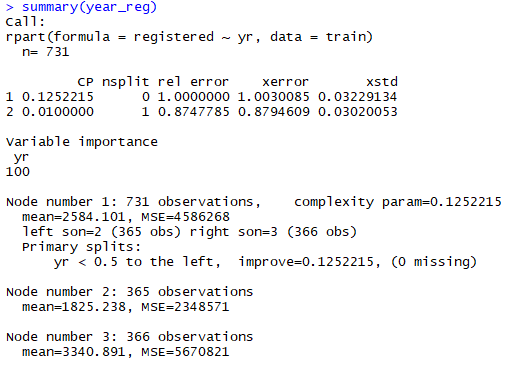


1. **Year Bins:**We had a hypothesis that bike demand will increase over time and we have proved it also.

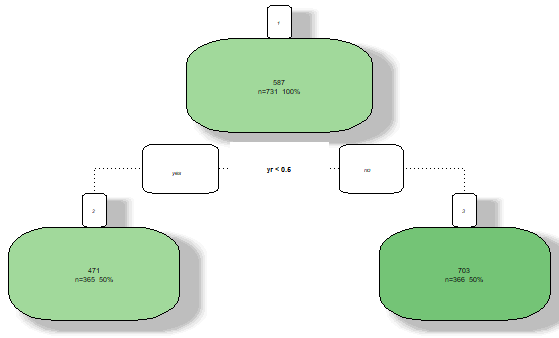
****

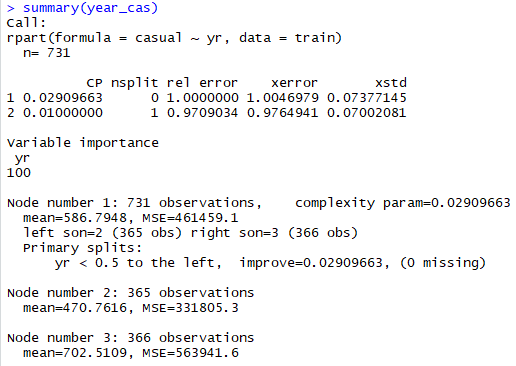
**Year\_reg:**

****

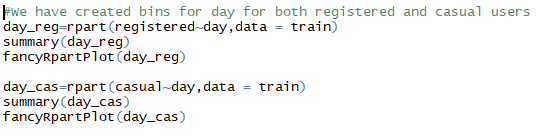
****

**Year\_cas:**

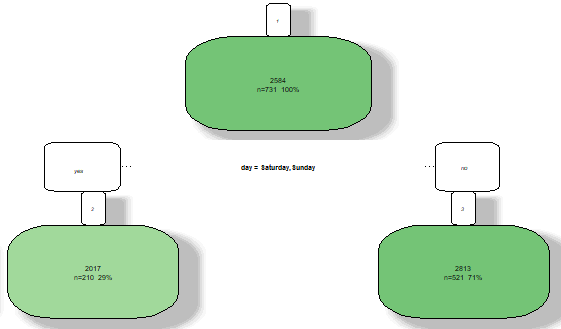
****

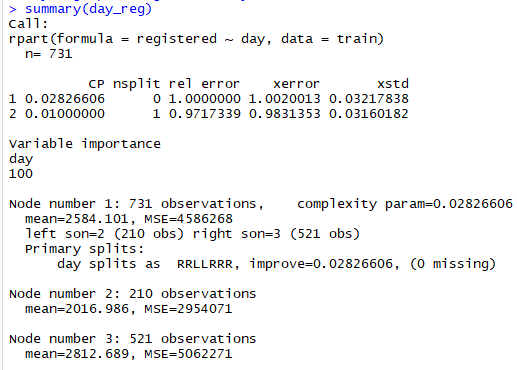
****

1. **Day Type:**Created a variable having categories like “weekday”, “weekend” and “holiday”.

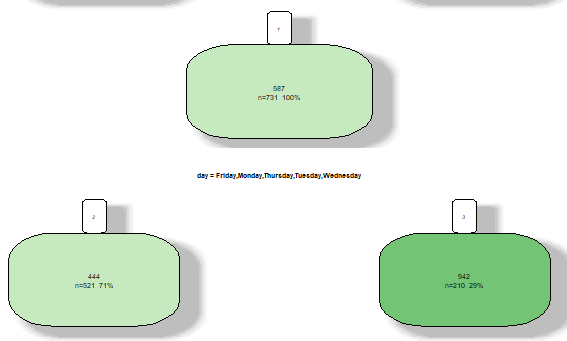
****

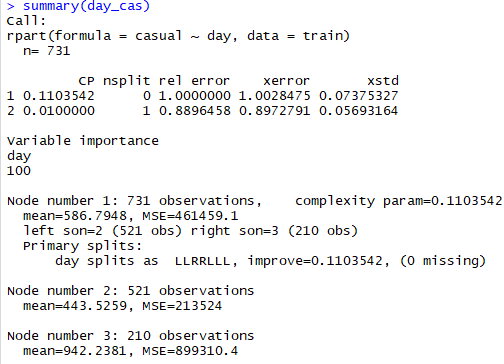
**Day\_Reg:**

****

****

**Day\_cas:**

****

****

**2.2 Modeling**

**2.2.1 Model Selection**

In our early stage of analysis during feature selection, we have come to understand that using decision tree the test result will be different. Therefore, we can neither combine the data sets nor use a single model for predicting variables. Hence, we need to analyzed the data sets separately and generate separate models for data sets.

The dependent variable can fall in either of the four categories:

1. Nominal

2. Ordinal

3. Interval

4. Ratio

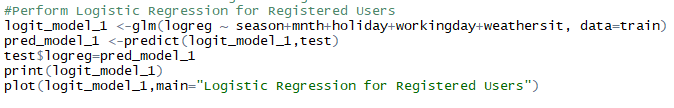
If the dependent variable, in our case Registered, Casual and Count, is Nominal the only predictive analysis that we can perform is Classification, and if the dependent variable is Interval or Ratio the normal method is to do a Regression analysis, or classification after binning. But the dependent variable we are dealing with is Interval or Ratio, for which classification and regression can be done, because the Registered, Casual and Count variable has intervals.

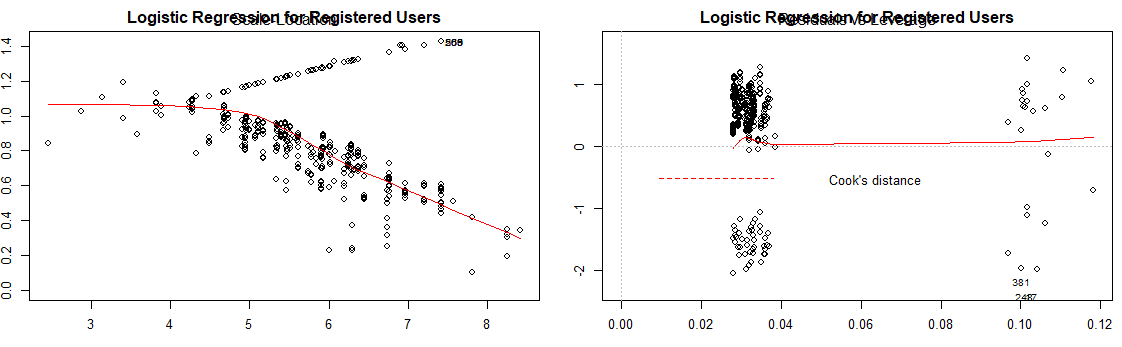
You always start your model building from the simplest to more complex. Therefore, we use Logistic Regression

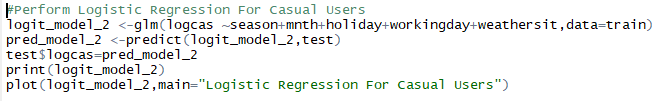
**2.2.2 Logistic Regression**

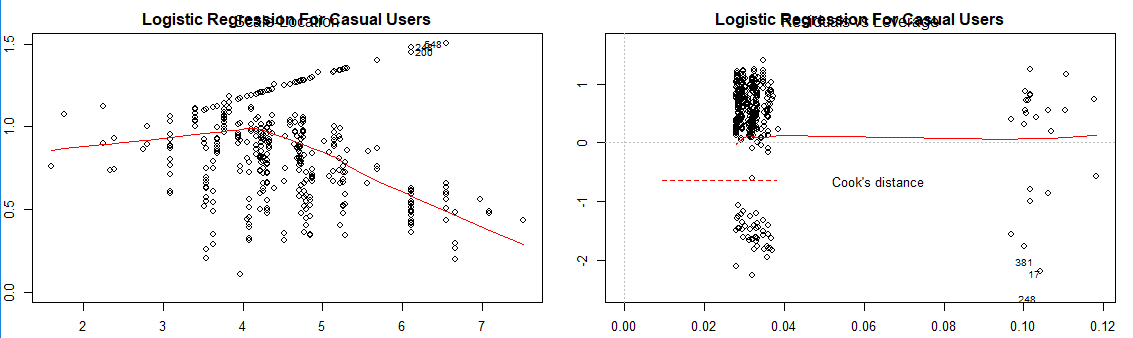
Logistic Regression is statistical model which input can be continuous or categorical. The input variables is in factor with 2 or more levels. That means you have only categorical variables to perform logistic Regression.

Here Registered and Casual users are predicted using Logistic Regression.







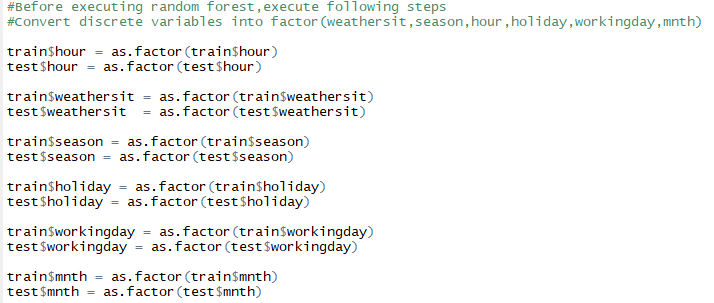


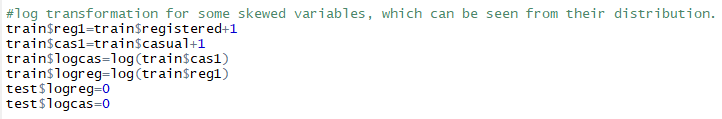
**2.2.3 Random Forest**

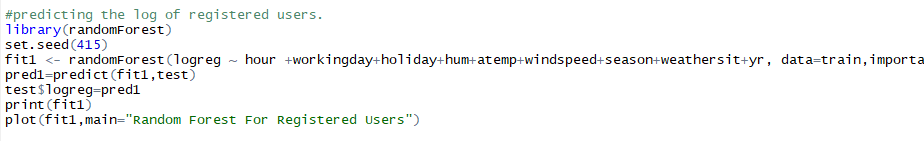
Random forest is a tree-based algorithm which involves building several trees (decision trees), then combining their output to improve generalization ability of the model. The method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner.

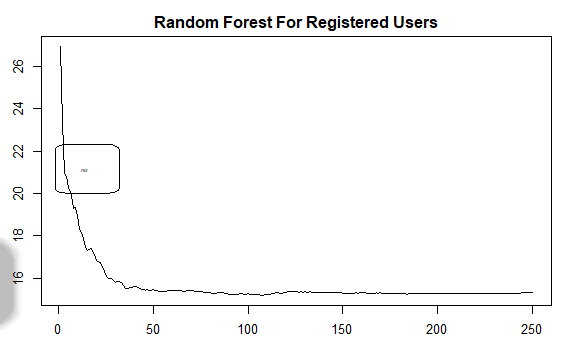
Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

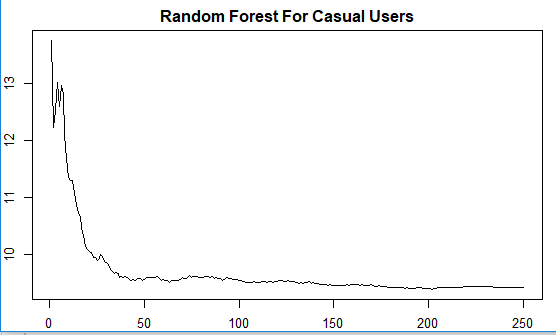
So, In our given dataset the dependent (target) variable is continuous so we solve this problem using regression.











**Chapter 3**

**Conclusion**

**3.1 Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. ROC Curve

2. Concordance & Discordance

3. ks-plot

4. ks-stat

5.Confusion Matrix

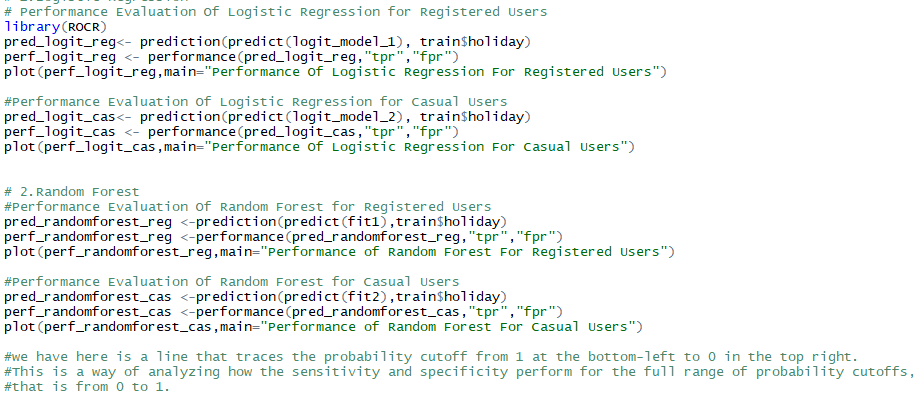
Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

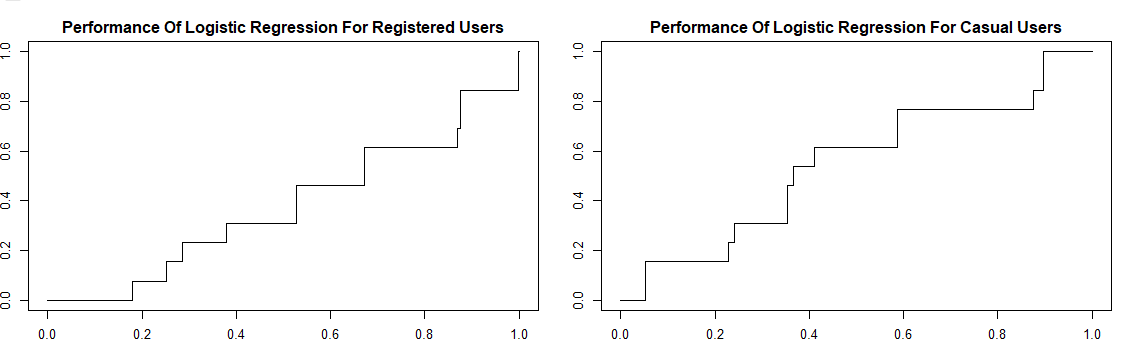
**3.1.1 ROC Curve**

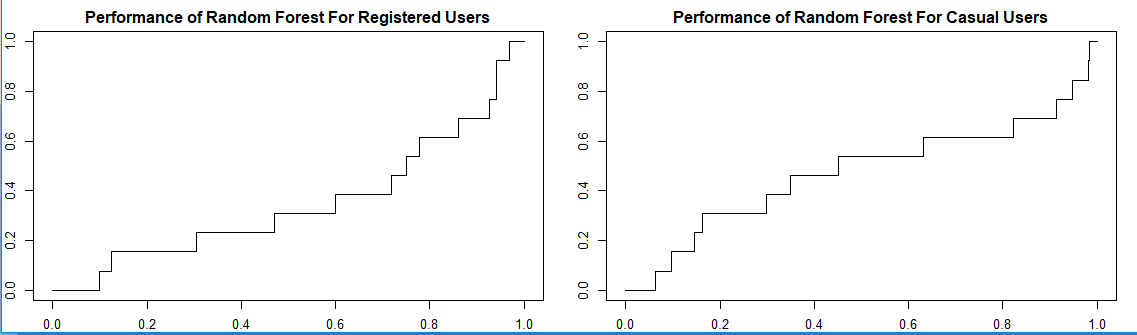
ROC Curve model is sort of a balance between predicting the one's accurately or the zeroes accurately.In other words sensitivity and specificity.

This is nicely captured by the 'Receiver Operating Characteristics' curve, also called as the ROC curve. In fact, the area under the ROC curve can be used as an evaluation metric to compare the efficacy of the models.

Let's plot the curve and the area using the plotROC from InformationValue package.

****

****

****

Instead, what we have here is a line that traces the probability cutoff from 1 at the bottom-left to 0 in the top right.

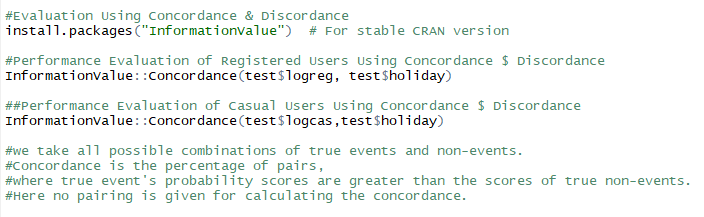
This is a way of analyzing how the sensitivity and specificity perform for the full range of probability cutoffs,that is from 0 to 1.

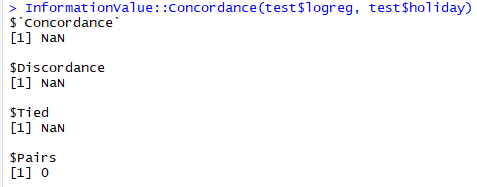
The ROC curve is the only metric that measures how well the model does for different values of prediction probability cutoffs. The [optimalCutoff](http://r-statistics.co/Information-Value-With-R.html" \l "3.1.%20optimalCutoff" \t "_blank) function from InformationValue can be used to know what cutoff gives the best sensitivity, specificity or both.

**3.1.2 Concordance & Discordance**

In an ideal model, the probability scores of all true 1's should be greater than the probability scores of ALL true 0's. Such a model is said to be perfectly concordant and this phenomenon can be measured by Concordance and Discordance.

For a perfect model, this will be 100%. So, the higher the concordance, the better is the quality of the model. This can be computed using the Concordance function in InformationValue package.



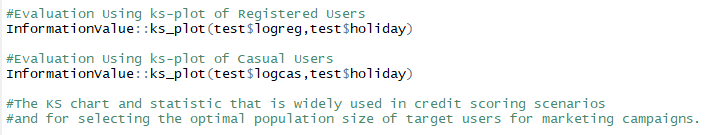


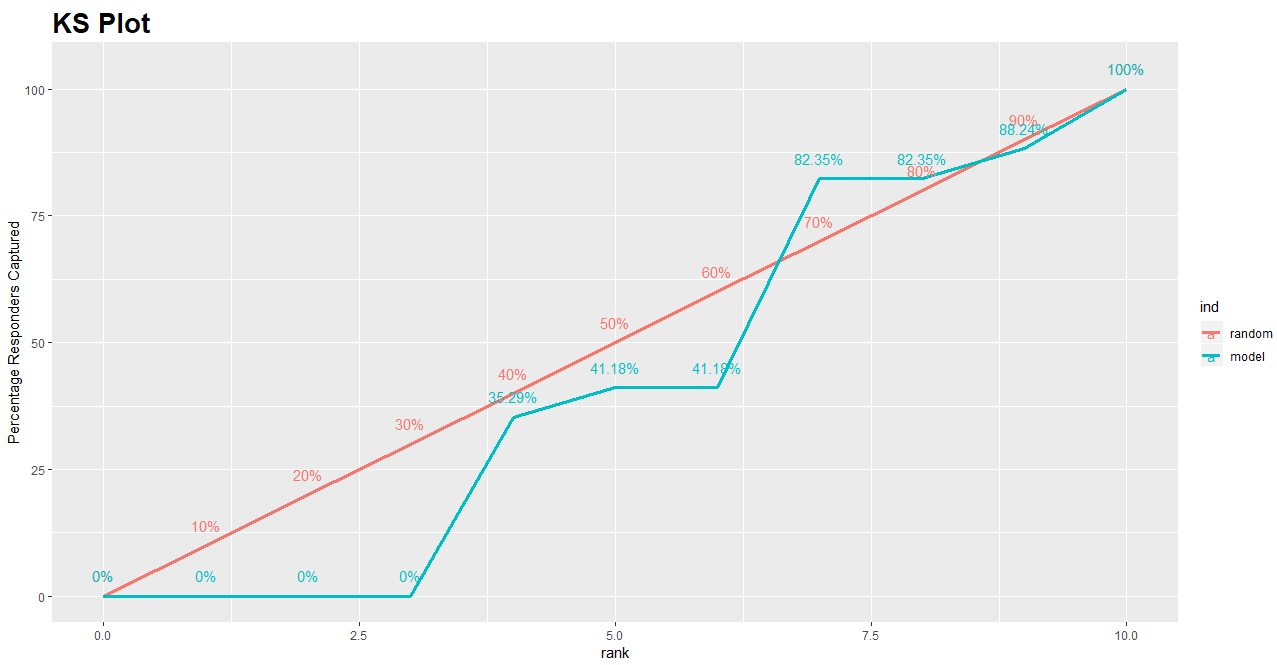
Here no paring is between two variables so their respective concordance and discordance is null.

**3.1.3 ks – plot**

The KS Chart is particularly useful in marketing campaigns and ads click predictions where you want to know the right population size to target to get the maximum response rate.

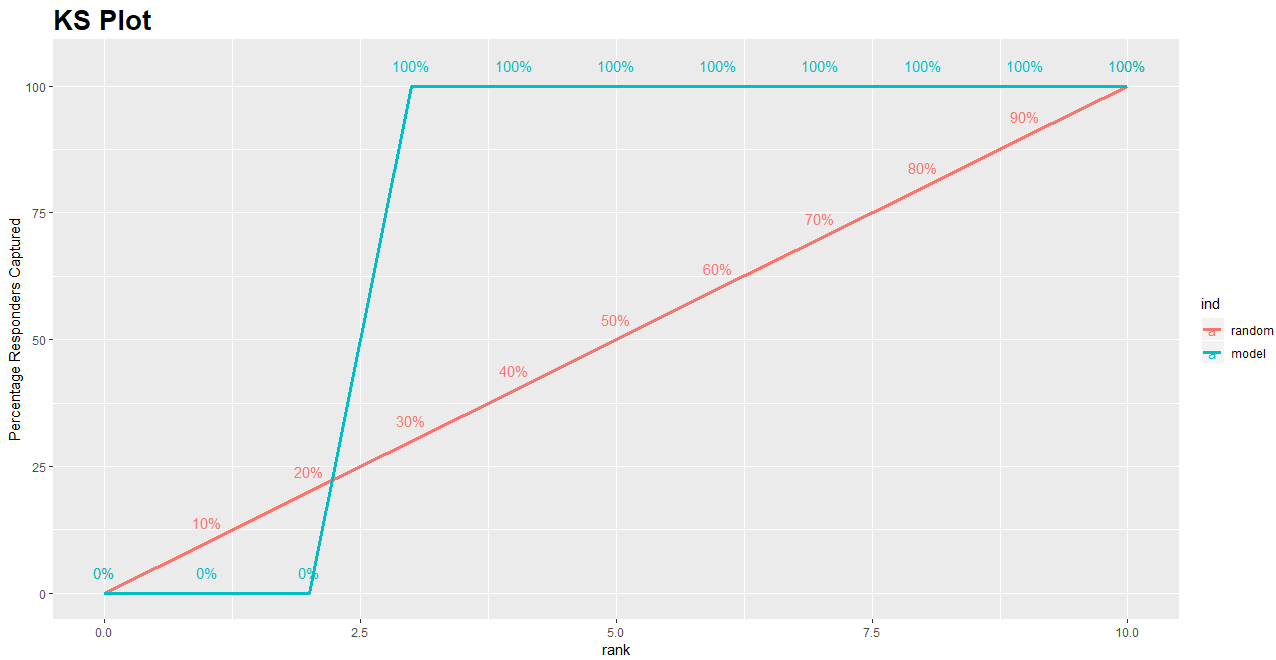
The KS chart below shows how this might look like. The length of the vertical dashed red line indicates the KS Statistic

****

****

**ks – plot for Registered Users**

By targeting the top 65% of the population, the model is able to cover 82.35% of responders (1's).

****

**ks –plot for Casual Users**

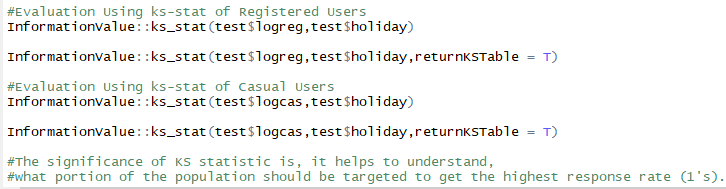
By targeting the top 20% of the population, the model is able to cover 100 % of responders (1's).

**3.1.4 ks – stat**

Ks-statistics and ks-chart are used to make decision like how to predict the bike rental count on the daily basis.

The significance of KS statistic is, it helps to understand, what portion of the population should be targeted to get the highest response rate (1's).

The KS statistic can be computed using the ks\_stat function in InformationValue package. By setting the returnKSTable = T, you can retrieve the table that contains the detailed decile level splits.

+

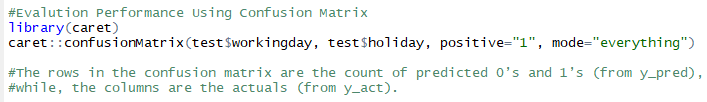


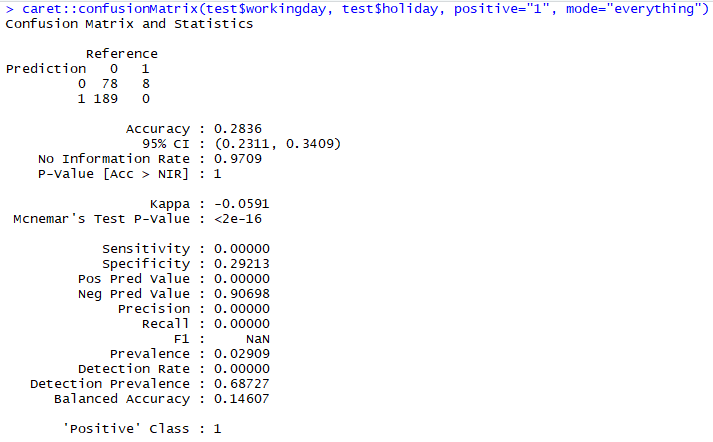


**3.1.5 Confusion Matrix**

The caret package provides the awesome confusionMatrix function for this. It takes in the predicted and actual values. And to avoid confusion, always specify the positive argument.

Otherwise, it is possible for ‘0’ to be taken as ‘positive’ or the ‘event’, and will cause a big mistake which may go unnoticed.

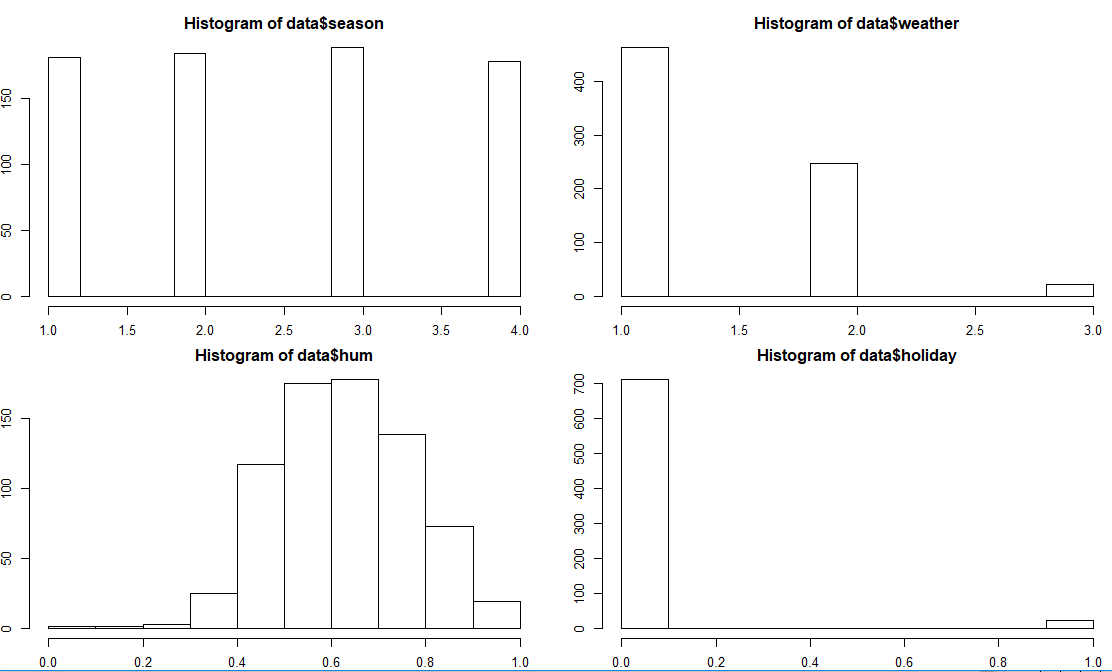


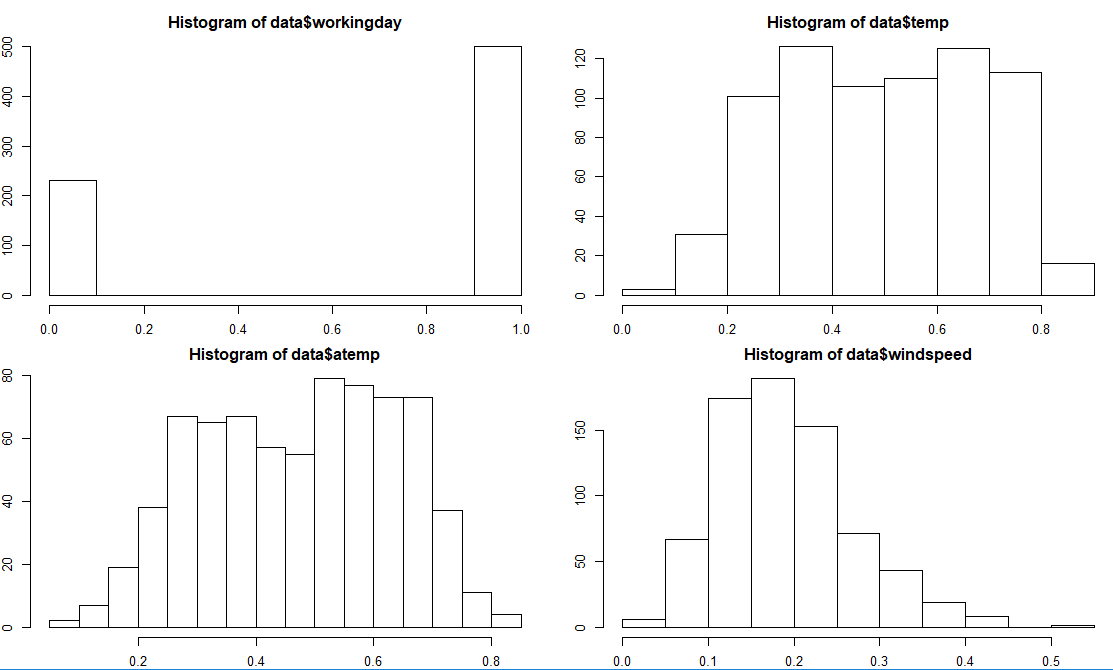


**3.2 Model Selection**

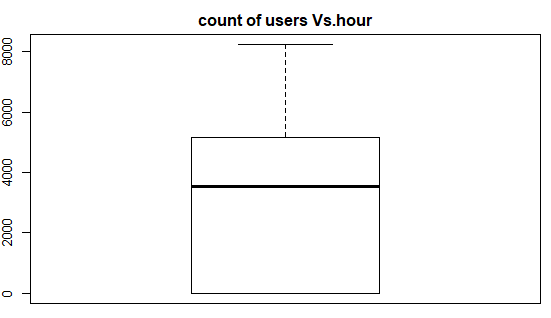
We can see that out of three models one models perform great accuracy work therefore we can select those models without any loss of information.

**Appendix A - Extra Figures**

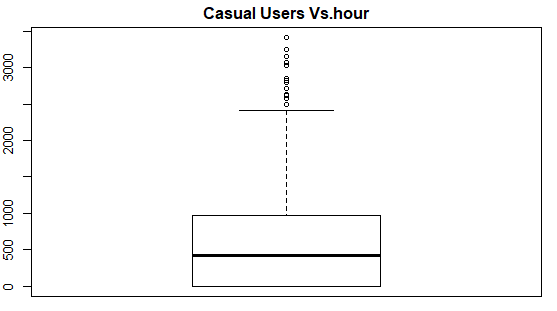
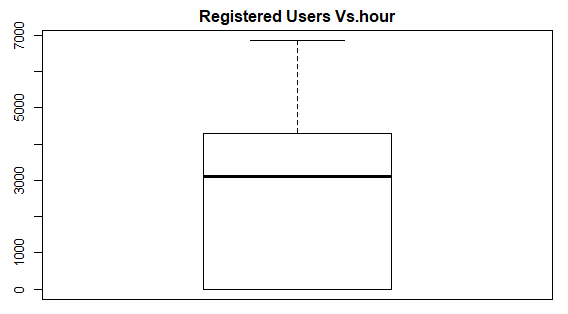
****

****

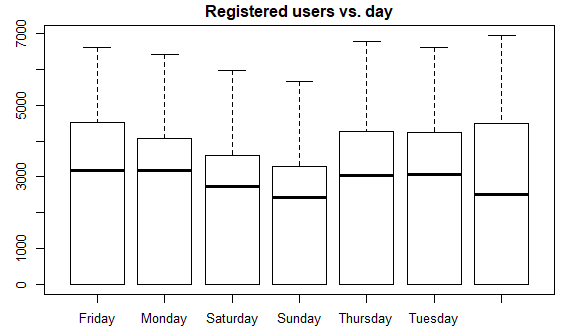
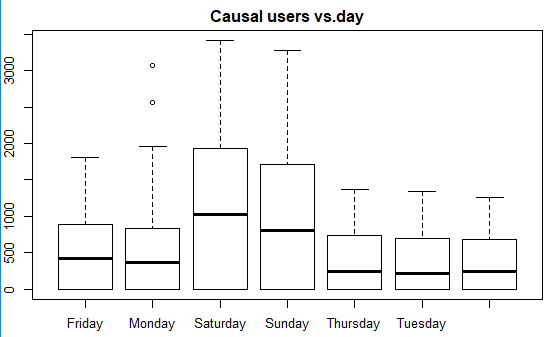
**Figure 1.1 Histogram For All Numerical Variables**

****

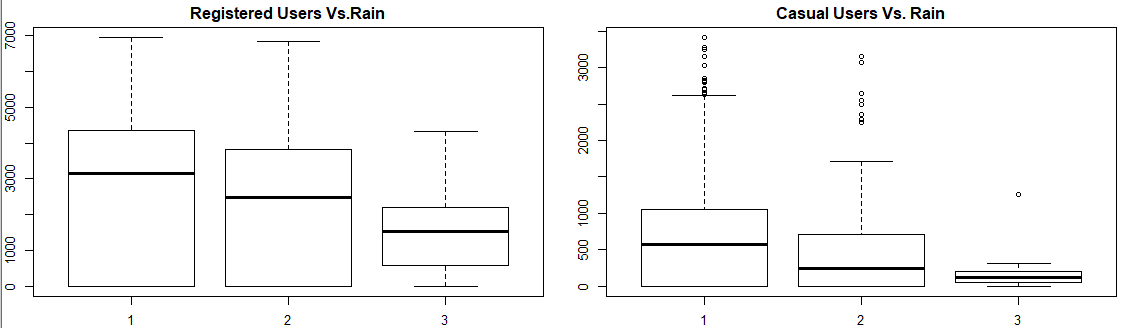
**Figure 2.1 Boxplot for Count of users Vs. Hour**

****

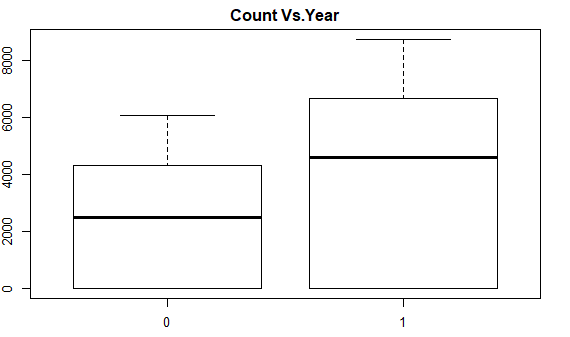
**Figure 2.2 Boxplot for Registered & Casual Users Vs. Hour**

** **

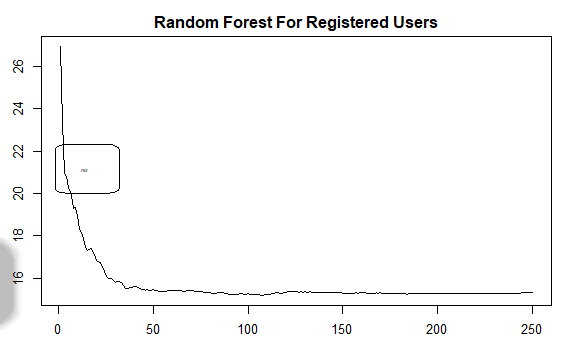
**Figure 2.3 Boxplot for Registered & Casual Users Vs. Day**

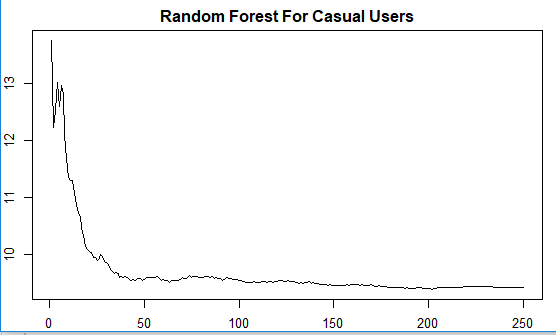
****

**Figure 2.4 Boxplot for Registered & Casual Users Vs.Rain**

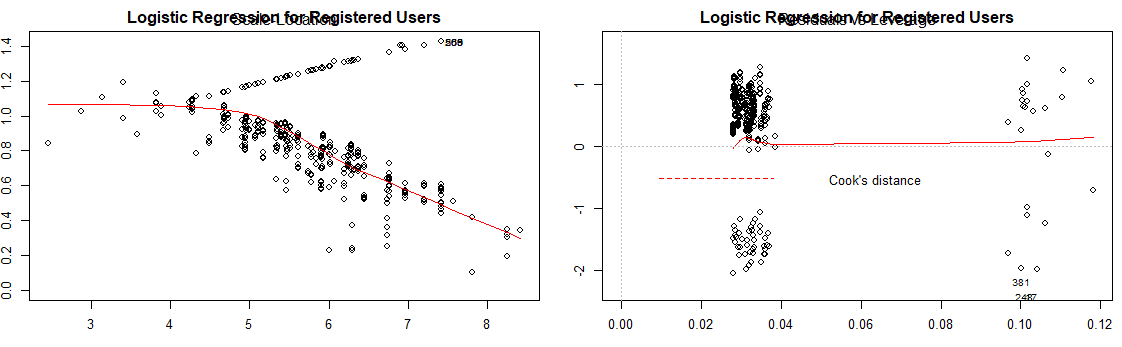
****

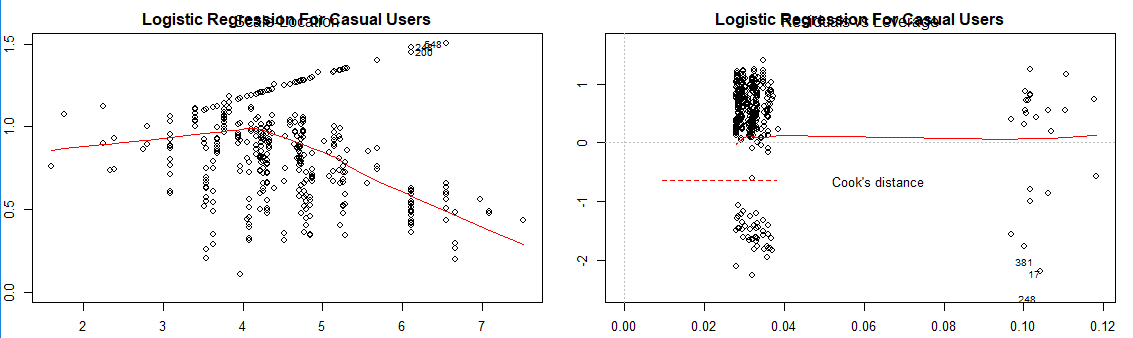
**Figure 2.5 Boxplot for Registered & Casual Users Vs. Year**

****

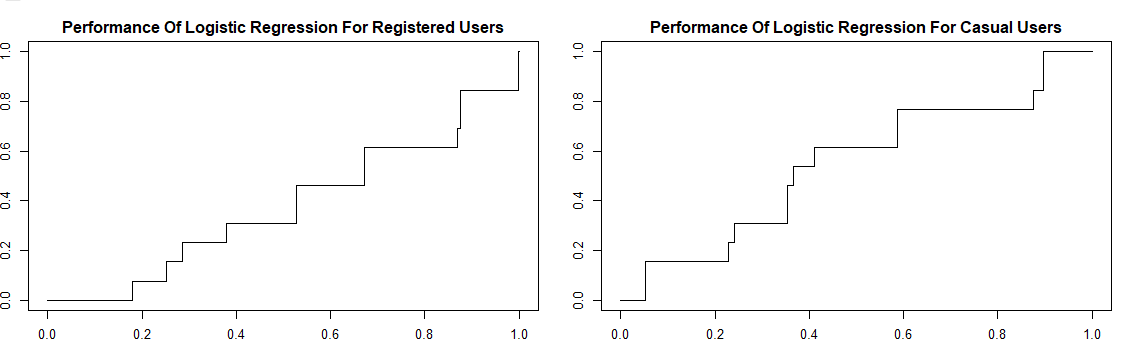
****

**Figure 2.6 Random Forest For Registered & Casual Users**

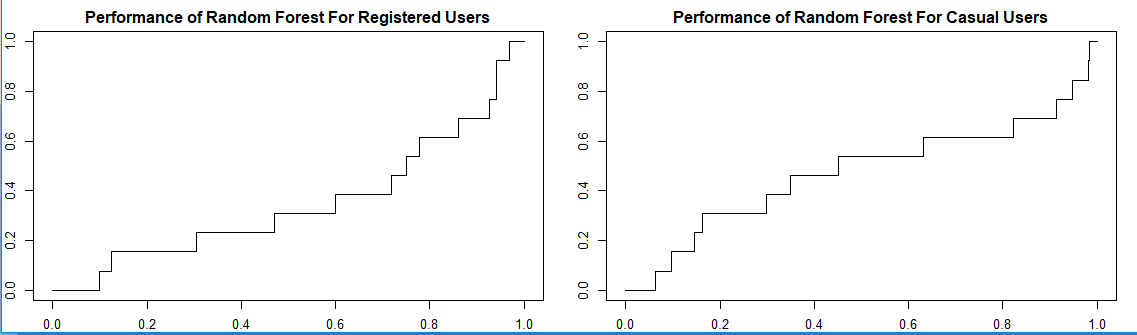
****

****

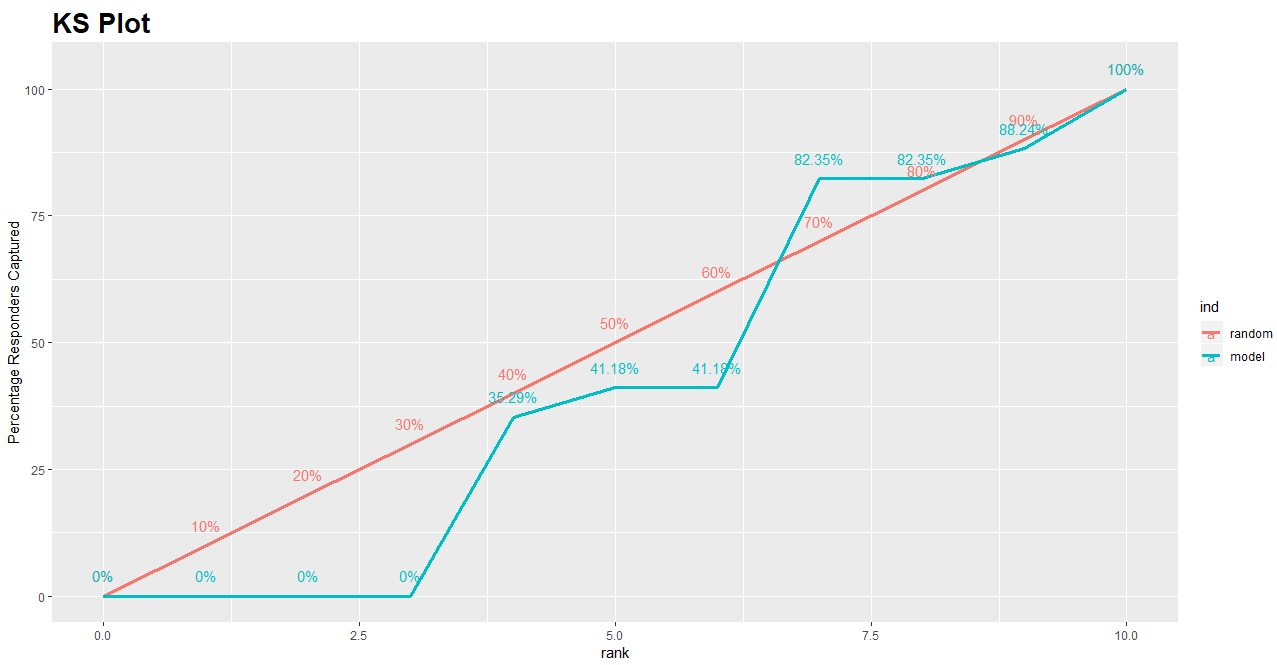
**Figure 2.7 Logistic Regression For Registered & Casual Users**

****

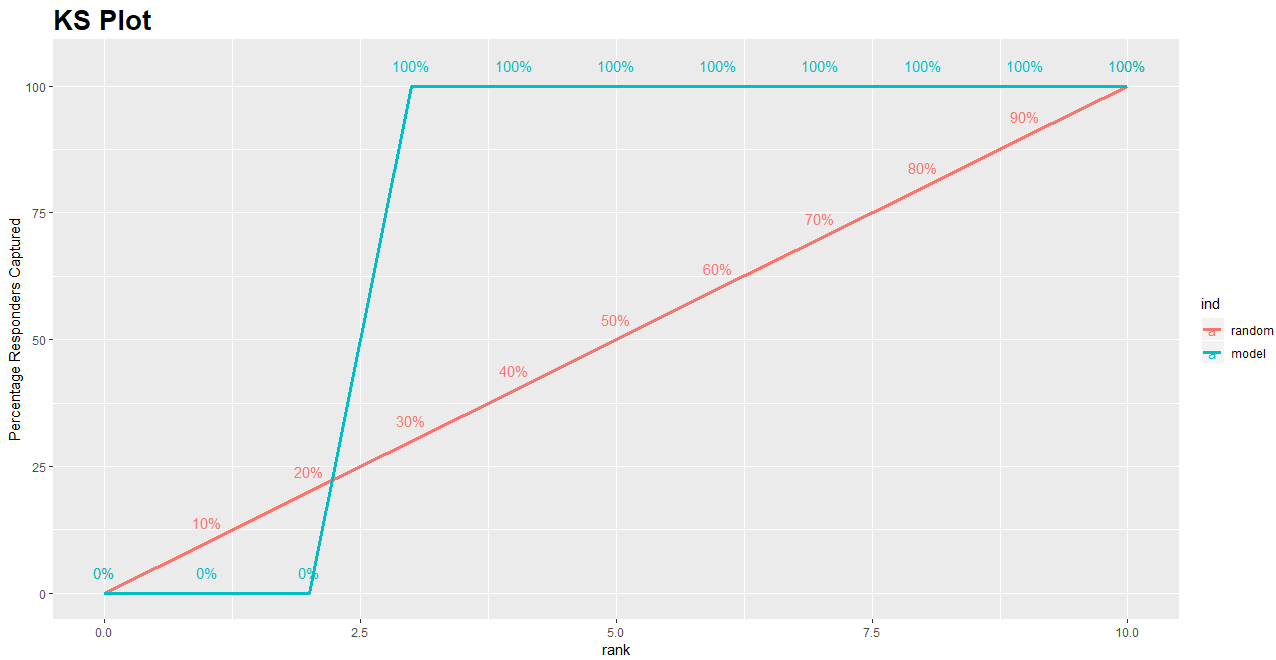
**Figure 3.1 Performance Of Logistic Regression Using ROC Curve**

****

**Figure 3.2 Performance Of Random Forest Using ROC Curve**

****

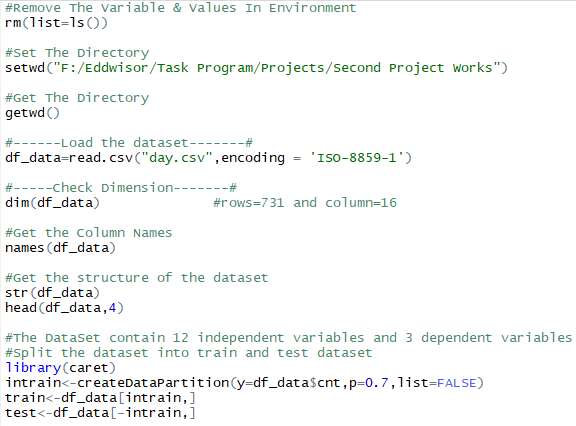
**Figure 3.3 ks-plot For Registered Users**

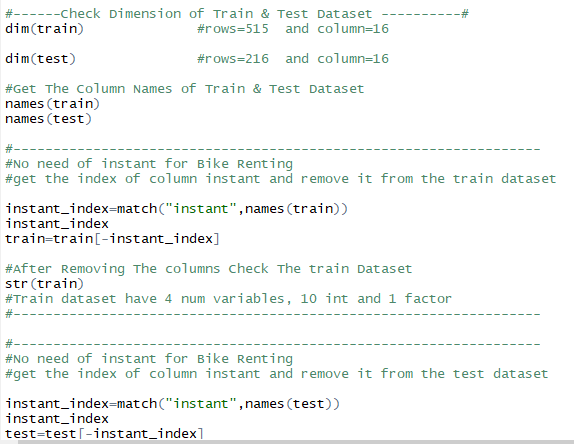
****

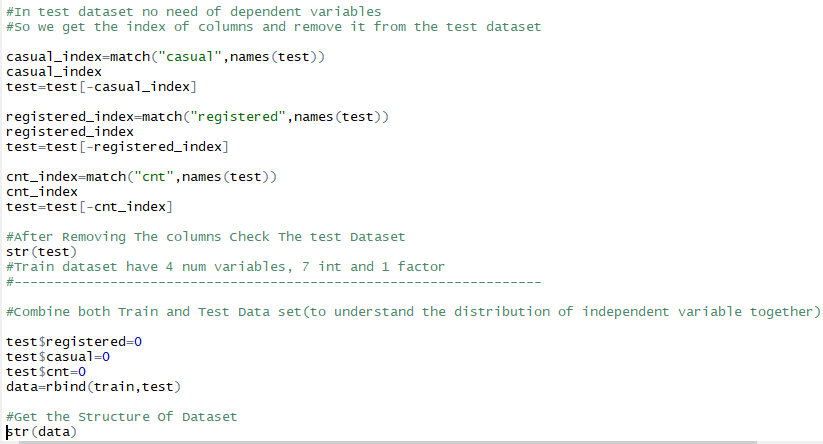
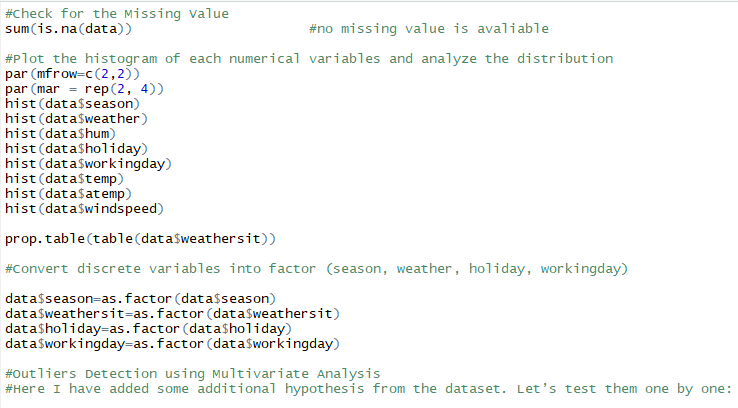
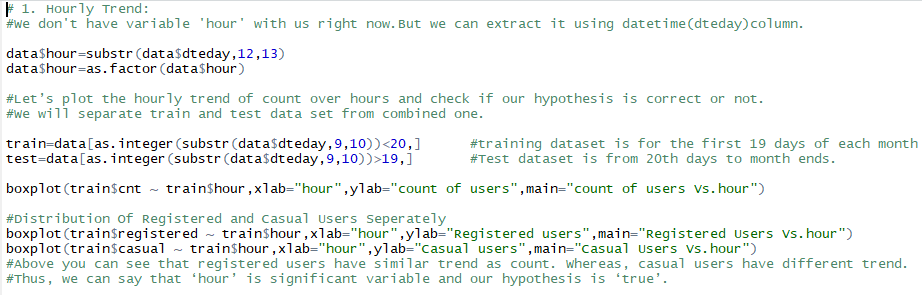
**Figure 3.4 ks-plot For Casual Users**

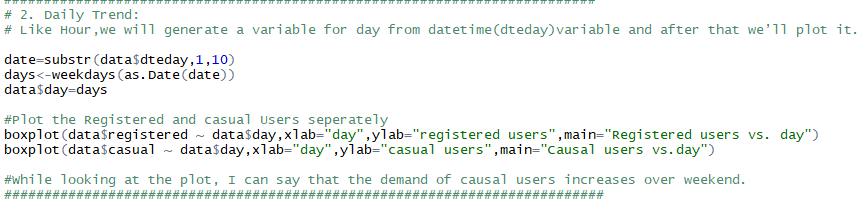
**Appendix B - R Code**

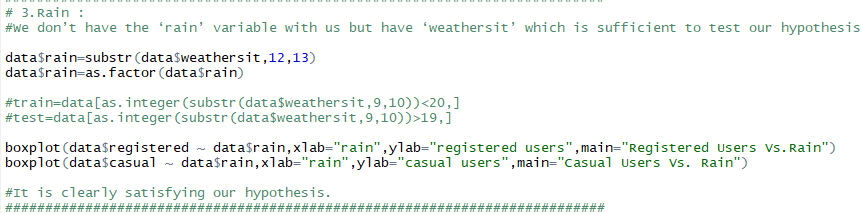
**Complete R File**

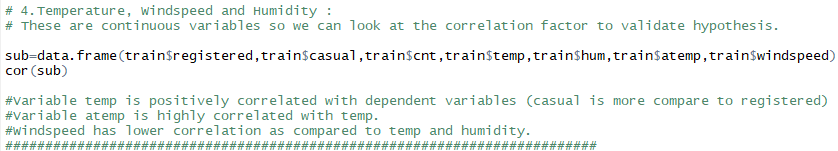
****

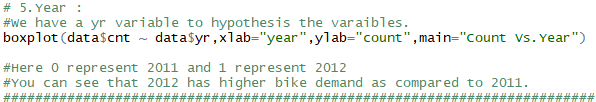
****

**  **

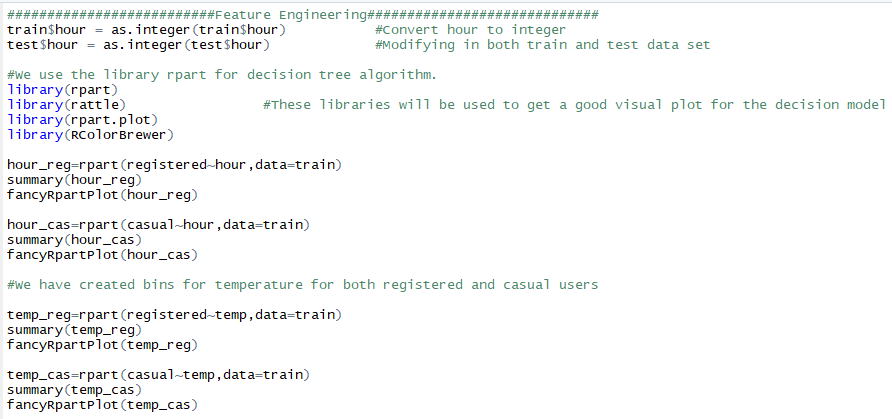
****

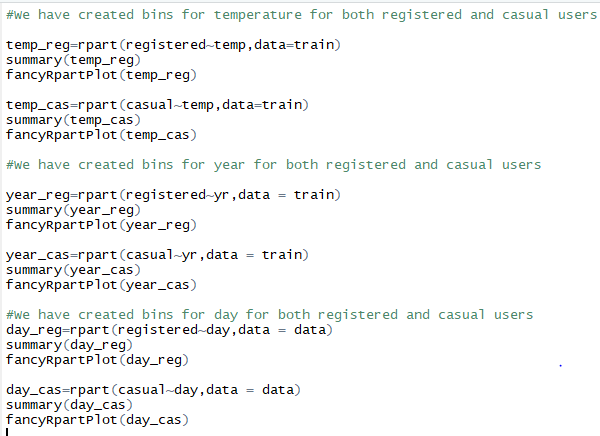
****

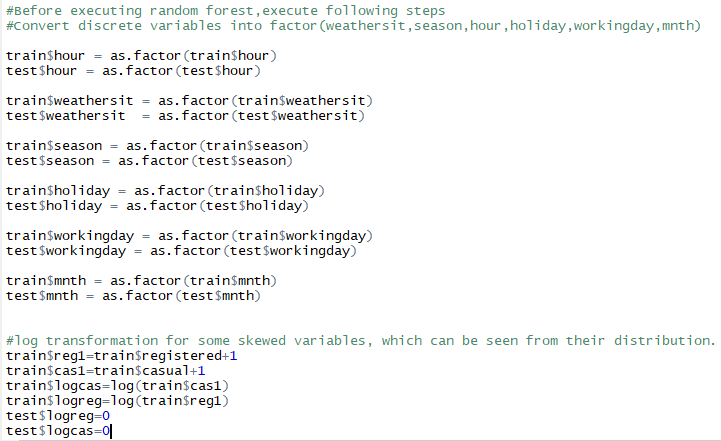
****

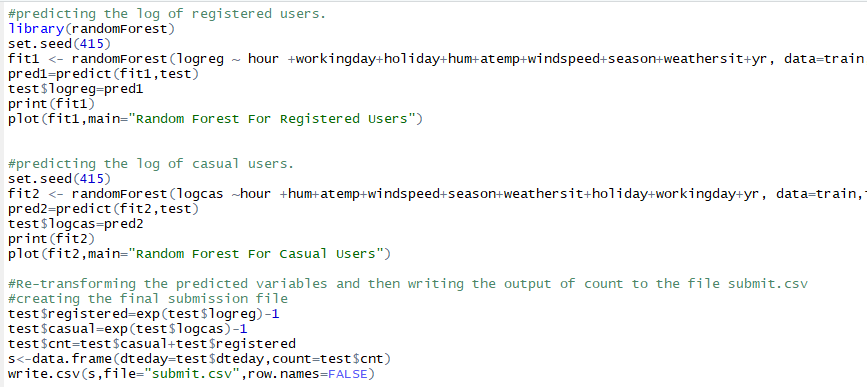
****

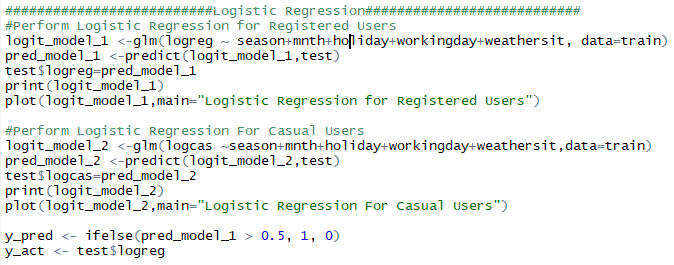
****

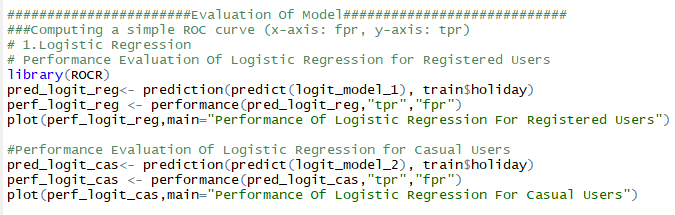
****

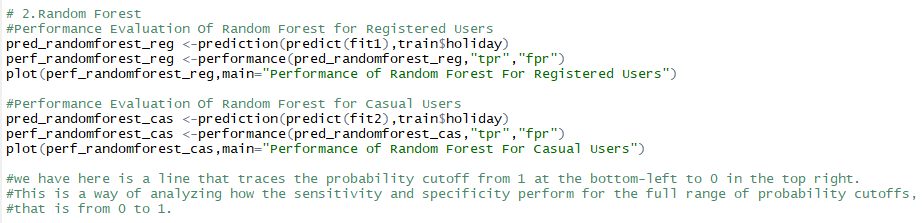
****

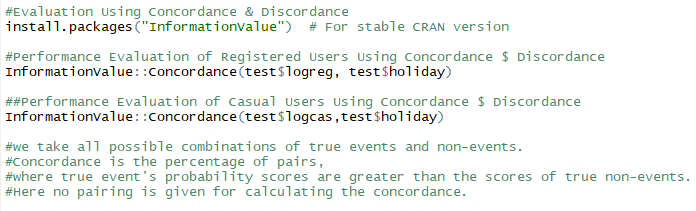
****

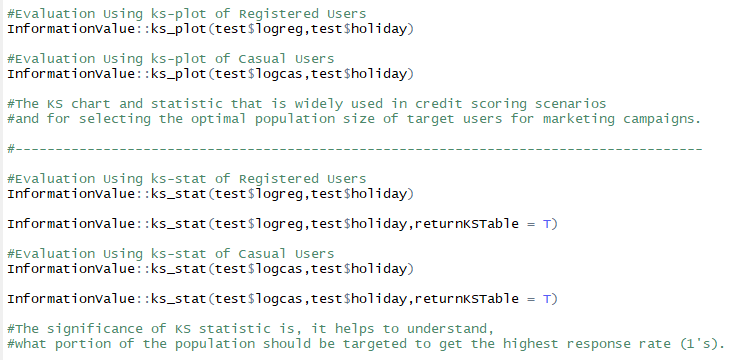
****

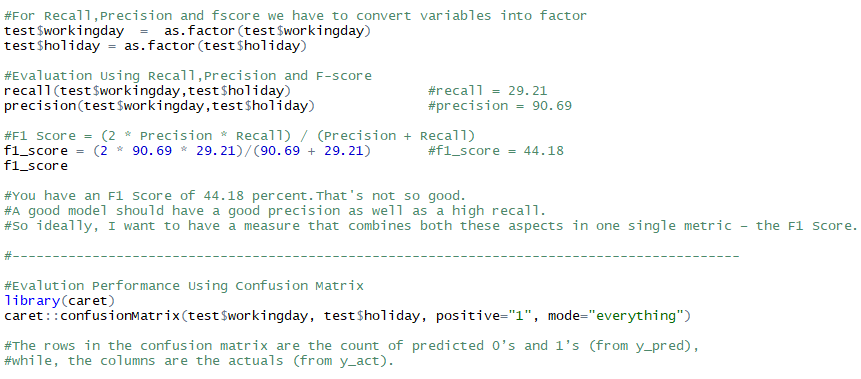
****

****

****

****

****

****