**Predicting Rental Bikes**

**Chandrakanth Bachu**

**August 2018**

**Contents**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **Introduction** | | | **3** |
|  | 1.1 | Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 3 |
|  | 1.2 | Data . | . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 4 |
| **2** | **Methodology** | | | **5** |
|  | 2.1 | Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 5 |
|  |  | 2.1.1 | Univariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 5 |
|  |  | 2.1.2 | Bivariate Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 7 |
|  |  | 2.1.3 |  |  |
|  | 2.2 | Preprocessing Techinues. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 8 |
|  |  | 2.2.1 | Missing value analysis | 8 |
|  |  | 2.2.2 | Outlier analysis | 8 |
|  |  | 2.2.3 | Feature Selection | 10 |
|  |  | 2.2.4 | Feature Scaling | 12 |
| **3** | **Modelling** | |  | **14** |
|  | 3.1 | Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 14 |
|  |  | [3.1.1 Evaluating Regression Model](#page16)  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 14 |
|  |  |  | |  |
|  | 3.2 | Decision Tree . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 15 |
|  | 3.3 | Random Forest | | 17 |
|  | 3.2 | Linear Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 19 |
|  |  |  | |  |
| **117Appendix A - Extra Figures** | | | | **21** |
| **Appendix B - R Code** | | | | **25** |
|  |  | [Univariate Analysis (Fig:](#page24)  [2.1 & 2.2)](#page6) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 26 |
|  |  | [Bivariate Analysis (Fig:](#page24)  [2.3)](#page7)  . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 26 |
|  |  | [Outlier Analysis (Fig:](#page24)  [2.4 & 2.5)](#page19)  . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 27 |
|  |  | [Feature Selection (Fig:](#page24)  [2.6)](#page8) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 27 |
|  |  | Complete R File . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 28 |
|  |  |  | |  |
|  |  |  | |  |
| [**Reference**](#page30)**s** | | |  | **29** |

1. **Problem Statement**

The Bike Rental Data contains the daily count of rental bikes between the year 2011 and 2012 with corresponding weather and seasonal information. We would like to predict the daily count of rental count in order to automate the system.

1. **Data**

Our task is to build Regression model which will give the daily count of rental bikes based on weather and season Given below is a sample of the data set that we are using to predict the count:

Table 1.1: Bike Rental Sample Data (Columns: 1-8)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| instant | dteday | Season | yr | mnth | holiday | weekday |
| 1 | 1/1/2011 | 1 | 0 | 1 | 0 | 6 |
| 2 | 1/2/2011 | 1 | 0 | 1 | 0 | 0 |
| 3 | 1/3/2011 | 1 | 0 | 1 | 0 | 1 |
| 4 | 1/4/2011 | 1 | 0 | 1 | 0 | 2 |
| 5 | 1/5/2011 | 1 | 0 | 1 | 0 | 3 |

Table 1.2: Bike Rental Sample Data (Columns: 9-14)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

Below are the variables we used to predict the count of bike rentals

Table 1.3: Bike Rental Predictors

|  |  |
| --- | --- |
| s.no | Variables |
| 1 | Dteday |
| 2 | Season |
| 3 | Yr |
| 4 | Mnth |
| 5 | Holiday |
| 6 | Weekday |
| 7 | workingday |
| 8 | weathersit |
| 9 | Temp |
| 10 | Atemp |
| 11 | Hum |
| 12 | windspeed |
| 13 | Casual |
| 14 | registered |

**Chapter 2**

**Methodology**

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**. To start this process we will first try and look at all the distributions of the Numeric variables. Most analysis like regression, require the data to be normally distributed.

**2.1.1 Univariate Analysis**

In Figure 2.1 and 2.2 we have plotted the probability density functions numeric variables present in the data including target variable cnt..

1. Target variable cnt is normally distributed
2. Independent variables like ‘temp’,’atemp’, and ‘regestered’ data is distributed normally.
3. Independent variable ‘casual’ data is slightly skewed to the right so, there is chances of getting outliers.
4. Other Independent variable ‘hum’ data is slightly skewed to the left , here data is already in normalize form so outliers are discarded.

Figure 2.1 Distribution of target variable (CNT) (python code in Appendix B)

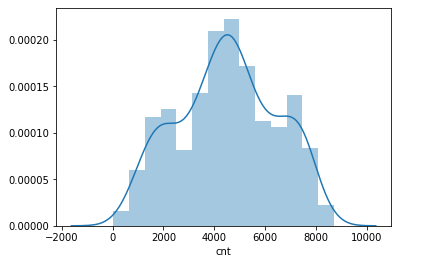
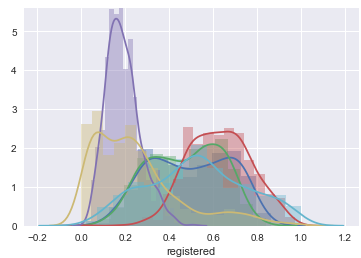
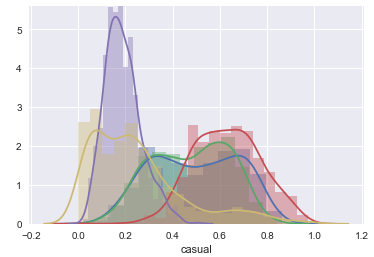
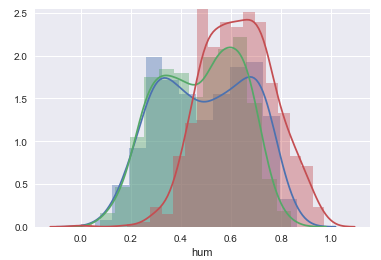
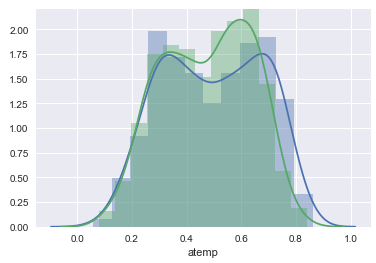
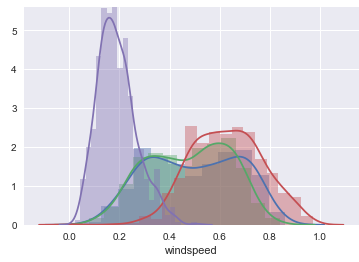
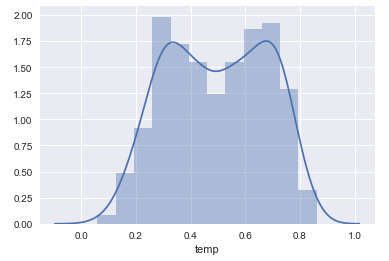


Figure 2.2 showing distribution of dependent variables (python code in Appendix B)



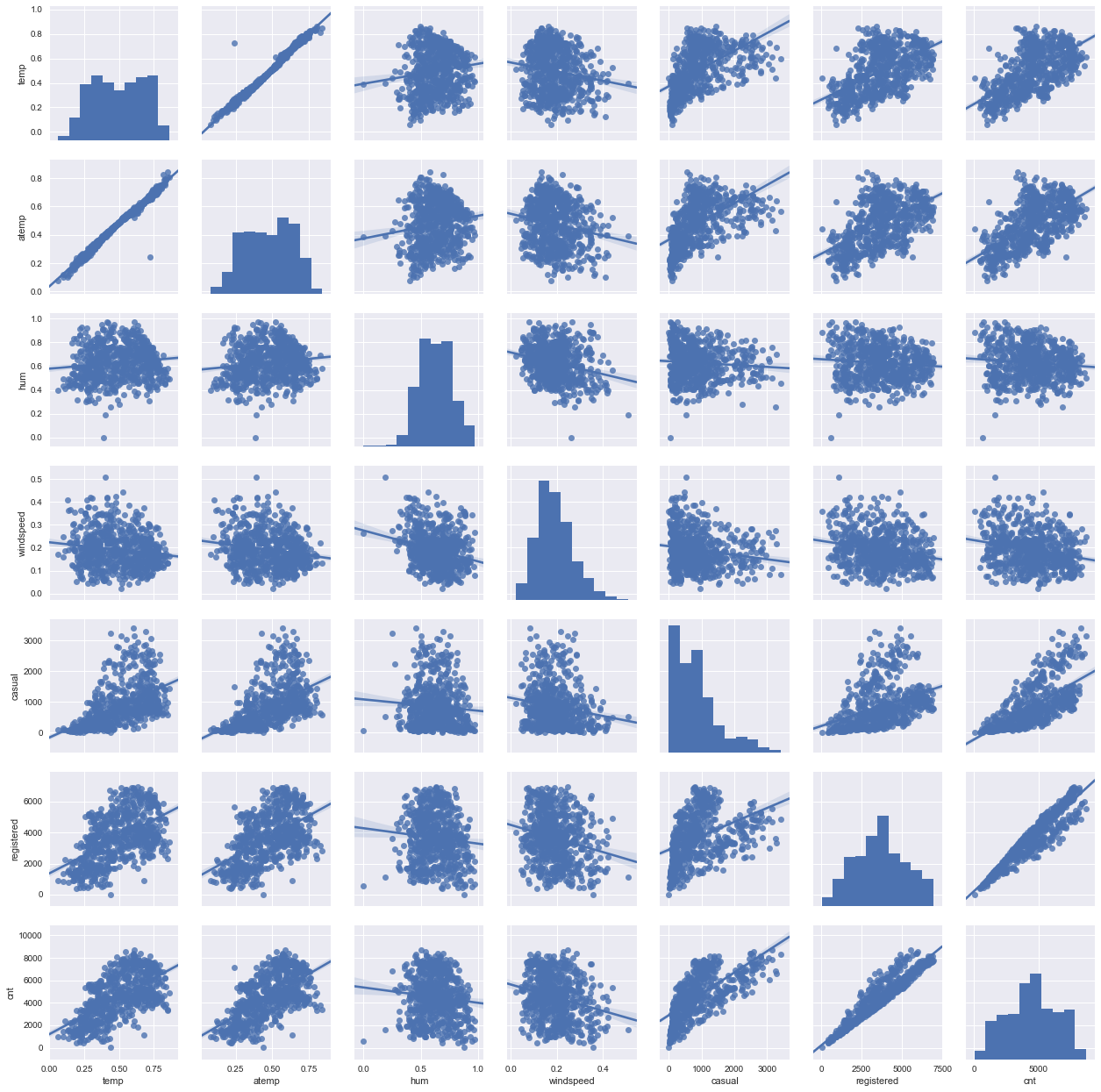
**2.1.2 Bivariate Analysis**

Ggpair function built upon [ggplot2](http://ggplot2.org/), [GGally](https://github.com/ggobi/ggally) provides templates for combining plots into a matrix through the ggpairs function. Such a matrix of plots can be useful for quickly exploring the relationships between multiple columns of data in a data frame. The lower and upper arguments to the ggpairs function specifies the type of plot or data in each position of the lower or upper diagonal of the matrix, respectively. For continuous X and Y data, one can specify the smooth option to include a regression line.

Below figures shows relationship between independent variables and also with numeric target variable using ggpair

1. Below ggpair graph is showing clearly that relationship between independent variables ‘temp’ and ‘atemp’ are very strong.
2. The relationship between ‘hum’ , ‘windspeed’ with target variable ‘cnt’ is less.

Figure 2.3 relationship between numeric variables (python code in Appendix B)



**2.2.1 Missing Value Analysis**

Missing values in data is a common phenomenon in real world problems. Knowing how to handle missing values effectively is a required step to reduce bias and to produce powerful models.

Below table illustrate no missing value present in the data.

2.1 missing values

|  |  |  |
| --- | --- | --- |
| s.no | Variables | missing values |
| 1 | dteday | 0 |
| 2 | season | 0 |
| 3 | yr | 0 |
| 4 | mnth | 0 |
| 5 | holiday | 0 |
| 6 | weekday | 0 |
| 7 | workingday | 0 |
| 8 | weathersit | 0 |
| 9 | temp | 0 |
| 10 | atemp | 0 |
| 11 | hum | 0 |
| 12 | windspeed | 0 |
| 13 | casual | 0 |
| 14 | registered | 0 |

**2.2.2 Outlier Analysis**

The Other steps of Preprocessing Technique is Outliers analysis , an outlier is an observation point that is distant from other observations. Outliers in data can distort predictions and affect the accuracy, if you don’t detect and handle them appropriately especially in regression models..

As we are observed in fig 2.2 the data is skewed so, there is chance of outlier in independent variable ‘casual’ , one of the best method to detect outliers is Boxplot

Fig 2.4 shows presence of Outliers in variable ‘casual’ and relationship between ‘casual’ and ‘cnt’ before removing Outliers

Fig 2.5 shows boxplot of ‘casual’ after removing outliers and relationship between ‘casual’ and ‘cnt’ after removing outliers

Boxplot :-  boxplot is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles

Figure 2.4 ‘casual’ Baxoplot and relation between ‘cnt’ and’ casual’ (R code in Appendix B)

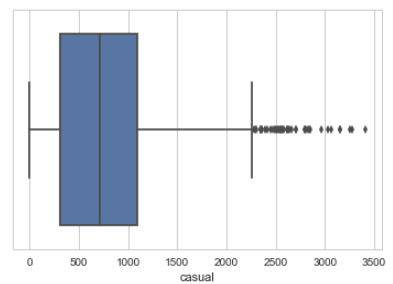
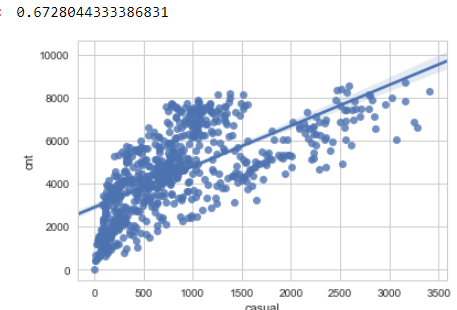
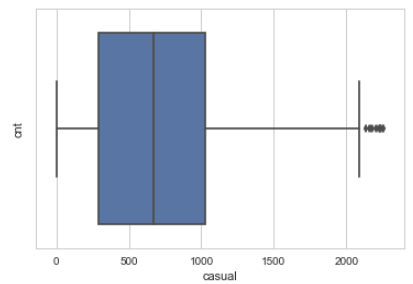
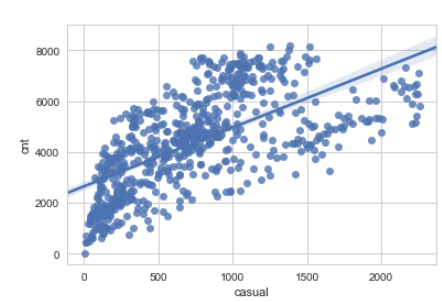
 

Figure 2.5 ‘casual’ Boxplot and relation between ‘casual’ and ‘cnt’ (R code in A  

Since there is significant difference between Pearson coefficient correlation between before and after outlier detection for ‘casual’ and ‘cnt’ and losing nearly 40 observation so, we are not going to treat the outliers.

Correlation before outliers : 0.67 and after outlier treatment is 0.64

**2.2.3 Features Selections**

Machine learning works on a simple rule – if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data.

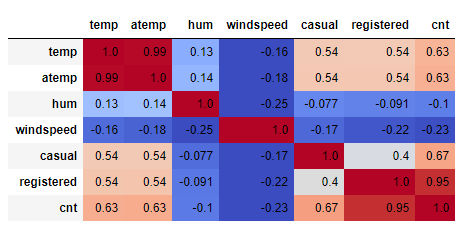
This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. I have myself witnessed feature subsets giving better results than complete set of feature for the same algorithm or – “Sometimes, less is better!”.

We should consider the selection of feature for model based on below criteria

1. The relationship between two independent variable should be less and
2. The relationship between Independent and Target variables should be high.

Below fig 2.6 illustrates that relationship between all numeric variables using Corrgram plot .

Figure 2.6 correlation plot of numeric variables (R code in Appendix B)



Color dark blue indicates there is strong positive relationship and if darkness is decreasing indicates relation between variables are decreasing.

Color dark Red indicates there is strong negative relationship and if darkness is decreasing indicates relationship between variables are decreasing.

Corrgram : it help us visualize the data in correlation matrices. correlograms are implimented through the **corrgram(x, order = , panel=, lower.panel=, upper.panel=, text.panel=, diag.panel=)**

**2.4.1 Dimensionality Reduction for numeric variables**

Above Fig 2.6 is showing

there is strong relationship between independent variables ‘temp’ and ‘atemp’ so considering any one feature enough to predict the better.

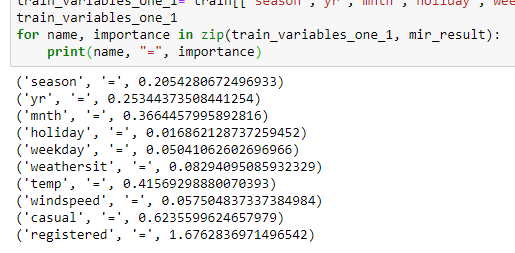
And it is also showing there is almost no relationship between independent variable ‘hum’ and dependent variable ‘cnt’. so, ‘hum’ is not so important to predict.

Subsetting two independent features ‘atemp’ and ‘hum’ from actual dataset.

**2.4.2 Dimensional Reduction using Random Forest Variable Importance**

There are several methods to check the relation between categorical variable , but here using Random Forest to get the importance of variables .

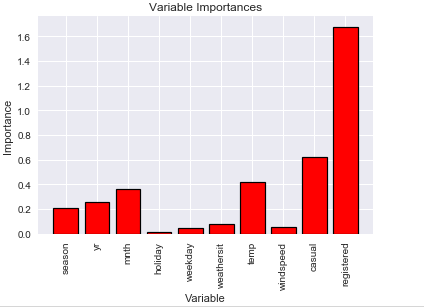
Figure 2.7 Variable Importance



The above figure shows that variabl’es ‘season’ ‘windspeed’ , ‘weekday ‘ ,‘weathersit’ and ‘holiday’ are less importance in predict the ‘cnt’ of Rental Bikes .

So these variable are removing while performaning Random Forest Model.

Figure 2.8 Variable Imporatnce Graph



**2.2.4 Features Scaling**

The word “normalization” is used informally in statistics, and so the term normalized data can have multiple meanings. In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to more easily compare data from different places. Some of the more common ways to normalize data include:

Transforming data using a [z-score](http://www.statisticshowto.com/probability-and-statistics/z-score/) or [t-score](http://www.statisticshowto.com/probability-and-statistics/t-distribution/t-score-formula/). This is usually called standardization. In the vast majority of cases, if a statistics textbook is talking about normalizing data, then this is the definition of “normalization” they are probably using.

[Rescaling data](http://www.statisticshowto.com/what-is-rescaling-data/) to have values between 0 and 1. This is usually called feature scaling. One possible formula to achieve this is.

[http://www.statisticshowto.com/wp-content/uploads/2015/11/normalize-data.png](http://www.statisticshowto.com/wp-content/uploads/2015/11/normalize-data.png)

In rental dataset numeric variables like ‘temp’ , ‘atem’ ,’hum’ and ‘ windspeed’ are in normalization form so , we have to Normalize two variables ‘casual’ and ‘registered’

After normalize ‘casual’ and ‘registered’ variables look like in table below where all values between 0 and 1

Table Normalization of ‘casual’ and ‘registered

|  |  |
| --- | --- |
| casual | registered |
| 0.037852113 | 0.09384926 |
| 0.034624413 | 0.17455963 |
| 0.025234742 | 0.21628646 |
| 0.042840376 | 0.21628646 |
| 0.019366197 | 0.12575801 |

**Chapter 3**

**Modelling**

**3.1 Model Selection**

In out earlier stage of analysis we have come to understand that few variables like ‘temp’ ,’casual,’registered ‘ are going to play key role in model development , for model development dependent variable may fall under below categories

1. Nominal
2. Ordinal
3. Interval
4. Ratio

In our case dependent variable is interval so, the predictive analysis that we can perform is **Regression** Analysis

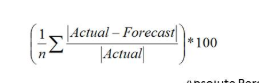
We will start our model building from Decision Tree .

**3.1.1 Evaluating Regression Model**

The main concept of looking at what is called **residuals** or difference between our predictions f(x[I,]) and actual outcomes y[i].

We are using two methods to evaluating performance of model

1. **MAPE** : (Mean Absolute Percent Error) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error.



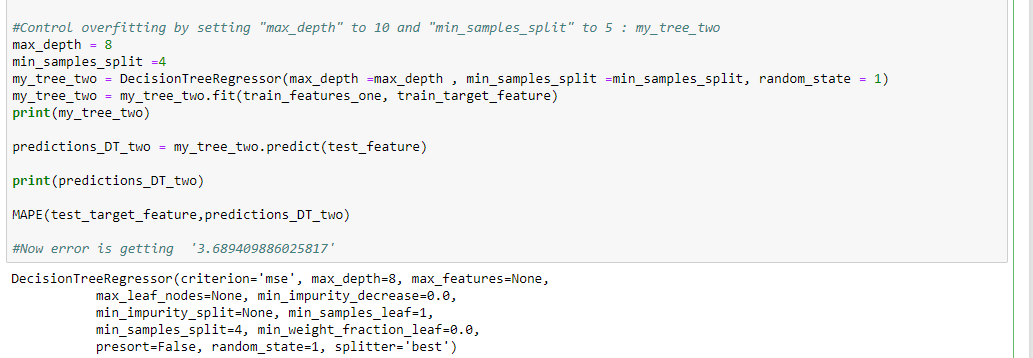
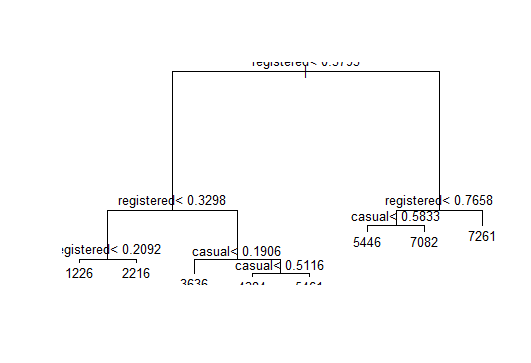
1. **RMSE :** (Root Mean Square Error) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled.



**3.2 Decision Tree**

A tree has many analogies in real life, and turns out that it has influenced a wide area of **machine learning**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

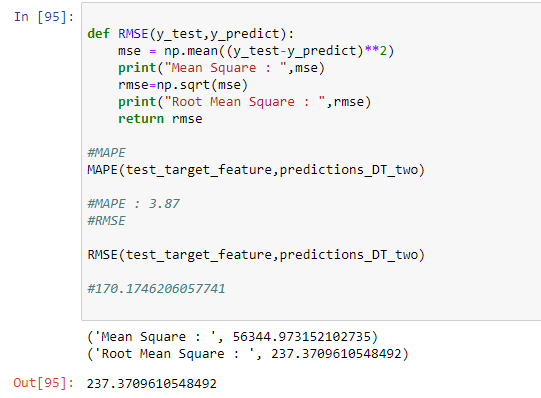
Figure 3.2.1 Decision Tree Algorithm

 Figure 3.2.2 Graphical Representation of Decision tree

Look at the above figure 3.2 here decision tree is using only two predictors variables to predict the model , which is not very impressive here the model is overfitted and biased towards only two predictors i.e ‘casual’ and ‘registered’ .

**3.2.1 Evaluation of Decision Tree Model**

Figure 3.2.3 Evaluation of Decision Tree using MAPE and RMSE



In Figure 3.2.3 Model Accuracy is 1- 3.8 = 0.962 which is nearly 96.2% it is quite good but RMSE is 237 which is very high so it’s clearly stating that our Decision Tree Model is Overfitted and it working well for training data but won’t predict good for new set of data. To overcome this overfit we have to tune the model using Random Forest.

**3.3 Random Forest**

Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

Random forest functions in below way

1. Draws a bootstrap sample from training data.
2. For each sample grow a decision tree and at each node of the tree
3. Ramdomly draws a subset of mtry variable and p total of features that are available
4. Picks the best variable and best split from the subset of mtry variable
5. Continues until the tree is fully grown.

As we saw in section 3.2 Decision tree is overfitting and its accuracy MAPE and RMSE is also poor in order to improve the performance of the model developing model using Random Forest.

Figure 3.3.1 Random Forest Implementation



Mtry : Number of variables to split at each node i.e. 7.

Nodesize : size of each node is 10

Our Random Forest model is looking quite good where it utilized maximum variables to predict the count values

**3.3.1 Evaluation of Random Forest**

Figure 3.2.2 Random Forest Evaluation

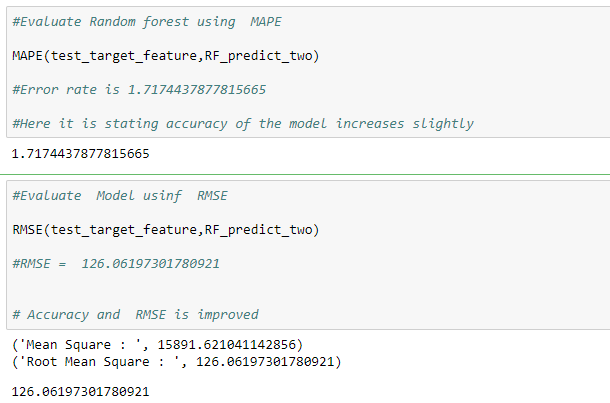


Fig 3.2.2 shows Random Forest model performs dramatically better than Decision tree on both training and test data and well also improve the Accuracy (MAPE = 1.71) and decrease the RMSE (126) of the model which is quite impressive.

Using Linear Regression we will predict the ‘cnt ‘ values and compare with Random Forest .

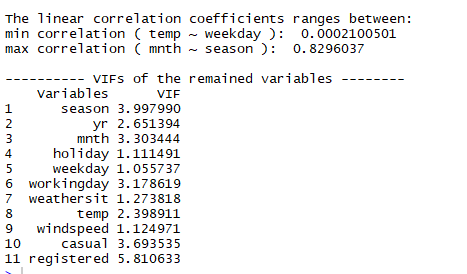
**3.4 Linear Regression**

[Multiple linear regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-multiple-linear-regression/) is the most common form of linear regression analysis.  As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.  The independent variables can be continuous or categorical.

**VIF ( Variance Inflation factor )** : It quantifies the multicollinearity between the independent variables.

As Linear regression will work well if multicollinearity between the Independent variables are less.

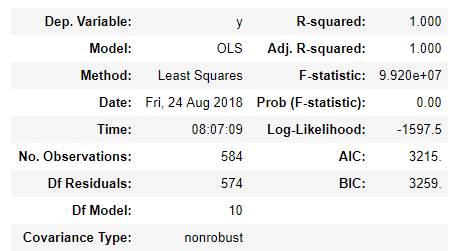
Figure 3.4.1 Multi collinearity between Independent variables



In the above figure it is showing there is strong correlation between two independent variable ‘mnth’ and ‘season’ so , it is enough to consider any one variable.

Figure 3.4.2 Multiple Linear Regression Model





Here :

**Residual standard error**: 3.231e-12 on 576 degrees of freedom

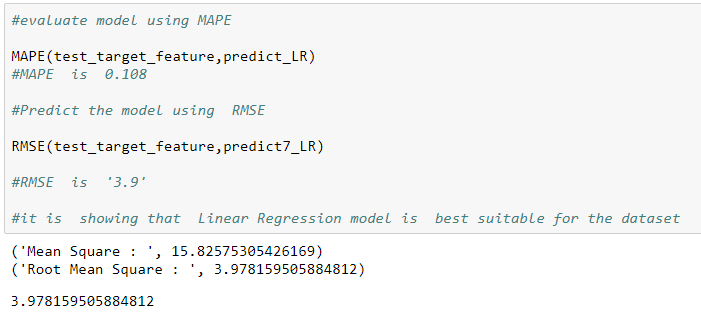
**Multiple R-squared**: 1, Adjusted R-squared: 1

Here residual Standard error is quite less so the distance between predicted values f(x[I,]) and actaual values f(x) are very less so this model is predicted almost accurate values.

And Multiple R-Square value is 1 so, we can explain about 100 % of the data using our multiple linear regression model. This is very impressive.

**3.4.2 Evaluation of Linear regression Model**

Figure 3.4.3 Evaluation of Regression Model



From above figure it is clearly showing that Model Accuracy is 99.9 % and RMSE is nearly equal to 3.9.

**Model Selection**

As we predicted counts for Bike Rental using three Models Decision Tree, Random Forest and Linear Regression as MAPE is high and RMSE is less for the Linear regression Model so conclusion is

**Conclusion**: - For the Bike Rental Data Linear Regression Model is best model to predict the count.

**Appendix A**

Figure 3.4 Relationship between Weekdays and cnt

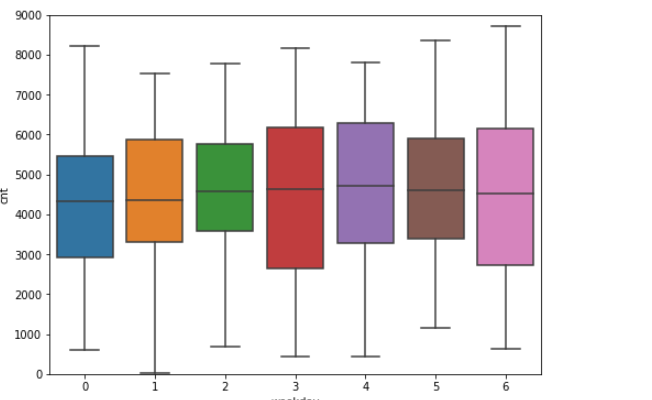
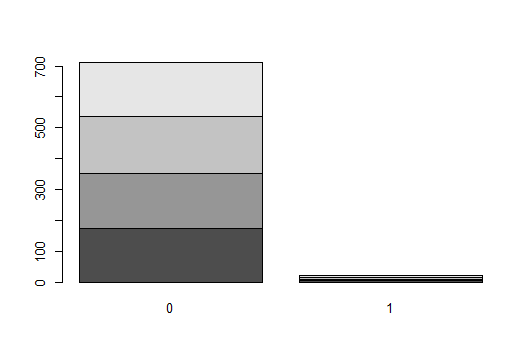


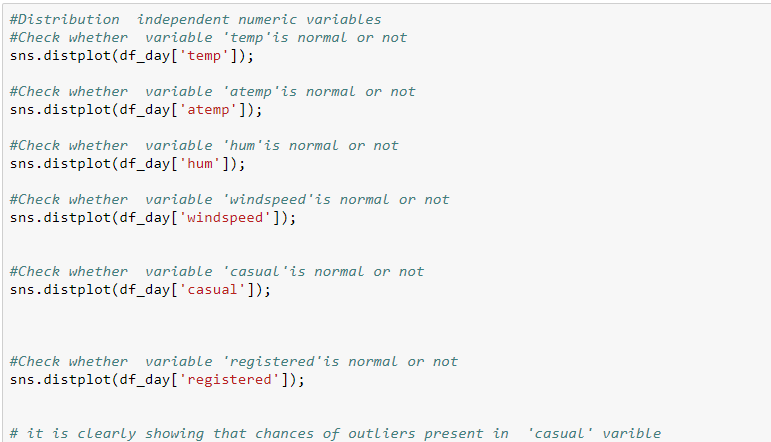
Figure 3.6 relationship between Holiday and Weathersit

Figure 3.8 shows relationship between ‘mnth’ and ‘holiday’



**Appendix B - Python Code**

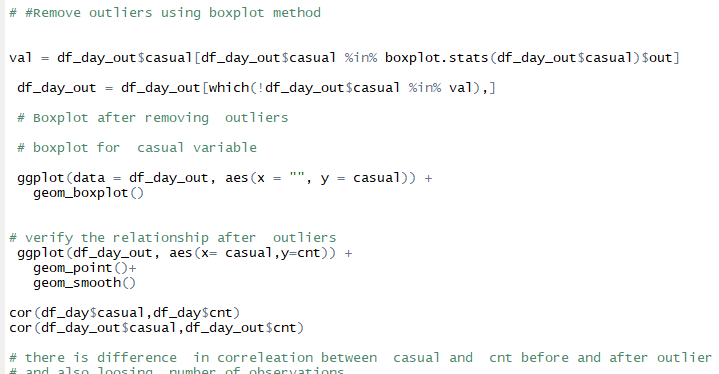
**Fig 2.1 and 2.2 Python Code**



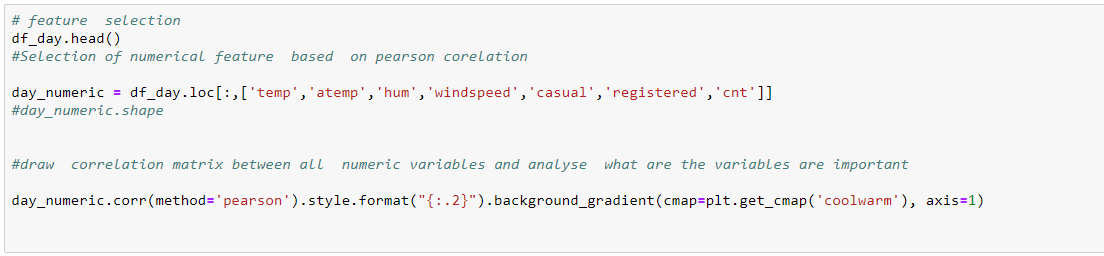
**Fig 2.3 Python Code**



**Fig 2.4 and 2.5 Python Code**



**Fig 2.6 Python Code**



**Complete Python File**

