**Predicting Bike Renting Counts**

**Yashwanth S G**

**12/11/2019**

**Contents:**

1. **Introduction** 
   1. Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   2. Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
2. **Methodology and pre-processing**

**2.1** Data analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

**2.1.1** Univariate analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

* + 1. Bivariate analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
    2. Missing value analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
    3. Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
    4. Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
    5. Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

1. **Modeling**
   1. Model Selection and Evaluating . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
   2. Decision tree . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .14
   3. Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
   4. Linear Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .16

**Appendix A** Extra figures

**Introduction:**

1. **Problem Statement:**

The Bike Rental Data contains the daily count of bikes rented between the year 2011 and 2012 with corresponding weather and seasonal information. The objective of this study is to predict the daily count of rental count. The prediction can helps to measure the business requirements. It also helps us to see which sort off weather and season is most profitable and in turn helps us to improve the needs and business perspectives.

1. **Data**

As our data is continuous, the main 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: 1-9)

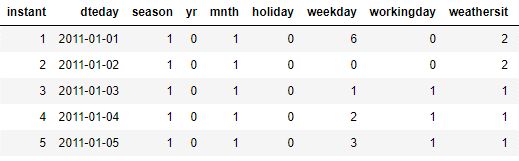
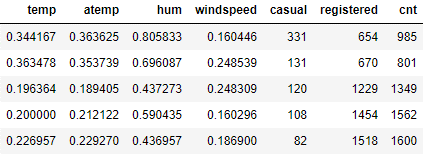


Table 1.2: Bike Rental Sample Data (Columns: 10-14)

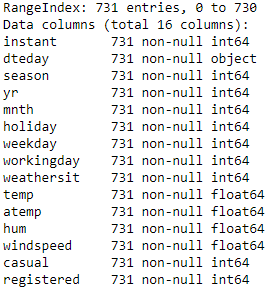


The dataset contains 731 observations and 16 variables of which first 15 variables are independent variable and 16th variable is dependent variable (predicting value ~ cnt).

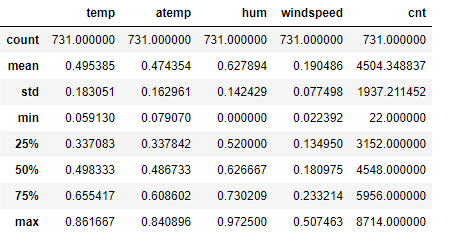
Below is the variable (dependent), which is to be predicted.



Below are the variables (independent) we used to predict the count of bike rentals.



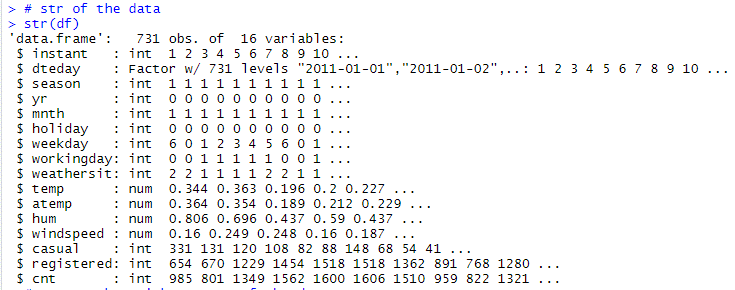
**Basic statistics of the dataset:**



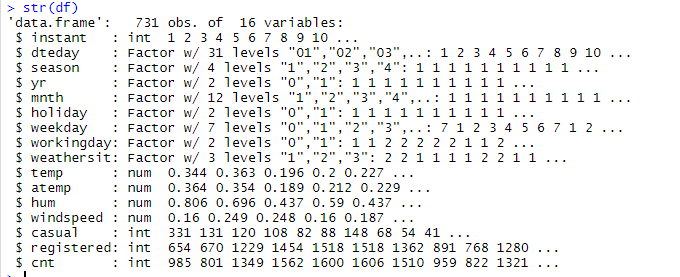
The average bike counts per day is 4504.

The minimum bikes rented per day is 22 and maximum bikes rented are 8714.

The structure of the dataset before processing:



The structure of the dataset after processing:



**Chapter 2**

**Methodology**

**2.1 Data Analysis:**

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 ‘registered’ 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)

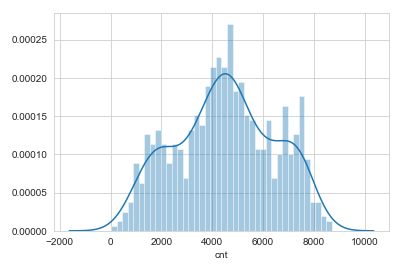
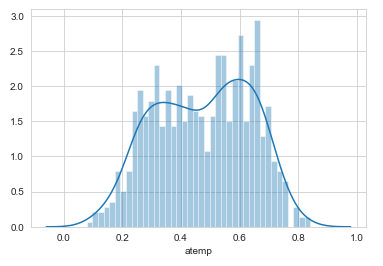
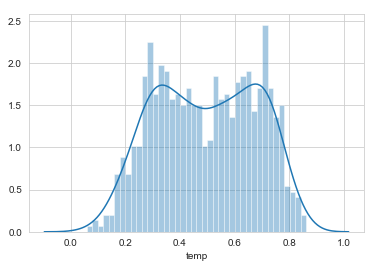
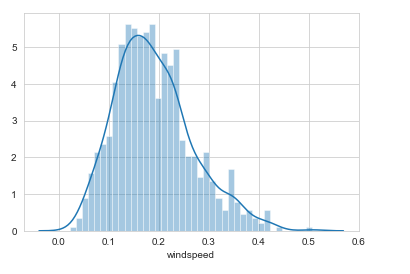
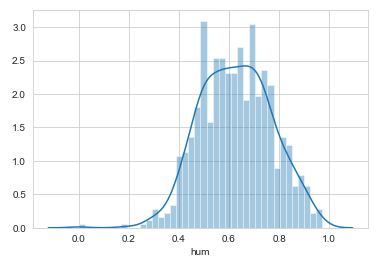
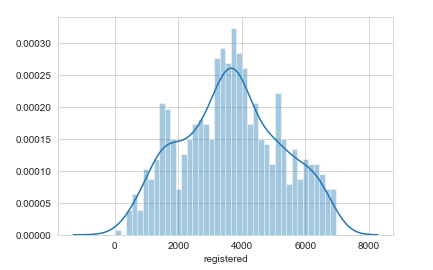
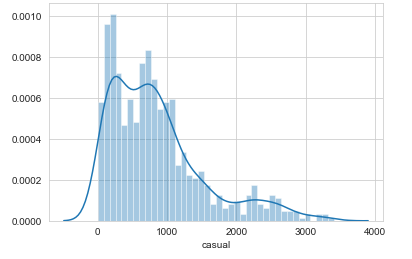


Figure 2.2 showing distribution of dependent variables.



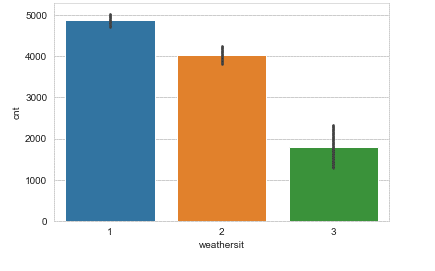




**2.1.2 Bivariate Analysis:**

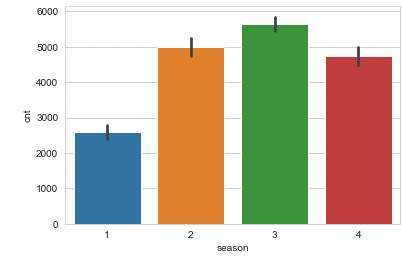
Bivariate analysis is plotting graphs with two variables. Bivariate analysis helps us to understand how two variables are related.

Figure 2.3 showing relation between weathersit vs cnt.



The above graph helps us to understand, weathersit 1 is have maximum average count of bike rents (i.e., Clear, Few clouds, partly cloudy, partly cloudy)

Figure 2.4 showing relation between season vs cnt.



The above graph helps us to understand, season 2 and 3 are have maximum average count of bike rents (i.e., fall and summer)

Figure 2.5 showing relation between temp and atemp vs cnt.

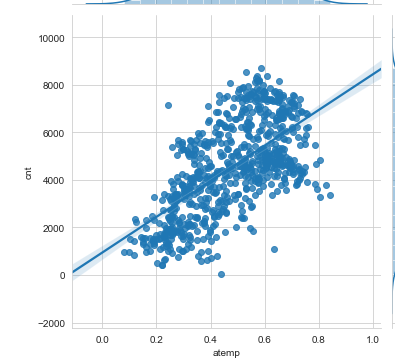
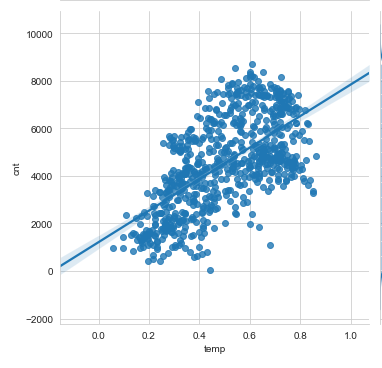


Figure 2.6 showing relation between hum and windspeed vs cnt.

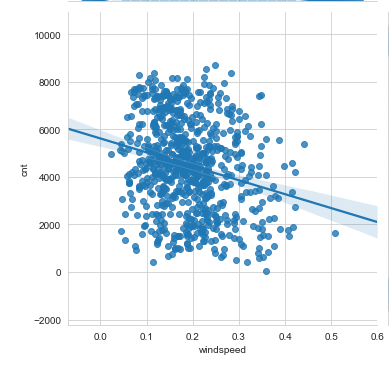
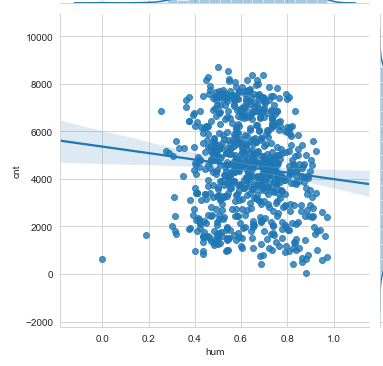
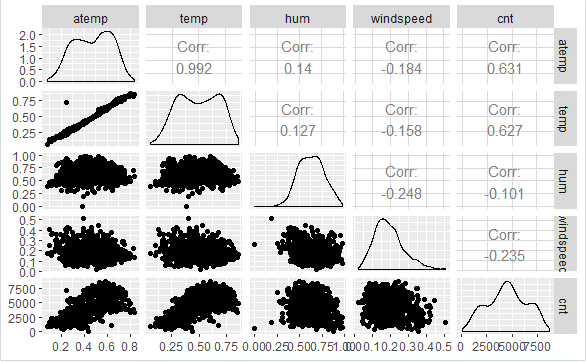


Figure 2.7 showing relation between hum and windspeed vs cnt.



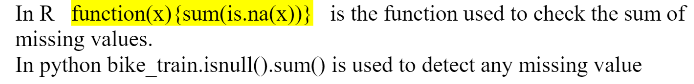
* From the above figures, temp and atemp are having correlation of 0.627 and 0.631 with cnt (Linear relation). To avoid multicollinearity atemp can be dropped from the dataset for modelling.
* Hum and cnt are having very less correlation with each other. So, hum can be removed from the dataset for modelling.
* Windspeed is having bit of negative linear relationship with cnt.

**2.1.3 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.

They are mean, median, KNN impute methods to deal with missing values.

Below table illustrate no missing value present in the data.





Note: Dataset is already uncoded and is in standard forms.

**2.1.4 Outlier Analysis:**

Outlier analysis is done to handle all the inconsistent observations in the dataset which are manmade or system errors. Outlier analysis is done only on the continuous variables. Not treating the outliers will lead to poor accuracy. Below shown figures are visualizations of numeric variable present in our dataset to detect outliers using boxplot.

Figure 2.8 showing boxplot to detect outliers ( casual and registered).

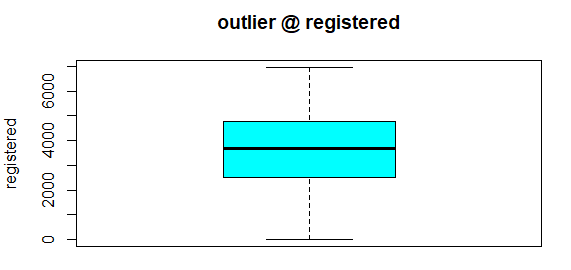
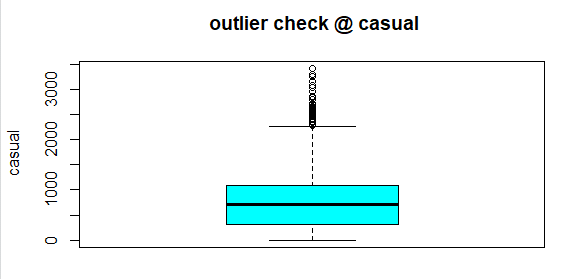


Figure 2.9 showing boxplot to detect outliers ( windspeed and cnt).

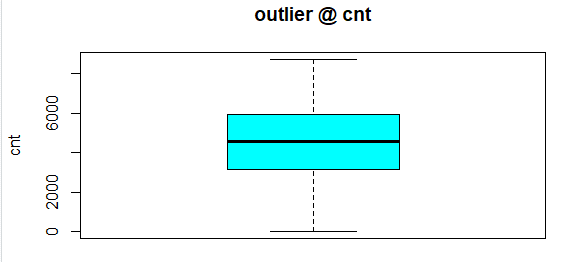
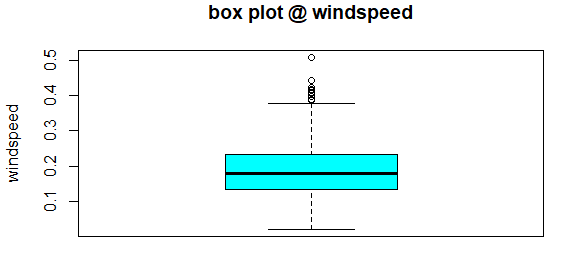
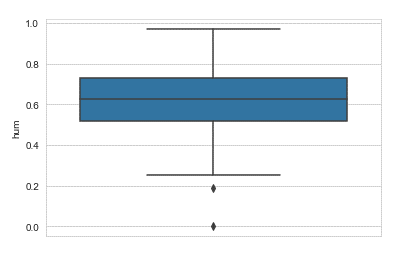


Figure 2.10 showing boxplot to detect outliers (hum).



As the variables “temp”, “atemp” are normally distributed, no need of checking for outliers.

According to above visualizations there are some outliers in windspeed, hum and casual variables.

As windspeed and hum defines the windspeed and humidity on a particular day. So, we can neglect these outliers because both these variables define environmental conditions. Due to drastic change in weather like heavy rain or strome. Casual variable is treated with outliers. The method used to remove outlier is boxplot method.

We can plot the relation between cnt and casual. If the Pearson coefficient correlation before and after outlier detect is significant we can leave the outliers as it is.

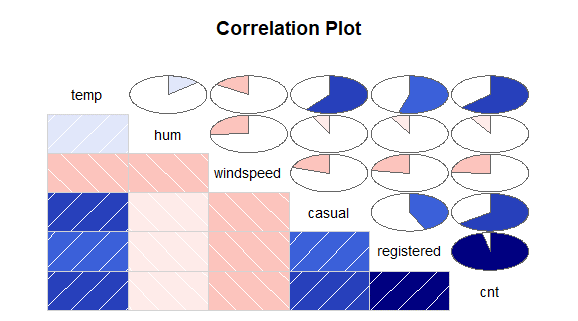
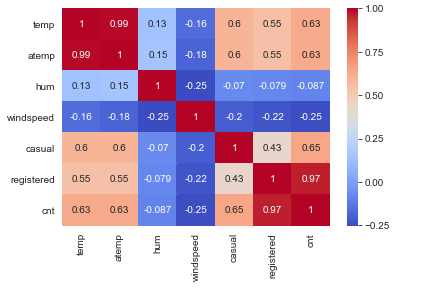
**2.1.5 Features Selection:**

Feature selection analysis is done to select relevant features (variables) to model construction**.** As the target variable is continuous we can go with correlation test.

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

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

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



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.

**Dimension reduction:**

Above Fig 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.

Sub setting two independent features ‘atemp’ and ‘hum’ from actual dataset and also removing instant variable as it is only a record number.

**2.1.6 Features Scaling:**

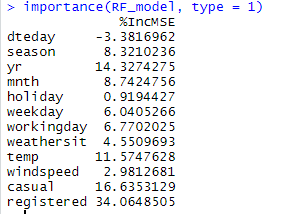
Feature scaling includes two functions normalization and standardization. It is done to reduce unwanted variation either within or between variables and to bring all of the variables into proportion with one another.

In the given dataset all the values are in normalized form.

**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 2.11 Variable Importance:



From the above the figure, we can see that casual, registered, temp, yr are very important variables and explain the target variable much. We can remove dteday as it is not explaining variance in target variable.

**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.

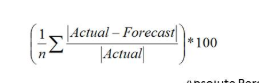
We will start our model building from Decision Tree.

**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.

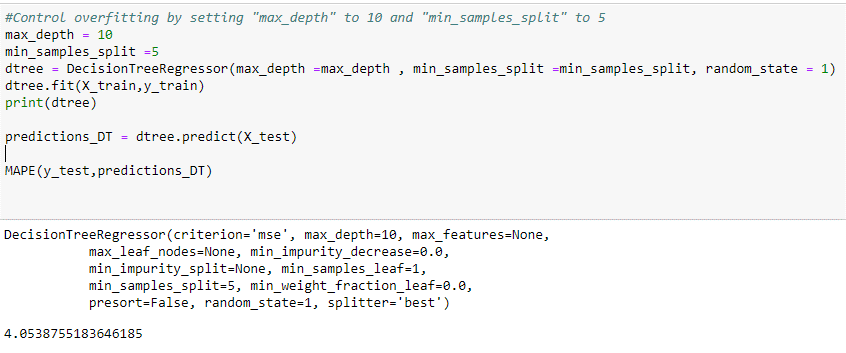


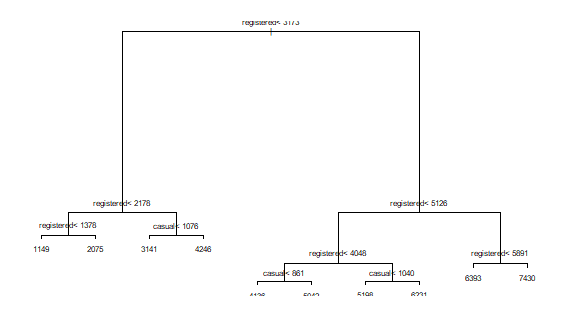
1. **RMSE:** (Root Mean Square Error) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled.



**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 3.1 Decision Tree Algorithm





Looking at the above figure 3.1 here decision tree is using only two predictors variables to predict the model , which is not very impressive here the model is over fitted and biased towards only two predictors i.e. ‘Casual’ and ‘registered’.

**Evaluation of Decision Tree Model:**

Figure 3.2 Decision Tree Algorithm

In Figure MAPE is 3.96 which is nearly 96 % it is quite good but RMSE is 185.85 which is very high so it’s clearly stating that our Decision Tree Model is Over fitted and it working well for training data but won’t predict good for new set of data. To overcome this over fit we have to tune the model using Random Forest.

**3.3 Random Forest**

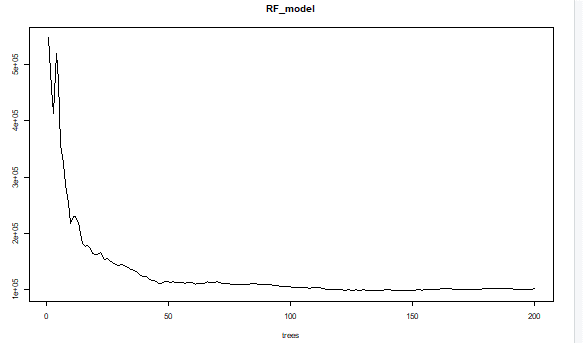
Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by 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 [over fitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

Random forest functions in below way

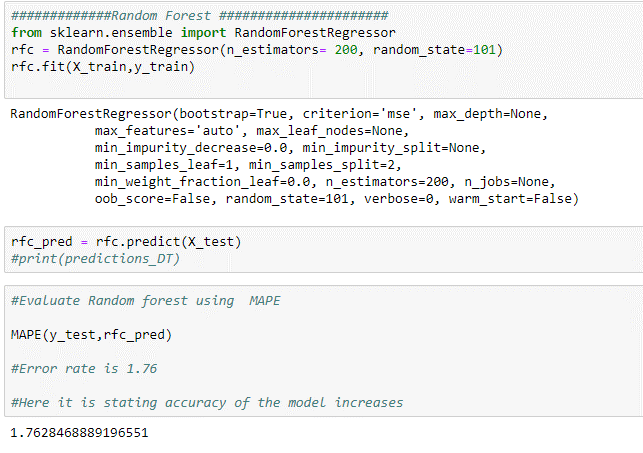
1. Draws a bootstrap sample from training data.
2. For each sample grow a decision tree and at each node of the tree
3. Randomly draws a subset of mtry variable from p total of features that are available
4. Picks the best variable and best split from the subset of mtry variable
5. Continues until the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

The figure below represents the curve of error rate as the number of trees increases. After 200 trees the error rate reaches to be constant.

In this model we are using 200 trees to predict the target variable.



As we saw in section 3.1 and 3.2 Decision tree is over fitting and its accuracy MAPE and RMSE is also poor in order to improve the performance of the model developing model using Random Forest.



**Evaluation of Random Forest**

Figure 3.3 Random Forest Evaluation

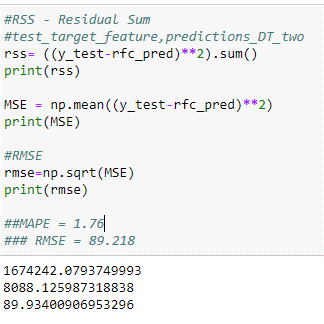


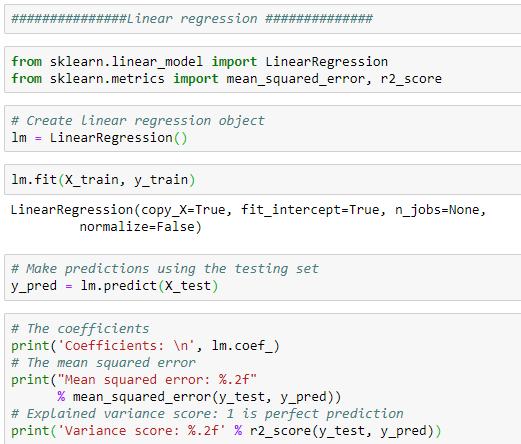
Fig 3.3 shows Random Forest model performs dramatically better than Decision tree on both training and test data and well also improve the Accuracy (MAPE = 1.76) and decrease the RMSE (89.218) of the model which is quite impressive and % of variance explained is 98.11 is good.

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

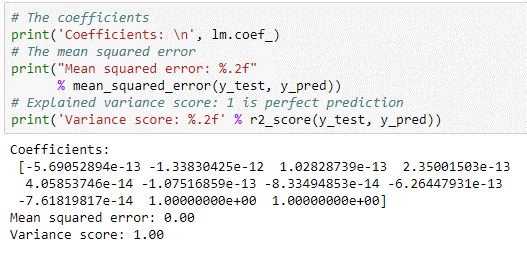
**3.4 Linear Regression**

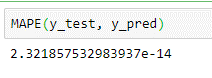
[Multiple linear regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-multiple-linear-regression/) is the most common form of linear regression analysis.  As a predictive analysis, the multiple linear regression 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.

Figure 3.4 Random Forest Evaluation



**Evaluation of Linear regression Model**





From above figure it is clearly showing that Model Accuracy is 99.9 % and R2 score is 1 which means perfect prediction.

**Model Selection**

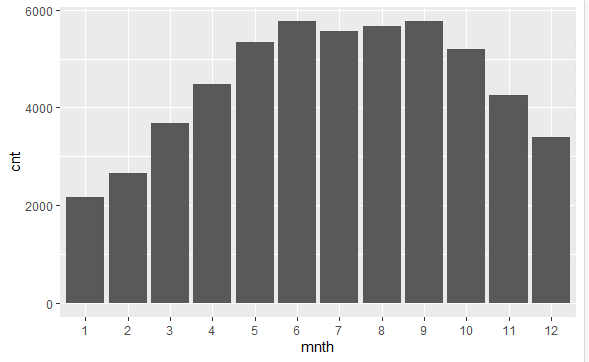
As we predicted counts for Bike Rental using three Models Decision Tree, Random Forest and Linear Regression as accuracy is high and RMSE is less for the linear regression Model.

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

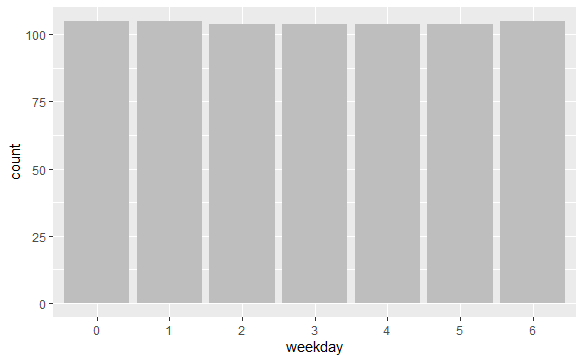
Further we can study, how casual or registered variable can be a good target variable.

Appendix – A: Extra Figures and visualizations.

Relation between mnth and cnt



Most number of rents are on holiday 0.



Most number of rents are on weathersit 1.

