**Krinal Chakraborty**

 chakraborty.krinal3@gmail.com

Bike Rent Count prediction

Contents

* 1. Problem statement Page 2
  2. Data Understanding Page 2
  3. Data Pre-Processing Page 3
  4. Missing Value Analysis Page 3
  5. Univariate Analysis Page 4
  6. Bivariate Analysis Page 6
  7. Time Series Visualization Page 8
  8. Feature Selection Page 9
     1. Correlation Analysis Page 9
     2. Chi-Square Test Page 9
     3. ANOVA Test Page 10
     4. Multicollinearity Test Page 10
  9. Feature Scaling Page 10
  10. Modelling Page 11
      1. Linear Regression Page 11
      2. Decision Tree Page 12
      3. Support Vector Machine Page 12
      4. K Nearest Neighbour Page 13
      5. Random Forrest Page 14
      6. Gradient Boosting Page 14
  11. Evaluation Page 15
  12. References Page 16

## Problem statement

The objective of this Case is to Predication of bike rental count on daily based on the

environmental and seasonal settings.

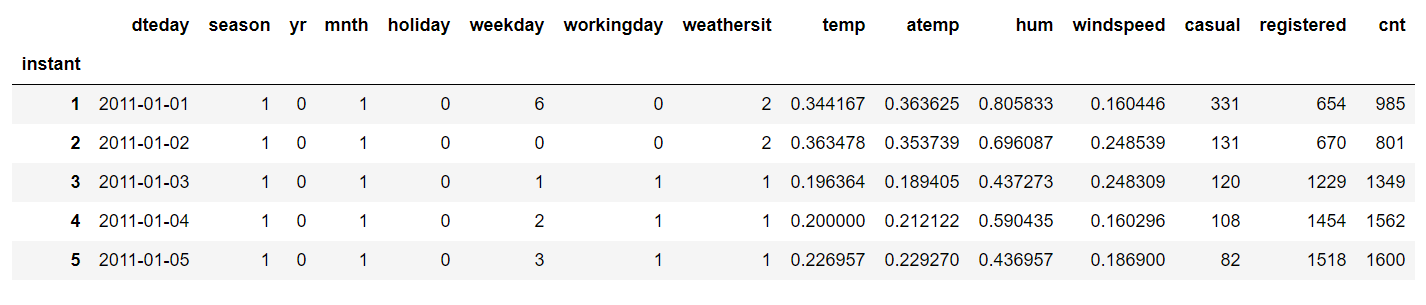
## Data Understanding

The data set consist of 731 observation recorded over a period of 2 years, between 2011 and 2012. It has 15 predictor variables and 1 target variable.

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

|  |  |
| --- | --- |
| Variable names | Description |
| 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 | Whether a day is holiday or not (extracted from Holiday Schedule) |
| Weekday | Day of the week |
| Workingday | If day is neither weekend nor holiday is 1, otherwise is 0. |
| weathersit | 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- t\_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 |

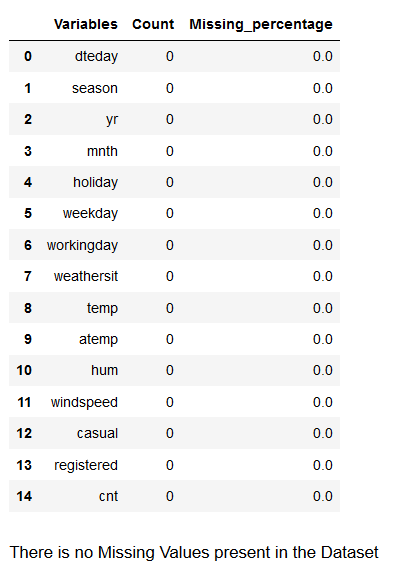
The data set consist of 7 continuous and 8 categorical variables. Sample data is shown below.



## Data Pre-Processing

Data pre-processing is the first stage of any type of project. In this stage we explore the data. This stage involves data cleaning, merging, sorting, Checking the missing values and outliers in the data, Imputing missing values if found by various methods such as mean, median, mode, KNN imputation, etc. We look at various plots of independent variables vs target variables. This stage is called as Exploratory Data Analysis.

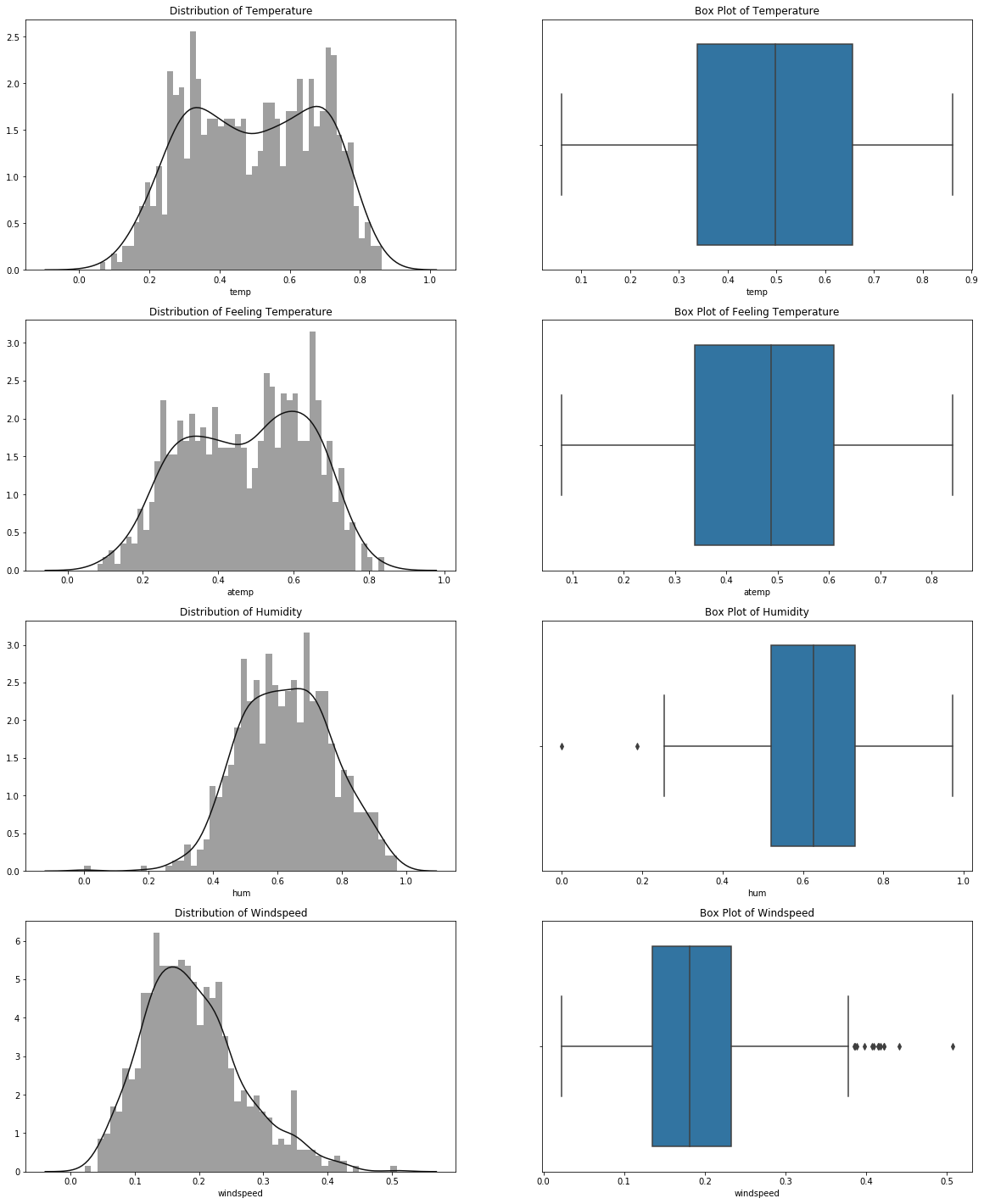
## Missing Value Analysis

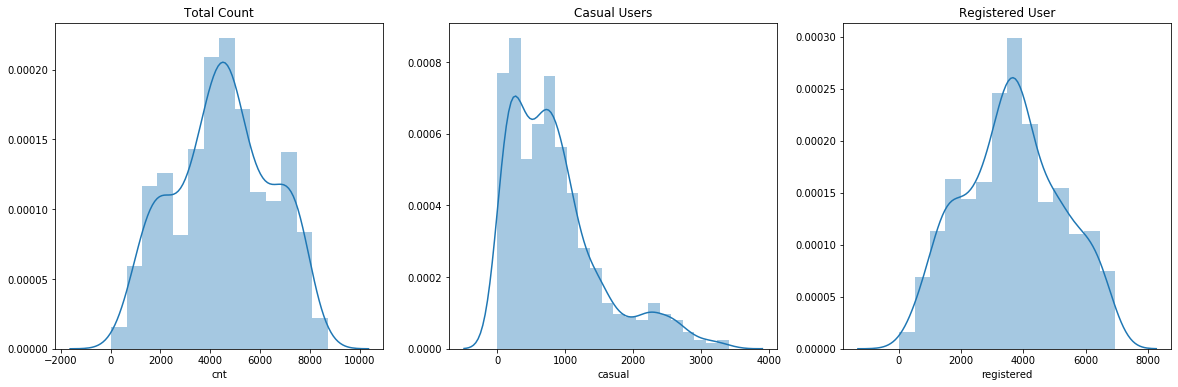
Missing value is availability of incomplete observations in the dataset. These Missing values affect the accuracy of model. So, it becomes important to check missing values in our given data.

## Univariate Analysis:

Distribution of continuous variables:

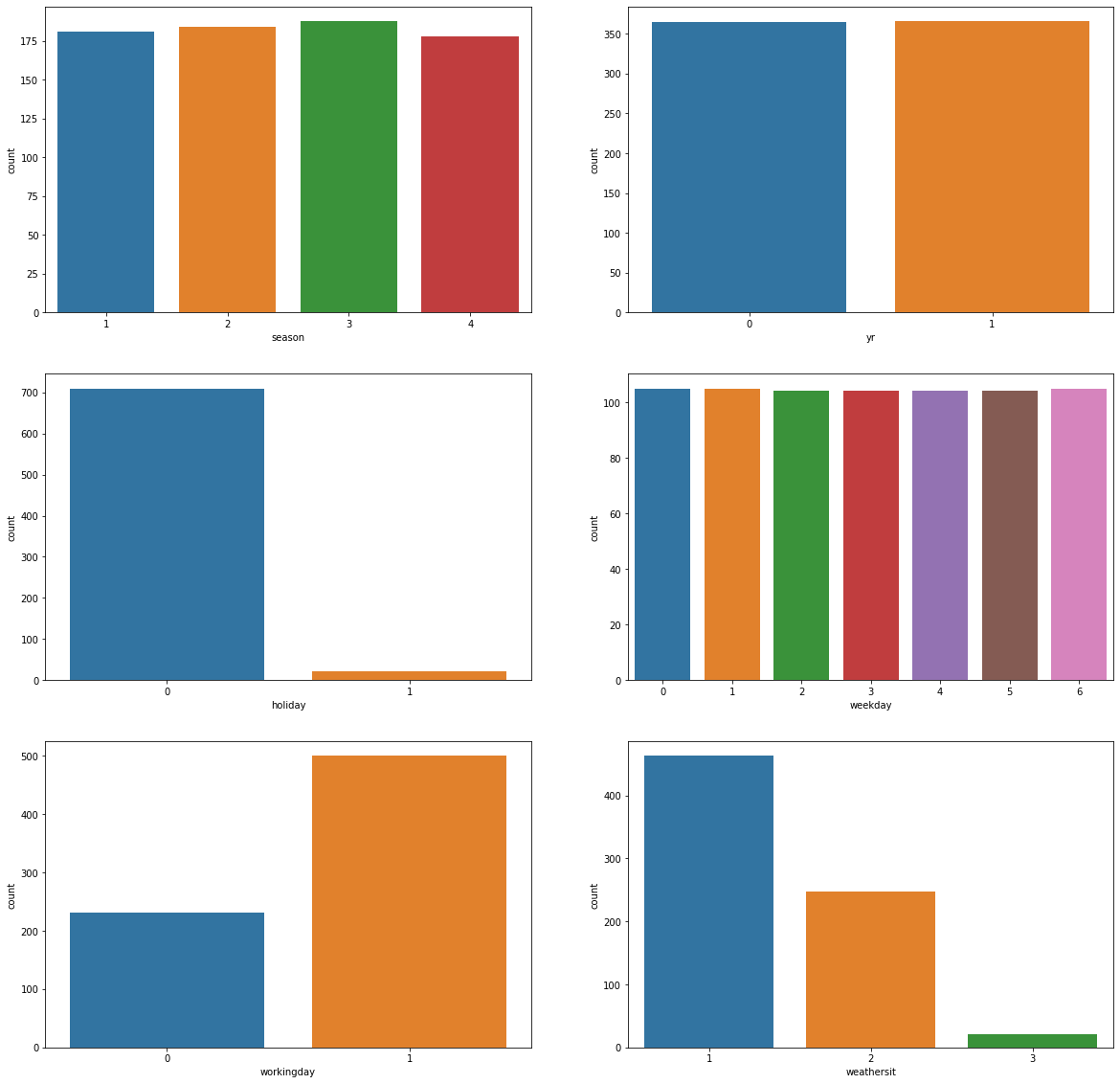
It can be observed from the below histograms is that temperature and feel temperature are normally distributed, where as the variables windspeed and humidity are slightly skewed. The skewness is likely because of the presence of outliers and extreme data in those variables.





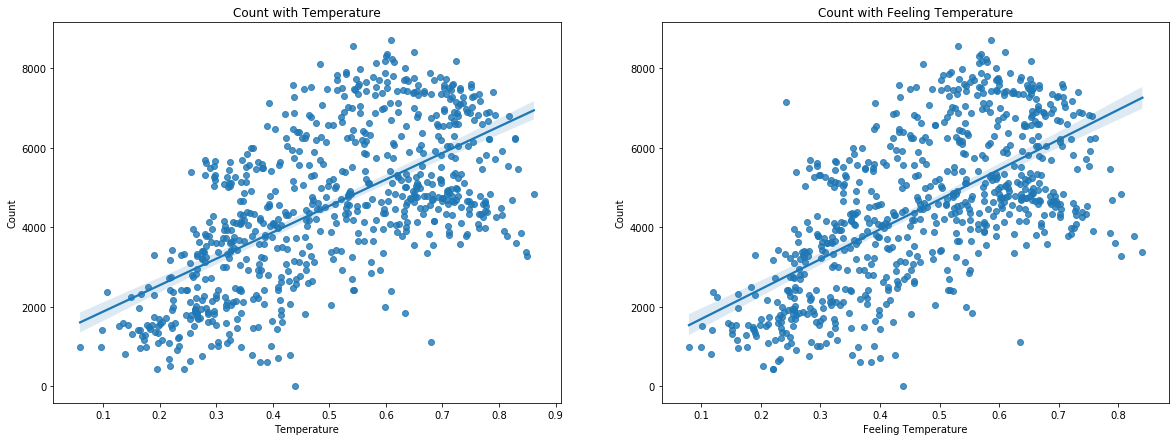
Distribution of categorical variables:

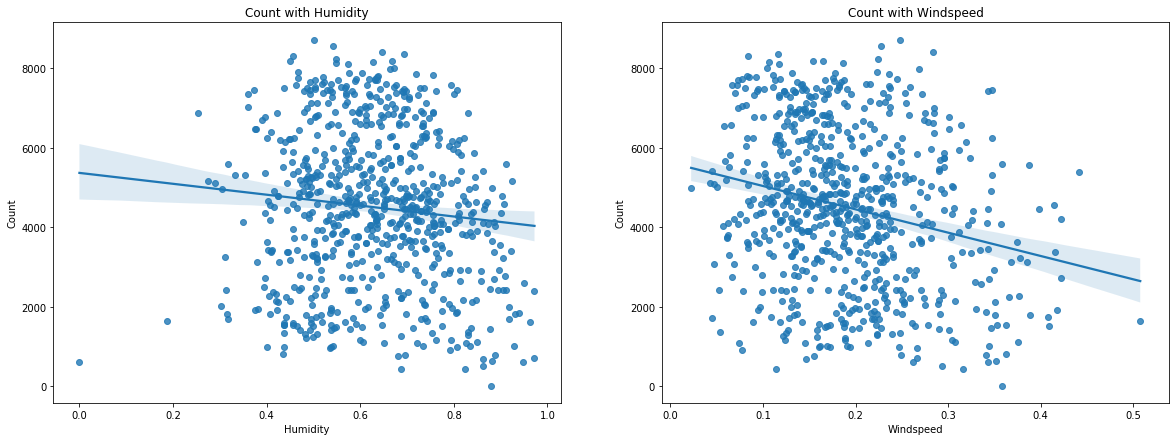
The distribution of categorical variables is as shown in the below figure:



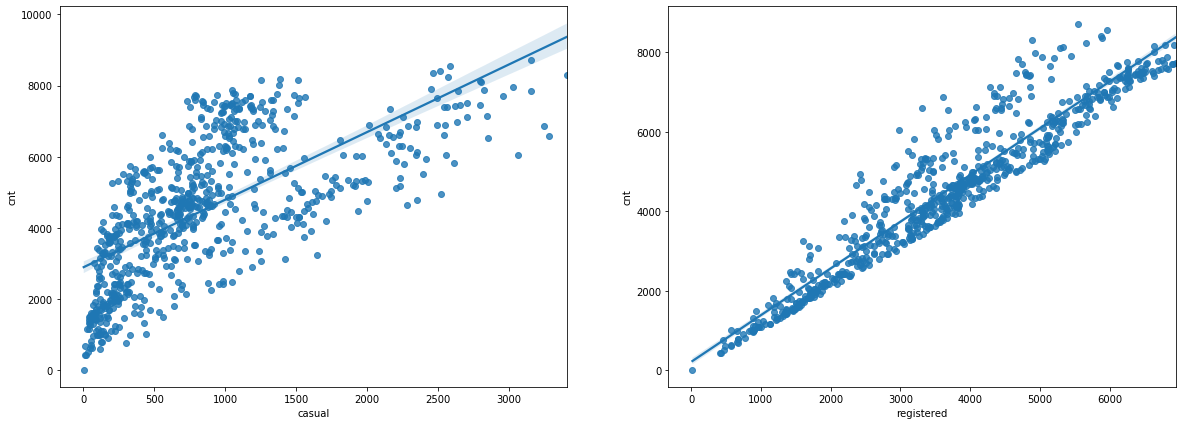
## Bivariate Analysis

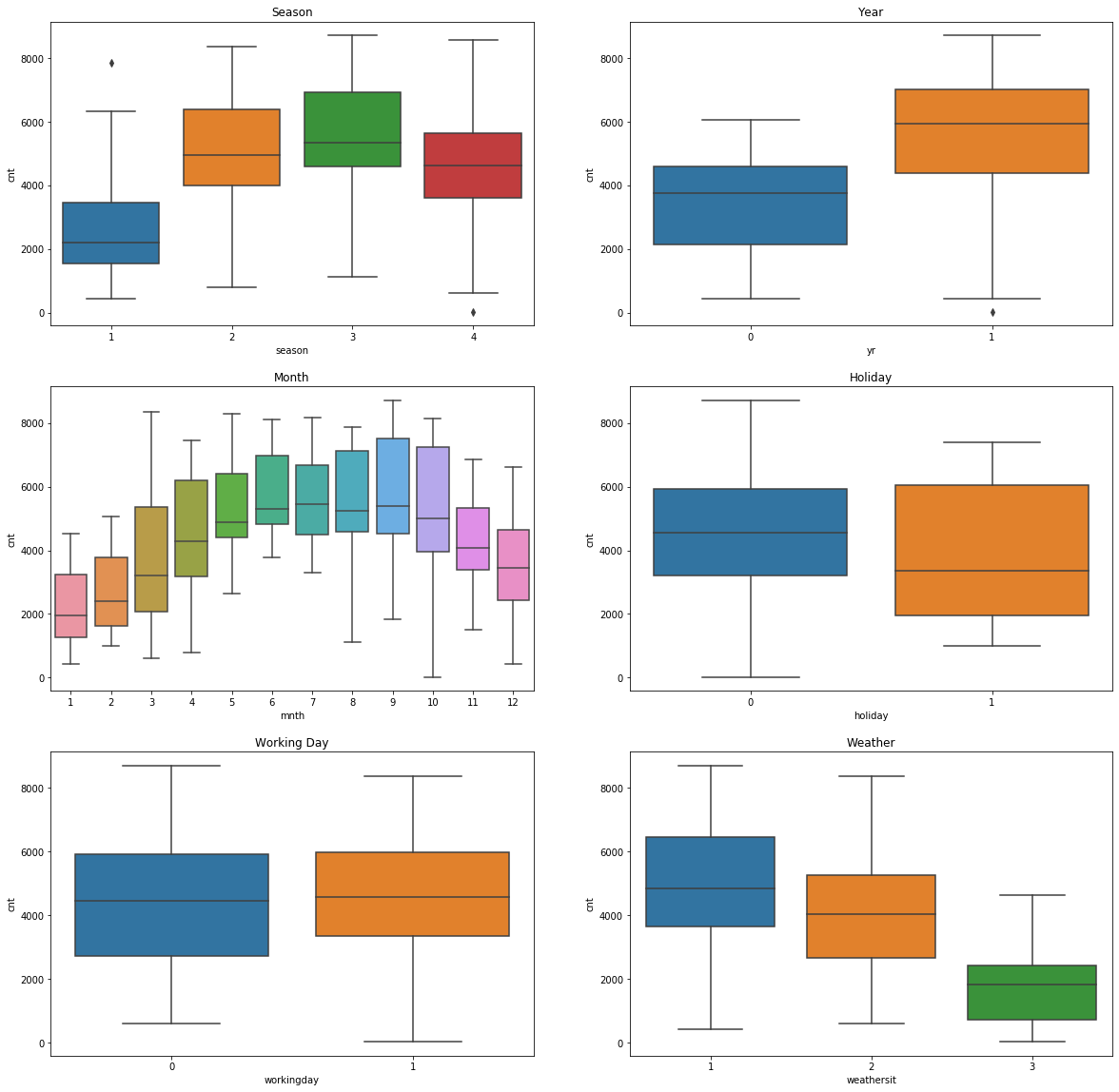
In bivariate analysis, we will look at the relationship between target variable and predictor.





Plot of Count with Casual Users and Registered Users:

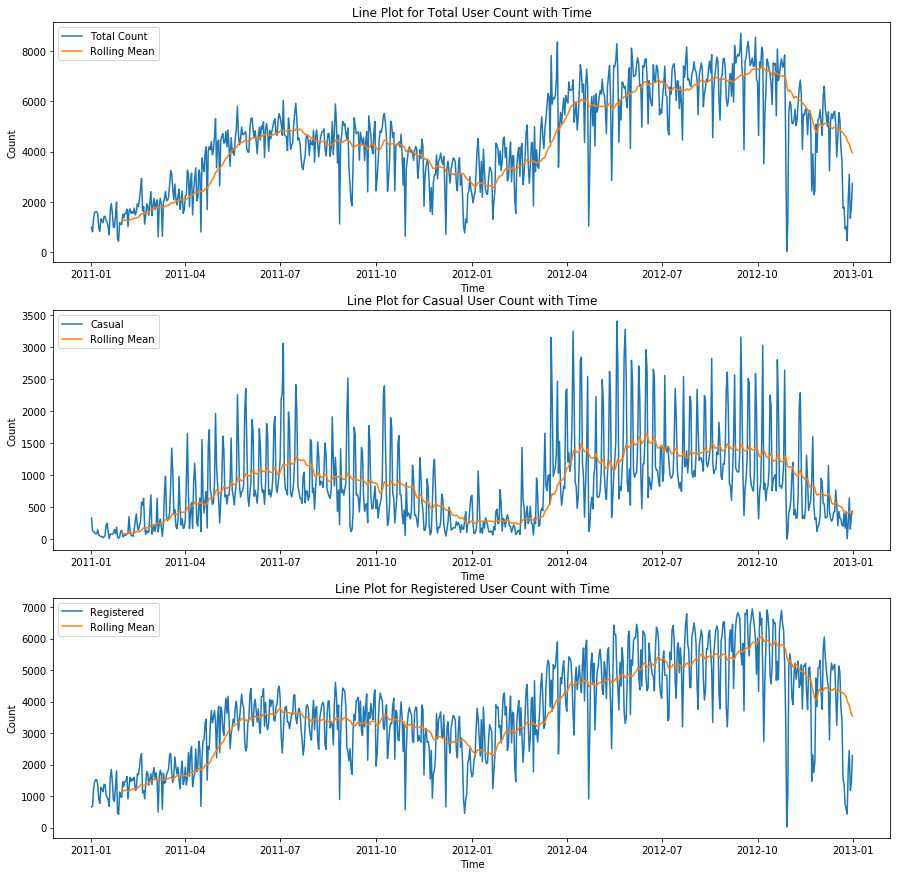


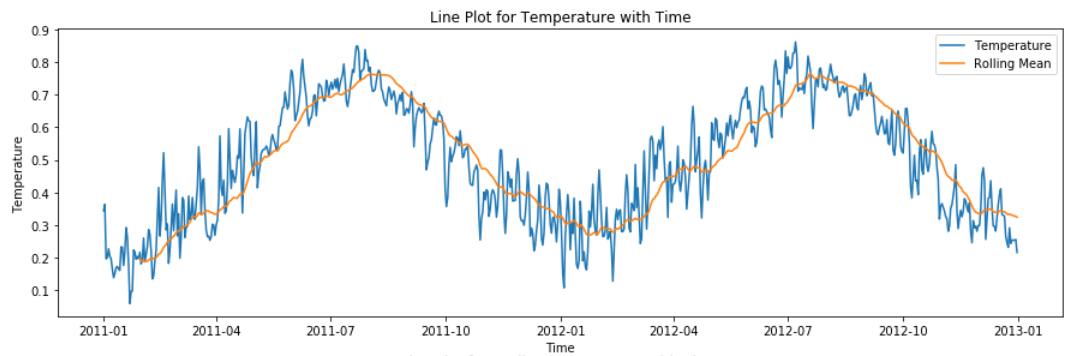


From the above plots, we can see that:

* ‘cnt’ with ‘temp’ and ‘atemp’ have strong and positive relationship. It means that as the temperature rises, the bike demand also increase.
* ‘hum’ has a negative linear relationship with ‘cnt’, but the relationship is quite flat
* ‘windspeed’ has negative linear relationship with ‘cnt’. With an increase in windspeed, bike count decreases.
* There is no significance difference between different seasons. The count is highest for fall season and lowest for spring season.
* Bike demand was higher in 2012 as compared with 2011.
* The count is maximum when weather situation is good and least when conditions are bad.
* We can see that count changes as the weather changes from cold to hot.

## Time Series Visualization:



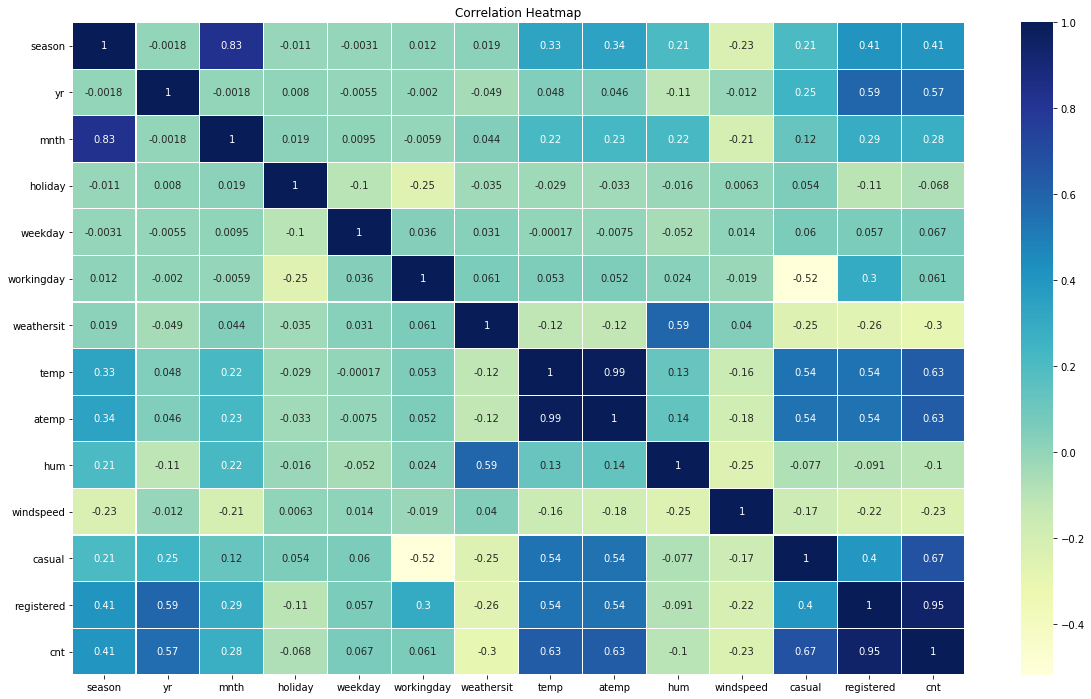


## Feature Selection

All the variables in our data may not be accurate enough to predict the target variable or contain redundant information and cause problem of multicollinearity, in such cases we need to analyse our data and select the dataset variables that can be most useful for our model. Feature selection helps by reducing time for computation of model and also reduces the complexity of the model.

* **Correlation Analysis:**

We plot the correlation heatmap to see the correlation of different continuous variable in our dataset and choose the features accordingly.



From above correlation plot we see that:

* 'temp' and 'atemp' are very highly correlated with each other.
* 'registered' and 'cnt' are highly correlated with each other.
* **Chi-Square test of independence:**

Chi-square test compares 2 categorical variables in a contingency table to see if they are related or not.

Null hypothesis: 2 variables are independent.

Alternate hypothesis: 2 variables are not independent.

If p-value is less than 0.05 then we reject the null hypothesis saying that 2 variables are dependent and if p-value is greater than 0.05 then we accept the null hypothesis saying that 2 variables are independent.

• If p-value<0.05 then remove the variable

• If p-value>0.05 then keep the variable

* **Anova Test:**

Analysis Of Variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent groups.

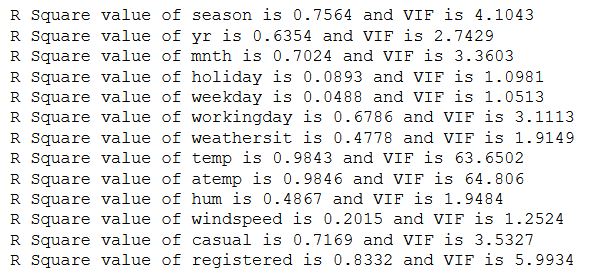
- Null Hypothesis: mean of all categories in a variable are same.

- Alternate Hypothesis: mean of at least one category in a variable is different.

If p-value is less than 0.05 then we reject the null hypothesis or else we accept the null hypothesis.

* **Multicollinearity:**

Multicollinearity exists whenever two or more of the predictors in a regression model are highly correlated. It is the condition when one predictor can be used to predict other. The basic problem is multicollinearity results in unstable estimation of coefficients which makes it difficult to access the effect of independent variable on dependent variable.



Therefore, analysing the different Feature Selection Techniques, we select the below features:

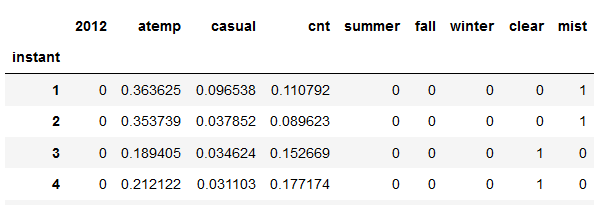


## Feature Scaling

Data Scaling methods are used when we want our variables in data to scaled on common ground. It is performed only on continuous variables.

• **Normalization**: Normalization refer to the dividing of a vector by its length. normalization normalizes the data in the range of 0 to 1. It is generally used when we are planning to use distance method for our model development purpose such as KNN. Normalizing the data improves convergence of such algorithms. Normalisation of data scales the data to a very small interval, where outliers can be loosed.

• **Standardization**: Standardization refers to the subtraction of mean from individual point and then dividing by its SD. Z is negative when the raw score is below the mean and Z is positive when above mean. When the data is distributed normally you should go for standardization.



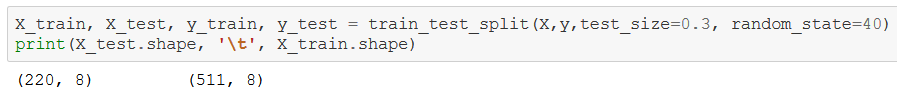
## Modelling

After a thorough pre-processing of data, we will use some regression machine learning models on our processed data to predict the target variable, as our target variable is a continuous data.

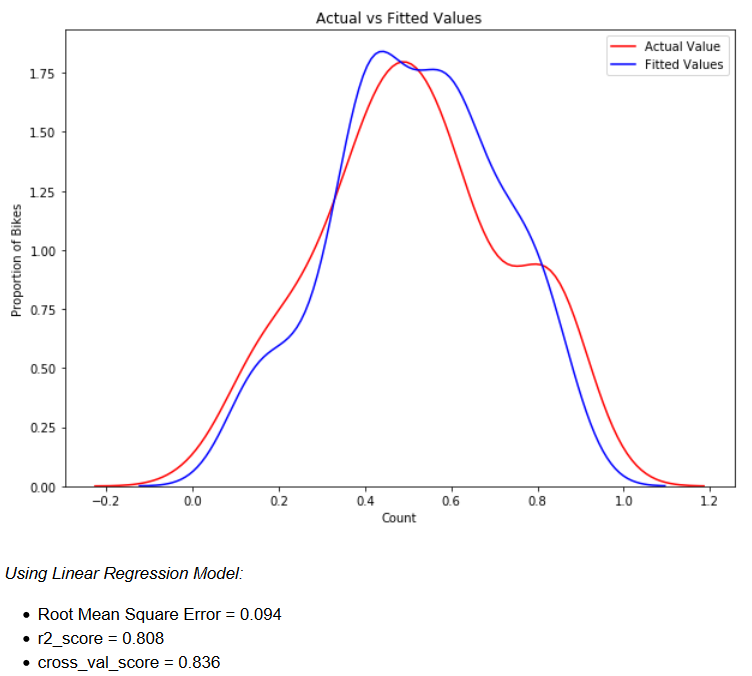
Following are the models which we have built:

* Linear Regression
* Decision Tree
* Random Forest
* K Nearest Neighbor
* Support Vector Machine
* Gradient Boosting

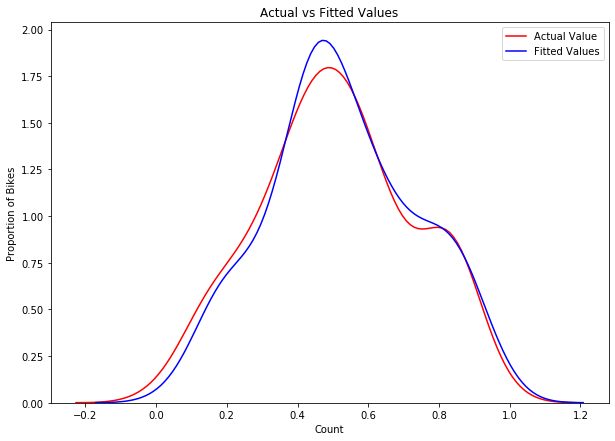
But, before running any model, we will split our data into two parts which is train and test data. Here in our case we have taken 70% of the data as our train data. Below is the snipped image of the split of train test.



# **Linear Regression**

Multiple Linear Regression is used as a predictive analysis, when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable. Below is the plot for Linear Regression Model with different accuracy score. 

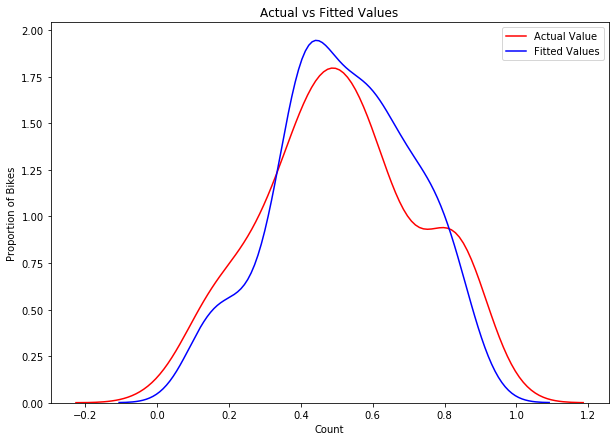
# **Decision Tree**

Decision Tree is a supervised learning predictive model that uses a set of binary rules to calculate the target value/dependent variable. 

*Using Decision Tree Model:*

* Root Mean Square Error = 0.094
* r2\_score = 0.825
* cross\_val\_score = 0.79

# **Support Vector Machine**

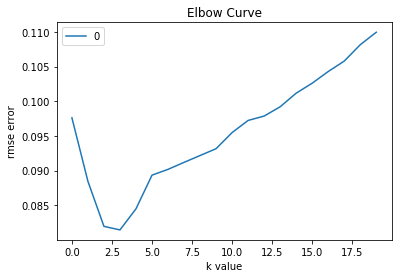
Support Vector Machine is a Supervised Learning technique used for Regregression as well as Classification problems.

*Using Support Vector Machine Model:*

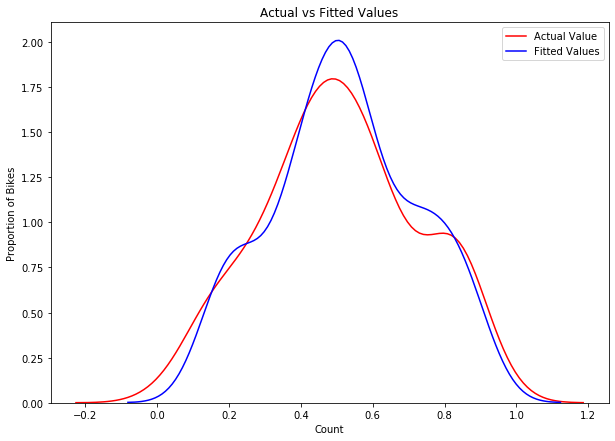
* Root Mean Square Error = 0.085
* r2\_score = 0.845
* cross\_val\_score = 0.835

# **K Nearest Neighbour**

The KNN model finds the nearest neighbours and tries to predict target value. The method goes as, for the value of new point to be assigned, this value is assigned on the basis of how closely this point resembles the other points in the training set. We use the elbow curve to find the optimum value of K for predicting the values in our data set. Below is the elbow curve.



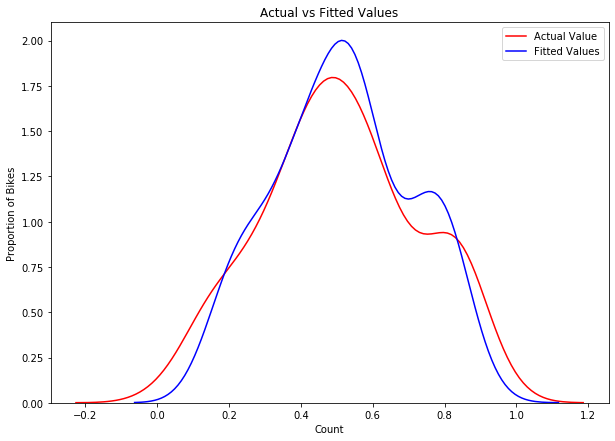
The value of K is selected to be 3 from the above elbow curve.



Using K Nearest Neighbour Model:

* Root Mean Square Error = 0.082
* r2\_score = 0.856
* cross\_val\_score = 0.846

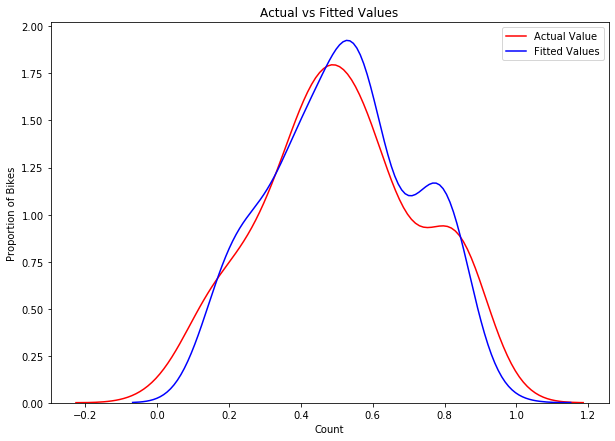
# **Random Forrest**

It is a process where the machine follows an ensemble learning method that operates by developing a number of decision trees and giving output as the mean prediction of the individual trees.

Using Random Forrest Model:

* Root Mean Square Error = 0.076
* r2\_score = 0.875
* cross\_val\_score = 0.876

# **Gradient Boosting**

Gradient Boosting is an ensemble Machine Learning Algorithm that uses the Boosting Algorithm by many weak learners using the gradient loss function.

Gradient Boosting Model:

* Root Mean Square Error = 0.072
* r2 score = 0.886
* cross\_val\_score = 0.874

## Evaluation

The various metrics used to evaluate the result of the predictions in our models are:

* Root Mean Squared Error - RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model. The errors are first squared before averaging which poses a high penalty on large errors.
* Coefficient of Determination or R² Error - R² is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. R² is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R² will always be less than or equal to 1.
* Cross-validation – cross\_val\_score is a resampling procedure used to evaluate machine learning models on a data sample for out of sample testing. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into.

# **Hyperparameter Optimization:**

To find the optimal hyperparameter we have used GridSearchCV for the models.

* Decision Tree:

{'max\_depth': 20, 'max\_features': 'auto', 'min\_samples\_split': 0.1}

* Support Vector Machine:

{'C' : 0.1, 'gamma' : 0.1, 'kernel' : 'linear'}

* Random Forrest:

{'max\_depth': 10, 'max\_features': 'sqrt', 'n\_estimators': 200}

* Gradient Boosting

{'learning\_rate':0.1,'max\_depth':5,'max\_features':'sqrt','n\_estimators':50}

So, applying different Machine Learning Model to predict the test values, we got the error metric as below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | R-Squared Score | Cross Validation Score |
| Linear Regression | 0.094 | 0.808 | 0.836 |
| Decision Tree | 0.094 | 0.825 | 0.790 |
| Random Forrest | 0.076 | 0.875 | 0.876 |
| K Nearest Neighbour | 0.082 | 0.856 | 0.846 |
| Support Vector Machine | 0.085 | 0.845 | 0.835 |
| Gradient Boosting | 0.072 | 0.886 | 0.874 |

So, from the above table, we can see the Root Mean Square Error is least and the R-squared score is highest for the model Gradient Boosting and Random Forrest Model. Thus we choose **Random Forrest Model** for predicting our target variable from the dataset given to us because it is giving a slightly better Cross Validation Score after Hyper parameters tuning.

## References

* [www.edwisor.com](http://www.edwisor.com)
* [www.cognitiveclass.ai](http://www.cognitiveclass.ai)
* [www.towardsdatascience.com](http://www.towardsdatascience.com)
* [www.stackoverflow.com](http://www.stackoverflow.com)
* [www.analyticsvidhya.com](http://www.analyticsvidhya.com)
* [www.medium.com](http://www.medium.com)
* [www.github.com](http://www.github.com)
* [www.youtube.com](http://www.youtube.com)