**Predication of bike rental count based on the environmental and seasonal settings**

**By**

**Amit Kumar Tiwari**

***Index***

1. **Introduction**
2. **Problem Statement**
3. **Data**
4. **Methodology**
5. **Pre-Processing**
6. **Modelling**
7. **Model Selection**
8. **Pre- processing**
9. **Data exploration and Cleaning (Missing Values and Outliers)**
10. **Creating some new variables from the given variables**
11. **Selection of variables**
12. **Some more data exploration**
13. **Dependent and Independent Variables**
14. **Uniqueness of Variables**
15. **Dividing the variables categories**
16. **Feature Scaling**
17. **Modelling**
18. **Linear Regression**
19. **Decision Tree**
20. **Random Forest**
21. **Conclusion**
22. **Model Evaluation**
23. **Mean Absolute Percentage Error (MAPE)**
24. **Accuracy**
25. **R Square**
26. **Cross Validation**
27. **Model Selection**
28. **References**
29. **Appendix**
30. ***Introduction***

***Problem Statement***

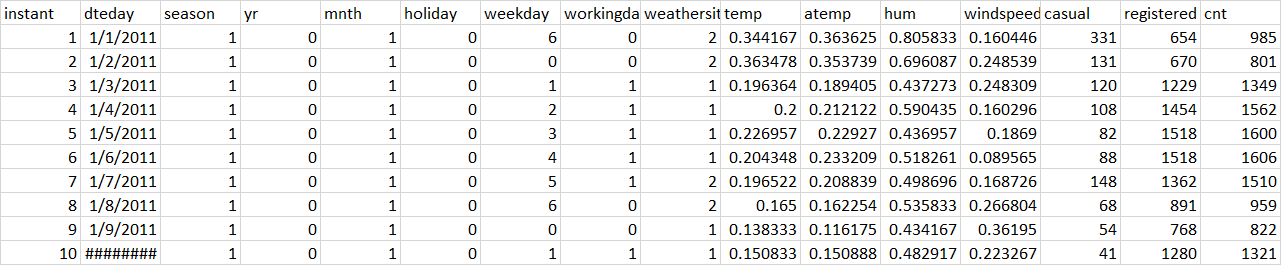
The project is about a bike rental company who has its historical data, and now our objective of this Project is to predict the bike rental count on daily basis, considering the environmental and seasonal settings. These predicted values will help the business to meet the demand on those particular days by maintain the amount of supply.

Nowadays there are number of bike renting companies like, Ola Bikes, Rapido etc. And these bike renting companies deliver services to lakhs of customers daily. Now it becomes really important to manage their data properly to come up with new business ideas to get best results. In this case we have to identify in which days there can be most demand, such that we have enough strategies met to deal with such demand.

***DATA***

The given dataset contains 16 variables and 731 observations. The “cnt” is the target variable and remaining all other variables are the independent variables.

Our objective is to develop a model that can determine the count for future test cases. And this model can be developed by the help of given data. A snapshot of the data is mentioned following.



1. ***METHODOLOGY***

After going through the dataset in detail and pre-understanding the data the next step is, Methodology that will help achieve our goal.

In Methodology following processes are followed:

**Pre-processing:** It includes missing value analysis, outlier analysis, feature selection and feature scaling.

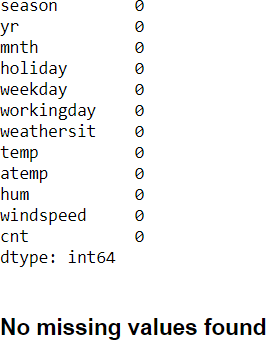
**Model development:** It includes identifying suitable Machine learning Algorithms and applying those algorithms in our given dataset.

***Pre-processing***

Here, we will use techniques like missing value analysis, outlier analysis, feature selection, feature scaling. These techniques are used to structure our data. Basically, pre-processing is done because and the model asks for structured data and preprocessing is used to structure the data we have got. As, normally the data we get can be messy i.e.: it can include many missing values, inconsistent values etc. And this things needs to be checked prior developing a model.

***Missing Value Analysis***

Missing value is availability of incomplete observations in the dataset. This is found because of reasons like, incomplete submission, wrong input, manual error etc. These Missing values affect the accuracy of model. So, it becomes important to check missing values in our given data.

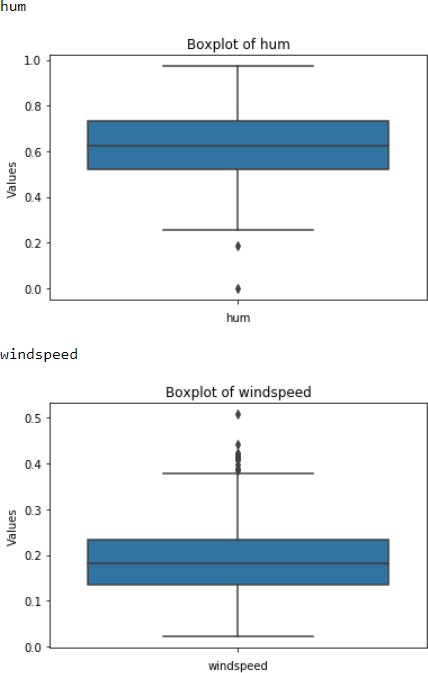
Here, in this project, after checking the data it is found that the data doesn’t consist any missing values.

As there are no missing values found in our given data, thus we don’t need to follow imputation processes here. So, we can directly move to our next step that is outlier analysis.

***Outlier Analysis***

Outlier is an abnormal observation that stands or deviates away from other observations. These happen because of manual error, poor quality of data and it is correct but exceptional data. But, it can cause an error in predicting the target variables. So we have to check for outliers in our data set and also remove or replace the outliers wherever required.

In this project, outliers are found in only two variables this are Humidity and wind speed, following are the box plots for both the variables and dots outside the quartile ranges are outliers.

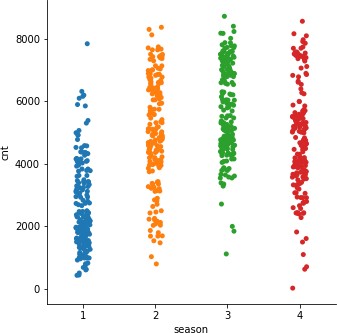


All this outliers mentioned above happened because of manual error, or interchange of data, or may be correct data but exceptional. But all these outliers can hamper our data model. So there is a requirement to eliminate or replace such outliers, and impute with proper methods to get better accuracy of the model. In this project, I used median method to impute the outliers in wind speed and humidity variables.

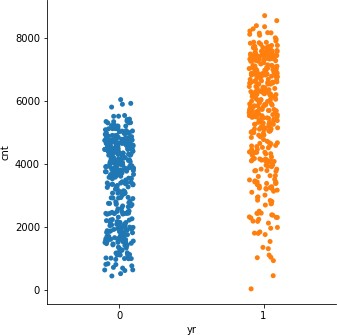
***Data Understanding***

Data Understand is a process where we know our data in a better way by the help of visual representations and come up with initial ideas to develop our model. Here, the specific variables are plotted with respect to the target variable. In some cases two variables are compared, whereas in some cases three variables are plotted together for our better understanding and visualization.

1. **Season**

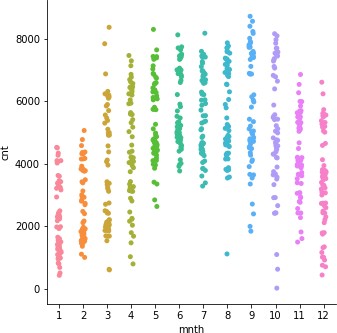


Here, it is found that in Season 2, 3 and 4 has the highest count

1. **Year**

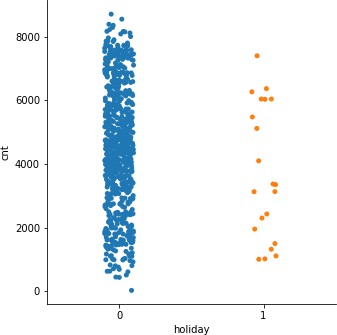
Here, it is found that in Year 1 has high count than 0

1. **Month**



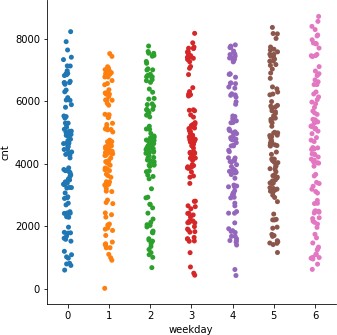
Here, it is observed that in Months 3 to 10 we got a good number of counts

1. **Holidays and Non-Holidays**



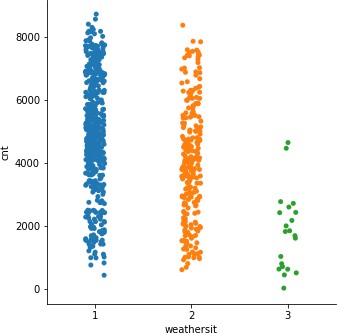
Here, it is found that, on holidays the count is higher when compared non-holidays

1. **Weekdays**



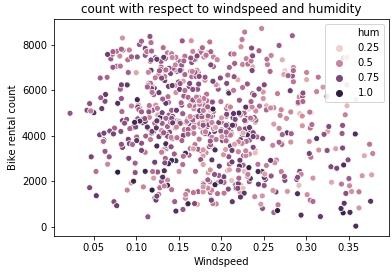
Here, it is observed that in weekdays, 0 and 6 i.e. Monday to Saturday the count is highest.

1. **Weather**



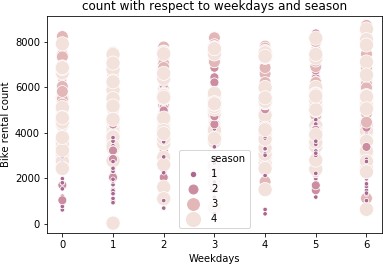
Here, in weather it is observed that, weather 1 has the highest count

1. **Wind speed and Humidity vs. count**



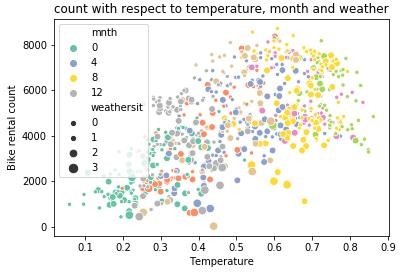
Here, it is found that in count vs. wind speed and humidity, Count is High in ranges of wind speed 0.10 to 0.25 and humidity 0.5 to 0.75

1. **Weekdays and Season vs. count**



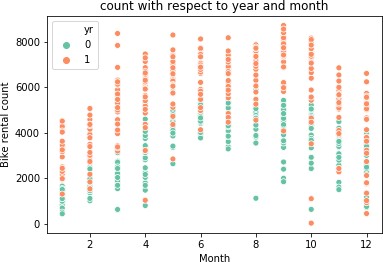
Here, it is observed that in count vs. weekdays and season, Count is high in 4th season and 1st and 6th of weekdays

1. **Temperature, month and weathers vs. count**



Here, it is found that in count vs. temperature, month and weather, Count is high in range temperature 0.5 to 0.8, in 8th month and weather is 0.

1. **Year and month vs. count**



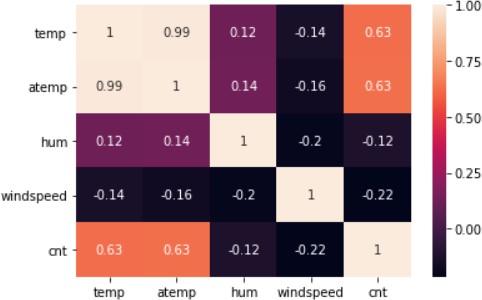
Here, it is found that count vs. respect to year and month, count is high in year 1, particularly from season 3 to 12 excluding 9th.

***Feature Selection***

Sometimes it happens that, all the variables in our data may not be accurate enough to predict the target variable, in such cases we need to analyze our data, understand our data and select the dataset variables that can be most useful for our model. In such cases we follow feature selection. Feature selection helps by reducing time for computation of model and also reduces the complexity of the model.

Here, in this project correlation analysis is done with numerical variables and ANOVA test is done with categorical variables to check if there is collinearity among the variables. And if there is any collinearity it’s better to drop such variables, else this redundant variables can hamper the accuracy of the model.

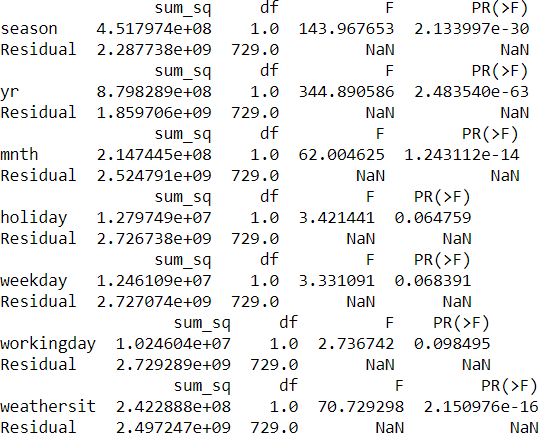
1. **Correlation Analysis for Numerical Variables.**



Plot: Correlation Analysis

Observing here, it is found that temperature and atemp are highly correlated with each other. So, in further processes we can drop atemp as it is similar to temperature.

1. **ANOVA Test for Categorical Variables**



Plot: ANOVA Test

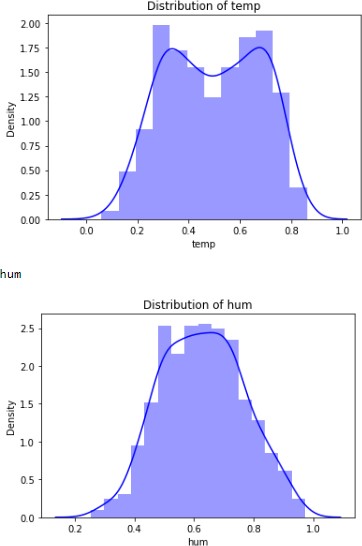
From the observations, it is found that the variables holiday, weekday, and working day has p value > 0.05. Here, null hypothesis is accepted. I.e. this variable has no dependency over target variable. So, in further processes these variables can be dropped before modeling. And this process of deducting the variables is also called as dimension reduction.

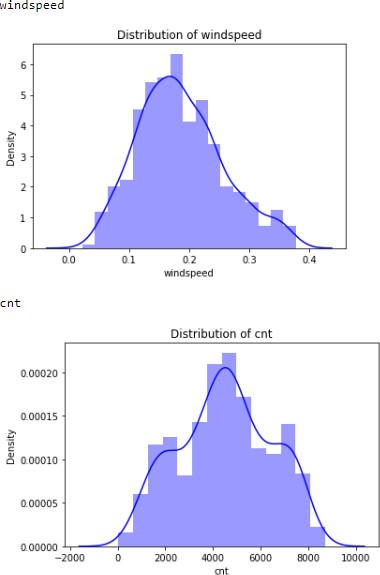
***Feature Scaling***

Here, In Feature Scaling ranges of variables are normalized or standardized, such that variables can be compared with same range. This is done for an unbiased and accurate model.

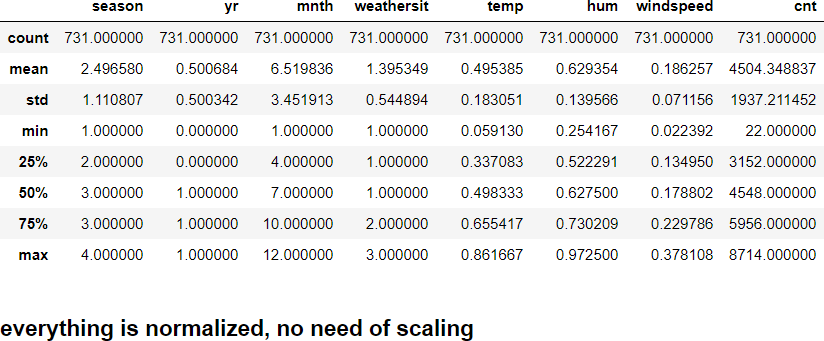
In this project, as the data are found as approximately symmetric. The feature scaling is not required. Following are the plots of approximately symmetric data visuals.

1. **Categorical Variables Distribution plot**





1. **For Numerical Variables Range check**



1. ***Model Development***

The next step after Exploratory Data Analysis and Data Pre-Processing is Model Development. Now we have our data ready to be implemented to develop a model. There are number of models and Machine learning algorithms that are used to develop model, some are like decision tree, random forest, SVM, KNN, Naïve Bayes, Linear regression, Logistic Regression etc. So, before implementing any model we have to choose precisely our model. So, the first step in Model Development is selection of model.

***Model Selection***

As per industry standards, there are four categories of models that are derived by classifying problem statement and goal of the project. These categories are:

1. Forecasting
2. Classification
3. Optimization
4. Unsupervised Learning

The process of selecting precise model depends on our goal and the problem statement. In this project the problem statement is to predict the bike rental count on daily basis, considering the environmental and seasonal settings. Thus, the problem statement is an identified as regression problem and falls under the category of forecasting, where we have to forecast a numeric data or continuous variable for the target.

In this project Decision Tree, Random Forest and Linear Regression are models selected for Model Development.

***Decision Tree***

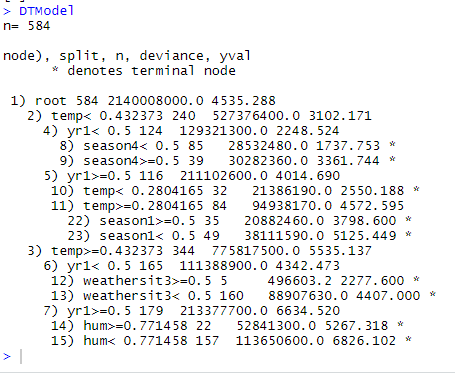
Decision Tree is a supervised learning predictive model that uses a set of binary rules to calculate the target value/dependent variable. Decision trees are divided into three main parts this are:

* Root Node : performs the first split
* Terminal Nodes : that predict the outcome, these are also called leaf nodes
* Branches : arrows connecting nodes, showing the flow from root to other leaves.

In this project Decision tree is applied in both R and Python, details are described following:

##### Decision Tree in R

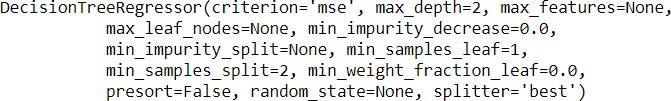
The Decision tree Method is used R with the structured data found after Data Preprocessing



The above plot shows the rules of splitting of trees. The main root splits into 2 nodes having temp < 0.432373 240 and temp >=0.432373 344 as its conditions. Nodes further split, The line with \* shows that it is the terminal node. These rules are then applied on the test data to predict values, And the MAPE, RSQUARE and Accuracy is noted below.

1. MAPE = 26.4225
2. RSQUARE = 0.7612102
3. ACCURACY = 73.51 %

##### Decision Tree in Python



The above fit plot shows the criterion that is used in developing the decision tree in Python. To develop the model in python, during modeling I have kept all the attributes at default, except the depth as 2. Although these attributes can be played around to derive better score of the model, which is called Hyper tuning of the model. After this the fit is used to predict in test data and the error rate, R-Square and accuracy is calculated.

1. MAPE = 36.948
2. RSQUARE = 0.6544
3. ACCURACY = 63.05 %

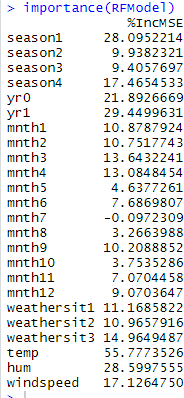
***Random Forest***

The next model to be followed in this project is Random forest. It is a process where the machine follows an ensemble learning method for classification and regression that operates by developing a number of decision trees at training time and giving output as the class that is the mode of the classes of all the individual decision trees.

In this project Random Forest is applied in both R and Python, details are described following.

##### Random Forest in R

In a RandomForest model the importance contributed by individual variables can be seen using importance function, it is mentioned below.

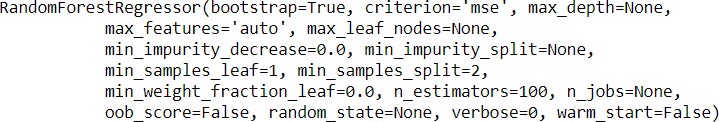


The above RF Model describes about the variable contributing most for predicting the target Variable. Few instances are like Temperature, humidity, season and year contributes most developing the model.

After the trained fit is used to predict the test data and error rate, accuracy and R-Square is noted.

1. MAPE = 19.32104
2. RSQUARE = 0.8685008
3. ACCURACY = 80.67 %

##### Random Forest in Python



Like the Decision tree above are all the criteria values that are used to develop the Random Forest model in python. Everything is kept default only except n\_estimators, which is tree numbers. Although these attributes can be altered to get a model with a better score. After this the error rate, R Square and accuracy of the model is noted.

1. MAPE = 20.4007
2. RSQUARE = 0.885114
3. ACCURACY = 79.05%

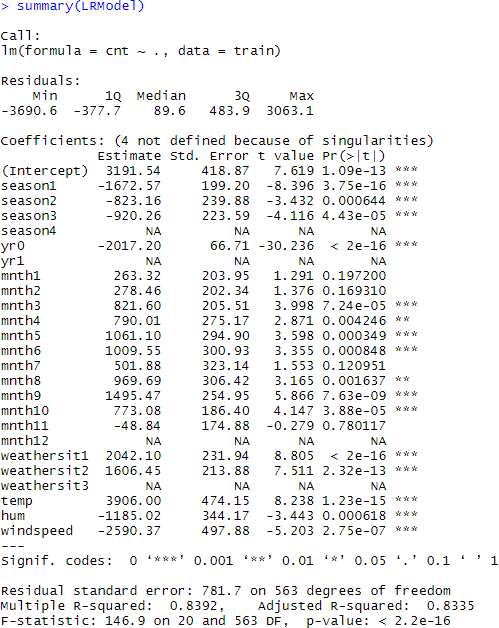
***Linear Regression***

The next method in the process is linear regression. It is used to predict the value of variable Y based on one or more input predictor variables X. The goal of this method is to establish a linear relationship between the predictor variables and the response variable. Such as, we can use this formula to estimate the value of the response Y, when only the predictors (X- Values) are known.

In this project Linear Regression is applied in both R and Python, details are described following.

##### Linear regression in R

After running the model the details I got are as follows

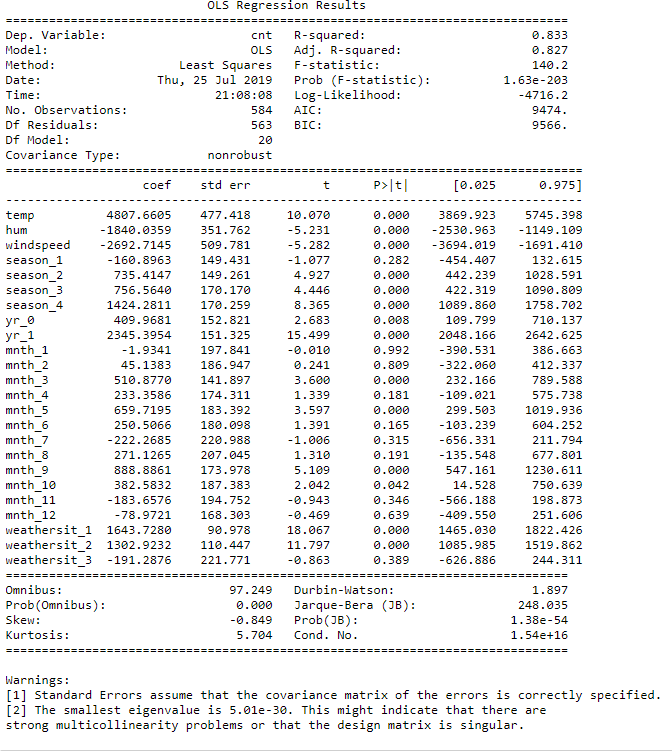


The above plot shows how the target variable count varies with change in each individual variable. The P- Value shows which values are significant in predicting the target variable. Here, we reject null hypothesis which is less than 0.05 and declare that the variable is significant for the model. F-Statistic explains about the quality of the model, and describes the relationship among predictor and target variables. The R squared and adjusted R squared values shows how much variance of the output variable is explained by the independent or input variables. Here the adjusted r square value is 83.35%, which indicated that 83% of the variance of count is explained by the input variables. This explains the model well enough. After this the error metrics and Accuracy is noted.

1. MAPE = 21.56792
2. RSQUARE = 0.8191175
3. ACCURACY = 78.44 %

##### Linear Regression in Python

After this the model is developed following details are found.



Here, F-Statistic explains about the quality of the model. AIC is Akkaine information criterion, if we have multiple models with same accuracy then we need to refer this to choose the best model. The table three values containing Omnibus and JB test are mostly required for time variance analysis. Here, as we are not using any time values in our project we can ignore this table 3. T-statistic explain how much statistically significant the coefficient is. It is also used to calculate the P –Value. And if P-Value is less than 0.05 we reject null hypothesis and say that the variable is significant. Here, all the variables are less than 0.05 and are significant. The R squared and adjusted R squared values show how much variance of the output variable is explained by the independent or input variables. Here the adjusted r square value is 82.7%, which explains that only 83% of the variance of count is explained by the input variables. This shows that the model is performing well. After this predictions are done and error metrics are calculated.

1. MAPE = 18.80069603
2. RSQUARE = 0.84360400
3. ACCURACY = 81.19 %
4. ***Model Summary***

From the above mentioned various models that can be developed for the given data. At first place, The Data is divided into train and test. Then the models are developed on the train data. After that the model is fit into it to test data to predict the target variable. After predicting the target variable in test data, the actual and predicted values of target variable are comparing to get the error and accuracy. And looking over the error and accuracy rates, the best model for the data is identified and it is kept for future usage.

1. ***Model evaluation***

So, now we have developed few models for predicting the target variable, now the next step is evaluate the models and identify which one to choose for deployment. To decide these, error metrics are used. In this project MAPE, R Square and Accuracy are used. And addition to this error metrics K-Fold Cross validation is also applied to identify the best model of all.

***Mean Absolute Error (MAE)***

MAE or Mean Absolute Error, it is one of the error measures that is used to calculate the predictive performance of the model. It is the sum of calculated errors. In this project we will apply this measure to our models.

* + 1. **In R :**

|  |  |
| --- | --- |
| Method | Mape Error( in Percentage) |
| Decision Tree | 26.4225 |
| Random Forest | 19.32104 |
| Linear Regression | 21.56792 |

* + 1. **In Python :**

|  |  |
| --- | --- |
| Method | Mape Error( in Percentage) |
| Decision Tree | 36.9480 |
| Random Forest | 20.9466 |
| Linear Regression | 18.8006 |

If we observe the above tables, we choose the model with lowest MAPE as a suitable Model. Here, from R we get Random Forest as a better model, whereas from Python we get Linear Regression as a better model. So following this we can conclude that Both Random Forest and Linear Regression can be used as model for this data, if you evaluate on the basis of MAPE. But we need more error metrics to cross check this. So, we go for R Square which is a better error metric.

***Accuracy***

The second matric to identify or compare for better model is Accuracy. It is the ratio of number of correct predictions to the total number of predictions made.

##### Accuracy= number of correct predictions / Total predictions made

It can also be calculated from MAE as Accuracy = 1- MAPE

1. In R

|  |  |
| --- | --- |
| Method | Accuracy (in Percentage) |
| Decision Tree | 73.57 |
| Random Forest | 80.67 |
| Linear Regression | 78.43 |

1. In Python

|  |  |
| --- | --- |
| Method | Accuracy (in Percentage) |
| Decision Tree | 63.051 |
| Random Forest | 79.053 |
| Linear Regression | 81.199 |

As, Accuracy derives from MAE/MAPE its observations also suggest same models as better models as suggested by MAPE. Here, the models with highest accuracy are chosen, and from the observations it is found that both Random Forest and Linear Regression are good models for the given data set.

***R Square***

R Square is another metric that helps us to know about the Correlation between original and predicted values.

1. In R

|  |  |
| --- | --- |
| Method | R – Square (in Percentage) |
| Decision Tree | 76.12 |
| Random Forest | 86.85 |
| Linear Regression | 81.91 |

1. In Python

|  |  |
| --- | --- |
| Method | R – Square (in Percentage) |
| Decision Tree | 65.44 |
| Random Forest | 88.43 |
| Linear Regression | 84.36 |

R Square is identified as a better error metric to evaluate models. If we observe the above tables, we choose the model with highest R Square as a suitable Model. Here, from both R and Python it is found that Random Forest is a best fit model for the given data.

***Cross Validation***

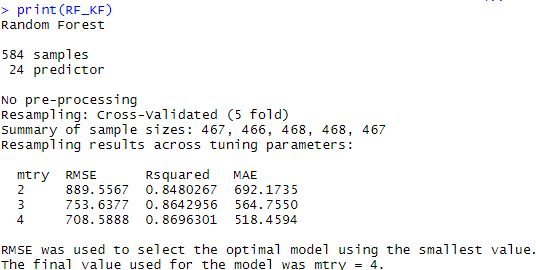
Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. Although we have followed above error metrics to identify a better model, there is always a chance that model is under fitting or over fitting the data. So, the problem with this evaluation technique is that it does not give an indication of how well the learner will generalize to an independent/ unseen data set. Getting this idea about our model is known as Cross Validation. So, it becomes important to cross validate our model in most cases. Cross – Validation are of different types. In this project K-Fold cross validation is used.

##### K-Fold Cross – Validation:

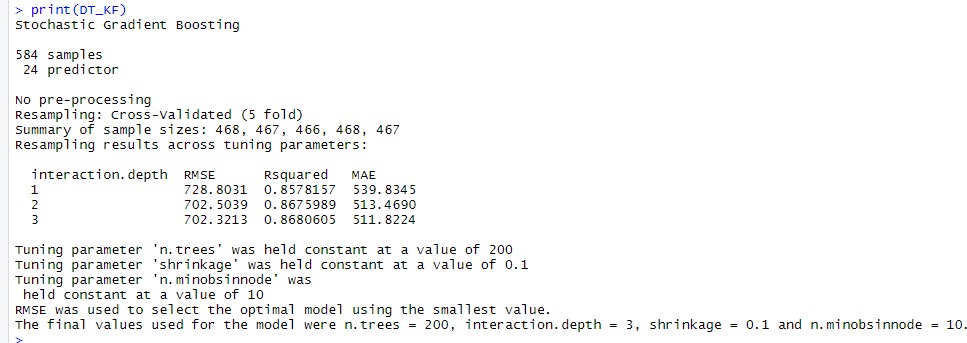
The procedure has a single parameter called k, that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Basically it distributes the data in various folds and averages the accuracy score of various folds to identify the best model. The model with highest cross validated average score of accuracy is termed as best model for the data.

##### In R: By the help of caret package in R the cross-validation is done for various model and results are plotted.

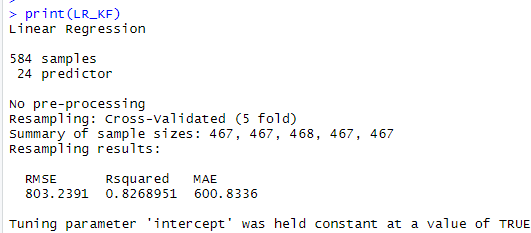
##### Random Forest: 5 folds are created and little hypertuning is done with mtry = 2,3,4 and the following observations are found, it says RF Model with 4 split is good with R-Square of 86.9 %



##### Decision Tree: 5 folds are created and little hyper tuning of interaction depth = 1,2,3 , and n.trees = 200, and the following observations are found, it says DT Model with interaction depth with 3 and 200 n.trees the model performs better as R-Square is 86.8 %

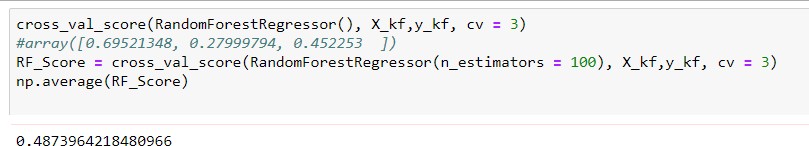


##### Linear Regression: 5 folds are created and the following observations are found for Linear Regression Cross Validation, it says LR Model performs well with as R-Square is 82.6 %

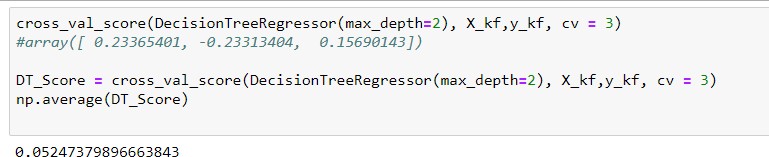


##### In Python: Here in python the cross\_val\_score function is imported from scikit learn library, which performs K Fold Cross Validation in various models. The details are noted below.

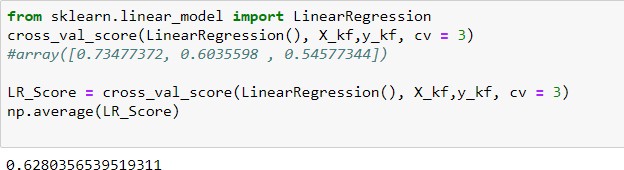
##### Random Forest: 3 Folds are created with n\_estimators = 100, and 3 folds scores are found and the average accuracy score of the model is found as 48.73 %. Thus, the model is not upto mark it can be tuned further, and if tuning also doesn’t improve the accuracy of the model, we will drop this model.



##### Decision Tree: 3 Folds are created with mac\_depth = 2, and 3 folds scores are found and the average accuracy score of the model is found as 5.24 %. Thus, the model is not upto mark it can be tuned further, and if tuning also doesn’t improve the accuracy of the model, we will drop this model.



##### Linear Regression: 3 Folds are created with no tuning, and 3 folds scores are found and the average accuracy score of the model is found as 62.80 %. Thus, the model is upto mark. it can also be tuned further to get better accuracy.



From the above cross-validation it is found that, in some cases Random Forest is a better model and in some other cases Linear Regression is a better model for the given data set. We can go with any one of them or both. Thus, these models can be used for further processes and this model can also be further tuned to get optimum results.

And also from all the criteria mentioned above, like MAPE, R Square, Accuracy and Cross- Validation, It is concluded that both the models Linear Regression and Random Forest are better for our given data set.

1. ***References***
2. For Data Cleaning and Model Development - <https://edwisor.com/career-data-scientist>
3. For other code related - Analyticsvidya.com
4. For Visualization & Model deployment – [Medium.com](https://towardsdatascience.com/)
5. Model evaluation – TDS & TAI