Bike Rental Prediction Project

Submitted by – *Kunal Choudhary*

Problem Statement:

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

Dataset:

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

instant: Record index

dteday: Date

season: Season (1: springer, 2: summer, 3: fall, 4: winter)

yr: Year (0: 2011, 1:2012)

mnth: Month (1 to 12)

hr: Hour (0 to 23)

holiday: weather day is holiday or not (extracted from Holiday Schedule)

weekday: Day of the week

workingday: If day is neither weekend nor holiday is 1, otherwise is 0.

weathersit: (extracted from Freemeteo)

1: Clear, Few clouds, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp: Normalized temperature in Celsius. The values are derived via

(t - t_min)/ (t_max -t_min), t_min=-8, t_max=+39 (only in hourly scale)

atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_maxt_min), t_min=-16, t_max=+50 (only in hourly scale)

hum: Normalized humidity. The values are divided to 100 (max)

windspeed: Normalized wind speed. The values are divided to 67 (max)

casual: count of casual users registered: count of registered users

cnt: count of total rental bikes including both casual and registered

Pre-Processing:

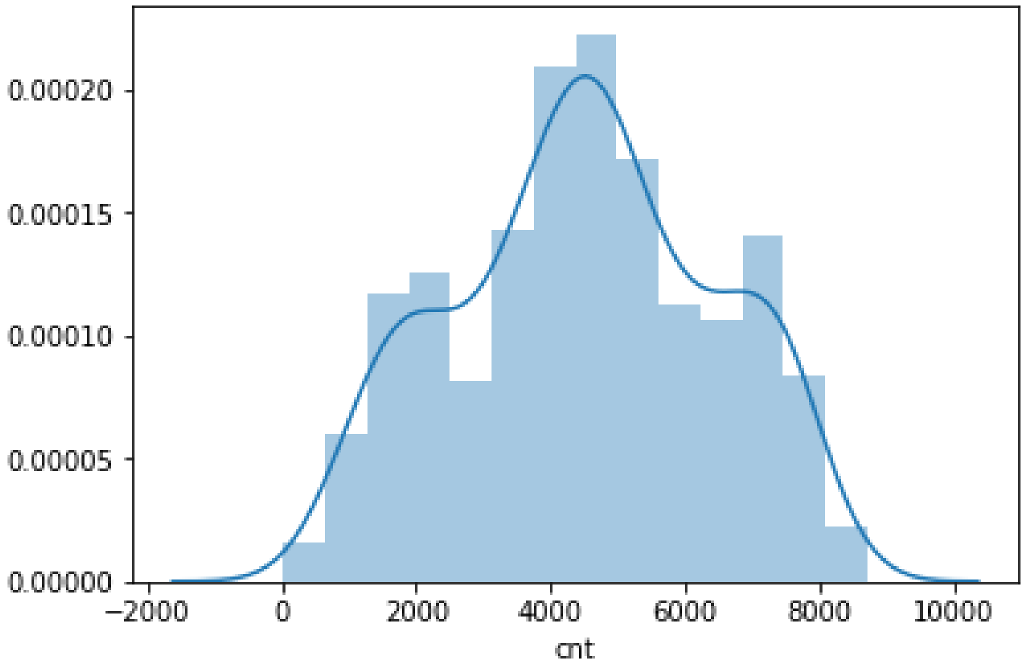
Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than just looking.

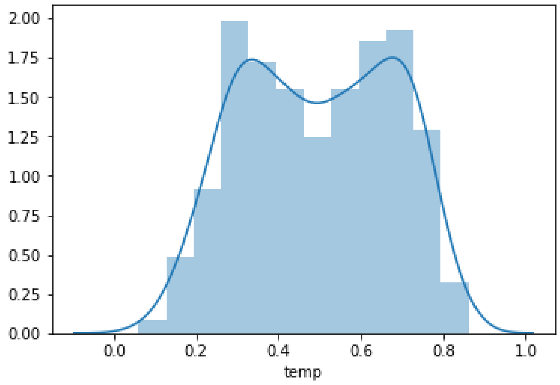
Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis (EDA). To start this process, we will first try and look at all the distributions of the numeric variables. Most analysis like regression, require the data to be normally distributed.

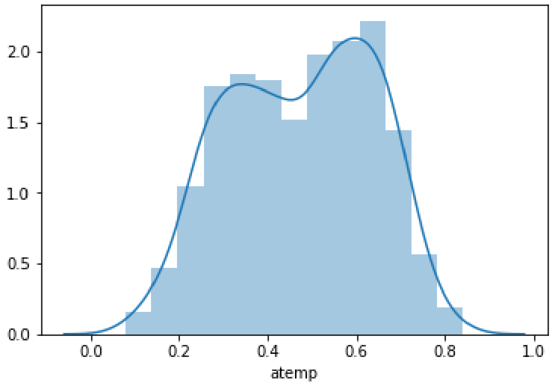
Univariate Analysis

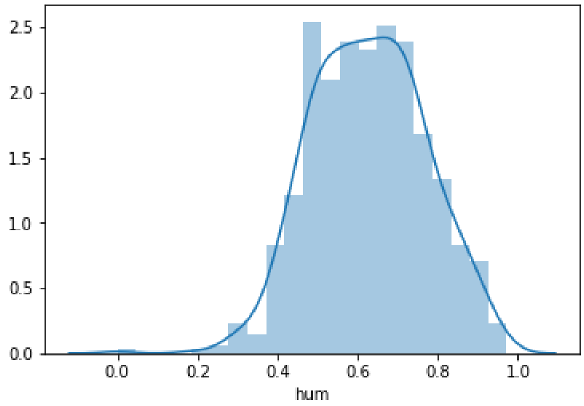
In Figure seen below we have plotted the probability density functions for numeric variables present in the dataset including target variable cnt.

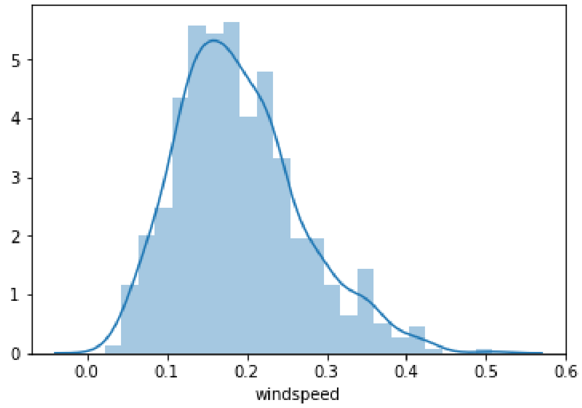
- Target variable *cnt* is normally distributed.
- Independent variables like *'temp','atemp'* are normally distributed.
- Independent variable 'casual' data is slightly skewed to the right so, there is chances of getting outliers.
- Other Independent variable *'hum'* data is slightly skewed to the left, here data is already in normalize form, so outliers are discarded.

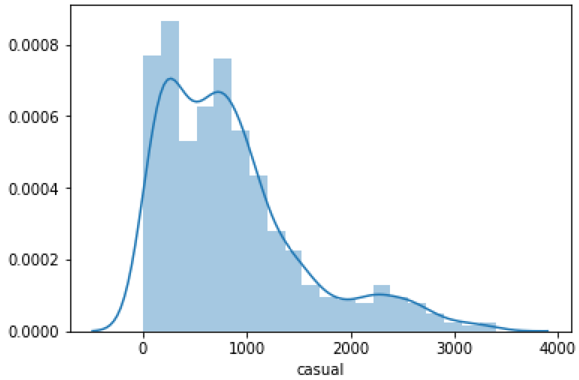


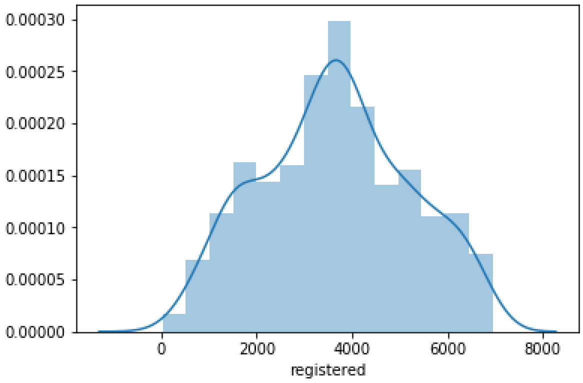








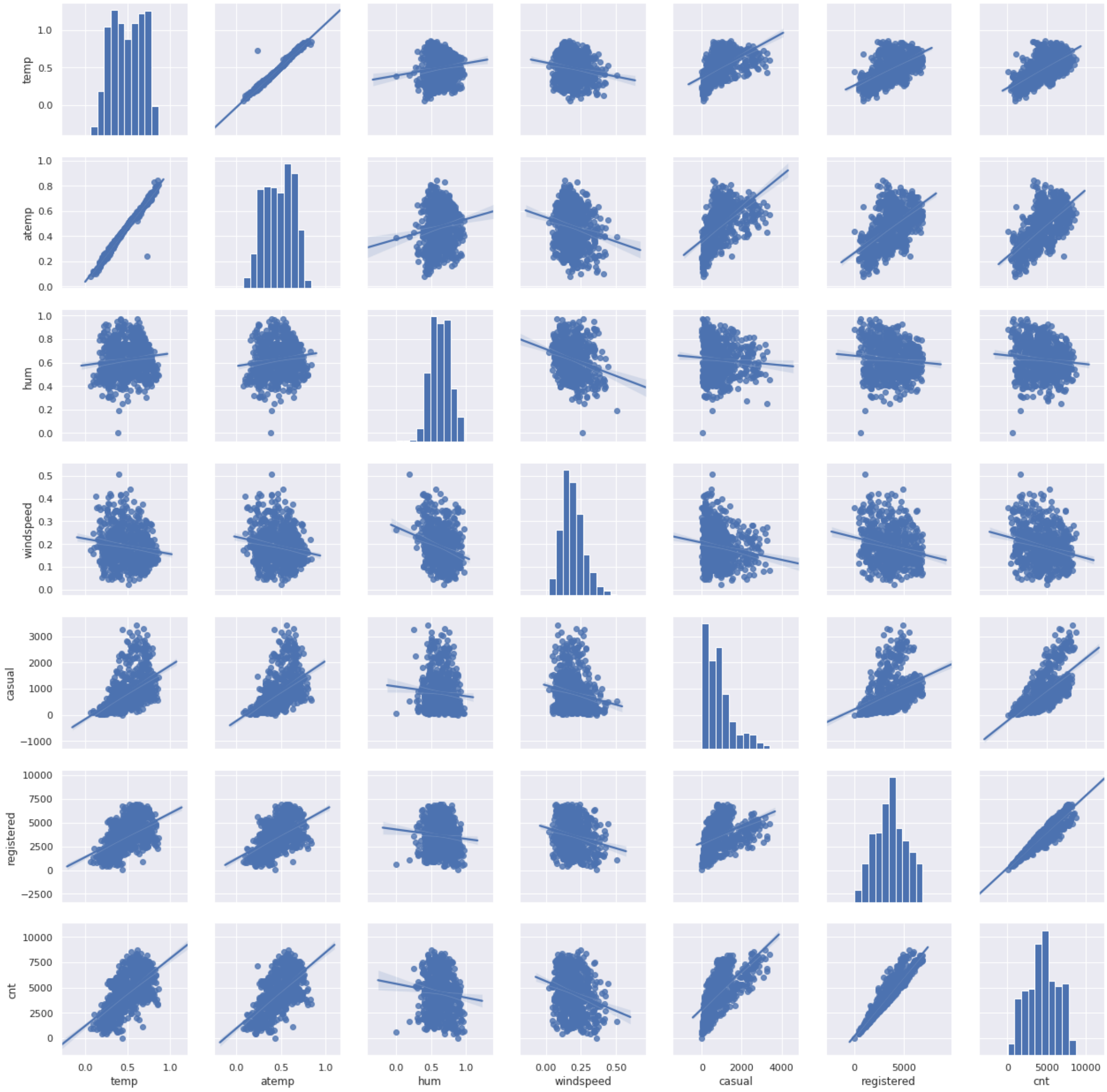




Bivariate Analysis

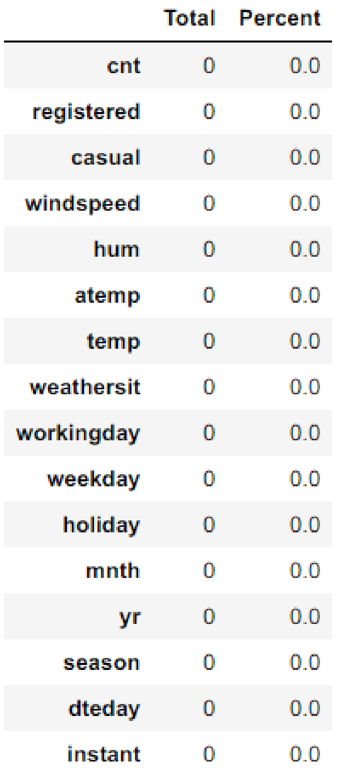
Pairplot is built in function in Seaborn python library, Seaborn provides templates for combining plots into a matrix through the pairplot function. Here a matrix of multiple plots can be useful for quickly exploring the relationships between multiple features in a data frame. Below figures shows relationship among various independent variables and with numeric target variable ‘cnt’ using pairplot.

- The *pairplot* graph below shows that relationship between independent variables ‘temp’ and ‘atemp’ is very strong.
- There is an inverse relationship between ‘hum’ as well as ‘windspeed’ with the target variable ‘cnt’.



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.



	Total	Percent
cnt	0	0.0
registered	0	0.0
casual	0	0.0
windspeed	0	0.0
hum	0	0.0
atemp	0	0.0
temp	0	0.0
weathersit	0	0.0
workingday	0	0.0
weekday	0	0.0
holiday	0	0.0
mnth	0	0.0
yr	0	0.0
season	0	0.0
dteday	0	0.0
instant	0	0.0

Outlier Analysis

The other steps of Pre-processing technique are 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 the data is skewed so, there is chance of outlier in independent variable 'casual', one of the best methods to detect outliers is Boxplot.

Fig below shows presence of Outliers in variable 'casual' and relationship between 'casual' and 'cnt' before removing outliers.

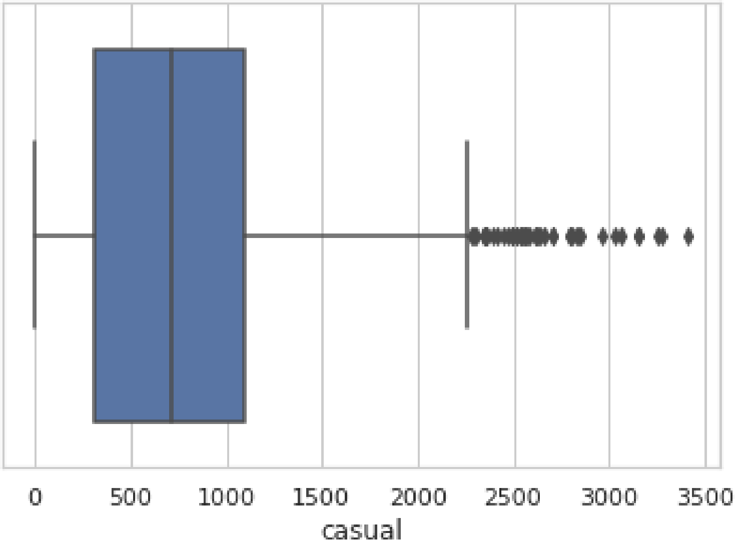
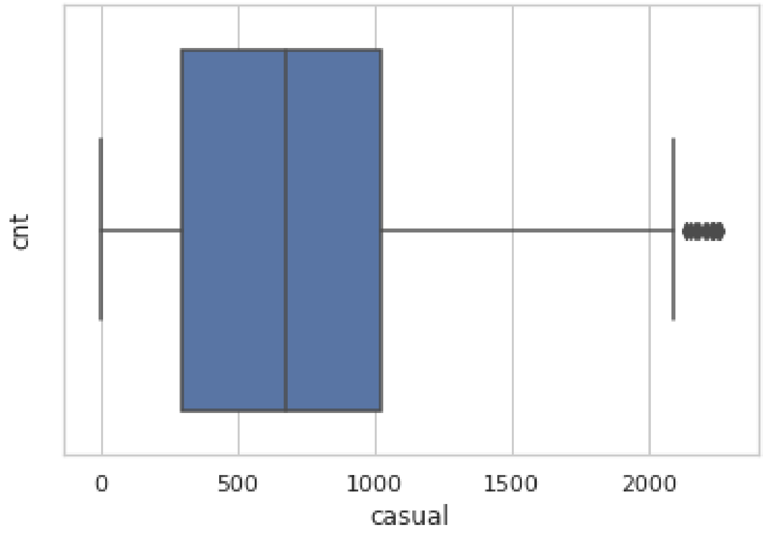


Fig below shows boxplot of ‘casual’ after removing outliers and relationship between ‘casual’ and ‘cnt’ after removing outliers.



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.

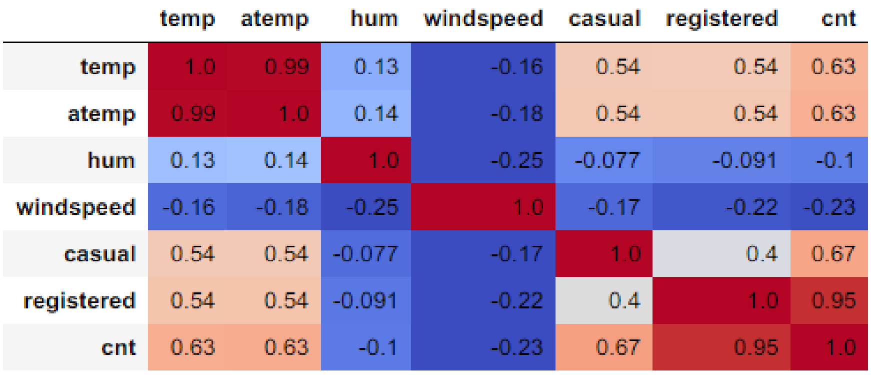
Features Selection

Machine learning works on a simple rule – GIGO (Garbage In Garbage Out). By garbage here, we mean noise in the dataset. This becomes even more important when the number of features are very large. You need not use every feature at our disposal for creating an algorithm; So, we can assist our algorithm by feeding in it, only those features that are important.

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

- The relationship between two independent variables should be less and
- The relationship between Independent and Target variables should be high.

Below fig illustrates that relationship between all numeric variables using *corr* method.



	temp	atemp	hum	windspeed	casual	registered	cnt
temp	1.0	0.99	0.13	-0.16	0.54	0.54	0.63
atemp	0.99	1.0	0.14	-0.18	0.54	0.54	0.63
hum	0.13	0.14	1.0	-0.25	-0.077	-0.091	-0.1
windspeed	-0.16	-0.18	-0.25	1.0	-0.17	-0.22	-0.23
casual	0.54	0.54	-0.077	-0.17	1.0	0.4	0.67
registered	0.54	0.54	-0.091	-0.22	0.4	1.0	0.95
cnt	0.63	0.63	-0.1	-0.23	0.67	0.95	1.0

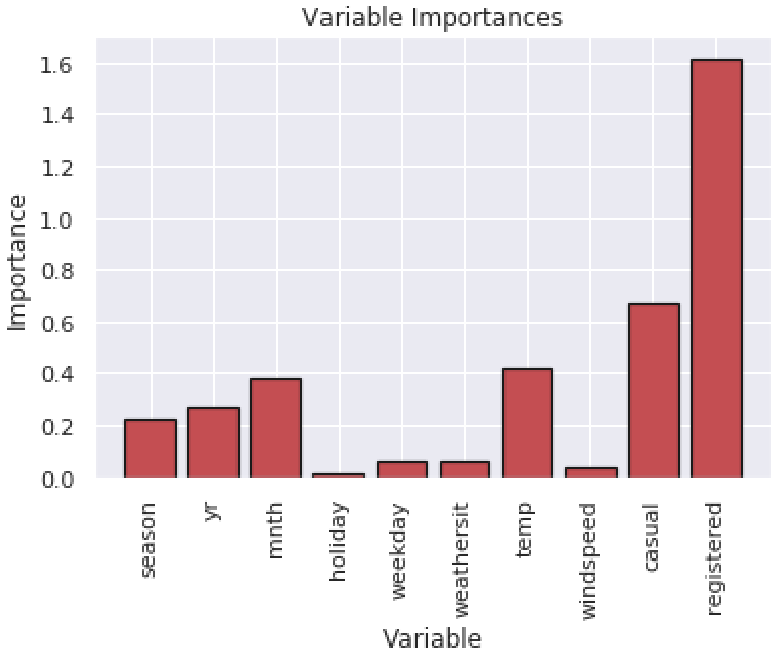
Dimensionality Reduction for numeric variables

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

Dimensional Reduction using Random Forest Variable Importance

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



The above figure shows that variables ‘season’ ‘windspeed’, ‘weekday ‘, ‘weathersit’ and ‘holiday’ are of less importance in predicting the ‘cnt’ of Rented bikes. So, these variables are to be removed while creating RandomForest Model.

Feature Scaling

In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to easily compare data coming from different sources. Some of the more common ways to normalize data include: Transforming data using a *z-score* or *t-score*, which is usually called standardization. Rescaling data to have values between 0 and 1 is usually called feature scaling. 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 normalization ‘casual’ and ‘registered’ variables look like in table below where all values between 0 and 1.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In rental dataset numeric variables like ‘temp’, ‘atem’,’hum’ and ‘windspeed’ are in normalization form so, we must mormalize two variables ‘casual’ and ‘registered’. After normalization ‘casual’ and ‘registered’ variables look like in table below where all values between 0 and 1.

Predictive Modelling

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.

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'.

Evaluation of Decision Tree Model

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 must tune the model using Random Forest.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random forest functions in below way

i. Draws a bootstrap sample from training data.

ii. For each sample grow a decision tree and at each node of the tree

   a. Randomly draws a subset of *mtry* variable and p total of features that are available

   b. Picks the best variable and best split from the subset of *mtry* variable

   c. 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.

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

nodesize: Size of each node is 10

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

Evaluation of Random Forest

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.

Linear Regression

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

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. 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 actual values f(x) are very less so this model is predicted almost accurate values. Multiple R-Squared value is 1 so, we can explain about 100 % of the data using our multiple linear regression model. This is very impressive.

Evaluation of Linear regression Model

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

Conclusion

As we predicted counts for Bike Rental using three Models. Among Decision Tree, Random Forest and Linear Regression it was noticed that the MAPE is higher and RMSE is lower for the Linear regression model, so we can conclude that for the Bike Rental Data Linear Regression model is best for the task of count prediction.