**Bike Rental Count Prediction**

Report By

**Kavin S**

[akkera2@gmail.com](mailto:akkera2@gmail.com)

9944573276

**Problem Statement**

Need to Predict the bike rental count on daily based on the environmental and seasonal settings.

**Software Requirements:**

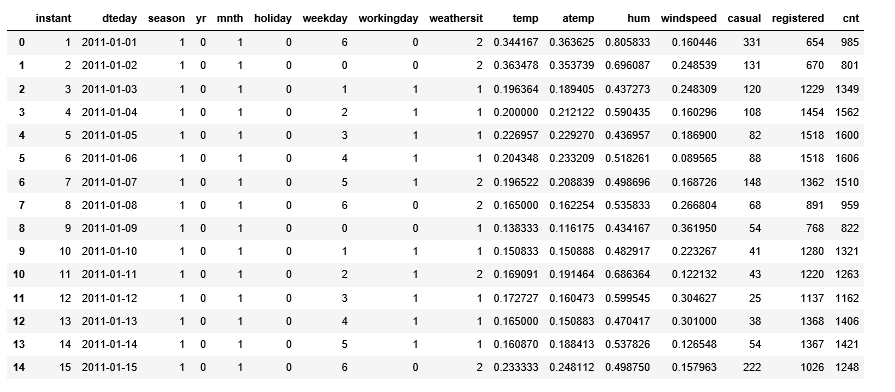
1) conda version 4.8.3

2) Python Version 3.6.10 with required Libraries

3) R version 3.6.3 with required Libraries

**Data:**

Our task is to build regression models which will predict the count of bike rented depending on various environmental and seasonal conditions. Please find below data that we are going to use to predict the count of bike rent.



**Steps involved to create Machine learning Model:**

**Data Exploration:**

It is the process of analysing thestructure of our data set it is also called as Exploratory data analysis and it is most important part Machine learning because based on the type of data in the dataset we decide the methods to analysis the data and appropriate model used to predict the output.

Steps to check the data;

1)Identify the predictor and target variable

2) Data type of variables and category of variables (Continuous or Categorical).

3)Based on the variables category we use different methods to identity relation between variables.

Checking on the training dataset provided, the dataset has target variable that need to be predicted so this project comes under supervised machine learning model. And the target variable has numerical value, we can use regression algorithm model to predict the target data.

**Data Types:**



**Missing Value Analysis:**

Missing values in data set are the common scenario which affects the conclusion we get from data. Missing values should be properly handles based on the % and type of the data. We need to decide whether the missing values to be removed or imputed based on % of missing data by variable

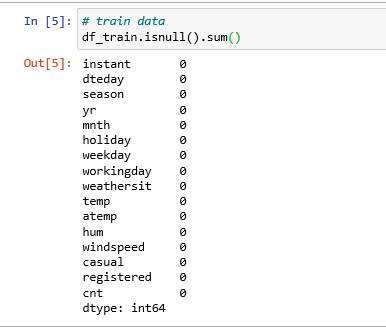
i) If a variable has less than 30%, Missing values can be easily treated using various methods like mean, median method, KNN method to impute missing value.

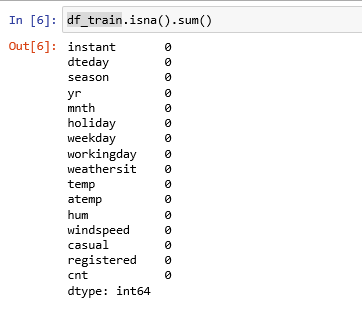
To check suitable impute method (Mean, Mode or KNN) method, first delete the any value from particular index and apply all three methods, check which methods gives nearest value to that deleted value and use the method which generated nearest value

ii)If a variable has more than 30% of its values missing, then those variable columns can be ignored. In our case, none of the columns have a high percentage of missing values.

Both methods have its own pros and cons based on the data set we handle.

Here our dataset doesn’t have any missing values in training dataset.so we have not used any method to solve the same.





**Outlier Analysis:**

Outliers are the data points that were stand out from rest of the data distribution in the data set. The easiest way to detect outliers is to create a graph. Plots such as Box plots, Scatterplots and Histograms can help to detect outliers. In this project I have used boxplot to visualize the outliers.

Box plot is convenient method that displays the distribution of data through their quartile by means of

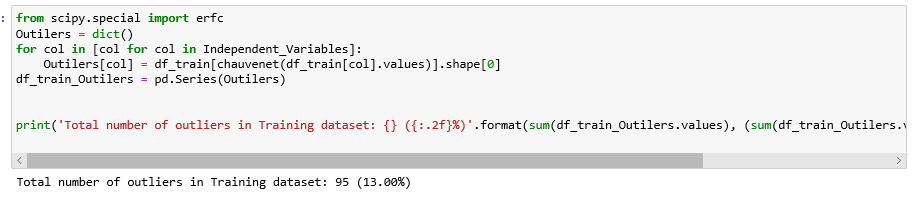
1. minimum value
2. Lower quartile
3. median
4. Higher quartile
5. Highest value

**Advantages of choosing Box plot:**

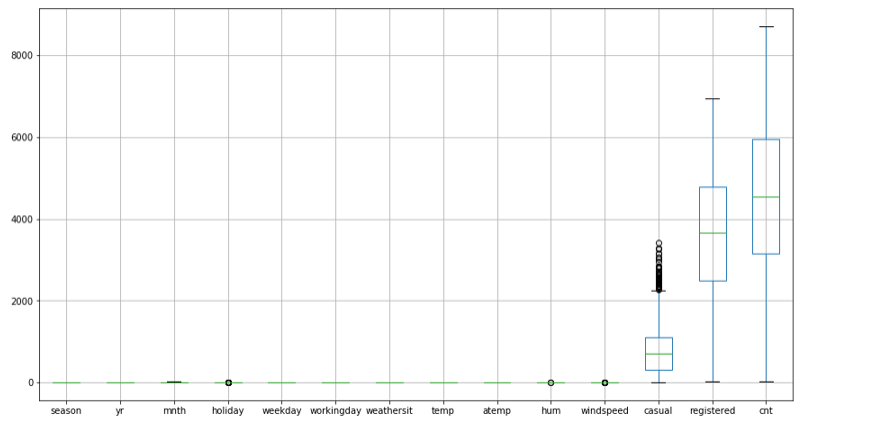
1. It will handle the large dataset by organizing data using above five key concepts which is not possible by other graphs.
2. It shows the simple summary of the distribution of results in graphical manner so that you can quickly view it and compare it with other data.
3. Most important thing, It will identify and display the outliers, which can be easy to determine and to remove those values.

In this project we have outliers in training dataset

**Outliers % in Training Dataset**:



**Box plot created for Training Data set.**



**Feature Selection:**

Feature selection is the process of reducing the independent variable which carries same information about the dependent variables. If we failed to remove those duplicate data, then it will affect the performance of our model.

Reducing the no of input variables will improve the computation cost and model performance.

The relationship between variable are considered as below.

1. Relation between two independent variables should be low
2. Relation between independent and dependent variables should be low

As our dependent variables are numerical, we can Correlation coefficient method to identify the relation between dependent variables.

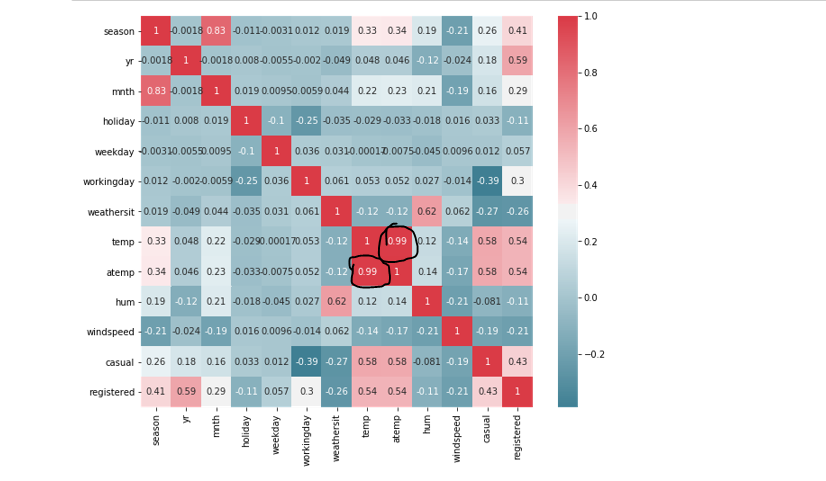
**Positive Correlation**: Means that if feature A increases then feature B also increases or if feature A decreases then feature B also decreases. Both features have a linear relationship.

**Negative Correlation**: Means that if feature A increases then feature B decreases and vice versa.**No Correlation**: No relationship between those two attributes.

If a Correlation coefficient value between two variables is greater than + -0.5 then we can remove any one variable that carries less information of target variable.

Here as far we checked the variables, we have high correlation between **temp and atemp** variables.so I decided to remove one variable among these variables based on the variable which carries less information about Target variable.

Please find the correlation heat map generated for training data set. Dark blue represents the high correlation. Here we can see the dark blue colour only for same variables, other than our variables didn’t have correlation.



**Information Gain:**

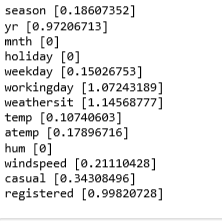
Mutual information is a measure between two (possibly multi-dimensional) random variables XX and YY, that defines the amount of information obtained about one random variable, through the other random variable.

it helps us to choose the right set of independent variables that carries the useful information about the target variables.so that we can neglect all the unwanted variable/noise from our dataset which will reduce the model performance.

It’s a kind of feature reduction process of reducing the number of variables from the dataset.

The Mutual Information value will not be negative and zero if the input and output variable are independent.

Comparing the information gain of **temp** and **atemp** variable, **temp** variable carries less information about target variable, so I will drop temp variable plus variable like **month, holiday, humidity** as it doesn’t carry any useful information about target variables.



**Regression model used:**

1. **Linear regression:**

Logistic regression is a Statistical algorithm that used when the dependent variable is numerical in nature. Independent variables can be numerical or categorical, but target variable should be numerical.

**Two Types:**

**Simple linear regression**-- if we have 1 dependent and 1 Independent variable then it is SLR

**Multiple Linear regression**-- if we have 1 dependent and more than 1 Independent variable then it is MLR

**Equation expressing this relation is the line**::

y=b0+b1x

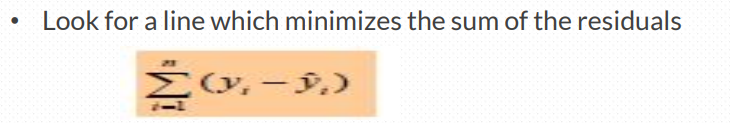
y---Predicted target variable

b0--intercept

b1--Coefficient of 1st independent variable x

if we have 2 independent variables then y=b0+b1x+b2x

**Error check**: Sum (Actual value -Predicted Value)



**Yi**= Actual value, **y~**= -Predicted Value

1. **Random Forest algorithm:**

Random forest is a supervised learning algorithm and it’s an ensemble technique that consists of many decision trees.

It will make the combination of weak learners to make the strong model. The main hypothesis is that when weak models are correctly combined, we can obtain more accurate and/or robust models. So I have tried out this algorithm expecting less error than linear regression.

**Advantage:**

It can be used for both classification and regression problem

It reduces misclassification error rate and improve the accuracy; we combine multiple decision tree to create a classifier, so it avoids overfitting of the model

**Disadvantage**:

As it will create large number of trees the real time prediction is too slow.

**3)Decision Tree Algorithm:**

A predictive model based on a branching series of Boolean tests. Can be used for both Regression and Classification. Each node in a tree represents attribute and Each Leaf Represents a class label

**Advantage:**

Easy to understand and interpret, perfect for visual representation.

Can work with numerical and categorical features.

Requires little data pre-processing: no need for one-hot encoding, dummy variables, and so on.

**Disadvantage**:

The main problem of the Decision Tree is it will lead to overfitting

Little bit of noise can make it unstable which leads to wrong predictions.

**Performance metrics for Regression model:**

These parameters are used to test the model developed on training Side,

In ML algorithm we freeze the model created on train based on test results, but in statistical model we will evaluate how the model calculated on trained data

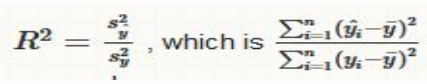
1. **R-Squared : should be >80%**

Gives us value of variance of dependent variables which explain all independent variables together.

R-Squared is the proportion of variance in the dependent variable which can be explained by the independent variables.

**Example :**If R-squared value is 0.86,then 86 % of variance target variable is explained by the all independent variables combinedly.

**Formula:**



**Yi~**= **Predicted value. ~ represents bar**

**Y~= Y mean of target/dependent variable**

**Yi = Actual value..** Output will varies from 0-1

1. **Adjusted R-square:**

If we keep on adding the independent variables to the model. some time it will be overfitting. That means output will be biased (one sided). if you add irrelevant variables to the model it will reduce the efficiency of strong predictors/independent variables

Adjustment of the R-squared that penalizes the addition of extraneous predictors to the model. i.e) it will reduce the unnecessary independent variables that doesn’t much information about Dependent variables.

**Formula :**

1 - ((1 - Rsq)((N - 1) /( N - k - 1))

where k is the number of predictors (Independent Variables). N- no of observations

Value is always less than or equal to adjusted R-Square

1. **MAPE (Mean Absolute Percentage Error) :**

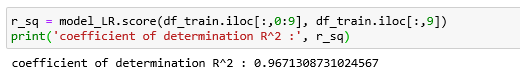
MAPE is the simplest form of regression error metric that calculate the residual for every data point, taking only the absolute value of each so that negative and positive residuals do not cancel out but with adjustments to convert everything into percentages.

mean(np.abs((y\_true - y\_pred) / y\_true))\*100

**Conclusion:**

**R square Value:**

I gotR square value 0.9671308731024567 ,which shows that 96 % of variance of dependent variables that I used in the algorithm explains about the Target variable. 96 % is good score.



**MAPE:**

Please find the mean absolute percentage error we got for three different algorithm I have tried in which Random forest got the lesser number of error percentage .so that we can go with **Random forest** model for prediction`

MAPE states that our model’s predictions are,

on average, 4.8 % off from actual value for Decision Tree Model

on average, 3.8 % off from actual value for Random Forest Model

on average, 7.2 % off from actual value for Linear Regression Model

