BIKE RENTING in R

CHANDRA MOHAN SRIVASTAVA

21/11/2019

**INDEX**

**1 Introduction   
 1.1 Problem Statement  
 1.2 Data**

**2. Methodology**

**2.1 Pre-Processing.**

**2.2 Missing Value Analysis**

**2.3 Data Visualization and Exploratory Data Analysis**

**2.4 Feature Selection**

**3. Model Selection and Evaluation**

**3.1 Linear Regression**

**3.2 Decision Tree**

**3.3 Random Forest**

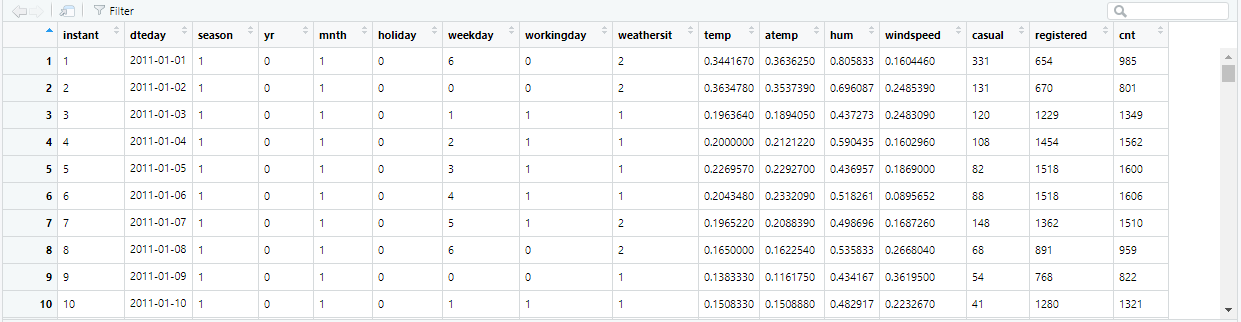
1. **Introduction** 
   1. **Problem Statement**

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

1.2 **Data**

The task is to build a regression model which will predict the number of bike rented on the basis of predictor variables. Following is a snippet of what the data  
looks like:



Description of the attributes :

instant – index which will be removed

dteday – date in yy/mm/dd format

yr- year : 0 = 2011 , 1=2012

mnth – month which ranges from 1 to 12 (January to December)

holiday – 0 = no holiday(including weekends) , 1 = holiday  
weekday – ranges from 0 to 6 where 0 represents Sunday and 6 Saturday and the numbers in between represents Monday to Friday in chronicle manner.  
workingday – 0 = not a working day and 1 = working day

weathersit – ranges from 1 to 4 .

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\_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

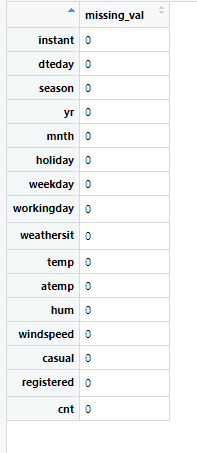
registered: count of registered users   
cnt: count of total rental bikes including both casual and registered

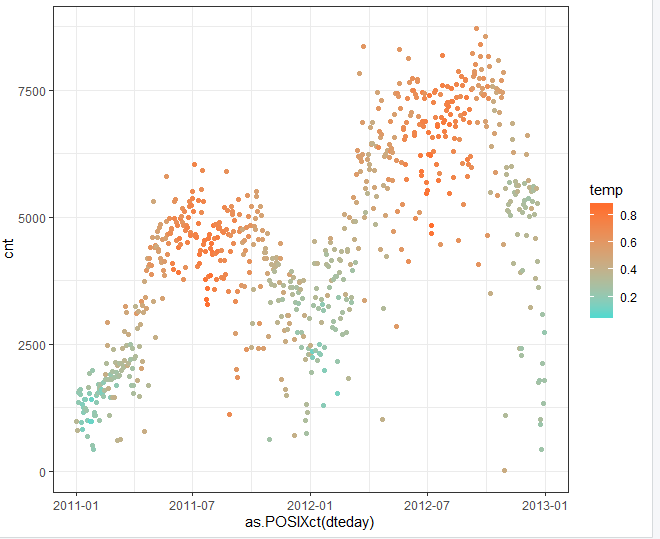
2 **Methodology**

**2.1 Preprocessing:** 1. It is imperative that before we dive into missing value analysis , first the   
 conversion of variables which are discrete needs to be converted from   
 numeric to factorial form. While doing so , we have converted season, yr,   
 holiday , weekday , workingday , weather sit from numeric to factor form.  
   
 2. Also, the variable instant has been deleted as it deems to be an extra column.

3. Eliminated casual and registered as these are directly contributing to cnt   
 or count – which we need to predict.

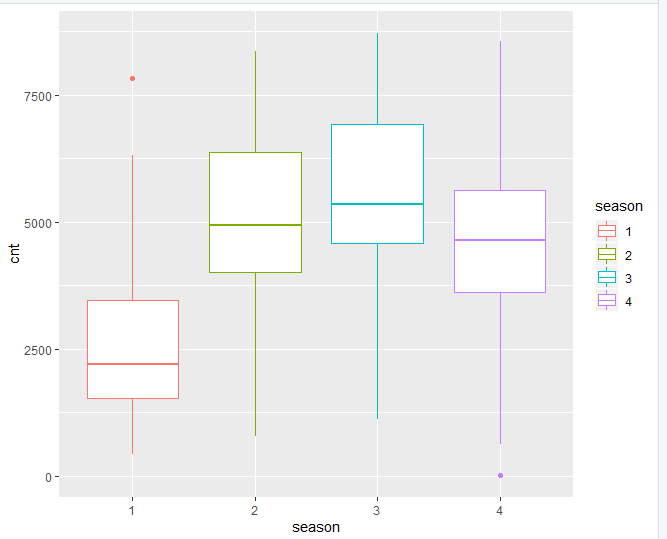
**2.2 Missing Value Analysis**:   
 It is important to get the hang of the data by checking the number of missing   
 values. In the data set presented , there is none as depicted below.

  
   
 **2.3 Data Visualization and Exploratory Data Analysis** One of the most intuitive way to get the feel of the data is by data visualization.  
 Given below are graphs depicting relationship between cnt i.e. bikes rented   
 and predictor variables.  
  
 cnt vs dteday(date) :



From the above graph , we can observe that the count increases with year in general.And is maximum steadily during the fall region and falls down during Winter and continues with the same trend only with increased count

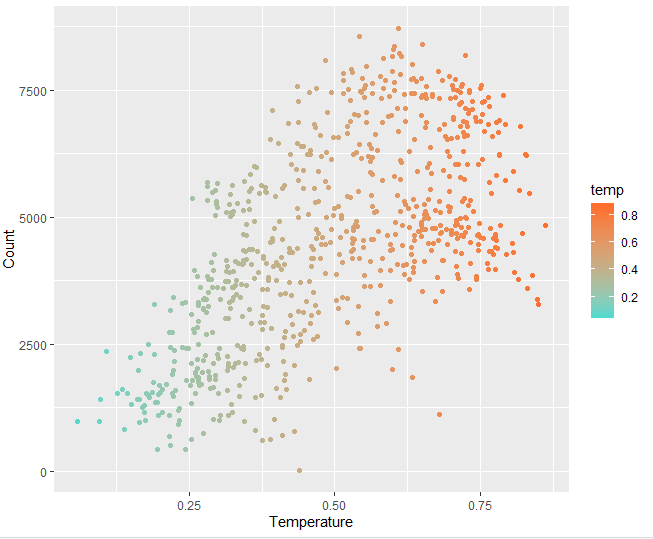
cnt vs season :



1-fall,2-spring,3-fall,4-winter

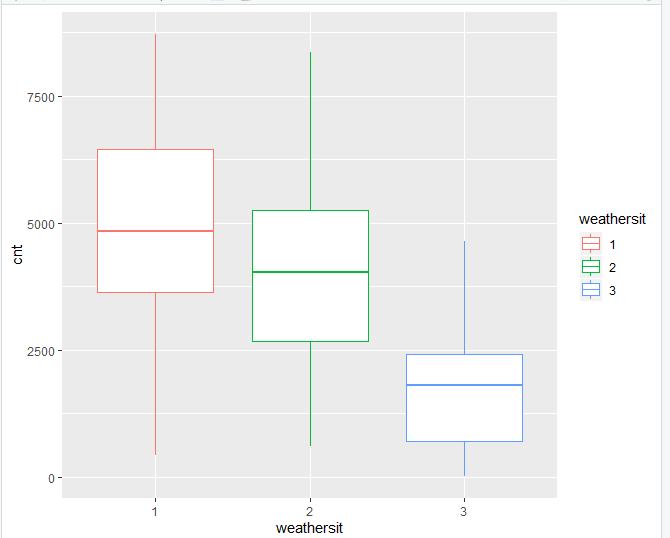
Observations: bike count is least during the spring , and peaks during the season of fall. Also , the other facet of the data-set is that average count or average cnt is steady throughout the season 2 , 3 and 4

Cnt vs temp :



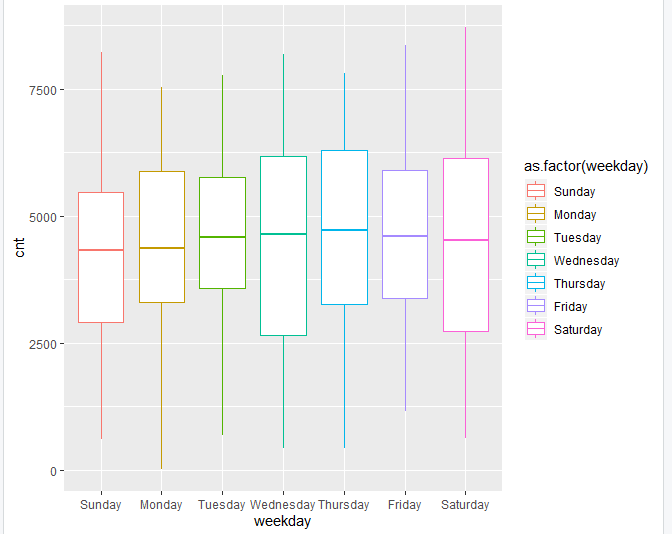
The graph shows that there is a linear relationship between temperature and bike count.The count increases steadily with temp till a certain temperature and then starts decreasing.

Cnt vs weathersit :



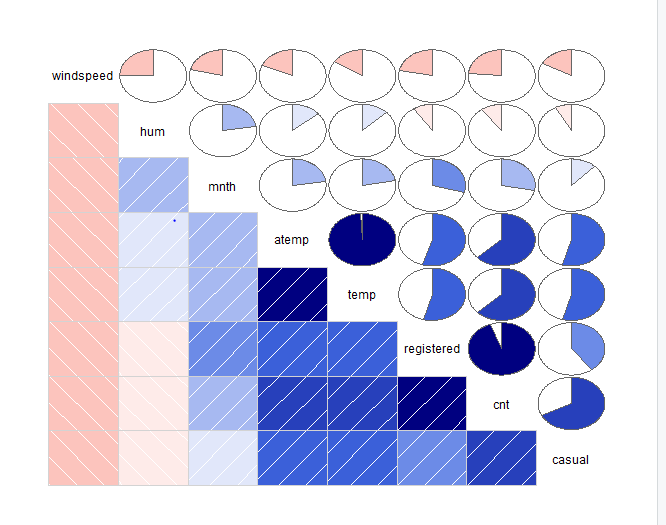
Bike count is maximum for weather 1 i.e. the most comfortable temperature and decreases as weathersit or weather becomes more harsh as seen with bikecount decreasing with increasing harshness of the weather conditions which makes sense. No record for weather no 4.

cnt vs weekday:



The maximum no bikecount belongs to the Fridays and Saturdays - makes sense as it's on the weekend.But at the sametime , we can also see that the average count is uniform throughout the days which implies that maybe the relation between workingday and count isn't too strong.

The categorical variables weekdays are converted later to numeric varables i.e. 1 – represents  
Sunday and 7 represents Saturday.  
  
2.4 **Feature Selection**:   
 Model selection is done by building a co-relation matrix. The idea of feature selection   
 is to detect multi-collinearity and get rid of the similar variables to remove redundancy  
 and improve its performance.

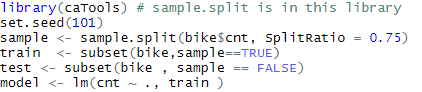
 atemp will be removed as it is highly co-related with temp. And so will be casual   
 and registered.

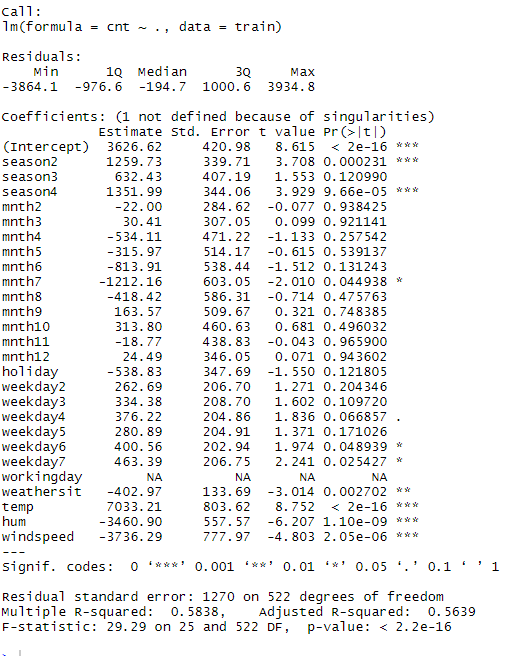
3. **Model Selection and Evaluation:**

**3.1 Model Selection** In this case , as we have to predict a continuous variable , the prediction will be   
 performed with the following models:

1. Linear Regression.  
 2. Decision Tree Regression.  
 3. Random Forest Regression

3.1.1 Linear Regression:  
 Following line of code is written to perform linear regression in R.



Below is the summary of the model 

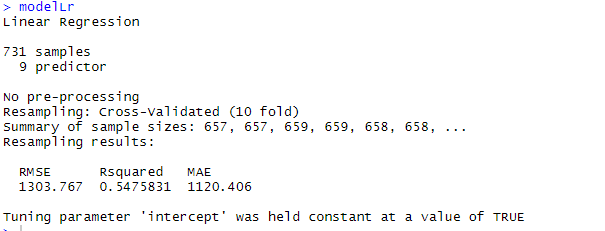
From the above summary , we can see that the cnt is heavily dependent   
 on environment condition i.e. temp , season , windspeed , hum and weather.

There’s some reliance on weekday 6 and weekday 7 i.e Friday and Saturday   
 respectively – which indicates that the bike count goes up during the weekends-  
 which is verified by the boxplot above(cnt vs weekday).

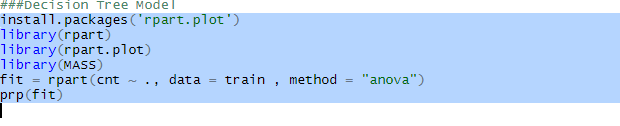
Once the model has been built , it is important to perform cross-validation technique  
 to get concrete result. Following line of code is used in R :



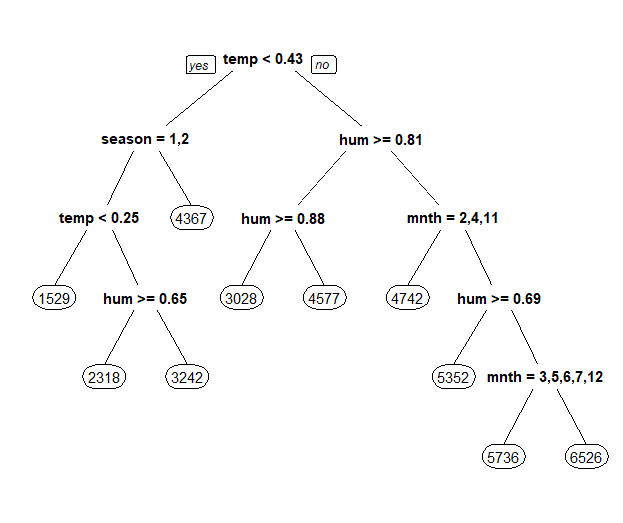
Summary of the model is given below :



3.2.2 **Decision Tree** Following line of code is used to perform Decision Tree in R :



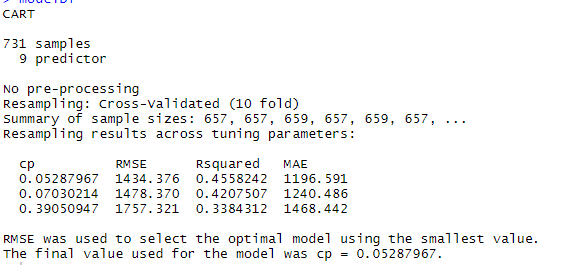
prp(fit) is used to create a tree.



From the above graph , we can again re-affirm that environmental condition and   
seasonality plays a pivotal role in deciding the bike count.

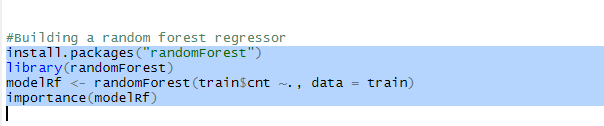
Again cross validation technique is performed to get a consolidated result.

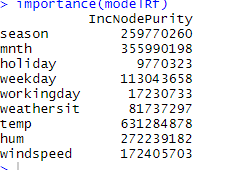




**3.3.3 Random Forest Regression**

Following line of code is used to generate a model with Random Forest

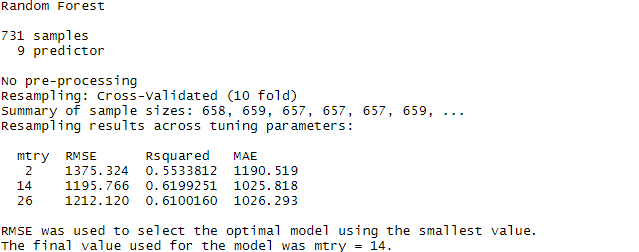
importance() function



From the above list, we can see that the most important predictor variable is temp.

Now , onto cross validation technique :





As the RMSE , MAE value is the lowest for Random Forest Regressor , it is evident that  
the most accurate prediction will be done by Random Forest Model.