R Notebook

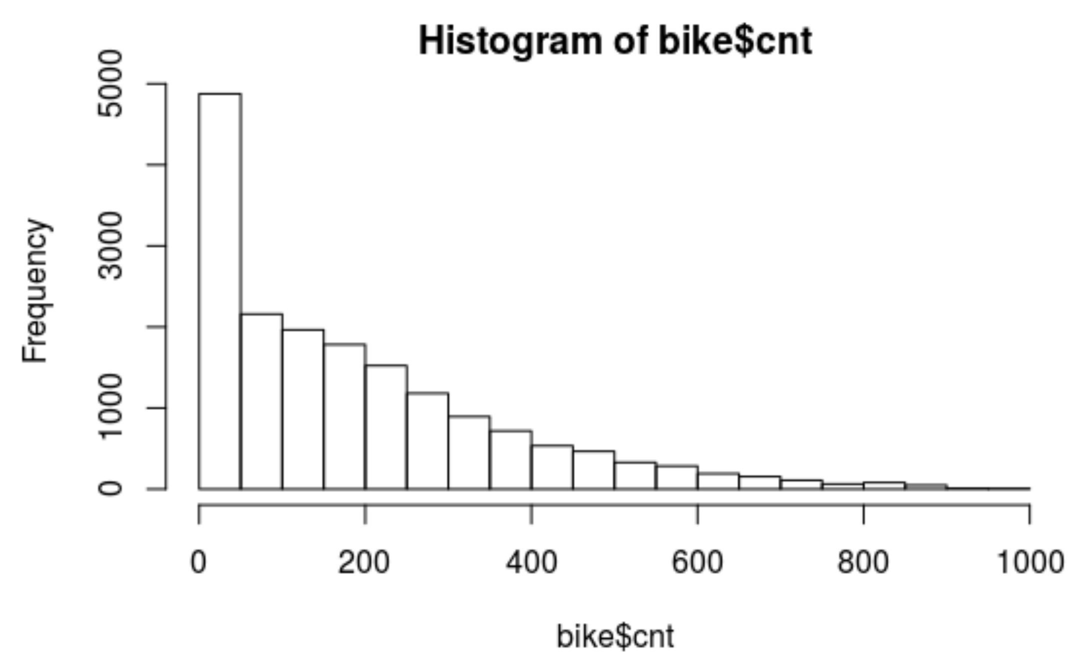
```
set.seed(101)
library(rpart)
library(rpart.plot)
library(rattle)
library(caret)
library(car)
library(randomForest)
```

```
bike <- read.csv("hour.csv", header = TRUE)
```

```
bike$instant<- NULL
bike$dteday <- NULL
bike$casual <-NULL
bike$registered<-NULL
bike$atemp<-NULL
```

remove instant, dteday, casual , registereded, and atemp to avoid multilinearity

```
hist(bike$cnt)
```



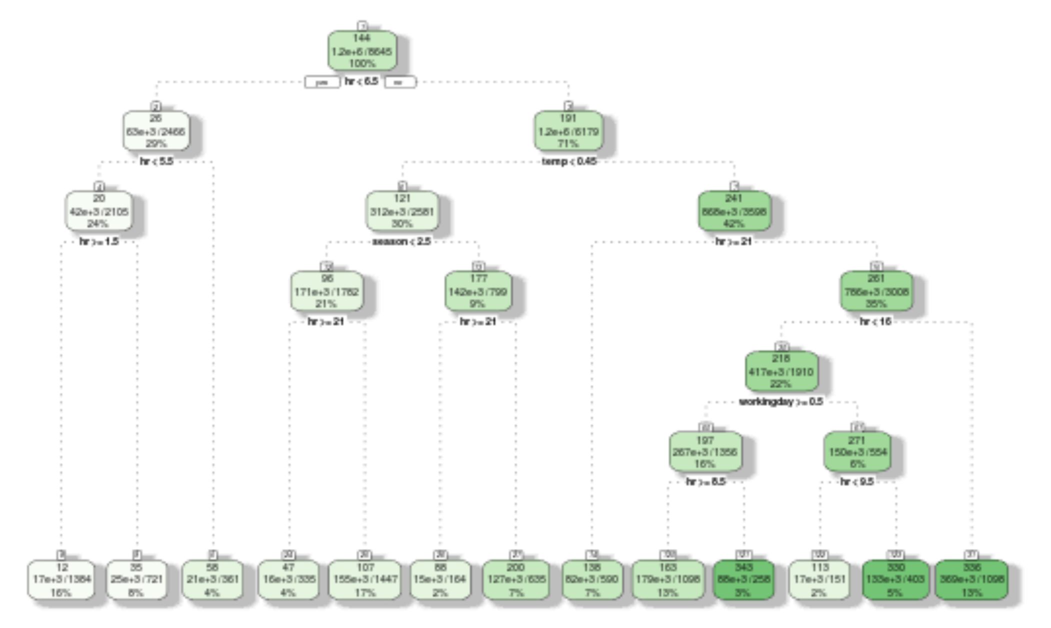
The counts of bike shared is not normal distributed. Linear Regression might not be the best model for such time series dataset.Let’s use the “hour.csv” to do Tree-based regression methods , polynomial regression or tree dicission regression. Idealy, for such time series counts prediction, a Poisson regression might be the best model.

We will still use “hour.csv” (bike_sub) to do the modelling.We use 2011 dataset as train data and 2012 dataset as test data.

```
bike2011 <- bike[1:8645,]
bike2012 <- bike[8646:17379,]
```

```
train=bike2011
test=bike2012
```

```
reg.tree <- rpart(cnt ~ ., method="poisson", data = train)
fancyRpartPlot(reg.tree, main="", sub="")
```



Poisson Regression Decision Tree model evaluation

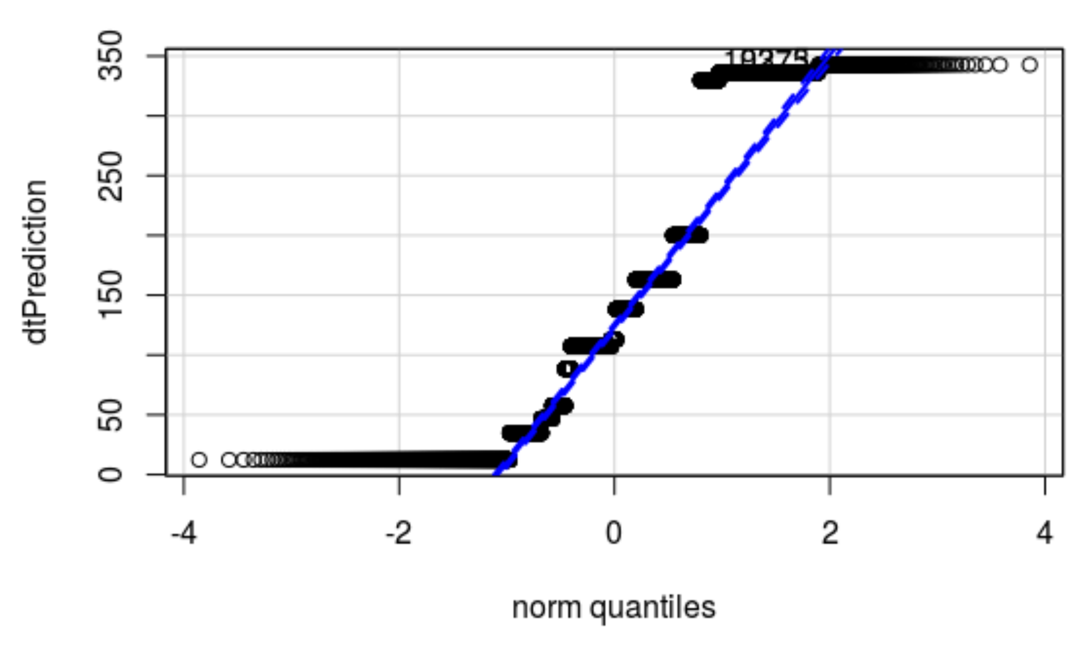
```
dtPrediction <- predict(reg.tree, test)
cor(dtPrediction,test$cnt)
```

```
[1] 0.8305135
```

As we can see, the accuracy of RT at 0.831, much higher than OLS Linear Regression.

```
qqPlot(dtPrediction, main="" )
```

```
 9275 10373
  630  1728
```



The plots show the presence of outliers and inaccuracy in the areas of low and high scores.

Let’s check if we use decision tree reg model to predict the bike shared on April 22 2012 base on the same weather condition.

```
predict(reg.tree, data.frame(season=2, yr=1,mnth=4, hr=8, holiday=0, weathersit = 3, temp=0.4 , hum =0.82, windsp
eed =0.2537, weekday=0,workingday=0))
```

```
       1
107.1923
```

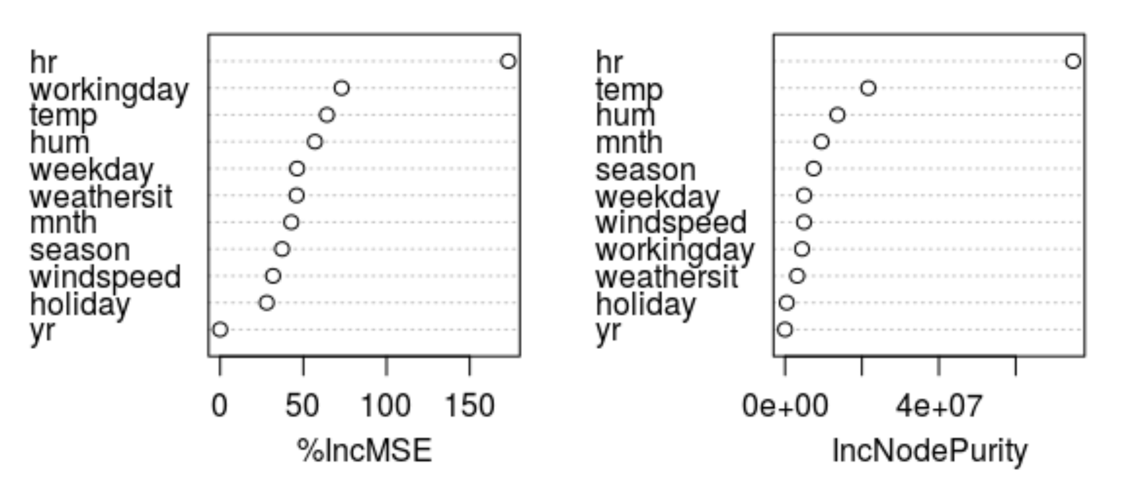
###The real counts of rent is 51, however on April 22 2012, the reg.tree predict 107. There was a historical snow in April.It was a kind of unsual weather condision in end of April. So that the Tree Model could not be accurately predict the right counts due to such special weather condition.It is understandable that the prediction count is much higher than the actual counts due to this unusual weather situation.

##Let’s see if Random Forest model works better.

Create Regression Forests, Random Forest Regressionmodel fit

```
fitRF <- randomForest(cnt ~ ., data=train, importance=TRUE, ntree=500)
```

```
varImpPlot(fitRF, main="")
```



Importance of the dataset attributes for the prediction of the “class” attributes shown in above figure.

Analyzing this chart, we conclude that the proper order of attributes importance for the prediction of the target attribute is: hr,workingday,temp,hum,weathersit,weekday,mnth,season,windspeed,holiday,yr. It contradicts the absolute values of the Linear Regression coeffitients assigned to independent attributes, i.e. temp at 0.35 seems to be more important than workingday at 0.10. This difference could be explained by limitation of OLS.

Random Forest Regression prediction and valuation

```
PredictionRF <- predict(fitRF, test)
cor(PredictionRF,test$cnt)
```

```
[1] 0.921053
```

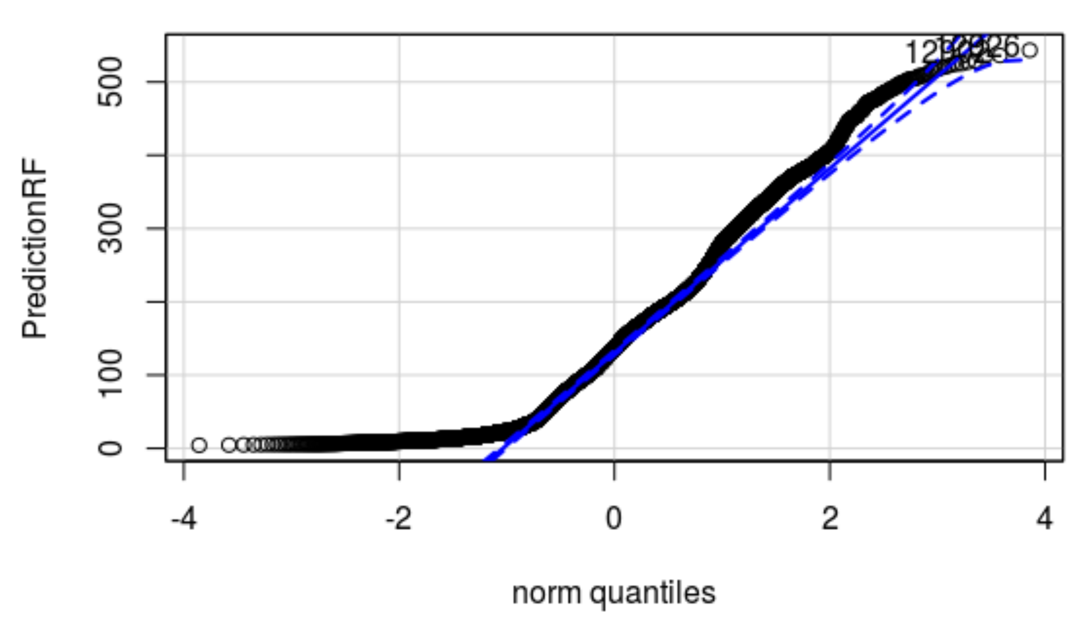
Code above applies fitted Random Forest Regressor model to the test data. The caculations show that RFR gives better accuracy at 0.9394 correlation to the actual counts.

Let’s visualize the results of the predictions, the code below generates a scatter plot of the Predictor vs Test values.

###Standalized Residuals visualization

```
qqPlot(PredictionRF, main="")
```

```
12926 12902
 4281  4257
```



The plots show the presence of outliers and inaccuracy in the areas of high scores.

Let’s check if we use Random Forest Reg model to predict the bike shared on April 22 2012 base on the same weather condition.

```
set.seed(101)
predict(fitRF, data.frame(season=2,yr=1,mnth=4, hr=8, holiday=0, weathersit = 3, temp=0.4 , hum =0.82, windspeed
=0.2537, weekday=0,workingday=0))
```

###The real counts of rent is 51, however on April 22 2012, the prediction is 111. There was a historical snow in April.It was a kind of unsual weather condision in end of April. So that the Forest Model could not accurately predicts the right counts due to such special weather condition.

Conclusion

Through exploring the Bike dataset and using several different methods of Linear Regression we developed 4 algorithms to predict number of bike shared using the dataset variables informatiom.

First we applied the Linear Regression OLS method and through several steps of correcting the model. The residuals is a little bit skwed however we did not adjust it because the count of the bike shares is not normaly distributed.

Next we applied tree-based regression methods. First we used a Regression Tree method which gave us accuracy of 0.83 correlation to the test$cnt. The final method was the Random Forest Regressor and achieved 90% better accuracy at 0.92 comparing to OLS.

The project was a success, however, none of the Linear Regression methods used would give us reliable precision. All the plots show the presence of outliers and inaccuracy in the areas of low and high scores. We conclude that the current problem could not be solved by the Linear Regression methods only, it looks that additional methods like Clustering is required to split the dataset into smaller sets to satisfy Linear Regression limitations. Or maybe Time Series Regression Forest methods will help.

We also use the algorithms to predict a real data base on the day of April 22 2012 information in Washingto DC. The result shows that Linear Regression has the best prediction 62 bike vs 51 actual bike. It might be the reason that LM OLS put temp and humidity more important than workingday. Moreover, on April 22 2012, there was a historical snow in April.It was a kind of unsual weather condision in end of April. So that the Tree or Forest Models could not be accurately predict the right counts due to such special weather condition.