Citibike_ML

Hongyang Yang

12/4/2017

Purpose: Is there a linear relationship bewteen the number of rents per day and the weather data in NYC? Which weather factor affect the number of rents the most?

```r
library(ggplot2)
library(glmnet)
```

```
## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-13
```

```r
library(ISLR)
library(tree)
library(randomForest)
```

```
## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(e1071)
library(MASS)
library(caret)
```

```
## Loading required package: lattice
```

```r
library(gbm)
```

```
## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##     cluster

## Loading required package: splines

## Loading required package: parallel

## Loaded gbm 2.1.3
```

```r
citibike_daily_weather=read.csv("citibike_daily_weather.csv")

sum(is.na(citibike_daily_weather))
```

```
## [1] 79
```

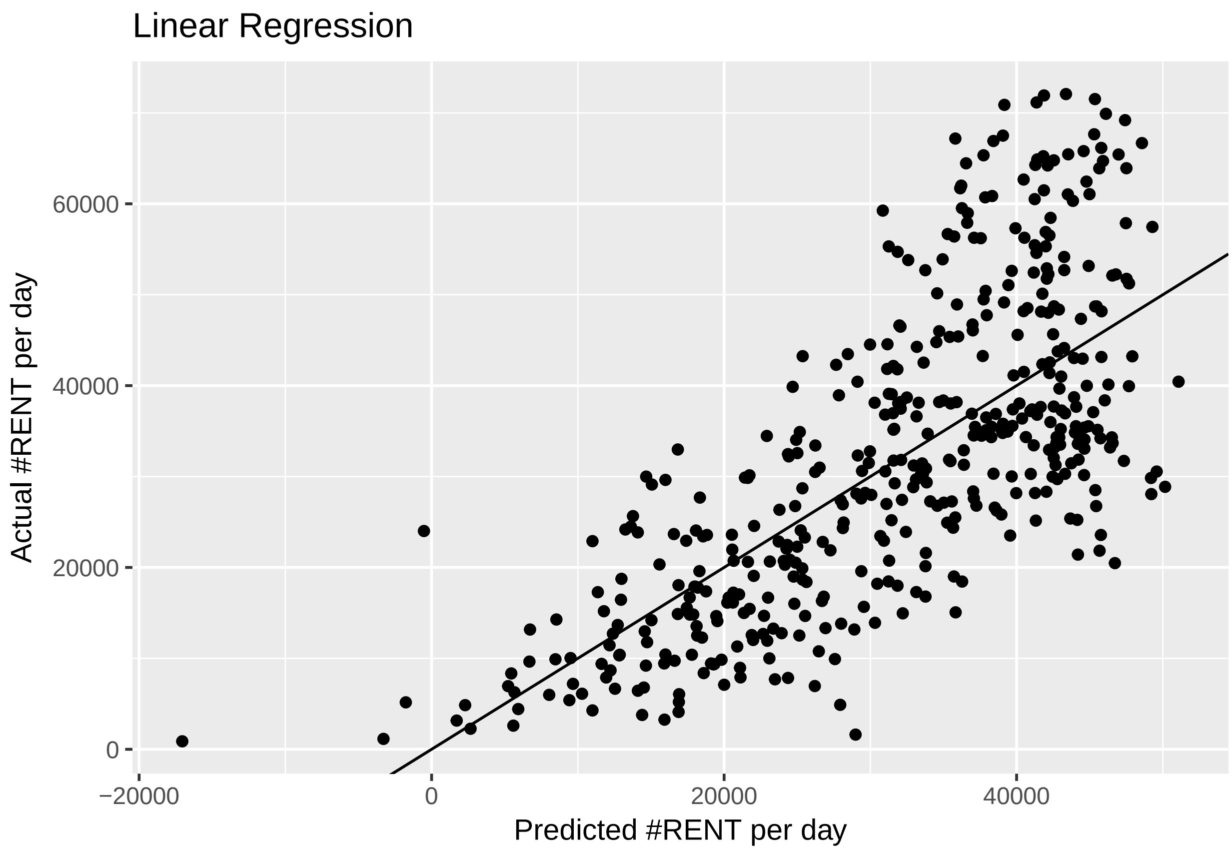
```r
citibike_daily_weather=na.omit(citibike_daily_weather)

set.seed(1)
train_ind=sample(1:nrow(citibike_daily_weather),0.7*nrow(citibike_daily_weather))
train=citibike_daily_weather[train_ind,c(2,5:14)]
test=citibike_daily_weather[-train_ind,c(2,5:14)]

test_x=test[,-1]
test_y=test[,1]
```

Model 1: Linear Regression

```r
###################################
#######Linear Regression##########
###################################

linear_model=lm(RENT~.,data=train)
summary(linear_model)
```

```
##
## Call:
## lm(formula = RENT ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31108  -7767  -1772   6449  33614
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5663.036   2585.314   2.190   0.0287 *
## AWND          -505.875    293.659  -1.723   0.0852 .
## PRCP        -10430.367   1059.245  -9.847  < 2e-16 ***
## SNOW          -539.047    574.435  -0.938   0.3483
## SNWD          -816.870    161.050  -5.072 4.65e-07 ***
## TMAX           464.017     68.750   6.749 2.46e-11 ***
## TMIN            33.289     72.534   0.459   0.6464
## WDF2           -11.762      5.347  -2.200   0.0281 *
## WDF5             4.624      5.485   0.843   0.3994
## WSF2           115.012    259.930   0.442   0.6582
## WSF5           -75.574    150.452  -0.502   0.6156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10810 on 1041 degrees of freedom
## Multiple R-squared:  0.5358, Adjusted R-squared:  0.5313
## F-statistic: 120.1 on 10 and 1041 DF,  p-value: < 2.2e-16
```

```r
linear_pre_y=predict(linear_model,test_x)
p.linear<-qplot((linear_pre_y), (test_y), xlab='Predicted #RENT per day',
              ylab='Actual #RENT per day', main='Linear Regression')
p.linear + geom_abline(slope=1, intercept=0)
```



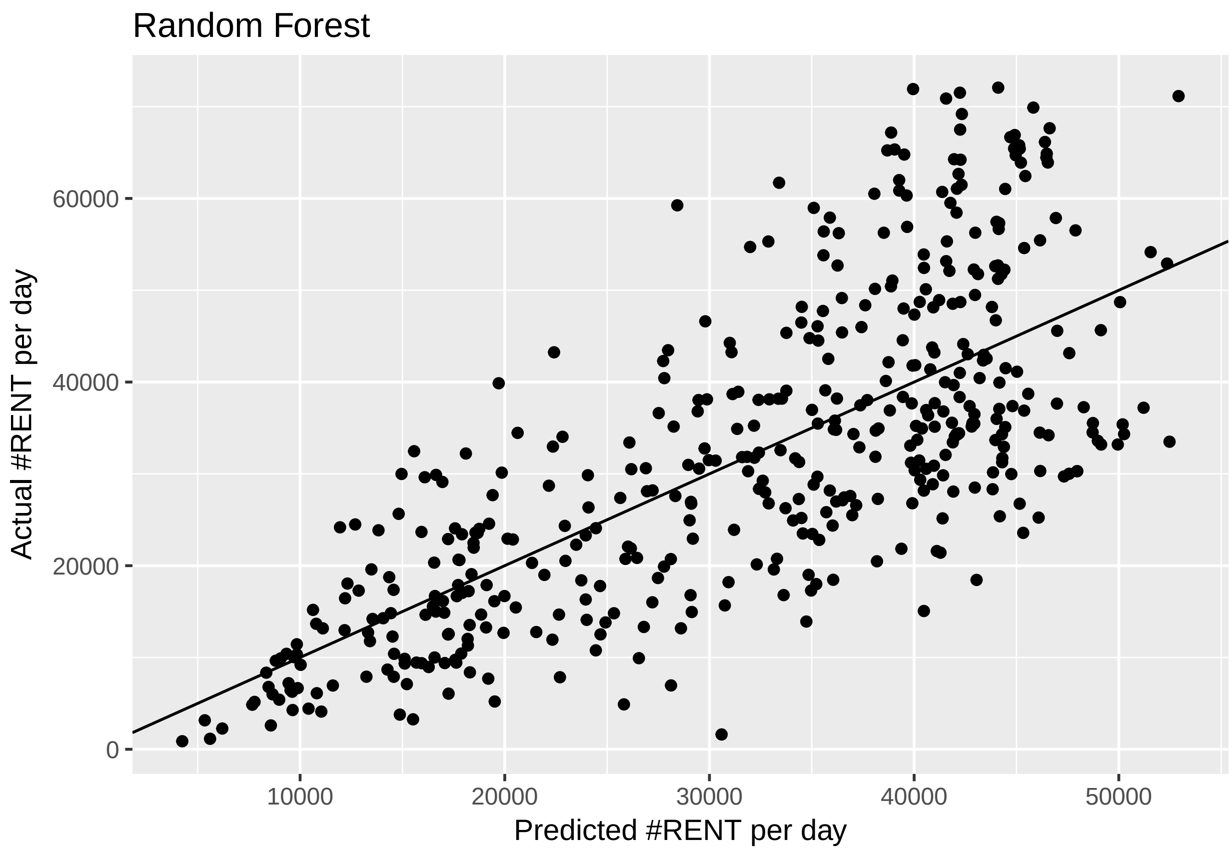
Model 2: Random Forest

```r
library(randomForest)
library(e1071)
library(MASS)
library(caret)


RF_Model=randomForest(RENT~.,data = na.omit(train) ,importance=TRUE, na.rm = TRUE)

RF_Model
```

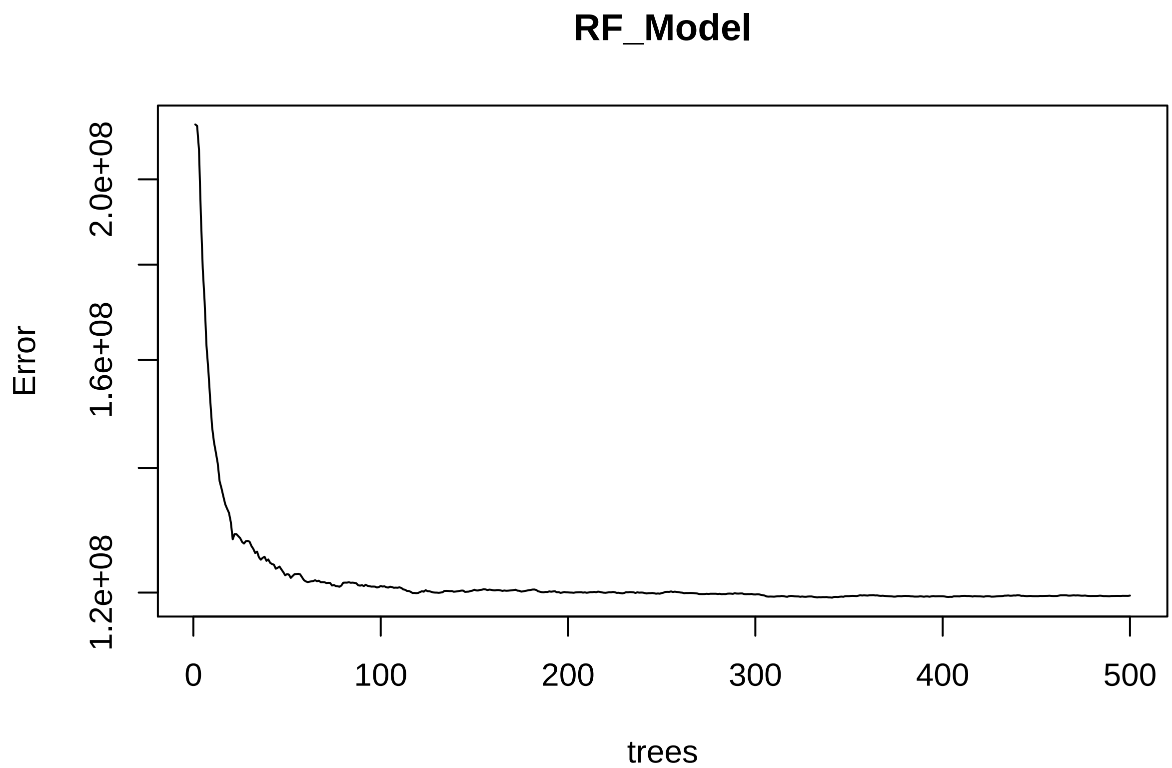
```
##
## Call:
##  randomForest(formula = RENT ~ ., data = na.omit(train), importance = TRUE,      na.rm = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
```

```
##           Mean of squared residuals: 119557939
##                     % Var explained: 52.04
```

```
yhat_bag=predict(RF_Model,test_x)
MSE_RF=mean((yhat_bag-test_y)^2,na.rm=TRUE)


#running the result
p.rf<-qplot((yhat_bag), (test_y), xlab='Predicted #RENT per day',
            ylab='Actual #RENT per day', main='Random Forest')
p.rf + geom_abline(slope=1, intercept=0)
```

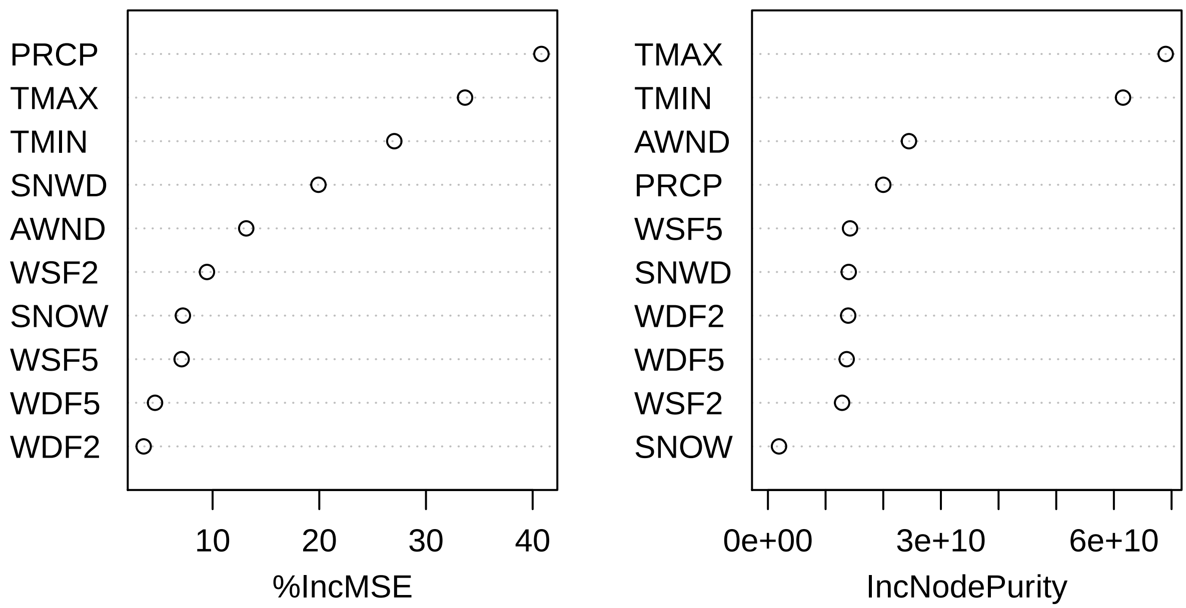



```
plot(RF_Model, log="y")
```




```
varImpPlot(RF_Model,main='Random Forest Importance Table')
```

Random Forest Importance Table

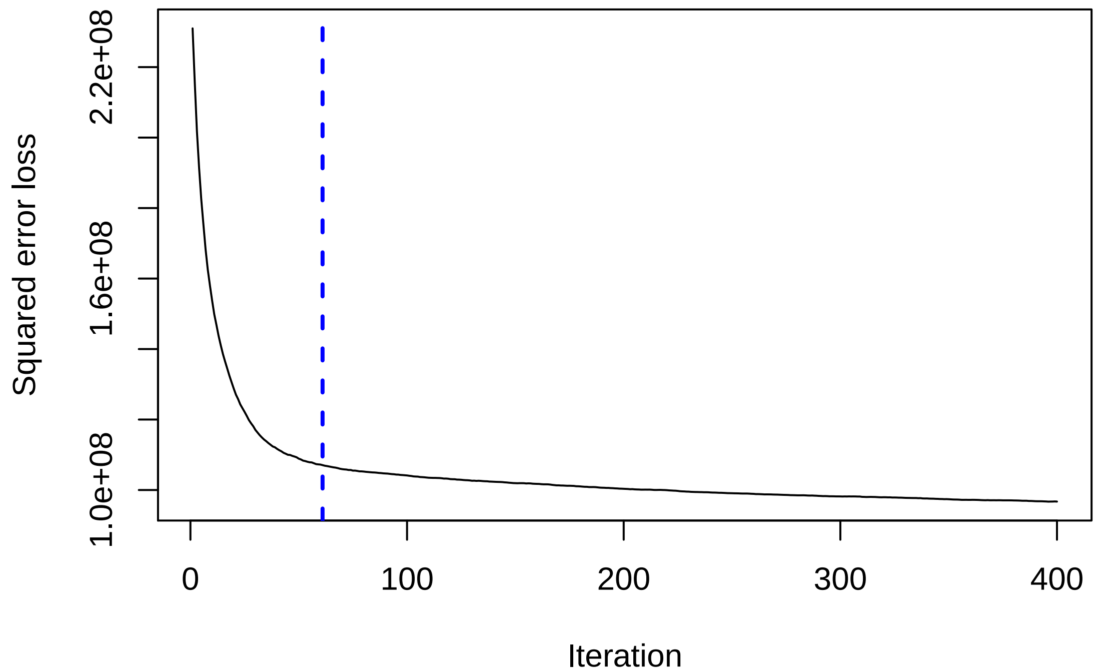



```
varImp(RF_Model)
```

```
##         Overall
## AWND 13.152700
## PRCP 40.822637
## SNOW  7.213752
## SNWD 19.920106
## TMAX 33.663150
## TMIN 27.032629
## WDF2  3.537843
## WDF5  4.603860
## WSF2  9.471154
## WSF5  7.097411
```

Model 3: GBM

```
#Generalized Boosted Regression Modeling
library(gbm)
gbm_model=gbm(RENT~.,data = train,dist="gaussian",n.tree = 400,shrinkage=0.1, cv.folds = 5)

best.iter <- gbm.perf(gbm_model,method="OOB")


gbm.perf(gbm_model,method="OOB")
```




```
## [1] 61
```

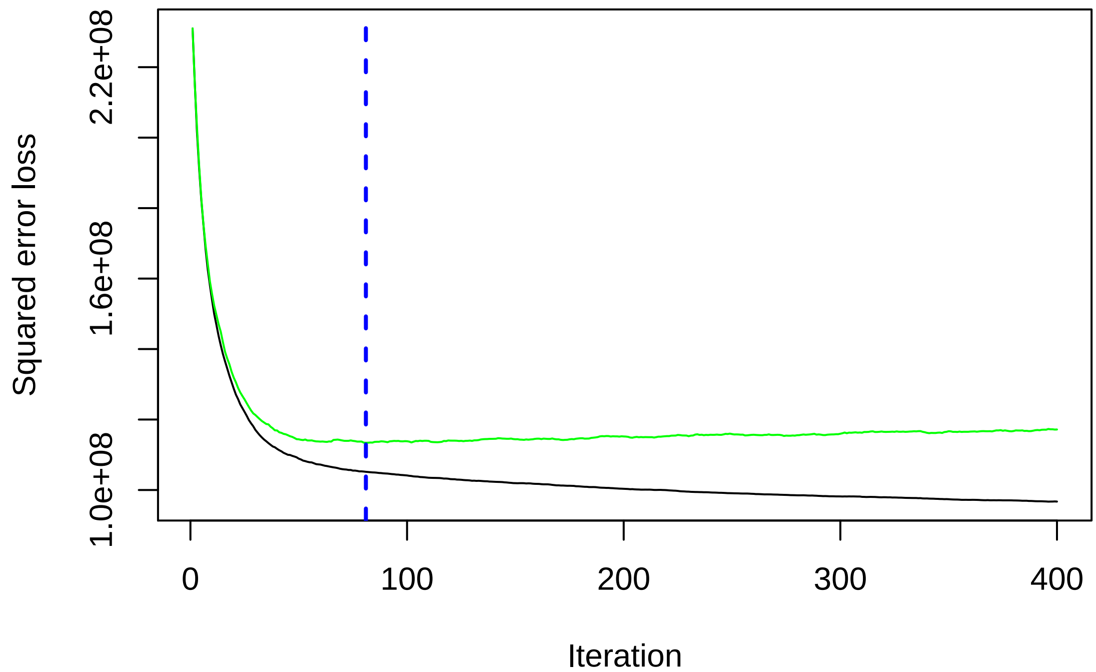
```r
print(best.iter)
```

```
## [1] 61
```

```r
best.iter <- gbm.perf(gbm_model,method="cv")
print(best.iter)
```

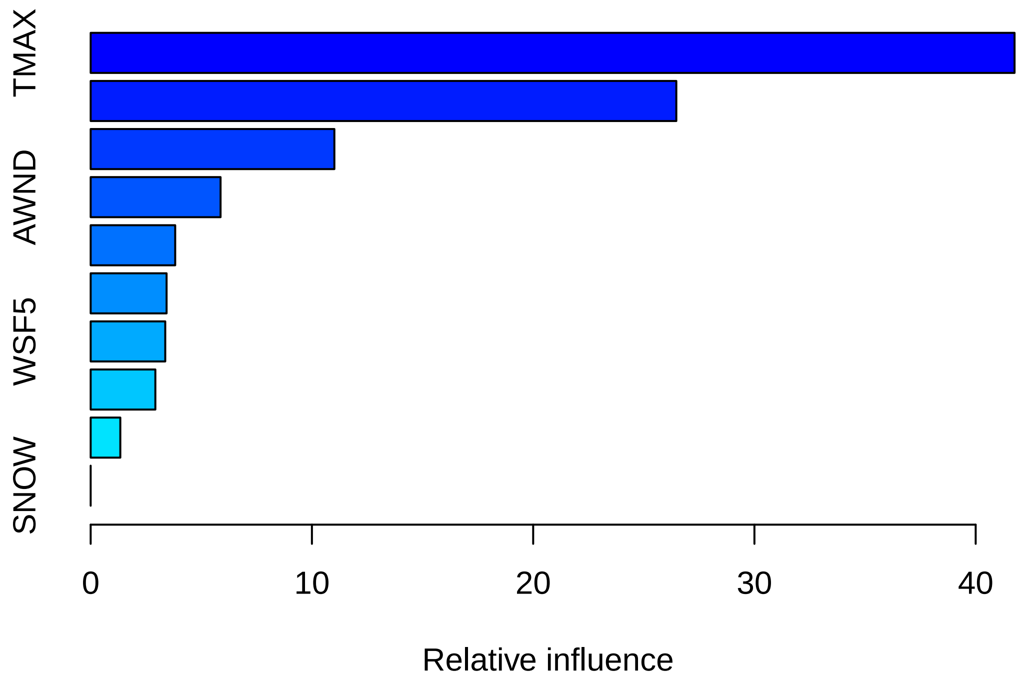
```
## [1] 81
```

```r
gbm.perf(gbm_model,method="cv")
```




```
## [1] 81
```

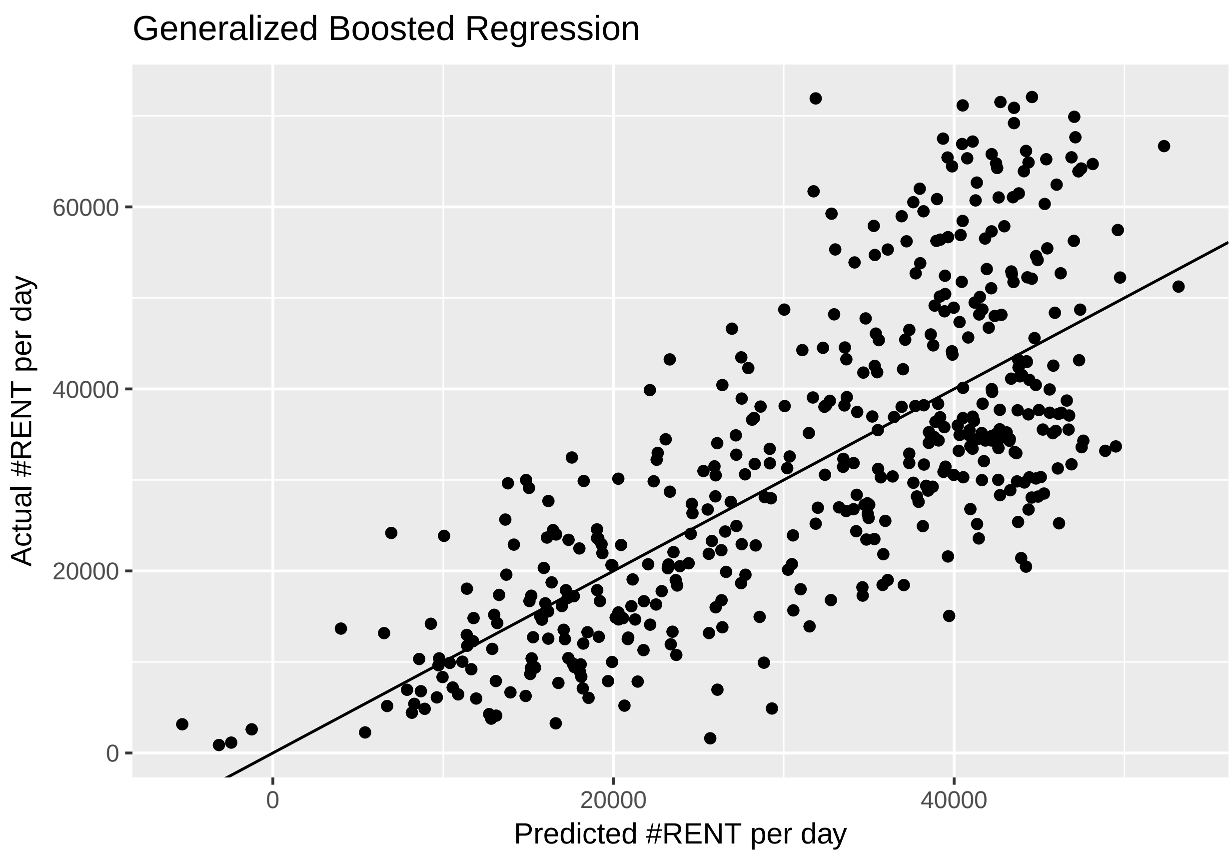
```r
sumary_GBM=summary(gbm_model)
```




```
sumary_GBM
```

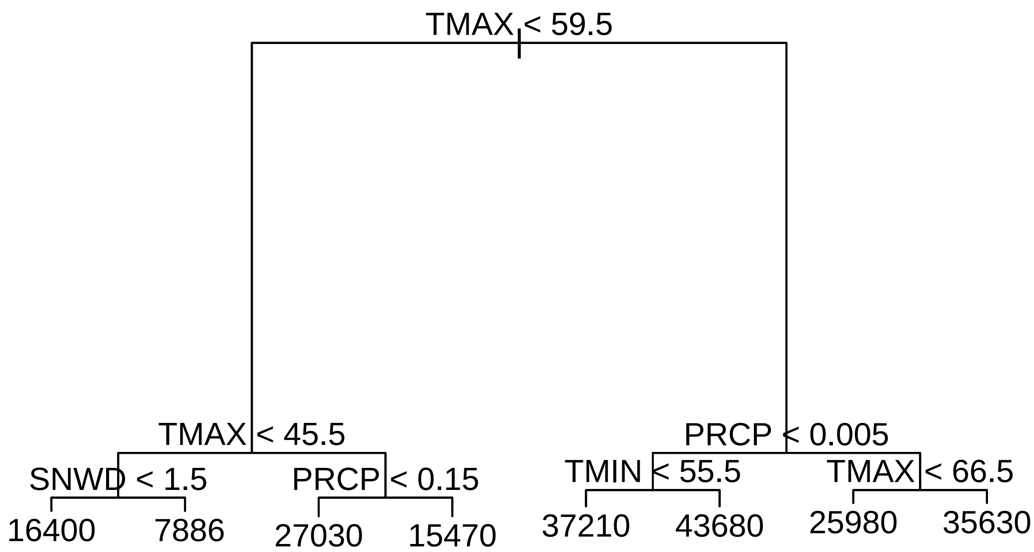
```
##       var   rel.inf
## TMAX TMAX 41.757137
## TMIN TMIN 26.469830
## PRCP PRCP 11.015002
## AWND AWND  5.870343
## WDF2 WDF2  3.821250
## WDF5 WDF5  3.431694
## WSF5 WSF5  3.370888
## SNWD SNWD  2.923930
## WSF2 WSF2  1.339927
## SNOW SNOW  0.000000
```

```
gbm_pred_y = predict(gbm_model, test, n.tree = 400, type = 'response')
MSE_gbm=mean((gbm_pred_y-test_y)^2,na.rm=TRUE)

p.rf<-qplot((gbm_pred_y), (test_y), xlab='Predicted #RENT per day',
            ylab='Actual #RENT per day', main='Generalized Boosted Regression')
p.rf +  geom_abline(slope=1, intercept=0)
```



Model 4: Regression Tree

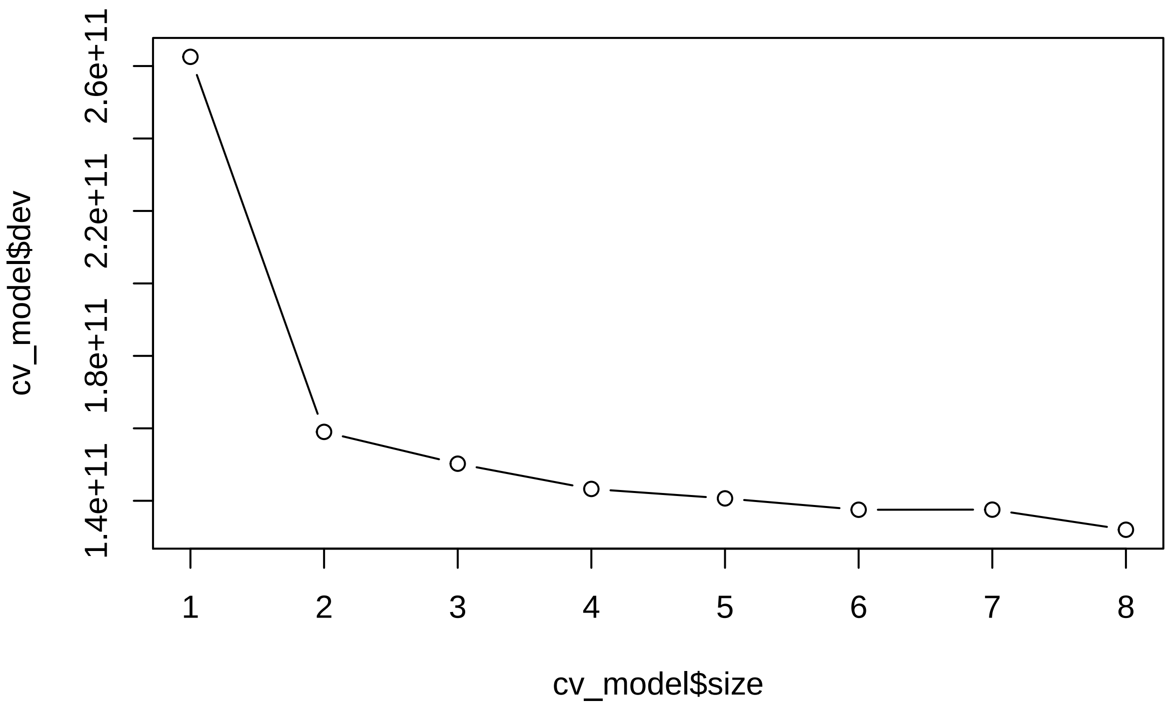
```
library(ISLR)
library(tree)
#set.seed(1)
tree_model=tree(RENT~.,data=train)
plot(tree_model)
text(tree_model,pretty=1)
```




```
tree_pred_y=predict(tree_model, test_x)

MSE_tree=mean((test_y-tree_pred_y)^2,na.rm=TRUE)


MSE_tree
```

```
## [1] 141486153
```

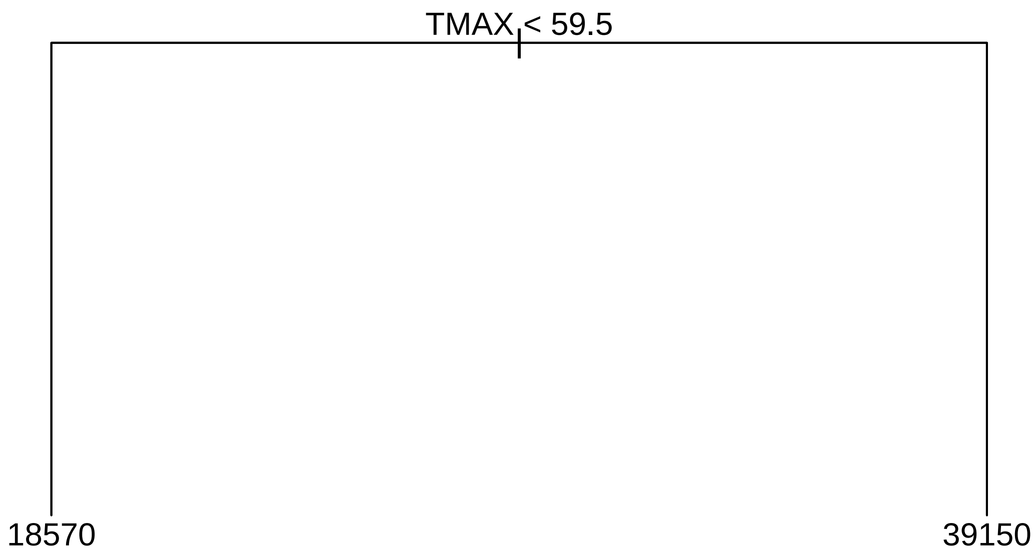
```
##### CROSS VALIDATION #####
cv_model=cv.tree(tree_model)
plot(cv_model$size,cv_model$dev,type='b')
```




```
bestSize=which.min(cv_model$dev)
print(bestSize)
```

```
## [1] 1
```

```
# Prune Tree
prune.tree=prune.tree(tree_model,best=2)
plot(prune.tree)
text(prune.tree,pretty=0)
```




```
pred.prune.tree = predict(prune.tree, newdata=test)
MSE_prune_tree=mean((test_y-pred.prune.tree)^2)
MSE_prune_tree
```

```
## [1] 165411575
```

Conclusion: Based on the 5 models we have, it can be concluded that there is a linear relationship between the number of rents per day and the weather data in NYC.