SUMMER PROJECT

VINEET

15/06/2021

**ROLL :MQMS2014**

**MODELLING USING BOSTON HOUSE PRICING DATA #importing the dataset**

**PROBLEM STATEMENT:**

**PROBLEM using boston housing dataset**

library(readxl)  
mydata <- read\_excel("C:/Users/Bineet/Downloads/Boston\_Housing\_Data 1set.xlsx")

## New names:  
## \* `` -> ...15

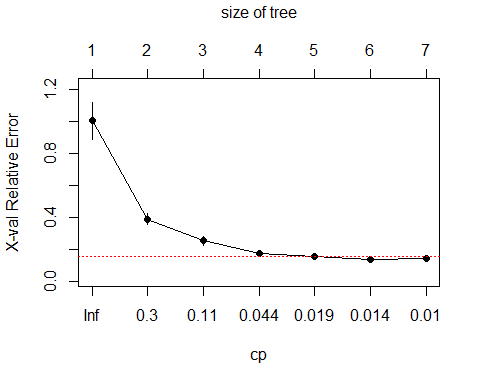
mydata <- mydata[,-15]  
set.seed(2)  
sampleid=sample(2,300,replace = TRUE,prob=c(0.8,0.2))  
training <- mydata[sampleid==1,]  
test <- mydata[sampleid==2,]

#CART MODEL

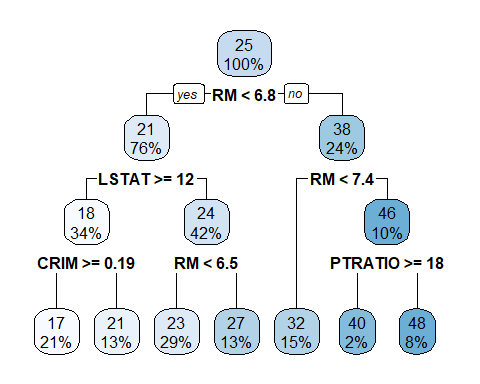
library(rpart)  
library(rpart.plot)  
attach(training)

## The following objects are masked from training (pos = 3):  
##   
## AGE, B, CHAS, CRIM, DIS, INDUS, LSTAT, MEDV, NOX, PTRATIO, RAD, RM,  
## TAX, ZN

set.seed(1)  
mymodel <- rpart(MEDV~.,data=training,method="anova",control = rpart.control(minsplit = 2))  
plotcp(mymodel,pch=19,col="red")



mymodel <- prune(mymodel,cp=0.01)  
rpart.plot(mymodel)



pred <- predict(mymodel)  
res <- training$MEDV-pred  
mse <- mean(res^2)  
mse #7.10

## [1] 7.103033

rmse=sqrt(mse)  
rmse #2.66

## [1] 2.665152

tss <- var(training$MEDV)\*(231-1)  
res\_ss <- sum(res^2)  
R\_sq <- 1-res\_ss/tss  
R\_sq #90.91%

## [1] 0.9091864

predtest <- predict(mymodel,newdata = test)  
res <- test$MEDV-predtest  
mse<- mean(res^2)  
mse #14.83

## [1] 14.83957

rmse <-sqrt(mse)  
rmse #3.85

## [1] 3.852216

tss <- var(test$MEDV)\*(69-1)  
res\_ss <- sum(res^2)  
R\_sq <- 1-res\_ss/tss  
R\_sq #81.54%

## [1] 0.8154488

FOR TRAINING DATA OF CART MSE=7.10 RMSE=2.66 R\_SQ=90.91% FOR TEST DATA OF CART ModEL MSE=14.83 RMSE=3.85 R\_SQ=81.54% CART MODEL IS ACCURATE AND GENERALIZABLE

#BAGGING MODEL

library(randomForest)  
set.seed(1)  
mymodel <- randomForest(MEDV~.,data=training,mtry=13,importance=TRUE)  
mymodel

##   
## Call:  
## randomForest(formula = MEDV ~ ., data = training, mtry = 13, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 13  
##   
## Mean of squared residuals: 7.594617  
## % Var explained: 90.29

predtest <- predict(mymodel,newdata = test)  
res <- test$MEDV-predtest  
mse <- mean(res^2)  
mse

## [1] 9.308411

rmse <- sqrt(mse)  
rmse

## [1] 3.050969

tss <- var(test$MEDV)\*(69-1)  
res\_ss <- sum(res^2)  
R\_sq <- 1-res\_ss/tss  
R\_sq

## [1] 0.8842367

FOR TRAINING DATA OF BAGGING MSE=7.59 RMSE=2.75 R\_SQ=90.29% FOR TEST DATA OF BAGGING ModEL MSE=9.30 RMSE=3.05 R\_SQ=88.42% MODEL IS ACCURATE AND GENERALIZABLE

#RANDOM FOREST MODEL

set.seed(1)  
library(randomForest)  
mymodel <- randomForest(MEDV~.,data=training,importance=TRUE)  
mymodel

##   
## Call:  
## randomForest(formula = MEDV ~ ., data = training, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 4  
##   
## Mean of squared residuals: 7.21054  
## % Var explained: 90.78

predtest <- predict(mymodel,newdata = test)  
res <- test$MEDV-predtest  
mse <- mean(res^2)  
mse

## [1] 6.718441

rmse <- sqrt(mse)  
rmse

## [1] 2.591996

tss <- var(test$MEDV)\*(69-1)  
res\_ss <- sum(res^2)  
R\_sq <- 1-res\_ss/tss  
R\_sq

## [1] 0.9164466

FOR TRAINING DATA OF RANDOM FOREST MSE=7.21 RMSE=2.75 R\_SQ=90.78% FOR TEST DATA OF RANDOM FOREST ModEL MSE=6.71 RMSE=2.59 R\_SQ=91.64% MODEL IS ACCURATE AND GENERALIZABLE



**ONLY THE CART ,BAGGING,RANDOM FOREST,SVM(LINEAR AND POLYNOMIAL) MODELS ARE ACCURATE.**

**ALSO ALL THE MODELS ARE GENERALIZABLE IF WE COMPARE THE RMSE (ROOT MEAN SQUARE ERROR) OF TEST AND TRAINING DATA OF ALL THE MODELS**

**THE GOOD MODELS ARE THE CART ,BAGGING,RANDOM FOREST,SVM(LINEAR AND POLYNOMIAL) MODELS.**

**THE BEST MODEL IS RANDOM FOREST MODEL AS IT HAS GOT LEAST RMSE AND MAXIMUM R\_SQ AND THAT TOO IN THE TEST DATA.**