Decision Tree

A cloth manufacturing company is interested to know about the segment or attributes causes high sale.

Approach - A decision tree can be built with target variable Sale (we will first convert it in categorical variable) & all other variable will be independent in the analysis.

**Ans :**

**R Code :**

## Decision Trees

########## Company Data Set #########

company <- read.csv('D:\\Data Science\\Excelr\\Assignments\\Assignment\\Decision Trees\\Company\_Data\_r.csv')

#install.packages("caret")

#install.packages("C50")

library(caret)

library(C50)

inTraininglocal <- createDataPartition(company$Target.Sales,p=.70,list=F)

training <- company[inTraininglocal,]

testing <- company[-inTraininglocal,]

#Model Building

model <- C5.0(training$Target.Sales~.,data=training)

#Generate the model summary

summary(model)

#Predict the test data set

pred <- predict.C5.0(model,testing[,-12])

a <- table(testing$Target.Sales,pred)

sum(diag(a))/sum(a)

plot(model)

#### Boosting

inTraininglocal <- createDataPartition(company$Target.Sales,p=.70,list=F)

training <- company[inTraininglocal,]

testing <- company[-inTraininglocal,]

#Model Building

model <- C5.0(training$Target.Sales~.,data=training,trials=25) #Trials - Boosting parameters

#Generate the model summary

summary(model)

#Predict the test data set

pred <- predict.C5.0(model,testing[,-12])

a <- table(testing$Target.Sales,pred)

sum(diag(a))/sum(a)

plot(model)

##### Bagging

acc <- c()

for (i in 1:100)

{

print(i)

##Data Partition

inTraininglocal <- createDataPartition(company$Target.Sales,p=.70,list=F)

training1 <- company[inTraininglocal,]

testing <- company[-inTraininglocal,]

## Model Building

fittree <- C5.0(training1$Target.Sales~.,data=training1)

#Predicting

pred <- predict.C5.0(fittree,testing[,-12])

a<-table(testing$Target.Sales,pred)

#Accuracy

acc<-c(acc,sum(diag(a))/sum(a))

}

summary(acc)

boxplot(acc)

##### Bagging and Boosting

acc <- c()

for (i in 1:100)

{

print(i)

##Data Partition

inTraininglocal <- createDataPartition(company$Target.Sales,p=.70,list=F)

training1 <- company[inTraininglocal,]

testing <- company[-inTraininglocal,]

## Model Building

fittree <- C5.0(training1$Target.Sales~.,data=training1,trials=20)

#Predicting

pred <- predict.C5.0(fittree,testing[,-12])

a<-table(testing$Target.Sales,pred)

#Accuracy

acc<-c(acc,sum(diag(a))/sum(a))

}

summary(acc)

boxplot(acc)

**Results :**

> model <- C5.0(training$Target.Sales~.,data=training)

> #Generate the model summary

> summary(model)

Call:

C5.0.formula(formula = training$Target.Sales ~ ., data = training)

C5.0 [Release 2.07 GPL Edition] Wed Apr 29 16:33:38 2020

-------------------------------

Class specified by attribute `outcome'

Read 281 cases (12 attributes) from undefined.data

Decision tree:

Sales > 9.95: High (56)

Sales <= 9.95:

:...Sales <= 4.99: Low (54)

Sales > 4.99: Medium (171)

Evaluation on training data (281 cases):

Decision Tree

----------------

Size Errors

3 0( 0.0%) <<

(a) (b) (c) <-classified as

---- ---- ----

56 (a): class High

54 (b): class Low

171 (c): class Medium

Attribute usage:

100.00% Sales

Time: 0.0 secs

> #Predict the test data set

> pred <- predict.C5.0(model,testing[,-12])

> a <- table(testing$Target.Sales,pred)

> sum(diag(a))/sum(a)

[1] 1

> model <- C5.0(training$Target.Sales~.,data=training,trials=25) #Trials - Boosting parameters

> #Generate the model summary

> summary(model)

Call:

C5.0.formula(formula = training$Target.Sales ~ ., data = training, trials = 25)

C5.0 [Release 2.07 GPL Edition] Wed Apr 29 16:33:55 2020

-------------------------------

Class specified by attribute `outcome'

Read 281 cases (12 attributes) from undefined.data

----- Trial 0: -----

Decision tree:

Sales > 9.95: High (56)

Sales <= 9.95:

:...Sales <= 4.97: Low (54)

Sales > 4.97: Medium (171)

\*\*\* boosting reduced to 1 trial since last classifier is very accurate

\*\*\* boosting abandoned (too few classifiers)

Evaluation on training data (281 cases):

Decision Tree

----------------

Size Errors

3 0( 0.0%) <<

(a) (b) (c) <-classified as

---- ---- ----

56 (a): class High

54 (b): class Low

171 (c): class Medium

Attribute usage:

100.00% Sales

Time: 0.0 secs

> #Predict the test data set

> pred <- predict.C5.0(model,testing[,-12])

> a <- table(testing$Target.Sales,pred)

> sum(diag(a))/sum(a)

[1] 0.9915966

Bagging

> summary(acc)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.9580 0.9916 0.9916 0.9921 1.0000 1.0000

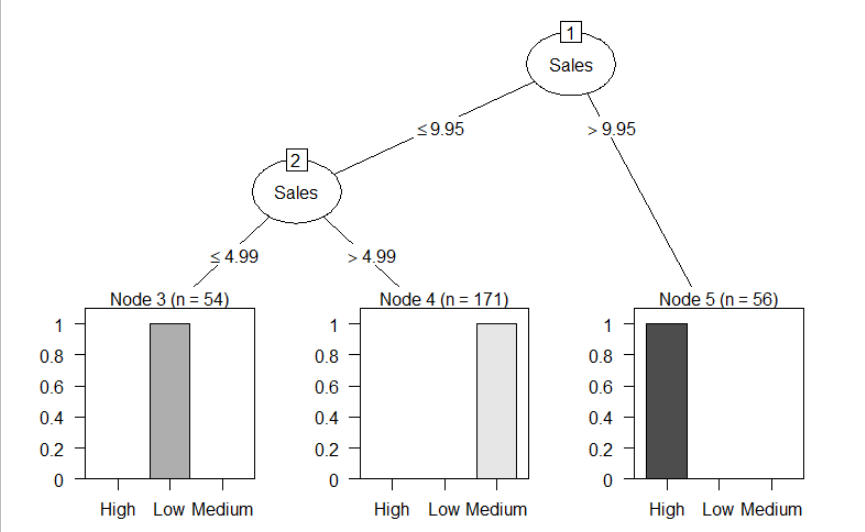
Bagging & Boosting

> summary(acc)

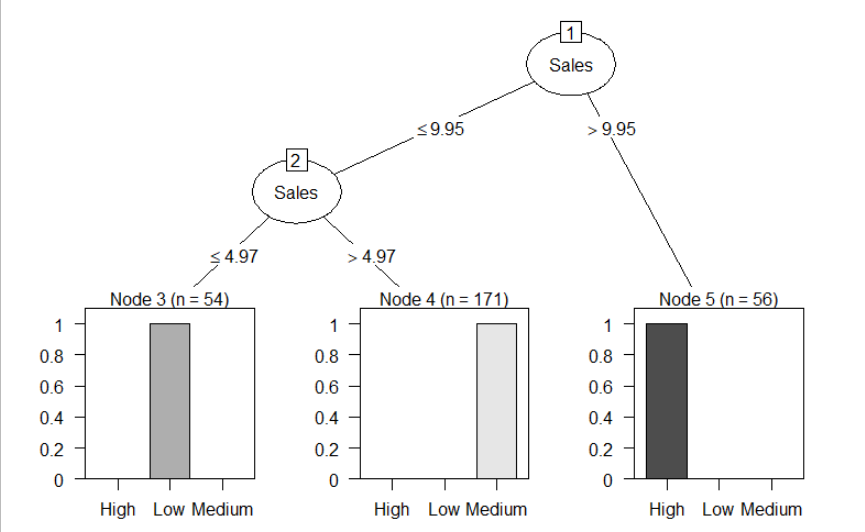
Min. 1st Qu. Median Mean 3rd Qu. Max.

0.9748 0.9916 1.0000 0.9941 1.0000 1.0000

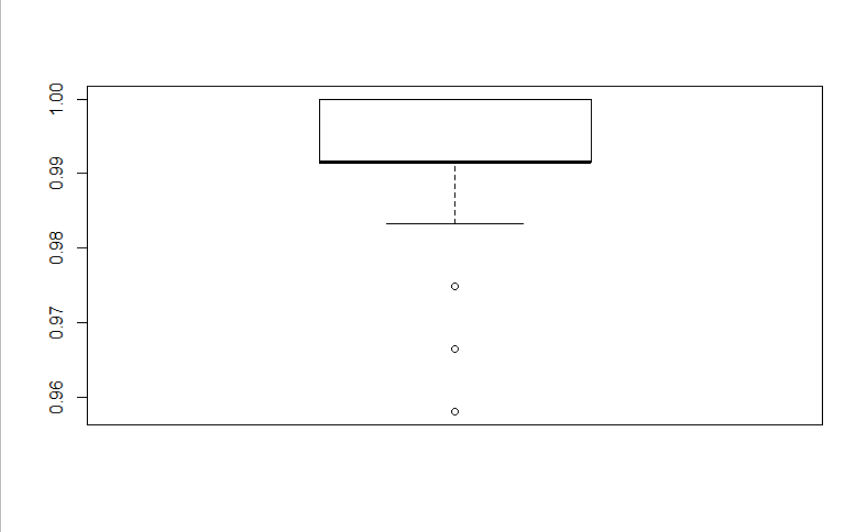
**Plots :**



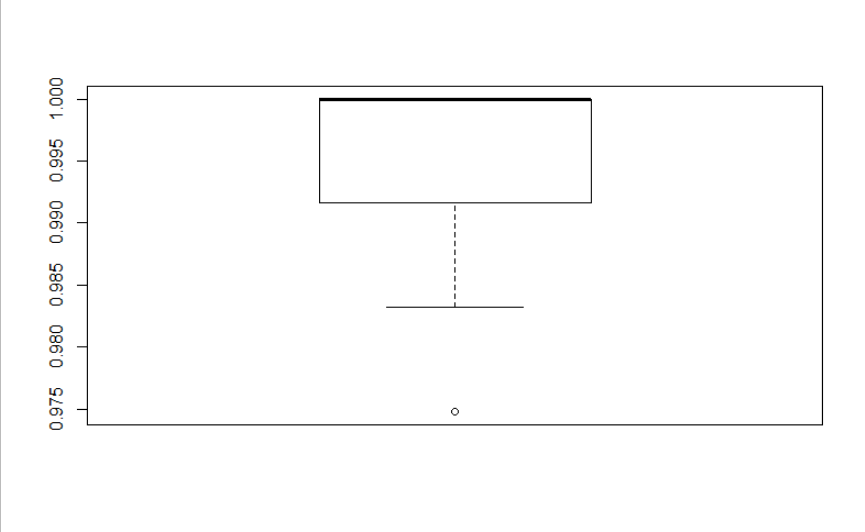
Boosting



Bagging



Bagging & Boosting



**Inference :**

Sales was found as the important feature and was considered as the root node.