**Decision Trees**

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

Iteration 1: Taking US as Variable

Codes:

Company\_Data

intraininglocal<-createDataPartition(Company\_Data$US,p=.70,list=F)

training<-Company\_Data[intraininglocal,]

testing<-Company\_Data[-intraininglocal,]

#Model Building

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

#generate summary

summary(model)

#predict for test data set

pred<-predict.C5.0(model,testing[,-11])

a<-table(testing$US,pred)

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

plot(model)

Results:

Call:

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

C5.0 [Release 2.07 GPL Edition] Fri Oct 23 02:12:41 2020

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

Class specified by attribute `outcome'

Read 281 cases (11 attributes) from undefined.data

Decision tree:

Advertising <= 1: No (109/16)

Advertising > 1: Yes (172/7)

Evaluation on training data (281 cases):

Decision Tree

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

Size Errors

2 23( 8.2%) <<

(a) (b) <-classified as

---- ----

93 7 (a): class No

16 165 (b): class Yes

Attribute usage:

100.00% Advertising

Time: 0.0 secs

> #predict for test data set

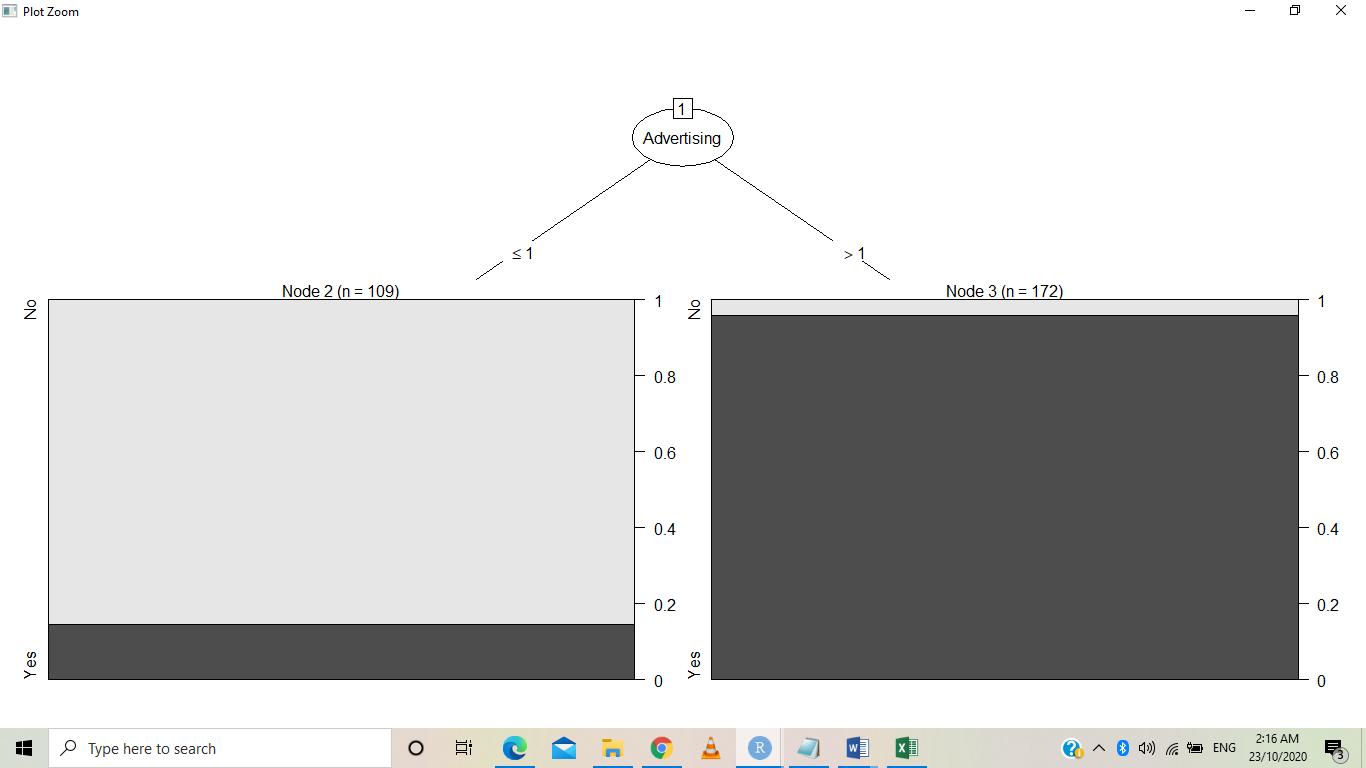
> pred<-predict.C5.0(model,testing[,-11])#5th column excluded which is a class variable

> a<-table(testing$US,pred)

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

[1] 0.8571429

Plot (Model)



Iteration 2: Considering Urban as a variable;

intraininglocal<-createDataPartition(Company\_Data$Urban,p=.70,list=F)

training<-Company\_Data[intraininglocal,]

testing<-Company\_Data[-intraininglocal,]

#Model Building

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

#generate summary

summary(model)

#predict for test data set

pred<-predict.C5.0(model,testing)#5th column excluded which is a class variable

a<-table(testing$Urban,pred)

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

plot(model)

US = No: Yes (101/36)

US = Yes:

:...Population > 380:

:...ShelveLoc = Medium: Yes (28/12)

: ShelveLoc = Good:

: :...Population <= 407: No (4)

: : Population > 407:

: : :...CompPrice <= 116: No (4/1)

: : CompPrice > 116: Yes (4)

: ShelveLoc = Bad:

: :...Sales > 8.41: No (2)

: Sales <= 8.41:

: :...Education > 12: Yes (6)

: Education <= 12:

: :...CompPrice <= 119: Yes (2)

: CompPrice > 119: No (3)

Population <= 380:

:...ShelveLoc = Bad: Yes (30/3)

ShelveLoc = Medium:

:...Age <= 32: No (5/1)

: Age > 32:

: :...Income <= 25: No (5/1)

: Income > 25: Yes (58/6)

ShelveLoc = Good:

:...CompPrice > 130: Yes (11)

CompPrice <= 130:

:...Advertising > 12: Yes (7)

Advertising <= 12:

:...Education > 17: Yes (2)

Education <= 17:

:...Education > 13: No (3)

Education <= 13:

:...Population <= 213: No (3)

Population > 213: Yes (3)

Evaluation on training data (281 cases):

Decision Tree

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

Size Errors

19 60(21.4%) <<

(a) (b) <-classified as

---- ----

26 57 (a): class No

3 195 (b): class Yes

Attribute usage:

100.00% US

64.06% Population

64.06% ShelveLoc

24.20% Age

22.42% Income

14.95% CompPrice

7.83% Education

6.41% Advertising

4.63% Sales

> #predict for test data set

> pred<-predict.C5.0(model,testing[,-11])#5th column excluded which is a class variable

Error in eval(predvars, data, env) : object 'US' not found

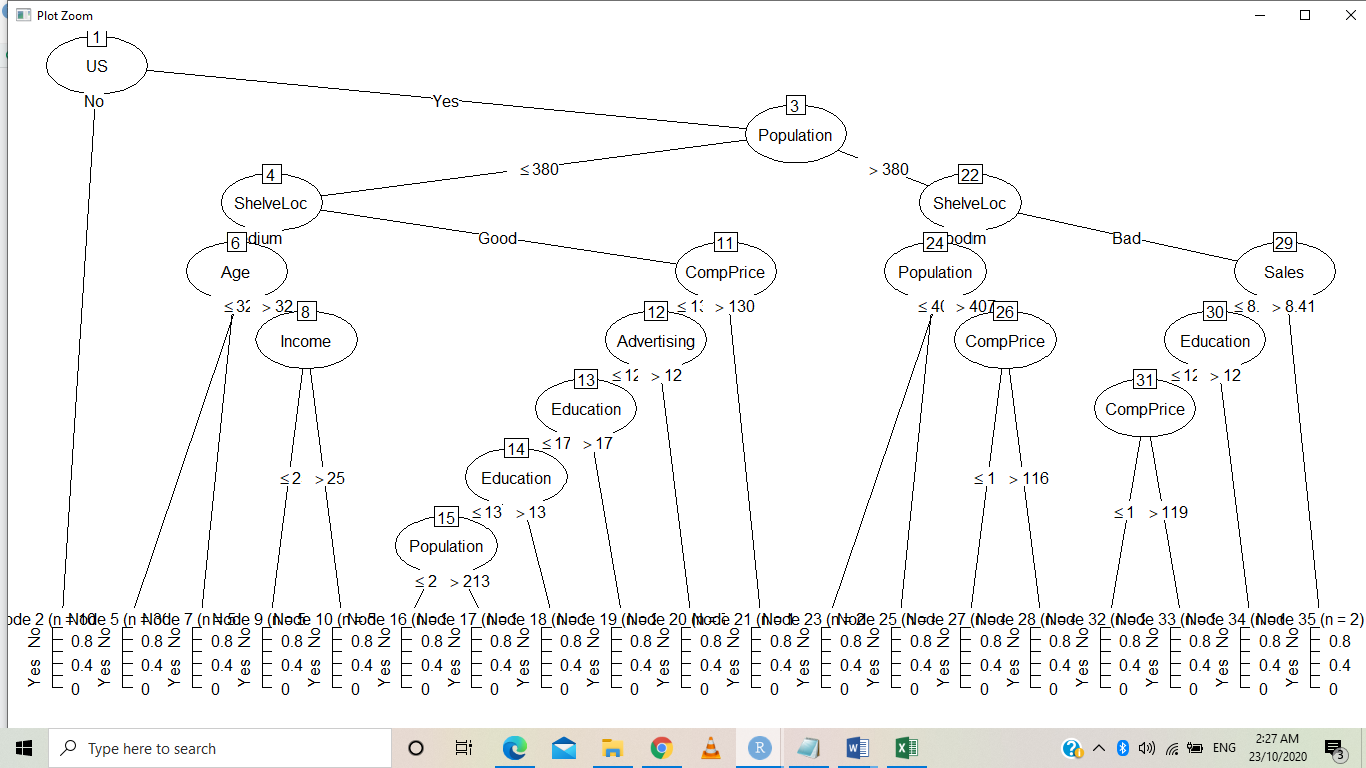
> #predict for test data set

> pred<-predict.C5.0(model,testing)#5th column excluded which is a class variable

> a<-table(testing$Urban,pred)

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

[1] 0.6806723



Conclusion: Iteration 1 best suits the for analysis.

1. Fraud Check

Statement :Use decision trees to prepare a model on fraud data treating those who have taxable\_income <= 30000 as "Risky" and others are "Good"

# Fraud Check

library(party)

library(C50)

df<-Fraud\_check

head(df)

str(df)

pairs(df)

#taxable income<=30000 as "Risky" and others are "Good"

type<-ifelse(df$Taxable.Income<=30000,"Risky","Good")

Fraud\_df<-data.frame(df,type)

plot(Fraud\_df$type)

**Solution In R:**

df<-Fraud\_check

> head(df)

Undergrad Marital.Status Taxable.Income City.Population Work.Experience Urban

1 NO Single 68833 50047 10 YES

2 YES Divorced 33700 134075 18 YES

3 NO Married 36925 160205 30 YES

4 YES Single 50190 193264 15 YES

5 NO Married 81002 27533 28 NO

6 NO Divorced 33329 116382 0 NO

> str(df)

'data.frame': 600 obs. of 6 variables:

$ Undergrad : Factor w/ 2 levels "NO","YES": 1 2 1 2 1 1 1 2 1 2 ...

$ Marital.Status : Factor w/ 3 levels "Divorced","Married",..: 3 1 2 3 2 1 1 3 3 1 ...

$ Taxable.Income : int 68833 33700 36925 50190 81002 33329 83357 62774 83519 98152 ...

$ City.Population: int 50047 134075 160205 193264 27533 116382 80890 131253 102481 155482 ...

$ Work.Experience: int 10 18 30 15 28 0 8 3 12 4 ...

$ Urban : Factor w/ 2 levels "NO","YES": 2 2 2 2 1 1 2 2 2 2 ...

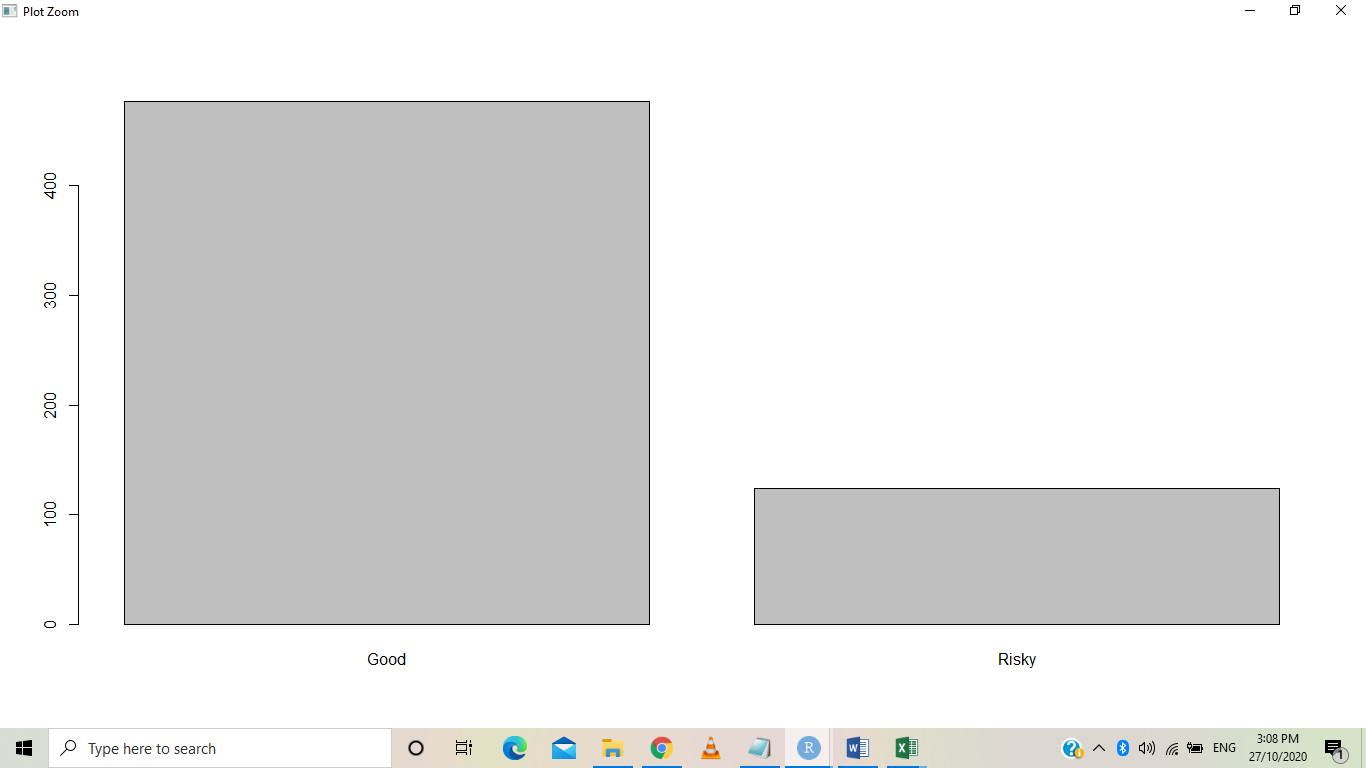
> pairs(df)

> #taxable income<=30000 as "Risky" and others are "Good"

> type<-ifelse(df$Taxable.Income<=30000,"Risky","Good")

> Fraud\_df<-data.frame(df,type)

> plot(Fraud\_df$type)



barplot(table(Fraud\_df$type))

> table(Fraud\_df$type)

Good Risky

476 124