decision Trees

By Saba Akram

2/11/2021

adult <- read.csv("D:/M.Sc in Banking and Financial Analytics/Sem 3/Data Analytic and Machine learning/Data Mining and Predictive Analysis/Data sets/adult/Clem3Training",stringsAsFactors = TRUE)  
  
## Collapsing some of the categories by giving them the same factor label.  
levels(adult$marital.status)

## [1] "Divorced" "Married-AF-spouse" "Married-civ-spouse"   
## [4] "Married-spouse-absent" "Never-married" "Separated"   
## [7] "Widowed"

levels(adult$workclass)

## [1] "?" "Federal-gov" "Local-gov" "Never-worked"   
## [5] "Private" "Self-emp-inc" "Self-emp-not-inc" "State-gov"   
## [9] "Without-pay"

levels(adult$marital.status)[2:4]<-"Married"  
levels(adult$marital.status)

## [1] "Divorced" "Married" "Never-married" "Separated"   
## [5] "Widowed"

levels(adult$workclass)[c(2,3,8)]<-"Gov"  
levels(adult$workclass)[c(5,6)]<-"Self"  
levels(adult$workclass)

## [1] "?" "Gov" "Never-worked" "Private" "Self"   
## [6] "Without-pay"

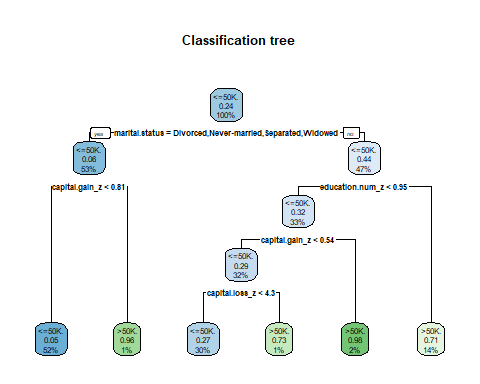
## Now as we can see that the levels "Married-AF-spouse", "Married-civ-spouse" and "Married-spouse-absent" of the marital.status are combined into a single level i.e. "married" and the levels "Self-emp-inc" and "Self-emp-not-inc" are combined into a single level i.e. "Self".

## Now we are standardizing the numeric variables to take them on the same scale.  
  
adult$age\_z<-(adult$age-mean(adult$age))/sd(adult$age)  
adult$education.num\_z<-(adult$education.num-mean(adult$education.num))/sd(adult$education.num)  
adult$capital.gain\_z<-(adult$capital.gain-mean(adult$capital.gain))/sd(adult$capital.gain)  
adult$capital.loss\_z<-(adult$capital.loss-mean(adult$capital.loss))/sd(adult$capital.loss)  
adult$hours.per.week\_z<-(adult$hours.per.week-mean(adult$hours.per.week))/sd(adult$hours.per.week)  
  
## Now I have standardizes the numeric variables in order to take them to the same scale.

## Now we use the predictors to classify whether or not a persons income is less than $50k.  
## it requires package rpart. Decision trees can be implemented by using the 'rpart' package in R. The 'rpart' package extends to Recursive Partitioning and Regression Trees which applies the tree-based model for regression and classification problems.  
  
library(rpart)  
cartfit<-rpart(income~age\_z+education.num\_z+capital.gain\_z+capital.loss\_z+hours.per.week\_z+race+sex+workclass+marital.status,data = adult)  
print(cartfit)

## n= 25000   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 25000 5984 <=50K. (0.76064000 0.23936000)   
## 2) marital.status=Divorced,Never-married,Separated,Widowed 13215 845 <=50K. (0.93605751 0.06394249)   
## 4) capital.gain\_z< 0.8082312 12986 625 <=50K. (0.95187125 0.04812875) \*  
## 5) capital.gain\_z>=0.8082312 229 9 >50K. (0.03930131 0.96069869) \*  
## 3) marital.status=Married 11785 5139 <=50K. (0.56393721 0.43606279)   
## 6) education.num\_z< 0.9458454 8296 2672 <=50K. (0.67791707 0.32208293)   
## 12) capital.gain\_z< 0.5352109 7894 2280 <=50K. (0.71117304 0.28882696)   
## 24) capital.loss\_z< 4.254168 7615 2076 <=50K. (0.72738017 0.27261983) \*  
## 25) capital.loss\_z>=4.254168 279 75 >50K. (0.26881720 0.73118280) \*  
## 13) capital.gain\_z>=0.5352109 402 10 >50K. (0.02487562 0.97512438) \*  
## 7) education.num\_z>=0.9458454 3489 1022 >50K. (0.29292061 0.70707939) \*

## plotting the decision trees  
## It requires "rpart.plot" package  
library(rpart.plot)  
rpart.plot(cartfit,main="Classification tree")



## C5.0  
library(C50)  
names(adult)

## [1] "age" "workclass" "demogweight" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country" "income" "age\_z"   
## [17] "education.num\_z" "capital.gain\_z" "capital.loss\_z" "hours.per.week\_z"

x<-adult[,c(2,6,9,10,16:20)]  
y<-adult$income  
c50fit1<-C5.0(x,y)  
summary(c50fit1)

##   
## Call:  
## C5.0.default(x = x, y = y)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu Mar 18 16:35:44 2021  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 25000 cases (10 attributes) from undefined.data  
##   
## Decision tree:  
##   
## capital.gain\_z > 0.7694287:  
## :...hours.per.week\_z > -0.4396555: >50K. (973/10)  
## : hours.per.week\_z <= -0.4396555:  
## : :...age\_z <= -0.8479775: <=50K. (6/1)  
## : age\_z > -0.8479775: >50K. (86/3)  
## capital.gain\_z <= 0.7694287:  
## :...marital.status in {Divorced,Never-married,Separated,Widowed}:  
## :...capital.loss\_z > 5.282196:  
## : :...capital.loss\_z <= 5.646056: <=50K. (38/12)  
## : : capital.loss\_z > 5.646056:  
## : : :...capital.loss\_z <= 7.270963: >50K. (28)  
## : : capital.loss\_z > 7.270963: <=50K. (8/1)  
## : capital.loss\_z <= 5.282196:  
## : :...capital.gain\_z > 0.4757047:  
## : :...capital.gain\_z <= 0.494004: >50K. (19)  
## : : capital.gain\_z > 0.494004: <=50K. (67/6)  
## : capital.gain\_z <= 0.4757047:  
## : :...education.num\_z <= 0.7503064: <=50K. (10336/211)  
## : education.num\_z > 0.7503064:  
## : :...education.num\_z <= 1.532462:  
## : :...hours.per.week\_z <= 0.2107898: <=50K. (1598/109)  
## : : hours.per.week\_z > 0.2107898:  
## : : :...sex = Female: <=50K. (314/49)  
## : : sex = Male:  
## : : :...education.num\_z <= 1.141384: <=50K. (328/83)  
## : : education.num\_z > 1.141384:  
## : : :...age\_z > 1.197644: <=50K. (14/2)  
## : : age\_z <= 1.197644:  
## : : :...age\_z <= -0.4096299: <=50K. (15/1)  
## : : age\_z > -0.4096299: >50K. (50/14)  
## : education.num\_z > 1.532462:  
## : :...age\_z <= -0.4826879: <=50K. (57/7)  
## : age\_z > -0.4826879:  
## : :...age\_z > 1.124586: <=50K. (22/3)  
## : age\_z <= 1.124586:  
## : :...marital.status in {Never-married,  
## : : Widowed}: >50K. (50/13)  
## : marital.status = Separated:  
## : :...sex = Female: <=50K. (6/2)  
## : : sex = Male: >50K. (7/1)  
## : marital.status = Divorced:  
## : :...education.num\_z <= 1.92354:  
## : :...sex = Female: <=50K. (8)  
## : : sex = Male: >50K. (10/4)  
## : education.num\_z > 1.92354:  
## : :...hours.per.week\_z <= 1.755597: >50K. (9/2)  
## : hours.per.week\_z > 1.755597: <=50K. (2)  
## marital.status = Married:  
## :...capital.loss\_z > 4.175664:  
## :...capital.loss\_z <= 4.711484: >50K. (439/10)  
## : capital.loss\_z > 4.711484:  
## : :...capital.loss\_z <= 5.175032: <=50K. (48)  
## : capital.loss\_z > 5.175032:  
## : :...education.num\_z > 0.7503064: >50K. (47/1)  
## : education.num\_z <= 0.7503064:  
## : :...education.num\_z <= -0.8140052: <=50K. (7)  
## : education.num\_z > -0.8140052:  
## : :...capital.loss\_z > 5.803064: <=50K. (7)  
## : capital.loss\_z <= 5.803064:  
## : :...capital.loss\_z > 5.745743: >50K. (9)  
## : capital.loss\_z <= 5.745743:  
## : :...age\_z <= 0.613181: <=50K. (10)  
## : age\_z > 0.613181: >50K. (13/4)  
## capital.loss\_z <= 4.175664:  
## :...capital.gain\_z > 0.5241912:  
## :...capital.gain\_z <= 0.7118593: >50K. (77)  
## : capital.gain\_z > 0.7118593:  
## : :...capital.gain\_z <= 0.7246822: >50K. (4)  
## : capital.gain\_z > 0.7246822: <=50K. (4)  
## capital.gain\_z <= 0.5241912:  
## :...education.num\_z > 0.7503064:  
## :...capital.loss\_z > 1.342044: <=50K. (25/5)  
## : capital.loss\_z <= 1.342044:  
## : :...capital.gain\_z > 0.2690694:  
## : :...capital.gain\_z <= 0.4023739: <=50K. (25)  
## : : capital.gain\_z > 0.4023739:  
## : : :...capital.gain\_z <= 0.4444489: >50K. (12/1)  
## : : capital.gain\_z > 0.4444489: <=50K. (17)  
## : capital.gain\_z <= 0.2690694:  
## : :...hours.per.week\_z <= -0.6835725:  
## : :...sex = Female: >50K. (69/28)  
## : : sex = Male:  
## : : :...education.num\_z <= 1.532462: <=50K. (140/31)  
## : : education.num\_z > 1.532462: >50K. (31/12)  
## : hours.per.week\_z > -0.6835725:  
## : :...age\_z <= -0.4096299:  
## : :...age\_z <= -0.9940934: <=50K. (50/14)  
## : : age\_z > -0.9940934: [S1]  
## : age\_z > -0.4096299:  
## : :...race in {Black,White}: >50K. (1747/507)  
## : race = Other: <=50K. (6/1)  
## : race = Amer-Indian-Eskimo:  
## : :...hours.per.week\_z <= 0.2920955: <=50K. (4)  
## : : hours.per.week\_z > 0.2920955: >50K. (3)  
## : race = Asian-Pac-Islander:  
## : :...sex = Male: >50K. (95/41)  
## : sex = Female:  
## : :...age\_z <= 0.1748335: >50K. (6/1)  
## : age\_z > 0.1748335: <=50K. (15/2)  
## education.num\_z <= 0.7503064:  
## :...capital.loss\_z > 3.68221: <=50K. (92)  
## capital.loss\_z <= 3.68221:  
## :...capital.gain\_z > 0.4444489: <=50K. (47)  
## capital.gain\_z <= 0.4444489:  
## :...capital.gain\_z > 0.4023739:  
## :...hours.per.week\_z <= 0.9425408: >50K. (41/5)  
## : hours.per.week\_z > 0.9425408: <=50K. (4/1)  
## capital.gain\_z <= 0.4023739:  
## :...capital.gain\_z > 0.2690694: <=50K. (131)  
## capital.gain\_z <= 0.2690694:  
## :...capital.gain\_z > 0.2543766: >50K. (50/2)  
## capital.gain\_z <= 0.2543766:  
## :...capital.gain\_z > -0.09184103: <=50K. (118)  
## capital.gain\_z <= -0.09184103:  
## :...hours.per.week\_z <= -0.5209612: <=50K. (736/92)  
## hours.per.week\_z > -0.5209612:  
## :...age\_z <= -0.2635141: <=50K. (2178/408)  
## age\_z > -0.2635141: [S2]  
##   
## SubTree [S1]  
##   
## race in {Asian-Pac-Islander,Other}: <=50K. (26/10)  
## race in {Amer-Indian-Eskimo,White}: >50K. (377/162)  
## race = Black:  
## :...education.num\_z <= 1.141384: >50K. (19/8)  
## education.num\_z > 1.141384: <=50K. (2)  
##   
## SubTree [S2]  
##   
## race in {Amer-Indian-Eskimo,Other}: <=50K. (55/7)  
## race in {Asian-Pac-Islander,Black,White}:  
## :...education.num\_z <= -0.4229273: <=50K. (2659/777)  
## education.num\_z > -0.4229273:  
## :...hours.per.week\_z > 0.2107898:  
## :...workclass in {?,Gov,Never-worked,Private}: >50K. (410.8/181.4)  
## : workclass = Without-pay: <=50K. (1)  
## : workclass = Self:  
## : :...education.num\_z <= 0.3592285: <=50K. (175.2/74.5)  
## : education.num\_z > 0.3592285:  
## : :...hours.per.week\_z <= 1.349069: >50K. (11/2)  
## : hours.per.week\_z > 1.349069: <=50K. (4)  
## hours.per.week\_z <= 0.2107898:  
## :...workclass in {?,Gov,Never-worked,Without-pay}: <=50K. (190/93.5)  
## workclass = Self:  
## :...sex = Female: >50K. (8.4/2.3)  
## : sex = Male: <=50K. (85.6/33.6)  
## workclass = Private:  
## :...age\_z > 0.613181:  
## :...age\_z <= 1.70905: >50K. (206.5/94.8)  
## : age\_z > 1.70905: <=50K. (20.7/3.7)  
## age\_z <= 0.613181:  
## :...sex = Male: <=50K. (337.4/118.7)  
## sex = Female:  
## :...education.num\_z <= -0.03184938: <=50K. (41.7/17)  
## education.num\_z > -0.03184938: >50K. (11.7/2)  
##   
##   
## Evaluation on training data (25000 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 78 3286(13.1%) <<  
##   
##   
## (a) (b) <-classified as  
## ---- ----  
## 17902 1114 (a): class <=50K.  
## 2172 3812 (b): class >50K.  
##   
##   
## Attribute usage:  
##   
## 100.00% capital.gain\_z  
## 95.74% marital.status  
## 95.74% capital.loss\_z  
## 92.81% education.num\_z  
## 52.65% hours.per.week\_z  
## 36.44% age\_z  
## 26.07% race  
## 6.43% sex  
## 5.90% workclass  
##   
##   
## Time: 0.4 secs

## C5.0 Pruned  
c50fit2<-C5.0(x,y,control = C5.0Control(CF=.1))  
summary(c50fit2)

##   
## Call:  
## C5.0.default(x = x, y = y, control = C5.0Control(CF = 0.1))  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu Mar 18 16:35:45 2021  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 25000 cases (10 attributes) from undefined.data  
##   
## Decision tree:  
##   
## capital.gain\_z > 0.7694287: >50K. (1065/18)  
## capital.gain\_z <= 0.7694287:  
## :...marital.status in {Divorced,Never-married,Separated,Widowed}:  
## :...capital.loss\_z > 5.282196:  
## : :...capital.loss\_z <= 5.646056: <=50K. (38/12)  
## : : capital.loss\_z > 5.646056: >50K. (36/7)  
## : capital.loss\_z <= 5.282196:  
## : :...capital.gain\_z > 0.4757047:  
## : :...capital.gain\_z <= 0.494004: >50K. (19)  
## : : capital.gain\_z > 0.494004: <=50K. (67/6)  
## : capital.gain\_z <= 0.4757047:  
## : :...education.num\_z <= 0.7503064: <=50K. (10336/211)  
## : education.num\_z > 0.7503064:  
## : :...education.num\_z <= 1.532462: <=50K. (2319/280)  
## : education.num\_z > 1.532462:  
## : :...age\_z <= -0.4826879: <=50K. (57/7)  
## : age\_z > -0.4826879:  
## : :...age\_z <= 1.124586: >50K. (92/34)  
## : age\_z > 1.124586: <=50K. (22/3)  
## marital.status = Married:  
## :...capital.loss\_z > 4.175664:  
## :...capital.loss\_z <= 4.711484: >50K. (439/10)  
## : capital.loss\_z > 4.711484:  
## : :...capital.loss\_z <= 5.175032: <=50K. (48)  
## : capital.loss\_z > 5.175032:  
## : :...education.num\_z <= 0.7503064: <=50K. (46/18)  
## : education.num\_z > 0.7503064: >50K. (47/1)  
## capital.loss\_z <= 4.175664:  
## :...capital.gain\_z > 0.5241912: >50K. (85/4)  
## capital.gain\_z <= 0.5241912:  
## :...education.num\_z <= 0.7503064:  
## :...capital.gain\_z > 0.4444489: <=50K. (47)  
## : capital.gain\_z <= 0.4444489:  
## : :...capital.gain\_z > 0.4023739: >50K. (45/8)  
## : capital.gain\_z <= 0.4023739:  
## : :...capital.gain\_z > 0.2690694: <=50K. (131)  
## : capital.gain\_z <= 0.2690694:  
## : :...capital.gain\_z <= 0.2543766: <=50K. (7342/1991)  
## : capital.gain\_z > 0.2543766: >50K. (50/2)  
## education.num\_z > 0.7503064:  
## :...capital.loss\_z > 1.342044: <=50K. (25/5)  
## capital.loss\_z <= 1.342044:  
## :...capital.gain\_z > 0.2690694:  
## :...capital.gain\_z <= 0.4023739: <=50K. (25)  
## : capital.gain\_z > 0.4023739:  
## : :...capital.gain\_z <= 0.4444489: >50K. (12/1)  
## : capital.gain\_z > 0.4444489: <=50K. (17)  
## capital.gain\_z <= 0.2690694:  
## :...hours.per.week\_z <= -0.6835725:  
## :...sex = Female: >50K. (69/28)  
## : sex = Male:  
## : :...education.num\_z <= 1.532462: <=50K. (140/31)  
## : education.num\_z > 1.532462: >50K. (31/12)  
## hours.per.week\_z > -0.6835725:  
## :...age\_z > -0.4096299: >50K. (1876/571)  
## age\_z <= -0.4096299:  
## :...age\_z <= -0.9940934: <=50K. (50/14)  
## age\_z > -0.9940934: >50K. (424/188)  
##   
##   
## Evaluation on training data (25000 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 30 3462(13.8%) <<  
##   
##   
## (a) (b) <-classified as  
## ---- ----  
## 18132 884 (a): class <=50K.  
## 2578 3406 (b): class >50K.  
##   
##   
## Attribute usage:  
##   
## 100.00% capital.gain\_z  
## 95.74% marital.status  
## 95.74% capital.loss\_z  
## 92.81% education.num\_z  
## 10.36% hours.per.week\_z  
## 10.08% age\_z  
## 0.96% sex  
##   
##   
## Time: 0.4 secs