Decision Tree - II

Learn Classification and Prediction via Data Mining

Contents

Conditional Inference Tree Algorithm

- Conditional Inference (CI) Tree algorithm can also be divided into three main steps:

Test for Association → Creating Binary Splits → Repeat Until No Variables are Left

Step 1: Test for Association

- The algorithm tests if any independent variables are associated with the given response variable, and chooses the variable that has the strongest association with the response, i.e. Variable with the smallest p-value based on permutation test is chosen

Conditional Inference Tree Algorithm

Step 2: Split Variables

- The algorithm makes a binary split in this variable, dividing the dataset into two subsets
- In case of a binary predictor with values A and B, one subset will contain all observations with value A, and the other will contain all cases with value B. If a variable has more levels, one group may have values A and B, and the other may contain observations with C
- If the variable is quantitative, the range of its values can be split into two, e.g. values from 0 to 100 can be split into two subsets: from 0 to 50 and from 51 to 100; OR 0-30 and 31 to 100, and so on.

Conditional Inference Tree Algorithm

Step 3: Repeat Until No Variables are Left

- The first two steps are repeated for each subset until there are no variables that are associated with the outcome at the pre-defined level of statistical significance. This is why the algorithm is called recursive.

Tests Used in CI Algorithm

- Conditional Inference algorithm can be used for Classification as well as Regression Models.
- Structure of the algorithm remains the same, tests used for checking variable association change as per variable type.

Dependent Variable	Independent Variables	Test
Categorical	Categorical	Chi-square
Continuous	Continuous	Correlation
Continuous	Categorical	ANOVA

Data Snapshot

EMPLOYEE CHURN DATA

Dependent Variable

Independent Variables

sn	status	function	exp	gender	source
1	1	CS	<3	M	external

Columns	Description	Type	Measurement	Possible values
sn	Serial Number	-	-	-
status	= 1 If the Employee Left Within 18 Months of Joining = 0 Otherwise	Binary	1,0	2
function	Employee Job Profile	Categorical	CS, FINANCE, MARKETING	3
exp	Experience in Years	Categorical	<3,3-5,>5	3
gender	Gender of the Employee	Categorical	M,F	2
source	Whether the Employee was Appointed via Internal or External Links	categorical	external, internal	2

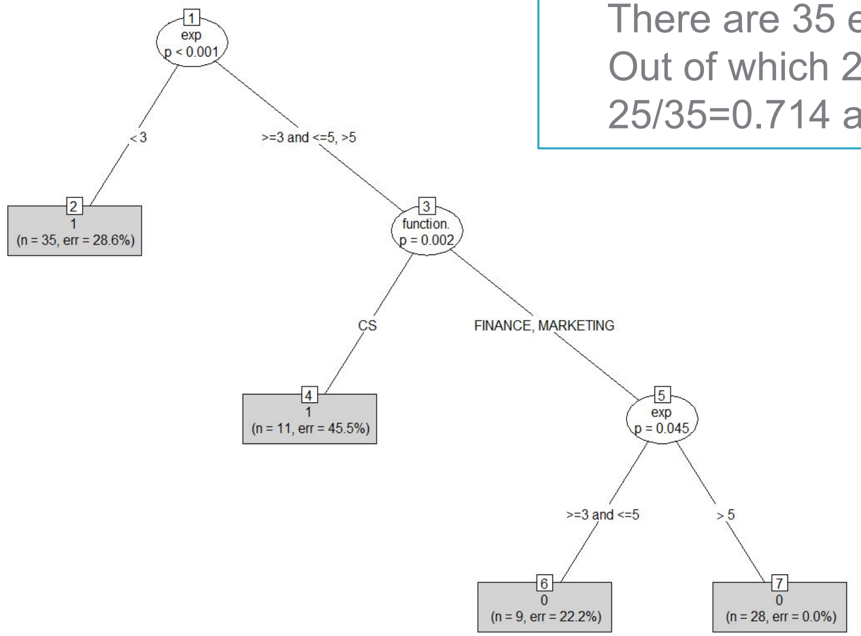
CHAID-like Implementation in Package “partykit”

```
# Decision Tree Using Package "partykit"
library(partykit)
empdata<-read.csv("EMPLOYEE CHURN DATA.csv",header=T)

empdata$status<-as.factor(empdata$status)
empdata$function.<-as.factor(empdata$function.)
empdata$exp<-as.factor(empdata$exp)
empdata$gender<-as.factor(empdata$gender)
empdata$source<-as.factor(empdata$source)

ctree<-partykit::ctree(formula=status~function.+exp+gender+sour
ce,
```

- We are instructing R to use the improved version of **ctree()** from package **“partykit”** by specifying **partykit::ctree()** in the command.
- formula= specifies dependent and independent variables

Decision Tree in Package "partykit"

```
plot(ctree,type="simple")
```



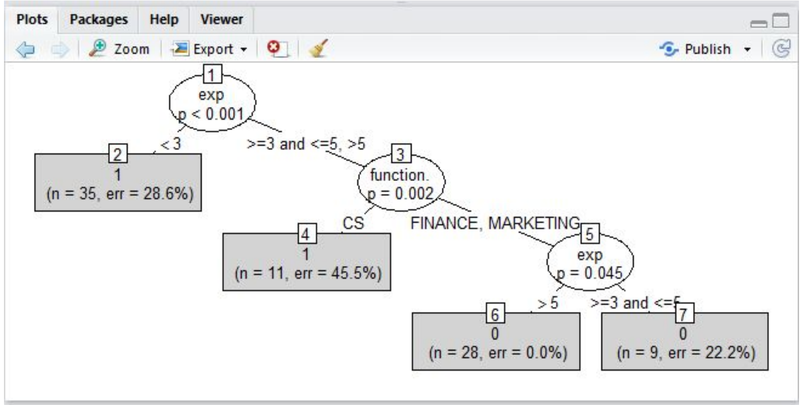
There are 35 employees with exp<3.
Out of which 25 Left within 18 months.
25/35=0.714 and Err=28.6%

Get an Edge!

- In case of large data, default tree plot may end up looking congested and difficult to interpret. Adjust the aesthetics of the tree plot for better results. Add argument **gp** (graphical parameter) in the **`plot()`** function.

```
plot(ctree,type="simple", gp=gpar(cex=0.8))
```

We have used **`gp=gpar()`** from package **`grid`** to decrease the text size



Data Snapshot

BANK LOAN

Column	Description	Type	Measurement	Possible Values
SN	Serial Number		-	-
AGE	Age Groups	Categorical	1(<28 years),2(28-40 years),3(>40 years)	3
EMPLOY	Number of years customer working at current employer	Continuous	-	Positive value
ADDRESS	Number of years customer staying at current address	Continuous	-	Positive value
DEBTINC	Debt to Income Ratio	Continuous	-	Positive value
CREDDEBT	Credit to Debit Ratio	Continuous	-	Positive value
OTHDEBT	Other Debt	Continuous	-	Positive value
DEFAULTER	Whether customer defaulted on loan	Binary	1(Defaulter), 0(Non-Defaulter)	2

Decision Tree for Continuous & Categorical Independent Variables

```
# ctree() for Continuous Independent Variables
```

```r
bankloan<-read.csv("BANK LOAN.csv",header=T)

str(bankloan)
```

- **str()** is used to check the structure of all variables.
- We convert DEFAULTER and AGE to factor variables using **as.factor()** as in our data these 2 variables are categorical.

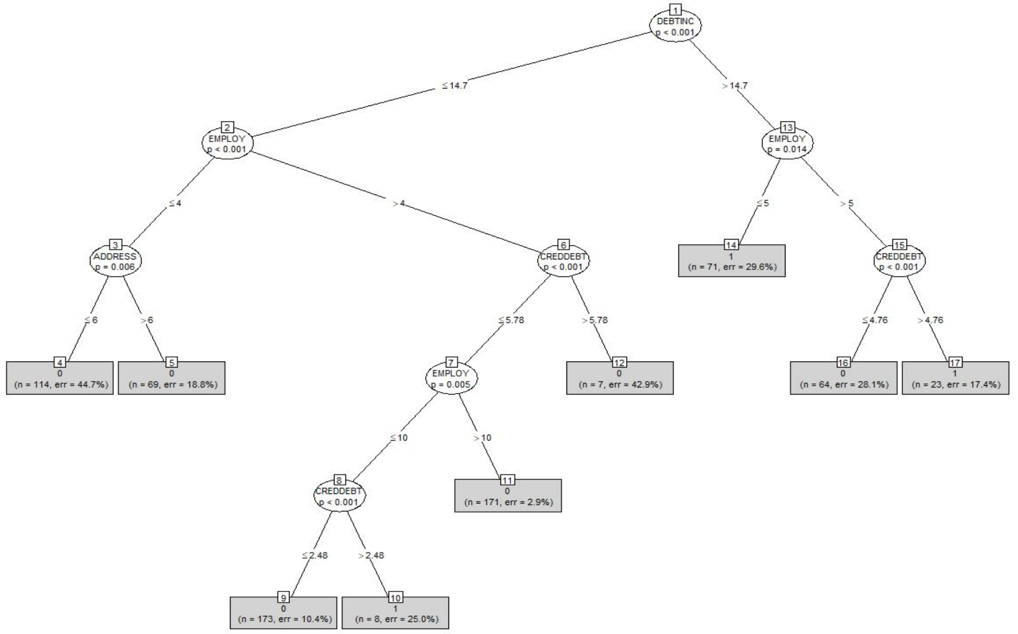
```
# Output
```

```
> str(bankloan)
'data.frame':   700 obs. of  8 variables:
 $ SN       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ AGE      : int  3 1 2 3 1 3 2 3 1 2 ...
 $ EMPLOY   : int  17 10 15 15 2 5 20 12 3 0 ...
 $ ADDRESS  : int  12 6 14 14 0 5 9 11 4 13 ...
 $ DEBTINC  : num  9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
 $ CREDDEBT : num  11.36 1.36 0.86 2.66 1.79 ...
 $ OTHDEBT  : num  5.01 4 2.17 0.82 3.06 ...
 $ DEFAULTER: int  1 0 0 0 1 0 0 0 1 0 ...
```

```r
bankloan$AGE<-as.factor(bankloan$AGE)
bankloan$DEFAULTER<-as.factor(bankloan$DEFAULTER)

bankctree<-partykit::ctree(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+
                           CREDDEBT+OTHDEBT, data=bankloan)
```

Decision Tree for Continuous & Categorical Independent Variables

```
plot(bankctree,type="simple", gp=gpar(cex=0.7))
```



Interpretation

- AGE and OTHDEBT do not appear in the tree.
- 114 customers with DEBTIC >14.7, employed for ≤ 5 years are mainly DEFAULTERS

Quick Recap

CI Tree

- **`partykit::ctree()`** in package "partykit" yields conditional inference trees for continuous & categorical independent variables