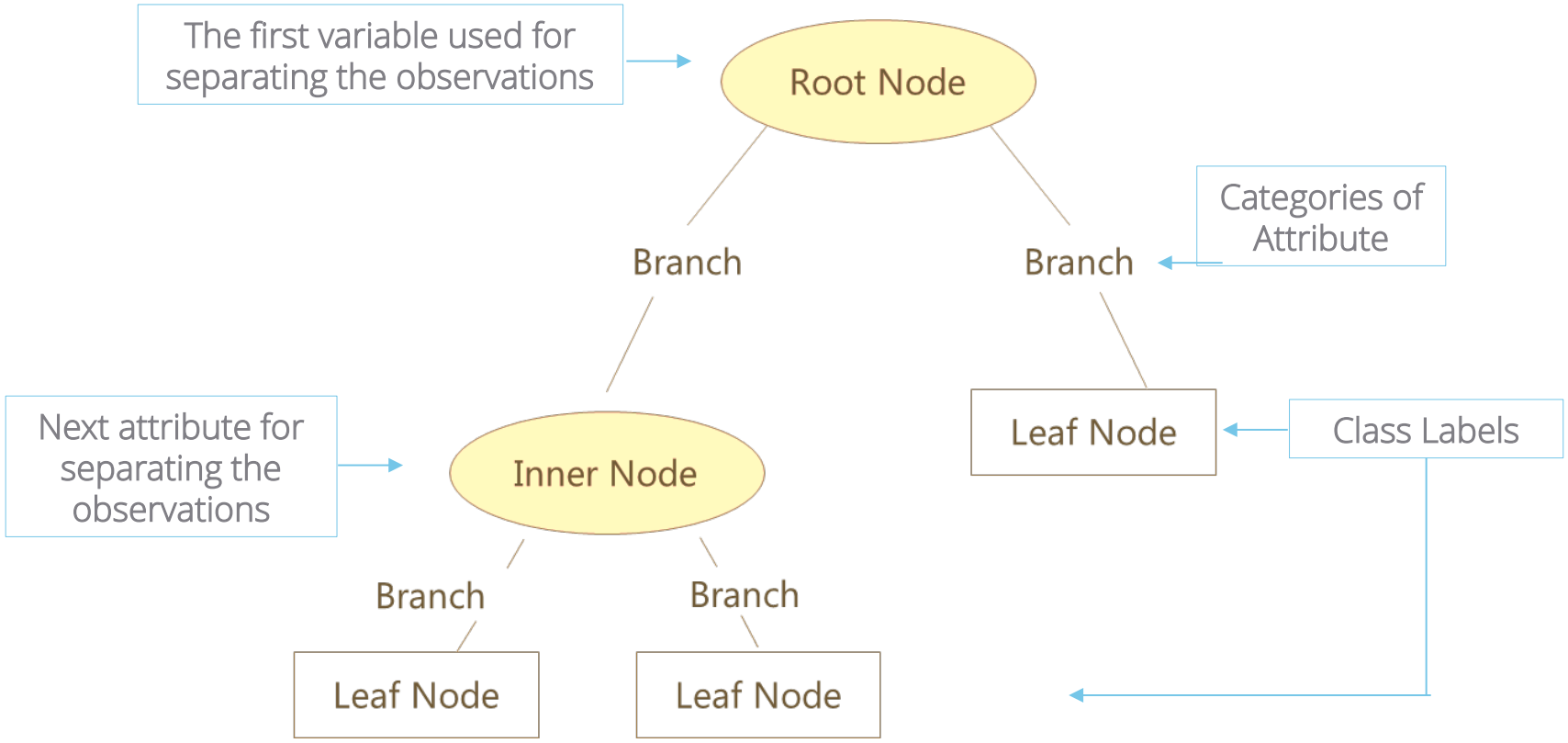
Decision Tree Method-WORKSHOP
ML ALGORITHM

Decision Tree – Basic Components

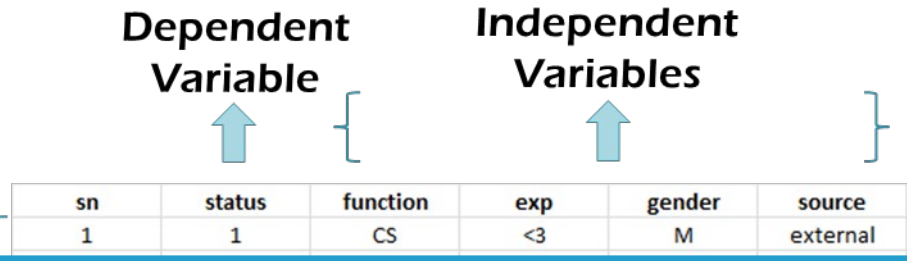


* Class labels show observations belong to which class. The leaf node also shows Number of observations and Error rate (Actual classification vs classification given by the tree)

DATA SCIENCE INSTITUTE

Data Snapshot

EMPLOYEE CHURN DATA

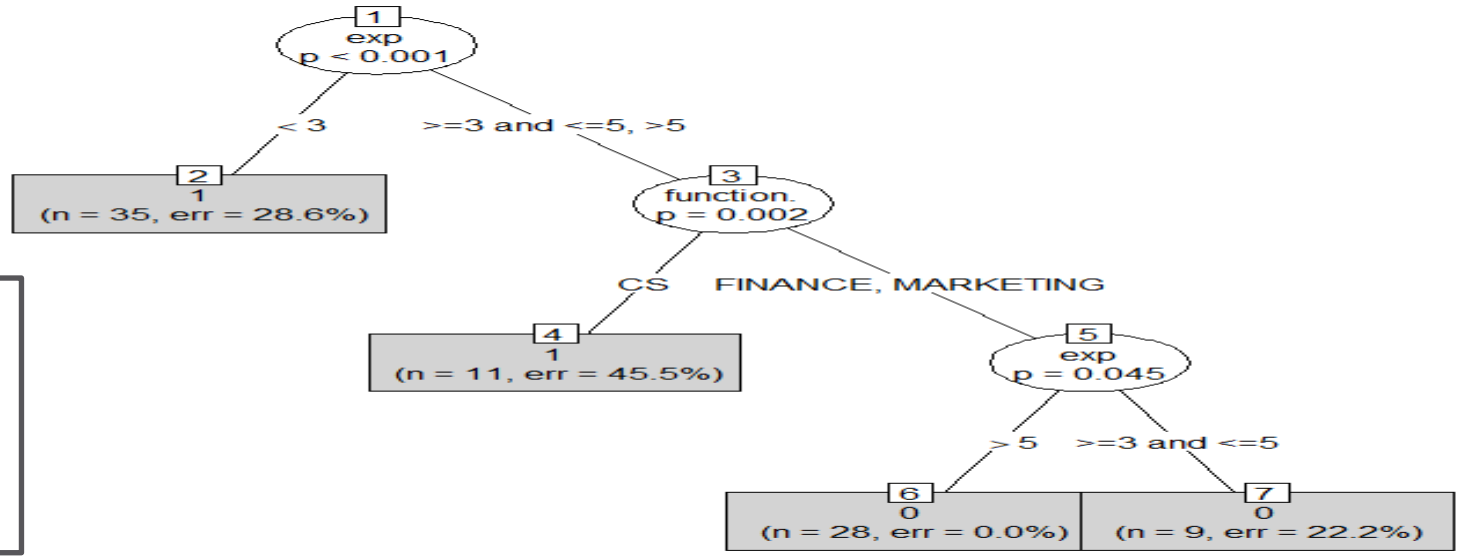


Columns	Description	Type	Measurement	Possible values
sn	Serial Number	-	-	-
status	= 1 If the Employee Left Within 18 Months of Joining = 0 Otherwise	Binary	1,0	2
function	Employee Job Profile	Categorical	CS, FINANCE, MARKETING	3
exp	Experience in Years	Categorical	<3,3-5,>5	3
gender	Gender of the Employee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

DATA SCIENCE INSTITUTE

Decision Tree in R....
CHAID-like Implementation in "partykit"

```
empdata<-read.csv(file.choose(),heade=T,stringsAsFactors =T)
install.packages("partykit")
library(partykit)
empdata$status<-as.factor(empdata$status) # classification problem
ctree<-partykit::ctree(formula=status~function.+exp+gender+source,
                       data=empdata)
plot(ctree,type="simple")
```

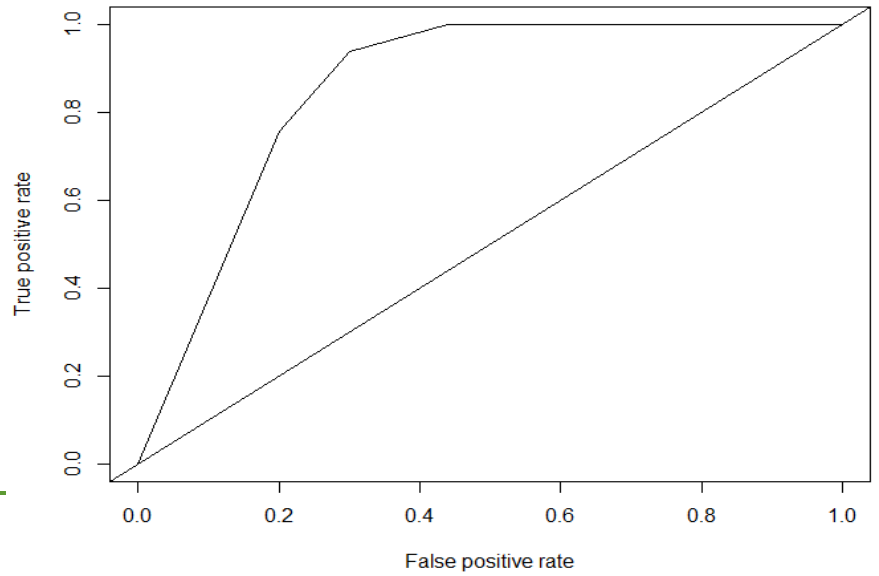


35 employees with exp<3.
Out of which 25 have Y=1 and 10 have Y=0
25/35=0.714

Decision Tree in R....
ROC Curve for "partykit" tree

```
predtree<-predict(ctree,empdata,type="prob")
library(ROCR)
pred<-prediction(predtree[,2],empdata$status)
perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)

## Area under ROC Curve in R (AUC)
auc<-performance(pred,"auc")
auc@y.values
```

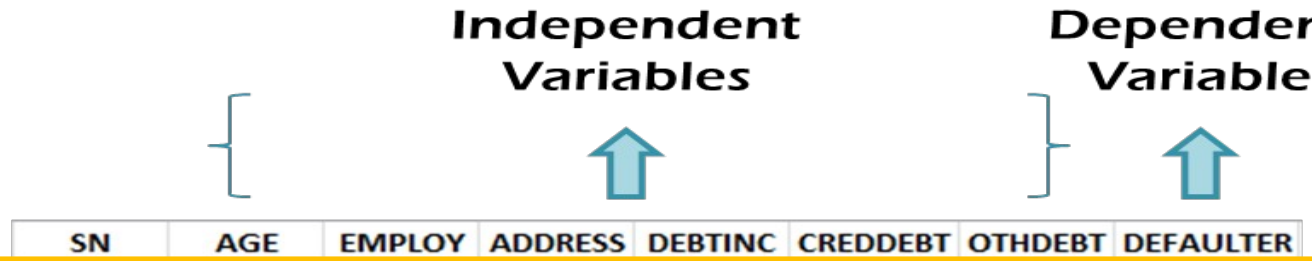



```
[[1]]
[1] 0.8563636
```

DATA SCIENCE INSTITUTE

BANK LOAN: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	Continuou	-	Positive value

DATA SCIENCE INSTITUTE

Decision Tree for Continuous & Categorical Independent Variables

```
# ctree() for Continuous Independent Variables

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 ...
```

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

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

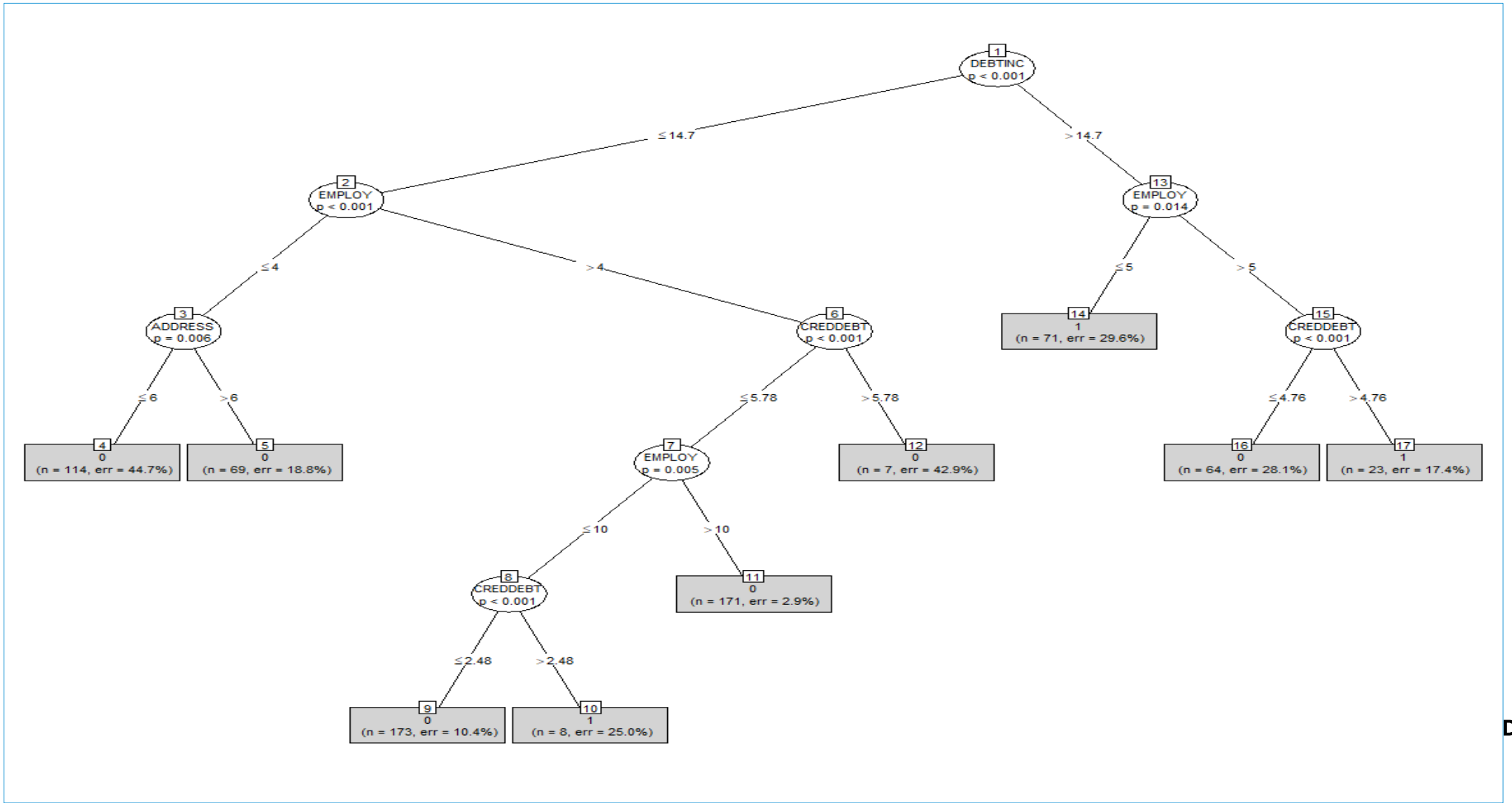
DATA SCIENCE INSTITUTE

Decision Tree for Continuous & Categorical Independent Variables

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

1: Defaulter

0: non-defaulter



DATA SCIENCE INSTITUTE

Note

- The “gp” argument is used to specify graphical parameters for customizing the appearance of the plot.
- These settings can control various aspects of the graphical output, such as colors, line types, line widths, font sizes, etc.
- For the “cex” parameter specifically, it controls the magnification of text and symbol sizes relative to the default size. The range for cex typically goes from 0 (which would effectively make text and symbols invisible) to any positive value greater than 0, where higher values increase the size of text and symbols.

Interpretation

- DEBTINC is the root node.
- This is a continuous variable. Ctree algorithm has automatically converted it into a categorical variable with 2 categories <=14.7 and >14.7.
- Customers with < 14.7 DEBTINC, and the number of years at the current employer are <=4, then

  a) if the number of years at the current address are <=6, there are 114 such customers.

  These customers are non-defaulters. The misclassification rate is 44.7%.
- b) if the number of years at the current address are > 6, there are 69 such customers.

  They should be classified as non-defaulters. The the misclassification rate is 18.8%.
- Out of 71 customers with DEBTINC > 14.7 and employed for <=5 years majority are defaulters.

- Out of 171 customers, with number of years with the current employer > 10, credit card date <=5.78 and

  Debt to income ratio <= 14.7, majority are non-defaulters.

- Age and OTHERDEBT do not appear in the Tree and hence are insignificant .

Logistic Regression in R

```
# Using glm function to develop binary logistic regression model

riskmodel<-glm(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT+OTHDEBT,
               family=binomial,data=data)
```

- **glm** is Generalized Linear Model. Logistic regression is type of GLM.
- LHS of ~ is dependent variable and independent variables on RHS are separated by '+'.
- **riskmodel** is the model object
- By setting the **family =binomial**, **glm()** fits a logistic regression model

Individual Hypothesis Testing in R

```
# Individual Testing
summary(riskmodel)

# Output:
```

- **summary()** function gives the output of glm.

```
Call:
glm(formula = DEFAULTER ~ AGE + EMPLOY + ADDRESS + DEBTINC +
    CREDDEBT + OTHDEBT, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3495  -0.6601  -0.2974   0.2509   2.8583

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78821    0.26407  -2.985  0.00284 **
AGE2         0.25202    0.26651   0.946  0.34433
AGE3         0.62707    0.36056   1.739  0.08201 .
EMPLOY      -0.26172    0.03188  -8.211  < 2e-16 ***
ADDRESS     -0.09964    0.02234  -4.459 8.22e-06 ***
DEBTINC      0.08506    0.02212   3.845  0.00012 ***
CREDDEBT     0.56336    0.08877   6.347 2.20e-10 ***
OTHDEBT      0.02315    0.05709   0.405  0.68517
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 553.41  on 692  degrees of freedom
AIC: 569.41

Number of Fisher Scoring iterations: 6
```

Interpretation :

- Since p-value is <0.05 for Employ, Address, Debtinc, Creddebt, these independent variables are statistically

DATA SCIENCE INSTITUTE

Quick Recap

CI Tree	• `partykit::ctree()` in package "partykit" yields conditional inference trees for continuous & categorical independent variables

THANK YOU!!

DATA SCIENCE
INSTITUTE