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 <- read\_csv(file.choose())

> head(Fraud\_check)

# A tibble: 6 x 6

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

<chr> <chr> <dbl> <dbl> <dbl> <chr>

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

#As per problem statement, treating those who have taxable\_income <= 30000 as "Risky" and others are "Good"

> Fraud\_check$Taxable.Income <- ifelse(Fraud\_check$Taxable.Income <=30000, yes = "Risky", no = "Good" )

> table(Fraud\_check$Taxable.Income)

Good Risky

476 124

> str(Fraud\_check)

Classes ‘spec\_tbl\_df’, ‘tbl\_df’, ‘tbl’ and 'data.frame': 600 obs. of 6 variables:

$ Undergrad : chr "NO" "YES" "NO" "YES" ...

$ Marital.Status : chr "Single" "Divorced" "Married" "Single" ...

$ Taxable.Income : chr "Good" "Good" "Good" "Good" ...

$ City.Population: num 50047 134075 160205 193264 27533 ...

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

$ Urban : chr "YES" "YES" "YES" "YES" ...

#Converting into dummy variables for chr columns.

> Fraud\_check$Undergrad <- ifelse(Fraud\_check$Undergrad=='YES',1,0)

> Fraud\_check$Urban <- ifelse(Fraud\_check$Urban=='YES',1,0)

> MS\_single <- ifelse(Fraud\_check$Marital.Status=="Single",1,0)

> MS\_divorced <- ifelse(Fraud\_check$Marital.Status=="Divorced",1,0)

> MS\_married <- ifelse(Fraud\_check$Marital.Status=="Married",1,0)

> Fr\_check <- Fraud\_check

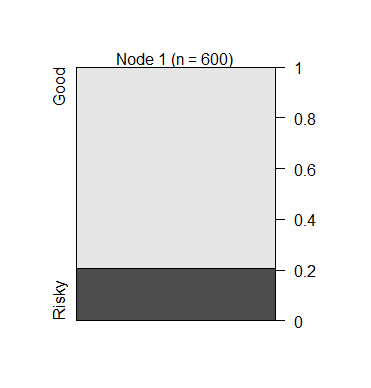
> Fr\_check <- cbind(Fr\_check,MS\_married,MS\_single,MS\_divorced)

#Build model using Party Library

> library(party)

> fc\_tree1 <- ctree(Fr\_check$Taxable.Income ~. , data = Fr\_check)

> plot(fc\_tree1)



#Build model using tree Library

> library(tree)

> fc\_tree <- tree(Fraud\_check$Taxable.Income ~. , data = Fraud\_check)

Warning messages:

1: In tree(Fraud\_check$Taxable.Income ~ ., data = Fraud\_check) :

NAs introduced by coercion

2: In tree(Fraud\_check$Taxable.Income ~ ., data = Fraud\_check) :

NAs introduced by coercion

> plot(fc\_tree)

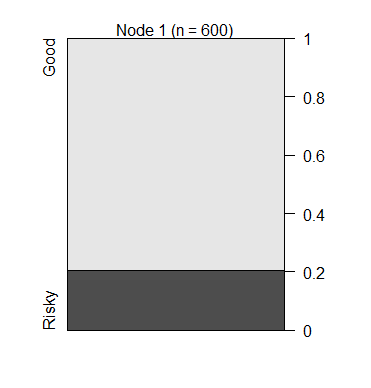
Error in plot.tree(fc\_tree) : cannot plot singlenode tree

#Build model using C50 Library

> library(C50)

> fc\_tree2 <- C5.0(Fr\_check[,-2],Fr\_check$Taxable.Income)

> plot(fc\_tree2)



#Build model using rpart Library

> library(rpart)

> fc\_tree3<- rpart(Fr\_check$Taxable.Income ~.,data= Fr\_check, method = 'anova')

> plot(fc\_tree3)

Error in plot.rpart(fc\_tree3) : fit is not a tree, just a root

As we can see from above 4 models, Due to lack of relevant information in data set we can conclude that unless and until a relevant variable is not introduced with this data, it won’t perform well. All the classification will be biased to one side.