NCD\_Cardiac

Capstone\_Hyderabad

Wednesday, July 05, 2017

### Load required libraries

library(caret)

## Warning: package 'caret' was built under R version 3.3.3

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.3.3

library(rpart)

## Warning: package 'rpart' was built under R version 3.3.3

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.3.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.3.2

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(ineq)

## Warning: package 'ineq' was built under R version 3.3.2

library(plyr)

## Warning: package 'plyr' was built under R version 3.3.2

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(scales)

## Warning: package 'scales' was built under R version 3.3.2

library(stringr)

## Warning: package 'stringr' was built under R version 3.3.3

library(rattle)

## Warning: package 'rattle' was built under R version 3.3.2

## Rattle: A free graphical interface for data mining with R.  
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

library(RColorBrewer)

## Warning: package 'RColorBrewer' was built under R version 3.3.2

library(lmtest)

## Warning: package 'lmtest' was built under R version 3.3.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.3.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(pscl)

## Warning: package 'pscl' was built under R version 3.3.3

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 3.3.3

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

## Classes and Methods for R developed in the

## Political Science Computational Laboratory

## Department of Political Science

## Stanford University

## Simon Jackman

## hurdle and zeroinfl functions by Achim Zeileis

require(caret)  
library(mlbench)

## Warning: package 'mlbench' was built under R version 3.3.3

library(ROSE)

## Warning: package 'ROSE' was built under R version 3.3.3

## Loaded ROSE 0.0-3

library(caTools)

## Warning: package 'caTools' was built under R version 3.3.3

### Load dataset

setwd("C:\\Material\\CAPStone\\Scripts")  
data <- read.csv("ncd\_latest\_062817.csv")  
  
source("required\_user\_functions.R")  
  
d\_data <- subset(data, Cardiac %in% c('No','Yes') & ALIVE...DEAD == 'ALIVE')  
  
data\_d <- subset(d\_data, select = c( "FINANCIAL.GROUP","FAMILY.HISTORY","STAPLE.FOOD","ALCOHOL","Known.H.o.Smoking","Cholestrol.High","PhysicalActivity","Obese","Diabetes","Cancer","Cardiac"))  
str(data\_d)

## 'data.frame': 9328 obs. of 11 variables:  
## $ FINANCIAL.GROUP : Factor w/ 3 levels "Higher Class",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ FAMILY.HISTORY : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ STAPLE.FOOD : Factor w/ 3 levels "Pizza","Rice",..: 3 3 3 1 3 3 3 3 3 3 ...  
## $ ALCOHOL : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 2 1 1 1 1 ...  
## $ Known.H.o.Smoking: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Cholestrol.High : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ PhysicalActivity : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Obese : num 20.7 24.1 22.6 19 24.4 ...  
## $ Diabetes : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Cancer : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Cardiac : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...

colnames(data\_d)[1] <- "Fin\_Group"   
colnames(data\_d)[2] <- "Family\_history"   
colnames(data\_d)[3] <- "Staple\_Food"   
colnames(data\_d)[4] <- "Alcohol"   
colnames(data\_d)[5] <- "Smoke"   
colnames(data\_d)[6] <- "High\_Cholestrol"   
colnames(data\_d)[7] <- "Phy\_Activity"   
colnames(data\_d)[8] <- "Obese"   
colnames(data\_d)[9] <- "Diabetes"   
colnames(data\_d)[10] <- "Cancer"   
colnames(data\_d)[11] <- "Cardiac"   
  
  
cardiac\_vector <- c("Fin\_Group","Family\_history","Staple\_Food","Alcohol","Smoke","High\_Cholestrol","Phy\_Activity","Obese","Diabetes","Cancer")  
   
  
data\_d$Obese <- ifelse(data\_d$Obese > 30,'Y','N')  
data\_d$Obese <- as.factor(data\_d$Obese)  
str(data\_d)

## 'data.frame': 9328 obs. of 11 variables:  
## $ Fin\_Group : Factor w/ 3 levels "Higher Class",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Family\_history : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Staple\_Food : Factor w/ 3 levels "Pizza","Rice",..: 3 3 3 1 3 3 3 3 3 3 ...  
## $ Alcohol : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 2 1 1 1 1 ...  
## $ Smoke : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ High\_Cholestrol: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Phy\_Activity : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Obese : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Diabetes : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Cancer : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Cardiac : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...

names(data\_d)

## [1] "Fin\_Group" "Family\_history" "Staple\_Food"   
## [4] "Alcohol" "Smoke" "High\_Cholestrol"  
## [7] "Phy\_Activity" "Obese" "Diabetes"   
## [10] "Cancer" "Cardiac"

nrow(data\_d)

## [1] 9328

count(data\_d, vars=c("Cardiac"))

## Warning: package 'bindrcpp' was built under R version 3.3.3

## # A tibble: 1 x 2  
## vars n  
## <chr> <int>  
## 1 Cardiac 9328

data\_d$Fin\_Group=as.numeric(factor(data\_d$Fin\_Group))-1  
data\_d$Family\_history=as.numeric(factor(data\_d$Family\_history))-1  
data\_d$Staple\_Food=as.numeric(factor(data\_d$Staple\_Food))-1  
data\_d$Alcohol=as.numeric(factor(data\_d$Alcohol))-1  
data\_d$Smoke=as.numeric(factor(data\_d$Smoke))-1  
data\_d$High\_Cholestrol=as.numeric(factor(data\_d$High\_Cholestrol))-1  
data\_d$Phy\_Activity=as.numeric(factor(data\_d$Phy\_Activity))-1  
data\_d$Obese=as.numeric(factor(data\_d$Obese))-1  
data\_d$Diabetes=as.numeric(factor(data\_d$Diabetes))-1  
data\_d$Cancer=as.numeric(factor(data\_d$Cancer))-1  
data\_d$Cardiac=as.numeric(factor(data\_d$Cardiac))  
data\_d$Cardiac=data\_d$Cardiac-1  
  
  
  
for (e in 1:length(cardiac\_vector)){  
eda\_functions("Cardiac",data\_d,data\_d[,"Cardiac"],data\_d[,e],cardiac\_vector[e])  
 }

##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Fin\_Group   
##   
## column 0 1 2  
## target   
## 0 1886 3072 1229  
## 1 855 1886 400  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1 2  
## 0 0.3048327 0.4965250 0.1986423  
## 1 0.2722063 0.6004457 0.1273480  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1 2  
## 0 0.6880700 0.6196047 0.7544506  
## 1 0.3119300 0.3803953 0.2455494  
##   
##   
## Performing Chi Square Test for Cardiac and Fin\_Group  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Fin\_Group  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Fin\_Group  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1 2  
## 0 1886 3072 1229  
## 1 855 1886 400  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 110.51, df = 2, p-value = 1.005e-24  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 855   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test  
##   
## data: tab1  
## X-squared = 110.51, df = 2, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.005183e-24   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Fin\_Group at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Family\_history   
##   
## column 0 1  
## target   
## 0 2716 3471  
## 1 1613 1528  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.4389850 0.5610150  
## 1 0.5135307 0.4864693  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.6273966 0.6943389  
## 1 0.3726034 0.3056611  
##   
##   
## Performing Chi Square Test for Cardiac and Family\_history  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Family\_history  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Family\_history  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 2716 3471  
## 1 1613 1528  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 46.55, df = 1, p-value = 8.935e-12  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 1528   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 46.25, df = 1, p-value = 1.041e-11  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.04092e-11   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Family\_history at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Staple\_Food   
##   
## column 0 1 2  
## target   
## 0 65 63 6059  
## 1 26 29 3086  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1 2  
## 0 0.010505899 0.010182641 0.979311460  
## 1 0.008277619 0.009232728 0.982489653  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1 2  
## 0 0.7142857 0.6847826 0.6625478  
## 1 0.2857143 0.3152174 0.3374522  
##   
##   
## Performing Chi Square Test for Cardiac and Staple\_Food  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Staple\_Food  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Staple\_Food  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1 2  
## 0 65 63 6059  
## 1 26 29 3086  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1.2724, df = 2, p-value = 0.5293  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 26   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test  
##   
## data: tab1  
## X-squared = 1.2724, df = 2, p-value = 0.5293  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.5292973   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are independent of Staple\_Food at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Alcohol   
##   
## column 0 1  
## target   
## 0 2556 3631  
## 1 1476 1665  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.4131243 0.5868757  
## 1 0.4699140 0.5300860  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.6339286 0.6856118  
## 1 0.3660714 0.3143882  
##   
##   
## Performing Chi Square Test for Cardiac and Alcohol  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Alcohol  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Alcohol  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 2556 3631  
## 1 1476 1665  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 27.378, df = 1, p-value = 1.673e-07  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 1476   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 27.147, df = 1, p-value = 1.885e-07  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.885113e-07   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Alcohol at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Smoke   
##   
## column 0 1  
## target   
## 0 3573 2614  
## 1 1680 1461  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.5775012 0.4224988  
## 1 0.5348615 0.4651385  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.6801828 0.6414724  
## 1 0.3198172 0.3585276  
##   
##   
## Performing Chi Square Test for Cardiac and Smoke  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Smoke  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Smoke  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 3573 2614  
## 1 1680 1461  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 15.397, df = 1, p-value = 8.714e-05  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 1461   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 15.224, df = 1, p-value = 9.548e-05  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 9.548498e-05   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Smoke at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs High\_Cholestrol   
##   
## column 0 1  
## target   
## 0 5632 555  
## 1 593 2548  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.91029578 0.08970422  
## 1 0.18879338 0.81120662  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.90473896 0.17885917  
## 1 0.09526104 0.82114083  
##   
##   
## Performing Chi Square Test for Cardiac and High\_Cholestrol  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of High\_Cholestrol  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of High\_Cholestrol  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 5632 555  
## 1 593 2548  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 4885, df = 1, p-value = 0  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 555   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 4882, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of High\_Cholestrol at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Phy\_Activity   
##   
## column 0 1  
## target   
## 0 2665 3522  
## 1 1459 1682  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.4307419 0.5692581  
## 1 0.4645018 0.5354982  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.6462173 0.6767871  
## 1 0.3537827 0.3232129  
##   
##   
## Performing Chi Square Test for Cardiac and Phy\_Activity  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Phy\_Activity  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Phy\_Activity  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 2665 3522  
## 1 1459 1682  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 9.627, df = 1, p-value = 0.001918  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 1459   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 9.4904, df = 1, p-value = 0.002065  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.002065477   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Phy\_Activity at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Obese   
##   
## column 0 1  
## target   
## 0 5310 877  
## 1 2718 423  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.8582512 0.1417488  
## 1 0.8653295 0.1346705  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.6614350 0.6746154  
## 1 0.3385650 0.3253846  
##   
##   
## Performing Chi Square Test for Cardiac and Obese  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Obese  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Obese  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 5310 877  
## 1 2718 423  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.8703, df = 1, p-value = 0.3509  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 423   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 0.81225, df = 1, p-value = 0.3675  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.3674571   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are independent of Obese at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Diabetes   
##   
## column 0 1  
## target   
## 0 5648 539  
## 1 2129 1012  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.91288185 0.08711815  
## 1 0.67780961 0.32219039  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.7262441 0.3475177  
## 1 0.2737559 0.6524823  
##   
##   
## Performing Chi Square Test for Cardiac and Diabetes  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Diabetes  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Diabetes  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 5648 539  
## 1 2129 1012  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 830.5, df = 1, p-value = 1.291e-182  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 539   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 828.76, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 3.016702e-182   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Diabetes at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.  
##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Cancer   
##   
## column 0 1  
## target   
## 0 22 6165  
## 1 613 2528  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target 0 1  
## 0 0.003555843 0.996444157  
## 1 0.195160777 0.804839223  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target 0 1  
## 0 0.03464567 0.70919130  
## 1 0.96535433 0.29080870  
##   
##   
## Performing Chi Square Test for Cardiac and Cancer  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cardiac are independent of Cancer  
##   
##   
## Alternative hypothesis : The relative proportions of Cardiac are dependent of Cancer  
##   
## Step 2: Formulate Analysis Plan   
##   
##   
## The analysis plan describes how to use sample data to accept or reject the null hypothesis.  
## Plan should have these elements:  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Use Chi Square test of independence  
## Significance level: We can use the standard 0.05 or 5% level  
## Test method: Chi Square test of independence  
## In our contingency table, we assume the expected frequency count for each cell of the table is at least 5.  
##   
## Step 3: Analyze Sample data   
##   
## column  
## target 0 1  
## 0 22 6165  
## 1 613 2528  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 9328   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1205.6, df = 1, p-value = 3.684e-264  
##   
##   
## If the counts are not large enough to use a Chi-square distribution,   
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE.  
## In our contingency table, if we do not have the expected frequency count for each cell of the table is at least 5,  
## we shall use the p-value based on simulation instead, including the argument simulate.p.value = TRUE  
## Minimum value 22   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 1202.6, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.668126e-263   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent of Cancer at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.

## un balanced data

table(data\_d$Cardiac)

##   
## 0 1   
## 6187 3141

### Using ROSE package Balanced the data

mydata1 <- ROSE(Cardiac ~ .,data=data\_d, seed = 123)$data  
  
table(mydata1$Cardiac)

##   
## 0 1   
## 4719 4609

### Split dataset into training dataset and testing dataset in the ratio 70 % : 30 %

set.seed(88)  
split <- sample.split(mydata1, SplitRatio = 0.70)  
d\_train <- subset(mydata1, split == TRUE)  
d\_test <- subset(mydata1, split == FALSE)

## CART

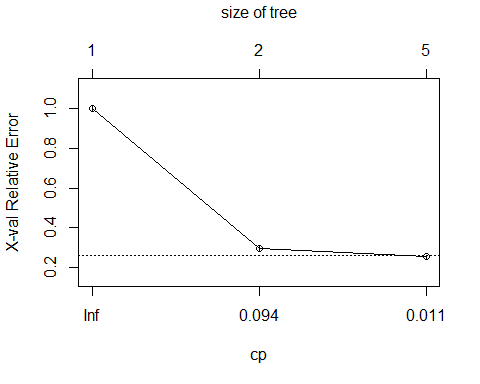
formula <- Cardiac ~ Fin\_Group + Family\_history + Staple\_Food + Alcohol + Smoke + High\_Cholestrol + Phy\_Activity + Obese + Diabetes + Cardiac  
title <- "Classification Tree for Cardiac"  
df <- d\_train  
fit<-NULL  
pfit <- cart\_fn(df, formula, title)

## Warning in model.matrix.default(attr(frame, "terms"), frame): the response  
## appeared on the right-hand side and was dropped

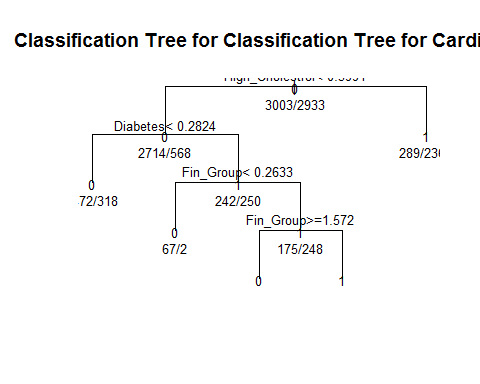
## Warning in model.matrix.default(attr(frame, "terms"), frame): problem with  
## term 10 in model.matrix: no columns are assigned

## Warning in cats \* (!isord): longer object length is not a multiple of  
## shorter object length

##   
## Classification tree:  
## rpart(formula = formula, data = d\_train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Diabetes Fin\_Group High\_Cholestrol  
##   
## Root node error: 2933/5936 = 0.4941  
##   
## n= 5936   
##   
## CP nsplit rel error xerror xstd  
## 1 0.707808 0 1.00000 1.00000 0.0131333  
## 2 0.012445 1 0.29219 0.29594 0.0092815  
## 3 0.010000 4 0.24651 0.25469 0.0087125



## Call:  
## rpart(formula = formula, data = d\_train, method = "class")  
## n= 5936   
##   
## CP nsplit rel error xerror xstd  
## 1 0.7078077 0 1.0000000 1.0000000 0.013133325  
## 2 0.0124446 1 0.2921923 0.2959427 0.009281521  
## 3 0.0100000 4 0.2465053 0.2546880 0.008712507  
##   
## Variable importance  
## High\_Cholestrol Diabetes Phy\_Activity Smoke   
## 61 18 8 6   
## Fin\_Group Family\_history   
## 5 1   
##   
## Node number 1: 5936 observations, complexity param=0.7078077  
## predicted class=0 expected loss=0.4941038 P(node) =1  
## class counts: 3003 2933  
## probabilities: 0.506 0.494   
## left son=2 (3282 obs) right son=3 (2654 obs)  
## Primary splits:  
## High\_Cholestrol < 0.3990847 to the left, improve=1513.12900, (0 missing)  
## Diabetes < 0.248414 to the left, improve= 334.84210, (0 missing)  
## Fin\_Group < 1.67346 to the right, improve= 28.51585, (0 missing)  
## Alcohol < 0.5936813 to the right, improve= 20.96015, (0 missing)  
## Family\_history < 0.5796423 to the right, improve= 15.79954, (0 missing)  
## Surrogate splits:  
## Diabetes < 0.3233981 to the left, agree=0.647, adj=0.210, (0 split)  
## Phy\_Activity < 0.4414039 to the right, agree=0.612, adj=0.132, (0 split)  
## Smoke < 0.4745713 to the left, agree=0.597, adj=0.099, (0 split)  
## Fin\_Group < 0.2342057 to the right, agree=0.570, adj=0.038, (0 split)  
## Family\_history < -0.02066023 to the right, agree=0.562, adj=0.021, (0 split)  
##   
## Node number 2: 3282 observations, complexity param=0.0124446  
## predicted class=0 expected loss=0.1730652 P(node) =0.5528976  
## class counts: 2714 568  
## probabilities: 0.827 0.173   
## left son=4 (2790 obs) right son=5 (492 obs)  
## Primary splits:  
## Diabetes < 0.2823959 to the left, improve=129.95330, (0 missing)  
## Phy\_Activity < 0.4953202 to the left, improve= 66.13164, (0 missing)  
## Smoke < 0.5216231 to the right, improve= 49.50656, (0 missing)  
## Fin\_Group < 0.3753933 to the left, improve= 49.02509, (0 missing)  
## Alcohol < 0.5473626 to the right, improve= 24.31715, (0 missing)  
## Surrogate splits:  
## Fin\_Group < 2.858901 to the left, agree=0.851, adj=0.004, (0 split)  
## Family\_history < 1.761441 to the left, agree=0.851, adj=0.004, (0 split)  
##   
## Node number 3: 2654 observations  
## predicted class=1 expected loss=0.1088922 P(node) =0.4471024  
## class counts: 289 2365  
## probabilities: 0.109 0.891   
##   
## Node number 4: 2790 observations  
## predicted class=0 expected loss=0.1139785 P(node) =0.4700135  
## class counts: 2472 318  
## probabilities: 0.886 0.114   
##   
## Node number 5: 492 observations, complexity param=0.0124446  
## predicted class=1 expected loss=0.4918699 P(node) =0.0828841  
## class counts: 242 250  
## probabilities: 0.492 0.508   
## left son=10 (69 obs) right son=11 (423 obs)  
## Primary splits:  
## Fin\_Group < 0.2632526 to the left, improve=36.849960, (0 missing)  
## Smoke < 0.6358504 to the right, improve=14.041140, (0 missing)  
## Phy\_Activity < 0.1864159 to the left, improve= 8.578750, (0 missing)  
## High\_Cholestrol < -0.1558211 to the right, improve= 7.958438, (0 missing)  
## Diabetes < 0.3795173 to the left, improve= 6.458416, (0 missing)  
## Surrogate splits:  
## Smoke < 1.393413 to the right, agree=0.866, adj=0.043, (0 split)  
## Staple\_Food < 2.252142 to the right, agree=0.864, adj=0.029, (0 split)  
##   
## Node number 10: 69 observations  
## predicted class=0 expected loss=0.02898551 P(node) =0.01162399  
## class counts: 67 2  
## probabilities: 0.971 0.029   
##   
## Node number 11: 423 observations, complexity param=0.0124446  
## predicted class=1 expected loss=0.4137116 P(node) =0.07126011  
## class counts: 175 248  
## probabilities: 0.414 0.586   
## left son=22 (131 obs) right son=23 (292 obs)  
## Primary splits:  
## Fin\_Group < 1.571571 to the right, improve=38.649810, (0 missing)  
## High\_Cholestrol < -0.1468522 to the right, improve= 9.020696, (0 missing)  
## Diabetes < 1.250126 to the left, improve= 7.421649, (0 missing)  
## Phy\_Activity < -0.04462664 to the left, improve= 6.381338, (0 missing)  
## Smoke < 0.06240359 to the right, improve= 4.780514, (0 missing)  
## Surrogate splits:  
## Alcohol < -0.5310228 to the left, agree=0.697, adj=0.023, (0 split)  
## Obese < 1.057082 to the right, agree=0.693, adj=0.008, (0 split)  
##   
## Node number 22: 131 observations  
## predicted class=0 expected loss=0.2671756 P(node) =0.02206873  
## class counts: 96 35  
## probabilities: 0.733 0.267   
##   
## Node number 23: 292 observations  
## predicted class=1 expected loss=0.2705479 P(node) =0.04919137  
## class counts: 79 213  
## probabilities: 0.271 0.729

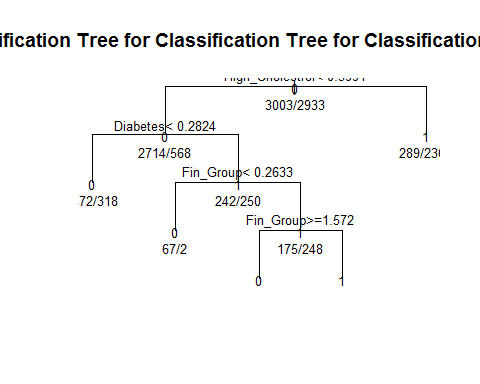


## Warning in model.matrix.default(Terms, m, contrasts): the response appeared  
## on the right-hand side and was dropped

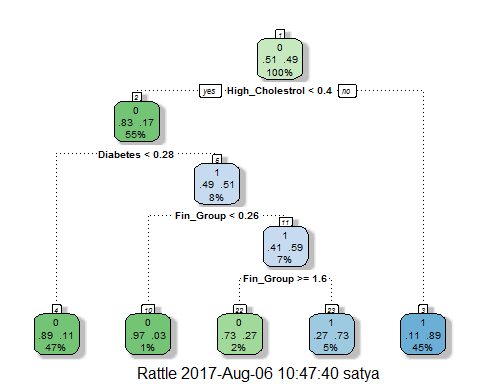
## Warning in model.matrix.default(Terms, m, contrasts): problem with term 10  
## in model.matrix: no columns are assigned

## Warning in train.default(x, y, weights = w, ...): You are trying to do  
## regression and your outcome only has two possible values Are you trying to  
## do classification? If so, use a 2 level factor as your outcome column.

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =  
## trainInfo, : There were missing values in resampled performance measures.



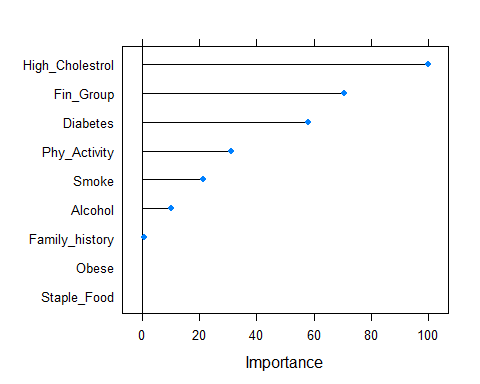
fancyRpartPlot(pfit)



varImp(object=fit)

## rpart variable importance  
##   
## Overall  
## High\_Cholestrol 100.0000  
## Fin\_Group 70.6780  
## Diabetes 58.1296  
## Phy\_Activity 31.2758  
## Smoke 21.5541  
## Alcohol 10.2609  
## Family\_history 0.8622  
## Staple\_Food 0.0000  
## Obese 0.0000

plot(varImp(fit))



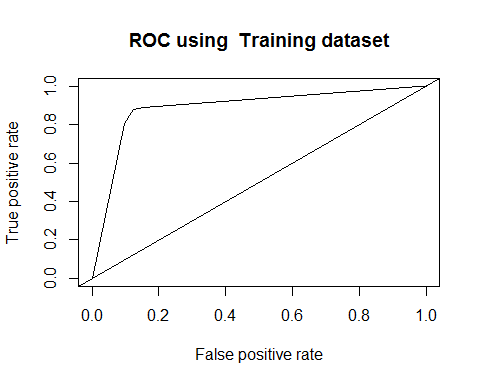
#### Interpretation of Variable Importance for Crdiac

###### Above graph shows the top 10 important variables.high cholestrol,Cancer and diabetes are the top influencing predictors.

###### Physical activity,smoking,Financial group and Alcohol comsumption also have significant importance.

### Performance measures using training data

target <- d\_train[,'Cardiac']  
title <- 'Training dataset'  
perf\_measures1 <- perf\_measures\_cart(d\_train,pfit,target,title)



print(perf\_measures1)

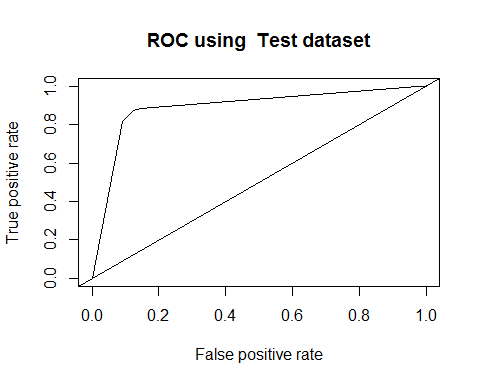
## Dataset KS auc gini OA  
## 1 Training dataset 0.7564194 0.8895511 0.3941448 0.8782008

cat("\n OA is overall accuracy \n")

##   
## OA is overall accuracy

### Performance measures using test data

target <- d\_test[,'Cardiac']  
title <- 'Test dataset'  
perf\_measures2 <- perf\_measures\_cart(d\_test,pfit,target,title)



print(perf\_measures2)

## Dataset KS auc gini OA  
## 1 Test dataset 0.7517413 0.890316 0.3960324 0.8758844

cat("\n OA is overall accuracy \n")

##   
## OA is overall accuracy

### Logistic Regression

strTarget<-"Cardiac"  
formula <- Cardiac ~ Fin\_Group + Family\_history +   
 Staple\_Food + Alcohol + Smoke + High\_Cholestrol + Phy\_Activity + Obese + Diabetes + Cancer  
logit=logistic(formula,d\_train)  
  
summary(logit)

##   
## Call:  
## glm(formula = formula, family = "binomial", data = d\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1203 -0.2925 -0.1055 0.3633 2.4457   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.29792 0.43676 5.261 1.43e-07 \*\*\*  
## Fin\_Group -0.14584 0.06886 -2.118 0.034181 \*   
## Family\_history -0.14808 0.09093 -1.629 0.103392   
## Staple\_Food 0.22496 0.19646 1.145 0.252170   
## Alcohol -0.29613 0.08991 -3.294 0.000989 \*\*\*  
## Smoke -0.05872 0.09525 -0.617 0.537553   
## High\_Cholestrol 5.74171 0.13847 41.466 < 2e-16 \*\*\*  
## Phy\_Activity 0.28604 0.09731 2.940 0.003287 \*\*   
## Obese -0.16826 0.12718 -1.323 0.185833   
## Diabetes 0.48186 0.11373 4.237 2.27e-05 \*\*\*  
## Cancer -5.78290 0.18769 -30.811 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8228.2 on 5935 degrees of freedom  
## Residual deviance: 2861.2 on 5925 degrees of freedom  
## AIC: 2883.2  
##   
## Number of Fisher Scoring iterations: 6

#### Interpretation of Variable Importance for cardiac based on Coefficients table shown above

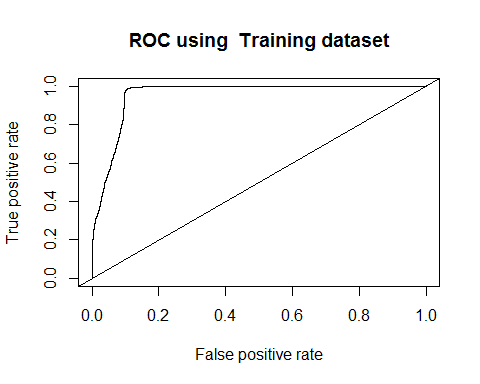
###### Smoking is the only significance predicator.Rest are not having any importance at significance level of 95%

perf <- pR2(logit)  
print(perf)

## llh llhNull G2 McFadden r2ML   
## -1430.5996257 -4114.1089184 5367.0185854 0.6522699 0.5951130   
## r2CU   
## 0.7935208

### Logistic Regression Performance measures using training data

target <- d\_train[,'Cardiac']  
title <- 'Training dataset'  
  
perf\_measures1 <- perf\_measures\_logistic(d\_train,logit,target,title)



print(perf\_measures1)

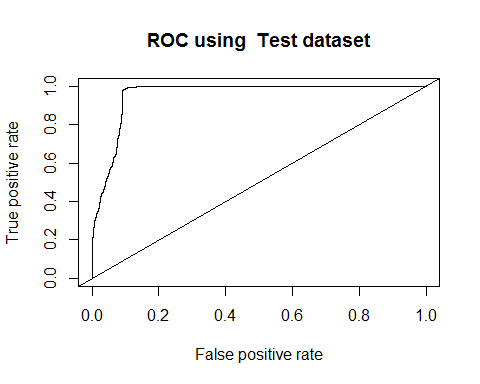
## Dataset auc McFadden OA  
## 1 Training dataset 0.9543515 0.1732045 0.5

cat("\n OA is overall accuracy \n")

##   
## OA is overall accuracy

### Logistic Regression Performance measures using test data

target <- d\_test[,'Cardiac']  
title <- 'Test dataset'  
perf\_measures2 <- perf\_measures\_logistic(d\_test,logit,target,title)



print(perf\_measures2)

## Dataset auc McFadden OA  
## 1 Test dataset 0.9556174 0.1732045 0.5

cat("\n OA is overall accuracy \n")

##   
## OA is overall accuracy

varImp(logit, scale = FALSE)

## Overall  
## Fin\_Group 2.117935  
## Family\_history 1.628629  
## Staple\_Food 1.145095  
## Alcohol 3.293652  
## Smoke 0.616518  
## High\_Cholestrol 41.465682  
## Phy\_Activity 2.939521  
## Obese 1.323006  
## Diabetes 4.236850  
## Cancer 30.811292

### End of R script

```