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

### 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

### Split the data into two sets each containing 1500 records of Cardiac and Not Cardiac

d\_d <- subset(data\_d,Cardiac == 'Yes')  
d\_n <- subset(data\_d,Cardiac == 'No')  
#  
NCD <- "Cardiac"  
displaySplitCount(d\_d$Cardiac,"1",NCD)

##   
## Dataset file 1 Cardiac Count :

## dd  
## No Yes   
## 0 3141

displaySplitCount(d\_n$Cardiac,"2",NCD)

##   
## Dataset file 2 Cardiac Count :

## dd  
## No Yes   
## 6187 0

set.seed(123)  
  
smp\_size1 = min(1000,nrow(d\_d))  
samp\_ind1 <- sample(seq\_len(nrow(d\_d)), size = smp\_size1)  
d1 <- d\_d[samp\_ind1,]  
  
smp\_size2 = min(1000,nrow(d\_n))  
samp\_ind2 <- sample(seq\_len(nrow(d\_n)), size = smp\_size2)  
d2 <- d\_n[samp\_ind2,]  
  
print(head(d1))

## Fin\_Group Family\_history Staple\_Food Alcohol Smoke  
## 3502 Lower Middle Class Yes Roti Yes Yes  
## 7726 Lower Middle Class Yes Roti Yes Yes  
## 4557 Lower Middle Class Yes Roti Yes No  
## 8543 Middle Class No Roti Yes Yes  
## 9046 Lower Middle Class Yes Roti Yes No  
## 2159 Lower Middle Class Yes Roti Yes No  
## High\_Cholestrol Phy\_Activity Obese Diabetes Cancer Cardiac  
## 3502 Yes Yes Y No Yes Yes  
## 7726 Yes No N Yes Yes Yes  
## 4557 Yes No N No Yes Yes  
## 8543 Yes No N Yes Yes Yes  
## 9046 Yes Yes N No Yes Yes  
## 2159 No Yes N Yes No Yes

print(nrow(d1))

## [1] 1000

print(head(d2))

## Fin\_Group Family\_history Staple\_Food Alcohol Smoke  
## 1925 Middle Class Yes Pizza No No  
## 6621 Higher Class No Roti Yes Yes  
## 1131 Lower Middle Class Yes Roti Yes No  
## 9785 Middle Class No Roti Yes No  
## 9741 Higher Class No Roti No Yes  
## 5052 Higher Class Yes Roti No Yes  
## High\_Cholestrol Phy\_Activity Obese Diabetes Cancer Cardiac  
## 1925 No Yes N Yes Yes No  
## 6621 No No N No Yes No  
## 1131 No Yes N No Yes No  
## 9785 No No N No Yes No  
## 9741 No No N No Yes No  
## 5052 No Yes N No Yes No

print(nrow(d2))

## [1] 1000

displaySplitCount(d1$Cardiac,"1",NCD)

##   
## Dataset file 1 Cardiac Count :

## dd  
## No Yes   
## 0 1000

displaySplitCount(d2$Cardiac,"2",NCD)

##   
## Dataset file 2 Cardiac Count :

## dd  
## No Yes   
## 1000 0

### Split dataset d1 into training dataset and testing dataset in the ratio 70 % : 30 %

set.seed(123)  
rnd <- sort(sample(nrow(d1),nrow(d1)\*.7))  
d1\_train <- d1[rnd,]  
d1\_test <- d1[-rnd,]  
  
print(nrow(d1\_train))

## [1] 700

print(nrow(d1\_test))

## [1] 300

### Split dataset d2 into training dataset and testing dataset

d2\_train <- d2[rnd,]  
d2\_test <- d2[-rnd,]  
  
print(nrow(d2\_train))

## [1] 700

print(nrow(d2\_test))

## [1] 300

d2\_train <- d2\_train[complete.cases(d2\_train),]

### Merge d1\_train (Cardiac dataset) and d2\_train (Non-Cardiac dataset) to form training dataset

d\_train <- bind\_rows(d1\_train,d2\_train)  
d\_test <- bind\_rows(d1\_test,d2\_test)  
  
print(nrow(d\_train))

## [1] 1400

print(nrow(d\_test))

## [1] 600

### Check the dataset the proportion of Cardiac and Non-Cardiac

Yes1\_train <- length(which(str\_trim(d1\_train$Cardiac) == 'Yes'))  
No1\_train <- length(which(str\_trim(d1\_train$Cardiac) == 'No'))  
  
cat("\n Cardiac Count Yes: ",Yes1\_train," No: ",No1\_train)

##   
## Cardiac Count Yes: 700 No: 0

Yes2\_train <- length(which(str\_trim(d2\_train$Cardiac) == 'Yes'))  
No2\_train <- length(which(str\_trim(d2\_train$Cardiac) == 'No'))  
  
cat("\n Cardiac Count Yes: ",Yes2\_train," No: ",No2\_train)

##   
## Cardiac Count Yes: 0 No: 700

Yes1\_test <- length(which(str\_trim(d1\_test$Cardiac) == 'Yes'))  
No1\_test <- length(which(str\_trim(d1\_test$Cardiac) == 'No'))  
  
cat("\n Cardiac Count Yes: ",Yes1\_test," No: ",No1\_test)

##   
## Cardiac Count Yes: 300 No: 0

Yes2\_test <- length(which(str\_trim(d2\_test$Cardiac) == 'Yes'))  
No2\_test <- length(which(str\_trim(d2\_test$Cardiac) == 'No'))  
  
###  
  
cat("\n Train dataset file Cardiac Count : ",nrow(d\_train))

##   
## Train dataset file Cardiac Count : 1400

table(d\_train$Cardiac)

##   
## No Yes   
## 700 700

cat("\n Test dataset file Cardiac Count : ",nrow(d\_test))

##   
## Test dataset file Cardiac Count : 600

table(d\_test$Cardiac)

##   
## No Yes   
## 300 300

### Remove unwanted variables

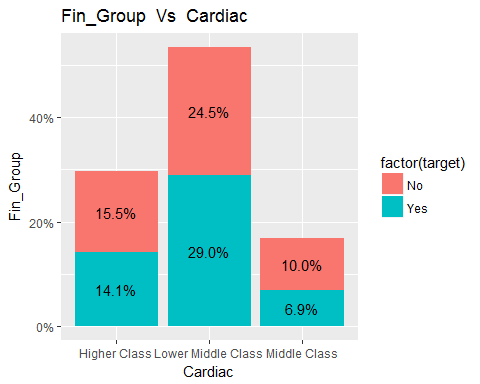
#### The following columns are not required for Diabetes - NCD:

1. "S.NO."
2. "Year"
3. "ALIVE...DEAD"
4. "RELIGION"
5. "Random.Blood.Sugar"
6. "HBA1C"
7. "Specialty.of.Treatment"
8. "Height"
9. "Weight"

for (e in 1:length(cardiac\_vector)){  
eda\_functions("Cardiac",d\_train,d\_train[,"Cardiac"],d\_train[,e],cardiac\_vector[e])  
 }

##   
##   
## Performing mlbenchxtab   
##   
## Cardiac Vs Fin\_Group   
##   
## column Higher Class Lower Middle Class Middle Class  
## target   
## No 217 343 140  
## Yes 198 406 96  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target Higher Class Lower Middle Class Middle Class  
## No 0.3100000 0.4900000 0.2000000  
## Yes 0.2828571 0.5800000 0.1371429  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target Higher Class Lower Middle Class Middle Class  
## No 0.5228916 0.4579439 0.5932203  
## Yes 0.4771084 0.5420561 0.4067797  
##   
##   
## 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 Higher Class Lower Middle Class Middle Class  
## No 217 343 140  
## Yes 198 406 96  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 14.372, df = 2, p-value = 0.000757  
##   
##   
## 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 198   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test  
##   
## data: tab1  
## X-squared = 14.372, df = 2, p-value = 0.000757  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.0007569848   
##   
##   
## <-- 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 No Yes  
## target   
## No 313 387  
## Yes 348 352  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.4471429 0.5528571  
## Yes 0.4971429 0.5028571  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.4735250 0.5236806  
## Yes 0.5264750 0.4763194  
##   
##   
## 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 No Yes  
## No 313 387  
## Yes 348 352  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 3.511, df = 1, p-value = 0.06097  
##   
##   
## 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 313   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 3.3131, df = 1, p-value = 0.06873  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.06872799   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are independent 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 Pizza Rice Roti  
## target   
## No 9 5 686  
## Yes 10 8 682  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target Pizza Rice Roti  
## No 0.012857143 0.007142857 0.980000000  
## Yes 0.014285714 0.011428571 0.974285714  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target Pizza Rice Roti  
## No 0.4736842 0.3846154 0.5014620  
## Yes 0.5263158 0.6153846 0.4985380  
##   
##   
## 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 Pizza Rice Roti  
## No 9 5 686  
## Yes 10 8 682  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.7566, df = 2, p-value = 0.685  
##   
##   
## 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 5   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test  
##   
## data: tab1  
## X-squared = 0.75664, df = 2, p-value = 0.685  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.6850129   
##   
##   
## <-- 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 No Yes  
## target   
## No 270 430  
## Yes 344 356  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.3857143 0.6142857  
## Yes 0.4914286 0.5085714  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.4397394 0.5470738  
## Yes 0.5602606 0.4529262  
##   
##   
## 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 No Yes  
## No 270 430  
## Yes 344 356  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 15.885, df = 1, p-value = 6.729e-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 270   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 15.459, df = 1, p-value = 8.431e-05  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 8.431219e-05   
##   
##   
## <-- 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 No Yes  
## target   
## No 390 310  
## Yes 376 324  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.5571429 0.4428571  
## Yes 0.5371429 0.4628571  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.5091384 0.4889590  
## Yes 0.4908616 0.5110410  
##   
##   
## 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 No Yes  
## No 390 310  
## Yes 376 324  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.565, df = 1, p-value = 0.4522  
##   
##   
## 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 310   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 0.48719, df = 1, p-value = 0.4852  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.4851843   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are independent 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 No Yes  
## target   
## No 627 73  
## Yes 129 571  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.8957143 0.1042857  
## Yes 0.1842857 0.8157143  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.8293651 0.1133540  
## Yes 0.1706349 0.8866460  
##   
##   
## 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 No Yes  
## No 627 73  
## Yes 129 571  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 713.1, df = 1, p-value = 4.138e-157  
##   
##   
## 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 73   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 710.29, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.733828e-156   
##   
##   
## <-- 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 No Yes  
## target   
## No 301 399  
## Yes 332 368  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.4300000 0.5700000  
## Yes 0.4742857 0.5257143  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.4755134 0.5202086  
## Yes 0.5244866 0.4797914  
##   
##   
## 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 No Yes  
## No 301 399  
## Yes 332 368  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 2.7711, df = 1, p-value = 0.09598  
##   
##   
## 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 301   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 2.5952, df = 1, p-value = 0.1072  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.1071876   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are independent 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 N Y  
## target   
## No 569 131  
## Yes 599 101  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target N Y  
## No 0.8128571 0.1871429  
## Yes 0.8557143 0.1442857  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target N Y  
## No 0.4871575 0.5646552  
## Yes 0.5128425 0.4353448  
##   
##   
## 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 N Y  
## No 569 131  
## Yes 599 101  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 4.65, df = 1, p-value = 0.03106  
##   
##   
## 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 101   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 4.345, df = 1, p-value = 0.03712  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.03711702   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cardiac are dependent 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 No Yes  
## target   
## No 630 70  
## Yes 474 226  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.9000000 0.1000000  
## Yes 0.6771429 0.3228571  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.5706522 0.2364865  
## Yes 0.4293478 0.7635135  
##   
##   
## 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 No Yes  
## No 630 70  
## Yes 474 226  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 104.26, df = 1, p-value = 1.775e-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 70   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 102.93, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 3.476727e-24   
##   
##   
## <-- 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 No Yes  
## target   
## No 3 697  
## Yes 134 566  
##   
##   
## Within each level of predictor variable values, % of Cardiac patients   
##   
## column  
## target No Yes  
## No 0.004285714 0.995714286  
## Yes 0.191428571 0.808571429  
##   
##   
## Cardiac Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.02189781 0.55186065  
## Yes 0.97810219 0.44813935  
##   
##   
## 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 No Yes  
## No 3 697  
## Yes 134 566  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 138.85, df = 1, p-value = 4.749e-32  
##   
##   
## 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 3   
##   
## <------ Chi square with simulation for p value --------->   
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##   
## data: tab1  
## X-squared = 138.85, df = NA, p-value = 0.0004998  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained with simulation for p-value : 0.0004997501   
##   
##   
## <-- 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.

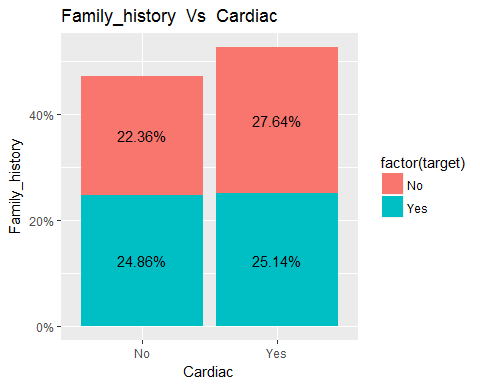
printggPlot2(d\_train,d\_train$Fin\_Group,d\_train$Cardiac,colnames(data\_d)[1],colnames(data\_d)[11])



### Interpretation

#### Lower Middle class people are more prone to Cardiac (29%)

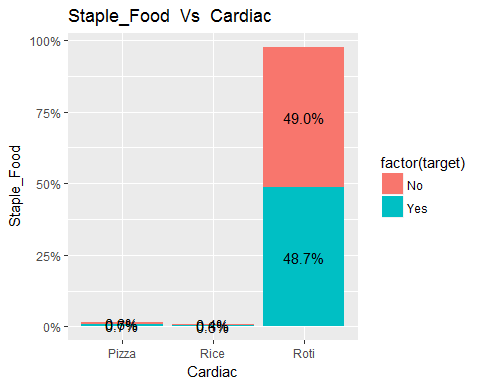
printggPlot2(d\_train,d\_train$Family\_history,d\_train$Cardiac,colnames(data\_d)[2],colnames(data\_d)[11])



### Interpretation

#### People having family history of Cardiac are more prone to Cardiac(25.14%)

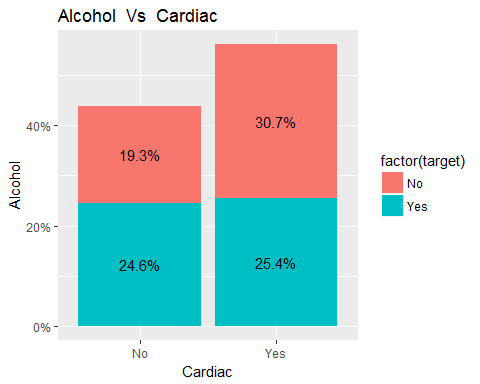
printggPlot2(d\_train,d\_train$Staple\_Food,d\_train$Cardiac,colnames(data\_d)[3],colnames(data\_d)[11])



### Interpretation

#### People having Staple Food are more prone to Cardiac(48.7%)

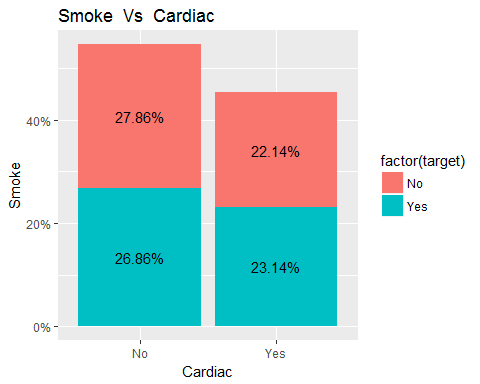
printggPlot2(d\_train,d\_train$Alcohol,d\_train$Cardiac,colnames(data\_d)[4],colnames(data\_d)[11])



### Interpretation

#### People having Alcohol habit are more prone to Cardiac(25.4%)

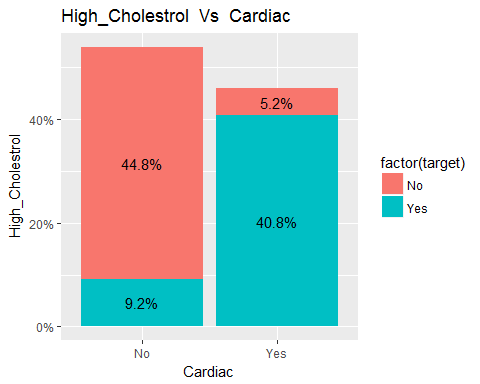
printggPlot2(d\_train,d\_train$Smoke,d\_train$Cardiac,colnames(data\_d)[5],colnames(data\_d)[11])



### Interpretation

#### People having Smoking habit are more prone to Cardiac(23.14%)

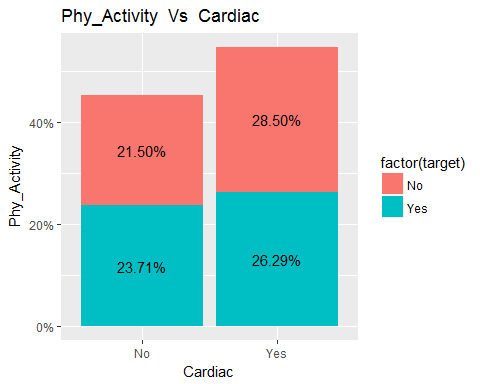
printggPlot2(d\_train,d\_train$High\_Cholestrol,d\_train$Cardiac,colnames(data\_d)[6],colnames(data\_d)[11])



### Interpretation

#### People having High\_Cholestrol are more prone to Cardiac(40.8%)

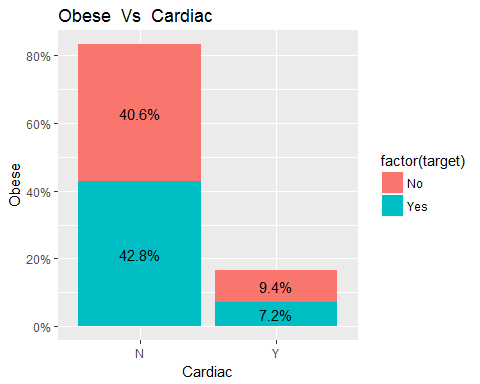
printggPlot2(d\_train,d\_train$Phy\_Activity,d\_train$Cardiac,colnames(data\_d)[7],colnames(data\_d)[11])



### Interpretation

#### People not having Phy\_Activity are more prone to Cardiac(23.71%)

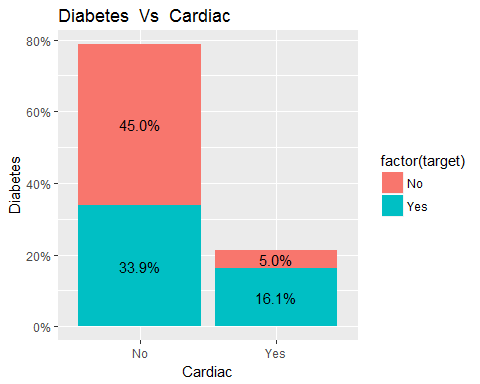
printggPlot2(d\_train,d\_train$Obese,d\_train$Cardiac,colnames(data\_d)[8],colnames(data\_d)[11])



### Interpretation

#### People having obese are less prone to Cardiac(7.2%)

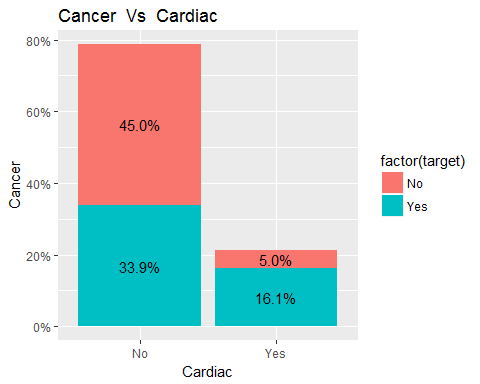
printggPlot2(d\_train,d\_train$Diabetes,d\_train$Cardiac,colnames(data\_d)[9],colnames(data\_d)[11])



### Interpretation

#### People having Dibetes are more prone to Cardiac(10.2%)

printggPlot2(d\_train,d\_train$Diabetes,d\_train$Cardiac,colnames(data\_d)[10],colnames(data\_d)[11])



### Interpretation

#### People having Cardiac are more prone to Cardiac(10.2%)

## CART

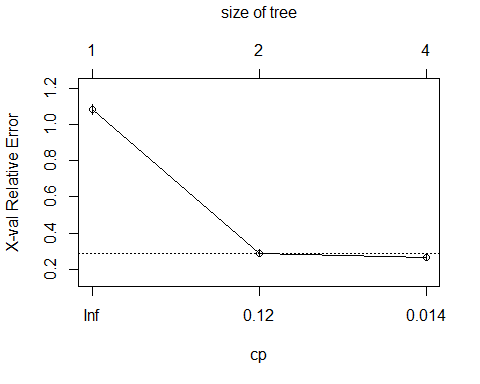
formula <- Cardiac ~ Fin\_Group + Family\_history + Staple\_Food + Alcohol + Smoke + High\_Cholestrol + Phy\_Activity + Obese + Diabetes + Cardiac  
title <- "Classification Tree for Cardiac"  
df <- d\_train  
fit<-NULL  
pfit <- cart\_fn(df, formula, title)

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

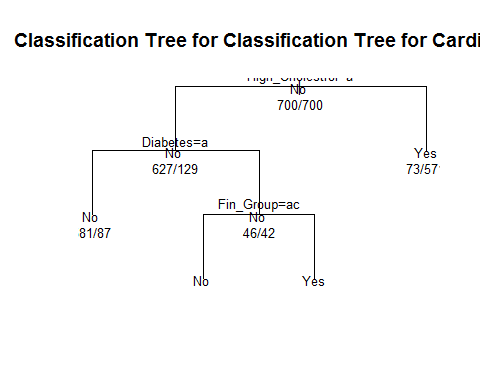
## Warning in model.matrix.default(attr(frame, "terms"), frame): problem with  
## term 10 in model.matrix: no columns are assigned

## 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: 700/1400 = 0.5  
##   
## n= 1400   
##   
## CP nsplit rel error xerror xstd  
## 1 0.71143 0 1.00000 1.08571 0.026628  
## 2 0.02000 1 0.28857 0.28857 0.018782  
## 3 0.01000 3 0.24857 0.26714 0.018184

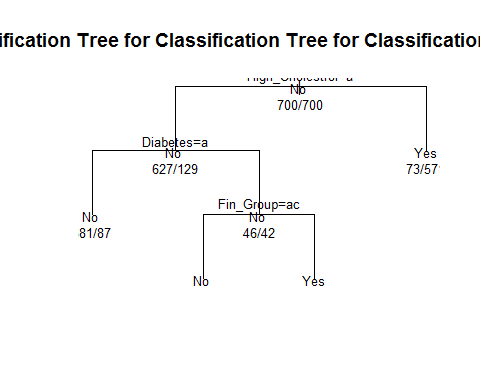


## Call:  
## rpart(formula = formula, data = d\_train, method = "class")  
## n= 1400   
##   
## CP nsplit rel error xerror xstd  
## 1 0.7114286 0 1.0000000 1.0857143 0.02662777  
## 2 0.0200000 1 0.2885714 0.2885714 0.01878201  
## 3 0.0100000 3 0.2485714 0.2671429 0.01818399  
##   
## Variable importance  
## High\_Cholestrol Diabetes Phy\_Activity Smoke   
## 61 15 10 6   
## Fin\_Group Family\_history   
## 6 3   
##   
## Node number 1: 1400 observations, complexity param=0.7114286  
## predicted class=No expected loss=0.5 P(node) =1  
## class counts: 700 700  
## probabilities: 0.500 0.500   
## left son=2 (756 obs) right son=3 (644 obs)  
## Primary splits:  
## High\_Cholestrol splits as LR, improve=356.573500, (0 missing)  
## Diabetes splits as LR, improve= 52.129850, (0 missing)  
## Alcohol splits as RL, improve= 7.942744, (0 missing)  
## Fin\_Group splits as LRL, improve= 5.697920, (0 missing)  
## Obese splits as RL, improve= 2.324929, (0 missing)  
## Surrogate splits:  
## Diabetes splits as LR, agree=0.626, adj=0.186, (0 split)  
## Phy\_Activity splits as RL, agree=0.614, adj=0.160, (0 split)  
## Smoke splits as LR, agree=0.581, adj=0.090, (0 split)  
## Family\_history splits as RL, agree=0.564, adj=0.051, (0 split)  
## Fin\_Group splits as RLL, agree=0.555, adj=0.033, (0 split)  
##   
## Node number 2: 756 observations, complexity param=0.02  
## predicted class=No expected loss=0.1706349 P(node) =0.54  
## class counts: 627 129  
## probabilities: 0.829 0.171   
## left son=4 (668 obs) right son=5 (88 obs)  
## Primary splits:  
## Diabetes splits as LR, improve=18.728780, (0 missing)  
## Phy\_Activity splits as LR, improve=17.873140, (0 missing)  
## Fin\_Group splits as LRR, improve=15.514650, (0 missing)  
## Smoke splits as RL, improve=15.476190, (0 missing)  
## Alcohol splits as RL, improve= 5.215097, (0 missing)  
##   
## Node number 3: 644 observations  
## predicted class=Yes expected loss=0.113354 P(node) =0.46  
## class counts: 73 571  
## probabilities: 0.113 0.887   
##   
## Node number 4: 668 observations  
## predicted class=No expected loss=0.1302395 P(node) =0.4771429  
## class counts: 581 87  
## probabilities: 0.870 0.130   
##   
## Node number 5: 88 observations, complexity param=0.02  
## predicted class=No expected loss=0.4772727 P(node) =0.06285714  
## class counts: 46 42  
## probabilities: 0.523 0.477   
## left son=10 (50 obs) right son=11 (38 obs)  
## Primary splits:  
## Fin\_Group splits as LRL, improve=20.464880000, (0 missing)  
## Phy\_Activity splits as LR, improve= 1.734732000, (0 missing)  
## Family\_history splits as RL, improve= 1.309091000, (0 missing)  
## Alcohol splits as LR, improve= 0.090909090, (0 missing)  
## Obese splits as RL, improve= 0.004068078, (0 missing)  
##   
## Node number 10: 50 observations  
## predicted class=No expected loss=0.18 P(node) =0.03571429  
## class counts: 41 9  
## probabilities: 0.820 0.180   
##   
## Node number 11: 38 observations  
## predicted class=Yes expected loss=0.1315789 P(node) =0.02714286  
## class counts: 5 33  
## probabilities: 0.132 0.868

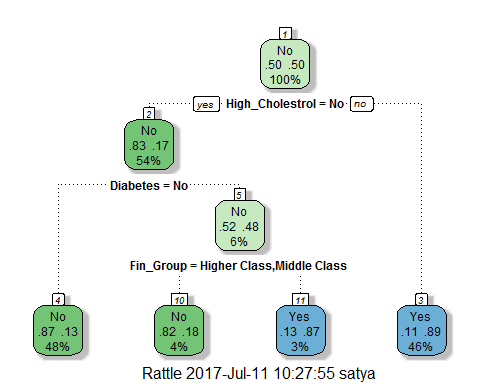


## Warning in model.matrix.default(Terms, m, contrasts): the response appeared  
## on the right-hand side and was dropped

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



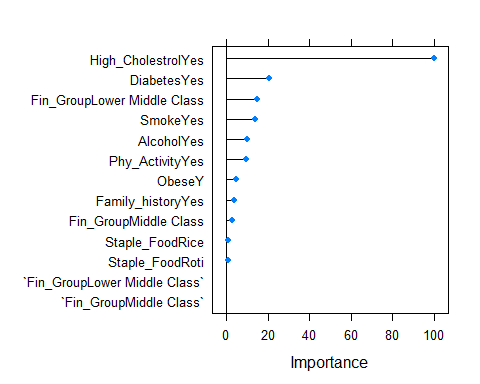
fancyRpartPlot(pfit)



varImp(object=fit)

## rpart variable importance  
##   
## Overall  
## High\_CholestrolYes 100.000  
## DiabetesYes 20.679  
## Fin\_GroupLower Middle Class 15.048  
## SmokeYes 14.087  
## AlcoholYes 9.943  
## Phy\_ActivityYes 9.461  
## ObeseY 4.654  
## Family\_historyYes 3.708  
## Fin\_GroupMiddle Class 3.023  
## Staple\_FoodRice 1.060  
## Staple\_FoodRoti 1.042  
## `Fin\_GroupMiddle Class` 0.000  
## `Fin\_GroupLower Middle Class` 0.000

plot(varImp(fit))



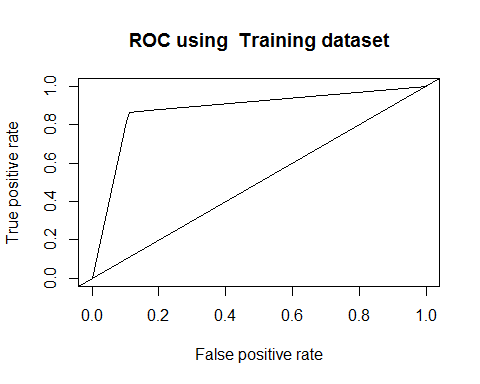
#### Interpretation of Variable Importance for Crdiac

###### 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)

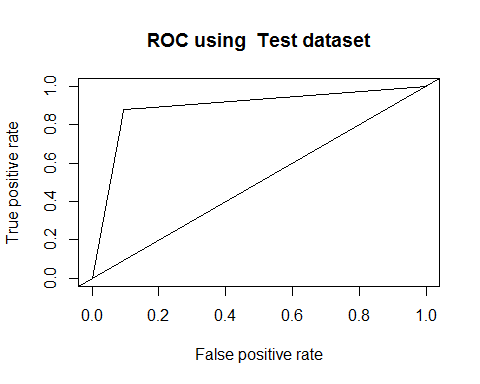
## Dataset KS auc gini OA  
## 1 Training dataset 0.7514286 0.8778653 0.3778653 0.8757143

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 0.7866667 0.8930222 0.3785439 0.8933333

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)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(logit)

##   
## Call:  
## glm(formula = formula, family = "binomial", data = d\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8061 0.0000 0.0000 0.4465 0.7771   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 20.3940 1085.8966 0.019 0.9850   
## Fin\_GroupLower Middle Class -0.2496 0.2912 -0.857 0.3914   
## Fin\_GroupMiddle Class -0.6060 0.3794 -1.597 0.1102   
## Family\_historyYes -0.1522 0.2489 -0.612 0.5408   
## Staple\_FoodRice 0.2869 1382.7541 0.000 0.9998   
## Staple\_FoodRoti -15.2793 1085.8962 -0.014 0.9888   
## AlcoholYes -0.4324 0.2539 -1.703 0.0885 .  
## SmokeYes 0.8510 0.3630 2.345 0.0191 \*  
## High\_CholestrolYes 34.7727 1343.0510 0.026 0.9793   
## Phy\_ActivityYes -0.3629 0.3273 -1.109 0.2675   
## ObeseY -0.3789 0.3073 -1.233 0.2176   
## DiabetesYes -0.8058 0.4508 -1.788 0.0738 .  
## CancerYes -37.2687 1343.0513 -0.028 0.9779   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1940.8 on 1399 degrees of freedom  
## Residual deviance: 466.6 on 1387 degrees of freedom  
## AIC: 492.6  
##   
## Number of Fisher Scoring iterations: 20

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

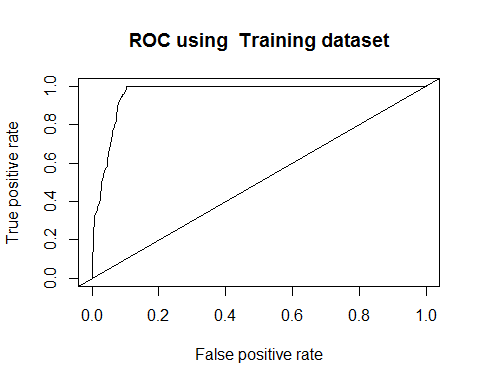
###### Smoking is the only significance predicator.Rest are not having any importance at significance level of 95%

perf <- pR2(logit)  
print(perf)

## llh llhNull G2 McFadden r2ML   
## -233.3010601 -970.4060528 1474.2099853 0.7595841 0.6511130   
## r2CU   
## 0.8681506

### 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)

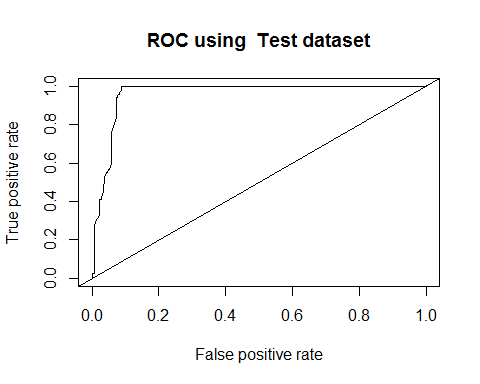
## Dataset auc McFadden OA  
## 1 Training dataset 0.9639296 0.1732045 0.6666667

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 0.9616444 0.1732045 0.3333333

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

##   
## OA is overall accuracy

varImp(logit, scale = FALSE)

## Overall  
## Fin\_GroupLower Middle Class 0.8570112021  
## Fin\_GroupMiddle Class 1.5970790474  
## Family\_historyYes 0.6116707001  
## Staple\_FoodRice 0.0002075173  
## Staple\_FoodRoti 0.0140707138  
## AlcoholYes 1.7032392542  
## SmokeYes 2.3445225345  
## High\_CholestrolYes 0.0258908616  
## Phy\_ActivityYes 1.1087382062  
## ObeseY 1.2328117039  
## DiabetesYes 1.7877154672  
## CancerYes 0.0277493132

### End of R script

```