NCD\_Cancer

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, Cancer %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"))  
str(data\_d)

## 'data.frame': 9328 obs. of 10 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 ...

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"   
   
  
  
cancer\_vector <- c("Fin\_Group","Family\_history","Staple\_Food","Alcohol","Smoke","High\_Cholestrol","Phy\_Activity","Obese","Diabetes")  
  
  
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 10 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 ...

names(data\_d)

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

nrow(data\_d)

## [1] 9328

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

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

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

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

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

##   
## Dataset file 1 Cancer Count :

## dd  
## No Yes   
## 0 8693

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

##   
## Dataset file 2 Cancer Count :

## dd  
## No Yes   
## 635 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  
## 3547 Middle Class No Roti Yes No  
## 8693 Lower Middle Class Yes Roti Yes No  
## 4782 Higher Class No Roti Yes No  
## 9655 Lower Middle Class No Roti No Yes  
## 10236 Higher Class No Rice Yes No  
## 473 Middle Class Yes Roti Yes No  
## High\_Cholestrol Phy\_Activity Obese Diabetes Cancer  
## 3547 No Yes N No Yes  
## 8693 No Yes N No Yes  
## 4782 Yes No Y No Yes  
## 9655 No Yes N No Yes  
## 10236 No No N No Yes  
## 473 No Yes Y Yes Yes

print(nrow(d1))

## [1] 1000

print(head(d2))

## Fin\_Group Family\_history Staple\_Food Alcohol Smoke  
## 2169 Lower Middle Class Yes Roti No No  
## 2383 Lower Middle Class Yes Roti Yes Yes  
## 2094 Lower Middle Class Yes Roti No No  
## 2554 Lower Middle Class No Roti No No  
## 2549 Middle Class No Roti No No  
## 2302 Lower Middle Class Yes Roti No No  
## High\_Cholestrol Phy\_Activity Obese Diabetes Cancer  
## 2169 No Yes Y Yes No  
## 2383 No Yes N No No  
## 2094 No Yes N Yes No  
## 2554 No Yes N No No  
## 2549 No Yes N No No  
## 2302 No Yes N No No

print(nrow(d2))

## [1] 635

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

##   
## Dataset file 1 Cancer Count :

## dd  
## No Yes   
## 0 1000

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

##   
## Dataset file 2 Cancer Count :

## dd  
## No Yes   
## 635 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] 183

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

### Merge d1\_train (Cancer dataset) and d2\_train (Non-Cancer 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] 1152

print(nrow(d\_test))

## [1] 483

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

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

##   
## Cancer Count Yes: 700 No: 0

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

##   
## Cancer Count Yes: 0 No: 452

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

##   
## Cancer Count Yes: 300 No: 0

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

##   
## Train dataset file Cancer Count : 1152

table(d\_train$Cancer)

##   
## No Yes   
## 452 700

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

##   
## Test dataset file Cancer Count : 483

table(d\_test$Cancer)

##   
## No Yes   
## 183 300

### Remove unwanted variables

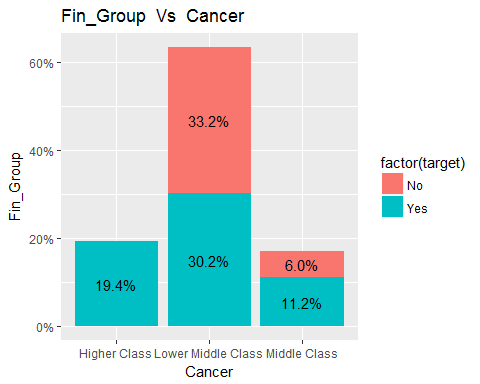
#### The following columns are not required for Cancer - NCD:

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(cancer\_vector)){  
 if(cancer\_vector[e]!="Cancer")  
eda\_functions("Cancer",d\_train,d\_train[,"Cancer"],d\_train[,e],cancer\_vector[e])  
 }

##   
##   
## Performing mlbenchxtab   
##   
## Cancer Vs Fin\_Group   
##   
## column Higher Class Lower Middle Class Middle Class  
## target   
## No 0 383 69  
## Yes 223 348 129  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target Higher Class Lower Middle Class Middle Class  
## No 0.0000000 0.8473451 0.1526549  
## Yes 0.3185714 0.4971429 0.1842857  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target Higher Class Lower Middle Class Middle Class  
## No 0.0000000 0.5239398 0.3484848  
## Yes 1.0000000 0.4760602 0.6515152  
##   
##   
## Performing Chi Square Test for Cancer and Fin\_Group  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Fin\_Group  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 0 383 69  
## Yes 223 348 129  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 198.68, df = 2, p-value = 7.211e-44  
##   
##   
## 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 0   
##   
## <------ Chi square with simulation for p value --------->   
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##   
## data: tab1  
## X-squared = 198.68, 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 Cancer 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   
##   
## Cancer Vs Family\_history   
##   
## column No Yes  
## target   
## No 140 312  
## Yes 338 362  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target No Yes  
## No 0.3097345 0.6902655  
## Yes 0.4828571 0.5171429  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.292887 0.462908  
## Yes 0.707113 0.537092  
##   
##   
## Performing Chi Square Test for Cancer and Family\_history  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Family\_history  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 140 312  
## Yes 338 362  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 33.91, df = 1, p-value = 5.777e-09  
##   
##   
## 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 140   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 33.199, df = 1, p-value = 8.319e-09  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 8.318748e-09   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs Staple\_Food   
##   
## column Pizza Rice Roti  
## target   
## No 4 3 445  
## Yes 11 10 679  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target Pizza Rice Roti  
## No 0.008849558 0.006637168 0.984513274  
## Yes 0.015714286 0.014285714 0.970000000  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target Pizza Rice Roti  
## No 0.2666667 0.2307692 0.3959075  
## Yes 0.7333333 0.7692308 0.6040925  
##   
##   
## Performing Chi Square Test for Cancer and Staple\_Food  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Staple\_Food  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 4 3 445  
## Yes 11 10 679  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 2.4771, df = 2, p-value = 0.2898  
##   
##   
## 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 = 2.4771, df = NA, p-value = 0.3078  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained with simulation for p-value : 0.3078461   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs Alcohol   
##   
## column No Yes  
## target   
## No 274 178  
## Yes 301 399  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target No Yes  
## No 0.6061947 0.3938053  
## Yes 0.4300000 0.5700000  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.4765217 0.3084922  
## Yes 0.5234783 0.6915078  
##   
##   
## Performing Chi Square Test for Cancer and Alcohol  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Alcohol  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 274 178  
## Yes 301 399  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 34.11, df = 1, p-value = 5.219e-09  
##   
##   
## 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 178   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 33.405, df = 1, p-value = 7.484e-09  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 7.483629e-09   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs Smoke   
##   
## column No Yes  
## target   
## No 407 45  
## Yes 374 326  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target No Yes  
## No 0.90044248 0.09955752  
## Yes 0.53428571 0.46571429  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.5211268 0.1212938  
## Yes 0.4788732 0.8787062  
##   
##   
## Performing Chi Square Test for Cancer and Smoke  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Smoke  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 407 45  
## Yes 374 326  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 168.65, df = 1, p-value = 1.456e-38  
##   
##   
## 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 45   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 166.98, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 3.376206e-38   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs High\_Cholestrol   
##   
## column No Yes  
## target   
## No 442 10  
## Yes 437 263  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target No Yes  
## No 0.97787611 0.02212389  
## Yes 0.62428571 0.37571429  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.50284414 0.03663004  
## Yes 0.49715586 0.96336996  
##   
##   
## Performing Chi Square Test for Cancer and High\_Cholestrol  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of High\_Cholestrol  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 442 10  
## Yes 437 263  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 189.91, df = 1, p-value = 3.333e-43  
##   
##   
## 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 10   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 187.96, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 8.884544e-43   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs Phy\_Activity   
##   
## column No Yes  
## target   
## No 16 436  
## Yes 324 376  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target No Yes  
## No 0.03539823 0.96460177  
## Yes 0.46285714 0.53714286  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.04705882 0.53694581  
## Yes 0.95294118 0.46305419  
##   
##   
## Performing Chi Square Test for Cancer and Phy\_Activity  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Phy\_Activity  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 16 436  
## Yes 324 376  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 241.24, df = 1, p-value = 2.114e-54  
##   
##   
## 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 16   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 239.19, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 5.918403e-54   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs Obese   
##   
## column N Y  
## target   
## No 415 37  
## Yes 598 102  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target N Y  
## No 0.91814159 0.08185841  
## Yes 0.85428571 0.14571429  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target N Y  
## No 0.4096742 0.2661871  
## Yes 0.5903258 0.7338129  
##   
##   
## Performing Chi Square Test for Cancer and Obese  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Obese  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 415 37  
## Yes 598 102  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 10.555, df = 1, p-value = 0.001159  
##   
##   
## 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 37   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 9.9619, df = 1, p-value = 0.001598  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.001598097   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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   
##   
## Cancer Vs Diabetes   
##   
## column No Yes  
## target   
## No 285 167  
## Yes 583 117  
##   
##   
## Within each level of predictor variable values, % of Cancer patients   
##   
## column  
## target No Yes  
## No 0.6305310 0.3694690  
## Yes 0.8328571 0.1671429  
##   
##   
## Cancer Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.3283410 0.5880282  
## Yes 0.6716590 0.4119718  
##   
##   
## Performing Chi Square Test for Cancer and Diabetes  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Cancer are independent of Diabetes  
##   
##   
## Alternative hypothesis : The relative proportions of Cancer 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 285 167  
## Yes 583 117  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1152   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 60.53, df = 1, p-value = 7.255e-15  
##   
##   
## 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 117   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 59.443, df = 1, p-value = 1.259e-14  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.258568e-14   
##   
##   
## <-- Decision: -->  
## The relative proportions of Cancer 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.

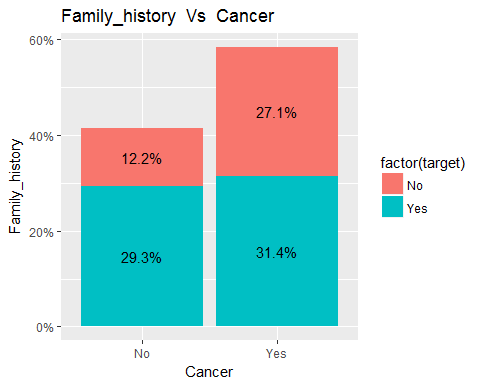
printggPlot2(d\_train,d\_train$Fin\_Group,d\_train$Cancer,colnames(data\_d)[1],colnames(data\_d)[10])



### Interpretation

#### Lower Middle class people are more prone to Cancer (30.2%)

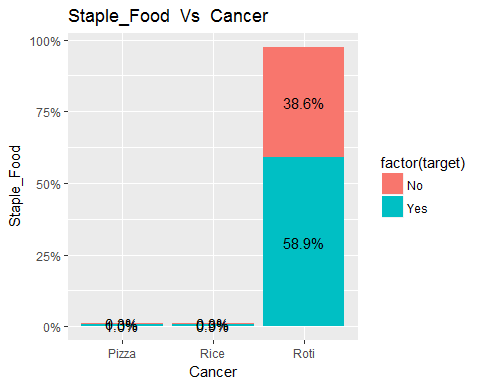
printggPlot2(d\_train,d\_train$Family\_history,d\_train$Cancer,colnames(data\_d)[2],colnames(data\_d)[10])



### Interpretation

#### People having family history of Cancer are more prone to Cancer(31.4%)

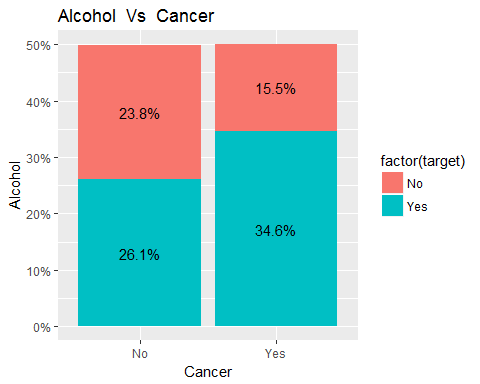
printggPlot2(d\_train,d\_train$Staple\_Food,d\_train$Cancer,colnames(data\_d)[3],colnames(data\_d)[10])



### Interpretation

#### People having Staple Food are more prone to Cancer(58.9%)

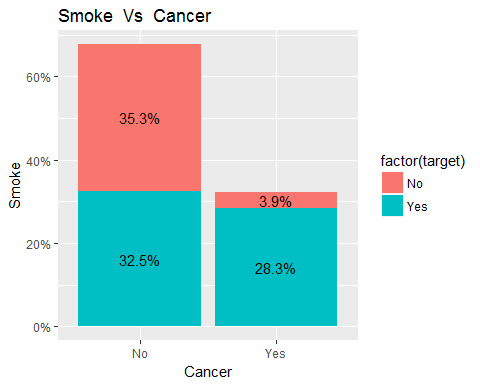
printggPlot2(d\_train,d\_train$Alcohol,d\_train$Cancer,colnames(data\_d)[4],colnames(data\_d)[10])



### Interpretation

#### People having Alcohol habit are more prone to Cancer(38.6%)

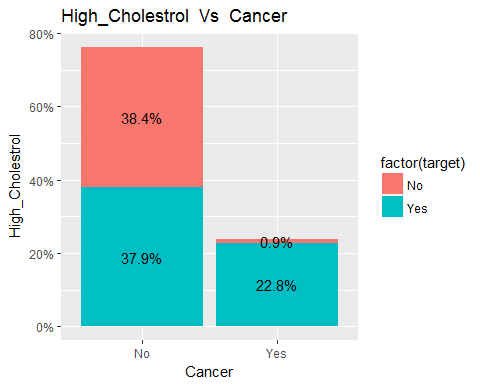
printggPlot2(d\_train,d\_train$Smoke,d\_train$Cancer,colnames(data\_d)[5],colnames(data\_d)[10])



### Interpretation

#### People having Smoking habit are more prone to Cancer(28.3%)

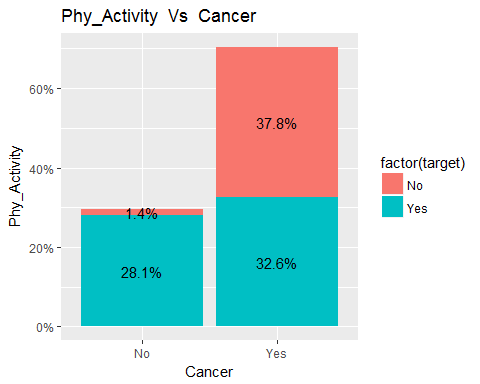
printggPlot2(d\_train,d\_train$High\_Cholestrol,d\_train$Cancer,colnames(data\_d)[6],colnames(data\_d)[10])



### Interpretation

#### People having High\_Cholestrol are more prone to Cancer(22.8%)

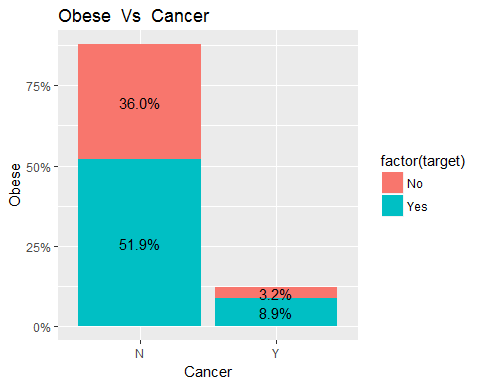
printggPlot2(d\_train,d\_train$Phy\_Activity,d\_train$Cancer,colnames(data\_d)[7],colnames(data\_d)[10])



### Interpretation

#### People not having Phy\_Activity are more prone to Cancer(32.6%)

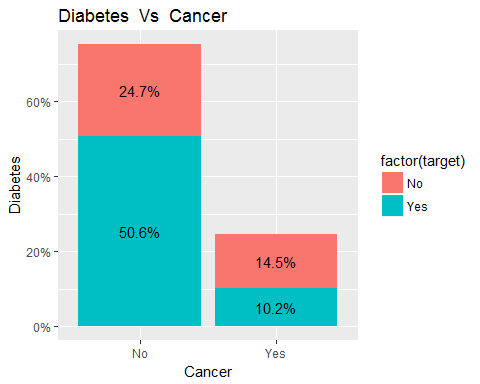
printggPlot2(d\_train,d\_train$Obese,d\_train$Cancer,colnames(data\_d)[8],colnames(data\_d)[10])



### Interpretation

#### People having obese are more prone to Cancer(8.9%)

printggPlot2(d\_train,d\_train$Diabetes,d\_train$Cancer,colnames(data\_d)[9],colnames(data\_d)[10])



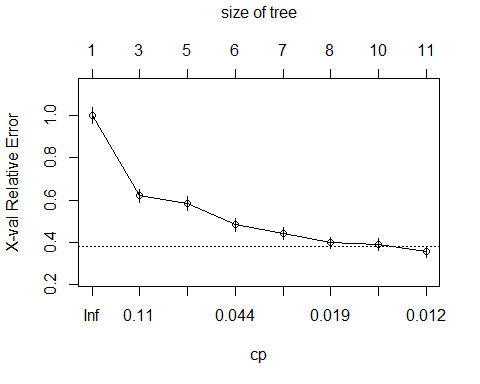
### Interpretation

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

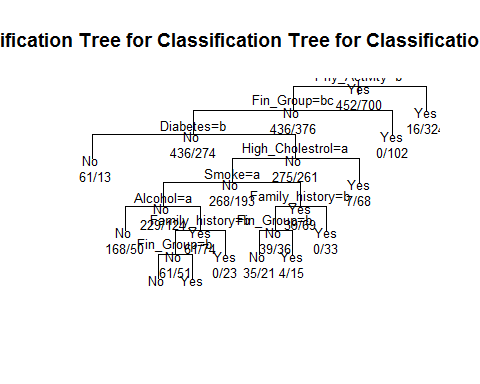
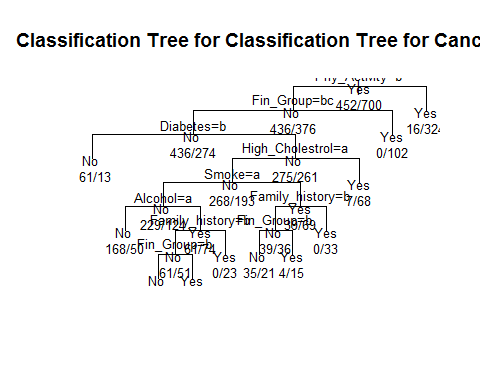
## CART

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

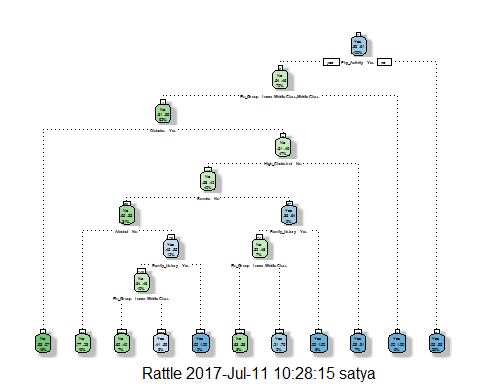
##   
## Classification tree:  
## rpart(formula = formula, data = d\_train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Alcohol Diabetes Family\_history Fin\_Group   
## [5] High\_Cholestrol Phy\_Activity Smoke   
##   
## Root node error: 452/1152 = 0.39236  
##   
## n= 1152   
##   
## CP nsplit rel error xerror xstd  
## 1 0.179204 0 1.00000 1.00000 0.036665  
## 2 0.067478 2 0.64159 0.61947 0.032209  
## 3 0.066372 4 0.50664 0.58407 0.031560  
## 4 0.028761 5 0.44027 0.48230 0.029413  
## 5 0.022124 6 0.41150 0.44027 0.028386  
## 6 0.015487 7 0.38938 0.39823 0.027265  
## 7 0.013274 9 0.35841 0.38938 0.027016  
## 8 0.010000 10 0.34513 0.35619 0.026037



## Call:  
## rpart(formula = formula, data = d\_train, method = "class")  
## n= 1152   
##   
## CP nsplit rel error xerror xstd  
## 1 0.17920354 0 1.0000000 1.0000000 0.03666516  
## 2 0.06747788 2 0.6415929 0.6194690 0.03220865  
## 3 0.06637168 4 0.5066372 0.5840708 0.03156049  
## 4 0.02876106 5 0.4402655 0.4823009 0.02941284  
## 5 0.02212389 6 0.4115044 0.4402655 0.02838627  
## 6 0.01548673 7 0.3893805 0.3982301 0.02726495  
## 7 0.01327434 9 0.3584071 0.3893805 0.02701572  
## 8 0.01000000 10 0.3451327 0.3561947 0.02603665  
##   
## Variable importance  
## Phy\_Activity Fin\_Group Diabetes High\_Cholestrol   
## 34 24 13 12   
## Family\_history Alcohol Smoke   
## 8 5 4   
##   
## Node number 1: 1152 observations, complexity param=0.1792035  
## predicted class=Yes expected loss=0.3923611 P(node) =1  
## class counts: 452 700  
## probabilities: 0.392 0.608   
## left son=2 (812 obs) right son=3 (340 obs)  
## Primary splits:  
## Phy\_Activity splits as RL, improve=115.02820, (0 missing)  
## High\_Cholestrol splits as LR, improve= 90.55238, (0 missing)  
## Fin\_Group splits as RLL, improve= 85.14194, (0 missing)  
## Smoke splits as LR, improve= 80.41918, (0 missing)  
## Diabetes splits as RL, improve= 28.86134, (0 missing)  
## Surrogate splits:  
## High\_Cholestrol splits as LR, agree=0.735, adj=0.103, (0 split)  
## Fin\_Group splits as RLL, agree=0.721, adj=0.056, (0 split)  
##   
## Node number 2: 812 observations, complexity param=0.1792035  
## predicted class=No expected loss=0.4630542 P(node) =0.7048611  
## class counts: 436 376  
## probabilities: 0.537 0.463   
## left son=4 (710 obs) right son=5 (102 obs)  
## Primary splits:  
## Fin\_Group splits as RLL, improve=67.26494, (0 missing)  
## High\_Cholestrol splits as LR, improve=57.20418, (0 missing)  
## Diabetes splits as RL, improve=56.70877, (0 missing)  
## Smoke splits as LR, improve=49.04186, (0 missing)  
## Alcohol splits as LR, improve=14.05841, (0 missing)  
##   
## Node number 3: 340 observations  
## predicted class=Yes expected loss=0.04705882 P(node) =0.2951389  
## class counts: 16 324  
## probabilities: 0.047 0.953   
##   
## Node number 4: 710 observations, complexity param=0.06747788  
## predicted class=No expected loss=0.3859155 P(node) =0.6163194  
## class counts: 436 274  
## probabilities: 0.614 0.386   
## left son=8 (174 obs) right son=9 (536 obs)  
## Primary splits:  
## Diabetes splits as RL, improve=44.643670, (0 missing)  
## High\_Cholestrol splits as LR, improve=41.373550, (0 missing)  
## Smoke splits as LR, improve=36.780560, (0 missing)  
## Alcohol splits as LR, improve=11.036040, (0 missing)  
## Fin\_Group splits as -LR, improve= 9.380433, (0 missing)  
##   
## Node number 5: 102 observations  
## predicted class=Yes expected loss=0 P(node) =0.08854167  
## class counts: 0 102  
## probabilities: 0.000 1.000   
##   
## Node number 8: 174 observations  
## predicted class=No expected loss=0.07471264 P(node) =0.1510417  
## class counts: 161 13  
## probabilities: 0.925 0.075   
##   
## Node number 9: 536 observations, complexity param=0.06747788  
## predicted class=No expected loss=0.4869403 P(node) =0.4652778  
## class counts: 275 261  
## probabilities: 0.513 0.487   
## left son=18 (461 obs) right son=19 (75 obs)  
## Primary splits:  
## High\_Cholestrol splits as LR, improve=30.724700, (0 missing)  
## Smoke splits as LR, improve=20.286440, (0 missing)  
## Alcohol splits as LR, improve=15.100770, (0 missing)  
## Fin\_Group splits as -LR, improve= 5.780747, (0 missing)  
## Obese splits as LR, improve= 1.566852, (0 missing)  
##   
## Node number 18: 461 observations, complexity param=0.06637168  
## predicted class=No expected loss=0.4186551 P(node) =0.4001736  
## class counts: 268 193  
## probabilities: 0.581 0.419   
## left son=36 (353 obs) right son=37 (108 obs)  
## Primary splits:  
## Smoke splits as LR, improve=13.681950, (0 missing)  
## Alcohol splits as LR, improve=11.789450, (0 missing)  
## Fin\_Group splits as -LR, improve= 5.733239, (0 missing)  
## Family\_history splits as LR, improve= 1.459594, (0 missing)  
## Staple\_Food splits as RRL, improve= 1.068223, (0 missing)  
##   
## Node number 19: 75 observations  
## predicted class=Yes expected loss=0.09333333 P(node) =0.06510417  
## class counts: 7 68  
## probabilities: 0.093 0.907   
##   
## Node number 36: 353 observations, complexity param=0.02876106  
## predicted class=No expected loss=0.3512748 P(node) =0.3064236  
## class counts: 229 124  
## probabilities: 0.649 0.351   
## left son=72 (218 obs) right son=73 (135 obs)  
## Primary splits:  
## Alcohol splits as LR, improve=16.9455600, (0 missing)  
## Family\_history splits as LR, improve= 7.8113230, (0 missing)  
## Fin\_Group splits as -LR, improve= 3.5682000, (0 missing)  
## Staple\_Food splits as RRL, improve= 1.2266060, (0 missing)  
## Obese splits as LR, improve= 0.6842935, (0 missing)  
## Surrogate splits:  
## Family\_history splits as LR, agree=0.666, adj=0.126, (0 split)  
## Fin\_Group splits as -LR, agree=0.643, adj=0.067, (0 split)  
##   
## Node number 37: 108 observations, complexity param=0.01548673  
## predicted class=Yes expected loss=0.3611111 P(node) =0.09375  
## class counts: 39 69  
## probabilities: 0.361 0.639   
## left son=74 (75 obs) right son=75 (33 obs)  
## Primary splits:  
## Family\_history splits as RL, improve=1.239333e+01, (0 missing)  
## Alcohol splits as RL, improve=2.963203e+00, (0 missing)  
## Fin\_Group splits as -LR, improve=2.048167e+00, (0 missing)  
## Obese splits as LR, improve=5.065856e-04, (0 missing)  
## Surrogate splits:  
## Staple\_Food splits as LRL, agree=0.713, adj=0.061, (0 split)  
##   
## Node number 72: 218 observations  
## predicted class=No expected loss=0.2293578 P(node) =0.1892361  
## class counts: 168 50  
## probabilities: 0.771 0.229   
##   
## Node number 73: 135 observations, complexity param=0.02212389  
## predicted class=Yes expected loss=0.4518519 P(node) =0.1171875  
## class counts: 61 74  
## probabilities: 0.452 0.548   
## left son=146 (112 obs) right son=147 (23 obs)  
## Primary splits:  
## Family\_history splits as RL, improve=11.3205000, (0 missing)  
## Fin\_Group splits as -LR, improve= 1.8293370, (0 missing)  
## Obese splits as LR, improve= 0.5979076, (0 missing)  
##   
## Node number 74: 75 observations, complexity param=0.01548673  
## predicted class=No expected loss=0.48 P(node) =0.06510417  
## class counts: 39 36  
## probabilities: 0.520 0.480   
## left son=148 (56 obs) right son=149 (19 obs)  
## Primary splits:  
## Fin\_Group splits as -LR, improve=4.87421100, (0 missing)  
## Alcohol splits as RL, improve=0.48761900, (0 missing)  
## Obese splits as RL, improve=0.02585859, (0 missing)  
##   
## Node number 75: 33 observations  
## predicted class=Yes expected loss=0 P(node) =0.02864583  
## class counts: 0 33  
## probabilities: 0.000 1.000   
##   
## Node number 146: 112 observations, complexity param=0.01327434  
## predicted class=No expected loss=0.4553571 P(node) =0.09722222  
## class counts: 61 51  
## probabilities: 0.545 0.455   
## left son=292 (80 obs) right son=293 (32 obs)  
## Primary splits:  
## Fin\_Group splits as -LR, improve=1.7160710, (0 missing)  
## Obese splits as LR, improve=0.1965056, (0 missing)  
##   
## Node number 147: 23 observations  
## predicted class=Yes expected loss=0 P(node) =0.01996528  
## class counts: 0 23  
## probabilities: 0.000 1.000   
##   
## Node number 148: 56 observations  
## predicted class=No expected loss=0.375 P(node) =0.04861111  
## class counts: 35 21  
## probabilities: 0.625 0.375   
##   
## Node number 149: 19 observations  
## predicted class=Yes expected loss=0.2105263 P(node) =0.01649306  
## class counts: 4 15  
## probabilities: 0.211 0.789   
##   
## Node number 292: 80 observations  
## predicted class=No expected loss=0.4 P(node) =0.06944444  
## class counts: 48 32  
## probabilities: 0.600 0.400   
##   
## Node number 293: 32 observations  
## predicted class=Yes expected loss=0.40625 P(node) =0.02777778  
## class counts: 13 19  
## probabilities: 0.406 0.594



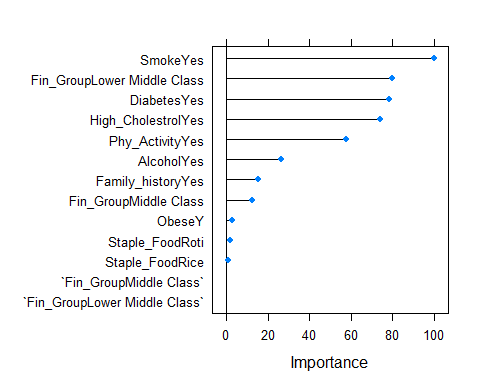
fancyRpartPlot(pfit)



varImp(object=fit)

## rpart variable importance  
##   
## Overall  
## SmokeYes 100.0000  
## Fin\_GroupLower Middle Class 79.9102  
## DiabetesYes 78.4327  
## High\_CholestrolYes 73.9414  
## Phy\_ActivityYes 57.5632  
## AlcoholYes 26.5463  
## Family\_historyYes 15.4103  
## Fin\_GroupMiddle Class 12.5619  
## ObeseY 2.8803  
## Staple\_FoodRoti 2.0141  
## Staple\_FoodRice 0.7388  
## `Fin\_GroupLower Middle Class` 0.0000  
## `Fin\_GroupMiddle Class` 0.0000

plot(varImp(fit))



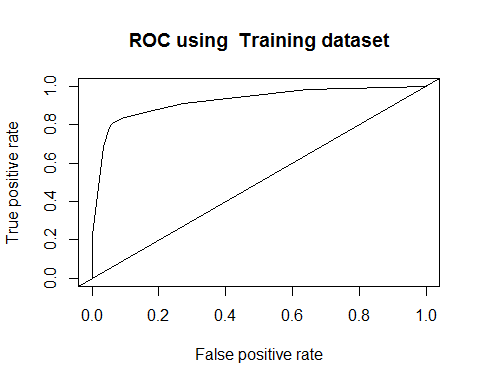
#### Interpretation of Variable Importance for Cancerc as shown in above graph

###### Alcohol,Smoke,High\_Cholestrol,Phy\_Activity,Diabetes are top influencing predictors Finacial group and and Obese also has significant.

###### significance level Considered is 95%.

### Performance measures using training data

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



print(perf\_measures1)

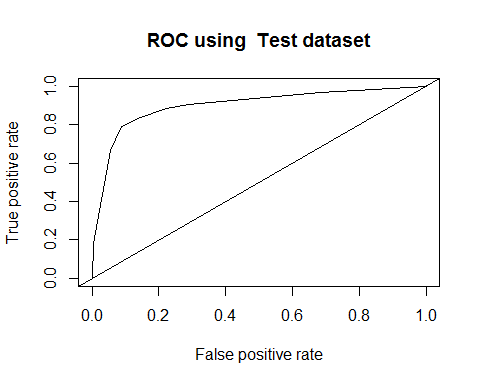
## Dataset KS auc gini OA  
## 1 Training dataset 0.7474083 0.9229693 0.3319134 0.8645833

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

##   
## OA is overall accuracy

### Performance measures using test data

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



print(perf\_measures2)

## Dataset KS auc gini OA  
## 1 Test dataset 0.7059016 0.8993625 0.3159412 0.8447205

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

##   
## OA is overall accuracy

### Logistic Regression

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

##   
## Call:  
## glm(formula = formula, family = "binomial", data = d\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.50059 -0.68333 0.00003 0.35524 2.42630   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 21.79692 583.10771 0.037 0.970182   
## Fin\_GroupLower Middle Class -19.32386 583.10714 -0.033 0.973563   
## Fin\_GroupMiddle Class -18.17708 583.10716 -0.031 0.975132   
## Family\_historyYes -0.13078 0.20131 -0.650 0.515919   
## Staple\_FoodRice -0.08446 1.25333 -0.067 0.946274   
## Staple\_FoodRoti -1.12013 0.74456 -1.504 0.132473   
## AlcoholYes 0.71306 0.19464 3.663 0.000249 \*\*\*  
## SmokeYes 1.14641 0.21943 5.224 1.75e-07 \*\*\*  
## High\_CholestrolYes 2.78988 0.37757 7.389 1.48e-13 \*\*\*  
## Phy\_ActivityYes -2.55784 0.31431 -8.138 4.02e-16 \*\*\*  
## ObeseY 0.83130 0.29751 2.794 0.005203 \*\*   
## DiabetesYes -2.26670 0.29937 -7.572 3.69e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1543.2 on 1151 degrees of freedom  
## Residual deviance: 762.8 on 1140 degrees of freedom  
## AIC: 786.8  
##   
## Number of Fisher Scoring iterations: 18

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

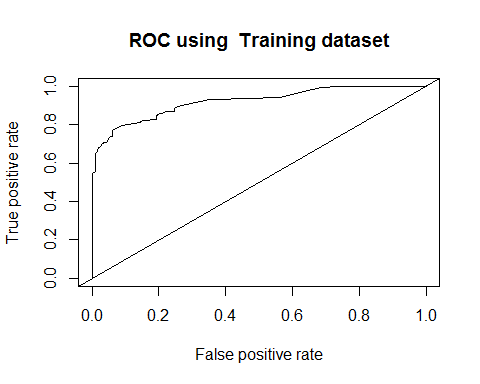
###### Alcohol,Smoke,High\_Cholestrol,Phy\_Activity,Diabetes and Obese predicator are highly significant at significance level of 95%.

perf <- pR2(logit)  
print(perf)

## llh llhNull G2 McFadden r2ML   
## -381.3991593 -771.6009973 780.4036761 0.5057042 0.4920812   
## r2CU   
## 0.6667355

### Logistic Regression Performance measures using training data

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



print(perf\_measures1)

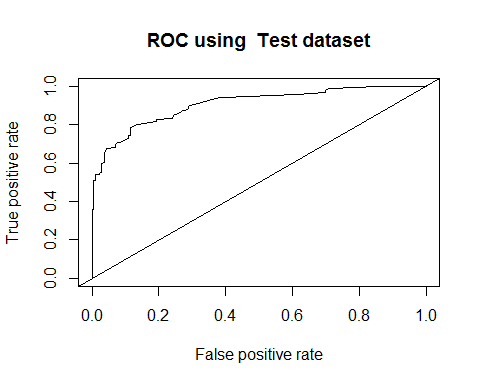
## Dataset auc McFadden OA  
## 1 Training dataset 0.9210635 0.1732045 0.9491525

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

##   
## OA is overall accuracy

### Logistic Regression Performance measures using test data

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



print(perf\_measures2)

## Dataset auc McFadden OA  
## 1 Test dataset 0.9060383 0.1732045 0.84

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

##   
## OA is overall accuracy

varImp(logit, scale = FALSE)

## Overall  
## Fin\_GroupLower Middle Class 0.03313947  
## Fin\_GroupMiddle Class 0.03117280  
## Family\_historyYes 0.64964903  
## Staple\_FoodRice 0.06738641  
## Staple\_FoodRoti 1.50442154  
## AlcoholYes 3.66343677  
## SmokeYes 5.22442048  
## High\_CholestrolYes 7.38912129  
## Phy\_ActivityYes 8.13804767  
## ObeseY 2.79417457  
## DiabetesYes 7.57165201

### End of R script

```