NCD\_Diabetes

Capstone\_Hyderabad

Wednesday, July 05, 2017

### Load required libraries

library(caret)

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

## Loading required package: lattice

## Loading required package: ggplot2

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

library(rpart)

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

library(ROCR)

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

## Loading required package: gplots

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

##   
## Attaching package: 'gplots'

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

library(ineq)

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

library(plyr)

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

library(dplyr)

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

##   
## Attaching package: 'dplyr'

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

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

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

library(ggplot2)  
library(scales)

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

library(stringr)

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

library(rattle)

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

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

library(RColorBrewer)

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

library(lmtest)

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

## Loading required package: zoo

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

##   
## Attaching package: 'zoo'

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

library(pscl)

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

## Loading required package: MASS

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

##   
## Attaching package: 'MASS'

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

## Classes and Methods for R developed in the

## Political Science Computational Laboratory

## Department of Political Science

## Stanford University

## Simon Jackman

## hurdle and zeroinfl functions by Achim Zeileis

require(caret)  
library(mlbench)

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

library(DMwR)

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

## Loading required package: grid

##   
## Attaching package: 'DMwR'

## The following object is masked from 'package:plyr':  
##   
## join

### Load dataset

setwd("C:\\Material\\CAPStone\\Scripts")  
data <- read.csv("ncd\_latest\_062817.csv")  
  
source("required\_user\_functions.R")  
  
d\_data <- subset(data, Diabetes %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"))  
str(data\_d)

## 'data.frame': 9328 obs. of 9 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 ...

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"   
  
  
diabetes\_vector <- c("Fin\_Group","Family\_history","Staple\_Food","Alcohol","Smoke","High\_Cholestrol","Phy\_Activity","Obese")  
  
  
  
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 9 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 ...

names(data\_d)

## [1] "Fin\_Group" "Family\_history" "Staple\_Food" "Alcohol"   
## [5] "Smoke" "High\_Cholestrol" "Phy\_Activity" "Obese"   
## [9] "Diabetes"

nrow(data\_d)

## [1] 9328

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

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

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

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

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

##   
## Dataset file 1 Diabetes Count :

## dd  
## No Yes   
## 0 1551

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

##   
## Dataset file 2 Diabetes Count :

## dd  
## No Yes   
## 7777 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  
## 2090 Lower Middle Class Yes Pizza No No  
## 8081 Middle Class Yes Roti Yes Yes  
## 3001 Higher Class No Roti Yes Yes  
## 9403 Middle Class No Pizza No Yes  
## 10072 Lower Middle Class Yes Roti Yes Yes  
## 323 Middle Class Yes Roti Yes No  
## High\_Cholestrol Phy\_Activity Obese Diabetes  
## 2090 No Yes N Yes  
## 8081 Yes No N Yes  
## 3001 Yes No N Yes  
## 9403 Yes No N Yes  
## 10072 Yes No N Yes  
## 323 No Yes N Yes

print(nrow(d1))

## [1] 1000

print(head(d2))

## Fin\_Group Family\_history Staple\_Food Alcohol Smoke  
## 3112 Lower Middle Class No Roti Yes Yes  
## 6509 Higher Class Yes Roti Yes No  
## 1747 Middle Class Yes Roti Yes No  
## 9277 Lower Middle Class No Roti Yes No  
## 9209 Lower Middle Class Yes Roti Yes Yes  
## 5246 Higher Class No Roti No No  
## High\_Cholestrol Phy\_Activity Obese Diabetes  
## 3112 Yes Yes N No  
## 6509 No Yes N No  
## 1747 No Yes N No  
## 9277 Yes Yes N No  
## 9209 Yes Yes N No  
## 5246 Yes Yes N No

print(nrow(d2))

## [1] 1000

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

##   
## Dataset file 1 Diabetes Count :

## dd  
## No Yes   
## 0 1000

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

##   
## Dataset file 2 Diabetes 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))  
rnd

## [1] 1 5 6 7 8 9 12 13 15 16 17 18 19 20  
## [15] 21 23 24 25 26 29 30 31 32 33 36 37 38 39  
## [29] 40 41 42 43 44 45 46 47 49 50 51 53 54 56  
## [43] 58 59 60 61 62 63 64 66 68 69 70 71 72 73  
## [57] 79 80 81 82 83 84 85 87 88 89 90 91 92 93  
## [71] 94 95 99 100 102 103 106 108 109 112 113 114 116 117  
## [85] 119 120 121 123 124 126 127 128 129 130 131 132 133 135  
## [99] 136 137 138 143 144 145 146 147 148 150 151 152 153 154  
## [113] 158 159 160 162 163 164 165 166 167 168 169 170 171 172  
## [127] 173 174 175 176 177 178 181 182 183 184 185 186 188 191  
## [141] 192 193 196 197 198 200 201 204 205 206 207 208 209 210  
## [155] 212 214 215 216 217 220 221 223 225 226 228 229 231 232  
## [169] 234 235 236 237 238 240 241 242 243 244 246 247 248 249  
## [183] 250 252 254 255 256 257 258 261 262 264 265 266 268 269  
## [197] 270 271 273 274 275 276 279 281 282 284 285 286 287 288  
## [211] 289 290 291 293 294 296 297 298 300 301 302 304 305 307  
## [225] 312 313 314 317 318 319 320 321 323 324 325 327 328 329  
## [239] 330 332 334 335 337 338 341 342 343 344 345 347 348 349  
## [253] 350 351 352 353 354 356 357 358 360 361 363 364 367 368  
## [267] 369 370 372 373 374 375 376 379 380 381 382 384 385 387  
## [281] 388 389 392 393 394 395 396 397 398 399 401 402 403 405  
## [295] 406 408 409 410 411 412 414 416 417 420 422 423 424 425  
## [309] 426 427 429 431 433 434 435 437 439 441 442 443 444 445  
## [323] 446 449 450 451 452 453 456 457 458 459 460 461 462 464  
## [337] 465 467 468 470 471 472 473 475 477 478 480 482 483 484  
## [351] 485 486 490 491 492 493 494 496 497 498 501 502 503 506  
## [365] 509 510 513 514 516 517 523 524 525 527 528 530 531 532  
## [379] 533 534 535 536 537 539 540 542 544 545 546 547 548 549  
## [393] 550 551 553 554 555 556 557 560 563 565 566 567 569 570  
## [407] 571 572 573 574 576 578 579 580 581 583 584 585 586 588  
## [421] 590 591 592 593 594 595 596 597 598 601 603 604 605 608  
## [435] 609 610 612 614 615 617 618 619 620 622 623 624 626 627  
## [449] 628 629 634 635 636 637 638 639 640 642 644 647 649 650  
## [463] 651 652 654 657 658 659 660 661 662 663 666 667 668 669  
## [477] 670 672 674 676 677 679 680 681 686 687 689 691 692 695  
## [491] 696 697 698 699 700 701 702 704 705 706 708 711 712 713  
## [505] 715 716 717 718 719 721 722 723 724 726 727 728 730 731  
## [519] 732 733 734 738 739 741 742 745 746 748 749 750 751 752  
## [533] 753 754 756 757 758 759 762 763 766 769 770 771 772 773  
## [547] 774 775 776 777 779 780 781 783 784 786 788 789 790 791  
## [561] 793 794 795 796 797 798 800 802 804 806 809 810 814 815  
## [575] 816 818 819 820 821 824 831 832 833 834 835 837 838 839  
## [589] 840 841 842 843 844 845 848 849 852 853 855 856 857 858  
## [603] 859 860 862 863 864 867 868 871 872 873 874 875 876 877  
## [617] 879 880 881 882 883 884 885 887 888 889 891 893 894 895  
## [631] 898 900 901 902 904 907 908 909 911 914 916 918 919 921  
## [645] 923 926 927 928 929 931 932 933 935 936 937 938 939 943  
## [659] 946 948 949 950 951 952 953 954 955 957 958 959 960 961  
## [673] 962 963 964 966 967 969 970 971 972 974 975 976 977 979  
## [687] 980 982 983 984 987 989 990 992 993 994 996 997 998 1000

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

### Merge d1\_train (Diabetes dataset) and d2\_train (Non-Diabetes 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 Diabetes and Non-Diabetes

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

##   
## Diabetes Count Yes: 700 No: 0

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

##   
## Diabetes Count Yes: 0 No: 700

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

##   
## Diabetes Count Yes: 300 No: 0

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

##   
## Train dataset file Diabetes Count : 1400

table(d\_train$Diabetes)

##   
## No Yes   
## 700 700

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

##   
## Test dataset file Diabetes Count : 600

table(d\_test$Diabetes)

##   
## 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"

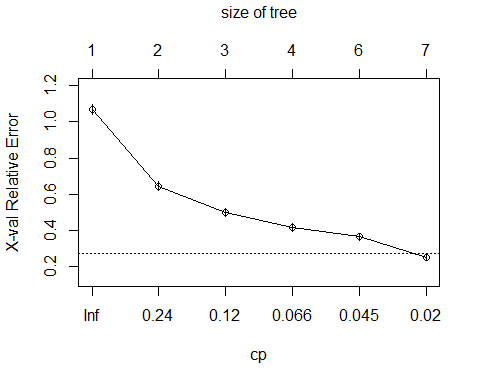
for (e in 1:length(diabetes\_vector)){  
eda\_functions("Diabetes",d\_train,d\_train[,"Diabetes"],d\_train[,e],diabetes\_vector[e])  
 }

##   
##   
## Performing mlbenchxtab   
##   
## Diabetes Vs Fin\_Group   
##   
## column Higher Class Lower Middle Class Middle Class  
## target   
## No 210 388 102  
## Yes 218 343 139  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target Higher Class Lower Middle Class Middle Class  
## No 0.3000000 0.5542857 0.1457143  
## Yes 0.3114286 0.4900000 0.1985714  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target Higher Class Lower Middle Class Middle Class  
## No 0.4906542 0.5307798 0.4232365  
## Yes 0.5093458 0.4692202 0.5767635  
##   
##   
## Performing Chi Square Test for Diabetes and Fin\_Group  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Fin\_Group  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 210 388 102  
## Yes 218 343 139  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 8.6, df = 2, p-value = 0.01357  
##   
##   
## 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 210   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test  
##   
## data: tab1  
## X-squared = 8.6002, df = 2, p-value = 0.01357  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.01356714   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes 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   
##   
## Diabetes Vs Family\_history   
##   
## column No Yes  
## target   
## No 363 337  
## Yes 263 437  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target No Yes  
## No 0.5185714 0.4814286  
## Yes 0.3757143 0.6242857  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.5798722 0.4354005  
## Yes 0.4201278 0.5645995  
##   
##   
## Performing Chi Square Test for Diabetes and Family\_history  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Family\_history  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 363 337  
## Yes 263 437  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 28.894, df = 1, p-value = 7.644e-08  
##   
##   
## 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 263   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 28.319, df = 1, p-value = 1.029e-07  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.028631e-07   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes 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   
##   
## Diabetes Vs Staple\_Food   
##   
## column Pizza Rice Roti  
## target   
## No 9 4 687  
## Yes 11 7 682  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target Pizza Rice Roti  
## No 0.012857143 0.005714286 0.981428571  
## Yes 0.015714286 0.010000000 0.974285714  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target Pizza Rice Roti  
## No 0.4500000 0.3636364 0.5018262  
## Yes 0.5500000 0.6363636 0.4981738  
##   
##   
## Performing Chi Square Test for Diabetes and Staple\_Food  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Staple\_Food  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 4 687  
## Yes 11 7 682  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1.0364, df = 2, p-value = 0.5956  
##   
##   
## 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 4   
##   
## <------ Chi square with simulation for p value --------->   
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##   
## data: tab1  
## X-squared = 1.0364, df = NA, p-value = 0.5562  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained with simulation for p-value : 0.5562219   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes 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   
##   
## Diabetes Vs Alcohol   
##   
## column No Yes  
## target   
## No 292 408  
## Yes 320 380  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target No Yes  
## No 0.4171429 0.5828571  
## Yes 0.4571429 0.5428571  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.4771242 0.5177665  
## Yes 0.5228758 0.4822335  
##   
##   
## Performing Chi Square Test for Diabetes and Alcohol  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Alcohol  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 292 408  
## Yes 320 380  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 2.276, df = 1, p-value = 0.1314  
##   
##   
## 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 292   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 2.1163, df = 1, p-value = 0.1457  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.1457379   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes are independent 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   
##   
## Diabetes Vs Smoke   
##   
## column No Yes  
## target   
## No 418 282  
## Yes 256 444  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target No Yes  
## No 0.5971429 0.4028571  
## Yes 0.3657143 0.6342857  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.6201780 0.3884298  
## Yes 0.3798220 0.6115702  
##   
##   
## Performing Chi Square Test for Diabetes and Smoke  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Smoke  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 418 282  
## Yes 256 444  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 75.09, df = 1, p-value = 4.505e-18  
##   
##   
## 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 256   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 74.162, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 7.195087e-18   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes 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   
##   
## Diabetes Vs High\_Cholestrol   
##   
## column No Yes  
## target   
## No 517 183  
## Yes 268 432  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target No Yes  
## No 0.7385714 0.2614286  
## Yes 0.3828571 0.6171429  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.6585987 0.2975610  
## Yes 0.3414013 0.7024390  
##   
##   
## Performing Chi Square Test for Diabetes and High\_Cholestrol  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of High\_Cholestrol  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 517 183  
## Yes 268 432  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 179.8, df = 1, p-value = 5.368e-41  
##   
##   
## 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 183   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 178.36, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 1.107847e-40   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes 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   
##   
## Diabetes Vs Phy\_Activity   
##   
## column No Yes  
## target   
## No 275 425  
## Yes 452 248  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target No Yes  
## No 0.3928571 0.6071429  
## Yes 0.6457143 0.3542857  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target No Yes  
## No 0.3782669 0.6315007  
## Yes 0.6217331 0.3684993  
##   
##   
## Performing Chi Square Test for Diabetes and Phy\_Activity  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Phy\_Activity  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 275 425  
## Yes 452 248  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 89.64, df = 1, p-value = 2.85e-21  
##   
##   
## 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 248   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 88.635, df = 1, p-value < 2.2e-16  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 4.748797e-21   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes 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   
##   
## Diabetes Vs Obese   
##   
## column N Y  
## target   
## No 608 92  
## Yes 590 110  
##   
##   
## Within each level of predictor variable values, % of Diabetes patients   
##   
## column  
## target N Y  
## No 0.8685714 0.1314286  
## Yes 0.8428571 0.1571429  
##   
##   
## Diabetes Vs predictor variable values   
##   
## column  
## target N Y  
## No 0.5075125 0.4554455  
## Yes 0.4924875 0.5445545  
##   
##   
## Performing Chi Square Test for Diabetes and Obese  
##   
## Step 1: State the hypothesis   
##   
##   
## Null hypothesis : The relative proportions of Diabetes are independent of Obese  
##   
##   
## Alternative hypothesis : The relative proportions of Diabetes 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 608 92  
## Yes 590 110  
## Call: xtabs(formula = ~(target + column), data = data1)  
## Number of cases in table: 1400   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1.8744, df = 1, p-value = 0.171  
##   
##   
## 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 92   
##   
## <------ Chi square without simulation for p value --------->   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tab1  
## X-squared = 1.6719, df = 1, p-value = 0.196  
##   
##   
##   
## Step 4: Interpret the results   
##   
##   
##   
## ### ---------------------------------------------------------  
## P value obtained without simulation for p-value : 0.1960004   
##   
##   
## <-- Decision: -->  
## The relative proportions of Diabetes are independent of Obese at 5% level of significance   
##   
##   
## This p-value is the probability of observing a sample statistic as extreme as the test statistic.   
##   
## We compare the p-value to the significance level and reject the null hypothesis when the p-value is less than the significance level else accept.

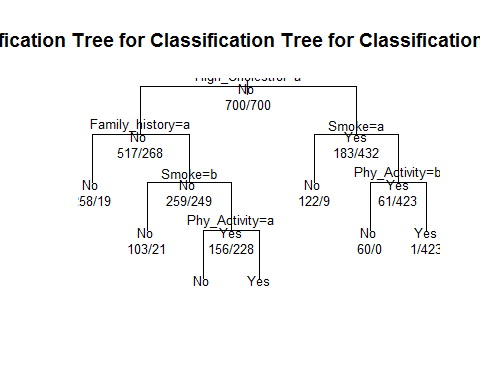
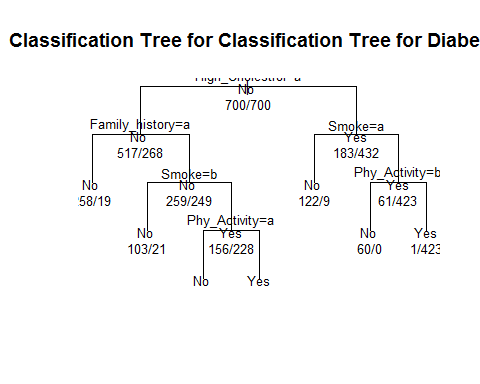
## CART

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

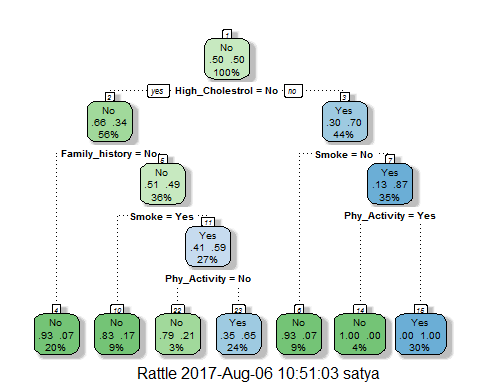
##   
## Classification tree:  
## rpart(formula = formula, data = d\_train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Family\_history High\_Cholestrol Phy\_Activity Smoke   
##   
## Root node error: 700/1400 = 0.5  
##   
## n= 1400   
##   
## CP nsplit rel error xerror xstd  
## 1 0.355714 0 1.00000 1.06857 0.026663  
## 2 0.161429 1 0.64429 0.64429 0.024978  
## 3 0.085714 2 0.48286 0.49857 0.023123  
## 4 0.051429 3 0.39714 0.41714 0.021717  
## 5 0.040000 5 0.29429 0.36571 0.020662  
## 6 0.010000 6 0.25429 0.25429 0.017807



## Call:  
## rpart(formula = formula, data = d\_train, method = "class")  
## n= 1400   
##   
## CP nsplit rel error xerror xstd  
## 1 0.35571429 0 1.0000000 1.0685714 0.02666322  
## 2 0.16142857 1 0.6442857 0.6442857 0.02497810  
## 3 0.08571429 2 0.4828571 0.4985714 0.02312342  
## 4 0.05142857 3 0.3971429 0.4171429 0.02171697  
## 5 0.04000000 5 0.2942857 0.3657143 0.02066193  
## 6 0.01000000 6 0.2542857 0.2542857 0.01780670  
##   
## Variable importance  
## Smoke Phy\_Activity High\_Cholestrol Family\_history   
## 38 31 17 14   
##   
## Node number 1: 1400 observations, complexity param=0.3557143  
## predicted class=No expected loss=0.5 P(node) =1  
## class counts: 700 700  
## probabilities: 0.500 0.500   
## left son=2 (785 obs) right son=3 (615 obs)  
## Primary splits:  
## High\_Cholestrol splits as LR, improve=89.898400, (0 missing)  
## Phy\_Activity splits as RL, improve=44.822400, (0 missing)  
## Smoke splits as LR, improve=37.543220, (0 missing)  
## Family\_history splits as LR, improve=14.447170, (0 missing)  
## Fin\_Group splits as LLR, improve= 3.430844, (0 missing)  
## Surrogate splits:  
## Phy\_Activity splits as RL, agree=0.743, adj=0.415, (0 split)  
## Smoke splits as LR, agree=0.734, adj=0.393, (0 split)  
## Family\_history splits as RL, agree=0.612, adj=0.117, (0 split)  
## Staple\_Food splits as LRL, agree=0.561, adj=0.002, (0 split)  
##   
## Node number 2: 785 observations, complexity param=0.05142857  
## predicted class=No expected loss=0.3414013 P(node) =0.5607143  
## class counts: 517 268  
## probabilities: 0.659 0.341   
## left son=4 (277 obs) right son=5 (508 obs)  
## Primary splits:  
## Family\_history splits as LR, improve=63.7138400, (0 missing)  
## Smoke splits as RL, improve=45.3644100, (0 missing)  
## Phy\_Activity splits as LR, improve=36.0856400, (0 missing)  
## Fin\_Group splits as LLR, improve=18.5328700, (0 missing)  
## Alcohol splits as RL, improve= 0.6359643, (0 missing)  
## Surrogate splits:  
## Phy\_Activity splits as LR, agree=0.652, adj=0.014, (0 split)  
##   
## Node number 3: 615 observations, complexity param=0.1614286  
## predicted class=Yes expected loss=0.297561 P(node) =0.4392857  
## class counts: 183 432  
## probabilities: 0.298 0.702   
## left son=6 (131 obs) right son=7 (484 obs)  
## Primary splits:  
## Smoke splits as LR, improve=133.7054000, (0 missing)  
## Phy\_Activity splits as RL, improve=132.8831000, (0 missing)  
## Alcohol splits as RL, improve= 0.9061767, (0 missing)  
## Obese splits as LR, improve= 0.7898675, (0 missing)  
## Fin\_Group splits as RLL, improve= 0.5847998, (0 missing)  
## Surrogate splits:  
## Phy\_Activity splits as RL, agree=0.793, adj=0.031, (0 split)  
##   
## Node number 4: 277 observations  
## predicted class=No expected loss=0.06859206 P(node) =0.1978571  
## class counts: 258 19  
## probabilities: 0.931 0.069   
##   
## Node number 5: 508 observations, complexity param=0.05142857  
## predicted class=No expected loss=0.4901575 P(node) =0.3628571  
## class counts: 259 249  
## probabilities: 0.510 0.490   
## left son=10 (124 obs) right son=11 (384 obs)  
## Primary splits:  
## Smoke splits as RL, improve=33.76448000, (0 missing)  
## Phy\_Activity splits as LR, improve=22.67744000, (0 missing)  
## Fin\_Group splits as LLR, improve= 7.51036700, (0 missing)  
## Alcohol splits as RL, improve= 3.21276800, (0 missing)  
## Staple\_Food splits as RRL, improve= 0.00197641, (0 missing)  
## Surrogate splits:  
## Phy\_Activity splits as LR, agree=0.795, adj=0.161, (0 split)  
##   
## Node number 6: 131 observations  
## predicted class=No expected loss=0.06870229 P(node) =0.09357143  
## class counts: 122 9  
## probabilities: 0.931 0.069   
##   
## Node number 7: 484 observations, complexity param=0.08571429  
## predicted class=Yes expected loss=0.1260331 P(node) =0.3457143  
## class counts: 61 423  
## probabilities: 0.126 0.874   
## left son=14 (60 obs) right son=15 (424 obs)  
## Primary splits:  
## Phy\_Activity splits as RL, improve=104.62870000, (0 missing)  
## Fin\_Group splits as RLR, improve= 0.54669420, (0 missing)  
## Obese splits as LR, improve= 0.07778531, (0 missing)  
## Staple\_Food splits as RRL, improve= 0.06443730, (0 missing)  
## Alcohol splits as RL, improve= 0.01197519, (0 missing)  
##   
## Node number 10: 124 observations  
## predicted class=No expected loss=0.1693548 P(node) =0.08857143  
## class counts: 103 21  
## probabilities: 0.831 0.169   
##   
## Node number 11: 384 observations, complexity param=0.04  
## predicted class=Yes expected loss=0.40625 P(node) =0.2742857  
## class counts: 156 228  
## probabilities: 0.406 0.594   
## left son=22 (48 obs) right son=23 (336 obs)  
## Primary splits:  
## Phy\_Activity splits as LR, improve=16.2976200, (0 missing)  
## Fin\_Group splits as LLR, improve= 3.9409090, (0 missing)  
## Alcohol splits as RL, improve= 2.0561390, (0 missing)  
## Obese splits as LR, improve= 0.2228395, (0 missing)  
##   
## Node number 14: 60 observations  
## predicted class=No expected loss=0 P(node) =0.04285714  
## class counts: 60 0  
## probabilities: 1.000 0.000   
##   
## Node number 15: 424 observations  
## predicted class=Yes expected loss=0.002358491 P(node) =0.3028571  
## class counts: 1 423  
## probabilities: 0.002 0.998   
##   
## Node number 22: 48 observations  
## predicted class=No expected loss=0.2083333 P(node) =0.03428571  
## class counts: 38 10  
## probabilities: 0.792 0.208   
##   
## Node number 23: 336 observations  
## predicted class=Yes expected loss=0.3511905 P(node) =0.24  
## class counts: 118 218  
## probabilities: 0.351 0.649



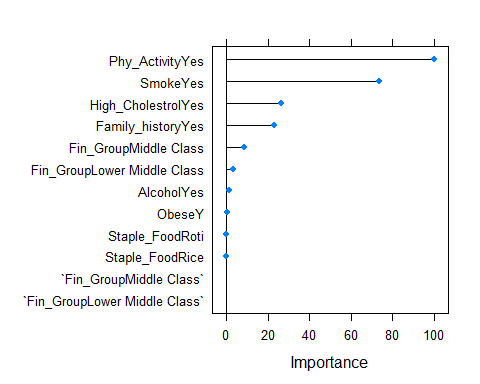
fancyRpartPlot(pfit)



varImp(object=fit)

## rpart variable importance  
##   
## Overall  
## Phy\_ActivityYes 100.00000  
## SmokeYes 73.40354  
## High\_CholestrolYes 26.35565  
## Family\_historyYes 22.91458  
## Fin\_GroupMiddle Class 8.64096  
## Fin\_GroupLower Middle Class 3.13072  
## AlcoholYes 1.20756  
## ObeseY 0.25437  
## Staple\_FoodRoti 0.13557  
## Staple\_FoodRice 0.03757  
## `Fin\_GroupMiddle Class` 0.00000  
## `Fin\_GroupLower Middle Class` 0.00000

plot(varImp(fit))



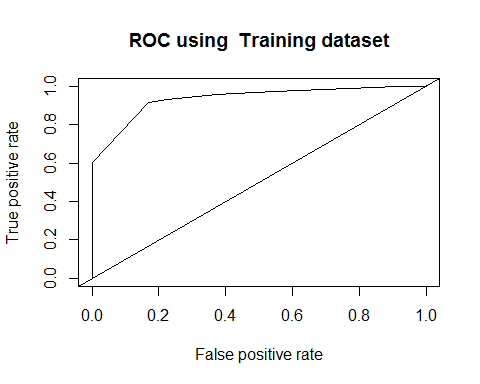
#### Interpretation of Variable Importance for diabetes

###### Above graph shows the top 10 important variables.Physical activity,smoking,high cholestrol and family history are the top influencing predictors.

###### Financial group and Alcohol comsumption also have significant importance.

### Performance measures using training data

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



print(perf\_measures1)

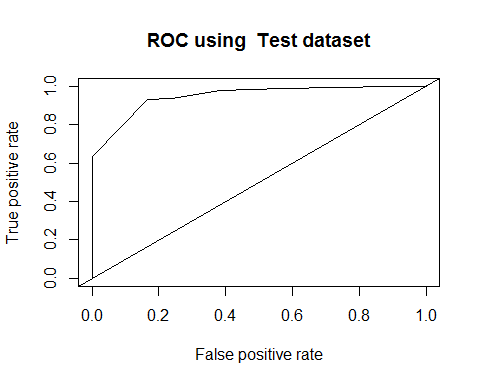
## Dataset KS auc gini OA  
## 1 Training dataset 0.7457143 0.9354122 0.4354122 0.8728571

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

##   
## OA is overall accuracy

### Performance measures using test data

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



print(perf\_measures2)

## Dataset KS auc gini OA  
## 1 Test dataset 0.7666667 0.9479611 0.4306094 0.8833333

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

##   
## OA is overall accuracy

### Logistic Regression

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

##   
## Call:  
## glm(formula = formula, family = "binomial", data = d\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.22864 -0.95941 -0.06611 1.00334 2.17366   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.318497 0.544110 -2.423 0.01538 \*   
## Fin\_GroupLower Middle Class -0.048254 0.140424 -0.344 0.73112   
## Fin\_GroupMiddle Class 0.837182 0.189552 4.417 1.00e-05 \*\*\*  
## Family\_historyYes 1.302732 0.138803 9.386 < 2e-16 \*\*\*  
## Staple\_FoodRice 0.702330 0.847708 0.829 0.40738   
## Staple\_FoodRoti 0.006088 0.507671 0.012 0.99043   
## AlcoholYes -0.255713 0.122956 -2.080 0.03755 \*   
## SmokeYes 0.423638 0.147235 2.877 0.00401 \*\*   
## High\_CholestrolYes 1.568759 0.149698 10.479 < 2e-16 \*\*\*  
## Phy\_ActivityYes -0.647083 0.150154 -4.309 1.64e-05 \*\*\*  
## ObeseY 0.174333 0.172493 1.011 0.31218   
## ---  
## 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: 1604.7 on 1389 degrees of freedom  
## AIC: 1626.7  
##   
## Number of Fisher Scoring iterations: 4

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

###### Physical activity,high cholestrol,family history and financial group are the top influencing predictors.

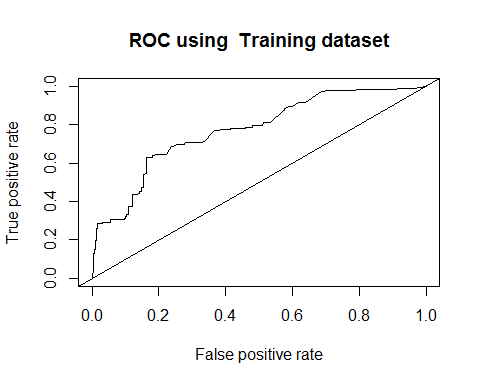
###### Smoking and Alcohol comsumption also have significant importance.significance level considered is 95%

perf <- pR2(logit)  
print(perf)

## llh llhNull G2 McFadden r2ML   
## -802.3273547 -970.4060528 336.1573962 0.1732045 0.2134606   
## r2CU   
## 0.2846141

### Logistic Regression Performance measures using training data

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



print(perf\_measures1)

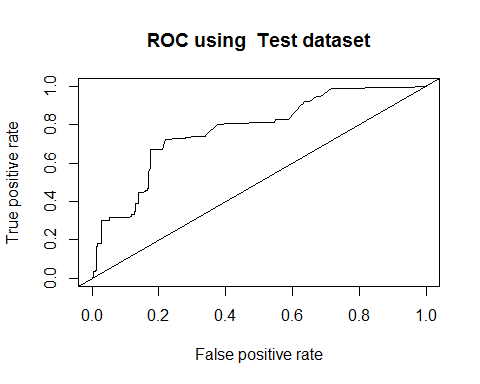
## Dataset auc McFadden OA  
## 1 Training dataset 0.7719918 0.1732045 0.483871

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

##   
## OA is overall accuracy

### Logistic Regression Performance measures using test data

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



print(perf\_measures2)

## Dataset auc McFadden OA  
## 1 Test dataset 0.7741556 0.1732045 0.25

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

##   
## OA is overall accuracy

varImp(logit, scale = FALSE)

## Overall  
## Fin\_GroupLower Middle Class 0.34363038  
## Fin\_GroupMiddle Class 4.41662832  
## Family\_historyYes 9.38551044  
## Staple\_FoodRice 0.82850474  
## Staple\_FoodRoti 0.01199143  
## AlcoholYes 2.07970128  
## SmokeYes 2.87729635  
## High\_CholestrolYes 10.47949576  
## Phy\_ActivityYes 4.30945466  
## ObeseY 1.01066723

### End of R script

```