Sai charan Talipineni sat122@pitt.edu

Homework 2

sai charan talipineni

31 January 2017

```
#**Question 1:**
setwd("D:/semester/2nd sem/DATA_MINING/hw2")
audit<-read.csv("audit.csv")
audit[1:2,]
```

```
##        ID Age Employment Education   Marital Occupation Income Gender
## 1 1004641  38    Private   College Unmarried    Service  81838 Female
## 2 1010229  35    Private Associate    Absent  Transport  72099   Male
##   Deductions Hours RISK_Adjustment TARGET_Adjusted
## 1          0    72               0               0
## 2          0    30               0               0
```

```
#RISK_Adjustment, TARGET_Adjusted are the response variables and the other
variables including Age, #Employment, Education, Marital, Occupation, Income,
Gender, Deductions, Hours are predictors.

#**Missing values:**

sapply(audit, function(x) sum(is.na(x)))
```

```
##              ID             Age      Employment       Education
##               0               0             100               0
##         Marital      Occupation          Income          Gender
##               0             101               0               0
##      Deductions           Hours RISK_Adjustment TARGET_Adjusted
##               0               0               0               0
```

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

mode_employ<-getmode(audit$Employment)
mode_occup<-getmode(audit$Occupation)

audit$Employment[is.na(audit$Employment)] <- mode_employ
audit$Occupation[is.na(audit$Occupation)] <- mode_occup
```

```
#**Question 2:**
#**(a)**
summary(audIt)
```

```
##        ID                Age              Employment        Education
##  Min.   :1004641   Min.   :17.00   Private   :1511   HSgrad    :660
##  1st Qu.:3437052   1st Qu.:28.00   Consultant: 148   College   :442
##  Median :5638451   Median :37.00   PSLocal   : 119   Bachelor  :345
##  Mean   :5624348   Mean   :38.62   SelfEmp   :  79   Master    :102
##  3rd Qu.:7876535   3rd Qu.:48.00   PSState   :  72   Vocational: 86
##  Max.   :9996101   Max.   :90.00   PSFederal :  69   Yr11      : 74
##                                    (Other)   :   2   (Other)   :291
##                   Marital           Occupation      Income
##  Absent               :669   Executive   :390   Min.   :   609.7
##  Divorced             :266   Professional:247   1st Qu.: 34433.1
##  Married              :917   Clerical    :232   Median : 59768.9
##  Married-spouse-absent: 22   Repair      :225   Mean   : 84688.5
##  Unmarried            : 67   Service     :210   3rd Qu.:113842.9
##  Widowed              : 59   Sales       :206   Max.   :481259.5
##                              (Other)     :490
##     Gender       Deductions          Hours       RISK_Adjustment
##  Female: 632   Min.   :   0.00   Min.   : 1.00   Min.   : -1453
##  Male  :1368   1st Qu.:   0.00   1st Qu.:38.00   1st Qu.:     0
##                Median :   0.00   Median :40.00   Median :     0
##                Mean   :  67.57   Mean   :40.07   Mean   :  2021
##                3rd Qu.:   0.00   3rd Qu.:45.00   3rd Qu.:     0
##                Max.   :2904.00   Max.   :99.00   Max.   :112243
##
##  TARGET_Adjusted
##  0:1537
##  1: 463
##
##
##
##
##
```

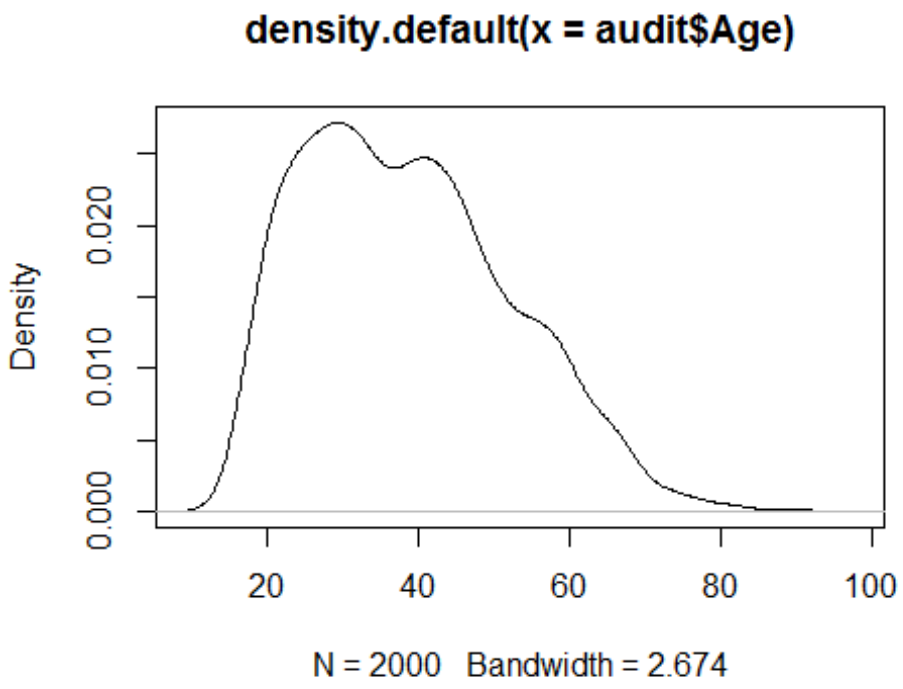
```
#From the above summary we can know that Age, Income, Deductions, Hours,
RISK_Adjustment are numerical variables. The summary table is as follows.

Age = c(summary(audit$Age), sd(audit$Age))
Income = c(summary(audit$Income), sd(audit$Income))
Deductions = c(summary(audit$Deductions), sd(audit$Deductions))
Hours = c(summary(audit$Hours), sd(audit$Hours))
RISK_Adjustment = c(summary(audit$RISK_Adjustment),
sd(audit$RISK_Adjustment))
result = rbind(Age, Income, Deductions, Hours,RISK_Adjustment)
result = as.data.frame(result)
colnames(result)[7] = c("sd")
result
```

```
##                     Min. 1st Qu. Median     Mean 3rd Qu.    Max.          sd
## Age                 17.0      28     37    38.62      48     90    13.58475
## Income             609.7   34430  59770 84690.00  113800 481300 69621.64450
## Deductions           0.0       0      0    67.57       0   2904   340.70470
## Hours                1.0      38     40    40.07      45     99    12.15372
## RISK_Adjustment  -1453.0       0      0  2021.00       0 112200  8341.87229
```

```
#**(b)**
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.3.2
```

```
plot(density(audit$Age))
```

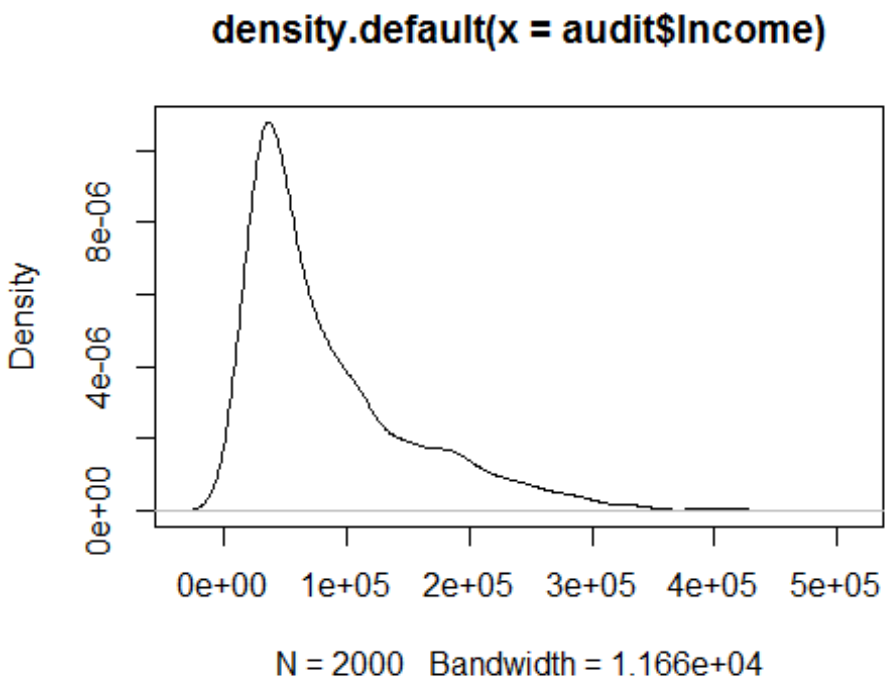



```
skewness(audit$Age)
```

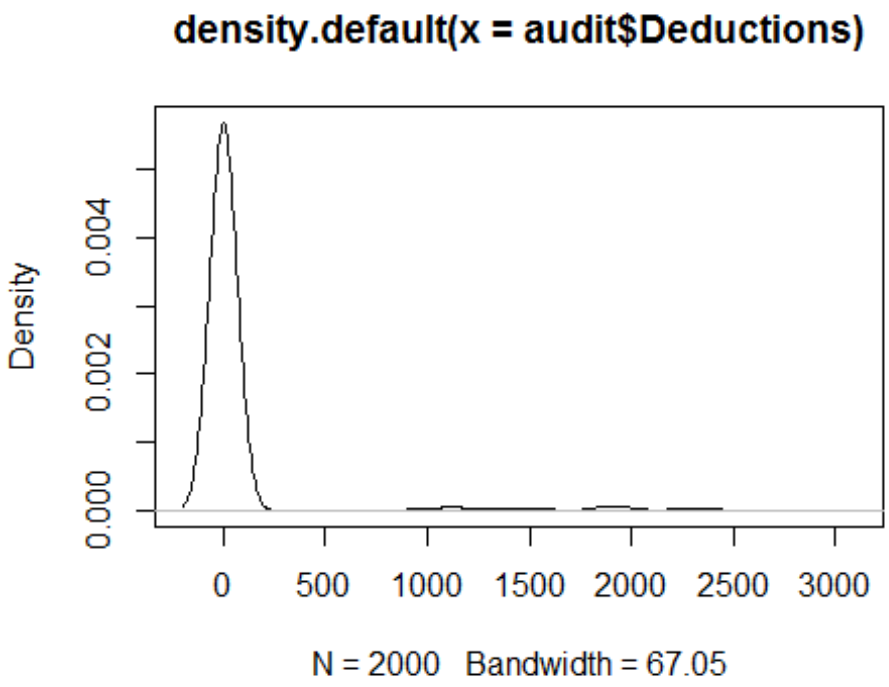
```
## [1] 0.4990696
```

```
#right skewed
```

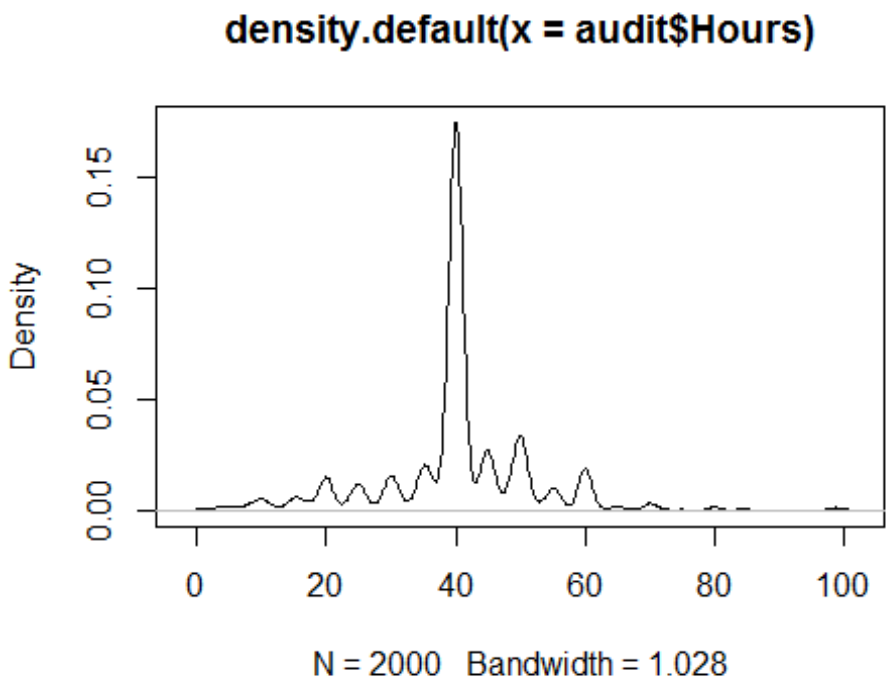
```
plot(density(audit$Income))
```



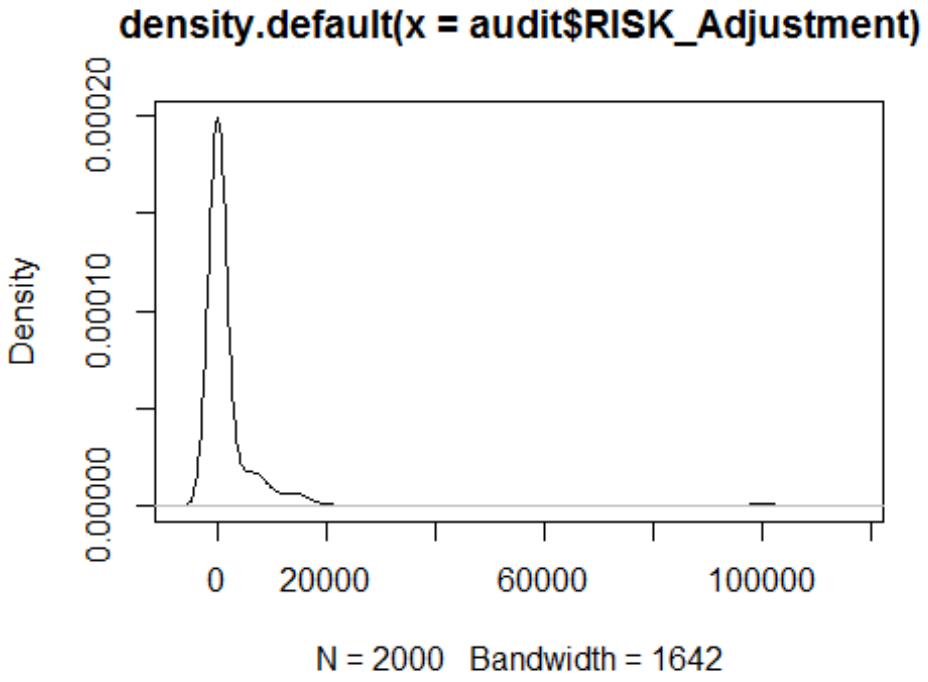

```
skewness(audit$Income)

## [1] 1.488821

#right skewed

plot(density(audit$Deductions))
```



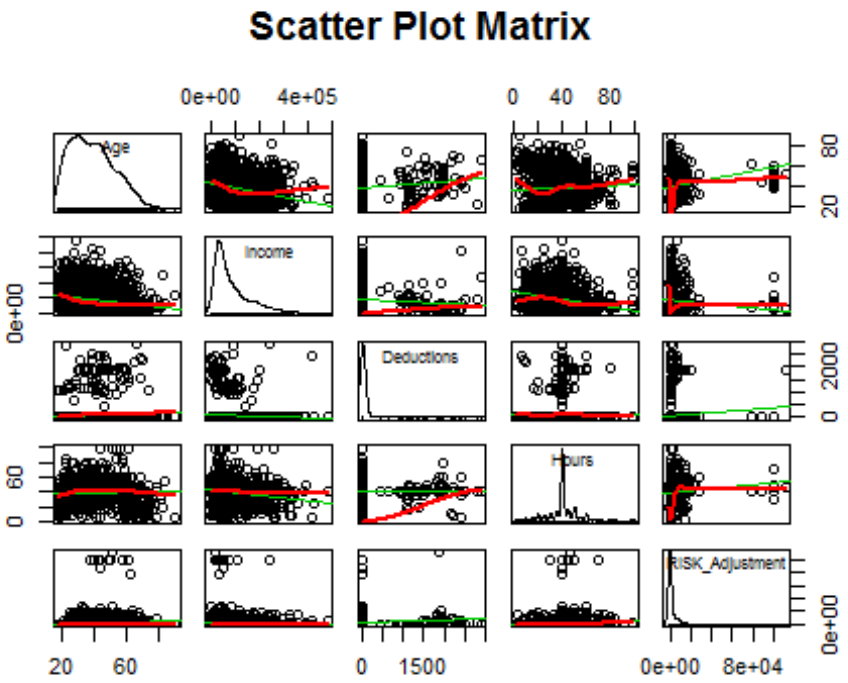

```
skewness(audit$Deductions)

## [1] 5.249432

#right skewed

plot(density(audit$Hours))
```



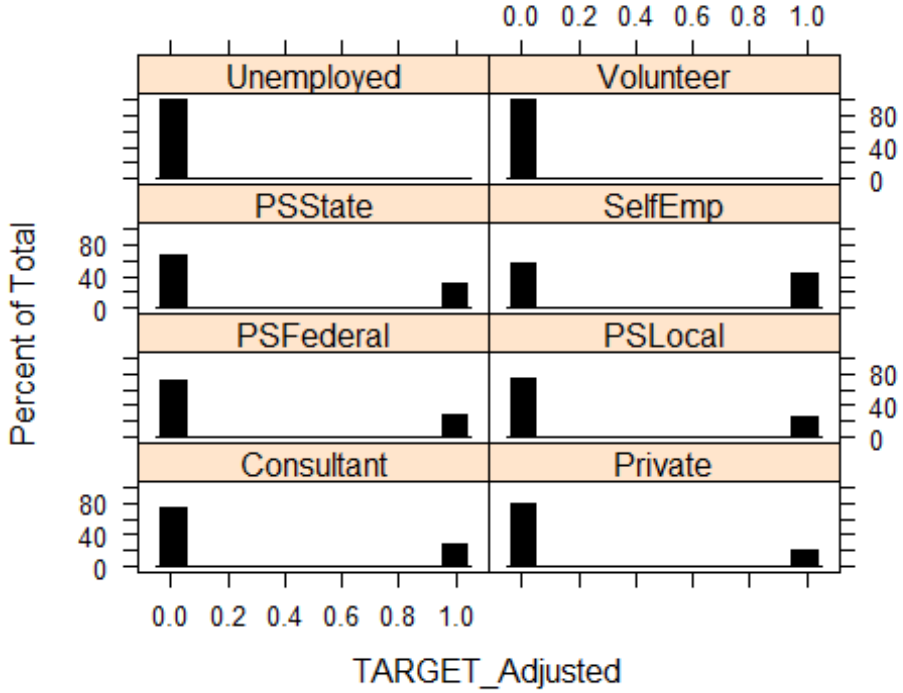

```
skewness(audit$Hours)

## [1] 0.1323312

#right skewed

plot(density(audit$RISK_Adjustment))
skewness(audit$RISK_Adjustment)

## [1] 9.591535

#right skewed


#correlation
library(car)

## Warning: package 'car' was built under R version 3.3.2
```



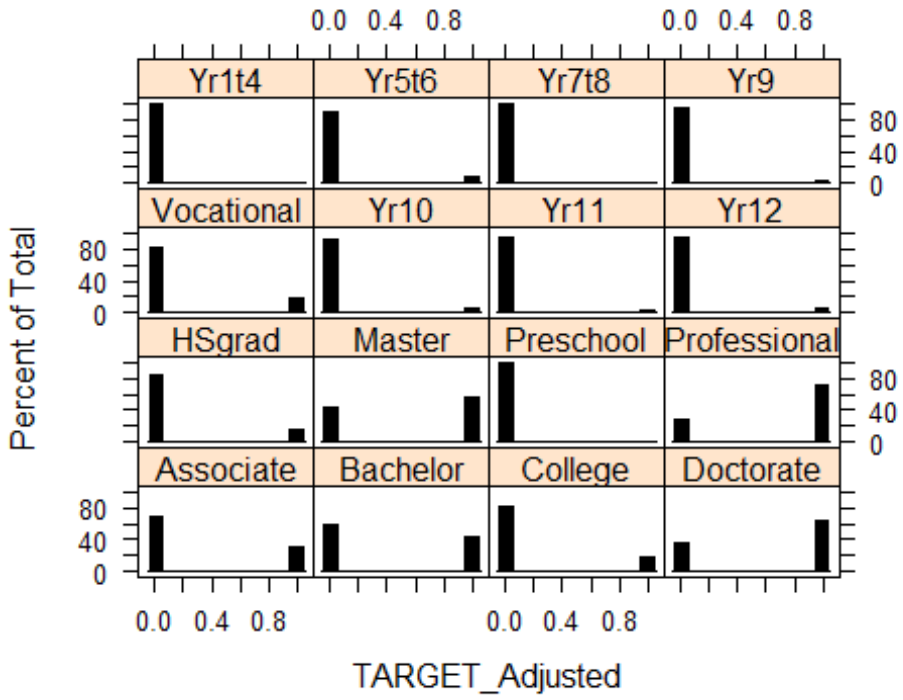

```
dt = audit[,c('Age','Income','Deductions','Hours','RISK_Adjustment')]
cor(dt)

##                          Age      Income  Deductions       Hours
## Age              1.00000000 -0.22686777  0.08399899  0.04236487
## Income          -0.22686777  1.00000000 -0.05734147 -0.21269065
## Deductions       0.08399899 -0.05734147  1.00000000  0.01365124
## Hours            0.04236487 -0.21269065  0.01365124  1.00000000
## RISK_Adjustment  0.12274079 -0.08339021  0.06559720  0.09060735
##                 RISK_Adjustment
## Age                  0.12274079
## Income              -0.08339021
## Deductions           0.06559720
## Hours                0.09060735
## RISK_Adjustment      1.00000000

#scatterplot
suppressWarnings(scatterplotMatrix(dt, spread = FALSE, lty.smooth = 2, main =
'Scatter Plot Matrix'))
```



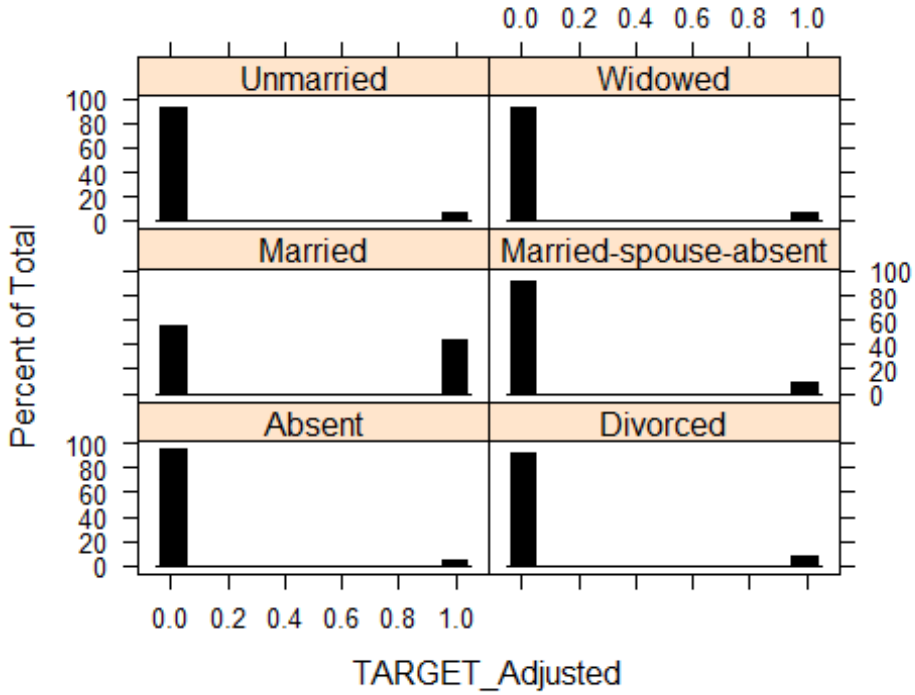

```
#**(c)**
library(lattice)
library(nutshell)

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

## Loading required package: nutshell.bbdb

## Warning: package 'nutshell.bbdb' was built under R version 3.3.2

## Loading required package: nutshell.audioscrobbler

## Warning: package 'nutshell.audioscrobbler' was built under R version 3.3.2

histogram(~TARGET_Adjusted|Employment,data=audit,layout=c(2,4),col="black")
```



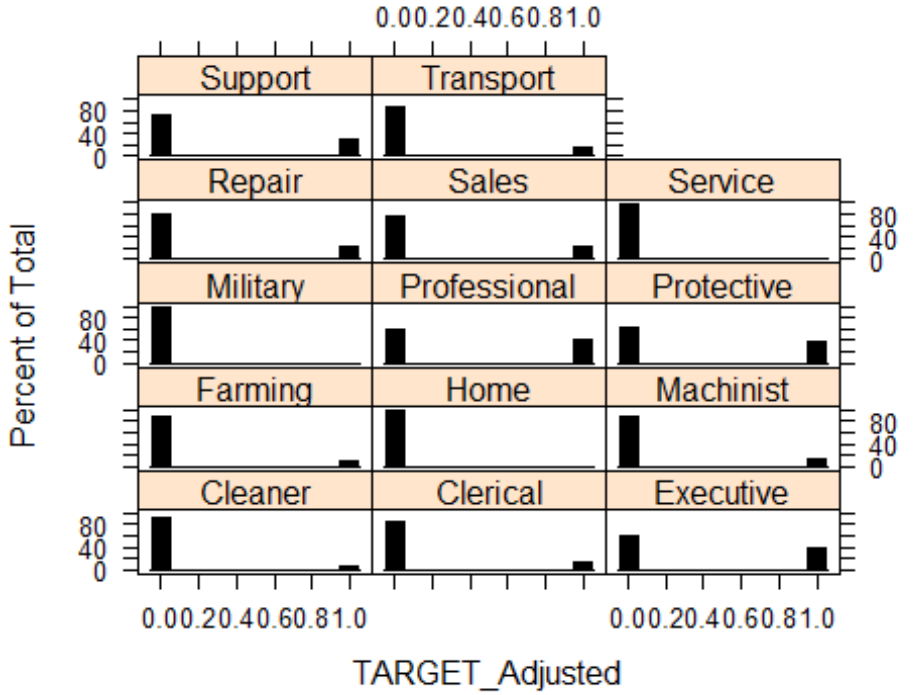

```
histogram(~TARGET_Adjusted|Education,data=audit,layout=c(4,4),col="black")
```



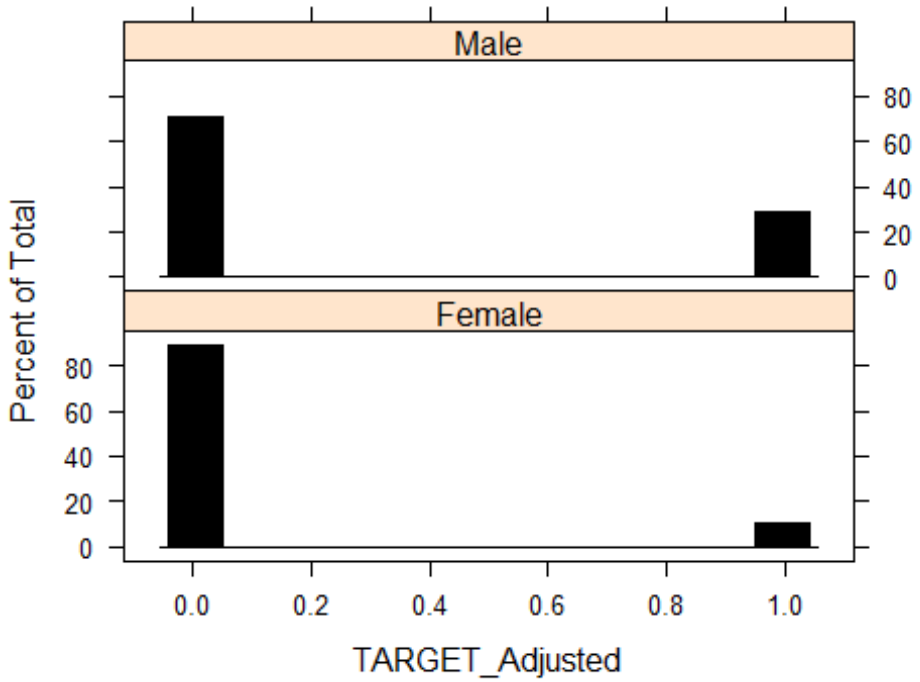

```
histogram(~TARGET_Adjusted|Marital,data=audit,layout=c(2,3),col="black")
```



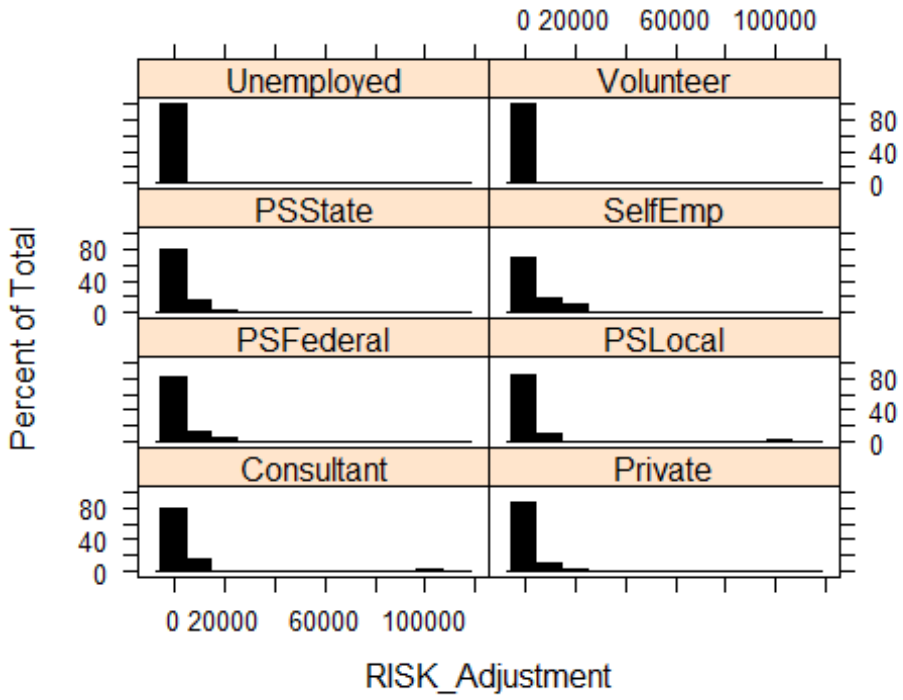

```
histogram(~TARGET_Adjusted|Occupation,data=audit,layout=c(3,5),col="black")
```



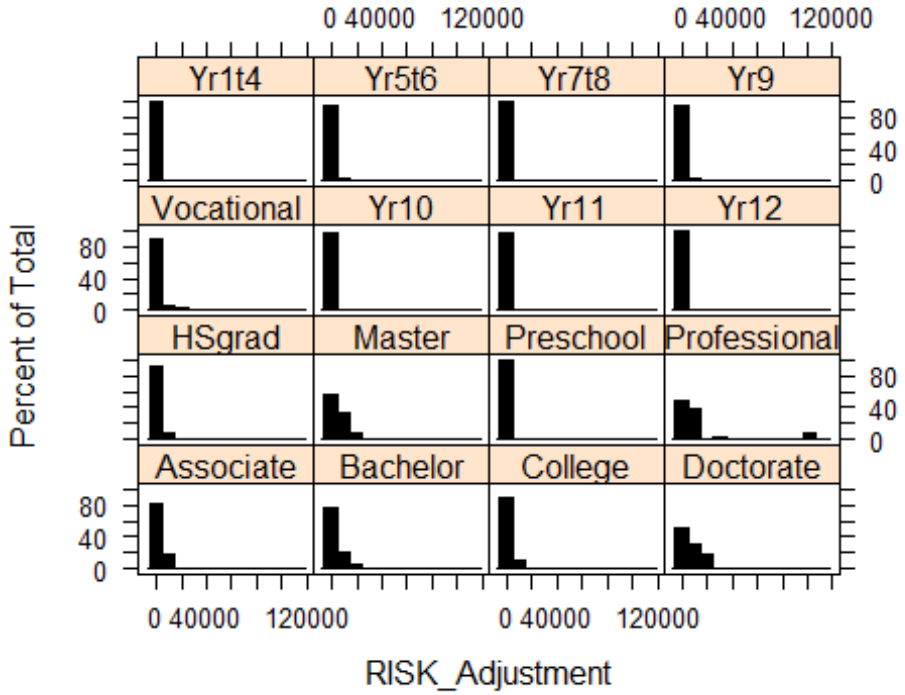

```
histogram(~TARGET_Adjusted|Gender,data=audit,layout=c(1,2),col="black")
```



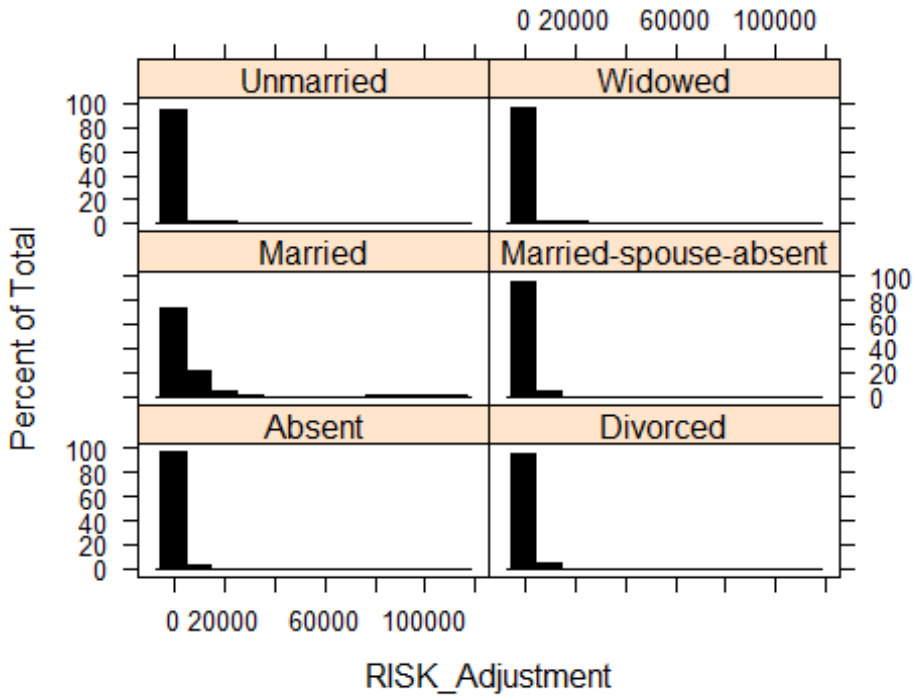

```
histogram(~RISK_Adjustment|Employment,data=audit,layout=c(2,4),col="black")
```



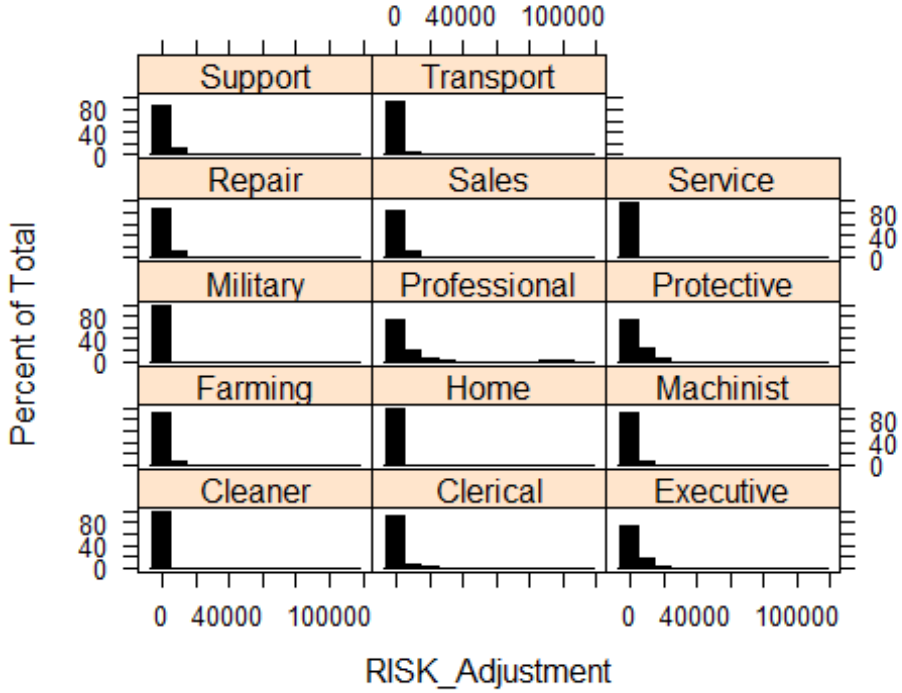

```
histogram(~RISK_Adjustment|Education,data=audit,layout=c(4,4),col="black")
```



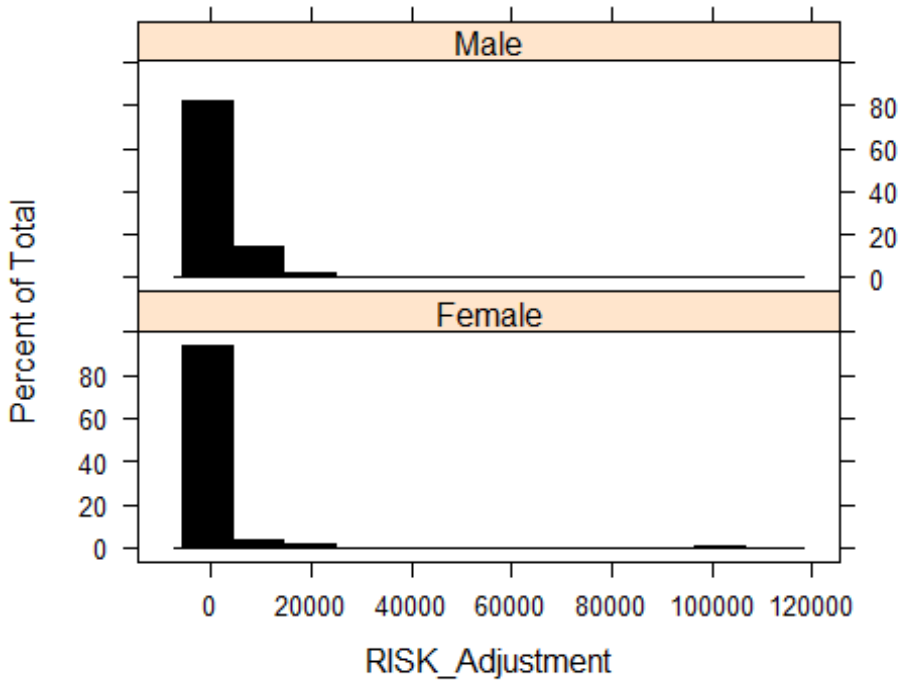

```
histogram(~RISK_Adjustment|Marital,data=audit,layout=c(2,3),col="black")
```




```
histogram(~RISK_Adjustment|Occupation,data=audit,layout=c(3,5),col="black")
```




```
histogram(~RISK_Adjustment|Gender,data=audit,layout=c(1,2),col="black")
```



```
#**Question 3:**
require(boot)

## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma

## The following object is masked from 'package:car':
##
##     logit

audit_t<-audit[,c(-1,-11)]
audit_r<-audit[,c(-1,-12)]

###############
xaudit_t <- model.matrix(TARGET_Adjusted~.,data=audit_t)[,-1]
dfxaudit_t<-as.data.frame(xaudit_t)
Audit_t<-data.frame(targetadj=audit_t$TARGET_Adjusted,dfxaudit_t)
audit_t_t<-Audit_t
audit_t_t<-audit_t_t[sample(nrow(audit_t_t)),]   #randomly shuffle data

#Create 10 equally size folds
folds <- cut(seq(1,nrow(audit_t_t)),breaks=10,labels=FALSE)
result<-NULL
temp<-NULL
#Perform 10 fold cross validation
for(i in 1:10){
    testIndexes <- which(folds==i,arr.ind=TRUE)
    testData <- audit_t_t[testIndexes, ]
    trainData <- audit_t_t[-testIndexes, ]
    m1 = glm(targetadj~.,family=binomial,data=trainData)
    ptest = predict(m1,newdata=data.frame(testData),type="response")
    temp<-cbind(ptest,testData$targetadj)
    result<-rbind(result,temp)

}

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
result<-as.data.frame(result)
names(result)<-c("ptest","ttest")
btest=floor(result$ptest+0.5)
conf.matrix = table(result$ttest,btest)
error=(conf.matrix[1,2]+conf.matrix[2,1])/2000
accuracy=1-error
accuracy
```

```
## [1] 0.837
```

```
precision=conf.matrix[1,1]/(conf.matrix[1,1]+conf.matrix[2,1])
precision
```

```
## [1] 0.8730746
```

```
Recall=conf.matrix[1,1]/(conf.matrix[1,1]+conf.matrix[1,2])
Recall
```

```
## [1] 0.9219258
```

```
F1score=2*precision*Recall/(precision+Recall)
F1score
```

```
## [1] 0.8968354
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.3.2
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

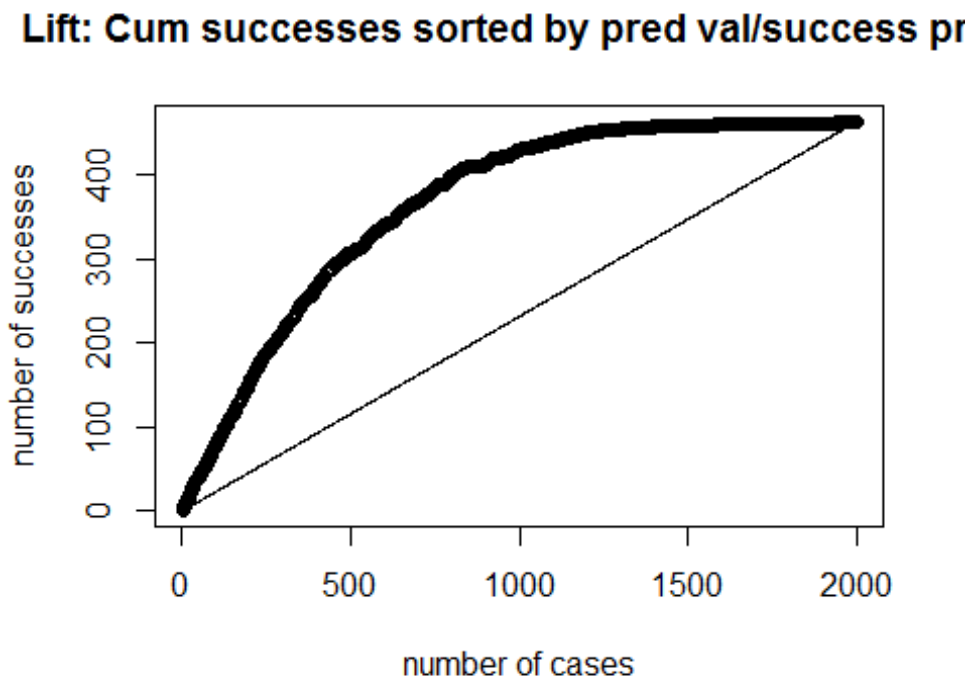
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
aucc=auc(result$ttest, result$ptest)
aucc
```

```
## Area under the curve: 0.8767
```

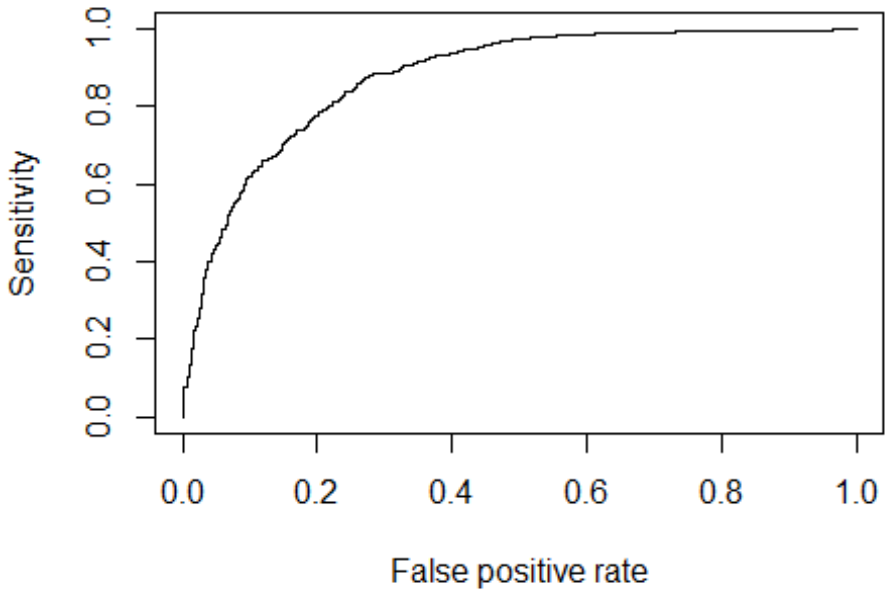
```
#liftchart
df <- result
rank.df=as.data.frame(df[order(result$ptest,decreasing=TRUE),])
colnames(rank.df) = c('predicted','actual')
baserate=mean(result$ttest)
ax=dim(result$ttest)
ay.base=dim(result$ttest)
ay.pred=dim(result$ttest)
ax[1]=1
ay.base[1]=baserate
```

```
ay.pred[1]=rank.df$actual[1]
for (i in 2:2000) {
  ax[i]=i
  ay.base[i]=baserate*i ## uniformly increase with rate xbar
  ay.pred[i]=ay.pred[i-1]+rank.df$actual[i]
}

df=cbind(rank.df,ay.pred,ay.base)
plot(ax,ay.pred,xlab="number of cases",ylab="number of successes",main="Lift:
Cum successes sorted by pred val/success prob")
points(ax,ay.base,type="l")
```




```
#roc
cut=1/2
truepos <- result$ttest==1 & result$ptest>=cut
trueneg <- result$ttest==0 & result$ptest<cut
# Sensitivity (predict default when it does happen)
sum(truepos)/sum(result$ttest==1)

## [1] 0.5550756

suppressWarnings( library(ROCR))

## Loading required package: gplots

##
## Attaching package: 'gplots'
```

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

data<-result
pred <- prediction(result$ptest,result$ttest)
perf <- performance(pred, "sens", "fpr")
plot(perf)
```




```
########################

###############
xaudit_t2 <-
model.matrix(TARGET_Adjusted~Age+Education+Income,data=audit_t)[,-1]
dfxaudit_t2<-as.data.frame(xaudit_t2)
Audit_t2<-data.frame(targetadj=audit_t$TARGET_Adjusted,dfxaudit_t2)
audit_t2_t2<-Audit_t2
audit_t2_t2<-audit_t2_t2[sample(nrow(audit_t2_t2)),]   #randomly shuffle data

#Create 10 equally size folds
folds <- cut(seq(1,nrow(audit_t2_t2)),breaks=10,labels=FALSE)
result<-NULL
temp<-NULL
testIndexes<-NULL
trainData<-NULL
ptest<-NULL
#Perform 10 fold cross validation
```

```
for(i in 1:10){
    testIndexes <- which(folds==i,arr.ind=TRUE)
    testData <- audit_t2_t2[testIndexes, ]
    trainData <- audit_t2_t2[-testIndexes, ]
    m2 =
glm(targetadj~Age+EducationBachelor+EducationCollege+EducationHSgrad+Educatio
nProfessional+EducationVocational+EducationYr10+EducationYr5t6+EducationYr7t8
+Income,family=binomial,data=trainData)
    ptest = predict(m2,newdata=data.frame(testData),type="response")
    temp<-cbind(ptest,testData$targetadj)
    result<-rbind(result,temp)

}
conf.matrix<-NULL
result<-as.data.frame(result)
names(result)<-c("ptest","ttest")
btest=floor(result$ptest+0.5)
conf.matrix = table(result$ttest,btest)
error=(conf.matrix[1,2]+conf.matrix[2,1])/2000
accuracy1=1-error
accuracy1
```

```
## [1] 0.783
```

```
precision1=conf.matrix[1,1]/(conf.matrix[1,1]+conf.matrix[2,1])
precision1
```

```
## [1] 0.8015309
```

```
Recall1=conf.matrix[1,1]/(conf.matrix[1,1]+conf.matrix[1,2])
Recall1
```

```
## [1] 0.9538061
```

```
F1score1=2*precision1*Recall1/(precision1+Recall1)
F1score1
```

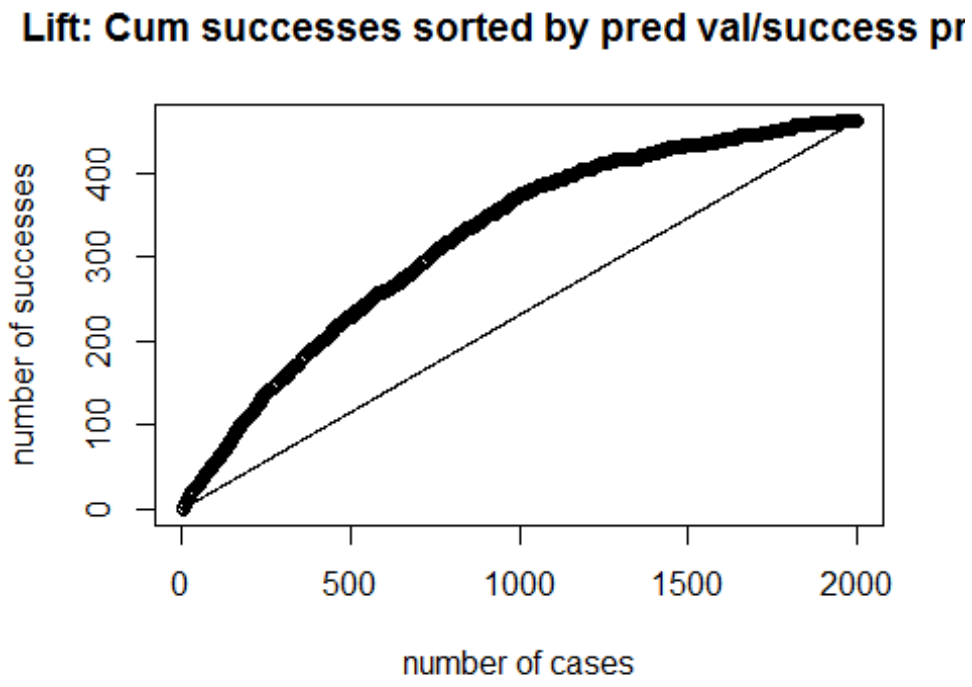
```
## [1] 0.8710636
```

```
auc1=auc(result$ttest, result$ptest)
auc1
```

```
## Area under the curve: 0.7555
```

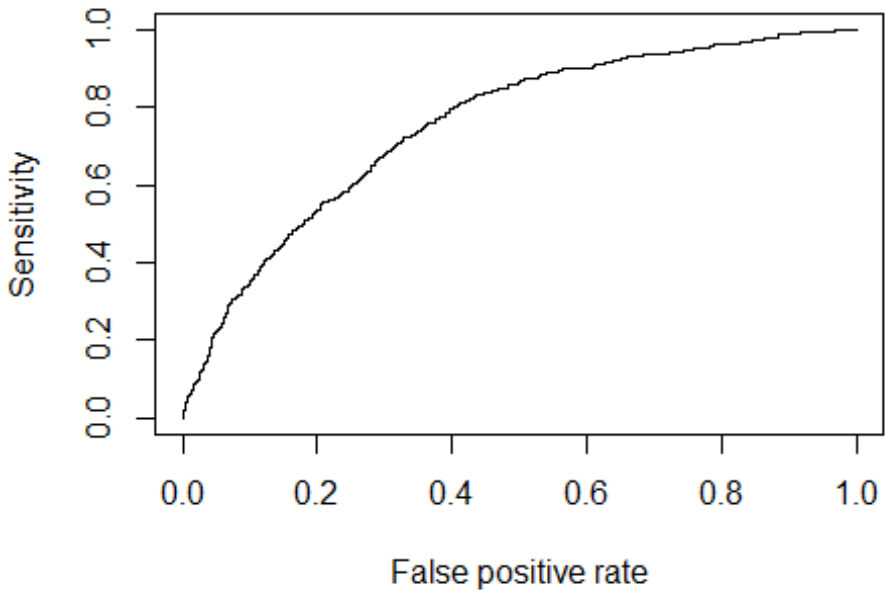
```
#liftchart
df <- result
rank.df=as.data.frame(df[order(result$ptest,decreasing=TRUE),])
colnames(rank.df) = c('predicted','actual')
baserate=mean(result$ttest)
ax=dim(result$ttest)
ay.base=dim(result$ttest)
ay.pred=dim(result$ttest)
```

```
ax[1]=1
ay.base[1]=baserate
ay.pred[1]=rank.df$actual[1]
for (i in 2:2000) {
  ax[i]=i
  ay.base[i]=baserate*i ## uniformly increase with rate xbar
  ay.pred[i]=ay.pred[i-1]+rank.df$actual[i]
}

df=cbind(rank.df,ay.pred,ay.base)
plot(ax,ay.pred,xlab="number of cases",ylab="number of successes",main="Lift:
Cum successes sorted by pred val/success prob")
points(ax,ay.base,type="l")
```




```
#roc
cut=1/2
truepos <- result$ttest==1 & result$ptest>=cut
trueneg <- result$ttest==0 & result$ptest<cut
# Sensitivity (predict default when it does happen)
sum(truepos)/sum(result$ttest==1)

## [1] 0.2159827

suppressWarnings( library(ROCR))
data<-result
pred <- prediction(result$ptest,result$ttest)
```

```
perf <- performance(pred, "sens", "fpr")
plot(perf)
```




```
########################
```

****(b)****

Based on the high accuracy , precision, f1score and low recall values model m1 found out to be the best one and following odds ratio has been done to it. The higher the value for the respective coefficient, mostly likely to be highly significant.

```
m1 = glm(targetadj~.,family=binomial,data=audit_t_t)
oddsratio<-exp(m1$coefficients)
oddsratio
```

```
##                   (Intercept)                           Age
##                  1.527647e-03                  1.027856e+00
##            EmploymentPrivate          EmploymentPSFederal
##                  1.292340e+00                  1.330603e+00
##            EmploymentPSLocal            EmploymentPSState
##                  1.084127e+00                  1.372124e+00
##            EmploymentSelfEmp         EmploymentUnemployed
##                  1.156585e+00                  6.587216e-06
##          EmploymentVolunteer            EducationBachelor
##                  2.315650e-08                  1.114277e+00
##             EducationCollege           EducationDoctorate
```

```
##                 4.213370e-01                 2.443580e+00
##              EducationHSgrad              EducationMaster
##                 3.086031e-01                 1.618500e+00
##           EducationPreschool       EducationProfessional
##                 1.639066e-07                 5.467535e+00
##          EducationVocational                EducationYr10
##                 3.741225e-01                 2.106021e-01
##                EducationYr11                EducationYr12
##                 1.826574e-01                 1.717699e-01
##               EducationYr1t4               EducationYr5t6
##                 3.736186e-08                 9.311356e-02
##               EducationYr7t8                 EducationYr9
##                 5.602639e-08                 5.171519e-02
##              MaritalDivorced               MaritalMarried
##                 9.892897e-01                 1.465774e+01
## MaritalMarried.spouse.absent             MaritalUnmarried
##                 1.372207e+00                 1.839084e+00
##                MaritalWidowed           OccupationClerical
##                 8.179417e-01                 3.206893e+00
##          OccupationExecutive            OccupationFarming
##                 3.852266e+00                 9.630579e-01
##               OccupationHome          OccupationMachinist
##                 3.820315e-06                 1.606481e+00
##           OccupationMilitary       OccupationProfessional
##                 2.184832e-06                 3.324473e+00
##         OccupationProtective              OccupationRepair
##                 6.210526e+00                 1.922544e+00
##              OccupationSales            OccupationService
##                 2.518477e+00                 6.807038e-01
##            OccupationSupport          OccupationTransport
##                 3.487010e+00                 1.262662e+00
##                       Income                   GenderMale
##                 1.000002e+00                 1.199613e+00
##                   Deductions                        Hours
##                 1.001051e+00                 1.037710e+00
```

```
#**Question 4**
leave.one.out <- function(formula, audit_r){
  n = length(audit_r$RISK_Adjustment)
  error = dim(n)
  for(k in 1:n){
    id = c(1:n)
    id.train = id[id != k]
    fit = lm(formula, data = audit_r[id.train, ])
    predicted = predict(fit)
    observation = audit_r$RISK_Adjustment[-id.train]
    error[k] = predicted - observation
  }
  me=mean(error)
  rmse = sqrt(mean(error^2))
```

```
  return(rmse)
}

#linear
formA<-RISK_Adjustment~.
formB<-RISK_Adjustment~Education+Income+Deductions+Hours
formC<-RISK_Adjustment~Employment+Income+Deductions

suppressWarnings(rmseA<-leave.one.out(formA, audit_r))
rmseA

## [1] 8390.117

suppressWarnings(rmseB<-leave.one.out(formB, audit_r))
rmseB

## [1] 8402.807

suppressWarnings(rmseC<-leave.one.out(formC, audit_r))
rmseC

## [1] 8359.086

#non-linear
formD<-RISK_Adjustment~poly(Age, degree = 2) + poly(Income, degree =
2)+Occupation
formE<-RISK_Adjustment~poly(Deductions, degree = 2) + poly(Income, degree =
3) +Education+Employment

suppressWarnings(rmseD<-leave.one.out(formD, audit_r))
rmseD

## [1] 8376.185

suppressWarnings(rmseE<-leave.one.out(formE, audit_r))
rmseE

## [1] 8415.921
```

(b)

Based on the low RMSE values the below model has found to be the best one and StepAIC has been applied to find significant predictors. All are found to be significant.

```
library(MASS)
fit = lm(RISK_Adjustment~poly(Age, degree = 2) + poly(Income, degree =
2)+Occupation, data = audit_r)
stepAIC(fit, direction="backward")

## Start:  AIC=36062.36
## RISK_Adjustment ~ poly(Age, degree = 2) + poly(Income, degree = 2) +
##     Occupation
```

```
##
##                             Df  Sum of Sq        RSS   AIC
## <none>                                    1.3306e+11 36062
## - poly(Income, degree = 2)  2  595916295 1.3366e+11 36067
## - Occupation               13 2135245816 1.3520e+11 36068
## - poly(Age, degree = 2)     2 1678817312 1.3474e+11 36083

##
## Call:
## lm(formula = RISK_Adjustment ~ poly(Age, degree = 2) + poly(Income,
##     degree = 2) + Occupation, data = audit_r)
##
## Coefficients:
##                (Intercept)     poly(Age, degree = 2)1
##                     1977.4                   32886.8
##    poly(Age, degree = 2)2  poly(Income, degree = 2)1
##                   -29423.6                  -18627.3
## poly(Income, degree = 2)2         OccupationClerical
##                    18751.1                    -302.4
##       OccupationExecutive          OccupationFarming
##                     919.2                   -1416.5
##            OccupationHome        OccupationMachinist
##                    -1154.3                   -1225.3
##        OccupationMilitary     OccupationProfessional
##                    -103.8                    1784.5
##      OccupationProtective           OccupationRepair
##                     556.3                    -613.2
##           OccupationSales          OccupationService
##                     350.8                    -946.8
##         OccupationSupport        OccupationTransport
##                    1012.0                   -1783.8
```