Banks\_logistic\_markdown

Raja Prashanth Mullapudi

September 18, 2018

bank<- read.csv(file.choose())  
View(bank)  
attach(bank)

summary(bank)

## age default balance housing   
## Min. :18.00 Min. :0.00000 Min. : -8019 Min. :0.0000   
## 1st Qu.:33.00 1st Qu.:0.00000 1st Qu.: 72 1st Qu.:0.0000   
## Median :39.00 Median :0.00000 Median : 448 Median :1.0000   
## Mean :40.94 Mean :0.01803 Mean : 1362 Mean :0.5558   
## 3rd Qu.:48.00 3rd Qu.:0.00000 3rd Qu.: 1428 3rd Qu.:1.0000   
## Max. :95.00 Max. :1.00000 Max. :102127 Max. :1.0000   
## loan duration campaign pdays   
## Min. :0.0000 Min. : 0.0 Min. : 1.000 Min. : -1.0   
## 1st Qu.:0.0000 1st Qu.: 103.0 1st Qu.: 1.000 1st Qu.: -1.0   
## Median :0.0000 Median : 180.0 Median : 2.000 Median : -1.0   
## Mean :0.1602 Mean : 258.2 Mean : 2.764 Mean : 40.2   
## 3rd Qu.:0.0000 3rd Qu.: 319.0 3rd Qu.: 3.000 3rd Qu.: -1.0   
## Max. :1.0000 Max. :4918.0 Max. :63.000 Max. :871.0   
## previous poutfailure poutother poutsuccess   
## Min. : 0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median : 0.0000 Median :0.0000 Median :0.0000 Median :0.00000   
## Mean : 0.5803 Mean :0.1084 Mean :0.0407 Mean :0.03342   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :275.0000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## poutunknown con\_cellular con\_telephone con\_unknown   
## Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.000   
## Median :1.0000 Median :1.0000 Median :0.00000 Median :0.000   
## Mean :0.8175 Mean :0.6477 Mean :0.06428 Mean :0.288   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.000   
## divorced married single joadmin.   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :1.0000 Median :0.0000 Median :0.0000   
## Mean :0.1152 Mean :0.6019 Mean :0.2829 Mean :0.1144   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## joblue.collar joentrepreneur johousemaid jomanagement   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :0.0000 Median :0.00000 Median :0.00000 Median :0.0000   
## Mean :0.2153 Mean :0.03289 Mean :0.02743 Mean :0.2092   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.0000   
## joretired joself.employed joservices jostudent   
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.00000 Median :0.00000 Median :0.00000 Median :0.00000   
## Mean :0.05008 Mean :0.03493 Mean :0.09188 Mean :0.02075   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000   
## jotechnician jounemployed jounknown y   
## Min. :0.000 Min. :0.00000 Min. :0.00000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.000   
## Median :0.000 Median :0.00000 Median :0.00000 Median :0.000   
## Mean :0.168 Mean :0.02882 Mean :0.00637 Mean :0.117   
## 3rd Qu.:0.000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.000   
## Max. :1.000 Max. :1.00000 Max. :1.00000 Max. :1.000

str(bank) #most variables are discrete, but are memtioned as int

## 'data.frame': 45211 obs. of 32 variables:  
## $ age : int 58 44 33 47 33 35 28 42 58 43 ...  
## $ default : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...  
## $ housing : int 1 1 1 1 0 1 1 1 1 1 ...  
## $ loan : int 0 0 1 0 0 0 1 0 0 0 ...  
## $ duration : int 261 151 76 92 198 139 217 380 50 55 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutfailure : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutother : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutsuccess : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutunknown : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ con\_cellular : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ con\_telephone : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ con\_unknown : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ divorced : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ married : int 1 0 1 1 0 1 0 0 1 0 ...  
## $ single : int 0 1 0 0 1 0 1 0 0 1 ...  
## $ joadmin. : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ joblue.collar : int 0 0 0 1 0 0 0 0 0 0 ...  
## $ joentrepreneur : int 0 0 1 0 0 0 0 1 0 0 ...  
## $ johousemaid : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ jomanagement : int 1 0 0 0 0 1 1 0 0 0 ...  
## $ joretired : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ joself.employed: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ joservices : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ jostudent : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ jotechnician : int 0 1 0 0 0 0 0 0 0 1 ...  
## $ jounemployed : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ jounknown : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ y : int 0 0 0 0 0 0 0 0 0 0 ...

typeof(bank)#o/p is it is a list

## [1] "list"

#glm function to implement logistic function (generalized linear model)  
logit<- glm(y ~ age + factor(default) + balance + factor(housing) + factor(loan)+duration+campaign+pdays+previous+factor(poutfailure)  
 +factor(poutother)+factor(poutsuccess)+factor(poutunknown)+factor(con\_cellular)+factor(con\_telephone)  
 +factor(con\_unknown)+factor(divorced)+factor(married)+factor(single)+factor(joadmin.)+factor(joblue.collar)+  
 factor(joentrepreneur)+factor(johousemaid)+factor(jomanagement)+factor(joretired)+factor(joself.employed)+  
 factor(joservices)+factor(jostudent)+factor(jotechnician)+factor(jounemployed)+factor(jounknown)  
 , family = binomial, data = bank)  
  
summary(logit)

##   
## Call:  
## glm(formula = y ~ age + factor(default) + balance + factor(housing) +   
## factor(loan) + duration + campaign + pdays + previous + factor(poutfailure) +   
## factor(poutother) + factor(poutsuccess) + factor(poutunknown) +   
## factor(con\_cellular) + factor(con\_telephone) + factor(con\_unknown) +   
## factor(divorced) + factor(married) + factor(single) + factor(joadmin.) +   
## factor(joblue.collar) + factor(joentrepreneur) + factor(johousemaid) +   
## factor(jomanagement) + factor(joretired) + factor(joself.employed) +   
## factor(joservices) + factor(jostudent) + factor(jotechnician) +   
## factor(jounemployed) + factor(jounknown), family = binomial,   
## data = bank)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.6748 -0.4060 -0.2731 -0.1625 3.4400   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.855e+00 2.447e-01 -15.754 < 2e-16 \*\*\*  
## age 3.567e-04 2.123e-03 0.168 0.866617   
## factor(default)1 -1.822e-01 1.613e-01 -1.130 0.258582   
## balance 1.872e-05 4.848e-06 3.860 0.000113 \*\*\*  
## factor(housing)1 -7.756e-01 3.953e-02 -19.618 < 2e-16 \*\*\*  
## factor(loan)1 -5.720e-01 5.811e-02 -9.843 < 2e-16 \*\*\*  
## duration 4.048e-03 6.264e-05 64.619 < 2e-16 \*\*\*  
## campaign -1.093e-01 9.905e-03 -11.037 < 2e-16 \*\*\*  
## pdays 1.441e-04 3.006e-04 0.479 0.631590   
## previous 1.042e-02 6.435e-03 1.620 0.105281   
## factor(poutfailure)1 2.563e-01 9.038e-02 2.835 0.004576 \*\*   
## factor(poutother)1 4.985e-01 1.028e-01 4.849 1.24e-06 \*\*\*  
## factor(poutsuccess)1 2.565e+00 8.318e-02 30.836 < 2e-16 \*\*\*  
## factor(poutunknown)1 NA NA NA NA   
## factor(con\_cellular)1 1.166e+00 5.762e-02 20.232 < 2e-16 \*\*\*  
## factor(con\_telephone)1 1.067e+00 8.825e-02 12.093 < 2e-16 \*\*\*  
## factor(con\_unknown)1 NA NA NA NA   
## factor(divorced)1 -1.762e-01 6.530e-02 -2.697 0.006986 \*\*   
## factor(married)1 -3.394e-01 4.448e-02 -7.631 2.32e-14 \*\*\*  
## factor(single)1 NA NA NA NA   
## factor(joadmin.)1 2.965e-01 2.263e-01 1.310 0.190041   
## factor(joblue.collar)1 -1.671e-01 2.256e-01 -0.741 0.458886   
## factor(joentrepreneur)1 -1.073e-01 2.463e-01 -0.436 0.663056   
## factor(johousemaid)1 -2.776e-01 2.495e-01 -1.113 0.265815   
## factor(jomanagement)1 2.380e-01 2.232e-01 1.067 0.286143   
## factor(joretired)1 6.315e-01 2.301e-01 2.745 0.006052 \*\*   
## factor(joself.employed)1 1.962e-02 2.392e-01 0.082 0.934620   
## factor(joservices)1 -3.688e-02 2.301e-01 -0.160 0.872676   
## factor(jostudent)1 8.087e-01 2.395e-01 3.377 0.000734 \*\*\*  
## factor(jotechnician)1 6.383e-02 2.247e-01 0.284 0.776342   
## factor(jounemployed)1 8.167e-02 2.397e-01 0.341 0.733322   
## factor(jounknown)1 NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 32631 on 45210 degrees of freedom  
## Residual deviance: 22640 on 45183 degrees of freedom  
## AIC: 22696  
##   
## Number of Fisher Scoring iterations: 6

#Significant terms -> Intercept, balance, factor(housing)1, factor(loan)1, duration, campaign, factor(poutfailure)1, factor(poutother)1  
# factor(poutsuccess)1, factor(con\_cellular)1 , factor(con\_telephone)1, factor(married)1, factor(divorced)1, factor(joretired)1 ,  
# factor(jostudent)1  
#not significant -> all the others  
  
#Null Deviance = 32,631, Residual Deviance = 22,640   
#Since Null > Residual the model variables are significant to predict the output

exp(coef(logit))

## (Intercept) age factor(default)1   
## 0.02118167 1.00035672 0.83339910   
## balance factor(housing)1 factor(loan)1   
## 1.00001872 0.46043208 0.56440759   
## duration campaign pdays   
## 1.00405573 0.89643728 1.00014415   
## previous factor(poutfailure)1 factor(poutother)1   
## 1.01047743 1.29208605 1.64624413   
## factor(poutsuccess)1 factor(poutunknown)1 factor(con\_cellular)1   
## 12.99778207 NA 3.20841569   
## factor(con\_telephone)1 factor(con\_unknown)1 factor(divorced)1   
## 2.90721486 NA 0.83848631   
## factor(married)1 factor(single)1 factor(joadmin.)1   
## 0.71218751 NA 1.34516468   
## factor(joblue.collar)1 factor(joentrepreneur)1 factor(johousemaid)1   
## 0.84612577 0.89824815 0.75757682   
## factor(jomanagement)1 factor(joretired)1 factor(joself.employed)1   
## 1.26874376 1.88051972 1.01981818   
## factor(joservices)1 factor(jostudent)1 factor(jotechnician)1   
## 0.96378877 2.24502192 1.06590788   
## factor(jounemployed)1 factor(jounknown)1   
## 1.08509561 NA

dim(bank)

## [1] 45211 32

table(y) #o/p is 0->39,922 and 1-> 5289

## y  
## 0 1   
## 39922 5289

#use function predict to get probability value for each row  
prob <- predict(logit, type=c("response"), bank)

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

prob<-as.data.frame(prob)  
  
#we select only elements that are more than 0.50  
confusion <- table(prob>0.50, bank$y)  
confusion

##   
## 0 1  
## FALSE 39013 3587  
## TRUE 909 1702

#Actual value of 0 in Y is 39922 and 1 for 5289 records  
#but predicted for 0 correctly in 39013 cases, and wrongly for 909 cases  
#similarly correct for 1 in 1702 cases and wrongly for 3587 cases  
#therefore model accuracy is 90.05%

#therefore model accuracy is 90.05%  
  
model\_accuracy<-sum(diag(confusion)/sum(confusion))  
model\_accuracy

## [1] 0.9005552

#ROC Curve  
#install.packages("ROCR")  
library(ROCR)

## Warning: package 'ROCR' was built under R version 3.4.4

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.4.4

##   
## Attaching package: 'gplots'

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

rocrpred <- prediction(prob, bank$y)  
rocrperf <- performance(rocrpred, 'tpr', 'fpr')  
plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))

