ADAPT Program Day 5 - Logistic Regression

# Loading packages

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(tidyr)  
library(ggplot2)  
library(data.table) #This is to use fread instead of read.csv (faster for big files)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

library(InformationValue)

# Multivariate Logistic Regression Model

## Loading files

df\_Data <- fread("./data/adult.csv", stringsAsFactors = T)  
  
summary(df\_Data)

## AGE WORKCLASS FNLWGT   
## Min. :17.00 Private :22696 Min. : 12285   
## 1st Qu.:28.00 Self-emp-not-inc: 2541 1st Qu.: 117827   
## Median :37.00 Local-gov : 2093 Median : 178356   
## Mean :38.58 ? : 1836 Mean : 189778   
## 3rd Qu.:48.00 State-gov : 1298 3rd Qu.: 237051   
## Max. :90.00 Self-emp-inc : 1116 Max. :1484705   
## (Other) : 981   
## EDUCATION EDUCATIONNUM MARITALSTATUS   
## HS-grad :10501 Min. : 1.00 Divorced : 4443   
## Some-college: 7291 1st Qu.: 9.00 Married-AF-spouse : 23   
## Bachelors : 5355 Median :10.00 Married-civ-spouse :14976   
## Masters : 1723 Mean :10.08 Married-spouse-absent: 418   
## Assoc-voc : 1382 3rd Qu.:12.00 Never-married :10683   
## 11th : 1175 Max. :16.00 Separated : 1025   
## (Other) : 5134 Widowed : 993   
## OCCUPATION RELATIONSHIP RACE   
## Prof-specialty :4140 Husband :13193 Amer-Indian-Eskimo: 311   
## Craft-repair :4099 Not-in-family : 8305 Asian-Pac-Islander: 1039   
## Exec-managerial:4066 Other-relative: 981 Black : 3124   
## Adm-clerical :3770 Own-child : 5068 Other : 271   
## Sales :3650 Unmarried : 3446 White :27816   
## Other-service :3295 Wife : 1568   
## (Other) :9541   
## SEX CAPITALGAIN CAPITALLOSS HOURSPERWEEK   
## Female:10771 Min. : 0 Min. : 0.0 Min. : 1.00   
## Male :21790 1st Qu.: 0 1st Qu.: 0.0 1st Qu.:40.00   
## Median : 0 Median : 0.0 Median :40.00   
## Mean : 1078 Mean : 87.3 Mean :40.44   
## 3rd Qu.: 0 3rd Qu.: 0.0 3rd Qu.:45.00   
## Max. :99999 Max. :4356.0 Max. :99.00   
##   
## NATIVECOUNTRY ABOVE50K   
## United-States:29170 Min. :0.0000   
## Mexico : 643 1st Qu.:0.0000   
## ? : 583 Median :0.0000   
## Philippines : 198 Mean :0.2408   
## Germany : 137 3rd Qu.:0.0000   
## Canada : 121 Max. :1.0000   
## (Other) : 1709

head(df\_Data)

## AGE WORKCLASS FNLWGT EDUCATION EDUCATIONNUM MARITALSTATUS  
## 1: 39 State-gov 77516 Bachelors 13 Never-married  
## 2: 50 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse  
## 3: 38 Private 215646 HS-grad 9 Divorced  
## 4: 53 Private 234721 11th 7 Married-civ-spouse  
## 5: 28 Private 338409 Bachelors 13 Married-civ-spouse  
## 6: 37 Private 284582 Masters 14 Married-civ-spouse  
## OCCUPATION RELATIONSHIP RACE SEX CAPITALGAIN CAPITALLOSS  
## 1: Adm-clerical Not-in-family White Male 2174 0  
## 2: Exec-managerial Husband White Male 0 0  
## 3: Handlers-cleaners Not-in-family White Male 0 0  
## 4: Handlers-cleaners Husband Black Male 0 0  
## 5: Prof-specialty Wife Black Female 0 0  
## 6: Exec-managerial Wife White Female 0 0  
## HOURSPERWEEK NATIVECOUNTRY ABOVE50K  
## 1: 40 United-States 0  
## 2: 13 United-States 0  
## 3: 40 United-States 0  
## 4: 40 United-States 0  
## 5: 40 Cuba 0  
## 6: 40 United-States 0

## Check BIAS of the data

table(df\_Data$ABOVE50K)

##   
## 0 1   
## 24720 7841

## Creating Training and Testing Datasets

positives <- df\_Data[which(df\_Data$ABOVE50K==1),]  
negatives <- df\_Data[which(df\_Data$ABOVE50K!=1),]  
  
set.seed(100)  
  
trIndex.pos <- sample(1:nrow(positives),0.7\*nrow(positives))  
trIndex.neg <- sample(1:nrow(negatives),0.7\*nrow(positives))  
   
trData <- rbind(positives[trIndex.pos,],negatives[trIndex.neg,])  
  
#Not Used Data is for Testing  
tsData <- rbind(positives[-trIndex.pos,],negatives[-trIndex.neg,])   
   
table(trData$ABOVE50K) #Balanced training Dataset

##   
## 0 1   
## 5488 5488

nrow(trData)/nrow(df\_Data)

## [1] 0.3370904

table(tsData$ABOVE50K) #Warning: It's more unbalanced than the original DS

##   
## 0 1   
## 19232 2353

## Selecting the significant factors

factor\_vars <- c("WORKCLASS", "EDUCATION", "MARITALSTATUS", "OCCUPATION", "RELATIONSHIP", "RACE", "SEX", "NATIVECOUNTRY")  
continuous\_vars <- c("AGE", "EDUCATIONNUM", "HOURSPERWEEK", "CAPITALGAIN", "CAPITALLOSS")  
  
#initialization for the IV results  
iv\_df <- data.frame(VARS=c(factor\_vars, continuous\_vars), IV=numeric(13))  
  
#Compute IV for categorical Variables  
for (i in factor\_vars) {  
iv\_df[iv\_df$VARS==i,"IV"] <- IV(df\_Data[[i]], df\_Data$ABOVE50K)  
}  
  
#Compute IV for numerical Variables  
for (i in continuous\_vars) {  
iv\_df[iv\_df$VARS==i,"IV"] <- IV(as.factor(df\_Data[[i]]), df\_Data$ABOVE50K)  
}  
  
arrange(iv\_df,-IV)

## VARS IV  
## 1 RELATIONSHIP 1.53560810  
## 2 MARITALSTATUS 1.33882907  
## 3 AGE 0.88214658  
## 4 OCCUPATION 0.77622839  
## 5 EDUCATION 0.74105372  
## 6 EDUCATIONNUM 0.74105372  
## 7 HOURSPERWEEK 0.49628770  
## 8 CAPITALGAIN 0.31266990  
## 9 SEX 0.30328938  
## 10 CAPITALLOSS 0.20749663  
## 11 WORKCLASS 0.16338802  
## 12 NATIVECOUNTRY 0.07939344  
## 13 RACE 0.06929987

## Training and testing the model

fmla <- "ABOVE50K ~ RELATIONSHIP + MARITALSTATUS + AGE + OCCUPATION + EDUCATION + HOURSPERWEEK + CAPITALGAIN + SEX"  
  
logitMod <- glm(fmla, data=trData, family = "binomial"(link="logit"))

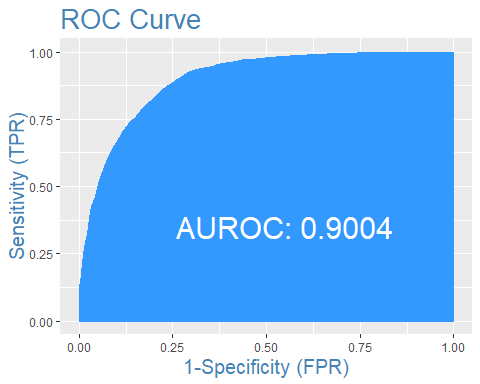
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

predicted <- predict(logitMod, tsData, type="response")  
  
summary(logitMod)

##   
## Call:  
## glm(formula = fmla, family = binomial(link = "logit"), data = trData)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.4221 -0.5073 -0.0001 0.6161 3.3528   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -7.211e+00 5.289e-01 -13.634 < 2e-16  
## RELATIONSHIPNot-in-family 5.771e-01 3.875e-01 1.489 0.136369  
## RELATIONSHIPOther-relative -5.559e-01 3.638e-01 -1.528 0.126496  
## RELATIONSHIPOwn-child -6.672e-01 3.794e-01 -1.759 0.078643  
## RELATIONSHIPUnmarried 5.166e-01 4.076e-01 1.267 0.205041  
## RELATIONSHIPWife 1.358e+00 1.542e-01 8.806 < 2e-16  
## MARITALSTATUSMarried-AF-spouse 3.342e+00 9.829e-01 3.400 0.000673  
## MARITALSTATUSMarried-civ-spouse 2.345e+00 3.890e-01 6.029 1.65e-09  
## MARITALSTATUSMarried-spouse-absent 1.977e-02 3.198e-01 0.062 0.950700  
## MARITALSTATUSNever-married -3.586e-01 1.236e-01 -2.901 0.003717  
## MARITALSTATUSSeparated -3.127e-01 2.247e-01 -1.391 0.164112  
## MARITALSTATUSWidowed 1.625e-01 2.146e-01 0.757 0.449024  
## AGE 2.967e-02 2.644e-03 11.220 < 2e-16  
## OCCUPATIONAdm-clerical 7.190e-01 1.711e-01 4.202 2.64e-05  
## OCCUPATIONArmed-Forces 3.449e-01 1.588e+00 0.217 0.828099  
## OCCUPATIONCraft-repair 8.930e-01 1.626e-01 5.493 3.96e-08  
## OCCUPATIONExec-managerial 1.607e+00 1.633e-01 9.842 < 2e-16  
## OCCUPATIONFarming-fishing -5.294e-01 2.227e-01 -2.377 0.017465  
## OCCUPATIONHandlers-cleaners 2.336e-01 2.268e-01 1.030 0.303037  
## OCCUPATIONMachine-op-inspct 4.410e-01 1.853e-01 2.380 0.017305  
## OCCUPATIONOther-service -5.314e-02 1.968e-01 -0.270 0.787171  
## OCCUPATIONPriv-house-serv -3.447e+00 2.564e+00 -1.344 0.178795  
## OCCUPATIONProf-specialty 1.406e+00 1.681e-01 8.364 < 2e-16  
## OCCUPATIONProtective-serv 1.182e+00 2.218e-01 5.329 9.86e-08  
## OCCUPATIONSales 9.970e-01 1.662e-01 5.999 1.98e-09  
## OCCUPATIONTech-support 1.523e+00 2.068e-01 7.366 1.76e-13  
## OCCUPATIONTransport-moving 6.009e-01 1.844e-01 3.259 0.001118  
## EDUCATION11th -9.098e-02 2.970e-01 -0.306 0.759352  
## EDUCATION12th -3.084e-01 4.067e-01 -0.758 0.448270  
## EDUCATION1st-4th -1.016e+00 5.484e-01 -1.853 0.063946  
## EDUCATION5th-6th -5.896e-01 4.216e-01 -1.398 0.162010  
## EDUCATION7th-8th -5.571e-01 3.295e-01 -1.691 0.090900  
## EDUCATION9th -2.404e-01 3.706e-01 -0.649 0.516538  
## EDUCATIONAssoc-acdm 1.153e+00 2.615e-01 4.410 1.04e-05  
## EDUCATIONAssoc-voc 1.193e+00 2.516e-01 4.741 2.13e-06  
## EDUCATIONBachelors 1.746e+00 2.310e-01 7.560 4.04e-14  
## EDUCATIONDoctorate 2.786e+00 3.427e-01 8.130 4.30e-16  
## EDUCATIONHS-grad 5.995e-01 2.227e-01 2.692 0.007106  
## EDUCATIONMasters 1.934e+00 2.495e-01 7.751 9.09e-15  
## EDUCATIONPreschool -1.236e+01 1.259e+02 -0.098 0.921735  
## EDUCATIONProf-school 3.188e+00 3.518e-01 9.061 < 2e-16  
## EDUCATIONSome-college 1.006e+00 2.267e-01 4.437 9.11e-06  
## HOURSPERWEEK 3.517e-02 2.682e-03 13.111 < 2e-16  
## CAPITALGAIN 3.186e-04 1.791e-05 17.788 < 2e-16  
## SEXMale 8.048e-01 1.096e-01 7.341 2.12e-13  
##   
## (Intercept) \*\*\*  
## RELATIONSHIPNot-in-family   
## RELATIONSHIPOther-relative   
## RELATIONSHIPOwn-child .   
## RELATIONSHIPUnmarried   
## RELATIONSHIPWife \*\*\*  
## MARITALSTATUSMarried-AF-spouse \*\*\*  
## MARITALSTATUSMarried-civ-spouse \*\*\*  
## MARITALSTATUSMarried-spouse-absent   
## MARITALSTATUSNever-married \*\*   
## MARITALSTATUSSeparated   
## MARITALSTATUSWidowed   
## AGE \*\*\*  
## OCCUPATIONAdm-clerical \*\*\*  
## OCCUPATIONArmed-Forces   
## OCCUPATIONCraft-repair \*\*\*  
## OCCUPATIONExec-managerial \*\*\*  
## OCCUPATIONFarming-fishing \*   
## OCCUPATIONHandlers-cleaners   
## OCCUPATIONMachine-op-inspct \*   
## OCCUPATIONOther-service   
## OCCUPATIONPriv-house-serv   
## OCCUPATIONProf-specialty \*\*\*  
## OCCUPATIONProtective-serv \*\*\*  
## OCCUPATIONSales \*\*\*  
## OCCUPATIONTech-support \*\*\*  
## OCCUPATIONTransport-moving \*\*   
## EDUCATION11th   
## EDUCATION12th   
## EDUCATION1st-4th .   
## EDUCATION5th-6th   
## EDUCATION7th-8th .   
## EDUCATION9th   
## EDUCATIONAssoc-acdm \*\*\*  
## EDUCATIONAssoc-voc \*\*\*  
## EDUCATIONBachelors \*\*\*  
## EDUCATIONDoctorate \*\*\*  
## EDUCATIONHS-grad \*\*   
## EDUCATIONMasters \*\*\*  
## EDUCATIONPreschool   
## EDUCATIONProf-school \*\*\*  
## EDUCATIONSome-college \*\*\*  
## HOURSPERWEEK \*\*\*  
## CAPITALGAIN \*\*\*  
## SEXMale \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 15216.0 on 10975 degrees of freedom  
## Residual deviance: 8396.7 on 10931 degrees of freedom  
## AIC: 8486.7  
##   
## Number of Fisher Scoring iterations: 12

## ROC Curve

plotROC(tsData$ABOVE50K,predicted)



## Confusion Matrix

cm <- confusionMatrix(tsData$ABOVE50K,predicted)  
colnames(cm) <- c("Actual 0", "Actual 1")  
rownames(cm) <- c("Predicted 0", "Predicted 1")  
  
fourfoldplot(as.matrix(cm))

