Predictive Model Using Logistic Regression

Author: Data For Excellence (DFE), GPE, PSA International

Updated on: 05 March 2020

Use of this document

This file is created for the sole use of PSA Data Analytics Technical Workshop participants for demonstration and learning about the predictive modelling.

Scripting language used

This document is created using R Markdown, a scripting language available as open source from R Foundation.

Dataset used in the model

The dataset, popularly known as “Adult” data, is publicly available in the UCI machine learning repository. The dataset is further modified for the purpose of making it useful for PSA training

End of Introduction Section

Start of Stage 1

1. The Business Understanding

- The income of a person is a function of many factors/attributes. Given enough data about these attributes, a supervised machine learning model could be developed.
- We want to predict who will earn more than 50k salary based on the 14 attributes of a person.
- The output is Yes/No or (1/0), where Yes or 1 indicate that the person will earn more than 50k. Since the output is a categorical variable, we will use Logistics Regression to predict if a person will earn 50k or not.

End of Stage 1

Start of Stage 2

2. Data Understanding

The dataset used in this project has 48,842 records and a binomial label indicating a salary of <50K or >50K USD. 76% of the records in the dataset have a class label of <50K.

#**Data fields**
AGE
WORKCLASS
FNLWGT
EDUCATION
EDUCATIONNUM
MARITALSTATUS
OCCUPATION
RELATIONSHIP
RACE

SEX
CAPITALGAIN
CAPITALLOSS
HOURSPERWEEK
NATIVECOUNTRY
ABOVE50K

Loading all the required packages

```
install.packages("InformationValue",repos = "http://cran.us.r-project.org")

## package 'data.table' successfully unpacked and MD5 sums checked
## package 'InformationValue' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\pramodkv\AppData\Local\Temp\RtmpqAWQYE\downloaded_packages

library(InformationValue)
```

May need to load more libraries/packages depending on local computer/server

Loading the file into R data-frame

```
#inputData <-
read.csv("https://github.com/laosze95/Training/raw/master/adult.csv")
inputData <- read.csv("D:/Data Analytics/Data Analytics Workshop/Data
Analytics Technical Workshop/Version 2/Data/adult.csv")
# From Pramod Verma Github Page, use
"https://github.com/laosze95/Training/raw/master/adult.csv"
# From internet use "http://rstatistics.net/wp-
content/uploads/2015/09/adult.csv"

head(inputData)

##   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
```

Looking at the structure of the data

```
dim(inputData)

## [1] 32561    15

class(inputData)

## [1] "data.frame"

str(inputData)

## 'data.frame':    32561 obs. of  15 variables:
##  $ AGE          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ WORKCLASS    : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5
7 5 5 ...
##  $ FNLWGT       : int  77516 83311 215646 234721 338409 284582 160187
209642 45781 159449 ...
##  $ EDUCATION    : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13
7 12 13 10 ...
##  $ EDUCATIONNUM : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ MARITALSTATUS: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..:
5 3 1 3 3 3 4 3 5 3 ...
##  $ OCCUPATION   : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11
5 9 5 11 5 ...
##  $ RELATIONSHIP : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2
1 6 6 2 1 2 1 ...
##  $ RACE         : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5
3 5 5 5 ...
##  $ SEX          : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1
2 ...
##  $ CAPITALGAIN  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ CAPITALLOSS  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ HOURSPERWEEK : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ NATIVECOUNTRY: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6
40 24 40 40 40 ...
##  $ ABOVE50K     : int  0 0 0 0 0 0 0 1 1 1 ...

summary(inputData)

##       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
```

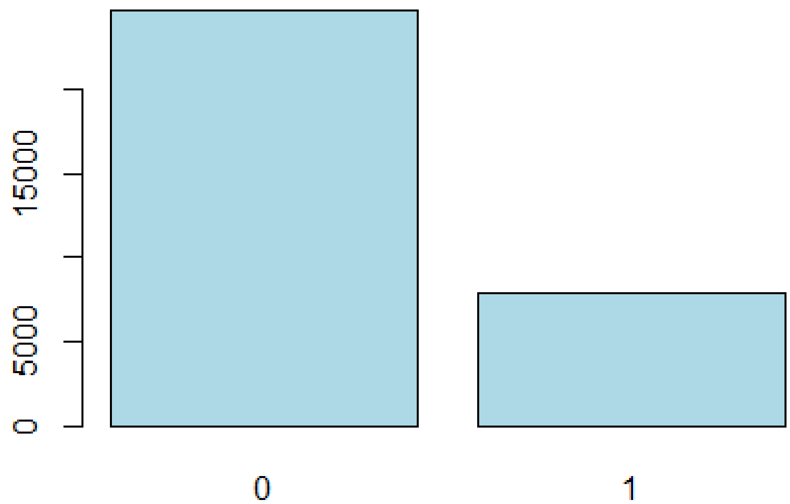
- There are 14 attributes consisting of eight categorical and six continuous attributes. The work class describes the type of employer such as self-employed or federal and occupation describes the employment type such as farming, clerical or managerial.
- Education contains the highest level of education attained such as high school or doctorate.
- The relationship attribute has categories such as unmarried or husband and marital status has categories such as married or separated.
- The other nominal attributes are country of residence, gender and race.
- The continuous attributes are age, hours worked per week, education number (numeric representation of the education attribute), capital gain and loss, and a weight

attribute which is a demographic score assigned to an individual based on information such as state of residence and type of employment.

- Some of the variables are not self-explanatory. The continuous variable fnlwgt represents final weight, which is the number of units in the target population that the responding unit represents.
- The variable education_num stands for the number of years of education in total, which is a continuous representation of the discrete variable education. The variable relationship represents the responding unit's role in the family.
- Capital_gain and capital_loss are income from investment sources other than wage/salary.
- For simplicity of this analysis, the weighting factor is discarded. Total number of years of education can represent by the highest education level completed. Role in the family can be assessed from gender and marital status. Thus, the following 3 variables are deleted education, relationship, and fnlwgt.

Checking the class bias of the data

```
table(inputData$ABOVE50K)

##
##     0     1
## 24720  7841

# histogram of age by income group
barplot(table(inputData$ABOVE50K), col = "lightblue")
```



Since there is a class bias, a condition observed when the proportion of events is much smaller than proportion of non-events. So we must sample the observations in approximately equal proportions to get better models.

End of Stage 2

Start of Stage 3

3. Data Preparation

First we want to clean up the data set to include only those variables which are importants From our data understanding, we know FNLWGT and RELATIONSHIP is not required.

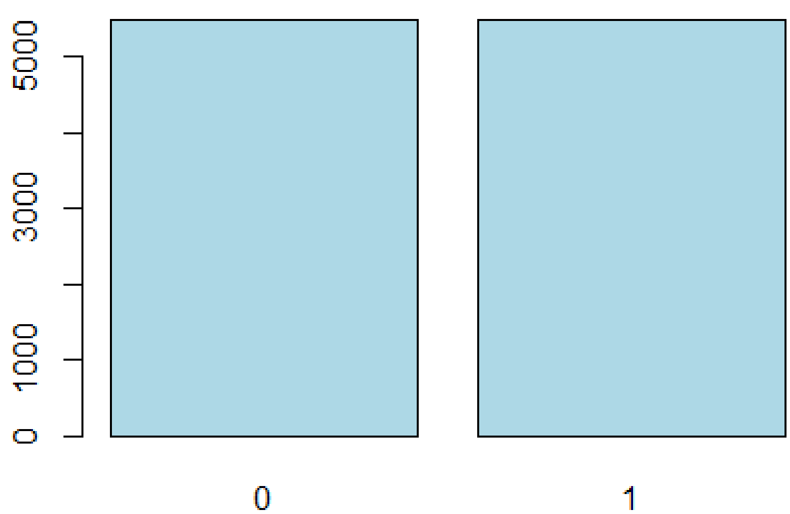
```
inputData$FNLWGT <- NULL
inputData$RELATIONSHIP <- NULL
head(inputData$FNLWGT)

## NULL

head(inputData$RELATIONSHIP)
```

```
## NULL
```

Creating two sets of data from given data * Training set - For training the model * Test set - For test and validation

Creating training data set

```
input_ones <- inputData[which(inputData$ABOVE50K == 1), ]  # all 1's
input_zeros <- inputData[which(inputData$ABOVE50K == 0), ]  # all 0's

set.seed(100)  # for repeatability of samples

input_ones_training_rows <- sample(1:nrow(input_ones), 0.7*nrow(input_ones))
# 1's for training
input_zeros_training_rows <- sample(1:nrow(input_zeros),
0.7*nrow(input_ones))  # 0's for training.

#Pick as many 0's as 1's
training_ones <- input_ones[input_ones_training_rows, ]
training_zeros <- input_zeros[input_zeros_training_rows, ]

# row bind the 1's and 0's
trainingData <- rbind(training_ones, training_zeros)

# Checking the bias on training data
barplot(table(trainingData$ABOVE50K),col = "lightblue")
```

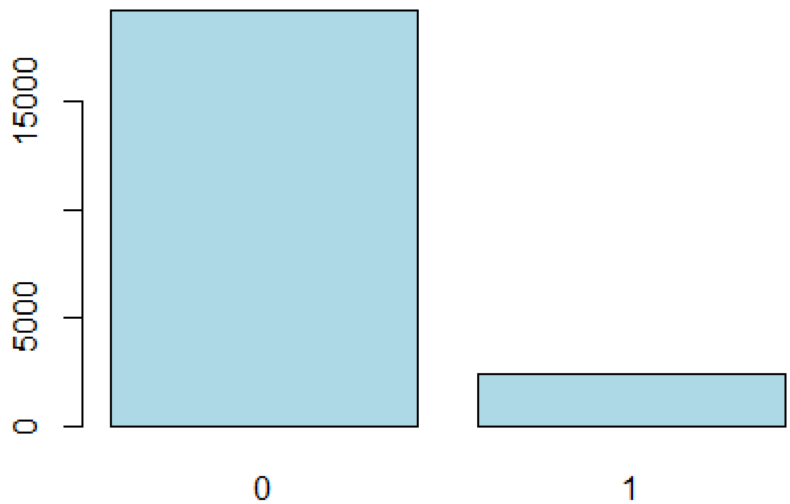


```r
head(trainingData)
```

```
##       AGE WORKCLASS     EDUCATION EDUCATIONNUM       MARITALSTATUS
## 15775  39   Private  Some-college           10  Married-civ-spouse
## 2031   32   Private    Assoc-acdm           12  Married-civ-spouse
## 14367  42   Private   Prof-school           15  Married-civ-spouse
## 15463  38   Private    Assoc-acdm           12  Married-civ-spouse
## 25637  35         ?  Some-college           10  Married-civ-spouse
## 29720  44   Private     Bachelors           13  Married-civ-spouse
##              OCCUPATION                RACE     SEX CAPITALGAIN CAPITALLOSS
## 15775      Adm-clerical               White    Male       15024           0
## 2031    Exec-managerial               White  Female           0        1485
## 14367    Prof-specialty               White    Male        7298           0
## 15463             Sales               White    Male           0           0
## 25637                 ?  Asian-Pac-Islander    Male           0           0
## 29720    Prof-specialty               White    Male           0           0
##       HOURSPERWEEK  NATIVECOUNTRY ABOVE50K
## 15775           40  United-States        1
## 2031            40  United-States        1
## 14367           35  United-States        1
## 15463           50  United-States        1
## 25637           40    Philippines        1
## 29720           40  United-States        1
```

Creating the test data set

```r
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros <- input_zeros[-input_zeros_training_rows, ]

# row bind the 1's and 0's
testData <- rbind(test_ones, test_zeros)

# We do not need to correct the bias on test data because model should take
care of future uncertainity
barplot(table(testData$ABOVE50K),col = "lightblue")
```




```
head(testData)
```

```
##    AGE     WORKCLASS     EDUCATION EDUCATIONNUM       MARITALSTATUS
## 10  42       Private     Bachelors           13  Married-civ-spouse
## 12  30     State-gov     Bachelors           13  Married-civ-spouse
## 21  40       Private     Doctorate           16  Married-civ-spouse
## 39  31       Private  Some-college           10  Married-civ-spouse
## 46  57   Federal-gov     Bachelors           13  Married-civ-spouse
## 56  43       Private  Some-college           10  Married-civ-spouse
##           OCCUPATION                RACE   SEX CAPITALGAIN CAPITALLOSS
## 10  Exec-managerial                White  Male        5178           0
## 12   Prof-specialty   Asian-Pac-Islander  Male           0           0
## 21   Prof-specialty                White  Male           0           0
## 39            Sales                White  Male           0           0
## 46   Prof-specialty                Black  Male           0           0
## 56     Tech-support                White  Male           0           0
##    HOURSPERWEEK  NATIVECOUNTRY ABOVE50K
## 10           40  United-States        1
## 12           40          India        1
## 21           60  United-States        1
## 39           38              ?        1
## 46           40  United-States        1
## 56           40  United-States        1
```

Feature Selection

- Now we want to know that out of 14 attributes, which are the most important one.There are many methods to find out the best attributes. We will use WOE (Weight of Evidence) method. The choice of feature selction is based on data types and model types.

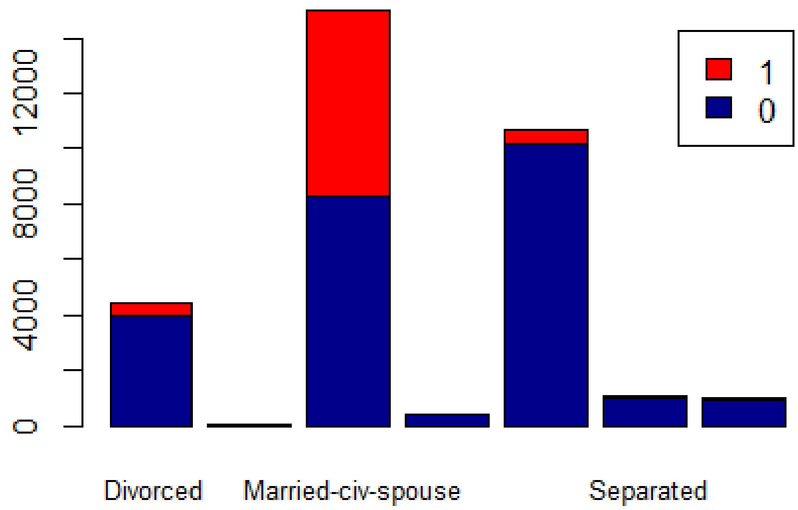
***Weight of evidence (WOE)** is a measure of how much the evidence supports or undermines a hypothesis. WOE measures the relative risk of an attribute of binning level. The value depends on whether the value of the target variable is a non-event or an event.

Compute Information Values (IV)

We will compute information values for both categorical and continuous variable. The continuous variable needs to be converted to categorical variable before we compute information value.

```
# segregate continuous and factor variables
factor_vars <- c ("WORKCLASS", "EDUCATION", "MARITALSTATUS", "OCCUPATION",
"RELATIONSHIP", "RACE", "SEX", "NATIVECOUNTRY")
continuous_vars <- c("AGE", "FNLWGT","EDUCATIONNUM", "HOURSPERWEEK",
"CAPITALGAIN", "CAPITALLOSS")


# initialization for the for IV results
iv_df <- data.frame(VARS=c(factor_vars, continuous_vars), IV=numeric(14))


# compute IV for categorical Variables

iv_df[iv_df$VARS == "WORKCLASS", "IV"] <- IV(X=inputData$WORKCLASS,
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "EDUCATION", "IV"] <- IV(X=inputData$EDUCATION,
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "MARITALSTATUS", "IV"] <- IV(X=inputData$MARITALSTATUS,
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "OCCUPATION", "IV"] <- IV(X=inputData$OCCUPATION,
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "RACE", "IV"] <- IV(X=inputData$RACE,
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "SEX", "IV"] <- IV(X=inputData$SEX,
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "NATIVECOUNTRY", "IV"] <- IV(X=inputData$NATIVECOUNTRY,
Y=inputData$ABOVE50K)[1]


# compute IV for Continuous Variables

iv_df[iv_df$VARS == "AGE", "IV"] <- IV(X=as.factor(inputData$AGE),
```

```
Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "EDUCATIONNUM", "IV"] <-
IV(X=as.factor(inputData$EDUCATIONNUM), Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "HOURSPERWEEK", "IV"] <-
IV(X=as.factor(inputData$HOURSPERWEEK), Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "CAPITALGAIN", "IV"] <-
IV(X=as.factor(inputData$CAPITALGAIN), Y=inputData$ABOVE50K)[1]
iv_df[iv_df$VARS == "CAPITALLOSS", "IV"] <-
IV(X=as.factor(inputData$CAPITALLOSS), Y=inputData$ABOVE50K)[1]


iv_df <- iv_df[order(-iv_df$IV), ]  # sort
#**Below are the significance of each variable in the prediction**
iv_df

##              VARS         IV
## 3  MARITALSTATUS 1.33882907
## 9            AGE 0.88214658
## 4     OCCUPATION 0.77622839
## 2      EDUCATION 0.74105372
## 11  EDUCATIONNUM 0.74105372
## 12  HOURSPERWEEK 0.49628770
## 13   CAPITALGAIN 0.31266990
## 7            SEX 0.30328938
## 14   CAPITALLOSS 0.20749663
## 1      WORKCLASS 0.16338802
## 8  NATIVECOUNTRY 0.07939344
## 6           RACE 0.06929987
## 5   RELATIONSHIP 0.00000000
## 10        FNLWGT 0.00000000

#**Let us check the most significant variable and dependent variable
relationship**
table(inputData$ABOVE50K,inputData$MARITALSTATUS )

##
##      Divorced  Married-AF-spouse  Married-civ-spouse  Married-spouse-
absent
##   0      3980                 13                8284
384
##   1       463                 10                6692
34
##
##      Never-married  Separated  Widowed
##   0          10192        959      908
##   1            491         66       85

barplot(table(inputData$ABOVE50K,inputData$MARITALSTATUS
),col=c("darkblue","red"),legend = TRUE, cex.names=0.8)
```



End of Stage 3

Start of Stage 4

4. Modelling

Building the Logistic Model using the most significant attributes which are

MARITALSTATUS
AGE
OCCUPATION
EDUCATION
EDUCATIONNUM
HOURSPERWEEK
CAPITALGAIN
SEX

However, we see that EDUCATION AND EDUCATIONNUM ARE HIGHLY CORELATED SO WE CAN PICK ONLY ONE

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

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

# predicted scores
predicted <- predict(logitMod, testData, type="response")
summary(logitMod)

## 
## Call:
## glm(formula = ABOVE50K ~ MARITALSTATUS + AGE + OCCUPATION + EDUCATION + 
##     HOURSPERWEEK + CAPITALGAIN + SEX, family = binomial(link = "logit"), 
##     data = trainingData)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.7945  -0.5328  -0.0001   0.6212   3.2815  
## 
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -6.635e+00  3.213e-01 -20.647  < 2e-16
***
## MARITALSTATUS Married-AF-spouse      3.858e+00  8.904e-01   4.333 1.47e-05
***
## MARITALSTATUS Married-civ-spouse     2.142e+00  9.274e-02  23.099  < 2e-16
***
## MARITALSTATUS Married-spouse-absent -9.390e-02  3.033e-01  -0.310  0.75684
## MARITALSTATUS Never-married         -5.820e-01  1.137e-01  -5.118 3.08e-07
***
## MARITALSTATUS Separated             -4.034e-01  2.185e-01  -1.846  0.06487
.
## MARITALSTATUS Widowed                5.930e-02  1.996e-01   0.297  0.76639
## AGE                                  3.113e-02  2.551e-03  12.203  < 2e-16
***
## OCCUPATION Adm-clerical              6.657e-01  1.622e-01   4.104 4.06e-05
***
## OCCUPATION Armed-Forces             -1.081e-01  1.529e+00  -0.071  0.94363
## OCCUPATION Craft-repair              7.760e-01  1.553e-01   4.995 5.87e-07
***
## OCCUPATION Exec-managerial           1.615e+00  1.559e-01  10.359  < 2e-16
***
## OCCUPATION Farming-fishing          -4.801e-01  2.219e-01  -2.164  0.03050
*
## OCCUPATION Handlers-cleaners         5.584e-02  2.221e-01   0.251  0.80146
## OCCUPATION Machine-op-inspct         3.753e-01  1.775e-01   2.114  0.03448
*
```

```
## OCCUPATION Other-service            2.932e-02  1.903e-01   0.154  0.87759
## OCCUPATION Priv-house-serv         -5.127e+00  1.957e+00  -2.619  0.00881
**
## OCCUPATION Prof-specialty           1.270e+00  1.591e-01   7.984 1.41e-15
***
## OCCUPATION Protective-serv          1.267e+00  2.166e-01   5.847 5.02e-09
***
## OCCUPATION Sales                    9.484e-01  1.585e-01   5.982 2.20e-09
***
## OCCUPATION Tech-support             1.275e+00  1.963e-01   6.497 8.17e-11
***
## OCCUPATION Transport-moving         5.313e-01  1.786e-01   2.976  0.00292
**
## EDUCATION 11th                      1.472e-01  2.946e-01   0.500  0.61738
## EDUCATION 12th                      1.006e+00  3.926e-01   2.563  0.01036
*
## EDUCATION 1st-4th                  -1.076e+00  6.485e-01  -1.659  0.09711
.
## EDUCATION 5th-6th                  -4.466e-01  4.135e-01  -1.080  0.28016
## EDUCATION 7th-8th                  -2.587e-01  3.174e-01  -0.815  0.41503
## EDUCATION 9th                      -1.417e-01  3.656e-01  -0.388  0.69828
## EDUCATION Assoc-acdm                1.614e+00  2.554e-01   6.319 2.64e-10
***
## EDUCATION Assoc-voc                 1.538e+00  2.394e-01   6.425 1.32e-10
***
## EDUCATION Bachelors                 2.122e+00  2.192e-01   9.679  < 2e-16
***
## EDUCATION Doctorate                 3.191e+00  3.263e-01   9.780  < 2e-16
***
## EDUCATION HS-grad                   9.401e-01  2.109e-01   4.458 8.29e-06
***
## EDUCATION Masters                   2.465e+00  2.391e-01  10.308  < 2e-16
***
## EDUCATION Preschool                -1.225e+01  1.816e+02  -0.067  0.94624
## EDUCATION Prof-school               2.789e+00  2.992e-01   9.321  < 2e-16
***
## EDUCATION Some-college              1.335e+00  2.148e-01   6.215 5.13e-10
***
## HOURSPERWEEK                        3.473e-02  2.634e-03  13.184  < 2e-16
***
## CAPITALGAIN                         3.407e-04  1.879e-05  18.130  < 2e-16
***
## SEX Male                            1.595e-01  7.558e-02   2.110  0.03489
*
## ---
## 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:  8599.7  on 10936  degrees of freedom
## AIC: 8679.7
##
## Number of Fisher Scoring iterations: 13
```

End of Stage 4

Start of Stage 5

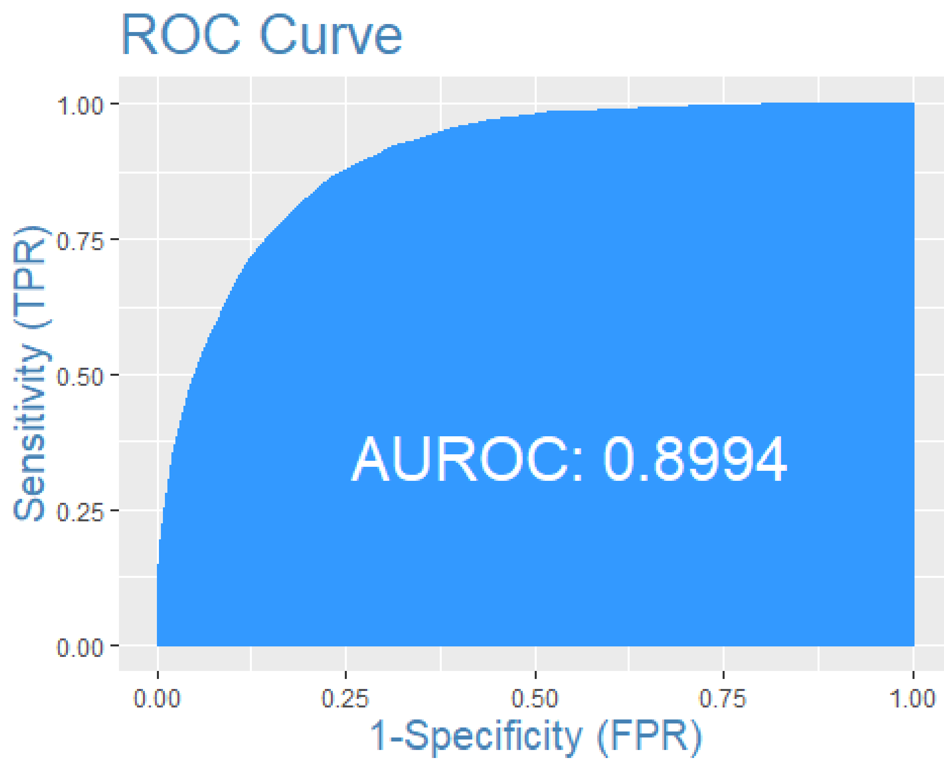
5. Evaluation

We need to evaluate the model using the test data. Evaluation checks a number of parameters for accuracy. In classication problems, we should be checking the followin parameters

- ROC: Receiver Operating Characteristics Curve traces the percentage of true positives accurately predicted by a given logit model as the prediction probability cutoff is lowered from 1 to 0. For a good model, as the cutoff is lowered, it should mark more of actual 1's as positives and lesser of actual 0's as 1's.

- So for a good model, the curve should rise steeply, indicating that the TPR (Y-Axis) increases faster than the FPR (X-Axis) as the cutoff score decreases. Greater the area under the ROC curve, better the predictive ability of the model.

```
# The model has area under ROC curve 89.7%, which is pretty good
plotROC(testData$ABOVE50K, predicted)
```



Specificity and Sensitivity

- Sensitivity (or True Positive Rate) is the percentage of 1's (actuals) correctly predicted by the model
- Specificity is the percentage of 0's (actuals) correctly predicted.
- Specificity can also be calculated as 1 - False Positive Rate.

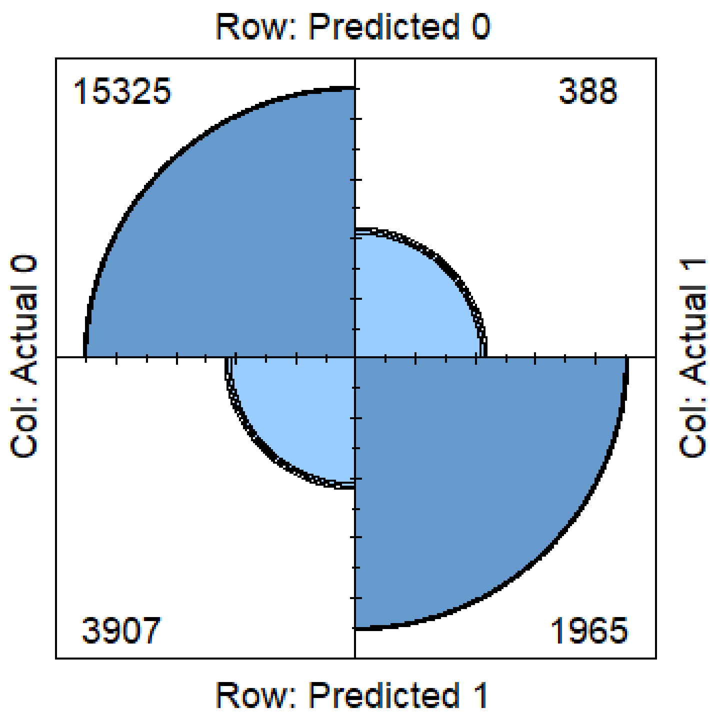
```
sensitivity(testData$ABOVE50K, predicted)

## [1] 0.8351041

specificity(testData$ABOVE50K, predicted)

## [1] 0.796849
```

The above numbers are calculated on the validation sample that was not used for training the model. So, a truth detection rate of 82% on test data is good.

Confusion Matrix

```
cm <- as.data.frame(confusionMatrix(testData$ABOVE50K, predicted))
colnames(cm) <- c("Actual 0", "Actual 1")
rownames(cm) <- c("Predicted 0", "Predicted 1")
cm

##             Actual 0 Actual 1
## Predicted 0    15325      388
## Predicted 1     3907     1965
```

```r
fourfoldplot(as.matrix(cm))
```



End of Stage 5

Start of Stage 6

6. Deployment

- Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.
- It often involves applying "live" models within an organization's decision making processes. For example, real-time personalization of Web pages or repeated scoring of marketing databases.
- Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable Data Analytics process across the enterprise.

- In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

- The deployment of the model will depend on the IT/product architecture, with which it needs to be integrated. The model could run outside the IT/product architecture. The output could be integrated with the system using API or similar interface.

- If the model needs to be integrated with a product (like GTOS), then the product should be able to support ML algorithms. Deployment is driven by IT and engineerng team with the support from the data scientist.

End of Stage 6

Start of Stage 7

7. Maintenance and Support

- A Data Analytic product could be created and deployed in less than a year. However, the maintenance and support of the product could run into years.

- This phase is very important because of changing nature of data and processes within an organisation. The data product may require fine tuning to accommodate the new realities.

Plan Maintenance and Support Roadmap

- Important if the Data Analytics results become part of the day-to-day business and IT environment

- Helps to avoid unnecessarily long periods of incorrect usage of Data Analytics results

- Needs a detailed plan on monitoring process

- Takes into account the specific type of deployment

End of the Script