IMPORTS

```
library(ggplot2)
library(caTools)
library(Amelia)
library(corrplot)
library(locfit)
library(dplyr)
library(ggthemes)
library (ROCR)
options(stringsAsFactors = TRUE)
```

DATA IMPORT AND INSPECTION

```
adult <- read.csv('C:/Users/adils/Desktop/ESCP/R & Business Analytics/Assignment 3/income.csv')

print(head(adult))
```

```
##   age   workclass    education education.num        marital.status
## 1  34 Federal-gov    Bachelors            13         Never-married
## 2  40     Private    Bachelors            13 Married-spouse-absent
## 3  37     Private      1st-4th             2    Married-civ-spouse
## 4  25     Private Some-college            10         Never-married
## 5  27     Private Some-college            10    Married-civ-spouse
## 6  40     Private      7th-8th             4    Married-civ-spouse
##        occupation   relationship               race    sex capital.gain
## 1 Exec-managerial      Unmarried Asian-Pac-Islander   Male         1471
## 2  Prof-specialty  Not-in-family Asian-Pac-Islander   Male        13550
## 3    Craft-repair        Husband Asian-Pac-Islander   Male            0
## 4    Craft-repair Other-relative Asian-Pac-Islander Female            0
## 5  Prof-specialty        Husband Asian-Pac-Islander   Male            0
## 6   Other-service        Husband Asian-Pac-Islander   Male            0
##   capital.loss hours.per.week native.country income
## 1            0             40       Cambodia  <=50K
## 2            0             40       Cambodia   >50K
## 3            0             40       Cambodia  <=50K
## 4            0             40       Cambodia  <=50K
## 5            0             40       Cambodia   >50K
## 6            0             42       Cambodia  <=50K
```

```
print (str(adult))
```

```
## 'data.frame':    7125 obs. of  14 variables:
##  $ age           : int  34 40 37 25 27 40 32 51 28 27 ...
##  $ workclass     : Factor w/ 8 levels "?","Federal-gov",..: 2 4 4 4 4 4 4 4 4 4 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 10 10 4 16 16 6 14 12 12 16 ...
##  $ education.num : int  13 13 2 10 10 4 1 9 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 4 3 5 3 3 3 3 5 5
...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 5 11 4 4 11 9 8 13 8 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 5 2 1 3 1 1 6 1 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 1 2 2 1 2 2 2 ...
##  $ capital.gain  : int  1471 13550 0 0 0 0 4508 0 0 0 ...
##  $ capital.loss  : int  0 0 0 0 0 0 0 0 0 1876 ...
##  $ hours.per.week: int  40 40 40 40 40 42 40 50 40 45 ...
##  $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 2 1 1 2 1 1 1 1 1 ...
## NULL
```

```
print(summary(adult))
```

```
##       age                   workclass           education    education.num
##  Min.   :17.00   Private         :5119   HS-grad     :2115   Min.   : 1.000
##  1st Qu.:28.00   Self-emp-not-inc: 521   Some-college:1452   1st Qu.: 9.000
##  Median :37.00   Local-gov       : 410   Bachelors   :1158   Median :10.000
##  Mean   :38.38   ?               : 407   Masters     : 363   Mean   : 9.717
##  3rd Qu.:47.00   State-gov       : 244   11th        : 276   3rd Qu.:12.000
##  Max.   :90.00   Self-emp-inc    : 237   Assoc-voc   : 268   Max.   :16.000
##                  (Other)         : 187   (Other)     :1493
##                marital.status           occupation           relationship
##  Divorced              : 855   Craft-repair   : 887   Husband       :2807
##  Married-AF-spouse     :   5   Prof-specialty : 874   Not-in-family :1775
##  Married-civ-spouse    :3263   Other-service  : 841   Other-relative: 354
##  Married-spouse-absent: 179   Exec-managerial: 806   Own-child     : 996
##  Never-married         :2343   Adm-clerical   : 780   Unmarried     : 814
##  Separated             : 263   Sales          : 750   Wife          : 379
##  Widowed               : 217   (Other)        :2187
##                  race          sex         capital.gain    capital.loss
##  Amer-Indian-Eskimo:  59   Female:2341   Min.   :    0   Min.   :   0.00
##  Asian-Pac-Islander: 702   Male  :4784   1st Qu.:    0   1st Qu.:   0.00
##  Black             : 655                 Median :    0   Median :   0.00
##  Other             : 139                 Mean   :  923   Mean   :  83.42
##  White             :5570                 3rd Qu.:    0   3rd Qu.:   0.00
##                                          Max.   :99999   Max.   :2603.00
##
##  hours.per.week        native.country   income
##  Min.   : 1.00   United-States:4355   <=50K:5536
##  1st Qu.:40.00   Mexico       : 643   >50K :1589
##  Median :40.00   Philippines  : 198
##  Mean   :40.37   Germany      : 137
##  3rd Qu.:45.00   Puerto-Rico  : 114
##  Max.   :99.00   Canada       : 107
##                  (Other)      :1571
```

CLEANING DATA

Cleaning workclass column

```
print(table(adult$workclass))
```

```
##
##               ?      Federal-gov        Local-gov          Private
##             407              185              410             5119
##     Self-emp-inc Self-emp-not-inc        State-gov      Without-pay
##             237              521              244                2
```

```
job.cleaning <- function(job){
  job <- as.character(job)
  if (job=='Never-worked' | job=='Without-pay'){
    return('Unemployed')
  }else if (job=='Local-gov' | job=='State-gov'){
    return("SL-gov")
  }else if (job=='Self-emp-inc' | job=='Self-emp-not-inc'){
    return("self-emp")
  }else{
    return(job)
  }
}

adult$workclass <- sapply(adult$workclass,job.cleaning)
```

Cleaning marital.status column

```
print(table(adult$marital.status))
```

```
##
##              Divorced     Married-AF-spouse    Married-civ-spouse
##                   855                     5                  3263
## Married-spouse-absent         Never-married             Separated
##                   179                  2343                   263
##               Widowed
##                   217
```

```
group_marital <- function(mar){
  mar <- as.character(mar)
  if (mar=='Separated' | mar=='Divorced' | mar=='Widowed'){
    return('Not-Married')
  }else if(mar=='Never-married'){
    return(mar)
  }else{
    return('Married')
  }
}
adult$marital.status <- sapply(adult$marital.status,group_marital)
```

Cleaning native.country column

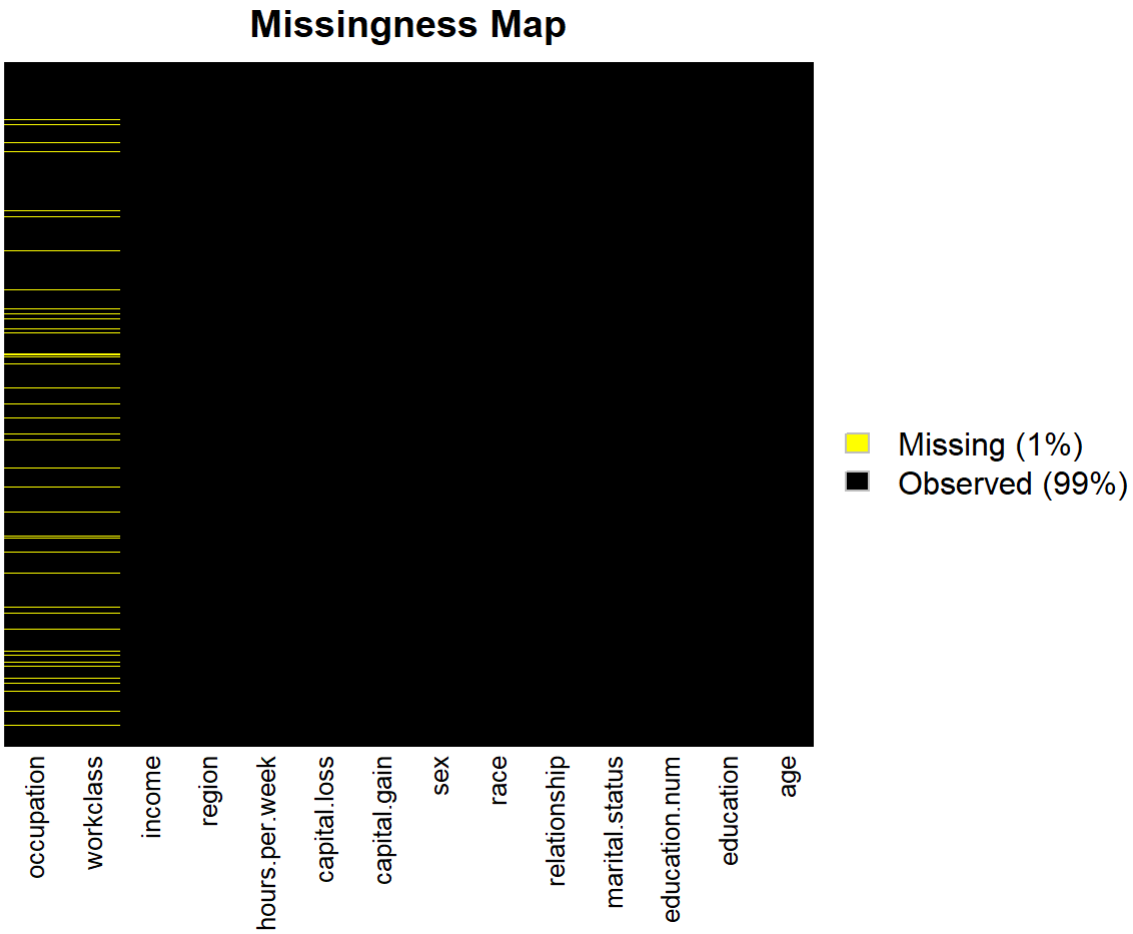
```
print(levels(adult$native.country))
```

```
##  [1] "Cambodia"                  "Canada"
##  [3] "China"                     "Columbia"
##  [5] "Cuba"                      "Dominican-Republic"
##  [7] "Ecuador"                   "El-Salvador"
##  [9] "England"                   "France"
## [11] "Germany"                   "Greece"
## [13] "Guatemala"                 "Haiti"
## [15] "Holand-Netherlands"        "Honduras"
## [17] "Hong"                      "Hungary"
## [19] "India"                     "Iran"
## [21] "Ireland"                   "Italy"
## [23] "Jamaica"                   "Japan"
## [25] "Laos"                      "Mexico"
## [27] "Nicaragua"                 "Outlying-US(Guam-USVI-etc)"
## [29] "Peru"                      "Philippines"
## [31] "Poland"                    "Portugal"
## [33] "Puerto-Rico"               "Scotland"
## [35] "South"                     "Taiwan"
## [37] "Thailand"                  "Trinadad&Tobago"
## [39] "United-States"             "Vietnam"
## [41] "Yugoslavia"
```

```
Asia <- c('China','Hong','India','Iran','Cambodia','Japan', 'Laos' ,
          'Philippines' ,'Vietnam' ,'Taiwan', 'Thailand')

North.America <- c('Canada','United-States','Puerto-Rico' )

Europe <- c('England' ,'France', 'Germany' ,'Greece','Holand-Netherlands','Hungary',
            'Ireland','Italy','Poland','Portugal','Scotland','Yugoslavia')

Latin.and.South.America <- c('Columbia','Cuba','Dominican-Republic','Ecuador',
                             'El-Salvador','Guatemala','Haiti','Honduras',
                             'Mexico','Nicaragua','Outlying-US(Guam-USVI-etc)','Peru',
                             'Jamaica','Trinadad&Tobago')
Other <- c('South')

group_country <- function(ctry){
  if (ctry %in% Asia){
    return('Asia')
  }else if (ctry %in% North.America){
    return('North.America')
  }else if (ctry %in% Europe){
    return('Europe')
  }else if (ctry %in% Latin.and.South.America){
    return('Latin.and.South.America')
  }else{
    return('Other')
  }
}

adult$native.country <- sapply(adult$native.country,group_country)
names(adult)[names(adult)=="native.country"] <- "region"
```

Dealing with missing data

```
adult[adult=="?"] <- NA
missmap(adult,y.at=c(1),y.labels = c(''),col=c('yellow','black'))
```




```
adult <- na.omit(adult)
```

Putting factor levels on the columns we changed

```
adult$workclass <- sapply(adult$workclass,factor)
adult$region <- sapply(adult$region,factor)
adult$marital.status <- sapply(adult$marital.status,factor)
adult$occupation <- sapply(adult$occupation,factor)
```

DATA EXPLORATION

```
plot(ggplot(adult,aes(age)) + geom_histogram(aes(fill=income),color='black',binwidth=1))
```


```
dev.off()
```

```
## null device
##           1
```

```
plot(ggplot(adult,aes(hours.per.week)) + geom_histogram())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
dev.off()
```

```
## null device
##           1
```

```
plot(ggplot(adult,aes(region)) + geom_bar(aes(fill=income),color='black')+theme_bw()+theme(axis.
text.x = element_text(angle = 90, hjust = 1)))
```

MODEL BUILDING

First Model

```
sample <- sample.split(adult$income, SplitRatio = 0.8421)
train <- subset(adult, sample == TRUE)
test <- subset(adult, sample == FALSE)
model <-  glm(income ~ ., family = binomial(logit), data = train)
```

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

```
print(summary(model))
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial(logit), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1294  -0.4956  -0.1924  -0.0312   3.6958
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -5.091e+00  9.799e-01  -5.196 2.04e-07 ***
## age                           2.181e-02  3.978e-03   5.482 4.21e-08 ***
## workclassPrivate             -6.421e-01  2.256e-01  -2.847 0.004418 **
## workclassself-emp            -9.305e-01  2.516e-01  -3.698 0.000218 ***
## workclassSL-gov              -8.593e-01  2.526e-01  -3.402 0.000669 ***
## workclassUnemployed          -1.443e+01  1.467e+03  -0.010 0.992150
## education11th                 9.102e-02  4.546e-01   0.200 0.841304
## education12th                 3.562e-01  5.603e-01   0.636 0.525000
## education1st-4th             -5.026e-01  6.640e-01  -0.757 0.449086
## education5th-6th             -1.364e-01  5.029e-01  -0.271 0.786272
## education7th-8th             -8.171e-01  5.016e-01  -1.629 0.103312
## education9th                 -5.573e-01  5.391e-01  -1.034 0.301304
## educationAssoc-acdm           1.004e+00  3.867e-01   2.596 0.009418 **
## educationAssoc-voc            8.911e-01  3.662e-01   2.433 0.014965 *
## educationBachelors            1.505e+00  3.261e-01   4.614 3.96e-06 ***
## educationDoctorate            2.262e+00  4.495e-01   5.033 4.83e-07 ***
## educationHS-grad              5.248e-01  3.168e-01   1.657 0.097569 .
## educationMasters              1.552e+00  3.541e-01   4.384 1.16e-05 ***
## educationPreschool           -1.890e+01  3.019e+02  -0.063 0.950070
## educationProf-school          2.421e+00  4.244e-01   5.705 1.16e-08 ***
## educationSome-college         7.245e-01  3.238e-01   2.237 0.025258 *
## education.num                        NA         NA      NA       NA
## marital.statusMarried         8.452e-01  3.702e-01   2.283 0.022413 *
## marital.statusNot-Married     6.198e-01  2.042e-01   3.035 0.002403 **
## occupationProf-specialty     -1.497e-01  1.532e-01  -0.977 0.328682
## occupationCraft-repair       -7.264e-01  1.552e-01  -4.682 2.84e-06 ***
## occupationOther-service      -1.508e+00  2.313e-01  -6.520 7.05e-11 ***
## occupationMachine-op-inspct  -1.252e+00  2.155e-01  -5.809 6.29e-09 ***
## occupationSales              -5.925e-01  1.563e-01  -3.790 0.000150 ***
## occupationFarming-fishing    -2.196e+00  3.352e-01  -6.551 5.73e-11 ***
## occupationAdm-clerical       -1.024e+00  1.839e-01  -5.567 2.59e-08 ***
## occupationHandlers-cleaners  -1.070e+00  2.819e-01  -3.797 0.000147 ***
## occupationTech-support        3.652e-02  2.360e-01   0.155 0.877035
## occupationTransport-moving   -8.451e-01  2.140e-01  -3.948 7.88e-05 ***
## occupationProtective-serv    -4.385e-01  3.049e-01  -1.438 0.150311
## occupationPriv-house-serv    -1.333e+01  2.981e+02  -0.045 0.964336
## occupationArmed-Forces       -1.459e+01  1.613e+03  -0.009 0.992786
## relationshipNot-in-family    -1.378e+00  3.575e-01  -3.855 0.000116 ***
## relationshipOther-relative   -1.494e+00  4.559e-01  -3.278 0.001047 **
## relationshipOwn-child        -2.702e+00  5.078e-01  -5.321 1.03e-07 ***
## relationshipUnmarried        -1.529e+00  4.080e-01  -3.748 0.000178 ***
## relationshipWife              1.280e+00  2.430e-01   5.268 1.38e-07 ***
## raceAsian-Pac-Islander        1.043e+00  7.538e-01   1.383 0.166587
```

```
## raceBlack                       1.256e+00  7.179e-01   1.750 0.080090 .
## raceOther                       3.780e-01  8.494e-01   0.445 0.656259
## raceWhite                       1.256e+00  6.992e-01   1.796 0.072473 .
## sexMale                         6.749e-01  1.899e-01   3.555 0.000379 ***
## capital.gain                    3.005e-04  2.466e-05  12.187  < 2e-16 ***
## capital.loss                    5.040e-04  9.125e-05   5.523 3.33e-08 ***
## hours.per.week                  3.260e-02  4.057e-03   8.036 9.32e-16 ***
## regionNorth.America            -5.722e-02  2.866e-01  -0.200 0.841759
## regionLatin.and.South.America -6.584e-01  3.165e-01  -2.080 0.037501 *
## regionEurope                    8.883e-02  3.151e-01   0.282 0.777973
## regionOther                    -7.447e-01  4.411e-01  -1.688 0.091389 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6115.6  on 5656  degrees of freedom
## Residual deviance: 3565.8  on 5604  degrees of freedom
## AIC: 3671.8
##
## Number of Fisher Scoring iterations: 15
```

New Model

```
new.step.model <- step(model)
```

```
## Start:  AIC=3671.79
## income ~ age + workclass + education + education.num + marital.status +
##     occupation + relationship + race + sex + capital.gain + capital.loss +
##     hours.per.week + region
##
##
## Step:  AIC=3671.79
## income ~ age + workclass + education + marital.status + occupation +
##     relationship + race + sex + capital.gain + capital.loss +
##     hours.per.week + region
##
##                  Df Deviance    AIC
## - race            4   3573.6 3671.6
## <none>                3565.8 3671.8
## - marital.status  2   3577.4 3679.4
## - workclass       4   3582.8 3680.8
## - sex             1   3578.7 3682.7
## - region          4   3586.7 3684.7
## - age             1   3595.9 3699.9
## - capital.loss    1   3596.7 3700.7
## - hours.per.week  1   3632.4 3736.4
## - relationship    5   3653.0 3749.0
## - occupation     13   3692.7 3772.7
## - education      15   3728.1 3804.1
## - capital.gain    1   3831.6 3935.6
##
## Step:  AIC=3671.58
## income ~ age + workclass + education + marital.status + occupation +
##     relationship + sex + capital.gain + capital.loss + hours.per.week +
##     region
##
##                  Df Deviance    AIC
## <none>                3573.6 3671.6
## - marital.status  2   3584.8 3678.8
## - workclass       4   3590.5 3680.5
## - sex             1   3586.6 3682.6
## - region          4   3598.2 3688.2
## - age             1   3604.5 3700.5
## - capital.loss    1   3604.8 3700.8
## - hours.per.week  1   3640.1 3736.1
## - relationship    5   3662.3 3750.3
## - occupation     13   3702.1 3774.1
## - education      15   3735.9 3803.9
## - capital.gain    1   3837.7 3933.7
```

```
print(summary(new.step.model))
```

```
##
## Call:
## glm(formula = income ~ age + workclass + education + marital.status +
##     occupation + relationship + sex + capital.gain + capital.loss +
##     hours.per.week + region, family = binomial(logit), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1118  -0.4964  -0.1925  -0.0337   3.7078
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -3.996e+00  6.409e-01  -6.235 4.50e-10 ***
## age                          2.205e-02  3.973e-03   5.550 2.86e-08 ***
## workclassPrivate            -6.430e-01  2.247e-01  -2.861 0.004220 **
## workclassself-emp           -9.317e-01  2.508e-01  -3.715 0.000204 ***
## workclassSL-gov             -8.531e-01  2.519e-01  -3.387 0.000707 ***
## workclassUnemployed         -1.446e+01  1.461e+03  -0.010 0.992103
## education11th                9.999e-02  4.537e-01   0.220 0.825562
## education12th                3.200e-01  5.604e-01   0.571 0.568014
## education1st-4th            -4.933e-01  6.632e-01  -0.744 0.456989
## education5th-6th            -1.696e-01  5.046e-01  -0.336 0.736765
## education7th-8th            -8.207e-01  5.005e-01  -1.640 0.101059
## education9th                -5.300e-01  5.389e-01  -0.984 0.325359
## educationAssoc-acdm          1.019e+00  3.860e-01   2.640 0.008296 **
## educationAssoc-voc           8.913e-01  3.657e-01   2.437 0.014791 *
## educationBachelors           1.503e+00  3.255e-01   4.617 3.88e-06 ***
## educationDoctorate           2.274e+00  4.495e-01   5.059 4.22e-07 ***
## educationHS-grad             5.273e-01  3.162e-01   1.667 0.095443 .
## educationMasters             1.560e+00  3.535e-01   4.414 1.02e-05 ***
## educationPreschool          -1.873e+01  3.000e+02  -0.062 0.950219
## educationProf-school         2.416e+00  4.235e-01   5.706 1.16e-08 ***
## educationSome-college        7.231e-01  3.232e-01   2.237 0.025283 *
## marital.statusMarried        8.038e-01  3.688e-01   2.180 0.029287 *
## marital.statusNot-Married    6.151e-01  2.038e-01   3.018 0.002541 **
## occupationProf-specialty    -1.466e-01  1.530e-01  -0.959 0.337757
## occupationCraft-repair      -7.314e-01  1.551e-01  -4.715 2.42e-06 ***
## occupationOther-service     -1.516e+00  2.307e-01  -6.572 4.96e-11 ***
## occupationMachine-op-inspct -1.274e+00  2.147e-01  -5.931 3.01e-09 ***
## occupationSales             -5.880e-01  1.562e-01  -3.765 0.000167 ***
## occupationFarming-fishing   -2.192e+00  3.345e-01  -6.551 5.71e-11 ***
## occupationAdm-clerical      -1.015e+00  1.836e-01  -5.529 3.22e-08 ***
## occupationHandlers-cleaners -1.061e+00  2.811e-01  -3.773 0.000161 ***
## occupationTech-support       4.922e-02  2.358e-01   0.209 0.834681
## occupationTransport-moving  -8.538e-01  2.135e-01  -4.000 6.34e-05 ***
## occupationProtective-serv   -4.391e-01  3.044e-01  -1.442 0.149209
## occupationPriv-house-serv   -1.339e+01  2.967e+02  -0.045 0.963995
## occupationArmed-Forces      -1.457e+01  1.615e+03  -0.009 0.992801
## relationshipNot-in-family   -1.419e+00  3.558e-01  -3.989 6.63e-05 ***
## relationshipOther-relative  -1.525e+00  4.554e-01  -3.349 0.000811 ***
## relationshipOwn-child       -2.729e+00  5.073e-01  -5.378 7.52e-08 ***
## relationshipUnmarried       -1.564e+00  4.070e-01  -3.842 0.000122 ***
## relationshipWife             1.277e+00  2.426e-01   5.265 1.40e-07 ***
```

```
## sexMale                         6.749e-01  1.896e-01   3.560 0.000371 ***
## capital.gain                    2.983e-04  2.455e-05  12.151  < 2e-16 ***
## capital.loss                    5.064e-04  9.130e-05   5.546 2.92e-08 ***
## hours.per.week                  3.249e-02  4.044e-03   8.034 9.40e-16 ***
## regionNorth.America             1.257e-01  1.439e-01   0.874 0.382214
## regionLatin.and.South.America -5.045e-01  1.960e-01  -2.574 0.010049 *
## regionEurope                    2.854e-01  1.921e-01   1.486 0.137397
## regionOther                    -7.643e-01  4.394e-01  -1.739 0.081950 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6115.6  on 5656  degrees of freedom
## Residual deviance: 3573.6  on 5608  degrees of freedom
## AIC: 3671.6
##
## Number of Fisher Scoring iterations: 15
```

```
test$predicted.income = predict(new.step.model, newdata=test, type="response")
print(table(test$income, test$predicted.income > 0.5))
```

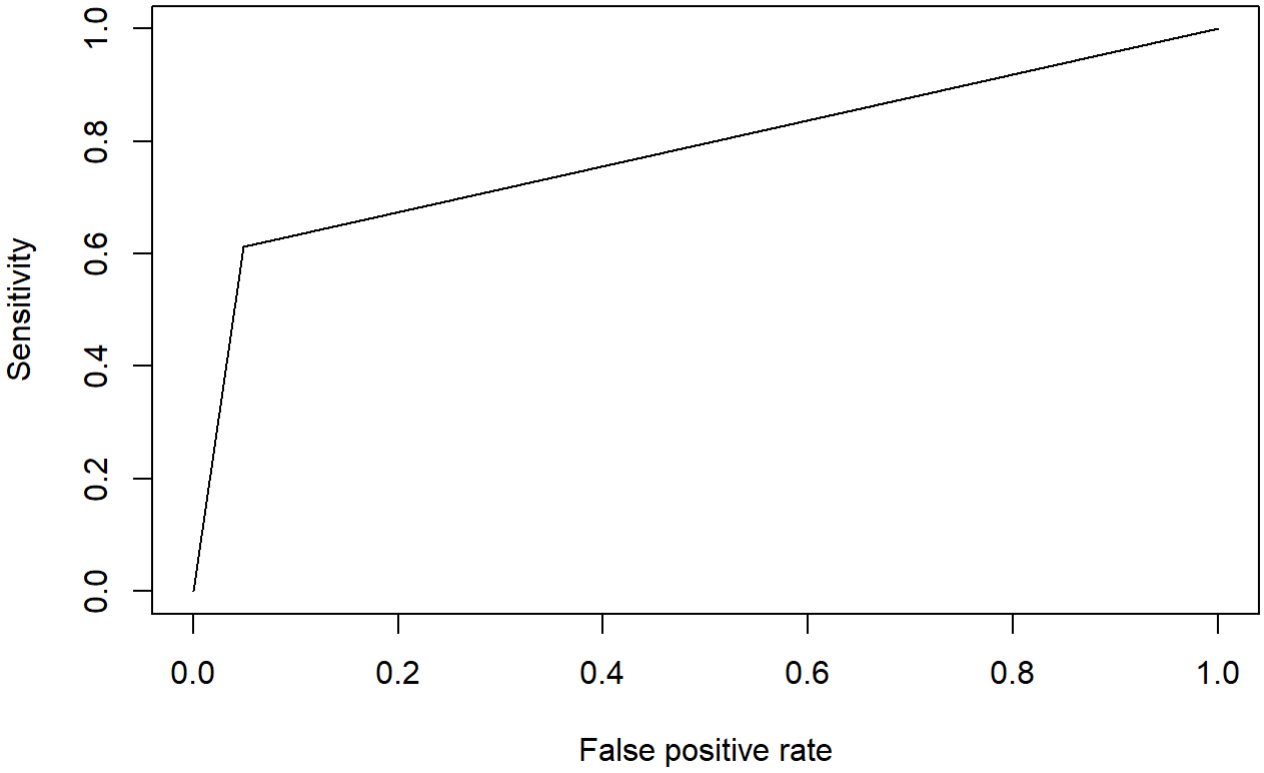
```
##
##         FALSE TRUE
##   <=50K   776   40
##   >50K     95  150
```

```
predicted <- ifelse(test$predicted.income > 0.5,1,0)
actual <- ifelse(test$income == ">50K",1,0)
misClasificError <- mean(predicted != actual)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.872761545711593"
```

ROC

```
data <- data.frame(predicted,actual)
pred <- prediction(data$predicted,data$actual)
perf <- performance(pred, "sens", "fpr")
plot(perf)
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate