Final Project on Introduction to Computer Language

Darix SAMANI SIEWE

2024-09-20

Introduction

In this project, you will import the `data.txt`, perform some data processing, visualization and modelling.

Required packages

```r
# list all packages here
install.packages("caret", dependencies = TRUE)
install.packages("ggplot2", dependencies = TRUE)
install.packages("moments", dependencies = TRUE)
install.packages("rpart", dependencies = TRUE)
install.packages("rpart.plot", dependencies = TRUE)

library(moments)
library(ggplot2)
library(rpart.plot)
library(caret)
library(rpart)
library(rpart.plot)
# library(caret)
```

Inport data

```r
data <- read.csv("./data.txt")
head(data)
```

```
##   X39         State.gov X77516  Bachelors X13          Never.married
## 1  50  Self-emp-not-inc  83311  Bachelors  13     Married-civ-spouse
## 2  38           Private 215646    HS-grad   9               Divorced
## 3  53           Private 234721       11th   7     Married-civ-spouse
## 4  28           Private 338409  Bachelors  13     Married-civ-spouse
## 5  37           Private 284582    Masters  14     Married-civ-spouse
## 6  49           Private 160187        9th   5  Married-spouse-absent
##         Adm.clerical  Not.in.family  White    Male X2174 X0 X40  United.States
## 1    Exec-managerial        Husband  White    Male     0  0  13  United-States
## 2  Handlers-cleaners  Not-in-family  White    Male     0  0  40  United-States
## 3  Handlers-cleaners        Husband  Black    Male     0  0  40  United-States
## 4     Prof-specialty           Wife  Black  Female     0  0  40           Cuba
## 5    Exec-managerial           Wife  White  Female     0  0  40  United-States
## 6      Other-service  Not-in-family  Black  Female     0  0  16        Jamaica
##   X..50K
## 1  <=50K
## 2  <=50K
## 3  <=50K
## 4  <=50K
## 5  <=50K
## 6  <=50K
```

Q1. What do you see when you look at the column name of data?

```
colnames(data)
```

```
##  [1] "X39"           "State.gov"     "X77516"        "Bachelors"    
##  [5] "X13"           "Never.married" "Adm.clerical"  "Not.in.family"
##  [9] "White"         "Male"          "X2174"         "X0"           
## [13] "X40"           "United.States" "X..50K"
```

Answer: The first row is consider as a cologne of our dataset(dataframe).

Q2. How can we solve the problem in Q1.? *Hint: explore the arguments of the `read.csv()` function.*

```
data <- read.csv("./data.txt", header=FALSE, sep=",")
head(data)
```

```
##   V1                V2     V3         V4 V5                  V6
## 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
##                   V7             V8     V9     V10  V11 V12 V13            V14
## 1       Adm-clerical  Not-in-family  White    Male 2174   0  40  United-States
## 2    Exec-managerial        Husband  White    Male    0   0  13  United-States
## 3  Handlers-cleaners  Not-in-family  White    Male    0   0  40  United-States
## 4  Handlers-cleaners        Husband  Black    Male    0   0  40  United-States
## 5     Prof-specialty           Wife  Black  Female    0   0  40           Cuba
## 6    Exec-managerial           Wife  White  Female    0   0  40  United-States
##      V15
## 1  <=50K
## 2  <=50K
## 3  <=50K
## 4  <=50K
## 5  <=50K
## 6  <=50K
```

Q3. How many rows and columns does `data` have?

```
cat("Nombers of rows:", nrow(data), "\n")
```

```
## Nombers of rows: 32561
```

```
cat("numbers of colums:", length(colnames(data)))
```

```
## numbers of colums: 15
```

Q4. Change the column names of `data` in this order: `age`, `workclass`, `fnlwgt`, `education`, `education_num`, `marital_status`, `occupation`, `relationship`, `race`, `sex`, `capital_gain`, `capital_loss`, `hours_per_week`, `native_country`, `class`.

Check the data names here

```
colnames(data) <- c("age", "workclass", "fnlwgt", "education", "education_num", "marital_status", "occu
head(data)
```

```
##   age         workclass fnlwgt  education education_num      marital_status
## 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 captal_gain capital_loss
## 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
##   hours_per_week native_country  class
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40           Cuba  <=50K
## 6             40  United-States  <=50K
```

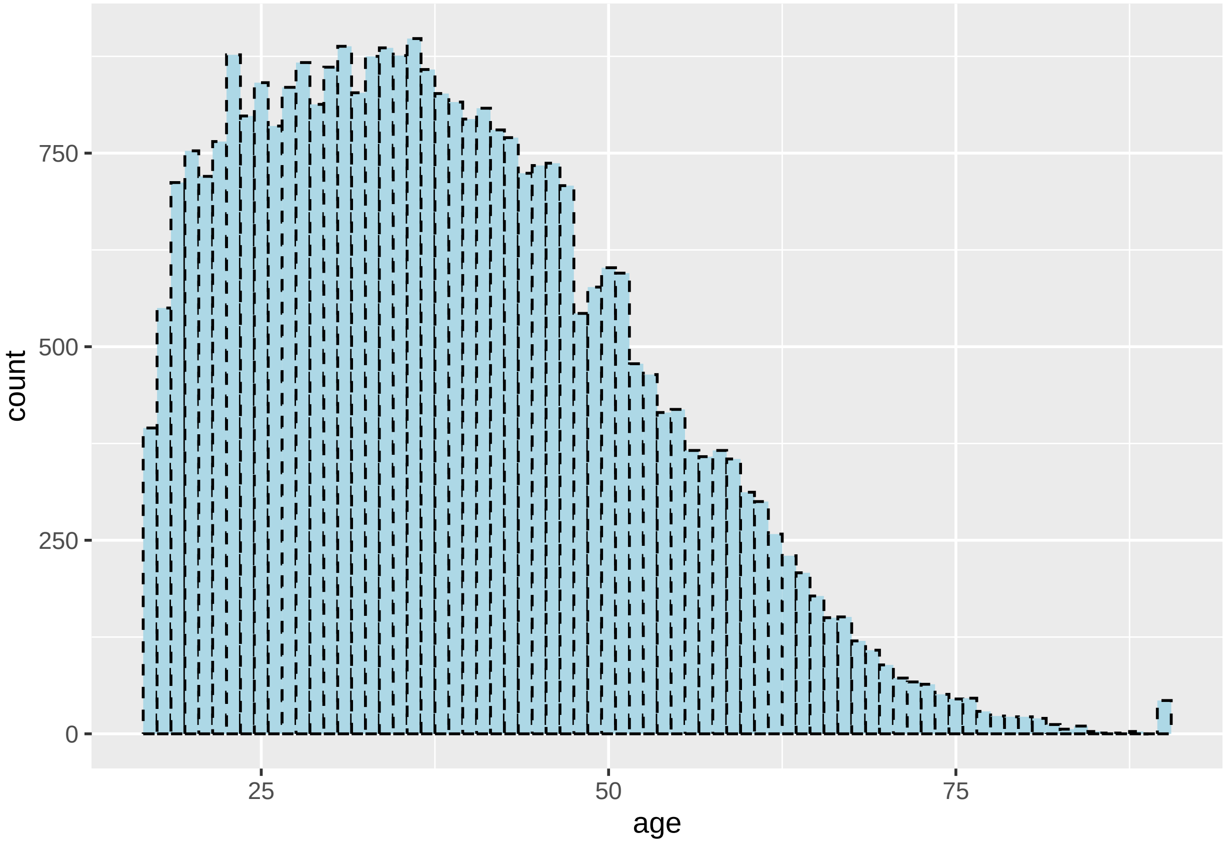
Q5. Replace all "?" in the data with `NA` (Not available)

```
 data[data==" ?"]<- NA
tail(data)
```

```
##       age     workclass fnlwgt     education education_num      marital_status
## 32556  22       Private 310152  Some-college            10       Never-married
## 32557  27       Private 257302    Assoc-acdm            12  Married-civ-spouse
## 32558  40       Private 154374       HS-grad             9  Married-civ-spouse
## 32559  58       Private 151910       HS-grad             9             Widowed
## 32560  22       Private 201490       HS-grad             9       Never-married
## 32561  52  Self-emp-inc 287927       HS-grad             9  Married-civ-spouse
##               occupation   relationship   race     sex captal_gain capital_loss
## 32556    Protective-serv  Not-in-family  White    Male           0            0
## 32557       Tech-support           Wife  White  Female           0            0
## 32558  Machine-op-inspct        Husband  White    Male           0            0
## 32559       Adm-clerical      Unmarried  White  Female           0            0
## 32560       Adm-clerical      Own-child  White    Male           0            0
## 32561    Exec-managerial           Wife  White  Female       15024            0
##       hours_per_week native_country  class
## 32556             40  United-States  <=50K
## 32557             38  United-States  <=50K
## 32558             40  United-States   >50K
## 32559             40  United-States  <=50K
## 32560             20  United-States  <=50K
## 32561             40  United-States   >50K
```

Q6. Plot the histogram of ages using ggplot2

```
ggplot(data, aes(x=age)) + geom_histogram(binwidth=1, color="black", fill="lightblue", linetype="dashed
```



Calculate the skewness of the variable `age` and comment about its distribution.

```
## R doen't have a native founction to compute the skewness of varibale, we need to load the package, mo
skewness(data$age)
```

```
## [1] 0.5587176
```

comments that a distribution is right skewed. A right skewed distribution would be biased towards higher values, such that the mean of the distribution will exceed the median of the distribution.

Q7. How many observation do we have for the `Private` category of the `workclass` variable?

```
length(which(data$workclass==" Private"))
```

```
## [1] 22696
```

Q8. How many `marital_status` are `Married-civ-spouse` for the `Private` workclass ?

```
length(which(data$marital_status == " Married-civ-spouse" & data$workclass == " Private"))
```

```
## [1] 9732
```

Q9. How many `marital_status` are `Married-civ-spouse` for the `Private` workclass and for each race?

```
x <- subset(data, marital_status == " Married-civ-spouse" & workclass == " Private")
tapply(x$marital_status, x[[9]], length)
```

```
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black               Other
##                  70                 336                 560                  82
##               White
##                 8684
```

Q10. How many `marital_status` are `Married-civ-spouse` for the `Private` workclass and for each sex?

```r
x <- subset(data, marital_status == " Married-civ-spouse" & workclass == " Private")
tapply(x$marital_status, x[[10]], length)
```

```
##  Female    Male
##    1064    8668
```

Q11. Recode the variable `class` to 0 if class is `<=50` and 1 else.

```r
data$class[data$class==" <=50K"] <- 0
data$class[data$class==" >50K"] <- 1
```

Q12. Replace NA with the mean if the variable is continuous and the mode if the variable is categorical.

```r
replace_na <- function (df){
  for (colname in colnames(df)){
    if (is.numeric((df[, colname]))){
      df[is.na(df[, colname]), colname] <- mean(data[, colname], na.rm = TRUE)
    }
    else {
     df[is.na(df[, colname]), colname] <- mode(data[, colname])
  }
  }
  df
}
data <- replace_na(data)
tail(data)
```

```
##       age     workclass fnlwgt     education education_num      marital_status
## 32556  22       Private 310152  Some-college            10       Never-married
## 32557  27       Private 257302    Assoc-acdm            12  Married-civ-spouse
## 32558  40       Private 154374       HS-grad             9  Married-civ-spouse
## 32559  58       Private 151910       HS-grad             9             Widowed
## 32560  22       Private 201490       HS-grad             9       Never-married
## 32561  52  Self-emp-inc 287927       HS-grad             9  Married-civ-spouse
##                occupation   relationship   race     sex captal_gain capital_loss
## 32556    Protective-serv  Not-in-family  White    Male           0            0
## 32557       Tech-support           Wife  White  Female           0            0
## 32558  Machine-op-inspct        Husband  White    Male           0            0
## 32559       Adm-clerical      Unmarried  White  Female           0            0
## 32560       Adm-clerical      Own-child  White    Male           0            0
## 32561    Exec-managerial           Wife  White  Female       15024            0
##       hours_per_week native_country class
## 32556             40  United-States     0
## 32557             38  United-States     0
## 32558             40  United-States     1
## 32559             40  United-States     0
## 32560             20  United-States     0
## 32561             40  United-States     1
```

Q.13 Split the data in train (80%) and test (20%) using the `caret` package. Set the seed to 20092024.

```r
set.seed(20092024)
library(caret)
trainIndex <- createDataPartition(data$class, p=0.8, list = FALSE)
```

```
trainData <- data[trainIndex,]


testData <- data[-trainIndex,]
```

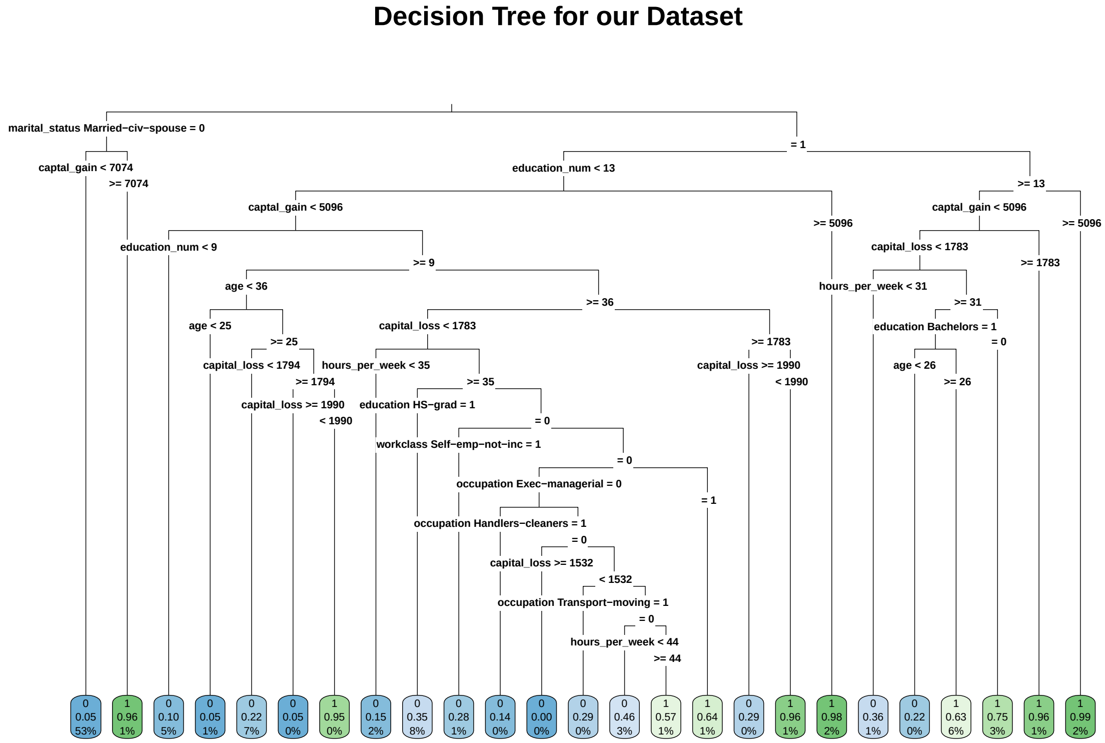
Q.14 Fit a decision tree with train set. What is the confusion matrix?

```
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

dtree_fit <- train(class ~., data = trainData, method = "rpart",
                   parms = list(split = "information"),
                   trControl=trctrl,
                   tuneLength = 10)

test_pred <- predict(dtree_fit, newdata = testData[, 1:14])
confusionMatrix(test_pred, as.factor(testData$class))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4675  667
##          1  269  901
##
##                Accuracy : 0.8563
##                  95% CI : (0.8475, 0.8647)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5696
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9456
##             Specificity : 0.5746
##          Pos Pred Value : 0.8751
##          Neg Pred Value : 0.7701
##              Prevalence : 0.7592
##          Detection Rate : 0.7179
##    Detection Prevalence : 0.8203
##       Balanced Accuracy : 0.7601
##
##        'Positive' Class : 0
##
```

a. Draw the decision three.

```
rpart.plot(dtree_fit$finalModel, type = 3, extra = "auto", main = "Decision Tree for our Dataset")
```



b. What is the accuracy of the model on the test set?

```
cm <- confusionMatrix(test_pred, as.factor(testData$class))

overall.accuracy <- cm$overall['Accuracy']
cat("accuracy on testing data : ", overall.accuracy, "\n")
```

```
## accuracy on testing data :  0.8562654
```

Q.15 Fit a generalized linear model with train set. What is the confusion matrix?

```
trainData$class <- as.numeric(trainData$class)
glm_model <- glm(class ~ ., data=trainData, family = "binomial")
predit_data_glm <- predict(glm_model, newdata=testData[, 1:14])

confusionMatrix(factor(predit_data_glm>0.5, levels = c(T,F), labels = c("1", "0")), as.factor(testData$c
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4758  821
##          1  186  747
##
##                Accuracy : 0.8454
##                  95% CI : (0.8363, 0.8541)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5092
##
##  Mcnemar's Test P-Value : < 2.2e-16
```

```
##
##             Sensitivity : 0.9624
##             Specificity : 0.4764
##          Pos Pred Value : 0.8528
##          Neg Pred Value : 0.8006
##              Prevalence : 0.7592
##          Detection Rate : 0.7307
##    Detection Prevalence : 0.8567
##       Balanced Accuracy : 0.7194
##
##        'Positive' Class : 0
##
```

a. Print the summary of the model.

```
summary(glm_model)
```

```
##
## Call:
## glm(formula = class ~ ., family = "binomial", data = trainData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1239  -0.4995  -0.1837  -0.0236   3.5956
##
## Coefficients: (2 not defined because of singularities)
##                                             Estimate Std. Error z value
## (Intercept)                               -6.467e+00  8.047e-01  -8.036
## age                                        2.455e-02  1.853e-03  13.250
## workclass Local-gov                       -6.294e-01  1.256e-01  -5.009
## workclass Never-worked                    -1.279e+01  4.849e+02  -0.026
## workclass Private                         -4.698e-01  1.049e-01  -4.480
## workclass Self-emp-inc                    -1.876e-01  1.373e-01  -1.366
## workclass Self-emp-not-inc                -9.499e-01  1.226e-01  -7.746
## workclass State-gov                       -7.896e-01  1.396e-01  -5.656
## workclass Without-pay                     -1.412e+01  3.841e+02  -0.037
## workclasscharacter                        -1.195e+00  1.582e-01  -7.557
## fnlwgt                                     7.696e-07  1.949e-07   3.949
## education 11th                             6.924e-02  2.350e-01   0.295
## education 12th                             4.359e-01  2.923e-01   1.491
## education 1st-4th                         -4.883e-01  5.397e-01  -0.905
## education 5th-6th                         -2.003e-01  3.549e-01  -0.564
## education 7th-8th                         -5.776e-01  2.633e-01  -2.194
## education 9th                             -4.295e-01  3.046e-01  -1.410
## education Assoc-acdm                       1.256e+00  1.957e-01   6.415
## education Assoc-voc                        1.265e+00  1.885e-01   6.709
## education Bachelors                        1.880e+00  1.747e-01  10.759
## education Doctorate                        3.119e+00  2.407e-01  12.961
## education HS-grad                          7.579e-01  1.702e-01   4.452
## education Masters                          2.272e+00  1.870e-01  12.152
## education Preschool                       -2.113e+01  3.331e+02  -0.063
## education Prof-school                      2.796e+00  2.229e-01  12.542
## education Some-college                     1.147e+00  1.727e-01   6.645
## education_num                                     NA         NA      NA
## marital_status Married-AF-spouse           2.685e+00  6.166e-01   4.354
## marital_status Married-civ-spouse       2.045e+00  2.909e-01    7.030
## marital_status Married-spouse-absent   -1.381e-01  2.620e-01   -0.527
## marital_status Never-married           -4.517e-01  9.797e-02   -4.610
## marital_status Separated               -2.233e-01  1.874e-01   -1.191
## marital_status Widowed                  4.210e-02  1.764e-01    0.239
## occupation Armed-Forces                -1.214e+00  1.552e+00   -0.782
## occupation Craft-repair                 4.811e-02  8.921e-02    0.539
## occupation Exec-managerial              7.286e-01  8.595e-02    8.477
## occupation Farming-fishing             -9.839e-01  1.556e-01   -6.324
## occupation Handlers-cleaners           -7.860e-01  1.630e-01   -4.823
## occupation Machine-op-inspct           -4.329e-01  1.158e-01   -3.740
## occupation Other-service               -9.443e-01  1.342e-01   -7.035
## occupation Priv-house-serv             -4.041e+00  1.706e+00   -2.368
## occupation Prof-specialty               4.537e-01  9.055e-02    5.010
## occupation Protective-serv              5.026e-01  1.390e-01    3.615
## occupation Sales                        2.478e-01  9.178e-02    2.700
## occupation Tech-support                 5.287e-01  1.231e-01    4.293
## occupation Transport-moving            -1.724e-01  1.113e-01   -1.548
## occupationcharacter                            NA         NA       NA
## relationship Not-in-family              4.162e-01  2.878e-01    1.446
## relationship Other-relative            -3.439e-01  2.672e-01   -1.287
## relationship Own-child                 -7.383e-01  2.839e-01   -2.600
## relationship Unmarried                  2.669e-01  3.064e-01    0.871
## relationship Wife                       1.386e+00  1.149e-01   12.060
## race Asian-Pac-Islander                 6.725e-01  2.945e-01    2.284
## race Black                              2.469e-01  2.538e-01    0.973
## race Other                             -3.072e-02  4.013e-01   -0.077
## race White                              4.531e-01  2.405e-01    1.884
## sex Male                                8.691e-01  8.871e-02    9.798
## captal_gain                             3.219e-04  1.159e-05   27.768
## capital_loss                            6.594e-04  4.152e-05   15.882
## hours_per_week                          2.988e-02  1.820e-03   16.415
## native_country Canada                  -3.874e-01  7.197e-01   -0.538
## native_country China                   -1.613e+00  7.463e-01   -2.161
## native_country Columbia                -1.482e+01  1.815e+02   -0.082
## native_country Cuba                    -4.733e-01  7.375e-01   -0.642
## native_country Dominican-Republic      -2.644e+00  1.234e+00   -2.143
## native_country Ecuador                 -1.076e+00  9.930e-01   -1.084
## native_country El-Salvador             -1.208e+00  8.481e-01   -1.424
## native_country England                 -6.405e-01  7.416e-01   -0.864
## native_country France                  -6.810e-01  9.015e-01   -0.755
## native_country Germany                 -4.693e-01  7.131e-01   -0.658
## native_country Greece                  -1.511e+00  8.840e-01   -1.709
## native_country Guatemala               -6.701e-01  1.002e+00   -0.669
## native_country Haiti                   -2.925e-01  9.842e-01   -0.297
## native_country Holand-Netherlands      -1.246e+01  1.455e+03   -0.009
## native_country Honduras                -1.940e+00  2.710e+00   -0.716
## native_country Hong                    -1.083e+00  9.298e-01   -1.165
## native_country Hungary                 -1.089e+00  1.204e+00   -0.904
## native_country India                   -1.681e+00  7.084e-01   -2.374
## native_country Iran                    -8.152e-01  7.830e-01   -1.041
## native_country Ireland                 -1.255e-01  9.283e-01   -0.135
## native_country Italy                   -2.566e-01  7.413e-01   -0.346
## native_country Jamaica                 -9.881e-01  8.276e-01   -1.194
## native_country Japan                      -6.194e-01  7.643e-01  -0.810
## native_country Laos                       -1.056e+00  1.094e+00  -0.966
## native_country Mexico                     -1.429e+00  6.991e-01  -2.045
## native_country Nicaragua                  -2.118e+00  1.275e+00  -1.661
## native_country Outlying-US(Guam-USVI-etc) -1.354e+01  4.256e+02  -0.032
## native_country Peru                       -1.679e+00  1.071e+00  -1.568
## native_country Philippines                -5.600e-01  6.802e-01  -0.823
## native_country Poland                     -7.682e-01  7.834e-01  -0.981
## native_country Portugal                   -1.183e+00  1.021e+00  -1.159
## native_country Puerto-Rico                -1.210e+00  7.903e-01  -1.531
## native_country Scotland                   -1.730e+00  1.326e+00  -1.305
## native_country South                      -1.928e+00  7.720e-01  -2.497
## native_country Taiwan                     -1.209e+00  8.263e-01  -1.463
## native_country Thailand                   -1.765e+00  1.116e+00  -1.582
## native_country Trinadad&Tobago            -1.307e+00  1.315e+00  -0.994
## native_country United-States              -6.747e-01  6.564e-01  -1.028
## native_country Vietnam                    -2.038e+00  8.701e-01  -2.343
## native_country Yugoslavia                 -5.600e-01  9.593e-01  -0.584
## native_countrycharacter                   -1.089e+00  6.667e-01  -1.633
##                                           Pr(>|z|)
## (Intercept)                               9.32e-16 ***
## age                                        < 2e-16 ***
## workclass Local-gov                       5.47e-07 ***
## workclass Never-worked                    0.978955
## workclass Private                         7.47e-06 ***
## workclass Self-emp-inc                    0.171845
## workclass Self-emp-not-inc                9.50e-15 ***
## workclass State-gov                       1.55e-08 ***
## workclass Without-pay                     0.970675
## workclasscharacter                        4.14e-14 ***
## fnlwgt                                    7.85e-05 ***
## education 11th                            0.768253
## education 12th                            0.135937
## education 1st-4th                         0.365531
## education 5th-6th                         0.572469
## education 7th-8th                         0.028244 *
## education 9th                             0.158558
## education Assoc-acdm                      1.41e-10 ***
## education Assoc-voc                       1.96e-11 ***
## education Bachelors                        < 2e-16 ***
## education Doctorate                        < 2e-16 ***
## education HS-grad                         8.51e-06 ***
## education Masters                          < 2e-16 ***
## education Preschool                       0.949417
## education Prof-school                      < 2e-16 ***
## education Some-college                    3.03e-11 ***
## education_num                                   NA
## marital_status Married-AF-spouse          1.34e-05 ***
## marital_status Married-civ-spouse         2.06e-12 ***
## marital_status Married-spouse-absent      0.598202
## marital_status Never-married              4.02e-06 ***
## marital_status Separated                  0.233531
## marital_status Widowed                    0.811341
## occupation Armed-Forces                   0.433940
## occupation Craft-repair                   0.589692
## occupation Exec-managerial                 < 2e-16 ***
## occupation Farming-fishing                2.55e-10 ***
## occupation Handlers-cleaners              1.41e-06 ***
## occupation Machine-op-inspct              0.000184 ***
## occupation Other-service                  1.99e-12 ***
## occupation Priv-house-serv                0.017876 *
## occupation Prof-specialty                 5.43e-07 ***
## occupation Protective-serv                0.000300 ***
## occupation Sales                          0.006929 **
## occupation Tech-support                   1.76e-05 ***
## occupation Transport-moving               0.121529
## occupationcharacter                             NA
## relationship Not-in-family                0.148107
## relationship Other-relative               0.198109
## relationship Own-child                    0.009310 **
## relationship Unmarried                    0.383734
## relationship Wife                          < 2e-16 ***
## race Asian-Pac-Islander                   0.022392 *
## race Black                                0.330715
## race Other                                0.938973
## race White                                0.059519 .
## sex Male                                   < 2e-16 ***
## captal_gain                                < 2e-16 ***
## capital_loss                               < 2e-16 ***
## hours_per_week                             < 2e-16 ***
## native_country Canada                     0.590413
## native_country China                      0.030658 *
## native_country Columbia                   0.934911
## native_country Cuba                       0.521048
## native_country Dominican-Republic         0.032108 *
## native_country Ecuador                    0.278400
## native_country El-Salvador                0.154407
## native_country England                    0.387726
## native_country France                     0.449994
## native_country Germany                    0.510456
## native_country Greece                     0.087397 .
## native_country Guatemala                  0.503744
## native_country Haiti                      0.766291
## native_country Holand-Netherlands         0.993170
## native_country Honduras                   0.474032
## native_country Hong                       0.243894
## native_country Hungary                    0.365842
## native_country India                      0.017618 *
## native_country Iran                       0.297867
## native_country Ireland                    0.892448
## native_country Italy                      0.729171
## native_country Jamaica                    0.232474
## native_country Japan                      0.417674
## native_country Laos                       0.334291
## native_country Mexico                     0.040879 *
## native_country Nicaragua                  0.096634 .
## native_country Outlying-US(Guam-USVI-etc) 0.974611
## native_country Peru                       0.116937
## native_country Philippines                0.410328
## native_country Poland                     0.326772
## native_country Portugal                   0.246629
## native_country Puerto-Rico                0.125878
## native_country Scotland                   0.191852
## native_country South                      0.012514 *
## native_country Taiwan                     0.143577
## native_country Thailand                   0.113656
## native_country Trinadad&Tobago            0.320089
## native_country United-States              0.304008
## native_country Vietnam                    0.019151 *
## native_country Yugoslavia                 0.559388
## native_countrycharacter                   0.102472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28759  on 26048  degrees of freedom
## Residual deviance: 16394  on 25950  degrees of freedom
## AIC: 16592
##
## Number of Fisher Scoring iterations: 14
```

b. Which variable(s) is/are not statistically significant? Explain why?

all variables that the p-value is greater than 0.05 is npt statistically significant

```
summary(glm_model)$coeff[-1,4] > 0.05
```

```
##                                        age
##                                      FALSE
##                        workclass Local-gov
##                                      FALSE
##                     workclass Never-worked
##                                       TRUE
##                          workclass Private
##                                      FALSE
##                     workclass Self-emp-inc
##                                       TRUE
##                 workclass Self-emp-not-inc
##                                      FALSE
##                        workclass State-gov
##                                      FALSE
##                      workclass Without-pay
##                                       TRUE
##                         workclasscharacter
##                                      FALSE
##                                     fnlwgt
##                                      FALSE
##                             education 11th
##                                       TRUE
##                             education 12th
##                                       TRUE
##                          education 1st-4th
##                                       TRUE
##                         education 5th-6th
##                                      TRUE
##                         education 7th-8th
##                                     FALSE
##                             education 9th
##                                      TRUE
##                      education Assoc-acdm
##                                     FALSE
##                       education Assoc-voc
##                                     FALSE
##                       education Bachelors
##                                     FALSE
##                       education Doctorate
##                                     FALSE
##                         education HS-grad
##                                     FALSE
##                         education Masters
##                                     FALSE
##                       education Preschool
##                                      TRUE
##                     education Prof-school
##                                     FALSE
##                    education Some-college
##                                     FALSE
##          marital_status Married-AF-spouse
##                                     FALSE
##         marital_status Married-civ-spouse
##                                     FALSE
##      marital_status Married-spouse-absent
##                                      TRUE
##              marital_status Never-married
##                                     FALSE
##                  marital_status Separated
##                                      TRUE
##                    marital_status Widowed
##                                      TRUE
##                   occupation Armed-Forces
##                                      TRUE
##                   occupation Craft-repair
##                                      TRUE
##                occupation Exec-managerial
##                                     FALSE
##                occupation Farming-fishing
##                                     FALSE
##              occupation Handlers-cleaners
##                                     FALSE
##              occupation Machine-op-inspct
##                                     FALSE
##                  occupation Other-service
##                                     FALSE
##                occupation Priv-house-serv
##                                     FALSE
##                 occupation Prof-specialty
##                                     FALSE
##                occupation Protective-serv
##                                      FALSE
##                          occupation Sales
##                                      FALSE
##                   occupation Tech-support
##                                      FALSE
##               occupation Transport-moving
##                                       TRUE
##                relationship Not-in-family
##                                       TRUE
##               relationship Other-relative
##                                       TRUE
##                    relationship Own-child
##                                      FALSE
##                    relationship Unmarried
##                                       TRUE
##                         relationship Wife
##                                      FALSE
##                   race Asian-Pac-Islander
##                                      FALSE
##                                race Black
##                                       TRUE
##                                race Other
##                                       TRUE
##                                race White
##                                       TRUE
##                                  sex Male
##                                      FALSE
##                               captal_gain
##                                      FALSE
##                              capital_loss
##                                      FALSE
##                            hours_per_week
##                                      FALSE
##                     native_country Canada
##                                       TRUE
##                      native_country China
##                                      FALSE
##                   native_country Columbia
##                                       TRUE
##                       native_country Cuba
##                                       TRUE
##         native_country Dominican-Republic
##                                      FALSE
##                    native_country Ecuador
##                                       TRUE
##                native_country El-Salvador
##                                       TRUE
##                    native_country England
##                                       TRUE
##                     native_country France
##                                       TRUE
##                    native_country Germany
##                                       TRUE
##                     native_country Greece
##                                      TRUE
##                  native_country Guatemala
##                                      TRUE
##                      native_country Haiti
##                                      TRUE
##         native_country Holand-Netherlands
##                                      TRUE
##                   native_country Honduras
##                                      TRUE
##                       native_country Hong
##                                      TRUE
##                    native_country Hungary
##                                      TRUE
##                      native_country India
##                                     FALSE
##                       native_country Iran
##                                      TRUE
##                    native_country Ireland
##                                      TRUE
##                      native_country Italy
##                                      TRUE
##                    native_country Jamaica
##                                      TRUE
##                      native_country Japan
##                                      TRUE
##                       native_country Laos
##                                      TRUE
##                     native_country Mexico
##                                     FALSE
##                  native_country Nicaragua
##                                      TRUE
## native_country Outlying-US(Guam-USVI-etc)
##                                      TRUE
##                       native_country Peru
##                                      TRUE
##                native_country Philippines
##                                      TRUE
##                     native_country Poland
##                                      TRUE
##                   native_country Portugal
##                                      TRUE
##                native_country Puerto-Rico
##                                      TRUE
##                   native_country Scotland
##                                      TRUE
##                      native_country South
##                                     FALSE
##                     native_country Taiwan
##                                      TRUE
##                   native_country Thailand
##                                      TRUE
##            native_country Trinadad&Tobago
##                                      TRUE
##              native_country United-States
##                                      TRUE
##                    native_country Vietnam
##                                     FALSE
##                 native_country Yugoslavia
##                                      TRUE
##                   native_countrycharacter
##                                      TRUE
```

c. What is the accuracy of the model on the test set?

```
cm <- confusionMatrix(factor(predit_data_glm>0.5, levels = c(T,F), labels = c("1", "0")), as.factor(tes
overall.accuracy <- cm$overall['Accuracy']
cat("accuracy on testing data : ", overall.accuracy, "\n")
```

```
## accuracy on testing data :  0.8453624
```