Project2_classification

hyeoncheol kim

2020 2 27

#Step1 Data exploration

1.Reading adult.csv file and used 6 R functions to explore data. adult.csv file has total 15 attributes. [age,sex,education.num,occupation,relationship,race,sex,capital.gain, captial.loss, hours.per.week, native.country,workclass,fnlwgt,education,income]

2.By using str(), I found some '?' data in workclass, occupation and native country columns which are needed to be cleaned.

```r
df<-read.csv("C://Users/compa/Desktop/2020 spring/CS 4375(ML)/Project2/adult.csv")
summary(df)
```

```
##       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     education.num                  marital.status
##  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         capital.gain    capital.loss     hours.per.week
##  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
##
##        native.country    income
##  United-States:29170   <=50K:24720
##  Mexico       :  643   >50K : 7841
##  ?            :  583
##  Philippines  :  198
##  Germany      :  137
##  Canada       :  121
##  (Other)      : 1709
```

```r
head(df)
```

```
##   age workclass fnlwgt    education education.num marital.status
## 1  90         ?  77053      HS-grad             9        Widowed
## 2  82   Private 132870      HS-grad             9        Widowed
## 3  66         ? 186061 Some-college            10        Widowed
## 4  54   Private 140359      7th-8th             4       Divorced
## 5  41   Private 264663 Some-college            10      Separated
## 6  34   Private 216864      HS-grad             9       Divorced
##          occupation  relationship  race    sex capital.gain capital.loss
## 1                 ? Not-in-family White Female            0         4356
## 2   Exec-managerial Not-in-family White Female            0         4356
## 3                 ?     Unmarried Black Female            0         4356
## 4 Machine-op-inspct     Unmarried White Female            0         3900
## 5    Prof-specialty     Own-child White Female            0         3900
## 6     Other-service     Unmarried White Female            0         3770
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             18  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States  <=50K
```

```r
tail(df)
```

```
##       age workclass fnlwgt    education education.num     marital.status
## 32556  53   Private 321865      Masters            14 Married-civ-spouse
## 32557  22   Private 310152 Some-college            10      Never-married
## 32558  27   Private 257302   Assoc-acdm            12 Married-civ-spouse
## 32559  40   Private 154374      HS-grad             9 Married-civ-spouse
## 32560  58   Private 151910      HS-grad             9            Widowed
## 32561  22   Private 201490      HS-grad             9      Never-married
##              occupation  relationship  race    sex capital.gain capital.loss
## 32556   Exec-managerial       Husband White   Male            0            0
## 32557   Protective-serv Not-in-family White   Male            0            0
## 32558      Tech-support          Wife White Female            0            0
## 32559 Machine-op-inspct       Husband White   Male            0            0
## 32560      Adm-clerical     Unmarried White Female            0            0
## 32561      Adm-clerical     Own-child White   Male            0            0
##       hours.per.week native.country income
## 32556             40  United-States   >50K
```

```
## 32557              40  United-States  <=50K
## 32558              38  United-States  <=50K
## 32559              40  United-States   >50K
## 32560              40  United-States  <=50K
## 32561              20  United-States  <=50K
```

```r
names(df)
```

```
##  [1] "age"            "workclass"      "fnlwgt"         "education"
##  [5] "education.num"  "marital.status" "occupation"     "relationship"
##  [9] "race"           "sex"            "capital.gain"   "capital.loss"
## [13] "hours.per.week" "native.country" "income"
```

```r
nrow(df)
```

```
## [1] 32561
```

```r
str(df)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 1 6 5 1 5 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 1 5 1 8 11 9 2 11 11 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 5 5 3 2 5 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
##  $ hours.per.week: int  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 1 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

#Step2 plotting

I created box plots of int type columns to see if they have outliers.

```r
par(mfrow =c(3,3))
boxplot(df$age, main = "Age")
boxplot(df$fnlwgt, main = "fnlwgt")
boxplot(df$education.num,main = "education.num")
boxplot(df$hours.per.week,main = "hours.per.week")
library(ggplot2)
ggplot(df, aes(age)) + geom_histogram(aes(fill = income), color = "black",binwidth = 1)
#the above histogram people around the age between 25-50 tend to earn more than 50k
```

#step3 Data cleaning

1. For better prediction, I used boxplot()out to find some extreme values in each column and removed these from the dataset.

2. After changing "?" value to NA, I used sapply function to see how many missing values in the dataset. Also I changed each NAs to most occurance value in each column.

3. Finally I changed independent values to numeric values and dependent values to factor.

```
#1
Outlier1 <- boxplot(df$age, plot = FALSE)$out #removing outliers for better model
df <- df[-which(df$age %in% Outlier1 ),]
Outlier2 <- boxplot(df$fnlwgt, plot = FALSE)$out
df <- df[-which(df$fnlwgt %in% Outlier2 ),]
Outlier3 <- boxplot(df$education.num, plot = FALSE)$out
df <- df[-which(df$education.num %in% Outlier3 ),]
Outlier4 <- boxplot(df$hours.per.week, plot = FALSE)$out
df <- df[-which(df$hours.per.week %in% Outlier4 ),]

#2
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
```

```r
df[df  == "?"] <- NA   #change '?' to NA

sapply(df,function(x) sum(is.na(x))) # To see NAs in each columns
```

```
##            age      workclass         fnlwgt      education  education.num
##              0            836              0              0              0
## marital.status     occupation   relationship           race            sex
##              0            838              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0            381              0
```

```r
df$occupation[is.na(df$occupation)] <- "Prof-specialty" #change Nas to prof-specialty
df$native.country[is.na(df$native.country)]<- "United-States" # change NAs to USA
df<-select(df,-c(workclass)) #workclass has too many NA values, so I removed it from dataframe.
#3
summary(df)
```

```
##       age            fnlwgt              education    education.num
##  Min.   :17.00   Min.   : 14878   HS-grad     :7758   Min.   : 5.00
##  1st Qu.:29.00   1st Qu.:117609   Some-college:4798   1st Qu.: 9.00
##  Median :37.00   Median :176711   Bachelors   :3923   Median :10.00
##  Mean   :38.65   Mean   :180635   Masters     :1227   Mean   :10.37
##  3rd Qu.:47.00   3rd Qu.:228612   Assoc-voc   :1065   3rd Qu.:13.00
##  Max.   :78.00   Max.   :416415   Assoc-acdm  : 756   Max.   :16.00
##                                   (Other)     :2431
##                marital.status           occupation           relationship
##  Divorced             : 3308   Prof-specialty :3795   Husband       :9354
##  Married-AF-spouse    :   13   Craft-repair   :3214   Not-in-family :5788
##  Married-civ-spouse   :10538   Exec-managerial:2959   Other-relative: 595
##  Married-spouse-absent:  259   Adm-clerical   :2860   Own-child     :2712
##  Never-married        : 6617   Sales          :2254   Unmarried     :2466
##  Separated            :  716   Other-service  :1711   Wife          :1043
##  Widowed              :  507   (Other)        :5165
##                  race            sex          capital.gain    capital.loss
##  Amer-Indian-Eskimo:  217   Female: 6961   Min.   :    0   Min.   :   0.00
##  Asian-Pac-Islander:  721   Male  :14997   1st Qu.:    0   1st Qu.:   0.00
##  Black             : 2170                  Median :    0   Median :   0.00
##  Other             :  168                  Mean   : 1024   Mean   :  90.11
##  White             :18682                  3rd Qu.:    0   3rd Qu.:   0.00
##                                            Max.   :99999   Max.   :4356.00
##
##  hours.per.week        native.country    income
##  Min.   :33.0   United-States:20330   <=50K:16170
##  1st Qu.:40.0   Mexico       :  238   >50K : 5788
```

```
##  Median :40.0   Philippines  :  150  
##  Mean   :41.6   Germany      :   98  
##  3rd Qu.:43.0   Canada       :   73  
##  Max.   :52.0   India        :   72  
##                 (Other)      :  997
```

```
df$native.country <- as.factor(ifelse(df$native.country=="United-States", "US","alien")) # Changed coun

df$hours.per.week <- as.factor(ifelse (df$hours.per.week>=40, ">=40", "<40"))
# changed hours level either >=40hours or <40 hours
```

#Step4 Logistc regression

After creating glm model, it turns out that the target is depending on all the features so, I used every feature for creating model.

1.Total number of train data is 16468 and test data is 5490. And my accuracy of the predcition with logistic regression came out as 84%. There waere some data that were not related to income when I made model with every model. They had high p-values and those data were excluded from the model for the best prediction.

2. Sensitivity is 0.9249 meaning that the model could find 92.49% of all predicted incomes that are >50K.
3. Specificity is 0.6102 meaning that the model could find 61.02% of all predicted incomes that are <=50K.
4. PPV is 0.8707, which means that out of all >50K income predictions 87.07% were true.
5. NPV is 0.7411, which means that out of all <=50K income predictions 74.11% were true.

```
set.seed(1234)
i <- sample(1:nrow(df), 0.75*nrow(df), replace =FALSE) #take sample of train and test dataset
train <- df[i,]
test<- df[-i, ]
glm1 <- glm(income ~ ., data = train, family=binomial)
```

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

```
summary(glm1)
```

```
## 
## Call:
## glm(formula = income ~ ., family = binomial, data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -5.3463  -0.5558  -0.2094   0.0796   3.3110  
## 
## Coefficients: (1 not defined because of singularities)
##                                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                       -7.870e+00  5.600e-01 -14.054  < 2e-16 ***
## age                                2.999e-02  2.269e-03  13.218  < 2e-16 ***
## fnlwgt                             9.731e-07  2.732e-07   3.562 0.000368 ***
## education11th                      6.730e-02  2.708e-01   0.249 0.803725    
## education12th                      4.811e-01  3.347e-01   1.437 0.150665    
## education9th                      -3.870e-01  3.363e-01  -1.151 0.249869
```

```
## educationAssoc-acdm                  1.599e+00  2.279e-01   7.018 2.26e-12 ***
## educationAssoc-voc                   1.374e+00  2.188e-01   6.281 3.37e-10 ***
## educationBachelors                   2.081e+00  2.034e-01  10.226  < 2e-16 ***
## educationDoctorate                   3.197e+00  2.998e-01  10.662  < 2e-16 ***
## educationHS-grad                     8.400e-01  1.975e-01   4.252 2.12e-05 ***
## educationMasters                     2.399e+00  2.172e-01  11.047  < 2e-16 ***
## educationProf-school                 3.029e+00  2.776e-01  10.911  < 2e-16 ***
## educationSome-college                1.232e+00  2.008e-01   6.135 8.52e-10 ***
## education.num                               NA         NA      NA       NA
## marital.statusMarried-AF-spouse      3.775e+00  7.942e-01   4.753 2.01e-06 ***
## marital.statusMarried-civ-spouse     2.217e+00  3.447e-01   6.433 1.25e-10 ***
## marital.statusMarried-spouse-absent -3.402e-01  3.276e-01  -1.038 0.299056
## marital.statusNever-married         -5.597e-01  1.166e-01  -4.799 1.60e-06 ***
## marital.statusSeparated             -5.476e-02  2.130e-01  -0.257 0.797115
## marital.statusWidowed                8.774e-02  2.112e-01   0.415 0.677906
## occupationArmed-Forces              -4.742e-01  1.741e+00  -0.272 0.785380
## occupationCraft-repair               6.249e-02  1.009e-01   0.619 0.535680
## occupationExec-managerial            8.016e-01  9.840e-02   8.147 3.73e-16 ***
## occupationFarming-fishing           -1.115e+00  2.174e-01  -5.128 2.93e-07 ***
## occupationHandlers-cleaners         -4.211e-01  1.752e-01  -2.403 0.016250 *
## occupationMachine-op-inspct         -2.893e-01  1.286e-01  -2.250 0.024471 *
## occupationOther-service             -7.680e-01  1.549e-01  -4.958 7.13e-07 ***
## occupationPriv-house-serv           -1.158e+01  1.377e+02  -0.084 0.932967
## occupationProf-specialty             4.050e-01  9.961e-02   4.066 4.78e-05 ***
## occupationProtective-serv            4.469e-01  1.597e-01   2.797 0.005150 **
## occupationSales                      3.563e-01  1.059e-01   3.364 0.000768 ***
## occupationTech-support               6.788e-01  1.428e-01   4.753 2.01e-06 ***
## occupationTransport-moving          -1.756e-01  1.356e-01  -1.295 0.195411
## relationshipNot-in-family            5.387e-01  3.421e-01   1.575 0.115299
## relationshipOther-relative          -4.688e-01  3.133e-01  -1.496 0.134528
## relationshipOwn-child               -5.638e-01  3.370e-01  -1.673 0.094324 .
## relationshipUnmarried                3.359e-01  3.626e-01   0.926 0.354214
## relationshipWife                     1.204e+00  1.381e-01   8.718  < 2e-16 ***
## raceAsian-Pac-Islander               8.422e-01  3.482e-01   2.419 0.015558 *
## raceBlack                            6.951e-01  3.290e-01   2.113 0.034629 *
## raceOther                           -6.408e-01  6.136e-01  -1.044 0.296335
## raceWhite                            9.103e-01  3.161e-01   2.879 0.003985 **
## sexMale                              8.544e-01  1.052e-01   8.124 4.52e-16 ***
## capital.gain                         3.424e-04  1.453e-05  23.559  < 2e-16 ***
## capital.loss                         5.882e-04  5.138e-05  11.449  < 2e-16 ***
## hours.per.week>=40                   3.287e-01  9.762e-02   3.368 0.000759 ***
## native.countryUS                     2.494e-01  1.064e-01   2.344 0.019055 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19012  on 16467  degrees of freedom
## Residual deviance: 11386  on 16421  degrees of freedom
## AIC: 11480
##
## Number of Fisher Scoring iterations: 13
```

```r
probs <- predict(glm1, newdata=test, type="response")
```

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

```r
pred <- ifelse(probs>0.5, ">50K", "<=50K")
acc <- mean(pred==test$income)
print(paste("accuracy = ", acc))
```

```
## [1] "accuracy =  0.836794171220401"
```

```r
table(pred, test$income)
```

```
##
## pred    <=50K >50K
##   <=50K  3714  560
##   >50K    336  880
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
confusionMatrix(as.factor(pred), test$income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  3714  560
##      >50K    336  880
##
##                Accuracy : 0.8368
##                  95% CI : (0.8267, 0.8465)
##     No Information Rate : 0.7377
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.556
##
##  Mcnemar's Test P-Value : 9.341e-14
##
##             Sensitivity : 0.9170
##             Specificity : 0.6111
##          Pos Pred Value : 0.8690
##          Neg Pred Value : 0.7237
##              Prevalence : 0.7377
##          Detection Rate : 0.6765
##    Detection Prevalence : 0.7785
##       Balanced Accuracy : 0.7641
##
##        'Positive' Class : <=50K
##
```

#step5 Naive bayes

1.By using Naive bayes algorithm, I could get 81% of accuracy. 2. Sensitivity is 0.9339 meaning that the model could find 93.39% of all predicted incomes that are >50K. 3. Specificity is 0.4752 meaning that the model could find 47.52% of all predicted incomes that are <=50K.

4. PPV is 0.8348, which means that out of all >50K income predictions 83.48% were true.
5. NPV is 0.7168, which means that out of all <=50K income predictions 71.68% were true.

```
library(e1071)
nb1<- naiveBayes(income~ ., data =  train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     <=50K      >50K
## 0.7359728 0.2640272
##
## Conditional probabilities:
##        age
## Y           [,1]     [,2]
##   <=50K 36.82756 12.25825
##   >50K  44.07521 10.10518
##
##        fnlwgt
## Y           [,1]     [,2]
##   <=50K 180548.4 87620.98
##   >50K  182036.5 85975.60
##
##        education
## Y              10th        11th        12th     1st-4th     5th-6th     7th-8th
##   <=50K 0.032508251 0.037623762 0.014768977 0.000000000 0.000000000 0.000000000
##   >50K  0.008509660 0.008279669 0.004369825 0.000000000 0.000000000 0.000000000
##        education
## Y               9th  Assoc-acdm   Assoc-voc   Bachelors   Doctorate     HS-grad
##   <=50K 0.021369637 0.033168317 0.046699670 0.139933993 0.003300330 0.400165017
##   >50K  0.003909844 0.037258510 0.049448022 0.289098436 0.033118675 0.224471021
##        education
## Y           Masters   Preschool Prof-school Some-college
##   <=50K 0.033003300 0.000000000 0.005198020  0.232260726
##   >50K  0.119825207 0.000000000 0.044158234  0.177552898
##
##        education.num
## Y            [,1]     [,2]
##   <=50K  9.931188 2.015793
##   >50K  11.578427 2.252533
##
##        marital.status
```

```
## Y          Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent
##   <=50K 0.185231023       0.000330033        0.342244224           0.014438944
##   >50K  0.061407544       0.001609936        0.855335787           0.003449862
##        marital.status
## Y       Never-married   Separated     Widowed
##   <=50K   0.389108911 0.040759076 0.027887789
##   >50K    0.059797608 0.008509660 0.009889604
##
##        occupation
## Y                   ? Adm-clerical Armed-Forces Craft-repair Exec-managerial
##   <=50K 0.0000000000 0.1514851485 0.0004125413 0.1506600660    0.0984323432
##   >50K  0.0000000000 0.0717571297 0.0002299908 0.1313247470    0.2428702852
##        occupation
## Y       Farming-fishing Handlers-cleaners Machine-op-inspct Other-service
##   <=50K    0.0245049505      0.0501650165      0.0824257426  0.1007425743
##   >50K     0.0096596136      0.0137994480      0.0365685373  0.0172493100
##        occupation
## Y       Priv-house-serv Prof-specialty Protective-serv        Sales
##   <=50K    0.0024752475   0.1430693069    0.0193069307 0.0959570957
##   >50K     0.0000000000   0.2564397424    0.0264489420 0.1200551978
##        occupation
## Y       Tech-support Transport-moving
##   <=50K 0.0290429043     0.0513201320
##   >50K  0.0402483901     0.0333486661
##
##        relationship
## Y           Husband Not-in-family Other-relative   Own-child   Unmarried
##   <=50K 0.304125413   0.320957096    0.034900990 0.164191419 0.144719472
##   >50K  0.760579577   0.105795768    0.004829807 0.008969641 0.029208832
##        relationship
## Y              Wife
##   <=50K 0.031105611
##   >50K  0.090616375
##
##        race
## Y       Amer-Indian-Eskimo Asian-Pac-Islander       Black       Other
##   <=50K        0.012376238        0.031600660 0.116831683 0.008745875
##   >50K         0.004139834        0.036108556 0.052207912 0.001379945
##        race
## Y             White
##   <=50K 0.830445545
##   >50K  0.906163753
##
##        sex
## Y          Female      Male
##   <=50K 0.3790429 0.6209571
##   >50K  0.1481141 0.8518859
##
##        capital.gain
## Y            [,1]       [,2]
##   <=50K  147.7374   862.6869
##   >50K  3612.3243 13528.7685
##
##        capital.loss
```

```
## Y              [,1]      [,2]
##   <=50K  55.16386 311.5989
##   >50K  176.29899 564.1271
##
##        hours.per.week
## Y               <40        >=40
##   <=50K 0.10693069 0.89306931
##   >50K  0.06094756 0.93905244
##
##        native.country
## Y             alien         US
##   <=50K 0.07830033 0.92169967
##   >50K  0.06347746 0.93652254
```

```r
pn <- predict(nb1, newdata=test, type="class")
table(pn, test$income)
```

```
##
## pn      <=50K >50K
##   <=50K  3765  755
##   >50K    285  685
```

```r
confusionMatrix(pn, test$income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  3765  755
##      >50K    285  685
##
##                Accuracy : 0.8106
##                  95% CI : (0.7999, 0.8209)
##     No Information Rate : 0.7377
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.453
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9296
##             Specificity : 0.4757
##          Pos Pred Value : 0.8330
##          Neg Pred Value : 0.7062
##              Prevalence : 0.7377
##          Detection Rate : 0.6858
##    Detection Prevalence : 0.8233
##       Balanced Accuracy : 0.7027
##
##        'Positive' Class : <=50K
##
```

#step6 Decision Tree

1.By using decision tree algorithm, I could get 83.65% of accuracy.

2. Sensitivity is 0.9454 meaning that the model could find 94.54% of all predicted incomes that are >50K.
3. Specificity is 0.5274 meaning that the model could find 52.74% of all predicted incomes that are <=50K.
4. PPV is 0.8503, which means that out of all >50K income predictions 85.03% were true.
5. NPV is 0.7727, which means that out of all <=50K income predictions 77.27% were true.

```r
library(caret)
library(rpart)
#install.packages("tree")
library(tree)

str(df)
```

```
## 'data.frame':    21958 obs. of  14 variables:
##  $ age           : int  66 41 34 38 45 38 51 46 57 22 ...
##  $ fnlwgt        : int  186061 264663 216864 150601 172274 164526 172175 45363 317847 119592 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 16 16 12 1 11 15 11 15 13 8 ...
##  $ education.num : int  10 10 9 6 16 15 16 15 14 12 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 6 1 6 1 5 5 1 1 5 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 11 11 9 2 11 11 11 11 5 7 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 5 4 5 5 5 2 2 2 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 3 5 5 5 3 5 5 5 5 3 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 2 1 2 2 2 2 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3770 3770 3004 2824 2824 2824 2824 2824 ...
##  $ hours.per.week: Factor w/ 2 levels "<40",">=40": 2 2 2 2 1 2 2 2 2 2 ...
##  $ native.country: Factor w/ 2 levels "alien","US": 2 2 2 2 2 2 2 2 2 2 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 2 2 2 2 2 2 ...
```

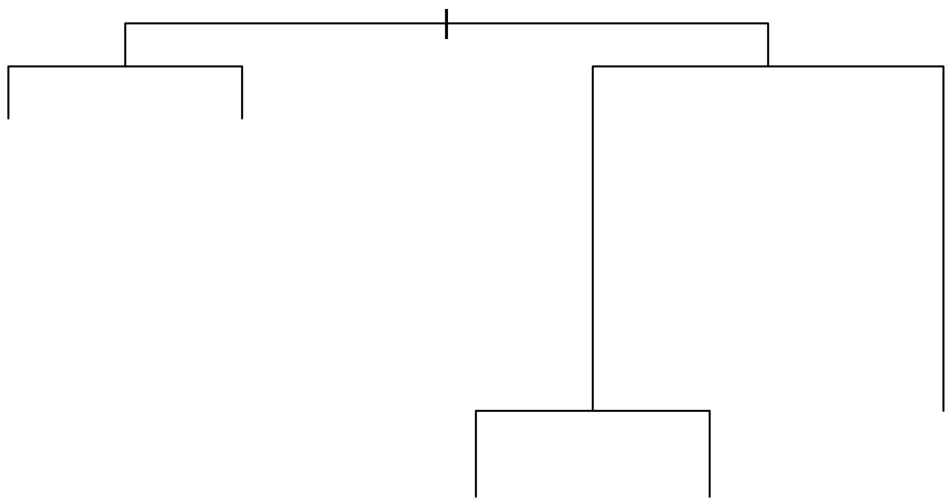
```r
tree1 <- rpart(income~ ., data = train)
summary(tree1)
```

```
## Call:
## rpart(formula = income ~ ., data = train)
##   n= 16468
##
##           CP nsplit rel error    xerror       xstd
## 1 0.13005980      0 1.0000000 1.0000000 0.01301026
## 2 0.06485741      2 0.7398804 0.7398804 0.01170146
## 3 0.03932843      3 0.6750230 0.6750230 0.01129513
## 4 0.01000000      4 0.6356946 0.6356946 0.01103018
##
## Variable importance
##   relationship marital.status   capital.gain            sex      education
##             25             25             12              9              9
##  education.num     occupation            age
##              9              6              5
##
## Node number 1: 16468 observations,    complexity param=0.1300598
```

```
##   predicted class=<=50K  expected loss=0.2640272  P(node) =1
##     class counts: 12120  4348
##    probabilities: 0.736 0.264
##   left son=2 (8704 obs) right son=3 (7764 obs)
##   Primary splits:
##       relationship   splits as  RLLLLR, improve=1328.6460, (0 missing)
##       marital.status splits as  LRRLLLL, improve=1318.6150, (0 missing)
##       capital.gain   < 5095.5  to the left,  improve= 841.7265, (0 missing)
##       education      splits as  LLL---LLLRRLR-RL, improve= 597.5427, (0 missing)
##       education.num  < 12.5    to the left,  improve= 597.5427, (0 missing)
##   Surrogate splits:
##       marital.status splits as  LRRLLLL, agree=0.993, adj=0.985, (0 split)
##       sex            splits as  LR, agree=0.696, adj=0.355, (0 split)
##       age            < 31.5    to the left,  agree=0.627, adj=0.209, (0 split)
##       occupation     splits as  -LLRRRLLLLLRLLR, agree=0.600, adj=0.151, (0 split)
##       capital.gain   < 2616    to the left,  agree=0.561, adj=0.069, (0 split)
##
## Node number 2: 8704 observations,    complexity param=0.03932843
##   predicted class=<=50K  expected loss=0.07433364  P(node) =0.5285402
##     class counts:  8057   647
##    probabilities: 0.926 0.074
##   left son=4 (8525 obs) right son=5 (179 obs)
##   Primary splits:
##       capital.gain  < 7073.5  to the left,  improve=298.25710, (0 missing)
##       education     splits as  LLL---LLLRRLR-RL, improve= 71.15785, (0 missing)
##       education.num < 12.5    to the left,  improve= 71.15785, (0 missing)
##       occupation    splits as  -LLLRLLLLLRRRRL, improve= 44.75386, (0 missing)
##       age           < 34.5    to the left,  improve= 37.40920, (0 missing)
##
## Node number 3: 7764 observations,    complexity param=0.1300598
##   predicted class=<=50K  expected loss=0.4766873  P(node) =0.4714598
##     class counts:  4063  3701
##    probabilities: 0.523 0.477
##   left son=6 (5427 obs) right son=7 (2337 obs)
##   Primary splits:
##       education     splits as  LLL---LLLRRLR-RL, improve=470.6033, (0 missing)
##       education.num < 12.5    to the left,  improve=470.6033, (0 missing)
##       occupation    splits as  -LRLRLLLLLRRRRL, improve=409.8580, (0 missing)
##       capital.gain  < 5095.5  to the left,  improve=364.2266, (0 missing)
##       age           < 35.5    to the left,  improve=163.5620, (0 missing)
##   Surrogate splits:
##       education.num < 12.5    to the left,  agree=1.000, adj=1.000, (0 split)
##       occupation    splits as  -LLLRLLLLLRLLLL, agree=0.778, adj=0.264, (0 split)
##       capital.gain  < 10585.5 to the left,  agree=0.714, adj=0.050, (0 split)
##       race          splits as  LRLLL, agree=0.705, adj=0.018, (0 split)
##       capital.loss  < 1894.5  to the left,  agree=0.703, adj=0.013, (0 split)
##
## Node number 4: 8525 observations
##   predicted class=<=50K  expected loss=0.05536657  P(node) =0.5176706
##     class counts:  8053   472
##    probabilities: 0.945 0.055
##
## Node number 5: 179 observations
##   predicted class=>50K   expected loss=0.02234637  P(node) =0.01086957
```

```
##     class counts:     4   175
##    probabilities: 0.022 0.978 
## 
## Node number 6: 5427 observations,    complexity param=0.06485741
##   predicted class=<=50K  expected loss=0.362447  P(node) =0.3295482
##     class counts:  3460  1967
##    probabilities: 0.638 0.362 
##   left son=12 (5137 obs) right son=13 (290 obs)
##   Primary splits:
##       capital.gain  < 5095.5  to the left,  improve=238.4038, (0 missing)
##       occupation    splits as  -RLLRLLLLLRRRRL, improve=102.0640, (0 missing)
##       age           < 35.5    to the left,  improve= 99.1668, (0 missing)
##       education     splits as  LLL---LRR--L---R, improve= 76.3936, (0 missing)
##       education.num < 9.5     to the left,  improve= 76.3936, (0 missing)
## 
## Node number 7: 2337 observations
##   predicted class=>50K   expected loss=0.2580231  P(node) =0.1419116
##     class counts:   603  1734
##    probabilities: 0.258 0.742 
## 
## Node number 12: 5137 observations
##   predicted class=<=50K  expected loss=0.3272338  P(node) =0.3119383
##     class counts:  3456  1681
##    probabilities: 0.673 0.327 
## 
## Node number 13: 290 observations
##   predicted class=>50K   expected loss=0.0137931  P(node) =0.01760991
##     class counts:     4   286
##    probabilities: 0.014 0.986
```

```r
plot(tree1)
```



```r
pred_dt<- predict(tree1, newdata = test, type = "class")
table(pred_dt, test$income)
```

```
##
## pred_dt <=50K >50K
##   <=50K  3849  700
##   >50K    201  740
```

```r
mean(pred_dt == test$income)
```

```
## [1] 0.8358834
```

```r
confusionMatrix(pred_dt, test$income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  3849  700
##      >50K    201  740
##
##                Accuracy : 0.8359
##                  95% CI : (0.8258, 0.8456)
##     No Information Rate : 0.7377
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5226
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9504
##             Specificity : 0.5139
##          Pos Pred Value : 0.8461
##          Neg Pred Value : 0.7864
##              Prevalence : 0.7377
##          Detection Rate : 0.7011
##    Detection Prevalence : 0.8286
##       Balanced Accuracy : 0.7321
##
##        'Positive' Class : <=50K
##
```

#Step7 Report

Based on the result of predicting Adult census income whether each individual make over 50k or not by given dataset with three algorithm(logistic regression, naive bayes and decision tree), Decision tree gives the best accuracy, logistic regression is the second and naive bayes.

And I assume that, since there are some of the features are dependent on each other such as education and education.num and also it is large space of dataset, these causes might leads poor prediction of naive bayes.