Logistic Regression Project

In this project I will be working with the UCI adult dataset. I will be attempting to predict if people in the data set belong in a certain class by salary, either making <=50k or >50k per year.

Importing the Data Set and Checking

```
adult <- read.csv("adult_sal.csv")
head(adult)
```

```
> head(adult)
  X age    type_employer fnlwgt education education_num           marital        occupation  relationship  race    sex capital_gain capital_loss
1 1  39        State-gov  77516 Bachelors           13      Never-married      Adm-clerical Not-in-family White   Male         2174            0
2 2  50 Self-emp-not-inc  83311 Bachelors           13 Married-civ-spouse   Exec-managerial       Husband White   Male            0            0
3 3  38          Private 215646   HS-grad            9          Divorced Handlers-cleaners Not-in-family White   Male            0            0
4 4  53          Private 234721      11th            7 Married-civ-spouse Handlers-cleaners       Husband Black   Male            0            0
5 5  28          Private 338409 Bachelors           13 Married-civ-spouse    Prof-specialty          Wife Black Female            0            0
6 6  37          Private 284582   Masters           14 Married-civ-spouse   Exec-managerial          Wife White Female            0            0
  hr_per_week       country income
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
```

Importing dplyr and Dropping the Repeated Index

```
library(dplyr)
adult <- select(adult,-X)
```

```
> library(dplyr)
> adult <- select(adult,-X)
> head(adult)
  age    type_employer fnlwgt education education_num           marital        occupation  relationship  race    sex capital_gain capital_loss
1  39        State-gov  77516 Bachelors           13      Never-married      Adm-clerical Not-in-family White   Male         2174            0
2  50 Self-emp-not-inc  83311 Bachelors           13 Married-civ-spouse   Exec-managerial       Husband White   Male            0            0
3  38          Private 215646   HS-grad            9          Divorced Handlers-cleaners Not-in-family White   Male            0            0
4  53          Private 234721      11th            7 Married-civ-spouse Handlers-cleaners       Husband Black   Male            0            0
5  28          Private 338409 Bachelors           13 Married-civ-spouse    Prof-specialty          Wife Black Female            0            0
6  37          Private 284582   Masters           14 Married-civ-spouse   Exec-managerial          Wife White Female            0            0
  hr_per_week       country income
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
```

Checking the Structure and Summary of the Dataset

```
str(adult)
Summary(adult)
```

```
> str(adult)
'data.frame':   32561 obs. of  15 variables:
 $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
 $ type_employer: chr  "State-gov" "Self-emp-not-inc" "Private" "Private" ...
 $ fnlwgt       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
 $ education    : chr  "Bachelors" "Bachelors" "HS-grad" "11th" ...
 $ education_num: int  13 13 9 7 13 14 5 9 14 13 ...
 $ marital      : chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
 $ occupation   : chr  "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...
 $ relationship : chr  "Not-in-family" "Husband" "Not-in-family" "Husband" ...
 $ race         : chr  "White" "White" "White" "Black" ...
 $ sex          : chr  "Male" "Male" "Male" "Male" ...
 $ capital_gain : int  2174 0 0 0 0 0 0 0 14084 5178 ...
 $ capital_loss : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hr_per_week  : int  40 13 40 40 40 40 16 45 50 40 ...
 $ country      : chr  "United-States" "United-States" "United-States" "United-States" ...
 $ income       : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
> summary(adult)
      age        type_employer          fnlwgt         education          education_num     marital           occupation
 Min.   :17.00   Length:32561       Min.   :  12285   Length:32561       Min.   : 1.00   Length:32561       Length:32561
 1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character   1st Qu.: 9.00   Class :character   Class :character
 Median :37.00   Mode  :character   Median : 178356   Mode  :character   Median :10.00   Mode  :character   Mode  :character
 Mean   :38.58                      Mean   : 189778                      Mean   :10.08
 3rd Qu.:48.00                      3rd Qu.: 237051                      3rd Qu.:12.00
 Max.   :90.00                      Max.   :1484705                      Max.   :16.00
 relationship          race               sex            capital_gain    capital_loss     hr_per_week      country            income
 Length:32561       Length:32561       Length:32561       Min.   :    0   Min.   :   0.0   Min.   : 1.00   Length:32561       Length:32561
 Class :character   Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :    0   Median :   0.0   Median :40.00   Mode  :character   Mode  :character
                                                          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
```

Data Cleaning

Inspecting the dataset, we can see that many of the variables are categorical which is good. However, many have to many factors and must be cleaned.

Employer column

```
table(adult$type-employer)
```

```
> table(adult$type_employer)

           ?      Federal-gov        Local-gov     Never-worked          Private     Self-emp-inc Self-emp-not-inc        State-gov
        1836              960             2093                7            22696             1116             2541             1298
 Without-pay
          14
```

Using a function to combine 'Never-worked' & 'Without-pay' into 'Unemployed'. Combining both self-employed inc/not inc values into 'Self-emp' and all government jobs into 'Government'

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

adult$type_employer <-sapply(adult$type_employer,unemployed)

table(adult$type_employer)
```

```
> unemployed <- function(job){
+   job <-as.character(job)
+   if (job=='Never-worked' | job=='Without-pay'){
+     return ('Unemployed')
+   }else if (job=='Self-emp-inc' | job=='Self-emp-not-inc'){
+     return('Self-emp')
+   }else if (job=='State-gov' | job =='Federal-gov' | job=='Local-gov'){
+     return('Government')
+   }else{
+     return(job)
+   }
+ }
> adult$type_employer <-sapply(adult$type_employer,unemployed)
> table(adult$type_employer)

          ? Government     Private    Self-emp Unemployed
       1836       4351      22696        3657         21
```

Marital Column

```
table(adult$marital)
```

```
> table(adult$marital)

          Divorced     Married-AF-spouse    Married-civ-spouse Married-spouse-absent         Never-married            Separated
              4443                   23                14976                   418                10683                 1025
           Widowed
               993
```

Want to combine values of 'separated', 'divorced' , and 'widowed' into 'Not-Married'

```
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$marital <-sapply(adult$marital,group_marital)

table(adult$marital)
```

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

      Married Never-married   Not-Married
        15417         10683          6461
```

Country Column

levels(adult$country)
table(adult$marital)

table(adult$country)
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 countries by region

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

adult$country <- sapply(adult$country,contry_grp)

table(adult$country)
```

```
> contry_grp <- function(cont){
+   if (cont %in% Asia){
+     return('Asia')
+   }else if (cont %in% North.America){
+     return('North.America')
+   }else if (cont %in% Europe){
+     return('Europe')
+   }else if (cont %in% Latin.and.South.America){
+     return('Latin.and.South.America')
+   }else{
+     return('Other')
+   }
+ }
> adult$country <- sapply(adult$country,contry_grp)
> table(adult$country)

                  Asia                 Europe Latin.and.South.America          North.America
                   671                    521                   1301                  29405
                 Other
                   663
```

Rename country to region.

```
names(adult)[names(adult)=="country"] <- "region"
```

Make sure any of the columns we changed have factor levels with factor()

```
adult$type_employer <- sapply(adult$type_employer,factor)
adult$country <- sapply(adult$country,factor)
adult$marital <- sapply(adult$marital,factor)
```

Missing Data

Turning values with "?" into Nan values

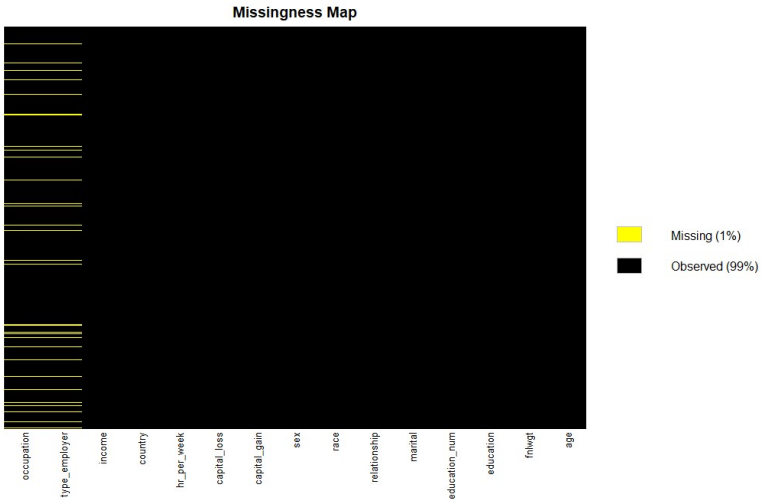
```
adult[adult == '?'] <- NA
```

```
> adult$type_employer <- sapply(adult$type_employer,factor)
> adult$country <- sapply(adult$country,factor)
> adult$marital <- sapply(adult$marital,factor)
> adult[adult == '?'] <- NA
> table(adult$type_employer)

Government   Self-emp     Private          ? Unemployed
      4351       3657       22696          0         21
```

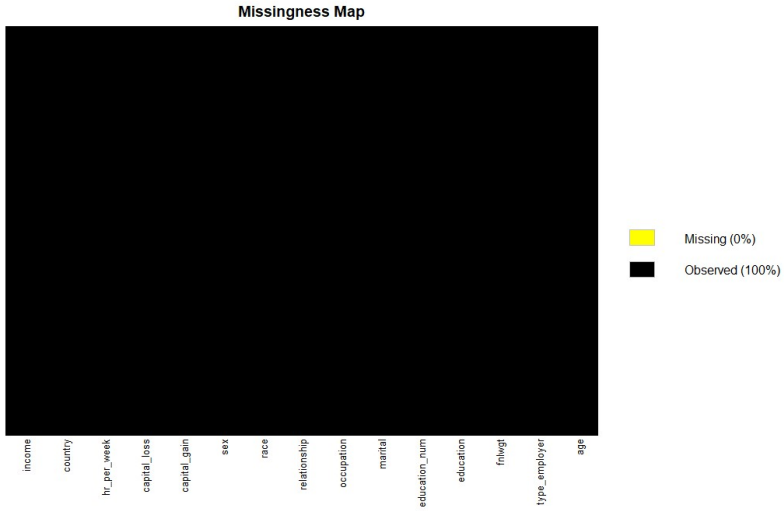
Using table() on a column with NA values should now not display those NA values. Instead, we see 0 for "?" need to refactor.

```
library(Amelia)
missmap(adult,y.at=c(1),y.labels = c(''),col=c('yellow','black'))
```



Omitting Nan values

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

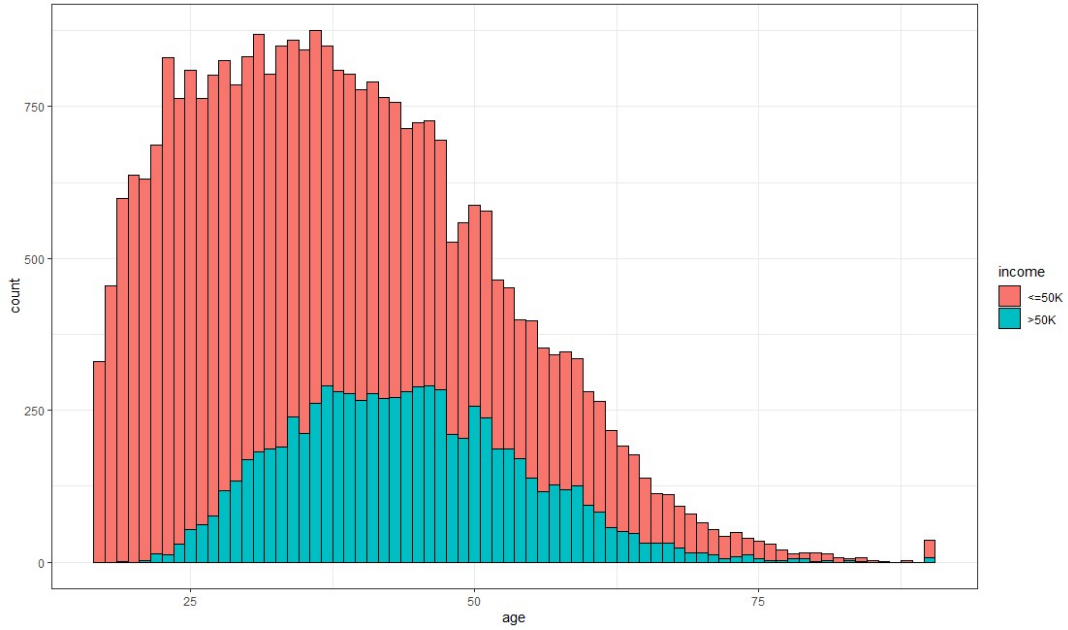


Looks good.

Exploratory Data Analysis

AGE

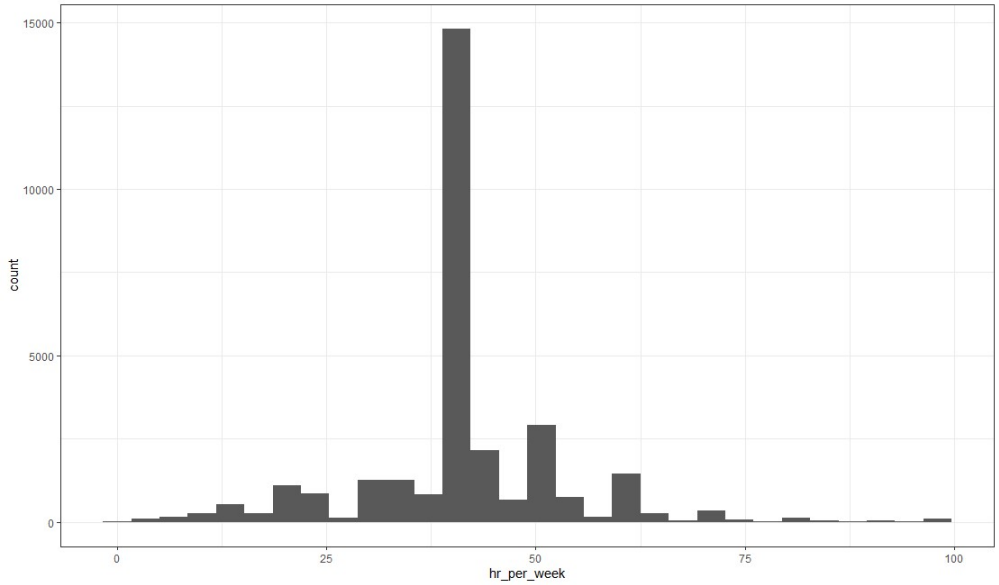
```
library(ggplot2)
ggplot(adult,aes(age)) +
geom_histogram(aes(fill=income),color='black',binwidth=1) + theme_bw()
```



From this chart we can see that through the distribution of ages we see that those making lower than 50k are mor prevalent. The age distribution tend to be most prevalent in those less than 60 years old.

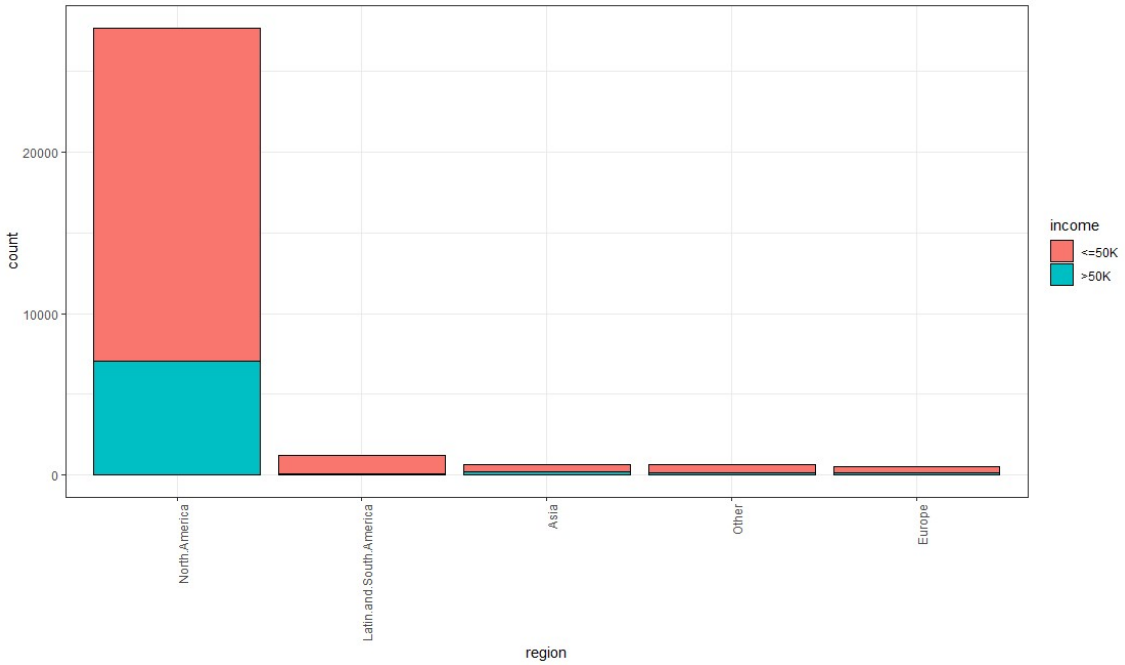
Work Hours per week

```
ggplot(adult,aes(hr_per_week)) + geom_histogram() + theme_bw()
```



Working 4 Working 40 hours per week is the most common in this data set.

Region



We can see through all the regions those who make less than 50k a more represented in the dataset. In addition, this dataset contains allot more samples of the local population.

Building the Logistic Regression Model

Import Library

```
library(caTools)
```

Randomly assigns a booleans to a new column "sample"

```
sample <- sample.split(adult$income, SplitRatio = 0.70) # SplitRatio = percent of sample==TRUE
```

Training Data

```
train = subset(adult, sample == TRUE)
```

Testing Data

```
test = subset(adult, sample == FALSE)

model = glm(income ~ ., family = binomial(logit), data = train)

summary(model)

new.step.model <- step(model)

summary(new.step.model)

test$predicted.income = predict(model, newdata=test, type="response")

table(test$income, test$predicted.income > 0.5)
```

```
Call:
glm(formula = income ~ ., family = binomial(logit), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1070  -0.5180  -0.1977   0.0000   3.5753

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -7.162e+00  4.224e-01 -16.955  < 2e-16 ***
age                            2.552e-02  2.004e-03  12.731  < 2e-16 ***
type_employerSelf-emp         -1.725e-01  8.307e-02  -2.077 0.037812 *
type_employerPrivate           5.690e-02  6.437e-02   0.884 0.376752
type_employerUnemployed       -1.365e+01  3.683e+02  -0.037 0.970433
fnlwgt                         5.427e-07  2.084e-07   2.605 0.009201 **
education11th                  2.122e-01  2.571e-01   0.825 0.409169
education12th                  3.773e-01  3.410e-01   1.106 0.268516
education1st-4th              -4.526e-01  6.067e-01  -0.746 0.455636
education5th-6th              -8.112e-02  3.979e-01  -0.204 0.838466
education7th-8th              -5.055e-01  2.882e-01  -1.754 0.079475 .
education9th                  -8.987e-03  3.191e-01  -0.028 0.977531
educationAssoc-acdm            1.266e+00  2.164e-01   5.851 4.88e-09 ***
educationAssoc-voc             1.457e+00  2.084e-01   6.991 2.73e-12 ***
educationBachelors             2.005e+00  1.938e-01  10.349  < 2e-16 ***
educationDoctorate             2.858e+00  2.635e-01  10.846  < 2e-16 ***
educationHS-grad               8.437e-01  1.889e-01   4.467 7.93e-06 ***
educationMasters               2.328e+00  2.062e-01  11.286  < 2e-16 ***
educationPreschool            -1.875e+01  1.645e+02  -0.114 0.909245
educationProf-school           2.812e+00  2.468e-01  11.394  < 2e-16 ***
educationSome-college          1.217e+00  1.915e-01   6.357 2.06e-10 ***
education_num                         NA         NA      NA       NA
maritalMarried                 1.275e+00  1.938e-01   6.581 4.67e-11 ***
maritalNot-Married             5.396e-01  9.941e-02   5.428 5.70e-08 ***
occupationExec-managerial      7.257e-01  9.029e-02   8.037 9.23e-16 ***
occupationHandlers-cleaners   -8.442e-01  1.723e-01  -4.901 9.53e-07 ***
occupationProf-specialty       4.334e-01  9.524e-02   4.551 5.34e-06 ***
occupationOther-service       -8.811e-01  1.380e-01  -6.386 1.71e-10 ***
occupationSales                2.442e-01  9.684e-02   2.521 0.011692 *
occupationCraft-repair        -8.840e-03  9.413e-02  -0.094 0.925185
occupationTransport-moving    -1.691e-01  1.181e-01  -1.432 0.152175
occupationFarming-fishing     -1.170e+00  1.614e-01  -7.246 4.28e-13 ***
occupationMachine-op-inspct   -2.655e-01  1.198e-01  -2.216 0.026688 *
occupationTech-support         6.609e-01  1.322e-01   4.999 5.75e-07 ***
occupationProtective-serv      4.474e-01  1.459e-01   3.066 0.002167 **
occupationArmed-Forces        -1.650e-01  1.828e+00  -0.090 0.928106
occupationPriv-house-serv     -3.621e+00  1.930e+00  -1.876 0.060701 .
relationshipNot-in-family     -8.591e-01  1.902e-01  -4.516 6.30e-06 ***
relationshipOther-relative    -1.085e+00  2.545e-01  -4.265 2.00e-05 ***
relationshipOwn-child         -1.803e+00  2.354e-01  -7.656 1.92e-14 ***
relationshipUnmarried         -1.030e+00  2.148e-01  -4.795 1.63e-06 ***
relationshipWife               1.469e+00  1.234e-01  11.906  < 2e-16 ***
raceAsian-Pac-Islander         6.113e-01  3.205e-01   1.907 0.056499 .
raceBlack                      4.586e-01  2.852e-01   1.608 0.107787
raceOther                      4.838e-02  4.226e-01   0.114 0.908851
raceWhite                      6.523e-01  2.715e-01   2.402 0.016304 *
sexMale                        9.037e-01  9.368e-02   9.647  < 2e-16 ***
capital_gain                   3.186e-04  1.270e-05  25.094  < 2e-16 ***
capital_loss                   6.555e-04  4.558e-05  14.382  < 2e-16 ***
hr_per_week                    2.917e-02  1.986e-03  14.686  < 2e-16 ***
regionLatin.and.South.America -5.907e-01  1.590e-01  -3.714 0.000204 ***
regionAsia                    -6.892e-02  2.036e-01  -0.339 0.734936
regionOther                   -4.205e-01  1.648e-01  -2.551 0.010741 *
regionEurope                   4.538e-02  1.550e-01   0.293 0.769630
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24138  on 21502  degrees of freedom
Residual deviance: 14033  on 21450  degrees of freedom
AIC: 14139

Number of Fisher Scoring iterations: 14
```

	FALSE	TRUE
<=50K	6376	544
>50K	874	1421

Accuracy

```
(6372+1423)/(6372+1423+548+872)
0.845903418339664
```

Recall

```
6732/(6372+548)
0.9728324
```

Precision

```
6732/(6372+872)
0.9293208
```

From the accuracy, recall and precision we can see that the model is working well.