Adult Classification Project

Kevin Torres

**Task:** I will attempt to predict if peple in the data set belong in a certain class based on salary. (either making <= 50k or > 50k)

**Dataset:** I will be using a dataset with information about adults including 15 variables.

Load csv file:

df = read.csv("adult\_sal.csv")  
adult <- as.data.frame(df)  
head(adult)

## X age type\_employer fnlwgt education education\_num marital  
## 1 1 39 State-gov 77516 Bachelors 13 Never-married  
## 2 2 50 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse  
## 3 3 38 Private 215646 HS-grad 9 Divorced  
## 4 4 53 Private 234721 11th 7 Married-civ-spouse  
## 5 5 28 Private 338409 Bachelors 13 Married-civ-spouse  
## 6 6 37 Private 284582 Masters 14 Married-civ-spouse  
## occupation relationship race sex capital\_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  
## 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

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

Remove the first column and show head, srt, and summary to get an idea of the data

adult <- select(adult, -X)  
head(adult)

## age type\_employer fnlwgt education education\_num marital  
## 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 capital\_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  
## 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

str(adult)

## 'data.frame': 32561 obs. of 15 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ type\_employer: Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ education\_num: int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital : Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...  
## $ occupation : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...  
## $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ 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 : Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...  
## $ income : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...

summary(adult)

## age type\_employer 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   
## 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 hr\_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   
##   
## country income   
## United-States:29170 <=50K:24720   
## Mexico : 643 >50K : 7841   
## ? : 583   
## Philippines : 198   
## Germany : 137   
## Canada : 121   
## (Other) : 1709

Clean the data - Find categories to combine to simplify the data

table(adult$type\_employer)

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

# Combine the employer type   
  
unemp <- function(job){  
 job <- as.character(job)  
 if (job == 'Never-worked' | job == 'Without-pay') {  
 return('Unemployed')  
 } else{  
 return(job)  
 }  
}  
  
adult$type\_employer <- sapply(adult$type\_employer, unemp)  
table(adult$type\_employer)

##   
## ? Federal-gov Local-gov Private   
## 1836 960 2093 22696   
## Self-emp-inc Self-emp-not-inc State-gov Unemployed   
## 1116 2541 1298 21

Do the same for Job

# Local and State level jobs  
  
sl\_gov <- function(job){  
 job <- as.character(job)  
 if (job == 'Local-gov' | job == 'State-gov') {  
 return('SL-gov')  
 } else{  
 return(job)  
 }  
}  
  
adult$type\_employer <- sapply(adult$type\_employer, sl\_gov)  
  
table(adult$type\_employer)

##   
## ? Federal-gov Private Self-emp-inc   
## 1836 960 22696 1116   
## Self-emp-not-inc SL-gov Unemployed   
## 2541 3391 21

# Self-employed  
  
self\_emp <- function(job){  
 job <- as.character(job)  
 if (job == 'Self-emp-inc' | job == 'Self-emp-not-inc') {  
 return('Self-emp')  
 } else{  
 return(job)  
 }  
}  
  
adult$type\_employer <- sapply(adult$type\_employer, self\_emp)  
  
table(adult$type\_employer)

##   
## ? Federal-gov Private Self-emp SL-gov Unemployed   
## 1836 960 22696 3657 3391 21

Do the same for Martial Status

not\_married <- function(status){  
 status <- as.character(status)  
 if (status == 'Divorced' | status == 'Separated' | status == 'Widowed') {  
 return('Not-Married')  
 } else{  
 return(status)  
 }  
}  
  
adult$marital <- sapply(adult$marital, not\_married)  
  
table(adult$marital)

##   
## Married-AF-spouse Married-civ-spouse Married-spouse-absent   
## 23 14976 418   
## Never-married Not-Married   
## 10683 6461

married <- function(status){  
 status <- as.character(status)  
 if (status == 'Married-spouse-absent' | status == 'Married-AF-spouse' | status == 'Married-civ-spouse') {  
 return('Married')  
 } else{  
 return(status)  
 }  
}  
  
adult$marital <- sapply(adult$marital, married)  
  
table(adult$marital)

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

Clean data - Country column

# Group the countries by continent  
Asia <- c('China', 'Hong', 'India', 'Iran', 'Cambodia', 'Japan', 'Laos', 'Philipines', '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', 'Trinidad&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$country <- sapply(adult$country, group\_country)  
  
table(adult$country)

##   
## Asia Europe Latin.and.South.America   
## 473 521 1282   
## North.America Other   
## 29405 880

Rename country column

# Now make each of the cleaned columns factors  
  
adult$type\_employer <- factor(adult$type\_employer)  
adult$marital <- factor(adult$marital)  
adult$country <- factor(adult$country)  
  
adult <- rename(adult, region = country)  
str(adult)

## 'data.frame': 32561 obs. of 15 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ type\_employer: Factor w/ 6 levels "?","Federal-gov",..: 5 4 3 3 3 3 3 4 3 3 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ education\_num: int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital : Factor w/ 3 levels "Married","Never-married",..: 2 1 3 1 1 1 1 1 2 1 ...  
## $ occupation : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...  
## $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ 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 ...  
## $ region : Factor w/ 5 levels "Asia","Europe",..: 4 4 4 4 3 4 3 4 4 4 ...  
## $ income : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...

Now that we have cleaned some of the data, lets check for NA values

library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.6, built: 2019-11-24)  
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

# First lets convert the responses with '?' to an NA response  
adult[adult == '?'] <- NA  
str(adult)

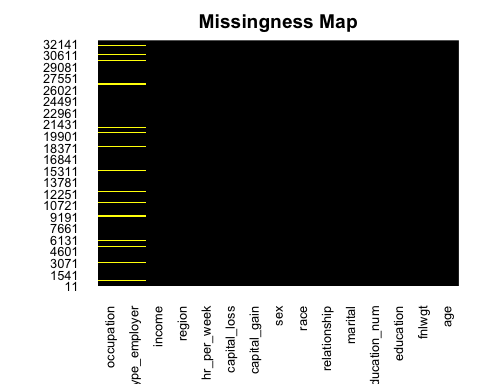
## 'data.frame': 32561 obs. of 15 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ type\_employer: Factor w/ 6 levels "?","Federal-gov",..: 5 4 3 3 3 3 3 4 3 3 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ education\_num: int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital : Factor w/ 3 levels "Married","Never-married",..: 2 1 3 1 1 1 1 1 2 1 ...  
## $ occupation : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...  
## $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ 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 ...  
## $ region : Factor w/ 5 levels "Asia","Europe",..: 4 4 4 4 3 4 3 4 4 4 ...  
## $ income : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...

table(adult$type\_employer)

##   
## ? Federal-gov Private Self-emp SL-gov Unemployed   
## 0 960 22696 3657 3391 21

Now we can see the missing data on the map

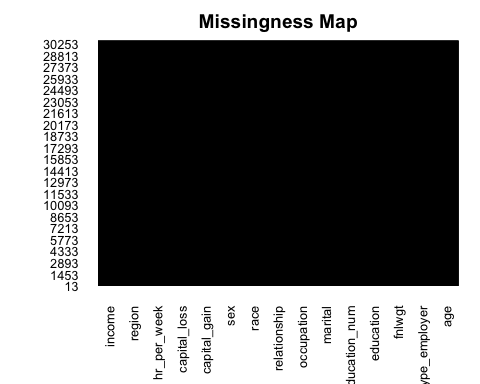
missmap(adult, main = 'Missingness Map', col = c("Yellow", 'Black'), legend = F)



# Remove NA values, even though this is not necessarily the best thing to do in most cases  
  
adult <- na.omit(adult)  
  
# Now that the data is clean, further explore it using visualization  
  
str(adult)

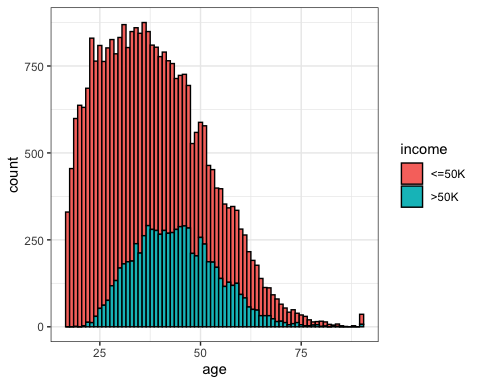
## 'data.frame': 30718 obs. of 15 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ type\_employer: Factor w/ 6 levels "?","Federal-gov",..: 5 4 3 3 3 3 3 4 3 3 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ education\_num: int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital : Factor w/ 3 levels "Married","Never-married",..: 2 1 3 1 1 1 1 1 2 1 ...  
## $ occupation : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...  
## $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ 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 ...  
## $ region : Factor w/ 5 levels "Asia","Europe",..: 4 4 4 4 3 4 3 4 4 4 ...  
## $ income : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...  
## - attr(\*, "na.action")= 'omit' Named int 28 62 70 78 107 129 150 155 161 188 ...  
## ..- attr(\*, "names")= chr "28" "62" "70" "78" ...

missmap(adult, main = 'Missingness Map', col = c("Yellow", 'Black'), legend = F)



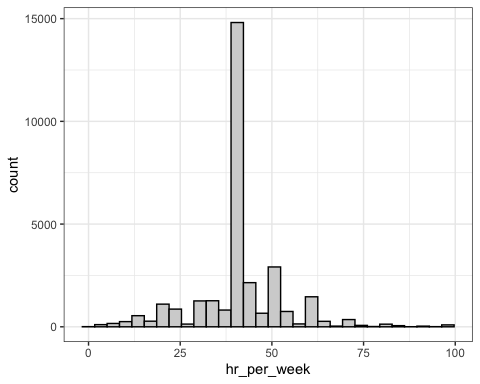
Exploratory Data Analysis

library(ggplot2)  
# Histogram of ages, colored by income  
  
ggplot(adult, aes(x = age, fill = income)) + geom\_histogram(binwidth = 1, color = 'black') + theme\_bw()

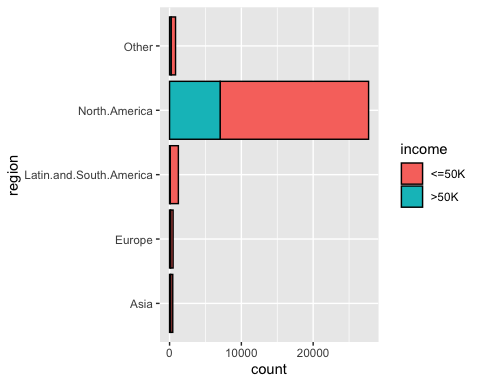


# Histogram of hours worked per week  
  
ggplot(adult, aes(x = hr\_per\_week)) + geom\_histogram(color = 'black', fill = 'light grey') + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Barplot of region  
  
ggplot(adult, aes(y = region, fill = income)) + geom\_bar(color = 'black')



Now it’s time to create a train and test data set

head(adult)

## age type\_employer fnlwgt education education\_num marital  
## 1 39 SL-gov 77516 Bachelors 13 Never-married  
## 2 50 Self-emp 83311 Bachelors 13 Married  
## 3 38 Private 215646 HS-grad 9 Not-Married  
## 4 53 Private 234721 11th 7 Married  
## 5 28 Private 338409 Bachelors 13 Married  
## 6 37 Private 284582 Masters 14 Married  
## occupation relationship race sex capital\_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  
## hr\_per\_week region income  
## 1 40 North.America <=50K  
## 2 13 North.America <=50K  
## 3 40 North.America <=50K  
## 4 40 North.America <=50K  
## 5 40 Latin.and.South.America <=50K  
## 6 40 North.America <=50K

library(caTools)  
  
sample <- sample.split(adult$income, SplitRatio = .7)  
train <- subset(adult, sample == T)  
test <- subset(adult, sample == F)

Build the Logistic Regression Model

model <- glm(income ~ ., family = binomial(link = 'logit'), data = train)

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

summary(glm(income ~ ., family = binomial(link = 'logit'), data = train))

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

##   
## Call:  
## glm(formula = income ~ ., family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.1833 -0.5196 -0.1955 0.0000 3.6149   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.569e+00 4.445e-01 -12.529 < 2e-16 \*\*\*  
## age 2.523e-02 1.991e-03 12.672 < 2e-16 \*\*\*  
## type\_employerPrivate -5.534e-01 1.137e-01 -4.867 1.13e-06 \*\*\*  
## type\_employerSelf-emp -8.666e-01 1.257e-01 -6.893 5.46e-12 \*\*\*  
## type\_employerSL-gov -8.009e-01 1.278e-01 -6.265 3.73e-10 \*\*\*  
## type\_employerUnemployed -1.558e+01 6.326e+02 -0.025 0.980347   
## fnlwgt 6.612e-07 2.086e-07 3.170 0.001524 \*\*   
## education11th 2.054e-01 2.507e-01 0.819 0.412675   
## education12th 6.285e-01 3.082e-01 2.039 0.041436 \*   
## education1st-4th -4.463e-01 5.432e-01 -0.822 0.411310   
## education5th-6th -3.184e-01 3.858e-01 -0.825 0.409218   
## education7th-8th -3.915e-01 2.839e-01 -1.379 0.167815   
## education9th -1.368e-01 3.148e-01 -0.434 0.663995   
## educationAssoc-acdm 1.390e+00 2.127e-01 6.535 6.35e-11 \*\*\*  
## educationAssoc-voc 1.303e+00 2.058e-01 6.330 2.45e-10 \*\*\*  
## educationBachelors 1.897e+00 1.910e-01 9.931 < 2e-16 \*\*\*  
## educationDoctorate 3.032e+00 2.637e-01 11.501 < 2e-16 \*\*\*  
## educationHS-grad 7.890e-01 1.859e-01 4.243 2.21e-05 \*\*\*  
## educationMasters 2.273e+00 2.044e-01 11.124 < 2e-16 \*\*\*  
## educationPreschool -2.029e+01 2.703e+02 -0.075 0.940171   
## educationProf-school 2.709e+00 2.429e-01 11.152 < 2e-16 \*\*\*  
## educationSome-college 1.182e+00 1.885e-01 6.274 3.52e-10 \*\*\*  
## education\_num NA NA NA NA   
## maritalNever-married -1.226e+00 2.051e-01 -5.975 2.30e-09 \*\*\*  
## maritalNot-Married -7.046e-01 2.049e-01 -3.439 0.000584 \*\*\*  
## occupationArmed-Forces -1.065e+00 1.610e+00 -0.661 0.508469   
## occupationCraft-repair 8.317e-02 9.468e-02 0.878 0.379685   
## occupationExec-managerial 8.462e-01 9.170e-02 9.228 < 2e-16 \*\*\*  
## occupationFarming-fishing -1.130e+00 1.699e-01 -6.650 2.92e-11 \*\*\*  
## occupationHandlers-cleaners -6.651e-01 1.681e-01 -3.958 7.57e-05 \*\*\*  
## occupationMachine-op-inspct -2.393e-01 1.202e-01 -1.990 0.046547 \*   
## occupationOther-service -7.497e-01 1.376e-01 -5.450 5.03e-08 \*\*\*  
## occupationPriv-house-serv -1.345e+01 1.982e+02 -0.068 0.945884   
## occupationProf-specialty 5.120e-01 9.725e-02 5.265 1.41e-07 \*\*\*  
## occupationProtective-serv 6.318e-01 1.476e-01 4.282 1.86e-05 \*\*\*  
## occupationSales 2.467e-01 9.764e-02 2.527 0.011515 \*   
## occupationTech-support 6.167e-01 1.324e-01 4.656 3.22e-06 \*\*\*  
## occupationTransport-moving -1.881e-01 1.184e-01 -1.590 0.111933   
## relationshipNot-in-family -9.621e-01 2.019e-01 -4.766 1.88e-06 \*\*\*  
## relationshipOther-relative -1.136e+00 2.595e-01 -4.377 1.20e-05 \*\*\*  
## relationshipOwn-child -1.989e+00 2.512e-01 -7.919 2.39e-15 \*\*\*  
## relationshipUnmarried -1.087e+00 2.249e-01 -4.834 1.34e-06 \*\*\*  
## relationshipWife 1.336e+00 1.251e-01 10.683 < 2e-16 \*\*\*  
## raceAsian-Pac-Islander 6.733e-01 3.174e-01 2.121 0.033897 \*   
## raceBlack 3.294e-01 2.885e-01 1.142 0.253506   
## raceOther 5.750e-02 4.243e-01 0.136 0.892199   
## raceWhite 5.798e-01 2.751e-01 2.108 0.035048 \*   
## sexMale 8.711e-01 9.524e-02 9.146 < 2e-16 \*\*\*  
## capital\_gain 3.298e-04 1.296e-05 25.457 < 2e-16 \*\*\*  
## capital\_loss 6.203e-04 4.462e-05 13.901 < 2e-16 \*\*\*  
## hr\_per\_week 3.053e-02 1.977e-03 15.443 < 2e-16 \*\*\*  
## regionEurope 3.242e-01 2.669e-01 1.215 0.224499   
## regionLatin.and.South.America -1.541e-01 2.640e-01 -0.584 0.559440   
## regionNorth.America 3.416e-01 2.168e-01 1.576 0.115088   
## regionOther 4.163e-02 2.215e-01 0.188 0.850940   
## ---  
## 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: 14049 on 21449 degrees of freedom  
## AIC: 14157  
##   
## Number of Fisher Scoring iterations: 15

# Here we see a lot of significant factors so there might be more factor groups we could create

Try to remove predictive variables from the model to attempt to delete variables that do not significantly add to the fit (use step function)

new.step.model <- step(model)

## Start: AIC=14156.67  
## income ~ age + type\_employer + fnlwgt + education + education\_num +   
## marital + occupation + relationship + race + sex + capital\_gain +   
## capital\_loss + hr\_per\_week + region

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

##   
## Step: AIC=14156.67  
## income ~ age + type\_employer + fnlwgt + education + marital +   
## occupation + relationship + race + sex + capital\_gain + capital\_loss +   
## hr\_per\_week + region

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

## Df Deviance AIC  
## <none> 14049 14157  
## - race 4 14064 14164  
## - region 4 14064 14164  
## - fnlwgt 1 14059 14165  
## - marital 2 14097 14201  
## - type\_employer 4 14117 14217  
## - sex 1 14135 14241  
## - age 1 14210 14316  
## - capital\_loss 1 14247 14353  
## - relationship 5 14301 14399  
## - hr\_per\_week 1 14295 14401  
## - occupation 13 14495 14577  
## - education 15 14725 14803  
## - capital\_gain 1 15298 15404

summary(new.step.model)

##   
## Call:  
## glm(formula = income ~ age + type\_employer + fnlwgt + education +   
## marital + occupation + relationship + race + sex + capital\_gain +   
## capital\_loss + hr\_per\_week + region, family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.1833 -0.5196 -0.1955 0.0000 3.6149   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.569e+00 4.445e-01 -12.529 < 2e-16 \*\*\*  
## age 2.523e-02 1.991e-03 12.672 < 2e-16 \*\*\*  
## type\_employerPrivate -5.534e-01 1.137e-01 -4.867 1.13e-06 \*\*\*  
## type\_employerSelf-emp -8.666e-01 1.257e-01 -6.893 5.46e-12 \*\*\*  
## type\_employerSL-gov -8.009e-01 1.278e-01 -6.265 3.73e-10 \*\*\*  
## type\_employerUnemployed -1.558e+01 6.326e+02 -0.025 0.980347   
## fnlwgt 6.612e-07 2.086e-07 3.170 0.001524 \*\*   
## education11th 2.054e-01 2.507e-01 0.819 0.412675   
## education12th 6.285e-01 3.082e-01 2.039 0.041436 \*   
## education1st-4th -4.463e-01 5.432e-01 -0.822 0.411310   
## education5th-6th -3.184e-01 3.858e-01 -0.825 0.409218   
## education7th-8th -3.915e-01 2.839e-01 -1.379 0.167815   
## education9th -1.368e-01 3.148e-01 -0.434 0.663995   
## educationAssoc-acdm 1.390e+00 2.127e-01 6.535 6.35e-11 \*\*\*  
## educationAssoc-voc 1.303e+00 2.058e-01 6.330 2.45e-10 \*\*\*  
## educationBachelors 1.897e+00 1.910e-01 9.931 < 2e-16 \*\*\*  
## educationDoctorate 3.032e+00 2.637e-01 11.501 < 2e-16 \*\*\*  
## educationHS-grad 7.890e-01 1.859e-01 4.243 2.21e-05 \*\*\*  
## educationMasters 2.273e+00 2.044e-01 11.124 < 2e-16 \*\*\*  
## educationPreschool -2.029e+01 2.703e+02 -0.075 0.940171   
## educationProf-school 2.709e+00 2.429e-01 11.152 < 2e-16 \*\*\*  
## educationSome-college 1.182e+00 1.885e-01 6.274 3.52e-10 \*\*\*  
## maritalNever-married -1.226e+00 2.051e-01 -5.975 2.30e-09 \*\*\*  
## maritalNot-Married -7.046e-01 2.049e-01 -3.439 0.000584 \*\*\*  
## occupationArmed-Forces -1.065e+00 1.610e+00 -0.661 0.508469   
## occupationCraft-repair 8.317e-02 9.468e-02 0.878 0.379685   
## occupationExec-managerial 8.462e-01 9.170e-02 9.228 < 2e-16 \*\*\*  
## occupationFarming-fishing -1.130e+00 1.699e-01 -6.650 2.92e-11 \*\*\*  
## occupationHandlers-cleaners -6.651e-01 1.681e-01 -3.958 7.57e-05 \*\*\*  
## occupationMachine-op-inspct -2.393e-01 1.202e-01 -1.990 0.046547 \*   
## occupationOther-service -7.497e-01 1.376e-01 -5.450 5.03e-08 \*\*\*  
## occupationPriv-house-serv -1.345e+01 1.982e+02 -0.068 0.945884   
## occupationProf-specialty 5.120e-01 9.725e-02 5.265 1.41e-07 \*\*\*  
## occupationProtective-serv 6.318e-01 1.476e-01 4.282 1.86e-05 \*\*\*  
## occupationSales 2.467e-01 9.764e-02 2.527 0.011515 \*   
## occupationTech-support 6.167e-01 1.324e-01 4.656 3.22e-06 \*\*\*  
## occupationTransport-moving -1.881e-01 1.184e-01 -1.590 0.111933   
## relationshipNot-in-family -9.621e-01 2.019e-01 -4.766 1.88e-06 \*\*\*  
## relationshipOther-relative -1.136e+00 2.595e-01 -4.377 1.20e-05 \*\*\*  
## relationshipOwn-child -1.989e+00 2.512e-01 -7.919 2.39e-15 \*\*\*  
## relationshipUnmarried -1.087e+00 2.249e-01 -4.834 1.34e-06 \*\*\*  
## relationshipWife 1.336e+00 1.251e-01 10.683 < 2e-16 \*\*\*  
## raceAsian-Pac-Islander 6.733e-01 3.174e-01 2.121 0.033897 \*   
## raceBlack 3.294e-01 2.885e-01 1.142 0.253506   
## raceOther 5.750e-02 4.243e-01 0.136 0.892199   
## raceWhite 5.798e-01 2.751e-01 2.108 0.035048 \*   
## sexMale 8.711e-01 9.524e-02 9.146 < 2e-16 \*\*\*  
## capital\_gain 3.298e-04 1.296e-05 25.457 < 2e-16 \*\*\*  
## capital\_loss 6.203e-04 4.462e-05 13.901 < 2e-16 \*\*\*  
## hr\_per\_week 3.053e-02 1.977e-03 15.443 < 2e-16 \*\*\*  
## regionEurope 3.242e-01 2.669e-01 1.215 0.224499   
## regionLatin.and.South.America -1.541e-01 2.640e-01 -0.584 0.559440   
## regionNorth.America 3.416e-01 2.168e-01 1.576 0.115088   
## regionOther 4.163e-02 2.215e-01 0.188 0.850940   
## ---  
## 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: 14049 on 21449 degrees of freedom  
## AIC: 14157  
##   
## Number of Fisher Scoring iterations: 15

Create a confuision matrix based on predicted values

test$predicted.income <- predict(model, 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

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

##   
## FALSE TRUE  
## <=50K 6412 508  
## >50K 875 1420

Get measurements based off confusion matrix

# Accuracy   
acc <- (6396+1409)/(6396+524+886+1409)  
acc

## [1] 0.8469886

# Recall  
rcl <- 6396/(6396+524)  
rcl

## [1] 0.9242775

# Precision  
prc <- 6396/(6396+886)  
prc

## [1] 0.8783301

We can determine if this is a good model after we find out the cost associated with accuracy, precision, and recall. Some models want to maximize one over the other.