FInal\_project

TarakRam Nunna

06/12/2021

# Setting working Directory

getwd()

## [1] "C:/Users/TARAKRAM/OneDrive/Desktop/ML\_code/64060\_tnunna/Final Project"

setwd("C:/Users/TARAKRAM/OneDrive/Desktop/ML\_code/64060\_tnunna/Final Project")

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

library(caTools)

## Warning: package 'caTools' was built under R version 4.1.2

library(Amelia)

## Warning: package 'Amelia' was built under R version 4.1.2

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.8.0, built: 2021-05-26)  
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

library(caret)

## Loading required package: lattice

adult\_income <- read.csv("AdultIncome.csv")  
View(adult\_income)  
str(adult\_income)

## 'data.frame': 32561 obs. of 15 variables:  
## $ age : int 90 82 66 54 41 34 38 74 68 41 ...  
## $ workclass : chr "?" "Private" "?" "Private" ...  
## $ fnlwgt : int 77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...  
## $ education : chr "HS-grad" "HS-grad" "Some-college" "7th-8th" ...  
## $ education.num : int 9 9 10 4 10 9 6 16 9 10 ...  
## $ marital.status: chr "Widowed" "Widowed" "Widowed" "Divorced" ...  
## $ occupation : chr "?" "Exec-managerial" "?" "Machine-op-inspct" ...  
## $ relationship : chr "Not-in-family" "Not-in-family" "Unmarried" "Unmarried" ...  
## $ race : chr "White" "White" "Black" "White" ...  
## $ sex : chr "Female" "Female" "Female" "Female" ...  
## $ 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: chr "United-States" "United-States" "United-States" "United-States" ...  
## $ income : chr "<=50K" "<=50K" "<=50K" "<=50K" ...

# cleaning the data

## combining factors of work class

table(adult\_income$workclass)

##   
## ? 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

adult\_income$workclass <- as.character(adult\_income$workclass)  
# combining work class of without-pay and never-worked as Unemployed  
adult\_income$workclass[adult\_income$workclass == "Without-pay" | adult\_income$workclass == "Never-worked"] <- "Unemployed"  
# combining work class of state-gov and local-gov as State/Local-gov  
adult\_income$workclass[adult\_income$workclass == "State-gov" | adult\_income$workclass == "Local-gov"] <- "State/Local-gov"  
# combining work class of self-emp-inc and self-emp-not-inc as Self-employed  
adult\_income$workclass[adult\_income$workclass == "Self-emp-inc" | adult\_income$workclass == "Self-emp-not-inc"] <- "Self-employed"  
# we are not combining federal work class and private work class because both are different work classes  
table(adult\_income$workclass)

##   
## ? Federal-gov Private Self-employed State/Local-gov   
## 1836 960 22696 3657 3391   
## Unemployed   
## 21

## Combining factors of marital status

table(adult\_income$marital.status)

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

adult\_income$marital.status <- as.character(adult\_income$marital.status)  
# Combining Married-AF-spouse, Married-civ-spouse and Married-spouse-absent as Married  
adult\_income$marital.status[adult\_income$marital.status == "Married-AF-spouse" | adult\_income$marital.status == "Married-civ-spouse" | adult\_income$marital.status == "Married-spouse-absent"] <- "Married"  
  
# Combining Divorced, Separated and Widowed as Not-Married  
adult\_income$marital.status[adult\_income$marital.status == "Divorced" | adult\_income$marital.status == "Separated" | adult\_income$marital.status == "Widowed"] <- "Not-Married"  
table(adult\_income$marital.status)

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

## Combining factors of Country

adult\_income$native.country <- as.character(adult\_income$native.country)  
# combining the below countries to North.America  
North.America <- c("Canada","Cuba","Dominican-Republic","El-Salvador","Guatemala","Haiti","Honduras","Jamaica","Mexico","Nicaragua","Outlying-US(Guam-USVI-etc)","Puerto-Rico","Trinadad&Tobago","United-States")  
# combining the below countries to Asia  
Asia <- c("Cambodia","China","Hong","India","Iran","Japan","Laos","Philippines","Taiwan","Thailand","Vietnam")  
# combining the below countries to South.America  
South.America <- c("Columbia","Ecuador","Peru")  
# combining the below countries to Europe  
Europe <- c("England", "France", "Germany", "Greece", "Holand-Netherlands", "Hungary", "Ireland", "Italy", "Poland", "Portugal", "Scotland", "Yugoslavia")  
# combining the below countries to others  
Others <- c("South","?")  
  
adult\_income$native.country[adult\_income$native.country %in% North.America] <- "North\_America"  
adult\_income$native.country[adult\_income$native.country %in% Asia] <- "Asia"  
adult\_income$native.country[adult\_income$native.country %in% South.America] <- "South\_America"  
adult\_income$native.country[adult\_income$native.country %in% Europe] <- "Europe"  
adult\_income$native.country[adult\_income$native.country %in% Others] <- "Others"  
table(adult\_income$native.country)

##   
## Asia Europe North\_America Others South\_America   
## 671 521 30588 663 118

# converting the below variables into factors  
adult\_income$workclass <- as.factor(adult\_income$workclass)  
adult\_income$marital.status <- as.factor(adult\_income$marital.status)  
adult\_income$native.country <- as.factor(adult\_income$native.country)  
str(adult\_income)

## 'data.frame': 32561 obs. of 15 variables:  
## $ age : int 90 82 66 54 41 34 38 74 68 41 ...  
## $ workclass : Factor w/ 6 levels "?","Federal-gov",..: 1 3 1 3 3 3 3 5 2 3 ...  
## $ fnlwgt : int 77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...  
## $ education : chr "HS-grad" "HS-grad" "Some-college" "7th-8th" ...  
## $ education.num : int 9 9 10 4 10 9 6 16 9 10 ...  
## $ marital.status: Factor w/ 3 levels "Married","Never-married",..: 3 3 3 3 3 3 3 2 3 2 ...  
## $ occupation : chr "?" "Exec-managerial" "?" "Machine-op-inspct" ...  
## $ relationship : chr "Not-in-family" "Not-in-family" "Unmarried" "Unmarried" ...  
## $ race : chr "White" "White" "Black" "White" ...  
## $ sex : chr "Female" "Female" "Female" "Female" ...  
## $ 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/ 5 levels "Asia","Europe",..: 3 3 3 3 3 3 3 3 3 4 ...  
## $ income : chr "<=50K" "<=50K" "<=50K" "<=50K" ...

# Now we deal with missing data

table(adult\_income$workclass)

##   
## ? Federal-gov Private Self-employed State/Local-gov   
## 1836 960 22696 3657 3391   
## Unemployed   
## 21

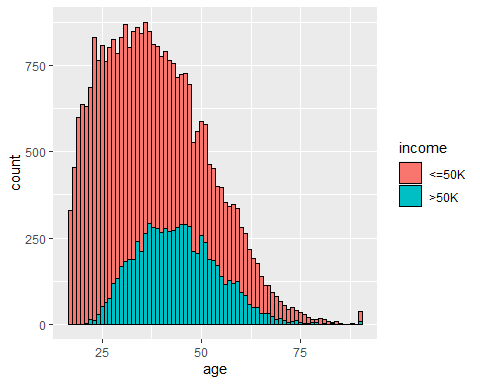
# assigning NA to the missing values  
adult\_income[adult\_income == "?"] <- NA  
# Converting Income to factors as well  
adult\_income$income <- as.factor(adult\_income$income)  
adult\_income[adult\_income$income == "<=50k"] <- "0"  
adult\_income[adult\_income$income == ">50k"] <- "1"  
table(adult\_income$workclass)

##   
## ? Federal-gov Private Self-employed State/Local-gov   
## 0 960 22696 3657 3391   
## Unemployed   
## 21

# omitting the NA values  
adult\_income <- na.omit(adult\_income)

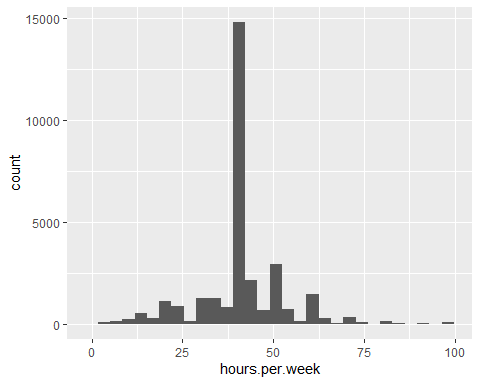
# Exploring and analysing data

# We have to understand the trends and representations of certain demographics  
ggplot(adult\_income, aes(age)) + geom\_histogram(aes(fill = income), color = "black", binwidth = 1)



ggplot(adult\_income, aes(hours.per.week)) + geom\_histogram()

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



ggplot(adult\_income, aes(age)) + geom\_histogram()

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



# Now we build our Logistic Regression Model

## The purpose of this model is to classify people into two groups, below 50K or above 50K

# now we split the data into 75% training and 25% testing  
Adult\_split <- sample.split(adult\_income$income, SplitRatio = 0.75)  
# we assigned training data to Adult\_train  
Adult\_Train <- subset(adult\_income, Adult\_split == TRUE,)  
# we assigned testing data to Adult\_test  
Adult\_Test <- subset(adult\_income, Adult\_split == FALSE)  
  
# Training the model  
Adult\_income\_model <- glm(income ~., family = binomial(), data = Adult\_Train)

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

# Predicting the Salary class  
Prediction <- predict(Adult\_income\_model, Adult\_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

# Creating a confusion matrix

table(Adult\_Test$income, Prediction >= 0.7)

##   
## FALSE TRUE  
## <=50K 5623 144  
## >50K 1164 748

# Accuracy  
(5610+724)/(5610+157+1188+724)

## [1] 0.824847

# REcall  
5627/(5627+1190)

## [1] 0.8254364

#precision  
5627/(5627+722)

## [1] 0.8862813