ALY 6015 Final Project Draft.R

# Intermediate Analytics  
# ALY 6015  
# Preliminary Analysis Group R Files  
# 02/12/2021  
# Team: Sunil Raj Thota, Nalini Macharla

# Get and set the working directories  
getwd()

## [1] "G:/NEU/Coursework/2021 Q1 Winter/ALY 6015 IA/Discussions & Assignments"

setwd('G:/NEU/Coursework/2021 Q1 Winter/ALY 6015 IA/Discussions & Assignments')  
getwd()

## [1] "G:/NEU/Coursework/2021 Q1 Winter/ALY 6015 IA/Discussions & Assignments"

# Installed the above packages into the work space  
install.packages("plyr")  
install.packages("dplyr")  
install.packages("tidyr")  
install.packages("tidyverse")  
install.packages("ggplot2")  
install.packages("e1071")  
install.packages("gmodels")  
install.packages("caret")  
install.packages("ROCR")  
install.packages("kableExtra")  
install.packages("rpart")  
install.packages("rpart.plot")  
install.packages("caTools")  
install.packages("ncvreg")  
install.packages("biglasso")  
install.packages("bigmemory")  
install.packages("glmnet")  
install.packages("lars")  
install.packages("randomForest")  
install.packages("rattle")  
install.packages("gridExtra")  
  
# Loaded the below libraries into the work space  
library(plyr)

library(dplyr)

library(tidyr)

library(tidyverse)

library(ggplot2)

require(e1071)

library(gmodels)

library(data.table)

library(caret)

library(ROCR)

library(kableExtra)

library(rpart)

library(rpart.plot)

library(caTools)

library(ncvreg)

library(biglasso)

library(bigmemory)

library(lars)

library(glmnet)

library(randomForest)

library(gridExtra)

library(rattle)

require(grDevices)

bankData <- read.csv("Bank Dataset.csv")

bankDataMain <- bankData  
  
View(bankData) # To View the bank Data set  
str(bankData) # To observe the structure of the Data set

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : chr "housemaid" "services" "services" "admin." ...  
## $ marital : chr "married" "married" "married" "married" ...  
## $ education : chr "basic.4y" "high.school" "high.school" "basic.6y" ...  
## $ default : chr "no" "unknown" "no" "no" ...  
## $ housing : chr "no" "no" "yes" "no" ...  
## $ loan : chr "no" "no" "no" "no" ...  
## $ contact : chr "telephone" "telephone" "telephone" "telephone" ...  
## $ month : chr "may" "may" "may" "may" ...  
## $ day\_of\_week : chr "mon" "mon" "mon" "mon" ...  
## $ duration : int 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : num 5191 5191 5191 5191 5191 ...  
## $ y : chr "no" "no" "no" "no" ...

head(bankData) # It shows first few rows in the Data set

## age job marital education default housing loan contact month  
## 1 56 housemaid married basic.4y no no no telephone may  
## 2 57 services married high.school unknown no no telephone may  
## 3 37 services married high.school no yes no telephone may  
## 4 40 admin. married basic.6y no no no telephone may  
## 5 56 services married high.school no no yes telephone may  
## 6 45 services married basic.9y unknown no no telephone may  
## day\_of\_week duration campaign pdays previous poutcome emp.var.rate  
## 1 mon 261 1 999 0 nonexistent 1.1  
## 2 mon 149 1 999 0 nonexistent 1.1  
## 3 mon 226 1 999 0 nonexistent 1.1  
## 4 mon 151 1 999 0 nonexistent 1.1  
## 5 mon 307 1 999 0 nonexistent 1.1  
## 6 mon 198 1 999 0 nonexistent 1.1  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 1 93.994 -36.4 4.857 5191 no  
## 2 93.994 -36.4 4.857 5191 no  
## 3 93.994 -36.4 4.857 5191 no  
## 4 93.994 -36.4 4.857 5191 no  
## 5 93.994 -36.4 4.857 5191 no  
## 6 93.994 -36.4 4.857 5191 no

tail(bankData) # It shows last few rows in the Data set

## age job marital education default housing loan contact  
## 41183 29 unemployed single basic.4y no yes no cellular  
## 41184 73 retired married professional.course no yes no cellular  
## 41185 46 blue-collar married professional.course no no no cellular  
## 41186 56 retired married university.degree no yes no cellular  
## 41187 44 technician married professional.course no no no cellular  
## 41188 74 retired married professional.course no yes no cellular  
## month day\_of\_week duration campaign pdays previous poutcome  
## 41183 nov fri 112 1 9 1 success  
## 41184 nov fri 334 1 999 0 nonexistent  
## 41185 nov fri 383 1 999 0 nonexistent  
## 41186 nov fri 189 2 999 0 nonexistent  
## 41187 nov fri 442 1 999 0 nonexistent  
## 41188 nov fri 239 3 999 1 failure  
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 41183 -1.1 94.767 -50.8 1.028 4963.6 no  
## 41184 -1.1 94.767 -50.8 1.028 4963.6 yes  
## 41185 -1.1 94.767 -50.8 1.028 4963.6 no  
## 41186 -1.1 94.767 -50.8 1.028 4963.6 no  
## 41187 -1.1 94.767 -50.8 1.028 4963.6 yes  
## 41188 -1.1 94.767 -50.8 1.028 4963.6 no

summary(bankData) # Provides the Descriptive Stats of the bank Data set

## age job marital education   
## Min. :17.00 Length:41188 Length:41188 Length:41188   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41188 Length:41188 Length:41188 Length:41188   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41188 Length:41188 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41188 Min. :-3.40000   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character 1st Qu.:-1.80000   
## Median :999.0 Median :0.000 Mode :character Median : 1.10000   
## Mean :962.5 Mean :0.173 Mean : 0.08189   
## 3rd Qu.:999.0 3rd Qu.:0.000 3rd Qu.: 1.40000   
## Max. :999.0 Max. :7.000 Max. : 1.40000   
## cons.price.idx cons.conf.idx euribor3m nr.employed   
## Min. :92.20 Min. :-50.8 Min. :0.634 Min. :4964   
## 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099   
## Median :93.75 Median :-41.8 Median :4.857 Median :5191   
## Mean :93.58 Mean :-40.5 Mean :3.621 Mean :5167   
## 3rd Qu.:93.99 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :94.77 Max. :-26.9 Max. :5.045 Max. :5228   
## y   
## Length:41188   
## Class :character   
## Mode :character

dim(bankData) # Shows the count of rows and columns in the dataset

## [1] 41188 21

sum(duplicated(bankDataMain)) # Check for duplicate records

## [1] 12

sum(!complete.cases(bankDataMain)) # Checking for Rows with missing Data

## [1] 0

all.empty <-  
 rowSums(is.na(bankDataMain)) == ncol(bankDataMain) # How many rows are completely went missing in all the cols  
sum(all.empty)

## [1] 0

sapply(bankDataMain, function(x)  
 sum(is.na(x))) # Missing values by variables

## age job marital education default   
## 0 0 0 0 0   
## housing loan contact month day\_of\_week   
## 0 0 0 0 0   
## duration campaign pdays previous poutcome   
## 0 0 0 0 0   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   
## 0 0 0 0 0   
## y   
## 0

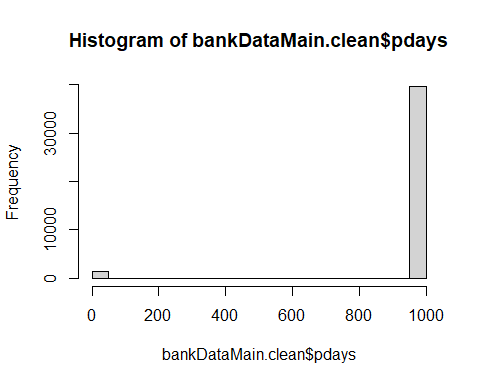
bankDataMain.clean <- bankDataMain[!all.empty,]  
bankDataMain.clean <- bankDataMain.clean %>% distinct  
# Remove rows with clos that has missing values  
  
nrow(bankDataMain.clean)

## [1] 41176

# Impute Missing Values - replace with average  
bankDataMain.clean$missing <- !complete.cases(bankDataMain.clean)  
  
bankDataMain.clean$age[is.na(bankDataMain.clean$age)] <-  
 mean(bankDataMain$age, na.rm = T)  
bankDataMain.clean$day[is.na(bankDataMain.clean$day)] <-  
 mean(bankDataMain$day, na.rm = T)

## Warning in mean.default(bankDataMain$day, na.rm = T): argument is not numeric or  
## logical: returning NA

bankDataMain.clean$duration[is.na(bankDataMain.clean$duration)] <-  
 mean(bankDataMain$duration, na.rm = T)  
bankDataMain.clean$previous[is.na(bankDataMain.clean$previous)] <-  
 mean(bankDataMain$previous, na.rm = T)  
bankDataMain.clean$campaign[is.na(bankDataMain.clean$campaign)] <-  
 mean(bankDataMain$campaign, na.rm = T)  
  
# Plotted histogram of pdays  
hist(bankDataMain.clean$pdays)



bankDataMain.clean$pdays[is.na(bankDataMain.clean$pdays)] <-  
 as.numeric(names(sort(-table(bankDataMain$pdays)))[1])  
  
bankDataMain.clean$balance[is.na(bankDataMain.clean$balance)] <-  
 as.numeric(names(sort(-table(  
 bankDataMain$balance  
 )))[1])

bankDataMain.clean <- bankDataMain.clean %>% distinct  
nrow(bankDataMain)

## [1] 41188

nrow(bankDataMain.clean)

## [1] 41176

# Remove duplicated rows and verify for deduplication  
sum(duplicated(bankDataMain.clean))

## [1] 0

sapply(bankDataMain.clean, function(x)  
 sum(is.na(x)))

## age job marital education default   
## 0 0 0 0 0   
## housing loan contact month day\_of\_week   
## 0 0 0 0 0   
## duration campaign pdays previous poutcome   
## 0 0 0 0 0   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   
## 0 0 0 0 0   
## y missing day   
## 0 0 0

levels(bankDataMain.clean$job)

## NULL

levels(bankDataMain.clean$marital)

## NULL

levels(bankDataMain.clean$education)

## NULL

levels(bankDataMain.clean$default)

## NULL

levels(bankDataMain.clean$loan)

## NULL

levels(bankDataMain.clean$contact)

## NULL

levels(bankDataMain.clean$poutcome)

## NULL

levels(bankDataMain.clean$y)

## NULL

levels(bankDataMain.clean$housing)

## NULL

levels(bankDataMain.clean$month)

## NULL

sum(bankDataMain.clean$missing)

## [1] 0

#Converting quantititative values to numeric class  
bankDataMain$age <- as.numeric(bankDataMain$age)  
bankDataMain$duration <- as.numeric(bankDataMain$duration)  
bankDataMain$campaign <- as.numeric(bankDataMain$campaign)  
bankDataMain$pdays <- as.numeric(bankDataMain$pdays)  
bankDataMain$previous <- as.numeric(bankDataMain$previous)  
bankDataMain$emp.var.rate <- as.numeric(bankDataMain$emp.var.rate)  
bankDataMain$cons.price.idx <-  
 as.numeric(bankDataMain$cons.price.idx)  
bankDataMain$cons.conf.idx <- as.numeric(bankDataMain$cons.conf.idx)  
bankDataMain$nr.employed <- as.numeric(bankDataMain$nr.employed)  
  
#checking classes of attributes after transformation  
sapply(bankDataMain, class)

## age job marital education default   
## "numeric" "character" "character" "character" "character"   
## housing loan contact month day\_of\_week   
## "character" "character" "character" "character" "character"   
## duration campaign pdays previous poutcome   
## "numeric" "numeric" "numeric" "numeric" "character"   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   
## "numeric" "numeric" "numeric" "numeric" "numeric"   
## y   
## "character"

summary(bankDataMain.clean)

## age job marital education   
## Min. :17.00 Length:41176 Length:41176 Length:41176   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41176 Length:41176 Length:41176 Length:41176   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41176 Length:41176 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41176 Min. :-3.40000   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character 1st Qu.:-1.80000   
## Median :999.0 Median :0.000 Mode :character Median : 1.10000   
## Mean :962.5 Mean :0.173 Mean : 0.08192   
## 3rd Qu.:999.0 3rd Qu.:0.000 3rd Qu.: 1.40000   
## Max. :999.0 Max. :7.000 Max. : 1.40000   
## cons.price.idx cons.conf.idx euribor3m nr.employed   
## Min. :92.20 Min. :-50.8 Min. :0.634 Min. :4964   
## 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099   
## Median :93.75 Median :-41.8 Median :4.857 Median :5191   
## Mean :93.58 Mean :-40.5 Mean :3.621 Mean :5167   
## 3rd Qu.:93.99 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :94.77 Max. :-26.9 Max. :5.045 Max. :5228   
## y missing day   
## Length:41176 Mode :logical Length:41176   
## Class :character FALSE:41176 Class :character   
## Mode :character Mode :character   
##   
##   
##

# Lets save the updated data in the below format  
write.csv(bankDataMain.clean, file = "Banks Data Cleaned.csv")  
  
bankDataCleaned <- bankDataMain.clean  
bankDataCleaned

# Conditionally formatting all "y" to 0, and 1  
bankDataCleaned$y <- ifelse(bankDataCleaned$y == "y", 1, 0)  
bankDataCleaned

str(bankDataCleaned)

## 'data.frame': 41176 obs. of 23 variables:  
## $ age : num 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : chr "housemaid" "services" "services" "admin." ...  
## $ marital : chr "married" "married" "married" "married" ...  
## $ education : chr "basic.4y" "high.school" "high.school" "basic.6y" ...  
## $ default : chr "no" "unknown" "no" "no" ...  
## $ housing : chr "no" "no" "yes" "no" ...  
## $ loan : chr "no" "no" "no" "no" ...  
## $ contact : chr "telephone" "telephone" "telephone" "telephone" ...  
## $ month : chr "may" "may" "may" "may" ...  
## $ day\_of\_week : chr "mon" "mon" "mon" "mon" ...  
## $ duration : num 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : num 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : num 5191 5191 5191 5191 5191 ...  
## $ y : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ missing : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ day : chr "mon" "mon" "mon" "mon" ...

nrow(bankDataCleaned)

## [1] 41176

ncol(bankDataCleaned)

## [1] 23

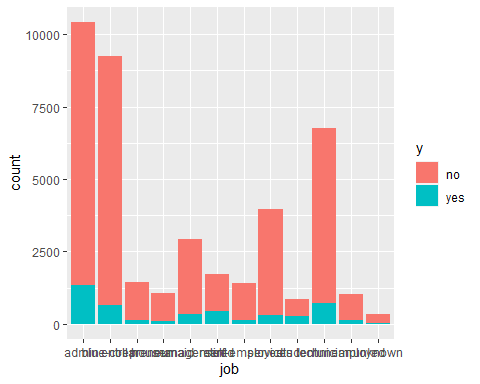
head(bankDataCleaned)

## age job marital education default housing loan contact month  
## 1 56 housemaid married basic.4y no no no telephone may  
## 2 57 services married high.school unknown no no telephone may  
## 3 37 services married high.school no yes no telephone may  
## 4 40 admin. married basic.6y no no no telephone may  
## 5 56 services married high.school no no yes telephone may  
## 6 45 services married basic.9y unknown no no telephone may  
## day\_of\_week duration campaign pdays previous poutcome emp.var.rate  
## 1 mon 261 1 999 0 nonexistent 1.1  
## 2 mon 149 1 999 0 nonexistent 1.1  
## 3 mon 226 1 999 0 nonexistent 1.1  
## 4 mon 151 1 999 0 nonexistent 1.1  
## 5 mon 307 1 999 0 nonexistent 1.1  
## 6 mon 198 1 999 0 nonexistent 1.1  
## cons.price.idx cons.conf.idx euribor3m nr.employed y missing day  
## 1 93.994 -36.4 4.857 5191 0 FALSE mon  
## 2 93.994 -36.4 4.857 5191 0 FALSE mon  
## 3 93.994 -36.4 4.857 5191 0 FALSE mon  
## 4 93.994 -36.4 4.857 5191 0 FALSE mon  
## 5 93.994 -36.4 4.857 5191 0 FALSE mon  
## 6 93.994 -36.4 4.857 5191 0 FALSE mon

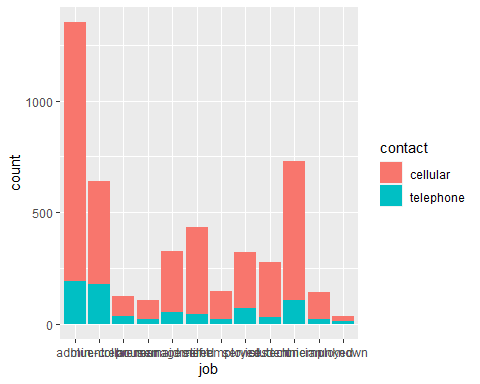
summary(bankDataCleaned)

## age job marital education   
## Min. :17.00 Length:41176 Length:41176 Length:41176   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41176 Length:41176 Length:41176 Length:41176   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41176 Length:41176 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41176 Min. :-3.40000   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character 1st Qu.:-1.80000   
## Median :999.0 Median :0.000 Mode :character Median : 1.10000   
## Mean :962.5 Mean :0.173 Mean : 0.08192   
## 3rd Qu.:999.0 3rd Qu.:0.000 3rd Qu.: 1.40000   
## Max. :999.0 Max. :7.000 Max. : 1.40000   
## cons.price.idx cons.conf.idx euribor3m nr.employed y   
## Min. :92.20 Min. :-50.8 Min. :0.634 Min. :4964 Min. :0   
## 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099 1st Qu.:0   
## Median :93.75 Median :-41.8 Median :4.857 Median :5191 Median :0   
## Mean :93.58 Mean :-40.5 Mean :3.621 Mean :5167 Mean :0   
## 3rd Qu.:93.99 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228 3rd Qu.:0   
## Max. :94.77 Max. :-26.9 Max. :5.045 Max. :5228 Max. :0   
## missing day   
## Mode :logical Length:41176   
## FALSE:41176 Class :character   
## Mode :character   
##   
##   
##

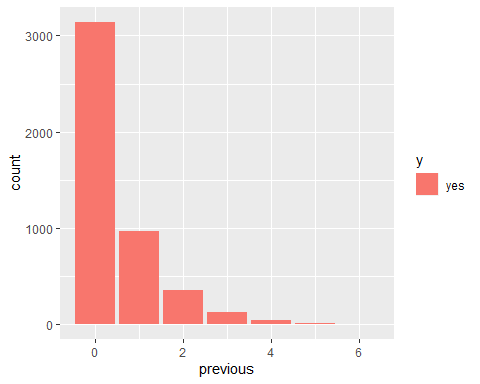
x <- filter(bankDataMain, y == "yes")  
  
# Age Distribution and Analysis  
ggplot(bankDataMain, aes(job)) + geom\_bar(aes(fill = y))



# Job Distribution and Analysis  
ggplot(x, aes(job)) + geom\_bar(aes(fill = contact))



# previous Distribution and Analysis  
ggplot(x, aes(previous)) + geom\_bar(aes(fill = y))



table(bankDataMain$poutcome, bankDataMain$y)

##   
## no yes  
## failure 3647 605  
## nonexistent 32422 3141  
## success 479 894

table(bankDataMain$contact, bankDataMain$y)

##   
## no yes  
## cellular 22291 3853  
## telephone 14257 787

table(bankDataMain$education)

##   
## basic.4y basic.6y basic.9y high.school   
## 4176 2292 6045 9515   
## illiterate professional.course university.degree unknown   
## 18 5243 12168 1731

table(bankDataMain$default)

##   
## no unknown yes   
## 32588 8597 3

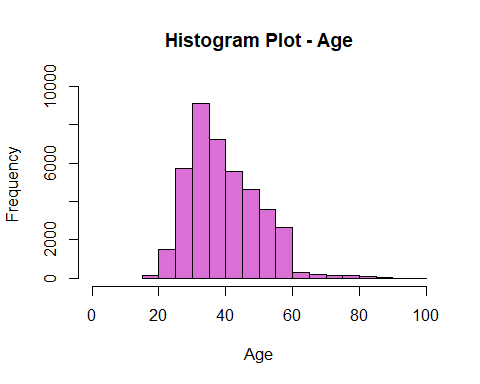
table(bankDataMain$housing)

##   
## no unknown yes   
## 18622 990 21576

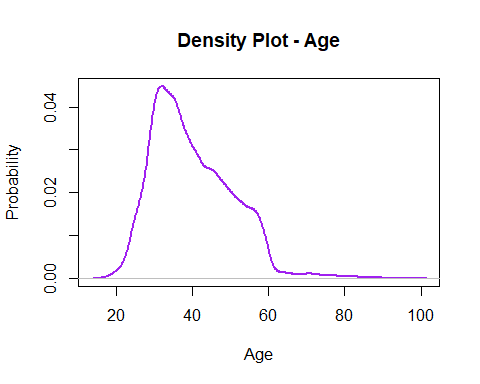
table(bankDataMain$month)

##   
## apr aug dec jul jun mar may nov oct sep   
## 2632 6178 182 7174 5318 546 13769 4101 718 570

# Age histogram  
hist(  
 bankDataMain$age,  
 main = "Histogram Plot - Age",  
 xlab = "Age",  
 ylab = "Frequency ",  
 border = "black",  
 xlim = c(0, 100),  
 ylim = c(0, 10000),  
 col = "orchid"  
)



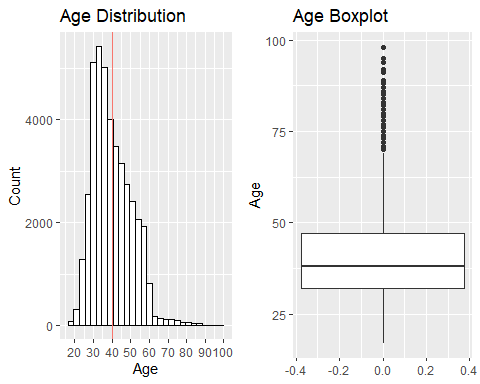
# Age Density Plot  
plot(  
 density(bankDataMain$age),  
 main = "Density Plot - Age",  
 xlab = "Age",  
 ylab = "Probability",  
 col = "purple",  
 lwd = 2.5,  
)



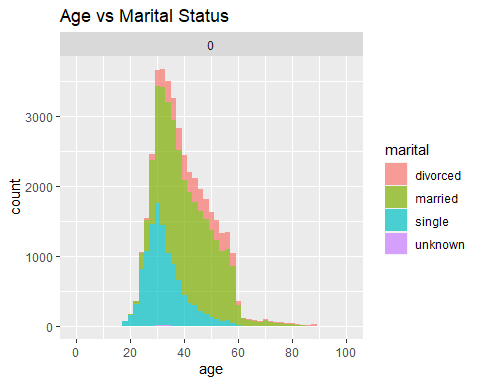
duration <- summary(bankDataMain$duration)  
duration

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 102.0 180.0 258.3 319.0 4918.0

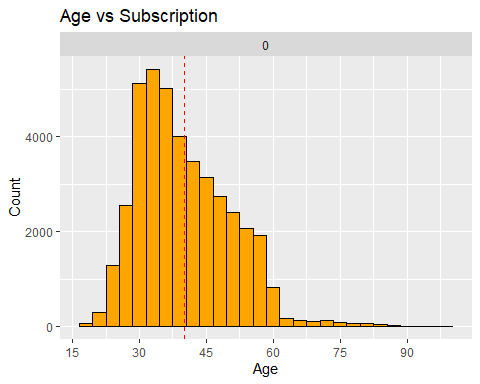
# Age ~ Marital Status Histogram  
ggPlot <- ggplot (bankDataCleaned)  
plot1 <- ggPlot + geom\_histogram(aes(x = age),  
 color = "black",  
 fill = "white",  
 binwidth = 3) +  
 ggtitle('Age Distribution') +  
 ylab('Count') +  
 xlab('Age') +  
 geom\_vline(aes(xintercept = mean(age), color = "tomato")) +  
 scale\_x\_continuous(breaks = seq(0, 100, 10)) +  
 theme(legend.position = "none")  
  
# Age ~ Marital Status Boxplot  
plot2 <- ggPlot + geom\_boxplot(aes(y = age)) +  
 ggtitle('Age Boxplot') +  
 ylab('Age')  
  
grid.arrange(plot1, plot2, ncol = 2, nrow = 1)



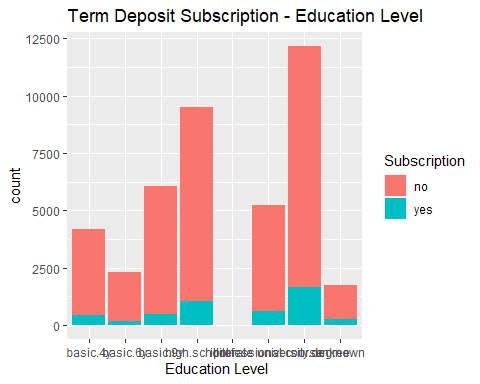
p3 <- ggplot(bankDataCleaned, aes(x = age, fill = marital)) +  
 geom\_histogram(binwidth = 2, alpha = 0.7) +  
 facet\_grid(cols = vars(y)) +  
 expand\_limits(x = c(0, 100)) +  
 scale\_x\_continuous(breaks = seq(0, 100, 20)) +  
 ggtitle("Age vs Marital Status")  
  
p3



meanAge <-  
 bankDataCleaned %>% group\_by(y) %>% summarize(grp.mean = mean(age))  
  
# Age ~ Subscription Status Histogram  
ggplot (bankDataCleaned, aes(x = age)) +  
 geom\_histogram(color = "black",  
 fill = "orange",  
 binwidth = 3) +  
 facet\_grid(cols = vars(y)) +  
 ggtitle('Age vs Subscription') + ylab('Count') + xlab('Age') +  
 scale\_x\_continuous(breaks = seq(0, 100, 15)) +  
 geom\_vline(  
 data = meanAge,  
 aes(xintercept = grp.mean),  
 color = "red",  
 linetype = "dashed"  
 )



# Education ~ Subscription Status Barplot  
ggplot(data = bankDataMain.clean, aes(x = education, fill = y)) +  
 geom\_bar() +  
 ggtitle("Term Deposit Subscription - Education Level") +  
 xlab(" Education Level") +  
 guides(fill = guide\_legend(title = "Subscription"))



bankDataMain.clean %>%  
 group\_by(education) %>%  
 summarize(pct.yes = mean(y == "yes") \* 100) %>%  
 arrange(desc(pct.yes))

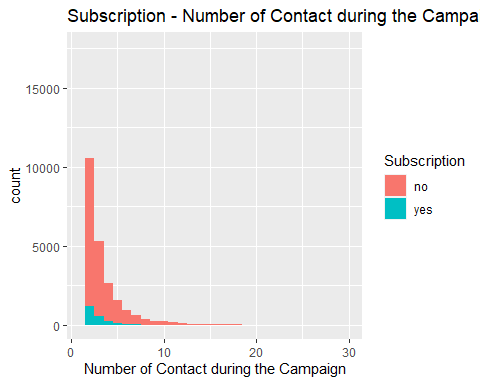
## # A tibble: 8 x 2  
## education pct.yes  
## <chr> <dbl>  
## 1 illiterate 22.2   
## 2 unknown 14.5   
## 3 university.degree 13.7   
## 4 professional.course 11.4   
## 5 high.school 10.8   
## 6 basic.4y 10.2   
## 7 basic.6y 8.21  
## 8 basic.9y 7.82

# Campaign ~ Subscription Status Histogram  
ggplot(data = bankDataMain.clean, aes(x = campaign, fill = y)) +  
 geom\_histogram() +  
 ggtitle("Subscription - Number of Contact during the Campaign") +  
 xlab("Number of Contact during the Campaign") +  
 xlim(c(min = 1, max = 30)) +  
 guides(fill = guide\_legend(title = "Subscription"))

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

## Warning: Removed 33 rows containing non-finite values (stat\_bin).

## Warning: Removed 4 rows containing missing values (geom\_bar).



bankDataMain.clean %>%  
 group\_by(campaign) %>%  
 summarize(contact.cnt = n(),  
 pct.con.yes = mean(y == "yes") \* 100) %>%  
 arrange(desc(contact.cnt)) %>%  
 head()

## # A tibble: 6 x 3  
## campaign contact.cnt pct.con.yes  
## <dbl> <int> <dbl>  
## 1 1 17634 13.0   
## 2 2 10568 11.5   
## 3 3 5340 10.7   
## 4 4 2650 9.40  
## 5 5 1599 7.50  
## 6 6 979 7.66

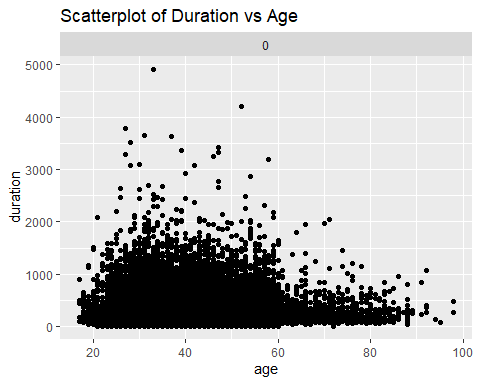
range(bankDataCleaned$duration)

## [1] 0 4918

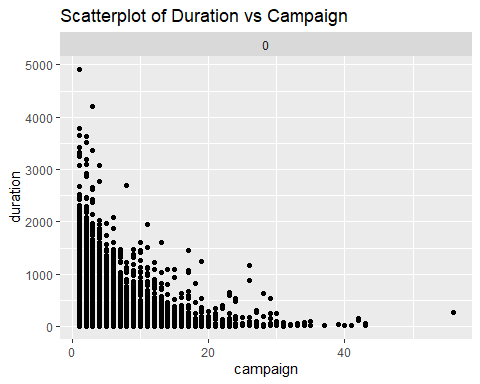
summary(bankDataCleaned$duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 102.0 180.0 258.3 319.0 4918.0

# Age ~ Duration Status Scatterplot  
ggplot(data = bankDataCleaned, aes(age, duration)) +  
 geom\_point() +  
 facet\_grid(cols = vars(y)) +  
 scale\_x\_continuous(breaks = seq(0, 100, 20)) +  
 ggtitle("Scatterplot of Duration vs Age")



# Campaign ~ Duration Status Scatterplot  
bankDataCleaned %>% filter(campaign < 63) %>%  
 ggplot(aes(campaign, duration)) +  
 geom\_point() +  
 facet\_grid(cols = vars(y)) +  
 ggtitle("Scatterplot of Duration vs Campaign")



ageTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$age)),  
 method = "pearson")  
ageTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$age))  
## t = 6.16, df = 41186, p-value = 7.342e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.02068781 0.03998505  
## sample estimates:  
## cor   
## 0.03033926

jobTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$job)),  
 method = "pearson")  
jobTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$job))  
## t = 5.1, df = 41186, p-value = 3.412e-07  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.01546842 0.03477124  
## sample estimates:  
## cor   
## 0.02512217

maritalTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$marital)),  
 method = "pearson")  
maritalTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$marital))  
## t = 9.3865, df = 41186, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.03656141 0.05583520  
## sample estimates:  
## cor   
## 0.04620261

eduTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$education)),  
 method = "pearson")  
eduTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$education))  
## t = 11.75, df = 41186, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.04816827 0.06741877  
## sample estimates:  
## cor   
## 0.05779889

housingTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$housing)),  
 method = "pearson")  
housingTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$housing))  
## t = 2.3445, df = 41186, p-value = 0.01906  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.00189439 0.02120683  
## sample estimates:  
## cor   
## 0.01155169

loanTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$loan)),  
 method = "pearson")  
loanTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$loan))  
## t = -0.99618, df = 41186, p-value = 0.3192  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.014565410 0.004749139  
## sample estimates:  
## cor   
## -0.004908593

housingLoanTermDeposit <-  
 cor.test(as.numeric(as.factor(bankDataMain$y)),  
 as.numeric(as.factor(bankDataMain$housing)) +  
 as.numeric(as.factor(bankDataMain$loan)),  
 method = "pearson")  
housingLoanTermDeposit

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(as.factor(bankDataMain$y)) and as.numeric(as.factor(bankDataMain$housing)) + as.numeric(as.factor(bankDataMain$loan))  
## t = 1.2733, df = 41186, p-value = 0.2029  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.003383843 0.015930411  
## sample estimates:  
## cor   
## 0.006273869

# Training and Testing the dataset  
set.seed(12345)  
sampleData <-  
 sample(  
 x = 1:nrow(bankDataMain),  
 size = 0.8 \* nrow(bankDataMain),  
 replace = F  
 )  
sampleData

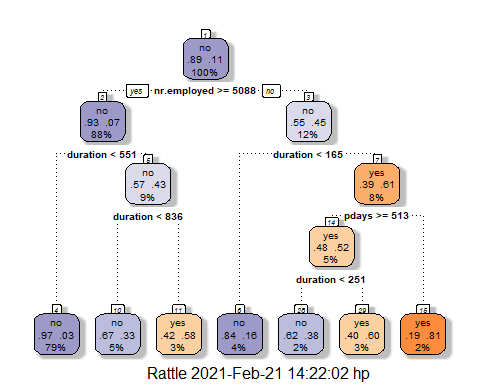
head(testData)

## age job marital education default housing loan contact  
## 4 40 admin. married basic.6y no no no telephone  
## 9 24 technician single professional.course no yes no telephone  
## 10 25 services single high.school no yes no telephone  
## 11 41 blue-collar married unknown unknown no no telephone  
## 13 29 blue-collar single high.school no no yes telephone  
## 17 35 blue-collar married basic.6y no yes no telephone  
## month day\_of\_week duration campaign pdays previous poutcome emp.var.rate  
## 4 may mon 151 1 999 0 nonexistent 1.1  
## 9 may mon 380 1 999 0 nonexistent 1.1  
## 10 may mon 50 1 999 0 nonexistent 1.1  
## 11 may mon 55 1 999 0 nonexistent 1.1  
## 13 may mon 137 1 999 0 nonexistent 1.1  
## 17 may mon 312 1 999 0 nonexistent 1.1  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 4 93.994 -36.4 4.857 5191 no  
## 9 93.994 -36.4 4.857 5191 no  
## 10 93.994 -36.4 4.857 5191 no  
## 11 93.994 -36.4 4.857 5191 no  
## 13 93.994 -36.4 4.857 5191 no  
## 17 93.994 -36.4 4.857 5191 no

sapply(bankDataMain, class)

## age job marital education default   
## "numeric" "character" "character" "character" "character"   
## housing loan contact month day\_of\_week   
## "character" "character" "character" "character" "character"   
## duration campaign pdays previous poutcome   
## "numeric" "numeric" "numeric" "numeric" "character"   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   
## "numeric" "numeric" "numeric" "numeric" "numeric"   
## y   
## "character"

bankCART <- rpart(y ~ ., trainData , method = 'class')  
  
par(mfrow = c(1, 1))  
fancyRpartPlot(bankCART ,  
 digits = 2 ,  
 palettes = c("Purples", "Oranges"))



cartPred <- predict(bankCART , testData , type = "class")  
cartProb <- predict(bankCART , testData , type = "prob")  
  
confusionMatrix(as.factor(testData$y), as.factor(cartPred))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 7020 295  
## yes 429 494  
##   
## Accuracy : 0.9121   
## 95% CI : (0.9058, 0.9181)  
## No Information Rate : 0.9042   
## P-Value [Acc > NIR] : 0.00734   
##   
## Kappa : 0.5284   
##   
## Mcnemar's Test P-Value : 7.697e-07   
##   
## Sensitivity : 0.9424   
## Specificity : 0.6261   
## Pos Pred Value : 0.9597   
## Neg Pred Value : 0.5352   
## Prevalence : 0.9042   
## Detection Rate : 0.8521   
## Detection Prevalence : 0.8880   
## Balanced Accuracy : 0.7843   
##   
## 'Positive' Class : no   
##

CrossTable(  
 testData$y,  
 cartPred,  
 prop.chisq = FALSE,  
 prop.c = FALSE,  
 prop.r = FALSE,  
 dnn = c('actual default', 'predicted default')  
)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 8238   
##   
##   
## | predicted default   
## actual default | no | yes | Row Total |   
## ---------------|-----------|-----------|-----------|  
## no | 7020 | 295 | 7315 |   
## | 0.852 | 0.036 | |   
## ---------------|-----------|-----------|-----------|  
## yes | 429 | 494 | 923 |   
## | 0.052 | 0.060 | |   
## ---------------|-----------|-----------|-----------|  
## Column Total | 7449 | 789 | 8238 |   
## ---------------|-----------|-----------|-----------|  
##   
##

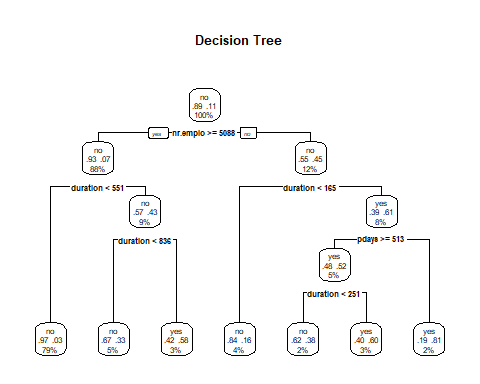
bank.knn <- train(  
 y ~ .,  
 data = trainData,  
 method = "knn",  
 maximize = TRUE,  
 trControl = trainControl(method = "cv", number = 10),  
 preProcess = c("center", "scale")  
)  
  
predictedkNN <- predict(bank.knn , newdata = testData)  
confusionMatrix(as.factor(predictedkNN) , as.factor(testData$y))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 7129 632  
## yes 186 291  
##   
## Accuracy : 0.9007   
## 95% CI : (0.894, 0.9071)  
## No Information Rate : 0.888   
## P-Value [Acc > NIR] : 0.0001041   
##   
## Kappa : 0.3674   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9746   
## Specificity : 0.3153   
## Pos Pred Value : 0.9186   
## Neg Pred Value : 0.6101   
## Prevalence : 0.8880   
## Detection Rate : 0.8654   
## Detection Prevalence : 0.9421   
## Balanced Accuracy : 0.6449   
##   
## 'Positive' Class : no   
##

### Cross table validation for KNN  
CrossTable(  
 testData$y,  
 predictedkNN,  
 prop.chisq = FALSE,  
 prop.c = FALSE,  
 prop.r = FALSE,  
 dnn = c('actual default', 'predicted default')  
)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 8238   
##   
##   
## | predicted default   
## actual default | no | yes | Row Total |   
## ---------------|-----------|-----------|-----------|  
## no | 7129 | 186 | 7315 |   
## | 0.865 | 0.023 | |   
## ---------------|-----------|-----------|-----------|  
## yes | 632 | 291 | 923 |   
## | 0.077 | 0.035 | |   
## ---------------|-----------|-----------|-----------|  
## Column Total | 7761 | 477 | 8238 |   
## ---------------|-----------|-----------|-----------|  
##   
##

# fit the decision tree classification  
decisionTree <-  
 rpart(formula = y ~ .,  
 data = trainData,  
 method = "class")  
  
# plot  
prp(  
 decisionTree,  
 type = 2,  
 extra = 104,  
 fallen.leaves = TRUE,  
 main = "Decision Tree"  
)

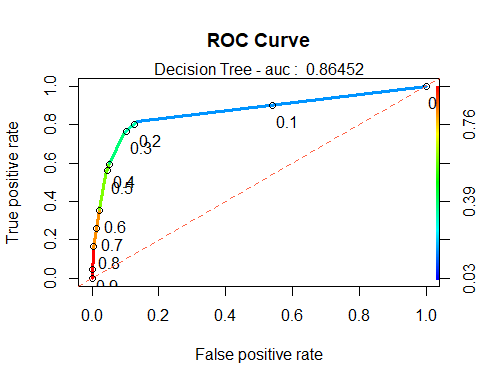


# predict test data by probability  
pred.DT <-  
 predict(decisionTree, newdata = testData[-21], type = 'prob')  
pred.DT

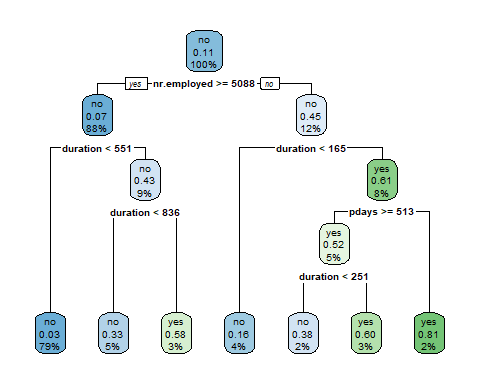
rocr.pred <-  
 prediction(predictions = pred.DT[, 2], labels = testData$y)  
rocr.perf <-  
 performance(rocr.pred, measure = "tpr", x.measure = "fpr")  
rocr.auc <- as.numeric(performance(rocr.pred, "auc")@y.values)  
  
# print ROC AUC  
rocr.auc

## [1] 0.8645178

# plot ROC curve  
plot(  
 rocr.perf,  
 lwd = 3,  
 colorize = TRUE,  
 print.cutoffs.at = seq(0, 1, by = 0.1),  
 text.adj = c(-0.2, 1.7),  
 main = 'ROC Curve'  
)  
mtext(paste('Decision Tree - auc : ', round(rocr.auc, 5)))  
abline(0, 1, col = "tomato", lty = 2)



rpart.plot(decisionTree)



pred <- predict(decisionTree, testData[-21], type = "class")  
confusionMatrix(as.factor(testData$y), as.factor(pred))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 7020 295  
## yes 429 494  
##   
## Accuracy : 0.9121   
## 95% CI : (0.9058, 0.9181)  
## No Information Rate : 0.9042   
## P-Value [Acc > NIR] : 0.00734   
##   
## Kappa : 0.5284   
##   
## Mcnemar's Test P-Value : 7.697e-07   
##   
## Sensitivity : 0.9424   
## Specificity : 0.6261   
## Pos Pred Value : 0.9597   
## Neg Pred Value : 0.5352   
## Prevalence : 0.9042   
## Detection Rate : 0.8521   
## Detection Prevalence : 0.8880   
## Balanced Accuracy : 0.7843   
##   
## 'Positive' Class : no   
##

# Logistic Regression Model  
logRegModel <-  
 glm(y ~ .,  
 family = binomial(link = "logit"),  
 data = bankDataCleaned)

## Warning: glm.fit: algorithm did not converge

logRegModel

##   
## Call: glm(formula = y ~ ., family = binomial(link = "logit"), data = bankDataCleaned)  
##   
## Coefficients:  
## (Intercept) age   
## -2.657e+01 4.538e-14   
## jobblue-collar jobentrepreneur   
## -7.872e-13 -4.600e-13   
## jobhousemaid jobmanagement   
## 1.148e-11 -3.485e-13   
## jobretired jobself-employed   
## -1.512e-12 -2.903e-13   
## jobservices jobstudent   
## -1.037e-13 6.082e-13   
## jobtechnician jobunemployed   
## -4.056e-14 -1.822e-13   
## jobunknown maritalmarried   
## -6.349e-13 7.420e-13   
## maritalsingle maritalunknown   
## 8.705e-13 1.739e-13   
## educationbasic.6y educationbasic.9y   
## -2.247e-12 -2.141e-12   
## educationhigh.school educationilliterate   
## -2.519e-12 -2.050e-12   
## educationprofessional.course educationuniversity.degree   
## -2.395e-12 -2.386e-12   
## educationunknown defaultunknown   
## -2.456e-12 -1.202e-12   
## defaultyes housingunknown   
## 1.914e-13 -8.029e-13   
## housingyes loanunknown   
## -6.502e-13 NA   
## loanyes contacttelephone   
## -3.491e-13 1.622e-13   
## monthaug monthdec   
## -7.899e-13 -1.486e-12   
## monthjul monthjun   
## -5.817e-13 -5.461e-13   
## monthmar monthmay   
## -2.669e-13 5.526e-13   
## monthnov monthoct   
## -9.521e-13 -1.249e-12   
## monthsep day\_of\_weekmon   
## -8.581e-13 1.621e-12   
## day\_of\_weekthu day\_of\_weektue   
## 1.083e-13 -1.409e-13   
## day\_of\_weekwed duration   
## -4.598e-14 1.404e-17   
## campaign pdays   
## -7.415e-14 7.324e-17   
## previous poutcomenonexistent   
## 3.624e-14 -4.560e-15   
## poutcomesuccess emp.var.rate   
## -3.429e-14 -4.242e-13   
## cons.price.idx cons.conf.idx   
## -3.112e-13 -1.384e-14   
## euribor3m nr.employed   
## 1.190e-12 -1.345e-14   
## missingTRUE daymon   
## NA NA   
## daythu daytue   
## NA NA   
## daywed   
## NA   
##   
## Degrees of Freedom: 41175 Total (i.e. Null); 41123 Residual  
## Null Deviance: 0   
## Residual Deviance: 2.389e-07 AIC: 106

summary(logRegModel)

##   
## Call:  
## glm(formula = y ~ ., family = binomial(link = "logit"), data = bankDataCleaned)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.409e-06 -2.409e-06 -2.409e-06 -2.409e-06 -2.409e-06   
##   
## Coefficients: (6 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.657e+01 4.306e+06 0 1  
## age 4.538e-14 2.189e+02 0 1  
## jobblue-collar -7.872e-13 6.539e+03 0 1  
## jobentrepreneur -4.600e-13 1.014e+04 0 1  
## jobhousemaid 1.148e-11 1.207e+04 0 1  
## jobmanagement -3.485e-13 7.634e+03 0 1  
## jobretired -1.512e-12 1.072e+04 0 1  
## jobself-employed -2.903e-13 1.018e+04 0 1  
## jobservices -1.037e-13 7.107e+03 0 1  
## jobstudent 6.082e-13 1.331e+04 0 1  
## jobtechnician -4.056e-14 6.297e+03 0 1  
## jobunemployed -1.822e-13 1.188e+04 0 1  
## jobunknown -6.349e-13 2.041e+04 0 1  
## maritalmarried 7.420e-13 5.784e+03 0 1  
## maritalsingle 8.705e-13 6.657e+03 0 1  
## maritalunknown 1.739e-13 4.027e+04 0 1  
## educationbasic.6y -2.247e-12 9.423e+03 0 1  
## educationbasic.9y -2.141e-12 7.452e+03 0 1  
## educationhigh.school -2.519e-12 7.710e+03 0 1  
## educationilliterate -2.050e-12 8.417e+04 0 1  
## educationprofessional.course -2.395e-12 8.669e+03 0 1  
## educationuniversity.degree -2.386e-12 7.868e+03 0 1  
## educationunknown -2.456e-12 1.059e+04 0 1  
## defaultunknown -1.202e-12 4.641e+03 0 1  
## defaultyes 1.914e-13 2.057e+05 0 1  
## housingunknown -8.029e-13 1.164e+04 0 1  
## housingyes -6.502e-13 3.587e+03 0 1  
## loanunknown NA NA NA NA  
## loanyes -3.491e-13 4.912e+03 0 1  
## contacttelephone 1.622e-13 6.908e+03 0 1  
## monthaug -7.899e-13 1.596e+04 0 1  
## monthdec -1.486e-12 2.930e+04 0 1  
## monthjul -5.817e-13 9.821e+03 0 1  
## monthjun -5.461e-13 1.583e+04 0 1  
## monthmar -2.669e-13 1.973e+04 0 1  
## monthmay 5.526e-13 9.260e+03 0 1  
## monthnov -9.521e-13 1.195e+04 0 1  
## monthoct -1.249e-12 1.848e+04 0 1  
## monthsep -8.581e-13 2.257e+04 0 1  
## day\_of\_weekmon 1.621e-12 5.599e+03 0 1  
## day\_of\_weekthu 1.083e-13 5.586e+03 0 1  
## day\_of\_weektue -1.409e-13 5.688e+03 0 1  
## day\_of\_weekwed -4.598e-14 5.669e+03 0 1  
## duration 1.404e-17 6.826e+00 0 1  
## campaign -7.415e-14 6.487e+02 0 1  
## pdays 7.324e-17 3.311e+01 0 1  
## previous 3.624e-14 8.646e+03 0 1  
## poutcomenonexistent -4.560e-15 1.157e+04 0 1  
## poutcomesuccess -3.429e-14 3.262e+04 0 1  
## emp.var.rate -4.242e-13 1.725e+04 0 1  
## cons.price.idx -3.112e-13 2.871e+04 0 1  
## cons.conf.idx -1.384e-14 9.902e+02 0 1  
## euribor3m 1.190e-12 1.426e+04 0 1  
## nr.employed -1.345e-14 3.435e+02 0 1  
## missingTRUE NA NA NA NA  
## daymon NA NA NA NA  
## daythu NA NA NA NA  
## daytue NA NA NA NA  
## daywed NA NA NA NA  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 0.0000e+00 on 41175 degrees of freedom  
## Residual deviance: 2.3889e-07 on 41123 degrees of freedom  
## AIC: 106  
##   
## Number of Fisher Scoring iterations: 25

## Accuracy: 0.9037907

## Precision: 0.5913163

## Recall: 0.5475431

## FScore: 0.5685885

## [1] Logistic Regression ROC Curve – AUC is 0.9078594

# Probability  
prob <-  
 (exp(logRegModel$coefficients[1])) / (1 + exp(logRegModel$coefficients[1]))  
prob

## (Intercept)   
## 2.900701e-12

# random forest  
rfModel <- train(y ~ .,  
 data = trainData,  
 method = "rf",  
 ntree = 20)  
# rpart.plot(rfModel)  
refPred <- predict(rfModel, testData)  
confusionMatrix(as.factor(testData$y), as.factor(refPred))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 7006 309  
## yes 428 495  
##   
## Accuracy : 0.9105   
## 95% CI : (0.9042, 0.9166)  
## No Information Rate : 0.9024   
## P-Value [Acc > NIR] : 0.006293   
##   
## Kappa : 0.5235   
##   
## Mcnemar's Test P-Value : 1.383e-05   
##   
## Sensitivity : 0.9424   
## Specificity : 0.6157   
## Pos Pred Value : 0.9578   
## Neg Pred Value : 0.5363   
## Prevalence : 0.9024   
## Detection Rate : 0.8504   
## Detection Prevalence : 0.8880   
## Balanced Accuracy : 0.7790   
##   
## 'Positive' Class : no