ML 1000 - Project

Ali El-Sharif, Neha Panchal, Sagnik Adusumilli, Sarmad Shubber

# Data Analysis

Source of data: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Extracted from <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>

## Data Dictionary

20 input variables and 1 output variable (desired target)

|  |  |
| --- | --- |
| Column Name | Column Description |
| age | age of the customer (numeric) |
| job | type of job (categorical: ‘admin.’,‘blue-collar’,‘entrepreneur’,‘housemaid’,‘management’,‘retired’,‘self-employed’,‘services’,‘student’,‘technician’,‘unemployed’,‘unknown’) |
| marital | marital status (categorical: ‘divorced’,‘married’,‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed) |
| education | education level of the customer (categorical: ‘basic.4y’,‘basic.6y’,‘basic.9y’,‘high.school’,‘illiterate’,‘professional.course’,‘university.degree’,‘unknown’) |
| default | has credit in default? (categorical: ‘no’,‘yes’,‘unknown’) |
| housing | has housing loan? (categorical: ‘no’,‘yes’,‘unknown’) |
| loan | has personal loan? (categorical: ‘no’,‘yes’,‘unknown’) |
| contact | contact communication type (categorical: ‘cellular’,‘telephone’) |
| month | last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’, …, ‘nov’, ‘dec’) |
| day\_of\_week | last contact day of the week (categorical: ‘mon’,‘tue’,‘wed’,‘thu’,‘fri’) |
| duration | last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=‘no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model |
| campaign | number of contacts performed during this campaign and for this client (numeric, includes last contact) |
| pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) |
| previous | number of contacts performed before this campaign and for this client (numeric) |
| poutcome | outcome of the previous marketing campaign (categorical: ‘failure’,‘nonexistent’,‘success’) |
| emp.var.rate | employment variation rate - quarterly indicator (numeric) |
| cons.price.idx | consumer price index - monthly indicator (numeric) |
| cons.conf.idx | consumer confidence index - monthly indicator (numeric) |
| euribor3m | euribor 3 month rate - daily indicator (numeric) |
| nr.employed | number of employees - quarterly indicator (numeric) |
| y | has the client subscribed a term deposit? (binary: ‘yes’,‘no’) |

## Data Exploartion

Summary:

## age job marital   
## Min. :17.00 admin. :10422 divorced: 4612   
## 1st Qu.:32.00 blue-collar: 9254 married :24928   
## Median :38.00 technician : 6743 single :11568   
## Mean :40.02 services : 3969 unknown : 80   
## 3rd Qu.:47.00 management : 2924   
## Max. :98.00 retired : 1720   
## (Other) : 6156   
## education default housing   
## university.degree :12168 no :32588 no :18622   
## high.school : 9515 unknown: 8597 unknown: 990   
## basic.9y : 6045 yes : 3 yes :21576   
## professional.course: 5243   
## basic.4y : 4176   
## basic.6y : 2292   
## (Other) : 1749   
## loan contact month day\_of\_week  
## no :33950 cellular :26144 may :13769 fri:7827   
## unknown: 990 telephone:15044 jul : 7174 mon:8514   
## yes : 6248 aug : 6178 thu:8623   
## jun : 5318 tue:8090   
## nov : 4101 wed:8134   
## apr : 2632   
## (Other): 2016   
## duration campaign pdays previous   
## Min. : 0.0 Min. : 1.000 Min. : 0.0 Min. :0.000   
## 1st Qu.: 102.0 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.000   
## Median : 180.0 Median : 2.000 Median :999.0 Median :0.000   
## Mean : 258.3 Mean : 2.568 Mean :962.5 Mean :0.173   
## 3rd Qu.: 319.0 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.000   
## Max. :4918.0 Max. :56.000 Max. :999.0 Max. :7.000   
##   
## poutcome emp.var.rate cons.price.idx cons.conf.idx   
## failure : 4252 Min. :-3.40000 Min. :92.20 Min. :-50.8   
## nonexistent:35563 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.7   
## success : 1373 Median : 1.10000 Median :93.75 Median :-41.8   
## Mean : 0.08189 Mean :93.58 Mean :-40.5   
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.4   
## Max. : 1.40000 Max. :94.77 Max. :-26.9   
##   
## euribor3m nr.employed y   
## Min. :0.634 Min. :4964 no :36548   
## 1st Qu.:1.344 1st Qu.:5099 yes: 4640   
## Median :4.857 Median :5191   
## Mean :3.621 Mean :5167   
## 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :5.045 Max. :5228   
##

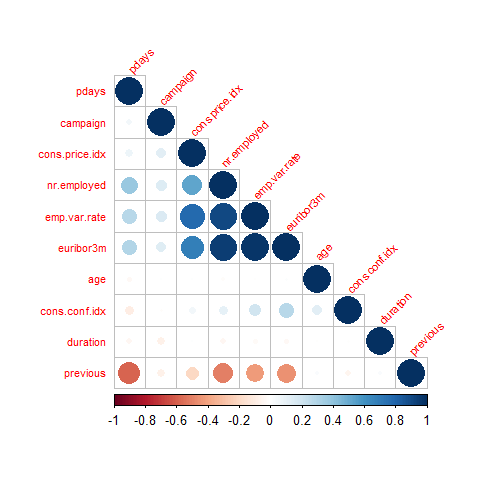
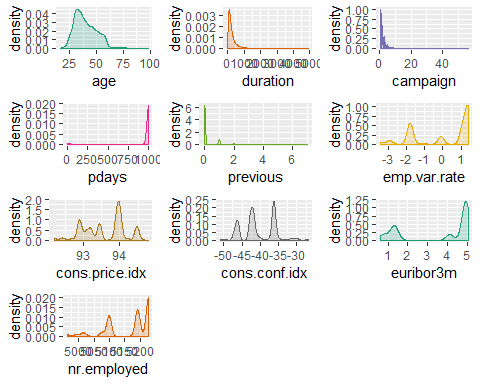
Structure:

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...  
## $ marital : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...  
## $ education : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...  
## $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...  
## $ housing : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...  
## $ loan : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...  
## $ contact : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...  
## $ month : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...  
## $ day\_of\_week : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ 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 : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ 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 : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

Missing Data:

## 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

Our dataset does not have any rows with columns that contain NA, so no rows will be omitted and no data imputing is necessary.

Correlation matrix:  Distribution of important features: 

## Data Preparation and Unsupervised Learning (Clustering)

#Encode columns in order to perform clustering on categorical values   
bankData1he <- bankData  
#install.packages(ade4)  
#One-hot-encoding of categorical features:  
library(ade4)

## Warning: package 'ade4' was built under R version 3.5.3

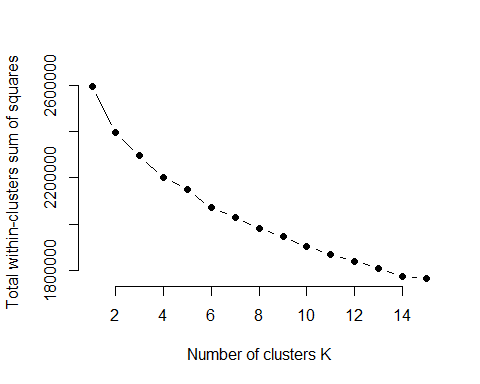
categorical\_columns = c('job', 'marital', 'education', 'default', 'housing',   
 'loan', 'contact', 'month', 'day\_of\_week', 'poutcome')  
for (f in categorical\_columns){  
 df\_all\_dummy = acm.disjonctif(bankData1he[f])  
 bankData1he[f] = NULL  
 bankData1he = cbind(bankData1he, df\_all\_dummy)  
}  
  
str(bankData1he)

## 'data.frame': 41188 obs. of 64 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ 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 ...  
## $ 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 : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ job.admin. : num 0 0 0 1 0 0 1 0 0 0 ...  
## $ job.blue-collar : num 0 0 0 0 0 0 0 1 0 0 ...  
## $ job.entrepreneur : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ job.housemaid : num 1 0 0 0 0 0 0 0 0 0 ...  
## $ job.management : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ job.retired : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ job.self-employed : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ job.services : num 0 1 1 0 1 1 0 0 0 1 ...  
## $ job.student : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ job.technician : num 0 0 0 0 0 0 0 0 1 0 ...  
## $ job.unemployed : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ job.unknown : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ marital.divorced : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ marital.married : num 1 1 1 1 1 1 1 1 0 0 ...  
## $ marital.single : num 0 0 0 0 0 0 0 0 1 1 ...  
## $ marital.unknown : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ education.basic.4y : num 1 0 0 0 0 0 0 0 0 0 ...  
## $ education.basic.6y : num 0 0 0 1 0 0 0 0 0 0 ...  
## $ education.basic.9y : num 0 0 0 0 0 1 0 0 0 0 ...  
## $ education.high.school : num 0 1 1 0 1 0 0 0 0 1 ...  
## $ education.illiterate : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ education.professional.course: num 0 0 0 0 0 0 1 0 1 0 ...  
## $ education.university.degree : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ education.unknown : num 0 0 0 0 0 0 0 1 0 0 ...  
## $ default.no : num 1 0 1 1 1 0 1 0 1 1 ...  
## $ default.unknown : num 0 1 0 0 0 1 0 1 0 0 ...  
## $ default.yes : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ housing.no : num 1 1 0 1 1 1 1 1 0 0 ...  
## $ housing.unknown : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ housing.yes : num 0 0 1 0 0 0 0 0 1 1 ...  
## $ loan.no : num 1 1 1 1 0 1 1 1 1 1 ...  
## $ loan.unknown : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ loan.yes : num 0 0 0 0 1 0 0 0 0 0 ...  
## $ contact.cellular : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ contact.telephone : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ month.apr : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.aug : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.dec : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.jul : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.jun : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.mar : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.may : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ month.nov : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.oct : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ month.sep : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ day\_of\_week.fri : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ day\_of\_week.mon : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ day\_of\_week.thu : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ day\_of\_week.tue : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ day\_of\_week.wed : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome.failure : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome.nonexistent : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ poutcome.success : num 0 0 0 0 0 0 0 0 0 0 ...

head(bankData1he)

## age duration campaign pdays previous emp.var.rate cons.price.idx  
## 1 56 261 1 999 0 1.1 93.994  
## 2 57 149 1 999 0 1.1 93.994  
## 3 37 226 1 999 0 1.1 93.994  
## 4 40 151 1 999 0 1.1 93.994  
## 5 56 307 1 999 0 1.1 93.994  
## 6 45 198 1 999 0 1.1 93.994  
## cons.conf.idx euribor3m nr.employed y job.admin. job.blue-collar  
## 1 -36.4 4.857 5191 no 0 0  
## 2 -36.4 4.857 5191 no 0 0  
## 3 -36.4 4.857 5191 no 0 0  
## 4 -36.4 4.857 5191 no 1 0  
## 5 -36.4 4.857 5191 no 0 0  
## 6 -36.4 4.857 5191 no 0 0  
## job.entrepreneur job.housemaid job.management job.retired  
## 1 0 1 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0  
## job.self-employed job.services job.student job.technician job.unemployed  
## 1 0 0 0 0 0  
## 2 0 1 0 0 0  
## 3 0 1 0 0 0  
## 4 0 0 0 0 0  
## 5 0 1 0 0 0  
## 6 0 1 0 0 0  
## job.unknown marital.divorced marital.married marital.single  
## 1 0 0 1 0  
## 2 0 0 1 0  
## 3 0 0 1 0  
## 4 0 0 1 0  
## 5 0 0 1 0  
## 6 0 0 1 0  
## marital.unknown education.basic.4y education.basic.6y education.basic.9y  
## 1 0 1 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 1 0  
## 5 0 0 0 0  
## 6 0 0 0 1  
## education.high.school education.illiterate education.professional.course  
## 1 0 0 0  
## 2 1 0 0  
## 3 1 0 0  
## 4 0 0 0  
## 5 1 0 0  
## 6 0 0 0  
## education.university.degree education.unknown default.no default.unknown  
## 1 0 0 1 0  
## 2 0 0 0 1  
## 3 0 0 1 0  
## 4 0 0 1 0  
## 5 0 0 1 0  
## 6 0 0 0 1  
## default.yes housing.no housing.unknown housing.yes loan.no loan.unknown  
## 1 0 1 0 0 1 0  
## 2 0 1 0 0 1 0  
## 3 0 0 0 1 1 0  
## 4 0 1 0 0 1 0  
## 5 0 1 0 0 0 0  
## 6 0 1 0 0 1 0  
## loan.yes contact.cellular contact.telephone month.apr month.aug  
## 1 0 0 1 0 0  
## 2 0 0 1 0 0  
## 3 0 0 1 0 0  
## 4 0 0 1 0 0  
## 5 1 0 1 0 0  
## 6 0 0 1 0 0  
## month.dec month.jul month.jun month.mar month.may month.nov month.oct  
## 1 0 0 0 0 1 0 0  
## 2 0 0 0 0 1 0 0  
## 3 0 0 0 0 1 0 0  
## 4 0 0 0 0 1 0 0  
## 5 0 0 0 0 1 0 0  
## 6 0 0 0 0 1 0 0  
## month.sep day\_of\_week.fri day\_of\_week.mon day\_of\_week.thu  
## 1 0 0 1 0  
## 2 0 0 1 0  
## 3 0 0 1 0  
## 4 0 0 1 0  
## 5 0 0 1 0  
## 6 0 0 1 0  
## day\_of\_week.tue day\_of\_week.wed poutcome.failure poutcome.nonexistent  
## 1 0 0 0 1  
## 2 0 0 0 1  
## 3 0 0 0 1  
## 4 0 0 0 1  
## 5 0 0 0 1  
## 6 0 0 0 1  
## poutcome.success  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

# Remove y column (column 11) before scaling and clustering  
#scale the variables  
scaled\_bd <- scale(bankData1he[,-c(11)])  
  
#Elbow Method for finding the optimal number of clusters  
set.seed(123)  
# Compute and plot wss for k = 2 to k = 15.  
k.max <- 15  
wss <- sapply(1:k.max,   
 function(k){kmeans(scaled\_bd, k, nstart=50,iter.max = 15 )$tot.withinss})  
  
plot(1:k.max, wss,  
 type="b", pch = 19, frame = FALSE,   
 xlab="Number of clusters K",  
 ylab="Total within-clusters sum of squares")



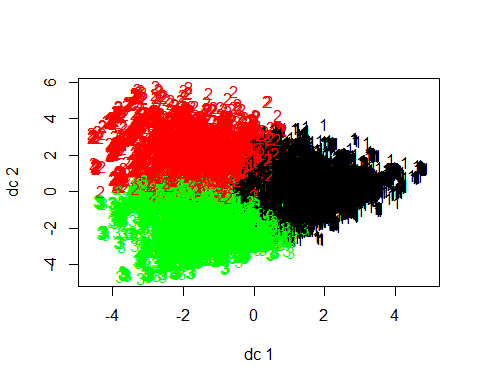
# k = 3, from the above plot we picked the elbow to be k=3 meaning 3 clusters  
library(cluster)  
clarax <- clara(scaled\_bd, 3, metric = "jaccard", stand = TRUE,   
 samples = 5, pamLike = TRUE)  
  
clarax

## Call: clara(x = scaled\_bd, k = 3, metric = "jaccard", stand = TRUE, samples = 5, pamLike = TRUE)   
## Medoids:  
## age duration campaign pdays previous emp.var.rate  
## 4152 -0.8659288 0.211027261 -0.204906 0.1954115 -0.34949 0.6480844  
## 21730 -0.4820977 -0.151516214 -0.204906 0.1954115 -0.34949 0.8390505  
## 13128 -1.4416755 0.002757605 0.156103 0.1954115 -0.34949 0.8390505  
## cons.price.idx cons.conf.idx euribor3m nr.employed job.admin.  
## 4152 0.7227137 0.8864358 0.7130278 0.3316759 1.7181248  
## 21730 -0.2274624 0.9512558 0.7735658 0.8451598 -0.5820158  
## 13128 0.5914166 -0.4747852 0.7729892 0.8451598 -0.5820158  
## job.blue-collar job.entrepreneur job.housemaid job.management  
## 4152 -0.5383105 -0.1914279 -0.1625264 -0.2764319  
## 21730 -0.5383105 -0.1914279 -0.1625264 -0.2764319  
## 13128 -0.5383105 -0.1914279 -0.1625264 -0.2764319  
## job.retired job.self-employed job.services job.student  
## 4152 -0.2087548 -0.1890298 -0.3265524 -0.1473249  
## 21730 -0.2087548 -0.1890298 -0.3265524 -0.1473249  
## 13128 -0.2087548 -0.1890298 -0.3265524 -0.1473249  
## job.technician job.unemployed job.unknown marital.divorced  
## 4152 -0.4424439 -0.1588697 -0.08986967 -0.3550923  
## 21730 2.2601187 -0.1588697 -0.08986967 -0.3550923  
## 13128 2.2601187 -0.1588697 -0.08986967 -0.3550923  
## marital.married marital.single marital.unknown education.basic.4y  
## 4152 0.8076278 -0.62493 -0.04411401 -0.3358951  
## 21730 0.8076278 -0.62493 -0.04411401 -0.3358951  
## 13128 -1.2381640 1.60014 -0.04411401 -0.3358951  
## education.basic.6y education.basic.9y education.high.school  
## 4152 -0.2427446 -0.4147377 1.8244625  
## 21730 -0.2427446 -0.4147377 -0.5480933  
## 13128 -0.2427446 -0.4147377 -0.5480933  
## education.illiterate education.professional.course  
## 4152 -0.02090935 -0.3819139  
## 21730 -0.02090935 2.6183280  
## 13128 -0.02090935 -0.3819139  
## education.university.degree education.unknown default.no  
## 4152 -0.6475236 -0.2094504 0.5137065  
## 21730 -0.6475236 -0.2094504 0.5137065  
## 13128 1.5443077 -0.2094504 0.5137065  
## default.unknown default.yes housing.no housing.unknown housing.yes  
## 4152 -0.5135933 -0.008534652 1.1008011 -0.1569315 -1.0488642  
## 21730 -0.5135933 -0.008534652 1.1008011 -0.1569315 -1.0488642  
## 13128 -0.5135933 -0.008534652 -0.9084073 -0.1569315 0.9533891  
## loan.no loan.unknown loan.yes contact.cellular contact.telephone  
## 4152 0.4617258 -0.1569315 -0.422867 -1.3182540 1.3182540  
## 21730 0.4617258 -0.1569315 -0.422867 0.7585608 -0.7585608  
## 13128 0.4617258 -0.1569315 -0.422867 0.7585608 -0.7585608  
## month.apr month.aug month.dec month.jul month.jun month.mar  
## 4152 -0.2612713 -0.4200709 -0.06662032 -0.4592472 -0.3850377 -0.1159054  
## 21730 -0.2612713 2.3804926 -0.06662032 -0.4592472 -0.3850377 -0.1159054  
## 13128 -0.2612713 -0.4200709 -0.06662032 2.1774234 -0.3850377 -0.1159054  
## month.may month.nov month.oct month.sep day\_of\_week.fri  
## 4152 1.411138 -0.3325284 -0.1331957 -0.1184603 -0.4843651  
## 21730 -0.708631 -0.3325284 -0.1331957 -0.1184603 -0.4843651  
## 13128 -0.708631 -0.3325284 -0.1331957 -0.1184603 -0.4843651  
## day\_of\_week.mon day\_of\_week.thu day\_of\_week.tue day\_of\_week.wed  
## 4152 1.9589757 -0.5145746 -0.4943882 -0.4960607  
## 21730 -0.5104585 -0.5145746 2.0226528 -0.4960607  
## 13128 -0.5104585 -0.5145746 -0.4943882 2.0158335  
## poutcome.failure poutcome.nonexistent poutcome.success  
## 4152 -0.3392864 0.3977011 -0.1856977  
## 21730 -0.3392864 0.3977011 -0.1856977  
## 13128 -0.3392864 0.3977011 -0.1856977  
## Objective function: 0.6940972  
## Clustering vector: Named int [1:41188] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 ...  
## - attr(\*, "names")= chr [1:41188] "1" "2" "3" "4" "5" "6" "7" ...  
## Cluster sizes: 19144 9339 12705   
## Best sample:  
## [1] 465 1660 3876 4152 5494 6104 6958 7195 7348 7964 8594   
## [12] 8900 9027 9071 13128 13939 14484 14877 14902 16499 17983 20459  
## [23] 20464 21359 21730 22591 24381 25020 26097 26263 27718 29677 29777  
## [34] 30342 30462 30789 31188 31577 33783 34729 34911 35863 36833 38241  
## [45] 38294 40270  
##   
## Available components:  
## [1] "sample" "medoids" "i.med" "clustering" "objective"   
## [6] "clusinfo" "diss" "call" "silinfo" "data"

# cluster sizes: 19144, 9339, 12705  
  
bankDataWithCluster <- cbind(bankData, cluster = clarax$clustering)  
#head(bankDataWithCluster, n = 4)  
  
# Medoids  
clarax$medoids

## age duration campaign pdays previous emp.var.rate  
## 4152 -0.8659288 0.211027261 -0.204906 0.1954115 -0.34949 0.6480844  
## 21730 -0.4820977 -0.151516214 -0.204906 0.1954115 -0.34949 0.8390505  
## 13128 -1.4416755 0.002757605 0.156103 0.1954115 -0.34949 0.8390505  
## cons.price.idx cons.conf.idx euribor3m nr.employed job.admin.  
## 4152 0.7227137 0.8864358 0.7130278 0.3316759 1.7181248  
## 21730 -0.2274624 0.9512558 0.7735658 0.8451598 -0.5820158  
## 13128 0.5914166 -0.4747852 0.7729892 0.8451598 -0.5820158  
## job.blue-collar job.entrepreneur job.housemaid job.management  
## 4152 -0.5383105 -0.1914279 -0.1625264 -0.2764319  
## 21730 -0.5383105 -0.1914279 -0.1625264 -0.2764319  
## 13128 -0.5383105 -0.1914279 -0.1625264 -0.2764319  
## job.retired job.self-employed job.services job.student  
## 4152 -0.2087548 -0.1890298 -0.3265524 -0.1473249  
## 21730 -0.2087548 -0.1890298 -0.3265524 -0.1473249  
## 13128 -0.2087548 -0.1890298 -0.3265524 -0.1473249  
## job.technician job.unemployed job.unknown marital.divorced  
## 4152 -0.4424439 -0.1588697 -0.08986967 -0.3550923  
## 21730 2.2601187 -0.1588697 -0.08986967 -0.3550923  
## 13128 2.2601187 -0.1588697 -0.08986967 -0.3550923  
## marital.married marital.single marital.unknown education.basic.4y  
## 4152 0.8076278 -0.62493 -0.04411401 -0.3358951  
## 21730 0.8076278 -0.62493 -0.04411401 -0.3358951  
## 13128 -1.2381640 1.60014 -0.04411401 -0.3358951  
## education.basic.6y education.basic.9y education.high.school  
## 4152 -0.2427446 -0.4147377 1.8244625  
## 21730 -0.2427446 -0.4147377 -0.5480933  
## 13128 -0.2427446 -0.4147377 -0.5480933  
## education.illiterate education.professional.course  
## 4152 -0.02090935 -0.3819139  
## 21730 -0.02090935 2.6183280  
## 13128 -0.02090935 -0.3819139  
## education.university.degree education.unknown default.no  
## 4152 -0.6475236 -0.2094504 0.5137065  
## 21730 -0.6475236 -0.2094504 0.5137065  
## 13128 1.5443077 -0.2094504 0.5137065  
## default.unknown default.yes housing.no housing.unknown housing.yes  
## 4152 -0.5135933 -0.008534652 1.1008011 -0.1569315 -1.0488642  
## 21730 -0.5135933 -0.008534652 1.1008011 -0.1569315 -1.0488642  
## 13128 -0.5135933 -0.008534652 -0.9084073 -0.1569315 0.9533891  
## loan.no loan.unknown loan.yes contact.cellular contact.telephone  
## 4152 0.4617258 -0.1569315 -0.422867 -1.3182540 1.3182540  
## 21730 0.4617258 -0.1569315 -0.422867 0.7585608 -0.7585608  
## 13128 0.4617258 -0.1569315 -0.422867 0.7585608 -0.7585608  
## month.apr month.aug month.dec month.jul month.jun month.mar  
## 4152 -0.2612713 -0.4200709 -0.06662032 -0.4592472 -0.3850377 -0.1159054  
## 21730 -0.2612713 2.3804926 -0.06662032 -0.4592472 -0.3850377 -0.1159054  
## 13128 -0.2612713 -0.4200709 -0.06662032 2.1774234 -0.3850377 -0.1159054  
## month.may month.nov month.oct month.sep day\_of\_week.fri  
## 4152 1.411138 -0.3325284 -0.1331957 -0.1184603 -0.4843651  
## 21730 -0.708631 -0.3325284 -0.1331957 -0.1184603 -0.4843651  
## 13128 -0.708631 -0.3325284 -0.1331957 -0.1184603 -0.4843651  
## day\_of\_week.mon day\_of\_week.thu day\_of\_week.tue day\_of\_week.wed  
## 4152 1.9589757 -0.5145746 -0.4943882 -0.4960607  
## 21730 -0.5104585 -0.5145746 2.0226528 -0.4960607  
## 13128 -0.5104585 -0.5145746 -0.4943882 2.0158335  
## poutcome.failure poutcome.nonexistent poutcome.success  
## 4152 -0.3392864 0.3977011 -0.1856977  
## 21730 -0.3392864 0.3977011 -0.1856977  
## 13128 -0.3392864 0.3977011 -0.1856977

# The mediods are 4152, 21730, 13128  
  
# plot the cluster solution  
library(fpc)  
plotcluster(scaled\_bd, clarax$clustering)



# display data around the mediods for each cluster  
bankDataWithCluster[4140:4160,]

## age job marital education default housing loan  
## 4140 52 management married basic.4y unknown no no  
## 4141 35 admin. single university.degree no no no  
## 4142 33 blue-collar married basic.9y no unknown unknown  
## 4143 46 blue-collar married basic.9y no yes no  
## 4144 26 services married high.school no no no  
## 4145 39 services married high.school unknown no no  
## 4146 47 admin. married university.degree no no no  
## 4147 26 blue-collar married basic.4y no no yes  
## 4148 34 blue-collar married professional.course no yes no  
## 4149 24 admin. single high.school no yes no  
## 4150 31 technician married professional.course no no no  
## 4151 26 admin. single high.school no no yes  
## 4152 31 admin. married high.school no no no  
## 4153 34 blue-collar married basic.6y no yes no  
## 4154 60 admin. married professional.course no yes no  
## 4155 32 blue-collar married basic.6y unknown yes no  
## 4156 24 blue-collar married basic.9y no yes no  
## 4157 28 admin. married high.school no no no  
## 4158 55 entrepreneur divorced university.degree no yes yes  
## 4159 38 admin. divorced high.school no no no  
## 4160 37 blue-collar married professional.course no yes no  
## contact month day\_of\_week duration campaign pdays previous  
## 4140 telephone may mon 88 35 999 0  
## 4141 telephone may mon 234 13 999 0  
## 4142 telephone may mon 215 3 999 0  
## 4143 telephone may mon 194 5 999 0  
## 4144 telephone may mon 394 2 999 0  
## 4145 telephone may mon 5 2 999 0  
## 4146 telephone may mon 408 2 999 0  
## 4147 telephone may mon 114 2 999 0  
## 4148 telephone may mon 210 2 999 0  
## 4149 telephone may mon 243 6 999 0  
## 4150 telephone may mon 180 6 999 0  
## 4151 telephone may mon 8 2 999 0  
## 4152 telephone may mon 313 2 999 0  
## 4153 telephone may mon 1622 2 999 0  
## 4154 telephone may mon 324 2 999 0  
## 4155 telephone may mon 205 3 999 0  
## 4156 telephone may mon 194 2 999 0  
## 4157 telephone may mon 165 5 999 0  
## 4158 telephone may mon 86 7 999 0  
## 4159 telephone may mon 160 2 999 0  
## 4160 telephone may mon 492 6 999 0  
## poutcome emp.var.rate cons.price.idx cons.conf.idx euribor3m  
## 4140 nonexistent 1.1 93.994 -36.4 4.858  
## 4141 nonexistent 1.1 93.994 -36.4 4.858  
## 4142 nonexistent 1.1 93.994 -36.4 4.858  
## 4143 nonexistent 1.1 93.994 -36.4 4.858  
## 4144 nonexistent 1.1 93.994 -36.4 4.858  
## 4145 nonexistent 1.1 93.994 -36.4 4.858  
## 4146 nonexistent 1.1 93.994 -36.4 4.858  
## 4147 nonexistent 1.1 93.994 -36.4 4.858  
## 4148 nonexistent 1.1 93.994 -36.4 4.858  
## 4149 nonexistent 1.1 93.994 -36.4 4.858  
## 4150 nonexistent 1.1 93.994 -36.4 4.858  
## 4151 nonexistent 1.1 93.994 -36.4 4.858  
## 4152 nonexistent 1.1 93.994 -36.4 4.858  
## 4153 nonexistent 1.1 93.994 -36.4 4.858  
## 4154 nonexistent 1.1 93.994 -36.4 4.858  
## 4155 nonexistent 1.1 93.994 -36.4 4.858  
## 4156 nonexistent 1.1 93.994 -36.4 4.858  
## 4157 nonexistent 1.1 93.994 -36.4 4.858  
## 4158 nonexistent 1.1 93.994 -36.4 4.858  
## 4159 nonexistent 1.1 93.994 -36.4 4.858  
## 4160 nonexistent 1.1 93.994 -36.4 4.858  
## nr.employed y cluster  
## 4140 5191 no 1  
## 4141 5191 no 1  
## 4142 5191 no 1  
## 4143 5191 no 1  
## 4144 5191 no 1  
## 4145 5191 no 1  
## 4146 5191 no 1  
## 4147 5191 no 1  
## 4148 5191 no 1  
## 4149 5191 no 1  
## 4150 5191 no 1  
## 4151 5191 no 1  
## 4152 5191 no 1  
## 4153 5191 yes 1  
## 4154 5191 no 1  
## 4155 5191 no 1  
## 4156 5191 no 1  
## 4157 5191 no 1  
## 4158 5191 no 1  
## 4159 5191 no 1  
## 4160 5191 no 1

bankDataWithCluster[21720:21740,]

## age job marital education default housing loan  
## 21720 36 technician single high.school no yes yes  
## 21721 29 technician single university.degree unknown no yes  
## 21722 52 admin. married high.school no yes no  
## 21723 34 admin. single university.degree no no no  
## 21724 47 blue-collar married basic.6y no yes no  
## 21725 31 admin. married university.degree unknown no no  
## 21726 56 blue-collar married basic.4y unknown yes no  
## 21727 43 management married university.degree no no no  
## 21728 41 technician single professional.course unknown yes no  
## 21729 39 admin. divorced university.degree no no no  
## 21730 35 technician married professional.course no no no  
## 21731 30 admin. single university.degree no yes no  
## 21732 29 admin. single university.degree no yes no  
## 21733 45 technician divorced professional.course no no no  
## 21734 30 admin. single university.degree no yes no  
## 21735 52 admin. divorced unknown no no no  
## 21736 40 technician divorced university.degree no no no  
## 21737 32 admin. married university.degree no unknown unknown  
## 21738 47 blue-collar married basic.6y no no no  
## 21739 36 technician married university.degree no yes no  
## 21740 55 housemaid married basic.4y no yes no  
## contact month day\_of\_week duration campaign pdays previous  
## 21720 cellular aug tue 629 7 999 0  
## 21721 cellular aug tue 137 2 999 0  
## 21722 cellular aug tue 231 4 999 0  
## 21723 cellular aug tue 119 6 999 0  
## 21724 cellular aug tue 308 2 999 0  
## 21725 cellular aug tue 609 4 999 0  
## 21726 cellular aug tue 147 3 999 0  
## 21727 cellular aug tue 123 3 999 0  
## 21728 cellular aug tue 57 5 999 0  
## 21729 cellular aug tue 199 4 999 0  
## 21730 cellular aug tue 219 2 999 0  
## 21731 cellular aug tue 386 3 999 0  
## 21732 cellular aug tue 112 5 999 0  
## 21733 cellular aug tue 706 3 999 0  
## 21734 cellular aug tue 173 9 999 0  
## 21735 cellular aug tue 228 5 999 0  
## 21736 cellular aug tue 207 3 999 0  
## 21737 cellular aug tue 347 2 999 0  
## 21738 cellular aug tue 256 2 999 0  
## 21739 cellular aug tue 85 4 999 0  
## 21740 cellular aug tue 117 3 999 0  
## poutcome emp.var.rate cons.price.idx cons.conf.idx euribor3m  
## 21720 nonexistent 1.4 93.444 -36.1 4.963  
## 21721 nonexistent 1.4 93.444 -36.1 4.963  
## 21722 nonexistent 1.4 93.444 -36.1 4.963  
## 21723 nonexistent 1.4 93.444 -36.1 4.963  
## 21724 nonexistent 1.4 93.444 -36.1 4.963  
## 21725 nonexistent 1.4 93.444 -36.1 4.963  
## 21726 nonexistent 1.4 93.444 -36.1 4.963  
## 21727 nonexistent 1.4 93.444 -36.1 4.963  
## 21728 nonexistent 1.4 93.444 -36.1 4.963  
## 21729 nonexistent 1.4 93.444 -36.1 4.963  
## 21730 nonexistent 1.4 93.444 -36.1 4.963  
## 21731 nonexistent 1.4 93.444 -36.1 4.963  
## 21732 nonexistent 1.4 93.444 -36.1 4.963  
## 21733 nonexistent 1.4 93.444 -36.1 4.963  
## 21734 nonexistent 1.4 93.444 -36.1 4.963  
## 21735 nonexistent 1.4 93.444 -36.1 4.963  
## 21736 nonexistent 1.4 93.444 -36.1 4.963  
## 21737 nonexistent 1.4 93.444 -36.1 4.963  
## 21738 nonexistent 1.4 93.444 -36.1 4.963  
## 21739 nonexistent 1.4 93.444 -36.1 4.963  
## 21740 nonexistent 1.4 93.444 -36.1 4.963  
## nr.employed y cluster  
## 21720 5228.1 no 2  
## 21721 5228.1 no 2  
## 21722 5228.1 no 2  
## 21723 5228.1 no 2  
## 21724 5228.1 no 2  
## 21725 5228.1 no 2  
## 21726 5228.1 no 2  
## 21727 5228.1 no 2  
## 21728 5228.1 no 2  
## 21729 5228.1 no 2  
## 21730 5228.1 no 2  
## 21731 5228.1 no 3  
## 21732 5228.1 no 2  
## 21733 5228.1 no 2  
## 21734 5228.1 no 2  
## 21735 5228.1 no 2  
## 21736 5228.1 no 2  
## 21737 5228.1 no 2  
## 21738 5228.1 no 2  
## 21739 5228.1 no 2  
## 21740 5228.1 no 2

bankDataWithCluster[13120:13140,]

## age job marital education default housing loan  
## 13120 37 management single university.degree no yes no  
## 13121 29 admin. single high.school no yes no  
## 13122 35 management married university.degree no no no  
## 13123 40 blue-collar divorced basic.9y no yes no  
## 13124 25 admin. married high.school no no no  
## 13125 57 retired married basic.4y unknown yes yes  
## 13126 32 services single high.school unknown yes no  
## 13127 34 admin. single university.degree unknown yes no  
## 13128 25 technician single university.degree no yes no  
## 13129 25 blue-collar married basic.9y no yes no  
## 13130 53 admin. married university.degree unknown no no  
## 13131 53 admin. married university.degree no yes yes  
## 13132 31 services married basic.9y no yes no  
## 13133 40 technician married professional.course no yes yes  
## 13134 57 blue-collar married basic.6y no yes no  
## 13135 31 services married basic.9y no no no  
## 13136 29 blue-collar married high.school no yes no  
## 13137 30 technician single university.degree no yes yes  
## 13138 35 self-employed single university.degree no yes no  
## 13139 53 admin. married university.degree no yes no  
## 13140 43 housemaid married basic.4y no yes no  
## contact month day\_of\_week duration campaign pdays previous  
## 13120 cellular jul wed 158 1 999 0  
## 13121 cellular jul wed 1272 1 999 0  
## 13122 cellular jul wed 182 1 999 0  
## 13123 cellular jul wed 207 3 999 0  
## 13124 cellular jul wed 95 1 999 0  
## 13125 cellular jul wed 253 1 999 0  
## 13126 cellular jul wed 72 1 999 0  
## 13127 cellular jul wed 132 1 999 0  
## 13128 cellular jul wed 259 3 999 0  
## 13129 cellular jul wed 261 3 999 0  
## 13130 cellular jul wed 815 1 999 0  
## 13131 cellular jul wed 205 1 999 0  
## 13132 cellular jul wed 537 1 999 0  
## 13133 cellular jul wed 323 1 999 0  
## 13134 cellular jul wed 64 1 999 0  
## 13135 cellular jul wed 679 1 999 0  
## 13136 telephone jul wed 42 1 999 0  
## 13137 cellular jul wed 63 1 999 0  
## 13138 cellular jul wed 2122 1 999 0  
## 13139 cellular jul wed 618 1 999 0  
## 13140 cellular jul wed 183 1 999 0  
## poutcome emp.var.rate cons.price.idx cons.conf.idx euribor3m  
## 13120 nonexistent 1.4 93.918 -42.7 4.962  
## 13121 nonexistent 1.4 93.918 -42.7 4.962  
## 13122 nonexistent 1.4 93.918 -42.7 4.962  
## 13123 nonexistent 1.4 93.918 -42.7 4.962  
## 13124 nonexistent 1.4 93.918 -42.7 4.962  
## 13125 nonexistent 1.4 93.918 -42.7 4.962  
## 13126 nonexistent 1.4 93.918 -42.7 4.962  
## 13127 nonexistent 1.4 93.918 -42.7 4.962  
## 13128 nonexistent 1.4 93.918 -42.7 4.962  
## 13129 nonexistent 1.4 93.918 -42.7 4.962  
## 13130 nonexistent 1.4 93.918 -42.7 4.962  
## 13131 nonexistent 1.4 93.918 -42.7 4.962  
## 13132 nonexistent 1.4 93.918 -42.7 4.962  
## 13133 nonexistent 1.4 93.918 -42.7 4.962  
## 13134 nonexistent 1.4 93.918 -42.7 4.962  
## 13135 nonexistent 1.4 93.918 -42.7 4.962  
## 13136 nonexistent 1.4 93.918 -42.7 4.962  
## 13137 nonexistent 1.4 93.918 -42.7 4.962  
## 13138 nonexistent 1.4 93.918 -42.7 4.962  
## 13139 nonexistent 1.4 93.918 -42.7 4.962  
## 13140 nonexistent 1.4 93.918 -42.7 4.962  
## nr.employed y cluster  
## 13120 5228.1 no 3  
## 13121 5228.1 no 3  
## 13122 5228.1 no 3  
## 13123 5228.1 no 3  
## 13124 5228.1 no 1  
## 13125 5228.1 no 3  
## 13126 5228.1 no 3  
## 13127 5228.1 no 3  
## 13128 5228.1 no 3  
## 13129 5228.1 no 3  
## 13130 5228.1 yes 3  
## 13131 5228.1 no 3  
## 13132 5228.1 yes 3  
## 13133 5228.1 no 3  
## 13134 5228.1 no 3  
## 13135 5228.1 yes 3  
## 13136 5228.1 no 3  
## 13137 5228.1 no 3  
## 13138 5228.1 yes 3  
## 13139 5228.1 no 3  
## 13140 5228.1 no 3

cluster1 <- bankDataWithCluster[bankDataWithCluster$cluster==1,]  
cluster2 <- bankDataWithCluster[bankDataWithCluster$cluster==2,]  
cluster3 <- bankDataWithCluster[bankDataWithCluster$cluster==3,]  
  
#Summary by cluster  
summary(cluster1)

## age job marital   
## Min. :18.0 admin. :6070 divorced: 2170   
## 1st Qu.:33.0 blue-collar :5242 married :13262   
## Median :39.0 services :2503 single : 3672   
## Mean :40.3 management :1209 unknown : 40   
## 3rd Qu.:47.0 technician : 829   
## Max. :92.0 entrepreneur: 709   
## (Other) :2582   
## education default housing loan   
## high.school :6821 no :14430 no :10525 no :15815   
## university.degree:4073 unknown: 4714 unknown: 498 unknown: 498   
## basic.9y :3274 yes : 0 yes : 8121 yes : 2831   
## basic.4y :2186   
## basic.6y :1297   
## unknown : 815   
## (Other) : 678   
## contact month day\_of\_week duration   
## cellular : 6925 may :10567 fri:3965 Min. : 0.0   
## telephone:12219 jun : 3714 mon:5750 1st Qu.: 103.0   
## nov : 1285 thu:3692 Median : 185.0   
## jul : 1194 tue:2893 Mean : 261.8   
## apr : 1013 wed:2844 3rd Qu.: 327.0   
## aug : 627 Max. :3785.0   
## (Other): 744   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.0 Min. :0.0000 failure : 1814   
## 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.0000 nonexistent:16858   
## Median : 2.000 Median :999.0 Median :0.0000 success : 472   
## Mean : 2.572 Mean :971.7 Mean :0.1487   
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.0000   
## Max. :56.000 Max. :999.0 Max. :7.0000   
##   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m   
## Min. :-3.4000 Min. :92.20 Min. :-50.80 Min. :0.635   
## 1st Qu.:-1.8000 1st Qu.:93.08 1st Qu.:-42.70 1st Qu.:1.344   
## Median : 1.1000 Median :93.99 Median :-41.80 Median :4.857   
## Mean : 0.1476 Mean :93.71 Mean :-40.29 Mean :3.689   
## 3rd Qu.: 1.4000 3rd Qu.:93.99 3rd Qu.:-36.40 3rd Qu.:4.864   
## Max. : 1.4000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr.employed y cluster   
## Min. :4964 no :17419 Min. :1   
## 1st Qu.:5099 yes: 1725 1st Qu.:1   
## Median :5191 Median :1   
## Mean :5165 Mean :1   
## 3rd Qu.:5228 3rd Qu.:1   
## Max. :5228 Max. :1   
##

summary(cluster2)

## age job marital   
## Min. :17.00 technician :3579 divorced: 976   
## 1st Qu.:34.00 blue-collar:1410 married :6993   
## Median :41.00 admin. :1310 single :1355   
## Mean :42.85 retired : 646 unknown : 15   
## 3rd Qu.:50.00 management : 623   
## Max. :98.00 services : 468   
## (Other) :1303   
## education default housing loan   
## professional.course:3695 no :7590 no :4947 no :7750   
## university.degree :2343 unknown:1746 unknown: 228 unknown: 228   
## basic.4y : 920 yes : 3 yes :4164 yes :1361   
## basic.9y : 877   
## high.school : 835   
## basic.6y : 351   
## (Other) : 318   
## contact month day\_of\_week duration   
## cellular :8215 aug :4684 fri:1638 Min. : 0.0   
## telephone:1124 may :1294 mon:1103 1st Qu.: 99.0   
## nov :1139 thu:1903 Median : 165.0   
## jun : 598 tue:3555 Mean : 240.1   
## jul : 532 wed:1140 3rd Qu.: 288.0   
## apr : 516 Max. :4199.0   
## (Other): 576   
## campaign pdays previous poutcome   
## Min. : 1.00 Min. : 0.0 Min. :0.0000 failure :1009   
## 1st Qu.: 1.00 1st Qu.:999.0 1st Qu.:0.0000 nonexistent:7928   
## Median : 2.00 Median :999.0 Median :0.0000 success : 402   
## Mean : 2.41 Mean :952.1 Mean :0.1933   
## 3rd Qu.: 3.00 3rd Qu.:999.0 3rd Qu.:0.0000   
## Max. :40.00 Max. :999.0 Max. :6.0000   
##   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m   
## Min. :-3.40000 Min. :92.20 Min. :-50.80 Min. :0.635   
## 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.00 1st Qu.:1.344   
## Median : 1.40000 Median :93.44 Median :-36.10 Median :4.960   
## Mean : 0.04309 Mean :93.33 Mean :-38.49 Mean :3.624   
## 3rd Qu.: 1.40000 3rd Qu.:93.44 3rd Qu.:-36.10 3rd Qu.:4.964   
## Max. : 1.40000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr.employed y cluster   
## Min. :4964 no :8131 Min. :2   
## 1st Qu.:5099 yes:1208 1st Qu.:2   
## Median :5228 Median :2   
## Mean :5171 Mean :2   
## 3rd Qu.:5228 3rd Qu.:2   
## Max. :5228 Max. :2   
##

summary(cluster3)

## age job marital   
## Min. :17.00 admin. :3042 divorced:1466   
## 1st Qu.:30.00 blue-collar :2602 married :4673   
## Median :35.00 technician :2335 single :6541   
## Mean :37.53 management :1092 unknown : 25   
## 3rd Qu.:44.00 services : 998   
## Max. :91.00 self-employed: 535   
## (Other) :2101   
## education default housing loan   
## university.degree :5752 no :10568 no :3150 no :10385   
## basic.9y :1894 unknown: 2137 unknown: 264 unknown: 264   
## high.school :1859 yes : 0 yes :9291 yes : 2056   
## basic.4y :1070   
## professional.course: 873   
## basic.6y : 644   
## (Other) : 613   
## contact month day\_of\_week duration   
## cellular :11004 jul :5448 fri:2224 Min. : 0.0   
## telephone: 1701 may :1908 mon:1661 1st Qu.: 104.0   
## nov :1677 thu:3028 Median : 183.0   
## apr :1103 tue:1642 Mean : 266.4   
## jun :1006 wed:4150 3rd Qu.: 330.0   
## aug : 867 Max. :4918.0   
## (Other): 696   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.0 Min. :0.0000 failure : 1429   
## 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.0000 nonexistent:10777   
## Median : 2.000 Median :999.0 Median :0.0000 success : 499   
## Mean : 2.677 Mean :956.2 Mean :0.1946   
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.0000   
## Max. :43.000 Max. :999.0 Max. :6.0000   
##   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m   
## Min. :-3.4000 Min. :92.20 Min. :-50.8 Min. :0.634   
## 1st Qu.:-1.8000 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.334   
## Median : 1.4000 Median :93.92 Median :-42.7 Median :4.864   
## Mean : 0.0114 Mean :93.54 Mean :-42.3 Mean :3.517   
## 3rd Qu.: 1.4000 3rd Qu.:93.92 3rd Qu.:-42.0 3rd Qu.:4.962   
## Max. : 1.4000 Max. :94.77 Max. :-26.9 Max. :4.970   
##   
## nr.employed y cluster   
## Min. :4964 no :10998 Min. :3   
## 1st Qu.:5099 yes: 1707 1st Qu.:3   
## Median :5228 Median :3   
## Mean :5168 Mean :3   
## 3rd Qu.:5228 3rd Qu.:3   
## Max. :5228 Max. :3   
##

## Clustering Analysis

To learn more about each cluster, we will run a Random Forest and get the important features (columns) of the cluster. We will also run a Decision Tree to know which are the most important features (columns).

######################################## Random Forest on cluster1 - most important features  
#19144 rows  
  
# Set a random seed  
set.seed(754)  
  
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

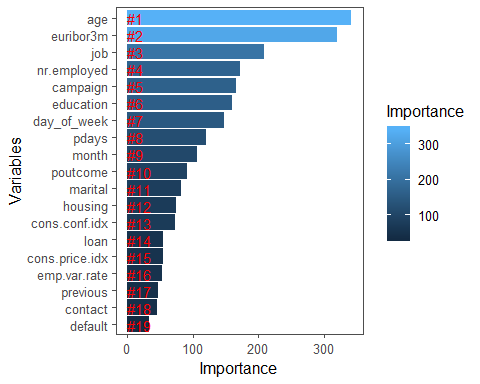
##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

# Build the model (note: not all possible variables are used)  
# remove duration column as per website recommendation  
rf\_model <- randomForest(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = cluster1)  
  
# Get importance  
importance <- importance(rf\_model)  
varImportance <- data.frame(Variables = row.names(importance),   
 Importance = round(importance[ ,'MeanDecreaseGini'],2))  
  
# Create a rank variable based on importance  
rankImportance <- varImportance %>%  
 mutate(Rank = paste0('#',dense\_rank(desc(Importance))))  
  
library(ggthemes)  
# Use ggplot2 to visualize the relative importance of variables  
ggplot(rankImportance, aes(x = reorder(Variables, Importance),   
 y = Importance, fill = Importance)) +  
 geom\_bar(stat='identity') +   
 geom\_text(aes(x = Variables, y = 0.5, label = Rank),  
 hjust=0, vjust=0.55, size = 4, colour = 'red') +  
 labs(x = 'Variables') +  
 coord\_flip() +   
 theme\_few()

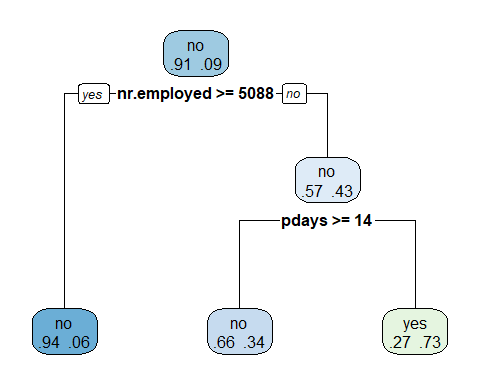


#Top 2 features: age, euribor3m  
#3rd feature: job  
#4th feature: nr.employed

##################################### Decision Tree on cluster1 - most important features  
#19144 rows  
  
set.seed(1984)  
  
library(rpart)  
dt\_model <- rpart(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = cluster1,  
 method = "class",  
 minsplit = 2,  
 minbucket = 1)  
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.5.3

rpart.plot(dt\_model, extra=4) # plot tree

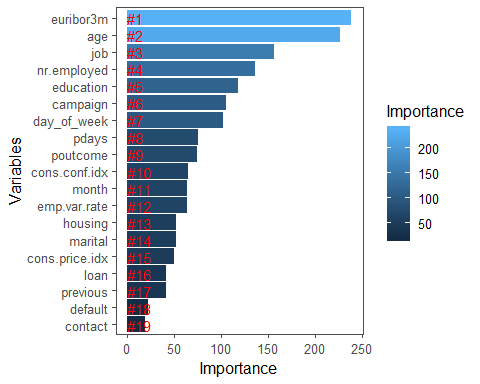


summary(dt\_model)

## Call:  
## rpart(formula = y ~ age + job + marital + education + default +   
## housing + loan + contact + month + day\_of\_week + campaign +   
## pdays + previous + poutcome + emp.var.rate + cons.price.idx +   
## cons.conf.idx + euribor3m + nr.employed, data = cluster1,   
## method = "class", minsplit = 2, minbucket = 1)  
## n= 19144   
##   
## CP nsplit rel error xerror xstd  
## 1 0.04985507 0 1.0000000 1.0000000 0.02296681  
## 2 0.01000000 2 0.9002899 0.9014493 0.02191191  
##   
## Variable importance  
## nr.employed euribor3m emp.var.rate cons.conf.idx cons.price.idx   
## 25 22 13 12 9   
## month pdays poutcome previous   
## 7 5 5 1   
##   
## Node number 1: 19144 observations, complexity param=0.04985507  
## predicted class=no expected loss=0.09010656 P(node) =1  
## class counts: 17419 1725  
## probabilities: 0.910 0.090   
## left son=2 (17466 obs) right son=3 (1678 obs)  
## Primary splits:  
## nr.employed < 5087.65 to the right, improve=416.7432, (0 missing)  
## euribor3m < 1.2395 to the right, improve=379.2457, (0 missing)  
## pdays < 512 to the right, improve=323.3952, (0 missing)  
## poutcome splits as LLR, improve=301.5721, (0 missing)  
## emp.var.rate < -0.15 to the right, improve=188.1518, (0 missing)  
## Surrogate splits:  
## euribor3m < 1.2395 to the right, agree=0.989, adj=0.874, (0 split)  
## emp.var.rate < -2.35 to the right, agree=0.957, adj=0.505, (0 split)  
## cons.conf.idx < -35.45 to the left, agree=0.954, adj=0.476, (0 split)  
## cons.price.idx < 92.7345 to the right, agree=0.945, adj=0.367, (0 split)  
## month splits as LLRLLRLLRR, agree=0.938, adj=0.297, (0 split)  
##   
## Node number 2: 17466 observations  
## predicted class=no expected loss=0.05776938 P(node) =0.9123485  
## class counts: 16457 1009  
## probabilities: 0.942 0.058   
##   
## Node number 3: 1678 observations, complexity param=0.04985507  
## predicted class=no expected loss=0.4266985 P(node) =0.08765148  
## class counts: 962 716  
## probabilities: 0.573 0.427   
## left son=6 (1300 obs) right son=7 (378 obs)  
## Primary splits:  
## pdays < 13.5 to the right, improve=88.30163, (0 missing)  
## poutcome splits as LLR, improve=84.15303, (0 missing)  
## contact splits as RL, improve=24.54211, (0 missing)  
## previous < 0.5 to the left, improve=21.91195, (0 missing)  
## euribor3m < 0.698 to the right, improve=14.14577, (0 missing)  
## Surrogate splits:  
## poutcome splits as LLR, agree=0.971, adj=0.870, (0 split)  
## previous < 1.5 to the left, agree=0.835, adj=0.267, (0 split)  
## age < 19.5 to the right, agree=0.775, adj=0.003, (0 split)  
##   
## Node number 6: 1300 observations  
## predicted class=no expected loss=0.3392308 P(node) =0.06790639  
## class counts: 859 441  
## probabilities: 0.661 0.339   
##   
## Node number 7: 378 observations  
## predicted class=yes expected loss=0.2724868 P(node) =0.01974509  
## class counts: 103 275  
## probabilities: 0.272 0.728

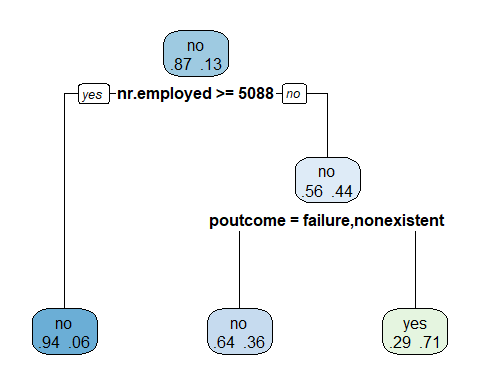
#variable importance: nr.employed, euribor3m, emp.var.rate, cons.conf.idx, cons.price.idx, month, pdays, poutcome, previous

######################################## Random Forest on cluster2 - most important features  
#9339 rows  
  
# Set a random seed  
set.seed(454)  
  
#library(randomForest)  
# Build the model (note: not all possible variables are used)  
# remove duration column as per website recommendation  
rf\_model <- randomForest(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = cluster2)  
  
# Get importance  
importance <- importance(rf\_model)  
varImportance <- data.frame(Variables = row.names(importance),   
 Importance = round(importance[ ,'MeanDecreaseGini'],2))  
  
# Create a rank variable based on importance  
rankImportance <- varImportance %>%  
 mutate(Rank = paste0('#',dense\_rank(desc(Importance))))  
  
#library(ggthemes)  
# Use ggplot2 to visualize the relative importance of variables  
ggplot(rankImportance, aes(x = reorder(Variables, Importance),   
 y = Importance, fill = Importance)) +  
 geom\_bar(stat='identity') +   
 geom\_text(aes(x = Variables, y = 0.5, label = Rank),  
 hjust=0, vjust=0.55, size = 4, colour = 'red') +  
 labs(x = 'Variables') +  
 coord\_flip() +   
 theme\_few()



#Top 2 features: euribor3m, age  
#3rd feature: job  
#4th feature: nr.employed

##################################### Decision Tree on cluster2 - most important features  
#9339 rows  
  
set.seed(1484)  
  
#library(rpart)  
dt\_model <- rpart(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = cluster2,  
 method = "class",  
 minsplit = 2,  
 minbucket = 1)  
#library(rpart.plot)  
rpart.plot(dt\_model, extra=4) # plot tree

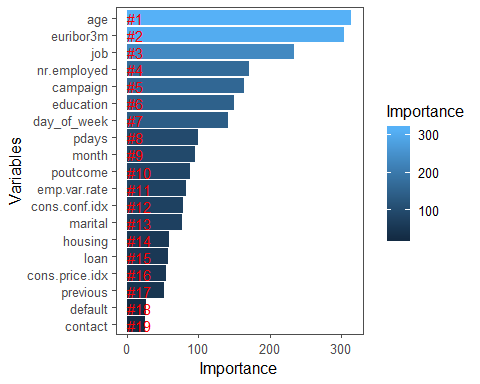


summary(dt\_model)

## Call:  
## rpart(formula = y ~ age + job + marital + education + default +   
## housing + loan + contact + month + day\_of\_week + campaign +   
## pdays + previous + poutcome + emp.var.rate + cons.price.idx +   
## cons.conf.idx + euribor3m + nr.employed, data = cluster2,   
## method = "class", minsplit = 2, minbucket = 1)  
## n= 9339   
##   
## CP nsplit rel error xerror xstd  
## 1 0.06043046 0 1.0000000 1.0000000 0.02684654  
## 2 0.01000000 2 0.8791391 0.8791391 0.02539695  
##   
## Variable importance  
## nr.employed euribor3m emp.var.rate cons.conf.idx cons.price.idx   
## 22 20 16 15 14   
## month poutcome pdays   
## 6 4 3   
##   
## Node number 1: 9339 observations, complexity param=0.06043046  
## predicted class=no expected loss=0.12935 P(node) =1  
## class counts: 8131 1208  
## probabilities: 0.871 0.129   
## left son=2 (7702 obs) right son=3 (1637 obs)  
## Primary splits:  
## nr.employed < 5087.65 to the right, improve=378.1792, (0 missing)  
## euribor3m < 1.265 to the right, improve=359.3009, (0 missing)  
## emp.var.rate < -0.65 to the right, improve=260.3180, (0 missing)  
## pdays < 513 to the right, improve=247.3454, (0 missing)  
## poutcome splits as LLR, improve=240.3226, (0 missing)  
## Surrogate splits:  
## euribor3m < 1.263 to the right, agree=0.988, adj=0.930, (0 split)  
## emp.var.rate < -2.35 to the right, agree=0.954, adj=0.737, (0 split)  
## cons.conf.idx < -35.45 to the left, agree=0.946, adj=0.690, (0 split)  
## cons.price.idx < 92.7345 to the right, agree=0.933, adj=0.617, (0 split)  
## month splits as LLRLLRLLRR, agree=0.870, adj=0.260, (0 split)  
##   
## Node number 2: 7702 observations  
## predicted class=no expected loss=0.06374968 P(node) =0.8247136  
## class counts: 7211 491  
## probabilities: 0.936 0.064   
##   
## Node number 3: 1637 observations, complexity param=0.06043046  
## predicted class=no expected loss=0.4379963 P(node) =0.1752864  
## class counts: 920 717  
## probabilities: 0.562 0.438   
## left son=6 (1281 obs) right son=7 (356 obs)  
## Primary splits:  
## poutcome splits as LLR, improve=64.89283, (0 missing)  
## pdays < 513 to the right, improve=60.88021, (0 missing)  
## nr.employed < 5049.85 to the right, improve=24.69366, (0 missing)  
## euribor3m < 0.738 to the right, improve=17.91087, (0 missing)  
## emp.var.rate < -2.35 to the left, improve=17.17858, (0 missing)  
## Surrogate splits:  
## pdays < 513 to the right, agree=0.977, adj=0.893, (0 split)  
## previous < 2.5 to the left, agree=0.797, adj=0.065, (0 split)  
## age < 90.5 to the left, agree=0.783, adj=0.003, (0 split)  
##   
## Node number 6: 1281 observations  
## predicted class=no expected loss=0.3637783 P(node) =0.1371667  
## class counts: 815 466  
## probabilities: 0.636 0.364   
##   
## Node number 7: 356 observations  
## predicted class=yes expected loss=0.2949438 P(node) =0.03811971  
## class counts: 105 251  
## probabilities: 0.295 0.705

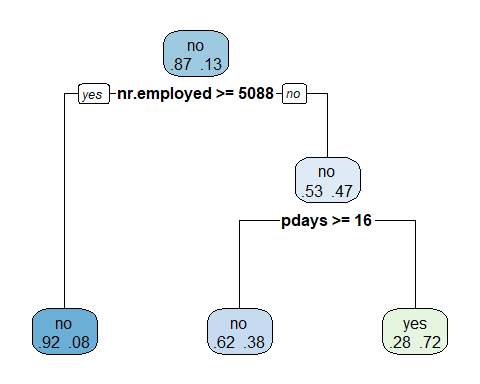
#variable importance: nr.employed, euribor3m, emp.var.rate, cons.conf.idx, cons.price.idx, month, poutcome, pdays

######################################## Random Forest on cluster3 - most important features  
#12705 rows  
  
# Set a random seed  
set.seed(760)  
  
#library(randomForest)  
# Build the model (note: not all possible variables are used)  
# remove duration column as per website recommendation  
rf\_model <- randomForest(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = cluster3)  
  
# Get importance  
importance <- importance(rf\_model)  
varImportance <- data.frame(Variables = row.names(importance),   
 Importance = round(importance[ ,'MeanDecreaseGini'],2))  
  
# Create a rank variable based on importance  
rankImportance <- varImportance %>%  
 mutate(Rank = paste0('#',dense\_rank(desc(Importance))))  
  
#library(ggthemes)  
# Use ggplot2 to visualize the relative importance of variables  
ggplot(rankImportance, aes(x = reorder(Variables, Importance),   
 y = Importance, fill = Importance)) +  
 geom\_bar(stat='identity') +   
 geom\_text(aes(x = Variables, y = 0.5, label = Rank),  
 hjust=0, vjust=0.55, size = 4, colour = 'red') +  
 labs(x = 'Variables') +  
 coord\_flip() +   
 theme\_few()



#Top 2 features: age, euribor3m  
#3rd feature: job  
#4th feature: nr.employed

##################################### Decision Tree on cluster3 - most important features  
#12705 rows  
  
set.seed(1254)  
  
#library(rpart)  
dt\_model <- rpart(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = cluster3,  
 method = "class",  
 minsplit = 2,  
 minbucket = 1)  
#library(rpart.plot)  
rpart.plot(dt\_model, extra=4) # plot tree



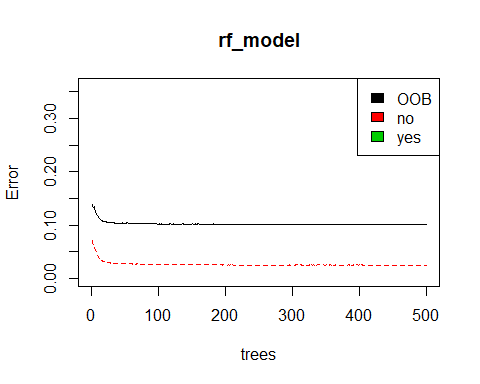
summary(dt\_model)

## Call:  
## rpart(formula = y ~ age + job + marital + education + default +   
## housing + loan + contact + month + day\_of\_week + campaign +   
## pdays + previous + poutcome + emp.var.rate + cons.price.idx +   
## cons.conf.idx + euribor3m + nr.employed, data = cluster3,   
## method = "class", minsplit = 2, minbucket = 1)  
## n= 12705   
##   
## CP nsplit rel error xerror xstd  
## 1 0.0559461 0 1.0000000 1.0000000 0.02251919  
## 2 0.0100000 2 0.8881078 0.8881078 0.02140545  
##   
## Variable importance  
## nr.employed euribor3m emp.var.rate cons.conf.idx cons.price.idx   
## 27 22 14 11 9   
## month pdays poutcome previous   
## 7 5 4 1   
##   
## Node number 1: 12705 observations, complexity param=0.0559461  
## predicted class=no expected loss=0.1343566 P(node) =1  
## class counts: 10998 1707  
## probabilities: 0.866 0.134   
## left son=2 (11056 obs) right son=3 (1649 obs)  
## Primary splits:  
## nr.employed < 5087.65 to the right, improve=428.4544, (0 missing)  
## euribor3m < 1.2395 to the right, improve=383.1642, (0 missing)  
## pdays < 510.5 to the right, improve=285.0412, (0 missing)  
## emp.var.rate < -0.6 to the right, improve=273.9247, (0 missing)  
## poutcome splits as LLR, improve=269.0593, (0 missing)  
## Surrogate splits:  
## euribor3m < 1.2395 to the right, agree=0.979, adj=0.835, (0 split)  
## emp.var.rate < -2.35 to the right, agree=0.937, adj=0.517, (0 split)  
## cons.conf.idx < -35.45 to the left, agree=0.925, adj=0.426, (0 split)  
## cons.price.idx < 92.778 to the right, agree=0.915, adj=0.343, (0 split)  
## month splits as LLRLLLLLRR, agree=0.905, adj=0.270, (0 split)  
##   
## Node number 2: 11056 observations  
## predicted class=no expected loss=0.08420767 P(node) =0.8702086  
## class counts: 10125 931  
## probabilities: 0.916 0.084   
##   
## Node number 3: 1649 observations, complexity param=0.0559461  
## predicted class=no expected loss=0.4705882 P(node) =0.1297914  
## class counts: 873 776  
## probabilities: 0.529 0.471   
## left son=6 (1216 obs) right son=7 (433 obs)  
## Primary splits:  
## pdays < 15.5 to the right, improve=73.37819, (0 missing)  
## poutcome splits as LLR, improve=67.91298, (0 missing)  
## emp.var.rate < -2.35 to the left, improve=18.92002, (0 missing)  
## cons.price.idx < 93.166 to the left, improve=18.92002, (0 missing)  
## nr.employed < 5013.1 to the right, improve=18.92002, (0 missing)  
## Surrogate splits:  
## poutcome splits as LLR, agree=0.971, adj=0.891, (0 split)  
## previous < 1.5 to the left, agree=0.802, adj=0.247, (0 split)  
## age < 17.5 to the right, agree=0.738, adj=0.002, (0 split)  
##   
## Node number 6: 1216 observations  
## predicted class=no expected loss=0.3815789 P(node) =0.09571035  
## class counts: 752 464  
## probabilities: 0.618 0.382   
##   
## Node number 7: 433 observations  
## predicted class=yes expected loss=0.2794457 P(node) =0.03408107  
## class counts: 121 312  
## probabilities: 0.279 0.721

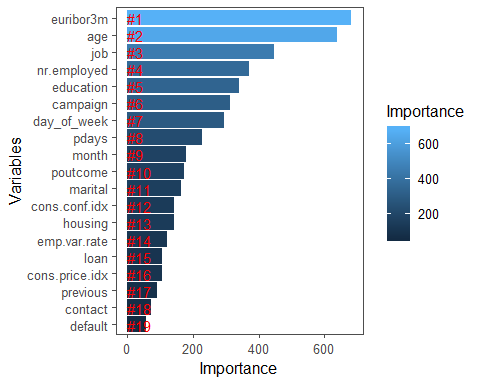
#variable importance: nr.employed, euribor3m, emp.var.rate, cons.conf.idx, cons.price.idx, month, pdays, poutcome, previous

## Supervised Learning (Prediction Models)

########################### Supervised Learning - prediction models  
  
######################################## Random Forest  
row\_count <- nrow(bankDataWithCluster)  
shuffled\_rows <- sample(row\_count)  
train <- bankDataWithCluster[head(shuffled\_rows,floor(row\_count\*0.75)),]  
test <- bankDataWithCluster[tail(shuffled\_rows,floor(row\_count\*0.25)),]  
  
# Set a random seed  
set.seed(754)  
  
#library(randomForest)  
# Build the model (note: not all possible variables are used)  
# remove duration column as per website recommendation  
rf\_model <- randomForest(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = train)  
  
# Show model error  
plot(rf\_model, ylim=c(0,0.36))  
legend('topright', colnames(rf\_model$err.rate), col=1:3, fill=1:3)



# Get importance  
importance <- importance(rf\_model)  
varImportance <- data.frame(Variables = row.names(importance),   
 Importance = round(importance[ ,'MeanDecreaseGini'],2))  
  
# Create a rank variable based on importance  
rankImportance <- varImportance %>%  
 mutate(Rank = paste0('#',dense\_rank(desc(Importance))))  
  
#library(ggthemes)  
# Use ggplot2 to visualize the relative importance of variables  
ggplot(rankImportance, aes(x = reorder(Variables, Importance),   
 y = Importance, fill = Importance)) +  
 geom\_bar(stat='identity') +   
 geom\_text(aes(x = Variables, y = 0.5, label = Rank),  
 hjust=0, vjust=0.55, size = 4, colour = 'red') +  
 labs(x = 'Variables') +  
 coord\_flip() +   
 theme\_few()



# Predict using the test set  
prediction <- predict(rf\_model, test)  
  
solution <- data.frame(test, subscribed = prediction)  
#solution  
  
library(caret)

## Loading required package: lattice

confusionMatrix(factor(solution$subscribed), solution$y)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 8879 833  
## yes 242 343  
##   
## Accuracy : 0.8956   
## 95% CI : (0.8895, 0.9014)  
## No Information Rate : 0.8858   
## P-Value [Acc > NIR] : 0.0008166   
##   
## Kappa : 0.3394   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9735   
## Specificity : 0.2917   
## Pos Pred Value : 0.9142   
## Neg Pred Value : 0.5863   
## Prevalence : 0.8858   
## Detection Rate : 0.8623   
## Detection Prevalence : 0.9432   
## Balanced Accuracy : 0.6326   
##   
## 'Positive' Class : no   
##

#Balanced Accuracy: 0.6326  
  
library(pROC)

## Type 'citation("pROC")' for a citation.

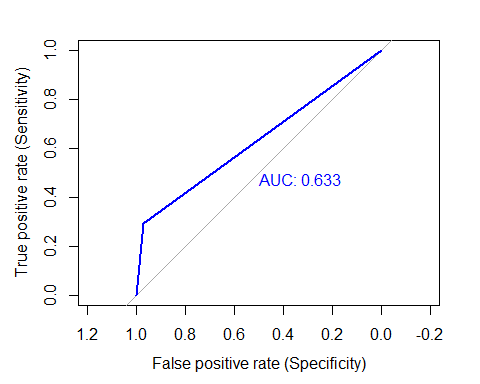
##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

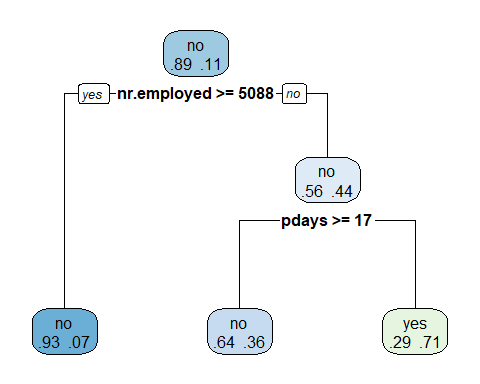
auc(as.numeric(solution$y), as.numeric(solution$subscribed))

## Area under the curve: 0.6326

#Area under the curve: 0.6326  
  
roc\_rose <- plot(roc(as.numeric(solution$y), as.numeric(solution$subscribed)), print.auc = TRUE, col = "blue", xlab = "False positive rate (Specificity)", ylab = "True positive rate (Sensitivity)")



##################################### Decision Tree  
row\_count <- nrow(bankDataWithCluster)  
shuffled\_rows <- sample(row\_count)  
train <- bankDataWithCluster[head(shuffled\_rows,floor(row\_count\*0.75)),]  
test <- bankDataWithCluster[tail(shuffled\_rows,floor(row\_count\*0.25)),]  
  
set.seed(1934)  
  
#library(rpart)  
dt\_model <- rpart(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data = train,  
 method = "class",  
 minsplit = 2,  
 minbucket = 1)  
#library(rpart.plot)  
rpart.plot(dt\_model, extra=4) # plot tree



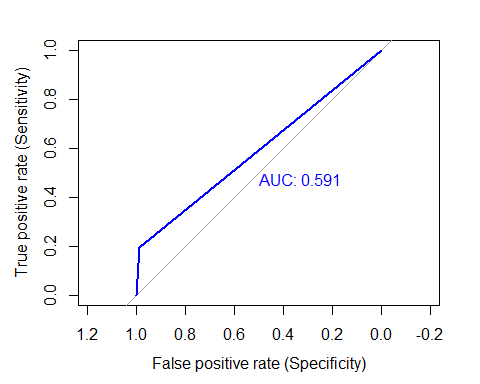
# Predict using the test set  
prediction <- predict(dt\_model, test)  
  
solution <- data.frame(test, subscribed = prediction)  
solution$subscribed <- ifelse(solution$subscribed.yes > solution$subscribed.no, "yes", "no")  
  
#solution  
  
#library(caret)  
confusionMatrix(factor(solution$subscribed), solution$y)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 9015 953  
## yes 100 229  
##   
## Accuracy : 0.8977   
## 95% CI : (0.8917, 0.9035)  
## No Information Rate : 0.8852   
## P-Value [Acc > NIR] : 2.738e-05   
##   
## Kappa : 0.2664   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9890   
## Specificity : 0.1937   
## Pos Pred Value : 0.9044   
## Neg Pred Value : 0.6960   
## Prevalence : 0.8852   
## Detection Rate : 0.8755   
## Detection Prevalence : 0.9680   
## Balanced Accuracy : 0.5914   
##   
## 'Positive' Class : no   
##

#nrow(solution[solution$y=='no',])  
#Balanced Accuracy: 0.5914  
  
#library(pROC)  
auc(as.numeric(solution$y), as.numeric(factor(solution$subscribed)))

## Area under the curve: 0.5914

#Area under the curve: 0.5914  
  
roc\_rose <- plot(roc(as.numeric(solution$y), as.numeric(factor(solution$subscribed))), print.auc = TRUE, col = "blue", xlab = "False positive rate (Specificity)", ylab = "True positive rate (Sensitivity)")



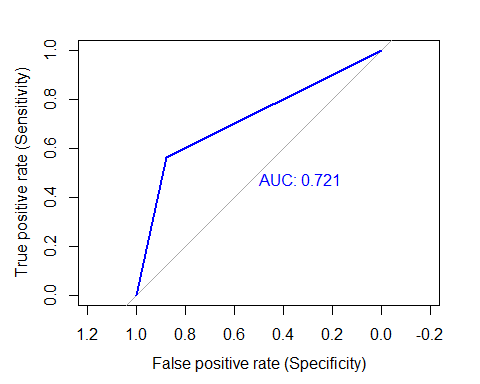
################################# Naive Bayes  
row\_count <- nrow(bankDataWithCluster)  
shuffled\_rows <- sample(row\_count)  
train <- bankDataWithCluster[head(shuffled\_rows,floor(row\_count\*0.75)),]  
test <- bankDataWithCluster[tail(shuffled\_rows,floor(row\_count\*0.25)),]  
  
library(e1071)  
  
nb\_model <- naiveBayes(y ~ age + job + marital + education + default +   
 housing + loan + contact +   
 month + day\_of\_week +  
 campaign + pdays + previous +  
 poutcome + emp.var.rate + cons.price.idx +  
 cons.conf.idx + euribor3m + nr.employed,  
 data=train)  
  
# Predict using the test set  
prediction <- predict(nb\_model, test)  
  
solution <- data.frame(test, subscribed = prediction)  
#solution  
  
#library(caret)  
confusionMatrix(factor(solution$subscribed), solution$y)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 8048 499  
## yes 1108 642  
##   
## Accuracy : 0.8439   
## 95% CI : (0.8368, 0.8509)  
## No Information Rate : 0.8892   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.358   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.8790   
## Specificity : 0.5627   
## Pos Pred Value : 0.9416   
## Neg Pred Value : 0.3669   
## Prevalence : 0.8892   
## Detection Rate : 0.7816   
## Detection Prevalence : 0.8300   
## Balanced Accuracy : 0.7208   
##   
## 'Positive' Class : no   
##

#Balanced Accuracy: 0.7129  
  
#library(pROC)  
auc(as.numeric(solution$y), as.numeric(solution$subscribed))

## Area under the curve: 0.7208

#Area under the curve: 0.7129  
  
roc\_rose <- plot(roc(as.numeric(solution$y), as.numeric(solution$subscribed)), print.auc = TRUE, col = "blue", xlab = "False positive rate (Specificity)", ylab = "True positive rate (Sensitivity)")



## Prediction Models Analysis

The higher the area under the curve (AUC), the better the prediction model. We tried 4 classification models: Random Forest, Decision Tree, Naive Bayes, and Support Vector Machine (SVM). SVM takes over 30 minutes to run so we are excluding its results. The order of models from highest AUC to lowest is: 1. Naive Bayes (AUC: 0.7129) 2. Random Forest (AUC: 0.6326) 3. Decision Tree (AUC: 0.5914)