r6

Simerpreet

November 20, 2020

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages -------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.3  
## v tibble 3.0.4 v stringr 1.4.0  
## v tidyr 1.0.0 v forcats 0.4.0  
## v readr 1.3.1

## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## -- Conflicts ----------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(e1071)

## Warning: package 'e1071' was built under R version 3.6.2

library(class)  
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.6.2

##   
## Attaching package: 'gridExtra'

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

library(summarytools)

## Warning: package 'summarytools' was built under R version 3.6.3

## Registered S3 method overwritten by 'pryr':  
## method from  
## print.bytes Rcpp

## For best results, restart R session and update pander using devtools:: or remotes::install\_github('rapporter/pander')

##   
## Attaching package: 'summarytools'

## The following object is masked from 'package:tibble':  
##   
## view

library(gt)

## Warning: package 'gt' was built under R version 3.6.3

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.3

## corrplot 0.84 loaded

library(janitor)

## Warning: package 'janitor' was built under R version 3.6.3

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(tidyselect)

## Warning: package 'tidyselect' was built under R version 3.6.3

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.6.3

## 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

library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:purrr':  
##   
## some

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

#full <- read\_delim(here::here("data", "bank-additional-full.csv"),';')  
full <- read.csv(file.choose(), sep=';')  
str(full)

## '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 ...

head(full)

## 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

nrow(full)

## [1] 41188

ncol(full)

## [1] 21

# Clean up column names  
full <- janitor::clean\_names(full)  
summary(full)

## 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   
##

#print(dfSummary(full, graph.magnif = 0.75), method = 'browser')  
str(full)

## '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 ...

# Check for missing values  
tibble(variable = names(colSums(is.na(full))),  
 missing = colSums(is.na(full))) %>%   
 gt() %>%   
 tab\_header(title = "Missing Values in Data")

Missing Values in Data

variable

missing

age

0

job

0

marital

0

education

0

default

0

housing

0

loan

0

contact

0

month

0

day\_of\_week

0

duration

0

campaign

0

pdays

0

previous

0

poutcome

0

emp\_var\_rate

0

cons\_price\_idx

0

cons\_conf\_idx

0

euribor3m

0

nr\_employed

0

y

0

#Looking at the dfsummary, there doesnt seem to be missing data in terms of just not having values. However, there are some fields that have explicit unknown or non-existent classes that could be considered as 'missing'. For example, loan and housing have 990 "unknown" values. And 'default' has 8597 "unknown" values representing 20.9%   
  
#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)  
  
# Remove missing values  
#```{r}  
#remove "unknowns" based on small sample sizes compared to full data set  
df <- full %>% filter(loan != "unknown")  
nrow(df)

## [1] 40198

#down to 40,198 obs  
df <- df %>% filter(marital != "unknown")  
nrow(df)

## [1] 40119

#down to 40,119 obs  
df <- df %>% filter(education != "unknown")  
nrow(df)

## [1] 38437

#down to 38,437 obs  
  
#Simer: remove unknowns from job  
df <- df %>% filter(job != "unknown")  
nrow(df)

## [1] 38245

#down to 38,245 obs  
#Remove column default from the analysis. REasons for removing the columns:   
#1) After cleaning up the data set of the other 'unknowns', this column has 7,757 unkown values as well. With only 3 values as 'yes',and 30,485 as 'no', it is difficult to impute values  
#2) Practically, column default would need to be used before the campagn, the sales person needs to decide if the person wilth a default needs to be approached to or not, not after the campaign,  
#so it appears that it is ok to let go of this column when predicting the outcome of the campaign.   
  
#Keeping default in for now  
#df <- df %>% dplyr::select(age:education,housing: y) #for some reason select(default is not working for me)  
#Closing Simer's changes  
str(df)

## 'data.frame': 38245 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 24 25 25 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 10 8 8 ...  
## $ marital : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 3 3 3 ...  
## $ education : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 6 4 4 ...  
## $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 1 1 1 ...  
## $ housing : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 3 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 261 149 226 151 307 198 139 380 50 222 ...  
## $ 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 ...

#recheck summary  
summary(df)

## age job marital   
## Min. :17.00 admin. :9937 divorced: 4302   
## 1st Qu.:32.00 blue-collar:8560 married :23183   
## Median :38.00 technician :6380 single :10760   
## Mean :39.86 services :3716 unknown : 0   
## 3rd Qu.:47.00 management :2728   
## Max. :98.00 retired :1577   
## (Other) :5347   
## education default housing   
## university.degree :11821 no :30485 no :17667   
## high.school : 9244 unknown: 7757 unknown: 0   
## basic.9y : 5856 yes : 3 yes :20578   
## professional.course: 5100   
## basic.4y : 4002   
## basic.6y : 2204   
## (Other) : 18   
## loan contact month day\_of\_week  
## no :32286 cellular :24441 may :12794 fri:7224   
## unknown: 0 telephone:13804 jul : 6630 mon:7927   
## yes : 5959 aug : 5822 thu:8011   
## jun : 4846 tue:7481   
## nov : 3898 wed:7602   
## apr : 2436   
## (Other): 1819   
## duration campaign pdays previous   
## Min. : 0.0 Min. : 1.000 Min. : 0.0 Min. :0.00   
## 1st Qu.: 102.0 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.00   
## Median : 180.0 Median : 2.000 Median :999.0 Median :0.00   
## Mean : 258.2 Mean : 2.567 Mean :963.5 Mean :0.17   
## 3rd Qu.: 319.0 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.00   
## Max. :4918.0 Max. :43.000 Max. :999.0 Max. :7.00   
##   
## poutcome emp\_var\_rate cons\_price\_idx cons\_conf\_idx   
## failure : 3936 Min. :-3.40000 Min. :92.20 Min. :-50.80   
## nonexistent:33066 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70   
## success : 1243 Median : 1.10000 Median :93.44 Median :-41.80   
## Mean : 0.08286 Mean :93.57 Mean :-40.54   
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.40   
## Max. : 1.40000 Max. :94.77 Max. :-26.90   
##   
## euribor3m nr\_employed y   
## Min. :0.634 Min. :4964 no :33987   
## 1st Qu.:1.344 1st Qu.:5099 yes: 4258   
## Median :4.857 Median :5191   
## Mean :3.623 Mean :5167   
## 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :5.045 Max. :5228   
##

#```  
#Simer changed the number of decreased 'yes' from 400 to 480 and the final number as well  
message ("Our yes group has decreased by about ~480 to 4,258.")

## Our yes group has decreased by about ~480 to 4,258.

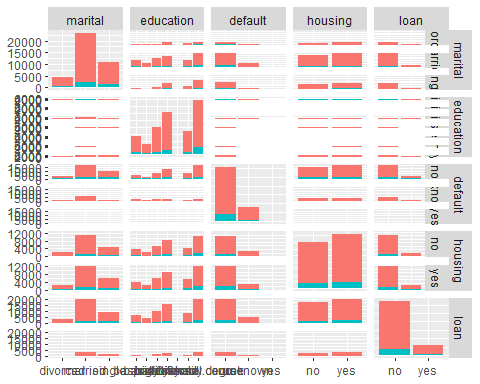
#Our yes group has decreased by about ~400 to 4,277.  
  
#```{r}  
#change some variables to factor  
#Simer removed default from the list of columns  
#cols <- c("job", "marital", "education", "default","housing","loan","contact","month","day\_of\_week","poutcome","y")  
cols <- c("job", "marital", "education", "housing","loan","contact","month","day\_of\_week","poutcome","y")  
df[cols] <- lapply(df[cols], factor)   
str(df)

## 'data.frame': 38245 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 24 25 25 ...  
## $ job : Factor w/ 11 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 10 8 8 ...  
## $ marital : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 3 3 3 ...  
## $ education : Factor w/ 7 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 6 4 4 ...  
## $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 1 1 1 ...  
## $ housing : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 1 2 2 2 ...  
## $ loan : Factor w/ 2 levels "no","yes": 1 1 1 1 2 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 261 149 226 151 307 198 139 380 50 222 ...  
## $ 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 ...

#make sure "success" level is defined as "yes"  
str(df$y)

## Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

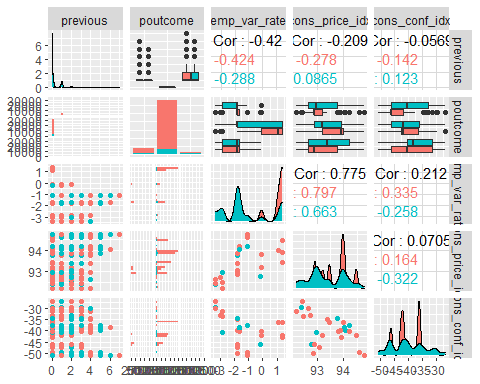
#```  
  
# Exploratory Data Analysis  
  
#ggpairs(df,columns=1:18, aes(colour=y))  
  
ggpairs(df,columns=3:7, aes(colour=y))



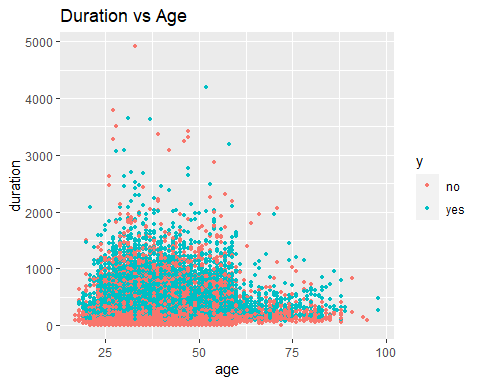
ggpairs(df, columns=14:18, aes(colour=y))

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

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



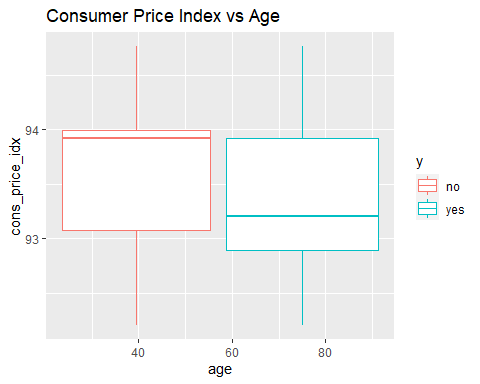
df\_yes <- df %>% filter(y=="yes")  
#summary(df\_yes)  
# Nothing interesting found in the below code so commenting it out  
# ggplot(bank\_additional\_full, aes(x=age, y=emp.var.rate)) +  
# geom\_point(size=1, shape="circle") +  
# ggtitle("Employment Variation Rate vs Age") +   
# facet\_wrap(~ y)  
ggplot(df, aes(x=age, y=duration, color = y)) + geom\_point(size=1, shape="circle") + ggtitle("Duration vs Age")



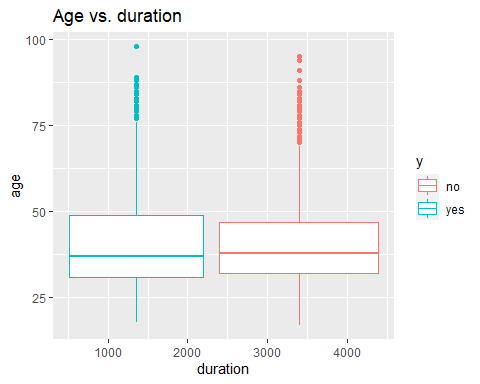
message("Duration vs Age: The duration of last contact (in seconds) was longer for ages 25-50. And it was understandably longer for 'yes'' vs for 'no'. ")

## Duration vs Age: The duration of last contact (in seconds) was longer for ages 25-50. And it was understandably longer for 'yes'' vs for 'no'.

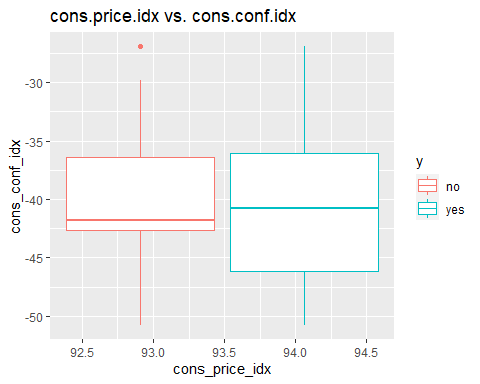
#ggplot(df, aes(x=age, y=cons\_price\_idx, color = y)) + geom\_point(size=1, shape="circle") + ggtitle("Consumer Price Index vs Age")  
  
  
#Checking collinearlity using box plots  
#Simer: Added box plot for Consumer Price Index vs Age  
ggplot(df, aes(x=age, y=cons\_price\_idx, color = y)) + geom\_boxplot() + ggtitle("Consumer Price Index vs Age")



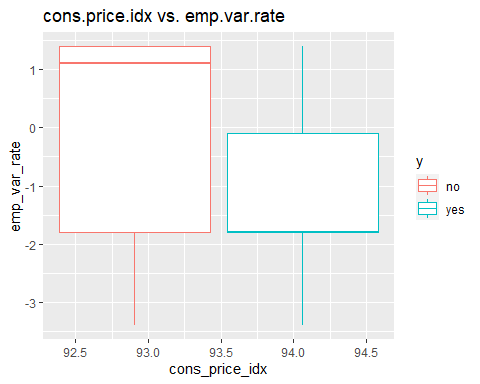
#Simer: Added box plot for Age vs. duration  
ggplot(df, aes(x=duration , y=age, color = y)) + geom\_boxplot() + ggtitle("Age vs. duration")



#Simer: Added box plot for cons.price.idx vs. cons.conf.idx  
ggplot(df, aes(x=cons\_price\_idx , y=cons\_conf\_idx, color = y)) + geom\_boxplot() + ggtitle("cons.price.idx vs. cons.conf.idx")

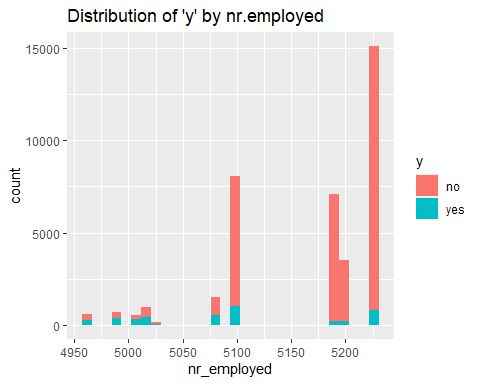


#Simer: Added box plot for cons.price.idx vs. emp.var.rate  
ggplot(df, aes(x=cons\_price\_idx , y=emp\_var\_rate, color = y)) + geom\_boxplot() + ggtitle("cons.price.idx vs. emp.var.rate")



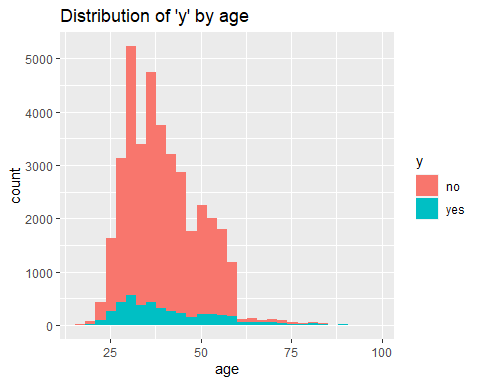
#Simer: Analysing nr.employed  
ggplot(df) + geom\_histogram(mapping = aes(x=nr\_employed, fill=y)) +ggtitle("Distribution of 'y' by nr.employed")

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

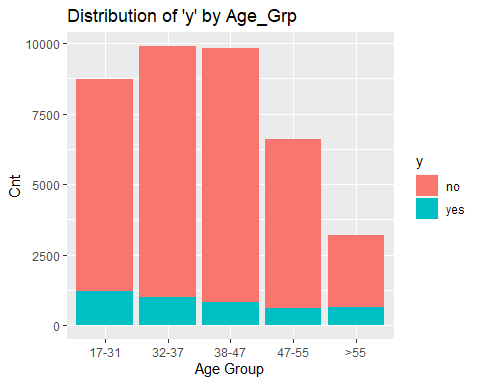


# ggplot(bank\_additional\_full, aes(x=age, y=education)) +  
# geom\_point(size=1, shape="circle") +   
# ggtitle("Education vs Age") +   
# facet\_wrap(~ y)  
  
  
#Analysing Age  
ggplot(df) + geom\_histogram(mapping = aes(x=age, fill=y)) +ggtitle("Distribution of 'y' by age")

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



#Creating new variables  
#Age\_Grp - split the data into age groups "17-31","32-37" ,"38-47", "47-55", ">55" (based in IQR)  
df$Age\_Grp <- cut(df$age, breaks = c(16,31,37,46,55,98), labels = c("17-31","32-37" ,"38-47", "47-55", ">55"))  
#validate the cut command  
#df %>% filter(!$Age\_Grp %in% c("17-31","32-37" ,"38-47", "47-55", ">55"))  
#df %>% filter(df$age==55)  
ggplot(df) + geom\_bar(mapping = aes(x=Age\_Grp, fill = y)) + ggtitle("Distribution of 'y' by Age\_Grp") + ylab("Cnt") + xlab("Age Group")

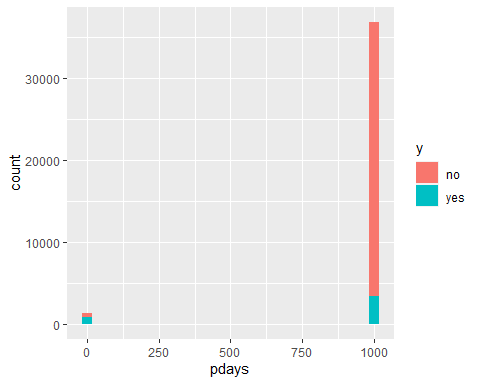


message("We will keep both age and age group in our model to see if one is selected over the other. We need to make sure to not use both in our model building.")

## We will keep both age and age group in our model to see if one is selected over the other. We need to make sure to not use both in our model building.

#Analyzing pdays  
ggplot(df) + geom\_histogram(mapping = aes(x=pdays, fill=y))

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



message("Analyzing 'pdays' ie., number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted")

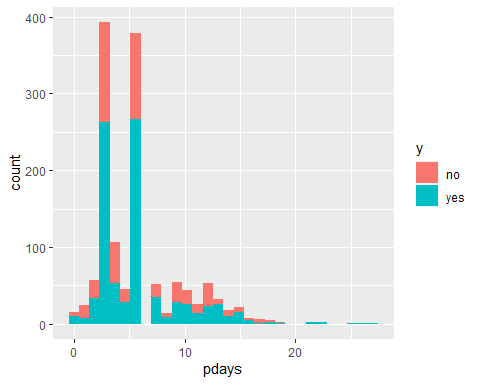
## Analyzing 'pdays' ie., number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted

message(" Most folks had no previous campaign but if they did, it looks like most who had a previous campaign decided to subscribe")

## Most folks had no previous campaign but if they did, it looks like most who had a previous campaign decided to subscribe

#zoom in for ones that were previously contacted  
df %>% filter(pdays < 999) %>% ggplot() + geom\_histogram(mapping = aes(x=pdays, fill=y))

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



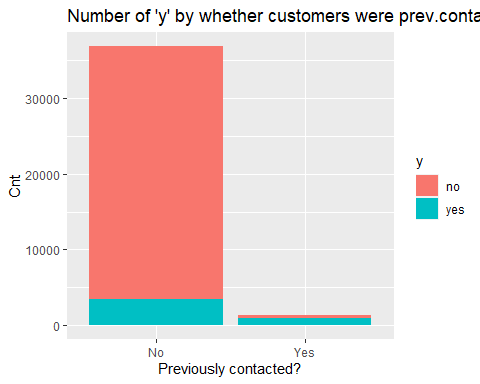
message(" Highest frequency appears to be less than 10 days since last contact. Let''s make this into a y/n variable instead due to the large gap between days contacted and the '999' variable.")

## Highest frequency appears to be less than 10 days since last contact. Let''s make this into a y/n variable instead due to the large gap between days contacted and the '999' variable.

message(" prevly\_Cntctd Yes/No. TO see the distribution or 'Y' on first time contact vs. a follow up")

## prevly\_Cntctd Yes/No. TO see the distribution or 'Y' on first time contact vs. a follow up

df$prevly\_Cntctd <- as.factor(case\_when(df$pdays==999 ~ "No", !df$pdays==999 ~ "Yes"))  
#Validate previously contacted variable  
#df %>% filter(!df$pdays==999)  
ggplot(df) + geom\_bar(mapping = aes(x=prevly\_Cntctd, fill = y)) + ggtitle("Number of 'y' by whether customers were prev.contacted or not") +  
 ylab("Cnt") + xlab("Previously contacted?")



message(" Same observation here as above: Most folks had no previous campaign but if they did, it looks like most who had a previous campaign decided to subscribe / likely to say 'Yes'. Since we have now created a new variable dependent on pdays, proceed to remove pdays to avoid issues with multicollinearity. Additionally, poutcome is dependent on whether or not someone has previously been contacted, and we have 'nonexistent' at 86% of the data. Remove this variable as well since it doesn''t add much value and is dependent on pdays/previously contacted.")

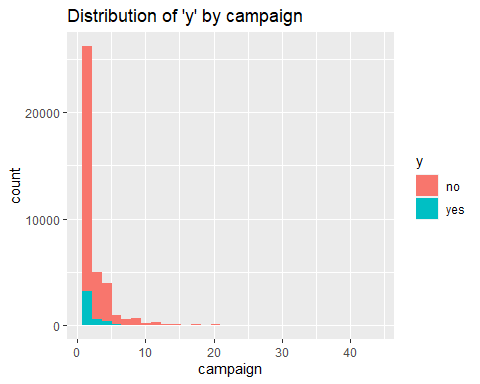
## Same observation here as above: Most folks had no previous campaign but if they did, it looks like most who had a previous campaign decided to subscribe / likely to say 'Yes'. Since we have now created a new variable dependent on pdays, proceed to remove pdays to avoid issues with multicollinearity. Additionally, poutcome is dependent on whether or not someone has previously been contacted, and we have 'nonexistent' at 86% of the data. Remove this variable as well since it doesn''t add much value and is dependent on pdays/previously contacted.

#Simer: Let's keep poutcome, these are independent variables. Let's check in refression or by VIF''s to check collinearily betweent he two  
#df <- df %>% select(-pdays)  
  
#df <- df %>% select(-pdays, -poutcome)  
summary(df)

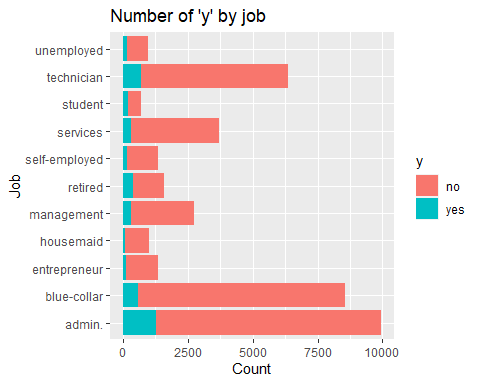
## age job marital   
## Min. :17.00 admin. :9937 divorced: 4302   
## 1st Qu.:32.00 blue-collar:8560 married :23183   
## Median :38.00 technician :6380 single :10760   
## Mean :39.86 services :3716   
## 3rd Qu.:47.00 management :2728   
## Max. :98.00 retired :1577   
## (Other) :5347   
## education default housing loan   
## basic.4y : 4002 no :30485 no :17667 no :32286   
## basic.6y : 2204 unknown: 7757 yes:20578 yes: 5959   
## basic.9y : 5856 yes : 3   
## high.school : 9244   
## illiterate : 18   
## professional.course: 5100   
## university.degree :11821   
## contact month day\_of\_week duration   
## cellular :24441 may :12794 fri:7224 Min. : 0.0   
## telephone:13804 jul : 6630 mon:7927 1st Qu.: 102.0   
## aug : 5822 thu:8011 Median : 180.0   
## jun : 4846 tue:7481 Mean : 258.2   
## nov : 3898 wed:7602 3rd Qu.: 319.0   
## apr : 2436 Max. :4918.0   
## (Other): 1819   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.0 Min. :0.00 failure : 3936   
## 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.00 nonexistent:33066   
## Median : 2.000 Median :999.0 Median :0.00 success : 1243   
## Mean : 2.567 Mean :963.5 Mean :0.17   
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.00   
## Max. :43.000 Max. :999.0 Max. :7.00   
##   
## emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m   
## Min. :-3.40000 Min. :92.20 Min. :-50.80 Min. :0.634   
## 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70 1st Qu.:1.344   
## Median : 1.10000 Median :93.44 Median :-41.80 Median :4.857   
## Mean : 0.08286 Mean :93.57 Mean :-40.54 Mean :3.623   
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.40 3rd Qu.:4.961   
## Max. : 1.40000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr\_employed y Age\_Grp prevly\_Cntctd  
## Min. :4964 no :33987 17-31:8741 No :36879   
## 1st Qu.:5099 yes: 4258 32-37:9893 Yes: 1366   
## Median :5191 38-47:9834   
## Mean :5167 47-55:6600   
## 3rd Qu.:5228 >55 :3177   
## Max. :5228   
##

#Analysing campaign  
ggplot(df) + geom\_histogram(mapping = aes(x=campaign, fill=y)) + ggtitle("Distribution of 'y' by campaign")

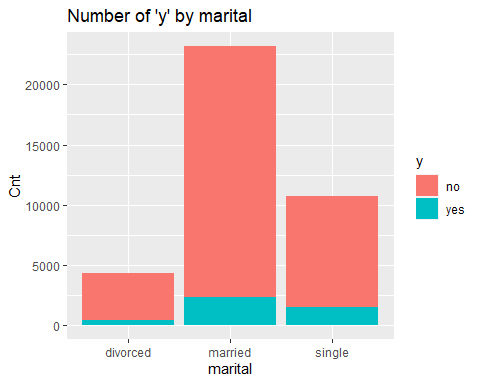
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Just visually, when we decided to stop contacting a person it didn''t affect our closing ratio which still dropped off precipitously   
  
# Ideally, the campaign would stop contacting people who are less likely to subscribe, and keep contacting people if they are more likely to subscribe. Then we should see the ratio of Yes to No go up as the number of no contacts goes up. Instead, it looks like the ratio stays the same and the number of Yes''s drops proportionately with the number of No''s.   
  
#Analyzing job  
ggplot(df) + geom\_bar(mapping = aes(x=job, fill = y)) + coord\_flip() + #Added coord flip here to make it more readable  
 ggtitle("Number of 'y' by job") + ylab("Count") + xlab("Job")



#"y" - has a client subscribed a term deposit? : admin, technician and blue collar jobs are the top 3 subscribers by volume   
  
  
#df2 <- df %>% group\_by(job) %>% count(y) %>% mutate(job\_conv = n/sum(n)) %>% filter(y == "yes")  
#ggplot(df2, aes(x=job, y=job\_conv)) + geom\_point() + coord\_flip()   
  
  
#Above, I looked at the ratio of "yes" vs "no" and see that students and retired persons convert at much higher rates than those of other professions. And 'blue collar' has the lowest conversion rate  
  
#So, if they were to want to improve the cost effectiveness of their campaigns they might want to target more 'students' and 'retirees'  
  
  
  
#Analyzing marital  
ggplot(data=df) + geom\_bar(mapping = aes(x=marital, fill = y)) + ggtitle("Number of 'y' by marital") + ylab("Cnt") + xlab("marital")



# More 'married' people are represented in the campaign  
#Visually looking, conversion rate seems to be higher for 'single' people  
  
  
#Analyzing duration and creating duration group variable  
summary(df$duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 102.0 180.0 258.2 319.0 4918.0

df$duration\_group <- cut(df$duration, breaks = c(-Inf,100,60,300,600,Inf), labels = c("0-30s", "30-60s", "1-5 min", "5-10min","10+ min"))  
# Check for missing values  
tibble(variable = names(colSums(is.na(df))),  
 missing = colSums(is.na(df))) %>%   
 gt() %>%   
 tab\_header(title = "Missing Values in Data")

Missing Values in Data

variable

missing

age

0

job

0

marital

0

education

0

default

0

housing

0

loan

0

contact

0

month

0

day\_of\_week

0

duration

0

campaign

0

pdays

0

previous

0

poutcome

0

emp\_var\_rate

0

cons\_price\_idx

0

cons\_conf\_idx

0

euribor3m

0

nr\_employed

0

y

0

Age\_Grp

0

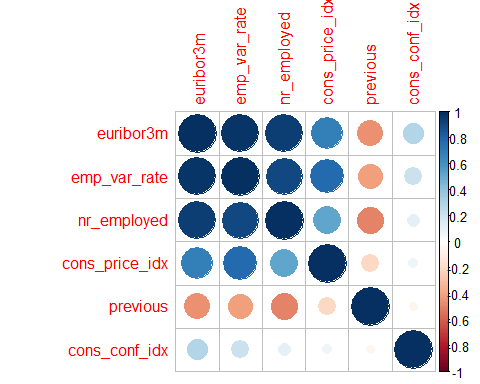
prevly\_Cntctd

0

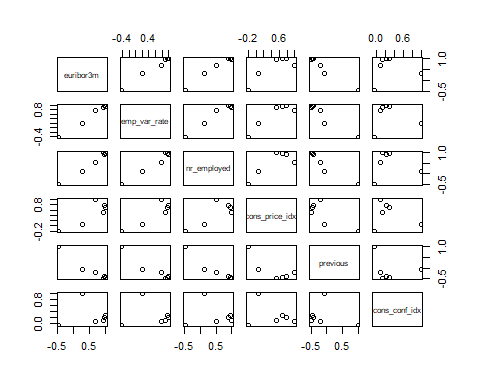
duration\_group

0

df3 <- df %>% group\_by(duration\_group) %>% count(y) %>% mutate(duration\_group\_conv = n/sum(n)) %>% filter(y == "yes")  
#ggplot(df3, aes(x=duration\_group, y=duration\_group\_conv)) + geom\_point() + facet\_wrap(~ y)  
  
  
#Looking above, clearly conversion rate goes up the longer the most recent call  
  
# Checking for correlation  
  
# Look for high correlation  
df.numeric <- df[ , sapply(df, is.numeric)]  
corrs = cor(df.numeric, use="everything") # Calculate correlations between all variables  
high\_corrs = findCorrelation(corrs, cutoff=abs(0.1)) # Find 'high' correlations among those variables (0.1 is not exactly "high" but there are so few numerics...)  
corrs = cor(df.numeric[,high\_corrs], use="everything") # get a data frame with only highly correlated variables  
#Create corrplot for numeric variables  
corrplot(corrs)



# pairs only on highly correlated variables...   
pairs(corrs,col=df$y)



#Based on the correlation plot above, we see high correlation between 'euribor3m' and 'emp\_var\_rate' and to a lesser degree with 'nr\_employed.' We also see 'nr\_employed' and 'emp\_var\_rate' also highly correlated, which makes sense since you would expect the number of employees to vary at the same time as the employment variation rate. Let us proceed with removal of 'emp\_var\_rate' as that appears to be correlated at a higher rate to euribor3m.  
  
  
#remove emp\_var\_rate  
summary(df)

## age job marital   
## Min. :17.00 admin. :9937 divorced: 4302   
## 1st Qu.:32.00 blue-collar:8560 married :23183   
## Median :38.00 technician :6380 single :10760   
## Mean :39.86 services :3716   
## 3rd Qu.:47.00 management :2728   
## Max. :98.00 retired :1577   
## (Other) :5347   
## education default housing loan   
## basic.4y : 4002 no :30485 no :17667 no :32286   
## basic.6y : 2204 unknown: 7757 yes:20578 yes: 5959   
## basic.9y : 5856 yes : 3   
## high.school : 9244   
## illiterate : 18   
## professional.course: 5100   
## university.degree :11821   
## contact month day\_of\_week duration   
## cellular :24441 may :12794 fri:7224 Min. : 0.0   
## telephone:13804 jul : 6630 mon:7927 1st Qu.: 102.0   
## aug : 5822 thu:8011 Median : 180.0   
## jun : 4846 tue:7481 Mean : 258.2   
## nov : 3898 wed:7602 3rd Qu.: 319.0   
## apr : 2436 Max. :4918.0   
## (Other): 1819   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.0 Min. :0.00 failure : 3936   
## 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.00 nonexistent:33066   
## Median : 2.000 Median :999.0 Median :0.00 success : 1243   
## Mean : 2.567 Mean :963.5 Mean :0.17   
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.00   
## Max. :43.000 Max. :999.0 Max. :7.00   
##   
## emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m   
## Min. :-3.40000 Min. :92.20 Min. :-50.80 Min. :0.634   
## 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70 1st Qu.:1.344   
## Median : 1.10000 Median :93.44 Median :-41.80 Median :4.857   
## Mean : 0.08286 Mean :93.57 Mean :-40.54 Mean :3.623   
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.40 3rd Qu.:4.961   
## Max. : 1.40000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr\_employed y Age\_Grp prevly\_Cntctd duration\_group   
## Min. :4964 no :33987 17-31:8741 No :36879 0-30s : 3985   
## 1st Qu.:5099 yes: 4258 32-37:9893 Yes: 1366 30-60s : 5337   
## Median :5191 38-47:9834 1-5 min:18524   
## Mean :5167 47-55:6600 5-10min: 7189   
## 3rd Qu.:5228 >55 :3177 10+ min: 3210   
## Max. :5228   
##

#df <- df %>% select(-emp\_var\_rate)  
#Simer: COmmented out the above step.   
  
  
# Run random forest on down-sampled data set to check for variable importance   
  
#I am running RF on a subset of the data to do a gross check for important variables and to determine if the new variables duration group and age group are deemed more important than the continuous variables of just raw duration and raw age.  
  
  
#move response variable to end of data set  
df <- df %>% relocate(y, .after = last\_col())  
#randomly sample 10k obs  
#sample10k <- sample\_n(df, 10000)  
#down sample to balance response  
#set.seed(1)  
#downsample <- downSample(x = sample10k[, -21],  
#y = sample10k$y)  
#table(downsample$Class)  
#RFcontrol <- rfeControl(functions=rfFuncs, method="cv", number=5, verbose = FALSE)  
#set.seed(123)  
#subsets <- c(1:5, 10, 15, 20)  
#RFresults <- rfe(downsample[,1:20], downsample[[21]], sizes=subsets, rfeControl=RFcontrol)  
#RFresults  
#varImp(RFresults)  
  
#Based on this, remove 'duration\_group' and 'age\_group' as the continuous version of those variables had higher importance on the final model.  
str(df)

## 'data.frame': 38245 obs. of 24 variables:  
## $ age : int 56 57 37 40 56 45 59 24 25 25 ...  
## $ job : Factor w/ 11 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 10 8 8 ...  
## $ marital : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 3 3 3 ...  
## $ education : Factor w/ 7 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 6 4 4 ...  
## $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 1 1 1 ...  
## $ housing : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 1 2 2 2 ...  
## $ loan : Factor w/ 2 levels "no","yes": 1 1 1 1 2 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 261 149 226 151 307 198 139 380 50 222 ...  
## $ 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 ...  
## $ Age\_Grp : Factor w/ 5 levels "17-31","32-37",..: 5 5 2 3 5 3 5 1 1 1 ...  
## $ prevly\_Cntctd : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ duration\_group: Factor w/ 5 levels "0-30s","30-60s",..: 3 3 3 3 4 3 3 4 1 3 ...  
## $ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

#Simer  
#RFcontrol <- rfeControl(functions=rfFuncs, method="cv", number=5, verbose = FALSE)  
#set.seed(123)  
#subsets <- c(1:5, 10, 15, 20)  
#RFresults <- rfe(df[,1:20], df[[21]], sizes=subsets, rfeControl=RFcontrol)  
#RFresults  
#varImp(RFresults)  
  
  
#df <- df %>% select(-duration\_group, -Age\_Grp)  
  
# Train/Test Split  
  
#Simer train/test split. Making sure the train and test set get enough 'yes' variables.   
summary(df)

## age job marital   
## Min. :17.00 admin. :9937 divorced: 4302   
## 1st Qu.:32.00 blue-collar:8560 married :23183   
## Median :38.00 technician :6380 single :10760   
## Mean :39.86 services :3716   
## 3rd Qu.:47.00 management :2728   
## Max. :98.00 retired :1577   
## (Other) :5347   
## education default housing loan   
## basic.4y : 4002 no :30485 no :17667 no :32286   
## basic.6y : 2204 unknown: 7757 yes:20578 yes: 5959   
## basic.9y : 5856 yes : 3   
## high.school : 9244   
## illiterate : 18   
## professional.course: 5100   
## university.degree :11821   
## contact month day\_of\_week duration   
## cellular :24441 may :12794 fri:7224 Min. : 0.0   
## telephone:13804 jul : 6630 mon:7927 1st Qu.: 102.0   
## aug : 5822 thu:8011 Median : 180.0   
## jun : 4846 tue:7481 Mean : 258.2   
## nov : 3898 wed:7602 3rd Qu.: 319.0   
## apr : 2436 Max. :4918.0   
## (Other): 1819   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.0 Min. :0.00 failure : 3936   
## 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.00 nonexistent:33066   
## Median : 2.000 Median :999.0 Median :0.00 success : 1243   
## Mean : 2.567 Mean :963.5 Mean :0.17   
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.00   
## Max. :43.000 Max. :999.0 Max. :7.00   
##   
## emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m   
## Min. :-3.40000 Min. :92.20 Min. :-50.80 Min. :0.634   
## 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70 1st Qu.:1.344   
## Median : 1.10000 Median :93.44 Median :-41.80 Median :4.857   
## Mean : 0.08286 Mean :93.57 Mean :-40.54 Mean :3.623   
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.40 3rd Qu.:4.961   
## Max. : 1.40000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr\_employed Age\_Grp prevly\_Cntctd duration\_group y   
## Min. :4964 17-31:8741 No :36879 0-30s : 3985 no :33987   
## 1st Qu.:5099 32-37:9893 Yes: 1366 30-60s : 5337 yes: 4258   
## Median :5191 38-47:9834 1-5 min:18524   
## Mean :5167 47-55:6600 5-10min: 7189   
## 3rd Qu.:5228 >55 :3177 10+ min: 3210   
## Max. :5228   
##

set.seed(1234)   
  
df\_yes <- df %>% filter(y=='yes')  
df\_No <- df %>% filter(y=='no')  
num\_rows\_yes <- nrow(df\_yes) #4,258  
num\_rows\_no <- nrow(df\_No) #33,987  
  
train\_idx\_yes <- sample(1:num\_rows\_yes, 0.8 \* num\_rows\_yes)  
train\_yes <- df\_yes[train\_idx\_yes, ]  
test\_yes <- df\_yes[-train\_idx\_yes, ]  
nrow(train\_yes) #3,406

## [1] 3406

nrow(test\_yes) #852

## [1] 852

train\_idx\_no <- sample(1:num\_rows\_no, 0.8 \* num\_rows\_no)  
train\_no <- df\_No[train\_idx\_no, ]  
test\_no <- df\_No[-train\_idx\_no, ]  
nrow(train\_no) #27,189

## [1] 27189

nrow(test\_no) #6798

## [1] 6798

train <- rbind(train\_yes, train\_no)  
test <- rbind(test\_yes, test\_no)  
  
nrow(train) #30,595

## [1] 30595

nrow(test) #7,650

## [1] 7650

nrow(train %>% filter(y=='yes')) #3,406

## [1] 3406

nrow(test %>% filter(y=='yes')) #852

## [1] 852

summary(train)

## age job marital   
## Min. :17.00 admin. :7922 divorced: 3418   
## 1st Qu.:32.00 blue-collar:6852 married :18575   
## Median :38.00 technician :5144 single : 8602   
## Mean :39.84 services :2955   
## 3rd Qu.:47.00 management :2166   
## Max. :98.00 retired :1267   
## (Other) :4289   
## education default housing loan   
## basic.4y :3178 no :24411 no :14161 no :25818   
## basic.6y :1751 unknown: 6181 yes:16434 yes: 4777   
## basic.9y :4714 yes : 3   
## high.school :7398   
## illiterate : 13   
## professional.course:4089   
## university.degree :9452   
## contact month day\_of\_week duration   
## cellular :19570 may :10144 fri:5795 Min. : 0.0   
## telephone:11025 jul : 5311 mon:6341 1st Qu.: 102.0   
## aug : 4700 thu:6407 Median : 179.0   
## jun : 3919 tue:6009 Mean : 257.3   
## nov : 3104 wed:6043 3rd Qu.: 319.0   
## apr : 1976 Max. :4918.0   
## (Other): 1441   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.0 Min. :0.000 failure : 3172   
## 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.000 nonexistent:26431   
## Median : 2.000 Median :999.0 Median :0.000 success : 992   
## Mean : 2.581 Mean :963.5 Mean :0.171   
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.000   
## Max. :43.000 Max. :999.0 Max. :7.000   
##   
## emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m   
## Min. :-3.40000 Min. :92.20 Min. :-50.80 Min. :0.634   
## 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70 1st Qu.:1.344   
## Median : 1.10000 Median :93.44 Median :-41.80 Median :4.857   
## Mean : 0.08521 Mean :93.57 Mean :-40.54 Mean :3.625   
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.40 3rd Qu.:4.961   
## Max. : 1.40000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr\_employed Age\_Grp prevly\_Cntctd duration\_group y   
## Min. :4964 17-31:7015 No :29502 0-30s : 3225 no :27189   
## 1st Qu.:5099 32-37:7927 Yes: 1093 30-60s : 4255 yes: 3406   
## Median :5191 38-47:7846 1-5 min:14820   
## Mean :5168 47-55:5290 5-10min: 5767   
## 3rd Qu.:5228 >55 :2517 10+ min: 2528   
## Max. :5228   
##

# Run Initial Logistic Regression  
#Simple regression model  
simple.log<-glm(y~.,family="binomial",data=train)  
summary(simple.log)

##   
## Call:  
## glm(formula = y ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.2555 -0.3019 -0.1699 -0.0747 3.8101   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.237e+02 5.113e+01 -4.376 1.21e-05 \*\*\*  
## age 6.312e-03 7.215e-03 0.875 0.381706   
## jobblue-collar -2.374e-01 9.380e-02 -2.531 0.011365 \*   
## jobentrepreneur -1.252e-01 1.442e-01 -0.868 0.385468   
## jobhousemaid -1.951e-01 1.741e-01 -1.121 0.262439   
## jobmanagement -4.918e-02 1.013e-01 -0.486 0.627280   
## jobretired -2.785e-02 1.425e-01 -0.195 0.845094   
## jobself-employed -8.184e-02 1.351e-01 -0.606 0.544544   
## jobservices -1.555e-01 1.012e-01 -1.537 0.124357   
## jobstudent 2.003e-01 1.442e-01 1.389 0.164982   
## jobtechnician 4.717e-02 8.248e-02 0.572 0.567409   
## jobunemployed -6.885e-02 1.585e-01 -0.434 0.664067   
## maritalmarried 1.548e-02 8.041e-02 0.193 0.847318   
## maritalsingle 3.812e-02 9.225e-02 0.413 0.679413   
## educationbasic.6y 2.410e-01 1.366e-01 1.764 0.077790 .   
## educationbasic.9y 3.172e-02 1.095e-01 0.290 0.771955   
## educationhigh.school 5.810e-02 1.080e-01 0.538 0.590620   
## educationilliterate 1.664e+00 8.360e-01 1.990 0.046552 \*   
## educationprofessional.course 1.227e-01 1.185e-01 1.035 0.300542   
## educationuniversity.degree 1.967e-01 1.087e-01 1.808 0.070529 .   
## defaultunknown -3.050e-01 7.805e-02 -3.908 9.32e-05 \*\*\*  
## defaultyes -7.098e+00 1.114e+02 -0.064 0.949192   
## housingyes -5.419e-03 4.810e-02 -0.113 0.910285   
## loanyes -2.510e-02 6.693e-02 -0.375 0.707649   
## contacttelephone -5.541e-01 9.352e-02 -5.925 3.12e-09 \*\*\*  
## monthaug 1.031e+00 1.490e-01 6.924 4.40e-12 \*\*\*  
## monthdec 2.616e-02 2.677e-01 0.098 0.922158   
## monthjul 2.773e-01 1.149e-01 2.414 0.015797 \*   
## monthjun -4.745e-01 1.551e-01 -3.060 0.002216 \*\*   
## monthmar 2.427e+00 1.778e-01 13.645 < 2e-16 \*\*\*  
## monthmay -3.575e-01 9.969e-02 -3.586 0.000335 \*\*\*  
## monthnov -4.124e-01 1.421e-01 -2.901 0.003718 \*\*   
## monthoct 1.987e-01 1.839e-01 1.081 0.279774   
## monthsep 3.898e-01 2.174e-01 1.793 0.073022 .   
## day\_of\_weekmon -8.544e-02 7.728e-02 -1.106 0.268899   
## day\_of\_weekthu 1.800e-02 7.509e-02 0.240 0.810596   
## day\_of\_weektue 4.057e-02 7.700e-02 0.527 0.598268   
## day\_of\_weekwed 7.237e-02 7.720e-02 0.937 0.348528   
## duration 1.966e-03 1.516e-04 12.966 < 2e-16 \*\*\*  
## campaign -2.259e-02 1.334e-02 -1.694 0.090304 .   
## pdays -2.100e-02 2.129e-02 -0.986 0.324032   
## previous -3.814e-02 7.313e-02 -0.522 0.602000   
## poutcomenonexistent 4.374e-01 1.151e-01 3.799 0.000146 \*\*\*  
## poutcomesuccess 8.901e-01 2.815e-01 3.162 0.001567 \*\*   
## emp\_var\_rate -1.855e+00 1.766e-01 -10.504 < 2e-16 \*\*\*  
## cons\_price\_idx 2.255e+00 3.068e-01 7.349 2.00e-13 \*\*\*  
## cons\_conf\_idx 2.636e-02 9.359e-03 2.817 0.004850 \*\*   
## euribor3m 4.077e-01 1.520e-01 2.682 0.007313 \*\*   
## nr\_employed 4.775e-03 3.704e-03 1.289 0.197298   
## Age\_Grp32-37 -2.228e-01 8.166e-02 -2.729 0.006353 \*\*   
## Age\_Grp38-47 -3.798e-01 1.202e-01 -3.159 0.001584 \*\*   
## Age\_Grp47-55 -2.672e-01 1.775e-01 -1.506 0.132181   
## Age\_Grp>55 -1.969e-01 2.498e-01 -0.788 0.430526   
## prevly\_CntctdYes -1.989e+01 2.104e+01 -0.945 0.344644   
## duration\_group30-60s 3.341e+00 1.011e+00 3.306 0.000945 \*\*\*  
## duration\_group1-5 min 4.678e+00 1.003e+00 4.666 3.08e-06 \*\*\*  
## duration\_group5-10min 5.690e+00 1.004e+00 5.665 1.47e-08 \*\*\*  
## duration\_group10+ min 6.850e+00 1.011e+00 6.772 1.27e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12078 on 30537 degrees of freedom  
## AIC: 12194  
##   
## Number of Fisher Scoring iterations: 10

exp(cbind("Odds ratio" = coef(simple.log), confint.default(simple.log, level = 0.95)))

## Odds ratio 2.5 % 97.5 %  
## (Intercept) 6.725005e-98 2.003684e-141 2.257127e-54  
## age 1.006332e+00 9.922005e-01 1.020664e+00  
## jobblue-collar 7.886577e-01 6.562197e-01 9.478242e-01  
## jobentrepreneur 8.823527e-01 6.650927e-01 1.170583e+00  
## jobhousemaid 8.227639e-01 5.849178e-01 1.157326e+00  
## jobmanagement 9.520051e-01 7.805780e-01 1.161080e+00  
## jobretired 9.725348e-01 7.354888e-01 1.285980e+00  
## jobself-employed 9.214232e-01 7.071350e-01 1.200649e+00  
## jobservices 8.559508e-01 7.019299e-01 1.043768e+00  
## jobstudent 1.221711e+00 9.208887e-01 1.620802e+00  
## jobtechnician 1.048297e+00 8.918223e-01 1.232227e+00  
## jobunemployed 9.334666e-01 6.841612e-01 1.273618e+00  
## maritalmarried 1.015604e+00 8.675116e-01 1.188976e+00  
## maritalsingle 1.038859e+00 8.670302e-01 1.244741e+00  
## educationbasic.6y 1.272476e+00 9.735348e-01 1.663213e+00  
## educationbasic.9y 1.032233e+00 8.329191e-01 1.279242e+00  
## educationhigh.school 1.059823e+00 8.576230e-01 1.309695e+00  
## educationilliterate 5.279967e+00 1.025735e+00 2.717862e+01  
## educationprofessional.course 1.130532e+00 8.962075e-01 1.426122e+00  
## educationuniversity.degree 1.217341e+00 9.836631e-01 1.506531e+00  
## defaultunknown 7.371258e-01 6.325607e-01 8.589759e-01  
## defaultyes 8.264439e-04 1.247128e-98 5.476660e+91  
## housingyes 9.945952e-01 9.051217e-01 1.092913e+00  
## loanyes 9.752131e-01 8.553220e-01 1.111909e+00  
## contacttelephone 5.745860e-01 4.783581e-01 6.901714e-01  
## monthaug 2.804782e+00 2.094622e+00 3.755714e+00  
## monthdec 1.026504e+00 6.074180e-01 1.734738e+00  
## monthjul 1.319511e+00 1.053489e+00 1.652706e+00  
## monthjun 6.222133e-01 4.591317e-01 8.432208e-01  
## monthmar 1.132094e+01 7.989286e+00 1.604195e+01  
## monthmay 6.994087e-01 5.752714e-01 8.503333e-01  
## monthnov 6.620835e-01 5.010991e-01 8.747862e-01  
## monthoct 1.219847e+00 8.507457e-01 1.749084e+00  
## monthsep 1.476700e+00 9.642813e-01 2.261418e+00  
## day\_of\_weekmon 9.181106e-01 7.890725e-01 1.068251e+00  
## day\_of\_weekthu 1.018158e+00 8.788217e-01 1.179586e+00  
## day\_of\_weektue 1.041405e+00 8.955249e-01 1.211048e+00  
## day\_of\_weekwed 1.075055e+00 9.240960e-01 1.250675e+00  
## duration 1.001967e+00 1.001670e+00 1.002265e+00  
## campaign 9.776592e-01 9.524299e-01 1.003557e+00  
## pdays 9.792227e-01 9.392030e-01 1.020948e+00  
## previous 9.625800e-01 8.340467e-01 1.110921e+00  
## poutcomenonexistent 1.548636e+00 1.235775e+00 1.940704e+00  
## poutcomesuccess 2.435287e+00 1.402646e+00 4.228168e+00  
## emp\_var\_rate 1.564696e-01 1.106936e-01 2.211757e-01  
## cons\_price\_idx 9.530979e+00 5.223914e+00 1.738918e+01  
## cons\_conf\_idx 1.026714e+00 1.008052e+00 1.045721e+00  
## euribor3m 1.503374e+00 1.116043e+00 2.025130e+00  
## nr\_employed 1.004787e+00 9.975191e-01 1.012107e+00  
## Age\_Grp32-37 8.002385e-01 6.818852e-01 9.391341e-01  
## Age\_Grp38-47 6.839849e-01 5.403803e-01 8.657521e-01  
## Age\_Grp47-55 7.655233e-01 5.406212e-01 1.083986e+00  
## Age\_Grp>55 8.212726e-01 5.033542e-01 1.339988e+00  
## prevly\_CntctdYes 2.308234e-09 2.824609e-27 1.886258e+09  
## duration\_group30-60s 2.825968e+01 3.899160e+00 2.048158e+02  
## duration\_group1-5 min 1.075126e+02 1.506790e+01 7.671250e+02  
## duration\_group5-10min 2.957753e+02 4.131777e+01 2.117322e+03  
## duration\_group10+ min 9.434254e+02 1.299464e+02 6.849375e+03

vif(simple.log)

## GVIF Df GVIF^(1/(2\*Df))  
## age 13.745120 1 3.707441  
## job 7.673807 10 1.107262  
## marital 1.482129 2 1.103371  
## education 3.334047 6 1.105557  
## default 1.141909 2 1.033732  
## housing 1.017174 1 1.008551  
## loan 1.008549 1 1.004265  
## contact 2.625712 1 1.620405  
## month 77.857569 9 1.273715  
## day\_of\_week 1.071447 4 1.008664  
## duration 4.484820 1 2.117739  
## campaign 1.061497 1 1.030290  
## pdays 71068.321281 1 266.586424  
## previous 4.639201 1 2.153880  
## poutcome 29.559795 2 2.331714  
## emp\_var\_rate 165.442323 1 12.862438  
## cons\_price\_idx 71.813071 1 8.474259  
## cons\_conf\_idx 5.314990 1 2.305426  
## euribor3m 140.016392 1 11.832852  
## nr\_employed 177.417100 1 13.319801  
## Age\_Grp 16.961935 4 1.424572  
## prevly\_Cntctd 70401.755004 1 265.333290  
## duration\_group 4.960031 4 1.221618

#Remove variables with high vifs and run the model again  
#pdays / prevly\_Cntctd  
#emp\_var\_rate/euribor3m/nr\_employed  
#age/ Age\_Grp  
#duration/duration\_group  
train\_simple <- train %>% dplyr::select(-age, -pdays,-emp\_var\_rate, -duration\_group )  
  
#Check vifs again  
simple.log<-glm(y~.,family="binomial",data=train\_simple)  
summary(simple.log)

##   
## Call:  
## glm(formula = y ~ ., family = "binomial", data = train\_simple)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.0394 -0.3043 -0.1906 -0.1343 3.1614   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.163e+02 2.875e+01 4.045 5.24e-05 \*\*\*  
## jobblue-collar -2.634e-01 9.337e-02 -2.821 0.004782 \*\*   
## jobentrepreneur -1.386e-01 1.442e-01 -0.961 0.336346   
## jobhousemaid -1.941e-01 1.728e-01 -1.123 0.261390   
## jobmanagement -2.254e-02 9.923e-02 -0.227 0.820324   
## jobretired 3.889e-02 1.315e-01 0.296 0.767384   
## jobself-employed -9.433e-02 1.339e-01 -0.705 0.481058   
## jobservices -1.525e-01 1.002e-01 -1.522 0.127997   
## jobstudent 1.862e-01 1.353e-01 1.376 0.168767   
## jobtechnician 1.459e-02 8.115e-02 0.180 0.857321   
## jobunemployed -6.369e-02 1.514e-01 -0.421 0.674078   
## maritalmarried 1.124e-02 7.965e-02 0.141 0.887815   
## maritalsingle 2.940e-02 9.078e-02 0.324 0.746019   
## educationbasic.6y 2.121e-01 1.371e-01 1.547 0.121909   
## educationbasic.9y 2.357e-02 1.091e-01 0.216 0.828942   
## educationhigh.school 4.754e-02 1.064e-01 0.447 0.655146   
## educationilliterate 1.550e+00 8.029e-01 1.931 0.053510 .   
## educationprofessional.course 9.065e-02 1.168e-01 0.776 0.437649   
## educationuniversity.degree 1.738e-01 1.070e-01 1.625 0.104197   
## defaultunknown -3.309e-01 7.983e-02 -4.146 3.39e-05 \*\*\*  
## defaultyes -7.495e+00 1.131e+02 -0.066 0.947143   
## housingyes -7.489e-03 4.745e-02 -0.158 0.874593   
## loanyes -7.859e-02 6.635e-02 -1.184 0.236234   
## contacttelephone -3.608e-01 8.078e-02 -4.466 7.97e-06 \*\*\*  
## monthaug 7.133e-02 1.211e-01 0.589 0.555914   
## monthdec -3.348e-01 2.525e-01 -1.326 0.184849   
## monthjul 2.945e-01 1.108e-01 2.658 0.007868 \*\*   
## monthjun 5.255e-01 1.079e-01 4.870 1.11e-06 \*\*\*  
## monthmar 1.311e+00 1.505e-01 8.711 < 2e-16 \*\*\*  
## monthmay -7.029e-01 8.983e-02 -7.825 5.08e-15 \*\*\*  
## monthnov -4.441e-01 1.424e-01 -3.119 0.001818 \*\*   
## monthoct -2.536e-01 1.788e-01 -1.418 0.156074   
## monthsep -6.822e-01 1.912e-01 -3.568 0.000360 \*\*\*  
## day\_of\_weekmon -1.076e-01 7.608e-02 -1.414 0.157333   
## day\_of\_weekthu 2.430e-02 7.404e-02 0.328 0.742799   
## day\_of\_weektue 3.863e-02 7.607e-02 0.508 0.611580   
## day\_of\_weekwed 6.336e-02 7.643e-02 0.829 0.407117   
## duration 4.727e-03 8.673e-05 54.501 < 2e-16 \*\*\*  
## campaign -4.283e-02 1.342e-02 -3.191 0.001416 \*\*   
## previous -2.909e-02 6.945e-02 -0.419 0.675258   
## poutcomenonexistent 4.529e-01 1.107e-01 4.090 4.32e-05 \*\*\*  
## poutcomesuccess 8.091e-01 2.464e-01 3.283 0.001027 \*\*   
## cons\_price\_idx -3.587e-01 1.581e-01 -2.268 0.023323 \*   
## cons\_conf\_idx 1.377e-02 8.960e-03 1.537 0.124291   
## euribor3m 1.673e-01 1.540e-01 1.087 0.277179   
## nr\_employed -1.683e-02 2.951e-03 -5.703 1.18e-08 \*\*\*  
## Age\_Grp32-37 -1.924e-01 6.917e-02 -2.781 0.005414 \*\*   
## Age\_Grp38-47 -2.656e-01 7.584e-02 -3.502 0.000461 \*\*\*  
## Age\_Grp47-55 -1.165e-01 8.587e-02 -1.356 0.174998   
## Age\_Grp>55 2.510e-02 1.133e-01 0.221 0.824746   
## prevly\_CntctdYes 1.003e+00 2.512e-01 3.995 6.48e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12734 on 30544 degrees of freedom  
## AIC: 12836  
##   
## Number of Fisher Scoring iterations: 10

exp(cbind("Odds ratio" = coef(simple.log), confint.default(simple.log, level = 0.95)))

## Odds ratio 2.5 % 97.5 %  
## (Intercept) 3.119590e+50 1.060453e+26 9.177064e+74  
## jobblue-collar 7.684087e-01 6.399033e-01 9.227205e-01  
## jobentrepreneur 8.705733e-01 6.562821e-01 1.154836e+00  
## jobhousemaid 8.235768e-01 5.869451e-01 1.155608e+00  
## jobmanagement 9.777150e-01 8.049161e-01 1.187610e+00  
## jobretired 1.039657e+00 8.034834e-01 1.345250e+00  
## jobself-employed 9.099828e-01 6.999664e-01 1.183012e+00  
## jobservices 8.585731e-01 7.055068e-01 1.044849e+00  
## jobstudent 1.204615e+00 9.240666e-01 1.570339e+00  
## jobtechnician 1.014697e+00 8.654845e-01 1.189634e+00  
## jobunemployed 9.382973e-01 6.973264e-01 1.262539e+00  
## maritalmarried 1.011300e+00 8.651280e-01 1.182169e+00  
## maritalsingle 1.029839e+00 8.619834e-01 1.230381e+00  
## educationbasic.6y 1.236276e+00 9.449221e-01 1.617464e+00  
## educationbasic.9y 1.023850e+00 8.267545e-01 1.267933e+00  
## educationhigh.school 1.048684e+00 8.512327e-01 1.291935e+00  
## educationilliterate 4.712884e+00 9.768398e-01 2.273789e+01  
## educationprofessional.course 1.094890e+00 8.708699e-01 1.376535e+00  
## educationuniversity.degree 1.189797e+00 9.647919e-01 1.467276e+00  
## defaultunknown 7.182608e-01 6.142352e-01 8.399040e-01  
## defaultyes 5.559322e-04 3.262056e-100 9.474414e+92  
## housingyes 9.925387e-01 9.043906e-01 1.089278e+00  
## loanyes 9.244150e-01 8.116816e-01 1.052806e+00  
## contacttelephone 6.971474e-01 5.950644e-01 8.167426e-01  
## monthaug 1.073933e+00 8.469997e-01 1.361668e+00  
## monthdec 7.154528e-01 4.361462e-01 1.173626e+00  
## monthjul 1.342444e+00 1.080379e+00 1.668077e+00  
## monthjun 1.691240e+00 1.368885e+00 2.089505e+00  
## monthmar 3.710245e+00 2.762369e+00 4.983374e+00  
## monthmay 4.951241e-01 4.151898e-01 5.904477e-01  
## monthnov 6.413907e-01 4.851770e-01 8.479008e-01  
## monthoct 7.760196e-01 5.466347e-01 1.101662e+00  
## monthsep 5.055058e-01 3.475171e-01 7.353196e-01  
## day\_of\_weekmon 8.979994e-01 7.735989e-01 1.042404e+00  
## day\_of\_weekthu 1.024593e+00 8.861983e-01 1.184600e+00  
## day\_of\_weektue 1.039385e+00 8.954207e-01 1.206495e+00  
## day\_of\_weekwed 1.065409e+00 9.171895e-01 1.237581e+00  
## duration 1.004738e+00 1.004567e+00 1.004909e+00  
## campaign 9.580756e-01 9.332046e-01 9.836095e-01  
## previous 9.713249e-01 8.477176e-01 1.112956e+00  
## poutcomenonexistent 1.572833e+00 1.265963e+00 1.954088e+00  
## poutcomesuccess 2.245907e+00 1.385533e+00 3.640547e+00  
## cons\_price\_idx 6.986076e-01 5.124227e-01 9.524413e-01  
## cons\_conf\_idx 1.013867e+00 9.962175e-01 1.031829e+00  
## euribor3m 1.182144e+00 8.741780e-01 1.598605e+00  
## nr\_employed 9.833119e-01 9.776414e-01 9.890153e-01  
## Age\_Grp32-37 8.249791e-01 7.203798e-01 9.447664e-01  
## Age\_Grp38-47 7.667189e-01 6.608123e-01 8.895989e-01  
## Age\_Grp47-55 8.900635e-01 7.521964e-01 1.053200e+00  
## Age\_Grp>55 1.025417e+00 8.211466e-01 1.280503e+00  
## prevly\_CntctdYes 2.727203e+00 1.666972e+00 4.461764e+00

vif(simple.log)

## GVIF Df GVIF^(1/(2\*Df))  
## job 6.408974 10 1.097335  
## marital 1.471680 2 1.101421  
## education 3.229283 6 1.102619  
## default 1.139480 2 1.033182  
## housing 1.013695 1 1.006824  
## loan 1.006201 1 1.003096  
## contact 1.882826 1 1.372161  
## month 14.549836 9 1.160388  
## day\_of\_week 1.065810 4 1.007999  
## duration 1.240907 1 1.113960  
## campaign 1.052855 1 1.026087  
## previous 4.493192 1 2.119715  
## poutcome 24.412119 2 2.222805  
## cons\_price\_idx 19.329202 1 4.396499  
## cons\_conf\_idx 5.242392 1 2.289627  
## euribor3m 141.924791 1 11.913219  
## nr\_employed 115.526284 1 10.748315  
## Age\_Grp 2.778832 4 1.136273  
## prevly\_Cntctd 10.983033 1 3.314066

#VIFs are still high for euribor3m and nr\_employed, but model shows euribor3m as insignificant. As euribor3m looks important practically, remove nr\_employed and see if things change.  
train\_simple\_2 <- train\_simple %>% dplyr::select(-nr\_employed )  
  
simple.log<-glm(y~.,family="binomial",data=train\_simple\_2)  
summary(simple.log)

##   
## Call:  
## glm(formula = y ~ ., family = "binomial", data = train\_simple\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.9913 -0.3099 -0.1921 -0.1326 3.2225   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.510e+01 5.087e+00 -8.866 < 2e-16 \*\*\*  
## jobblue-collar -2.780e-01 9.326e-02 -2.981 0.002876 \*\*   
## jobentrepreneur -1.522e-01 1.440e-01 -1.057 0.290618   
## jobhousemaid -1.930e-01 1.730e-01 -1.115 0.264787   
## jobmanagement -2.514e-02 9.908e-02 -0.254 0.799717   
## jobretired 4.293e-02 1.314e-01 0.327 0.743977   
## jobself-employed -9.090e-02 1.337e-01 -0.680 0.496743   
## jobservices -1.498e-01 1.000e-01 -1.498 0.134110   
## jobstudent 1.898e-01 1.352e-01 1.404 0.160433   
## jobtechnician 2.268e-02 8.101e-02 0.280 0.779504   
## jobunemployed -4.812e-02 1.510e-01 -0.319 0.749995   
## maritalmarried 1.298e-02 7.955e-02 0.163 0.870406   
## maritalsingle 3.074e-02 9.064e-02 0.339 0.734466   
## educationbasic.6y 2.054e-01 1.370e-01 1.499 0.133761   
## educationbasic.9y 1.469e-02 1.090e-01 0.135 0.892783   
## educationhigh.school 3.676e-02 1.063e-01 0.346 0.729443   
## educationilliterate 1.513e+00 8.063e-01 1.876 0.060618 .   
## educationprofessional.course 8.686e-02 1.167e-01 0.744 0.456851   
## educationuniversity.degree 1.790e-01 1.069e-01 1.675 0.093975 .   
## defaultunknown -3.415e-01 7.973e-02 -4.283 1.85e-05 \*\*\*  
## defaultyes -7.487e+00 1.131e+02 -0.066 0.947210   
## housingyes 4.404e-04 4.738e-02 0.009 0.992584   
## loanyes -7.928e-02 6.621e-02 -1.197 0.231189   
## contacttelephone -3.878e-01 8.100e-02 -4.787 1.69e-06 \*\*\*  
## monthaug 2.243e-01 1.189e-01 1.886 0.059298 .   
## monthdec 1.105e-01 2.405e-01 0.459 0.646071   
## monthjul 3.652e-01 1.107e-01 3.298 0.000975 \*\*\*  
## monthjun 3.898e-01 1.064e-01 3.664 0.000248 \*\*\*  
## monthmar 1.665e+00 1.406e-01 11.840 < 2e-16 \*\*\*  
## monthmay -5.883e-01 8.713e-02 -6.752 1.46e-11 \*\*\*  
## monthnov 3.654e-02 1.151e-01 0.317 0.750883   
## monthoct 3.639e-01 1.427e-01 2.551 0.010755 \*   
## monthsep -5.162e-02 1.565e-01 -0.330 0.741555   
## day\_of\_weekmon -9.855e-02 7.596e-02 -1.297 0.194492   
## day\_of\_weekthu 8.130e-03 7.395e-02 0.110 0.912458   
## day\_of\_weektue 5.723e-02 7.590e-02 0.754 0.450812   
## day\_of\_weekwed 7.288e-02 7.629e-02 0.955 0.339445   
## duration 4.712e-03 8.656e-05 54.434 < 2e-16 \*\*\*  
## campaign -4.589e-02 1.343e-02 -3.416 0.000635 \*\*\*  
## previous -2.130e-02 6.924e-02 -0.308 0.758347   
## poutcomenonexistent 4.855e-01 1.104e-01 4.398 1.09e-05 \*\*\*  
## poutcomesuccess 8.004e-01 2.458e-01 3.256 0.001129 \*\*   
## cons\_price\_idx 4.853e-01 5.595e-02 8.674 < 2e-16 \*\*\*  
## cons\_conf\_idx 5.205e-02 6.011e-03 8.659 < 2e-16 \*\*\*  
## euribor3m -7.038e-01 2.195e-02 -32.069 < 2e-16 \*\*\*  
## Age\_Grp32-37 -1.990e-01 6.903e-02 -2.883 0.003945 \*\*   
## Age\_Grp38-47 -2.709e-01 7.572e-02 -3.577 0.000347 \*\*\*  
## Age\_Grp47-55 -1.174e-01 8.575e-02 -1.370 0.170803   
## Age\_Grp>55 1.450e-02 1.133e-01 0.128 0.898160   
## prevly\_CntctdYes 1.028e+00 2.506e-01 4.101 4.10e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12767 on 30545 degrees of freedom  
## AIC: 12867  
##   
## Number of Fisher Scoring iterations: 10

exp(cbind("Odds ratio" = coef(simple.log), confint.default(simple.log, level = 0.95)))

## Odds ratio 2.5 % 97.5 %  
## (Intercept) 2.591932e-20 1.212127e-24 5.542414e-16  
## jobblue-collar 7.573234e-01 6.308151e-01 9.092024e-01  
## jobentrepreneur 8.588174e-01 6.476019e-01 1.138921e+00  
## jobhousemaid 8.245149e-01 5.873681e-01 1.157408e+00  
## jobmanagement 9.751746e-01 8.030499e-01 1.184192e+00  
## jobretired 1.043862e+00 8.067900e-01 1.350597e+00  
## jobself-employed 9.131127e-01 7.025557e-01 1.186774e+00  
## jobservices 8.608526e-01 7.076126e-01 1.047278e+00  
## jobstudent 1.209012e+00 9.275284e-01 1.575918e+00  
## jobtechnician 1.022939e+00 8.727608e-01 1.198958e+00  
## jobunemployed 9.530190e-01 7.088546e-01 1.281286e+00  
## maritalmarried 1.013062e+00 8.668153e-01 1.183983e+00  
## maritalsingle 1.031222e+00 8.633718e-01 1.231705e+00  
## educationbasic.6y 1.227964e+00 9.388755e-01 1.606066e+00  
## educationbasic.9y 1.014799e+00 8.196002e-01 1.256488e+00  
## educationhigh.school 1.037448e+00 8.423413e-01 1.277746e+00  
## educationilliterate 4.539932e+00 9.347428e-01 2.204989e+01  
## educationprofessional.course 1.090742e+00 8.676692e-01 1.371165e+00  
## educationuniversity.degree 1.195989e+00 9.699854e-01 1.474651e+00  
## defaultunknown 7.107121e-01 6.078878e-01 8.309292e-01  
## defaultyes 5.604074e-04 3.151791e-100 9.964382e+92  
## housingyes 1.000441e+00 9.117117e-01 1.097805e+00  
## loanyes 9.237821e-01 8.113484e-01 1.051796e+00  
## contacttelephone 6.785443e-01 5.789310e-01 7.952976e-01  
## monthaug 1.251481e+00 9.912383e-01 1.580048e+00  
## monthdec 1.116792e+00 6.969934e-01 1.789434e+00  
## monthjul 1.440783e+00 1.159674e+00 1.790034e+00  
## monthjun 1.476731e+00 1.198788e+00 1.819116e+00  
## monthmar 5.284881e+00 4.011881e+00 6.961814e+00  
## monthmay 5.552822e-01 4.681136e-01 6.586829e-01  
## monthnov 1.037221e+00 8.277294e-01 1.299733e+00  
## monthoct 1.438868e+00 1.087905e+00 1.903051e+00  
## monthsep 9.496909e-01 6.988022e-01 1.290655e+00  
## day\_of\_weekmon 9.061512e-01 7.808086e-01 1.051615e+00  
## day\_of\_weekthu 1.008163e+00 8.721381e-01 1.165403e+00  
## day\_of\_weektue 1.058900e+00 9.125403e-01 1.228733e+00  
## day\_of\_weekwed 1.075597e+00 9.262167e-01 1.249068e+00  
## duration 1.004723e+00 1.004552e+00 1.004893e+00  
## campaign 9.551429e-01 9.303212e-01 9.806269e-01  
## previous 9.789226e-01 8.546908e-01 1.121212e+00  
## poutcomenonexistent 1.625000e+00 1.308846e+00 2.017521e+00  
## poutcomesuccess 2.226351e+00 1.375206e+00 3.604288e+00  
## cons\_price\_idx 1.624684e+00 1.455942e+00 1.812983e+00  
## cons\_conf\_idx 1.053428e+00 1.041089e+00 1.065912e+00  
## euribor3m 4.946826e-01 4.738540e-01 5.164267e-01  
## Age\_Grp32-37 8.195595e-01 7.158462e-01 9.382989e-01  
## Age\_Grp38-47 7.627221e-01 6.575318e-01 8.847406e-01  
## Age\_Grp47-55 8.891947e-01 7.516407e-01 1.051922e+00  
## Age\_Grp>55 1.014602e+00 8.126171e-01 1.266792e+00  
## prevly\_CntctdYes 2.794727e+00 1.710209e+00 4.566984e+00

vif(simple.log)

## GVIF Df GVIF^(1/(2\*Df))  
## job 6.389262 10 1.097166  
## marital 1.469759 2 1.101061  
## education 3.223042 6 1.102441  
## default 1.137007 2 1.032621  
## housing 1.013097 1 1.006527  
## loan 1.006256 1 1.003123  
## contact 1.893549 1 1.376063  
## month 5.171191 9 1.095580  
## day\_of\_week 1.056497 4 1.006893  
## duration 1.232599 1 1.110225  
## campaign 1.050912 1 1.025140  
## previous 4.487638 1 2.118405  
## poutcome 24.296882 2 2.220177  
## cons\_price\_idx 2.485647 1 1.576593  
## cons\_conf\_idx 2.356210 1 1.534995  
## euribor3m 2.888869 1 1.699667  
## Age\_Grp 2.773307 4 1.135991  
## prevly\_Cntctd 11.002778 1 3.317044

#poutcome and prevly\_Cntctd have higher vifs but let's keeo both of them.  
  
#Remove statistically insignificant variables and run the model again  
train\_simple\_3 <- train\_simple\_2 %>% dplyr::select(-marital,-day\_of\_week, -loan, -housing,-previous )  
  
#Check model again  
simple.log<-glm(y~.,family="binomial",data=train\_simple\_3)  
summary(simple.log)

##   
## Call:  
## glm(formula = y ~ ., family = "binomial", data = train\_simple\_3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.0104 -0.3102 -0.1920 -0.1327 3.2406   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.439e+01 4.926e+00 -9.011 < 2e-16 \*\*\*  
## jobblue-collar -2.732e-01 9.301e-02 -2.937 0.003315 \*\*   
## jobentrepreneur -1.507e-01 1.436e-01 -1.050 0.293791   
## jobhousemaid -1.867e-01 1.727e-01 -1.081 0.279701   
## jobmanagement -2.418e-02 9.852e-02 -0.245 0.806087   
## jobretired 4.433e-02 1.313e-01 0.338 0.735705   
## jobself-employed -9.209e-02 1.336e-01 -0.689 0.490588   
## jobservices -1.560e-01 9.988e-02 -1.562 0.118369   
## jobstudent 1.976e-01 1.340e-01 1.475 0.140328   
## jobtechnician 2.203e-02 8.098e-02 0.272 0.785604   
## jobunemployed -5.093e-02 1.510e-01 -0.337 0.735837   
## educationbasic.6y 2.086e-01 1.368e-01 1.524 0.127425   
## educationbasic.9y 1.443e-02 1.089e-01 0.132 0.894601   
## educationhigh.school 4.111e-02 1.062e-01 0.387 0.698609   
## educationilliterate 1.539e+00 8.034e-01 1.916 0.055384 .   
## educationprofessional.course 9.046e-02 1.167e-01 0.775 0.438181   
## educationuniversity.degree 1.796e-01 1.066e-01 1.684 0.092092 .   
## defaultunknown -3.428e-01 7.965e-02 -4.304 1.68e-05 \*\*\*  
## defaultyes -7.432e+00 1.131e+02 -0.066 0.947607   
## contacttelephone -3.855e-01 8.081e-02 -4.771 1.83e-06 \*\*\*  
## monthaug 2.328e-01 1.181e-01 1.972 0.048602 \*   
## monthdec 1.023e-01 2.405e-01 0.425 0.670594   
## monthjul 3.766e-01 1.102e-01 3.416 0.000636 \*\*\*  
## monthjun 4.001e-01 1.055e-01 3.792 0.000150 \*\*\*  
## monthmar 1.671e+00 1.400e-01 11.939 < 2e-16 \*\*\*  
## monthmay -5.774e-01 8.644e-02 -6.680 2.39e-11 \*\*\*  
## monthnov 4.631e-02 1.143e-01 0.405 0.685480   
## monthoct 3.747e-01 1.423e-01 2.633 0.008462 \*\*   
## monthsep -3.693e-02 1.560e-01 -0.237 0.812925   
## duration 4.717e-03 8.649e-05 54.535 < 2e-16 \*\*\*  
## campaign -4.788e-02 1.346e-02 -3.557 0.000375 \*\*\*  
## poutcomenonexistent 5.079e-01 7.470e-02 6.799 1.05e-11 \*\*\*  
## poutcomesuccess 8.307e-01 2.354e-01 3.529 0.000417 \*\*\*  
## cons\_price\_idx 4.777e-01 5.403e-02 8.841 < 2e-16 \*\*\*  
## cons\_conf\_idx 5.245e-02 5.996e-03 8.747 < 2e-16 \*\*\*  
## euribor3m -7.014e-01 2.161e-02 -32.453 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.052e-01 6.702e-02 -3.062 0.002196 \*\*   
## Age\_Grp38-47 -2.838e-01 7.153e-02 -3.968 7.25e-05 \*\*\*  
## Age\_Grp47-55 -1.332e-01 7.986e-02 -1.668 0.095233 .   
## Age\_Grp>55 -1.227e-03 1.082e-01 -0.011 0.990950   
## prevly\_CntctdYes 9.892e-01 2.336e-01 4.235 2.29e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12775 on 30554 degrees of freedom  
## AIC: 12857  
##   
## Number of Fisher Scoring iterations: 10

exp(cbind("Odds ratio" = coef(simple.log), confint.default(simple.log, level = 0.95)))

## Odds ratio 2.5 % 97.5 %  
## (Intercept) 5.284448e-20 3.390229e-24 8.237022e-16  
## jobblue-collar 7.609579e-01 6.341421e-01 9.131344e-01  
## jobentrepreneur 8.600664e-01 6.490958e-01 1.139607e+00  
## jobhousemaid 8.297225e-01 5.914914e-01 1.163904e+00  
## jobmanagement 9.761058e-01 8.047053e-01 1.184014e+00  
## jobretired 1.045332e+00 8.080786e-01 1.352244e+00  
## jobself-employed 9.120261e-01 7.019457e-01 1.184980e+00  
## jobservices 8.555726e-01 7.034538e-01 1.040586e+00  
## jobstudent 1.218457e+00 9.370297e-01 1.584408e+00  
## jobtechnician 1.022273e+00 8.722384e-01 1.198116e+00  
## jobunemployed 9.503477e-01 7.069532e-01 1.277540e+00  
## educationbasic.6y 1.231956e+00 9.421272e-01 1.610945e+00  
## educationbasic.9y 1.014538e+00 8.194760e-01 1.256031e+00  
## educationhigh.school 1.041962e+00 8.462285e-01 1.282968e+00  
## educationilliterate 4.660660e+00 9.651792e-01 2.250541e+01  
## educationprofessional.course 1.094682e+00 8.708904e-01 1.375982e+00  
## educationuniversity.degree 1.196786e+00 9.710469e-01 1.475003e+00  
## defaultunknown 7.097759e-01 6.071880e-01 8.296967e-01  
## defaultyes 5.921922e-04 3.213586e-100 1.091278e+93  
## contacttelephone 6.800841e-01 5.804664e-01 7.967978e-01  
## monthaug 1.262189e+00 1.001430e+00 1.590845e+00  
## monthdec 1.107695e+00 6.913993e-01 1.774644e+00  
## monthjul 1.457279e+00 1.174104e+00 1.808752e+00  
## monthjun 1.491935e+00 1.213214e+00 1.834689e+00  
## monthmar 5.318101e+00 4.042197e+00 6.996740e+00  
## monthmay 5.613487e-01 4.738635e-01 6.649856e-01  
## monthnov 1.047396e+00 8.371138e-01 1.310502e+00  
## monthoct 1.454552e+00 1.100524e+00 1.922469e+00  
## monthsep 9.637462e-01 7.098101e-01 1.308528e+00  
## duration 1.004728e+00 1.004557e+00 1.004898e+00  
## campaign 9.532462e-01 9.284262e-01 9.787298e-01  
## poutcomenonexistent 1.661848e+00 1.435505e+00 1.923880e+00  
## poutcomesuccess 2.294934e+00 1.446856e+00 3.640116e+00  
## cons\_price\_idx 1.612320e+00 1.450313e+00 1.792425e+00  
## cons\_conf\_idx 1.053849e+00 1.041536e+00 1.066308e+00  
## euribor3m 4.958740e-01 4.753062e-01 5.173318e-01  
## Age\_Grp32-37 8.144587e-01 7.142054e-01 9.287846e-01  
## Age\_Grp38-47 7.529018e-01 6.544157e-01 8.662095e-01  
## Age\_Grp47-55 8.752557e-01 7.484424e-01 1.023556e+00  
## Age\_Grp>55 9.987735e-01 8.079189e-01 1.234714e+00  
## prevly\_CntctdYes 2.689077e+00 1.701285e+00 4.250395e+00

vif(simple.log)

## GVIF Df GVIF^(1/(2\*Df))  
## job 6.093349 10 1.094568  
## education 3.173198 6 1.101010  
## default 1.134480 2 1.032046  
## contact 1.885721 1 1.373216  
## month 4.936162 9 1.092752  
## duration 1.231096 1 1.109548  
## campaign 1.045889 1 1.022687  
## poutcome 10.699311 2 1.808585  
## cons\_price\_idx 2.319706 1 1.523058  
## cons\_conf\_idx 2.347912 1 1.532290  
## euribor3m 2.802923 1 1.674193  
## Age\_Grp 2.282237 4 1.108652  
## prevly\_Cntctd 9.561123 1 3.092107

#simple model -1   
simple.log<-glm(y~job+education+default+contact+month+duration+campaign+poutcome+cons\_price\_idx+cons\_conf\_idx+euribor3m+Age\_Grp+prevly\_Cntctd,family="binomial",data=train)  
#simple.log<-glm(y~.,family="binomial",data=train\_simple\_3)  
summary(simple.log)

##   
## Call:  
## glm(formula = y ~ job + education + default + contact + month +   
## duration + campaign + poutcome + cons\_price\_idx + cons\_conf\_idx +   
## euribor3m + Age\_Grp + prevly\_Cntctd, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.0104 -0.3102 -0.1920 -0.1327 3.2406   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.439e+01 4.926e+00 -9.011 < 2e-16 \*\*\*  
## jobblue-collar -2.732e-01 9.301e-02 -2.937 0.003315 \*\*   
## jobentrepreneur -1.507e-01 1.436e-01 -1.050 0.293791   
## jobhousemaid -1.867e-01 1.727e-01 -1.081 0.279701   
## jobmanagement -2.418e-02 9.852e-02 -0.245 0.806087   
## jobretired 4.433e-02 1.313e-01 0.338 0.735705   
## jobself-employed -9.209e-02 1.336e-01 -0.689 0.490588   
## jobservices -1.560e-01 9.988e-02 -1.562 0.118369   
## jobstudent 1.976e-01 1.340e-01 1.475 0.140328   
## jobtechnician 2.203e-02 8.098e-02 0.272 0.785604   
## jobunemployed -5.093e-02 1.510e-01 -0.337 0.735837   
## educationbasic.6y 2.086e-01 1.368e-01 1.524 0.127425   
## educationbasic.9y 1.443e-02 1.089e-01 0.132 0.894601   
## educationhigh.school 4.111e-02 1.062e-01 0.387 0.698609   
## educationilliterate 1.539e+00 8.034e-01 1.916 0.055384 .   
## educationprofessional.course 9.046e-02 1.167e-01 0.775 0.438181   
## educationuniversity.degree 1.796e-01 1.066e-01 1.684 0.092092 .   
## defaultunknown -3.428e-01 7.965e-02 -4.304 1.68e-05 \*\*\*  
## defaultyes -7.432e+00 1.131e+02 -0.066 0.947607   
## contacttelephone -3.855e-01 8.081e-02 -4.771 1.83e-06 \*\*\*  
## monthaug 2.328e-01 1.181e-01 1.972 0.048602 \*   
## monthdec 1.023e-01 2.405e-01 0.425 0.670594   
## monthjul 3.766e-01 1.102e-01 3.416 0.000636 \*\*\*  
## monthjun 4.001e-01 1.055e-01 3.792 0.000150 \*\*\*  
## monthmar 1.671e+00 1.400e-01 11.939 < 2e-16 \*\*\*  
## monthmay -5.774e-01 8.644e-02 -6.680 2.39e-11 \*\*\*  
## monthnov 4.631e-02 1.143e-01 0.405 0.685480   
## monthoct 3.747e-01 1.423e-01 2.633 0.008462 \*\*   
## monthsep -3.693e-02 1.560e-01 -0.237 0.812925   
## duration 4.717e-03 8.649e-05 54.535 < 2e-16 \*\*\*  
## campaign -4.788e-02 1.346e-02 -3.557 0.000375 \*\*\*  
## poutcomenonexistent 5.079e-01 7.470e-02 6.799 1.05e-11 \*\*\*  
## poutcomesuccess 8.307e-01 2.354e-01 3.529 0.000417 \*\*\*  
## cons\_price\_idx 4.777e-01 5.403e-02 8.841 < 2e-16 \*\*\*  
## cons\_conf\_idx 5.245e-02 5.996e-03 8.747 < 2e-16 \*\*\*  
## euribor3m -7.014e-01 2.161e-02 -32.453 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.052e-01 6.702e-02 -3.062 0.002196 \*\*   
## Age\_Grp38-47 -2.838e-01 7.153e-02 -3.968 7.25e-05 \*\*\*  
## Age\_Grp47-55 -1.332e-01 7.986e-02 -1.668 0.095233 .   
## Age\_Grp>55 -1.227e-03 1.082e-01 -0.011 0.990950   
## prevly\_CntctdYes 9.892e-01 2.336e-01 4.235 2.29e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12775 on 30554 degrees of freedom  
## AIC: 12857  
##   
## Number of Fisher Scoring iterations: 10

exp(cbind("Odds ratio" = coef(simple.log), confint.default(simple.log, level = 0.95)))

## Odds ratio 2.5 % 97.5 %  
## (Intercept) 5.284448e-20 3.390229e-24 8.237022e-16  
## jobblue-collar 7.609579e-01 6.341421e-01 9.131344e-01  
## jobentrepreneur 8.600664e-01 6.490958e-01 1.139607e+00  
## jobhousemaid 8.297225e-01 5.914914e-01 1.163904e+00  
## jobmanagement 9.761058e-01 8.047053e-01 1.184014e+00  
## jobretired 1.045332e+00 8.080786e-01 1.352244e+00  
## jobself-employed 9.120261e-01 7.019457e-01 1.184980e+00  
## jobservices 8.555726e-01 7.034538e-01 1.040586e+00  
## jobstudent 1.218457e+00 9.370297e-01 1.584408e+00  
## jobtechnician 1.022273e+00 8.722384e-01 1.198116e+00  
## jobunemployed 9.503477e-01 7.069532e-01 1.277540e+00  
## educationbasic.6y 1.231956e+00 9.421272e-01 1.610945e+00  
## educationbasic.9y 1.014538e+00 8.194760e-01 1.256031e+00  
## educationhigh.school 1.041962e+00 8.462285e-01 1.282968e+00  
## educationilliterate 4.660660e+00 9.651792e-01 2.250541e+01  
## educationprofessional.course 1.094682e+00 8.708904e-01 1.375982e+00  
## educationuniversity.degree 1.196786e+00 9.710469e-01 1.475003e+00  
## defaultunknown 7.097759e-01 6.071880e-01 8.296967e-01  
## defaultyes 5.921922e-04 3.213586e-100 1.091278e+93  
## contacttelephone 6.800841e-01 5.804664e-01 7.967978e-01  
## monthaug 1.262189e+00 1.001430e+00 1.590845e+00  
## monthdec 1.107695e+00 6.913993e-01 1.774644e+00  
## monthjul 1.457279e+00 1.174104e+00 1.808752e+00  
## monthjun 1.491935e+00 1.213214e+00 1.834689e+00  
## monthmar 5.318101e+00 4.042197e+00 6.996740e+00  
## monthmay 5.613487e-01 4.738635e-01 6.649856e-01  
## monthnov 1.047396e+00 8.371138e-01 1.310502e+00  
## monthoct 1.454552e+00 1.100524e+00 1.922469e+00  
## monthsep 9.637462e-01 7.098101e-01 1.308528e+00  
## duration 1.004728e+00 1.004557e+00 1.004898e+00  
## campaign 9.532462e-01 9.284262e-01 9.787298e-01  
## poutcomenonexistent 1.661848e+00 1.435505e+00 1.923880e+00  
## poutcomesuccess 2.294934e+00 1.446856e+00 3.640116e+00  
## cons\_price\_idx 1.612320e+00 1.450313e+00 1.792425e+00  
## cons\_conf\_idx 1.053849e+00 1.041536e+00 1.066308e+00  
## euribor3m 4.958740e-01 4.753062e-01 5.173318e-01  
## Age\_Grp32-37 8.144587e-01 7.142054e-01 9.287846e-01  
## Age\_Grp38-47 7.529018e-01 6.544157e-01 8.662095e-01  
## Age\_Grp47-55 8.752557e-01 7.484424e-01 1.023556e+00  
## Age\_Grp>55 9.987735e-01 8.079189e-01 1.234714e+00  
## prevly\_CntctdYes 2.689077e+00 1.701285e+00 4.250395e+00

vif(simple.log)

## GVIF Df GVIF^(1/(2\*Df))  
## job 6.093349 10 1.094568  
## education 3.173198 6 1.101010  
## default 1.134480 2 1.032046  
## contact 1.885721 1 1.373216  
## month 4.936162 9 1.092752  
## duration 1.231096 1 1.109548  
## campaign 1.045889 1 1.022687  
## poutcome 10.699311 2 1.808585  
## cons\_price\_idx 2.319706 1 1.523058  
## cons\_conf\_idx 2.347912 1 1.532290  
## euribor3m 2.802923 1 1.674193  
## Age\_Grp 2.282237 4 1.108652  
## prevly\_Cntctd 9.561123 1 3.092107

#Prediction using simple model  
fit.pred.simple<-predict(simple.log,newdata=test,type="response")  
  
  
# Feature selection using step  
library(MASS)

##   
## Attaching package: 'MASS'

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

library(car) # for VIF  
full.log<-glm(y~.,family="binomial",data=train)  
step.log<-full.log %>% stepAIC(trace=FALSE)  
summary(step.log)

##   
## Call:  
## glm(formula = y ~ education + default + contact + month + duration +   
## campaign + pdays + poutcome + emp\_var\_rate + cons\_price\_idx +   
## cons\_conf\_idx + euribor3m + Age\_Grp + duration\_group, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.2519 -0.3031 -0.1705 -0.0750 3.8087   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.877e+02 1.250e+01 -15.010 < 2e-16 \*\*\*  
## educationbasic.6y 2.114e-01 1.345e-01 1.572 0.115943   
## educationbasic.9y 1.672e-02 1.065e-01 0.157 0.875206   
## educationhigh.school 1.301e-01 9.685e-02 1.343 0.179246   
## educationilliterate 1.689e+00 8.246e-01 2.049 0.040471 \*   
## educationprofessional.course 2.337e-01 1.051e-01 2.224 0.026141 \*   
## educationuniversity.degree 2.930e-01 9.348e-02 3.134 0.001725 \*\*   
## defaultunknown -3.212e-01 7.760e-02 -4.139 3.49e-05 \*\*\*  
## defaultyes -7.067e+00 1.114e+02 -0.063 0.949397   
## contacttelephone -5.259e-01 8.898e-02 -5.910 3.41e-09 \*\*\*  
## monthaug 9.770e-01 1.316e-01 7.426 1.12e-13 \*\*\*  
## monthdec -6.727e-02 2.539e-01 -0.265 0.791041   
## monthjul 2.773e-01 1.140e-01 2.432 0.014999 \*   
## monthjun -3.655e-01 1.256e-01 -2.911 0.003600 \*\*   
## monthmar 2.318e+00 1.454e-01 15.942 < 2e-16 \*\*\*  
## monthmay -4.148e-01 9.062e-02 -4.578 4.70e-06 \*\*\*  
## monthnov -4.888e-01 1.270e-01 -3.849 0.000118 \*\*\*  
## monthoct 7.863e-02 1.504e-01 0.523 0.601161   
## monthsep 2.254e-01 1.625e-01 1.387 0.165352   
## duration 1.964e-03 1.514e-04 12.973 < 2e-16 \*\*\*  
## campaign -2.395e-02 1.334e-02 -1.795 0.072605 .   
## pdays -8.536e-04 2.495e-04 -3.421 0.000624 \*\*\*  
## poutcomenonexistent 4.811e-01 7.720e-02 6.233 4.58e-10 \*\*\*  
## poutcomesuccess 1.011e+00 2.502e-01 4.040 5.36e-05 \*\*\*  
## emp\_var\_rate -1.736e+00 1.380e-01 -12.575 < 2e-16 \*\*\*  
## cons\_price\_idx 1.909e+00 1.290e-01 14.799 < 2e-16 \*\*\*  
## cons\_conf\_idx 1.920e-02 6.709e-03 2.862 0.004215 \*\*   
## euribor3m 5.626e-01 1.046e-01 5.378 7.55e-08 \*\*\*  
## Age\_Grp32-37 -2.179e-01 6.639e-02 -3.282 0.001030 \*\*   
## Age\_Grp38-47 -3.370e-01 7.015e-02 -4.804 1.56e-06 \*\*\*  
## Age\_Grp47-55 -1.744e-01 7.830e-02 -2.227 0.025915 \*   
## Age\_Grp>55 -4.849e-03 8.850e-02 -0.055 0.956306   
## duration\_group30-60s 3.347e+00 1.010e+00 3.312 0.000925 \*\*\*  
## duration\_group1-5 min 4.682e+00 1.002e+00 4.671 3.00e-06 \*\*\*  
## duration\_group5-10min 5.695e+00 1.004e+00 5.671 1.42e-08 \*\*\*  
## duration\_group10+ min 6.857e+00 1.011e+00 6.781 1.20e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12101 on 30559 degrees of freedom  
## AIC: 12173  
##   
## Number of Fisher Scoring iterations: 10

#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
vif(step.log)

## GVIF Df GVIF^(1/(2\*Df))  
## education 1.257653 6 1.019288  
## default 1.130537 2 1.031149  
## contact 2.381103 1 1.543082  
## month 15.526052 9 1.164582  
## duration 4.479531 1 2.116490  
## campaign 1.056804 1 1.028010  
## pdays 9.750051 1 3.122507  
## poutcome 11.097757 2 1.825193  
## emp\_var\_rate 101.317447 1 10.065657  
## cons\_price\_idx 12.669027 1 3.559358  
## cons\_conf\_idx 2.738421 1 1.654818  
## euribor3m 66.475840 1 8.153272  
## Age\_Grp 1.271286 4 1.030458  
## duration\_group 4.932460 4 1.220767

#Remove variables with high vifs and run the model again  
train\_step <- train %>% dplyr::select(-emp\_var\_rate )  
#Check vifs again  
full.log<-glm(y~.,family="binomial",data=train\_step)  
step.log<-full.log %>% stepAIC(trace=FALSE)  
summary(step.log)

##   
## Call:  
## glm(formula = y ~ education + default + contact + month + duration +   
## campaign + pdays + poutcome + cons\_price\_idx + euribor3m +   
## nr\_employed + Age\_Grp + duration\_group, family = "binomial",   
## data = train\_step)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1547 -0.3048 -0.1731 -0.0774 4.2696   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.661e+02 1.818e+01 9.136 < 2e-16 \*\*\*  
## educationbasic.6y 2.065e-01 1.341e-01 1.540 0.123546   
## educationbasic.9y 2.196e-02 1.062e-01 0.207 0.836236   
## educationhigh.school 1.282e-01 9.663e-02 1.326 0.184693   
## educationilliterate 1.637e+00 8.190e-01 1.998 0.045682 \*   
## educationprofessional.course 2.043e-01 1.050e-01 1.946 0.051645 .   
## educationuniversity.degree 2.870e-01 9.323e-02 3.078 0.002083 \*\*   
## defaultunknown -3.344e-01 7.755e-02 -4.312 1.62e-05 \*\*\*  
## defaultyes -7.209e+00 1.114e+02 -0.065 0.948403   
## contacttelephone -1.806e-01 7.846e-02 -2.302 0.021330 \*   
## monthaug 2.213e-01 1.212e-01 1.826 0.067840 .   
## monthdec -4.329e-01 2.631e-01 -1.645 0.099960 .   
## monthjul 4.087e-01 1.110e-01 3.683 0.000231 \*\*\*  
## monthjun 7.039e-01 1.047e-01 6.726 1.75e-11 \*\*\*  
## monthmar 1.455e+00 1.513e-01 9.619 < 2e-16 \*\*\*  
## monthmay -7.464e-01 9.124e-02 -8.180 2.83e-16 \*\*\*  
## monthnov -4.435e-01 1.368e-01 -3.243 0.001184 \*\*   
## monthoct -2.827e-01 1.782e-01 -1.586 0.112699   
## monthsep -6.881e-01 1.920e-01 -3.583 0.000339 \*\*\*  
## duration 1.978e-03 1.513e-04 13.074 < 2e-16 \*\*\*  
## campaign -3.011e-02 1.346e-02 -2.237 0.025276 \*   
## pdays -9.761e-04 2.454e-04 -3.978 6.94e-05 \*\*\*  
## poutcomenonexistent 4.893e-01 7.666e-02 6.383 1.74e-10 \*\*\*  
## poutcomesuccess 9.081e-01 2.460e-01 3.691 0.000223 \*\*\*  
## cons\_price\_idx -6.586e-01 9.767e-02 -6.743 1.55e-11 \*\*\*  
## euribor3m 4.094e-01 1.018e-01 4.020 5.81e-05 \*\*\*  
## nr\_employed -2.203e-02 1.948e-03 -11.308 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.303e-01 6.615e-02 -3.481 0.000499 \*\*\*  
## Age\_Grp38-47 -3.205e-01 6.976e-02 -4.594 4.35e-06 \*\*\*  
## Age\_Grp47-55 -1.747e-01 7.791e-02 -2.243 0.024898 \*   
## Age\_Grp>55 2.340e-02 8.785e-02 0.266 0.789924   
## duration\_group30-60s 3.492e+00 1.010e+00 3.457 0.000547 \*\*\*  
## duration\_group1-5 min 4.787e+00 1.002e+00 4.776 1.79e-06 \*\*\*  
## duration\_group5-10min 5.783e+00 1.004e+00 5.760 8.42e-09 \*\*\*  
## duration\_group10+ min 6.921e+00 1.011e+00 6.844 7.69e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12214 on 30560 degrees of freedom  
## AIC: 12284  
##   
## Number of Fisher Scoring iterations: 10

#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
vif(step.log)

## GVIF Df GVIF^(1/(2\*Df))  
## education 1.261288 6 1.019533  
## default 1.127903 2 1.030547  
## contact 1.826971 1 1.351655  
## month 8.571127 9 1.126770  
## duration 4.470715 1 2.114407  
## campaign 1.055349 1 1.027302  
## pdays 9.495974 1 3.081554  
## poutcome 10.742043 2 1.810388  
## cons\_price\_idx 7.171800 1 2.678022  
## euribor3m 63.342540 1 7.958803  
## nr\_employed 49.809135 1 7.057559  
## Age\_Grp 1.262000 4 1.029514  
## duration\_group 4.881183 4 1.219174

#euribor and nr\_employed are both statistically significant in the model but have high VIFs.Removing nr\_employed  
train\_step\_2 <- train\_step %>% dplyr::select(-nr\_employed )  
  
full.log<-glm(y~.,family="binomial",data=train\_step\_2)  
step.log<-full.log %>% stepAIC(trace=FALSE)  
summary(step.log)

##   
## Call:  
## glm(formula = y ~ education + default + contact + month + duration +   
## campaign + pdays + poutcome + cons\_price\_idx + cons\_conf\_idx +   
## euribor3m + Age\_Grp + prevly\_Cntctd + duration\_group, family = "binomial",   
## data = train\_step\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1223 -0.3099 -0.1796 -0.0791 4.2809   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.682e+01 2.066e+01 -0.814 0.415574   
## educationbasic.6y 1.959e-01 1.340e-01 1.462 0.143785   
## educationbasic.9y 1.307e-02 1.062e-01 0.123 0.902015   
## educationhigh.school 1.218e-01 9.648e-02 1.262 0.206872   
## educationilliterate 1.587e+00 8.241e-01 1.926 0.054104 .   
## educationprofessional.course 2.134e-01 1.048e-01 2.036 0.041781 \*   
## educationuniversity.degree 3.001e-01 9.308e-02 3.224 0.001263 \*\*   
## defaultunknown -3.482e-01 7.749e-02 -4.494 7.00e-06 \*\*\*  
## defaultyes -7.182e+00 1.113e+02 -0.065 0.948566   
## contacttelephone -2.498e-01 8.318e-02 -3.004 0.002667 \*\*   
## monthaug 3.693e-01 1.214e-01 3.041 0.002358 \*\*   
## monthdec 1.032e-01 2.523e-01 0.409 0.682612   
## monthjul 4.683e-01 1.121e-01 4.177 2.95e-05 \*\*\*  
## monthjun 4.636e-01 1.077e-01 4.303 1.68e-05 \*\*\*  
## monthmar 1.907e+00 1.487e-01 12.824 < 2e-16 \*\*\*  
## monthmay -5.888e-01 8.847e-02 -6.656 2.82e-11 \*\*\*  
## monthnov 1.512e-01 1.160e-01 1.304 0.192382   
## monthoct 4.658e-01 1.477e-01 3.154 0.001610 \*\*   
## monthsep 7.788e-02 1.626e-01 0.479 0.631925   
## duration 1.990e-03 1.517e-04 13.122 < 2e-16 \*\*\*  
## campaign -3.278e-02 1.347e-02 -2.434 0.014913 \*   
## pdays -3.130e-02 2.049e-02 -1.528 0.126623   
## poutcomenonexistent 5.228e-01 7.672e-02 6.815 9.46e-12 \*\*\*  
## poutcomesuccess 7.528e-01 2.609e-01 2.885 0.003913 \*\*   
## cons\_price\_idx 4.706e-01 5.596e-02 8.411 < 2e-16 \*\*\*  
## cons\_conf\_idx 5.587e-02 6.157e-03 9.075 < 2e-16 \*\*\*  
## euribor3m -7.304e-01 2.268e-02 -32.207 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.394e-01 6.598e-02 -3.628 0.000286 \*\*\*  
## Age\_Grp38-47 -3.266e-01 6.962e-02 -4.691 2.72e-06 \*\*\*  
## Age\_Grp47-55 -1.764e-01 7.777e-02 -2.268 0.023319 \*   
## Age\_Grp>55 6.815e-03 8.809e-02 0.077 0.938332   
## prevly\_CntctdYes -2.996e+01 2.027e+01 -1.478 0.139414   
## duration\_group30-60s 3.467e+00 1.010e+00 3.432 0.000600 \*\*\*  
## duration\_group1-5 min 4.757e+00 1.002e+00 4.746 2.08e-06 \*\*\*  
## duration\_group5-10min 5.736e+00 1.004e+00 5.714 1.11e-08 \*\*\*  
## duration\_group10+ min 6.852e+00 1.011e+00 6.777 1.23e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12256 on 30559 degrees of freedom  
## AIC: 12328  
##   
## Number of Fisher Scoring iterations: 10

#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
vif(step.log)

## GVIF Df GVIF^(1/(2\*Df))  
## education 1.255921 6 1.019171  
## default 1.125557 2 1.030011  
## contact 2.051259 1 1.432222  
## month 5.332276 9 1.097448  
## duration 4.461173 1 2.112149  
## campaign 1.053991 1 1.026641  
## pdays 67048.651107 1 258.937543  
## poutcome 12.271026 2 1.871631  
## cons\_price\_idx 2.435220 1 1.560519  
## cons\_conf\_idx 2.322638 1 1.524020  
## euribor3m 3.150444 1 1.774949  
## Age\_Grp 1.267147 4 1.030038  
## prevly\_Cntctd 66533.221257 1 257.940344  
## duration\_group 4.837996 4 1.217820

#pdays and prevly\_Cntctd are have high VIFs.Removing pdays  
train\_step\_3 <- train\_step\_2 %>% dplyr::select(-pdays )  
#Check vifs again  
full.log<-glm(y~.,family="binomial",data=train\_step\_3)  
step.log<-full.log %>% stepAIC(trace=FALSE)  
summary(step.log)

##   
## Call:  
## glm(formula = y ~ education + default + contact + month + duration +   
## campaign + poutcome + cons\_price\_idx + cons\_conf\_idx + euribor3m +   
## Age\_Grp + prevly\_Cntctd + duration\_group, family = "binomial",   
## data = train\_step\_3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1202 -0.3098 -0.1797 -0.0794 4.2796   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.740e+01 5.188e+00 -9.136 < 2e-16 \*\*\*  
## educationbasic.6y 1.963e-01 1.340e-01 1.465 0.142812   
## educationbasic.9y 1.156e-02 1.062e-01 0.109 0.913274   
## educationhigh.school 1.225e-01 9.646e-02 1.270 0.204228   
## educationilliterate 1.585e+00 8.241e-01 1.924 0.054405 .   
## educationprofessional.course 2.112e-01 1.049e-01 2.015 0.043943 \*   
## educationuniversity.degree 3.019e-01 9.305e-02 3.245 0.001175 \*\*   
## defaultunknown -3.474e-01 7.747e-02 -4.484 7.33e-06 \*\*\*  
## defaultyes -7.185e+00 1.113e+02 -0.065 0.948549   
## contacttelephone -2.483e-01 8.316e-02 -2.986 0.002824 \*\*   
## monthaug 3.659e-01 1.215e-01 3.012 0.002593 \*\*   
## monthdec 9.410e-02 2.524e-01 0.373 0.709268   
## monthjul 4.643e-01 1.121e-01 4.140 3.47e-05 \*\*\*  
## monthjun 4.658e-01 1.077e-01 4.324 1.53e-05 \*\*\*  
## monthmar 1.902e+00 1.487e-01 12.794 < 2e-16 \*\*\*  
## monthmay -5.912e-01 8.845e-02 -6.684 2.32e-11 \*\*\*  
## monthnov 1.514e-01 1.160e-01 1.306 0.191621   
## monthoct 4.642e-01 1.477e-01 3.142 0.001678 \*\*   
## monthsep 7.487e-02 1.626e-01 0.460 0.645286   
## duration 1.990e-03 1.517e-04 13.121 < 2e-16 \*\*\*  
## campaign -3.293e-02 1.347e-02 -2.445 0.014501 \*   
## poutcomenonexistent 5.216e-01 7.668e-02 6.802 1.03e-11 \*\*\*  
## poutcomesuccess 8.921e-01 2.442e-01 3.654 0.000258 \*\*\*  
## cons\_price\_idx 4.633e-01 5.574e-02 8.313 < 2e-16 \*\*\*  
## cons\_conf\_idx 5.590e-02 6.158e-03 9.078 < 2e-16 \*\*\*  
## euribor3m -7.286e-01 2.264e-02 -32.177 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.398e-01 6.598e-02 -3.635 0.000278 \*\*\*  
## Age\_Grp38-47 -3.273e-01 6.963e-02 -4.701 2.59e-06 \*\*\*  
## Age\_Grp47-55 -1.754e-01 7.776e-02 -2.256 0.024049 \*   
## Age\_Grp>55 8.717e-03 8.803e-02 0.099 0.921121   
## prevly\_CntctdYes 1.001e+00 2.420e-01 4.137 3.51e-05 \*\*\*  
## duration\_group30-60s 3.471e+00 1.010e+00 3.436 0.000590 \*\*\*  
## duration\_group1-5 min 4.753e+00 1.002e+00 4.743 2.11e-06 \*\*\*  
## duration\_group5-10min 5.732e+00 1.004e+00 5.710 1.13e-08 \*\*\*  
## duration\_group10+ min 6.849e+00 1.011e+00 6.773 1.26e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12258 on 30560 degrees of freedom  
## AIC: 12328  
##   
## Number of Fisher Scoring iterations: 10

#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
vif(step.log)

## GVIF Df GVIF^(1/(2\*Df))  
## education 1.253772 6 1.019025  
## default 1.125504 2 1.029999  
## contact 2.050805 1 1.432063  
## month 5.318947 9 1.097296  
## duration 4.462848 1 2.112545  
## campaign 1.053971 1 1.026631  
## poutcome 10.718947 2 1.809414  
## cons\_price\_idx 2.416085 1 1.554376  
## cons\_conf\_idx 2.325063 1 1.524816  
## euribor3m 3.140346 1 1.772102  
## Age\_Grp 1.266583 4 1.029981  
## prevly\_Cntctd 9.474336 1 3.078041  
## duration\_group 4.831243 4 1.217608

full.log<-glm(y~education+default+contact+month+duration+campaign+poutcome+cons\_price\_idx+euribor3m+Age\_Grp,family="binomial",data=train)  
#full.log<-glm(y~.,family="binomial",data=train\_step\_3)  
step.log<-full.log %>% stepAIC(trace=FALSE)  
summary(step.log)

##   
## Call:  
## glm(formula = y ~ education + default + contact + month + duration +   
## campaign + poutcome + cons\_price\_idx + euribor3m + Age\_Grp,   
## family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.0685 -0.3203 -0.1932 -0.1258 3.3168   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.204e+01 4.578e+00 -6.999 2.57e-12 \*\*\*  
## educationbasic.6y 1.575e-01 1.345e-01 1.171 0.241590   
## educationbasic.9y -2.286e-02 1.058e-01 -0.216 0.828866   
## educationhigh.school 1.186e-01 9.546e-02 1.242 0.214074   
## educationilliterate 1.518e+00 7.971e-01 1.905 0.056813 .   
## educationprofessional.course 2.020e-01 1.035e-01 1.952 0.050974 .   
## educationuniversity.degree 3.017e-01 9.194e-02 3.281 0.001034 \*\*   
## defaultunknown -3.705e-01 7.917e-02 -4.680 2.87e-06 \*\*\*  
## defaultyes -7.446e+00 1.134e+02 -0.066 0.947626   
## contacttelephone -1.170e-01 7.205e-02 -1.624 0.104435   
## monthaug 8.656e-01 9.836e-02 8.801 < 2e-16 \*\*\*  
## monthdec 7.077e-01 2.310e-01 3.064 0.002185 \*\*   
## monthjul 7.383e-01 1.028e-01 7.181 6.92e-13 \*\*\*  
## monthjun 6.555e-01 1.016e-01 6.451 1.11e-10 \*\*\*  
## monthmar 1.819e+00 1.328e-01 13.697 < 2e-16 \*\*\*  
## monthmay -5.020e-01 8.429e-02 -5.956 2.59e-09 \*\*\*  
## monthnov 4.117e-01 1.069e-01 3.853 0.000117 \*\*\*  
## monthoct 1.012e+00 1.262e-01 8.021 1.05e-15 \*\*\*  
## monthsep 6.077e-01 1.392e-01 4.366 1.26e-05 \*\*\*  
## duration 4.730e-03 8.650e-05 54.685 < 2e-16 \*\*\*  
## campaign -5.222e-02 1.344e-02 -3.885 0.000103 \*\*\*  
## poutcomenonexistent 4.152e-01 7.100e-02 5.849 4.95e-09 \*\*\*  
## poutcomesuccess 1.817e+00 9.783e-02 18.568 < 2e-16 \*\*\*  
## cons\_price\_idx 3.187e-01 4.920e-02 6.478 9.29e-11 \*\*\*  
## euribor3m -7.118e-01 2.123e-02 -33.520 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.370e-01 6.476e-02 -3.660 0.000252 \*\*\*  
## Age\_Grp38-47 -3.125e-01 6.894e-02 -4.533 5.81e-06 \*\*\*  
## Age\_Grp47-55 -1.389e-01 7.664e-02 -1.812 0.069918 .   
## Age\_Grp>55 9.290e-02 8.505e-02 1.092 0.274703   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12895 on 30566 degrees of freedom  
## AIC: 12953  
##   
## Number of Fisher Scoring iterations: 10

#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
vif(step.log)

## GVIF Df GVIF^(1/(2\*Df))  
## education 1.249616 6 1.018743  
## default 1.122014 2 1.029199  
## contact 1.488326 1 1.219970  
## month 2.605011 9 1.054631  
## duration 1.229352 1 1.108761  
## campaign 1.044743 1 1.022127  
## poutcome 1.292286 2 1.066202  
## cons\_price\_idx 1.917062 1 1.384580  
## euribor3m 2.681983 1 1.637676  
## Age\_Grp 1.254327 4 1.028730

#poutcome and prevly\_Cntctd have high VIFs but these are not interchangable.  
#Remove statistically insignificant variables and run the model again  
#str(train\_step\_3)  
#Run step model again  
#full.log<-glm(y~education+default+contact+month+duration+campaign+poutcome+cons\_price\_idx+euribor3m+Age\_Grp,family="binomial",data=train)  
#full.log<-glm(y~.,family="binomial",data=train\_step\_3)  
  
#step.log<-full.log %>% stepAIC(trace=FALSE)  
#summary(step.log)  
#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
#vif(step.log)   
  
#education is border line and contact became insignificant. Remove contact from the model  
  
#Run step model again  
full.log<-glm(y~education+default+month+duration+campaign+poutcome+cons\_price\_idx+euribor3m+Age\_Grp,family="binomial",data=train)  
#full.log<-glm(y~.,family="binomial",data=train)  
step.log<-full.log %>% stepAIC(trace=FALSE)  
summary(step.log)

##   
## Call:  
## glm(formula = y ~ education + default + month + duration + campaign +   
## poutcome + cons\_price\_idx + euribor3m + Age\_Grp, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.0873 -0.3199 -0.1930 -0.1273 3.3021   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.031e+01 4.443e+00 -6.821 9.01e-12 \*\*\*  
## educationbasic.6y 1.561e-01 1.345e-01 1.160 0.245892   
## educationbasic.9y -2.180e-02 1.058e-01 -0.206 0.836652   
## educationhigh.school 1.183e-01 9.548e-02 1.239 0.215409   
## educationilliterate 1.509e+00 8.000e-01 1.886 0.059328 .   
## educationprofessional.course 2.040e-01 1.035e-01 1.972 0.048661 \*   
## educationuniversity.degree 3.036e-01 9.195e-02 3.302 0.000959 \*\*\*  
## defaultunknown -3.697e-01 7.917e-02 -4.670 3.02e-06 \*\*\*  
## defaultyes -7.420e+00 1.133e+02 -0.065 0.947806   
## monthaug 8.755e-01 9.810e-02 8.925 < 2e-16 \*\*\*  
## monthdec 6.884e-01 2.305e-01 2.987 0.002819 \*\*   
## monthjul 7.646e-01 1.014e-01 7.540 4.70e-14 \*\*\*  
## monthjun 6.239e-01 9.967e-02 6.259 3.87e-10 \*\*\*  
## monthmar 1.811e+00 1.327e-01 13.647 < 2e-16 \*\*\*  
## monthmay -5.192e-01 8.364e-02 -6.208 5.37e-10 \*\*\*  
## monthnov 4.214e-01 1.067e-01 3.950 7.81e-05 \*\*\*  
## monthoct 9.994e-01 1.257e-01 7.950 1.87e-15 \*\*\*  
## monthsep 5.997e-01 1.389e-01 4.317 1.58e-05 \*\*\*  
## duration 4.734e-03 8.645e-05 54.760 < 2e-16 \*\*\*  
## campaign -5.337e-02 1.343e-02 -3.974 7.06e-05 \*\*\*  
## poutcomenonexistent 4.089e-01 7.090e-02 5.767 8.06e-09 \*\*\*  
## poutcomesuccess 1.815e+00 9.782e-02 18.558 < 2e-16 \*\*\*  
## cons\_price\_idx 3.002e-01 4.777e-02 6.284 3.29e-10 \*\*\*  
## euribor3m -7.190e-01 2.073e-02 -34.688 < 2e-16 \*\*\*  
## Age\_Grp32-37 -2.372e-01 6.475e-02 -3.663 0.000249 \*\*\*  
## Age\_Grp38-47 -3.138e-01 6.893e-02 -4.552 5.32e-06 \*\*\*  
## Age\_Grp47-55 -1.392e-01 7.665e-02 -1.816 0.069378 .   
## Age\_Grp>55 8.786e-02 8.499e-02 1.034 0.301213   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21372 on 30594 degrees of freedom  
## Residual deviance: 12898 on 30567 degrees of freedom  
## AIC: 12954  
##   
## Number of Fisher Scoring iterations: 10

#education becomes statistically significant after removing contact. VIFs look good.  
#exp(cbind("Odds ratio" = coef(step.log), confint.default(step.log, level = 0.95)))  
vif(step.log)

## GVIF Df GVIF^(1/(2\*Df))  
## education 1.248514 6 1.018668  
## default 1.122475 2 1.029305  
## month 2.155722 9 1.043597  
## duration 1.229923 1 1.109019  
## campaign 1.041924 1 1.020747  
## poutcome 1.287428 2 1.065199  
## cons\_price\_idx 1.818918 1 1.348673  
## euribor3m 2.558493 1 1.599529  
## Age\_Grp 1.252317 4 1.028524

#Predicting using step   
fit.pred.step<-predict(step.log,newdata=test,type="response")  
test$y[1:15]

## [1] yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes  
## Levels: no yes

fit.pred.step[1:15]

## 1 18 20 33 44 52   
## 0.73592551 0.08788080 0.90491831 0.16539058 0.01439464 0.09420273   
## 53 60 66 70 73 74   
## 0.02916903 0.01758288 0.05961790 0.04584461 0.06567052 0.38434974   
## 76 77 81   
## 0.05719251 0.12381212 0.36912697

# Feature selection using lasso  
library(glmnet)

## Warning: package 'glmnet' was built under R version 3.6.3

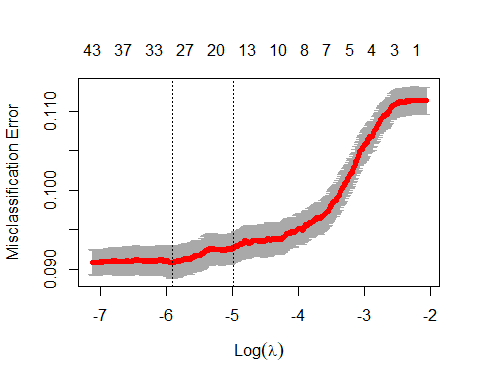
## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.0-2

dat.train.x <- model.matrix(y~.,train)  
dat.train.y<-train[,24]  
cvfit <- cv.glmnet(dat.train.x, dat.train.y, family = "binomial", type.measure = "class", nlambda = 1000)  
plot(cvfit)



coef(cvfit, s = "lambda.min")

## 59 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 49.5620211829  
## (Intercept) .   
## age .   
## jobblue-collar -0.1506260202  
## jobentrepreneur .   
## jobhousemaid .   
## jobmanagement .   
## jobretired 0.0414155763  
## jobself-employed .   
## jobservices -0.0023372335  
## jobstudent 0.2047583017  
## jobtechnician .   
## jobunemployed .   
## maritalmarried .   
## maritalsingle 0.0198707496  
## educationbasic.6y .   
## educationbasic.9y .   
## educationhigh.school .   
## educationilliterate .   
## educationprofessional.course .   
## educationuniversity.degree 0.0818136185  
## defaultunknown -0.2101440216  
## defaultyes .   
## housingyes .   
## loanyes .   
## contacttelephone -0.1880337178  
## monthaug 0.0687506262  
## monthdec .   
## monthjul 0.1077500896  
## monthjun 0.1961392520  
## monthmar 1.4307978934  
## monthmay -0.7271024791  
## monthnov -0.2214063196  
## monthoct .   
## monthsep -0.1510007365  
## day\_of\_weekmon -0.0150505562  
## day\_of\_weekthu .   
## day\_of\_weektue .   
## day\_of\_weekwed .   
## duration 0.0023179076  
## campaign -0.0012354086  
## pdays -0.0008806163  
## previous .   
## poutcomenonexistent 0.2370231029  
## poutcomesuccess 0.7879870635  
## emp\_var\_rate -0.1748221885  
## cons\_price\_idx .   
## cons\_conf\_idx 0.0202226454  
## euribor3m .   
## nr\_employed -0.0100509553  
## Age\_Grp32-37 -0.0171681388  
## Age\_Grp38-47 -0.1081115791  
## Age\_Grp47-55 .   
## Age\_Grp>55 0.0178967651  
## prevly\_CntctdYes .   
## duration\_group30-60s -0.4908353460  
## duration\_group1-5 min 0.2851034146  
## duration\_group5-10min 1.1294970376  
## duration\_group10+ min 2.0409912469

#CV misclassification error rate is little below .1  
print("CV Error Rate:")

## [1] "CV Error Rate:"

cvfit$cvm[which(cvfit$lambda==cvfit$lambda.min)]

## [1] 0.09083184

#Optimal penalty  
print("Penalty Value:")

## [1] "Penalty Value:"

cvfit$lambda.min

## [1] 0.00272658

finalmodel<-glmnet(dat.train.x, dat.train.y, family = "binomial",lambda=cvfit$lambda.min)  
finalmodel$call

## glmnet(x = dat.train.x, y = dat.train.y, family = "binomial",   
## lambda = cvfit$lambda.min)

dat.test.x<-model.matrix(y~.,test)  
fit.pred.lasso <- predict(finalmodel, newx = dat.test.x, type = "response")  
  
test$y[1:15]

## [1] yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes  
## Levels: no yes

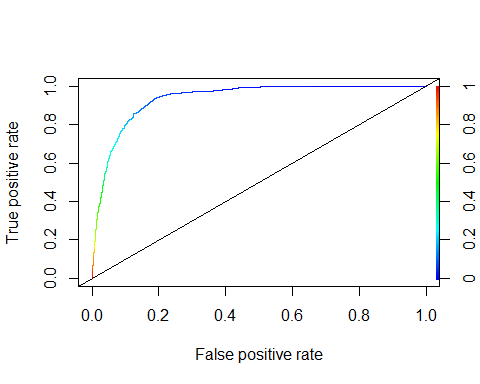
fit.pred.lasso[1:15]

## [1] 0.52450430 0.18328097 0.71559202 0.24956342 0.03644571 0.19901951  
## [7] 0.05408934 0.03435125 0.20931555 0.17117718 0.17459407 0.38854155  
## [13] 0.15794007 0.21608078 0.37449534

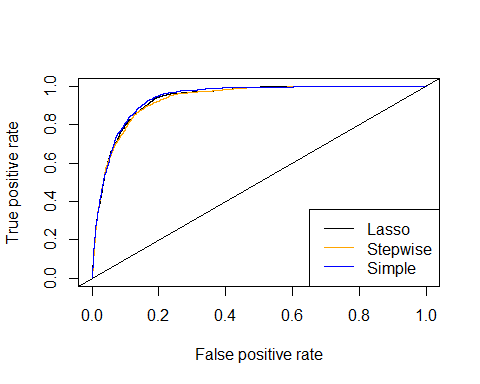
#ROCR  
library(ROCR)

## Warning: package 'ROCR' was built under R version 3.6.3

results.lasso<-prediction(fit.pred.lasso, test$y,label.ordering=c("no","yes"))  
roc.lasso = performance(results.lasso, measure = "tpr", x.measure = "fpr")  
plot(roc.lasso,colorize = TRUE)  
abline(a=0, b= 1)



results.step<-prediction(fit.pred.step, test$y,label.ordering=c("no","yes"))  
roc.step = performance(results.step, measure = "tpr", x.measure = "fpr")  
  
  
simple.log<-glm(y~.,family="binomial",data=train)  
fit.pred.origin<-predict(simple.log,newdata=test,type="response")  
results.origin<-prediction(fit.pred.origin,test$y,label.ordering=c("no","yes"))  
roc.origin=performance(results.origin,measure = "tpr", x.measure = "fpr")  
  
plot( roc.lasso)  
plot(roc.step,col="orange", add = TRUE)  
plot(roc.origin,col="blue",add=TRUE)  
legend("bottomright",legend=c("Lasso","Stepwise","Simple"),col=c("black","orange","blue"),lty=1,lwd=1)  
abline(a=0, b= 1)



#Playing with different cut offs  
cutoff<-0.5  
class.lasso<-factor(ifelse(fit.pred.lasso>cutoff,"yes","no"),levels=c("no","yes"))  
class.step<-factor(ifelse(fit.pred.step>cutoff,"yes","no"),levels=c("no","yes"))  
class.simple<-factor(ifelse(fit.pred.simple>cutoff,"yes","no"),levels=c("no","yes"))  
  
#Confusion Matrix for Lasso  
conf.lasso<-table(class.lasso,test$y)  
print("Confusion matrix for LASSO")

## [1] "Confusion matrix for LASSO"

conf.lasso

##   
## class.lasso no yes  
## no 6612 485  
## yes 186 367

#Confusion Matrix for step  
conf.step<-table(class.step,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.step

##   
## class.step no yes  
## no 6621 488  
## yes 177 364

#Confusion Matrix for simple  
conf.simple<-table(class.simple,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.simple

##   
## class.simple no yes  
## no 6622 487  
## yes 176 365

#Accuracy of LASSO and Stepwise  
print("Overall accuracy for LASSO and Stepwise respectively")

## [1] "Overall accuracy for LASSO and Stepwise respectively"

sum(diag(conf.lasso))/sum(conf.lasso)

## [1] 0.9122876

sum(diag(conf.step))/sum(conf.step)

## [1] 0.9130719

print("Alternative calculations of accuracy")

## [1] "Alternative calculations of accuracy"

Acc\_LASSO\_0.5 <- mean(class.lasso==test$y)  
Acc\_STEP\_0.5 <-mean(class.step==test$y)  
Acc\_SIMPLE\_0.5<-mean(class.simple==test$y)  
  
  
cutoff<-0.1  
class.lasso<-factor(ifelse(fit.pred.lasso>cutoff,"yes","no"),levels=c("no","yes"))  
class.step<-factor(ifelse(fit.pred.step>cutoff,"yes","no"),levels=c("no","yes"))  
class.simple<-factor(ifelse(fit.pred.simple>cutoff,"yes","no"),levels=c("no","yes"))  
  
#Confusion Matrix for Lasso  
conf.lasso<-table(class.lasso,test$y)  
print("Confusion matrix for LASSO")

## [1] "Confusion matrix for LASSO"

conf.lasso

##   
## class.lasso no yes  
## no 5674 80  
## yes 1124 772

#Confusion Matrix for step  
conf.step<-table(class.step,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.step

##   
## class.step no yes  
## no 5666 87  
## yes 1132 765

#Confusion Matrix for simple  
conf.simple<-table(class.simple,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.simple

##   
## class.simple no yes  
## no 5695 90  
## yes 1103 762

#Accuracy of LASSO and Stepwise  
print("Overall accuracy for LASSO and Stepwise respectively")

## [1] "Overall accuracy for LASSO and Stepwise respectively"

sum(diag(conf.lasso))/sum(conf.lasso)

## [1] 0.8426144

sum(diag(conf.step))/sum(conf.step)

## [1] 0.8406536

print("Alternative calculations of accuracy")

## [1] "Alternative calculations of accuracy"

Acc\_LASSO\_0.1 <- mean(class.lasso==test$y)  
Acc\_STEP\_0.1 <-mean(class.step==test$y)  
Acc\_SIMPLE\_0.1<-mean(class.simple==test$y)  
  
cutoff<-0.15  
class.lasso<-factor(ifelse(fit.pred.lasso>cutoff,"yes","no"),levels=c("no","yes"))  
class.step<-factor(ifelse(fit.pred.step>cutoff,"yes","no"),levels=c("no","yes"))  
class.simple<-factor(ifelse(fit.pred.simple>cutoff,"yes","no"),levels=c("no","yes"))  
  
#Confusion Matrix for Lasso  
conf.lasso<-table(class.lasso,test$y)  
print("Confusion matrix for LASSO")

## [1] "Confusion matrix for LASSO"

conf.lasso

##   
## class.lasso no yes  
## no 5938 122  
## yes 860 730

#Confusion Matrix for step  
conf.step<-table(class.step,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.step

##   
## class.step no yes  
## no 5976 140  
## yes 822 712

#Confusion Matrix for simple  
conf.simple<-table(class.simple,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.simple

##   
## class.simple no yes  
## no 5990 143  
## yes 808 709

#Accuracy of LASSO and Stepwise  
print("Overall accuracy for LASSO and Stepwise respectively")

## [1] "Overall accuracy for LASSO and Stepwise respectively"

sum(diag(conf.lasso))/sum(conf.lasso)

## [1] 0.871634

sum(diag(conf.step))/sum(conf.step)

## [1] 0.8742484

print("Alternative calculations of accuracy")

## [1] "Alternative calculations of accuracy"

Acc\_LASSO\_0.15 <- mean(class.lasso==test$y)  
Acc\_STEP\_0.15 <-mean(class.step==test$y)  
Acc\_SIMPLE\_0.15<-mean(class.simple==test$y)  
  
cutoff<-0.2  
class.lasso<-factor(ifelse(fit.pred.lasso>cutoff,"yes","no"),levels=c("no","yes"))  
class.step<-factor(ifelse(fit.pred.step>cutoff,"yes","no"),levels=c("no","yes"))  
class.simple<-factor(ifelse(fit.pred.simple>cutoff,"yes","no"),levels=c("no","yes"))  
  
#Confusion Matrix for Lasso  
conf.lasso<-table(class.lasso,test$y)  
print("Confusion matrix for LASSO")

## [1] "Confusion matrix for LASSO"

conf.lasso

##   
## class.lasso no yes  
## no 6129 176  
## yes 669 676

#Confusion Matrix for step  
conf.step<-table(class.step,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.step

##   
## class.step no yes  
## no 6176 212  
## yes 622 640

#Confusion Matrix for simple  
conf.simple<-table(class.simple,test$y)  
print("Confusion matrix for Stepwise")

## [1] "Confusion matrix for Stepwise"

conf.simple

##   
## class.simple no yes  
## no 6210 210  
## yes 588 642

#Accuracy of LASSO and Stepwise  
print("Overall accuracy for LASSO and Stepwise respectively")

## [1] "Overall accuracy for LASSO and Stepwise respectively"

sum(diag(conf.lasso))/sum(conf.lasso)

## [1] 0.8895425

sum(diag(conf.step))/sum(conf.step)

## [1] 0.8909804

print("Alternative calculations of accuracy")

## [1] "Alternative calculations of accuracy"

Acc\_LASSO\_0.2 <- mean(class.lasso==test$y)  
Acc\_STEP\_0.2 <-mean(class.step==test$y)  
Acc\_SIMPLE\_0.2<-mean(class.simple==test$y)  
  
  
Accuracy\_simple<- data.frame("Cut off"= c("0.1", "0.15","0.2","0.5"),"Simple\_Accuracy"=c(Acc\_SIMPLE\_0.1,Acc\_SIMPLE\_0.15,Acc\_SIMPLE\_0.2,Acc\_SIMPLE\_0.5 ) )  
Accuracy\_step<- data.frame("Step Cut off"= c("0.1", "0.15","0.2","0.5"),"Step\_Accuracy"=c(Acc\_STEP\_0.1,Acc\_STEP\_0.15,Acc\_STEP\_0.2,Acc\_STEP\_0.5 ) )  
Accuracy\_lasso<- data.frame("LASSO Cut off"= c("0.1", "0.15","0.2","0.5"),"LASSO\_Accuracy"=c(Acc\_LASSO\_0.1,Acc\_LASSO\_0.15,Acc\_LASSO\_0.2,Acc\_LASSO\_0.5 ) )  
 Accuracy <- cbind(Accuracy\_simple,Accuracy\_step$Step\_Accuracy,Accuracy\_lasso$LASSO\_Accuracy)   
 Accuracy

## Cut.off Simple\_Accuracy Accuracy\_step$Step\_Accuracy  
## 1 0.1 0.8440523 0.8406536  
## 2 0.15 0.8756863 0.8742484  
## 3 0.2 0.8956863 0.8909804  
## 4 0.5 0.9133333 0.9130719  
## Accuracy\_lasso$LASSO\_Accuracy  
## 1 0.8426144  
## 2 0.8716340  
## 3 0.8895425  
## 4 0.9122876