Project\_2\_Simer\_v1

Simerpreet

November 8, 2020

I did some EDA, created variables “age\_group” and “previously contacted”, ran logistic regression on the full data set. I tried balancing the data sets but there are so many variables to be considered. let’s discuss more on that.

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- 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(tidyr)  
library(dplyr)  
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

bank\_addl\_full <- read.csv(file.choose(), sep=';')  
nrow(bank\_addl\_full) #41,188

## [1] 41188

ncol(bank\_addl\_full)

## [1] 21

#head(bank\_addl\_full)  
summary(bank\_addl\_full)

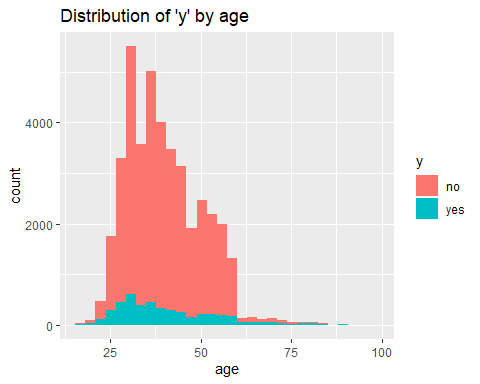
## age job marital   
## Min. :17.00 admin. :10422 divorced: 4612   
## 1st Qu.:32.00 blue-collar: 9254 married :24928   
## Median :38.00 technician : 6743 single :11568   
## Mean :40.02 services : 3969 unknown : 80   
## 3rd Qu.:47.00 management : 2924   
## Max. :98.00 retired : 1720   
## (Other) : 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   
##

str(bank\_addl\_full)

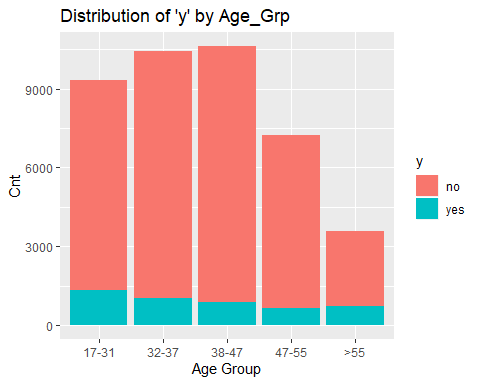
## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...  
## $ marital : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...  
## $ education : Factor w/ 8 levels "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 ...

#Analysing the entire data set  
#Analysing Age  
ggplot(bank\_addl\_full) + geom\_histogram(mapping = aes(x=age, fill=y)) +ggtitle("Distribution of 'y' by age")

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

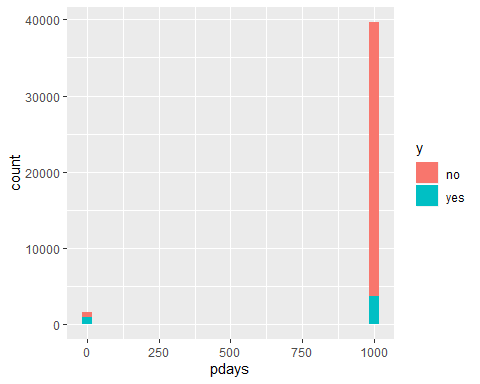


#Creating new variables  
#Age\_Grp - split the data into age groups "17-31","32-37" ,"38-47", "47-55", ">55" (based in IQR)  
bank\_addl\_full$Age\_Grp <- cut(bank\_addl\_full$age, breaks = c(16,31,37,46,55,98), labels = c("17-31","32-37" ,"38-47", "47-55", ">55"))  
  
#validate the cut command  
#bank\_addl\_full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

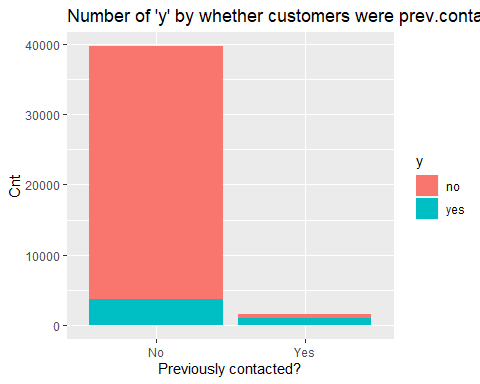


#Analysing pdays  
ggplot(bank\_addl\_full) + geom\_histogram(mapping = aes(x=pdays, fill=y))

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

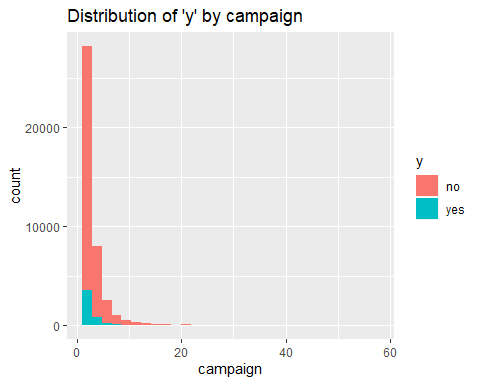


#prevly\_Cntctd Yes/No. TO see the distribution or 'Y' on first time contact vs. a follow up  
bank\_addl\_full$prevly\_Cntctd <- case\_when(bank\_addl\_full$pdays==999 ~ "No", !bank\_addl\_full$pdays==999 ~ "Yes")  
  
#Validate previously contacted variable  
#bank\_addl\_full %>% filter(!bank\_addl\_full$pdays==999)  
  
ggplot(bank\_addl\_full) + 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?")

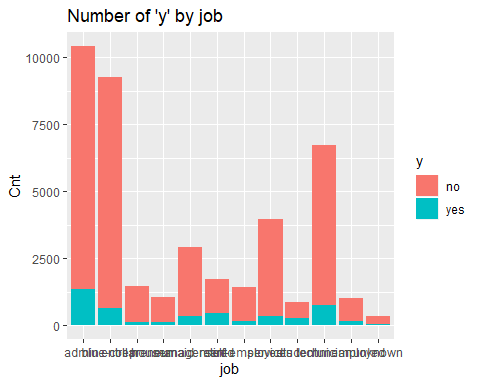


#Analysing campaign  
ggplot(bank\_addl\_full) + geom\_histogram(mapping = aes(x=campaign, fill=y)) +ggtitle("Distribution of 'y' by campaign")

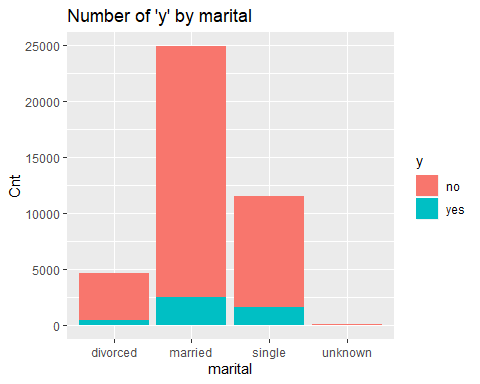
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Analysing job  
ggplot(bank\_addl\_full) + geom\_bar(mapping = aes(x=job, fill = y)) + ggtitle("Number of 'y' by job") +  
 ylab("Cnt") + xlab("job")

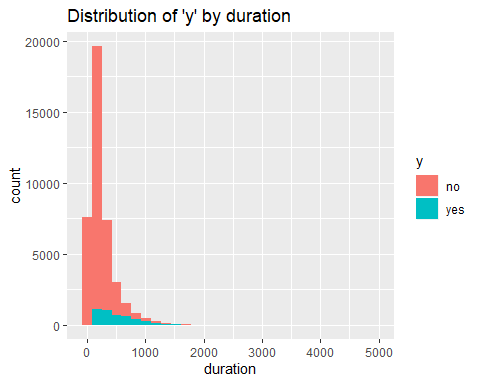


#Analysing marital  
ggplot(bank\_addl\_full) + geom\_bar(mapping = aes(x=marital, fill = y)) + ggtitle("Number of 'y' by marital") +  
 ylab("Cnt") + xlab("marital")



#Analysing duration  
ggplot(bank\_addl\_full) + geom\_histogram(mapping = aes(x=duration, fill=y)) +ggtitle("Distribution of 'y' by duration")

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



#bank\_addl\_full %>% filter(bank\_addl\_full$duration>4000)  
  
#Running logistic regression on full data set  
simple.log<-glm(y~.,family="binomial",data=bank\_addl\_full)  
summary(simple.log)

##   
## Call:  
## glm(formula = y ~ ., family = "binomial", data = bank\_addl\_full)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.9972 -0.2973 -0.1854 -0.1328 3.4077   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.035e+02 4.226e+01 -4.814 1.48e-06 \*\*\*  
## age 6.003e-03 5.828e-03 1.030 0.302958   
## jobblue-collar -2.296e-01 7.998e-02 -2.871 0.004096 \*\*   
## jobentrepreneur -1.769e-01 1.260e-01 -1.404 0.160261   
## jobhousemaid -7.525e-02 1.492e-01 -0.504 0.614127   
## jobmanagement -5.309e-02 8.555e-02 -0.621 0.534888   
## jobretired 6.175e-02 1.184e-01 0.522 0.601903   
## jobself-employed -1.713e-01 1.178e-01 -1.453 0.146103   
## jobservices -1.493e-01 8.618e-02 -1.733 0.083183 .   
## jobstudent 1.404e-01 1.136e-01 1.235 0.216721   
## jobtechnician -8.639e-03 7.115e-02 -0.121 0.903362   
## jobunemployed 2.195e-02 1.280e-01 0.171 0.863871   
## jobunknown -1.327e-01 2.407e-01 -0.552 0.581227   
## maritalmarried -1.143e-02 6.876e-02 -0.166 0.867975   
## maritalsingle -5.261e-03 7.925e-02 -0.066 0.947070   
## maritalunknown 1.965e-02 4.169e-01 0.047 0.962414   
## educationbasic.6y 1.719e-01 1.213e-01 1.417 0.156425   
## educationbasic.9y 2.307e-02 9.593e-02 0.240 0.809943   
## educationhigh.school 7.426e-02 9.273e-02 0.801 0.423239   
## educationilliterate 1.193e+00 7.576e-01 1.575 0.115362   
## educationprofessional.course 1.453e-01 1.021e-01 1.423 0.154646   
## educationuniversity.degree 2.215e-01 9.284e-02 2.386 0.017035 \*   
## educationunknown 1.661e-01 1.200e-01 1.384 0.166403   
## defaultunknown -2.851e-01 6.765e-02 -4.214 2.51e-05 \*\*\*  
## defaultyes -7.343e+00 1.136e+02 -0.065 0.948449   
## housingunknown -9.111e-02 1.400e-01 -0.651 0.515142   
## housingyes -4.297e-03 4.140e-02 -0.104 0.917317   
## loanunknown NA NA NA NA   
## loanyes -4.978e-02 5.759e-02 -0.864 0.387387   
## contacttelephone -6.386e-01 7.686e-02 -8.309 < 2e-16 \*\*\*  
## monthaug 8.474e-01 1.207e-01 7.020 2.22e-12 \*\*\*  
## monthdec 3.019e-01 2.090e-01 1.444 0.148665   
## monthjul 1.086e-01 9.651e-02 1.125 0.260448   
## monthjun -5.371e-01 1.264e-01 -4.249 2.15e-05 \*\*\*  
## monthmar 1.979e+00 1.446e-01 13.691 < 2e-16 \*\*\*  
## monthmay -4.552e-01 8.274e-02 -5.502 3.75e-08 \*\*\*  
## monthnov -4.320e-01 1.213e-01 -3.561 0.000370 \*\*\*  
## monthoct 1.772e-01 1.541e-01 1.150 0.250330   
## monthsep 3.568e-01 1.799e-01 1.984 0.047296 \*   
## day\_of\_weekmon -1.206e-01 6.620e-02 -1.822 0.068490 .   
## day\_of\_weekthu 4.786e-02 6.415e-02 0.746 0.455604   
## day\_of\_weektue 9.237e-02 6.596e-02 1.400 0.161390   
## day\_of\_weekwed 1.646e-01 6.576e-02 2.503 0.012305 \*   
## duration 4.717e-03 7.471e-05 63.136 < 2e-16 \*\*\*  
## campaign -4.075e-02 1.157e-02 -3.522 0.000429 \*\*\*  
## pdays -3.150e-02 1.747e-02 -1.804 0.071288 .   
## previous -7.641e-02 6.015e-02 -1.270 0.203907   
## poutcomenonexistent 4.067e-01 9.504e-02 4.279 1.87e-05 \*\*\*  
## poutcomesuccess 7.896e-01 2.294e-01 3.442 0.000576 \*\*\*  
## emp.var.rate -1.743e+00 1.422e-01 -12.254 < 2e-16 \*\*\*  
## cons.price.idx 2.170e+00 2.526e-01 8.589 < 2e-16 \*\*\*  
## cons.conf.idx 1.964e-02 7.770e-03 2.527 0.011492 \*   
## euribor3m 3.333e-01 1.302e-01 2.560 0.010482 \*   
## nr.employed 5.280e-03 3.119e-03 1.693 0.090470 .   
## Age\_Grp32-37 -2.486e-01 6.959e-02 -3.572 0.000354 \*\*\*  
## Age\_Grp38-47 -3.575e-01 1.003e-01 -3.565 0.000363 \*\*\*  
## Age\_Grp47-55 -3.260e-01 1.460e-01 -2.233 0.025561 \*   
## Age\_Grp>55 -1.295e-01 2.034e-01 -0.637 0.524373   
## prevly\_CntctdYes -3.020e+01 1.726e+01 -1.750 0.080197 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28999 on 41187 degrees of freedom  
## Residual deviance: 17044 on 41130 degrees of freedom  
## AIC: 17160  
##   
## Number of Fisher Scoring iterations: 10

# Balancing the data set - more work needs to be done here  
bank\_bal\_yes <- bank\_addl\_full %>% filter(bank\_addl\_full$y=="yes")  
summary(bank\_bal\_yes)

## age job marital   
## Min. :17.00 admin. :1352 divorced: 476   
## 1st Qu.:31.00 technician : 730 married :2532   
## Median :37.00 blue-collar: 638 single :1620   
## Mean :40.91 retired : 434 unknown : 12   
## 3rd Qu.:50.00 management : 328   
## Max. :98.00 services : 323   
## (Other) : 835   
## education default housing loan   
## university.degree :1670 no :4197 no :2026 no :3850   
## high.school :1031 unknown: 443 unknown: 107 unknown: 107   
## professional.course: 595 yes : 0 yes :2507 yes : 683   
## basic.9y : 473   
## basic.4y : 428   
## unknown : 251   
## (Other) : 192   
## contact month day\_of\_week duration   
## cellular :3853 may :886 fri: 846 Min. : 37.0   
## telephone: 787 aug :655 mon: 847 1st Qu.: 253.0   
## jul :649 thu:1045 Median : 449.0   
## jun :559 tue: 953 Mean : 553.2   
## apr :539 wed: 949 3rd Qu.: 741.2   
## nov :416 Max. :4199.0   
## (Other):936   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0 Min. :0.0000 failure : 605   
## 1st Qu.: 1.000 1st Qu.:999 1st Qu.:0.0000 nonexistent:3141   
## Median : 2.000 Median :999 Median :0.0000 success : 894   
## Mean : 2.052 Mean :792 Mean :0.4927   
## 3rd Qu.: 2.000 3rd Qu.:999 3rd Qu.:1.0000   
## Max. :23.000 Max. :999 Max. :6.0000   
##   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m   
## Min. :-3.400 Min. :92.20 Min. :-50.80 Min. :0.634   
## 1st Qu.:-1.800 1st Qu.:92.89 1st Qu.:-46.20 1st Qu.:0.849   
## Median :-1.800 Median :93.20 Median :-40.40 Median :1.266   
## Mean :-1.233 Mean :93.35 Mean :-39.79 Mean :2.123   
## 3rd Qu.:-0.100 3rd Qu.:93.92 3rd Qu.:-36.10 3rd Qu.:4.406   
## Max. : 1.400 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr.employed y Age\_Grp prevly\_Cntctd   
## Min. :4964 no : 0 17-31:1344 Length:4640   
## 1st Qu.:5018 yes:4640 32-37:1036 Class :character   
## Median :5099 38-47: 881 Mode :character   
## Mean :5095 47-55: 638   
## 3rd Qu.:5191 >55 : 741   
## Max. :5228   
##

bank\_bal\_no <- bank\_addl\_full %>% filter(bank\_addl\_full$y=="no" , !bank\_addl\_full$housing=="unknown", !marital=="unknown", !default=="unknown", !job=="unknown",!education=="unknown" )  
summary(bank\_bal\_no)

## age job marital   
## Min. :17.00 admin. :7521 divorced: 3143   
## 1st Qu.:32.00 blue-collar :5223 married :15436   
## Median :37.00 technician :4832 single : 8050   
## Mean :38.84 services :2599 unknown : 0   
## 3rd Qu.:45.00 management :2025   
## Max. :95.00 entrepreneur: 988   
## (Other) :3441   
## education default housing   
## university.degree :8870 no :26626 no :12250   
## high.school :6765 unknown: 0 unknown: 0   
## basic.9y :3896 yes : 3 yes :14379   
## professional.course:3783   
## basic.4y :2054   
## basic.6y :1253   
## (Other) : 8   
## loan contact month day\_of\_week  
## no :22446 cellular :17170 may :9033 fri:5058   
## unknown: 0 telephone: 9459 jul :4569 mon:5573   
## yes : 4183 aug :4140 thu:5516   
## jun :3162 tue:5166   
## nov :3131 wed:5316   
## apr :1647   
## (Other): 947   
## duration campaign pdays previous   
## Min. : 0.0 Min. : 1.000 Min. : 0.0 Min. :0.0000   
## 1st Qu.: 95.0 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.0000   
## Median : 163.0 Median : 2.000 Median :999.0 Median :0.0000   
## Mean : 220.3 Mean : 2.593 Mean :981.4 Mean :0.1489   
## 3rd Qu.: 277.0 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.0000   
## Max. :4918.0 Max. :43.000 Max. :999.0 Max. :7.0000   
##   
## poutcome emp.var.rate cons.price.idx cons.conf.idx   
## failure : 2953 Min. :-3.4000 Min. :92.20 Min. :-50.80   
## nonexistent:23264 1st Qu.:-1.8000 1st Qu.:93.08 1st Qu.:-42.70   
## success : 412 Median : 1.1000 Median :93.44 Median :-41.80   
## Mean : 0.1157 Mean :93.55 Mean :-40.71   
## 3rd Qu.: 1.4000 3rd Qu.:93.99 3rd Qu.:-36.40   
## Max. : 1.4000 Max. :94.77 Max. :-26.90   
##   
## euribor3m nr.employed y Age\_Grp   
## Min. :0.634 Min. :4964 no :26629 17-31:6650   
## 1st Qu.:1.365 1st Qu.:5099 yes: 0 32-37:7531   
## Median :4.857 Median :5196 38-47:6617   
## Mean :3.674 Mean :5171 47-55:4102   
## 3rd Qu.:4.961 3rd Qu.:5228 >55 :1729   
## Max. :5.045 Max. :5228   
##   
## prevly\_Cntctd   
## Length:26629   
## Class :character   
## Mode :character   
##   
##   
##   
##

nrow(bank\_bal\_no)

## [1] 26629

bank\_bal <- rbind(bank\_bal\_yes,bank\_bal\_no)  
nrow(bank\_bal)

## [1] 31269

summary(bank\_bal)

## age job marital   
## Min. :17.00 admin. :8873 divorced: 3619   
## 1st Qu.:31.00 blue-collar:5861 married :17968   
## Median :37.00 technician :5562 single : 9670   
## Mean :39.15 services :2922 unknown : 12   
## 3rd Qu.:46.00 management :2353   
## Max. :98.00 retired :1293   
## (Other) :4405   
## education default housing   
## university.degree :10540 no :30823 no :14276   
## high.school : 7796 unknown: 443 unknown: 107   
## professional.course: 4378 yes : 3 yes :16886   
## basic.9y : 4369   
## basic.4y : 2482   
## basic.6y : 1441   
## (Other) : 263   
## loan contact month day\_of\_week  
## no :26296 cellular :21023 may :9919 fri:5904   
## unknown: 107 telephone:10246 jul :5218 mon:6420   
## yes : 4866 aug :4795 thu:6561   
## jun :3721 tue:6119   
## nov :3547 wed:6265   
## apr :2186   
## (Other):1883   
## duration campaign pdays previous   
## Min. : 0.0 Min. : 1.000 Min. : 0.0 Min. :0.0000   
## 1st Qu.: 105.0 1st Qu.: 1.000 1st Qu.:999.0 1st Qu.:0.0000   
## Median : 185.0 Median : 2.000 Median :999.0 Median :0.0000   
## Mean : 269.7 Mean : 2.513 Mean :953.3 Mean :0.1999   
## 3rd Qu.: 332.0 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.0000   
## Max. :4918.0 Max. :43.000 Max. :999.0 Max. :7.0000   
##   
## poutcome emp.var.rate cons.price.idx cons.conf.idx   
## failure : 3558 Min. :-3.40000 Min. :92.20 Min. :-50.80   
## nonexistent:26405 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70   
## success : 1306 Median : 1.10000 Median :93.44 Median :-41.80   
## Mean :-0.08451 Mean :93.52 Mean :-40.58   
## 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 Age\_Grp   
## Min. :0.634 Min. :4964 no :26629 17-31:7994   
## 1st Qu.:1.313 1st Qu.:5099 yes: 4640 32-37:8567   
## Median :4.856 Median :5191 38-47:7498   
## Mean :3.444 Mean :5160 47-55:4740   
## 3rd Qu.:4.961 3rd Qu.:5228 >55 :2470   
## Max. :5.045 Max. :5228   
##   
## prevly\_Cntctd   
## Length:31269   
## Class :character   
## Mode :character   
##   
##   
##   
##

#Running logistic regression on full data set  
simple.log<-glm(y~.,family="binomial",data=bank\_bal)

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

summary(simple.log)

##   
## Call:  
## glm(formula = y ~ ., family = "binomial", data = bank\_bal)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.8711 -0.3203 -0.1934 -0.1365 3.4623   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.511e+02 4.601e+01 -5.457 4.85e-08 \*\*\*  
## age 3.496e-03 6.680e-03 0.523 0.600768   
## jobblue-collar -2.051e-01 9.085e-02 -2.258 0.023959 \*   
## jobentrepreneur -1.671e-01 1.395e-01 -1.198 0.231027   
## jobhousemaid -6.769e-03 1.675e-01 -0.040 0.967768   
## jobmanagement -2.155e-02 9.197e-02 -0.234 0.814762   
## jobretired 1.329e-01 1.355e-01 0.980 0.326895   
## jobself-employed -9.592e-02 1.254e-01 -0.765 0.444369   
## jobservices -1.701e-01 9.552e-02 -1.781 0.074961 .   
## jobstudent 2.008e-01 1.269e-01 1.582 0.113635   
## jobtechnician 4.458e-02 7.622e-02 0.585 0.558615   
## jobunemployed 3.017e-02 1.384e-01 0.218 0.827414   
## jobunknown 1.658e+01 4.534e+02 0.037 0.970829   
## maritalmarried -1.059e-02 7.533e-02 -0.141 0.888157   
## maritalsingle -2.661e-02 8.610e-02 -0.309 0.757292   
## maritalunknown 1.721e+01 8.832e+02 0.019 0.984457   
## educationbasic.6y 1.580e-01 1.447e-01 1.092 0.274867   
## educationbasic.9y 4.591e-02 1.115e-01 0.412 0.680575   
## educationhigh.school 1.150e-01 1.063e-01 1.082 0.279192   
## educationilliterate 1.748e+00 8.518e-01 2.052 0.040140 \*   
## educationprofessional.course 1.629e-01 1.151e-01 1.415 0.157172   
## educationuniversity.degree 2.597e-01 1.064e-01 2.442 0.014612 \*   
## educationunknown 1.790e+01 1.943e+02 0.092 0.926613   
## defaultunknown 1.875e+01 1.538e+02 0.122 0.902965   
## defaultyes -1.335e+01 2.281e+03 -0.006 0.995330   
## housingunknown 1.822e+01 2.965e+02 0.061 0.950999   
## housingyes -2.089e-02 4.505e-02 -0.464 0.642800   
## loanunknown NA NA NA NA   
## loanyes -5.640e-02 6.238e-02 -0.904 0.365973   
## contacttelephone -6.614e-01 8.291e-02 -7.977 1.50e-15 \*\*\*  
## monthaug 8.371e-01 1.301e-01 6.432 1.26e-10 \*\*\*  
## monthdec 2.208e-01 2.269e-01 0.973 0.330613   
## monthjul 2.553e-02 1.066e-01 0.239 0.810739   
## monthjun -6.805e-01 1.348e-01 -5.048 4.46e-07 \*\*\*  
## monthmar 2.061e+00 1.551e-01 13.287 < 2e-16 \*\*\*  
## monthmay -4.433e-01 8.937e-02 -4.961 7.02e-07 \*\*\*  
## monthnov -4.753e-01 1.339e-01 -3.551 0.000384 \*\*\*  
## monthoct 2.366e-01 1.687e-01 1.403 0.160673   
## monthsep 4.543e-01 1.964e-01 2.313 0.020720 \*   
## day\_of\_weekmon -6.987e-02 7.326e-02 -0.954 0.340258   
## day\_of\_weekthu 1.166e-01 7.118e-02 1.638 0.101358   
## day\_of\_weektue 1.572e-01 7.316e-02 2.148 0.031685 \*   
## day\_of\_weekwed 2.341e-01 7.274e-02 3.219 0.001288 \*\*   
## duration 4.564e-03 8.417e-05 54.219 < 2e-16 \*\*\*  
## campaign -4.052e-02 1.297e-02 -3.123 0.001789 \*\*   
## pdays -2.171e-02 1.857e-02 -1.169 0.242341   
## previous -6.062e-02 6.461e-02 -0.938 0.348134   
## poutcomenonexistent 4.551e-01 1.027e-01 4.431 9.38e-06 \*\*\*  
## poutcomesuccess 8.252e-01 2.459e-01 3.355 0.000793 \*\*\*  
## emp.var.rate -1.888e+00 1.512e-01 -12.491 < 2e-16 \*\*\*  
## cons.price.idx 2.426e+00 2.740e-01 8.854 < 2e-16 \*\*\*  
## cons.conf.idx 2.043e-02 8.429e-03 2.424 0.015344 \*   
## euribor3m 2.917e-01 1.465e-01 1.991 0.046518 \*   
## nr.employed 7.998e-03 3.450e-03 2.318 0.020435 \*   
## Age\_Grp32-37 -2.343e-01 7.527e-02 -3.113 0.001855 \*\*   
## Age\_Grp38-47 -3.230e-01 1.118e-01 -2.889 0.003869 \*\*   
## Age\_Grp47-55 -2.712e-01 1.657e-01 -1.637 0.101637   
## Age\_Grp>55 -1.346e-01 2.324e-01 -0.579 0.562556   
## prevly\_CntctdYes -2.050e+01 1.836e+01 -1.117 0.264073   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26260 on 31268 degrees of freedom  
## Residual deviance: 13881 on 31211 degrees of freedom  
## AIC: 13997  
##   
## Number of Fisher Scoring iterations: 16

#PC With scaled data -- Still need to make this work  
#pc.result<-prcomp(bank\_addl\_full[,],scale.=TRUE)  
#pc.scores<-pc.result$x  
#pairs(pc.scores)  
#cor(pc.scores)  
#var.pca<-apply(pc.scores,2,var)  
#var.pca  
#sum(var.pca)  
#pc.result$rotation  
#scree plot   
#par(mfrow=c(1,2))  
#eigenvals<-(pc.result$sdev)^2  
#plot(1:17,eigenvals/sum(eigenvals),type="l",main="Scree Plot",ylab="Prop. Var. Explained")  
#cumulative.prop<-cumsum(eigenvals/sum(eigenvals))  
#plot(1:17,cumulative.prop,type="l",main="Cumulative proportion",ylim=c(0,1))  
#par(mfrow=c(1,1))