```r
#loading the libraries
library(forecast)
library(readr)
library(ggplot2)
library(lattice)
library(plyr)
library(dplyr)
library(caret)
library(mlbench)
library(foreign)
library(ggplot2)
library(reshape)
#loading the data
alldata <- read.csv('../bank-additional-full.csv', stringsAsFactors
age = alldata$age
job = alldata$job
marital = alldata$marital
```

12:17 (Top Level) R Script

Console Terminal ×

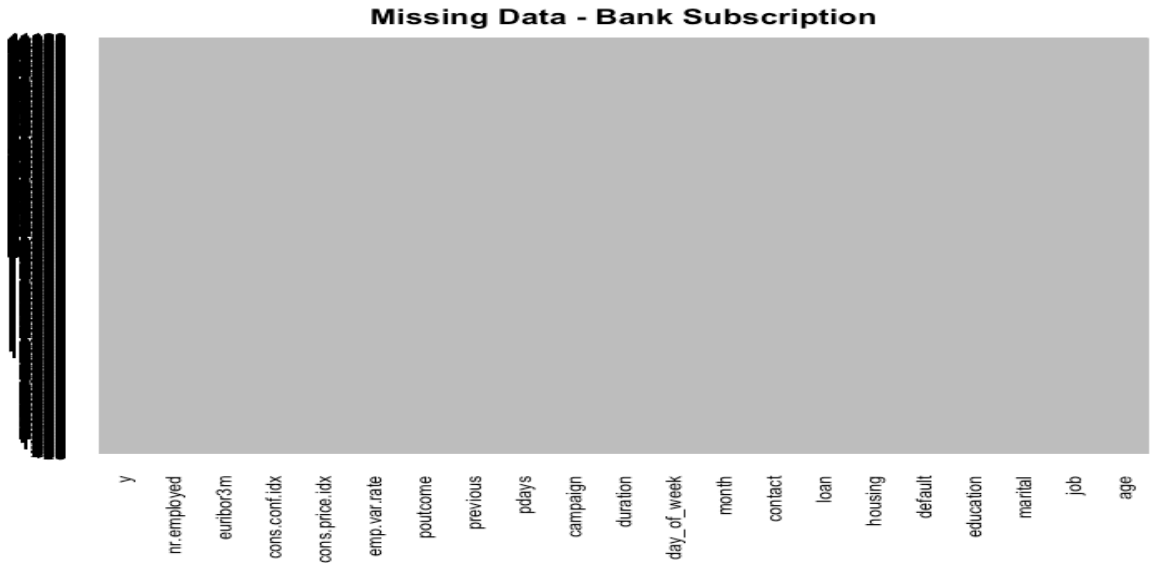
~/Desktop/final proj/

```
> library(forecast)
> library(readr)
> library(ggplot2)
> library(lattice)
> library(plyr)
> library(dplyr)
> library(caret)
> library(mlbench)
> library(foreign)
> library(ggplot2)
> library(reshape)
>
```

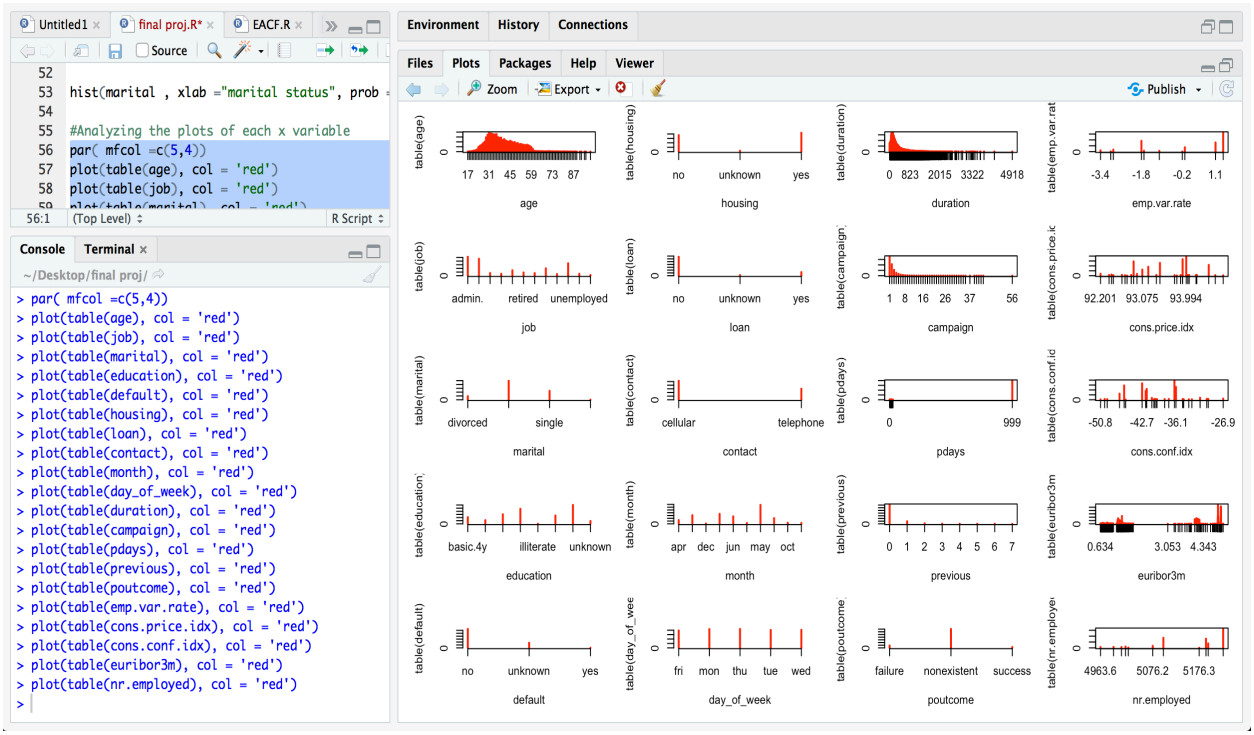
```
> alldata <- read.csv('../bank-additional-full.csv', stringsAsFactors = T, sep = ';')
> age = alldata$age
> job = alldata$job
> marital = alldata$marital
> education = alldata$education
> default = alldata$default
> housing = alldata$housing
> loan = alldata$loan
> contact = alldata$contact
> month = alldata$month
> day_of_week = alldata$day_of_week
> duration = alldata$duration
> campaign = alldata$campaign
> pdays = alldata$pdays
> previous = alldata$previous
> poutcome = alldata$poutcome
> emp.var.rate = alldata$emp.var.rate
> cons.price.idx = alldata$cons.price.idx
> cons.conf.idx = alldata$cons.conf.idx
> euribor3m= alldata$euribor3m
> nr.employed = alldata$nr.employed
> term_deposit = alldata$y
>
```

Missing Values

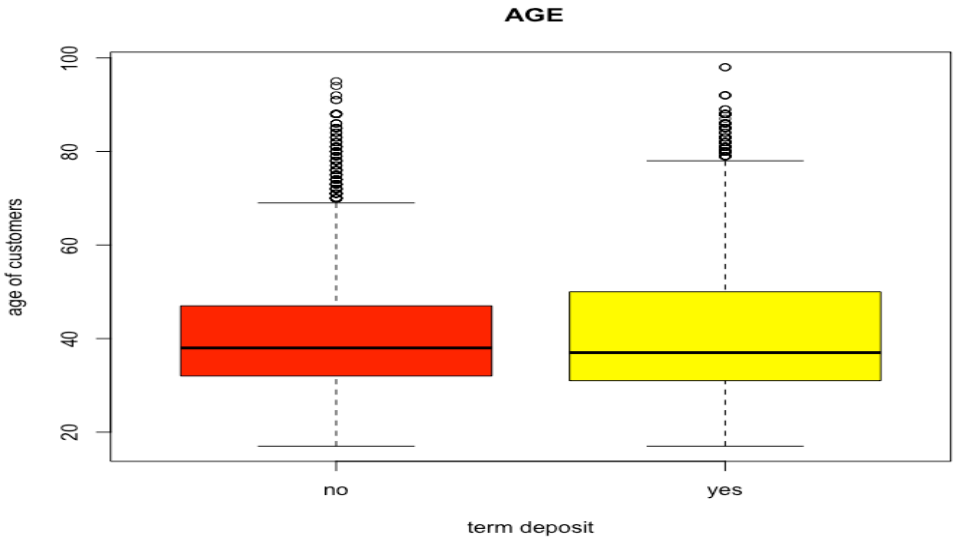
```
> missmap(alldata,main="Missing Data - Bank Subscription", col=c("re
d","grey"),legend=FALSE)
>
```



spread of data in each of the x variables.

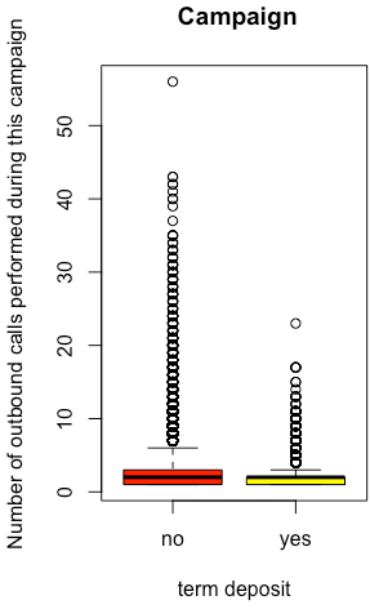


1. **Age**

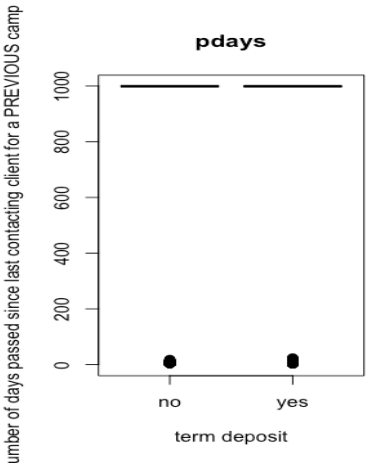


2. Campaign

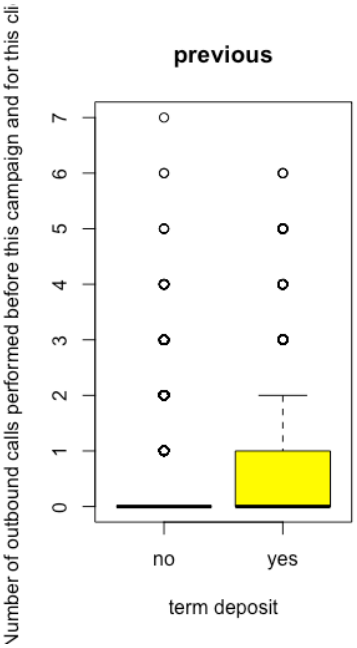
.



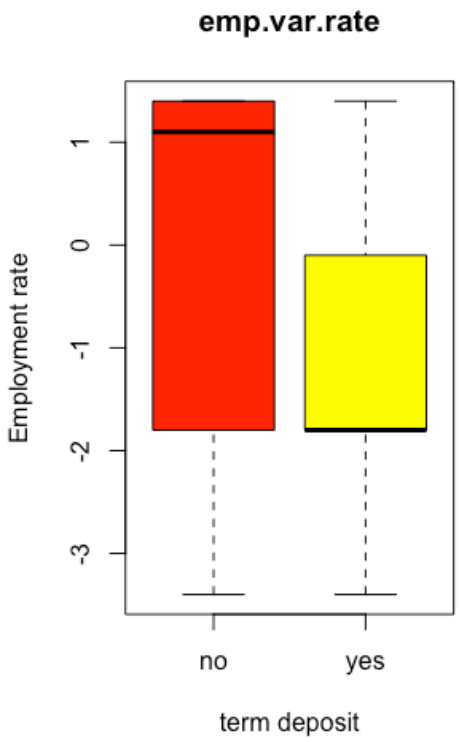
3. pdays



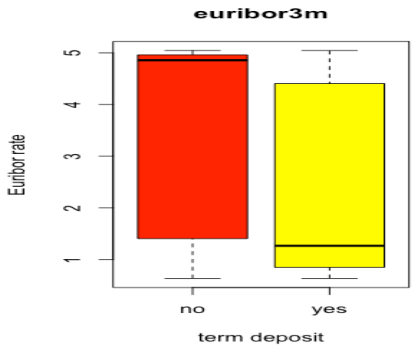
4. previous



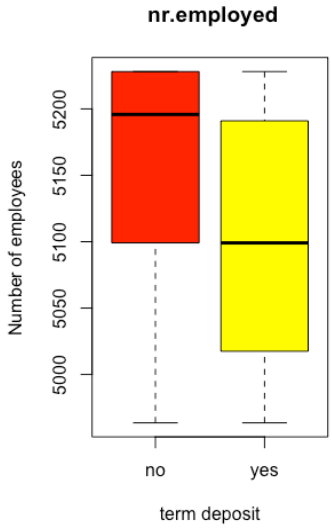
5. emp.var.rate



6. euribor3m

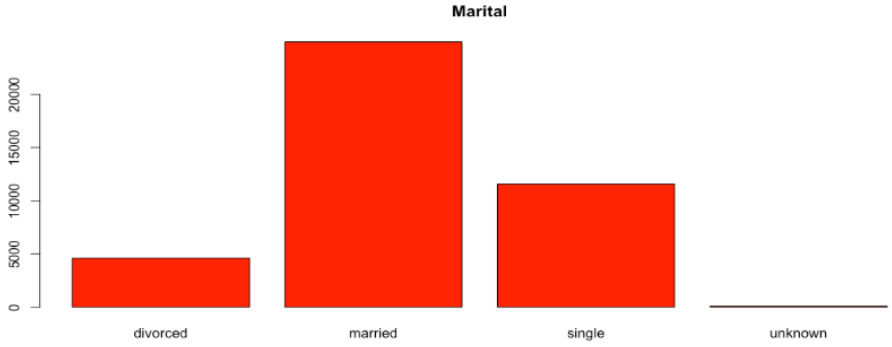


7. nr.employed

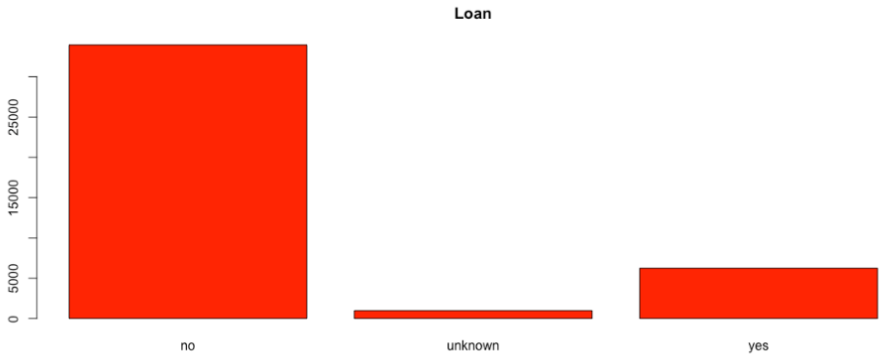


Bar plots for categorical variables:

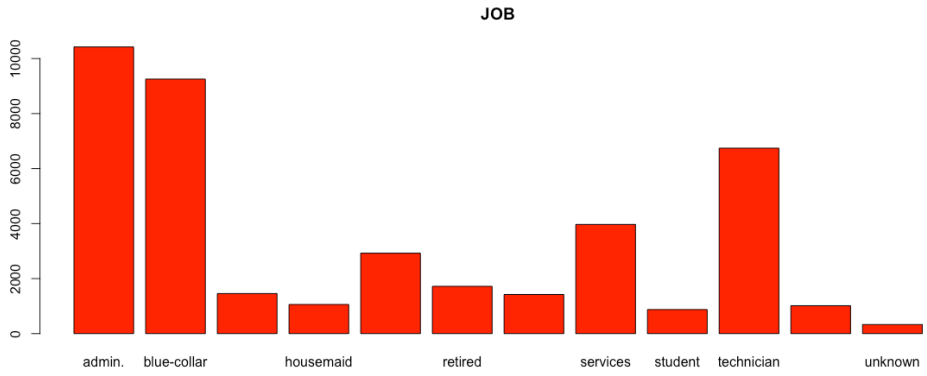
1. **Marital**



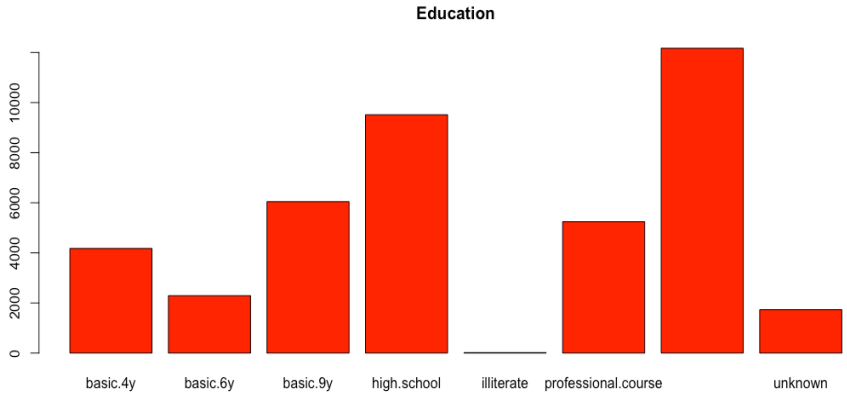
2. Loan


Loan
0
5000
15000
25000
no
unknown
yes

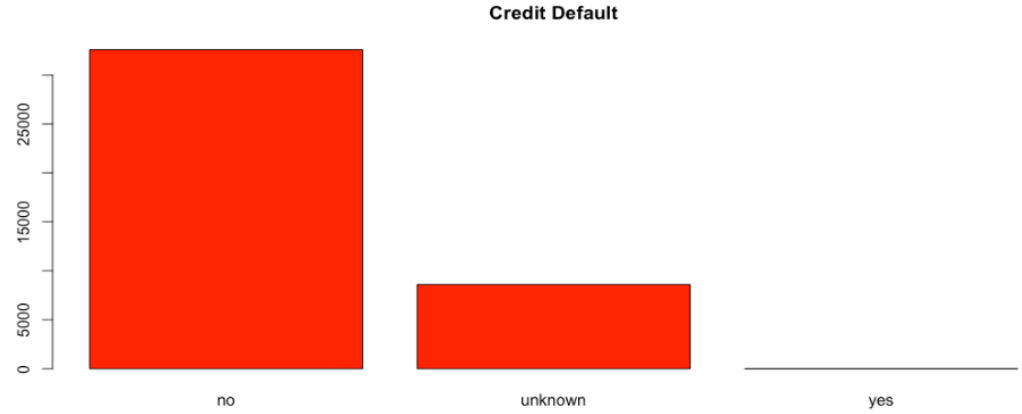
3. Job:


JOB
0
2000
4000
6000
8000
10000
admin.
blue-collar
housemaid
retired
services
student
technician
unknown

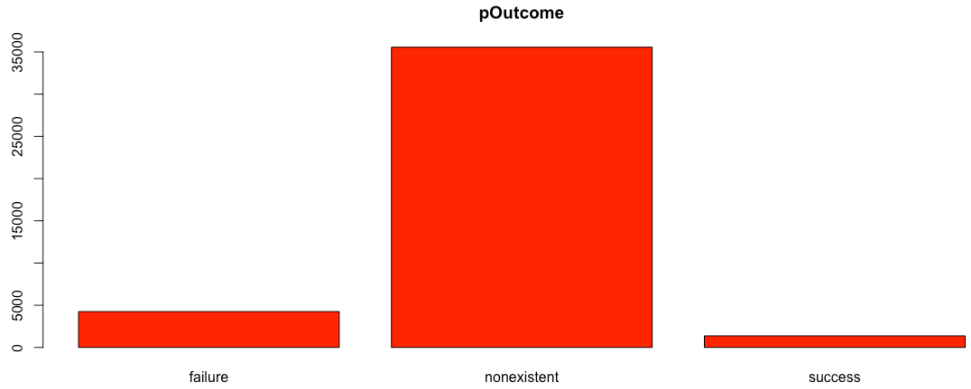
4. Education:



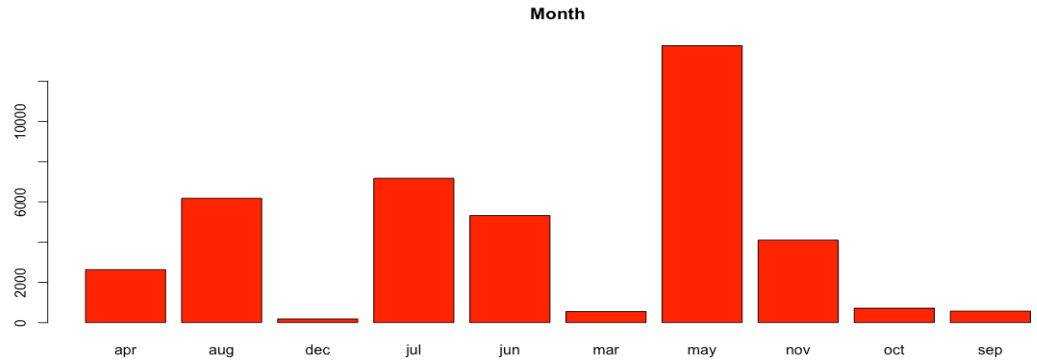
5. Default:



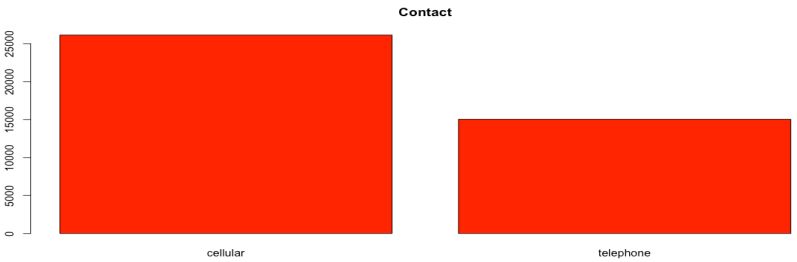
6. pOutcome:



7. Month:



8. Contact:

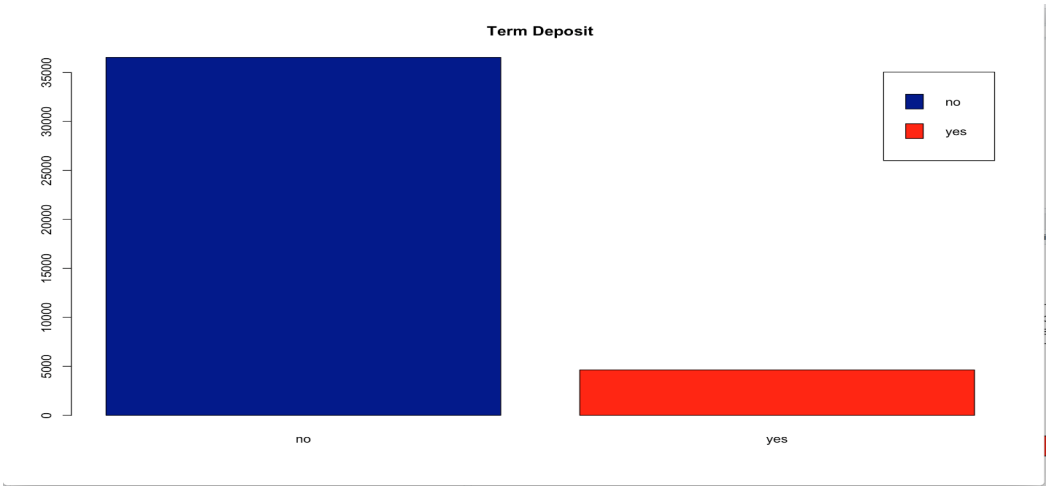


Correlation:

```
> alldata.cont<-data.frame(alldata$age,alldata$campaign,alldata$pdays,alldata$previous,alldata$emp.var.rate,alldata$cons.price.idx,alldata$cons.conf.idx, alldata$euribor3m,
alldata$nr.employed)
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> cor(alldata.cont)
                      alldata.age alldata.campaign alldata.pdays alldata.previous alldata.emp.var.rate alldata.cons.price.idx
alldata.age           1.0000000000      0.00459358   -0.03436895      0.02436474       -0.0003706855           0.000856715
alldata.campaign      0.0045935805      1.00000000    0.05258357     -0.07914147        0.1507538056           0.127835912
alldata.pdays        -0.0343689512      0.05258357    1.00000000     -0.58751386        0.2710041743           0.078889109
alldata.previous      0.0243647409     -0.07914147   -0.58751386      1.00000000       -0.4204891094          -0.203129967
alldata.emp.var.rate -0.0003706855      0.15075381    0.27100417     -0.42048911        1.0000000000           0.775334171
alldata.cons.price.idx 0.0008567150     0.12783591    0.07888911     -0.20312997        0.7753341708           1.000000000
alldata.cons.conf.idx  0.1293716142    -0.01373310   -0.09134235     -0.05093635        0.1960412681           0.058986182
alldata.euribor3m     0.0107674295      0.13513251    0.29689911     -0.45449365        0.9722446712           0.688230107
alldata.nr.employed  -0.0177251319      0.14409489    0.37260474     -0.50133293        0.9069701013           0.522033977
                     alldata.cons.conf.idx alldata.euribor3m alldata.nr.employed
alldata.age                     0.12937161       0.01076743        -0.01772513
alldata.campaign               -0.01373310       0.13513251         0.14409489
alldata.pdays                  -0.09134235       0.29689911         0.37260474
alldata.previous               -0.05093635      -0.45449365        -0.50133293
alldata.emp.var.rate            0.19604127       0.97224467         0.90697010
alldata.cons.price.idx          0.05898618       0.68823011         0.52203398
alldata.cons.conf.idx           1.00000000       0.27768622         0.10051343
alldata.euribor3m               0.27768622       1.00000000         0.94515443
alldata.nr.employed             0.10051343       0.94515443         1.00000000
```

DATA BALANCE TESTING

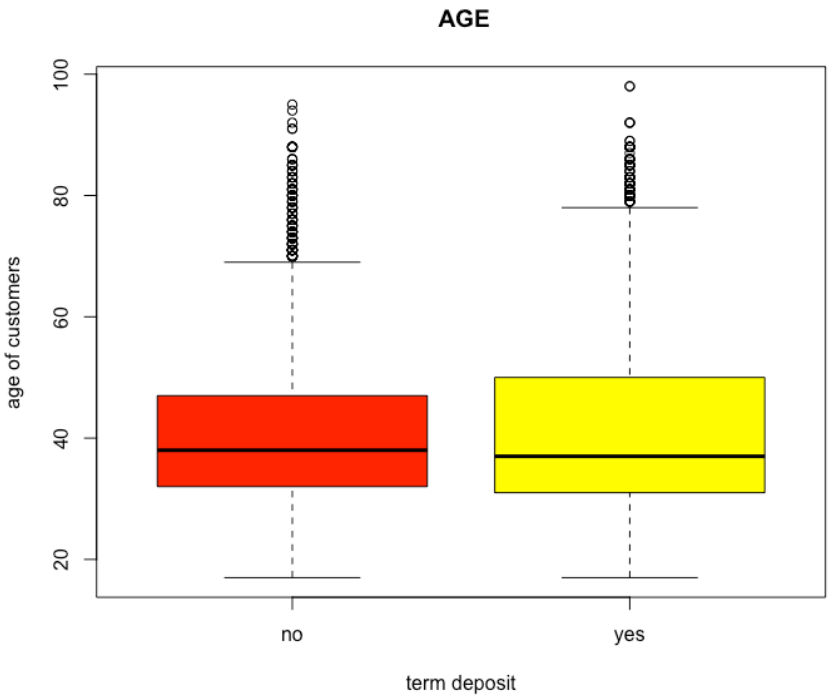
```
> table(alldata$y)

    no    yes
36548   4640
```




```
> alldata_n=upSample(x, y, list = FALSE, yname = "y")
> table(alldata_n$y)

   no   yes
36548 36548
>
```

HYPOTHESIS TESTING:

1. Two sample One-tailed Hypothesis Testing




```
> hypo1 <- z.test(age_y, age_n, alternative= "less", mu = 0, sigma.x=sd(age_y), sigma.y=sd(age_n), conf.level=0.95)
> hypo1

        Two-sample z-Test

data:  age_y and age_n
z = 4.7795, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       NA 1.346781
sample estimates:
mean of x mean of y
 40.91315  39.91119
```

Model 1: Logistic Regression

a) Full model:

```
> summary(fullu)

Call:
glm(formula = Class ~ ., family = binomial(), data = datasetu)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.87945  -0.85952   0.07715   0.80935   2.05917

Coefficients: (10 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -4.408e-01  1.532e+01  -0.029 0.977044
age                        -5.232e-03  1.122e-02  -0.466 0.641027
jobadmin.                  -4.209e-03  4.512e-02  -0.093 0.925674
`jobblue-collar`           -4.351e-02  4.350e-02  -1.000 0.317112
jobentrepreneur             6.224e-03  2.070e-02   0.301 0.763669
jobhousemaid               -3.156e-02  1.858e-02  -1.699 0.089307 .
jobmanagement              -1.137e-02  2.762e-02  -0.412 0.680635
jobretired                  6.367e-02  2.215e-02   2.874 0.004047 **
`jobself-employed`         -1.535e-02  2.055e-02  -0.747 0.454938
emp.var.rate               -2.468e+00  1.099e-01 -22.462  < 2e-16 ***
cons.price.idx              1.152e+00  6.954e-02  16.571  < 2e-16 ***
cons.conf.idx               8.165e-02  1.980e-02   4.124 3.73e-05 ***
euribor3m                   6.676e-01  1.075e-01   6.207 5.38e-10 ***
nr.employed                 2.969e-01  1.040e-01   2.856 0.004291 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101333  on 73095  degrees of freedom
Residual deviance:  78139  on 73043  degrees of freedom
AIC: 78245

Number of Fisher Scoring iterations: 25
```

b) Stepwise Forward Model:

Building a base model with one x-variable and building a stepwise forward model :

```
> #Building a base model with one x-variable and building a step-wise forward model using it: upsampli
ng
> baseu = glm(class~age, datasetu, family = binomial())
> forwardu = step(baseu, scope = list(upper = fullu, lower = ~1), direction = 'forward', trace = F)
> summary(forwardu)

Call:
glm(formula = Class ~ age + nr.employed + monthmay + pdays +
    poutcomefailure + monthmar + contactcellular + monthnov +
    campaign + defaultno + day_of_weekmon + jobretired + emp.var.rate +
    euribor3m + cons.price.idx + educationuniversity.degree +
    monthjun + monthaug + monthapr + jobstudent + cons.conf.idx +
    day_of_weekwed + educationbasic.4y + maritalunknown + maritalsingle +
    previous + educationbasic.9y + jobhousemaid + monthdec +
    educationilliterate + housingno + day_of_weektue + maritaldivorced +
    poutcomenonexistent + jobadmin. + jobentrepreneur + educationbasic.6y +
    `jobblue-collar` + housingunknown + jobservices, family = binomial(),
    data = datasetu)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.87644  -0.86009  -0.05521   0.81072   2.07771

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.438914   0.009141 -48.016  < 2e-16 ***
age                       -0.005015   0.011040  -0.454 0.649656
nr.employed                0.299134   0.083971   3.562 0.000368 ***
monthmay                  -0.289625   0.015286 -18.947  < 2e-16 ***
pdays                     -0.228261   0.027452  -8.315  < 2e-16 ***
poutcomefailure           -0.206170   0.044477  -4.635 3.56e-06 ***
monthmar                   0.152997   0.008711  17.565  < 2e-16 ***
contactcellular            0.303586   0.015946  19.039  < 2e-16 ***
monthnov                  -0.198289   0.012146 -16.325  < 2e-16 ***
campaign                  -0.088899   0.010873  -8.176 2.94e-16 ***
defaultno                  0.071389   0.010186   7.008 2.41e-12 ***
day_of_weekmon            -0.080131   0.009770  -8.201 2.37e-16 ***
jobretired                 0.071682   0.009861   7.269 3.62e-13 ***
emp.var.rate              -2.440700   0.098078 -24.885  < 2e-16 ***
euribor3m                  0.643467   0.095719   6.722 1.79e-11 ***
cons.price.idx             1.146279   0.059725  19.193  < 2e-16 ***
educationuniversity.degree 0.041085   0.010147   4.049 5.14e-05 ***
monthjun                  -0.280262   0.019998 -14.014  < 2e-16 ***
monthaug                   0.177570   0.017881   9.931  < 2e-16 ***
monthapr                  -0.053222   0.010350  -5.142 2.71e-07 ***
jobstudent                 0.032371   0.008498   3.809 0.000139 ***
cons.conf.idx              0.084095   0.018924   4.444 8.84e-06 ***
day_of_weekwed             0.028868   0.009503   3.038 0.002384 **
educationbasic.4y         -0.026555   0.011129  -2.386 0.017026 *
maritalunknown             0.026685   0.007718   3.457 0.000546 ***
maritalsingle              0.029027   0.009983   2.908 0.003642 **
previous                  -0.057250   0.020207  -2.833 0.004609 **
educationbasic.9y         -0.013308   0.010868  -1.225 0.220761
jobhousemaid              -0.025197   0.009867  -2.554 0.010660 *
monthdec                   0.017927   0.008113   2.210 0.027134 *
educationilliterate        0.016203   0.007370   2.199 0.027906 *
housingno                  0.017908   0.008887   2.015 0.043895 *
day_of_weektue            -0.020986   0.009671  -2.170 0.029999 *
maritaldivorced           -0.019623   0.009271  -2.117 0.034290 *
poutcomenonexistent       -0.101359   0.049515  -2.047 0.040654 *
jobadmin.                  0.013545   0.010197   1.328 0.184071
jobentrepreneur            0.013636   0.009147   1.491 0.136040
educationbasic.6y          0.024767   0.009866   2.510 0.012065 *
`jobblue-collar`          -0.027343   0.012841  -2.129 0.033222 *
housingunknown            -0.015256   0.008996  -1.696 0.089917 .
jobservices               -0.015674   0.010161  -1.543 0.122912
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101333  on 73095  degrees of freedom
Residual deviance:  78146  on 73055  degrees of freedom
AIC: 78228

Number of Fisher Scoring iterations: 5
```

c) Stepwise Backward Model:

Building a stepwise backward model :

```
> backwardu = step(fullu, direction = 'backward', trace = F)
There were 50 or more warnings (use warnings() to see the first 50)
> summary(backwardu)

Call:
glm(formula = Class ~ `jobblue-collar` + jobhousemaid + jobretired +
    `jobself-employed` + jobservices + jobstudent + maritaldivorced +
    maritalmarried + maritalsingle + educationbasic.4y + educationbasic.6y +
    educationilliterate + educationuniversity.degree + defaultunknown +
    housingno + housingunknown + contactcellular + monthapr +
    monthaug + monthdec + monthjun + monthmar + monthmay + monthnov +
    day_of_weekfri + day_of_weekmon + day_of_weekthu + day_of_weektue +
    campaign + pdays + previous + poutcomefailure + poutcomenonexistent +
    emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m +
    nr.employed, family = binomial(), data = datasetu)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.87619  -0.86006  -0.05503   0.80913   2.07987

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.438855   0.009141 -48.012  < 2e-16 ***
`jobblue-collar`     -0.041662   0.011147  -3.738 0.000186 ***
jobhousemaid         -0.029602   0.009697  -3.053 0.002268 **
jobretired            0.064923   0.008758   7.413 1.23e-13 ***
`jobself-employed`   -0.013070   0.008870  -1.473 0.140618
jobservices          -0.021147   0.009689  -2.183 0.029066 *
jobstudent            0.029952   0.008209   3.649 0.000264 ***
maritaldivorced      -0.209271   0.055790  -3.751 0.000176 ***
maritalmarried       -0.293765   0.085659  -3.429 0.000605 ***
```

```
maritalsingle               -0.239197   0.078937  -3.030 0.002444 **
educationbasic.4y           -0.022169   0.010385  -2.135 0.032787 *
educationbasic.6y            0.028597   0.009405   3.041 0.002362 **
educationilliterate          0.016901   0.007372   2.292 0.021879 *
educationuniversity.degree   0.046353   0.009824   4.718 2.38e-06 ***
defaultunknown              -0.071765   0.010098  -7.107 1.19e-12 ***
housingno                    0.017924   0.008886   2.017 0.043681 *
housingunknown              -0.015445   0.008994  -1.717 0.085921 .
contactcellular              0.302913   0.015943  19.000  < 2e-16 ***
monthapr                    -0.053941   0.010348  -5.213 1.86e-07 ***
monthaug                     0.177288   0.017860   9.927  < 2e-16 ***
monthdec                     0.017709   0.008113   2.183 0.029049 *
monthjun                    -0.281430   0.019975 -14.089  < 2e-16 ***
monthmar                     0.153092   0.008707  17.582  < 2e-16 ***
monthmay                    -0.289583   0.015288 -18.942  < 2e-16 ***
monthnov                    -0.198476   0.012126 -16.368  < 2e-16 ***
day_of_weekfri              -0.035107   0.011021  -3.185 0.001446 **
day_of_weekmon              -0.109617   0.011254  -9.741  < 2e-16 ***
day_of_weekthu              -0.023597   0.011025  -2.140 0.032337 *
day_of_weektue              -0.050132   0.011060  -4.533 5.82e-06 ***
campaign                    -0.088609   0.010867  -8.154 3.53e-16 ***
pdays                       -0.228440   0.027447  -8.323  < 2e-16 ***
previous                    -0.057455   0.020210  -2.843 0.004471 **
poutcomefailure             -0.204981   0.044465  -4.610 4.03e-06 ***
poutcomenonexistent         -0.100058   0.049497  -2.022 0.043227 *
emp.var.rate                -2.446752   0.098013 -24.964  < 2e-16 ***
cons.price.idx               1.149875   0.059665  19.272  < 2e-16 ***
cons.conf.idx                0.084068   0.018884   4.452 8.52e-06 ***
euribor3m                    0.641371   0.095667   6.704 2.03e-11 ***
nr.employed                  0.304499   0.083894   3.630 0.000284 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101333  on 73095  degrees of freedom
Residual deviance:  78148  on 73057  degrees of freedom
AIC: 78226

Number of Fisher Scoring iterations: 5
```

Performing cross-validation:

```
> cv.glm(fullu, data = datasetu, K=10)$delta
[1] 0.1787567 0.1787413
There were 29 warnings (use warnings() to see them)
> cv.glm(forwardu, data = datasetu, K=10)$delta
[1] 0.1787156 0.1787043
> cv.glm(backwardu, data = datasetu, K=10)$delta
[1] 0.1786781 0.1786692
```

Model 2: Naïve Bayes

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   32.00   38.00   40.02   47.00   98.00
> summary(campaign)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.568   3.000  56.000
> summary(pdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   999.0   999.0   962.5   999.0   999.0
> summary(previous)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.173   0.000   7.000
> summary(emp.var.rate)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.40000 -1.80000  1.10000  0.08189  1.40000  1.40000
> summary(cons.conf.idx)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -50.8   -42.7   -41.8   -40.5   -36.4   -26.9
> summary(cons.price.idx)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  92.20   93.08   93.75   93.58   93.99   94.77
> summary(euribor3m)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.634   1.344   4.857   3.621   4.961   5.045
> summary(nr.employed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4964    5099    5191    5167    5228    5228
>
```

```
> data_nb = alldata_n
> data_nb$age <- cut(data_nb$age, breaks = c(-Inf,32,38,47,Inf), labels= c("young", "middle-aged", "old", "very-old"), right = FALSE)
> data_nb$campaign <- cut(data_nb$campaign, breaks = c(-Inf,2.568,Inf), labels= c("c1", "c2"), right = FALSE)
> data_nb$pdays <- cut(data_nb$pdays, breaks = c(-Inf,962.5,Inf), labels= c("p1", "p2"), right = FALSE)
> data_nb$previous <- cut(data_nb$previous, breaks = c(-Inf,0.173,Inf), labels= c("prev1", "prev2"), right = FALSE)
> data_nb$emp.var.rate <- cut(data_nb$emp.var.rate, breaks = c(-Inf,0.08189,Inf), labels= c("emp1", "emp2"), right = FALSE)
> data_nb$cons.conf.idx <- cut(data_nb$cons.conf.idx, breaks = c(-Inf,-40.5,Inf), labels= c("conf1", "conf2"), right = FALSE)
> data_nb$cons.price.idx <- cut(data_nb$cons.price.idx, breaks = c(-Inf,93.58,Inf), labels= c("cons_price1", "cons_price2"), right = FALSE
> data_nb$euribor3m <- cut(data_nb$euribor3m, breaks = c(-Inf,3.621,Inf), labels= c("low", "high"), right = FALSE)
> data_nb$nr.employed <- cut(data_nb$nr.employed, breaks = c(-Inf,5167, Inf), labels= c("n1", "n2"), right = FALSE)
```

```
> str(data_nb)
'data.frame':   73096 obs. of  20 variables:
 $ age           : Factor w/ 4 levels "young","middle-aged",..: 4 4 2 3 4 3 4 3 1 1 ...
 $ 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 ...
 $ campaign      : Factor w/ 2 levels "c1","c2": 1 1 1 1 1 1 1 1 1 1 ...
 $ pdays         : Factor w/ 2 levels "p1","p2": 2 2 2 2 2 2 2 2 2 2 ...
 $ previous      : Factor w/ 2 levels "prev1","prev2": 1 1 1 1 1 1 1 1 1 1 ...
 $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ emp.var.rate  : Factor w/ 2 levels "emp1","emp2": 2 2 2 2 2 2 2 2 2 2 ...
 $ cons.price.idx: Factor w/ 2 levels "cons_price1",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ cons.conf.idx : Factor w/ 2 levels "conf1","conf2": 2 2 2 2 2 2 2 2 2 2 ...
 $ euribor3m     : Factor w/ 2 levels "low","high": 2 2 2 2 2 2 2 2 2 2 ...
 $ nr.employed   : Factor w/ 2 levels "n1","n2": 2 2 2 2 2 2 2 2 2 2 ...
 $ y             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

We built the Naïve Bayes model here and performed 10-fold cross validation on this to test our model.

```
> x=data_nb[1:19]
> str(x)
'data.frame':   73096 obs. of  19 variables:
 $ age           : Factor w/ 4 levels "young","middle-aged",..: 4 4 2 3 4 3 4 3 1 1 ...
 $ 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 ...
 $ campaign      : Factor w/ 2 levels "c1","c2": 1 1 1 1 1 1 1 1 1 1 ...
 $ pdays         : Factor w/ 2 levels "p1","p2": 2 2 2 2 2 2 2 2 2 2 ...
 $ previous      : Factor w/ 2 levels "prev1","prev2": 1 1 1 1 1 1 1 1 1 1 ...
 $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ emp.var.rate  : Factor w/ 2 levels "emp1","emp2": 2 2 2 2 2 2 2 2 2 2 ...
 $ cons.price.idx: Factor w/ 2 levels "cons_price1",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ cons.conf.idx : Factor w/ 2 levels "conf1","conf2": 2 2 2 2 2 2 2 2 2 2 ...
 $ euribor3m     : Factor w/ 2 levels "low","high": 2 2 2 2 2 2 2 2 2 2 ...
 $ nr.employed   : Factor w/ 2 levels "n1","n2": 2 2 2 2 2 2 2 2 2 2 ...
> y=data_nb$y
> str(y)
 Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> nbmodel = train(x,y,'nb',trControl=trainControl(method='cv',number=10),na.action=na.pass)
There were 50 or more warnings (use warnings() to see the first 50)
> print(nbmodel)
Naive Bayes

73096 samples
   19 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 65786, 65786, 65786, 65788, 65786, 65786, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.7160718  0.4321435
   TRUE      0.7160718  0.4321435

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was
 held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
>
```

Model 3: K- Nearest Neighbor

```
D:/R Workspace/DA_Project/
> data_knn=dummy.data.frame(alldata_n,names = c("job","marital", "education","default",
+                                               "housing","loan","contact","month","day_of_week","poutcome"))
> head(data_knn)
  age jobadmin. jobblue-collar jobentrepreneur jobhousemaid jobmanagement jobretired jobself-employed jobservices
jobstudent jobtechnician
1  56         0              0               0            1             0          0                0           0
         0              0
2  57         0              0               0            0             0          0                0           1
         0              0
3  37         0              0               0            0             0          0                0           1
         0              0
4  40         1              0               0            0             0          0                0           0
         0              0
5  56         0              0               0            0             0          0                0           1
         0              0
6  45         0              0               0            0             0          0                0           1
         0              0
  jobunemployed jobunknown maritaldivorced maritalmarried maritalsingle maritalunknown educationbasic.4y education
basic.6y
1             0          0               0              1             0              0                1
         0
2             0          0               0              1             0              0                0
         0
3             0          0               0              1             0              0                0
         0
4             0          0               0              1             0              0                0
         1
5             0          0               0              1             0              0                0
         0
6             0          0               0              1             0              0                0
         0
  educationbasic.9y educationhigh.school educationilliterate educationprofessional.course educationuniversity.degr
ee educationunknown
1                 0                    0                    0                           0
 0                0
2                 0                    1                    0                           0
 0                0
3                 0                    1                    0                           0
 0                0
4                 0                    0                    0                           0
```

```
> library(dummies)
dummies-1.5.6 provided by Decision Patterns

> data_knn=alldata_n
> data_knn=dummy.data.frame(data_knn,names=c("job","marital","education","default","housing","loan","contact", "month", "day_of_week","poutcome")
> num.vars <- sapply(data_knn,is.numeric)
> data_knn[num.vars] <- lapply(data_knn[num.vars],scale)
> head(data_knn)
         age  jobadmin. jobblue-collar jobentrepreneur jobhousemaid jobmanagement jobretired jobself-employed jobservices jobstudent
1  1.29648896 -0.6099976     -0.4793564      -0.1793503    6.2845731    -0.2745019 -0.2610638       -0.1863422  -0.3032788 -0.1969686
2  1.37945351 -0.6099976     -0.4793564      -0.1793503   -0.1591176    -0.2745019 -0.2610638       -0.1863422   3.2972514 -0.1969686
3 -0.27983735 -0.6099976     -0.4793564      -0.1793503   -0.1591176    -0.2745019 -0.2610638       -0.1863422   3.2972514 -0.1969686
4 -0.03094372  1.6393282     -0.4793564      -0.1793503   -0.1591176    -0.2745019 -0.2610638       -0.1863422  -0.3032788 -0.1969686
5  1.29648896 -0.6099976     -0.4793564      -0.1793503   -0.1591176    -0.2745019 -0.2610638       -0.1863422   3.2972514 -0.1969686
6  0.38387899 -0.6099976     -0.4793564      -0.1793503   -0.1591176    -0.2745019 -0.2610638       -0.1863422   3.2972514 -0.1969686
  jobtechnician jobunemployed  jobunknown maritaldivorced maritalmarried maritalsingle maritalunknown educationbasic.4y
1    -0.4391539    -0.1681094 -0.08935449      -0.3473554      0.8527189    -0.6721246    -0.04654239         3.0362482
2    -0.4391539    -0.1681094 -0.08935449      -0.3473554      0.8527189    -0.6721246    -0.04654239        -0.3293493
3    -0.4391539    -0.1681094 -0.08935449      -0.3473554      0.8527189    -0.6721246    -0.04654239        -0.3293493
4    -0.4391539    -0.1681094 -0.08935449      -0.3473554      0.8527189    -0.6721246    -0.04654239        -0.3293493
5    -0.4391539    -0.1681094 -0.08935449      -0.3473554      0.8527189    -0.6721246    -0.04654239        -0.3293493
6    -0.4391539    -0.1681094 -0.08935449      -0.3473554      0.8527189    -0.6721246    -0.04654239        -0.3293493
```

```
> x=data_knn[c(1:62)]
> y=data_knn$y
```

```
> knnmodel = train(x,y,'knn',trControl=trainControl(method='cv',number=10),tuneGrid=expand.grid(k = 1:20))
> print(knnmodel)
k-Nearest Neighbors

73096 samples
   62 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 65786, 65788, 65786, 65787, 65786, 65786, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9513653  0.9027306
   2  0.9156315  0.8312630
   3  0.8856026  0.7712051
   4  0.8591032  0.7182063
   5  0.8375562  0.6751122
   6  0.8182802  0.6365603
   7  0.8039429  0.6078858
   8  0.7923418  0.5846835
   9  0.7838460  0.5676920
  10  0.7763217  0.5526433
  11  0.7716565  0.5433129
  12  0.7691666  0.5383331
  13  0.7687972  0.5375944
  14  0.7674702  0.5349403
  15  0.7668955  0.5337912
  16  0.7659516  0.5319033
  17  0.7668683  0.5337366
  18  0.7647751  0.5295502
  19  0.7636671  0.5273341
  20  0.7612183  0.5224367

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 1.
>
```

5.2. Evaluations and Results

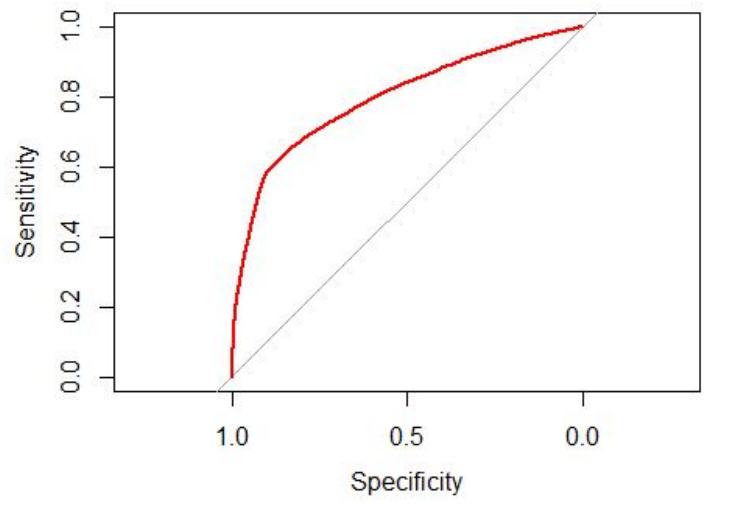
Model 1. Logistic Regression

a) Full Model

```
> pred_fullu = predict(fullu, datasetu[,-63])
Warning message:
In predict.lm(object, newdata, se.fit, scale = 1, type = ifelse(type ==  :
  prediction from a rank-deficient fit may be misleading
```

```
> head(lapply(pred_fullu, as.numeric))
$`1`
[1] -1.610399

$`2`
[1] -1.608782

$`3`
[1] -1.444105

$`4`
[1] -1.212664

$`5`
[1] -1.393244

$`6`
[1] -1.637224
```

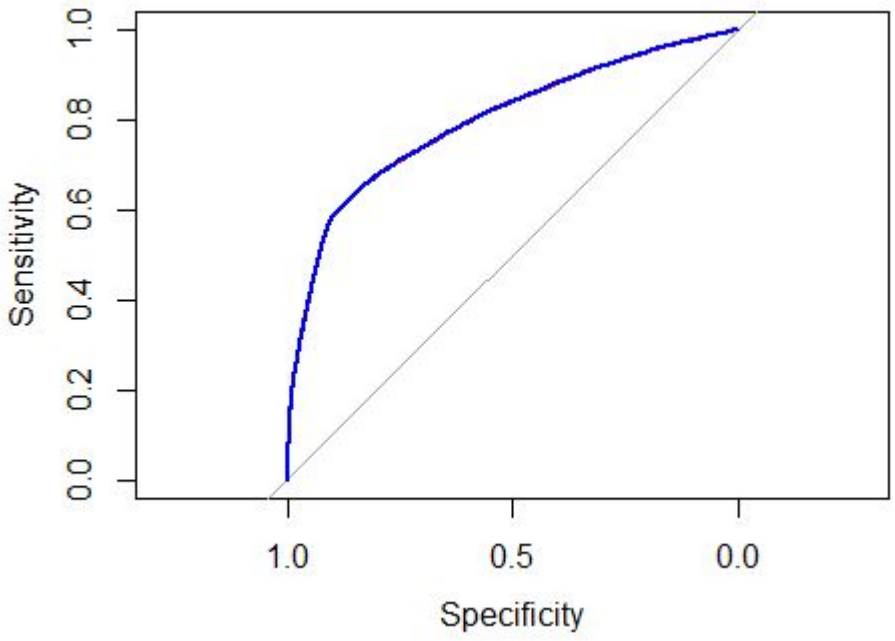
```
> roc_fullu <- roc(Class ~ as.numeric(pred_fullu), data = datasetu)
> plot(roc_fullu, col = "red")
> auc(roc_fullu)
Area under the curve: 0.7956
```



b) Logistic Stepwise Forward Model

```
> pred_forwardu = predict(forwardu, datasetu[,-63])
> head(lapply(pred_forwardu, as.numeric))
$`1`
[1] -1.613461

$`2`
[1] -1.595602

$`3`
[1] -1.446321

$`4`
[1] -1.219471

$`5`
[1] -1.419483

$`6`
[1] -1.627434
```

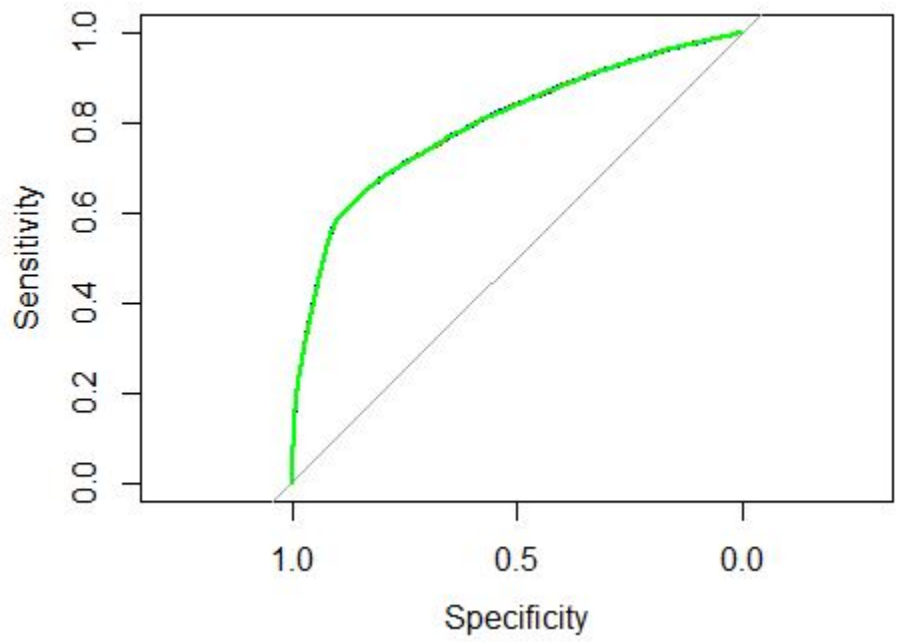
```
> roc_forwardu <- roc(Class ~ as.numeric(pred_forwardu), data = datasetu)
> lines(roc_forwardu, col = "blue")
> auc(roc_forwardu)
Area under the curve: 0.7958
```



c) Stepwise Backward Model

```
> pred_backwardu = predict(backwardu, datasetu[,-63])
> head(lapply(pred_backwardu, as.numeric))
$`1`
[1] -1.605483

$`2`
[1] -1.593344

$`3`
[1] -1.45277

$`4`
[1] -1.220352

$`5`
[1] -1.416757

$`6`
[1] -1.593344
```

```
> roc_backwardu <- roc(Class ~ as.numeric(pred_backwardu), data = datasetu)
> lines(roc_backwardu, col = "green")
> auc(roc_backwardu)
Area under the curve: 0.7957
```



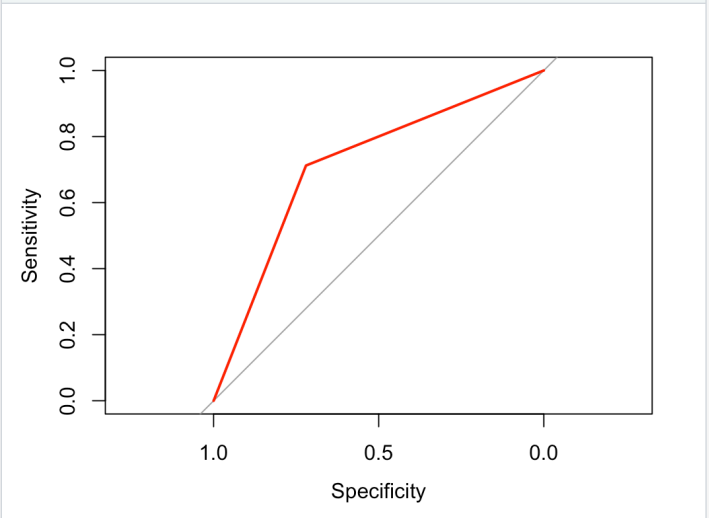
Model 2. Naïve Bayes

Confusion Matrix:

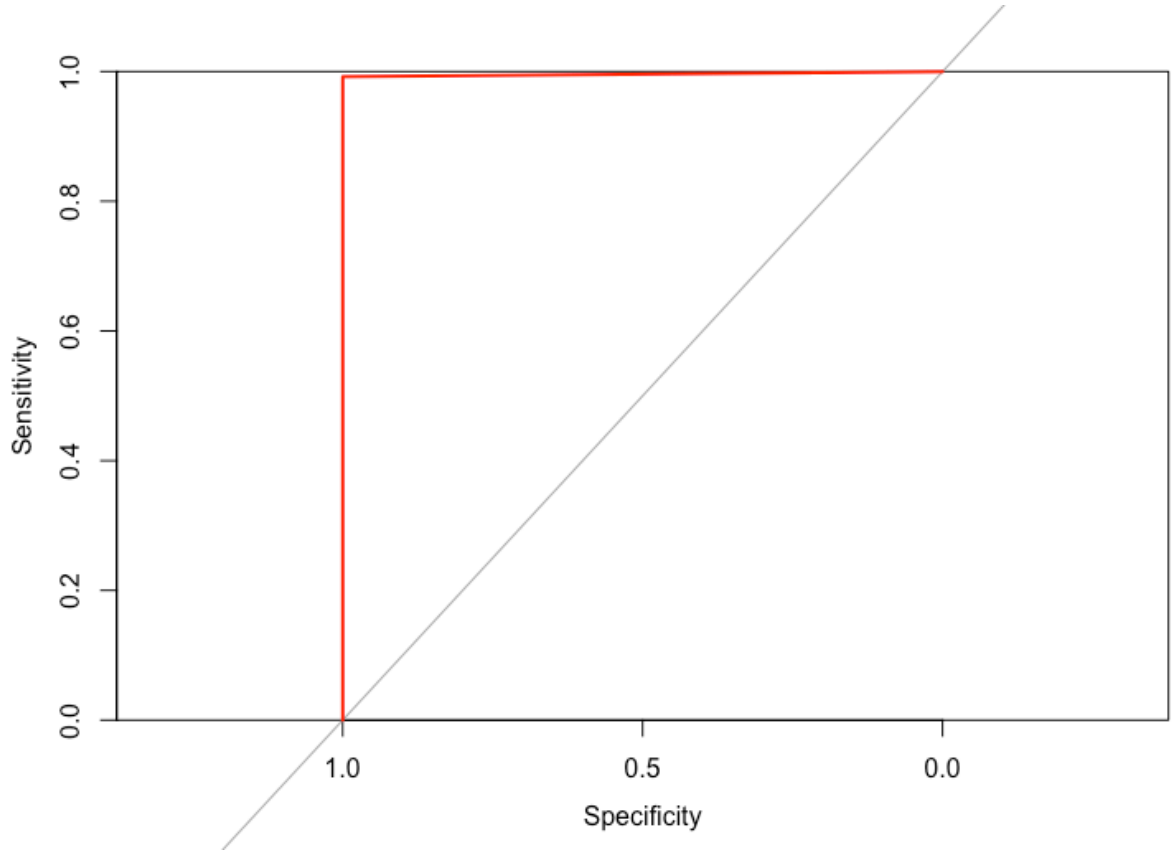
```
> cm_nb <- predict(nbmodel,x)
There were 50 or more warnings (use warnings() to see the first 50)
> tab1 <- table(cm_nb, data_nb$y)
> confusionMatrix(tab1)
Confusion Matrix and Statistics


cm_nb    no   yes
  no  26318 10519
  yes 10230 26029

               Accuracy : 0.7161
                 95% CI : (0.7129, 0.7194)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.4323
 Mcnemar's Test P-Value : 0.04557

            Sensitivity : 0.7201
            Specificity : 0.7122
         Pos Pred Value : 0.7144
         Neg Pred Value : 0.7179
             Prevalence : 0.5000
         Detection Rate : 0.3600
   Detection Prevalence : 0.5040
      Balanced Accuracy : 0.7161

       'Positive' Class : no
```

ROC and AUC:

```
> roc_nb <- roc(y ~ as.numeric(cm_nb), data = data_nb)
> plot(roc_nb, col = "red")
> auc(roc_nb)
Area under the curve: 0.7161
>
```



Model 3. K-Nearest Neighbor



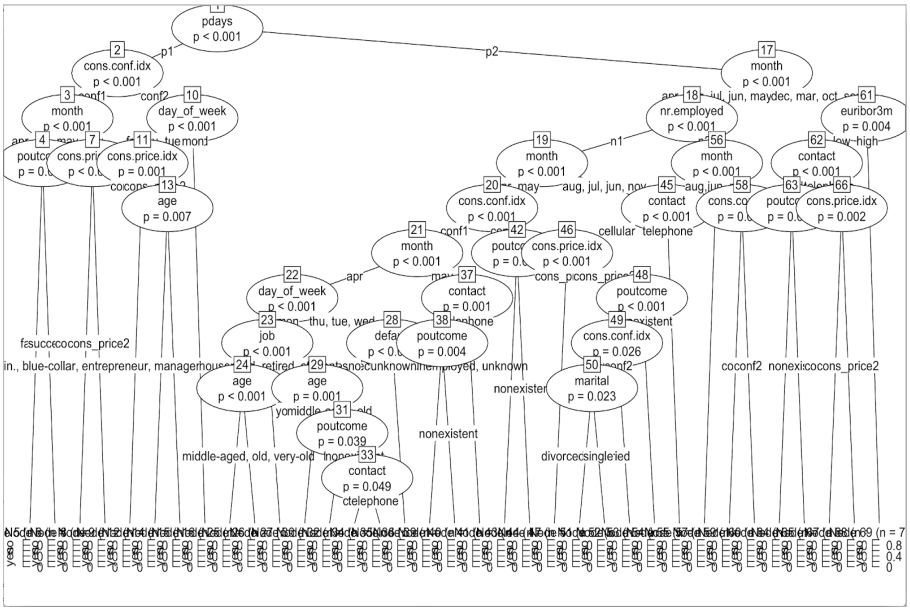

```
roc_knn <- roc(y ~ as.numeric(cm_knn), data = data_knn)

plot(roc_knn, col = "red")
```

Decision Tree:

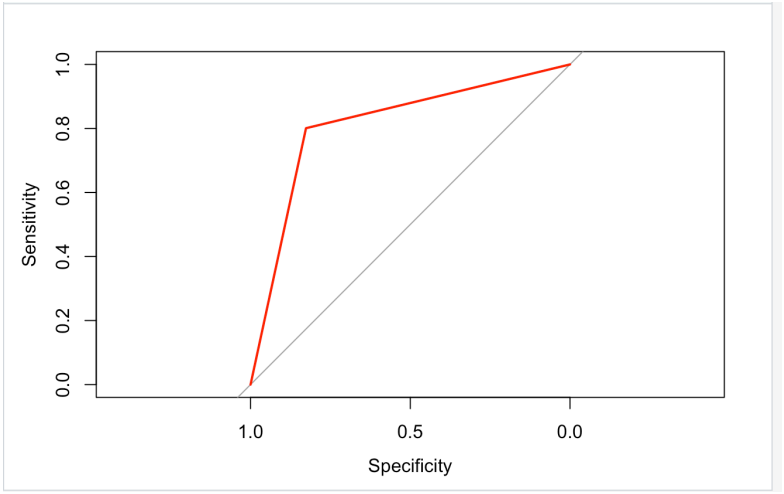
```
> tree_model <- ctree(y ~.,data = data_nb)
> ctrl_rf <- trainControl(method = "cv", number = 10)
> predict_tree <- predict(tree_model, data=data_nb, trControl=ctrl_rf)
> table(predict_tree)
predict_tree
   no   yes
39413  1775
```

```
> confusionMatrix(table(predict_tree, data_nb$y))
Confusion Matrix and Statistics


predict_tree    no   yes
         no  30196  7280
         yes  6352 29268

               Accuracy : 0.8135
                 95% CI : (0.8107, 0.8163)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.627
 Mcnemar's Test P-Value : 2.028e-15

            Sensitivity : 0.8262
            Specificity : 0.8008
         Pos Pred Value : 0.8057
         Neg Pred Value : 0.8217
             Prevalence : 0.5000
         Detection Rate : 0.4131
   Detection Prevalence : 0.5127
      Balanced Accuracy : 0.8135

       'Positive' Class : no
```

```
> print(tree_model)

Model formula:
y ~ age + job + marital + education + default + housing + loan +
    contact + month + day_of_week + campaign + pdays + previous +
    poutcome + emp.var.rate + cons.price.idx + cons.conf.idx +
    euribor3m + nr.employed

Fitted party:
[1] root
|   [2] pdays in p1
|   |   [3] cons.conf.idx in conf1
|   |   |   [4] month in apr, jun, mar, oct
|   |   |   |   [5] poutcome in failure: no (n = 20, err = 30.0%)
|   |   |   |   [6] poutcome in success: yes (n = 205, err = 36.6%)
|   |   |   [7] month in may, nov
|   |   |   |   [8] cons.price.idx in cons_price1: no (n = 199, err = 28.1%)
|   |   |   |   [9] cons.price.idx in cons_price2: yes (n = 44, err = 36.4%)
|   |   [10] cons.conf.idx in conf2
|   |   |   [11] day_of_week in fri, thu, tue, wed
|   |   |   |   [12] cons.price.idx in cons_price1: yes (n = 427, err = 31.1%)
|   |   |   |   [13] cons.price.idx in cons_price2
|   |   |   |   |   [14] age in young, middle-aged, old: yes (n = 291, err = 24.7%)
|   |   |   |   |   [15] age in very-old: yes (n = 137, err = 8.8%)
|   |   |   [16] day_of_week in mon: yes (n = 192, err = 43.2%)
|   [17] pdays in p2
|   |   [18] month in apr, aug, jul, jun, may, nov
|   |   |   [19] nr.employed in n1
|   |   |   |   [20] month in apr, may
|   |   |   |   |   [21] cons.conf.idx in conf1
|   |   |   |   |   |   [22] month in apr
|   |   |   |   |   |   |   [23] day_of_week in fri, mon
|   |   |   |   |   |   |   |   [24] job in admin., blue-collar, entrepreneur, management, self-employed, services
, technician, unemployed, unknown
|   |   |   |   |   |   |   |   |   [25] age in young: no (n = 162, err = 17.3%)
|   |   |   |   |   |   |   |   |   [26] age in middle-aged, old, very-old: no (n = 984, err = 5.7%)
|   |   |   |   |   |   |   |   [27] job in housemaid, retired, student: no (n = 83, err = 27.7%)
|   |   |   |   |   |   |   [28] day_of_week in thu, tue, wed
|   |   |   |   |   |   |   |   [29] default in no
|   |   |   |   |   |   |   |   |   [30] age in young, very-old: no (n = 566, err = 33.4%)
```




```
> roc_tree <- roc(y ~ as.numeric(predict_tree), data = data_nb)
> plot(roc_tree, col='red')
> auc(roc_tree)
Area under the curve: 0.8135
>
```



Random Forest:

```
> bank_rf = data_nb
> ctrl_rf <- trainControl(method = "cv", number = 10)
> model_rf2 <- randomForest(y~., data = data_nb, do.trace = TRUE, importance = TRUE, ntree = 500, mtry = 6, forest = TRUE, trControl = ctrl_rf)
ntree      OOB      1      2
    1:  14.55% 18.32% 10.79%
    2:  14.40% 17.78% 11.02%
    3:  14.06% 17.26% 10.88%
    4:  13.83% 16.89% 10.77%
    5:  13.56% 15.97% 11.15%
    6:  13.19% 15.31% 11.07%
    7:  12.65% 14.63% 10.67%
    8:  12.39% 14.28% 10.50%
    9:  11.89% 13.77% 10.02%
   10:  11.55% 13.69%  9.42%
   11:  11.31% 13.41%  9.21%
   12:  11.03% 13.12%  8.94%
   13:  10.74% 13.08%  8.41%
   14:  10.43% 12.73%  8.13%
   15:  10.19% 12.64%  7.75%
   16:  10.08% 12.53%  7.64%
   17:   9.99% 12.39%  7.58%
   18:   9.85% 12.21%  7.49%
   19:   9.70% 12.07%  7.34%
   20:   9.61% 12.08%  7.14%
   21:   9.44% 12.02%  6.85%
   22:   9.44% 11.96%  6.91%
   23:   9.30% 11.83%  6.77%
```

```
> tab3 <- table(predict_rf2, data_nb$y)
> confusionMatrix(tab3)
Confusion Matrix and Statistics


predict_rf2    no   yes
        no  32669  1786
        yes  3879 34762

               Accuracy : 0.9225
                 95% CI : (0.9205, 0.9244)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.845
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8939
            Specificity : 0.9511
         Pos Pred Value : 0.9482
         Neg Pred Value : 0.8996
             Prevalence : 0.5000
         Detection Rate : 0.4469
   Detection Prevalence : 0.4714
      Balanced Accuracy : 0.9225

       'Positive' Class : no
```

```
> predict_rf2 <- predict(model_rf2, data = data_nb, na.action = na.pass)
> roc_rf2 <- roc(y ~ as.numeric(predict_rf2), data = data_nb)
> plot(roc_rf2, col = "red")
> auc(roc_rf2) # Area under curve
Area under the curve: 0.9225
>
```