Subscription_prediction.R

harshitmehta

2020-04-02

```
# loading all the libraries

library(dplyr)

##
## 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(ggplot2)
library(lattice)
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 3.0-2

library(ROSE)

## Loaded ROSE 0.0-3

bank=read.table("bank.csv",sep=";",header=TRUE)

head(bank)

##   age         job marital education default balance housing loan  contact
day
## 1  30  unemployed married   primary      no    1787      no   no cellular
19
## 2  33    services married secondary      no    4789     yes  yes cellular
11
## 3  35  management  single  tertiary      no    1350     yes   no cellular
16
## 4  30  management married  tertiary      no    1476     yes  yes  unknown
3
## 5  59 blue-collar married secondary      no       0     yes   no  unknown
5
## 6  35  management  single  tertiary      no     747      no   no cellular
```

```
23
##   month duration campaign pdays previous poutcome  y
## 1   oct       79        1    -1        0  unknown no
## 2   may      220        1   339        4  failure no
## 3   apr      185        1   330        1  failure no
## 4   jun      199        4    -1        0  unknown no
## 5   may      226        1    -1        0  unknown no
## 6   feb      141        2   176        3  failure no

######################### Data Cleaning & Preparation
##############################

# to check if there are any missing values
any(is.na(bank))

## [1] FALSE

# Thus we have no missing values in the data set.

colnames(bank)

##  [1] "age"       "job"       "marital"   "education" "default"   "balance"
##  [7] "housing"   "loan"      "contact"   "day"       "month"
"duration"
## [13] "campaign"  "pdays"     "previous"  "poutcome"  "y"

glimpse(bank)

## Observations: 4,521
## Variables: 17
## $ age       <int> 30, 33, 35, 30, 59, 35, 36, 39, 41, 43, 39, 43, 36, 20,
31,…
## $ job       <fct> unemployed, services, management, management, blue-
collar, …
## $ marital   <fct> married, married, single, married, married, single,
married…
## $ education <fct> primary, secondary, tertiary, tertiary, secondary,
tertiary…
## $ default   <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no,
no,…
## $ balance   <int> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88,
9374, 2…
## $ housing   <fct> no, yes, yes, yes, yes, no, yes, yes, yes, yes, yes,
yes, n…
## $ loan      <fct> no, yes, no, yes, no, no, no, no, no, yes, no, no, no,
no, …
## $ contact   <fct> cellular, cellular, cellular, unknown, unknown,
cellular, c…
## $ day       <int> 19, 11, 16, 3, 5, 23, 14, 6, 14, 17, 20, 17, 13, 30, 29,
29…
## $ month     <fct> oct, may, apr, jun, may, feb, may, may, may, apr, may,
```

```
apr,…
## $ duration  <int> 79, 220, 185, 199, 226, 141, 341, 151, 57, 313, 273,
113, 3…
## $ campaign  <int> 1, 1, 1, 4, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 5, 1, 1,
1,…
## $ pdays     <int> -1, 339, 330, -1, -1, 176, 330, -1, -1, 147, -1, -1, -1,
-1…
## $ previous  <int> 0, 4, 1, 0, 0, 3, 2, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 2, 0,
1,…
## $ poutcome  <fct> unknown, failure, failure, unknown, unknown, failure,
other…
## $ y         <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, yes,
no…

summary(bank)

##       age                 job          marital         education
default
##  Min.   :19.00   management :969   divorced: 528   primary  : 678   no
:4445
##  1st Qu.:33.00   blue-collar:946   married :2797   secondary:2306   yes:
76
##  Median :39.00   technician :768   single  :1196   tertiary :1350
##  Mean   :41.17   admin.     :478                   unknown  : 187
##  3rd Qu.:49.00   services   :417
##  Max.   :87.00   retired    :230
##                  (Other)    :713
##     balance      housing     loan           contact          day
##  Min.   :-3313   no :1962   no :3830   cellular :2896   Min.   : 1.00
##  1st Qu.:   69   yes:2559   yes: 691   telephone: 301   1st Qu.: 9.00
##  Median :  444                         unknown  :1324   Median :16.00
##  Mean   : 1423                                          Mean   :15.92
##  3rd Qu.: 1480                                          3rd Qu.:21.00
##  Max.   :71188                                          Max.   :31.00
##
##      month         duration       campaign          pdays
##  may    :1398   Min.   :   4   Min.   : 1.000   Min.   : -1.00
##  jul    : 706   1st Qu.: 104   1st Qu.: 1.000   1st Qu.: -1.00
##  aug    : 633   Median : 185   Median : 2.000   Median : -1.00
##  jun    : 531   Mean   : 264   Mean   : 2.794   Mean   : 39.77
##  nov    : 389   3rd Qu.: 329   3rd Qu.: 3.000   3rd Qu.: -1.00
##  apr    : 293   Max.   :3025   Max.   :50.000   Max.   :871.00
##  (Other): 571
##     previous          poutcome      y
##  Min.   : 0.0000   failure: 490   no :4000
##  1st Qu.: 0.0000   other  : 197   yes: 521
##  Median : 0.0000   success: 129
##  Mean   : 0.5426   unknown:3705
##  3rd Qu.: 0.0000
```

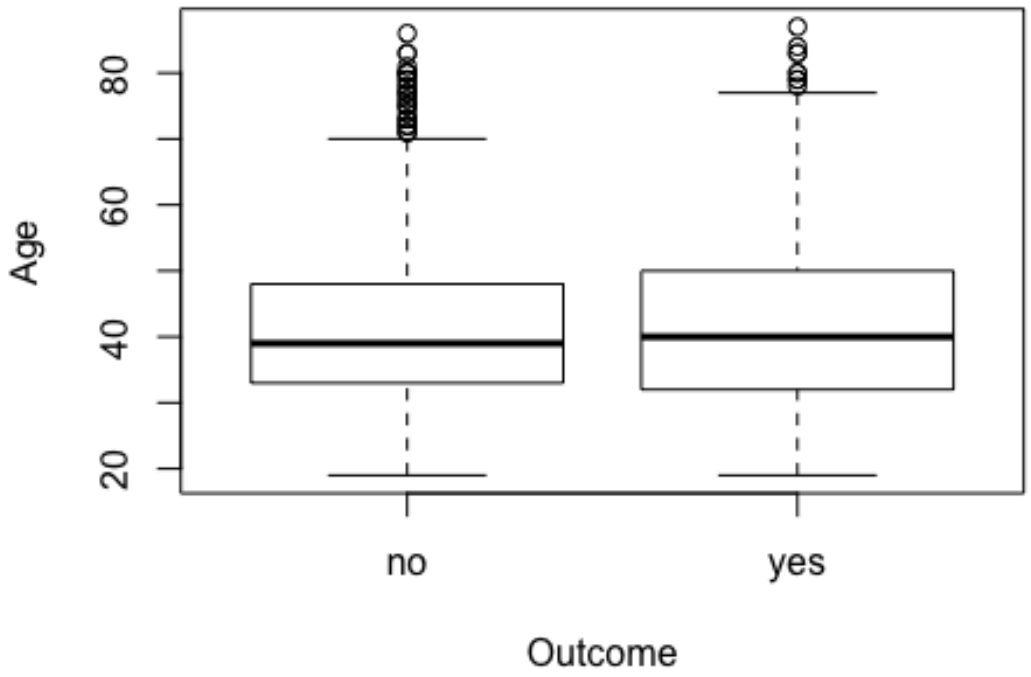
```
##  Max.   :25.0000
##
```

```
# The following variables need to be removed from the dataset as they are not
useful
# for analysis purpose :
# pdays : 75% of the values are -1 (not previously contacted)
# previous : 75% of the values are 0 (75% of clients never contacted before)
# poutcome : 87% of the observations fall in unknown/other category
# durartion : can be known only after making the call - not useful for
prediction purposes

dropped_cols = c("pdays", "previous", "poutcome", "duration")
bank_df = bank[,! (names(bank) %in% dropped_cols)]

summary(bank_df)
```

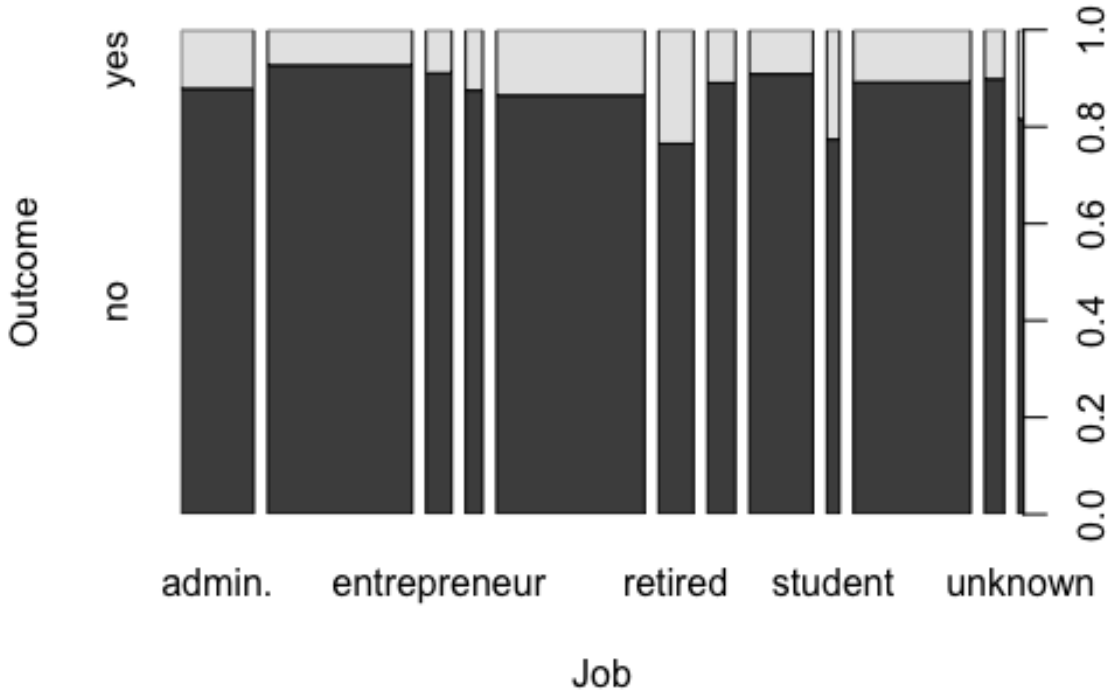
```
##       age                 job          marital         education
default
##  Min.   :19.00   management :969   divorced: 528   primary  : 678   no
:4445
##  1st Qu.:33.00   blue-collar:946   married :2797   secondary:2306   yes:
76
##  Median :39.00   technician :768   single  :1196   tertiary :1350
##  Mean   :41.17   admin.     :478                   unknown  : 187
##  3rd Qu.:49.00   services   :417
##  Max.   :87.00   retired    :230
##                  (Other)    :713
##     balance      housing     loan           contact          day
##  Min.   :-3313   no :1962   no :3830   cellular :2896   Min.   : 1.00
##  1st Qu.:   69   yes:2559   yes: 691   telephone: 301   1st Qu.: 9.00
##  Median :  444                         unknown  :1324   Median :16.00
##  Mean   : 1423                                          Mean   :15.92
##  3rd Qu.: 1480                                          3rd Qu.:21.00
##  Max.   :71188                                          Max.   :31.00
##
##      month         campaign        y
##  may    :1398   Min.   : 1.000   no :4000
##  jul    : 706   1st Qu.: 1.000   yes: 521
##  aug    : 633   Median : 2.000
##  jun    : 531   Mean   : 2.794
##  nov    : 389   3rd Qu.: 3.000
##  apr    : 293   Max.   :50.000
##  (Other): 571
```

```
########################################## EDA
####################################
```

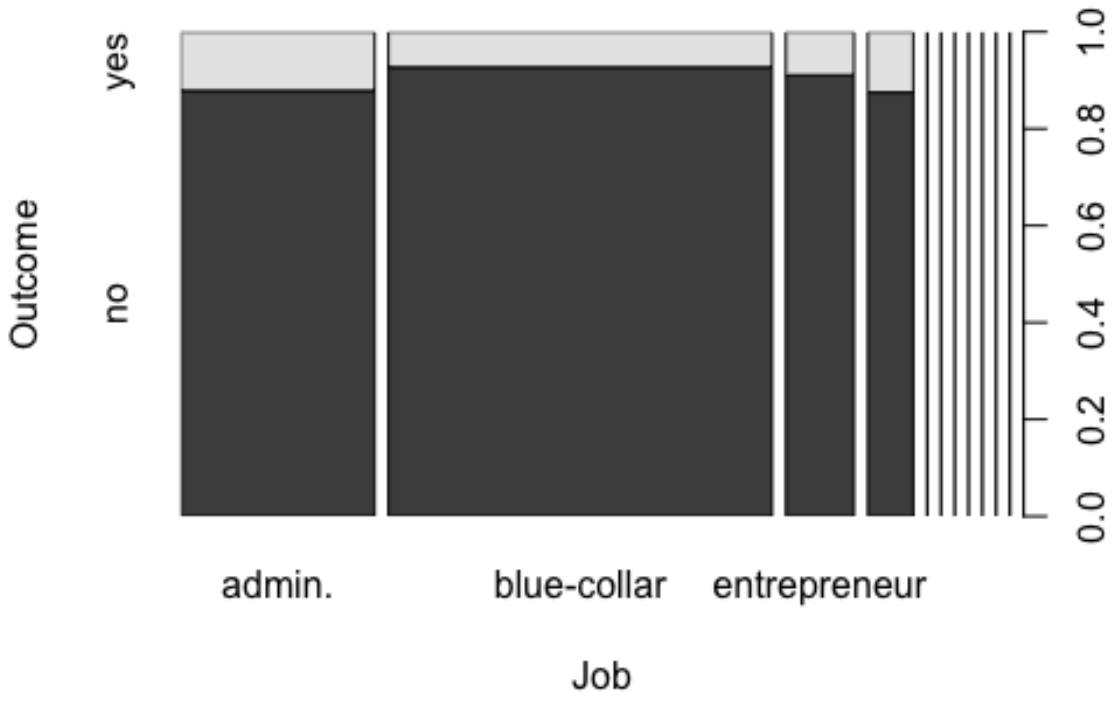
```
plot(bank_df$y, bank_df$age, xlab = "Outcome", ylab = "Age")
```



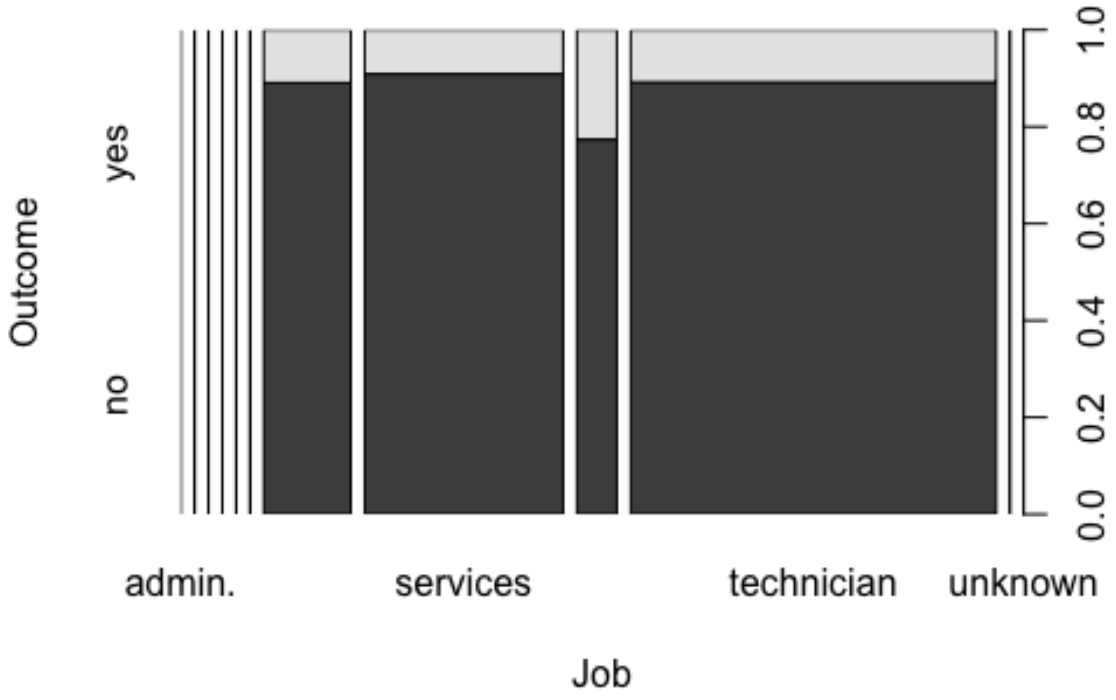

```r
#mosaicplot(job~y, data = bank_df, xlab = "Job", ylab = "Outcome")
spineplot(y~job, data = bank_df, xlab = "Job", ylab = "Outcome")
```



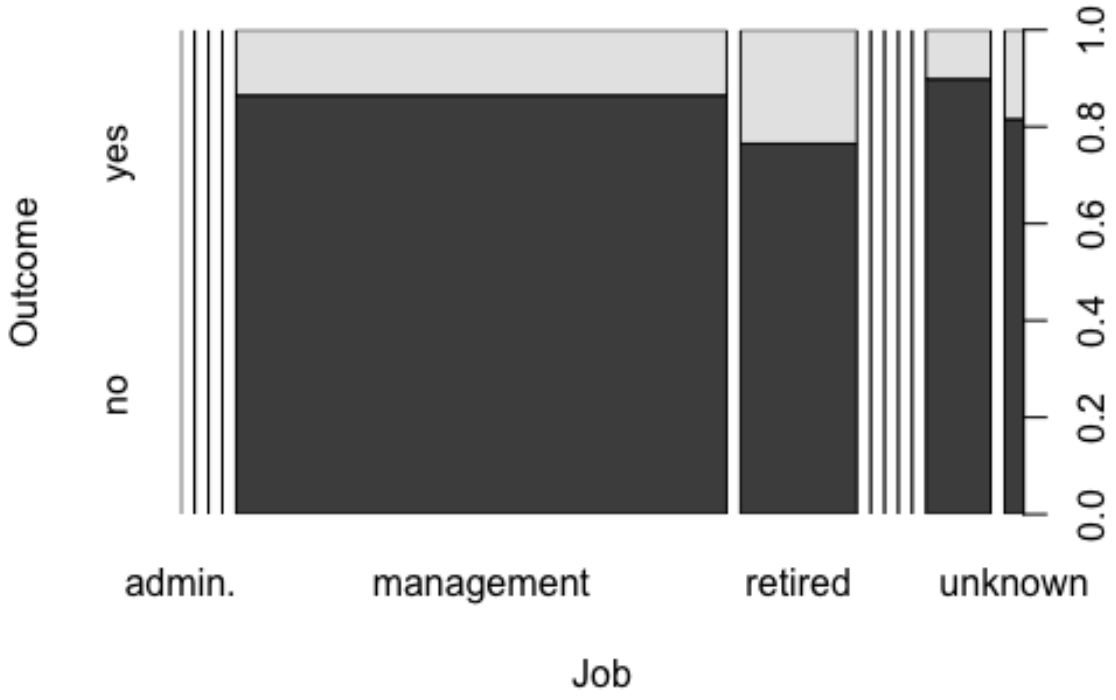

```
job_set1 = c("admin.","blue-collar","entrepreneur","housemaid")
job_set2 = c("self-employed","services","student","technician")
job_set3 = c("management","retired","unemployed","unknown")
spineplot(y~job, data = bank_df[(bank_df$job %in% job_set1),], xlab = "Job",
ylab = "Outcome")
```



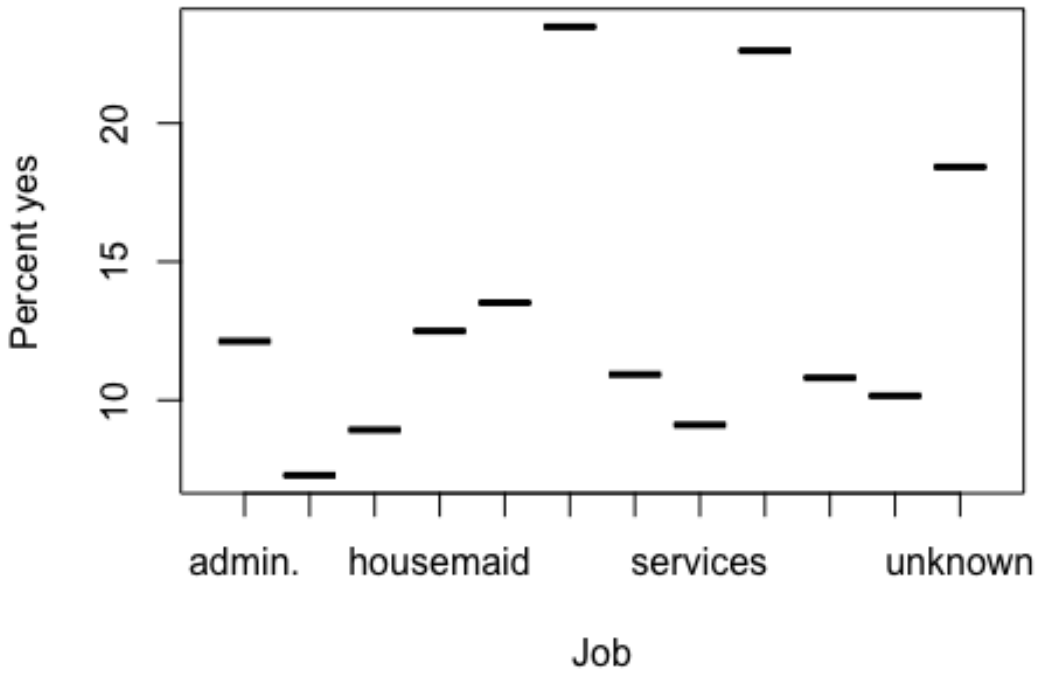

```r
spineplot(y~job, data = bank_df[(bank_df$job %in% job_set2),], xlab = "Job",
ylab = "Outcome")
```




```r
spineplot(y~job, data = bank_df[(bank_df$job %in% job_set3),], xlab = "Job",
ylab = "Outcome")
```



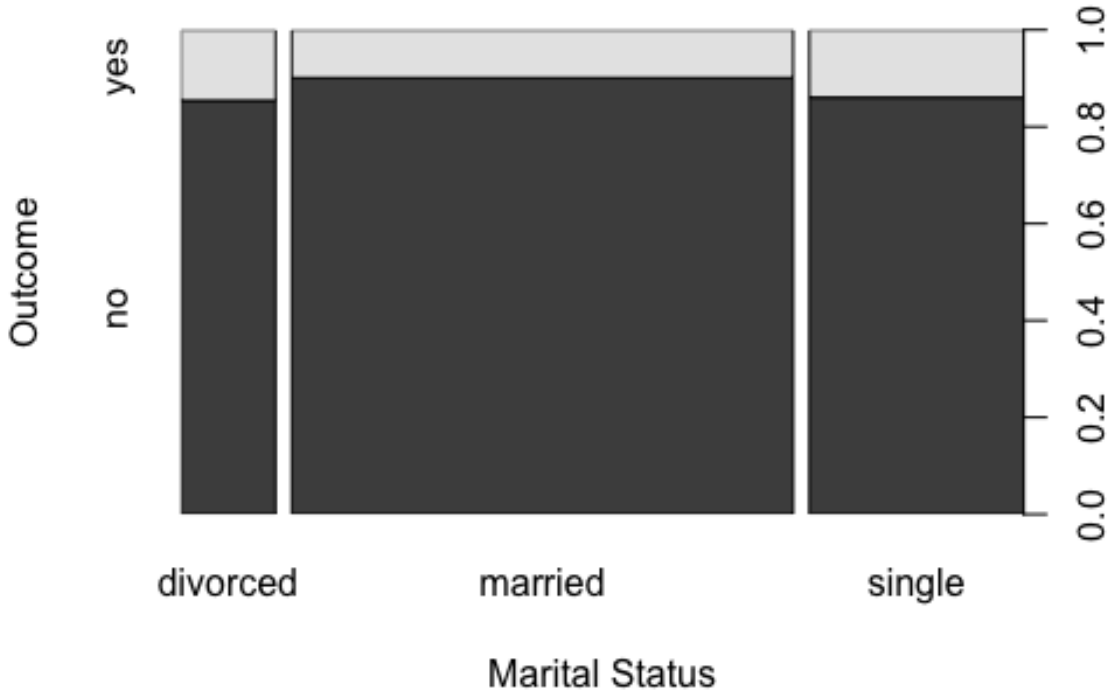

```
job_effect = table(bank_df$job,bank_df$y)
job_frame = as.data.frame(job_effect)
t1 = job_frame[1:12,]
t2 = job_frame[13:24,]
job_frame = merge(t1, t2, by="Var1")
names(job_frame)[names(job_frame) == "Var1"] <- "Job"
names(job_frame)[names(job_frame) == "Freq.x"] <- "no"
names(job_frame)[names(job_frame) == "Freq.y"] <- "yes"
job_frame = job_frame[,c(1,3,5)]
job_frame$percent_yes =
round(job_frame$yes/(job_frame$yes+job_frame$no),4)*100
plot(job_frame$Job, job_frame$percent_yes, xlab = "Job", ylab = "Percent
yes")
```



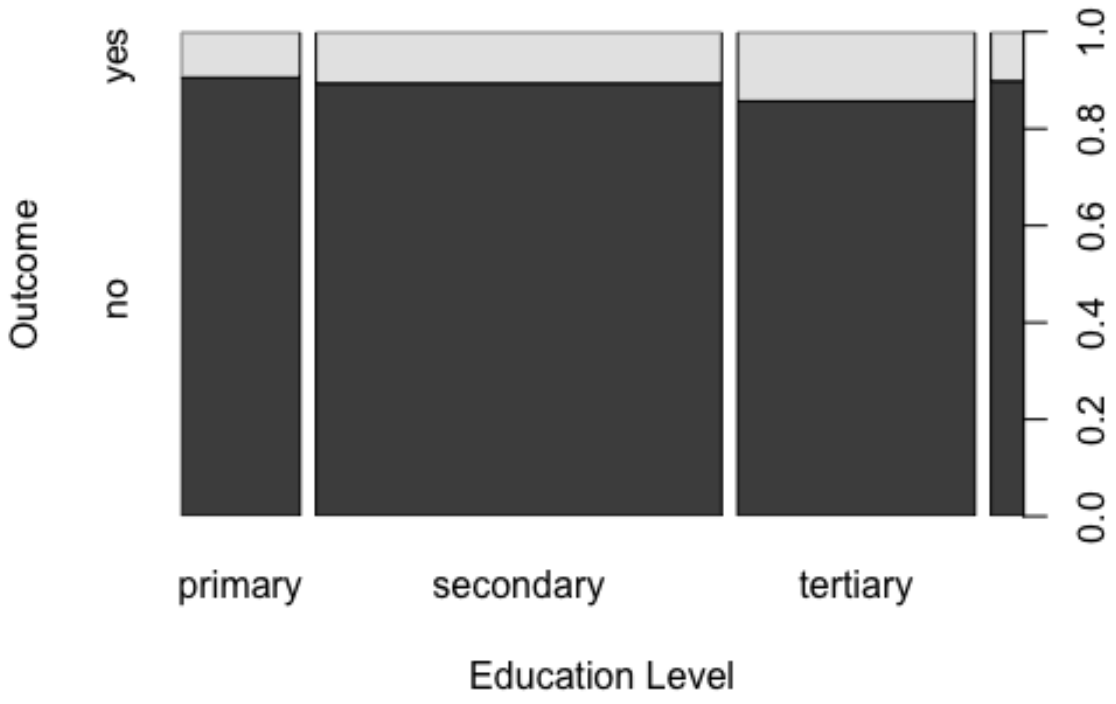

```
# Students and retired people are more likely to subscribe to the bank
product.
# Blue-collar proffessionals least likely to subscribe

# martital status effect on subscription
table(bank_df$marital,bank_df$y)
```

```
##
##              no  yes
##   divorced  451   77
##   married  2520  277
##   single   1029  167
```

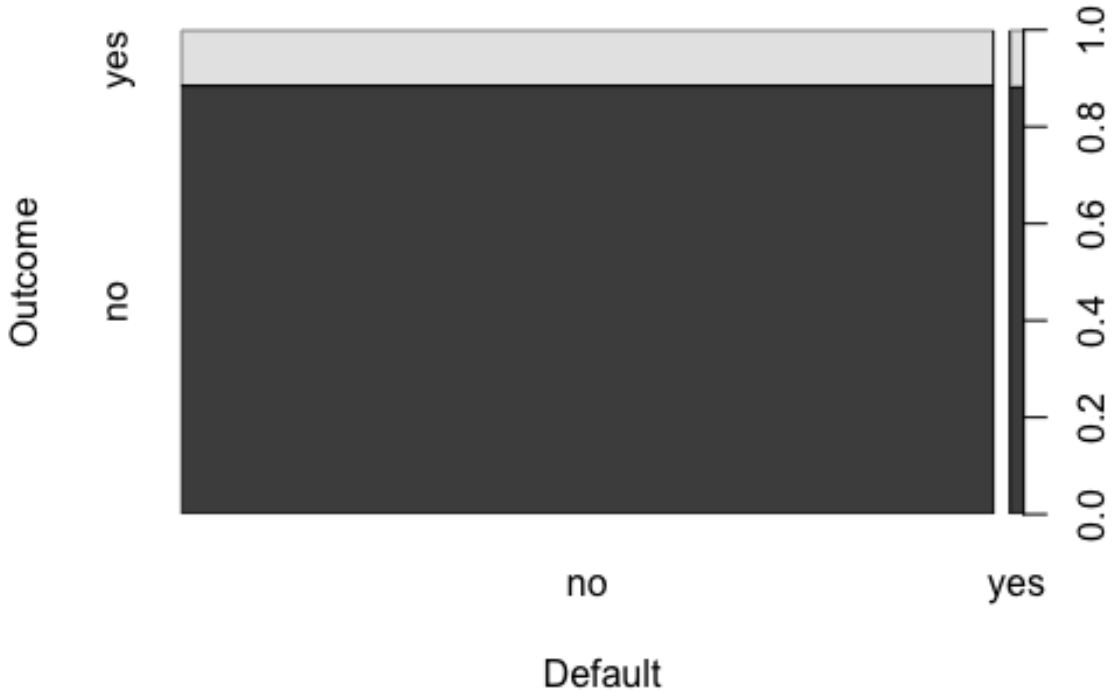
```
spineplot(y~marital, data = bank_df, xlab = "Marital Status", ylab =
"Outcome")
```



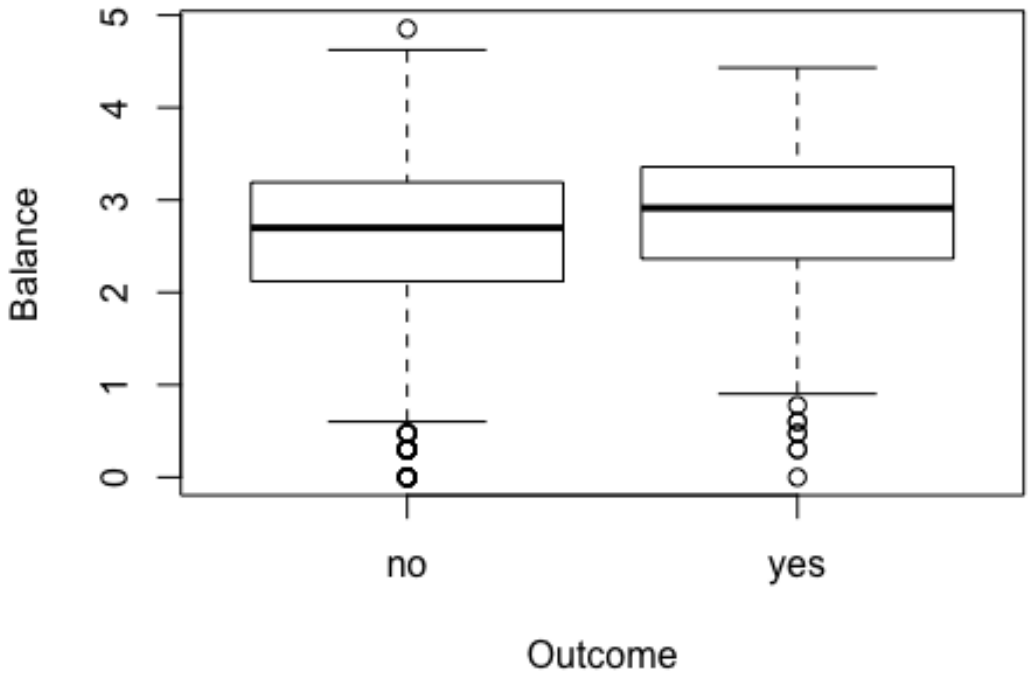

```
# married people less likely to invest in long-term deposits

# education effect on subscription
table(bank_df$education,bank_df$y)

##
##                no  yes
##   primary     614   64
##   secondary  2061  245
##   tertiary   1157  193
##   unknown     168   19

spineplot(y~education, data = bank_df, xlab = "Education Level", ylab =
"Outcome")
```




```
# Customers with tertiary level of education most likely to invest in long-
term deposits

# default effect on subscription
table(bank_df$default,bank_df$y)

##
##          no  yes
##    no  3933  512
##    yes   67    9

spineplot(y~default, data = bank_df, xlab = "Default", ylab = "Outcome")
```




```
# default does not look like an important variable

# bank balance effect on subscription
plot(bank_df$y, log10(bank_df$balance), xlab = "Outcome", ylab = "Balance")

## Warning in is.factor(y): NaNs produced

## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
z$out[z$group
## == : Outlier (-Inf) in boxplot 1 is not drawn

## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
z$out[z$group
## == : Outlier (-Inf) in boxplot 2 is not drawn
```




```
summary(bank_df[bank_df$y=="yes",]$balance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1206     171     710    1572    2160   26965
```

```
summary(bank_df[bank_df$y=="no",]$balance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3313.0    61.0   419.5  1403.2  1407.0 71188.0
```

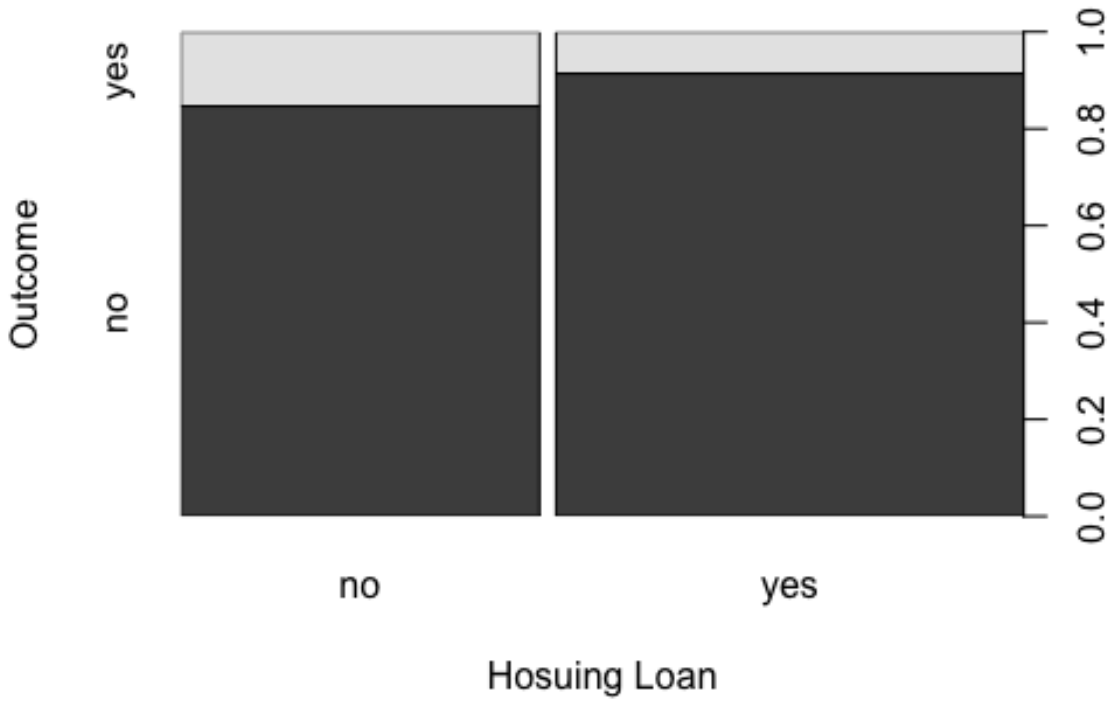
```
# people subscribing to long term generally have more bank balance

# housing loan effect on subscription
summary(bank_df$housing)
```

```
##   no  yes
## 1962 2559
```

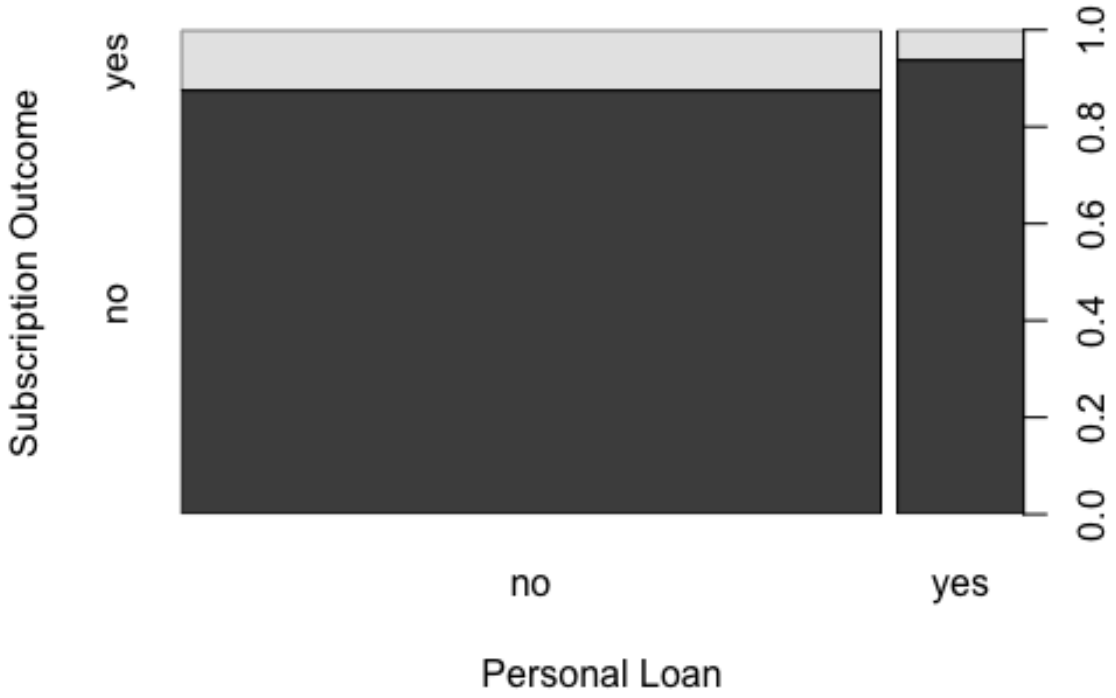
```
table(bank_df$housing,bank_df$y)
```

```
##
##         no  yes
##   no  1661  301
##   yes 2339  220
```

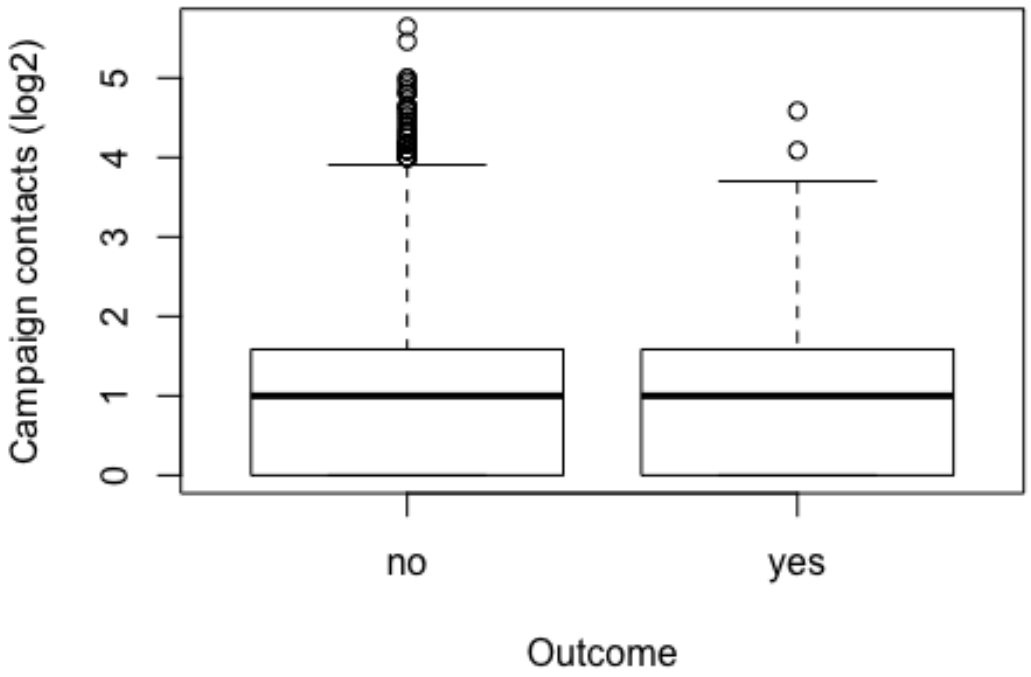
```
spineplot(y~housing, data = bank_df, xlab = "Hosuing Loan", ylab = "Outcome")
```




```
# people who do not have housing loan are more likely to invest in long-term
loans

# personal loan effect on subscription
summary(bank_df$loan)

##   no  yes
## 3830  691

table(bank_df$loan,bank_df$y)

##
##         no  yes
##   no  3352  478
##   yes  648   43

spineplot(y~loan, data = bank_df, xlab = "Personal Loan", ylab =
"Subscription Outcome")
```




```
# people who do not have personal loan are more likely to invest in long-term
loans

# number of contacts persormed's effect on subscription
summary(bank_df$campaign)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.794   3.000  50.000

plot(bank_df$y, log2(bank_df$campaign), xlab = "Outcome", ylab = "Campaign
contacts (log2)")
```




```r
summary(bank_df[bank_df$y=="yes",]$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.267   3.000  24.000
```

```r
summary(bank_df[bank_df$y=="no",]$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.862   3.000  50.000
```

```r
# 50% of the customers who subscribed did so after 2 calls.


#################################### Test/Train Split
##############################

set.seed(1)
train = sample(1:nrow(bank_df), 3164)

######################################### Modelling
#################################
```

```r
# Logistic Regression 1
glm.fit <- glm(y~., data = bank_df, subset = train, family = binomial)
glm.probs = predict(glm.fit, newdata = bank_df[-train,], type="response")
glm.pred = ifelse(glm.probs>0.5, "yes","no")
actual = bank_df[-train,]$y
# predicting almost all as "no"
mean(glm.pred==actual)
```

```
## [1] 0.8747237
```

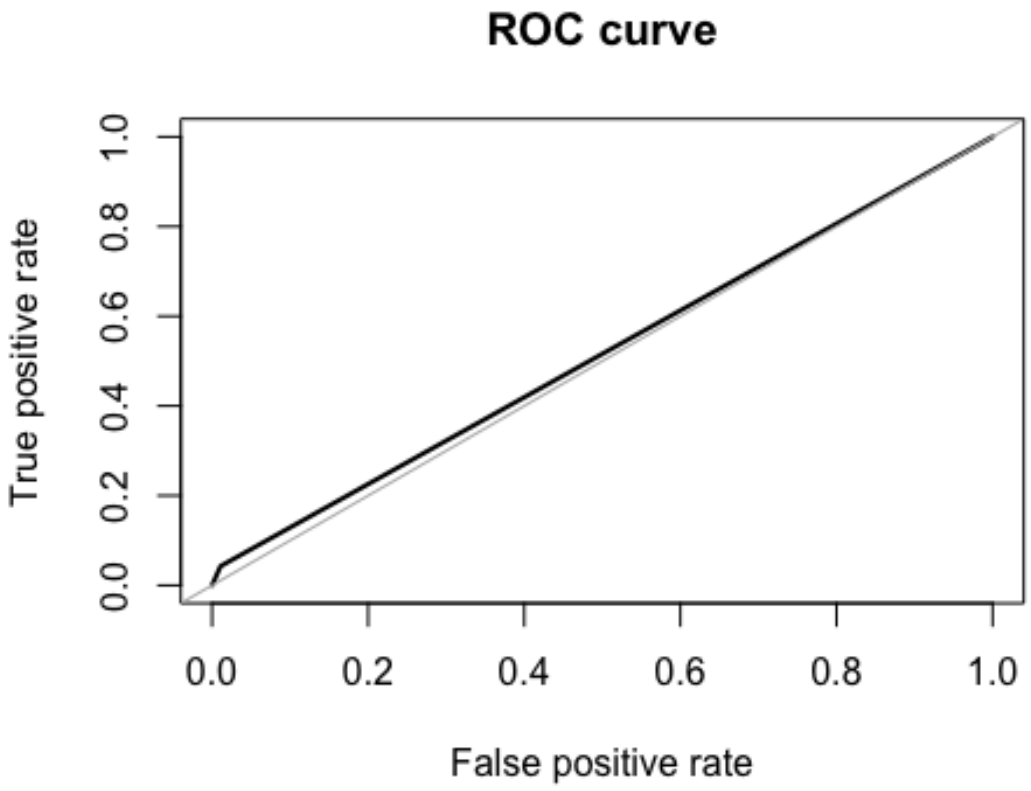
```r
# 87% accuracy - but it is basically labelling everything as "no"
confusion_matrix1 <- table(glm.pred, actual)
confusion_matrix1
```

```
##         actual
## glm.pred   no  yes
##      no  1180  157
##      yes   13    7
```

```r
cat("Accuracy of Logistic Regression : ",((confusion_matrix1[1,"no"] +
confusion_matrix1[2,"yes"])/1357),"\n" )
```

```
## Accuracy of Logistic Regression :  0.8747237
```

```r
roc.curve(bank_df[-train,]$y, glm.pred, plotit = TRUE)
```




```
## Area under the curve (AUC): 0.516

# Classification Trees

library(rpart)
library(rpart.plot)
tree_fit1 <- rpart(y~., method = "class", data = bank_df, subset = train,
control = rpart.control(maxdepth = 20, cp=0.0018727))
summary(tree_fit1)

## Call:
## rpart(formula = y ~ ., data = bank_df, subset = train, method = "class",
##     control = rpart.control(maxdepth = 20, cp = 0.0018727))
##   n= 3164
##
##            CP nsplit rel error   xerror       xstd
## 1 0.014005602      0 1.0000000 1.000000 0.04985042
## 2 0.009803922      5 0.9299720 1.000000 0.04985042
## 3 0.008403361      7 0.9103641 1.000000 0.04985042
## 4 0.007002801     10 0.8851541 1.000000 0.04985042
## 5 0.005602241     19 0.8207283 1.000000 0.04985042
## 6 0.003361345     21 0.8095238 1.042017 0.05075078
## 7 0.002801120     26 0.7927171 1.064426 0.05122005
```

```
## 8 0.002334267     27 0.7899160 1.072829 0.05139412
## 9 0.001872700     34 0.7731092 1.081232 0.05156717
##
## Variable importance
##     month       age   balance       day       job education  campaign
contact
##        31        18        11        11         9         5         5
5
##   marital      loan   housing
##         3         1         1
##
## Node number 1: 3164 observations,    complexity param=0.0140056
##   predicted class=no   expected loss=0.1128319  P(node) =1
##     class counts:  2807   357
##    probabilities: 0.887 0.113
##   left son=2 (3015 obs) right son=3 (149 obs)
##   Primary splits:
##       month   splits as  LLRLLLLRLLRR, improve=32.709330, (0 missing)
##       age     < 60.5    to the left,   improve=15.312530, (0 missing)
##       contact splits as  RRL,          improve=11.147380, (0 missing)
##       job     splits as  LLLLLRLLRLLR, improve=10.303440, (0 missing)
##       housing splits as  RL,           improve= 6.995371, (0 missing)
##
## Node number 2: 3015 observations,    complexity param=0.008403361
##   predicted class=no   expected loss=0.09684909  P(node) =0.9529077
##     class counts:  2723   292
##    probabilities: 0.903 0.097
##   left son=4 (2944 obs) right son=5 (71 obs)
##   Primary splits:
##       age     < 60.5    to the left,   improve=14.120080, (0 missing)
##       job     splits as  LLLLLRLLRLLL, improve= 7.756695, (0 missing)
##       contact splits as  RRL,          improve= 7.062799, (0 missing)
##       month   splits as  RR-RLLR-LL--, improve= 6.675861, (0 missing)
##       housing splits as  RL,           improve= 4.490232, (0 missing)
##
## Node number 3: 149 observations,    complexity param=0.0140056
##   predicted class=no   expected loss=0.4362416  P(node) =0.04709229
##     class counts:    84    65
##    probabilities: 0.564 0.436
##   left son=6 (133 obs) right son=7 (16 obs)
##   Primary splits:
##       marital splits as  RLL,          improve=5.075245, (0 missing)
##       day     < 16.5    to the left,   improve=2.959400, (0 missing)
##       age     < 54.5    to the right,  improve=2.331499, (0 missing)
##       job     splits as  RLLLRLRRLRRR, improve=2.070409, (0 missing)
##       month   splits as  --R----L--RL, improve=1.954535, (0 missing)
##   Surrogate splits:
##       job splits as  LLLLLLLRLLLL, agree=0.899, adj=0.062, (0 split)
##
## Node number 4: 2944 observations,    complexity param=0.007002801
```

```
##   predicted class=no   expected loss=0.08933424  P(node) =0.9304678
##     class counts:  2681   263
##    probabilities: 0.911 0.089
##   left son=8 (1101 obs) right son=9 (1843 obs)
##   Primary splits:
##       contact splits as  RLL,          improve=5.709255, (0 missing)
##       month   splits as  RR-RLLR-LL--, improve=5.583282, (0 missing)
##       marital splits as  RLR,          improve=2.615451, (0 missing)
##       housing splits as  RL,           improve=2.352601, (0 missing)
##       loan    splits as  RL,           improve=2.245519, (0 missing)
##   Surrogate splits:
##       month     splits as  RR-RRRL-LR--, agree=0.797, adj=0.458, (0 split)
##       day       < 3.5     to the left,   agree=0.632, adj=0.016, (0 split)
##       education splits as  RRRL,         agree=0.627, adj=0.004, (0 split)
##       campaign  < 24.5    to the right,  agree=0.627, adj=0.003, (0 split)
##       balance   < -1356.5 to the left,   agree=0.627, adj=0.002, (0 split)
##
## Node number 5: 71 observations,    complexity param=0.008403361
##   predicted class=no   expected loss=0.4084507  P(node) =0.02243995
##     class counts:    42    29
##    probabilities: 0.592 0.408
##   left son=10 (51 obs) right son=11 (20 obs)
##   Primary splits:
##       month   splits as  LL-RRLR-LL--, improve=3.249075, (0 missing)
##       balance < 1480.5  to the left,   improve=1.403958, (0 missing)
##       marital splits as  RLL,          improve=1.352283, (0 missing)
##       age     < 65.5    to the right,  improve=1.115742, (0 missing)
##       job     splits as  RLRRLRR----R, improve=1.011498, (0 missing)
##   Surrogate splits:
##       day < 22.5    to the left,   agree=0.775, adj=0.20, (0 split)
##       job splits as  LLLLLLR----L, agree=0.732, adj=0.05, (0 split)
##
## Node number 6: 133 observations,    complexity param=0.0140056
##   predicted class=no   expected loss=0.3909774  P(node) =0.0420354
##     class counts:    81    52
##    probabilities: 0.609 0.391
##   left son=12 (8 obs) right son=13 (125 obs)
##   Primary splits:
##       campaign < 3.5     to the right,  improve=2.602346, (0 missing)
##       age      < 54.5    to the right,  improve=2.505693, (0 missing)
##       job      splits as  RLLLRLLLLRRR, improve=2.244133, (0 missing)
##       day      < 16.5    to the left,   improve=1.628638, (0 missing)
##       balance  < 475.5   to the left,   improve=1.326298, (0 missing)
##
## Node number 7: 16 observations
##   predicted class=yes  expected loss=0.1875  P(node) =0.00505689
##     class counts:     3    13
##    probabilities: 0.188 0.812
##
## Node number 8: 1101 observations
```

```
##   predicted class=no   expected loss=0.04904632  P(node) =0.3479772
##     class counts:  1047    54
##    probabilities: 0.951 0.049
##
## Node number 9: 1843 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.1134021  P(node) =0.5824905
##     class counts:  1634   209
##    probabilities: 0.887 0.113
##   left son=18 (1797 obs) right son=19 (46 obs)
##   Primary splits:
##       month   splits as  LL-LLLR-LL--, improve=9.745503, (0 missing)
##       balance < 1923.5  to the left,   improve=3.232182, (0 missing)
##       loan    splits as  RL,           improve=2.766993, (0 missing)
##       marital splits as  RLR,          improve=2.292004, (0 missing)
##       day     < 1.5     to the right,  improve=1.982283, (0 missing)
##   Surrogate splits:
##       day < 1.5     to the right, agree=0.98, adj=0.196, (0 split)
##
## Node number 10: 51 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.3137255  P(node) =0.01611884
##     class counts:    35    16
##    probabilities: 0.686 0.314
##   left son=20 (31 obs) right son=21 (20 obs)
##   Primary splits:
##       contact   splits as  LRL,         improve=2.283365, (0 missing)
##       education splits as  LRRL,        improve=1.470653, (0 missing)
##       age       < 73.5    to the left,  improve=1.339163, (0 missing)
##       marital   splits as  RLL,         improve=1.278245, (0 missing)
##       balance   < 1561    to the left,  improve=1.227865, (0 missing)
##   Surrogate splits:
##       age       < 73.5    to the left,   agree=0.725, adj=0.30, (0 split)
##       job       splits as  LLRLLL-----R, agree=0.647, adj=0.10, (0 split)
##       education splits as  LLLR,         agree=0.647, adj=0.10, (0 split)
##       day       < 4.5     to the right,  agree=0.647, adj=0.10, (0 split)
##       balance   < 698.5   to the right,  agree=0.627, adj=0.05, (0 split)
##
## Node number 11: 20 observations,    complexity param=0.005602241
##   predicted class=yes  expected loss=0.35  P(node) =0.006321113
##     class counts:     7    13
##    probabilities: 0.350 0.650
##   left son=22 (8 obs) right son=23 (12 obs)
##   Primary splits:
##       campaign < 1.5     to the right,  improve=2.0166670, (0 missing)
##       age      < 73.5    to the right,  improve=1.0560440, (0 missing)
##       balance  < 1269    to the left,   improve=0.5343434, (0 missing)
##       day      < 22.5    to the right,  improve=0.5343434, (0 missing)
##       month    splits as  ---RL-R-----, improve=0.1329670, (0 missing)
##   Surrogate splits:
##       job       splits as  RL-RLRL----R, agree=0.75, adj=0.375, (0 split)
##       education splits as  LRRR,         agree=0.70, adj=0.250, (0 split)
```

```
##       balance   < 4643    to the right,  agree=0.70, adj=0.250, (0 split)
##       contact   splits as  RL-,          agree=0.65, adj=0.125, (0 split)
##       day       < 27.5    to the right,  agree=0.65, adj=0.125, (0 split)
##
## Node number 12: 8 observations
##   predicted class=no   expected loss=0  P(node) =0.002528445
##     class counts:     8     0
##    probabilities: 1.000 0.000
##
## Node number 13: 125 observations,    complexity param=0.0140056
##   predicted class=no   expected loss=0.416  P(node) =0.03950695
##     class counts:    73    52
##    probabilities: 0.584 0.416
##   left son=26 (33 obs) right son=27 (92 obs)
##   Primary splits:
##       age     < 54.5    to the right,  improve=2.7017440, (0 missing)
##       job     splits as  RLLLRLLLLRRR, improve=2.4091600, (0 missing)
##       day     < 15.5    to the left,   improve=2.1050480, (0 missing)
##       balance < 184     to the left,   improve=1.5619410, (0 missing)
##       month   splits as  --L----R--RL, improve=0.9775584, (0 missing)
##   Surrogate splits:
##       job       splits as  RRRRRLRRRRRR, agree=0.880, adj=0.545, (0 split)
##       education splits as  LRRR,         agree=0.744, adj=0.030, (0 split)
##
## Node number 18: 1797 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.1051753  P(node) =0.567952
##     class counts:  1608   189
##    probabilities: 0.895 0.105
##   left son=36 (1490 obs) right son=37 (307 obs)
##   Primary splits:
##       month   splits as  RL-RLL--LL--, improve=3.370259, (0 missing)
##       balance < 1923.5  to the left,   improve=2.867053, (0 missing)
##       marital splits as  RLR,          improve=2.437670, (0 missing)
##       loan    splits as  RL,           improve=2.281270, (0 missing)
##       age     < 22.5    to the right,  improve=1.748160, (0 missing)
##   Surrogate splits:
##       day < 3.5     to the right, agree=0.85, adj=0.121, (0 split)
##       age < 20.5    to the right, agree=0.83, adj=0.007, (0 split)
##
## Node number 19: 46 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.4347826  P(node) =0.01453856
##     class counts:    26    20
##    probabilities: 0.565 0.435
##   left son=38 (38 obs) right son=39 (8 obs)
##   Primary splits:
##       age      < 50.5    to the left,   improve=3.753432, (0 missing)
##       job      splits as  LR--LRRLLLR-, improve=3.226343, (0 missing)
##       campaign < 1.5     to the left,   improve=1.345538, (0 missing)
##       balance  < 1956.5  to the left,   improve=1.290014, (0 missing)
##       day      < 4.5     to the left,   improve=1.285619, (0 missing)
```

```
##   Surrogate splits:
##       job       splits as  LL--LRLLLLL-, agree=0.87, adj=0.25, (0 split)
##       education splits as  RLLR,         agree=0.87, adj=0.25, (0 split)
##
## Node number 20: 31 observations
##   predicted class=no   expected loss=0.1935484  P(node) =0.009797724
##     class counts:    25     6
##    probabilities: 0.806 0.194
##
## Node number 21: 20 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.5  P(node) =0.006321113
##     class counts:    10    10
##    probabilities: 0.500 0.500
##   left son=42 (9 obs) right son=43 (11 obs)
##   Primary splits:
##       balance   < 808.5   to the left,   improve=2.5252530, (0 missing)
##       month     splits as  RR---R--LL--, improve=0.9890110, (0 missing)
##       age       < 68.5    to the right,  improve=0.4166667, (0 missing)
##       education splits as  LRLL,         improve=0.4166667, (0 missing)
##       day       < 8.5     to the right,  improve=0.4166667, (0 missing)
##   Surrogate splits:
##       education splits as  LRRL,         agree=0.85, adj=0.667, (0 split)
##       day       < 16      to the right,  agree=0.70, adj=0.333, (0 split)
##       month     splits as  RR---R--LL--, agree=0.70, adj=0.333, (0 split)
##       age       < 76      to the right,  agree=0.65, adj=0.222, (0 split)
##       marital   splits as  LR-,          agree=0.60, adj=0.111, (0 split)
##
## Node number 22: 8 observations
##   predicted class=no   expected loss=0.375  P(node) =0.002528445
##     class counts:     5     3
##    probabilities: 0.625 0.375
##
## Node number 23: 12 observations
##   predicted class=yes  expected loss=0.1666667  P(node) =0.003792668
##     class counts:     2    10
##    probabilities: 0.167 0.833
##
## Node number 26: 33 observations,    complexity param=0.00280112
##   predicted class=no   expected loss=0.2424242  P(node) =0.01042984
##     class counts:    25     8
##    probabilities: 0.758 0.242
##   left son=52 (24 obs) right son=53 (9 obs)
##   Primary splits:
##       education splits as  LLRR,         improve=2.4267680, (0 missing)
##       job       splits as  RL-LLR---L-L, improve=1.6864300, (0 missing)
##       balance   < 2832    to the left,   improve=1.4848480, (0 missing)
##       age       < 58.5    to the left,   improve=1.0442890, (0 missing)
##       month     splits as  --R----R--LL, improve=0.4848485, (0 missing)
##   Surrogate splits:
##       balance < 21.5    to the right,  agree=0.788, adj=0.222, (0 split)
```

```
##       day     < 25.5    to the left,   agree=0.788, adj=0.222, (0 split)
##       job     splits as  LL-LLL---R-L, agree=0.758, adj=0.111, (0 split)
##
## Node number 27: 92 observations,    complexity param=0.0140056
##   predicted class=no   expected loss=0.4782609  P(node) =0.02907712
##     class counts:    48    44
##    probabilities: 0.522 0.478
##   left son=54 (65 obs) right son=55 (27 obs)
##   Primary splits:
##       age     < 41.5    to the left,   improve=6.856633, (0 missing)
##       job     splits as  RLLRR-LLLRRR, improve=2.261083, (0 missing)
##       balance < 5963.5  to the right,  improve=2.186853, (0 missing)
##       day     < 27.5    to the right,  improve=1.854621, (0 missing)
##       month   splits as  --L----R--RL, improve=1.424407, (0 missing)
##   Surrogate splits:
##       contact splits as  LRL,          agree=0.761, adj=0.185, (0 split)
##       job     splits as  LLLLL-LLLLLR, agree=0.717, adj=0.037, (0 split)
##
## Node number 36: 1490 observations,    complexity param=0.002334267
##   predicted class=no   expected loss=0.09127517  P(node) =0.4709229
##     class counts:  1354   136
##    probabilities: 0.909 0.091
##   left son=72 (1460 obs) right son=73 (30 obs)
##   Primary splits:
##       age     < 25.5    to the right,  improve=1.883657, (0 missing)
##       job     splits as  LLLLLLLLRLLL, improve=1.551930, (0 missing)
##       marital splits as  RLR,          improve=1.479498, (0 missing)
##       balance < 450.5   to the left,   improve=1.471665, (0 missing)
##       day     < 28.5    to the right,  improve=1.285421, (0 missing)
##
## Node number 37: 307 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.1726384  P(node) =0.09702908
##     class counts:   254    53
##    probabilities: 0.827 0.173
##   left son=74 (268 obs) right son=75 (39 obs)
##   Primary splits:
##       day     < 20.5    to the left,   improve=11.957500, (0 missing)
##       housing splits as  RL,           improve= 3.698866, (0 missing)
##       balance < 1886    to the left,   improve= 3.204366, (0 missing)
##       job     splits as  RLLRRLRLRRRR, improve= 2.492309, (0 missing)
##       loan    splits as  RL,           improve= 1.183209, (0 missing)
##
## Node number 38: 38 observations,    complexity param=0.007002801
##   predicted class=no   expected loss=0.3421053  P(node) =0.01201011
##     class counts:    25    13
##    probabilities: 0.658 0.342
##   left son=76 (29 obs) right son=77 (9 obs)
##   Primary splits:
##       job      splits as  LR--L-RLLLR-, improve=2.4845740, (0 missing)
##       balance  < 1294.5  to the left,   improve=1.5237250, (0 missing)
```

```
##       age      < 39      to the right,  improve=1.3211720, (0 missing)
##       campaign < 1.5     to the left,   improve=1.1052630, (0 missing)
##       day      < 13.5    to the left,   improve=0.7690313, (0 missing)
##
## Node number 39: 8 observations
##   predicted class=yes  expected loss=0.125  P(node) =0.002528445
##     class counts:     1     7
##    probabilities: 0.125 0.875
##
## Node number 42: 9 observations
##   predicted class=no   expected loss=0.2222222  P(node) =0.002844501
##     class counts:     7     2
##    probabilities: 0.778 0.222
##
## Node number 43: 11 observations
##   predicted class=yes  expected loss=0.2727273  P(node) =0.003476612
##     class counts:     3     8
##    probabilities: 0.273 0.727
##
## Node number 52: 24 observations
##   predicted class=no   expected loss=0.125  P(node) =0.007585335
##     class counts:    21     3
##    probabilities: 0.875 0.125
##
## Node number 53: 9 observations
##   predicted class=yes  expected loss=0.4444444  P(node) =0.002844501
##     class counts:     4     5
##    probabilities: 0.444 0.556
##
## Node number 54: 65 observations,    complexity param=0.009803922
##   predicted class=no   expected loss=0.3538462  P(node) =0.02054362
##     class counts:    42    23
##    probabilities: 0.646 0.354
##   left son=108 (26 obs) right son=109 (39 obs)
##   Primary splits:
##       balance   < 1047    to the right,  improve=2.261538, (0 missing)
##       job       splits as  LLLRR-LLLLRR, improve=1.625516, (0 missing)
##       contact   splits as  RLL,          improve=1.231013, (0 missing)
##       day       < 15.5    to the left,   improve=1.199911, (0 missing)
##       education splits as  LRRL,         improve=0.783683, (0 missing)
##   Surrogate splits:
##       job       splits as  RRRRL-LRRRLR, agree=0.646, adj=0.115, (0 split)
##       education splits as  LRRL,         agree=0.646, adj=0.115, (0 split)
##       month     splits as  --L----L--RR, agree=0.646, adj=0.115, (0 split)
##       day       < 30.5    to the right,  agree=0.631, adj=0.077, (0 split)
##       age       < 39.5    to the right,  agree=0.615, adj=0.038, (0 split)
##
## Node number 55: 27 observations
##   predicted class=yes  expected loss=0.2222222  P(node) =0.008533502
##     class counts:     6    21
```

```
##    probabilities: 0.222 0.778
##
## Node number 72: 1460 observations,    complexity param=0.002334267
##   predicted class=no   expected loss=0.08767123  P(node) =0.4614412
##     class counts:  1332   128
##    probabilities: 0.912 0.088
##   left son=144 (1200 obs) right son=145 (260 obs)
##   Primary splits:
##       balance < 2078.5  to the left,   improve=1.6320620, (0 missing)
##       day     < 26.5    to the right,  improve=1.1936180, (0 missing)
##       marital splits as  RLR,          improve=1.1416150, (0 missing)
##       loan    splits as  RL,           improve=1.1340950, (0 missing)
##       month   splits as  -R--LL--RL--, improve=0.9684128, (0 missing)
##   Surrogate splits:
##       day < 2.5     to the right, agree=0.823, adj=0.008, (0 split)
##
## Node number 73: 30 observations
##   predicted class=no   expected loss=0.2666667  P(node) =0.009481669
##     class counts:    22     8
##    probabilities: 0.733 0.267
##
## Node number 74: 268 observations,    complexity param=0.003361345
##   predicted class=no   expected loss=0.119403  P(node) =0.08470291
##     class counts:   236    32
##    probabilities: 0.881 0.119
##   left son=148 (143 obs) right son=149 (125 obs)
##   Primary splits:
##       job       splits as  LLLRRLRLRRLR, improve=2.4693140, (0 missing)
##       balance   < 1886    to the left,   improve=2.0227690, (0 missing)
##       housing   splits as  RL,           improve=1.9722310, (0 missing)
##       education splits as  LLRL,         improve=1.1795530, (0 missing)
##       age       < 48.5    to the left,   improve=0.7026036, (0 missing)
##   Surrogate splits:
##       education splits as  LLRL,        agree=0.754, adj=0.472, (0 split)
##       marital   splits as  LLR,         agree=0.575, adj=0.088, (0 split)
##       age       < 34.5    to the right, agree=0.567, adj=0.072, (0 split)
##       housing   splits as  RL,          agree=0.567, adj=0.072, (0 split)
##       balance   < 1771    to the left,  agree=0.560, adj=0.056, (0 split)
##
## Node number 75: 39 observations,    complexity param=0.007002801
##   predicted class=yes  expected loss=0.4615385  P(node) =0.01232617
##     class counts:    18    21
##    probabilities: 0.462 0.538
##   left son=150 (20 obs) right son=151 (19 obs)
##   Primary splits:
##       job       splits as  LR--LLRRLRRL, improve=2.9161940, (0 missing)
##       balance   < 550     to the right,  improve=1.0989010, (0 missing)
##       age       < 31.5    to the left,   improve=0.9365634, (0 missing)
##       day       < 25      to the right,  improve=0.7018568, (0 missing)
##       education splits as  RRLL,         improve=0.5909646, (0 missing)
```

```
##   Surrogate splits:
##       campaign  < 1.5     to the right,  agree=0.692, adj=0.368, (0 split)
##       education splits as  RRLL,         agree=0.641, adj=0.263, (0 split)
##       day       < 22.5    to the right,  agree=0.615, adj=0.211, (0 split)
##       balance   < 146     to the left,   agree=0.590, adj=0.158, (0 split)
##       month     splits as  R--L--------, agree=0.590, adj=0.158, (0 split)
##
## Node number 76: 29 observations
##   predicted class=no   expected loss=0.2413793  P(node) =0.009165613
##     class counts:    22     7
##    probabilities: 0.759 0.241
##
## Node number 77: 9 observations
##   predicted class=yes  expected loss=0.3333333  P(node) =0.002844501
##     class counts:     3     6
##    probabilities: 0.333 0.667
##
## Node number 108: 26 observations
##   predicted class=no   expected loss=0.1923077  P(node) =0.008217446
##     class counts:    21     5
##    probabilities: 0.808 0.192
##
## Node number 109: 39 observations,    complexity param=0.009803922
##   predicted class=no   expected loss=0.4615385  P(node) =0.01232617
##     class counts:    21    18
##    probabilities: 0.538 0.462
##   left son=218 (28 obs) right son=219 (11 obs)
##   Primary splits:
##       balance  < 475.5   to the left,   improve=3.897602, (0 missing)
##       job      splits as  LLLRR--LLLLR, improve=3.081167, (0 missing)
##       campaign < 1.5     to the left,   improve=2.884615, (0 missing)
##       contact  splits as  RLL,          improve=1.732830, (0 missing)
##       day      < 15.5    to the left,   improve=1.449833, (0 missing)
##   Surrogate splits:
##       job   splits as  LLLLR--LLLLL, agree=0.769, adj=0.182, (0 split)
##       age   < 24.5    to the right,  agree=0.744, adj=0.091, (0 split)
##       month splits as  --R----L--LL, agree=0.744, adj=0.091, (0 split)
##
## Node number 144: 1200 observations
##   predicted class=no   expected loss=0.07666667  P(node) =0.3792668
##     class counts:  1108    92
##    probabilities: 0.923 0.077
##
## Node number 145: 260 observations,    complexity param=0.002334267
##   predicted class=no   expected loss=0.1384615  P(node) =0.08217446
##     class counts:   224    36
##    probabilities: 0.862 0.138
##   left son=290 (101 obs) right son=291 (159 obs)
##   Primary splits:
##       balance   < 4759    to the right, improve=2.6138690, (0 missing)
```

```
##       loan      splits as  RL,          improve=1.4492710, (0 missing)
##       day       < 18.5    to the right, improve=1.4381230, (0 missing)
##       age       < 58.5    to the left,  improve=1.2108930, (0 missing)
##       education splits as  RLLL,        improve=0.7819005, (0 missing)
##   Surrogate splits:
##       job     splits as  RRRLRRRLRRRR, agree=0.627, adj=0.04, (0 split)
##       marital splits as  LRR,          agree=0.619, adj=0.02, (0 split)
##       age     < 58.5    to the right,  agree=0.615, adj=0.01, (0 split)
##       day     < 4.5     to the left,   agree=0.615, adj=0.01, (0 split)
##
## Node number 148: 143 observations
##   predicted class=no   expected loss=0.05594406  P(node) =0.04519595
##     class counts:   135     8
##    probabilities: 0.944 0.056
##
## Node number 149: 125 observations,    complexity param=0.003361345
##   predicted class=no   expected loss=0.192  P(node) =0.03950695
##     class counts:   101    24
##    probabilities: 0.808 0.192
##   left son=298 (103 obs) right son=299 (22 obs)
##   Primary splits:
##       balance  < 1978.5  to the left,   improve=2.5165680, (0 missing)
##       housing  splits as  RL,           improve=2.2628960, (0 missing)
##       campaign < 3.5     to the left,   improve=1.3947250, (0 missing)
##       month    splits as  L--R--------, improve=0.8823607, (0 missing)
##       day      < 17.5    to the left,   improve=0.8599331, (0 missing)
##
## Node number 150: 20 observations,    complexity param=0.005602241
##   predicted class=no   expected loss=0.35  P(node) =0.006321113
##     class counts:    13     7
##    probabilities: 0.650 0.350
##   left son=300 (10 obs) right son=301 (10 obs)
##   Primary splits:
##       balance  < 619     to the right,  improve=2.5000000, (0 missing)
##       marital  splits as  LLR,          improve=0.9000000, (0 missing)
##       job      splits as  R---LL--R--L, improve=0.5343434, (0 missing)
##       age      < 34.5    to the right,  improve=0.2919192, (0 missing)
##       campaign < 1.5     to the right,  improve=0.2919192, (0 missing)
##   Surrogate splits:
##       day     < 29.5    to the left,   agree=0.75, adj=0.5, (0 split)
##       housing splits as  LR,           agree=0.70, adj=0.4, (0 split)
##       loan    splits as  LR,           agree=0.65, adj=0.3, (0 split)
##       age     < 36      to the right,  agree=0.60, adj=0.2, (0 split)
##       job     splits as  R---LR--R--L, agree=0.60, adj=0.2, (0 split)
##
## Node number 151: 19 observations
##   predicted class=yes  expected loss=0.2631579  P(node) =0.006005057
##     class counts:     5    14
##    probabilities: 0.263 0.737
##
```

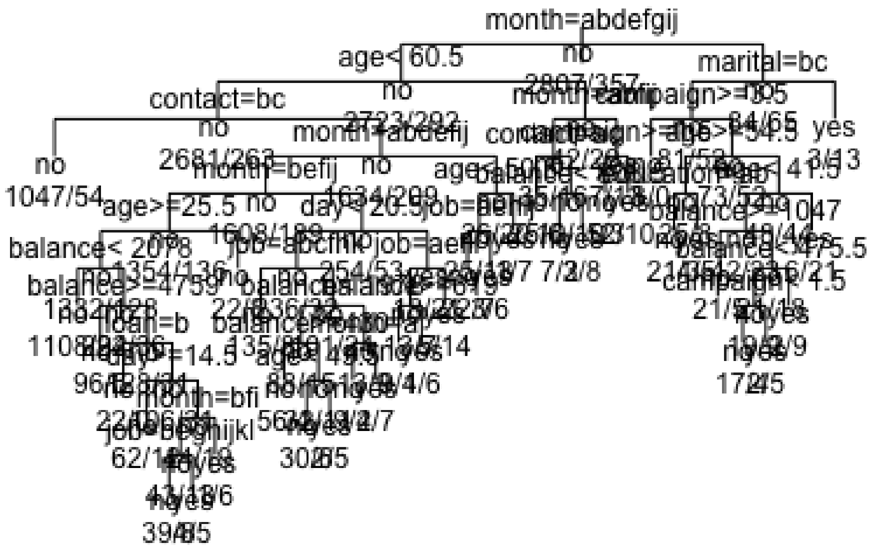
```
## Node number 218: 28 observations,    complexity param=0.008403361
##   predicted class=no   expected loss=0.3214286  P(node) =0.008849558
##     class counts:    19     9
##    probabilities: 0.679 0.321
##   left son=436 (21 obs) right son=437 (7 obs)
##   Primary splits:
##       campaign < 1.5     to the left,   improve=2.8809520, (0 missing)
##       job      splits as  LRLRR--LLLLR, improve=2.0642860, (0 missing)
##       housing  splits as  LR,           improve=1.4540520, (0 missing)
##       balance  < 3.5     to the right,  improve=0.7142857, (0 missing)
##       day      < 12      to the right,  improve=0.6428571, (0 missing)
##   Surrogate splits:
##       job splits as  LLLLR--LLLLR, agree=0.821, adj=0.286, (0 split)
##
## Node number 219: 11 observations
##   predicted class=yes  expected loss=0.1818182  P(node) =0.003476612
##     class counts:     2     9
##    probabilities: 0.182 0.818
##
## Node number 290: 101 observations
##   predicted class=no   expected loss=0.04950495  P(node) =0.03192162
##     class counts:    96     5
##    probabilities: 0.950 0.050
##
## Node number 291: 159 observations,    complexity param=0.002334267
##   predicted class=no   expected loss=0.1949686  P(node) =0.05025284
##     class counts:   128    31
##    probabilities: 0.805 0.195
##   left son=582 (22 obs) right son=583 (137 obs)
##   Primary splits:
##       loan    splits as  RL,          improve=1.9411470, (0 missing)
##       day     < 14.5    to the right, improve=1.4621900, (0 missing)
##       marital splits as  LLR,         improve=1.1382370, (0 missing)
##       age     < 38.5    to the right, improve=1.0569420, (0 missing)
##       balance < 4105.5  to the left,  improve=0.9500607, (0 missing)
##   Surrogate splits:
##       job splits as  RRLRRRRRRRR, agree=0.868, adj=0.045, (0 split)
##
## Node number 298: 103 observations,    complexity param=0.003361345
##   predicted class=no   expected loss=0.1456311  P(node) =0.03255373
##     class counts:    88    15
##    probabilities: 0.854 0.146
##   left son=596 (60 obs) right son=597 (43 obs)
##   Primary splits:
##       balance  < 430     to the right, improve=1.7923080, (0 missing)
##       age      < 49.5    to the left,  improve=1.2371290, (0 missing)
##       day      < 17.5    to the left,  improve=0.9570753, (0 missing)
##       housing  splits as  RL,          improve=0.9099141, (0 missing)
##       campaign < 3.5     to the left,  improve=0.8978258, (0 missing)
##   Surrogate splits:
```

```
##       education splits as  LRLL,         agree=0.670, adj=0.209, (0 split)
##       default   splits as  LR,           agree=0.621, adj=0.093, (0 split)
##       age       < 42.5    to the left,   agree=0.612, adj=0.070, (0 split)
##       day       < 2.5     to the right,  agree=0.612, adj=0.070, (0 split)
##       job       splits as  ---LL-L-LL-R, agree=0.592, adj=0.023, (0 split)
##
## Node number 299: 22 observations,    complexity param=0.003361345
##   predicted class=no   expected loss=0.4090909  P(node) =0.006953224
##     class counts:    13     9
##    probabilities: 0.591 0.409
##   left son=598 (11 obs) right son=599 (11 obs)
##   Primary splits:
##       month    splits as  L--R--------, improve=2.272727, (0 missing)
##       housing  splits as  RL,           improve=2.020979, (0 missing)
##       campaign < 1.5     to the left,   improve=1.603030, (0 missing)
##       day      < 7       to the left,   improve=1.455411, (0 missing)
##       balance  < 4602.5  to the right,  improve=1.455411, (0 missing)
##   Surrogate splits:
##       education splits as  LLRR,        agree=0.682, adj=0.364, (0 split)
##       campaign  < 1.5     to the left,  agree=0.682, adj=0.364, (0 split)
##       housing   splits as  RL,          agree=0.636, adj=0.273, (0 split)
##       day       < 3.5     to the right, agree=0.636, adj=0.273, (0 split)
##       age       < 35.5    to the left,  agree=0.591, adj=0.182, (0 split)
##
## Node number 300: 10 observations
##   predicted class=no   expected loss=0.1  P(node) =0.003160556
##     class counts:     9     1
##    probabilities: 0.900 0.100
##
## Node number 301: 10 observations
##   predicted class=yes  expected loss=0.4  P(node) =0.003160556
##     class counts:     4     6
##    probabilities: 0.400 0.600
##
## Node number 436: 21 observations
##   predicted class=no   expected loss=0.1904762  P(node) =0.006637168
##     class counts:    17     4
##    probabilities: 0.810 0.190
##
## Node number 437: 7 observations
##   predicted class=yes  expected loss=0.2857143  P(node) =0.002212389
##     class counts:     2     5
##    probabilities: 0.286 0.714
##
## Node number 582: 22 observations
##   predicted class=no   expected loss=0  P(node) =0.006953224
##     class counts:    22     0
##    probabilities: 1.000 0.000
##
## Node number 583: 137 observations,    complexity param=0.002334267
```

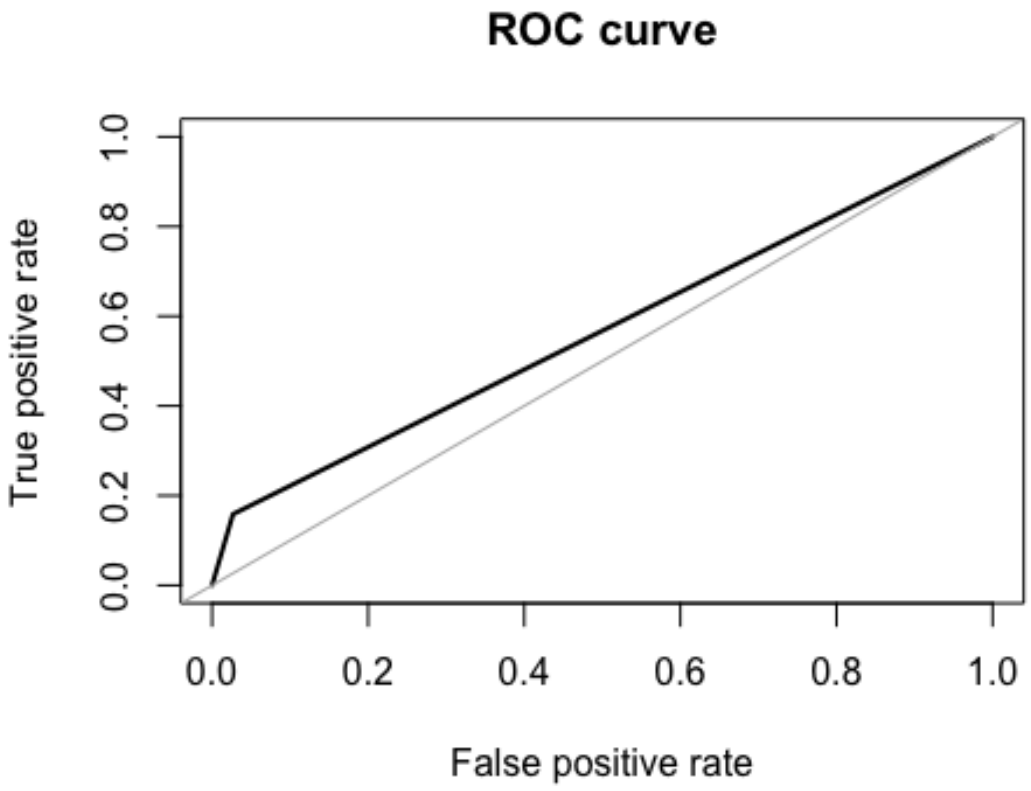
```
##   predicted class=no   expected loss=0.2262774  P(node) =0.04329962
##     class counts:   106    31
##    probabilities: 0.774 0.226
##   left son=1166 (74 obs) right son=1167 (63 obs)
##   Primary splits:
##       day     < 14.5    to the right,  improve=1.3230120, (0 missing)
##       balance < 3333.5  to the left,   improve=0.9726582, (0 missing)
##       age     < 38.5    to the right,  improve=0.8612080, (0 missing)
##       job     splits as  RRRRRRLRRRLL, improve=0.6407781, (0 missing)
##       housing splits as  LR,           improve=0.6344781, (0 missing)
##   Surrogate splits:
##       month     splits as  -R--LL--RL--, agree=0.715, adj=0.381, (0 split)
##       job       splits as  LLLRLLLRLLRR, agree=0.606, adj=0.143, (0 split)
##       education splits as  RRLL,         agree=0.591, adj=0.111, (0 split)
##       campaign  < 2.5     to the right,  agree=0.591, adj=0.111, (0 split)
##       age       < 28.5    to the right,  agree=0.562, adj=0.048, (0 split)
##
## Node number 596: 60 observations
##   predicted class=no   expected loss=0.06666667  P(node) =0.01896334
##     class counts:    56     4
##    probabilities: 0.933 0.067
##
## Node number 597: 43 observations,    complexity param=0.003361345
##   predicted class=no   expected loss=0.255814  P(node) =0.01359039
##     class counts:    32    11
##    probabilities: 0.744 0.256
##   left son=1194 (36 obs) right son=1195 (7 obs)
##   Primary splits:
##       age       < 49.5    to the left,   improve=3.5149500, (0 missing)
##       education splits as  RLRL,         improve=1.5547020, (0 missing)
##       job       splits as  ---RR-L-LL-R, improve=0.6633974, (0 missing)
##       housing   splits as  RL,           improve=0.5304641, (0 missing)
##       day       < 17.5    to the left,   improve=0.4990772, (0 missing)
##   Surrogate splits:
##       education splits as  RLLL,         agree=0.884, adj=0.286, (0 split)
##       job       splits as  ---LL-L-LL-R, agree=0.860, adj=0.143, (0 split)
##       marital   splits as  RLL,          agree=0.860, adj=0.143, (0 split)
##       balance   < -268    to the right,  agree=0.860, adj=0.143, (0 split)
##
## Node number 598: 11 observations
##   predicted class=no   expected loss=0.1818182  P(node) =0.003476612
##     class counts:     9     2
##    probabilities: 0.818 0.182
##
## Node number 599: 11 observations
##   predicted class=yes  expected loss=0.3636364  P(node) =0.003476612
##     class counts:     4     7
##    probabilities: 0.364 0.636
##
## Node number 1166: 74 observations
```

```
##   predicted class=no   expected loss=0.1621622  P(node) =0.02338812
##     class counts:    62    12
##    probabilities: 0.838 0.162
##
## Node number 1167: 63 observations,    complexity param=0.002334267
##   predicted class=no   expected loss=0.3015873  P(node) =0.0199115
##     class counts:    44    19
##    probabilities: 0.698 0.302
##   left son=2334 (56 obs) right son=2335 (7 obs)
##   Primary splits:
##       month     splits as  -L--RL--LR--, improve=4.8611110, (0 missing)
##       job       splits as  RLRLLRRRLLLL, improve=2.6213150, (0 missing)
##       balance   < 3688.5  to the left,   improve=2.1170410, (0 missing)
##       campaign  < 1.5     to the right,  improve=1.2908450, (0 missing)
##       education splits as  RLLL,         improve=0.8381441, (0 missing)
##
## Node number 1194: 36 observations
##   predicted class=no   expected loss=0.1666667  P(node) =0.011378
##     class counts:    30     6
##    probabilities: 0.833 0.167
##
## Node number 1195: 7 observations
##   predicted class=yes  expected loss=0.2857143  P(node) =0.002212389
##     class counts:     2     5
##    probabilities: 0.286 0.714
##
## Node number 2334: 56 observations,    complexity param=0.002334267
##   predicted class=no   expected loss=0.2321429  P(node) =0.01769912
##     class counts:    43    13
##    probabilities: 0.768 0.232
##   left son=4668 (47 obs) right son=4669 (9 obs)
##   Primary splits:
##       job      splits as  RLRRLRLLLLLL, improve=2.2432460, (0 missing)
##       balance  < 3688.5  to the left,   improve=1.8418370, (0 missing)
##       campaign < 1.5     to the right,  improve=0.7124003, (0 missing)
##       age      < 48      to the left,   improve=0.5833333, (0 missing)
##       marital  splits as  -LR,          improve=0.5833333, (0 missing)
##   Surrogate splits:
##       balance < 4152    to the left,  agree=0.893, adj=0.333, (0 split)
##
## Node number 2335: 7 observations
##   predicted class=yes  expected loss=0.1428571  P(node) =0.002212389
##     class counts:     1     6
##    probabilities: 0.143 0.857
##
## Node number 4668: 47 observations
##   predicted class=no   expected loss=0.1702128  P(node) =0.01485461
##     class counts:    39     8
##    probabilities: 0.830 0.170
##
```

```
## Node number 4669: 9 observations
##   predicted class=yes  expected loss=0.4444444  P(node) =0.002844501
##     class counts:     4     5
##    probabilities: 0.444 0.556

printcp(tree_fit1)

##
## Classification tree:
## rpart(formula = y ~ ., data = bank_df, subset = train, method = "class",
##     control = rpart.control(maxdepth = 20, cp = 0.0018727))
##
## Variables actually used in tree construction:
##  [1] age       balance   campaign  contact   day       education job
##  [8] loan      marital   month
##
## Root node error: 357/3164 = 0.11283
##
## n= 3164
##
##          CP nsplit rel error xerror     xstd
## 1 0.0140056      0   1.00000 1.0000 0.049850
## 2 0.0098039      5   0.92997 1.0000 0.049850
## 3 0.0084034      7   0.91036 1.0000 0.049850
## 4 0.0070028     10   0.88515 1.0000 0.049850
## 5 0.0056022     19   0.82073 1.0000 0.049850
## 6 0.0033613     21   0.80952 1.0420 0.050751
## 7 0.0028011     26   0.79272 1.0644 0.051220
## 8 0.0023343     27   0.78992 1.0728 0.051394
## 9 0.0018727     34   0.77311 1.0812 0.051567

plot(tree_fit1, uniform = TRUE)
text(tree_fit1, all=TRUE, cex=0.75, splits=TRUE, use.n=TRUE, xpd = TRUE)
```




```
library(maptree)

## Loading required package: cluster

tree_pred = predict(tree_fit1, bank_df[-train,], type="class")
confusion_matrix2 <- table(tree_pred, actual = bank_df[-train,]$y)
confusion_matrix2

##           actual
## tree_pred   no  yes
##        no  1161  138
##        yes   32   26

cat("Accuracy of CT : ",((confusion_matrix2[1,"no"] +
confusion_matrix2[2,"yes"])/1357) )

## Accuracy of CT :  0.8747237

roc.curve(bank_df[-train,]$y, tree_pred, plotit = TRUE)
```




```
## Area under the curve (AUC): 0.566
```

```
######## Random Forests
```

```
library(randomForest)
```

```
## randomForest 4.6-14
```

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

```
##
## Attaching package: 'randomForest'
```

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

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

```
rf_fit <- randomForest(y~., data = bank_df, subset = train)
rf_fit
```

```
##
## Call:
##  randomForest(formula = y ~ ., data = bank_df, subset = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 11.76%
## Confusion matrix:
##       no yes class.error
## no  2761  46   0.0163876
## yes  326  31   0.9131653
```

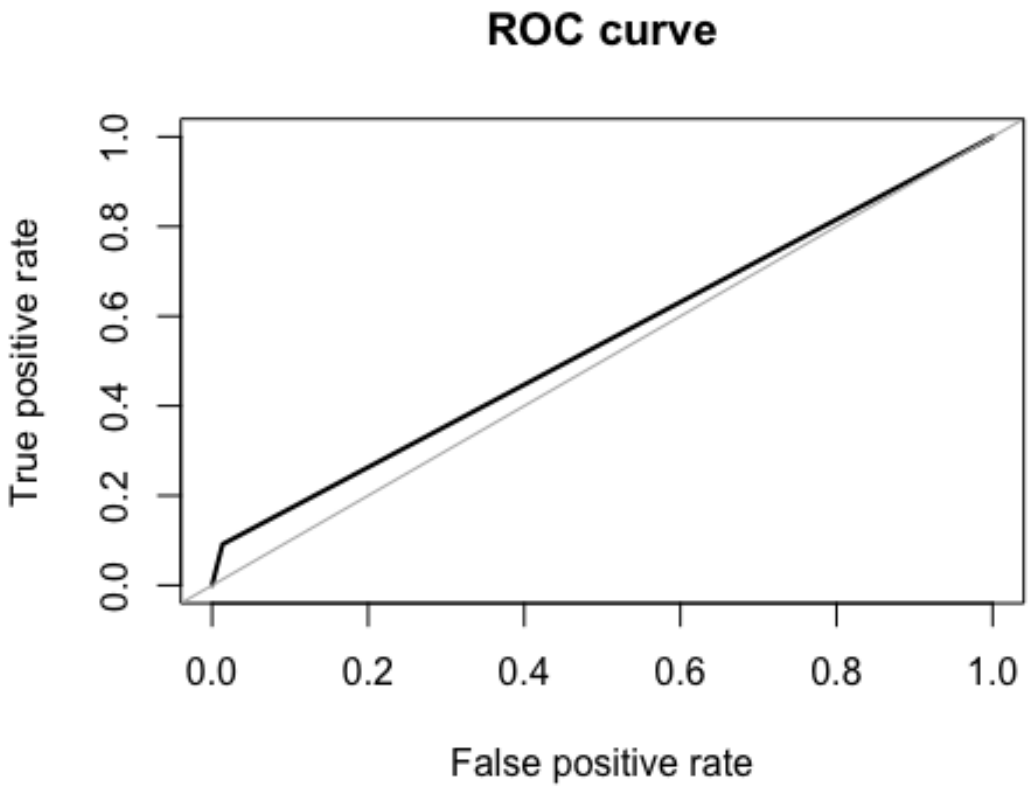
```
confusion_matrix3 <- table( predicted = predict(rf_fit, newdata = bank_df[-
train,], type = "class"),
                            actual = bank_df[-train,]$y)
confusion_matrix3
```

```
##          actual
## predicted   no  yes
##       no  1177  149
##       yes   16   15
```

```
cat("Accuracy of RF : ",((confusion_matrix3[1,"no"] +
confusion_matrix3[2,"yes"])/1357) )
```

```
## Accuracy of RF :  0.8784083
```

```
roc.curve(bank_df[-train,]$y, predict(rf_fit, newdata = bank_df[-train,],
type = "class"), plotit = TRUE)
```




```
## Area under the curve (AUC): 0.539

################################## Over Sampling
#####################################

#over_sampled_data <- ovun.sample(y~., data = bank_df[train,], method =
"both", N=4000,
#                                 p=0.5, seed = 1)$data
#table(over_sampled_data$y)

rose_data <- ROSE(y~., data = bank_df[train,], seed = 1)$data
table(rose_data$y)

##
##   no  yes
## 1642 1522

# Logistic Regression 2
glm.fit_2 <- glm(y~., data = rose_data, family = binomial)
glm.probs_2 = predict(glm.fit_2, newdata = bank_df[-train,], type="response")
glm.pred_2 = ifelse(glm.probs_2>0.5, "yes","no")
actual = bank_df[-train,]$y
mean(glm.pred_2==actual)
```

```
## [1] 0.678703
```

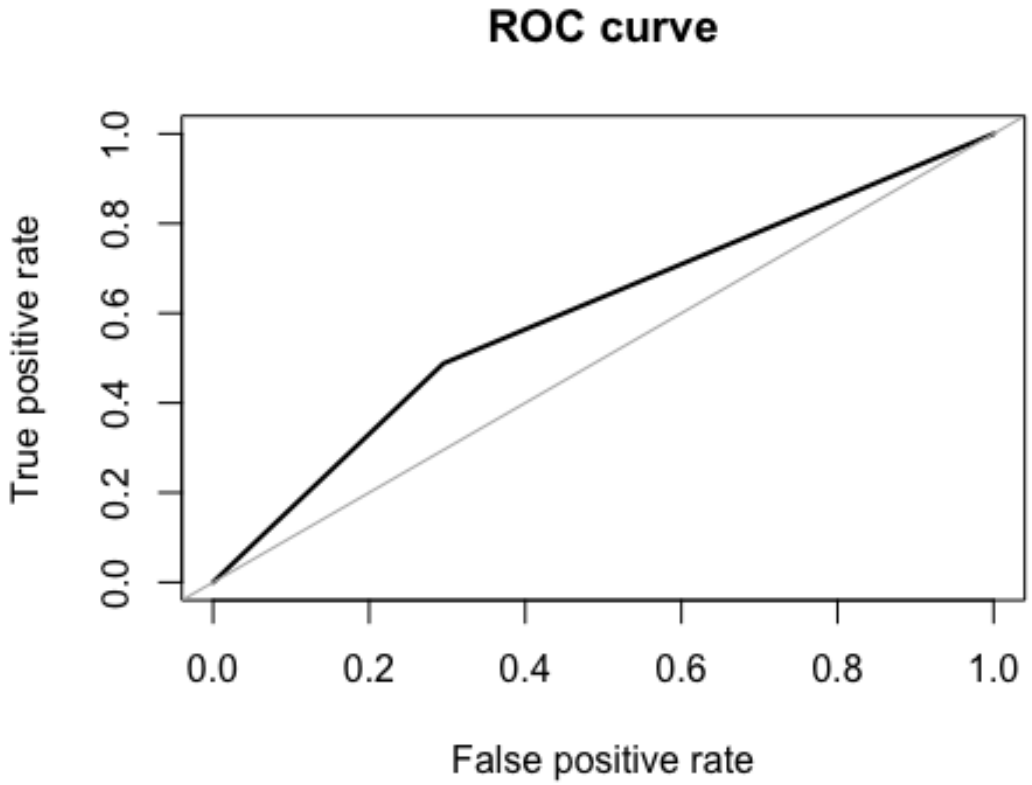
```
confusion_matrix4 <- table(glm.pred_2, actual)
confusion_matrix4
```

```
##           actual
## glm.pred_2  no yes
##        no  841  84
##        yes 352  80
```

```
cat("Accuracy of Logistic Regression 2 : ",((confusion_matrix4[1,"no"] +
confusion_matrix4[2,"yes"])/1357) )
```

```
## Accuracy of Logistic Regression 2 :  0.678703
```

```
r1 = roc.curve(bank_df[-train,]$y, glm.pred_2, plotit = TRUE)
```




```
r1
```

```
## Area under the curve (AUC): 0.596
```

```
# Classification Tree 2

tree_fit2 <- rpart(y~., method = "class", data = rose_data, control =
rpart.control(maxdepth = 20, cp=0.0026281))
```

```r
#summary(tree_fit2)
printcp(tree_fit2)

##
## Classification tree:
## rpart(formula = y ~ ., data = rose_data, method = "class", control =
rpart.control(maxdepth = 20,
##     cp = 0.0026281))
##
## Variables actually used in tree construction:
##  [1] age       balance   campaign  contact   day       education housing
##  [8] job       loan      marital   month
##
## Root node error: 1522/3164 = 0.48104
##
## n= 3164
##
##           CP nsplit rel error  xerror     xstd
## 1  0.2227332      0   1.00000 1.00000 0.018465
## 2  0.0558476      1   0.77727 0.77727 0.017881
## 3  0.0096364      2   0.72142 0.72208 0.017596
## 4  0.0082129     11   0.61104 0.67083 0.017278
## 5  0.0055848     13   0.59461 0.63469 0.017020
## 6  0.0050372     17   0.56833 0.61564 0.016873
## 7  0.0045992     24   0.52365 0.61761 0.016889
## 8  0.0043802     25   0.51905 0.61629 0.016878
## 9  0.0042707     28   0.50591 0.60381 0.016778
## 10 0.0036137     30   0.49737 0.59855 0.016734
## 11 0.0032852     32   0.49014 0.60118 0.016756
## 12 0.0029566     34   0.48357 0.60118 0.016756
## 13 0.0026281     38   0.47175 0.59921 0.016740
## 14 0.0026281     43   0.45861 0.59724 0.016723

plot(tree_fit2, uniform = TRUE)
text(tree_fit2, all=TRUE, cex=0.75, splits=TRUE, use.n=TRUE, xpd = TRUE)
```

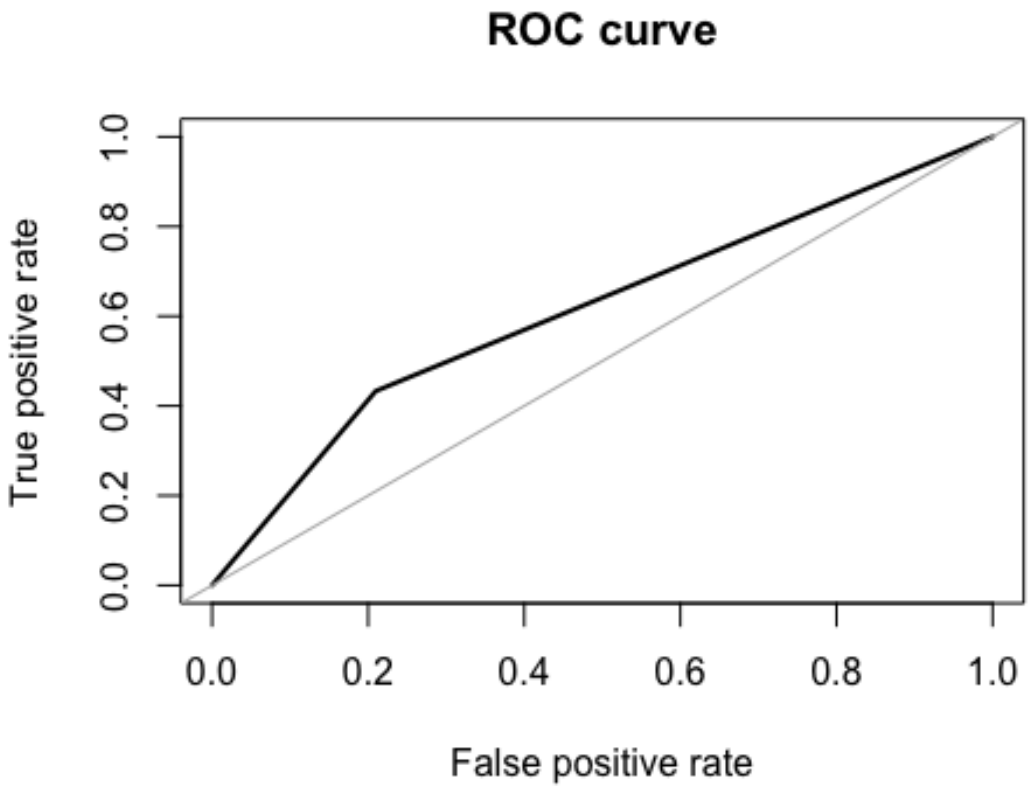

```
library(maptree)
tree_pred_2 = predict(tree_fit2, bank_df[-train,], type="class")
confusion_matrix5 <- table(tree_pred_2, actual = bank_df[-train,]$y)
confusion_matrix5
```

```
##            actual
## tree_pred_2  no yes
##         no  943  93
##         yes 250  71
```

```
cat("Accuracy of CT 2 : ",((confusion_matrix5[1,"no"] +
confusion_matrix5[2,"yes"])/1357),"\n" )
```

```
## Accuracy of CT 2 :  0.7472366
```

```
r2 =roc.curve(bank_df[-train,]$y, tree_pred_2, plotit = TRUE)
```




```
r2

## Area under the curve (AUC): 0.612

# Random Forests 2

library(randomForest)
set.seed(1)
rf_fit2 <- randomForest(y~., data = rose_data, ntree = 500)
rf_fit2

##
## Call:
##  randomForest(formula = y ~ ., data = rose_data, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 15.49%
## Confusion matrix:
##       no  yes class.error
## no  1395  247   0.1504263
## yes  243 1279   0.1596583
```

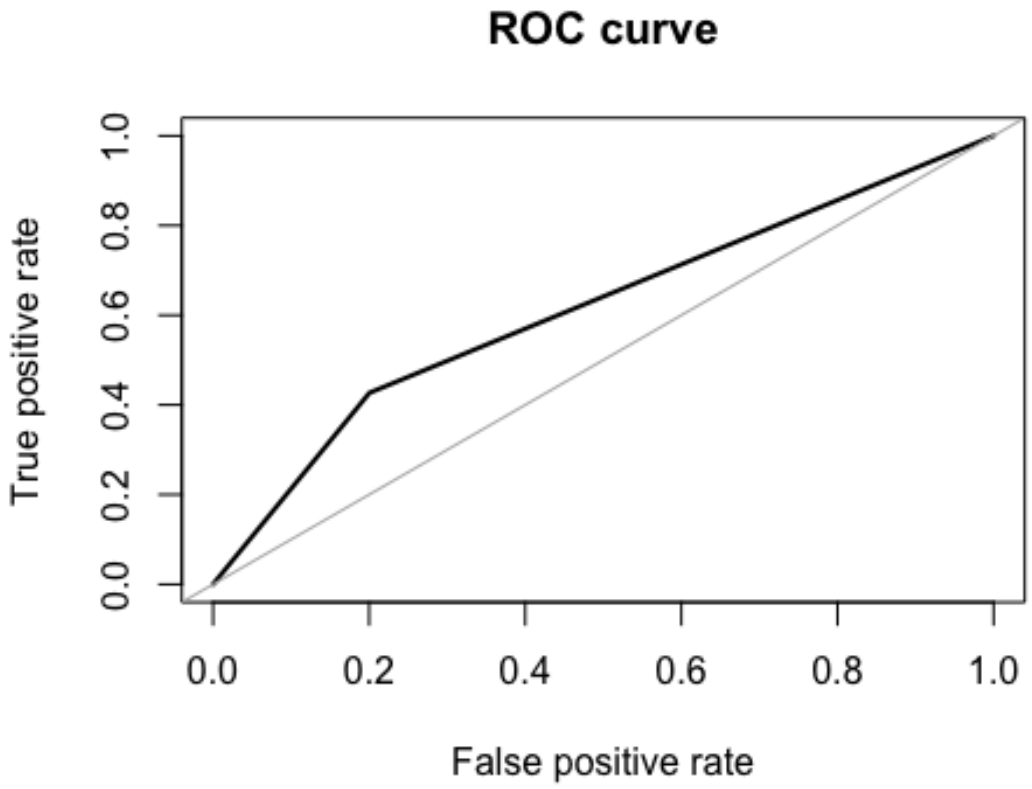
```
confusion_matrix6 <- table( predicted = predict(rf_fit2, newdata = bank_df[-
train,], type = "class"),
                            actual = bank_df[-train,]$y)
confusion_matrix6
```

```
##          actual
## predicted  no yes
##       no  955  94
##       yes 238  70
```

```
cat("Accuracy of RF 2 : ",((confusion_matrix6[1,"no"] +
confusion_matrix6[2,"yes"])/1357),"\n" )
```

```
## Accuracy of RF 2 :  0.7553427
```

```
r3 = roc.curve(bank_df[-train,]$y, predict(rf_fit2, newdata = bank_df[-
train,], type = "class"), plotit = TRUE)
```




```
r3
```

```
## Area under the curve (AUC): 0.613
```

```
##################################  RESULTS
#########################################
```

```r
cat("\n\n Model Performance : \n\n")
```

```
##
##
##  Model Performance :
```

```r
cat("AUC of Logistic Regression : ", r1$auc,"\n")
```

```
## AUC of Logistic Regression :  0.5963752
```

```r
cat("AUC of Classification Tree : ", r2$auc,"\n")
```

```
## AUC of Classification Tree :  0.6116855
```

```r
cat("AUC of RF : ", r3$auc,"\n")
```

```
## AUC of RF :  0.613247
```