Bank Telemarketing

Matt Goldsmith

June 28, 2020

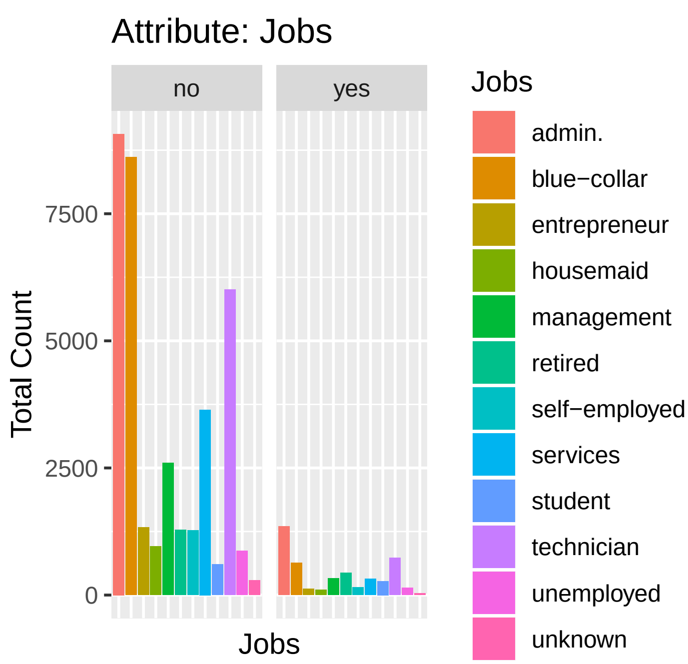
Background

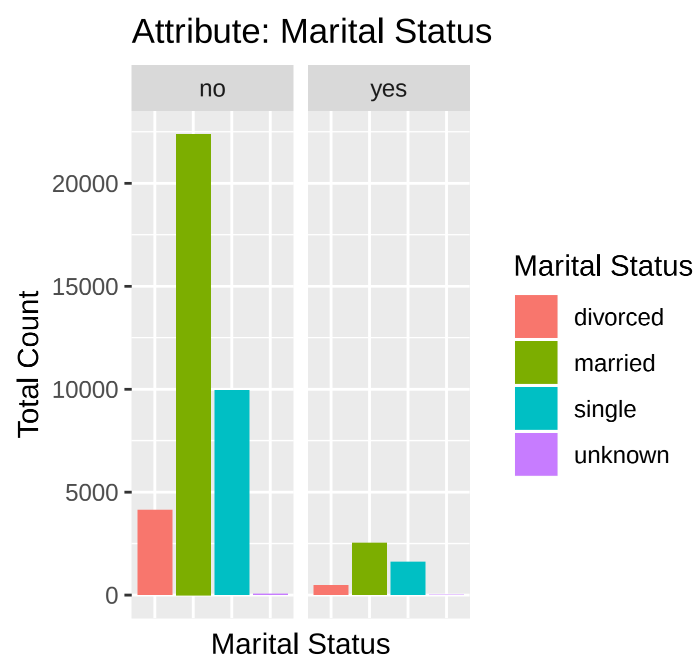
The purpose of this project is to examine the success of telemarketing in banking. The dataset used contains 41,188 unique instances with 20+ attributes per instance. The goal is to accurately classify a binary response variable (if the client subscribed to a bank term deposit).

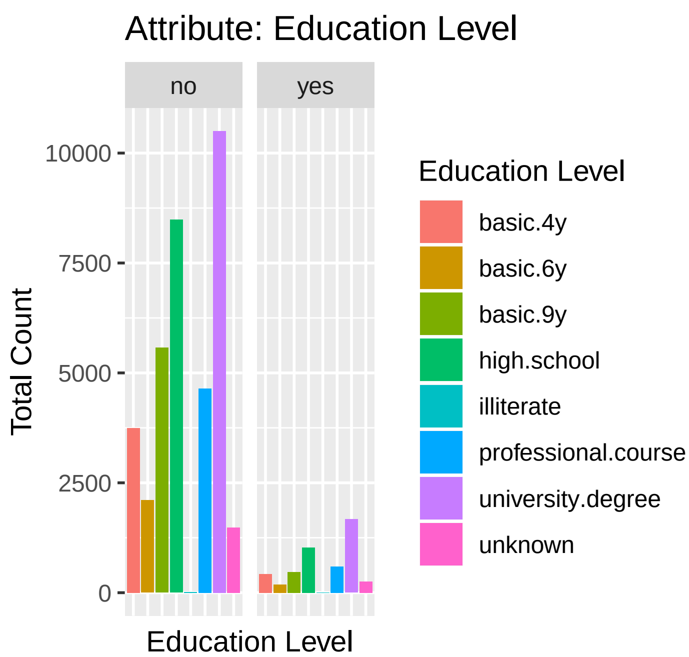
Dataset citation: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing.

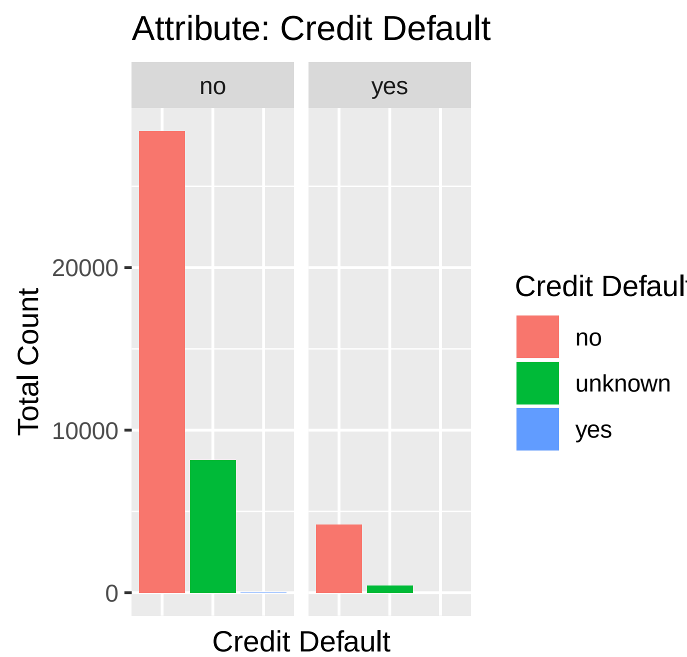
R Packages: tidyverse, corrplot

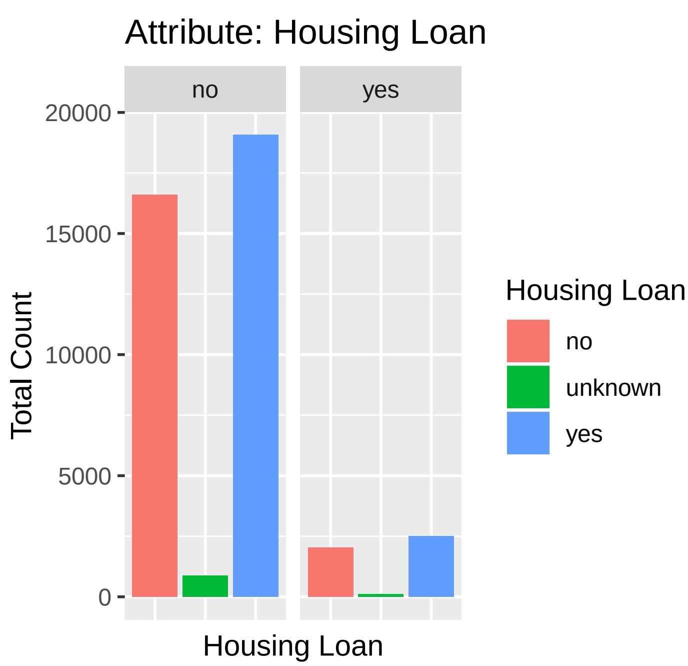
Data Visualization

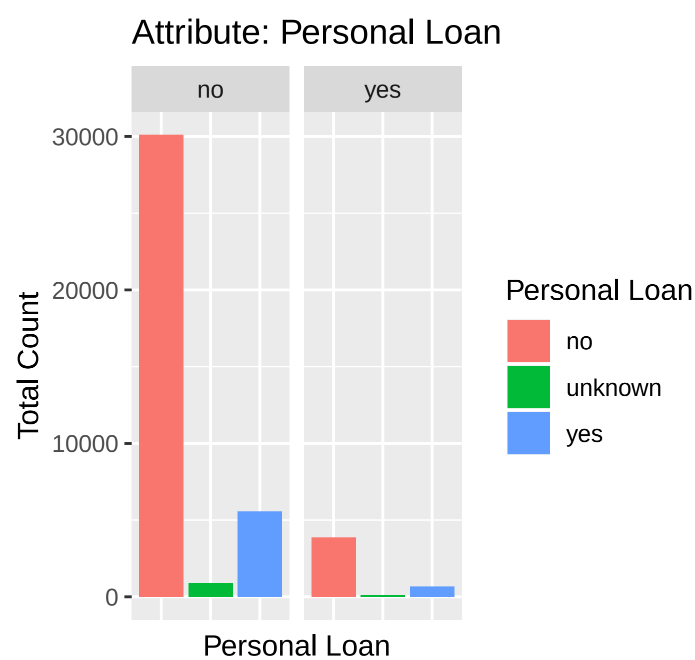


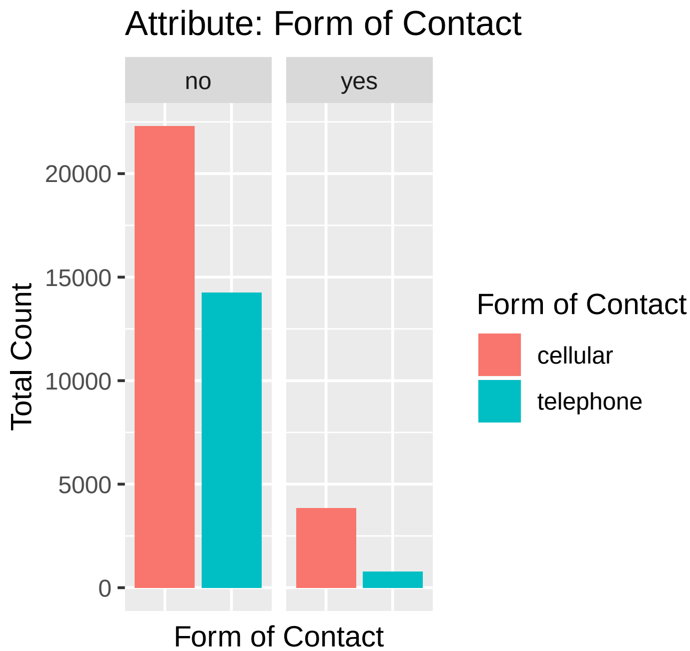


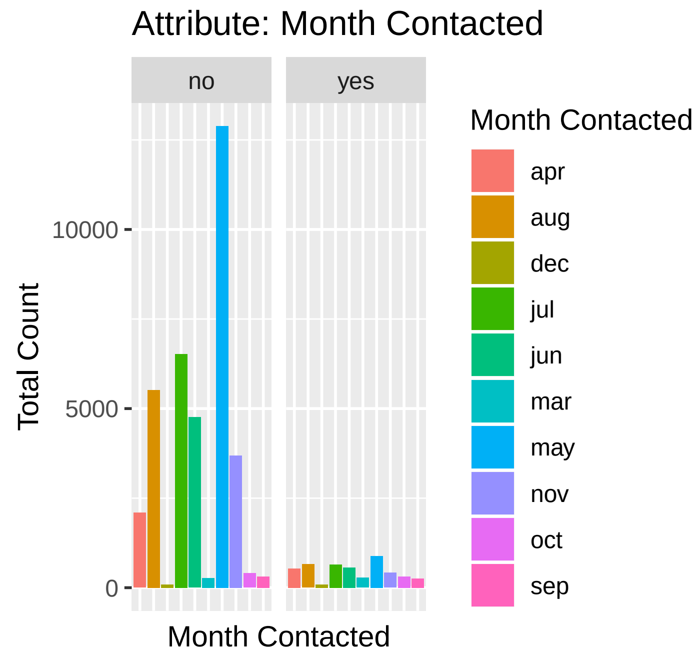

Attribute: Education Level
no
yes
Total Count
10000
7500
5000
2500
0
Education Level
Education Level
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown

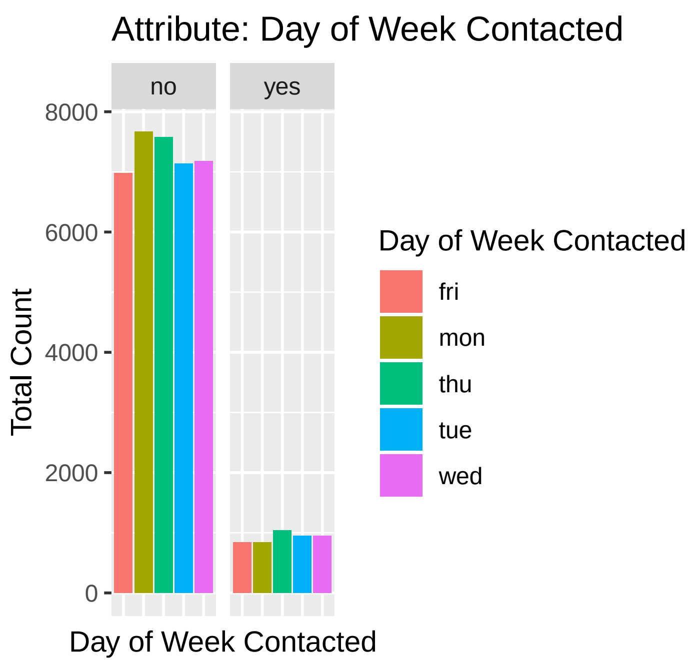

Attribute: Credit Default
no
yes
Total Count
20000
10000
0
Credit Default
Credit Default
no
unknown
yes

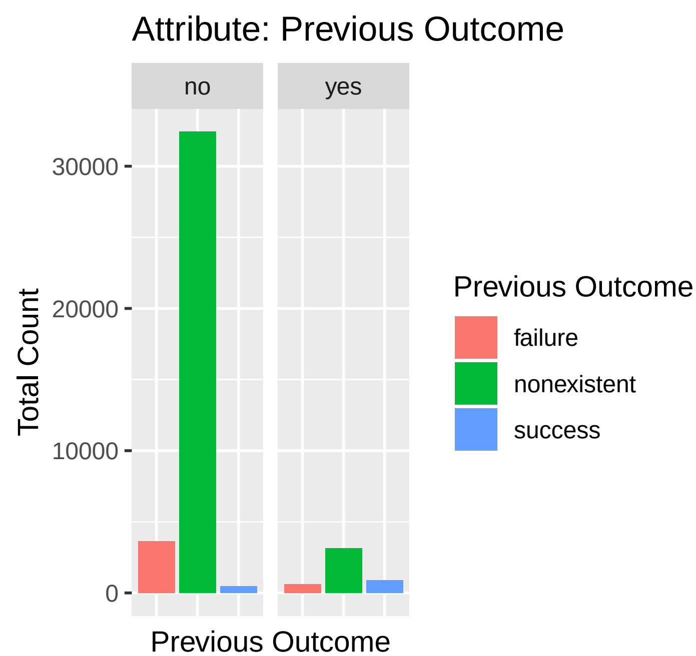

Attribute: Housing Loan
no
yes
Total Count
20000
15000
10000
5000
0
Housing Loan
Housing Loan
no
unknown
yes


Attribute: Personal Loan
no
yes
Total Count
30000
20000
10000
0
Personal Loan
Personal Loan
no
unknown
yes









In the charts above, we can see that the distributions are similar between the two responses. There are some slight differences between the two responses, but overall the disributions are close enough.

Data Cleaning

```
bank_data <- bank_data %>%
    mutate(response = ifelse(y=="yes", 1, 0)) %>%
    mutate(isDefault = ifelse(default == "yes", 1, 0)) %>%
    mutate(isHouseLoan = ifelse(housing == "yes", 1, 0)) %>%
    mutate(isLoan = ifelse(loan == "yes", 1, 0))
```

```
{
    set.seed(234234)
    x <- sample(nrow(bank_data), 0.75*nrow(bank_data), replace=F)
    train <- bank_data[x,]
    test <- bank_data[-x,]
}
```

Step-wise

Full Model

```
f <- "response ~ campaign + previous + age + marital + education + isDefault + isHouseLoan
+ isLoan + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + nr.employed"

m1 <- glm(f, data=train, family="binomial")
```

Backwards Stepwise

```
backward <- step(m1)
```

```
## Start:  AIC=20405.08
## response ~ campaign + previous + age + marital + education +
##     isDefault + isHouseLoan
##
##               Df Deviance   AIC
## - isDefault    1    20373 20403
## - isHouseLoan  1    20375 20405
## <none>              20373 20405
## - age          1    20431 20461
## - marital      3    20451 20477
## - education    7    20469 20487
## - campaign     1    20501 20531
## - previous     1    21317 21347
##
## Step:  AIC=20403.32
## response ~ campaign + previous + age + marital + education +
##     isHouseLoan
##
##               Df Deviance   AIC
## - isHouseLoan  1    20375 20403
## <none>              20373 20403
## - age          1    20432 20460
## - marital      3    20451 20475
## - education    7    20469 20485
## - campaign     1    20501 20529
## - previous     1    21317 21345
##
## Step:  AIC=20403.07
## response ~ campaign + previous + age + marital + education
##
##             Df Deviance   AIC
## <none>            20375 20403
```

```
## - age         1    20433 20459
## - marital     3    20453 20475
## - education   7    20471 20485
## - campaign    1    20503 20529
## - previous    1    21322 21348
```

```r
summary(backward)
```

```
##
## Call:
## glm(formula = response ~ campaign + previous + age + marital +
##     education, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7855  -0.4919  -0.4278  -0.3647   2.7185
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.96728    0.13367 -22.198  < 2e-16 ***
## campaign                      -0.10644    0.01068  -9.967  < 2e-16 ***
## previous                       0.86748    0.02817  30.793  < 2e-16 ***
## age                            0.01486    0.00193   7.701 1.35e-14 ***
## maritalmarried                 0.06138    0.06315   0.972 0.331003
## maritalsingle                  0.44712    0.07051   6.341 2.28e-10 ***
## maritalunknown                 0.17118    0.41812   0.409 0.682252
## educationbasic.6y             -0.09011    0.10811  -0.834 0.404559
## educationbasic.9y             -0.18078    0.08482  -2.131 0.033076 *
## educationhigh.school           0.10196    0.07522   1.355 0.175288
## educationilliterate            1.27433    0.60178   2.118 0.034208 *
## educationprofessional.course  0.20411    0.08113   2.516 0.011877 *
## educationuniversity.degree    0.33978    0.07115   4.776 1.79e-06 ***
## educationunknown               0.35916    0.10254   3.503 0.000461 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21803  on 30890  degrees of freedom
## Residual deviance: 20375  on 30877  degrees of freedom
## AIC: 20403
##
## Number of Fisher Scoring iterations: 5
```

```r
f <- "response ~ campaign + previous + age + marital + education + emp.var.rate
+ cons.price.idx + cons.conf.idx + nr.employed"
```

Forward Stepwise

```r
empty <- glm(response ~ 1, data=train, family="binomial")
forward <- step(empty, scope=list(lower=formula(empty),
                                  upper=formula(m1)),direction = "forward")
```

```
## Start:  AIC=21804.65
## response ~ 1
```

```
##
##               Df Deviance   AIC
## + previous     1    20713 20717
## + campaign     1    21610 21614
## + education    7    21658 21674
## + marital      3    21723 21731
## + age          1    21780 21784
## + isHouseLoan  1    21797 21801
## <none>              21803 21805
## + isDefault    1    21802 21806
##
## Step:  AIC=20716.92
## response ~ previous
##
##               Df Deviance   AIC
## + campaign     1    20587 20593
## + education    7    20600 20618
## + marital      3    20660 20670
## + age          1    20698 20704
## + isHouseLoan  1    20710 20716
## <none>              20713 20717
## + isDefault    1    20713 20719
##
## Step:  AIC=20592.93
## response ~ previous + campaign
##
##               Df Deviance   AIC
## + education    7    20472 20492
## + marital      3    20533 20545
## + age          1    20572 20580
## + isHouseLoan  1    20585 20593
## <none>              20587 20593
## + isDefault    1    20587 20595
##
## Step:  AIC=20491.86
## response ~ previous + campaign + education
##
##               Df Deviance   AIC
## + marital      3    20433 20459
## + age          1    20453 20475
## + isHouseLoan  1    20470 20492
## <none>              20472 20492
## + isDefault    1    20472 20494
##
## Step:  AIC=20459.4
## response ~ previous + campaign + education + marital
##
##               Df Deviance   AIC
## + age          1    20375 20403
## <none>              20433 20459
## + isHouseLoan  1    20432 20460
## + isDefault    1    20433 20461
##
## Step:  AIC=20403.07
```

```
## response ~ previous + campaign + education + marital + age
##
##                Df Deviance   AIC
## <none>               20375 20403
## + isHouseLoan  1     20373 20403
## + isDefault    1     20375 20405
```

As we can see, the stepwise function forward and backward resulted in the same set of explanatory variables. In the next code section, we will run this model and test its accuracy.

```r
m2 <- glm(f, data=train, family="binomial")

summary(m2)
```

```
##
## Call:
## glm(formula = f, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6684  -0.5132  -0.3190  -0.2728   2.8032
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.9170783  0.1357782 -21.484  < 2e-16 ***
## campaign                    -0.0610500  0.0107505  -5.679 1.36e-08 ***
## previous                     0.4196931  0.0282103  14.877  < 2e-16 ***
## age                          0.0104821  0.0018887   5.550 2.86e-08 ***
## maritalmarried               0.0555574  0.0649320   0.856 0.392205
## maritalsingle                0.2862781  0.0728982   3.927 8.60e-05 ***
## maritalunknown              -0.0005232  0.4275020  -0.001 0.999024
## educationbasic.6y           -0.0927052  0.1111165  -0.834 0.404108
## educationbasic.9y           -0.2075353  0.0874714  -2.373 0.017663 *
## educationhigh.school         0.0362807  0.0778548   0.466 0.641212
## educationilliterate          1.1706455  0.6386680   1.833 0.066810 .
## educationprofessional.course  0.1982120  0.0838795   2.363 0.018125 *
## educationuniversity.degree   0.2633155  0.0737678   3.570 0.000358 ***
## educationunknown             0.3350680  0.1061169   3.158 0.001591 **
## emp.var.rate                -0.4796350  0.0126351 -37.960  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21803  on 30890  degrees of freedom
## Residual deviance: 18883  on 30876  degrees of freedom
## AIC: 18913
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(m2)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
```

```
## Response: response
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                         30890      21803
## campaign      1   192.90     30889      21610
## previous      1  1022.82     30888      20587
## age           1    15.30     30887      20572
## marital       3   100.25     30884      20471
## education     7    96.31     30877      20375
## emp.var.rate  1  1492.33     30876      18883
```

```r
pred2 <- predict(m2, newdata = test, type = "response")

cutoff <- seq(from=0.001, to=.8, by=.001)
cutoff_pred <- as.data.frame(NULL)
count <- 1
for (i in cutoff) {
    pred2_check <- ifelse(pred2 >= i, 1, 0)
    correct <- ifelse(pred2_check==test$response, 1, 0)
    cutoff_pred[count,1] <- i
    cutoff_pred[count,2] <- sum(correct)/length(correct)
    count <- count + 1
}

cutoff_pred_sort <- cutoff_pred[order(cutoff_pred$V2, decreasing = TRUE),]

pred2_check <- ifelse(pred2 >= cutoff_pred_sort[1,1], 1, 0)
correct <- ifelse(pred2_check==test$response, 1, 0)
sum(correct)/length(correct)
```

```
## [1] 0.8915218
```

Conclusion

The stepwise model resulted in an accuracy of 89.4%. From the ANOVA table above, we can conclude that the 'previous' and 'emp.var.rate' have the biggest impact on the model, and the 'cons.price.inx' also has a significant impact but to a lesser extent. This result makes sense, because the 'previous' variable represents the number of times this client was previously contacted and the 'emp.var.rate' is a measure of the variation in employment rate. Both of these factors logically would have an impact on whether or not a client desires to subscribe to a term deposit.