Stat 36858 Project

Introduction

A term deposit is a deposit that a bank or a financial instituion offers with a fixed rate in which your money will be returned back at a specific maturity time. The dataset contains banking marketing campaign information and we can use it to optimize marketing campaigns to attract more customers for term deposit subscription. Goal of analysis is to identify factors which affect the campaign results in order to find out ways to make marketing campaigns more efficient.

```
# Set up
setwd("~/Documents/Fall 2019/STAT 36858")
library(tidyverse); library(VGAM)

## ── Attaching packages
─────────────────────────────────────────────────────────────────────
tidyverse 1.2.1 ──

## ✓ ggplot2 3.2.1     ✓ purrr   0.3.2
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   0.8.3     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0

## ── Conflicts
──────────────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

## Loading required package: stats4

## Loading required package: splines

##
## Attaching package: 'VGAM'

## The following object is masked from 'package:tidyr':
##
##     fill

# Data import
bank = read_csv("bank.csv")

## Parsed with column specification:
## cols(
##   age = col_double(),
##   job = col_character(),
```

```
##   marital = col_character(),
##   education = col_character(),
##   default = col_character(),
##   balance = col_double(),
##   housing = col_character(),
##   loan = col_character(),
##   contact = col_character(),
##   day = col_double(),
##   month = col_character(),
##   duration = col_double(),
##   campaign = col_double(),
##   pdays = col_double(),
##   previous = col_double(),
##   poutcome = col_character(),
##   deposit = col_character()
## )

# Trimming data
bank = bank %>% select(-pdays) %>% mutate(default = as.integer(default ==
"yes"), housing = as.integer(housing == "yes"), loan = as.integer(loan ==
"yes"), deposit = as.integer(deposit == "yes"))
```

The input variables of the data are as below.
a) age: age of the customer
b) job: type of job (one of 'admin', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', or 'unknown')
c) marital: marital status (one of 'divorced', 'married', 'single' or 'unknown') (note: 'divorced' includes being widowed)
d) education: education status (one of 'primary', 'secondary', 'tertiary' or 'unknown')
e) default: whether the customer has credit in default (one of 'yes', 'no' or 'unknown')
f) housing: wheter the customer has housing loan (one of 'yes', 'no' or 'unknown')
g) loan: wheter the customer has personal loan (one of 'yes', 'no' or 'unknown')
h) balance: balance of the individual
i) contact: contact communication type (one of 'cellular', 'telephone' or 'unknown')
j) month: last contact month of year (one of 'jan', 'feb', ... or 'dec')
k) day: last contact day of week (one of 'mon', 'tue', 'wed', 'thu' or 'fri')
l) duration: last contact duration
m) campaign: number of contacts performed during this campaign to this client
n) pdays: number of days that passed by after the client was last contacted from a previous campaign
o) previous: number of contacts performed before this campaign and for this client
p) deposit: whether the customer subscribed a term deposit (one of 'yes' or 'no')

Fortunately, it is known that there are no missing values in the dataset and thus we proceed with the given one. We set binary variable 'deposit' as our response variable since it is the desired target and only focus on age, balance, month and campaign as our explanatory variables.

Exploring the Basic Statistics

```
attach(bank)
mean(age); min(age); max(age)

## [1] 41.23195

## [1] 18

## [1] 95

mean(campaign); median(campaign)

## [1] 2.508421

## [1] 2

mean(deposit)

## [1] 0.4738398

detach(bank)
```

The mean age is approximately 41 years old. The youngest is 18 years old and the oldest is 95 years old. From now on, we categorize age into four categories including clients whose age is below 30, in between 30 and 50, in between 50 and 70 and above 70. Meanwhile, we also split twelve months into four categories by season.

```
# Factorizing 'age'
age1 = bank %>% rename(age_int = "age") %>% filter(age_int < 30) %>%
mutate(age = 1)
age2 = bank %>% rename(age_int = "age") %>% filter(30 <= age_int, age_int <
50) %>% mutate(age = 2)
age3 = bank %>% rename(age_int = "age") %>% filter(50 <= age_int, age_int <
70) %>% mutate(age = 3)
age4 = bank %>% rename(age_int = "age") %>% filter(age_int > 70) %>%
mutate(age = 4)
bank = rbind(age1, age2, age3, age4) %>% select(-age_int) %>% select(age,
everything())
bank$age = as.factor(bank$age)

# Factorizing 'month'
month1 = bank %>% rename(month_chr = "month") %>% filter(month_chr %in%
c("mar", "apr", "may")) %>% mutate(month = "spring")
month2 = bank %>% rename(month_chr = "month") %>% filter(month_chr %in%
c("jun", "jul", "aug")) %>% mutate(month = "summer")
month3 = bank %>% rename(month_chr = "month") %>% filter(month_chr %in%
c("sep", "oct", "nov")) %>% mutate(month = "fall")
month4 = bank %>% rename(month_chr = "month") %>% filter(month_chr %in%
c("dec", "jan", "feb")) %>% mutate(month = "winter")
bank = rbind(month1, month2, month3, month4) %>% select(-month_chr) %>%
select("age", "job", "marital", "education", "default", "balance", "housing",
```

```
"loan", "contact", "day", "month", everything())
bank$month = as.factor(bank$month)
```

Fitting Logistic Regression Model by Variable

a) Age

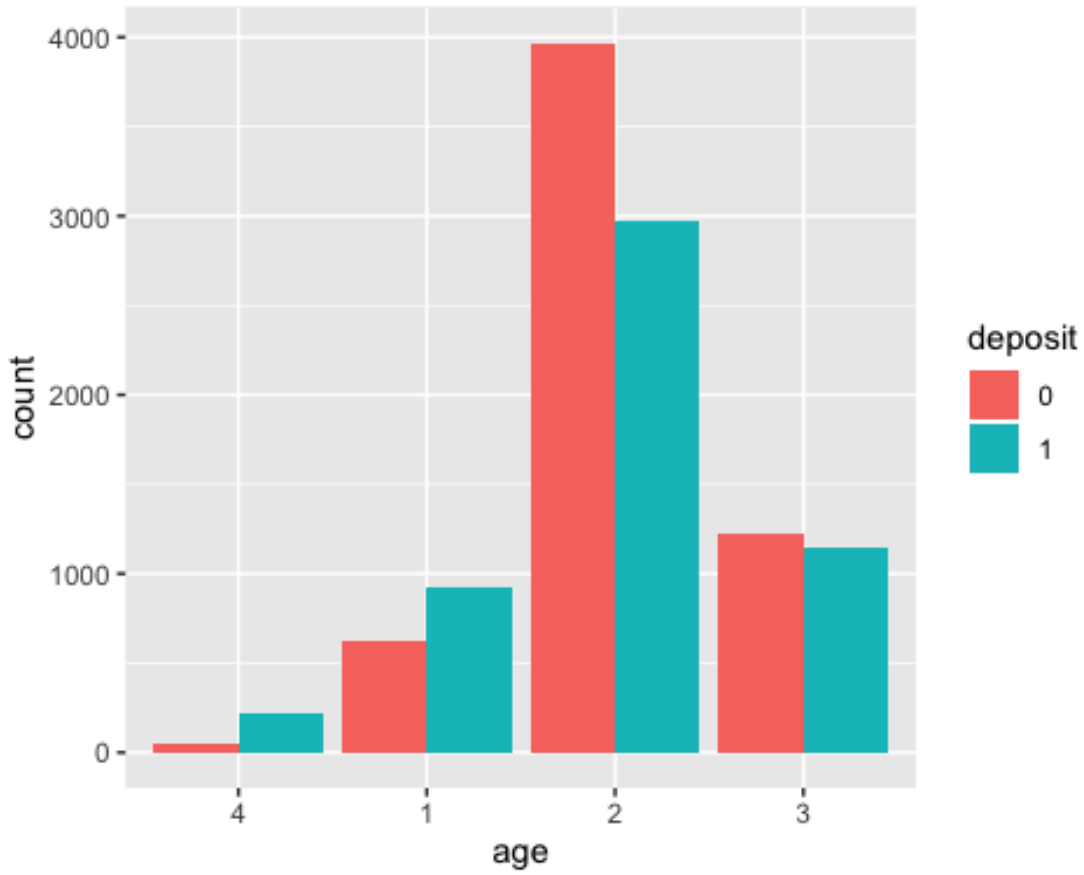
```
(bank_age = bank %>% group_by(age, deposit) %>% summarize(n = n()) %>%
spread(key = deposit, value = n) %>% rename(yes = '1', no = '0') %>%
select(age, yes, no) %>% replace_na(list(no = 0)))
```

```
## # A tibble: 4 x 3
## # Groups:   age [4]
##   age     yes    no
##   <fct> <int> <dbl>
## 1 1       928   623
## 2 2      2976  3970
## 3 3      1150  1219
## 4 4       218    55
```

```
bank$age = relevel(bank$age, ref = "4")
age_glm = glm(cbind(yes, no) ~ age, data = bank_age, family = binomial)
summary(age_glm)
```

```
##
## Call:
## glm(formula = cbind(yes, no) ~ age, family = binomial, data = bank_age)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.39849    0.05180   7.694 1.43e-14 ***
## age2        -0.68667    0.05719 -12.007  < 2e-16 ***
## age3        -0.45675    0.06613  -6.907 4.94e-12 ***
## age4         0.97868    0.15954   6.135 8.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.7745e+02  on 3  degrees of freedom
## Residual deviance: 3.4173e-13  on 0  degrees of freedom
## AIC: 38.881
##
## Number of Fisher Scoring iterations: 3
```

```
ggplot(bank, aes(age, fill = as.factor(deposit))) + geom_bar(position =
"dodge") + labs(fill = "deposit")
```



All estimates for group 1, 2 and 3 are negative meaning group 4 had the most deposit subscription. This can also be noticed from the bar chart since group 4 is the group with the biggest rate of subscription. It can be intuitively understood that the elderly are more likely to save money by term deposit after retiring. However, in bar chart we can also see group 1 having more subscribed count than unsubscribed count. This is also very natural since a lot of people in group 1 would be students.

b) Balance

```
bal_glm = glm(deposit ~ balance, data = bank, family = binomial)
summary(bal_glm)
```

```
##
## Call:
## glm(formula = deposit ~ balance, family = binomial, data = bank)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.768  -1.107  -1.091   1.240   1.343
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.966e-01  2.173e-02  -9.047   <2e-16 ***
```

```
## balance      6.008e-05  7.283e-06   8.250   <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: 15410  on 11138  degrees of freedom
## Residual deviance: 15329  on 11137  degrees of freedom
## AIC: 15333
##
## Number of Fisher Scoring iterations: 3
```

```
cor(bank$balance, bank$deposit)
```

```
## [1] 0.08120869
```

The estimate for balance is nearly zero. By looking at the correlation between balance and deposit, it also has a value near zero. We can conclude that balance does not matter on term deposit subscription.

c) Month

```
(bank_month = bank %>% group_by(month, deposit) %>% summarize(n = n()) %>%
spread(key = deposit, value = n) %>% rename(yes = '1', no = '0') %>%
select(month, yes, no))
```

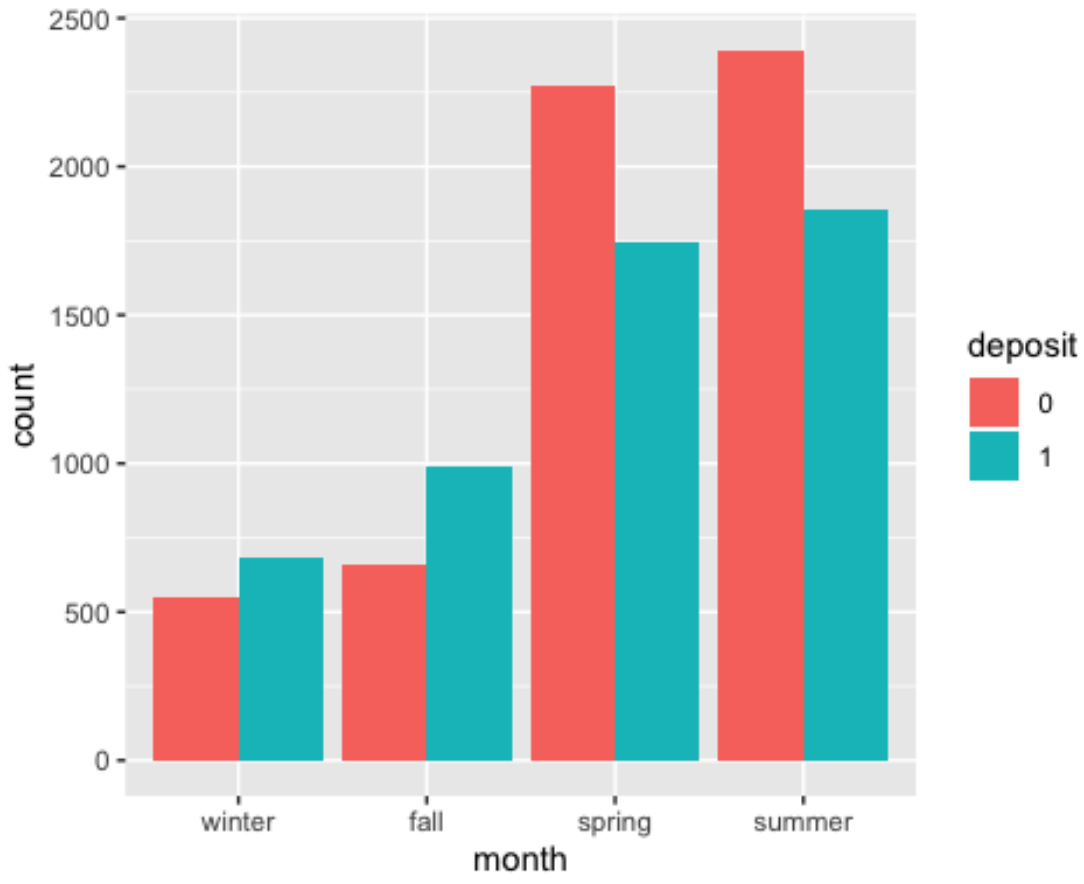
```
## # A tibble: 4 x 3
## # Groups:   month [4]
##   month    yes    no
##   <fct>  <int> <int>
## 1 fall     989   656
## 2 spring  1747  2270
## 3 summer  1856  2394
## 4 winter   680   547
```

```
bank$month = relevel(bank$month, ref = "winter")
month_glm = glm(cbind(yes, no) ~ month, data = bank_month, family = binomial)
summary(month_glm)
```

```
##
## Call:
## glm(formula = cbind(yes, no) ~ month, family = binomial, data =
bank_month)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.41053    0.05035   8.153 3.55e-16 ***
## monthspring -0.67241    0.05957 -11.288  < 2e-16 ***
## monthsummer -0.66508    0.05909 -11.255  < 2e-16 ***
```

```
## monthwinter -0.19289    0.07638  -2.525   0.0116 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.8710e+02  on 3  degrees of freedom
## Residual deviance: 9.2593e-13  on 0  degrees of freedom
## AIC: 40.891
##
## Number of Fisher Scoring iterations: 2
```

```
ggplot(bank, aes(month, fill = as.factor(deposit))) + geom_bar(position =
"dodge") + labs(fill = "deposit")
```



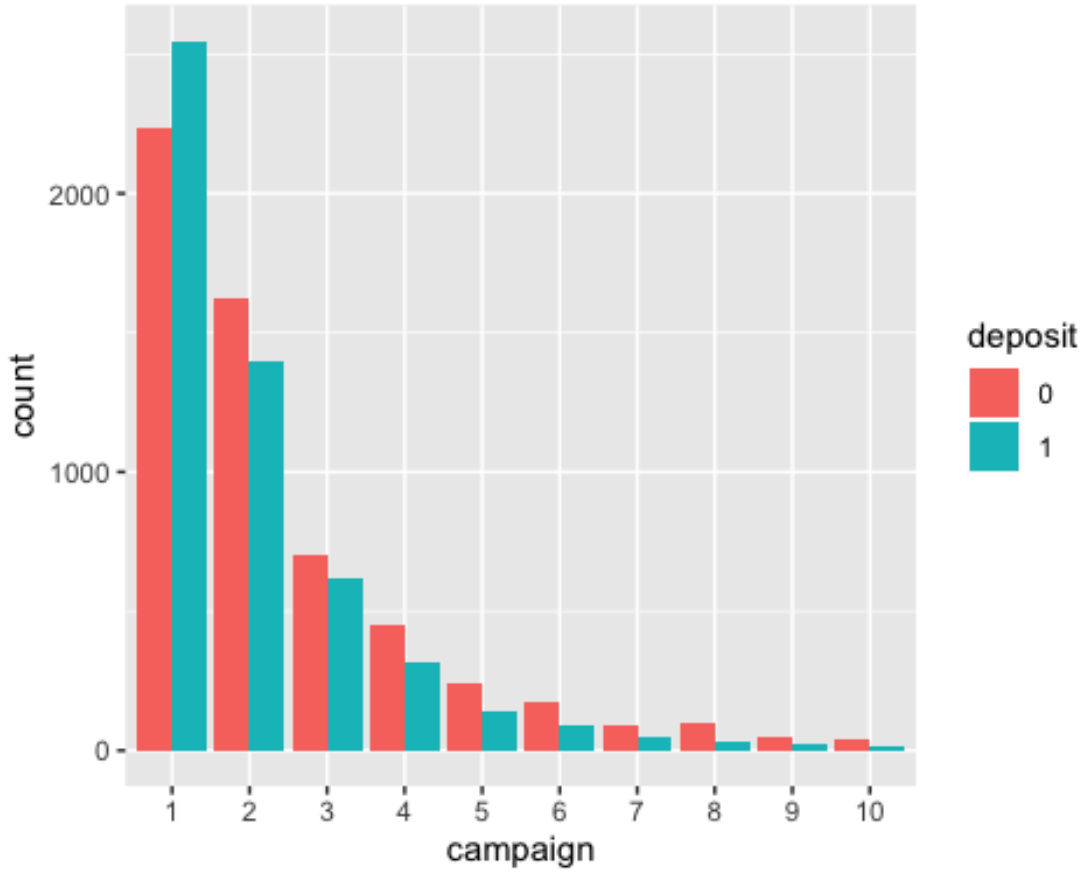
All estimates have small p-value and the estimates for fall and winter turn out to be positive. It can also be seen from the bar chart that it is more likely of customers to subscribe for a term deposit than to decline in fall and winter compared to spring and summer.

d) Campaign

```
camp_glm = glm(deposit ~ campaign, data = bank, family = binomial)
summary(camp_glm)
```

```
##
## Call:
## glm(formula = deposit ~ campaign, family = binomial, data = bank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2058  -1.1540  -0.8659   1.1493   2.7354
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.188659   0.028999   6.506 7.73e-11 ***
## campaign    -0.122058   0.009401 -12.983  < 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: 15410  on 11138  degrees of freedom
## Residual deviance: 15199  on 11137  degrees of freedom
## AIC: 15203
##
## Number of Fisher Scoring iterations: 4
```

```
bank %>% filter(campaign <= 10) %>% ggplot(aes(as.factor(campaign), fill =
as.factor(deposit))) + geom_bar(position = "dodge") + labs(x = "campaign",
fill = "deposit")
```



The p-value turns out to be very small and thus the variable can be said significant. The estimate is negative meaning the it has anti-effect on convincing customers to subscribe. However, looking at the bar chart we can roughly say up to 3 campaign calls is effective. If it exceeds the number then clients tend to decline the offer making it inadequate.

Fitting Multiple Logistic Regression Model

Assuming there are no interactions between any two variables, we fit the model using all dependent variables only with main effects.

```
bank_full = glm(deposit ~ age + balance + month + campaign, data = bank,
family = binomial)
summary(bank_full)

##
## Call:
## glm(formula = deposit ~ age + balance + month + campaign, family =
binomial,
##     data = bank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.6190  -1.0714  -0.8224   1.1874   2.6582
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.579e+00  1.624e-01   9.721  < 2e-16 ***
## age1        -7.479e-01  1.621e-01  -4.613 3.98e-06 ***
## age2        -1.442e+00  1.552e-01  -9.289  < 2e-16 ***
## age3        -1.264e+00  1.587e-01  -7.962 1.69e-15 ***
## balance      5.185e-05  7.307e-06   7.095 1.29e-12 ***
## monthfall    1.763e-01  7.820e-02   2.254   0.0242 *
## monthspring -4.111e-01  6.718e-02  -6.119 9.40e-10 ***
## monthsummer -2.764e-01  6.731e-02  -4.106 4.03e-05 ***
## campaign    -1.072e-01  9.575e-03 -11.200  < 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: 15410  on 11138  degrees of freedom
## Residual deviance: 14762  on 11130  degrees of freedom
## AIC: 14780
##
## Number of Fisher Scoring iterations: 4
```

We process both backward and forward elimination to select a logistic regression model.

```
step(bank_full, direction = "backward")
```

```
## Start:  AIC=14780.42
## deposit ~ age + balance + month + campaign
##
##            Df Deviance   AIC
## <none>           14762 14780
## - balance   1    14820 14836
## - month     3    14873 14885
## - campaign  1    14915 14931
## - age       3    14995 15007

##
## Call:  glm(formula = deposit ~ age + balance + month + campaign, family =
binomial,
##     data = bank)
##
## Coefficients:
## (Intercept)         age1         age2         age3      balance
##   1.579e+00   -7.479e-01   -1.442e+00   -1.264e+00    5.184e-05
##   monthfall  monthspring  monthsummer     campaign
##   1.763e-01   -4.111e-01   -2.764e-01   -1.072e-01
##
## Degrees of Freedom: 11138 Total (i.e. Null);  11130 Residual
```

```
## Null Deviance:       15410
## Residual Deviance: 14760     AIC: 14780
```

```
step(bank_full, direction = "forward")
```

```
## Start:  AIC=14780.42
## deposit ~ age + balance + month + campaign

##
## Call:  glm(formula = deposit ~ age + balance + month + campaign, family =
binomial,
##     data = bank)
##
## Coefficients:
## (Intercept)         age1         age2         age3      balance
##   1.579e+00   -7.479e-01   -1.442e+00   -1.264e+00    5.184e-05
##   monthfall  monthspring  monthsummer     campaign
##   1.763e-01   -4.111e-01   -2.764e-01   -1.072e-01
##
## Degrees of Freedom: 11138 Total (i.e. Null);  11130 Residual
## Null Deviance:       15410
## Residual Deviance: 14760     AIC: 14780
```

Both results from the two steps are identical and happen to be the full model. Yet, the estimate for balance 5.185e-05 (0.00005185) is almost zero and hence balance would not give a big difference in deposit variable.

Conclusion

The next marketing campaign of the bank should target potential clients in their 20s or younger and 70s or older. It will be great if for the next campaign the bank addressed these two categories and therefore, increase the likelihood of more term deposits suscriptions. In addition, customers opted to suscribe term deposits during the seasons of fall and winter. The next marketing campaign should focus its activity throughout these seasons. Finally, a policy should be implemented that states that no more than 3 calls should be applied to the same potential customer in order to save time and effort in getting new potential clients. Too much calls would actually result in people to decline opening a term deposit.