Machine Learning

Bach, Helena * Fleming, María† Karampelas, Petro‡ Romero, Pablo José§

30/12/2021

Introduction

In this document the following R packages are made use of:

```
library(corrplot)
library(caTools)
library(biotools)
library(ggplot2)
library(MASS)
library(xtable)
library(dplyr)
library(tidyr)
library(caret)
library(kableExtra)
```

Descriptive Analysis

The data set used in this project collects information about a Portuguese bank marketing campaign aiming to get customers to subscribe to a term deposit. The data set contains 45.211 observations and 17 different variables: 11 categorical and 6 numerical.

Firs of all, the data will be uploaded, the variables classified, the train and test sample created and a exploratory analysis will be done.

```
bank <- read.csv(file="bank-full.csv", header=TRUE, sep=";")
summary(bank)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.94
##  3rd Qu.:48.00
##  Max.   :95.00
```

*Università di Bologna, xxx
†Università di Bologna, xxx
‡Università di Bologna, xxx
§Universidad Nacional de Córdoba, xxx

```
##    default            balance         housing             loan
##  Length:45211       Min.   : -8019   Length:45211       Length:45211
##  Class :character   1st Qu.:    72   Class :character   Class :character
##  Mode  :character   Median :   448   Mode  :character   Mode  :character
##                     Mean   :  1362
##                     3rd Qu.:  1428
##                     Max.   :102127
##    contact               day           month              duration
##  Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0
##  Mode  :character   Median :16.00   Mode  :character   Median : 180.0
##                     Mean   :15.81                      Mean   : 258.2
##                     3rd Qu.:21.00                      3rd Qu.: 319.0
##                     Max.   :31.00                      Max.   :4918.0
##     campaign          pdays          previous          poutcome
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   Length:45211
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   Class :character
##  Median : 2.000   Median : -1.0   Median :  0.0000   Mode  :character
##  Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
##  3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :63.000   Max.   :871.0   Max.   :275.0000
##       y
##  Length:45211
##  Class :character
##  Mode  :character
##
##
##
```

```
str(bank)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
##  $ marital  : chr  "married" "single" "married" "married" ...
##  $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
##  $ default  : chr  "no" "no" "no" "no" ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : chr  "yes" "yes" "yes" "yes" ...
##  $ loan     : chr  "no" "no" "yes" "no" ...
##  $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : chr  "may" "may" "may" "may" ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ y        : chr  "no" "no" "no" "no" ...
```

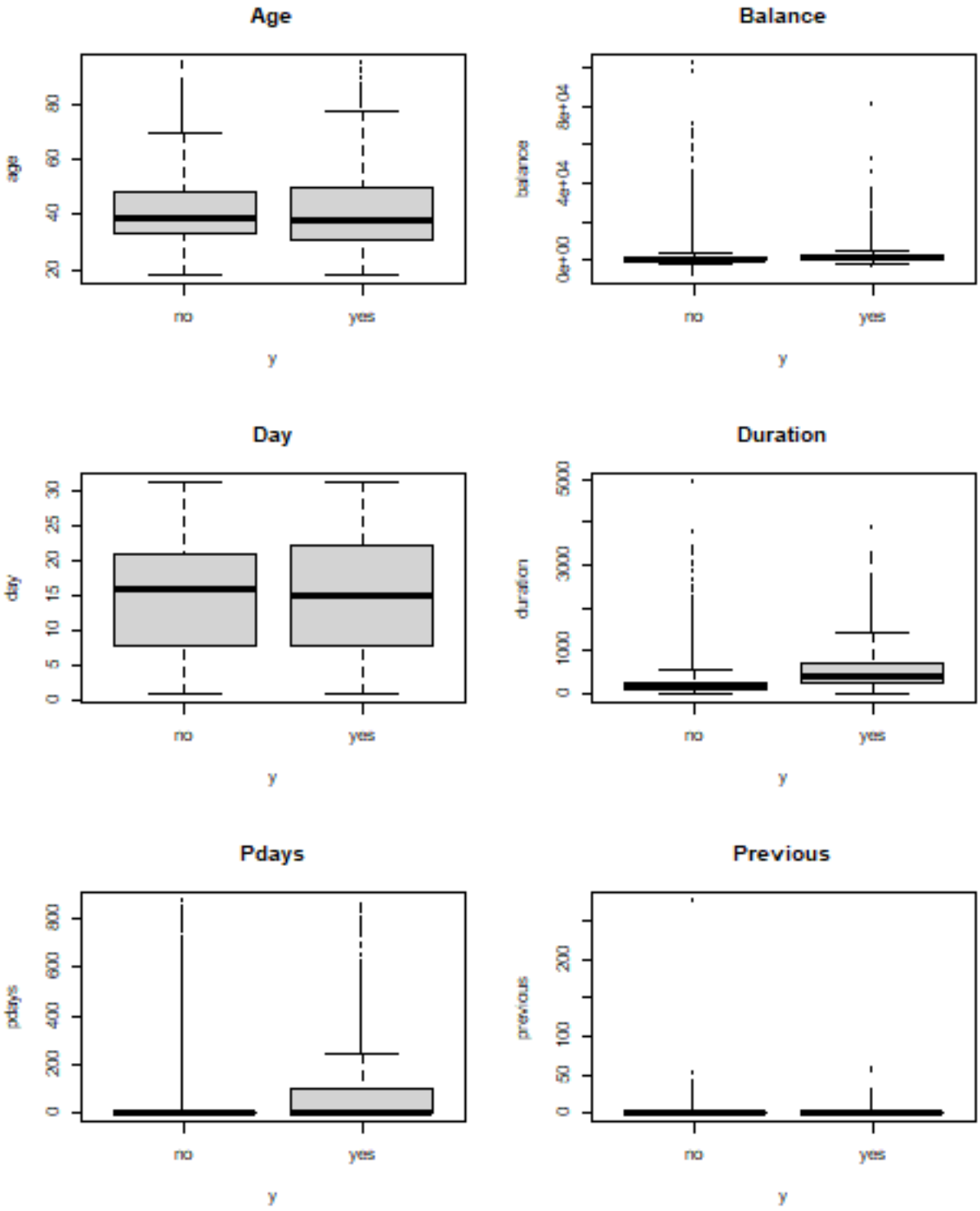
```
#variables that should be treated as categorical

bank$marital <- as.factor(bank$marital)
```

```
bank$education <- as.factor(bank$education)
bank$default <- as.factor(bank$default)
bank$housing <- as.factor(bank$housing)
bank$job <- as.factor(bank$job)
bank$y <- as.factor(bank$y)

#create test and training sample

set.seed(101)
sample = sample.split(bank[,1], SplitRatio = .75)
train <- bank[sample, ]
test <- bank[-sample, ]
```

Figure 1 shows the box plot of the numerical variables. It can be noticed that *age* follows a very similar distribution for clients who subscribed to a term deposit and for those who did not, indicating that the age of the customer won't play a significant role when predicting the behavior of clients. On the other hand, from the box-plot of *duration* it can be inferred that high duration (long contact duration in second) might have a positive impact on the subscription rate.

```
#########DESCRIPTIVE ANALYSIS#########

par(mfrow=c(3,2))
boxplot(age ~ y, data = bank, main="Age")
boxplot(balance ~ y, data = bank, main="Balance")
boxplot(day ~ y, data = bank, main="Day")
boxplot(duration ~ y, data = bank, main="Duration")
boxplot(pdays ~ y, data = bank, main="Pdays")
boxplot(previous ~ y, data = bank, main="Previous")
```

Figure 1: Boxplot



Logit

The logistic regression models the probability of belonging to one class, under the assumption that the dependent variable follows a binomial distribution. It is a classification technique that allows to predict a dichotomous variable.

Given X (explanatory variables) we can represent the probability that the client has subscribed to a term deposit (Y) as $p(X) = P(y = yes|X)$, using the logistic function to ensure that the output lies between 0 and 1.

$$p(X) = \frac{e^{\beta_0+\beta_1 \cdot X}}{1+e^{\beta_0+\beta_1 \cdot X}}$$

```
mylogit <-  glm(y ~., data = train,
                family = "binomial")

tabl <- summary(mylogit)

kable(tabl$coefficients,
      caption = "\\label{fig:logit}Logit summary",
      format ="latex",
      align = "c",
      table.envir = "figure")
```

From Table 2 (included in the annex) it can be inferred which variables are significant when determining the outcome of the term deposit subscription. Working as a housemaid, entrepreneur, manual labor or being self-employed has a significant negative effect on the probability of subscribing to the term deposit, while being retired or a student increases the probability. In terms of education, completing a secondary or tertiary level of education, increases the probability of subscribing. Clients who are married, have a housing or personal loan are less likely to subscribe to the term deposit. Accounts with a large average yearly balance have higher odds of accepting the marketing offer. In line with Figure 1, duration has a positive significant effect, so clients that were contacted for a longer time (seconds) are more likely to subscribe.

To evaluate the performance of the logistic model defined previously, we create a confusion matrix. The confusion matrix compares the predicted values obtained from the model to the actual values.

```
probabilities <- mylogit %>% predict(test, type = "response")
test$predicted <- ifelse(probabilities > 0.5, 'yes', 'no')

#Confusion matrix
cmatrix <- confusionMatrix(table(test$y, test$predicted))

kable(cmatrix$table,
      caption = "\\label{fig:confmat}Confusion Matrix",
      format ="latex",
      align = "c",
      table.envir = "figure")
```

Figure 2: Logit summary

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.3582949	0.2094553	-11.2591801	0.0000000
age	-0.0017381	0.0025344	-0.6857833	0.4928498
jobblue-collar	-0.2387594	0.0836332	-2.8548396	0.0043059
jobentrepreneur	-0.3649721	0.1456083	-2.5065335	0.0121921
jobhousemaid	-0.4067260	0.1546476	-2.6300182	0.0085380
jobmanagement	-0.1218363	0.0846113	-1.4399522	0.1498809
jobretired	0.2846790	0.1125541	2.5292632	0.0114302
jobself-employed	-0.3105765	0.1306347	-2.3774419	0.0174332
jobservices	-0.1555906	0.0965301	-1.6118352	0.1069978
jobstudent	0.3568839	0.1260357	2.8316098	0.0046314
jobtechnician	-0.1191150	0.0793501	-1.5011328	0.1333212
jobunemployed	-0.0907000	0.1263874	-0.7176347	0.4729825
jobunknown	-0.0974263	0.2531100	-0.3849168	0.7002991
maritalmarried	-0.2080183	0.0676321	-3.0757341	0.0020998
maritalsingle	0.0529894	0.0772368	0.6860648	0.4926722
educationsecondary	0.1690260	0.0741658	2.2790292	0.0226653
educationtertiary	0.3826250	0.0861538	4.4411864	0.0000089
educationunknown	0.2088004	0.1194621	1.7478383	0.0804920
defaultyes	-0.1552601	0.1945568	-0.7980191	0.4248594
balance	0.0000144	0.0000057	2.5238903	0.0116064
housingyes	-0.7003516	0.0504801	-13.8738171	0.0000000
loanyes	-0.4120969	0.0688614	-5.9844443	0.0000000
contacttelephone	-0.1030498	0.0849289	-1.2133653	0.2249902
contactunknown	-1.6796573	0.0845413	-19.8678808	0.0000000
day	0.0089008	0.0028706	3.1006521	0.0019309
monthaug	-0.7936763	0.0910539	-8.7165533	0.0000000
monthdec	0.6795486	0.2008848	3.3827780	0.0007176
monthfeb	-0.1524005	0.1020823	-1.4929183	0.1354586
monthjan	-1.2886923	0.1392258	-9.2561333	0.0000000
monthjul	-0.8746982	0.0888761	-9.8417738	0.0000000
monthjun	0.5227409	0.1069242	4.8888925	0.0000010
monthmar	1.5856344	0.1375100	11.5310511	0.0000000
monthmay	-0.4261599	0.0826435	-5.1566020	0.0000003
monthnov	-0.8833143	0.0967266	-9.1320730	0.0000000
monthoct	0.8607510	0.1233629	6.9773902	0.0000000
monthsep	0.9448199	0.1358427	6.9552481	0.0000000
duration	0.0041583	0.0000745	55.8342222	0.0000000
campaign	-0.0830747	0.0114597	-7.2492867	0.0000000
pdays	-0.0002593	0.0003524	-0.7358027	0.4618508
previous	0.0110347	0.0070987	1.5544682	0.1200728
poutcomeother	0.1581791	0.1037001	1.5253514	0.1271715
poutcomesuccess	2.2450502	0.0943497	23.7949916	0.0000000
poutcomeunknown	-0.1402661	0.1063776	-1.3185685	0.1873134

Figure 3: Confusion Matrix

	no	yes
no	38937	984
yes	3450	1839