

علي جمال برغوث 201910317

هبة وائل عوض 201912014

شذى باسل رداد 201910532

Introduction

- Bank Marketing Data Set, This data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

- the tool used is (R).

Data set

- link for the dataset :

<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

- about dataset :

Number of Instances: 45211

Number of Attributes: 16 + output attribute

- Attribute Information:

Input variables:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical:

'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')

6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')

7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

8 - contact: contact communication type (categorical: 'cellular', 'telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is

obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

- attached needed library :

library(dplyr)

library(Hmisc)

library(e1071)

Problem definition

Determine the categories of customers who can deposit in the bank

Data preparation .

cleaning dataset :

Step 1: Reading dataset from CSV file.

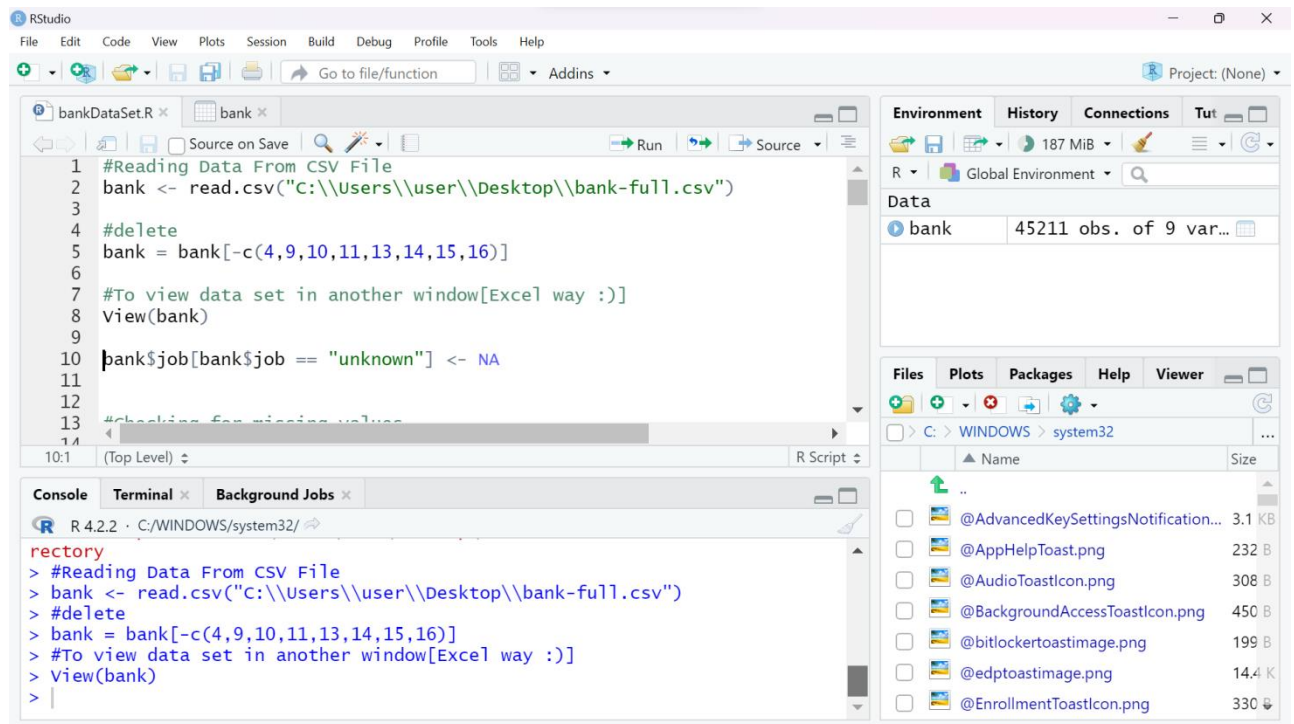
Step 2: Delete this columns (education, day and month), because they are not useful for this analysis.

And delete this columns (pdays, previous, poutcome), because they are duplicated.

And delete contact column because it is not important and contain Missing value.

By using dplyr library

Step 3: view data set in another window[Excel way :)]



The screenshot displays the RStudio environment. The main editor window shows an R script with the following code:

```
1 #Reading Data From CSV File
2 bank <- read.csv("C:\\Users\\user\\Desktop\\bank-full.csv")
3
4 #delete
5 bank = bank[-c(4,9,10,11,13,14,15,16)]
6
7 #To view data set in another window[Excel way :)]
8 view(bank)
9
10 bank$job[bank$job == "unknown"] <- NA
11
12
13 #checking for missing values
14
```

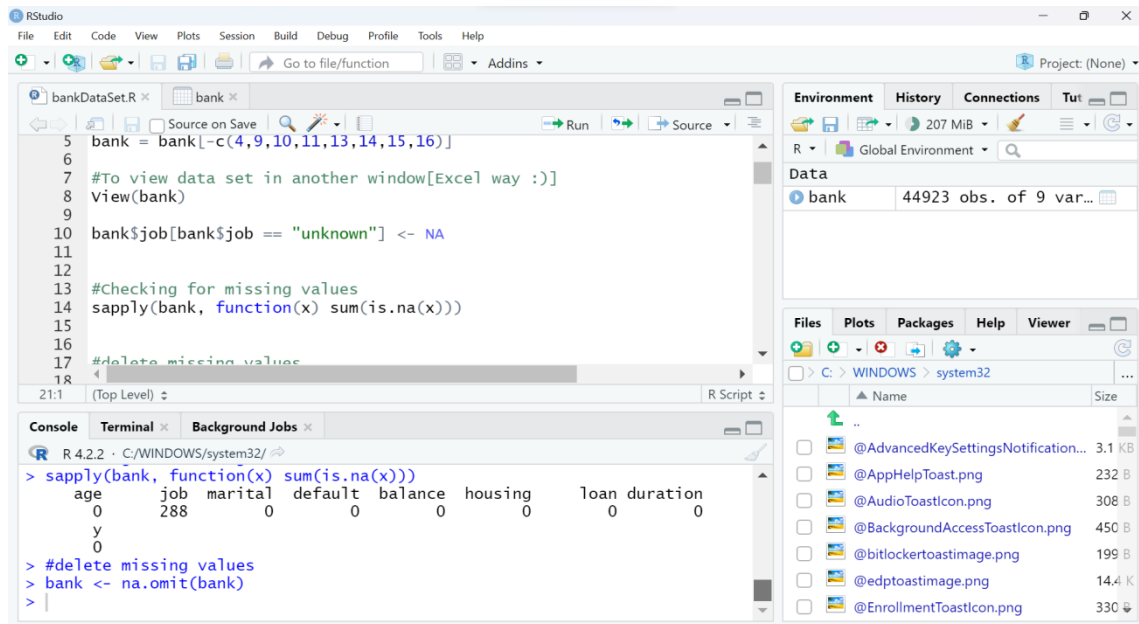
The console window at the bottom shows the execution of the script:

```
R 4.2.2 > C:/WINDOWS/system32/
rectory
> #Reading Data From CSV File
> bank <- read.csv("C:\\Users\\user\\Desktop\\bank-full.csv")
> #delete
> bank = bank[-c(4,9,10,11,13,14,15,16)]
> #To view data set in another window[Excel way :)]
> view(bank)
>
```

The Environment pane on the right shows the Global Environment with 187 MiB of memory used. The Data pane shows the 'bank' object with 45211 observations and 9 variables. The Files pane shows the current directory structure, including system32 and various system files.

Step 4: Convert unknown values to (NA).

Step 5: checking the NA values and delete missing values



Normalization

Step 1: Factoring columns

(default ,housing ,loan ,marital ,job) variable

Step 2: Removing Outliers

(Removing values higher than $q3 + 1.5 \cdot iqr$)

(Removing values lower than $q1 - 1.5 \cdot iqr$)

Tack out outliers in duration column .

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

bankDataSet.R* x bank x

Source on Save Run Source

```

23
24 #Writing Data to CSV File
25 write.csv(bank,"C:\\Users\\user\\Desktop\\bank-withoutMissingValues
26
27 install.packages("e1071")
28 library(e1071)
29
30 #Factoring default,housing,loan,marital,job variable
31 bank$default <- as.factor(bank$default)
32 bank$housing <- as.factor(bank$housing)
33 bank$loan <- as.factor(bank$loan)
34 bank$marital <- as.factor(bank$marital)
35 bank$job <- as.factor(bank$job)
36
37 #Removing Outliers
38 ##Calculating IQR
39 ###Summary gives 6 values : min, q1, median, mean, q3, max
40 summary = summary(bank$duration)
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

30:44 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · C:/WINDOWS/system32/

Environment History Connections Tutori

R Global Environment

Data

Variable	Value
bank	41706 obs. of 9 vari...
nb	List of 5
testingD...	12512 obs. of 9 vari...
training...	29194 obs. of 9 vari...

Values

Variable	Value
accuracy	90.8807544757033
iqr	216
misclass...	0.0911924552429667
predicti...	'table' int [1:2, 1:2]...
q1	103
q3	319
s	int [1:29194] 28448 14...

Files Plots Packages Help Viewer

C > WINDOWS > system32

Name Size

@AdvancedKeySettingsNotification... 3.1 KB

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

bankDataSet.R* x bank x

Source on Save Run Source

```

37 #Removing Outliers
38 ##Calculating IQR
39 ###Summary gives 6 values : min, q1, median, mean, q3, max
40 summary = summary(bank$duration)
41
42 ####Save q1 and q3 as numeric values
43 q1 = as.numeric(summary(bank$duration)[2])
44 q3 = as.numeric(summary(bank$duration)[5])
45
46 ##Find IQR
47 iqr <- q3 - q1
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

49:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · C:/WINDOWS/system32/

```

> ###Summary gives 6 values : min, q1, median, mean, q3, max
> summary = summary(bank$duration)
> ####Save q1 and q3 as numeric values
> q1 = as.numeric(summary(bank$duration)[2])
> q3 = as.numeric(summary(bank$duration)[5])
> ##Find IQR
> iqr <- q3 - q1
>

```

Environment History Connections Tut

R Global Environment

Data

Variable	Value
bank	44923 obs. of 9 var...

Values

Variable	Value
iqr	216
q1	103
q3	319

Files Plots Packages Help Viewer

C > WINDOWS > system32

Name Size

@AdvancedKeySettingsNotification... 3.1 KB

@AppHelpToast.png 232 B

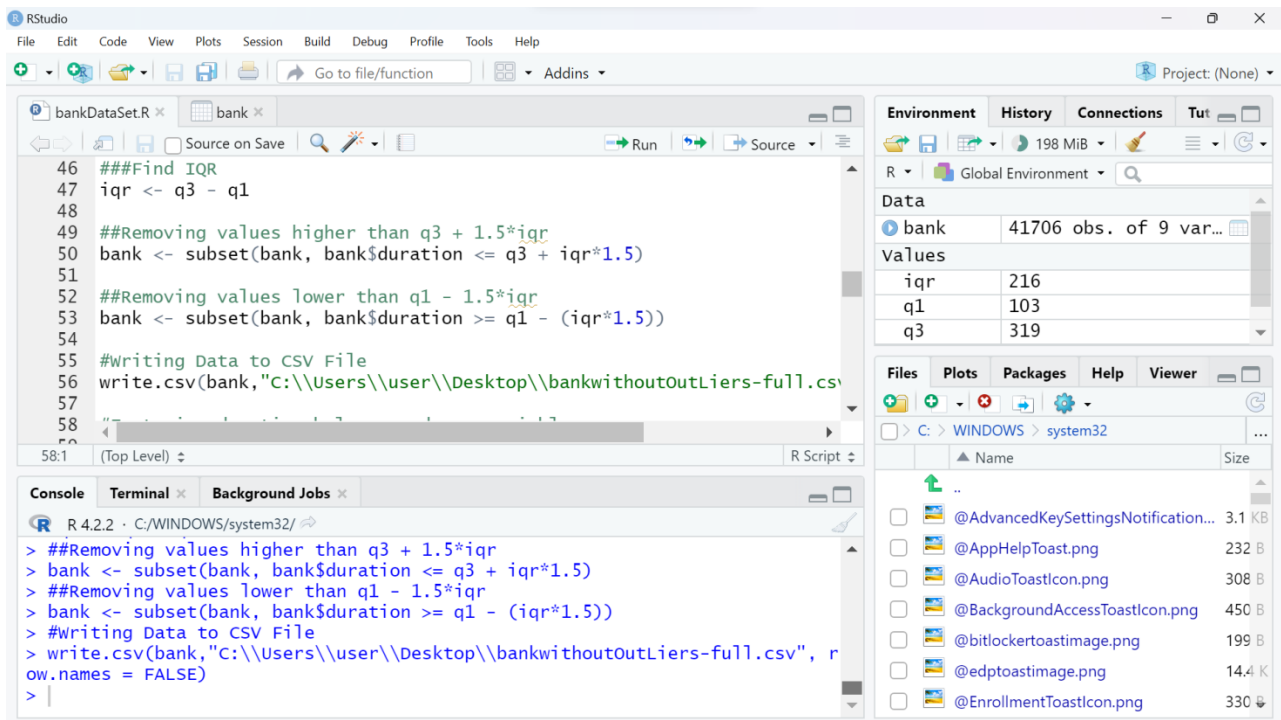
@AudioToastIcon.png 308 B

@BackgroundAccessToastIcon.png 450 B

@bitlockertoastimage.png 199 B

@edptoastimage.png 14.4 K

@EnrollmentToastIcon.png 330 B



Step 3: convert data type of this columns (duration ,balance and age) from numeric to ordinal

by dividing the range of the numerical variable into bins and assigning values to each bin.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections

R 209 MiB

Global Environment

Data

bank 41706 obs. of 10 va...

Values

iqr	216
q1	103
q3	319

Files Plots Packages Help Viewer

C:\WINDOWS\system32

Name Size

- @AdvancedKeySettingsNotification... 3.1 KB
- @AppHelpToast.png 232 B
- @AudioToastIcon.png 308 B
- @BackgroundAccessToastIcon.png 450 B
- @bitlockertoastimage.png 199 B
- @edpttoastimage.png 14.4 K
- @EnrollmentToastIcon.png 330 B

```

56 write.csv(bank,"C:\\Users\\user\\Desktop\\bankwithoutOutLiers-full.csv")
57
58 #Factoring duration,balance and age variables
59 #0 - 200 = Short, 200-400 = Normal, <400 = Long
60 ##duration
61 bank$duration2[bank$duration < 200] <- "Short"
62 bank$duration2[bank$duration >=200 & bank$duration < 400] <- "Normal"
63 bank$duration2[bank$duration >= 400] <- "Long"
64 bank$duration <- bank$duration2
65 bank$duration <- as.factor(bank$duration)
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80

```

Console Terminal Background Jobs

R 4.2.2 · C:/WINDOWS/system32/

```

> #0 - 200 = Short, 200-400 = Normal, <400 = Long
> ##duration
> bank$duration2[bank$duration < 200] <- "Short"
> bank$duration2[bank$duration >=200 & bank$duration < 400] <- "Normal"
> bank$duration2[bank$duration >= 400] <- "Long"
> bank$duration <- bank$duration2
> bank$duration <- as.factor(bank$duration)
>

```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections

R 223 MiB

Global Environment

Data

bank 41706 obs. of 11 va...

\$ age	: int	58 44 33 4...
\$ job	: Factor w/ 11 le...	
\$ marital	: Factor w/ 3 lev...	
\$ default	: Factor w/ 2 lev...	

Files Plots Packages Help Viewer

C:\WINDOWS\system32

Name Size

- @AdvancedKeySettingsNotification... 3.1 KB
- @AppHelpToast.png 232 B
- @AudioToastIcon.png 308 B
- @BackgroundAccessToastIcon.png 450 B
- @bitlockertoastimage.png 199 B
- @edpttoastimage.png 14.4 K
- @EnrollmentToastIcon.png 330 B

```

68 #>10000 = Low, 10000-60000 = Normal, <60000 = High
69 ##balance
70 bank$balance2[bank$balance < 10000] <- "Low"
71 bank$balance2[bank$balance >=10000 & bank$balance < 60000] <- "Normal"
72 bank$balance2[bank$balance >= 60000] <- "High"
73 bank$balance <- bank$balance2
74 bank$balance <- as.factor(bank$balance)
75
76 #>25 = Young, 25-60 = Mature, <60 = Old
77 ##age
78 bank$age2[bank$age < 25] <- "Young"
79 bank$age2[bank$age >=25 & bank$age < 60] <- "Mature"
80

```

Console Terminal Background Jobs

R 4.2.2 · C:/WINDOWS/system32/

```

> #>10000 = Low, 10000-60000 = Normal, <60000 = High
> ##balance
> bank$balance2[bank$balance < 10000] <- "Low"
> bank$balance2[bank$balance >=10000 & bank$balance < 60000] <- "Normal"
> bank$balance2[bank$balance >= 60000] <- "High"
> bank$balance <- bank$balance2
> bank$balance <- as.factor(bank$balance)
>

```

Studio

Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

bankDataSet.R x bank x

Source on Save Run Source

```
77 ##age
78 bank$age2[bank$age < 25] <- "Young"
79 bank$age2[bank$age >=25 & bank$age < 60] <- "Mature"
80 bank$age2[bank$age >= 60] <- "old"
81 bank$age <- bank$age2
82 bank$age <- as.factor(bank$age)
83
84
85 ##delete the temporary age2,duratin2,balance2
86 #delete
87 bank = bank[-c(10,11,12)]
88
89
```

85:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 C:/WINDOWS/system32/

```
#>25 = Young, 25-60 = Mature, <60 = old
##age
bank$age2[bank$age < 25] <- "Young"
bank$age2[bank$age >=25 & bank$age < 60] <- "Mature"
bank$age2[bank$age >= 60] <- "old"
bank$age <- bank$age2
bank$age <- as.factor(bank$age)
```

Environment History Connections Tut

R Global Environment 194 MiB

Data

bank 41706 obs. of 12 variables

\$ age	: Factor w/ 3 levels
\$ job	: Factor w/ 11 levels
\$ marital	: Factor w/ 3 levels
\$ default	: Factor w/ 2 levels
\$ balance	: Factor w/ 3 levels
\$ housing	: Factor w/ 2 levels
\$ loan	: Factor w/ 2 levels
\$ duration	: Factor w/ 3 levels
\$ y	: chr "no" "no" ...
\$ duration2	: chr "Normal" "..."
\$ balance2	: chr "Low" "Low..."
\$ age2	: chr "Mature" "..."

Values

iar 216

Files Plots Packages Help Viewer

C:/WINDOWS/system32

Name Size

data mining model

classification:

we used the Naive Bayes classifier, which is quite faster in comparison to other classification algorithms.

The classification goal is to predict if the client will subscribe a term deposit (variable y).

Results:

The accuracy is more than 80%, so we can accept the results.

The screenshot displays the RStudio interface. The script editor on the left contains R code for data sampling, model building, and prediction. The Environment pane on the right shows the objects created, including 'bank', 'nb', 'testingData', and 'trainingData', along with their sizes and data types. The Console at the bottom shows the R version and the current directory.

```
87 bank = bank[-c(10,11,12)]
88
89
90 #Random sampling data -> 70% for training and rest for testing
91 s <- sample(nrow(bank), nrow(bank)*0.7)
92 trainingData <- bank[s,]
93 testingData <- bank[-s,]
94
95
96 #Building Naive Bayes Model
97 nb <- naiveBayes(y~., bank, laplace = 1)
98 #Prediction
99 predictionTable <- table(predict(nb, testingData),testingData$y)
100 ##Confusion Matrix
101 misclassification = 1 - sum(diag(predictionTable)) / sum(predictionTable)
102 ##Accuracy
103 accuracy = (1-misclassification)*100
104 ##Accuracy is nearly 90%.
105
106
```

Environment

Object	Size
bank	41706 obs. of 9 vari...
nb	List of 5
testingD...	12512 obs. of 9 vari...
training...	29194 obs. of 9 vari...

Values

Variable	Value
accuracy	90.8807544757033
iqr	216
misclass...	0.0911924552429667
predicti...	'table' int [1:2, 1:2]...
q1	103
q3	319
s	int [1:29194] 28448 14...

Files | **Plots** | **Packages** | **Help** | **Viewer** | **Project**

C: > WINDOWS > system32

..

@AdvancedKeySettingsNotification... 3.1 KB