Report I

Anna Szymanek (230042), Patryk Wielopolski (234891)

Introduction

In this report we will be analysing the data [1] related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

This specific problem is extremly important for call centers to manage their limited human resoruces. In order to make such a campaign succesful we have to point out people who are most likely to subscribe bank term deposit.

Our goal of this project is to conduct data analysis to better understand available data for modeling and then create classifier which will be correctly assigning classes.

Structer of this document is as follows. In Section 2 we will describe methods used to perform whole modeling process. In the next section we will analyse our results and sum up whole project in the last section.

Methods

This section is divided to 3 parts - data description, data analysis, classification.

In first subsection we will describe dataset from official documentation which will give us general overview about dataset and information about variables meaning and types. Also we might there spot some interesting facts which could be interesting for us from modeling perspective, i.e. variable coding or they might give us real-world application context.

Then in the next subsection we will conduct statistical analysis of given dataset. After this part we would like to know basic properties of variables / features (range, properties, distribution), find all missing values and outliers, be familiar with correlations between features in the dataset and give initial assessment of discriminative ability of consecutive features (i.e. ability to separate objects from different classes).

In the last subsection we will describe our approach to the modeling task. We will define our objective, describe when, where, and how was the study done, what materials were used and who was included in the study groups. Also we will describe methods and algorithms used in the project.

Data description

Telemarketing is one of the forms used to encourage clients to buy new bank's product. If we imagine real-world scenario it may be very hard to decide to which customer we should call in order to achive our goal (in this case bank term deposit subscription) because it's not possible to call them all as we have limited human resources of telemarketers. Such in this case we could use historical data about calls done in previous marketing campaigns to formulate conclusions about what type of clients are our target group and what part of the day / week / year is a good time for such a projects. Moreover after that we can create classification models which will learn to give us good recomendations which clients we should call in the first order.

Variable name	Type	Description
Bank client data		
Age	Numeric	Age of client
Job	Categorical	Type of job
Marital	Categorical	Marital status of client
Education	Categorical	Education status of client
Default	Categorical	Has credit in default?
Housing	Categorical	Has housing loan?
Loan	Categorical	Has personal loan?
Variables related with the last contact of the current campaign		
Contact	Categorical	Contact communication type
Month	Categorical	Last contact month of year
Day of week	Categorical	Last contact day of the week
Duration	Numeric	Last contact duration, in seconds
Other attributes		
Campaign	Categorical	Number of contacts performed during this campaign
Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign
Previous	Numeric	Number of contacts performed before this campaign
Poutcome	Categorical	Outcome of the previous marketing campaign
Social and economic context attributes		
Emp.var.rate	Numeric	Employment variation rate - quarterly indicator
Cons.price.idx	Numeric	Consumer price index - monthly indicator
Cons.conf.idx	Numeric	Consumer confidence index - monthly indicator
Euribor3m	Numeric	Euribor 3 month rate - daily indicator
Nr. employed	Numeric	Number of employees - quarterly indicator
Outcome variable		
y	Categorical	Has the client subscribed a term deposit?

Table 1: Input variables.

In general our dataset can be splitted in five categories: bank client data, last contact of the current campaign, other attributes, social and economic context, outcome as we can observer in Table 1. First category describes general information about client - age, job, etc. Second category describes how the last contact with client was performed. There is also important note that Duration 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 we will discard it because our intention is to have a realistic predictive model. Third category in general tell us about previous campaigns and previous contacts with given person. There is also a note from dataset authors that Pdays equal 999 means client was not previously contacted. We will also note that in our data. Fourth category is about social and economic context attributes such as employment rate. This might give us information how was economy in this time and might be driving factor for some people. The last category is our outcome variable, i.e. flag if the client subsribed a term deposit.

Data analysis

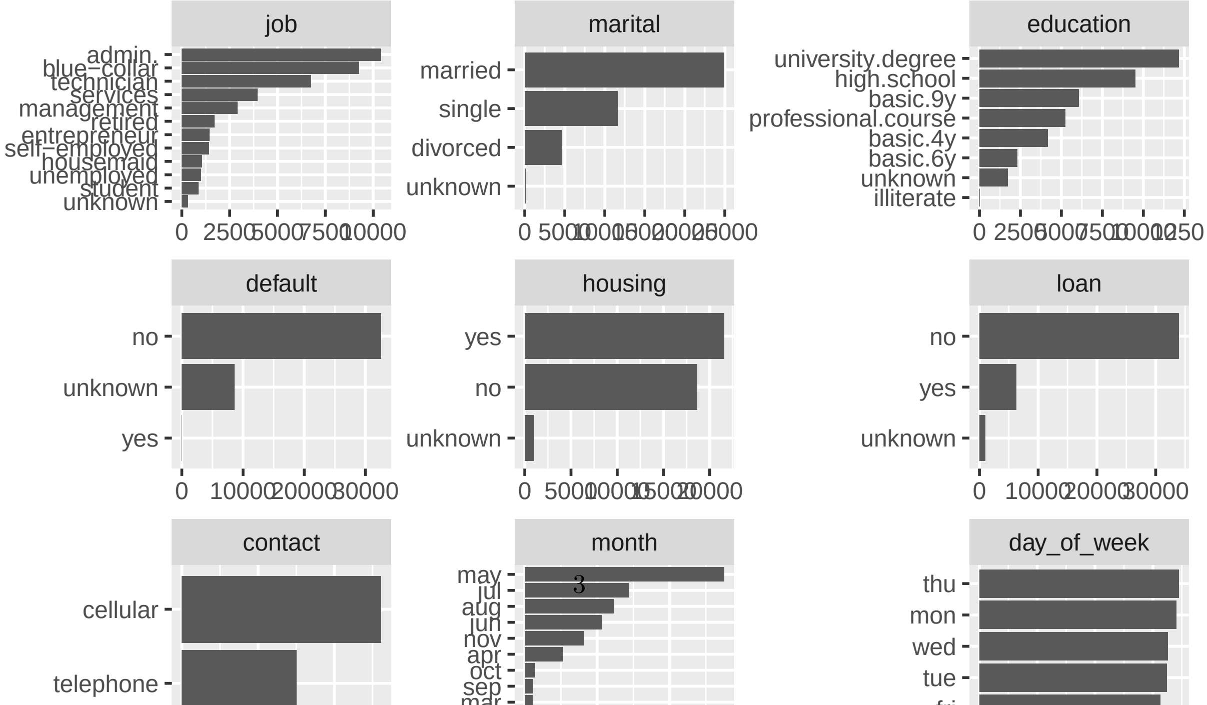
Input variables 1. There is 41188 rows. As we can observe in 3. There is only one column with a big amout of missing values - nr.empolyed. It has about 81% of missing values. It's very big amount and we've decided to remove this column from our analysis.

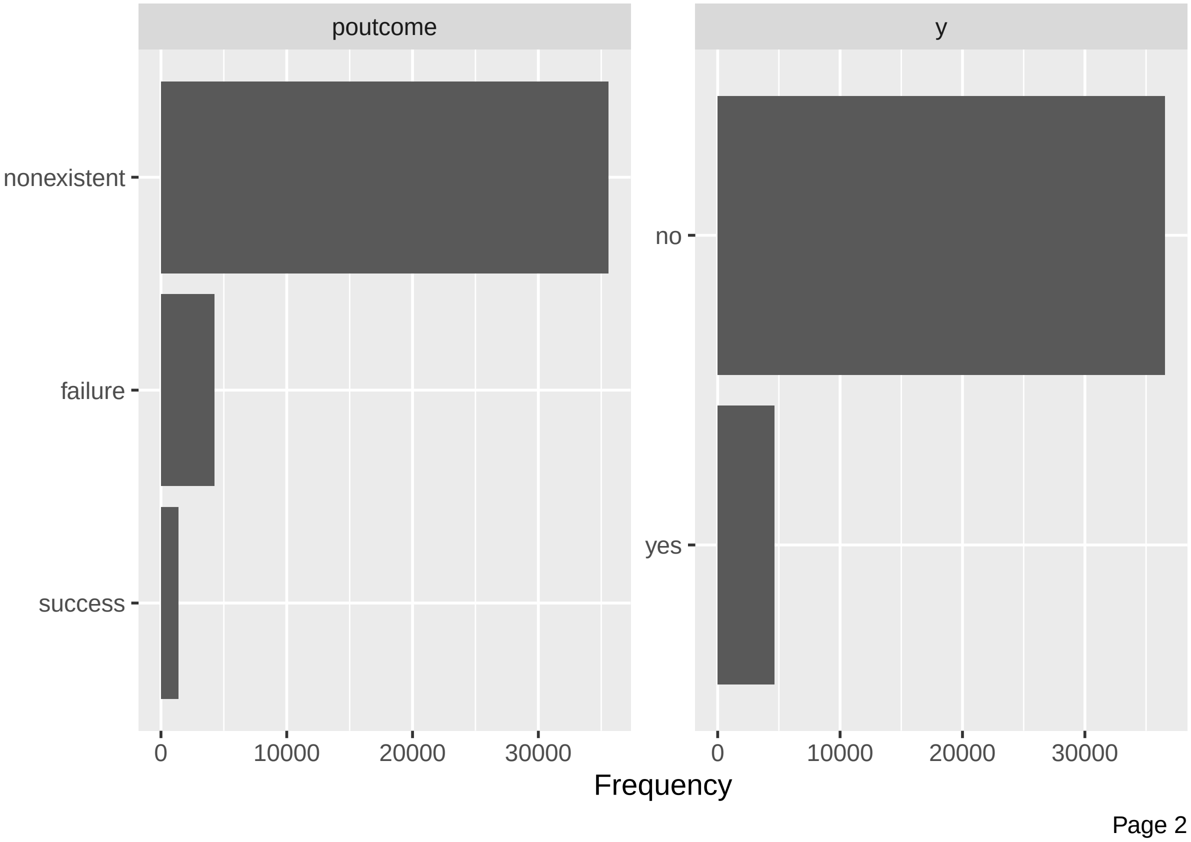
	Info
rows	41188.00
columns	21.00
discrete_columns	11.00
continuous_columns	10.00
all_missing_columns	0.00
total_missing_values	73098.00
complete_rows	0.00
total_observations	864948.00
memory_usage	6763504.00

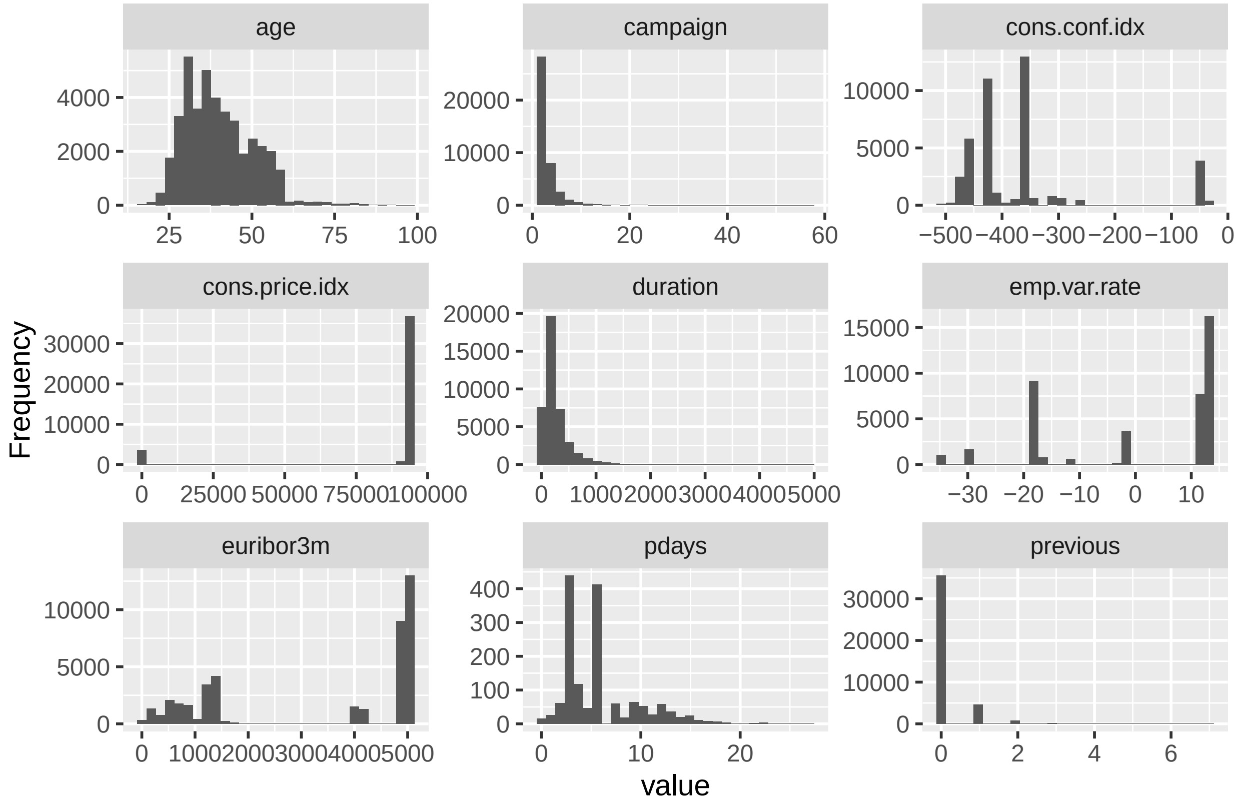
Table 2: Basic summary about dataset.

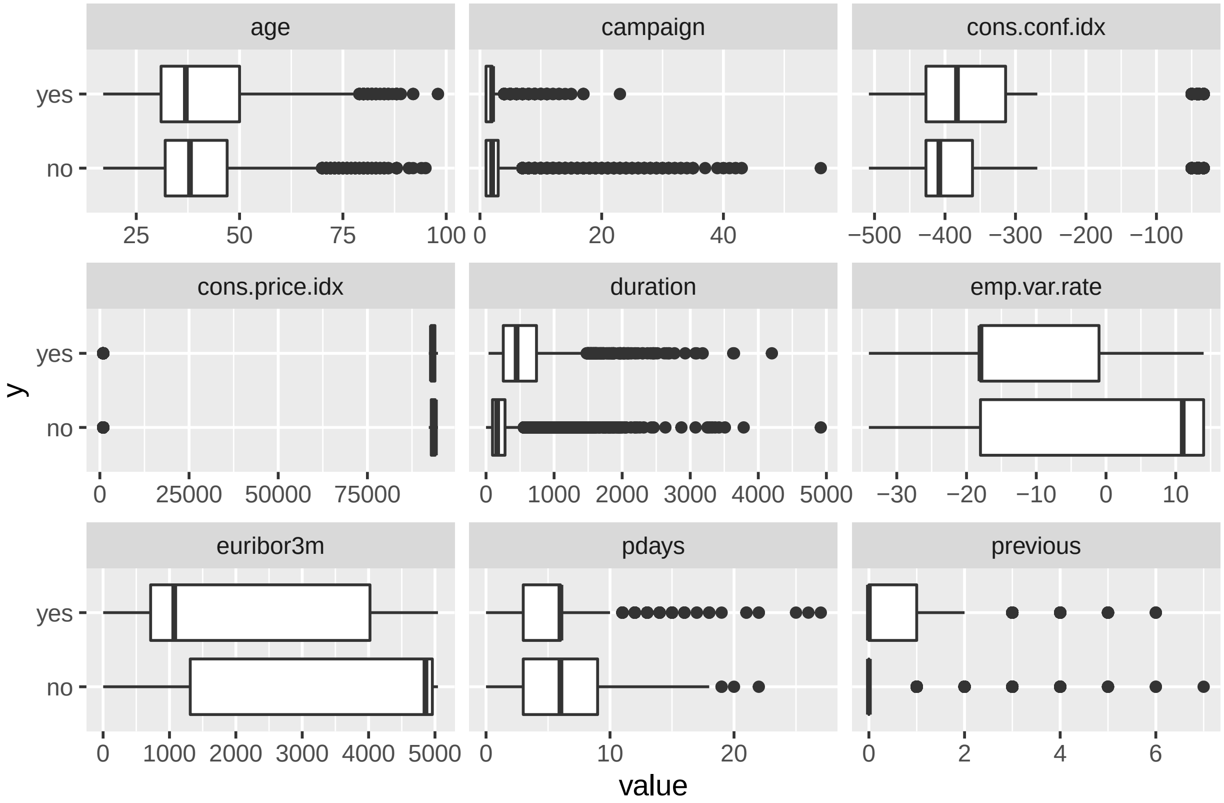
	feature	num_missing	pct_missing
1	age	0	0.00
2	job	0	0.00
3	marital	0	0.00
4	education	0	0.00
5	default	0	0.00
6	housing	0	0.00
7	loan	0	0.00
8	contact	0	0.00
9	month	0	0.00
10	day_of_week	0	0.00
11	duration	0	0.00
12	campaign	0	0.00
13	pdays	39673	0.96
14	previous	0	0.00
15	poutcome	0	0.00
16	emp.var.rate	0	0.00
17	cons.price.idx	0	0.00
18	cons.conf.idx	0	0.00
19	euribor3m	0	0.00
20	nr.employed	33425	0.81
21	y	0	0.00

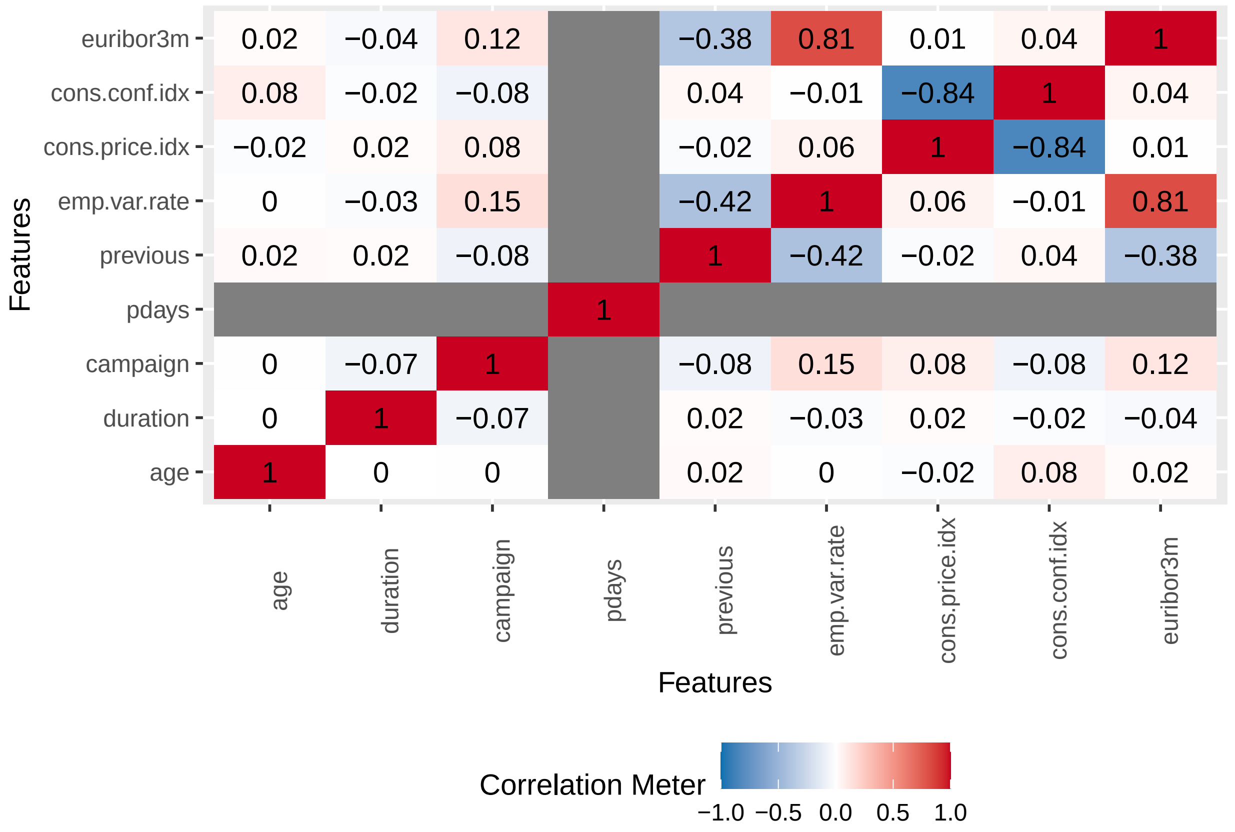
Table 3: Basic summary about missing values in dataset.




poutcome
y
nonexistent
failure
success
no
yes
0
10000
20000
30000
0
10000
20000
30000
Frequency
Page 2


age
campaign
cons.conf.idx
cons.price.idx
duration
emp.var.rate
euribor3m
pdays
previous
Frequency
value
0
2000
4000
25
50
75
100
0
10000
20000
0
20
40
60
0
5000
10000
−500
−400
−300
−200
−100
0
0
10000
20000
30000
0
25000
50000
75000
100000
0
5000
10000
15000
20000
0
1000
2000
3000
4000
5000
0
5000
10000
15000
−30
−20
−10
0
10
0
5000
10000
0
1000
2000
3000
4000
5000
0
100
200
300
400
0
10
20
0
10000
20000
30000
0
2
4
6


age
campaign
cons.conf.idx
cons.price.idx
duration
emp.var.rate
euribor3m
pdays
previous
yes
no
y
value
25
50
75
100
0
20
40
−500
−400
−300
−200
−100
0
25000
50000
75000
0
1000
2000
3000
4000
5000
−30
−20
−10
0
10
0
1000
2000
3000
4000
5000
0
10
20
0
2
4
6



Classification

Task description

When, where, and how was the study done? What materials were used or who was included in the study groups (patients, etc.)

- Methods and algorithms used in the project What methods / algorithms have been used? Which tasks these methods / algorithms were used for? (e.g. preliminary analysis and visualization, classifiaction, prediction, cluster analysis, etc.)
- Classification ** Objective - classification rule ** Methods: LDA, QDA, KNN, other ** Assessment of accuracy: Holdout / Cross Validation # End of the task description

Linear Discriminant Analysis

Quadratic Discriminant Analysis

k-Nearest Neighbours

Decision Tree

Results

Task description

What answer was found to the research question; what did the study find? Was the tested hypothesis true?

Results presented in the form of corresponding tables, graphs and diagrams. Note that only the most important results should be included in the report, whereas additional results can be added as attachments.

** Mehtods *** Summary statistics *** Plots # End of the task description

Discussion

Task description

What might the answer imply and why does it matter? How does it fit in with what other researchers have found? What are the perspectives for future research?

- Conclusions Precise conclusions: what can be concluded from the analyses carried out? How these conclusions could be put into practice? (e.g. development of a new / better strategy in the company, new / better diagnostic methods, etc.)
- Further research suggestions Short information on further possible directions of research (what could / should be further studied and what additional methods / algorithms could be used?) # End of the task description

References

[1] P. Cortez S. Moro and P. Rita. A data-driven approach to predict the success of bank telemarketin. *Decision Support Systems*, 2014.