Data Science Internship- Data Glacier

Data Science Project: Bank Marketing Campaign

Group Name: Greek Scientist

Name: Nikolaos Sintoris

Email: nikolaossintoris@gmail.com

Country: Greece

College/Company: Data Glacier

Specialization: Data Science

1. Problem Description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

2. Business Understanding

In the Business Understanding phase, we basically aim to:

- Determine Business Objectives: In our case, the fact that ABC Bank wants to sell its deposit product to customers.
- Assess Situations: In our case, better understanding of the customers intentions.
- Determine Data Science Goal: In our case, the goal is a binary classification. Do, the customers plan to buy the product or not?
- Produce Project Plan: In our case, the use of a machine learning model that will aid in shortlisting customers, who have higher chances of buying the product is going to save resources and time as well.

3. Project Lifecycle along with Deadlines

The project lifecycle of the bank marketing campaign consists of:

- Understanding the business problem (Week 7 of virtual internship)
- Data understanding (Week 8 of virtual internship)
- Data cleansing and transformation (Week 9 of virtual internship)
- Exploratory Data Analysis (EDA) (Week 10 of virtual internship)

- EDA presentation (Week 11 of virtual internship)
- Model selection and model building (Week 12 of virtual internship)
- Propose solution and model deployment (Week 13 of virtual internship)

4. Data Understanding (Week 8)

Data Information: The data is related with 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 (or not) subscribed.

Summary of the data: The bank-ful.csv file, consists of 45211 observations (rows) and 17 at tributes (columns). The size of the file is 4503 KB.
Data Types:

Input variables:

bank client data:

1. age (numeric)

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

3. marital: marital status (categorical: "married", "divorced", "single", note: "divorced" means divorced or widowed)

4. education: (categorical: "unknown", "secondary", "primary", "tertiary")

5. default: has credit in default? (binary: "yes", "no")

6. balance: average yearly balance, in euros (numeric)

7. housing: has housing loan? (binary: "yes", "no")

8. loan: has personal loan? (binary: "yes", "no")

related with the last contact of the current campaign:

9. contact: contact communication type (categorical: "unknown", "telephone", "cellular")

10. day: last contact day of the month (numeric)

11. month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12. duration: last contact duration, in seconds (numeric)

other attributes:

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

14. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15. previous: number of contacts performed before this campaign and for this client (numeric)

16. poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

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

Missing Values: In our data, missing values is the value "unknown" in the categorical attributes, so job has 288 missing values, education has 1857 missing values, contact has 13020 missing values and poutcome has 36959 missing values as they are shown in the screenshot below:

	Column name	Frequency
0	age	0
1	job	288
2	marital	0
3	education	1857
4	default	0
5	balance	0
6	housing	0
7	loan	0
8	contact	13020
9	day	0
10	month	0
11	duration	0
12	campaign	0
13	pdays	0
14	previous	0
15	poutcome	36959
16	y	0

We can deal with these missing values with a lot of techniques. We can use the mode technique, where we replace the missing values with the most frequent. Another technique is to replace them with a random one. Alternatively, we can remove the observations with the missing values, but we avoid this technique because we are going to miss a lot of information.

Outliers: In our dataset, we have 7 numerical attributes. I plotted a box plot and we can point out that the most outliers in the dataset are int the attributes age, balance, duration and campaign. A way to deal with these problem is to replace the outliers with the mean value of

every attribute. Alternatively, we can remove any observation that has outliers, but we avoid this technique because we are going to miss a lot of information.

5. Data Cleansing and Transformation (Week 9)

The first thing I did, was to convert “y” attribute from categorical to numerical because it will help me in the future. Then for the missing values I used 2 techniques. When the most frequent word has by far more data points, I am going to replace the missing values with the most frequent one. Otherwise, I am going to replace the missing values with a random value. Based on the above, for the “job” attribute I am going to use the second technique and the first technique for the other attributes (“education”, “contact”, “poutcome”). I avoid removing any observations with missing values because it may result in loss of information.