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-additional-ful.csv file, consists of 41188 observations (row s) and 21 attributes (columns). The size of the file is 6.6 MB.
Data Types:

Input variables:

bank client data:

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")

related with the last contact of the current campaign:

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", …, “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")

social and economic context attributes

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")

Missing Values: In our data, missing values is the value "unknown" in the categorical attributes, so job has 330 missing values, marital has 1731, default has 8597, housing has 990 and loan has 990.as they are shown in the screenshot below:

	Column name	# of missing values
0	age	0
1	job	330
2	marital	80
3	education	1731
4	default	8597
5	housing	990
6	loan	990
7	contact	0
8	month	0
9	day_of_week	0
10	duration	0
11	campaign	0
12	pdays	0
13	previous	0
14	poutcome	0
15	emp.var.rate	0
16	cons.price.idx	0
17	cons.conf.idx	0
18	euribor3m	0
19	nr.employed	0
20	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 10 numerical attributes. I plotted a box plot and we can point out that the most outliers in the dataset are int the attributes age, duration, previous 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 attributes *"job", "education"* and *"housing"* I am going to use the second technique and the first technique for the other attributes *("marital", "default", "loan"*). I avoid removing any observations with missing values because it may result in loss of information.

6. Exploratory Data Analysis (Week 10)

Github link: https://github.com/NikolaosSintoris/Data-Glacier-Virtual-Internship/tree/main/Week%2010

I performed Exploratory Data Analysis on the data. First I examined the pearson correlation coefficient between the attributes and the result was that there was no correlation between them. So I examined the relationship between the output variable ("y") and some of the rest attributes.

Total number of consumers by age: In the graph we can see that for the ages 20 to 60, there is a significant difference between those who applied and those who did not. We also observe that at every age, those who did not apply far outnumbered those who did.

Total number of consumers by job: In the graph we can see that the jobs admin, blu-collar and technician have the highest number of consumers. We also observe that at every job, those who did not apply far outnumbered those who did.

Total number of consumers by education: In the graph we can see that university degree has the highest number of consumers in both categories. We also observe that at every education, those who did not apply far outnumbered those who did.

Total number of consumers by previous outcome: In the graph we can see that nonexistent poutcome has the highest number of consumers in both categories. We also observe that those who have previous campaign and it was successful they subscribe in a term deposit.

Total number of consumers by month: In the graph we can see in May we have the most contacts in both cases. Generally, months in summer have the most contacts in both cases. We also observe that December, March, October and September we have small differences between those 2 cases.

Additionally, from the data description we know that the attribute duration 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.

So, based on the above the EDA showed that age, duration, month and poutcome may have important role for the bank to sell his term deposit to the customers.