Week 8: Deliverables

Group Name: JSSN.

	Name	Email	Country	University/Company	Specialization
1	Peter Okwukogu	peter.okwukogu@gmail.com	Nigeria	Colab Kaduna	Data Science
2	Almudena Zhou Ramírez	almu180@gmail.com	Spain	UNED	Data Science
3	Shoug Alotaibi	Shouga.1417@gmail.com	Saudi Arabia	KFUPM	Data Science
4	Mha Luqman Salameh	mahasa179@gamil.com	Jordan	AABU	Data Science

Github link:

https://github.com/Astharen/DataScientistsGroupProject.git

Problem Description :

ABC Bank wants to sell it's term deposit product to customers. The bank aspire to use an ML Model to focus on shortlisted customers which they are often interested in bank's campaigns. The Bank desires to save time costs for the employees and resources in these campaigns that will achieve only with marvelous ML model.

Data understanding:

Here the details of the dataset collected:

1. It has 41,188 observations
2. It has 21 columns
3. 5 of 21 the attributes have numerical data types, 11 of the 21 attributes are objects(strings) data types, and the remaining 5 are floats.

Clients data and their type:

Attributes Information:

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 a housing loan? (categorical: 'no','yes','unknown')

7 - loan: has a 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','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')

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')

Problems with data:

- NA values: There are no direct missing or null values

```
In [6]: df.isnull().sum()

Out[6]: age          0
        job          0
        marital      0
        education    0
        default      0
        balance      0
        housing      0
        loan         0
        contact      0
        day          0
        month        0
        duration     0
        campaign     0
        pdays        0
        previous     0
        poutcome     0
        y            0
        dtype: int64
```

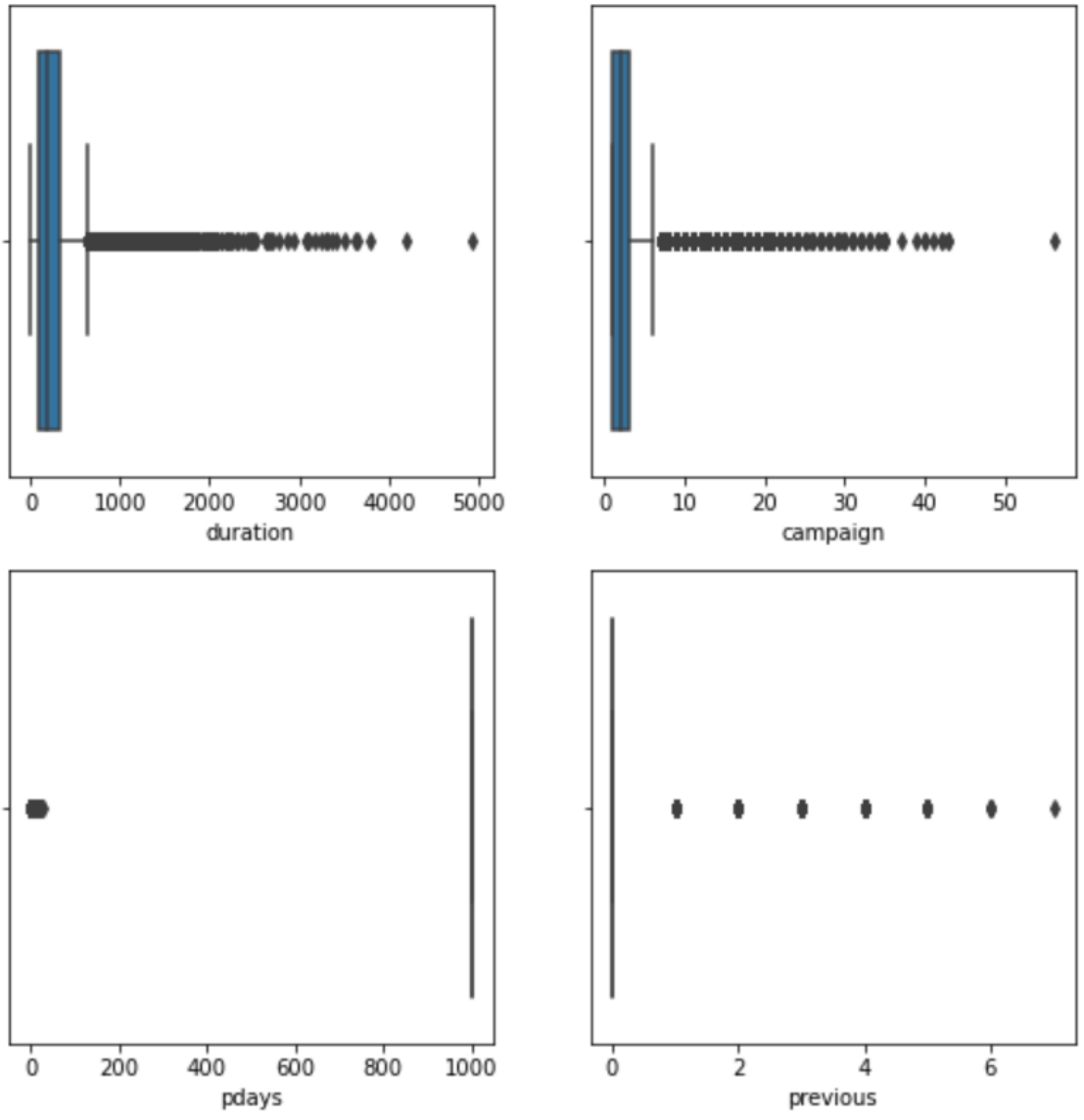
All the variables with NAs are categorical variables:

- job: unknown values will be considered NA. We will use the mode to replace them, since the proportion of unknown is considerably small.
- marital: the same method as at the job column is applied.
- education: knn to replace the unknown values, since it is an ordinal variable.
- default, housing and loan: either ignore the variable if we can prove that it's not significant enough or keep it as if unknown values were another

category. Default has more unknown values than yes so it is probable that it will be discarded. Housing has a similar frequency in both so we can compare between the frequencies they have of each y category. Loan has a few more yes than unknown but the solution of using the mode can be considered.

- poutcome: considering the frequency table, it will be ignored or we will consider nonexistent like another value.

- Outliers: We detected some outliers on duration, campaign, pdays and previous features.



Some outliers exist because of variability while others exist because of mistakes during data imputation. Outliers that exist because of variability will be left alone, those that exist because of anomalies will be corrected or removed based on the business understanding

- Skewness: All the numerical features tend to have positive skewness except pdays has negative skewness.