Group Name: The Greeks

Name: Michalis Galanakis

Email: mihalisgalanakis@hotmail.com

Country: Greece

College/Company: Athens University of Economics and Business

Specialization: Data Science

Problem Description: ABC Bank wants to sell its term deposit product to customers. Before launching the product, they want to develop a model which will help them understand whether a particular customer plans to buy their product or not (based on customer's past interaction with the bank or other Financial Institution).

Data Understanding: 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 ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Data types: In the dataset we have 10 numerical and 11 categorical features. Specifically, we have 5 features with integer variables, 11 features with string variables and 5 features with float variables.

Summary of the data: As regards the data volume and the total number of variables in the additional bank full data set, there are 5,56 MB of data, 41188 observations of 21 variables.

The data attributes are shown below:

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")
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 of the data set as well as a proposed approach:

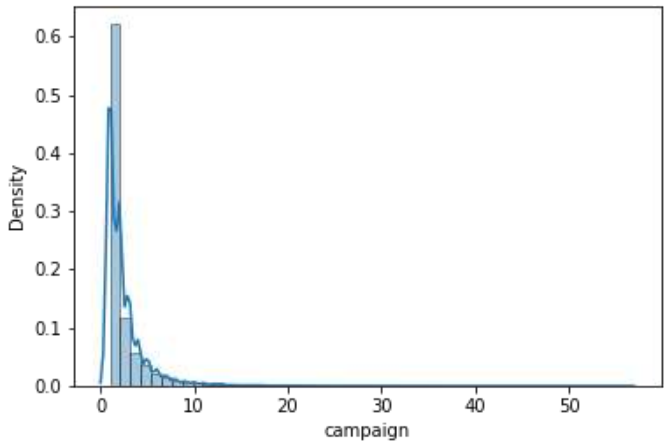
- **Missing values**: There are several missing values in some categorical attributes, all coded with the "unknown" label, as they are shown in the screenshot below.

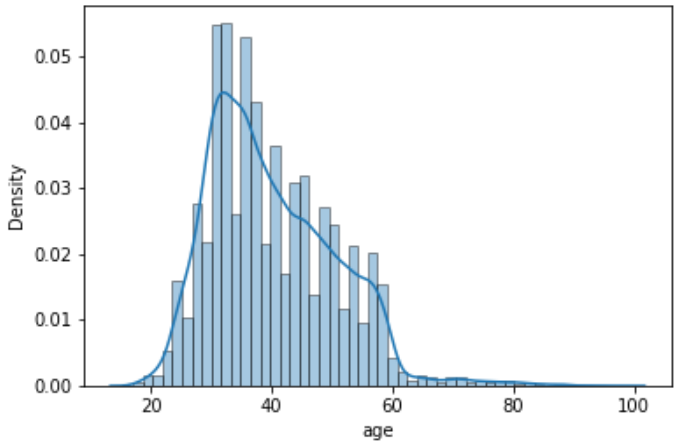
```
df.isnull().sum()
```

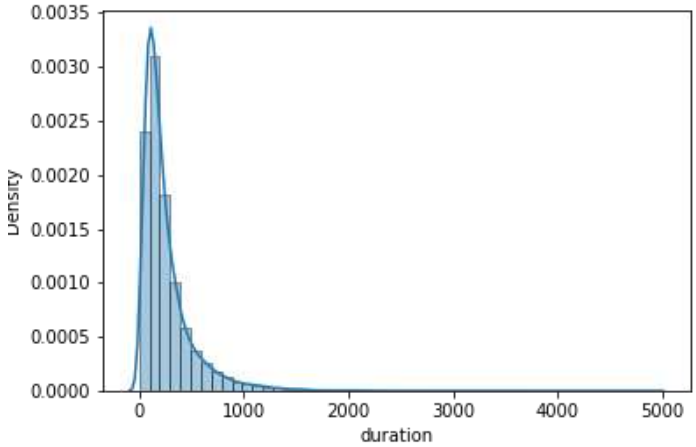
```
age                 0
job               330
marital            80
education        1731
default          8597
housing           990
loan              990
contact             0
month               0
day_of_week         0
duration            0
campaign            0
pdays               0
previous            0
poutcome            0
emp.var.rate        0
cons.price.idx      0
cons.conf.idx       0
euribor3m           0
nr.employed         0
y                   0
dtype: int64
```

These missing values will be treated with imputation techniques. Due to the fact that they are categorical attributes we will be using the mode technique, which is finding the most frequent value and filling the missing ones. Another strategy to apply is to fill the missing values with a random label from the existing ones. On the whole, we avoid removing missing observations, (unless it is necessary) because it could result in a model with bias and loss of information.
Alternatively, we can develop a model able to predict these missing values.

- **Skeweness:** Data is skewed when its distribution curve is asymmetrical (as compared to a normal distribution that is perfectly symmetrical) and skewness is the measure of asymmetry. The skewness for a normal distribution is 0. We have strong evidence of non-normality in the data set since the skewness of each feature is higher or lower than zero. Skewness in most of the attributes can be detected as well by plotting histograms, thus visualizing the data to validate their key characteristics. The feature of cons.price.idx has -0.230888 value of skewness, which is the closest value to zero. The following graphs showcase these issues in a few of the attributes.







An efficient way in order to handle the skewed data would be the use of transformations such as the log transformation which transforms skewed distribution to a normal one. As another option, we could remove outliers (tempting but now always the best decision) or normalize the skewed attributes instead.

Outliers: We have 10 numerical columns. By plotting boxplots, it is visible that the columns of Age, duration, previous and campaign have the most outliers in the dataset. The columns of pdays and cons.conf.idx have a few outliers. Finally, the rest of the numerical columns do not have any outliers. An efficient way to tackle this problem would be finding out which values are higher or lower than 1,5*IQR and then remove these values (should there be any). Last but not least, we could try imputation. By imputing the outliers we ensure that no data is lost. As impute values, we can choose between the mean, median, mode and boundary values.