Week 8 Data Glacier

Team Name: Data 4 Science

Team Member Details:

- **Amshumann Singh**
  amshumannsingh@gmail.com
  Ireland
  Dublin City University
  Data Science

- **Hamza Al Hajj Chehade**
  hamza.h.chehade@gmail.com
  Lebanon
  Lebanese University – Faculty of Sciences
  Data Science

- **Vaishnavi Dixit**
  withmylaptop@gmail.com
  India
  SIES GST, Mumbai University
  Data Science

Problem Statement:

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).

Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers whose chances of buying the product is more. This will save resource and their time (which is directly involved in the cost (resource billing)).

Data Understanding:

Attributes:

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','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')

In [206]:
```
data.describe()
```

Out[206]:

	age	duration	campaign	pdays	previous	EVR	CPI	CCI	euribor	no_emp
count	41176.00000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	40.02380	258.315815	2.567879	962.464810	0.173013	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	10.42068	259.305321	2.770318	186.937102	0.494964	1.570883	0.578839	4.627860	1.734437	72.251364
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

We have 11 categorical variables (job, marital, education, default, housing, loan contact, month, day_of_week, poutcome, y) including the target variable and 10 numerical variables (other variables).

We have from the mean age is about 40 years old (the minimum age is 17 years old and the maximum age is 98 years old).

As the data information said that it is better to drop the duration attribute since it is highly correlated in whether a potential client will buy a term deposit. Also, duration is obtained after the call is made to the potential client so if the target client has never received calls then this feature is not useful. The reason why duration is highly correlated with opening a term deposit is because the more the bank talks to a target client the higher the probability the target client will open a term deposit since a higher duration means a higher interest (commitment) from the potential client.

A campaign makes at least 1 call which might be the last call so the client may accept to buy a term deposit or he might have no interest to do so. However, the largest number of calls done by a campaign is about 56 calls including the last call. This might show that the client is very interested to buy a term deposit.

The pdays variable shows the number of days passed by after the client was last contacted from the previous campaign. It is numeric but the value 999 shows that the client is not previously contacted.

The EVR (employment variation rate) variable is numeric variable of unique values (1.1, 1.4, -0.1, -0.2, -1.8, -2.9, -3.4, -3.0, -1.7, -1.1) it is indicated every 3 months (quarterly indicator)

The CPI (consumer price index) variable is numeric variable of distinct values of a range between 92 and 95. It is calculated every month.

The CCI (consumer confidence index) variable is a numeric variable of negative distinct values. It is calculated every month.

The Euribor the approximation Euro Interbank Offered Rate it is 3-month rate based on the interest rates at which a panel of European banks borrow funds from one another. The average of the Euribor is

approximately around 3.621 where the minimum Euribor is 0.634 and the maximum is 5.045. It is calculated daily.

The no_emp (number of employee) variable is numeric variable indicated every 3 months (quarterly indicator) the minimum number of employees in 4963 while the maximum number of employees is 5228.

Type of Data:

The data is a mix of qualitative type due to the presence of numerical attributes and qualitative type due to the presence of categorical variables. The data shows a classification goal to predict if a customer open a term deposit (says yes) or not (says no).

The Problems in the Data:

- *Duplication:*
  These duplicates don't show that there exist clients with the same details, it shows that duplication happened while entering the data.
- *NA values:*
  No NA values are found.

```
data.isnull().sum()
```

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
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
```

- *Outliers:*

  Some numerical columns have outliers (especially 'pdays', 'campaign' and 'previous' columns).

Approaches:

pdays' holds the number of days that passed by after the client was last contacted from a previous campaign Looking closer into 'pdays' data we can see that:

- only 1.2% of values above 400. They are possibly outliers, so we should consider imputing something (possibly mean value) instead of these values.
- -1 possibly means that the client wasn't contacted before or stands for missing data.

Since we are not sure exactly what -1 means I suggest to drop this column, because -1 makes more than 50% of the values of the column.

'campaign' holds the number of contacts performed during this campaign and for this client (numeric, includes last contact) Numbers for 'campaign' above 34 are clearly noise, so I suggest to impute them with average campaign values while data cleaning.

'previous' holds the number of contacts performed before this campaign and for this client (numeric) Numbers for 'previous' above 34 are also really strange, so I suggest to impute them with average campaign values while data cleaning.

Note that instead of managing the data using the mean approach we can use the following approaches: Mode and Median

GitHub Link:
https://github.com/HamzaAlHajjChehade/Bank_Marketing_Campaign