BANK MARKETING CAMPAIGN

Group Name

Sparagua

Name, Email:

- Daniela Alvarez

  daniela.alvarez04@gmail.com
- Akhil Nair

  akhil.nair1908@gmail.com

Country

- Peru
- India

College/Company

- Daniela: Datacamp, Kaggle Learn, Universidad de Piura (Peru)
- Akhil: SIESGST Nerul

Specialization:

Data Science

Problem description

Portuguese bank is having a decrease in its revenue. The bank wants to be able to predict which clients are most likely to subscribe a term deposit so they can focus marketing efforts and resources on them and avoid wasting money and time on clients that will probably not subscribe.

Data understanding

Data collection: What type of data you have got for analysis?

In this project we have been given 4 datasets:

- Bank
- Bank-full
- Bank-additional
- Bank-additional-full

As Bank-additional-full.csv includes all information of Bank-full.csv plus 4 additional features, we decided to use the most complete dataset (Bank-additional-full.csv). The test dataset for that would be Bank-additional.csv. We found a problem when trying to read the data in csv format so we decided to save it into excel format so Pandas could read it correctly.

- Bank-additional: obtained using Pandas with read_excel function.
- Bank-additional-full: obtained using Pandas with read_excel function.

Features identified as not relevant:

- Duration: Duration highly affects the output target, yet it is not known before a call is performed. This input should be discarded for a realistic predictive model.

Data Description

The shape of the train (bank-additional-full) data is: (41188,21)
The shape of the test data (bank-additional) data is: (4119,21)

The data includes numerical data (float64, int64) and categorical data (object).

In the following picture we can see there are no null values at all so we'll verify that with some exploratory data analysis and visualizations.

```
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

Let's explain the meaning of each attribute:

1. Age (numeric)
2. Job: type of job (categorical)
3. Marital: marital status (categorical)
4. Education (categorical)
5. Default: has credit in default? (categorical)
6. Housing: has housing loan? (categorical)
7. Loan: has personal loan? (categorical)
8. contact: contact communication type (categorical)
9. month: last contact month of year (categorical)
10. day_of_week: last contact day of the week (categorical)
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.

12 campaign: number of contacts performed during this campaign and for this client (numeric)
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)

Important note: this attribute has 3 categories: “failure”, “success” and “non-existent”. 86% of the data falls into “non-existent” category.

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)

Target variable:

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

Basic statistics of numeric attributes:

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
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

<u>Data Exploration: What are the problems in the data (number of NA values, outliers , skewed etc)</u>

- Even tough there are no missing values, every feature has an “unknown” category which we should figure out how to deal with.

- In ‘poutcome’ feature there is a third category called ‘non-existent’. We should find out what does it mean.

- According to the statistics of numeric attributes there are some outliers in ‘Age’, ‘Duration’ and ‘Campaign’ features because the maximum value is much higher than the mean value and the third quartile. We should figure out how to deal with those outliers.

Data Quality: What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

- As there are no null values, we will consider ‘unknown’ values as null and we will try to impute them using the mean in numerical columns and using the most common value in categorical features.

- We will not drop outliers as they seem like possible values to happen.

Github Repo link

https://github.com/danielaaz04/Bank-Marketing-Campaign