BANK MARKETING CAMPAIGN

NAME: CATHERINE SANDA

EMAIL: sandacate@gmail.com

COUNTRY: KENYA

SPECIALIZATION: DATA SCIENCE

GITHUB REPO LINK: https://github.com/cate6495/bankweek8

PROBLEM DESCRIPTION

We are given data related to direct marketing campaigns i.e. phone calls of a bank in Portugal.

The classification goal is to predict whether a client will subscribe or not(yes/no) to a term deposit (variable y).

DATA UNDERSTANDING

The dataset consists of data of direct marketing campaigns of a Portuguese banking institution. The dataset was picked from UCI Machine Learning Repository. There are four variants of the datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with fewer inputs).

4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with fewer inputs).

I chose the bank-additional-full.csv. This dataset has 21 columns and 41188 rows.

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’; 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’

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.

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')

TYPE OF DATA FOR ANALYSIS

There were 10 Categorical variables and 10 numerical variables

There is a mix of float, object and integer datatypes.

There are no missing values in the dataset.

There are outliers in a few columns in our dataset.

Skewness was present in the age attribute of the dataset.

APPROACHES TO PROBLEMS IN THE DATASET

If there were any missing values in our dataset, we would have removed the rows if the number of missing values was less than the amount of data we have. Alternatively, we would have used imputation to handle the missing values.

In the case of outliers, we will drop them since outliers can skew statistical measures and data distributions causing a misleading representation of the underlying data and relationships.