Proposal: Bank Marketing Classification

Introduction:

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Business Problem:

predict the success of a bank marketing campaign based on the features we have in out data. This project should therefore help us be better able to identify potential customers and refine our the focus of future campaigns.

Data description:

Bank client data:
Age (numeric)
Job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
Marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

Contact: contact communication type (categorical:'cellular','telephone')
Month: last contact month of year (categorical: 'jan', 'feb', 'mar',…, 'nov', 'dec')
Dayofweek: last contact day of the week (categorical:'mon','tue','wed','thu','fri')
Duration: last contact duration, in seconds (numeric). Important
note: this attribute highly affects the output target (e.g., ifduration=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:

Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
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)
Previous: number of contacts performed before this campaign and for this client (numeric)
Poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

Emp.var.rate: employment variation rate - quarterly indicator (numeric)
Cons.price.idx: consumer price index - monthly indicator (numeric)
Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
Euribor3m: euribor 3 month rate - daily indicator (numeric)
Nr.employed: number of employees - quarterly indicator (numeric)
Output variable (desired target): y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Size of Data:

- Number of rows: 41188
- Number of columns: 21

Dataset sourec:

from Kaggle website (https://www.kaggle.com/henriqueyamahata/bank-marketing?select=bank-additional-full.csv)

Algorithms:

- Logistic Regression
- K-Nearest Neighbors
- Decision Trees
- Random Forest
- XGBoost

Tools:

Softwares:

1. VScode
2. Jupyter
3. Github
4. PowerPoint
5. Zoom

Languages & Library:

- Python
- Pandas
- numpy
- seaborn
- plotly

goals:

1. Look at counts of categorical variables and distributions of continuous variables and correlations.
2. Find out what factors affect on the y (target variable).
3. Try different models to see which performs best.

Team Members:

- Shaima Alzahrani
- Raghad Alnasser
- Nasser Alquraini