Group Name: Glacier Analysis Group

Team Members Details:

Name, Email, Country, College/Company, Specialization (Data Science, NLP, Data Analyst)

1. Tarusmita Boro, tarusmita.boro@gmail.com, United Kingdom, Data Science
2. Jonas Lütticken, jonaslutticken@gmail.com, Germany, Data Science
3. Noella Mutuku, noellamutuku@gmail.com, Kenya, Data Glacier, Data Science
4. Jeffrey Joseph, jeffreyjoseph21506@gmail.com, United Kingdom, Data Science

Problem Description:

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps 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.

Data understanding:

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.
Training data contains 45211 instances, that includes 17 features in total: 16 predictable features and 1 target variable ‘y’.
Some columns have missing values and they have been represented in the data as ‘NaN’. The year dataset was created is 2014, which was last updated on 18th August 2023.
The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. It required contacting the same client more than once to confirm whether they subscribed to the product, i.e. term deposit (Yes) or not (No).
There are four datasets:

- bank-additional-full.csv (41188) and 20 inputs, ordered by date (from May 2008 to November 2010)
- bank-additional.csv - that contains 10% of the examples (4119), randomly selected, 20 inputs.
- bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs.
- bank.csv – that contains only 10% of the examples selected and 17 input features. (older version of this dataset with less inputs)

The smaller datasets (10%) were provided to test computationally demanding machine learning algorithms like SVM.
The goal is to build a classifier to predict if the client will subscribe (yes/no) to a term deposit (variable y in our dataset).

1. age: numeric value. Age of the customer.

2. Job: type of job with the values admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services.
3. Marital: marital status. The values include married, divorced, single. Please note, divorced means divorced or widowed.
4. Education: the values could be unknown, secondary, primary, tertiary.
5. Default: has the credit in default. This would be binary: yes or no.
6. Balance: numeric value. Average yearly balance, in euros.
7. Housing: Has housing loan? Binary: yes or no.
8. Loan: has any personal loan? Binary: yes or no.
9. Contact: related with the last contact of only the current campaign. Contact communication type: unknown, telephone or cellular.
10. Day: Last contact day of the month (numeric)
11. Month: Last contact month of the year: jan, feb, mar, …, nov, dec.
12. Duration: Last contact duration, in seconds (numeric).
13. Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15. previous: number of contacts performed before this campaign and for this client (numeric)
16. poutcome: outcome of the previous marketing campaign. Categorical: unknown, other, failure, success.
17. Y: Output variable (desired target). Has the client subscribed a term deposit? (binary: yes or no)

Type of Data:

We have four datasets related to banking data:

- Bank-additional-full.csv: This dataset contains all examples (41,188) and includes 20 input features. The data is ordered by date, spanning from May 2008 to November 2010.
- Bank-additional.csv: This dataset represents a random sample of 10% (4,119 examples) from the Bank-additional-full.csv dataset. It also contains 20 input features.
- Bank-full.csv: An older version of the dataset, Bank-full.csv includes all examples and 17 input features. Similar to the previous datasets, it is ordered by date.
- Bank.csv: This dataset is derived from Bank-full.csv and consists of a random 10% sample with 17 input features.

We have different types of data in these datasets ranging from numeric/binary data like age, housing, loan, etc. to text data like job, marital status, etc.

Data Issues:

The datasets exhibit some missing values. Notably, columns related to ‘Job’, ‘marital’, ‘Education’,

‘default’ ‘Housing’ and ‘Loan’ contain unknown values.

Approach to Addressing Data Issues:

To handle these missing values, we tried dropping the row which contains unknown values from the education and job column as we could see that most of the clients have not subscribed. Another approach we tried was to employ one-hot encoding, to handle these missing values. This technique converts the “unknown” elements into numeric values, allowing us to incorporate them effectively in our analysis.

This is our initial analysis of the data. As we progress further on Exploratory Data Analysis, we might try some other approaches and will decide the best approach.

Github Repo link

https://github.com/TaruIndia/predict-termdepositsubscription