WEEKLY DELIVERABLE 2

Group Name: Bank Marketing DS 01

Specialization: Data Science

Team Members:

- **Ines Perko**
    - Master of Mathematics
    - Freie Universität Berlin, Germany
    - Specialization: Finite Element Methods
    - Email: ines.perko93@gmail.com

- **Suvansh Vaid**
    - Master of Data Science
    - Monash University, Melbourne, Australia
    - Specialization: Data Science
    - Email: suvanshvaid@gmail.com

- **Zeynep Başak Eken**
    - Bachelor's in Economics, Minor: Software Development
    - Bilkent University, Ankara, Turkey
    - Specialization: Data Analysis
    - Email: zeynepbasakeken@gmail.com

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).

Business understanding:

- **Business Objective:**

The Bank wants to shortlist customers whose chances of buying the product are more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only on

those customers. This will save their resource and their time (which is directly involved in the cost (resource billing).

- **Success Criteria:**

The success criteria for this business problem would be based on how much maximum number of customers we are able to predict who have subscribed to the product.

Data Understanding:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

The dataset “bank-full.csv” contains a total of 45211 rows and 17 variables (columns) including the 16 input variables such as bank client data, the data related to the last contact of the current campaign, the data of attributes related to the campaign details and social and economic context attributes. The output variable y has two binary options: yes and no, saying if the client subscribed to a term deposit.

Information on each column:

<u>Input variables:</u>

bank client data:
1 - age (numeric)
2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", , “blue-collar","self-employed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary: "yes","no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes","no")
8 - loan: has personal loan? (binary: "yes","no")

related with the last contact of the current campaign:
9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)

other attributes:
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")

Output variable (desired target):
17 - y - has the client subscribed a term deposit? (binary: "yes","no")

Checking Data types: There are two types of variables for analysis, categorical and numerical. There are 10 categorical features: 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome and 10 numerical features: 'age', 'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed'. In those 10 numerical features, 3 are discrete numerical features**:** 'previous', 'emp.var.rate', 'nr.employed' and the rest 7 of it are continuous numerical features.

Checking for outliers: Duration and campaign are heavily skewed towards left and seem to have some outliers, age as well. Our strategy for dealing with the outliers would be not to remove any of the outliers but instead replace them by either mean, median or mode. Also, we would use 3 different outlier detection strategies such as 3 sigma, hampel correction and boxplot rule, to make sure no outliers are left undetected.

Checking for NA values: There are no NA values in the dataset. However, there are a few Unknown values under a few categorical variables, which suggests the missing information which may affect the analysis. To deal with this, we would first check whether removing rows with the unknown values makes any difference. We could also simply ignore this and carry forward the unknown factor while building the model.

GitHub Repo link:

https://github.com/SuvanshVaid27/Bank-Marketing-Project