WEEKLY DELIVERABLE 3 (Week 9)

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.

GitHub Repo link:

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

Data Cleaning and Transformation

1. 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.
For detecting the outliers and replacing them, we used the following 3 strategies each:
Ines – **Boxplot rule; replacing with mean**
Basak – **3 sigma; replacing with mode**
Suvansh – **Hampel correction; replacing with median**

2. Imbalanced Data: Since the output variable y for our classification is highly imbalanced with majority of rows having `no` responses, we have tried 3 different strategies each to make the data balanced, as given below:
Suvansh - **Over Sampling**
Ines **- Under Sampling**
Basak – **SMOTE**

3. 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.