Team Details:

Group Name: Data Detectives

Name:	Bisma Azeem	Elif Nur Kemiksiz
Email:	bismazeem1304@gmail.com	lose.yourself.elif@gmail.com
Country:	Saudi Arabia	Turkey
College/Company:	Virtual University of Pakistan	Marmara University
Specialization:	Data Science	Data Science

Problem Description:

ABC Bank intends to develop a model that can predict which clients are most likely to buy their new term deposit product. They will be able to target their marketing activities more successfully as a result, concentrating on clients who are more likely to make a purchase. The project will involve building and evaluating machine learning models using customer data. The model's success will be measured by its ability to accurately predict buyers and the resulting cost savings from optimized marketing efforts.

Data Understanding:

1. Data Description:

- Source of Data:
  The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The dataset can be found here
- Purpose of Data Collection:
  Come up with a data-driven approach to predict the success of bank telemarketing campaigns for selling term deposits, with the goal of optimizing marketing resources by focusing on customers with a higher likelihood of subscribing.
  Link to the Documentation of this dataset.

2. Dataset Overview:

- Dataset name: bank-additional-full
- File Format: .csv
- File Size : 5MB
- Records Count: 41188
- Attributes Count: 21
- Target Variable: ‘Outcome/y’

3. Data Exploration:

Col_Name	Data_Type	Col_Group	NA_Vals	Description
age	int64	Continuous	0	Age of customer
job	object	Categorical	330	Type of job
martial	object	Categorical	80	Marital status
education	object	Categorical	1596	Level of education
default	object	Categorical	8597	Has credit in default?
housing	object	Categorical	990	Has housing loan?
loan	object	Categorical	990	Has personal loan?
contact	object	Categorical	0	Contact communication type
month	object	Categorical	0	Last contacted month
day_of_week	object	Categorical	0	Last contacted day
duration	int64	continuous	0	last contact duration, in seconds
campaign	int64	continuous	0	number of contacts performed during this campaign
pdays	int64	continuous	0	number of days that passed by after the client was last contacted from a previous campaign
previous	int64	continuous	0	number of contacts

				performed before this campaign
poutcome	object	Categorical	35217	outcome of the previous marketing campaign
emp.var.rate	float64	continuous	0	employment variation rate - quarterly indicator
cons.price.idx	float64	continuous	0	consumer price index - monthly indicator
cons.conf.idx	float64	continuous	0	consumer confidence index - monthly indicator
euribor3m	float64	continuous	0	euribor 3 month rate - daily indicator
nr.employed	float64	continuous	0	number of employees - quarterly indicator
outcome\y	object	categorical	0	has the client subscribed a term deposit?

4. Initial Data Quality Assessment:

- **Data Accuracy:** Checked dataset for errors in data entry, typos and inconsistency (incorrect ages and job titles etc) and found none.
- **Data Completeness:** There are a few categorical columns that have some NA values as “unknown”.
    - We'll exclude 'Job' and 'Marital' features with unknown values because they represent less than 1% of the data. This minimal impact allows us to focus on the remaining data with minimal loss of information.
    - To ensure consistency, we'll remove the 'housing' and 'loan' columns entirely. Since both these features have unknown values for the same records, dropping 990 rows seems like a more robust approach compared to imputation techniques. Imputation might introduce bias, and removing these columns avoids that potential issue
    - Instead of dropping the 8,597 rows with unknown values in 'default', which would result in significant data loss, we'll utilize KNN imputation to predict these unknowns. KNN imputation fills missing defaults by finding similar data points, potentially improving model performance for complete data needs. This approach allows us to leverage the existing data to fill in the missing values and maintain a more comprehensive dataset for analysis.

- The ‘education’ and ‘poutcome’ features have a significant portion of ‘unknown’ and ‘nonexistent’ values respectively. We’ll deal with them by encoding them as a feature category as they might be helpful for the model in analyzing patterns.
- The majority of our numerical features exhibit positive skewness. Additionally, features like 'emp.var.rate,' 'cons.price.idx,' and 'cons.conf.idx' demonstrate significant imbalance. To address these issues, we'll employ transformation techniques for the skewed features and explore appropriate methods to handle the imbalanced features
- By addressing all potential issues, we ensure the data is more reliable and consistent for exploratory data analysis (EDA). This, in turn, can lead to the development of a more accurate model.