Assignment I

Dataset information

Load dataset_DT.csv

The data is ordered by date (day, month)

Features:

- age --> age
- job --> type of Job
- marital --> marital status
- education --> highest education finished
- default --> already has credit in default?
- balance --> account balance
- housing --> taken housing loan?
- loan --> taken personal loan?
- contact --> communication via...
- day --> day of last contact
- month --> month of last contact
- duration --> duration of last contact
- campaign --> number of contacts made to the client during the campaign
- pdays --> number of days that passed by after the client was last contacted from a previous campaign (999 means client wasn't previously contacted)
- previous --> number of contacts performed before this campaign and for this client
- poutcome --> outcome of the previous marketing campaign

Target variable:

- y --> has the client subscribed a term deposit?

--> Programming Assignment Details

1) For this assignment use **ONLY R**

2) You can use libraries:

- **dplyr**: Data manipulation.
- **tidyr**: Data tidying.
- **stringr**: String manipulation.
- **caret**: Model training and preprocessing.
- **data.table**: Fast data manipulation.

3) Make sure to write about 2-3 lines to explain any kind of visualization.

I. Hints -

1. Apart from null values, the dataset consists of "unknown" (string) values in multiple columns. You need to handle them as a part of null values.
2. There might be columns with redundant data, i.e., information from a column might also be available from another column. If there are such columns, you can drop them.
3. Unwanted data can reduce the model's accuracy.

II. Tasks

1. Import the libraries and load the dataset (from the csv file) [5 points]
2. **Data Pre-processing[15pts]**
    A. Handling Missing Values:
        - Description: Identifying and addressing missing data using techniques like imputation, deletion, or filling with default values.
        - Objective: Ensure the dataset is complete and accurate

B. Removing Duplicates:

- Description: Identifying and removing duplicate records to avoid redundancy.
- Objective: Maintain data integrity and avoid biased analysis.

C. Handling Outliers

- Description: Identifying and managing outliers that may skew the results.
- Objective: Ensure that outliers do not disproportionately influence the analysis.

3. Data Transformation[15pts]

A. Normalization and Scaling:

- Description: Adjusting the range of numerical features to a common scale (e.g., scaling between 0 and 1).
- Objectiv**e**: Ensure that features contribute equally to the model.

B. Encoding Categorical Variables:

- Description: Converting categorical variables into numerical format using techniques such as one-hot encoding or label encoding.
- Objective: Make categorical data usable for machine learning algorithms

4. Data Reduction[15pts]

A. Feature Selection:

- Description: Determine the Six most influential attributes on target attribute (with explanation). You do not necessarily need to drop the remaining features. Your task is just to determine and show the ***Six*** most influential attributes with detailed explanation.
- Objective: Reduce the dimensionality of the dataset and avoid overfitting.

B. Dimensionality Reduction:
    - Description: Using techniques like Principal Component Analysis (PCA) to reduce the number of features while preserving important information(>=90%).
    - Objective: Improve model performance and reduce computational complexity

C. Aggregation:
    - To compress the data, numerous columns might be combined into one feature.

5. **Visualization[15pts]**
    A. Perform 2 visualizations of the features with respect to target variable with detailed explanation.
    B. Perform the box plots before and after the data preparation.

6. **DEMO [15pts]**
7. **Good luck**