Group Project

Group name: One-Man-Army
Name: Denis Remo
Email address: denisrusa@duck.com
Country: U.S.A
College/Company: University of Suffolk
Specialization: Data Science
GitHub Repo link:
https://github.com/DenisSRemo/virtual_internship/tree/main/Group%20project/group%20project%20data
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 help 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

What type of data you have got for analysis?

The datasets is structured as tabular data in CSV format, containing various features for analysis. These features include demographic information, financial and categorical variables, as well as numerical variables. The target variable for classification appears to be denoted by "y." The values within each row are delimited by semicolons. This format is conducive to a diverse range of analytical techniques, making it suitable for exploratory data analysis, statistical modeling, and machine learning applications.

What are the problems in the data (number of NA values, outliers , skewed etc)

1. **Missing Values:**
    - While the provided sample data doesn't overtly indicate the presence of missing values, it is noteworthy that categorical columns, denoted as 'unknown' (e.g., default, education), may signify instances of unreported or undefined data. A comprehensive evaluation of the entire dataset is imperative to discern and appropriately address any latent missing values.
2. **Outliers:**
    - Preliminary scrutiny has hinted at potential outliers. A more thorough examination is requisite to discern the context of these outliers and determine whether corrective actions are warranted.
3. **Skewed Data:**
    - Notably, the 'duration' variable demonstrates right skewness, indicating a concentration of lower values. Consideration of suitable transformation techniques, such as logarithmic transformations, may be prudent to ameliorate skewness and enhance the variable's distributional characteristics.

4. **Imbalanced Classification:**
    - An evident observation is the potential imbalance within the classification variable 'y,' where instances predominantly exhibit a 'no' outcome. Addressing the imbalance in binary classification scenarios is paramount, and strategies such as resampling or employing specialized evaluation metrics may be employed to ensure model robustness and efficacy.
5. **Data Format:**
    - A distinctive characteristic of the dataset is the amalgamation of all values into a single column, demarcated by semicolons. This format implies a delimited structure, necessitating careful consideration during data parsing and formatting to align with the requirements of analytical tools and methodologies.

This evaluative summary is grounded in the provided dataset. A comprehensive analysis of the entire dataset is strongly recommended to gain a nuanced understanding of its intricacies and potential analytical considerations.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

1. **Handling Missing Values:**
    - Utilize appropriate imputation techniques based on the nature of the missing data. This may involve replacing missing values with mean, median, or mode for numerical variables and employing methods like mode or forward-fill for categorical variables. The choice of imputation method depends on the distribution and context of the data.
2. **Addressing Outliers:**
    - Employ statistical methods such as the IQR (Interquartile Range) or Z-score to identify and subsequently handle outliers. Depending on the extent and impact of outliers, options include either transforming the data, Winsorizing (capping) extreme values, or considering specialized outlier detection algorithms. The chosen approach should align with the dataset characteristics and the specific goals of the analysis.
3. **Dealing with Skewed Data:**
    - Apply appropriate transformations to mitigate skewness in the data distribution. Common techniques involve logarithmic transformations for positively skewed data or power transformations. The goal is to achieve a more symmetrical distribution, facilitating improved model performance and interpretability.
4. **Mitigating Imbalanced Data:**
    - Employ strategies tailored to handle imbalanced classification, especially when dealing with a disproportionate distribution of classes. Techniques may include oversampling the minority class, undersampling the majority class, or using advanced sampling methods like SMOTE (Synthetic Minority Over-sampling Technique). Additionally, utilizing evaluation metrics such as precision, recall, and F1-score instead of accuracy is crucial to accurately assess model performance in the presence of imbalanced data.
5. **Data Parsing and Formatting:**
    - Given the specific data format of semicolon-separated values, employ parsing techniques to structure the data appropriately. Utilize libraries or functions

compatible with the chosen analytical tools, ensuring seamless integration and accurate representation during subsequent analyses.

These approaches are grounded in established best practices within the field of data preprocessing and analysis, aiming to enhance the overall quality and reliability of the dataset for downstream analytical tasks.