Bank Marketing (Campaign)2

Team member's details : Team Me, Francisco Lopez, panch.lopez.21.1998@icloud.com, United States, Eastern University, Data Science

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).

Data understanding:

- **Attributes:** The dataset contains various attributes such as demographic information, financial details, contact methods, and campaign metrics.
- **Data Types:** Both numerical and categorical features are present.
- **Missing Values:** Several columns have missing values, notably 'job', 'education', 'contact', and 'outcome'.
- **Other Issues:**

  -Some categorical variables contain inconsistent or unknown values.

  -'Duration' may be correlated with the target variable but could lead to data leakage.

  -'Pdays' has many records with a value of -1, indicating no previous contact.

What type of data you have got for analysis:

The data consists of both numerical and categorical variables. Numerical attributes include age, balance, duration, campaign, pdays, and previous. Categorical attributes include job, marital status, education, default, housing loan, personal loan, contact method, day of the week, and month.

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

The problems in the data include:

1. **Missing Values: Several columns have missing values, such as 'job', 'education', 'contact', and 'poutcome'.**
2. **Outliers: Outliers may exist in numerical columns like 'age', 'balance', 'duration', 'campaign', 'pdays', and 'previous'.**

Bank Marketing (Campaign)2

3. **Skewed Distribution: The distribution of numerical variables might be skewed, affecting the performance of some machine learning algorithms.**

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

addressing data issues:

1. **For missing values:**
    - **Use mode imputation for categorical features like 'job', 'education', 'contact', and 'poutcome'.**
    - **Consider dropping columns with a high proportion of missing values if they're not critical.**
2. **For outliers:**
    - **Apply Winsorization to replace extreme outliers with less extreme values.**
    - **Prefer robust algorithms like Random Forest or Gradient Boosting for outlier-resilient model training.**
3. **For skewed distributions:**
    - **Utilize transformations like logarithmic or square root transformations for skewed numerical variables.**
    - **Apply binning to convert continuous variables into categorical ones, enhancing interpretability**

Github Repo link: https://github.com/Panch2/Bank-Marketing-Campaign-.git