Project : Bank Marketing
Name : Apoorv Ajay Thite
University : The Pennsylvania State University
Email : apoorvthite21@gmail.com
Country : United States of America
Specialisation : Data Science
Internship Batch : LISUM34
Date : 23rd July, 2024

Data Understanding and Preliminary Analysis

Problem Description

ABC Bank aims to predict whether a customer will subscribe to their term deposit product based on their past interactions with the bank or other financial institutions. The data is sourced from direct marketing campaigns conducted by a Portuguese banking institution, primarily through phone calls. The classification goal is to predict if a client will subscribe ('yes') or not ('no') to a term deposit, helping the bank to focus their marketing efforts efficiently and save resources.

Data Understanding

The dataset provided for analysis includes information about customers and their interactions with the bank during the marketing campaigns. It contains both numerical and categorical variables, such as age, job, marital status, education, balance, housing, loan, contact, and campaign details. The target variable is 'y', which indicates whether the client subscribed to the term deposit.

Problems in the Data

Upon initial inspection, the dataset reveals several common data quality issues:

1. **Missing Values**:
   - The dataset doesn’t contain any missing values which need to be identified and appropriately handled to ensure robust model performance.
2. **Outliers**:
   - Outliers are present in several numerical columns. These can skew the results and affect the accuracy of the machine learning models.
3. **Skewness**:
   - Several numerical columns exhibit skewness, which can impact the distribution of the data and the performance of the models.

Approaches to Overcome Data Problems

To address the identified issues, the following approaches will be applied:

1. **Handling Missing Values**:
    - Missing values will be identified using the isnull() function and then handled through imputation. Since there are non here, we wouldn’t have to worry about handling missing values.

2. **Managing Outliers**:
    - Outliers will be detected using the Interquartile Range (IQR) method. Depending on their impact, outliers may be removed or transformed to reduce their influence on the model. Techniques such as log transformation or capping at a certain threshold might be applied.

3. **Addressing Skewness**:
    - To handle skewness in numerical columns, data transformations such as log transformation, square root transformation, or Box-Cox transformation will be employed. These transformations help normalize the data distribution, improving model performance and interpretability.

By implementing these preprocessing steps, the dataset will be prepared for effective modelling, ensuring that the machine learning models can learn from the data without being misled by missing values, outliers, or skewed distributions. This comprehensive approach will help in developing robust predictive models to assist ABC Bank in optimizing their marketing strategies and resource allocation.