Group Name: GIG group

Name: Rupert Tawiah-Quashie

Email: rupertquash@gmail.com

Country: USA

College: Hampshire College

Specialization: Data Science

Problem Description:

- ABC Bank currently sells term deposits (fixed-term savings accounts) through generalized marketing campaigns via channels like telemarketing, email, etc.
- They want to improve the efficiency of marketing by targeting customers more likely to subscribe to the term deposit product.
- The bank has data on ~45,000 customers with details on demographics, account history, previous marketing contacts, economic indicators, and most importantly the label of whether the customer subscribed to a term deposit in the past campaign.
- The goal is to build a predictive model using this data to estimate the probability that each customer will subscribe to a term deposit.

Approach to Data Cleaning and Transformation

Data Import

- Use fetch_ucirepo() to load bank marketing dataset and access the features dataframe X

Missing Value Handling

Method 1:

- Identify numeric and categorical columns in X
- Use SimpleImputer with 'median' strategy to impute numeric columns
- Impute categorical missing values with mode
- Confirm no more missing values using .isnull().sum()

Method 2:

- Load different dataset (heart disease) using fetch_openml()
- Identify and print missing value counts using .isnull().sum()
- Drop rows with ANY missing values using .dropna()
- Confirm cleaned dataframe has no missing values

Outlier Identification

Method 1:

- Calculate z-scores for each numeric column using scipy.stats

- Count outliers with z>3 standard deviations
- Clip outliers by capping to +3 to -3 range
- Filter dataframe to remove rows with clipped outliers

Method 2:

- Calculate IQR and define outlier bounds
- Identify outliers based on IQR threshold
- Set identified outliers to NaN to remove them
- Confirm summary stats after outlier handling