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.

Data Understanding:

- The data comes from a Portuguese banking institution and relates to direct marketing campaigns conducted through phone calls. The classification goal is to predict whether a client will subscribe (labeled as "yes") or not subscribe (labeled as "no") to a term deposit bank product.
- The full dataset (bank-additional-full.csv) contains 41,188 examples with 20 input features for each example. The examples span marketing campaign contacts made from May 2008 to November 2010, ordered chronologically by date. This full set allows analysis across the entire time period.
- There are also sampled subsets with 4,119 examples (bank-additional.csv) and 17 input features (bank-full.csv and bank.csv). These smaller sets were created by random sampling to enable the testing of more computationally intensive machine learning algorithms.
- By combining analysis of the full dataset with the experimentally more flexible sampled subsets, both broad and deep investigation of the marketing response modeling can be supported. The multiple datasets provide the ability to understand campaign dynamics over time as well as tune predictive models.
- The outcome variable to predict is a binary "yes/no" label indicating a term deposit subscription. Given this clear target variable and substantial provided datasets, the banking

marketing problem appears well-posed for precision modeling of client response using machine learning approaches. Both input pattern analysis and tuned model development can be explored.

What type of data you have got for analysis

- The data captures whether clients subscribed (labeled "yes") or did not subscribe (labeled "no") to term deposit bank products based on phone call marketing campaigns.
- There are a total of 41,188 client examples in the full dataset (bank-additional-full.csv), with 20 input features provided for each client.
- The client input features provided likely cover demographics like age and job information, banking history with the institution, contact details, previous campaign characteristics if applicable, and term deposit specific features relevant to response modeling.
- The full set of 41,188 examples spans all marketing campaign contacts made over multiple years, from May 2008 to November 2010. This data is ordered chronologically.
- In addition to the full dataset, smaller 10% sampled datasets are provided for more flexible analysis with computationally intensive methods. These contain 4,119 examples and 17 input features.

Potential data issues:

1. Missing values:
    - The "unknown" category for categorical variables like job, education, contact, and previous outcome could indicate missing values. This needs to be confirmed.
    - If they are missing values, it needs to be determined what percentage are missing.
2. Skewed distributions:
    - Numeric variables like age, balance, duration could be highly skewed rather than normal distributions. Exploratory data analysis is required.
3. Outliers:
    - Extreme values in numeric variables like balance, age, duration, could distort analysis. These need to be detected and handled.
4. Other data quality checks:
    - Are there illegal values or errors for appropriate ranges, formats, etc?
    - Are there enough examples of each class? The number of "yes" vs "no" term deposit subscriptions should be checked.

Handling errors:

Missing values ("unknown" categories):

- Impute missing values using mean, median or model prediction values. This retains sample size.
- Or exclude missing value rows if proportion is small. This avoids assumptions.

    - Understanding missing value causes could inform strategies.
2. Skewed distributions:
    - Try log transforms for positive highly skewed numeric variables. Normalizes distributions.
    - Alternatively, create buckets like quartile bins. Reduces the impact of outliers.
3. Outliers:
    - Windsorizing - Cap outliers to threshold values derived from the median and percentiles. Reduces distorting effects of extremes.
    - Assess model performance with and without outliers. Remove if degrades performance.
4. Data errors:
    - Identify illegal values, unreasonable ranges, etc. Mark for further follow-up.
    - Check the balance of subscription classes. Oversample minority class if needed.