Group Name: The Banker
Name: Manil Shangle
Email: manilshangle@gmail.com
College: University of Texas at Austin
Specialization: Data Science

Problem Description
ABC Bank wants to promote its new term deposit product. To make its marketing efforts more efficient and cost-effective, it needs a machine learning model that can predict whether a customer will subscribe to the term deposit based on their personal, interaction, and socio-economic data. This model will help the bank focus its resources on customers with a higher likelihood of subscribing.

Data Understanding
The dataset contains information on customers contacted during a marketing campaign, including:

- Demographic attributes (age, job, marital status, education, etc.)
- Banking information (balance, loan, housing)
- Marketing campaign data (number of contacts, last contact duration, contact method)
- Economic indicators (employment variation rate, consumer confidence index, etc.)
- Target variable: whether the customer subscribed to a term deposit (yes or no)

Type of Data for Analysis
We are working with tabular data that includes:

- Categorical features (job, education, contact, etc.)
- Numerical features (age, balance, duration, etc.)
- Binary features (loan, default, housing)
- Target variable (binary classification)

Problems in the Data

- Missing/NA values: Some fields contain 'unknown' instead of standard missing values.
- Outliers: Observed in numerical columns like balance and duration.
- Imbalanced classes: The target variable has significantly more 'no' responses than 'yes'.
- Skewed distributions: Features like balance and duration are right-skewed

Approaches to Handle Data Issues

- Missing/NA values: Replace 'unknown' with NaN, then apply imputation strategies (e.g., mode for categorical, median for numerical).
- Outliers: Use IQR-based filtering or log transformation for high-skew variables like balance.
- Skewness: Apply transformations (e.g., Box-Cox, log) to normalize distributions.

- Imbalanced classes: Use SMOTE (Synthetic Minority Over-sampling Technique) or apply class weights during model training to mitigate imbalance.

Each of these techniques ensures that our model is robust and generalizable, especially for real-world deployment where interpretability and efficiency are crucial.

GitHub Repository Link:
https://github.com/ManilShangle/DataGlacierProject