Project: Bank Marketing (Campaign)

Name: Project Week 9 (Bank Marketing Campaign)
Week 7: Deliverables
Name: Priyanjali Patel
Email: priyanjalipatel@gmail.com
Country: India
Batch Code:LISUM45
Specialization: Data Science
Submission Date: 6 July 2025
Submitted to: Data Glacier
(Individual project)

Table of Contents

1. Problem Description

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help 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).

- **Approaches to Handle Data Issues**

To ensure data quality and improve model performance, the following approaches are applied:

🔹 1. Handling Missing/Unknown Values (NA or 'unknown')

- **Categorical columns** like job, marital, education, default, etc., contain 'unknown' instead of actual NAs.
- **Approach**: Treat 'unknown' as a separate category or impute using the median, mean or mode when appropriate.
- **Why**: This preserves data volume and avoids biased removal of potentially informative rows.

🔹 2. Handling Outliers

- Columns like 'age', 'campaign', 'pdays','previous','emp.var.rate' have extreme values.
- **Approach**:
  - Cap values at the 95th percentile (Winsorizing) for campaign and previous. (The term "quantile(0.95)" refers to the 95th percentile of a dataset, which is the value below which 95% of the data points fall. In simpler terms, it's the point where 95% of the values are less than or equal to that specific value)
  - Treat pdays = 999 as a special case (e.g., create a binary feature: previously contacted or not).
  - Visualize and evaluate outliers in age for domain consistency.
- **Why**: Outliers can distort model learning, lead to overfitting, and affect interpretability.

🔹 3. Encoding Categorical Variables

- Features like job, education, marital, contact, etc., are categorical.
- **Approach**: Use **one-hot encoding** or **label encoding** depending on the model used.
- **Why**: ML algorithms require numerical inputs.

🔹 4. Class Imbalance in Target Variable

- Target (y) is imbalanced (no >> yes).
- **Approach**: Try **SMOTE**, **class weighting**, or **undersampling** techniques during model training.
- **Why**: To avoid biasing the model toward the majority class and improve recall for minority class.

These preprocessing steps ensure that the dataset is clean and suitable for training accurate and generalizable machine learning models.

Imputation : Mean/Median, Segmented Imputation

In data science, imputation refers to the process of replacing missing values in a dataset with substituted values.

Feature	95th Percentile Capping	IQR-Based Clipping
Based on	**Fixed percentile (usually 95%)**	**Spread of data (Q1, Q3, IQR)**
Affects	**Top 5% only**	**Anything beyond 1.5×IQR from Q1/Q3**
Symmetry	**Can cap one or both ends (e.g., 5% & 95%)**	**Always caps both tails (low & high)**
Sensitivity to skewed data	**Low**	**High — may wrongly cap in skewed data**
Risk of over-correction	**Lower**	**Higher in small or skewed datasets**

Typical use	**ML pipelines, winsorizing**	**Statistical analysis, boxplot filtering**

Final Recommendation:

  - **Use IQR or 95th percentile to identify outliers**
    Then use median to replace them for robustness

(bank marketing project):

The dataset is structured/tabular, not NLP-focused. So this step is not mandatory unless:

- **We add text fields (e.g., customer feedback, call transcripts)**
- **We work with an extended dataset containing textual info**

I focus on instead:

- **Handle outliers and missing values in numerical columns**
- **Use techniques like IQR, capping, median imputation**
- **Clean categorical columns (e.g., `'unknown'`, whitespace)**
- **Encode them (LabelEncoding, OneHotEncoding, or WOE)**

5. Github Repo link

https://github.com/priyanjalipatel/Data_Glacier_Final_Project/tree/main