Week 8- Bank Marketing Campaign Project

Team member's details :

Name	Email	Country	College/Company	Specialization
Leena Ganta	leenarganta@gmail.com	USA	North Carolina State University	NLP
Najma Abdi	abdinajma225@gmail.com	KENYA	UNIVERSITY OF NAIROBI	DATA SCIENCE
Adama	adamal.sall@gmail.com	SAUDI ARABIA	University Of Suffolk	BSc Honours Computer Science(AI and Data Science)

Problem description:

ABC Bank is planning to launch a new term deposit product and wants to predict which customers are likely to subscribe based on their past interactions with previous marketing campaigns. By developing a machine learning model, the bank aims to identify high-probability prospects in advance, allowing the marketing team to focus its efforts where they are most likely to succeed. This targeted approach will help reduce overall campaign costs while improving conversion rates.

However, the dataset exhibits a class imbalance (most customers don’t subscribe) and certain highly predictive features, such as ‘duration', can’t be used for real-time targeting.

The goal is to develop and compare ML models (with and without the 'duration' feature) in order to:

- Improve the precision of campaign targets.
- Reduce the operational costs of telemarketing and messaging.
- Convert ML model metrics into insights that business stakeholders can understand.

The main goal is to enable cost-efficient, data-driven campaign targeting with measurable business impact.

Data understanding:

The Bank Marketing Dataset comprises 41,188 records and 21 characteristics, with the primary objective of predicting whether a client will subscribe to a term deposit. The dataset encompasses:

- Client information (e.g., age, occupation, marital status),
- Contact details (e.g., month of contact, contact method, duration),
- Campaign performance metrics (e.g., number of contacts, days since last contact), and
- Socioeconomic indicators (e.g., employment rate, consumer confidence).

Although no explicit missing values are present, certain categorical fields—such as *default*, *education*, and *housing*—utilize the placeholder 'unknown' to denote unavailable data. Additionally, numerical features like *duration*, *campaign*, and *previous* exhibit significant skewness and outliers. While *duration* is highly predictive, it introduces data leakage and should be excluded from modeling.

Overall, the dataset offers valuable insights but necessitates comprehensive preprocessing, including categorical encoding, imputation of "unknown" values, and mitigation of distributional irregularities to ensure robust model performance.

What type of data do you have for analysis?

The analysed data set is the Bank Marketing Dataset (bank-additional-full.csv) which contains 41,188 records with 21 mixed-type features (numerical and categorical), where the target variable (y) indicates term deposit subscription.

Features comprise client profile (age, job type, marital status, education), campaign details (contact method, timing, interaction frequency), and socio-economic context (employment variation, consumer indices, Euribor rates). The analysis identifies key decision factors, optimizes marketing strategies, and ensures GDPR-compliant handling of sensitive data.

What are the problems in the data (number of NA values, outliers , skewed etc):

The dataset contains various data quality concerns that necessitate preprocessing. Although there are no genuine NaN values, missing data in category columns is marked as 'unknown'. This is most evident in the 'default' area, which contains 8,597 'unknown' entries, followed by the 'education' field, which has 1,731 entries. Significant amounts of 'unknown' entries can also be found in the 'housing', 'loan', 'employment', and marital status' fields.

Outlier analysis using the IQR approach finds a substantial number of extreme values across important numerical features: Previous (5,625 outliers), Duration (2,963), Campaign (2,406), Pdays (1,515), and Age (469). Furthermore, distribution analysis reveals substantial skewness: positive for 'length' (3.26), 'campaign' (4.76), and 'prior' (3.83), and severely negative for 'pdays' (-4.92). If these concerns are not addressed correctly by imputation, outlier treatment, or transformation approaches, the analytical conclusions may be misinterpreted.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

To address dataset challenges, we implement targeted preprocessing steps. For categorical variables containing 'unknown' values (e.g., 'default', 'education'), we treat these as valid categories or impute using mode substitution, balancing information preservation with data integrity.

Numerical outliers identified through IQR analysis are managed via Winsorization or logarithmic transformation, maintaining extreme value patterns while stabilizing model

performance. We apply power transformations (log1p or square root) to heavily skewed features (duration, pdays) to approximate normal distributions.

The duration feature is excluded entirely to prevent data leakage, as it's only available post-call. For categorical encoding, we employ one-hot encoding for nominal variables and ordinal encoding for ranked categories, ensuring optimal model compatibility. These methods collectively maintain statistical rigor while optimizing predictive performance.

Github Repo link: https://github.com/leenarganta/bank_marketing_campaign