NAME: VIVIAN KERUBO MOSOMI

Email: kerubomosmi7@gmail.com

Country: Kenya

College: Jomo Kenyatta University of Agriculture and Technology (JKUAT)

Specialization: Data Science

1. Introduction

Background

The banking sector has always sought efficient ways to attract customers for term deposits. Traditional marketing strategies often rely on mass advertising, but machine learning can offer a more targeted approach. By analyzing customer data, banks can identify individuals who are most likely to purchase a term deposit, thereby optimizing their marketing efforts.

Problem Statement

The bank wants to improve its marketing campaign efficiency by predicting which customers are likely to subscribe to a term deposit. This study aims to build a machine learning model that accurately classifies whether a customer will purchase a term deposit based on various attributes such as age, job, marital status, and past interactions.

2. Dataset Overview

Data Source

The dataset used in this study contains customer information collected from a bank's marketing campaigns. The key features include:

- **Demographic Attributes:** Age, Job, Marital Status, Education.
- **Financial Attributes:** Balance, Loan Status, Housing Loan Status.
- **Previous Contact Details:** Number of contacts, last contact duration.
- **Economic Indicators:** Consumer confidence index, euribor three-month rate.
- **Target Variable:** purchased_deposit (Yes/No)

Data Distribution

- **No Purchase:** 88.73%
- **Yes Purchase:** 11.26%

- The dataset exhibits a class imbalance, requiring resampling techniques to improve model learning.

---

3. Data Preprocessing

Handling Missing Values

- Checked for missing values and found none in critical features.
- Standardized categorical variables by encoding them into numerical values.

Feature Engineering

- One-hot encoding applied to categorical features such as job, marital status, and education.
- Cyclical encoding applied to time-related features (month, day of the week).

Class Imbalance Handling

- Used **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the dataset, ensuring equal representation of both classes.

---

4. Exploratory Data Analysis (EDA)

- **Correlation Analysis:** Identified key influencing features such as duration of the last contact and euribor rate.
- **Class Distribution Visualization:** Highlighted the imbalance before and after applying SMOTE.
- **Feature Importance:** Used Random Forest and XGBoost to rank the most impactful features.

5. Model Selection & Training

Models I Implemented Include:

1. **Logistic Regression** (Baseline Model)
2. **Random Forest Classifier**
3. **XGBoost Classifier**

Model Training

- First split data into **80% training** and **20% testing**.
- Performed HyperParameter Tuning using Bayesian Optimization.

---

6. Model Evaluation & Results

Evaluation Metrics

I used the following metrics:

- **Accuracy**: Measures overall correctness.
- **Precision**: Ensures fewer false positives (important for marketing costs).
- **Recall**: Ensures fewer false negatives (important for capturing potential customers).
- **F1-Score**: Balances Precision and Recall.

7. Performance Comparison

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Logistic Regression	90.36	65.50	38.52	48.51
Random Forest	90.51	62.42	48.92	54.85
XGBoost	91.27	66.45	52.42	58.61
XGBoost(After Tuning)	91.54	68.87	51.49	58.93

6.3 Best Model Selection

XGBoost emerged as the best-performing model, achieving the highest F1-score.