# BANK TERM DEPOSIT SUBSCRIPTION PREDICTION REPORT

Juliet Fafali Kukuia June 2025

#### 1. Problem Statement

The objective of this project is to develop a machine learning model that predicts whether a client will subscribe to a term deposit based on their demographic, financial, and campaign interaction data. This enables the bank to optimise targeting strategies for future marketing campaigns, reduce customer acquisition costs, and improve ROI.

### 2. Methodology Summary

The most feature-rich dataset with all examples (41188) and 20 inputs, was selected to maximise model performance and insight generation. After preprocessing to treat outliers, encode categorical variables, and drop information-leaking features (like duration), three models were built: Logistic Regression, Random Forest, and XGBoost. These models were trained on a stratified 80:20 train-test split. Their performance was compared using precision, recall, F1-score, ROC-AUC, confusion matrices, and ROC curves. XGBoost was chosen for deployment via a Streamlit app due to its balanced performance across all metrics.

### 3. Exploratory Data Analysis (EDA) Findings

 Imbalanced Classes: The target variable y was imbalanced, with a significantly larger number of 'no' responses compared to 'yes'.

- Outliers Detected: Several numerical features such as age, campaign, previous, and pdays contained outliers. These were treated through winsorization and clipping.
- Missing/duplicate Values: 12 duplicate values were removed from the dataset
- Correlation Insights: Variables like duration, euribor3m, and emp.var.rate showed strong signals in relation to subscription rates.
- Data Preprocessing: Dummy encoding was applied to categorical variables. The column duration, which is known to cause target leakage, was removed before model training.

## 4. Modelling Methods and Evaluation

Three classifiers were trained and compared:

• Logistic Regression, Random Forest and XGBoost Classifier

Each model was evaluated using:

• Precision, Recall, F1-Score, ROC-AUC Score and Confusion Matrix

The results indicated:

Model	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Logistic	0.35	0.65	0.45	0.798	Good recall; correctly
Regression					identified 603
					subscribers and 6,208
					non-subscribers.
Random	0.55	0.29	0.38	0.776	High precision;
Forest					predicted 7,088 non-
					subscribers but only
					269 subscribers.
XGBoost	0.36	0.60	0.45	0.787	Best trade-off; 557
					subscribers and 6,325
					non-subscribers
					correctly identified.

#### 5. Final Model Selection

XGBoost was selected as the final model due to its superior balance of recall and F1-score, making it the most appropriate choice for identifying potential term deposit subscribers with reduced false negatives.

# 6. Insights and Recommendations

- Customer Profile: Customers likely to subscribe are older, contacted fewer times, had recent contact, and had favourable economic indicators (e.g., low emp.var.rate, high euribor3m).
- Marketing Focus: Optimise call campaigns around customers with similar profiles, and reduce repeated contacts which showed diminishing returns.
- Feature Importance: The top 10 influential features included euribor3m,
  nr.employed, emp.var.rate, and age.
- Operational Impact: Improved campaign targeting can reduce costs and improve conversion rates significantly.

### 7. Appendix

- Project Notebook: Azubi Africa TMP.ipynb
- Streamlit App Code: xgboost \_app.py
- Model File: xgb\_subscription\_model.pkl
- Feature Columns: xgb model features.pkl