📊 Marketing Analytics & Modeling Report

Project Summary

This analysis aims to assist a marketing team in identifying clients most likely to subscribe to a term deposit based on the bank-additional-full.csv dataset. Both bank-full.csv and bank-additional-full.csv were initially reviewed, but the modeling focused on the richer bank-additional-full.csv.

Data Preparation

- Data was loaded and checked for missing values (none found).
- Preprocessing included:
    - Mapping categorical time features like month to numeric and quarterly representations.
    - Creating new features like contacted_before (based on pdays) and campaign_vs_previous (a proxy for contact saturation).
    - Outlier removal using IQR for duration and campaign.
    - One-hot encoding for all categorical variables including engineered features.

Modeling Workflow

- **Model**: XGBoost Classifier was chosen due to its strength with tabular data and ability to handle class imbalance.
- **Imbalance Handling**: scale_pos_weight was calculated from the data to address the skewed target.
- **Hyperparameter Tuning**: Performed with GridSearchCV across 5 folds and multiple metrics, ultimately selecting:
    - max_depth=5, learning_rate=0.1, n_estimators=100, subsample=0.8, colsample_bytree=1
- **Evaluation**:
    - ROC AUC: 0.955
    - Accuracy: 88%
    - Precision (Class 1): 0.40
    - Recall (Class 1): 0.90
    - F1 Score (Class 1): 0.56

Feature Analysis

- Most impactful features:

- emp.var.rate, nr.employed, and euribor3m—all tied to **macroeconomic context**.
- duration was highly predictive but only useful for understanding outcomes, not pre-call strategy.
- Seasonal indicators like month_oct and job categories such as job_blue-collar provided additional discriminatory power.

Threshold Optimization

An **interactive threshold slider** was implemented to help marketers adjust sensitivity:

- At **higher thresholds**, precision increases (fewer false positives), but recall drops.
- At **lower thresholds**, recall increases—maximizing captured subscribers—but with lower precision.

Given the goal of a **telemarketing campaign**, **higher recall is strategically preferred**, enabling the capture of most willing subscribers even at the cost of some inefficiency.

Conclusion & Next Steps

This modeling pipeline:

- Effectively identifies the profile of likely subscribers.
- Provides actionable segmentation insights to support **targeted, data-driven outreach**.
- Balances trade-offs between over-contacting and missing potential customers via a tunable decision threshold.

Recommendations:

- Integrate model outputs into CRM systems to score leads.
- Schedule campaigns during months with higher success likelihood.
- Use economic indicators to time major campaigns.
- Conduct A/B tests to compare results with and without model-informed targeting.