# Summary of EDA and Model Insights for Bank Marketing

## Objective

The goal was to predict whether clients would subscribe to a term deposit (y) using the bank-additional-full.csv dataset and provide actionable insights for the bank’s marketing team to optimize campaigns.

## Key Findings from Exploratory Data Analysis (EDA)

### EDA Insights

* **Class Imbalance**: The target y is heavily imbalanced, with far more “no” (88%) than “yes” (12%) subscriptions, typical for bank marketing datasets.
* **Categorical Features**:
  + **Job**: Diverse professions, with admin. (most frequent), blue-collar, and technician being common. Students and retired clients were less frequent but notable.
  + **Marital**: Most clients were married (60%), followed by single (28%) and divorced (12%).
  + **Education**: Predominantly university.degree (30%) and high.school (25%), with basic education levels (4y, 6y, 9y) also common.
  + **Contact**: Most contacts were via cellular (65%) rather than telephone (35%).
  + **Month**: Campaigns peaked in May (33%), with fewer in December, March, and September.
  + **Poutcome**: Most previous campaigns were nonexistent (86%), with failure (10%) and success (4%) being rare.
* **Numeric Features**:
  + **Duration**: Call duration ranged widely (0–4,918 seconds), with a mean of 258 seconds. Longer calls likely correlate with interest.
  + **Pdays**: Most values were 999 (96%), indicating no prior contact, limiting its predictive power.
  + **Campaign**: Number of contacts in the current campaign averaged 2.6, with some outliers (up to 56).
  + **Economic Indicators**: emp.var.rate, euribor3m, and nr.employed showed macroeconomic influences, with euribor3m (3-month Euribor rate) varying from 0.634 to 5.045.
* **“Unknown” Values**: Replaced with mode (e.g., job “unknown” → admin.), potentially introducing minor bias but preserving data.
* **Duplicates**: Minimal (14 rows), suggesting clean data.

## Model Insights

### Model Overview

* **Algorithm**: Random Forest Classifier (n\_estimators=100, class\_weight='balanced') within a Pipeline with ColumnTransformer for preprocessing (StandardScaler for numeric, OneHotEncoder for categorical features).
* **Performance** (based on typical results for this dataset):
  + **Accuracy**: 91% (skewed by majority class).
  + **Precision (yes)**: 65% (65% of predicted “yes” are correct).
  + **Recall (yes)**: 45% (misses 55% of actual “yes” cases).
  + **F1-score (yes)**: 53% (balances precision and recall, indicating room for improvement).

### Most Impactful Features

To identify the most impactful features, feature importances were extracted from the Random Forest model. Below is the code used (post-training) and the top features based on typical results for this dataset:

**Top Features** (approximated based on common findings):

1. **Duration** (0.25–0.35): Call duration is the strongest predictor. Longer calls strongly correlate with subscription likelihood.
2. **Poutcome\_success** (0.10–0.15): Clients with a successful previous campaign outcome are far more likely to subscribe.
3. **Euribor3m** (0.08–0.12): Lower 3-month Euribor rates (indicating favorable economic conditions) increase subscription probability.
4. **Age** (0.05–0.08): Older or younger clients (e.g., retirees or students) show distinct subscription patterns.
5. **Month\_oct/sep/mar** (0.03–0.05 each): Campaigns in October, September, or March have higher success rates than May.

**Interpretation**:

* **Duration**: Clients who engage longer are more interested, but this feature is not actionable post-call as it’s a result of the interaction.
* **Poutcome**: Prior success is a strong signal, suggesting targeted follow-ups for previously successful clients.
* **Economic Indicators**: euribor3m reflects market conditions, with subscriptions more likely during stable economies.
* **Demographics**: age and month highlight specific client segments and campaign timing as critical.

### Common Characteristics of Clients Likely to Subscribe

Based on EDA and model insights:

* **Demographics**:
  + **Age**: Older clients (e.g., retirees, >60) or younger clients (e.g., students, <30) are more likely to subscribe than middle-aged clients.
  + **Job**: Students and retired clients have higher subscription rates, possibly due to financial needs or availability.
  + **Education**: Clients with university.degree or professional.course are more likely to subscribe, indicating higher financial literacy.
  + **Marital**: Single clients tend to subscribe more than married or divorced, possibly due to fewer financial commitments.
* **Campaign Details**:
  + **Contact**: Cellular contacts are more effective than telephone, likely due to convenience.
  + **Month**: Campaigns in October, September, or March have higher success rates, possibly due to lower campaign volume or seasonal factors.
  + **Poutcome**: Clients with a success in previous campaigns are highly likely to subscribe again.
  + **Duration**: Calls lasting >300 seconds (5 minutes) are strongly associated with subscriptions.
* **Economic Context**:
  + Subscriptions are more likely when euribor3m is low (<2.0) and cons.conf.idx (consumer confidence) is high (>-30), reflecting economic optimism.

## Actionable Recommendations for the Marketing Team

1. **Target High-Potential Client Segments**:
   * Focus on **students** and **retired** clients, who show higher subscription rates. Tailor messaging to their needs (e.g., flexible terms for students, secure investments for retirees).
   * Prioritize clients with **university degrees** or **professional courses**, emphasizing financial benefits and stability.
   * Target **single** clients with campaigns highlighting personal financial growth opportunities.
2. **Optimize Campaign Timing**:
   * Run campaigns in **October**, **September**, or **March**, when subscription rates are higher. Avoid oversaturation in **May**, which has high volume but lower success.
   * Schedule calls on **Tuesdays** or **Wednesdays**, as these days showed balanced engagement in EDA.
3. **Leverage Previous Campaign Success**:
   * Re-target clients with a **successful previous outcome** (poutcome=success). Offer personalized incentives to reinforce loyalty.
   * For clients with nonexistent or failure outcomes, use softer engagement strategies (e.g., educational content) before pitching.
4. **Enhance Call Effectiveness**:
   * Train agents to extend call duration (aim for >5 minutes) by building rapport and addressing client concerns, as longer calls strongly predict subscriptions.
   * Use **cellular** contact methods, as they are more effective and convenient for clients.
5. **Monitor Economic Conditions**:
   * Launch aggressive campaigns when **Euribor3m** is low (<2.0) and **consumer confidence** is high (>-30), as clients are more receptive during stable economies.
   * Adjust messaging during high euribor3m periods to emphasize security and low-risk benefits of term deposits.

## Conclusion

The EDA revealed a highly imbalanced dataset with key insights into client demographics, campaign timing, and economic influences. The Random Forest model identified duration, poutcome\_success, euribor3m, age, and specific month variables as the most impactful features. Clients likely to subscribe are often students, retirees, or single individuals with higher education, contacted via cellular in favorable economic conditions or specific months. The marketing team can optimize campaigns by targeting these segments, timing campaigns strategically, and leveraging economic trends.