Project: Customer Product Analysis

Author: Phan Thi Dinh

1. Problem Statement

The bank aims to increase the subscription rate of term deposits through outbound calling campaigns. The current success rate is low (~11%) while calling costs are high. Data analysis is needed to:

- Understand the characteristics of potential customers.
- Identify factors influencing subscription decisions.
- Build a predictive model to optimize the calling list.

2. Data Overview

- **Source:** Bank Marketing (UCI).
- **Size:** 41,188 customers with multiple features (demographics, contact history, communication channel, macroeconomic conditions).
- **Target:** Whether the customer subscribed (yes) or not (no).

3. Methodology

- **EDA:** Distribution of age, number of contacts, subscription rate by occupation, education, and communication channel.
- **Feature Engineering:** Age grouping, categorical encoding, normalization of quantitative variables.
- **Modeling:** Logistic Regression to explain factors affecting subscription likelihood.
- **Evaluation:** Accuracy, ROC–AUC, Confusion Matrix, Classification Report.

Exploratory Data Analysis (EDA)

2.1 Overall Subscription Rate

Only ≈11.3% of customers subscribed, indicating a significant class imbalance. This implies that predictive models and marketing strategies must carefully account for the cost of misclassification and optimize accuracy for the minority class.

2.2 Age

- Most customers are aged 30–45.
- Students and retirees show higher-than-average subscription rates.
- This reflects differences in financial goals: students focus on short-term savings, while retirees prioritize safety and fixed income.

2.3 Occupation and Education

- Blue-collar and services occupations show significantly lower subscription rates.

- Conversely, students, retirees, and management groups exhibit higher rates.
- Higher education (university or above) correlates with stronger subscription likelihood, suggesting financial literacy is a key factor.

2.4 Contact History and Communication Channel

- Customers with prior positive campaign outcomes (poutcome = success) are much more likely to subscribe.
- Comparing channels: mobile calls (cellular) outperform landline calls (telephone).
- The *campaign* variable (number of calls) shows a negative relationship: more calls → lower success rate → evidence of campaign fatigue.

2.5 Macroeconomic Variables

- Indicators such as **emp.var.rate** (employment variation rate) negatively affect subscription likelihood.
- **cons.price.idx** (consumer price index) correlates positively with subscription.
- This suggests customer financial decisions are sensitive to business cycles and inflation expectations.

3. Modeling Results

3.1 Overall Performance

- **Accuracy:** 0.7876
- **ROC AUC:** 0.7767 → fairly good discrimination ability.
- **Confusion Matrix:**
  - TP = 967, FP = 2200, TN = 8765, FN = 425
- **Classification Report:**
  - Precision (class=1): 0.305
  - Recall (class=1): 0.695
  - F1-score (class=1): 0.424

3.2 Interpretation

- The model identifies nearly 70% of actual potential customers (high recall).
- However, precision is only 30%: for every 10 predicted subscribers, only 3 actually subscribe.
- This results in many false positives, leading to wasted calling costs.

3.3 Cost–Benefit Analysis (Business Implication)

- Total predicted “YES” customers to call = 3167 (≈25.6% of the dataset).
- With precision = 30.5%, on average, ~3.3 calls are needed per successful subscription.

- **Break-even formula:** R > 3.275 × Ccall

  Where **R** = net revenue per subscriber, **Ccall** = cost per call
- Example: If Ccall = $1, then revenue per subscriber must exceed $3.3 to break even.

4. Key Insights & Recommended Actions

1. **Leverage customers with prior successful contact history**
   - Highest likelihood group.
   - **Action:** Develop dedicated retargeting campaigns and prioritize resources.
2. **Segment by occupation and age**
   - Retirees and students have distinct financial motivations.
   - **Action:** Personalize messages (safety for retirees, savings/interest for students).
3. **Optimize channel and frequency**
   - Mobile outperforms landline.
   - Limit maximum calls to 3 per customer to avoid fatigue.
4. **Monitor economic cycles**
   - When unemployment rises, campaigns should be adjusted or postponed.
   - Incentives (e.g., bonus rates) can mitigate defensive customer behavior.

5. Advanced & Extended Analysis

5.1 Rationale

The current model focuses on prediction but not profit optimization. Next steps must address class imbalance, cost-sensitive learning, and deeper causal analysis of calling effects.

5.2 Proposed Extended Models

- Logistic Regression with interaction terms (occupation × age × contact history).
- Decision Trees & Ensembles (XGBoost, LightGBM) for nonlinear effects.
- Uplift Modeling to identify customers truly influenced by calls.
- Probability Calibration to optimize decision thresholds for profit.

5.3 Expected Outcomes

- Higher precision@k for targeted customer groups.
- Clear decision rules: only call top X% customers above optimal probability threshold.
- Increased ROI through reduced false positives.