Author: KELVIN OFFEI ANIM

Dataset: bank-additional-full.csv

Objective

The goal of this project is to develop a predictive model to determine whether a client would subscribe to a term deposit, based on the features provided in the dataset. This will support the marketing team in targeting potential customers more effectively during future campaigns.

1. **Exploratory Data Analysis (EDA)**
   Findings from the exploratory data analysis are highlighted below:
   - Dataset had 41,188 rows and 21 input features.
   - The target variable y (yes/no) was highly imbalanced (approx. 88% no, 12% yes).
   - Several numeric features (e.g., age, campaign, pdays) showed skewed distributions.
   - Categorical variables like job, month, poutcome, and education showed strong associations with the target variable.
   - The feature, "duration" was dropped to avoid data leakage, as it directly affects the outcome.
   - Missing values were encoded explicitly (e.g., unknown, 999 in pdays was handled with a binary flag).

2. Feature Engineering

- Binary encoding was applied to features like 'default', 'housing', 'loan', and 'y'.
- A new feature pdays_was_999 was created to capture the unique interpretation of pdays == 999.
- All remaining categorical features were one-hot encoded
- Feature importance analysis (via Random Forest) was used to drop 18 features with very low predictive power (< 0.5%).

3. Model Building

The model used was a Random Forest Classifier, embedded in a pipeline that included:

- StandardScaler to normalize the features.
- SMOTE to handle class imbalance by oversampling the minority class in the training set.
- RandomForestClassifier as the main classifier.

Two models were trained:

- Baseline model with all features included.

- Reduced model after dropping the features with low importance (<0.005).

The pipeline was trained on an 80/20 train-test split.

4. Model Evaluation

Full Feature Model Results(with all features)

- **Accuracy:** 89%
- **Precision(yes):** 0.50
- **Recall (yes):** 0.37
- **F1 Score (yes):** 0.42
- **ROC AUC:** 0.773

Reduced Feature Model(after dropping low-importance features)

- **Accuracy:** 88%
- **Precision(yes):** 0.48
- **Recall (yes):** 0.37
- **F1 Score (yes):** 0.42
- **ROC AUC:** 0.771

5. Key Insights & Recommendations

- Important predictors included: month, contact type, previous outcome, job, and pdays_was_999.
- Clients more likely to subscribe are:
    - Clients contacted during certain months like March, December, and October.
    - Clients who had a history of previous contact.
- **Recommendations for Marketing:**
    - Calls should be focused on segments with a history of prior contact.
    - Emphasis should be made on outreach during more productive months.
    - Messages should be tailored based on employment and economic indicators.