Bank Term Deposit Subscription Prediction

Team 9: Srinivas, Ravi Programme : LTC - IIT M Batch 1 - Sep 2025

Introduction

This project addresses the prediction of term deposit subscription using both synthetic and real-world banking datasets. The goal is to build robust machine learning models that generalize well from synthetic data to real data, leveraging advanced techniques such as feature engineering, data balancing, and transfer learning.

Data Understanding & Exploration

Datasets Used
Synthetic Data: Provided for the P5S8 challenge, split into train and test sets.
Real Data: Extracted from the UCI Bank Marketing dataset (bank-full.csv).

Initial Exploration
Inspected column names, shapes, and feature types (numerical/categorical).
Checked for missing values and target variable distribution.
Compared feature distributions between synthetic and real datasets.

Key Observations
Both datasets have similar feature distributions.
The target variable y is encoded as 0/1 in synthetic data and as yes/no in real data (converted to 0/1).
Features pdays and previous showed mean shifts between datasets, indicating potential domain differences.

Data Preprocessing - Feature Engineering
Combined synthetic train and test sets for unified training.
Standardized numerical features using Z-score normalization.
One-hot encoded categorical features, except for education, which was ordinally encoded.

Handling Imbalanced Data
Applied SMOTE (SMOTENC for categorical features) to balance the target classes, improving minority class representation.
Feature Alignment - Ensured consistent feature sets between train and test by dropping/renaming columns as needed.

Exploratory Data Analysis
Compared summary statistics and value proportions for all features.
Visualized distributions and highlighted significant differences (>10% in means/proportions).
Used correlation matrices and KL divergence to quantify dataset similarity.
Performed chi-square tests for categorical variables to assess statistical differences.

Model Development

CatBoost Classifier
Trained on synthetic data, evaluated on real test data.
Initial results showed underperformance on minority class due to imbalance.

XGBoost Classifier
Hyperparameter tuning via GridSearchCV (parameters: n_estimators, max_depth, learning_rate, subsample, colsample_bytree).
Achieved ROC AUC > 90% on test data after balancing and feature engineering.

Transfer Learning
Reused engineered features and tuned hyperparameters from synthetic models.
Retrained on real data, leading to further improvement in ROC AUC (up to 95%).

Results & Learnings

Feature Distribution: Synthetic and real datasets are closely matched, but some features (e.g., pdays, previous) require normalization.
Imbalanced Classes: SMOTE significantly improved minority class prediction.
Model Performance: XGBoost with transfer learning outperformed CatBoost, achieving high ROC AUC.
Transfer Learning: Leveraging synthetic data for feature engineering and hyperparameter tuning accelerated model development and improved generalization.