Report on Bank Marketing Data Set Analysis and Model Deployment

Objective

This project focused on developing and deploying a machine learning model to predict whether a client will subscribe to a term deposit. The analysis made use of the "bank-additional.csv" dataset, containing 4119 instances (10% of the full dataset) and 20 input features. This report covers data analysis, preprocessing, feature engineering, model selection, and deployment.

Dataset Overview

The dataset originates from the UCI Machine Learning Repository. It includes demographic, economic, and campaign-related features aimed at predicting client subscriptions to term deposits.

Key Features:
- Demographic: Age, job type, marital status, education, etc.
- Economic: Employment variation rate, consumer price index, Euribor 3-month rate, etc.
- Campaign: Number of contacts during this campaign, number of days since last contact, etc.

The target variable `y` indicates whether a client subscribed to a term deposit ("yes" or "no").

Data Cleaning and Preprocessing

1. Handling Missing Values: No missing values were identified in the dataset.
2. Encoding Categorical Variables: Label encoding was applied to convert categorical features into numeric formats suitable for machine learning models.
3. Addressing Data Imbalance: Oversampling and undersampling techniques were employed to balance the dataset, addressing the 89.1% ("no") and 10.9% ("yes") distribution of the target variable.
4. Feature Scaling: Standardization was applied to numerical features for uniformity.

Exploratory Data Analysis (EDA)

1. Target Variable Distribution: The target variable was heavily imbalanced ("no": 89.1%, "yes": 10.9%).
2. Feature Insights:
   - Numerical Features: Histograms highlighted key distributions.
   - Categorical Features: Count plots provided us with insights into feature distributions.
3. Correlation Analysis: A heatmap revealed important correlations, especially between `emp.var.rate` and `euribor3m` (correlation: 0.97).

Feature Engineering

Feature Importance: “euribor3m” (0.197171) was identified as key feature influencing the target variable.

Model Selection

Three supervised learning models were evaluated:

1. Logistic Regression
2. Decision Tree
3. XGBoost
4. Random Forest

Evaluation Metrics:

- Accuracy
- F1-score
- RMSE
- ROC AUC (to address the imbalanced dataset)

Chosen Model: Random Forest has the highest ROC-AUC (0.750) and a low RMSE (0.323). It balances the training and test scores well, indicating no significant overfitting. It achieves a strong F1-score (0.90) for the subscribed class, making it well-suited for an imbalanced classification problem. Thus, Random Forest would be the final model.

Hyperparameter Tuning

GridSearchCV was used for hyperparameter optimization, focusing on maximizing the Random Forest model's performance. The optimized parameters improved prediction accuracy and robustness.

Deployment

A Streamlit web application was developed to deploy the model. Users can input client data to receive predictions ("yes" or "no") on term deposit subscriptions.

Deployment Link: https://banktermpredict-jdv3pnrdwvwhbgq6j2tmgx.streamlit.app/

Key Findings and Recommendations

1. Influential Features:

- Euribor 3-month rate highly impact predictions.
- Attributes like `age` and `nr.employed` also correlate with the target variable.

2. Marketing Strategy:

- Focus efforts on clients around 41 years old when Euribor rates are low.
- Target clients with recent prior contacts for improved campaign effectiveness.

Conclusion

The project successfully demonstrates machine learning's potential in optimizing marketing strategies. The deployed model helps decision-making by providing real-time predictions, improving marketing campaign efficiency.