Project Report: Predictive Modeling for Term Deposit Subscription

1. Introduction

This report presents a comprehensive analysis and predictive modeling project aimed at identifying clients likely to subscribe to a term deposit, based on data collected during a marketing campaign by a Portuguese bank. The goal was to build a reliable classification model, supported by data exploration, feature engineering, and model evaluation.

2. Exploratory Data Analysis (EDA)

To begin the analysis, the dataset was examined for structure, completeness, and basic patterns:

- Missing Values Check: No missing values were found in the dataset, indicating clean and complete data.
- Target Variable Distribution: The target variable ‘y’ (indicating subscription to a term deposit) was found to be highly imbalanced, with a dominant majority of "no" responses.
- Correlation Analysis: A heatmap was used to explore correlations among numerical variables. The target variable was encoded to numeric format (yes = 1, no = 0) to evaluate its correlation with other features.
- Mutual Information: Mutual information scores were computed to quantify the dependency between each feature and the target. The top 15 most informative features were visualized to guide model input selection.

3. Data Preprocessing and Feature Engineering

To prepare the data for machine learning, the following preprocessing steps were undertaken:

- Categorical Encoding: All categorical variables were transformed using one-hot encoding.
- Numerical Feature Scaling: StandardScaler was applied to normalize numerical columns, ensuring uniformity in scale across features.
- Class Imbalance Handling: Synthetic Minority Oversampling Technique (SMOTE) was used to address the significant imbalance in the target variable. This step was crucial in improving the model’s ability to detect the minority class (clients who subscribed).

4. Model Development

The predictive model was built using the following approach:

- Train-Test Split: The dataset was split into training and testing sets, with 90% reserved for testing to rigorously assess model performance.
- Model Selection: A Random Forest Classifier was selected due to its ability to handle non-linear data and provide feature importance insights. (class_weight='balanced') was used to further address the class imbalance.
- Model Training and Prediction: The model was trained on the training subset. Predictions

were made on the test set, and the results were evaluated using several performance metrics.

5. Model Evaluation

The model was evaluated based on the following metrics:

- Accuracy: Approximately 90%, indicating overall correctness of predictions.
- Precision: High precision, reflecting the model's ability to minimize false positives.
- Recall: High recall, crucial for ensuring actual subscribers are not missed.
- F1 Score: Balanced and high, indicating strong overall model performance.
- Confusion Matrix: Used to further break down true positives, false positives, true negatives, and false negatives.

The use of SMOTE significantly enhanced the model's ability to generalize across both classes, particularly the underrepresented "yes" class.

6. Feature Importance Analysis

The Random Forest model provided feature importance rankings, which were visualized to gain business insights:

- Most Influential Features:
  - 'duration' (length of the last contact) was the single most impactful predictor.
  - Other top features included 'euribor3m', 'nr.employed', 'pdays', 'cons.price.idx', and 'emp.var.rate'.

These findings suggest that both client behavior (e.g., call duration) and external economic indicators strongly influence the likelihood of a subscription.

7. Model Deployment and Artifacts

The following components were saved to facilitate deployment and reproducibility:

- Trained model file (model.pkl)
- Scaler used for preprocessing (scaler.pkl)
- List of feature names (feature_names.pkl)
- List of numeric features (numeric_features.pkl)

To enhance accessibility and usability of the model, I developed and deployed an interactive Streamlit web application. This application allows users to input client data through a user-friendly interface and instantly receive a prediction on whether the client is likely to subscribe to a term deposit. This makes the model not only technically sound but also practical for business users who may not have a technical background.

The Streamlit app demonstrates how machine learning can be integrated into real-time decision-support tools, providing valuable insights during marketing or client interaction activities.

8. Conclusion

This project successfully delivers a robust predictive model for identifying clients likely to subscribe to a term deposit. Through a structured pipeline that included EDA, careful preprocessing, SMOTE balancing, and Random Forest modeling, the solution achieved strong predictive performance and offered meaningful business insights.

The project demonstrates the value of machine learning in marketing strategy optimization and lays the foundation for data-driven targeting of potential clients in future campaigns.