Group Name: Data Adventurers

Name: Shatabdi Pal

Email: shatabdi@pdx.edu

Country: USA

College/ Company: Portland State University

Specialization: Data Science

Problem description:

ABC Bank aims to optimize its marketing efforts to promote a new term deposit product by developing a predictive model. The model will analyze customer data from past interactions to forecast the likelihood of each customer subscribing to the term deposit. By accurately identifying potential buyers, the bank can focus its marketing resources on high-probability customers, enhancing efficiency and reducing costs. This project involves creating and evaluating models with and without the 'duration' feature, a critical element that will ensure practical applicability and performance.

GitHub Repo link: https://bitbucket.org/shatabdi_workpace1/bankmarketing/src/main/

EDA performed on the data

I conducted a comprehensive Exploratory Data Analysis (EDA) to uncover the underlying patterns and relationships within the dataset. The below steps are used for EDA.

Dataset Overview:

- The dataset consists of categorical and numerical features related to a Portuguese banking institution's direct marketing campaigns (phone calls).

- The dataset has 21 features, including the label. The target variable y is binary (‘yes’ and ‘no’).

Missing Value:

- The dataset does not contain any null values.

- Categorical attributes like 'job,' 'marital,' 'education,' 'default,' 'housing,' and 'loan' have 'unknown' values, which can be considered missing.

- During the data cleaning process, categorical variables with 'unknown' values were carefully imputed with the most frequent category, ensuring the integrity of the dataset.

Descriptive Statistics:

- Summary statistics (mean, median, mode, standard deviation) were calculated for numeric features to understand the distributions.
- Frequency counts were computed for categorical features to identify the most common categories.

Data Visualization:

- Histogram and Box plots were created for numeric features to visualize distributions and identify outliers.

- Bar charts were used for categorical features to show the frequency of each category.

- Pair plots and correlation matrices were generated to explore relationships between numeric features and the target variable.

- The heat map is used to show the correlation between features.

Outlier Detection:

- Outliers were identified using box plots.

- The impact of outliers on the model was considered, and the IQR method was used to handle those.

Feature Relationships and Feature Engineering:

- They analyzed relationships between features and the target variable using group by operations.

- I visualized these relationships with bar charts and box plots to identify significant patterns.

- Features like 'duration,' 'campaign,' and 'previous' have a skewness greater than 1. A square root transformation is applied to normalize skewed features.

- New features, such as Total Contacts, Economic Indicators Ratio, and Interaction between features, were created using some existing features that will be used during model training.

Data Imbalance:

- The target variable was imbalanced, with a higher proportion of 'no' responses than 'yes.'

- Techniques such as SMOTE (Synthetic Minority Oversampling Technique) were considered to handle this imbalance during model training.

Final Recommendation

Based on the EDA and subsequent model development, the following recommendations are made:

1. Model Choice:

After evaluating various models, logistic regression, ensemble model using random forest classifier, gradient boosting classifier, voting classifier, and Adaboost using decision tree classifier can be used for implementation.

Here are the reasons for model choice

Logistic Regression: Simple and interpretable, effective for linearly separable data, providing probabilistic outputs.

Random Forest Classifier: Robust to overfitting, handles non-linear relationships, and ranks feature importance well.

Gradient Boosting Classifier: This classifier builds models sequentially to correct errors. It is effective for complex data and often yields high accuracy.

Voting Classifier: Combines multiple models to improve generalizability and robustness, leveraging the strengths of each.

Adaboost using a Decision Tree Classifier: This classifier focuses on misclassified instances, enhances weak learners, and improves prediction accuracy.

2. Feature Selection:

Excluding the 'duration' feature from the final model is not practical for real-time prediction, and its inclusion would lead to unrealistic performance expectations. Focus on features related to client demographics, past interactions, and economic indicators.

3. Handling Imbalance:

Apply SMOTE techniques to balance the target variable during training.

4. Model Evaluation:

Evaluation of the final model using the best performance metrics (accuracy, precision, recall, F1 score, and ROCAUC) can be done. Converting ML metrics into business metrics to quantify the potential impact on marketing efforts and cost savings.

5. Deployment and Monitoring:

The model deployment can be done in a cloud environment using tools like Flask, Azure, or Heroku. Implement a monitoring system to track model performance and periodically retrain with new data to maintain accuracy.

6. Business Communication:

Preparing a comprehensive presentation for nontechnical stakeholders, explaining the model's impact on marketing efficiency and cost savings. Highlighting the key features influencing the prediction to provide insights into customer behavior.

By following these recommendations, ABC Bank can effectively target customers for their term deposit product, optimize marketing resources, and achieve better conversion rates.