Report on Bank Marketing Data Set Analysis

and Model Deployment

Object

This project focused on constructing and deploying a machine learning model to predict whether a client will subscribe to a term deposit, using the Bank Marketing Data Set. Specifically, the "bank-additional.csv" file was utilized, containing 10% of the examples (4119 instances) randomly selected from the full dataset. The dataset includes 20 input features and a target feature. The process covered extensive data analysis, data cleaning and preprocessing, feature engineering, model selection and optimization, and finally, deployment via Streamlit.

Dataset Overview

The dataset was sourced from the UCI Machine Learning Repository, containing 4119 instances with 20 feature values. These included age, job type, marital status, education level, economic indicators, and more. The main objective is to use feature values to predict whether a client will subscribe to a term deposit, or not.

Data Cleaning and Preprocessing

- No missing value is found in the dataset.
- Converted categorical variables into a suitable format for machine learning models by applying Label Encoding.
- Checked for data imbalances and found that the data set was imbalanced, then applied oversampling and undersampling techniques.
- Standardized numerical features to ensure uniformity by feature scaling.

Exploratory Data Analysis (EDA)

- Conducted exploratory data analysis to understand the distribution of variables.
- To evaluate imbalances within the target variable, two distinct visualizations were utilized: a bar chart and a pie chart. The analysis reveals a noticeable imbalance in the target variable, with the distribution of "no" constituting 89.1%, while the distribution of "yes" accounts for only 10.9%.
- The distribution of numerical features in the dataset was illustrated using the histograms and the distribution of categorical features was depicted through the counterplots.
- The distribution of each feature value relative to the target variable 'y' was visualized using boxplots and legends.
- We have discussed the correlations between variables shown by the correlation matrix. We have realized that some variables seem to influence each other like emp.var.rate and euribor3m with a correlation of 0.97.

Feature Engineering

Added 'previous_contact' and 'unemployed', derived from existing data, to provide more insights into the clients' profiles.

Model Selection

We have evaluated 3 different supervised learning models: Logistic Regression, Random Forest, and Neural Networks. Then we compared the results using evaluation metrics such as accuracy, and f1-score. Also, we have used ROC AUC to get a more accurate result for the performance of the model since the target variable is imbalanced. Based on the metrics, Random Forest was the most appropriate model.

Hyperparameter Tuning

Performed hyperparameter tuning using GridSearchCV for the Random Forest model to find the most efficient hyperparameters and maximize the performance.

Creating Pipeline

We have wrapped the preprocessing, feature engineering, and hyperparameter tuning using pipelines to increase readability and consistency. This helps in organizing and structuring the code, making it easier to manage, reproduce, and deploy machine learning models.

Model Deployment

We have created an additional Python script to deploy our model using Streamlit. Our simple web application enables users to interact with the model and predict term deposit subscriptions. The interface allows for easy input of client data and displays the prediction outcome.
Interact with the deployed model here: https://bank-deposit-prediction.streamlit.app/

Instructions on How to Interact with the Deployed Model

Users should enter the input features of a client into the respective input fields. For example, the user should enter a number indicating the age of the client to the “Age” input field. After entering all the relevant inputs, the user can predict if the client will issue a deposit or not by clicking on the button titled “Bank Marketing Prediction”. The result will be shown under the button as “1” indicating “yes” and “0” indicating “no”.

Key Findings and Recommendations

- **Influential Features**: According to the heatmap, economic context attributes like pdays (number of days that passed by after the client was last contacted from a previous campaign) and nr.employed (number of employees) were highly correlated with the target variable.

- **Targeting Strategy**: Recommend focusing marketing efforts on clients identified as likely subscribers to improve campaign effectiveness. According to the feature importance bar chart, “age” and “euribor3m” attributes are highly important and affect the target variable. Clients who issue deposits have a mean age of 41, and a mean euribor3m of 2.14. Therefore, marketing efforts should be towards clients of age around 41 and when the Euribor 3 months rate is low.

Conclusion

The project successfully demonstrates the application of machine learning in a practical banking scenario, enhancing the efficiency of marketing strategies. The deployed model on Streamlit serves as an effective tool for real-time predictions, aiding decision-making processes.