Objective

This project aims to develop a machine learning model to predict whether a bank client will subscribe to a term deposit. The dataset utilized is the Bank Marketing Data Set, available at https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Specifically, the "bank-additional.csv" file with 10% of the examples (4119), randomly selected from the entire dataset and containing 20 input variables, will be used. The data pertains to direct marketing campaigns (phone calls) conducted by a Portuguese banking institution. The classification objective is determining if a client subscribes to a term deposit (variable y). These marketing campaigns involved multiple phone calls to some clients to ascertain whether they would subscribe to the bank's term deposit product ('yes') or not ('no').

Summary of Dataset

This dataset, obtained from the UCI Machine Learning Repository, comprises 4119 instances, each characterized by 20 features. These features encompass variables such as age, occupation, marital status, education level, and economic metrics. The primary aim is to utilize these feature values to forecast whether a client will opt to subscribe to a term deposit.

Data Cleaning and Data Preprocessing

No missing values were detected within the dataset.

Categorical variables were transformed into a format compatible with machine learning models using Label Encoding.

Overwhelming and undersampling techniques were implemented to address data imbalances after identifying the dataset's imbalance.

Numerical features were standardized through feature scaling to ensure consistency across the dataset.

Exploratory Data Analysis (EDA)

Exploratory data analysis was performed to grasp the distribution of variables.

We employed a bar chart to assess imbalances within the target variable, revealing a significant imbalance.

Histograms were utilized to showcase the distribution of numerical and categorical features within the dataset.

Boxplots were generated to visualize the distribution of each feature value concerning the target variable 'y'.

Feature Engineering

Enhancements were made by introducing additional features such as 'previous campaign success rate', 'age education level', and 'pdays group'. These features, derived from existing data, offer deeper insights into the profiles of clients.

Model Selection

We assessed three supervised learning models: Logistic Regression, Random Forest, and Neural Networks. Subsequently, we compared their performances utilizing evaluation metrics like accuracy and f1-score. Additionally, we employed ROC AUC to obtain a more precise evaluation of model performance due to the imbalanced nature of the target variable. Random Forest emerged as the most suitable model based on these metrics.

Hyperparameter Tuning

We used GridSearchCV to conduct hyperparameter tuning for the Random Forest model. This approach allowed us to identify the optimal hyperparameters and enhance performance to its fullest potential.

Creating Pipeline

Using pipelines, we encapsulated the preprocessing, feature engineering, and hyperparameter tuning stages to enhance readability and maintain consistency in the code. Such structuring aids in organizing the codebase, simplifying management, facilitating reproducibility, and easing the deployment of machine learning models.

Model Deployment

We have developed an extra Python script to deploy our model utilizing Streamlit. This straightforward web application empowers users to engage with the model and forecast term deposit subscriptions. The interface facilitates effortless input of client data and showcases the prediction results. Access and interact with the deployed model through this link:

Guidance for Interacting with the Deployed Model

Users need to input the client's details into the corresponding fields. For instance, the client's age should be entered into the "Age" field. Once all relevant inputs are provided, users can predict whether the client will subscribe to a deposit by clicking the "Bank Marketing Prediction" button. The outcome will be displayed beneath the button as either "1" for "yes" or "0" for "no."

Key Findings

The dataset presents comprehensive insights into the demographics of individuals targeted by marketing campaigns, encompassing age, occupation, marital status, education level, and loan status. It also outlines the various communication channels for marketing outreach, including phone and cellular methods. Additionally, past campaign outcomes, particularly subscription rates to term deposits, offer valuable insights into historical performance. Economic indicators such as the consumer price index and employment variation rate further contextualize the effectiveness of marketing efforts.

Recommendations

To enhance campaign effectiveness, leveraging demographic insights for targeted segmentation is recommended. Optimization of communication channels should be prioritized, focusing on those that demonstrate the highest response rates. Analyzing past campaign outcomes can inform strategies for future success. Adapting marketing approaches in line with economic indicators can ensure relevance and resonance with consumers. Personalization of messaging based on audience preferences is vital to fostering deeper engagement. Balancing contact frequency is crucial to avoid audience fatigue. Lastly, continuous experimentation and refinement based on insights from the dataset are essential for maximizing campaign effectiveness over time.

Conclusion

In summary, this project showcases the practical implementation of machine learning within a banking context, improving the effectiveness of marketing tactics. The model deployed via Streamlit is a valuable resource for making real-time predictions, thereby supporting decision-making procedures.