Predicting Term Deposit Subscription Using Machine Learning

(ADA 442 Project Report)

ADA 442: Statistical Learning

Dr. Hakan Emekçi

May 18, 2025

Group 13

Umay Çelik

Esra İkbal Narin

Mustafa Karakuş

Efe Emür

INTRODUCTION

This project aims to build a machine learning model that predicts whether a client will subscribe to a term deposit or not. The dataset of this project contains client information from a Portuguese bank's telemarketing campaigns, and it includes random selected 10% (4119 instances) clients with 20 input features and a target feature in the sample. The project follows a complete machine learning pipeline including data cleaning, feature engineering, preprocessing, model training, evaluation, and deployment to create a model that provides practical and realistic insights about clients' subscription tendencies.

DATA PREPARATION

For the beginning; "pandas", "numpy", "matplotlib", "seaborn", "scikit-learn" and "imblearn" libraries were used for data processing, visualization and modeling; "pickle" was used for model registration. The dataset was loaded from the 'bank-additional.csv' file, and its structure consisting of 4119 rows and 21 columns was examined. For a first look, general information about the data structure was obtained through observations. Variable names were sorted. (e.g. "age" → "Age"). The structure of the dataset was reviewed again and checked again for deficiencies. The relationships between numerical variables were analyzed with the correlation matrix.

Outliers were detected; the most outliers were observed in the "PreviousCampaignContacts" and "CallDuration" variables. Winsorize applied to outliers. The "PreviousCampaignOutcome" column with a lot of missing data was deleted, and other missing values were filled with "mode".

Data Preprocessing

The dataset was imbalanced, as it includes 89% "No" and 11% "Yes". Data splitted into train and test sets (80/20). To handle imbalance, four resampling methods were tested using logistic regression and F1 score: no resampling, SMOTE, Random Under Sampling, and SMOTE-Tomek. SMOTE gave the best F1 score (0.467) and was chosen for further modeling.

Feature Engineering and Selection

'CallDuration' was divided into four categories: 'Very Short', 'Short', 'Medium', 'Long' based on predefined time intervals so, the model captured non-linear relations. A new categorical variable 'Season' was created. Education levels were ordinally encoded. Original columns 'CallDuration', 'LastContactMonth', and 'Education' were dropped to avoid redundancy. Low-variance features like 'Credit' were removed as it caused high imbalance. Similarly, 'PreviousContactDay' was changed with a binary indicator called 'previous_contact'.

Encoding and Future Selection

‘SubscribedTermDeposit’ was transformed into binary form by label encoding. One-hot encoding was applied to categorical features to enable numerical processing. Three different feature selection methods were used: SelectKBest to identify the top features with the strongest individual relationship, Recursive Feature Elimination (RFE) with Random Forest, to iteratively remove the least important features and, feature importances. Results were combined, and only the most relevant predictors were kept for creating final feature set.

MODELING AND EVALUATION

Hyperparameter Tuning

Hyperparameter tuning was performed for Logistic Regression, Random Forest, Gradient Boosting, and KNN using GridSearchCV with five-fold cross-validation and F1 score. Each model had a separate parameter grid. Gradient Boosting showed the best performance (F1=0.5813), followed by Random Forest (F1=0.5043), Logistic Regression (F1=0.4808), and KNN (F1=0.4597). Therefore, Gradient Boosting was selected as the final model.

Deployment

Final evaluation was conducted on the best models using metrics such as accuracy, AUC, precision, recall, F1-score, and confusion matrix. **Gradient Boosting performed the highest F1-score (0.5689).** According to this strong performance and interpretability it was chosen for deployment. Feature importance was visualized, and SHAP values were used to explain individual contributions. A complete inference pipeline was built, and all necessary files—model, features, label encoder, and evaluation info—were saved using pickle for deployment.

RECOMMENDATIONS

While testing the Streamlit, the model mostly predicted “no,”. A “yes” only appeared when the Consumer Confidence Index was approaching high values. This indicates that the model gives this variable too much weight, probably due to positive (“yes”) examples in the training set. To fix this, creating a wider variety of “yes” cases using SMOTE and adding different combinations of features, such as Call Duration Category, Euribor 3M and previous contact. Adding new features based on customer behavior and adjusting the sampling method could improve the model’s potential subscribers and provide more balanced predictions.

CONCLUSION

This project performed a machine learning model to predict term deposit subscription. After different stages, Gradient Boosting was selected for deployment due to its strong F1-score and interpretability. While testing the Streamlit, the model showed tendency to predict “no,” which highlights potential issues with class imbalance. However, the model serves as an effective tool for real-time predictions.

Streamlit cloud address:

https://app2-ada442.streamlit.app/