ADA 442 Statistical Learning | Final Project Assignment

Instructor: Dr. Hakan Emekci

Banu Genç

Orhan Can Öcal

Ozan Alp Sarıdoğan

Yusuf Eren Pekmezci

Introduction

This project aims to develop a classification model to predict whether a customer will subscribe to a term deposit or not. The dataset used in this project belongs to a marketing study led by the Portuguese banking institute. A random sample consisting of 4119 instances, approximately %10 of the full dataset, was used for this project. Various techniques, including data analysis, data cleaning, preprocessing, model training and evaluation, were applied. Finally, a user-friendly UI is developed using streamlit to allow for testing the model in a simple environment.

Data Cleaning

As a first step, exploratory data analysis was conducted to understand the structure and quality of the dataset. The data was analyzed for missing values, duplicate rows, and the number of unique values in each column. No missing values and duplicate rows were found, but the missing values are recorded as unknown. The distribution of the target variable showed that the dataset is highly imbalanced. Moreover, the distribution of each future were analyzed to detect any patterns that could manipulate the model's performance. These findings informed the preprocessing steps to ensure that the dataset was properly prepared for modeling.

Data Preprocessing

During the preprocessing phase, categorical and numerical features were distinguished. Missing values ('unknown') converted to NaN. For categorical features, missing values were imputed using the simple imputer, while for numerical features, the median was used to preserve the distribution and reduce the influence of potential outliers. To prepare categorical variables for ML models, label encoding was applied, transforming them into numerical values the algorithms can interpret. Given the significant dataset imbalance in the target variable (y). The SMOTE (Synthetic Minority Over-sampling Technique) method was used to synthetically generate new instances of the minority class. Additionally, to improve the data quality, outlier detection and removal were performed using the Isolation Forest algorithm. This approach takes the %1 of the most anomalous samples based on the distribution of numerical features and removes outliers from the dataset to train a more robust model.

Feature Engineering

Not all 21 variables were utilized for training the models. A correlation matrix was constructed to assess the relationships among features, and independent variables were selected based on their importance scores. Additionally, to enhance model accuracy, a new feature named economic_index was engineered. This feature represents a composite of four key macroeconomic indicators: employment variation rate (emp.var.rate), consumer price index (cons.price.idx), number of employed individuals (nr.employed), and the 3-month EURIBOR interest rate (euribor3m). While each of these variables holds individual significance, their isolated impact on customer decision-making may be limited. By aggregating them into a single representative feature, the model is better equipped to capture underlying economic influences, thereby improving its predictive performance.

Model Comparison

Given the dataset's imbalance, accuracy alone is insufficient for evaluation. Instead, recall and F1-score are emphasized to assess how well models identify the minority class—subscribers. XGBoost strikes the best balance with a precision of 0.64, recall of 0.48, and an F1-score of 0.55. While Logistic Regression had the highest recall (0.88), its low precision (0.45) leads to many false positives. Other models like KNN and Random Forest also had low recall, making them less effective. In conclusion, XGBoost is chosen as the most suitable model, offering strong performance in identifying true subscribers while keeping false positives at a reasonable level.

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.91	0.64	0.48	0.55
Random Forest	0.91	0.67	0.27	0.38
Logistic Reg.	0.87	0.45	0.88	0.59
KNN	0.91	0.71	0.28	0.40
Decision Tree	0.88	0.45	0.37	0.40

Recommendation

Our suggestion is that the bank should not be deceived by the accuracy score of the models as they are working with unbalanced data. Recall and precision scores are the scores that better represent subscriber capture among the models. The model may correctly know that 910 out of 1000 customers will not subscribe, but a model that does not capture any subscribers and seems to be successful is of no use to the bank. At this stage, we recommend the pipeline of the XGBoost model to the bank's account manager and marketing team. With the help of Stream-lit, the marketing team can enter customer data into the model and get instant predictions, moreover, it can help campaign planning with the instant output of the model.