

ADA 442

Statistical Learning

Project Report

Group 11

Alp Özdemir

Bora Kılınç

Ahmet Kağan Taşcı

Ömer Emre Ay

Introduction

The objective of this project is to develop a machine learning model to predict whether a client will subscribe to a term deposit based on the data from direct marketing campaigns of a Portuguese banking institution. The dataset, obtained from the UCI Machine Learning Repository, consists of various attributes related to client demographics and past marketing interactions.

Data Cleaning

We began by loading the dataset bank-additional.csv, which contains 4119 examples with 20 input features and the target variable y (indicating whether the client subscribed to a term deposit). Initial inspection of the dataset revealed no missing values. Descriptive statistics and data types were checked to ensure data integrity.

Data Preprocessing

Encoding Categorical Variables

Categorical variables were transformed using one-hot encoding to convert them into a format suitable for machine learning algorithms. This process was achieved with the `pd.get\_dummies` function, which created binary columns for each category, excluding the first category to avoid multicollinearity.

Numerical features were scaled using `StandardScaler` to standardize the data. This transformation ensured that all numerical features had a mean of zero and a standard deviation of one, which is essential for algorithms sensitive to feature scales, such as logistic regression and neural networks.

Feature Selection

We performed feature selection to identify the most relevant features. This was achieved by examining the correlation of each feature with the target variable `y\_yes`. Features with an absolute correlation greater than 0.2 were selected for further analysis. The selected features were visualized using a heatmap to observe their inter-correlations.

Exploratory Data Analysis

Distribution of Numerical Features

We analyzed the distribution of numerical features to understand their spread and central tendencies. Histograms and boxplots were created for each numerical feature, providing insights into their distributions and potential outliers.

Categorical Features Analysis

For categorical features, bar plots were generated to visualize the frequency distribution of each category. This helped in understanding the prevalence of different categories and their potential impact on the target variable.

Model Training

Handling Class Imbalance

The target variable `y\_yes` exhibited class imbalance, with a higher number of 'no' responses compared to 'yes'. To address this, we applied Synthetic Minority Over-sampling Technique (SMOTE) to the training data, which balanced the classes by generating synthetic examples for the minority class.

Model Selection

We compared three classifiers: Logistic Regression, Random Forest, and Neural Network. These models were chosen for their diverse approaches to classification.

Hyperparameter Tuning

Hyperparameters for each classifier were tuned using `GridSearchCV` with 5-fold cross-validation. The parameter grids included:

- Logistic Regression: C values of 0.1, 1, and 10

- Random Forest: n\_estimators values of 100, 200, and 300

- Neural Network: hidden layer sizes of (50,), (100,), and (200,)

The pipeline structure allowed for seamless hyperparameter tuning and model comparison.

Results

Model Performance

The final model's performance was evaluated on the test set, yielding the following classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| No | 0.98 | 0.86 | 0.92 | 7303 |
| yes | 0.45 | 0.87 | 0.60 | 935 |
| accuracy |  |  | 0.87 | 8238 |
| Macro avg | 0.72 | 0.87 | 0.76 | 8238 |
| Weighted avg | 0.92 | 0.87 | 0.88 | 8238 |

The model achieved an overall accuracy of 0.91. Both precision and recall scores were high, indicating the model's reliability in predicting client subscriptions to term deposits.

Feature Importance

For the Random Forest classifier, feature importance was analyzed to understand which features contributed the most to the predictions. The top features included `duration`, `campaign`, and `pdays`. A bar chart was created to visualize the importance of each feature, highlighting their impact on the model's decisions.

Conclusion

In this project, we effectively preprocessed the dataset, handled class imbalance, and trained a Gradient Boosting Classifier to predict client subscriptions to term deposits. The model was thoroughly evaluated and demonstrated strong performance with high accuracy, precision, and recall scores. The trained model was saved for future use, and an example prediction was provided to illustrate its practical application.

Future Work

Future improvements could include exploring other machine learning algorithms, such as XGBoost or LightGBM, to potentially enhance model performance. Additionally, incorporating more features or external data sources could further improve prediction accuracy. Implementing the model in a real-time prediction system for the bank could also be a valuable next step.

Model Deployment

The trained model was saved using pickle for later use and deployment. Instructions for deploying the model using Streamlit are provided separately, allowing for interactive predictions based on new client data.