ADS_542 Term Project Report

Duygu Halim Kırlı

Data

- The data is related with Portuguese banking institution. There are 20 variables related to clients and camping information. Target variable y - has the client subscribed a term deposit? (binary: 'yes', 'no'). The main data set randomly is separated by % 10 and constituted as new data set “Bank-Additional”. This dataset covers 4119 rows and 21 columns.
- Data set contains both numeric variables like “age”, ”duration”, ”eurobor3m” and categorical variables like “job”, “marital”, “education”. The data has no null data so there is no need to impute the null data.

Data Visualising

- Data is an imbalanced one. Clients having no deposit subscription is almost %90 and having deposit subscription is only % 10.
- For Exploratory Data Analysis, Numeric and categorical variables are plotted separately. And then numeric varibles distribution and categorical variables distribution by target variable are analysed.
- In following part, correlation of numeric variables are analysed by heatmap. In this section, categorical variables are transformed numeric ones by label encoder and they are added to heatmap also. According to correlation matrix, “y” is mostly related to “duration” and has negative high correlation with “nr.employed”, “pdays”, “euribor3m” and “emp.var.rate”.

Data Preprocessing

- There is no empty cell so no need to imputation.
- “Duration” is dropped since it is highly correlated with target variable to avoid multicorrenality.
- A new varible is added in terms of feature engineering. “Risky” variable contains high risk group in job who is student and unemployed and also default is yes.
- Two encoding type one-hot encoding and label encoding are tried.
- The data is splitted two part as %80 train and %20 test.
- Since data is imbalanced, by using SMOTE technique, data is made balanced.
- Then since each variable have different scale, by Standart Scale method, the data is scaled again.

Feature Selection

- In this part, embedded methods are chosen and Lasso technique is used to define important variables.

Model Selection

- Since data set is not small, Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) are not used. 4 model Logistic Regression, Random Forest, Neural Network and XGBoost are used.

Model	Accuracy	Precision (1)	Recall (1)	F1 (1)
Random Forest	0.893	0.50	0.38	0.43
Logistic Regression	0.811	0.31	0.57	0.40
XGBoost	0.875	0.42	0.29	0.34
Neural Network	0.850	0.35	0.41	0.38

- According to the result, best model is RandomForest which has highest accuracy. However, like other ones, this model has problems in terms of low precision, recall and F1 score. Also, ROC AUC matrix is evaluated to choose the best model.

Definition of Pipeline

- The previous stages of a pipeline are listed in order: label encoder, splitting data, sampling by SMOTE, scaling by StandartScaler and feature selecting by Lasso. And a pipeline is created.

Hyperparameter and GridSearchCV

- By GridSearchCV, hyperparameter turning is performed to find best parameters maximizing for RandomForest model.

Evaluation

- Although model accuracy is high enough, model is not very good to in minority class. To make better, random forest model is used with “class weight=balanced” and get a bit better result. But it still have imbalanced class problem.

```
Accuracy: 0.8932038834951457
              precision    recall  f1-score   support

           0       0.92      0.96      0.94       732
           1       0.53      0.37      0.44        92

    accuracy                           0.89       824
   macro avg       0.73      0.66      0.69       824
weighted avg       0.88      0.89      0.88       824
```

Streamlit

Links: https://ads542bankmarketing-ifkfrxornaghodu5zweob9.streamlit.app/

https://github.com/DuyguHalim/ADS542_bank_marketing