ADA 442 Statistical Learning

Author: Özge Sena Karabıyık & Korhan Deniz Akın & Kerem Irmak

Created: 6 June 2023

Final Project Assignment

Introduction

This project aims to develop a machine learning model that predicts whether bank customers will subscribe to a term deposit or not. The dataset used for this project is provided from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. The dataset has 20 attributes, including target variable, and 4119 instances. This project involves data cleaning, data preprocessing, feature selection, model selection, hyperparameter tuning, evaluation and deployment. For these steps, Python was used as the programming language, Jupyter Notebook and Visual Studio Code were used as the environment, and Streamlit with python code support was used for distribution.

Project Steps

1. Data Cleaning

The data cleaning process involves data formatting, handling missing values and dealing with duplicates. Formatting data ensures consistency in data, formatting done by using Excel VBA and convert resulting file to format .csv in order to use it properly rest of the project. The existence of missing values and duplicate values was determined by codes, since they did not exist, no special steps were applied for these processes.

2. Data Preprocessing

In this project, ordinal encoding for variables with ordered categories, one-hot encoding for variables without an order, and standard scaling and min-max scaling for numerical variables are used. These methods guarantee appropriate categorical variable handling and numerical variable standardization, enabling accurate model training and evaluation. These methods are implemented with the pipeline, using the methods of including the necessary libraries, with the support of the ColumnTransformer() method. Creating new features not be done because it was not found appropriate for this project since it can introduce complexity and lead to overfitting and also lead to noise in data which already has much attributes. Effectively using existing features rather than adding new features has been a priority.

3. Feature Selection

As there are lots of features on the given dataset, it was necessary to apply some feature selection before training our model. However, since we are not able to know which feature might be useful, and which might not, manually doing so was not the best option. We have used Scikit Learn’s SelectFromModel algorithm which is useful to extract the best features from the trained model. It is doing so by assigning feature importance values to each feature. If the feature importance values are lower than some threshold these features are removed from the model. Additionally, regularization has been applied to some of our experimented models.

4. Model Selection

Experimented on 3 different models as it is hard to predict which one is the most accurate. Logistic Regression, Decision Tree Classifier, and Random Forest Classifier algorithms have been used. Pipelines are used to apply pre-processing, which includes data encoding and scaling. After finding the best model, finding its hyperparameters was crucial to obtain better performance.

Hyperparameter tuning, which means finding the best parameters on such a model, has been applied with the help of the Grid Search CV algorithm. It applies the combination of some given parameters such as learning rate, n-degree polynomials, etc. After applying grid search with the pipeline we obtained the best parameters for the best model.

Experiments among three models (as stated above) gave the result that Logistic Regression is the most accurate. Also, as a result of hyperparameter tuning:

logisticregression__C (inverse of regularization strength): 0.1

logisticregression__penalty (l1 means Lasso, l2 means Ridge regression): l2. Meanly applying Ridge Regression produced better accuracy.

logisticregression__solver: liblinear were results.

5. Evaluation

To evaluate our model we used: accuracy, precision, f1 score, and confusion matrix metrics. Accuracy refers to the ratio of true predictions and all predictions. Precision is the ratio of true positives and the sum of true positives and false positives. F1 score was the another metric that we used however, it is not the most necessary. Lastly, a confusion matrix is observed with true positives and true negatives on the main diagonal and false positives and false negatives on the second diagonal.

As a result of these performance metrics for the best model observed is:

Accuracy: 0.9085760517799353,

Precision: 0.6125,

F1 Score: 0.46445497630331756

Confusion Matrix:

```
[[1074   31]
 [  82   49]].
```

6. Streamlit And Publishing

1- After training the Model the first thing we do is to install Streamlit. We do that by running ‘pip install streamlit’ in our command prompt.

2- Then we import any libraries that are to be used. In our case these are streamlit, pandas, numpy, pickle, sklearn, matplotlib. pyplot and warnings(to be able to ignore warnings)

3- We design and implement the user interface that is going to help us take inputs from the user. We used some dropdown menus and some basic number input functions for these.

4- We then load the trained model with the help of pickle and push our input into the model.

5- And with the help of a button that says press me, you get the expected outcome.

You can visit the application here : https://kdakn-ada442-streamlit-stuff-twv5n2.streamlit.app/

Conclusion

In this project, we develop a machine learning model to predict whether a bank client will subscribe to a term deposit. The above-mentioned steps were implemented in the project, and the model with the highest accuracy value, selected by comparing the accuracy values, was supported by the interface. Project was completed by developing an application that informs the user whether subscribe to a term deposit or not, if the user enters the required values.