Friday, July 1, 2022

Project One-pager:

# Data Summary:

The data consists of 52 columns, one of them being the target variable (the dependent variable) consisting of ‘0’s and ‘1’s. (Column AZ). The rest of the data are the remaining 51 columns, and those are my features (the independent variables). The whole data sits in a .csv file. The data types are as follows: **dtypes: float64(48), int64(1), object(3)**.

# General Method:

Since, the target variable is binary, therefore this is a **binary classification challenge**. And, since we have dependent and independent variables, our learning type is supervised. I used Spyder IDE (Python 3.9) for the project.

# Classification Algorithm Used:

After testing out six alternative classification algorithms, I decided that the **logistic regression** algorithm performed the best (f1 score: ‘0’= 0.98, ‘1’=0.99). Detailed discussions and the method I followed are in the comments in the code (.py file).

# Deliverables:

1- .py file, 2- pickle file, 3- one-pager (this document), 4- Seven different plots

# Model Building Steps (Summarized)

1- Read in the data,

2- EDA process and data standardization by column dropping, encoding and scaling,

1. Checked to see if there is any feature with missing values (none came up)
2. I dropped the empty columns from my dataframe
3. Found the object datatype columns and converted them to numbers with LabelEncoder
4. I dropped Feature\_50 because it had identical values
5. I dropped several identical columns
6. I did outliers analysis using boxplots
7. Split the dataframe to dependent and independent variables
8. Scaled my data using RobustScaler

3- Split the data into train and test subsets, and train the model,

1. Split my data into train and test subsets
2. Trained 6 different classification algorithms and compared their performances

4- Checked performance metrics of the models and compare,

1. Created confusion matrices for these 6 algorithms
2. Compared their performance metrics

5- Pick the best performing algorithm,

1. Decided logistic regression was the top performer. Trained my model with it
2. Created confusion matrix and classification report
3. Plotted ROC.

6- Import the model to a pickle file,

1. Imported my model to a pickle file.

Thanks.

Ubeydullah Keleş