BFSI Project Solution

28/08/2023

Smit Kalathiya

Overview

Given a BFSI masked dataset,which contains more than 40 variables. Task is to predict the dependent Variable accurately. This type of problem is a classic machine learning problem of Binary classification.

Solution

1. **Pre-processing the data.**
    a. Remove columns which have a large number of missing values.
    b. Impute the missing values using 'SimpleImputer' or 'IterativeImputer'.
        i. Iterative Imputer should be used as it is Multivariate so it will fill missing values better than SimpleImputer , hence increasing accuracy.
2. **Feature Selection**
    a. Select feature using filter method or wrapper method.
        i. Filter method should be used as it is computationally cheaper for high dimensional datasets.
3. **Identify Class imbalance .**
    a. If imbalance is found then handle it using Downsampling or Upsampling.
4. **Model selection**
    a. RidgeClassifierCV
    b. LogisticRegressionCV

c. SVC

d. MLPClassifier (**Selected**)

e. RandomForestClassifier

Model Selection is done using the metric roc_auc_score and the difference between the scores of train dataset and validation dataset. The model with high auc_roc_score and low spread in score is selected.

5. **Predicting probability of the test set.**