PART 1: DATASET INFORMATION

Topic: EPILEPTIC SEIZURE RECOGNITION

Dataset: Epileptic Seizure Recognition Data Set

Source: UCI Machine Learning Repository, Life Sciences Datasets

Dataset information: Each file is a recording of brain activity for 23.6 seconds. There are 500 individuals in the datasets so that the dataset contains 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column, explanatory variables Xs), the last column represents the label y {1,2,3,4,5}.

The response variable is y in column 179, the Explanatory variables X1, X2, ..., X178

y contains the category of the 178-dimensional input vector. Specifically y in {1, 2, 3, 4, 5}:

5 - eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

4 - eyes closed, means when they were recording the EEG signal the patient had their eyes closed

3 - Yes they identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area

2 - They recorder the EEG from the area where the tumor was located

1 - Recording of seizure activity

All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure. Only subjects in class 1 have epileptic seizure. Our motivation for creating this version of the data was to simplify access to the data via the creation of a .csv version of it. Although there are 5 classes most authors have done binary classification, namely class 1 (Epileptic seizure) against the rest. (Dataset information link: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition)

PART 2: HYPOTHESIS TESTING

The aim of this study is to analyze how likely recorded brain activities lead to epileptic seizure.

PART 3: EXPLORATORY DATA ANALYSIS

Number of Columns: 180 [179X variables, y]
Number of Rows: 11500

Data types: Integer (all of the variables)
Missing values: No
Value counts of y : There are 5 categories. Each category has 2300 counts

PART 4: MODELING

A. RECURSIVE FEATURE ELIMINATION

Aim: To conduct feature engineering and eliminate unnecassary X variables, I performed a Recursive Feature Elimination model.

Results: The RFE eliminated only five(5) X variables which is not enough to conduct the study.

B. RIP CORRELATION - PREDICTIVE POWER SCORE

Aim: To perform RIP correlation to select variables which are highly correlated with y value.

Results: I selected 53 variables which are correlated 0.12 and 0.13 since those are higher scores.

C. MODELING USING PYCARET

Aim: Compare machine learning algorithms on the epileptic seizure dataset

Results: The finding showed that CatBoost Classifier is the best model which explains the relationships between brain activities (Xs) and epileptic seizure signals (y) with the best accuracy of 0.64

D. GRID SEARCH HYPERPARAMETERS

Aim: Tuning hyperparameters to achieve best accuracy score.

Result: The result of the grid search classifier showed the best score across all searched params is 0.2 which is below the modeling without tuning, which is an unexpected score.