Classification Project Write Up

Josh Jingtian Wang

04/19/2022

Abstract:

A stroke is a medical condition that cuts off the oxygen supply to the brain. In 2015, stroke was the second most frequent cause of death after coronary artery disease. Due to its deadly and sudden nature, it is important to be able to predict the onset of strokes. Here I propose a classification model to predict the onset of strokes using features that can easily be collected from patients.

Design:

My work allows customers to input simple body metrics such as height, weight, health history, smoking habits, etc. and outputs the probability of the onset of stroke. This work will allow for self-diagnosis by patients and assist in the medical diagnosis by doctors.

Data Description:

The dataset was downloaded from here: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>. The data contains 11 features and 5110 rows, with 249 positive rows and 4861 negative rows.

Algorithm:

* The dataset was downloaded from here: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>
* Categorical variables were one-hot encoded. Missing values were filled in by kNNimputing. Interaction features were generated
* Undersampling, oversampling, and class weights were used to deal with class imbalance of the dataset.
* For probabilistic output, KNN, RandomForest, XGBoost and SVC were calibrated with CalibratedClassifierCV().
* RandomSearchCV and GridSearchCV were used to tune the hyperparameters of Logistic Regression, KNN, RandomForest, XGBoost and SVC.

Tools:

Pandas, sklearn, imblearn, matplotlib

Communication:

Please refer to the slides.