Stroke Prediction

Abstract

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. The purpose of this project was to use classification models to predict whether a given person had a stroke or not, and find the main risk factor for stroke to create a prediction system to predict the stroke in its early stages. I worked with data provided by Kaggle, first I dealt with outliers and missing data, then I attempted to perform extensive data visualization, exploratory data analysis, feature Engineering, deal with imbalanced data, hyperparameter tuning of the model, and finally fit a model to make predictions on features.

1. Design

The data is provided by Kaggle, this dataset is used to predict whether a patient is likely to get a stroke or not. Data visualization and data analysis help in finding the factor that has the most impact on stroke, finding the most age had a stroke and finding the impact of diseases on stroke. This analysis helps real-world problems as strokes account for as much as 11% of all deaths in the world. Also, classifying statuses accurately via machine learning models would enable predicting if a particular person has a stroke in the early stages.

2. Data

The dataset contains 5110 recorders with 12 features for each, 7 of which are categorical. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

3. Algorithms

3.1 Feature Engineering

Encoding the categorical values
Splitting the dataset into the Training set and Test set
Handling Imbalance data using SMOTE
Feature selection with target and select random forest classification

3.2 Models

Logistic Regression, Decision Tree, K Nearest Neighbour, and Random Forest Classifier were used before settling on the random forest as the model with the strongest performance.

3.3 Model Evaluation and Selection

The data was split into training and test data and the training data is used to create models that predict whether or not an entry from the test data will suffer a stroke. I tuned the hyperparameters with the help of GridSearch to get the best model which is Random Forest Classifier.

Before resample data
Since the data is imbalanced it could achieve high accuracy, if the classifier always returns 0 (4861/5110 = 95.13%). We can see that accuracy is not a useful metric in the context of strongly imbalanced data. So, in this case I focused on AUC instead of accuracy.
Results:

- AUC: 0.75
- F1: 0.97 micro, 0 macro
- Precision: 0.95 micro, 0 macro
- recall: 1 micro, 0 macro
- Mean squared error: 0.05
- Root mean squared error: 0.22

After resample data
In this case I focused on accuracy.
Results:

- Accuracy: 0.89
- F1: 0.94 micro, 0.12 macro
- precision: 0.96 micro, 0.10 macro
- recall: 0.93 micro, 0.16 macro
- Mean squared error: 0.11
- Root mean squared error: 0.33

4. Tools

- NumPy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- imblearn for imbalanced data

5. Communication

I examined Stroke Prediction Dataset. Firstly, I made Exploratory Data Analysis, Visualization, then I applied Machine Learning algorithms to this dataset.

The most prominent results:

- Age has the most correlation with the target
- People with glucose greater than 150 are less impact on stroke
- People with age more than 64 are the most people who suffer a stroke
- In unbalanced data we have AUC score of 75% in Random Forest, which is good
- After resample data we have the highest accuracy of 89% in Random Forest

In **addition**, you can see my presentation slide here and project code here