Stroke prediction model

Abstract

The goal of this project was to use classification models to predict the occurrence of a stroke in order to help the people to be aware of their likelihood of suffering from a stroke. I worked with data found in Kaggle where I did many preprocessing to it. Moreover, I mainly focused on two algorithms, but made oversampling technique on them which made them 4 separated algorithms. That is because of an imbalance issue which reside in the original data. The results in terms of precession were random forest 0, logisitic regression 0, oversampled random forest 0.995217, oversampled logestic regression 0.976598.

Design

The data is provided by Kaggle and presents a binary-class incident status for stroke. Classifying statuses accurately via machine learning models would help in increasing the awareness and the commitment to following an early corrective approach in people's lives such as changing the diet plan, taking care of the mental and psychological health, avoid smoking … etc. Furthermore, enabling people from being aware of their cases and illnesses in an early stage helps in reducing the number of fatalities annually and relieve the burden on the medical sector.

Data

The dataset contains 43401 patient entries with 11 features for each, 4 of which are categorical. Some of the features are gender, smoking status, ever married, glocuse, bmi … etc. My model will predict whether this entry/case/patient is most likely to be diagnosed with stroke or not. Thus, my model will produce either yes/no , 0/1 classes which is a classification problem.

Algorithms

Data preprocessing

- deal with missing values
- deal with redunduncy
- deal with outliers
- One-hot encoding (same as label encoder, but with order between values)
- dummy variables
- standardization
- EDA

Exploratory Data Analysis (EDA)

- understand the relationship between features by plotting some plots
- understand the relationship between features and the target by plotting some plots
- visualize the counts for some features
- visualize the percentages of the target classes before and after oversampling

- visualize box plots
- Data distribution

Machine Learning algorithms (classifiers)

- Random Forest
- Logistic regression

Data resampling techniques

- Oversampling SMOTE technique

Results & comparison

Model/metric	accuracy	precession	recall
Random forest	0.982796	0	0
Logestic regression	0.983026	0	0
Oversampled random forest	0.988503	0.995217	0.981974
Oversampled logestic regression	0.940519	0.976598	0.903992

Tools

- Pandas
- matplotlib
- seaborn
- sklearn

Communication
EDA and some metric visualizations