Factors Influencing Stroke

In the analysis of understanding the causes of strokes, I chose the first option to build Ml model to predict the occurrence of stroke or not. The best model was Logistic regression with an assurance of 84% and AUC of 0.803

Tuah Ebenezer Kwadwo

2020.

Data And Methodology

The dataset was derived from Kaggle with 11 separate features. Some numerical and others alphabetical.
Here is the metadata:

Exploratory data analysis was done as well and visualization of the data.It is noticed that Age, bmi dynamics affected the occurrence of stroke.

DATA DICTIONARY

Variable	Definition
id	Patient ID
gender	Gender of Patient
age	Age of Patient
hypertension	0 - no hypertension, 1 - suffering from hypertension
heart_disease	0 - no heart disease, 1 - suffering from heart disease
ever_married	Yes/No
work_type	Type of occupation
Residence_type	Area type of residence (Urban/ Rural)
avg_glucose_level	Average Glucose level (measured after meal)
bmi	Body mass index
smoking_status	patient's smoking status
stroke	0 - no stroke, 1 - suffered stroke

Fig. 1 The metadata of the dataset.

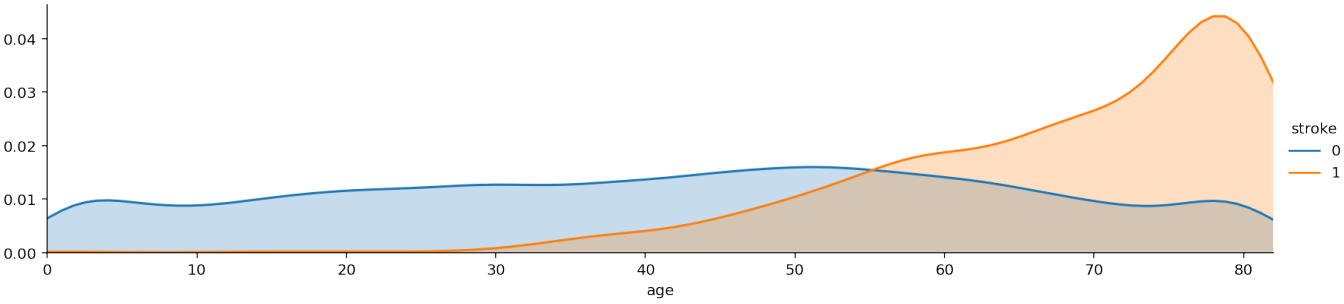

Fig 2. The age distribution of occurrence of stroke

The data needed to balanced using SMOTE since it was disproportional. Other steps taking in the analysis or study was to reduce the significance of the numerical data through polynomial featuring.

In building the model, the accuracy score used was the most important part since we wanted to deal with the True Positives and the False Positives. You don't want to tell someone who has stroke not to have it. It will be terrible in such an instance. That is why Logistic regression was finally selected which gave score of 84% and AUC of 80.3%

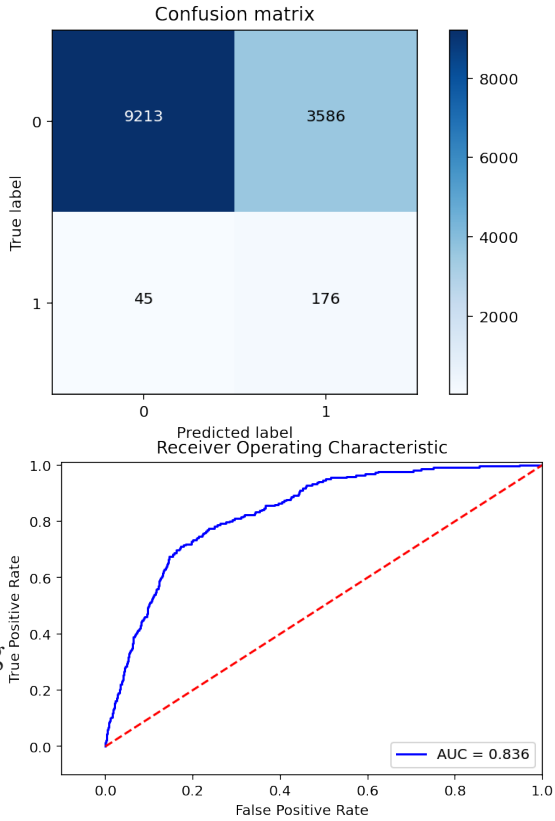

Fig. 3 Model prediction and

In terms of features of factors that affected the model were age, hypertension heart_disease which obviously makes sense. I expected to have had BMI and Smoking play a major part of the factors affecting strokes occurrence. However that is one of the downside of «black box» modelling. It is must be noted that the polynomial featuring helped over under-fitting of the model by rolling it over power 2.

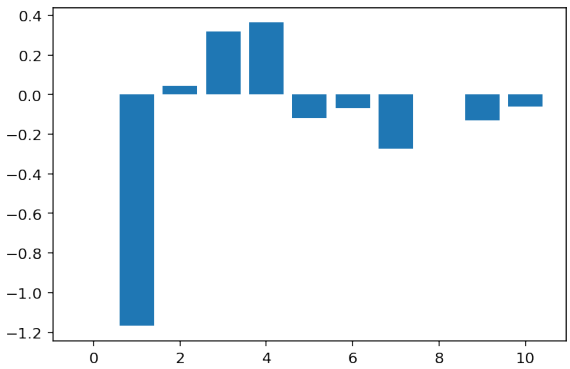

Fig 4. Model selection.

References.

1. https://vas3k.ru/blog/machine_learning/
2. Python Data Science Handbook Essential Tools for Working with Data, Jake VanderPlas