Healthcare-Stroke- Prediction

Abstract:

According to the World Health Organization (WHO) stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths.
This dataset is use to predict whether a patient is likely to get 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

Design:

In order to determine the types of transaction statue, data was download from Kaggle. Then, multiple models were implement to get the best one to make a clear classification

Data Description:

The original source for this data is here, and we have taken from kaggle, this data set is name as healthcare-dataset-stroke-data contain 12 column and 5111 rows.

Features:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not
*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Algorithms:

- Exploratory Data Analysis.
- Building multiple models and finding out the well-suited one for this specific dataset.

Cleaning:

Drop null values

Feature Engineering:

Dummy variable

Model Building:

Around five models were tried and played with to get the best model that goes hand in hand with the dataset. After performing simple train and validation on the models one was be chose for further investigation. Models trained was:

- Logistic regression (Baseline)
- KNN
- Decision trees
- Random forest
- XGB Classifier

The Best Models: Logistic regression

Dealing with Class Imbalance by pipline with gridsearch for the best models and over sampling.

Tools Description

The main technologies and libraries that will be used:

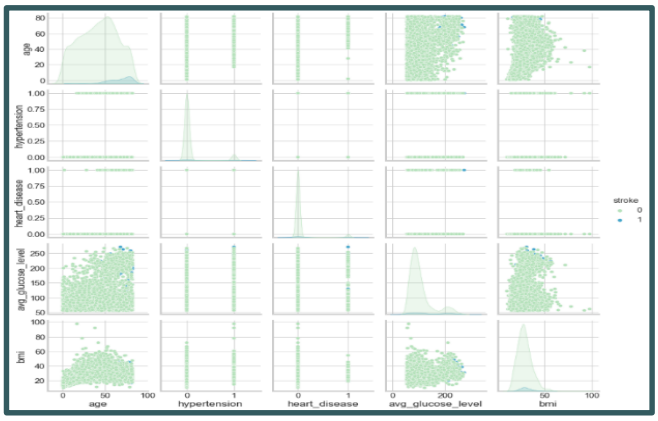
1. Python
2. Jupyter Notebook

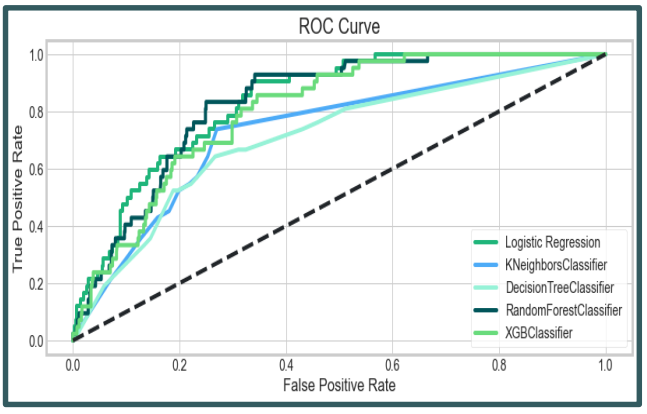
Libraries:

1. Pandas
2. Matplotlib
3. Seaborn
4. Numpy
5. Sklearn

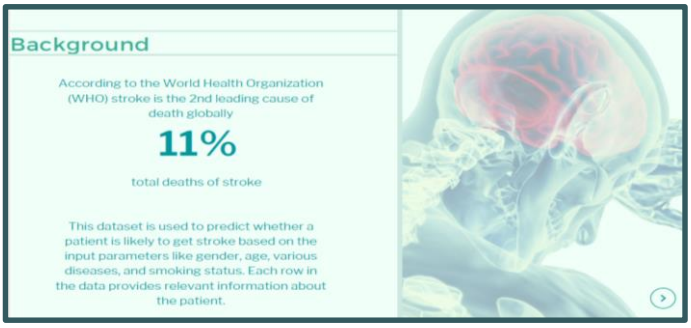
Communication:

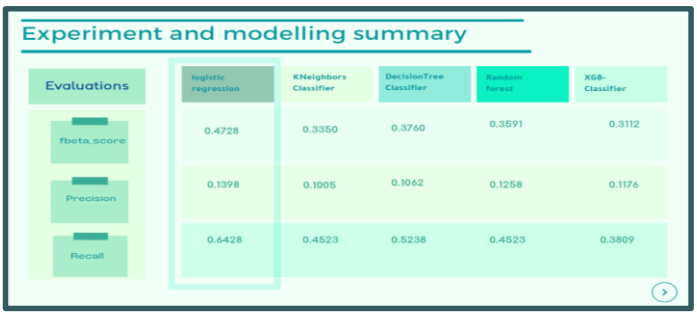
Charts:

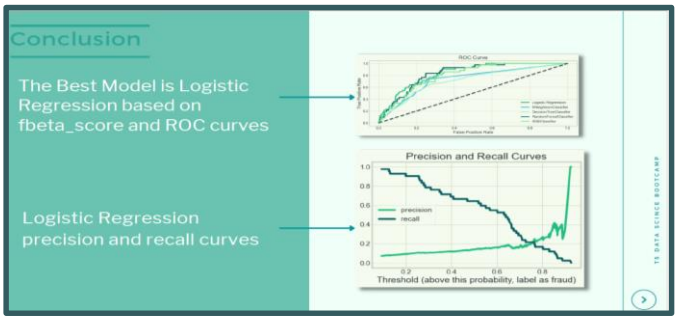




Presentation snips:







Contributors: Tahani Alqahtani

Amani albalawi