Stroke Prediction

Dataset description:

The data set used was the stroke prediction dataset which can be found at [Kaggle](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset).

The attributes of patients that contributed to the data collection are as follows:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| 1. id | Unique patient identifier |
| 1. gender | Values include “male”, “female” and “other” |
| 1. age | The age of the patient |
| 1. hypertension | 0 if the patient doesn’t have hypertension and 1 if the patient does |
| 1. heart\_disease | 0 if the patient doesn’t have heart disease and 1 if the patient does |
| 1. ever\_married | Accepted values are “Yes” or “No” |
| 1. work\_type | Values include “children”, “Govt\_jov”, “Never\_worked”, “Private” and “Self-employed” |
| 1. Residence\_type | Accepted values are “Rural” or “Urban” |
| 1. avg\_glucose\_level | Average glucose level in patient’s blood |
| 1. smoking\_status | Values include “formerly smoked”, “never smoked”, “smokes”, and “Unknown” |
| 1. bmi | Body Mass Index of patient |
| 1. stroke | The target variable.  1 if the patient had stroke and 0 if they didn’t |

Main Objective of Analysis:

The objective here is to train a model that best predicts the likelihood of a patient getting stroke considering the afore mentioned features.

More emphasis will be placed in getting as many right guesses for stroke patients even at the cost of a few false alarms.

But we would like to minimize the number of false alarms as much as possible.

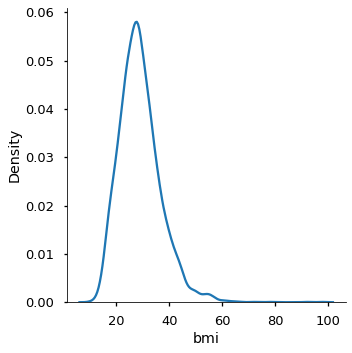
Summary of EDA:

1. Handling of missing data:

All features but “bmi” (which had 201 missing) had no missing values for any patient.

Missing data was therefore filled with the mean bmi after a careful inspection of the distribution. Since the bmi data followed a normal distribution (meaning most of the patient’s bmi were around the mean bmi), and the mean of both patient’s with stroke and those without were very close, the mean bmi proved to be the best estimate.

1. **Distribution of BMI of patients**



1. **Distribution of BMI for patients with stroke and those without**

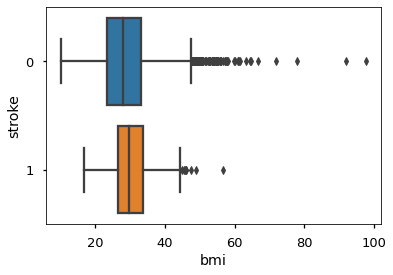


Figure a) figure showing that the distribution of the bmi of the patients is normal. It can be seen that most of the patients have BMI around the mean value.  
b) figure showing distribution of bmi of patients with and without stroke. It can be seen that the mean values of both boxplots (the middle line in both the orange and blue box) are very close to each other.

1. Removing unnecessary features:

Since the “id” feature is unique to each patient, it will not help in our prediction, therefore, it was removed from the dataset.

1. Checking normality of continuous features:

After a careful inspection of the distribution of all the continuous data (age, average\_glucose\_level, bmi) it was observed that most of them had roughly normal distribution hence no transformation was needed in that respect.

1. Checking scale of continuous features:

After an inspection of the standard deviation of the continuous features, it was observed that the features were measured on different scales.

Hence sklearn’s **StandardScaler** was used to normalize the features so that each had a standard deviation of 1.

1. Visualizing categorical features:

Because the data was unbalanced, that is a ‘**95:4’** ratio for ‘**no stroke : stroke’**, observing the frequency of various features among patients with stroke and those without stroke will not give much insight into which categorical features play more important roles in the stroke status of a patient.

1. Resampling of data:

New datasets were created with both **over sampling** and **under sampling** for the models to be trained on.

Four new datasets were created with the training split of the data. Each new different dataset was created with a different type of resampling technique;

1. Random Over Sampling
2. SMOTE
3. ADASYN
4. Random Under Sampling

Training Models:

Six different models were trained with and without resampled training data.

The results are as shown in the table below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Sampling** | **AUC** | **Accuracy** | **Recall** | **Precision** | **F -score** |
| Logistic Regression | Without resampling | 0.510 | 0.952 | 0.020 | 1.00 | 0.039 |
| Class Weighting | 0.673 | 0.846 | 0.480 | 0.155 | 0.234 |
| Random Oversampling | 0.693 | 0.792 | 0.620 | 0.120 | 0.201 |
| SMOTE | 0.657 | 0.764 | 0.540 | 0.110 | 0.183 |
| ADASYN | 0.675 | 0.761 | 0.580 | 0.115 | 0.192 |
| Under sampling | 0.717 | 0.732 | 0.700 | 0.119 | 0.204 |
| Random Forest | Without resampling | 0.500 | 0.951 | 0.000 | 0.000 | 0.000 |
| Class Weighting | 0.700 | 0.736 | 0.660 | 0.115 | 0.196 |
| Random Oversampling | 0.703 | 0.743 | 0.660 | 0.118 | 0.201 |
| SMOTE | 0.701 | 0.755 | 0.640 | 0.121 | 0.204 |
| ADASYN | 0.704 | 0.743 | 0.660 | 0.119 | 0.201 |
| Under sampling | 0.722 | 0.706 | 0.740 | 0.114 | 0.198 |
| Gradient Boosting Classifier | Without resampling | 0.508 | 0.949 | 0.02 | 0.250 | 0.037 |
| Random Oversampling | 0.504 | 0.942 | 0.02 | 0.091 | 0.033 |
| SMOTE | 0.542 | 0.940 | 0.100 | 0.238 | 0.141 |
| ADASYN | 0.530 | 0.936 | 0.080 | 0.174 | 0.110 |
| Under sampling | 0.702 | 0.704 | 0.700 | 0.109 | 0.188 |
| XGBClassifier | Without resampling | 0.499 | 0.949 | 0.000 | 0.000 | 0.000 |
| Random Oversampling | 0.566 | 0.932 | 0.160 | 0.222 | 0.186 |
| SMOTE | 0.556 | 0.932 | 0.140 | 0.206 | 0.167 |
| ADASYN | 0.533 | 0.923 | 0.100 | 0.128 | 0.112 |
| Under sampling | 0.720 | 0.719 | 0.720 | 0.117 | 0.201 |
| SVC | Without resampling | 0.500 | 0.951 | 0.000 | 0.000 | 0.000 |
| Class Weighting | 0.704 | 0.690 | 0.720 | 0.106 | 0.185 |
| Random Oversampling | 0.657 | 0.798 | 0.5 | 0.121 | 0.195 |
| SMOTE | 0.602 | 0.803 | 0.380 | 0.101 | 0.159 |
| ADASYN | 0.592 | 0.801 | 0.360 | 0.095 | 0.151 |
| Under sampling | 0.699 | 0.716 | 0.68 | 0.110 | 0.190 |
| KNN | Without resampling | 0.498 | 0.948 | 0.000 | 0.000 | 0.000 |
| Random Oversampling | 0.525 | 0.889 | 0.120 | 0.080 | 0.096 |
| SMOTE | 0.619 | 0.781 | 0.440 | 0.101 | 0.164 |
| ADASYN | 0.579 | 0.812 | 0.32 | 0.092 | 0.143 |
| Under sampling | 0.725 | 0.712 | 0.740 | 0.116 | 0.201 |
| Voting Classifier | Under sampling | 0.736 | 0.732 | 0.740 | 0.123 | 0.212 |

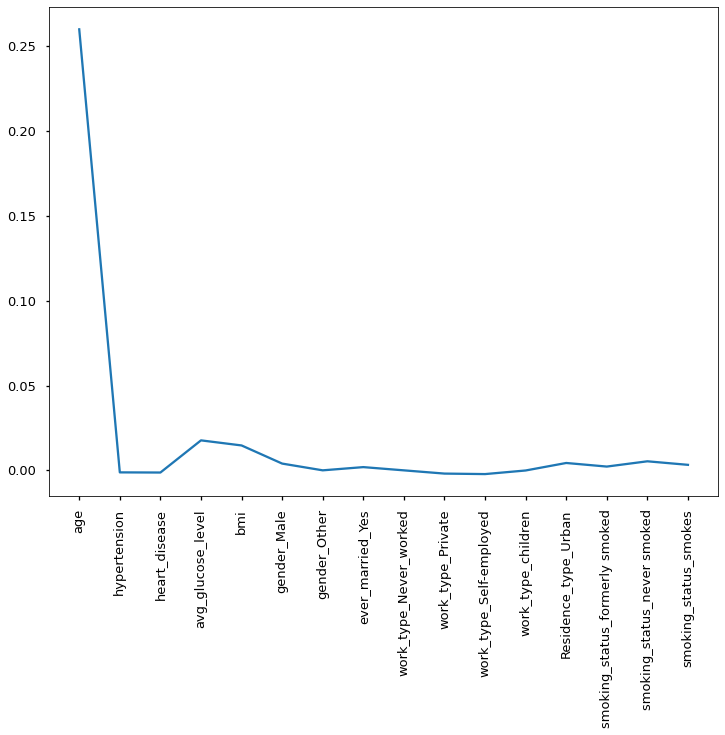
Observation:

The ideal model for prediction would be one with high **recall**, **precision** and consequently a high **f1 score**. And one that maintains a relatively high accuracy.

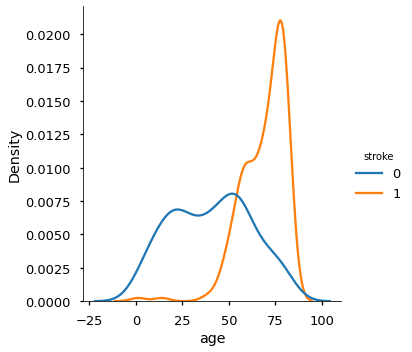
From the table above, it can be observed that the Voting Classifier model trained (using the following models: logistic regression, xgbclassifier, knn, svc) on the under sampled data meets the criteria best.

Observing relevant features in model:

1. **Importance of features**



1. **Distribution of age for patients with and without stroke**



From figure **a** above, it can be observed that age plays a major role in determining the likelihood of an individual getting stroke.

The figure below, figure **b** affirms that and shows that **the older a person, the more likely they are to get stroke**.

Suggestions for making better predictions:

The major issue faced here was lack of sufficient data. More specifically data on patients who have had stroke before. Therefore, to be able to improve upon the model’s accuracy and precision, getting more data, about patients that have had stroke before, that can be augmented into our current dataset, would prove very useful.