Springboard Data Science
Capstone Project 2
Stroke Prediction
By Paul Kim
March, 2022

1.Introduction

1) Background

According to WHO, stroke is the second leading cause of death globally with an annual mortality rate of 5.5 million. Not only the high mortality rate, but it also results in morbidity in 50 percent of the patients. The symptoms may include trouble speaking, paralysis, blindness, headache and trouble walking. Stroke occurs when blood flow to the brain is interrupted or blocked, and brain cells begin to die in a short period. It is a medical emergency and quick treatment is crucial to one's life.

2) Problem Statement

Since stroke is a crucial disease threatening one's life, early action must be taken to reduce brain damage. However, preventing the possibilities would be the first action to fight the disease. Therefore, understanding the reasons and causes that increase the risk of stroke is essential. The project aims to identify people with high risk of stroke according to the features with machine learning techniques.

3) Overview

The project's goal was to develop a model that successfully detects people with high risk of stroke using their features. The model can be used in the health industry where all the necessary features of the patients are accessible and inform them of getting medical checked due to high risk of stroke according to the model developed.

To develop the model, samples with features and stroke status were collected and several machine learning algorithms were used. Finally, the tuned Random Forest Classification model was able to achieve an ROC-AUC score of 0.80.
Implementation details can be found in the link below

(https://github.com/paulkimDSN/Capstone-2-Strok-prediction).

2. Approach

1. Data Wrangling

The raw dataset called ‘Stroke Prediction Dataset’ was retrieved from the website kaggle in a csv format.

The dataset has 12 columns and 5110 samples. The columns include seven ‘int64’ or ‘float64’ data type columns and five ‘object’ data type columns. The attribute information of the dataset is listed below.

1) id: unique identifier
2) gender: "Male", "Female"
3) age: age of the patient
4) hypertension: 0 :patient without hypertension, 1: patient with hypertension
5) heart_disease: patient without heart disease, 1: patient with 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: patient with stroke 0: patient without stroke

The columns’ unique variables were checked for consistency and data cleaning. From the process, one sample with ‘Other’ variable was observed from the gender column. Simply, the whole row was removed from the dataset since it was a one patient with ‘Other’ variable.

After, columns with missing values are identified and number of missing values were counted There were 201 samples with missing bmi values that is about 3.93 percent of the whole sample size. To impute the missing values, bmi distribution with average and median bmi marked was plotted.

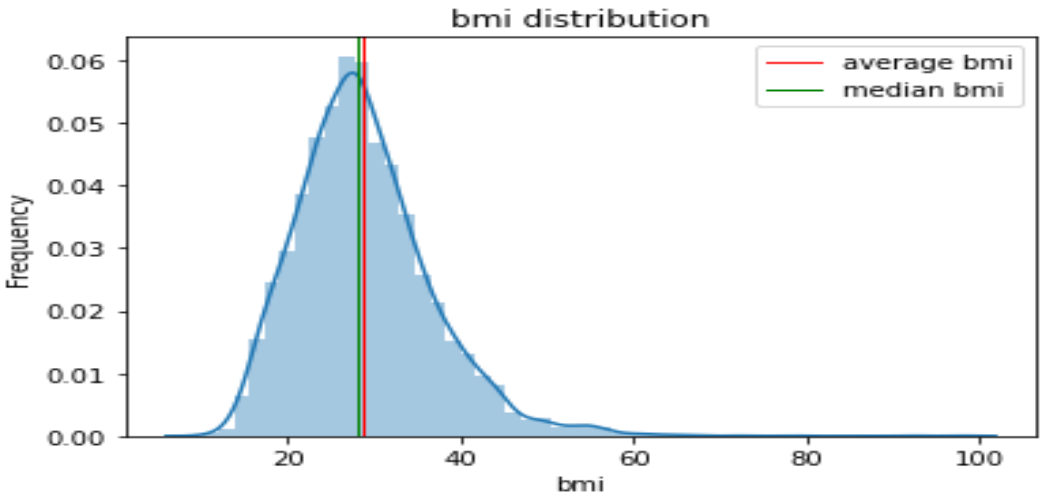

Figure 1: BMI distribution of the sample with average and median

As the graph shows above, median bmi is more appropriate to be used as a value to be imputed for missing values. Now, the dataset is ready for exploration and the final shape of the data is 5099 rows and 12 columns.

2. Exploratory Data Analysis

a) Data distribution

The dataset was explored and visualized to see the overall picture of the distribution of the cleaned dataset. First, the proportion of the patients with stroke(target variable) was calculated. As the result, 4.87 percent of the sample are the patients with stroke and 95.13 percent of the sample are patients without stroke cases.

After, the age distribution of the sample was plotted with an average age of 43 marked.

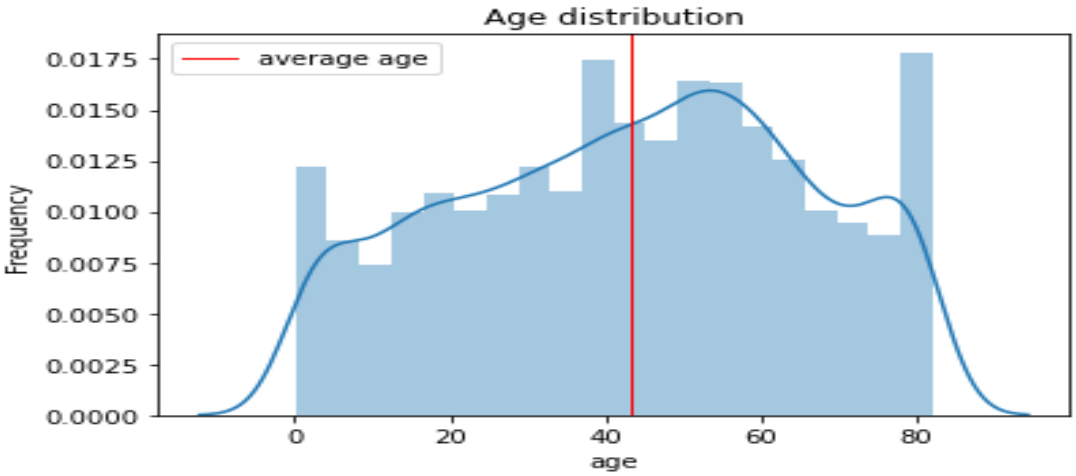

Figure 2: Age distribution of the sample with average age.

b) Stroke vs Features

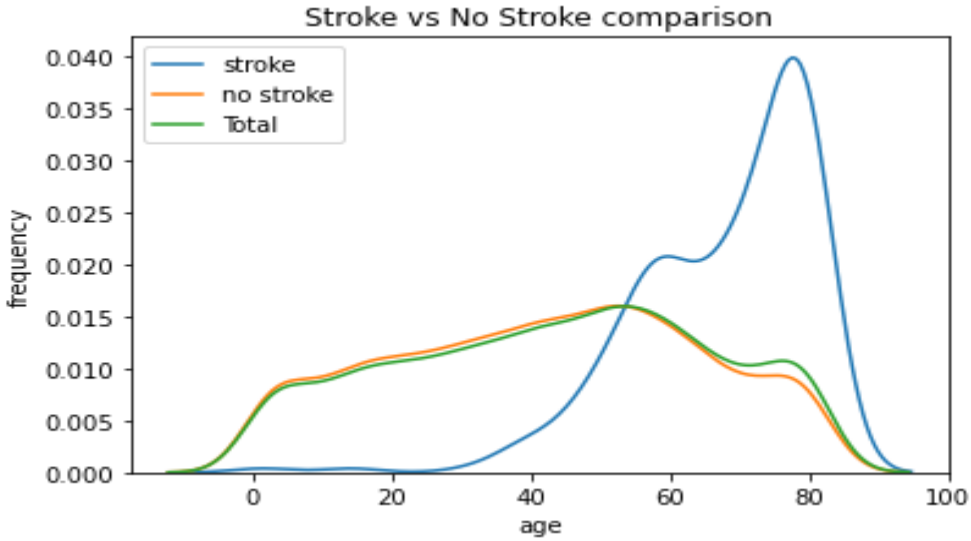

Figure 3: KDE plot of patients with stroke vs without stroke by age

In figure 3, patients with and without stroke are plotted by age and it clearly shows how stroke occurrence starts to increase from 40 and peaks at age of 80. Stroke occurrence graph is above the ‘no stroke’ until approximately at 50 and it flips afterwards.

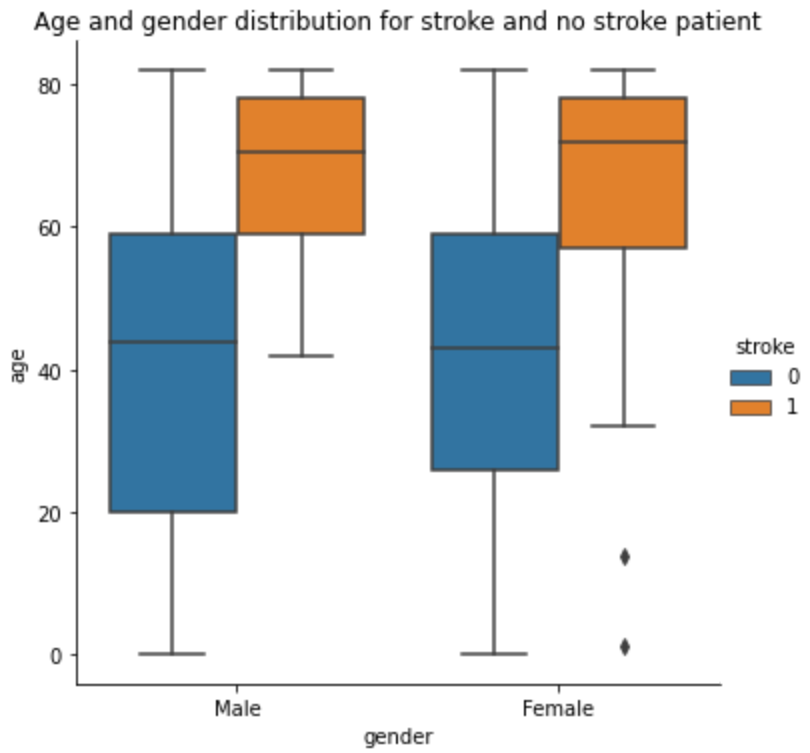

Figure 4: Box plot of patients with stroke and without stroke by gender and age

Another graph that explains the relationship between features and stroke occurrence was plotted. In figure 4 above, it shows the distribution of patients with stroke within age groups and by genders. As it was mentioned above, most of the stroke patients are older than 60. In terms of gender, they have similar distributions however, both of the outliers are from females. Therefore, females are more exposed to the risk of stroke in the early age compare to male.

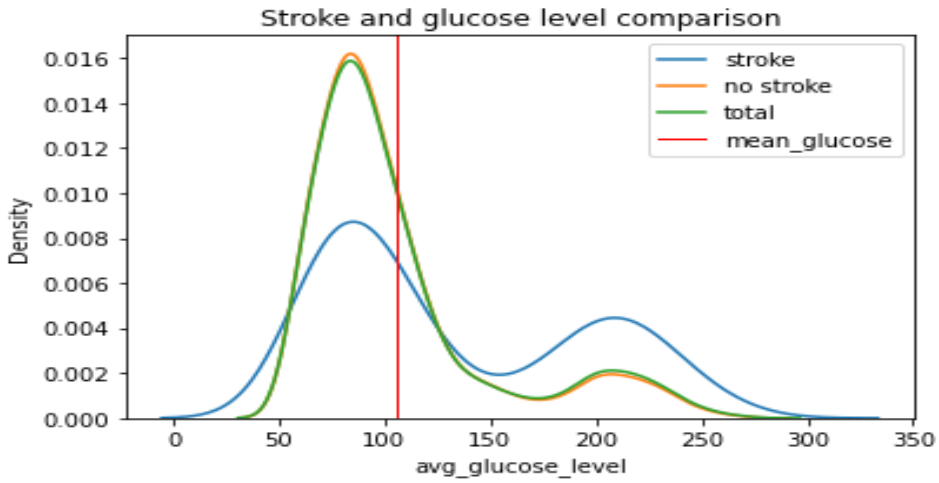

Figure 5: KDE plot of patients with stroke and without stroke by average glucose level

Stroke occurrence over glucose level was visualized. As the graph shows above, stroke and glucose level does not have as strong relationship as age and stroke however, it clearly shows that there are more patients with stroke than without stroke once the average glucose level is above 150.

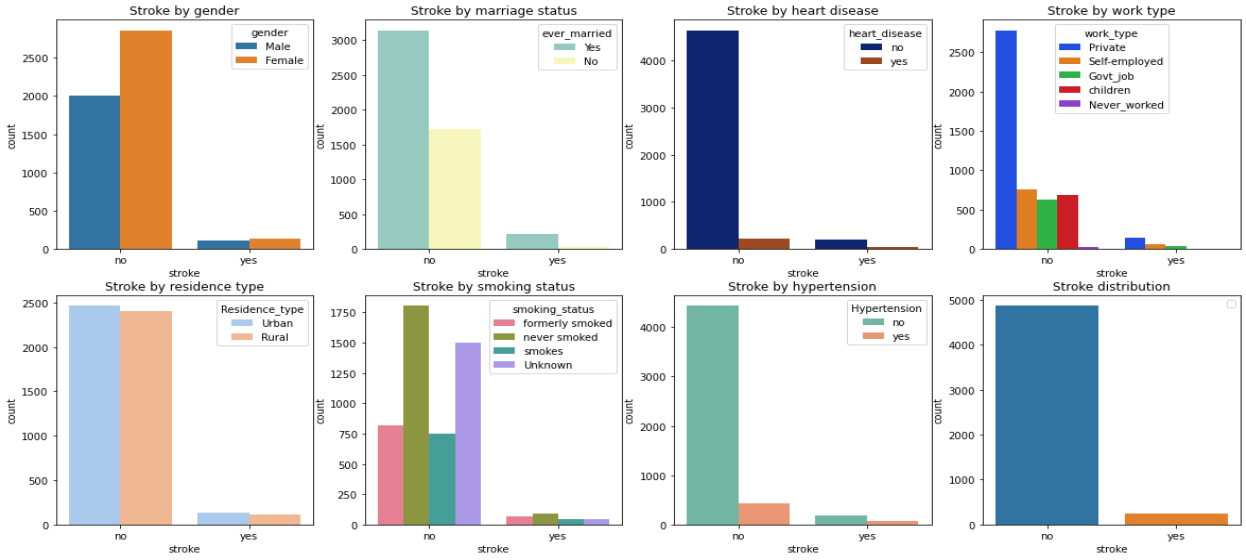

Figure 6: Histogram of patients with stroke and without stroke by each feature

Lastly, proportion of patients with and without stroke by all the categorical features are visualized with bar charts. Most of the proportion was similar however, who have never been married and works with children had significantly low risk of the stroke.

c) Correlation

Correlations between each column are calculated and interesting features are visualized to see the relationship.Below, heatmap of each column are plotted with brightness of the color by the strength of correlation.

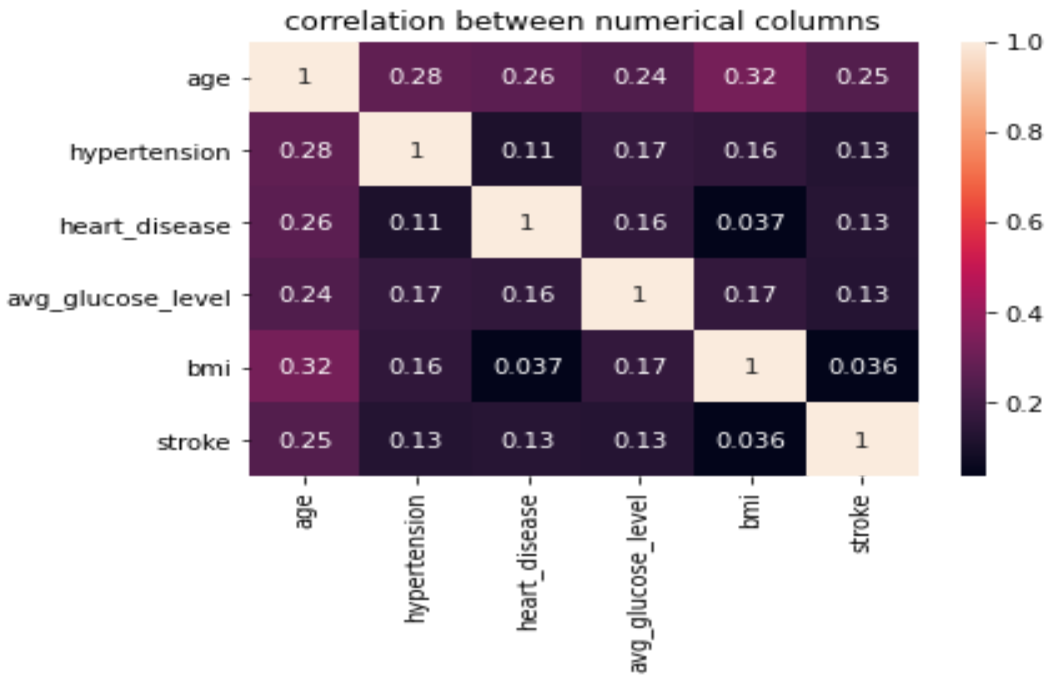

Figure 7: Correlation heatmap of columns

Overall, the correlation between each feature was not strong however, most of the strong correlation occurred against the age column. Therefore, correlation between age and other columns were visualized to see how strong they are.

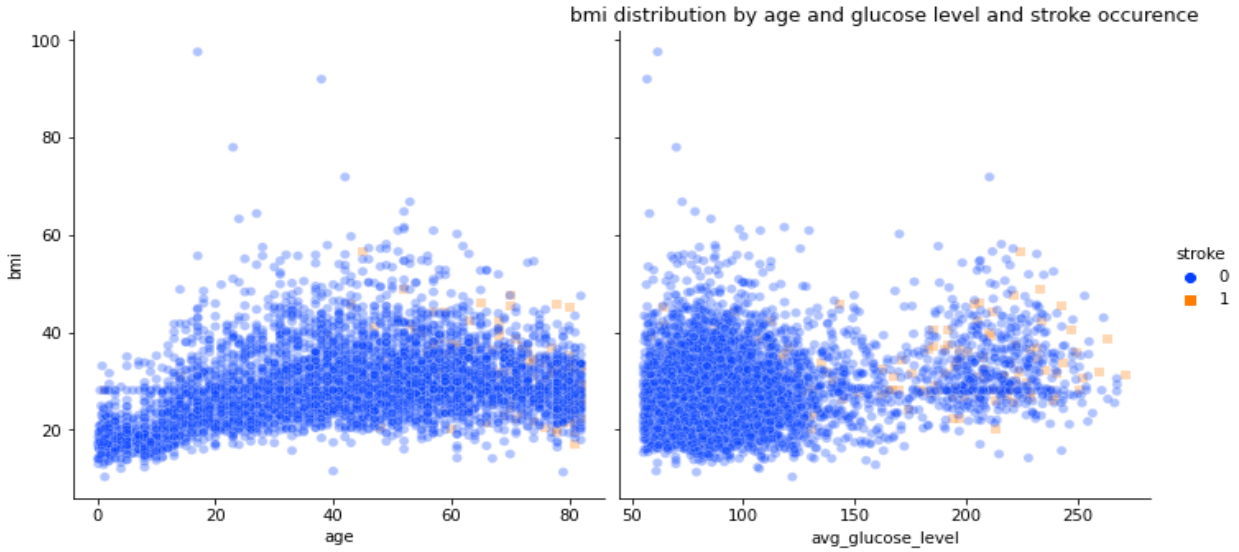

Figure 8: Stroke occurrence over bmi by age and average glucose level.

As the figure 8 shows, age and bmi has weak positive correlation. However, bmi and average glucose level is also very weakly correlated. BMI also has no to slight impact to risk of stroke compare to age and average glucose level

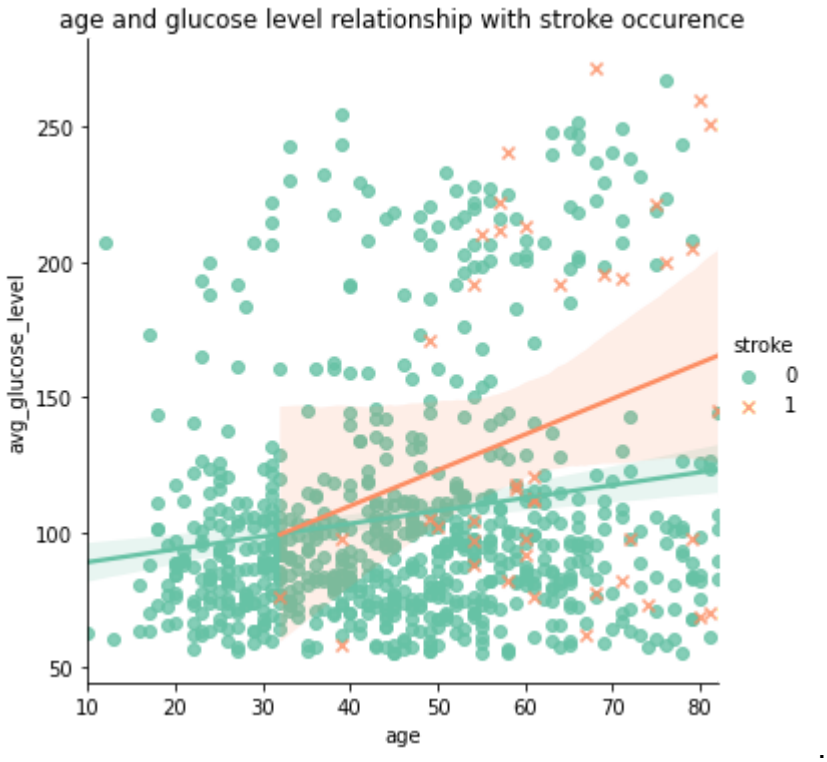

Figure 9: age vs glucose level with stroke occurrence

Although age and average glucose level are weakly correlated, the positive correlation is present as figure 9 shows. Also, patients with stroke have stronger positive correlation in terms of age and glucose level compared to patients without stroke.

d) Baseline Modeling

For the baseline modeling logistic regression classification model was used. As a result, it returned an accuracy score of 95.24 percent and 0.84 of ROC-AUC.

	precision	recall	f1-score	support
0	0.95	1.00	0.98	1458
1	1.00	0.03	0.05	75
accuracy			0.95	1533
macro avg	0.98	0.51	0.51	1533
weighted avg	0.95	0.95	0.93	1533

Figure 10: classification report of baseline model

After, the confusion matrix of the baseline model was visualized and realized it encountered a problem. The model does a horrible job of predicting the ‘True’ variable even if the accuracy score of the model is high.

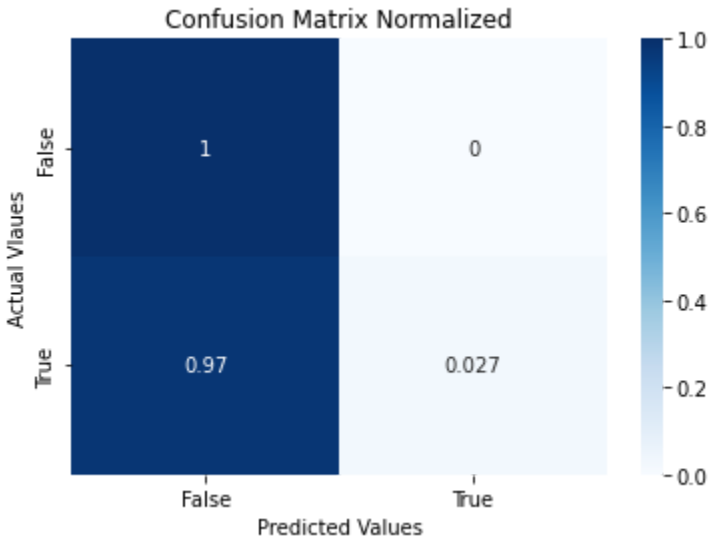

Figure 11: Confusion Matrix of baseline model(Normalized)

The model had a poor performance on predicting the ‘True’ value was due to imbalance classes of the dataset. The original dataset has proportion of 95.13 percent of patients with stroke and only 4.87 percent of patients without stroke.

e) Extended Modeling

Imbalanced Classification was dealt with oversampling (SMOTE and ADASYN) and undersampling(Random Undersampling). After, three different machine learning classification models: Random Forest Classifier, XGBOOST and LGBM were used.

The models were evaluated with accuracy score, precision score, recall score and F1 score as shown in figure 12. Hyperparameter was conducted with top performed classification models and resampling methods. In conclusion , Random Forest, LGBM and XGBoost with SMOTE are selected.

	Accuracy_score	Precision_score	Recall	F1Score
LGBM ADASYN	0.923679	0.208333	0.200000	0.204082
LGBM SMOTE	0.916504	0.202247	0.240000	0.219512
Random Forest SMOTE	0.913894	0.139241	0.146667	0.142857
Random Forest ADASYN	0.911937	0.142857	0.160000	0.150943
XGBoost Adasyn	0.888454	0.152174	0.280000	0.197183
XGBoost SMOTE	0.882583	0.147651	0.293333	0.196429
Random Forest Undersample	0.702544	0.124260	0.840000	0.216495
LGBM Undersample	0.699935	0.109533	0.720000	0.190141
XGBoost Undersample	0.694716	0.115460	0.786667	0.201365

Figure 12: Results of each model performance

3. Findings

After hyperparameter tuning is performed another table showing precision, recall, and f1 score for each class is created. The metric that was focused on when building and tuning my model was the recall score for positive class. By optimizing the recall score for positive class, the number of false negative is minimized.

	precision, class:negative	precision, class:positive	recall, class:negative	recall, class:positive	f1, class:negative	f1, class:positive	ROC_AUC	total negative	total positive
Random Forest	0.98	0.11	0.68	0.75	0.80	0.19	0.80	1458	75
XGBoost	0.96	0.17	0.92	0.31	0.94	0.22	0.61	1458	75
LGBM	0.98	0.11	0.70	0.72	0.82	0.19	0.71	1458	75

Figure 13: Model Performance after tuning

As figure 13 shows Random Forest Classifier is the best model for our case since it has the highest recall score for positive class.

The original goal of the case was to identify the patients with risk of stroke by the features. Therefore, false negatives would cause the most dangerous problem where

patients with the risk of stroke may be labeled as ‘No Stroke’. Optimizing recall score may increase the false positive however, it is not as dangerous as false negative.

4. Future Work

- I would like to find another dataset where it has more variables directly related to the symptoms or cause of stroke to merge the dataset and develop a model that predicts with more variables.

- Since patients with stroke are mostly distributed in the age of 40 or older, developing two different models for different age groups would lead to a better performance model. Since patients with stroke in the early age are rare therefore, the causes of stroke in the early age may be different from another age group.

5. Recommendations for the clients

- In the health industry, it is likely to have patients basic health information. As the features used in the model developed are also simple health information that can be collected during the general check up of the patients, the model can be used on every patient who visit the clinic.

- The model has 0.80 of ROC-AUC, therefore it may not be high enough to be used for final decision of patients’ condition. However, the model can be a good start for elderly patients who might have risk conditions that may lead to stroke. When the model was designed, I focused on minimizing the false negative by optimizing the recall score. The model will have a higher number of false positives however, it is more important to make sure not a single patient leaves the clinic with the negative result but actually with a high risk of stroke.

- Therefore, the model can be used for general check up on risk of stroke for patients and if the model shows, the patient's condition may lead to possible stroke, more detailed examination can be conducted.

- If the budget for the industry is limited for examinations, people with older age can take the priority over the younger patients since age is the most important feature that affects the stroke as figure 14 shows.

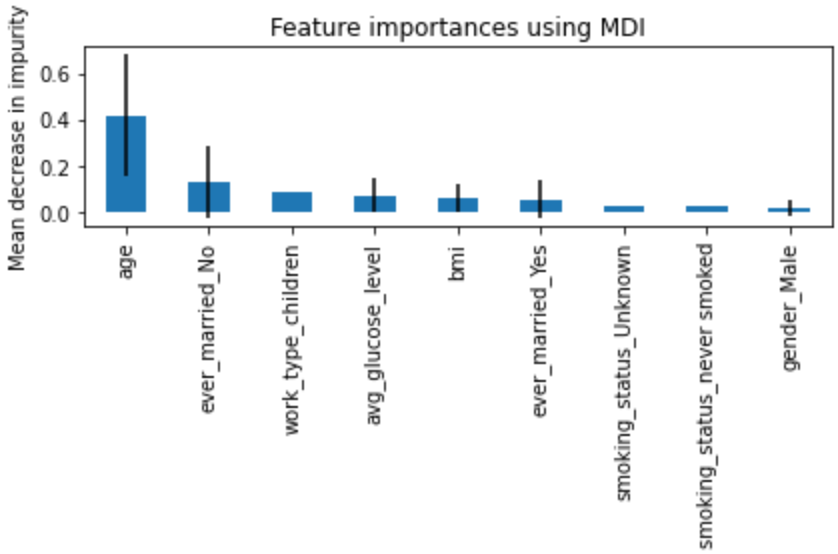

Figure 14: histogram of feature importance