Stroke Prediction

Jiahui Zhang, Data Science Institute, Brown University

December 8, 2023

1. Introduction

According to the World Health Organization (WHO), stroke is one of the leading causes of death globally. In this context, early prediction and intervention for populations at high risk of this disease are crucial. The project utilizes a dataset from Kaggle containing health risk factors and whether there is a stroke. The features, along with their respective data types and brief descriptions, are presented in Table 1. The objective through this project is to predict the stroke conditions (whether a patient has a stroke) based on the provided health risk factors.

Feature Name	Data Type	Description
id	int64	unique identifier
gender	object	"Male", "Female" or "Other"
age	float64	age of the patient
hypertension	int64	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart disease	int64	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever married	object	"No" or "Yes"
work type	object	"children", "Govt job", "Never worked", "Private" or "Self-employed"
Residence type	object	"Rural" or "Urban"
avg glucose level	float64	average glucose level in blood
bmi	float64	body mass index
smoking status	object	"formerly smoked", "never smoked", "smokes" or "Unknown"
stroke	int64	1 if the patient had a stroke or 0 if not

Table 1: Data Types and Descriptions of Features

2. Exploratory Data Analysis

The EDA is structured to examine various aspects of the dataset, starting with class imbalances, the relationship between critical health indicators and stroke occurrence, and the demographic distribution of the affected individuals.

2.1 Class Imbalance

The first step in Exploratory Data Analysis (EDA) is to understand the distribution of the target variable, as depicted in the histogram in Figure 1. This histogram clearly shows that the dataset is imbalanced: 5% of the values are 1, while 95% are 0.

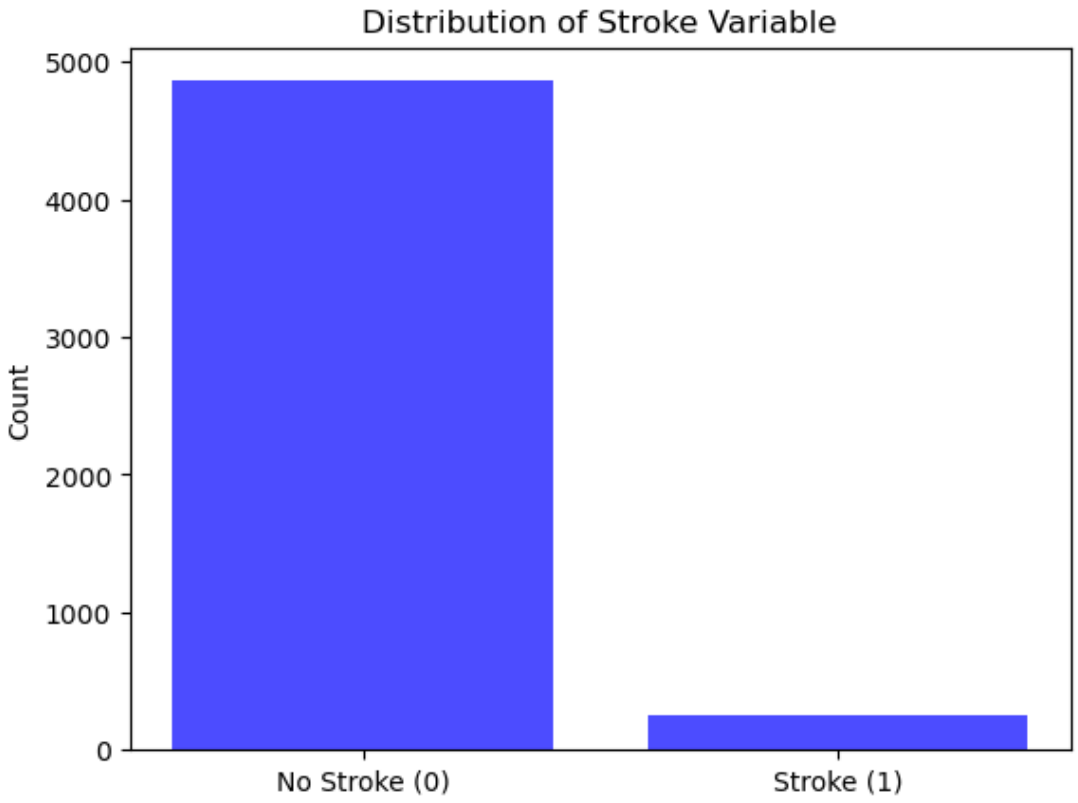

Figure1: Distribution of Stroke Variable

2.2 Stroke and Average Glucose Level

Figure 2 illustrates the relationship between stroke incidence and average glucose levels. It reveals that the mean average glucose level in stroke patients is approximately 25 mg/dL lower than in the non-stroke population. However, the upper quartile of average glucose levels in the stroke group is significantly higher compared to those without stroke.

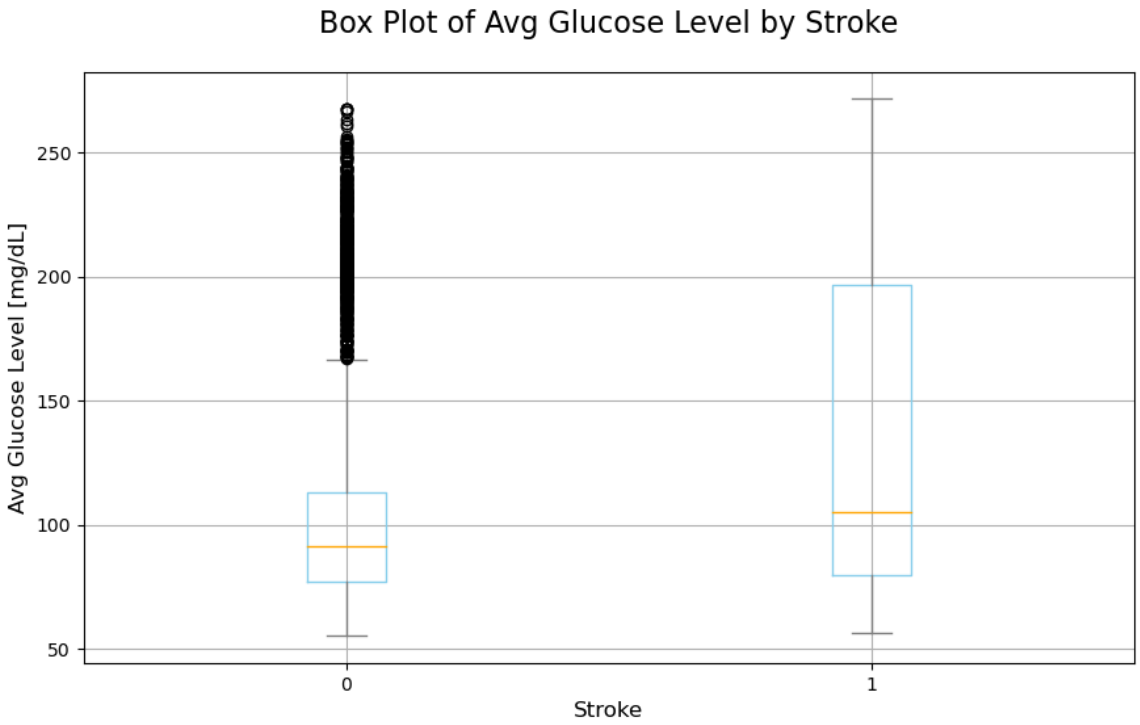

Figure2: Box Plot of Average Glucose Level by Stroke

2.3 Distribution of age by stroke patients

The histogram in Figure 3 displays the age distribution in the dataset, differentiated by the occurrence of strokes. It shows that stroke patients are primarily distributed in the age group above 40, and there is a noticeable increase in stroke incidence among individuals aged 60 and older.

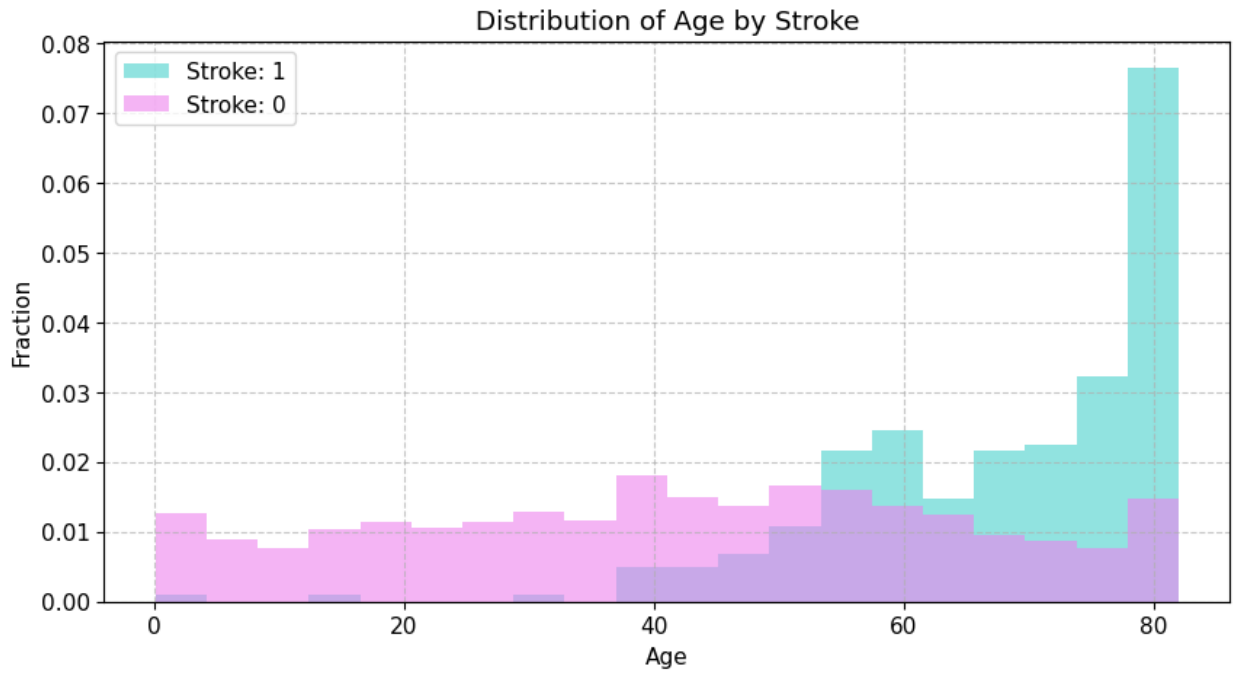

Figure3: Distribution of Age by Stroke

3. Methodology

3.1 Splitting

Two splitting methods were used on the dataset.

- Use simply splitting method to split data into other and test.
- Use StratifiedKFold to split other data into train and validation.
    - To deal with the imbalance data, StratifiedKFold Method was used to make sure in every random state, both points with target variable 0 and 1 can be trained and learned by the model.

3.2 Preprocessing

The features were preprocessed by OneHot Encoder, Standard Scaler, and MinMax Encoder. Preprocessing made the features get ready to be trained.

3.3 Hyperparameter Tuning and Cross Validation

To address the issue of imbalanced data, algorithms with weight parameters were employed for training. The table presents the algorithms used, along with the specific parameters for each, in the data training process.

Algorithm	Parameters
Random Forest	N_estimators [50, 100, 200]; Max_depth [None, 10, 20]; min _sample_split [2, 5, 10]; min _sample_leaf [1, 2, 4]; Class_weight ['balanced', 'balanced_subsample', None]
Support Vector Machines (SVM)	C [0.1, 1, 10, 100]; Kernel ['linear', 'rbf']; Gamma ['scale', 'auto', 0.1, 0.01, 0.001]; class_wieght [None, 'balanced', {0: 1, 1: 5}]
XGBoost	Learning_rate [0.01, 0.1, 0.2]; N_estimators [50, 100, 200]; Max_depth [3, 5, 7]; Subsample [0.8, 1.0]; Colsample_bytree [0.8, 1.0]; Gamma [0, 1, 5]; Scale_pos_weight [72]
K-Nearest-Neighbors (KNN)	N_neighbors [3, 5, 7, 10]; Weights ['uniform', 'distance']; Metric ['euclidean', 'manhattan', 'minkowski']; P [1, 2]

Table2: Algorithms and parameters

3.4 Model Pipeline

In the entire model pipeline, for each randomly selected statement, the data was initially divided into two parts: data_other and data_test. Within data_other, a further split was made into data_train and data_validation. Each model, fitted with various parameters, was trained using data_train and then evaluated using data_validation. The best-performing models for each algorithm were subsequently selected and then used to evaluate data_test, with the final output being the best models and their respective test scores.

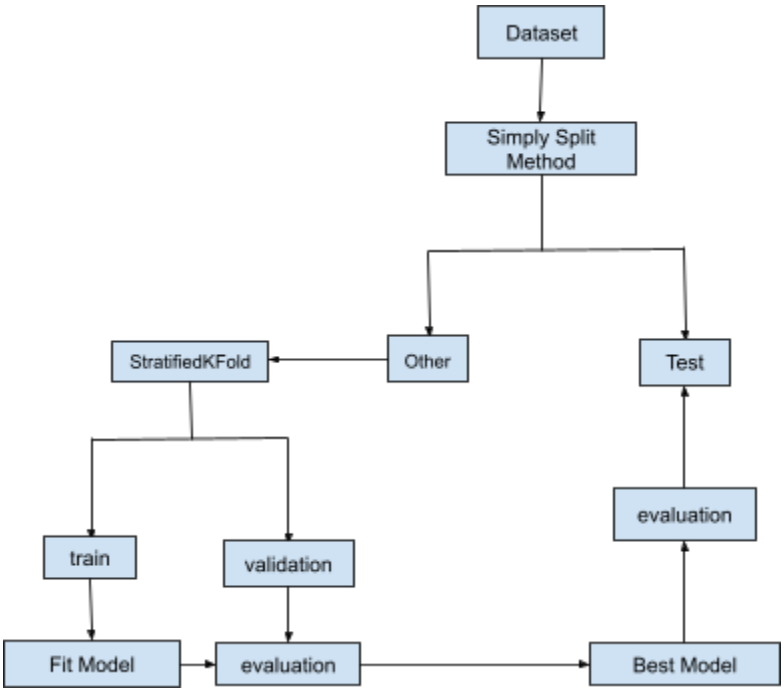

Figure4: Model Pipeline

4. Results

4.1 Accuracy

Table 3 presents the accuracy of the baseline model (which predicts all points with a target variable of 0) alongside the accuracy scores of the best models. The baseline model's accuracy stands at 0.95732, while the accuracies of the four algorithms are marginally lower than this baseline figure. Among these, the XGBoost model, despite having the lowest test mean accuracy, is selected as the final model for further analysis, including calculating feature importance. The rationale behind this choice will be elaborated upon in the section discussing the confusion matrix.

Model	Test Mean Score	Standard Deviation
Baseline	0.95732	

Random Forest	0.9569	0.0005
Support Vector Machines (SVM)	0.95702	0.0004
XGBoost	0.929	0.0074
K-Nearest-Neighbors (KNN)	0.9566	0.0008

Table3: Accuracy Scores and Standard Deviation for Each Model

4.2 Confusion Matrix

The initial baseline confusion matrix heatmap represents a scenario where all data points are predicted to have a target variable value of 0. This results in 42 false negatives, indicating that 42 instances with an actual target variable of 1 were incorrectly predicted as 0. The confusion matrix heatmaps for the Random Forest, SVM, and KNN models show a similar pattern, as they seldom predict the rare value 1 for the target variable. The primary objective of disease prediction models is to minimize false negatives and maximize true positives. In comparison with the baseline and the three aforementioned models, the XGBoost model's heatmap displays a relatively higher true positive rate and a lower false negative rate. Consequently, despite its marginally lower accuracy, the XGBoost model was selected as the final model for its improved performance in these critical areas.

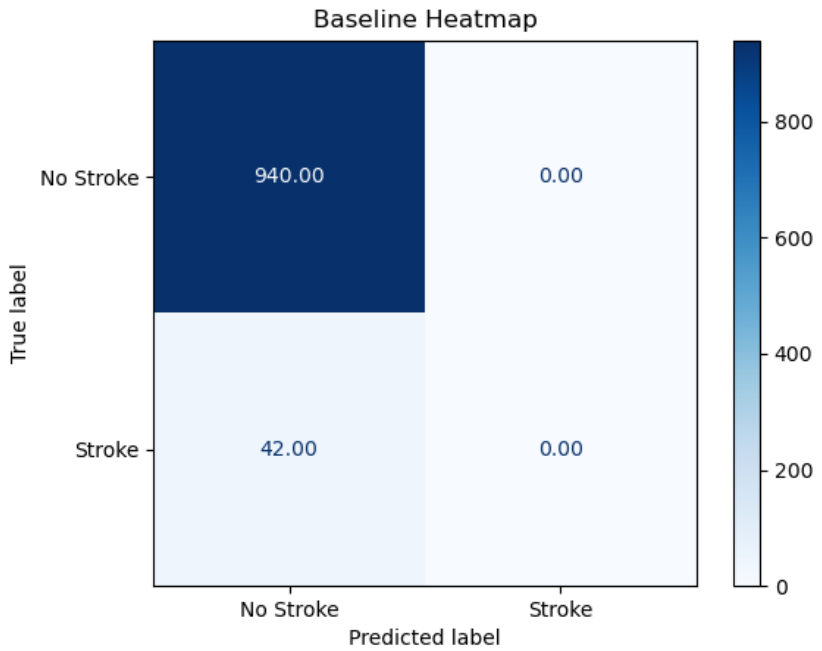

Figure5: Confusion Matrix Heatmap for Baseline

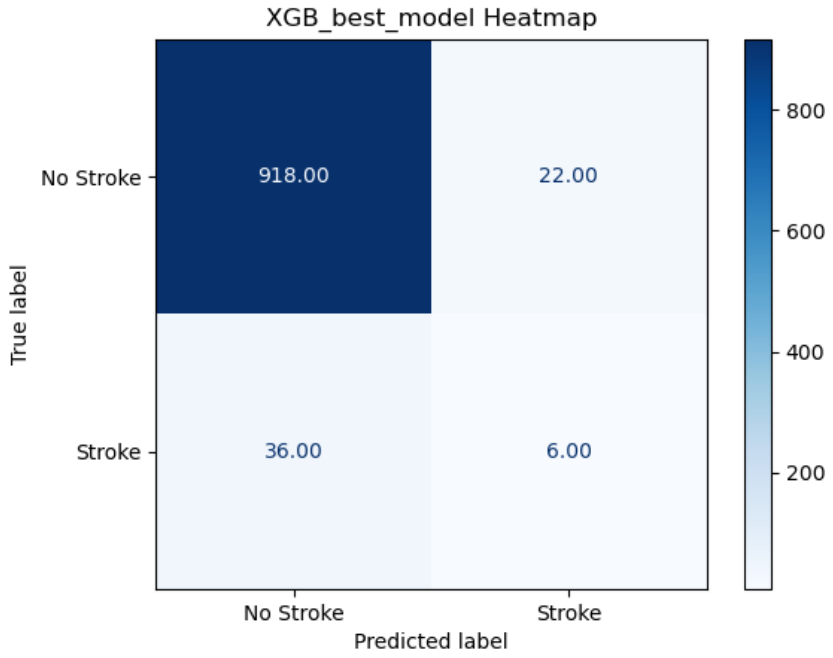

Figure6: Confusion Matrix Heatmap for Best CGBoost Model

4.3 Importance

4.3.1 Permutation Feature Importance

The plot illustrates how accuracy fluctuates following the permutation of different features. Notably, there are significant changes in accuracy after permuting the 'age' and 'work_type' features, suggesting that these two attributes play a relatively important role in the XGBoost model's performance.

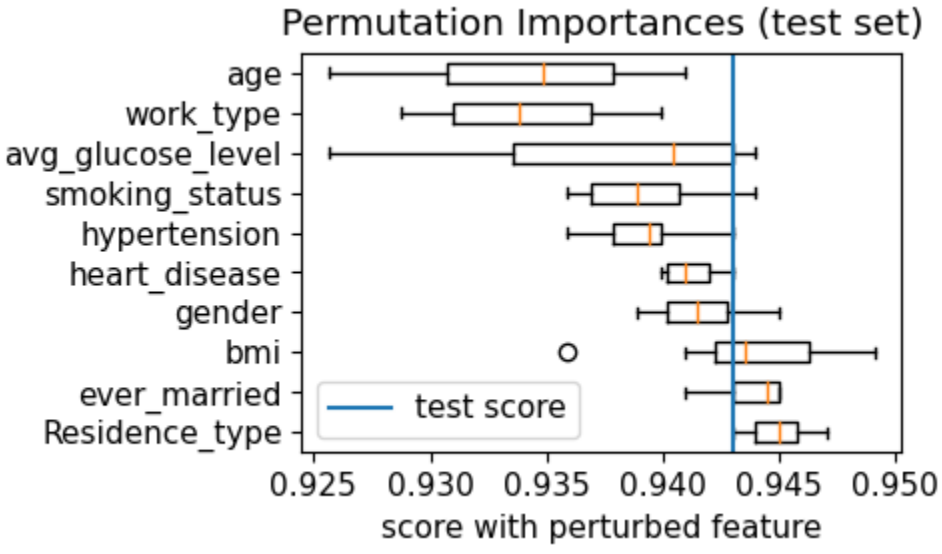

Figure7: Permutation Feature Importance

4.3.2 XGBoost Importance Metric – Total Gain

The figure displays the cumulative gain (improvement in accuracy) attributed to each feature across all splits in all trees where the feature is used. According to the figure, the total gains for the features 'mbi', 'average glucose level', and 'age' are significantly higher compared to the gains of other features.

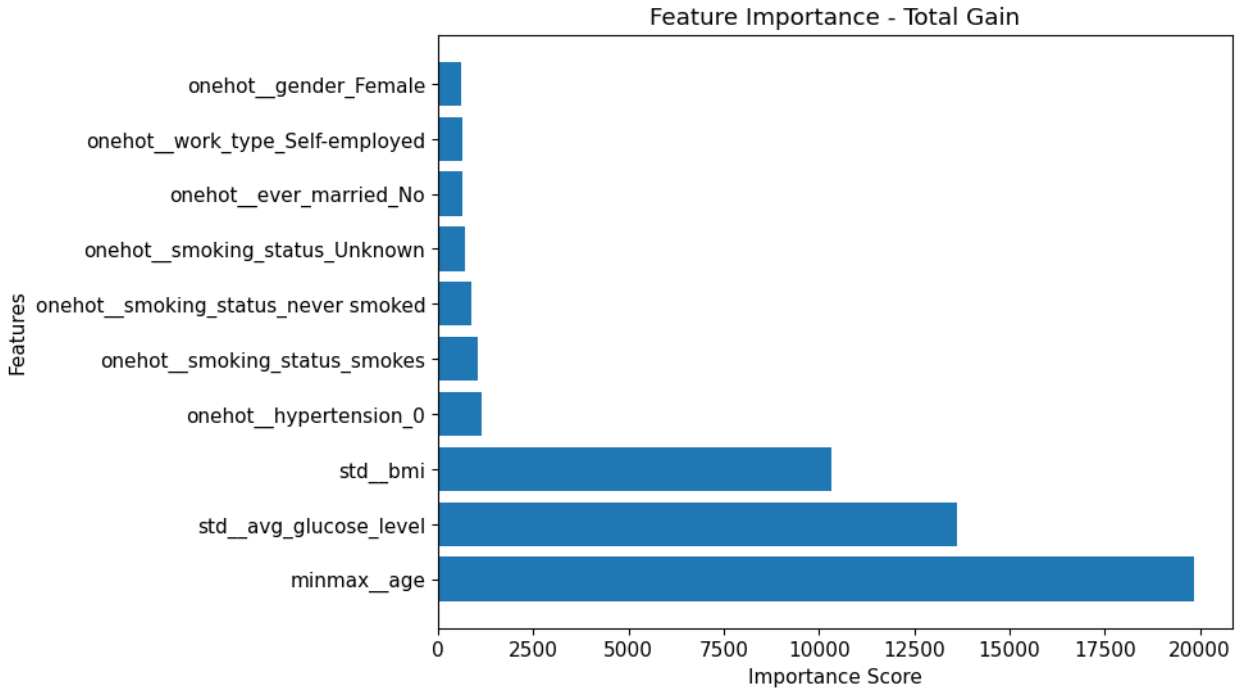

Figure8: Feature Importance - Total Gain

4.3.3 Local SHAP Force Plot

Figures 9 and 10 display local SHAP force plots for a data point with stroke values of 1 and 0, respectively. In Figure 9, the features 'age' and 'average glucose level' significantly contribute positively to predicting that a patient has a stroke, whereas the 'work type' feature has a notable negative contribution to the prediction of stroke 1. Conversely, Figure 10 illustrates that the 'average glucose level' and 'bmi' features have a substantial negative impact on predicting stroke 1.

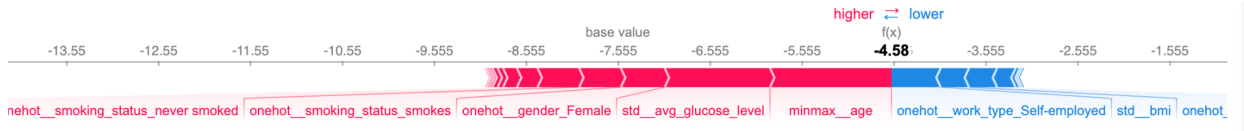

Figure9: Force Plot 1

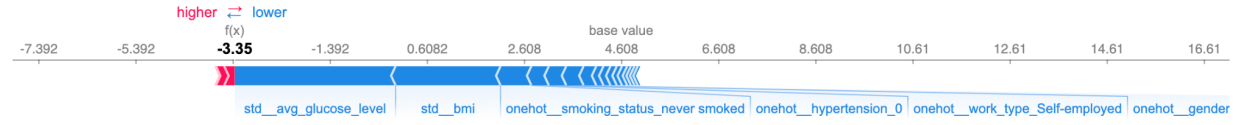

Figure10: Force Plot 2

5. Outlook

For future improvements, the focus will be on enhancing both the quality and quantity of the data. Crucial factors such as family history and diabetes, currently missing from the dataset, will be included to enable more precise predictions. Additionally, expanding the dataset, which currently comprises approximately 6,000 data points, and incorporating more instances with a target variable of 1, will help achieve a more balanced dataset and potentially enhance the model's predictive performance.

6. References

[1] Data Source - https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data
[2] Previous Work - https://www.kaggle.com/code/juliohernandezpabon/stroke-predictio, https://www.kaggle.com/code/danielshola/stroke-prediction-svc-and-logistic-regression
[3] GitHub Repository - https://github.com/jiahuz4-1976260/Data1030-Project.git