Stroke prediction

Capstone Project 1

Yuliya Selevich

Introduction

A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or ruptures. When that happens, part of the brain cannot get the blood and thus oxygen it needs, so it and brain cells die. According to the American Stroke Association, it is the No. 5 cause of death and a leading cause of disability in the United States. Therefore prediction of a stroke has key importance.

Data Wrangling

The dataset used for this project was downloaded from kaggle.com. The dataset is in tabular form and contains 5110 rows with 12 columns.

The dataset was relatively clean with no duplicate rows and erroneous values. The following steps were taken to transform the data into a more readily used format:

- 'id' column was dropped as it contained irrelevant information.
- The rows that contained age values below 17 were dropped. Even though a minor can have a stroke, it is very rare and normally caused by completely different reasons than those being studied in this project. Therefore it was appropriate to drop these rows.

Exploratory Data Analysis

Target variable

The clear choice for the target variable in this dataset is 'stroke'. The variable has two classes: 1 – for cases when a stroke occurred, and 0 – when there was no stroke in a patient's history. The target variable is highly imbalanced with only ~ 5.72 % of samples having a value of 1.

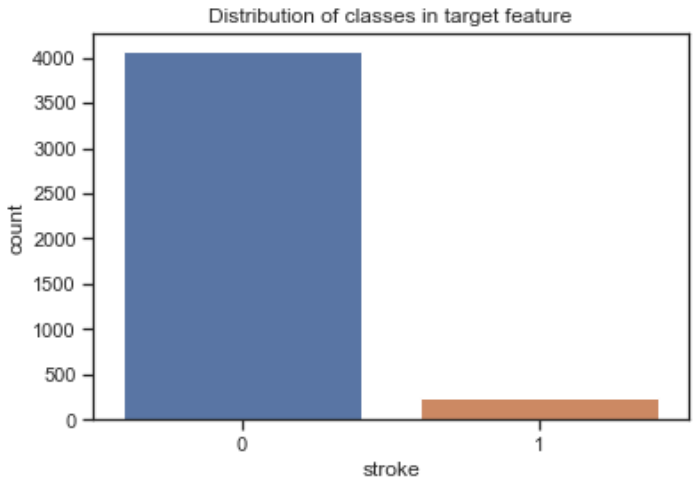

Figure 1 Distribution of classes in target feature

Predictor variables

There are 10 predictor variables, 3 of which are continuous and 7 are categorical.

Continuous variables

1. Average glucose level

Average glucose level doesn't have any missing values, however, the distribution of values is skewed to the right and bimodal, with a second significant group with elevated glucose levels most likely representative of diabetic and/or pre-diabetic patients.

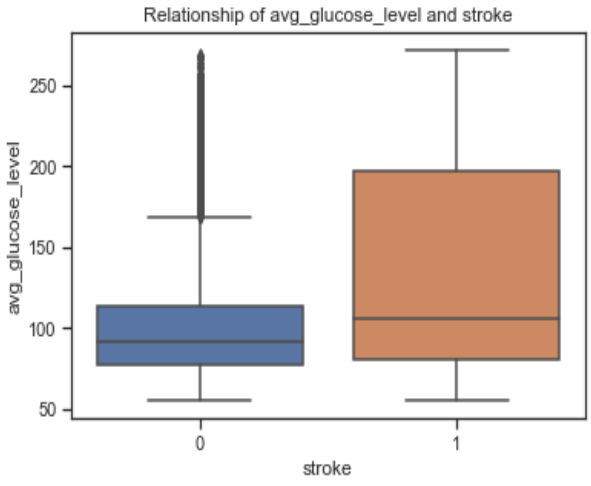

Figure 2 Relationship of avg_glucose_level and stroke

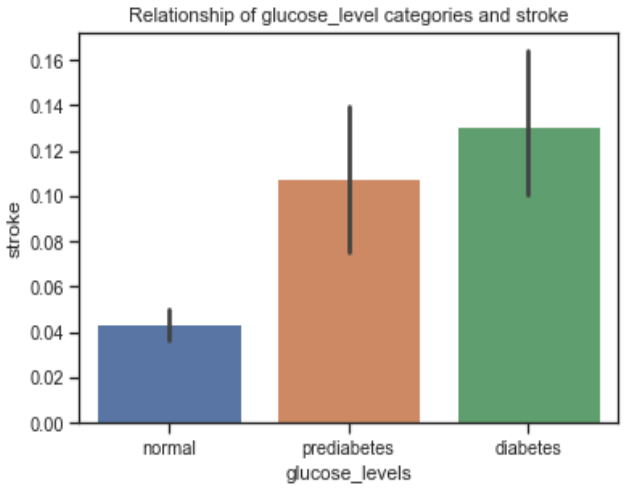

Figure 3 Relationship of glucose level categories and stroke

If we divide this variable into categories based on ranges corresponding to normal glucose level, prediabetes, and diabetes, we can see a stronger relationship between prediabetes and diabetes categories with stroke occurrence compared to normal glucose levels.

2. BMI

The distribution of Body Mass Index (BMI) values is skewed to the right as well. And, similar to the average glucose level variable, outlier values are not unnatural and can't be discarded. Therefore they will need scaling.

The BMI variable is the only variable in the whole dataset that has missing values. Even though it doesn't have a lot of missing values, deleting corresponding rows would cause a further imbalance of the imbalanced target variable. Therefore missing values will have to be dealt with in the data pre-processing step.

A comparison of stroke occurrence and continuous BMI values doesn't give a clear picture of their relationship with the median values being somewhat similar. However, dividing BMI values into categories helps to see that there is a stronger relationship between being in the overweight and obesity ranges and an increased likelihood of a stroke compared to the underweight and normal ranges.

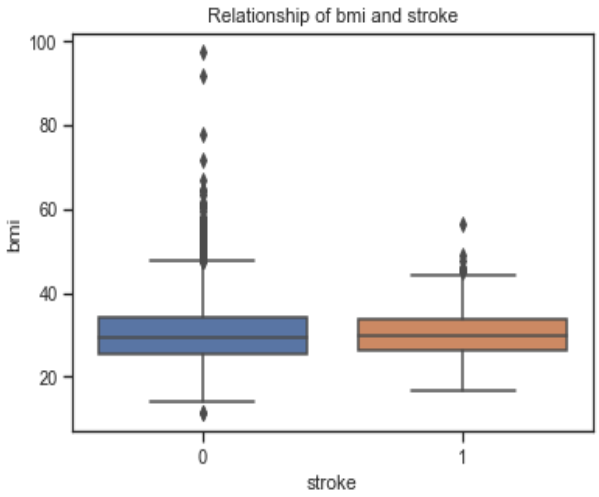

Figure 4 Relationship of bmi and stroke

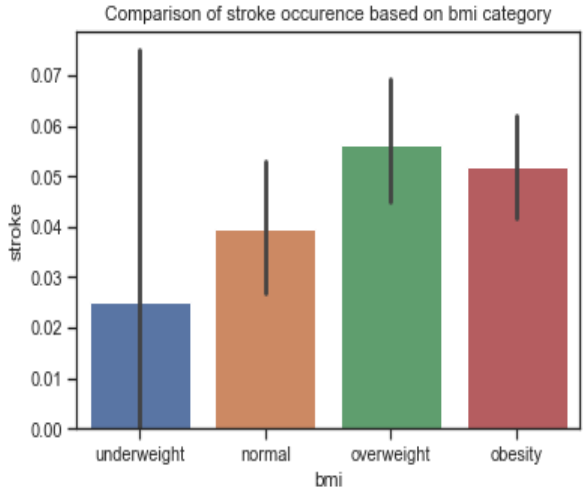

Figure 5 Relationship of stroke occurrence based on bmi category

3. Age

The age variable has a uniform distribution. There is a clear increased strength of a relationship between stroke occurrence with higher age. The median value for age for cases with no stroke is about 50 years, and the median value for cases when a stroke occurred is closer to 70 years.

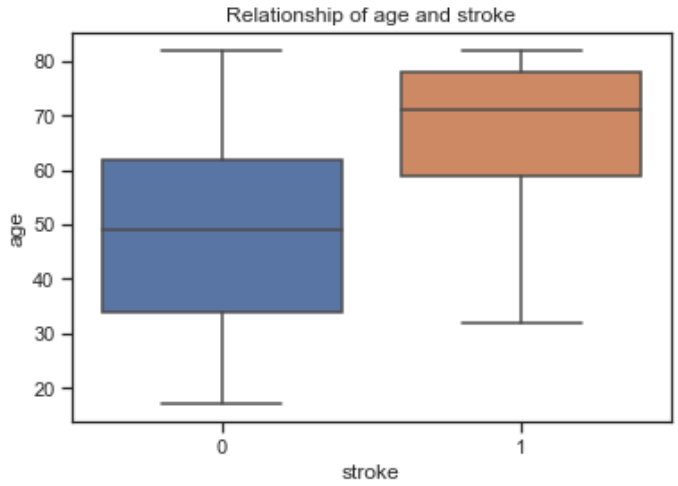

Figure 6 Relationship of age and stroke

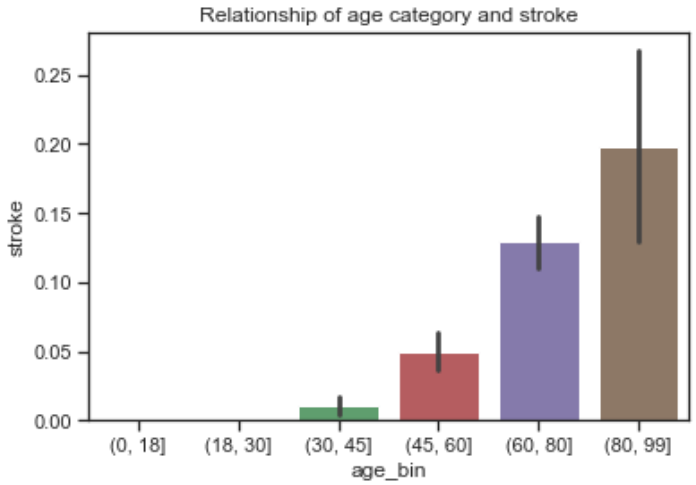

Figure 7 Relationship of age categories and stroke

Categorical variables

All categorical variables are pretty self-explanatory and don't require explicit description.

1. Gender

'Male' gender shows a slightly stronger relationship with class 1 of the 'stroke' variable. There is strong evidence that the occurrence of stroke is higher in men than in women in all age classes, and women are, on average, several years older than men when they suffer their first stroke (Wyller TB, 1999).

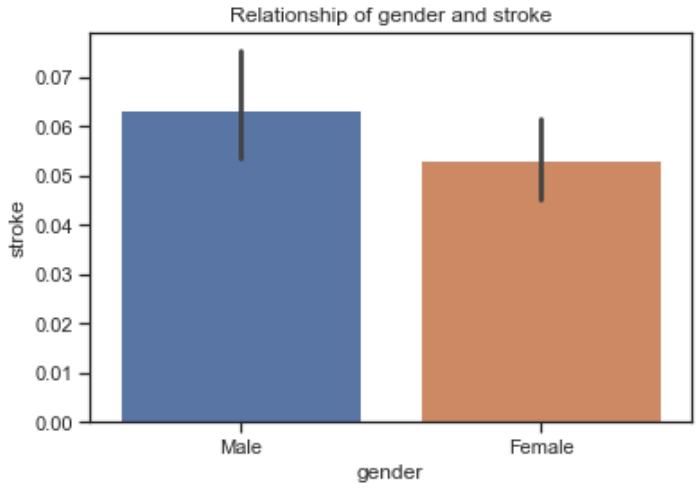

Figure 8 Relationship of gender and stroke

2. Hypertension and heart_disease

The plot below shows that patients with hypertension or heart disease are far more likely to have a stroke. According to the CDC, high blood pressure is a leading cause of heart disease and stroke because it damages the lining of the arteries, making them more susceptible to the buildup of plaque, which narrows the arteries leading to the heart and brain.

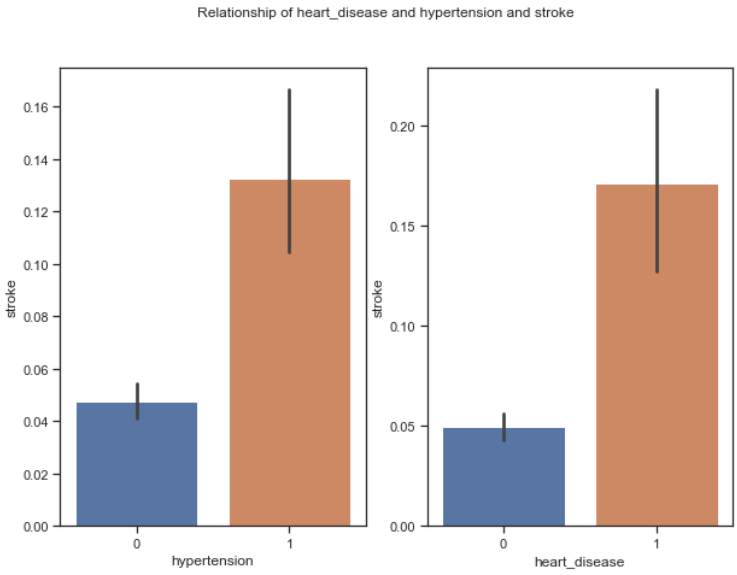

Figure 9 Relationship of heart_disease and hypertension and stroke

3. Work type

'Self-employed' category shows a somewhat stronger relationship with stroke compared to the 'private' and 'government job' categories. 'Never_worked' category shows

no relationship with stroke. Self-employed individuals have a higher prevalence of cardiovascular risks, including stroke, in the contemporary US population. Self-employed individuals may encounter adversity such as financial management, work-related stress, unstable working schedule, and a lack of essential health insurance, leading to cardiovascular risk (Krittanawong et. al., 2020).

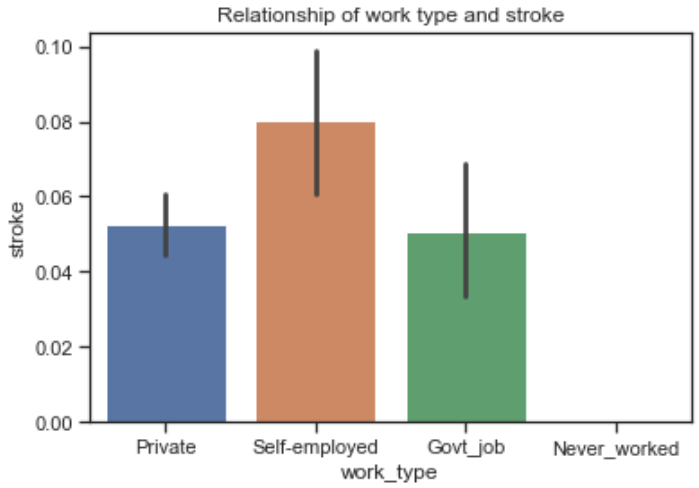

Figure 10 Relationship of work type and stroke

4. Ever married

Married people show a higher likelihood of developing a stroke which may be related to the fact that married people are also more likely to be of a higher age which also showed a strong relationship with stroke occurrence.

Figure 11 Relationship of marriage status and stroke

5. Residence type

Residence type doesn't seem to have a large effect on the development of stroke, with people of both 'urban' and 'rural' residences having a somewhat similar likelihood of having a stroke.

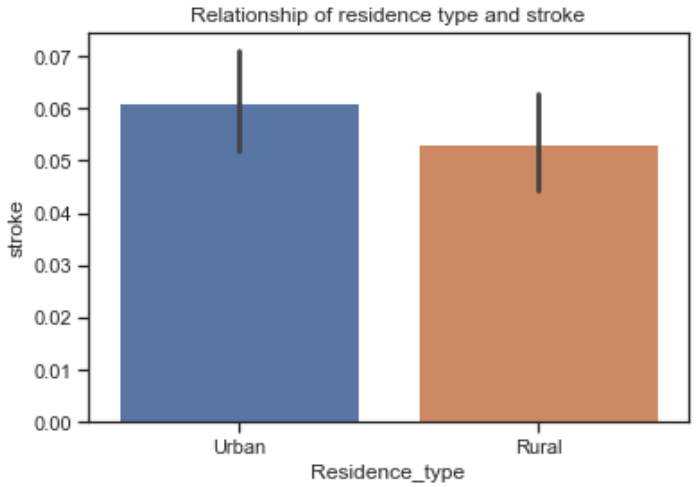

Figure 12 Relationship of residence type and stroke

6. Smoking status

Patients that formerly smoked show a surprisingly stronger relationship with stroke likelihood compared to current smokers. However, the evidence suggests that smoking discontinuance results in a considerable reduction in stroke risk across gender, race, and age (Shah and Cole, 2010). Therefore we can assume that a stronger relationship, in this case, may as well be related to a prevalence of patients of higher age in the 'formerly smoked' category.

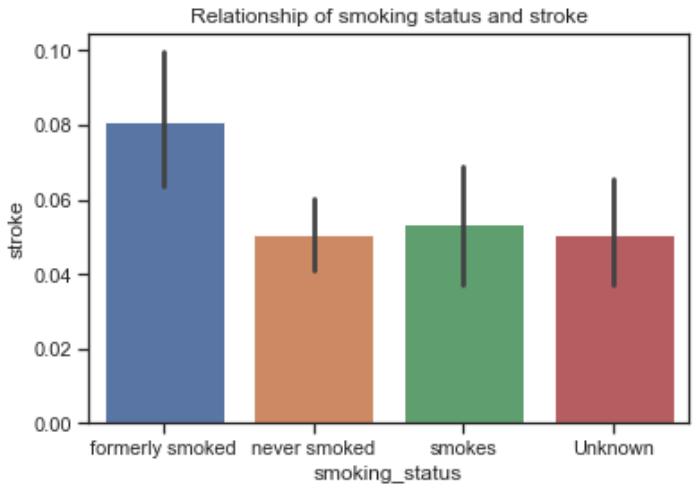

Figure 13 Relationship of smoking status and stroke

7. Statistical testing

Chi-squared test was used to evaluate the relationship strength of every independent variable and the target variable. The null hypothesis states that there is no statistically significant relationship between the two variables. Critical values were calculated for every variable for a significance level of 0.05. Next Chi-squared score was calculated for every variable using the 'SelectKBest' class of the scikit-learn. Critical values and Chi-squared scores of every pair of variables were compared. If the Chi-squared score is higher than a critical value, the null hypothesis can be rejected. Only five variables had scores higher than their corresponding critical values: age, heart disease, hypertension, BMI, and marriage status. Thus these variables show a strong statistically significant relationship with our target variable.

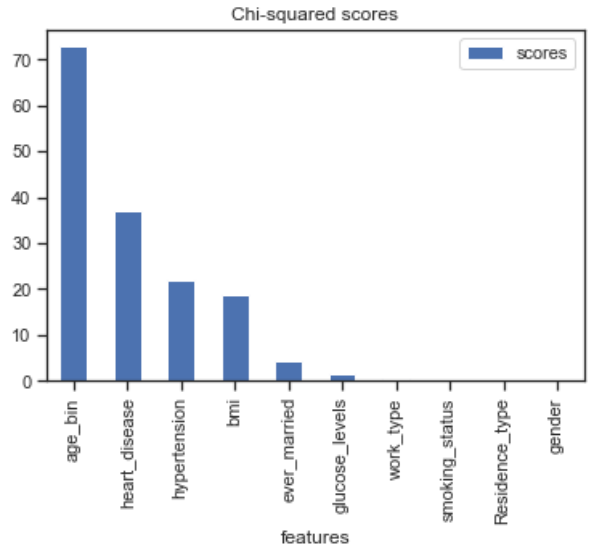

Figure 14 Chi-squared scores

Pre-processing

Age, BMI, and 'avg_glucose_level' are the only continuous variables in the dataset. 'avg_glucose_level' values have bimodal distribution and BMI distribution is skewed to the right which can affect the performance of a selected model. Therefore it was reasonable to transform these features into categories. The BMI and glucose levels variables were split based on the ranges universally accepted by the medical community. Missing values in the BMI feature were transformed into a separate category 'Unknown'. The age variable was divided into 6 bins.

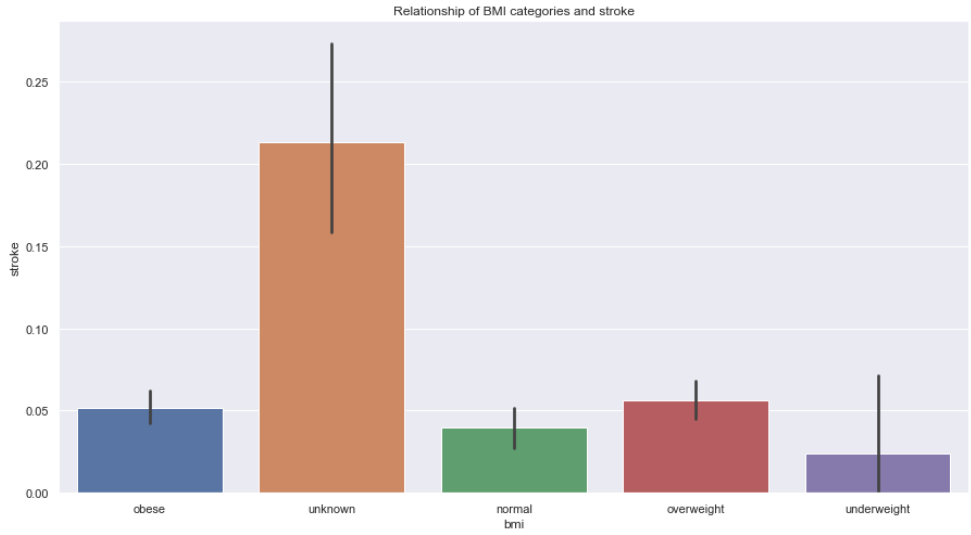

Figure 15 Relationship of BMI categories and stroke

Next, the whole data set was split into X_train, X_test, y_train, and y_test sets with X_test and y_test intended for model performance validation. Next get_dummies function was used to encode all categorical variables.

Modeling

Taking into consideration class imbalance in the target variable, StratifiedKFold was selected for cross-validation. StratifiedKFold cross-validator ensures that each fold of the dataset has the same proportion of observations with a given label, therefore it is appropriate to use it for this imbalanced dataset.

Three models were compared: Logistic Regression, Random Forest Classifier, and XGBoost Classifier. GridSearchCV was used to find optimal hyperparameters for each model.

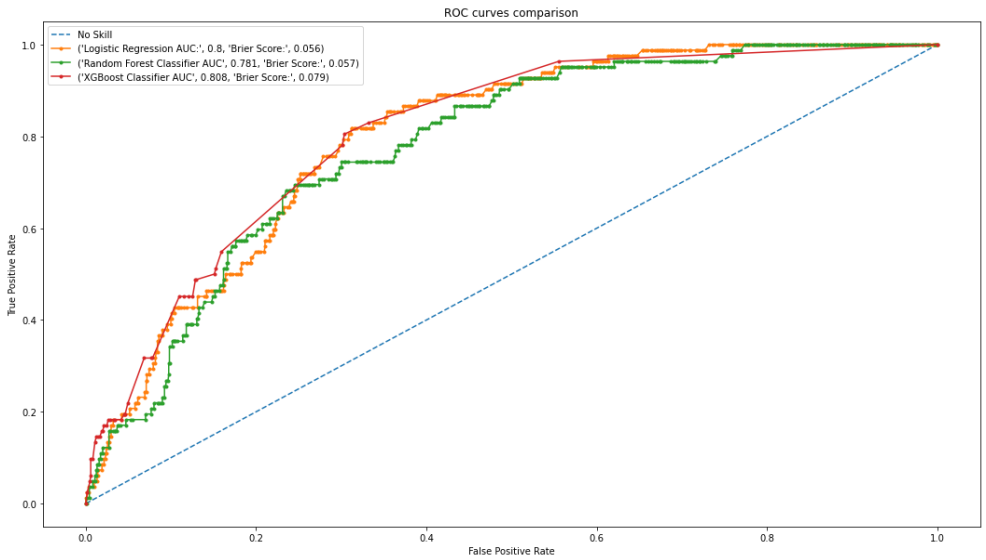

Figure 16 Comparison of ROC curves of all models

ROC-AUC and F-score were used as evaluation metrics. The Brier score that measures the mean squared difference between the predicted probability and the actual outcome, was also calculated for every model. Even though XGBoost Classifier has a slightly higher AUC score than Logistic Regression, I will use Logistic Regression as my final model as it has a better Brier score.

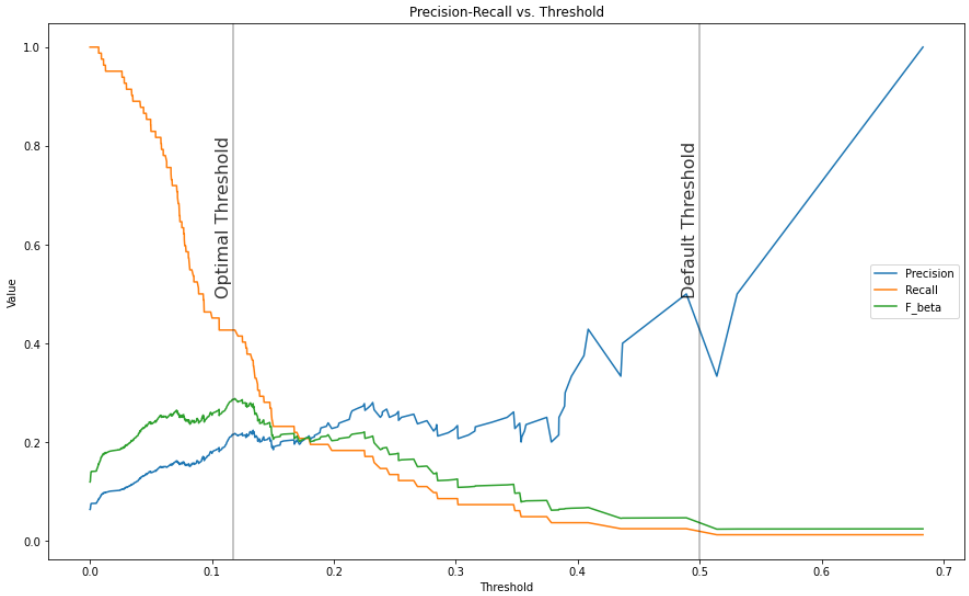

Figure 17 Precision-Recall vs. Threshold

If a Type I error (false positive) occurs in the case of stroke prediction, it may even be beneficial for a patient as it may encourage them to do proper medical examination and evaluation of overall health condition. But Type II error (false negative) is more dangerous, as the potentially life-threatening condition may get overlooked. Therefore it's crucial to emphasize minimizing the number of Type II errors. Thus it is reasonable to favor higher recall over precision and adjust a threshold used for classification. Therefore optimal threshold of 0.117 (compared to a default of 0.5) was calculated that increased recall from 0.01 to 0.43 without significantly reducing precision and overall accuracy.

Conclusion

While the hyperparameter tuning improved the performance of the final model and threshold adjusting increased recall, low precision and recall for positive class compared to significantly higher precision and recall for negative class in the target variable mean that the model will likely make both Type I and Type II error when making a prediction. Thus a single model cannot be used as the sole predictor of stroke occurrence, but some of its insights such as the strong effect of some variables on the outcome may be helpful for estimating personal health risks.

References

1. Wyller TB. Stroke and gender. J Gend Specif Med. 1999 May-Jun;2(3):41-5. PMID: 11252851.
2. Krittanawong C, Kumar A, Wang Z, Baber U, Bhatt DL. Self-employment and cardiovascular risk in the US general population. Int J Cardiol Hypertens. 2020 Jun 11;6:100035. doi: 10.1016/j.ijchy.2020.100035. PMID: 33442670; PMCID: PMC7287446.
3. Shah RS, Cole JW. Smoking and stroke: the more you smoke the more you stroke. Expert Rev Cardiovasc Ther. 2010 Jul;8(7):917-32. doi: 10.1586/erc.10.56. PMID: 20602553; PMCID: PMC2928253.