Stroke Prediction

Capstone Project Final Report

Szeling Annie Hsu

I. Introduction

Stroke is the No. 5 cause of death and a leading cause of serious long-term disability in the United States. According to CDC, stroke kills nearly 150,000 of the 860,000 Americans who die of cardiovascular disease each year. Stroke causes brain tissue to die and in turn lead to brain damage, disability, and death. Every year, about 800,000 people in the United States have a stroke and about 25% of the cases are recurrent. People who had a stroke before have a higher risk of having a recurrent stroke.

There are many factors contribute to the risk of stroke. Some are beyond one's control like gender and age. Others can be controlled like high blood pressure, diabetes, obesity and smoking habits. In this study, we will analyze 10 factors in relation to stroke and to predict the likelihood of a patient having a stroke.

Health care providers can be benefited from this study to assess the risk of their patients having stroke and to educate them on how to change their lifestyles and unhealthy habits to prevent deaths and long team disabilities.

II. Data Source

This dataset was chose from Kaggle.com for this study. The data file "healthcare-dataset-stroke-data.csv" has 5110 entries of patients' information. Each patient has the following information:

1. id: an unique identifier for the patient
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: "0" if the patient does not have hypertension, "1" if the patient has hypertension

5. heart_disease: “0” if the patient does not have any heart disease, “1” if the patient has a heart disease
6. ever_married: if the patient ever married “No” or “Yes”
7. work_type: “children”, “Govt_job”, “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: “0” if the patient does not have stroke, “1” if the patient has stroke

III. Data Wrangling

1) “bmi” has 201 missing values and was imputed with the median BMI.
2) “id” was dropped as this information is not needed for our analysis.
3) The “Other” type of “gender” was dropped as there was only one entry and we treated it as the patient’s information was not available.

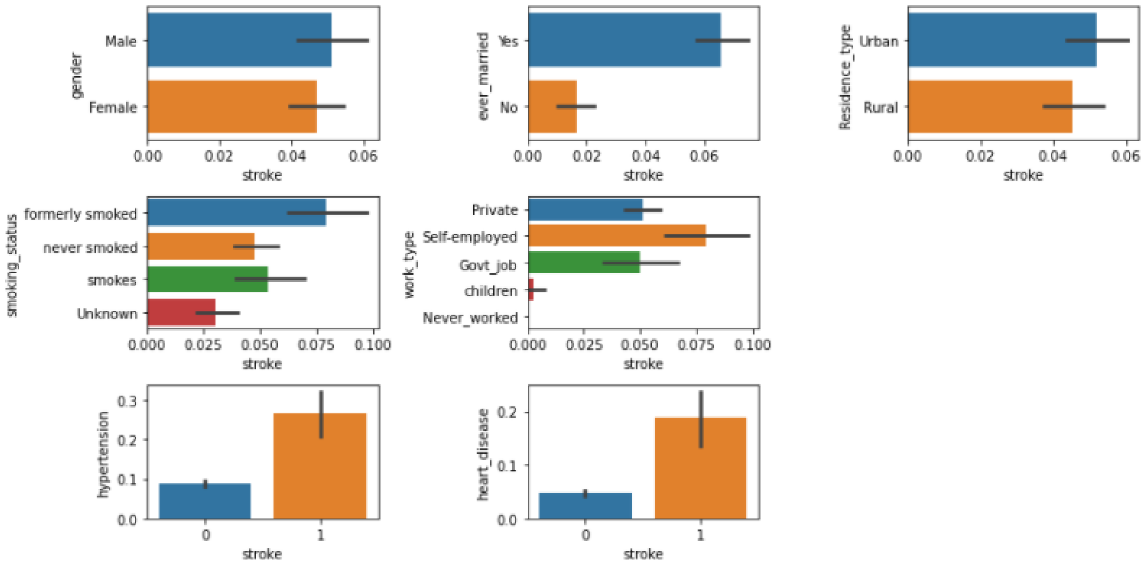
The cleaned dataset was stored in “healthcare-dataset-stroke-data-cleaned.csv”.

IV. Exploratory Data Analysis

There are 7 categorical variables and 3 numerical variables.

1) Categorical variables

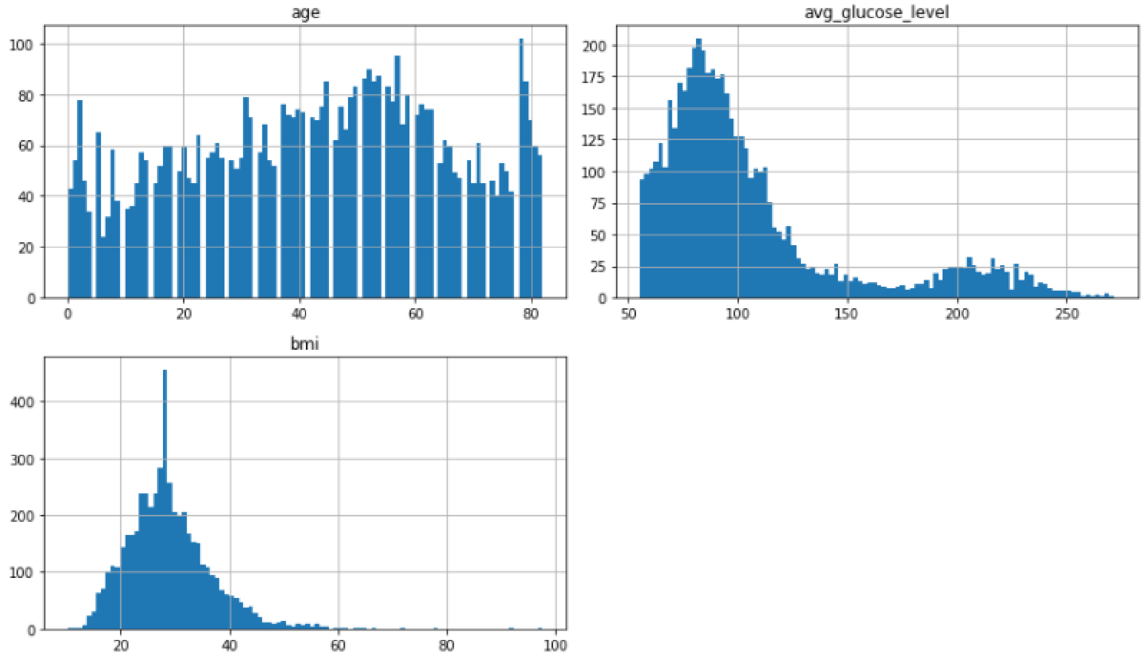
From the plots of each categorical variable vs the target variable “stroke”, we see that male and female seem to have about the same chance getting a stroke. If a person is ever married, the chance of having stroke seem to be much higher. The residence type does not seem to make a different. If a person smokes or formerly smoked has a higher chance to get stroke also. Self-employed people has a much higher chance getting stroke. One possible reason maybe they endue higher stress. If a person has hypertension or heart disease, the chance of having stroke seems to grow triple. The plots below show each of the categorical variables vs the target variable.



Plots of the seven categorical variables vs target variable

2) Numerical variables

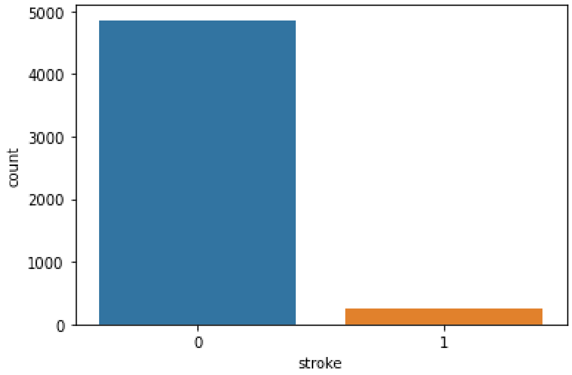
From the plots of the distributions of the three numerical variables, we see that stroke can happen in all ages. We will explore if the age, average glucose level and the BMI contribute to the chance of having stroke.



Plots of the three numerical variables

3) Target variable

There are 249 patients had stroke and 4860 patients did not have stroke before. We have an imbalanced dataset. Imbalanced dataset is very normal in medical data sets as most people are healthy whereas there is a much smaller number of people are having certain kind of medical condition.



Plot of the target variable

V. Models

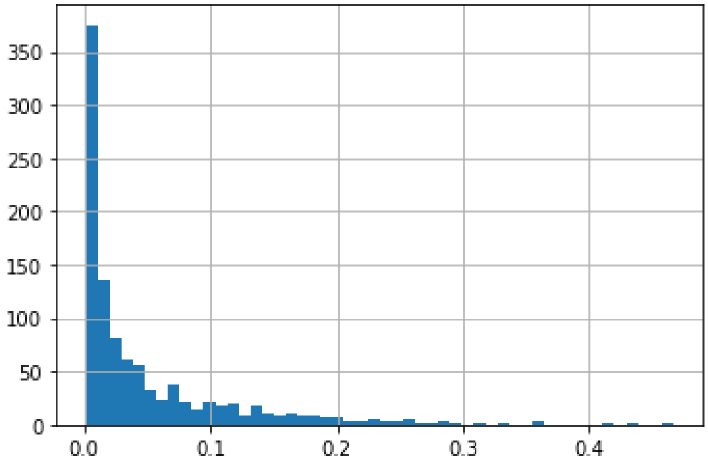
To solve this binary classification problem, we built several models to compare the performance. Predictive models built include Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, Gradient Boosting and XGBoost. The train-test split is 80% and 20%.

VI. Analysis

We first built a model with the Logistic Regression algorithm. Both the train and test accuracy of the model reached 95%. We also use the confusion matrix and the classification report to evaluate the precision and recall rates of the model. The precision rate refers to the ratio of the results which are relevant and the recall rate refers to the ratio of total relevant results correctly classified.

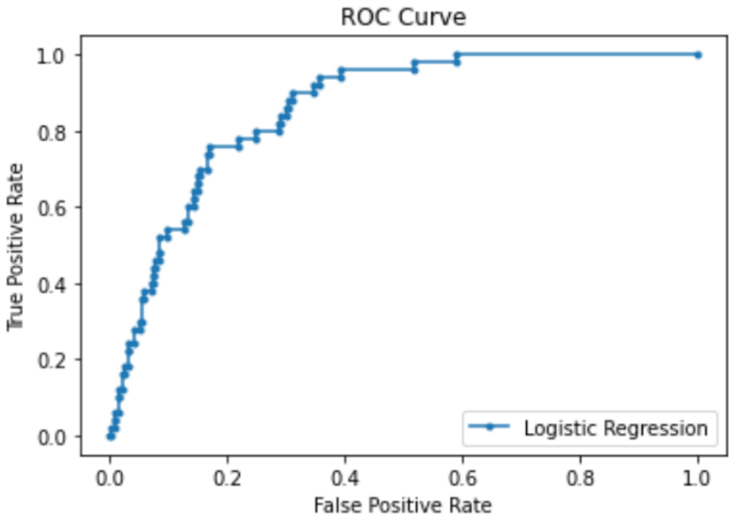
$$\textbf{\textit{precision}} = \frac{\textit{true positives}}{\textit{true positives} + \textit{false positives}} \qquad \textbf{\textit{recall}} = \frac{\textit{true positives}}{\textit{true positives} + \textit{false negatives}}$$

The classification report showed that both the precision and the recall rates for the minority class (class “1”), positive for stroke, are 0. None of the positive cases was captured. This problem is caused by our imbalanced dataset. Many machine learning algorithms have the default threshold set at 0.5 for interpreting probabilities to class labels. All values equal to or greater than the threshold are mapped to one class and all other values are mapped to another class. The plot of the distribution of the predicted probabilities for the minority class shows all of the probabilities are less than 0.5. This was why all the positive class samples were mislabeled. The imbalanced dataset led to learning bias on the majority class (class “0”).



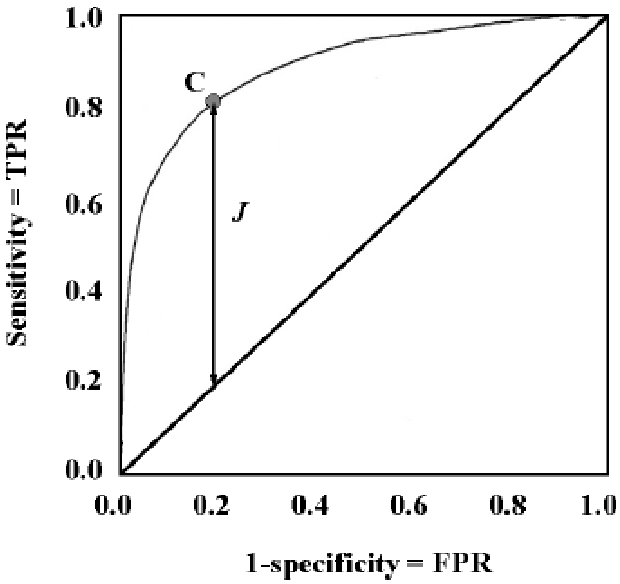
Plot of the distribution of the predicted probabilities for the minority class

To evaluate the diagnostic ability of a binary classifier, a receiver operating characteristic (ROC) curve is usually used. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. The area under the ROC curve is usually used to compare the performance of different classifiers. The area under the ROC curve for our model is 0.86 and the ROC curve for the minority class is shown below.



ROC curve for the minority class

In order to improve the performance of our model, we used a method called Threshold-moving to find the optimal threshold to map the probabilities to class labels. To find this optimal threshold that results in the best balance of the true positive rate and the false positive rate, we chose to use the Youden's J statistic to help finding the optimal threshold. The Youden's index equivalents to the vertical distance above the diagonal no discrimination (chance) line to the ROC curve for a single decision threshold. The maximum Youden's index gives the best balance of the true positive rate and the false positive rate.



A Youden's index represents on a ROC curve

The optimal threshold for our model was founded to be 0.08. With this optimal threshold, the recall rate has a significant improvement from 0% to 74%.

In medical setting, the diagnostic odds ratio (DOR) is usually used to measure the effectiveness of a diagnostic test. It is the ratio of the Positive Likelihood Ratio LR+ (the probability of detection) to the Negative Likelihood Ratio LR- (the probability of false alarm). The calculation involves all four values from the confusion matrix – the true positives, the false positives, the true negatives, and the false negatives. It is considered to be a more well-balanced metrics to measure a model as all four values are taken into account, where as the calculation for the precision rate and the recall rate do not take the true negatives into account. The model DOR score is found to be 14.0. By current standards, a DOR of 10.00 is considered to be a very good test.

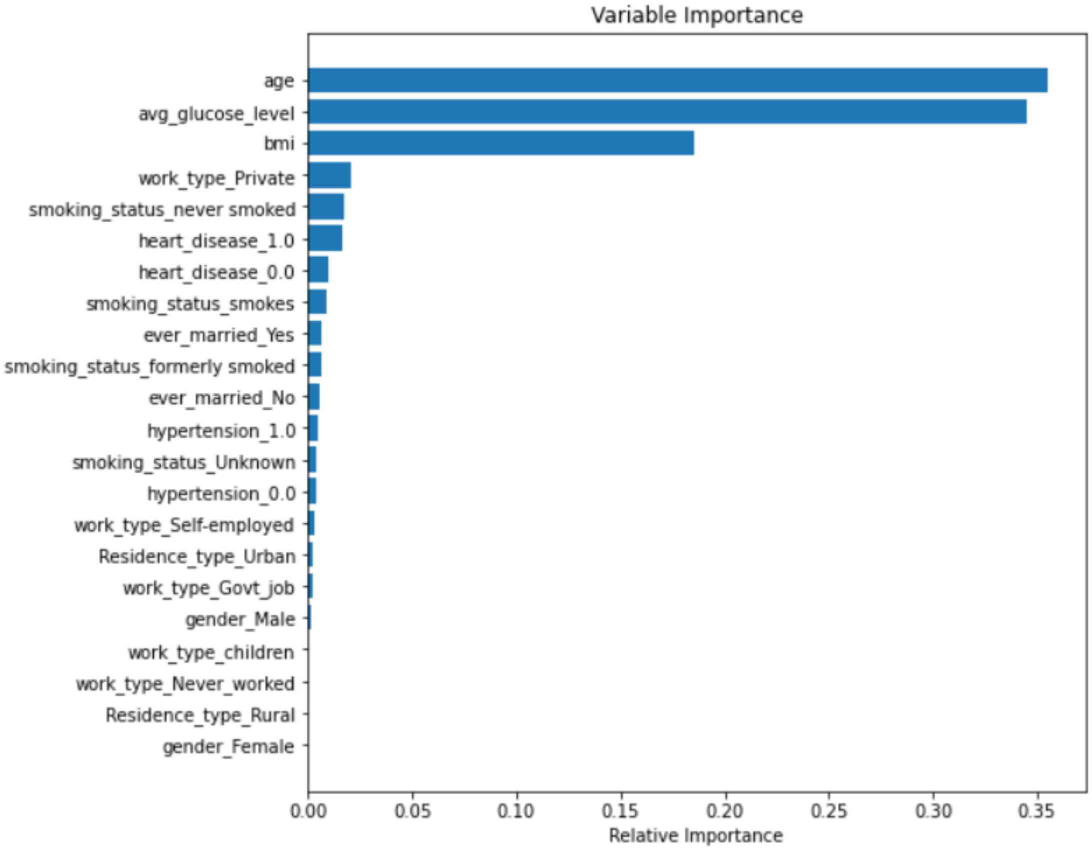
After choosing the metrics to measure our model, we built more models using other algorithms. We chose to use K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, and XGBoost algorithms. We also tried to tune our Logistic Regression, Random Forest and Gradient Boosting models to improve model performance. The following chart is the model performance comparison.

Model	ROC AUC score	DOR score
Tuned Gradient Boosting	0.876	19.483
Tuned Logistic Regression	0.859	16.506
Tuned Random Forest	0.858	15.833
Gradient Boosting	0.862	15.081
Logistic Regression	0.859	14.023
Random Forest	0.790	8.023
XGBoost	0.809	4.985
K-Neighbors	0.709	4.358
Decision Tree	0.568	NaN

Model performance comparison chart

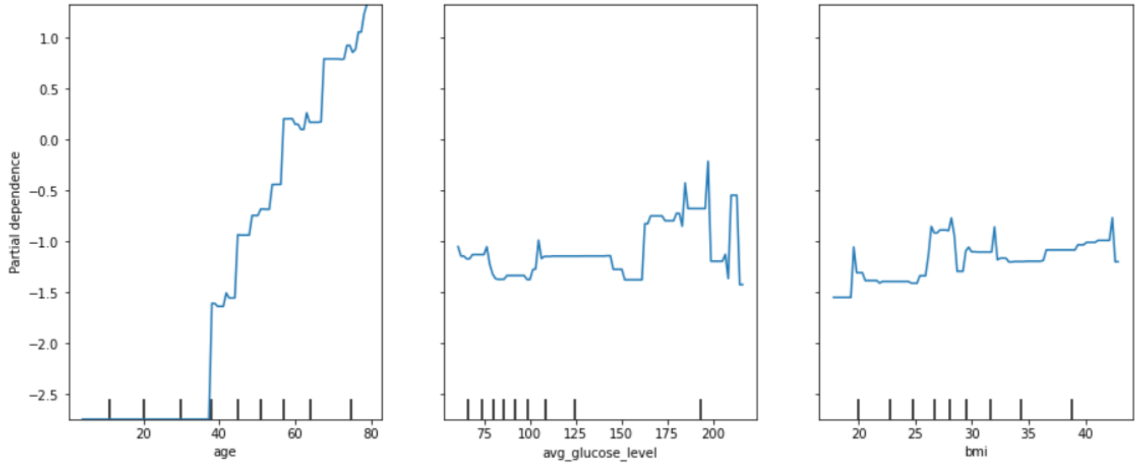
The model performance comparison chart shows that the tuned Gradient Boosting model yields the highest ROC AUC score and DOR score.

Among all of the features, we found that the age, average glucose level and the BMI are the top three most important features. Just the age and the average glucose level contribute almost 70% of the importance.



Feature Importance plot

From the partial dependence plot of the three most important features, we found that the risk of having stroke increases by age after the age of 40. If a person has pre-diabetes or diabetes (average glucocse level above 150 mg/dL), the risk will also increase.



Partial dependence plot of the three most important features

VII. Conclusion

This project is to evaluate 10 features contributing to the risk of having stroke and to build predictive models to determine if a patient has stroke according to those features. The age, the average glucose level, and the BMI were found to be the top three most important features contributing to the risk of having stroke. Six predictive models were built and the tuned Gradient Boosting model yielded the best AUC ROC score and DOR score.

For future improvement, we can try other methods to deal with the imbalanced dataset problem. We can collect more samples to balance the majority and the minority classes. Another method we can try is to oversample the minority class. We can try the Synthetic Minority Oversampling Technique (SMOTE) to synthesize new samples from the existing ones.

VIII. References

1. Know the Facts About Stroke: https://www.cdc.gov/stroke/facts_stroke.htm
2. About Stroke: https://www.stroke.org/en/about-stroke