Stroke Prediction

Using a dataset of patient attributes to determine the likelihood of stroke. Preventative measures could be taken to save patients from the detrimental side effects of having a stroke.

Fay Dennis

September 2021

Final Report

Stroke Prediction

Problem Statement

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient. What patterns in a patient's attributes make them more or less likely to have a stroke? How can we track that and lower overall healthcare costs to the patient, and what measures could be taken to prevent a patient from getting a stroke?

By using ## data, I hope to see if there is a way to predict if a patient will or will not have a stroke. Using a variety of models to see how well they can predict I hope to make as best of accurate probability as I can.

The data set starts with 5110 observations and 12 attributes.

Data Wrangling

The raw dataset from Kaggle contains 5110 observations and 12 attributes. The user who posted the data is named fdesoriano on Kaggle it is called the 'Stroke Prediction Dataset'. It is fairly sized, but needs a little cleaning. There were some missing values in the 'bmi' column, about 4% of the observations. In this instance I decided to fill the missing values with the mean of that column. The 'patient_id' column was an unnecessary one, so I decided to drop that. But with so few attributes in the dataset, I didn't really want to drop too many columns.

There was also some categorical data, for instance Male/Female column that I turned into numerical data, 0=Female, 1=Male.

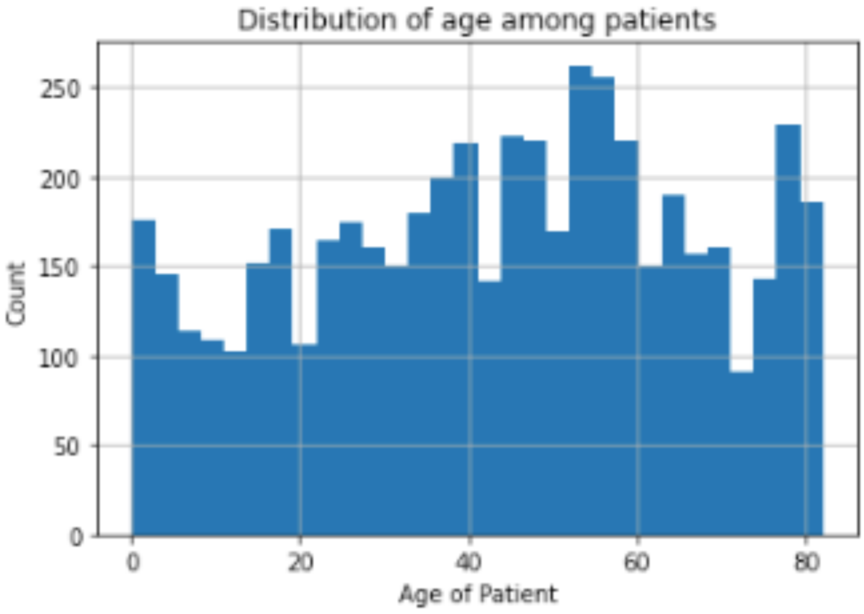
Exploratory Data Analysis

I started doing research about what attributes(or factors) lead to a stroke. This lead me to the Mayo Clinic to do some research about what factors really do contribute to a stroke. And according to the Mayo Clinic there are many factors. So I took the factors that paralleled with what they said, and what attributes I had in my dataset, as follows:

- Age
- Hypertension (another name for High Blood Pressure)
- Heart Disease
- Average Glucose Level (which can be an indicator of Diabetes)
- BMI or Body Mass Index (which can be an indicator of obesity, and a patients physique)
- Smoking Status
- Stroke (If they have had a stroke before or not)

I then did some EDA about what each of these factors looked like graphed, some graphs I did were just simple value-counts, some were histograms, some were count plots.

Some of the insights I learned from my dataset were very interesting, such as when I plotted the different age ranges from my dataset.



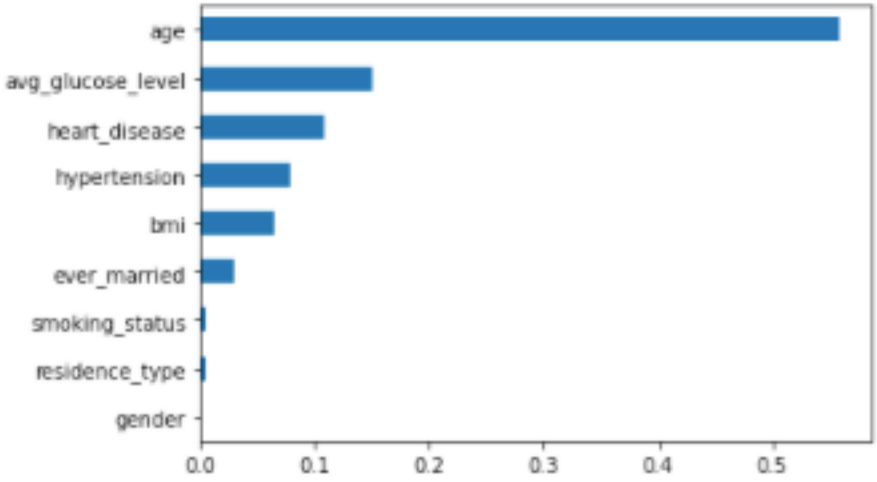
Heart Disease and Hypertension (aka blood pressure) were also interesting because of how little percentage of patients in my dataset had those factors. Only 5.4% of patients have been diagnosed with heart disease, and only 9.7% of patients were diagnosed with hypertension.

My hypothesis is that age is going to be a huge factor of stroke, and I think its going to be important to keep all age ranges in my dataset.

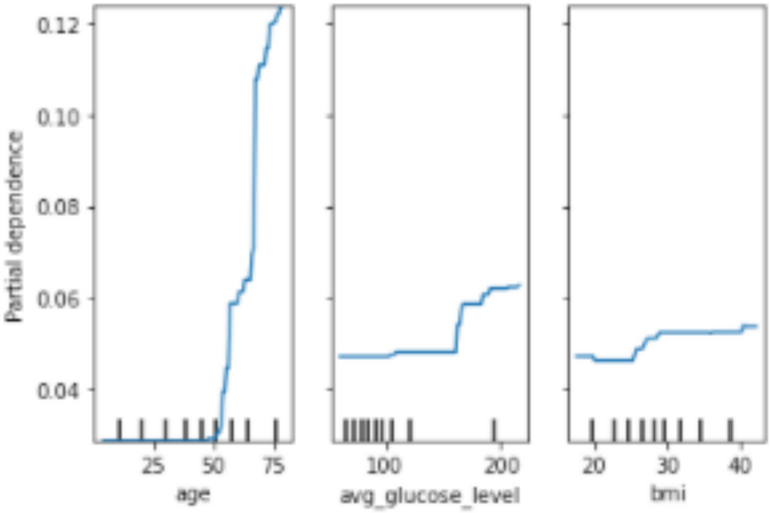
Model Selection

In order to see if we can determine a stroke many happen or not I tested 3 different machine learning classification models: Logistic Regression, Random Forest Classifier and Gradient Boosting. The problem with this was I found out that with the train/test split we imbalanced dataset. So with this I did a precision recall plot to try and find and ideal threshold number for each model.

For the Logistic Regression threshold, we found the best threshold was 0.35, for Random Forest the best threshold was around 0.11, for Gradient Boosting the best threshold was 0.06. I then looked at feature importance for each model, and saw that age was in fact the variable that had the biggest impact on stroke prediction. Here is the feature importance plot that I ran with Random Forest model.



I then also plotted the partial dependence plot to see the direction and when we'd expect increases in stroke likelihood for a given variable. As you can see after age 50, your chance of stroke goes up! This was a super interesting thing to plot.



Comparing the algorithms the train/test scores where both very similar, all of them were in the 79% to 84% of the ROC-AUC scores. All together Random Forest turned out to be the best model selection with its high scores, even though Logistic Regression had very similar training and test scores.

Conclusion

Logistic Regression is the best model for my stroke prediction capstone, although Random Forest is a very close second. Random Forest tended to overfit on the training set. But all in all a simple model worked better on this dataset because it was so small (only 5000 patients). I did try hyperparameter tuning on the Random Forest model, but it didn't seem to do any better and took a long time to run. I would have liked to add more data, especially on patients that were diagnosed to have already had a stroke. As it stood my dataset only had about 4.9% of the patients did already have a stroke.

I would like to take this project further with more data and even different patient datasets, because as it stands the kaggle dataset was too small. I think it would be an awesome idea to take this further, for the benefit of public health knowledge.

Even though many people know the risk factors of stroke, to see their actual different attributes of their health plotted visually would be more eye-opening. I hope that the public should have access to their data, and make these life changing decisions more clear!

My Process

I started with the dataset from Kaggle, which was already pretty clean to be honest. There were 5110 observations and 12 attributes, which I did end up dropping 3 of the columns (‘id’, ’Residence_type’, and ‘work_type’) because they were not important factors of stroke or stress, and couldn’t be quantified anyway. I did find that the ‘bmi’ column had 201 counts of missing values, about 3.9%. I filled the missing values with the mean of the column because BMI is an important factor of a person’s physique, and an important factor of obesity. I did also have to deal with some categorical features of the dataset such as gender, ever-married, smoking_status. I thought it was important to translate those categorical features to numerical features, so I did that. For instance with gender, Females became 0, and Males became 1.

It was at this point in my project that I started to do lots of research about what factors make up a stroke, and I copied them from the Mayo Clinic onto my jupyter notebook so that I could compare them to the attributes that I had in my dataset. After this I plotted different attributes to get some visualization on my data. This is when I found out that my data had age ranges from 0 to 80 years old! This was a very interesting part of my EDA exploratory data analysis. The correlation heatmap showed that there was higher correlation between ‘age’ and ‘ever_married’. I also looked into the ‘smoking_status’ attribute, which is interesting to think of because there are children/young people in the dataset. There are four responses: ‘Unknown’, ‘never_smoked’, ‘formerly_smoked’, ‘smokes’. Smoking is an important factor when looking for stroke because it does things to your body like, thickening and narrowing of blood vessels, increase buildup of plaque, and makes blood more likely to clot. 67% are ‘Unknown’ and ‘never_smoked’, but I don’t know how to qualify this.

In my preprocessing/training and modeling part of my project I started to split the data into X equaling the dataset without the stroke column, and y equaling the stroke column. I then decided I would use three different models for this project: Logistic Regression, Random Forest, and Gradient Boosting. For each of the models I fit the model, and looked at the shape. I then predicted on the test set for scoring and diagnostics. The threshold for each model was going to have to be figured out, so a precision recall plot was the best way to visualize the best threshold. It was also important to plot the feature importance in each model so see when we’d expect increases in stroke likelihood for a given variable. Age was by far the most important variable and that also ran true when I used a partial dependence plot. The likelihood of stroke after age 50, shot nearly straight up. What I found was Logistic Regression and Random Forest were the two best performing models.

Comparing the performance of LR and RF. Logistic Regression had a ROC-AUC train score of 83.87% and a ROC-AUC test score of 84.11%. Random Forest had a ROC-AUC train score of 84.23% and a ROC-AUC test score of 82.72%.