Brooke Hanson

Final Report:
Analysis and Modeling of Stroke Data Set

Problem Statement:

In the United States someone dies from a stroke every 4 minutes. This startling statistic from the CDC sheds light on the prevalence and frequency of this ailment. With strokes being one of the leading causes of death in the United States it begs the question, is it possible to predict if an individual will have a stroke? What factors impact stroke occurrence the most? These questions attempt to be answered by the analysis and modeling of stroke data.

By using a data set from Kaggle.com which included demographic, health, and stroke information, I created a tool that can help individuals and their healthcare providers to identify the probability of future stroke occurrence. I created this tool with the hopes of mitigating risks and decreasing fatalities associated with strokes. This tool was created by utilizing exploratory data analysis followed by machine learning algorithms to identify the best predictive model for the variable of interest.

Data Wrangling & Cleaning:

For this data set I found, there was not a huge necessity for an extensive data wrangling process. The data set contained 5110 rows of data with 11 features. There were some NA values in BMI which I filled using the median value, as well as a value in Gender valued ‘Other’ which I substituted ‘Female’ for as it was the gender with a higher count. The final data set I analyzed still had 5110 entries with 10 predictive features and stroke being the response variable.

Exploratory Data Analysis:

In the dataset, many of the features used to predict are categorical opposed to numeric such as:

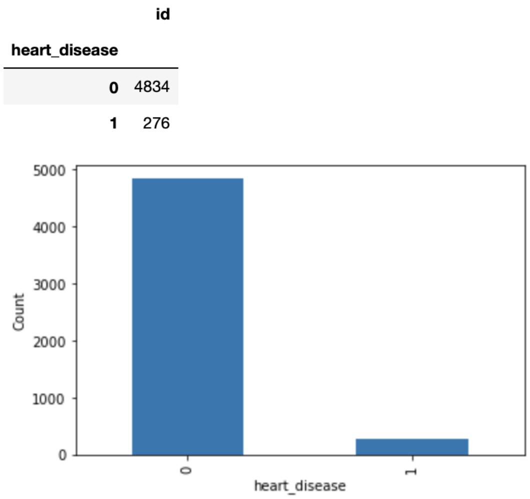
- Hypertension: Yes or No
- Heart Disease: Yes or No
- Ever_married: Yes or No
- Work_type: Private, Self Employed, Government ect.
- Residence_type: Urban or Rural
- Smoking_status: smokes, formerly_smoked, never smoked
- Gender: Male or Female
- Stroke: Yes or No

Brooke Hanson

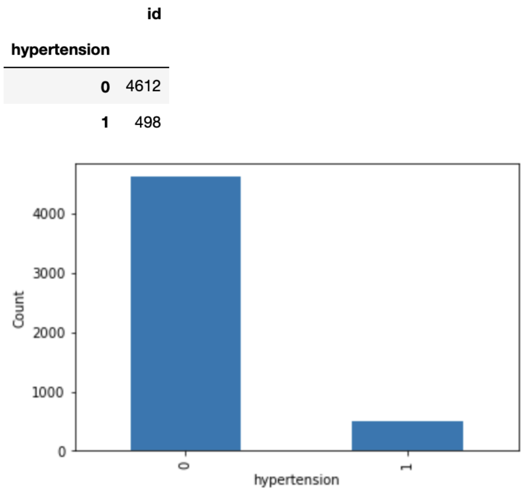
The numeric features included:

- Age
- BMI
- Average Glucose Level

One of the first things I wanted to visualize was the count of each categorical variable to identify if there are large discrepancies in the demographics. The largest discrepancies were found with the pre-existing health conditions: heart disease and hypertension as seen below:

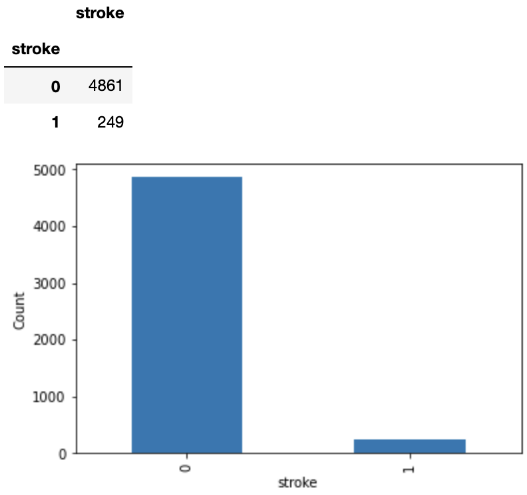


	id
heart_disease	
0	4834
1	276



	id
hypertension	
0	4612
1	498

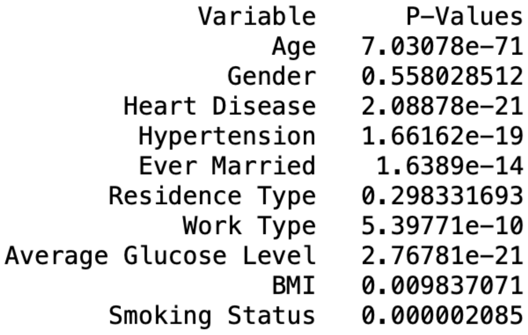
Another large discrepancy to note is that of stroke itself, in this data set the majority of individuals have not experienced a stroke. This is an important distinction to note as it will affect what performance metric to use. The count is seen below:



	stroke
stroke	
0	4861
1	249

With the majority of the variables being categorical, it was best to use a chi^2 test in order to test the independence of variables, and for the numeric variables I used a t-test to identify which of those variables are independent. The results of these test are listed below, if the p-value listed is below 0.05 then we can reject the null hypothesis that the variable is independent of stroke occurrence.

As can be seen from the table to the right, the only variables that are statistically independent from stroke occurrence are Gender and Residence Type. This is a good indication that we have variables that will be useful in predicting stroke occurrence in other adults.



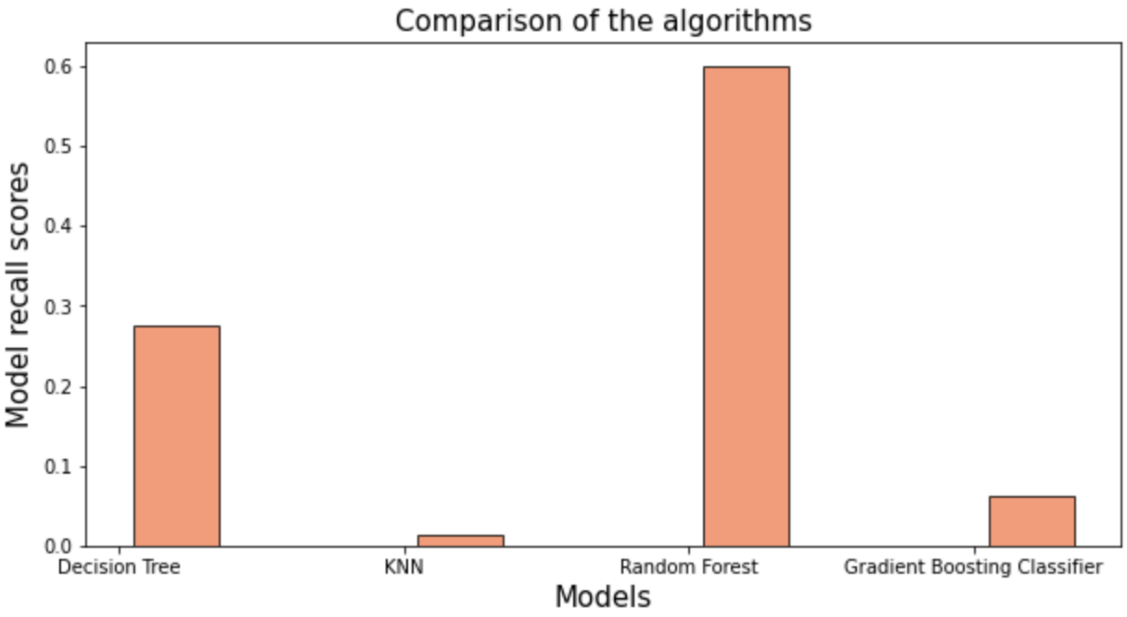
```
              Variable     P-Values
                   Age  7.03078e-71
                Gender  0.558028512
         Heart Disease  2.08878e-21
          Hypertension  1.66162e-19
          Ever Married   1.6389e-14
        Residence Type  0.298331693
             Work Type  5.39771e-10
 Average Glucose Level  2.76781e-21
                   BMI  0.009837071
        Smoking Status  0.000002085
```

In Depth Analysis and Modeling:

Once the relationships between the variables had been identified, it is time to move on to using supervised learning to create a predictive model for the data. Since the response variable in this case is stroke, and the outcome is either 1 for stroke or 0 for non stroke patients, it is best to use a classification model. It is also important to note that since there is a discrepancy in the count of 1s and 0s, and the outcome of interest is 1’s, the performance metric I chose to use is recall opposed to accuracy, since we want to make sure we are focused on best predicting who will have a stroke opposed to who won’t.

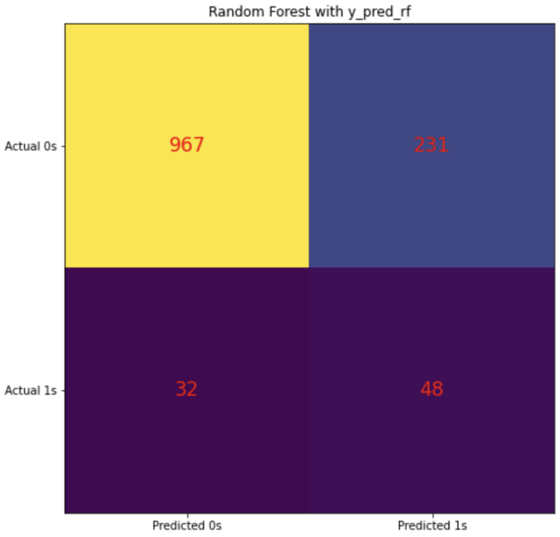
In total I tested and tuned 4 models to compare and contrast predictive power. The models I tested were Decision Tree Classifier, K Nearest Neighbor, Gradient Boosting Classifier, and Random Forest Classifier. For each of these models, to determine which hyperparameters to use, I did a grid search and used the recall performance metric as the scorer. This process indicated which hyperparameters were best for each model considering the scoring function.

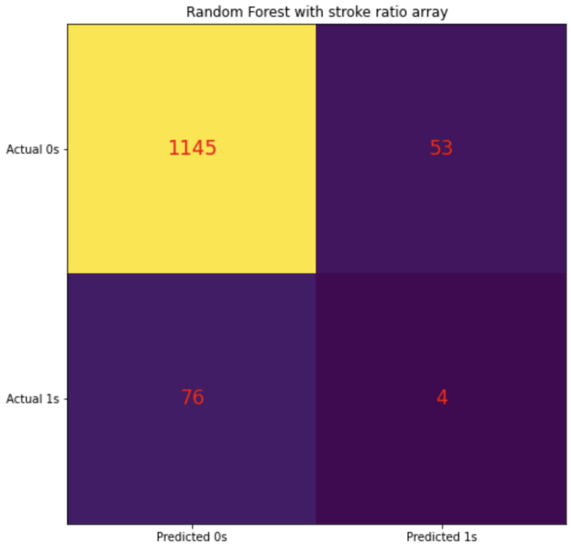
Once each model had been tuned for optimal performance, I compared the recall scores of each model to decide which one to use, the results are seen below:



This plot lets us know that the Random Forest model had the best recall score and that will be the final model used.

As I mentioned above there is a significant difference in the amount of individuals in the data that have experienced strokes compared to those that have not. Due to this difference I wanted to verify that the model was actually predicting well opposed to just following the discrepancies found in the data. I achieved this by comparing the confusion matrix of the tuned Random Forest model to a confusion matrix using the same model, but a binary array with the same ratio of stroke patients. Seen below:





From these confusion matrices, we can see that the predicted 1’s and actual 1’s is much higher in the left confusion matrix with our predictions generated from the Random

Forest model, compared to the basic array generated from the stroke ratio. This lets us know that the tuned Random Forest model is actually making accurate predictions for the variable of interest, opposed to just following the count discrepancies found in the data.

Creating a function for future predictions:

The main benefit of being able to identify factors for prediction and creating a good predictive model, is to be able to use that on new data. I wanted to create a function that would allow medical professionals or individuals to fill in their own information and get an idea of the probability of future strokes. This included standardizing the new line of data, processing the data, then using the model to predict the outcome and the probabilities associated with the resulting prediction. This interactive tool may be accessed in this link.

Takeaways and Improvements:

As said by George Box, “All models are wrong, but some are useful”, the same applies in this scenario. This data allowed me to create a predictive function but there is always room for improvements. In this case, 5110 is not a large enough sample size to be truly representative of the entire population so more data would most likely improve the model. There are also a variety of other demographics that could have been included, some that come to mind would be geographic location, family history of stroke, race, and income level. There is no sure fire way to predicted if someone will have a stroke but the information I have processed and provided may help mitigate some risk factors and spur health changes.