For this project, I attempted to build a classification model that would take in a patient's attributes and determine whether the patient was, or was not, likely to have a stroke in the future. A stroke is typically caused by either insufficient blood flow to the brain, or excessive bleeding within the brain. Because of this, the best indicator of stroke risk is high blood pressure, but in the dataset that I utilized, there was no blood pressure variable. This provided me with an opportunity to find other prevalent factors of stroke risk that may be less obvious. According to the CDC, more than 795,000 Americans have a stroke every year. There is tremendous value to be found by looking through data and identifying patients who could be at risk, so that we can find ways to take preemptive action.

1. Data

The dataset I used was found on Kaggle.com, and was shared by user fedesoriano. Here is a link to the Kaggle page:
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

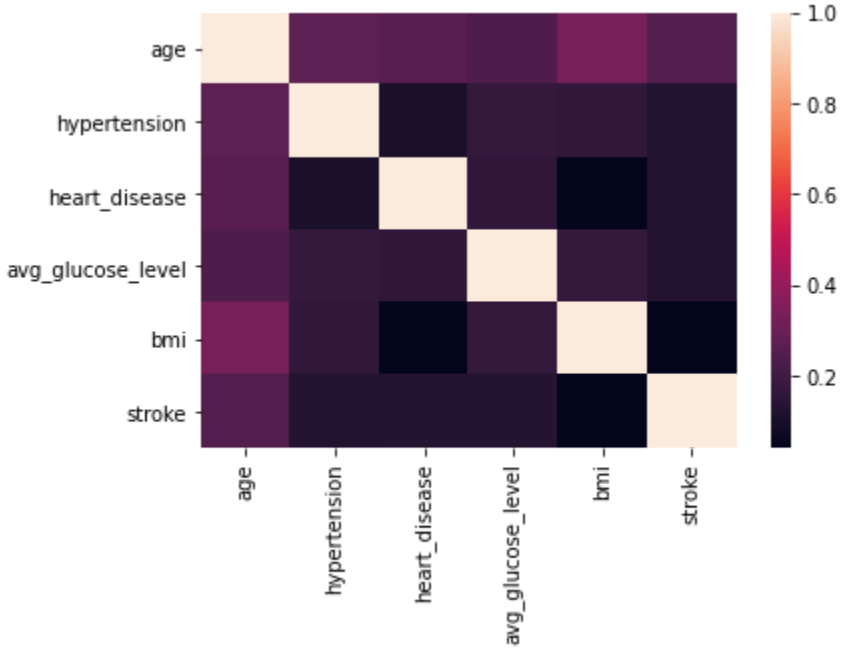
2. Data Cleaning

I started by using pandas' read_csv method to import the kaggle dataset as a pandas dataframe. I inspected the data using both the .info method and the .head method. The data had 12 columns (including an id column) and 5110 entries. There were 10 predictor variables that I could use to go along with my target variable, 'stroke', which was an indicator variable in which a value of 0 indicates that a patient did not have a stroke and a value of 1 indicates that a patient did have a stroke. Notably, the only

feature that contained null values was bmi. I decided that this alone did not significantly dilute the quality of an entry, so I did not drop observations with null bmi values. At the same time, there was another issue with the bmi column:
The maximum and minimum values appeared to be out of the range of plausibility. After doing some research, I discovered that the NHSN had defined BMIs below 12 or above 60 as outliers. I decided to drop patients with BMIs in that outlier range from my dataframe.

3. EDA

I begun my exploratory data analysis by generating a heatmap:



The heatmap indicates that age is by far the greatest factor that we have in predicting stroke likelihood. I further inspected the data by dividing it into two sets; “stroke” in which all of the stroke variables were equal to 0, and “nostroke” in which all of the stroke variables were equal to 1. I followed this by generating histograms that

explored the relationship between each feature variable and the stroke/nostroke datasets. I determined that age, average glucose level, hypertension, heart disease, marriage status and smoking status appeared to be relevant factors in predicting stroke likelihood.

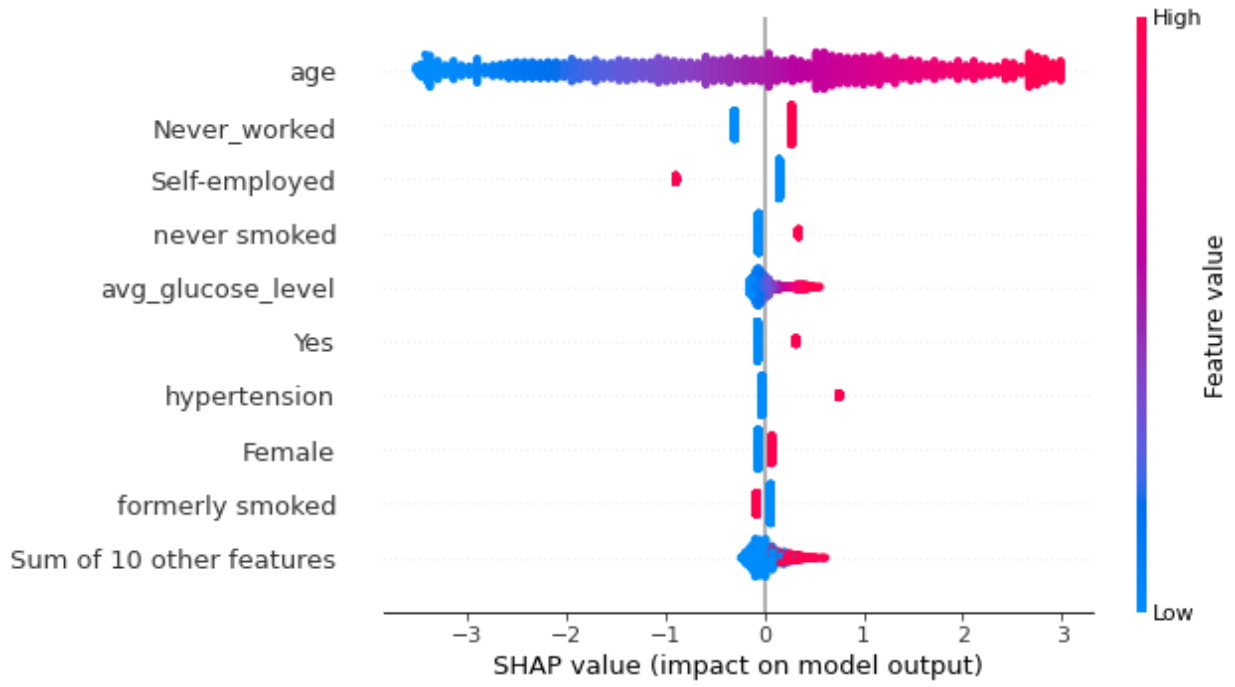
4. Data Preprocessing

The preprocessing phase mainly consisted of the creation of dummy variables. I ended up converting my 7 categorical variables to 16 dummy variables. Next, I divided the data into a training/testing set at a 70/30 ratio and scaled the features using a MinMax Scaler.

5. Data Modeling

I decided to try out two types of models: Logistic Regression and Random Forest. As stated previously, the features were scaled with MinMaxScaler. Additionally, I imputed the median bmi for observations that were missing a bmi value. To tune the models, I utilized GridSearchCV in order to figure out the appropriate weight in which to oversample the positive stroke observations. This was necessary because the stroke dataset is heavily imbalanced toward negative observations. The scoring metric I used for grid search was roc_auc. In the end, I decided that Logistic Regression was superior to Random Forest for this project. You can find the metrics for all tested models here:
https://github.com/ryanoh999/Stroke-Prediction/blob/main/Stroke%20Capstone%20Model%20Metrics.pdf

This is a beeswarm plot for the final Logistic Regression model:



It is evident that age was by far the most important factor in the model. Those who were self-employed were significantly less likely to have a stroke. Perhaps being self-employed lowers one's stress level (or people who are inclined to be self-employed or capable of being self-employed are less stressed than people who are not). We can also see that those with hypertension were more likely to have a stroke which is unsurprising.

6. Recommendations

- Doctors should use this model to assess their patients' risk and take preemptive action if risk is indicated

- Patients should be made aware of the factors that lead to stroke risk and act accordingly; lowering stroke risk is the one of the best ways to extend your mortality
- Doctors should give roughly equal focus to a patient's personal life as they do to tangible health metrics when assessing their stroke risk

7. Future Research

I think there are several other relevant features that could be added to the dataset, and potentially improve the model. I'm intrigued by how self-employed appears to be a relevant factor. It would also be interesting to see how effective taking the preemptive measures that the model suggests would be with regards to improving stroke prevention.