Hello and welcome to the presentation on Predicting Stroke Risk

Today we will be discussing the use of machine learning techniques to classify whether or not an individual is at risk of having a stroke based on a subset of personal traits.

-- 2

Each year, millions die as the direct result of a stroke. According to the World Health Organization stroke is the second leading cause of death worldwide, accounting for around 11% of total deaths. Many of the top causes of death are directly related to cardiovascular disease. These 15 million deaths from stroke make up 1 out of every 6 deaths related to cardiovascular disease.

In the United States alone, there are around 800 thousand strokes each year. This amounts to one stroke every 40 seconds, with a death every 3 ½ minutes, on average.

-- 3

Of the approximately 15 million annual strokes worldwide, roughly 1/3 result in death. Of those who survive a stroke, only around half fully recover, while others are left permanently disabled. Because of these significant numbers, stroke is one of the leading causes of long-term disability in the US.

Many risk factors for stroke are already known. Most of these, as you might expect, are related to other cardiovascular and health issues, such as high blood pressure or cholesterol, obesity and diabetes, and just old age. Of course, there are a lot of health risks that increase with age. There are also other, less obvious risk factors. For instance, being black doubles your risk of stroke relative to being white, and in the US your mortality risk from having a stroke increases if you live in the South. Of course, these are not likely causal relationships, but are interesting correlations nonetheless. One risk factor that I had not considered that actually makes sense is if you have had a stroke before, you are at increased risk for another stroke.

-- 4

The dataset I used for stroke prediction was found on Kaggle. There were a couple of similar datasets available, and I selected the one with more entries coming in at over 40,000 rows, compared to around 5,000 for the other dataset.

The dataset contains 10 features with information about the individual, mostly health-related, but also things like the type of work they do or whether they have ever been married. Half of the features were recorded into the dataset as categorical and the other half as numeric, though several of these were just binary yes/no type categories.

The target variable was whether or not the individual had experienced a stroke with 1 being a positive stroke case and 0 being negative. The target data was incredibly imbalanced however, with positive stroke cases only making up about 2% of the data.

-- 5

When I first looked at the distributions I noticed that of the numeric features, only BMI and glucose really appeared skewed, with glucose actually showing a somewhat bimodal distribution. With the target variable being so imbalanced, it was difficult to compare distributions between the stroke and non-stroke data, so I balanced by resampling the positive stroke data and compared distributions. There were some apparent differences between the stroke and non-stroke data, as demonstrated by the glucose histogram here.

Something I noticed was that a large portion of the data consisted of children, who are at lower risk of stroke anyway. Many of the differences in the feature variables between the stroke and non-stroke data appeared to be related to this ‘children’ group because, for instance, at a lower age they are at a lower risk for hypertension and heart disease, will not have been married or had a job, and are unlikely to have smoked.

One interesting difference was that those who reported being self-employed appeared to be at a much higher risk of stroke.

-- 6

In processing the data, only two of the features contained any Null values: smoking status and BMI. Because they made up such a large chunk of the smoking status group and it is a categorical feature, I converted them to their own category of ‘Unknown’. With BMI, a much smaller proportion was Null and it is a numerical value. I tried something I have not used before and attempted using logistic regression to predict the missing values based on the other data. While this was an interesting learning experience for me, I ultimately ended up using median to impute the Null values because it was a simpler process and I appeared to get slightly better results.

Encoding was pretty straightforward. The binary values were all converted to 0s and 1s and multiclass features were one-hot encoded. The one exception is gender, where I chose to eliminate the 11 instances of ‘other’ and then just convert the remaining male and female classes to binary.

Following the train-test split, I transformed the numerical age, BMI and glucose features using Box-Cox transformation, then scaled all of the feature to the same -1 to 1 range.

Balancing the data in regard to the target variable proved a little trickier. I tried oversampling the positive stroke data to match the non-stroke set but I feel this overrepresented the few positive stroke cases. To help with this issue, I tried oversampling using SMOTE and then tried a combination of oversampling using SMOTE to 10% of the non-stroke size, then sampling the non-stroke data down to double the stroke data size. Interestingly, even with much less data, this proved to be as effective or better than the previous two efforts. Ultimately however, I decided to leave the data imbalanced and set class\_weight to ‘balanced’ for any models that offered the option.

-- 7

I compared 15 different classification models and as expected, those that did not offer class weighting did not prove effective. Initially I was using accuracy for evaluating models, but with such a large imbalance, this was not the appropriate metric as some models were predicting everything as non-stroke and getting 98% accuracy. Recall was useful for increasing true positive stroke results but could lead to the opposite problem where 100% recall could be reached by predicting everything as positive for stroke. I found a nice balance using Matthews Correlation Coefficient and AUC.

I used GridSearchCV for tuning parameters on my top models, scoring with a combination of the aforementioned metrics with some models showing more improvement than others with tuned parameters. I also attempted to utilize the Voting Classifier with the now-tuned top models, weighting them based on one of the metrics, but I didn’t see a whole lot of further improvement.

-- 8

With the huge imbalance in the target data, it was difficult to establish a really good model. Without having a clear metric like accuracy I needed to find a balance in the model predictions that would both correctly predict the majority of positive stroke cases, but without getting an unnecessary number of false positives.

Including more positive stroke cases as well as additional predictive features I think could both prove to be useful additions to the model. I also wonder how many of the people in the negative stroke data are indeed at high risk for stroke but just haven’t had a stroke yet and if this is adding some bias to the model.

-- 9

I think this kind of model could be incredibly helpful in providing patients some control over their own personal healthcare, giving them a deeper level of understanding that they can use when discussing their health with their physician.

I could see an app that would ask you some health-related questions that could be used for predicting risk of a variety of health outcomes. You could record and update your weight and blood pressure or link fitness apps to record activity levels over time and it could provide suggestions on how to lower your highest risks as well as feedback on how your health risks are impacted based on changes you make.

All in all I think there is usefulness and potential in a predictive model, not just for stroke, but for many common health issues.