Scott Breitbach
DSC680, Weeks 5-7
05-July-2022

Project 2: Milestone 1

Topic:

Predict whether an individual will experience a stroke based on several health-related factors.

Business Problem:

Stroke is one of the leading causes of death worldwide. In the US alone, almost 800,000 people each year suffer a stroke and it is the leading cause of long-term disability in the US.

Time is a critical factor for improving outcomes in the event of a stroke. Being able to predict whether an individual is likely to have a stroke can help in promoting awareness of the symptoms to support prompt action as well as preventative measures that can be taken to improve risk factors associated with stroke.

Datasets:

I was able to find a couple of datasets on Kaggle to support this project, both with identical columns. However, while one appeared to be already cleaned up, containing no NAs or missing values, at around 5,000 records, it was significantly smaller than the other dataset, which contained over 40,000 entries. The target variable, 'stroke' is a binary outcome and is very imbalanced in both sets, with the 'True' outcome consisting of around 5% of the smaller dataset and only around 2% of the larger dataset. For now, I am planning to utilize the larger dataset for my modeling.

In addition to the target 'stroke' column and the unique 'id' column, the data consists of 10 other features. Three of these are binary yes/no variables, 'hypertension', 'heart_disease', and 'ever_married'. Two are binary categories, 'gender' and 'residence_type' (Rural or Urban). Two are categorical, 'work_type' (5 categories) and 'smoking_status' (3-4 categories), and three are numeric: 'age', 'avg_glucose_level', and 'bmi'.

Methods:

I plan to start with the larger dataset and clean up any outliers from the numeric features and remove or impute any null values. The minimum value in the 'age' variable is 0.08, which seems low to include in this study, so I may add a cutoff on the lower end of the age range. The 'smoking_status' variable appears to have a large number of missing values (~30% of the data) so I will likely encode that as it's own "Unknown" category. Following this, I will encode the categorical data and transform and/or scale the numeric data.

I will perform some feature selection and try a variety of algorithms to see which works best with my dataset according to several evaluation metrics.

Ethical Considerations:

When dealing with any medical data, it is important that identifying patient information is scrubbed out. It appears that this is already taken care of with this dataset as it uses an ID code for each record and the personal identifying information is broad enough (gender, age, urban/rural) that it is extremely unlikely to be traceable to any individual.

While race was not included in this dataset, it is also an important factor to consider as race is a known risk factor for stroke, with black people at nearly twice the risk of stroke as well as having the highest death rates from stroke. When race and race-related factors are included in your data, it is important to consider carefully any biases that may be introduced into your dataset.

Challenges / Issues:

I don't know if it will even be a relevant variable, but the 'work_type' categories are somewhat vague. If there isn't a clear effect on the target variable, I may remove this feature. As mentioned above in the Methods section, dealing with outliers and Null values is often challenging, but my biggest concern is whether there is enough information in the provided features to reasonably predict the target outcome.

One question I have about the dataset is the inherent limitation of collecting this type of data. The people included are from a variety of age groups and the outcome is whether or not they have ever had a stroke, but many of these people may someday have a stroke and have just not had one yet. If they have a stroke later, but their data is used to train a model with a no-stroke outcome, then the data is incorrect.

References:

Amal, Lirilkumar. "Heart Stroke." Kaggle, 26 Oct. 2020, https://www.kaggle.com/datasets/lirilkumaramal/heart-stroke.

CDC. "Know Your Risk for Stroke." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 12 Apr. 2022, https://www.cdc.gov/stroke/risk_factors.htm.

CDC. "Stroke Facts." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 5 Apr. 2022, https://www.cdc.gov/stroke/facts.htm.

Fedesoriano. "Stroke Prediction Dataset." Kaggle, 26 Jan. 2021, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

Reinberg, Steven. "1 In 4 People Over 25 Will Be Hit by Stroke." WebMD, WebMD, 20 Dec. 2018, https://www.webmd.com/stroke/news/20181220/1-in-4-people-over-25-will-be-hit-by-stroke.

Stroke Awareness Foundation. "Stroke Facts & Statistics." Stroke Awareness Foundation, 23 Jan. 2021, https://www.strokeinfo.org/stroke-facts-statistics/.