Cerebral Stroke ROC Project

I chose to do my ROC Project on a cerebral stroke dataset with data sourced from a publication titled *A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets* by Tianyu Liu et al. 2019. The dataset originally contained 11 independent variables, with stroke as the dependent variable. A cerebral stroke occurs when blood supplied to part of the brain is interrupted or reduced thus preventing the brain tissue from getting oxygen and nutrients, brain cells within minutes begin to die. Every year, over 795,000 Americans have a stroke, this project will investigate what variables are good predictors of having a stroke according to this particular dataset. Before I began my initial exploration, I dropped the independent variable “ID” from the dataset as it merely served as a distinguishing factor between each observation and not essential for my calculations. I selected to keep Gender in my dataset due to a study published in the *Journal of Cerebral Blood Flow and Metabolism* by Claire L Gibson 2013, which stated that “Evidence exists to suggest that gender influences many aspects of ischemic stroke including stroke risk/incidence, diagnosis, symptoms, treatment and outcomes” (Tiwari,S. 2021). I also kept Age in my dataset due to a publication by the *CDC* which stated that “The chance of having a stroke about doubles every 10 years after age 55. Although stroke is common among older adults, many people younger than 65 years also have strokes. In fact, about one in seven strokes occur in adolescents and young adults ages 15 to 49” (CDC, 2022). I kept Hypertension in my dataset due to a study published in the *European Cardiology Review* by Mauricio Wajngarten and Gisele Sampaio Silva 2019, which stated that “Hypertension is the most prevalent risk factor for stroke. Stroke causes and haemodynamic consequences are heterogeneous which makes the management of blood pressure in stroke patients complex requiring an accurate diagnosis and precise definition of therapeutic goals” (Wajngarten & Silva, 2019). I kept Heart Disease in my dataset due to a study published in the *HHS Public Access* by Zhili Chen et al. 2017 which stated that “Approximately 20% of ischemic strokes are caused by cardiac disease, the major risk factor being atrial fibrillation” (Chen et al., 2017). I also kept Marriage in my dataset due to a study published in the *Journal of Neurology* by Qi Liu et al. 2018 which concluded that “Marital status was associated with all adverse stroke outcomes in patients with acute ischemic stroke” (Liu et al., 2018). I chose to keep the Work-type variable in my dataset due to a publication by the *American Academy of Neurology* which concluded “The analysis found that people with high stress jobs had a 22 percent higher risk of stroke than those with low stress jobs. Women with high stress jobs had a 33 percent higher risk of stroke than women with low stress jobs. People with high stress jobs were 58 percent more likely to have an ischemic stroke than those with low stress jobs” (Dingli , 2015). I also kept Residence-type in my dataset due to a publication in *Circulation: Cardiovascular Quality and Outcomes* in 2019 which stated that “Stroke incidence and mortality are higher in rural than in urban areas in North America” (Kapral et al., 2019). I kept Average Glucose Levels in my dataset due to a publication by the *American Stroke Association* which states that “A fasting blood glucose (sugar) level of 126 milligrams per deciliter (mg/dL) or higher is dangerous. People who have diabetes are 2 times as likely to have a stroke compared to people who do not have diabetes” (ASA, 2023). I also kept BMI in my dataset due to a publication by the *Bariatric Department at Lafayette General Medical Center* which stated that “The link between excess weight and an increased likelihood of stroke is unmistakable. Repeated studies estimate that each unit increase in body mass

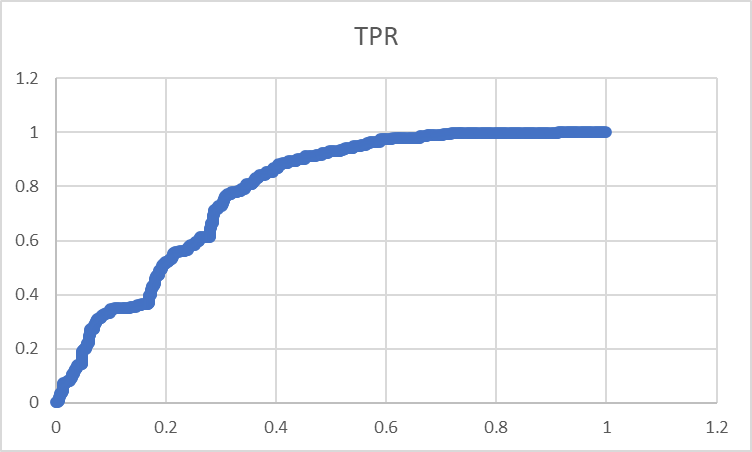
index (BMI) increases the risk of stroke by 5 percent” (Ochsner 2019). For my final independent variable, I decided to keep Smoking Status in my dataset due to a publication in the journal *Medicine* which stated that “Based on the dose-response meta-analysis, the risk of stroke increased by 12% for each increment of 5 cigarettes per day” (Biqi et al., 2019).

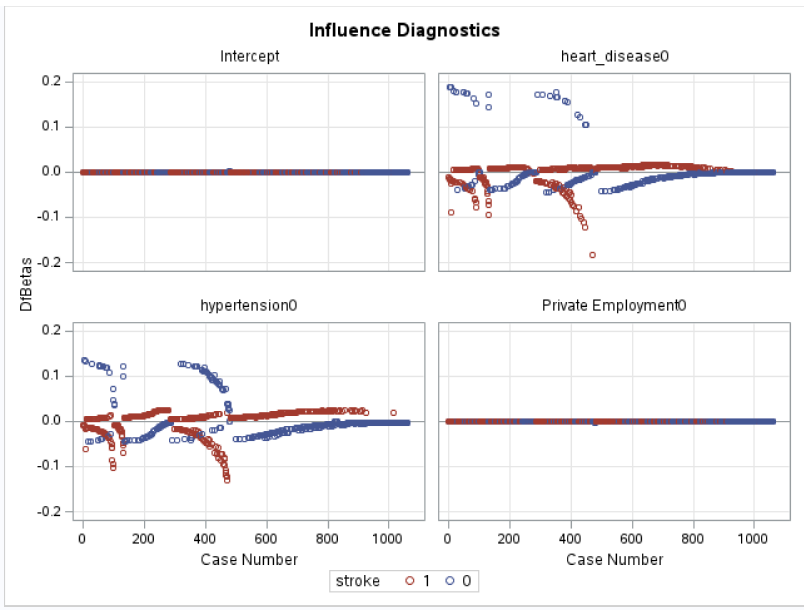
I set-up my data by filtering through the original dataset until I had 548 condition positive observations and 512 condition negative observations for a total of 1060 observations and a 51.7% and 48.3% split of the data, neither of which percentages exceed 70% indicating a balanced dataset. I normalized my continuous independent variables: Residence, Gender, and Age in order to better run a logit calculation for each record. Next, I calculated e^L for reach record, followed by calculating P(x) for each record to determine the probability of stroke occurring. Then I calculated P(x)^Y \* (1-P(x)^(1-Y) to determine the conditional probability for each record. After that I took the natural logarithm of each record using the LN() function and summed them up for the objective function. Finally, I ran an Evolutionary solving method on my data changing my independent variables as well as the intercept value with the goal of maxing out my Pct OK value with constraints on each value that was being changed to not exceed or equate to 5 or equate to or drop below -5, with a tolerance of 0.2. It is important that the Pct OK value gets as close as possible to a value of 1. Another calculated field is “OK?” which is: IF(ABS(P(x)-Actual)<0.2,1,0) in order to check if the difference between the p(x) column and the actual outcome column is less than the tolerance (0.2).

My base Evolutionary Model, with all of my initial variables I had chosen to keep, reached a Pct OK value of 0.639622642. For an attempt to raise the Pct OK value I chose to drop Gender from my calculations which brought the value up to 0.651886792 indicating that Gender was not a good predicting variable for my model and was hindering it, these results are shown on my second sheet. Next, I chose to drop the Married variable from my calculations which further increased my Pct OK value to 0.666037736 indicating that Marriage was also not a good predicting variable for my model, these results are shown on my third sheet. In my next four sheets of data, I consecutively dropped BMI, Residence, Glucose, and Smoking variables from my model all of which have no impact on the Pct OK value which remains at 0.666037736 despite each variable being dropped and the Evolutionary solving method being performed again on the remaining variables. This indicates that none of those variables are significant to my particular dataset when predicting Cerebral Stroke so I removed them from my model as indicated by the results displayed in each of these four sheets. The next thing I tried to do was improve my model by removing the Work\_type variable from my calculations which actually decreased my Pct OK value to 0.637735849 as shown by the results in my ninth sheet, which indicated that it was an important predictor to my dataset and should not be removed, so I reinserted it into my model. I then dropped Age to try and improve my model but once again ended up decreasing my Pct OK value to 0.658490566 as shown by the results in my tenth sheet, which indicated that it was also an important predictor to my dataset and should not be removed, so it was also reinserted. Next, I removed Hypertension from my model and this dropped my Pct OK value to 0.633018868 as shown by the results in my eleventh sheet, which indicated that it was an important predictor to my dataset and should not be removed, so it was reinserted. Finally, I chose to remove Heart Disease from my model and dropped my Pct OK value to 0.623584906 as shown by the results in my twelfth sheet, which indicated that it as well as an important predicator to my dataset and should not be removed, so it was also reinserted back

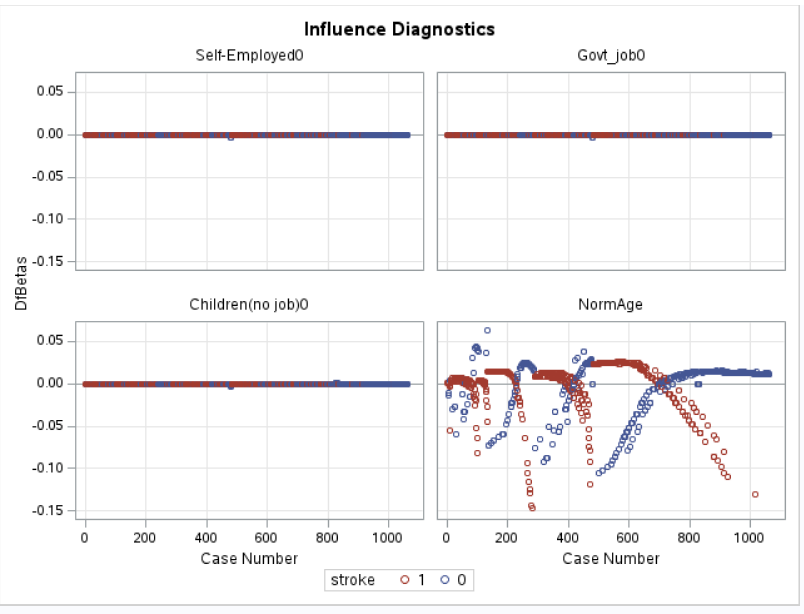
into my model. My final model now consisted of the independent variables: Hypertension, Heart\_disease, Employment(Private Employment, Self-Employed, Govt\_job, Children), and NormAge(Age), with stroke being the dependent variable. If you look at the values under each of these variables; Hypertension and Heart Disease play by far the most significant predicting factors to Cerebral Stroke emphasized by their holding of the highest values, which are just under 5. The next most important variable is a sub-variable of Employment which is Self-Employed; its value is also quite high and thus indicates it is an important predictor, though the other sub-variables for Employment prove the opposite, they are all negative values and are thus poor predictors for Cerebral Stroke with this dataset; but removing Employment altogether from the model lowered the Pct OK score due to the one sub-variable “Self-Employed” outweighing the overall hindering effects of the rest, thus they were all kept in the model to maximize the Pct OK value. Due to all of these aforementioned calculations the sheet I decided to run the rest of my calculations on as my optimal model is the eighth sheet titled “Model Drop Smoking(final model)” which encompasses the best Pct OK value and the most appropriate variables. A Log Likelihood was also calculated (which holds a value of -1197.331196 in my final model) with the initial focus to attempt to maximize this value as a metric to assess the optimization of my data, but through my calculations and deciding to switch from GRG Nonlinear to the more appropriate Evolutionary solving method for my data, I chose to focus on maximizing Pct OK rather than this Log Likelihood value.

Lastly, using my optimal model on my eighth sheet, I calculated an AUC value which accounted for the area under my curve consisting of True Positive Rate over False Positive Rate. In order to do this, I first had to sort my data in descending order by Logit and add a column named Rank ordered in ascending order by the logit rank. The sum of the positive outcomes was equated into the True Positives column, and the False Positives column was equated to the Rank – True Positives. Each column was added for TPR and FPR by dividing each TP and FP entry by the last entry in each TP/FP column. The goal was to maximize this value to indicate a better performing model. The True Positive Rate measures the probability that an actual positive observation will be classified as positive, while the False Positive Rate measures the probability that a negative observation will be classified as positive. My value of AUC ended up being 0.78221587 which was as optimal as I could achieve with the model I finalized with my particular dataset. The graph displayed shows TPR over FPR for the AUC curve. The best shape of the curve is one that is closer to the top left corner of the plot thus indicating a better performing model, in contrast a poor model would have a curve which is closer to the diagonal line. An AUC value of as close to 1 as possible is what is desired and my curve has a decent rise toward to the top left initially before curving out which indicates that my model is performing well.





According to the above heart\_disease and hypertension plots, it can be ascertained that the positive and negative DFBeta values (as well as outliers) indicate that these individuals possess strong influences on the relationship between heart disease and the risk of stroke, as well as hypertension and risk of stroke. These plots also indicate that Private Employment has little effect on the risk of stroke due to their being no outliers and no significant level of values out of the zero range.



According to the above NormAge plot, the positive and negative DFBeta values (as well as outliers) indicate that these individuals have a strong influence on the relationship between an individual’s age and their risk of stroke. The positive DfBetas suggest a positive influence in that older individuals are more likely to have a stroke, and the negative DfBetas suggest a negative influence in that younger individuals are less likely to have a stroke according to the particular set of data. These plots also indicate that Self-Employed, Govt\_job, and Children(no job) employment categories have little influence on the risk of stroke due to their being no outliers and no significant level of values out of the zero range.

In conclusion, the variables that were left in the final model through the performed binary logistic regression analysis have proven to be the best predictors of stroke within and according to my particular dataset. These variables that translate into whether or not an individual has hypertension or heart disease, what their employment is, and how old they are have the strongest association with the probability of having a stroke. Further study and documentation of these predictor variables can potentially provide insights into the magnitude and direction of their effects on the risk of stroke.

References

American Stroke Association. (n.d.). *Diabetes and stroke prevention*. www.stroke.org. Retrieved February 6, 2023, from https://www.stroke.org/en/about-stroke/stroke-risk-factors/diabetes-and-stroke-prevention#:~:text=A%20fasting%20blood%20glucose%20(sugar,who%20do%20not%20have%20diabetes.

Bariatric Department at Lafayette General Medical Center. (2019, March 26). *How obesity affects stroke risk*. Ochsner Lafayette General. Retrieved February 6, 2023, from https://ochsnerlg.org/about-us/news/how-obesity-affects-stroke-risk#:~:text=The%20link%20between%20excess%20weight,of%20stroke%20by%205%20percent.

Biqi, P., Xiao, J., Liu, J., Shaohong, Q., Qiuping, Z., & Mingwo, P. (2019). *The relationship between smoking and stroke: A meta-analysis : Medicine*. LWW.com. Retrieved February 6, 2023, from https://journals.lww.com/md-journal/fulltext/2019/03220/the\_relationship\_between\_smoking\_and\_stroke\_\_a.22.aspx

Centers for Disease Control and Prevention. (2022, April 12). *Know your risk for stroke*. Centers for Disease Control and Prevention. Retrieved February 6, 2023, from https://www.cdc.gov/stroke/risk\_factors.htm#:~:text=The%20chance%20of%20having%20a,65%20years%20also%20have%20strokes.&text=In%20fact%2C%20about%20one%20in,adults%20ages%2015%20to%2049.

Chen, Z., Venkat, P., Seyfried, D., Chopp, M., Yan, T., & Chen, J. (2017, August 4). *Brain-heart interaction: Cardiac complications after stroke*. Circulation research. Retrieved February 6, 2023, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553569/#:~:text=Approximately%2020%25%20of%20ischemic%20strokes,factor%20being%20atrial%20fibrillation9>.

Dingli , X. (2015, October 14). *CAN WORK STRESS BE LINKED TO STROKE?* AAN Publications. Retrieved February 6, 2023, from https://www.aan.com/PressRoom/Home/PressRelease/1412#:~:text=The%20analysis%20found%20that%20people,women%20with%20low%20stress%20jobs.

Gibson, C. L. (2013, September). *Cerebral ischemic stroke: Is gender important?* Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism. Retrieved February 6, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3764377/

Kapral, M. K., Austin, P. C., Jeyakumar, G., Hall, R., Khan, A. M., Jin, A. Y., Martin, C., Manuel, D., Silver, F. L., Swartz, R. H., & Tu, J. V. (2019, February 14). *Rural-urban differences in stroke risk factors, incidence, and ...* The CANHEART Stroke Study. Retrieved February 6, 2023, from https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.118.004973

Liu, Q., Wang, X., Wang, Y., Wang, C., Zhao, X., Liu, L., Li, Z., Meng, X., Guo, L., & Wang, Y. (2018, April). *Association between marriage and outcomes in patients with acute ischemic stroke*. Journal of Neurology. Retrieved February 6, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5878185/

Liu, Tianyu; Fan, Wenhui; Wu, Cheng (2019), “Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets”, Mendeley Data, V1, doi: 10.17632/x8ygrw87jw.1

Tiwari, S. (2021, August 22). *Cerebral stroke prediction-imbalanced dataset*. Kaggle. Retrieved February 6, 2023, from https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset

Wajngarten, M., & Silva, G. S. (2019, July 11). *Hypertension and stroke: Update on treatment*. European cardiology. Retrieved February 6, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6659031/