Analyzing Predictors of Strokes

Erez S. Sarousi

Bellevue University

DSC 680: Applied Data Science

Professor Williams

May 10, 2022

Topic

This topic will focus on strokes and potential causal connections between strokes, and other health and lifestyle attributes. This project aims to uncover relationships that could prevent strokes for others.

Business Problem

A stroke is medically defined as an incident where a medical emergency where blood supply to the brain is blocked in a way that suffocates the brain and starves it from essential nutrients (Brown, 2022). It can be noticed by, among others, difficulty speaking, facial drooping on one side, an inability to raise arms above their head, and trouble walking. Strokes are ranked as the fifth most common cause of death in the United States (CDC, 2022). It killed over 160,000 people in 2020, and both afflicted someone in the United States every 40 seconds and killed them every 3.5 minutes (CDC, 2022).

Taking steps to reduce the amount of strokes the United States faces can save the country up to $53 billion a year and almost 200,000 lives a year. Doing so would most likely make the country healthier as a whole and could also potentially lower the risk of complications from both the stroke and procedures taken to remedy it (Cleveland Clinic, 2022).

Data Explanation

The dataset for this project came from Kaggle (Link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset). This is a

highly-rated dataset with a usability score of 10/10. This project utilized predictive analytics, involving logistic regression, and a random forest classifier.

This dataset came with a number of features about each test subject. These include: ID (Numerical identifier), Gender Male, Gender Female, Age, BMI, Hypertension (If the test subject had been diagnosed with this condition), Heart Disease (if the test subject had been diagnosed with this disease), Average Glucose Level, Ever Married (if the subject had been married before), Work Type Children (Working with children), Work Type Government (Working for the government), Work Type Self-Employed (Working for themselves), Work Type Never Worked, Work Type Private (Working for a private company), Residence Type Urban (living in an urban environment), Residence Type Rural (living in a rural environment), Smoker Status Never Smokes (Never have smoked), Smoker Status Former Smoker (Used to smoke, but no longer), Smoker Status Smokes (Currently Smoked), Stroke (If the test subject had a stroke - this is the target variable.)

Methods

The data was prepared by refactoring binary strings into categorical integers. The data frame was also subsetted to remove unnecessary columns and otherwise unhelpful data (such as responses like, “Unknown”). All now-numeric columns were reclassified as numeric so that it could be appropriately processed. After data cleaning, descriptive statistics were visualized.

A correlation matrix was implemented to see how all columns are correlated with one another. Due to the binary nature of the target variable, logistic regression was

implemented. To check the best accuracy rates for the data, a multitude of different classification models were tested such as the Decision Tree, Support Vector Machines, K-Nearest Neighbors, and Naive Bayes.

Analysis

Correlation analysis found that the significant correlations with the target variable are: Age (coef ~ 0.242, $p < 0.001$), Hypertension (coef ~ 0.144, $p < 0.001$), Heart Disease (coef ~ 0.139, $p < 0.001$), Ever Married (coef ~ 0.072, $p < 0.001$), Average Glucose Level (coef ~ 0.141, $p < 0.001$), Work Type Children (coef ~ -0.034, p ~ 0.050), Work Type Self-Employed (coef ~ 0.050, p ~ 0.003), Smoker Status Never Smoked (coef ~ -0.035, p ~ 0.041), and Smoker Status Formerly Smoked (coef ~ 0.040, p ~ 0.020),

Non-significant relations are as follows: BMI (coef ~ 0.012, p ~ 0.497), Work Type Government Job (coef ~ -0.015, p ~ 0.390), Work Type Never Worked (coef ~ -0.015, p ~ 0.377), Work Type Private (coef ~ -0.018, p ~ 0.290), Residence Type Rural (coef ~ -0.006, p ~ 0.726), Residence Type Urban (coef ~ 0.006, p ~ 0.726), Smoking Status Smokes (coef ~ 0.001, p ~ 0.960), Gender Female (coef ~ -0.012, p 0.468), and Gender Male (coef ~ 0.012, p ~ 0.468).

Due to the unexpected link between being married and stroke, mediation analysis was conducted with age as a mediating factor. Because logistic regression between all three variables all had a p value of less than .05, age was successfully concluded to be a mediating factor. The same is true for stroke and being

self-employed. Interaction analysis concluded that hypertension is indeed a mediating factor between being self-employed and the target variable of having a stroke.

Five different classification models were tested with a train test split of 70%-30%. These models, along with their model accuracies are: Logistic Regression (94.742%), Decision Tree (46.206%), Random Forest (94.811%), Support Vector Machine (SVM) (95.428%), K-Nearest Neighbors (KNN) (80%), and Naive Bayes (94.785%).

Conclusion

The statistical tests conclude that there is a close relationship with having a stroke and the following factors:

- Older Age
- Hypertension (High blood pressure)
- Cardiomyopathy (Heart disease)
- Higher average glucose level
- Being a former smoker

Furthermore, the most accurate classification model to test this predictive analytic project is the Support Vector Machine model.

Assumptions

The assumptions on this dataset are extremely minor. Some of the minor assumptions come from some of the variables such as “Hypertension”. These variables do not specify the stages of hypertension, if the person ever had been diagnosed as hypertension, but does not suffer from it any longer or what their average blood

pressure is like. Therefore, some assumptions had to be made in order to reach a reasonable conclusion.

Another major assumption is coming to conclusions based on the data stemming from the work type. The results of this project conclude that being self-employed is a predictor of stroke, but it seems more likely that self-employed people face stress levels that could be a predictor of stroke (Madell, 2020).

Limitations

As suggested in the assumptions section, the limitations come from a lot of the dummy variables. For instance, hypertension can be interpreted as a measure of stress since it plays a large role in the condition (Harvard Health, 2020). However, the information stops there; there is not any scale that could provide more insight such as average blood pressure.

The same is true for heart disease; the data shows whether the patient had been diagnosed with the disease, but not a range of data showing any specific characteristic used in the diagnosis of heart disease such as the electrical output from the heart during an electrocardiogram (EKG).

Challenges

There are some challenges in regards to this project. Firstly, this project will be done with R, a coding language that offers a less extensive library and support network than that of Python. Furthermore, there were some complications when trying to run classification models with all variables since some variables were perfectly correlated

with each other (such as different genders or residence types). Some of these variables had to be dropped so that the predictive analytic models could properly run.

Future Uses / Additional Applications

Many of the variables that are deemed to be predictors of a stroke (hypertension, average glucose level, etc.) are similar to that of a heart attack (Roland, 2018). Furthermore, it seems reasonable that many of these variables also seem to be indicative of overall health. Therefore, it seems likely that many of the variables tested here could be used as potential predictors for other illnesses.

Recommendations

Those who are concerned about their risk of stroke should first of all consult with a trusted healthcare provider. With that said, recommendation for people to prevent the chances of having a stroke are to:

- Take steps to lower their blood pressure.
- Do their best to treat their heart disease.
- Do their best to control their blood sugar level.
- Not to start smoking, if they hadn't already.
- Quit smoking, if they have already started.

Implementation Plan

In order to implement these changes to prevent a stroke, one should consult their primary care physician and have a conversation with them on their risk for having a

stroke. If they are found to be at high risk, they should follow all medical advice. This could include changes to their diet, activity level, and medications. This could also include testing, including bloodwork.

Ethical Assessment

Ethical considerations for this project are minimal. While each row does indicate a person and different aspects about them such as their BMI, glucose level, and many others, their names or other identifying features have been redacted; it has been replaced with an ID.

Ten Questions From The Audience

1. If work type children has a p value of .05, why is it marked as statistically significant?
    a. The p value of this variable has a p value of slightly less than .05 and only reaches that number due to a rounding error.
2. Why is ever_married not considered a potential predictor of stroke considering its high correlation coefficient?
    a. Ever_married was not considered a potential predictor of stroke because, while the correlation suggests that it may be, age is a mediating factor. In other words, Ever_married only tests that way because age causes both marriage and a stroke. As someone gets older, the more likely they are to get married. As someone gets older, the more likely they are to be afflicted by a stroke.
3. Why was mediation analysis not implemented on all variables?
    a. Mediation analysis was only implemented on variables deemed to be an anomaly. Mediation analysis was warranted because there wasn't a clear and direct connection between being married and suffering from a stroke.
4. With a p value of less than .05, why hasn't the never_smoker variable been considered a potential predictor of stroke?
    a. Never having smoked wasn't considered to be a potential predictor of stroke because the correlation coefficient was negative. In other

words, the more someone is to have never smoked, the less likely they actually are to suffer a stroke.

5. Are there any hypotheses why being a former smoker could be highly correlated to having a stroke?
    a. There are some hypotheses as to why being a former smoker could be highly correlated to suffering from a stroke. The strongest hypothesis is that by the time a smoker quits smoking, the damage has already been done and the body is predisposed for the stroke occurring.
6. How does an SVM differ from that of other classification models?
    a. The Support Vector Machine differs from that of other classification models by being more efficient in high-dimensional spaces that does better compared to other models as the number of features increase (Ghandi, 2018). Because the approach to this predictive model resulted in extra variables, the model was higher than that of other models such as logistic regression or random forest.
7. Why were classification models considered as opposed to other predictive analytic models?
    a. The project here actually encompasses both classification and other models, such as regression models. However, it's important to recognize that the target variable here is binary in nature - either there's a stroke or there isn't. In that respect, the one regression model used, logistic regression, outputs in a binary fashion, so it

matches the project. With that being said, whether something is a stroke or not is a classification method, therefore a classification model is appropriate.

8. How large is the concern that by omitting some variables in the regression models would affect the accuracy of the model?
    a. The concern is negligible at best. The only variable that was removed altogether was ID, and that's just because it doesn't say anything about the patient, other than who the patient is - but the whole point of an ID is to protect patient privacy, so it does no good there. However, other variables weren't deleted, but were rather created into dummy variables. This was done so that further analysis could be done. This method was done in order to optimize accuracy, not reduce it.
9. How large is the concern that by removing rows that had undesirable aspects (such as BMI equalling "unknown"), the results from the data would be less meaningful?
    a. Much like the previous question, the concern here is negligible. Before data cleaning, there were 5110 rows. After data cleaning, there were 3425 rows. Considering the fact that a sample size of 40 is considered to be sufficient for most quantitative studies, there is not much concern that much accuracy has been lost.
10. Why was this project completed on R as opposed to Python?

a. There are a lot of benefits of using R over Python. While both programming languages can complete the objectives of this project and while Python has a lot to offer with its benefits in machine learning, R is a more specialized programming language in that it is more geared toward statistical analyses (Watson, Yee, 2017).

References

Brown, R. (2022, January 20). *Stroke - Symptoms and causes*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

Budiu, R., & Moran, K. (2021, July 25). How Many Participants for Quantitative Usability Studies: A Summary of Sample-Size Recommendations. Nielsen Norman Group. https://www.nngroup.com/articles/summary-quant-sample-sizes/

CDC. (2022, January 13). *FastStats*. Leading Causes of Death. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

CDC. (2022b, April 5). *Stroke Facts | cdc.gov*. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/facts.htm

Cleveland Clinic. (2022, April 12). *After Your Stroke: How to Handle 5 Common Complications*. https://health.clevelandclinic.org/after-your-stroke-how-to-handle-5-common-complications/

Gandhi, R. (2018, July 5). Support Vector Machine — Introduction to Machine Learning Algorithms. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gupta, S. (2022, February 25). Regression Vs. Classification In Machine Learning: What's The Difference? Springboard Blog. https://www.springboard.com/blog/data-science/regression-vs-classification/

Harvard Health. (2020, March 25). 7 ways to reduce stress and keep blood pressure down.

https://www.health.harvard.edu/heart-health/7-ways-to-reduce-stress-and-keep-blood-pressure-down

Higuera, V. (2018, September 29). How Is Heart Disease Diagnosed? Healthline. https://www.healthline.com/health/heart-disease/tests-diagnosis#physical-exam-and-bloodtests

Madell, R. (2020, December 28). Stress Management for the Self-Employed. FlexJobs Job Search Tips and Blog. https://www.flexjobs.com/blog/post/stress-management-self-employed/

Roland, J. (2018, September 29). Is It a Stroke or a Heart Attack? Healthline. https://www.healthline.com/health/stroke-vs-heart-attack

Watson, S., & Yee, D. (2017, February 28). Why You Should Become a UseR: A Brief Introduction to R. Association for Psychological Science - APS. https://www.psychologicalscience.org/observer/why-you-should-become-a-user-a-brief-introduction-to-r

Appendix