Sam Loyd

Reducing Stroke Risk

June 2021

Executive Summary

The purpose of this project was to analyze a stroke related dataset to determine if it could be used to train a machine learning model to target individuals for risk reduction strategies by their health care providers. In addition, identifying correlations between stroke risk and other dimensions was a desired secondary outcome. Not only are strokes a leading cause of death in the United States, but the long-term consequences to a survivor can be life altering. The cost of associated medical treatments burdens society. The dataset used in this work was obtained under a personal license freely distributable for any non-commercial ventures. The initial data analysis found that the target dimension of having had a stroke was considerably imbalanced towards patients with no stroke history. This problem makes training a model more challenging. Accuracy becomes less of a primary metric as always predicting no would yield a high accuracy, but entirely miss the goal of risk mitigation. Assuming everyone is at risk would waste too many resources. Ultimately, a model was found that was moderately successful at finding at risk individuals while minimizing the targeting of others who were not. This model could be used to target resources without being overly burdensome on the entire population.

Abstract

In 2019, the World Health Organization reported stroke as the second leading cause of death (World Health Organization, 2020). Even for those who survive a stroke, there is often a slow recovery with life-long consequences (CDC, n.d.). Being able to predict that likelihood, would allow for recommendations on preventative measures by health care workers. This research sought to explore that potential with the goal of providing a tool for identifying that risk by using a machine learning binary classification model based on the dimensions found within the stroke prediction dataset. The research also sought to determine if there were any significant correlations found within those dimensions to having had a stroke.
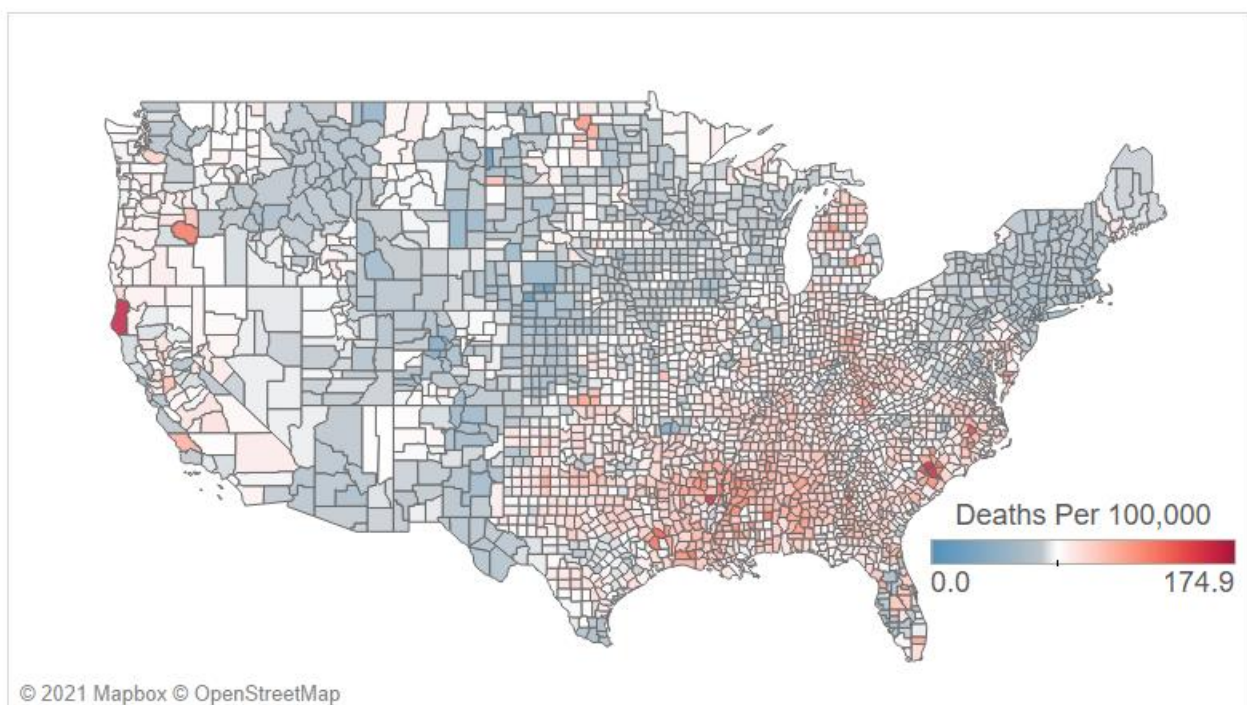
Introduction

Strokes affect almost 800,000 Americans each year according to the Center for Disease Control (CDC, n.d.). Most strokes are ischemic and are caused by vascular blockage. Based on data from 2015, 46 billion dollars was spent on strokes related costs for the previous year (CDC, n.d.). Slightly over 12 percent of strokes are fatal within the first month which increases to 25 percent during the first year (Higuera, 2019). Even for those that survive, loss of motor skills, vision loss and mood swings can lead to disability (Higuera, 2019). The Southern United States is particularly impacted by stroke mortalities. Figure 1 shows how the Southern United States has a high concentration of stroke related deaths.

Figure 1

Center for Disease Control

With all that in mind, a dataset was acquired from Kaggle with related dimensions (2020). This was used to train a model to target those individuals at risk. That dataset was processed using a systemized methodology addressing all stages of the data science process. The dataset consisted of eleven variables. These included an identity key, gender, age, hypertension status, heart disease status, if the patient ever been married, type of work, type of residence, average glucose level, body mass index (BMI), smoking status, and the target stroke variable.
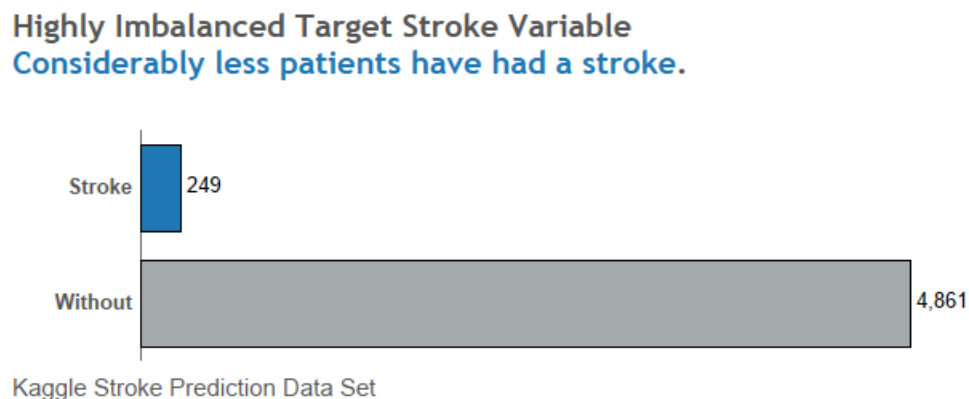
Methods

The CRISP-DM methodology was implemented for this project (Siegel, 2016). As such, six iterative stages were followed.

- Business (Domain) Understanding

- Data Understanding

- Data Preparation

- Modeling

- Evaluation

- Deployment

In the context of this paper, the first four will be covered in the current methods section. The evaluation and deployment phases of CRISP-DM can be found in the results and conclusion sections. Domain knowledge was acquired first. This involved researching stroke related publications referenced above. The data was analyzed in the data understanding phase and explored using statistical measures. In traditional data science, this is also referred to as exploratory data analysis. The primary concern with this dataset was a large imbalance between those who had had a stroke and those who had not as noted in Figure 2. Less than five percent were labeled as stroke victims. Missing data was also discovered for the body mass index variable. Marked skew was noted on the numeric variables for body

mass index and average glucose levels (see Appendix for reference histograms) causing them to deviate

from a normal distribution. Outliers outside normal ranges were noted as the likely cause. The data was

then cleansed and prepared accordingly. Missing body mass index data was replaced with a nearest

neighbor model. Given the non-normal distribution of the data, the numeric dimensions were

normalized before model training. However, removing patients with missing values or removing those

with outliers exacerbated the imbalanced target stroke variable problem. This was tested and decided

against.

Figure 2

Highly Imbalanced Target Stroke Variable
Considerably less patients have had a stroke.

Stroke — 249
Without — 4,861

Kaggle Stroke Prediction Data Set

To address the imbalance concern, synthetic oversampling techniques were applied.

Oversampling involves redrawing from the underrepresented value randomly until an adequate balance

for training is achieved. In synthetic oversampling these values are created artificially based on closeness

to nearest neighbors in the training data. Oversampling was applied to the training data after the data

was split into training, test and validation sets to avoid leaking information into the test and validation

data which could provide a false measure of accuracy. Multiple models were enhanced through

aggregation techniques and then combined into a blended final model. Overall accuracy is not the ideal

metric for tuning best fit in an imbalanced dataset. As such, focusing on relevant results using recall was
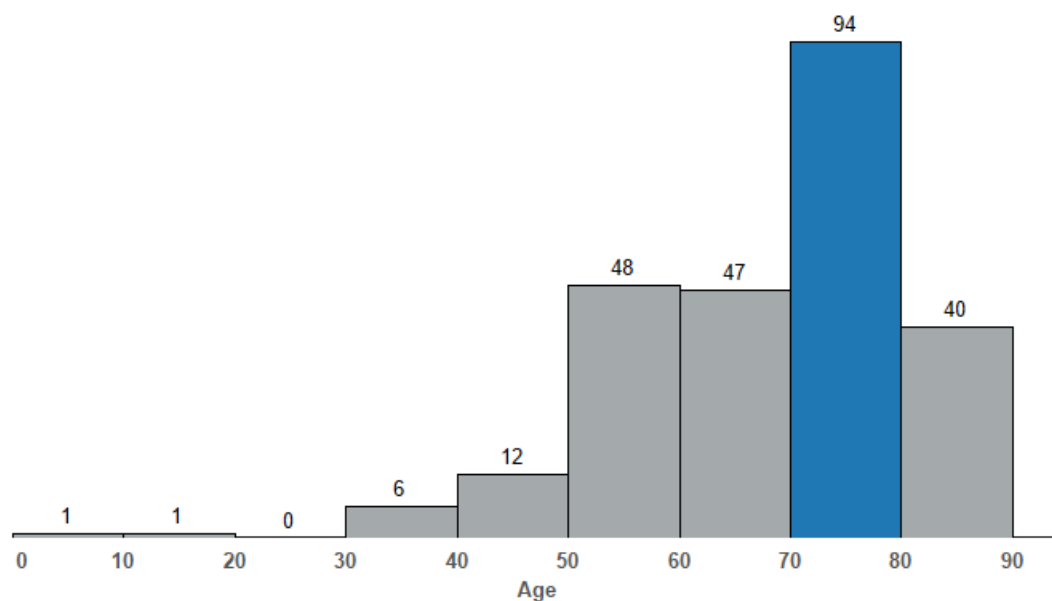
used to tune the blended model.

Results

The model was moderately successful at predicting stroke risk. It was 79 percent accurate at

predicting individuals who had not had a stroke. It was 75 percent accurate at predicting patients who

had. This left the overall accuracy at 79 percent. These measures were in line with more advanced

evaluation metrics for this model's performance. This result struck a reasonable risk versus reward

threshold. Only age showed a moderate degree of correlation across three statistical methods. Age was

noted as the primary feature in multiple models blended to make up the final model. This was also in

line with total counts for strokes for each decade of life as shown in Figure 3.

Figure 3

Strokes Per Decade of Life
Strokes trended higher after 30 until peaking in the 70s.
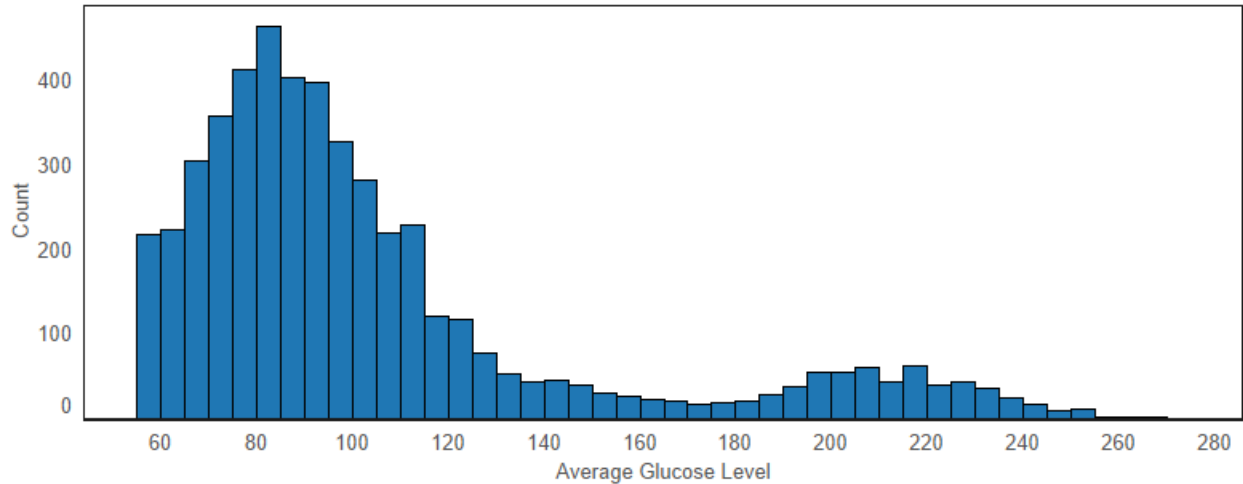
Kaggle Stroke Prediction Data Set

Conclusions

This model could be used as a guide in a clinical environment for targeting those more likely to have a stroke. While not highly accurate, a reasonable balance was truck between predicting those with a stroke history and those without. The final model was processed for deployment and stored in a format for cross platform portability. Further training against more data would likely improve the predictive power. Given the variance shown in the heat map from Figure 1, it would also be worth researching models that included a geographic dimension.

Appendix

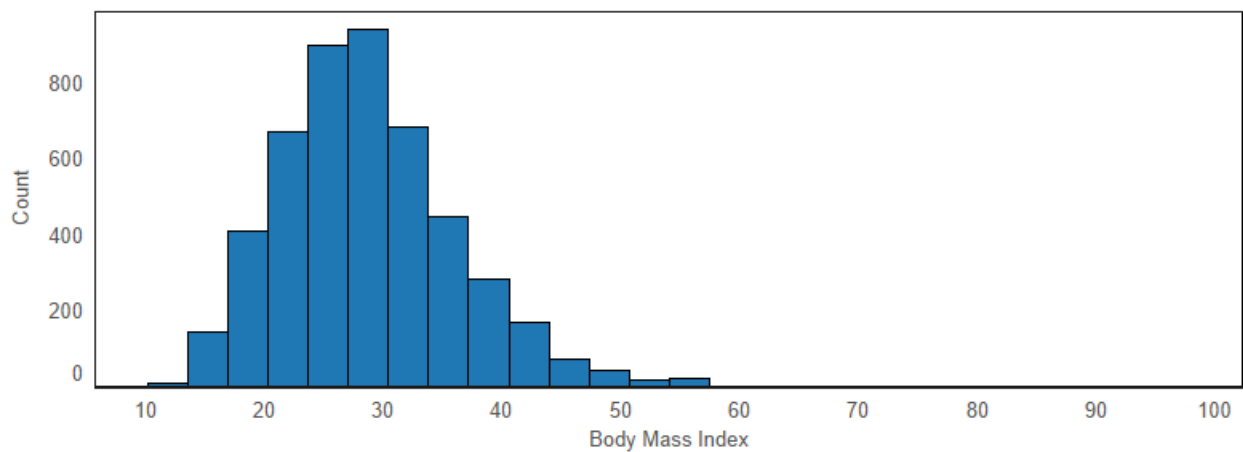## Average Glucose Level Histogram
Approximate kurtosis of 1.7 and skew of 1.6

Kaggle Stroke Prediction Dataset

Body Mass Index Histogram
Approximate kurtosis of 3.3 and skew of 1.1

Kaggle Stroke Prediction Dataset

References

CDC. (n.d.). Stroke Facts. Retrieved from https://www.cdc.gov/stroke/facts.htm

Higuera, V. (2019, June 10). Stroke Severity and Mortality: Types, Treatments, and Symptoms. Retrieved from https://www.healthline.com/health/can-you-die-from-a-stroke

Kaggle. (2020). Stroke Predictions Dataset. Retrieved from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata

Siegel, E. (2016). Predictive Analytics. Hoboken, NJ: Wiley.

WHO? (2020, December 9). The Top 10 Causes of Death. Retrieved from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death