Optimizing Stroke Prediction Models: A Comparative Study on Class Imbalance Techniques:

Group 2 - Brendan Baker, Landon Carpenter, Brian Wimmer Georgetown University Data Science & Analytics

ANLY 512: Statistical Learning

Prof. Nakul Padalkar

May 5, 2023

Abstract

This research paper aims to explore risk factors for stroke patients and create machine learning models to predict the risk of a stroke. Gender, age, hypertension, heart disease, marital status, work type, residence type, BMI, and smoking status were used in the exploration and modeling of this paper. Next, an exploratory data analysis was used to explore the data visually. Correlation plots, jitter and mosaic plots, and histograms were crafted to pull insights from certain variables of interest. Multiple statistical models were utilized to predict stroke risk from the various clinical features, with a research emphasis on class imbalance strategies. Feature selection, oversampling, performance metrics, ensemble methods, and cost-sensitive classification and thresholding were among the methods practiced. The cost-sensitive logistic model was concluded to be the best model for predicting strokes based on its F1 score (.30). While all of the models demonstrated high accuracy, they were found to perform relatively poorly when looking at the precision, recall, and F1 score. Additional models using other methods and techniques could be created in the future to try to discover a better overall model, however, a high class imbalance can be difficult when creating a predictive model. In conclusion, many medical datasets suffer from class imbalance and further techniques could be utilized to alleviate those concerns.

Keywords: stroke, class imbalance, classification, heart disease, hypertension, bmi, glucose, smoking, blood, machine learning, logistic regression, classification, prediction, clinical data, statistical modeling

Introduction

Strokes are a potentially deadly or debilitating disease that over 795,000 people in the United States experience each year. In fact, every 40 seconds someone in the United States has a stroke, and every 3.5 minutes someone will die of a stroke. Those lucky enough to survive, if they have a stroke, are left with physical, cognitive, emotional, and financial challenges. There are two main categories of strokes: hemorrhagic and ischemic.

Hemorrhagic strokes are caused by bleeding in the brain, while ischemic strokes occur when a blood vessel is blocked by a clot. Beyond these two categories, there are many more subtypes of strokes that can occur. Given the complexity of stroke and its significant impact on individuals and society, there is a growing interest in leveraging data science and machine learning approaches to better understand the disease, identify risk factors, and develop more effective prevention and treatment strategies (CDC, 2020).

The purpose of this research project is to explore what factors, if any, contribute to a patient having a stroke. Our goal is to create various visualizations and models to gather insights from the data and establish some reasoning behind strokes.

Methods

Data

The data was gathered from Kaggle and no main source was given, as it is protected health information (PHI) and meant to be used for educational purposes only. It includes various clinical features and data-points for predicting strokes:

variable	levels / description	type
id	unique identification number	-
gender	male, female, or other	categorical
age	age of patient	numerical
hypertension	0 for no, 1 for yes	binary
heart_disease	0 for no, 1 for yes	binary
every_married	yes or no	categorical
work_type	Children, Govt_jov, Never_worked, Private, or Self-employed	categorical
residence_type	rural or urban	categorical
avg_glucose_level	average blood glucose	numerical
bmi	body mass index	numerical
smoking_status	formerly smoked, never smoked, smokes, or Unknown	categorical
stroke	0 for no, 1 for yes	binary; target

Table 1: Variables and Datatypes for the Stroke Dataset

Exploratory Data Analysis

To begin exploring the data set and how certain factors might play a role in stroke risk, various visualizations were created. The purpose of these visualizations was to get a better understanding of the data and the types of patients within the data set. R was used with ‘ggplot’ to plot all visualizations.

Figure 1: Correlation Plot. The correlation plot was created using only the numerical variables in the dataset (age, average glucose level, heart disease, stroke, bmi, and hypertension). Among numerical variables, heart disease and age, as well as hypertension and age, seem to be the highest correlated at 0.3.

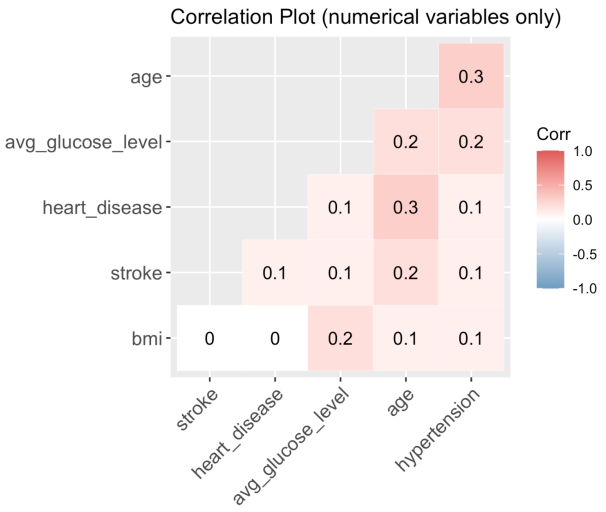

Figure 1: Correlation Plot

Figure 2: BMI, Glucose Level, and Age vs. Stroke Status. Jitter plots of BMI, average glucose level, and age were created to show the relationship between them and stroke status. Additionally, the mean value for each was plotted using the black diamond. Among those who had a stroke, the mean average glucose level and age was much higher than those who did not have a stroke. BMI was about the same for both groups.

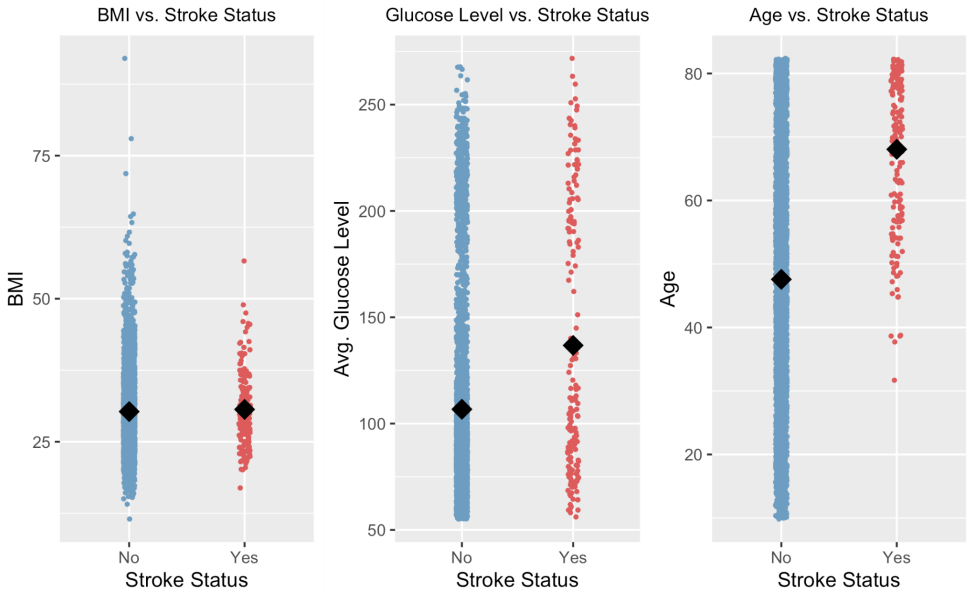

Figure 2: BMI, Glucose Level, and Age vs. Stroke Status

Figure 3: Mosaic Plot - Gender & Smoking Status. A mosaic plot was created to show the total patients by smoking status and gender. From the plot, we can see more females overall in the

dataset, as each smoking status is over 50% female. However, males make up a larger proportion of those who smoke or have formerly smoked, in comparison to females.

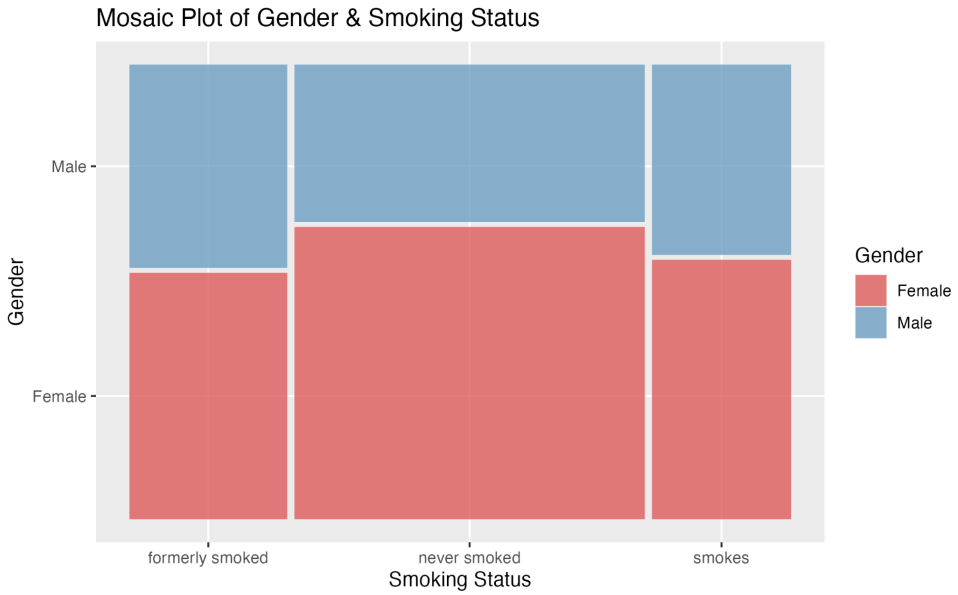

Figure 3: Mosaic Plot - Gender & Smoking Status

Figure 4: Histograms of Age, BMI, & Average Glucose Level. The plots show the distribution of observations for age, BMI, and average glucose level. Also visible on the plot are the average values for each: age (48.65), BMI (30.29), and average glucose level (108.31). From these averages, we can see that the typical patient in the dataset is nearly 50 years old, classified as obese (CDC, 2021), and is prediabetic (Cleveland Clinic, 2022).

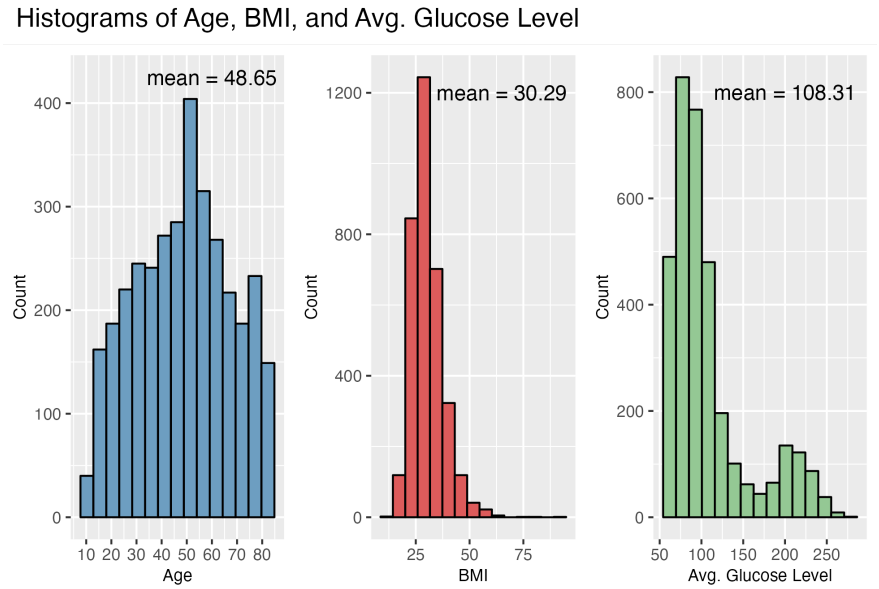

Figure 4: Histograms of Age, BMI, & Average Glucose Level

Statistical Models

Predicting stroke risk from the available features in the dataset is a binary classification problem. The primary goal of this study is to compare different classification models and determine which model has the most utility. While there are many types of classification models available, this study focused on models that could account for class imbalance. The stroke dataset suffered from a significant class imbalance - with only 146 observed strokes compared to 2594 negative stroke results. Since just 5% of the observations belong to the positive class, alternative models needed to be considered.

Several approaches to handling class imbalance have been suggested in the literature. These methods include feature selection, relying on different performance metrics, sampling methods, cost sensitive (weighted) classification, and ensemble methods (Abd Elrahman & Abraham, 2013; Japkowicz & Stephen, 2002). Each of these methods were utilized here.

Least Absolute Shrinkage and Selection Operator (LASSO) Model

For feature selection, a Least Absolute Shrinkage and Selection Operator (LASSO) model was used. A lasso model adds a penalty to the loss function that is equal to the sum of the absolute values for the variable coefficients, multiplied by a lambda parameter that controls the penalty. In this case, LASSO was applied to a logistic regression, as in the formula below:

$$L(w) = -\sum_{i=1}^{n}[y_i \log(p(y_i|x_i, w)) + (1 - y_i) \log(1 - p(y_i|x_i, w))] + \lambda \|w\|_1$$

Equation 1: Logistic Regression Cost Function with LASSO Regularization.

Applying LASSO for an optimal lambda value leaves only the features that are most useful in predicting the target. The stats package in R was used for LASSO logistic regression.

Performance Metrics

Many classification models are first evaluated in terms of their accuracy score. In cases of class imbalance, however, accuracy is a poor measure of model quality. Consider that in the current study, a model that predicts every value to be the negative class would be 95% accurate, but would have no utility. The current study used F1 score as a primary performance metric, which is the harmonic mean of precision (proportion of true positives out of all positive predictions), and recall (proportion of true positives out of all positive instances).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \qquad Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Equation 2: Precision, Recall, and F1 Score

Precision, recall, and F1 score are ideal for evaluating models that work with a class imbalance as they all focus on the model's ability to predict the minority class. A model that is built for

predicting strokes should primarily be evaluated on how well it can determine when a stroke has occurred to be useful.

Oversampling

Oversampling techniques are another method that has been proposed to maneuver the class imbalance problem. One such method is the Random Over-Sampling Examples (ROSE) package in R, which is specifically designed for binary classification (Lunardon, Menardi, & Torelli, 2014). The procedure for generating one new artificial sample is as follows:

1. *Select* $y^* = Y_j$ *with probability* π_j
2. *Select* $(x_i, y_i) \in T_n$, *such that* $y_i = y^*$, *with probability* $\frac{1}{n_j}$.
3. Sample x^* from $KH_j(\cdot, x_i)$, with KH_j a probability distribution centered at x_i and covariance matrix H_j

Algorithm 1: Random Over-Sampling Examples (ROSE)

Repeated sampling in this manner is conducted until the class balance is approximately equal. The newly sampled data is then used as input into the model.

Cost Sensitive Classification and Thresholding

Cost sensitive classification assigns different penalties for different types of misclassification errors. In the current study, we assigned different cost ratios to a cost matrix for a logistic regression that penalized misclassification of the minority (stroke positive) class to varying degrees. The cost matrix was passed as the weights argument to the glm function from the stats package in R.

Thresholding, a similar approach, works with the probability outputs from a logistic regression. In this technique, a specific evaluation metric is optimized by changing the decision boundary for classification of the positive class. In this case, we iterated through different thresholds of the logistic regression to optimize the F1 score.

Ensemble Methods (Random Forest with Stratified Random Sampling)

Finally, we attempted to use ensemble methods as a solution to the class imbalance problem. Ensemble methods such as random forest are another common method used in the literature for this type of problem (Liu & Zhou, 2013). Random Forest (RF) models combine many weak learners (decision trees) that 'vote' on the most probable class. RF models can also be combined with a sampling method so that each weak learner works on a different subsample of the training data. Here, Stratified Random Sampling was used to create a balanced random sample for each tree that was constructed as part of the model. The RandomForest package in R was used for analyses.

Results & Discussion

Model Evaluation Results

The best model from each category is listed in table 2 below. Overall, the cost-sensitive logistic model had the best F1 score at .300. The model with the best recall was the logistic model that used data oversampled with the ROSE algorithm. Of the RF models, the base model had the highest precision of all models, and the RF model that utilized stratified random sampling and with an adjusted cost matrix had the highest F1 score.

Model Type	Accuracy	Precision	Recall	F1 Score	Notes/Best Parameters
Logistic - Base	.950	N/A	0	N/A	No predicted positives
Logistic - ROSE	**.764**	**.128**	**.647**	**.213**	**Oversampling**
Logistic - LASSO and Threshold Tuning	.842	.170	.559	.260	Threshold .10
Logistic - Cost Sensitive	**.900**	**.404**	**.239**	**.300**	**5:1 cost ratio**
Random Forest - Base	.832	.411	.128	.195	Mtry = 4, ntree = 900
Random Forest - Stratified Random Sampling	.775	.130	.618	.214	Sample size = 100
Random Forest - Stratified + Cost Sensitive	.802	.141	.588	.228	10:1 cost ratio

Table 2: Model Evaluation Metrics. The best model for each category, optimized by F1 score, is listed.

Discussion and Future Directions

In general, all of the models that were built here performed poorly in terms of F1 score. While many of the models had high accuracy, none of them were very accurate at predicting the positive class. It would not be advisable to use these models in any clinical setting. Some utility could be gained by using one of the high recall models, which could flag patients for further stroke screening. In this case the model would overpredict the amount of strokes but would capture many of the actual strokes.

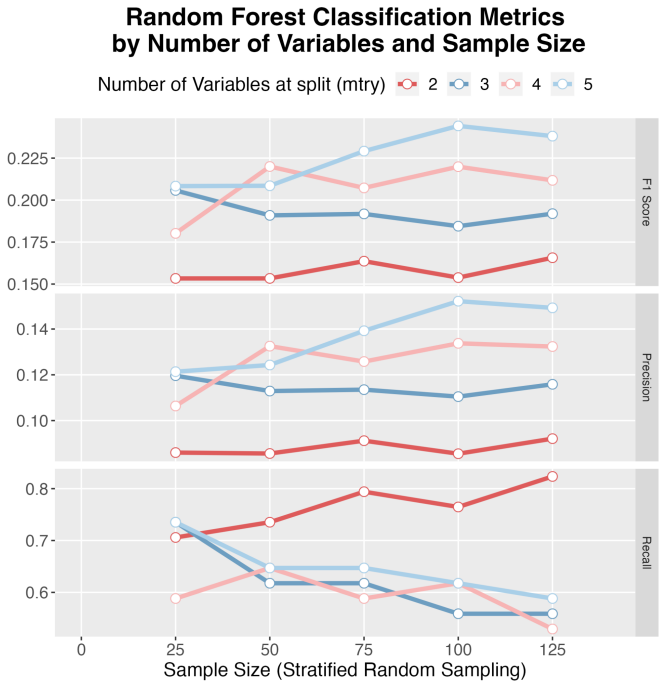

Figure 5: Hyperparameter Tuning for F1 Score Optimization. Precision, Recall, and F1 Score are shown as a function of two hyperparameters - mtry (number of variables sampled at each split, and sample size (stratified random sampling).

A better dataset that has more instances of stroke - even if still unbalanced - could permit better classification by oversampling from a more diverse range of observations and then allow for more complex models such as neural networks to be trained. Overall, much more work is needed to be able to classify positive stroke cases precisely and deliver utility to physicians and medical practitioners.

Conclusions

Strokes affect the lives of thousands of people and their families each year in the US. Our goal was to explore the risk factors and to what extent they can be used in predicting strokes. After gathering and cleaning the data, we created various visualizations - discovering that there was no major clear correlation between any of the features. We also noticed that among those who had a stroke, they were often older and had higher average glucose levels.

We also created various models to try to predict a stroke, based on the features available to us. These models included logistic regression and random forest models with various cost-sensitive and sampling techniques. While some models exhibited good accuracy, this was primarily due to the data class imbalance allowing them to label very few strokes, and overall when considering other performance metrics the models performed poorly. The overall poor performance of these models highlights the difficulties a data scientist might encounter when working with unbalanced classes. Classifying medical data can often suffer from class imbalances due to low prevalence of a condition. Some of the methods applied here may be considered for use in such cases.

References

1. Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*(2013), 332-340.

2. CDC. (2020, January 31). *Stroke Facts | cdc.gov*. Www.cdc.gov. https://www.cdc.gov/stroke/facts.htm#:~:text=Stroke%20Statistics

3. CDC. (2021, June 7). *Defining Adult Overweight and Obesity*. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=Adult%20Body%20Mass%20Index&text=If%20your%20BMI%20is%20less

4. Cleveland Clinic. (2022, November 16). *Blood Glucose (Sugar) Test: Levels & What They Mean*. Cleveland Clinic; Cleveland Clinic. https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test#:~:text=A%20healthy%20(normal)%20fasting%20blood

5. fedesoriano. (n.d.). *Stroke Prediction Dataset*. Www.kaggle.com; Kaggle. Retrieved March 25, 2023, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?page=2

6. Liu, X. Y., & Zhou, Z. H. (2013). Ensemble methods for class imbalance learning. *Imbalanced learning: Foundations, algorithms, and applications*, 61-82.

7. Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R journal*, *6*(1).

8. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429-449.