The Impact of Physiological and Lifestyle Factors on Stroke

Utkarshna Sinha, Joseph Punnapuzha, Nahush Bhat, Dachuan Zhang

Summary

Overview

According to the CDC, someone has a stroke every 40 seconds and dies of it every 4 minutes in the US. Suffering from a stroke can lead to lasting brain damage, long-term disability, or even death. We wanted to understand if certain physiological and lifestyle factors could impact the occurrence of stroke as well as investigate if we can accurately predict whether a person is likely to have a stroke. Relationships or findings between the different factors and the incidence of stroke were examined using visualizations in R. Furthermore, the project aimed to develop a supervised machine learning model to predict stroke likelihood in a patient given these factors.

Goals

Through analysis of this dataset, our goals were twofold:

1. To find any apparent relationships between different physiological or lifestyle factors provided, and the incidence of stroke.
2. Develop a model that could accurately predict what factors may lead to a stroke.

The identification and analysis of factors linked with stroke can lead to enhanced preventative healthcare for patients who are determined to be at a higher risk. The exploration of this dataset, and a strong predictive model may lead to better healthcare outcomes for patients.

With regards to our second goal of developing a stroke prediction model, our hypotheses are as follows:

Null Hypothesis (H_0): There is no significant relationship between stroke status and any factors.
Alternative Hypothesis (H_a): There exists a relationship between stroke status and a combination of physiological and lifestyle factors.

Dataset Description

A publicly available Kaggle dataset that has tracked the incidence of stroke with other factors was used for our analysis. In total, 12 attributes are tracked, including physiological factors such as a patient's BMI, average glucose level, smoking status, the presence of heart disease and/or hypertension. It also includes a patient's demographic data, including their age, gender, the type of work they do, their residence type, and marital status. There are over 43,000 observations of unique patients in the data where 42,617 have not had a stroke, and 783 have had a stroke.

Methods

Data Preprocessing

The dataset being used was already in a structured format, and tidy. However, various adjustments were made to deal with missingness in the dataset, and to make the data more suitable for our analysis.

- ID: Unique IDs for each observation not meaningful for our analysis, and therefore, removed.
- Age groups: A new column for age group was added to enhance the age analysis of the data.

- Gender: 11 patients were categorized as 'Other' in the gender column. These were dropped from the gender analysis since it was too few to provide significant insight. (11 vs ~43K records)
- Smoking status: Smoking status column had about 13,292 (~30%) missing values. Since this was a significant proportion, a new value of "unknown" was used to impute the NA values.
- BMI: There were 1,458 records (~3%) which were marked NA in the BMI column. Since 140 of those observations were patients who suffered a stroke, this represented a large percentage of the 783 total stroke patients in the entire dataset. The NA values in the BMI column were imputed with the overall mean BMI.

Exploratory Data Analysis using corrgram and ggplot2

Preliminary Analysis

We used the corrgram package to identify any preliminary relationships and get a quick understanding of the dataset.

Exploratory Data Visualizations

We used ggplot2 in R to create various visualizations in order to understand the data, inspect relationships between variables, and conduct other exploratory data analyses. We inspected each of the variables of the dataset, especially against the stroke variable, which indicated if a patient had stroke or not. For the data visualizations, we created a column for age groups to visualize data by age groups categorically rather than continuously.

Predictive Modeling Using the caret package

Since the stroke response variable is categorical, we opted to use supervised learning classification techniques including k-nearest neighbors (knn) from the kknn package, and logistic regression from the glm package. Since we were dealing with at most 11 influencing variables, a combination of those factors could be used by either model determine whether a person was likely to have a stroke.

Classification Techniques

The train function from the caret package were used to train the model on the training data. The model was then used for testing to determine its accuracy, sensitivity, and specificity. The following steps were performed to develop and evaluate the model:

1. Predictors that were considered irrelevant or showed no effect on stroke incidence according to the exploratory data analysis were removed from the dataset so only relevant potential predictors could be analyzed for predictive potential.
2. The dataset was split into 80% training and 20% test set for the classification algorithms used which were k-nearest neighbors, and logistic regression methods.
```
index <- createDataPartition(stroke_data2$stroke, p = 0.8, list = FALSE)
stroke_t <- stroke_data2[as.integer(index),]
stroke_test <- stroke_data2[-as.integer(index),]
```
3. ROSE sampling was used which generated synthetic balanced samples for both stroke positive and stroke negative categories in the training set. This technique allowed us to lower the impact of the imbalance in the class distribution on our model estimate and evaluation.
```
stroke_train <- ROSE(stroke ~ ., data = stroke_t, seed=22)$data
```

4. Although the training set was resampled using ROSE sampling, the original test set was used for testing without resampling.
5. The knn and logistic regression models were fitted to the resulting training dataset. The fitting was done stepwise starting with the strongest predictor and adding additional predictors to inspect the change in accuracy, sensitivity and specificity of the resulting confusion matrix when the trained model was tested on the test set. The caret package was used with median imputation of missing values.

```
s_fit1 <- train(stroke ~ age, data=stroke_train,
                method="glm", family=binomial(link="logit"),
                preProcess="medianImpute",
                trControl=trainControl(method="none"),
                na.action=na.pass)
```

6. A confusion matrix of the prediction was analyzed for p-value (alpha = 0.05), accuracy, sensitivity and specificity.

```
confusionMatrix(predict(s_fit1, stroke_test, na.action=na.pass),
                stroke_test$stroke)
```

7. The ROC curve was plotted and the area under the curve was used to determine the strength of the classifier.

Results

Exploratory Data Analysis

Preliminary Correlations

The preliminary correlational data (Figure 1) showed correlations between stroke status and some of the candidate factors including age, heart disease, hypertension and average glucose levels. This plot allowed us to get a quick understanding of each of the variables based on the whole dataset.

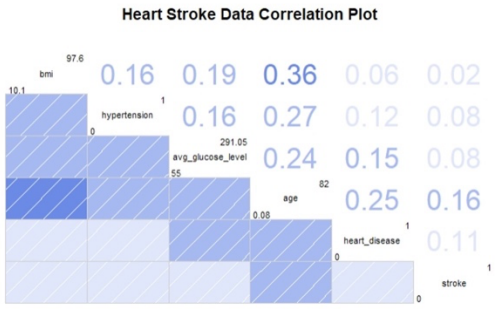

Figure 1. This correlational plot shows the relationship between heart stroke status, BMI, hypertension, average glucose level, age, and heart disease.

Further EDA

After our preliminary look at the data, a deeper analysis was conducted with various visualizations to inspect relationships between variables in the data set and stroke. Variables including gender, body mass index, residence type, marital status, and work type showed no impact on stroke incidence after analysis was conducted using correlation data and visualizations. Notable results were found between the following variables and stroke: (1) age, (2) hypertension, (3) heart disease, and (4) average glucose levels.

Age

A key relationship discovered was that as age increases, the count of stroke patients increased, regardless of gender, as seen in Figure 2.1. Thus, older patients were more susceptible to having stroke. Furthermore, analysis between stroke status (whether a patient had stroke or no stroke) and age was conducted. In Figure 2.2, the left pair of boxes show the median age of those who have had stroke, split by gender whereas the right pair of boxes show the median age of those who do not have stroke. Although there appears to be no difference between the genders, the median age of those who had a stroke is roughly 25 years greater than those who did not. Thus, it appears that age may have an impact on stroke.

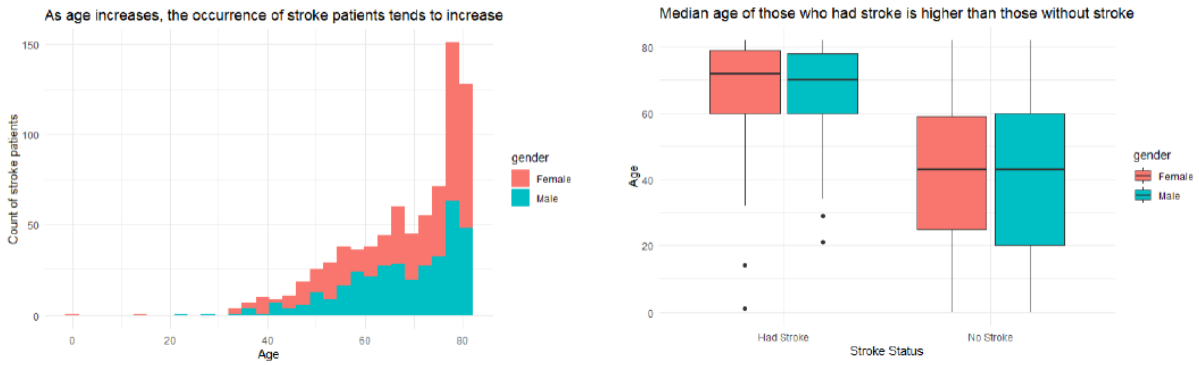

Figure 2.1 (left) shows increase in occurrence of stroke patients as age increases. Figure 2.2 (right) illustrates stroke status (had stroke and no stroke) vs. Age.

Hypertension

Visualizations of the data show that the percentage of people with hypertension is higher in those who have stroke. Since age analysis revealed that a higher number of older people tend to have strokes, people 50 and older were examined first. In Figure 3.1, which shows the percentage of hypertension in people 50 and older without stroke, about 18% of patients are positive for hypertension. In Figure 3.2, of people 50 and older who had a stroke, 27% were hypertensive. This shows that more people who have had a stroke also suffer from hypertension. Further analysis was conducted on those under the age of 50. In that analysis, the trend was similar to patients 50 and older; 11% of people with stroke under 50 were hypertensive compared to 3% who didn't have stroke. Based on this analysis, hypertension should be looked at as having an impact on stroke.

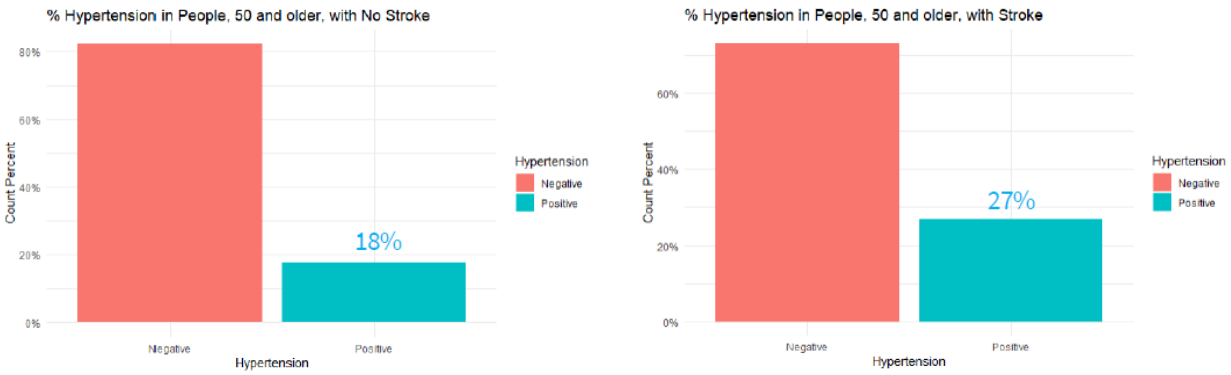

Figure 3.1 (left) shows percent hypertension in people 50 & older who don't have stroke. Figure 3.2 (right) illustrates percent hypertension in people 50 & older who have stroke.

Heart Disease

The percentage of people with heart disease is higher in those with stroke, as indicated by our analysis of patients who suffered from heart disease against those who did not. Figure 4.1 demonstrates the percent heart disease among people 50 and older without stroke. In this chart, 10% of that group were positive for heart disease as a similar examination was done on those 50 and older with stroke in Figure 4.2. In this figure, 25% were positive for heart disease, increasing from the 10% in those without stroke. Therefore, heart disease is a variable we want to consider as impacting the incidence of stroke.

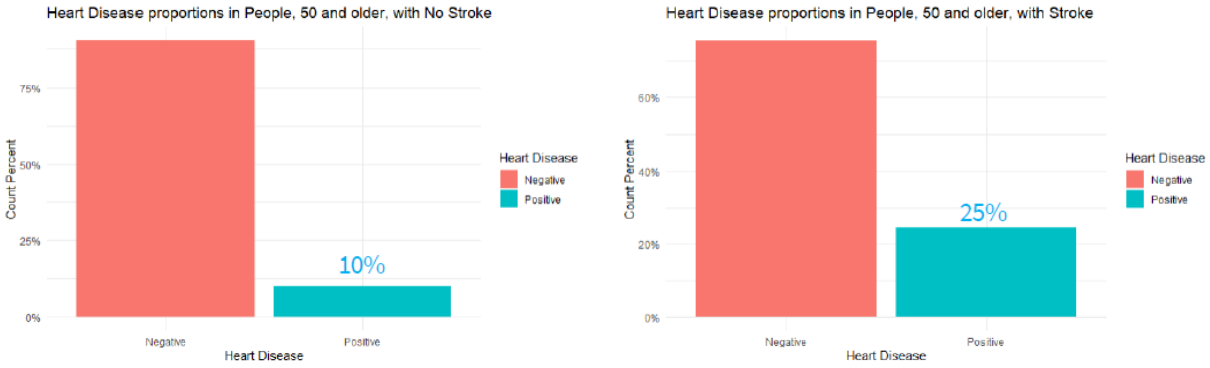

Figure 4.1 (left) shows percent heart disease in people 50 & older who don't have stroke. Figure 4.2 (right) illustrates percent heart disease in people 50 & older who have stroke.

Average Glucose Level

The next variable worth considering as having an impact on stroke from our analysis was the Average Glucose Level. In Figure 5.1, stroke status was compared against average glucose level, faceted by age groups. The median glucose levels are not significantly different for stroke status in any age group. However, on closer inspection at those 50 and over, the median glucose level for those with stroke is slightly higher than those without stroke. Even more worth noting is that the interquartile range is more significant for those who had a stroke than those without, indicating a higher variance in glucose levels. Therefore, the average glucose level may need to be considered impacting stroke based on the dataset.

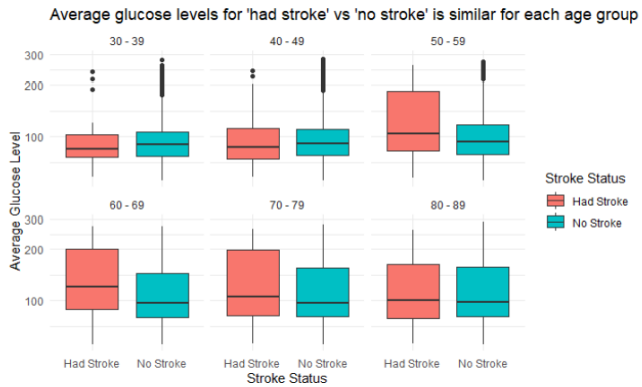

Figure 5.1 shows stroke status against average glucose level faceted by age group.

Supervised Machine Learning Classification

For our predictive analysis, we aimed to investigate which of the candidate predictors could be used to predict stroke incidence to a high level of sensitivity and specificity.

ROSE Sampling

Visualization of the samples that had stroke against those that didn't reveal a significant imbalance in the two groups as only 1.8% of the dataset represented stroke positive patients (Figure 6.1). Only 783 patients recorded had a stroke, whereas 42,617 did not. Due to this imbalance, modeling results were poor as there was a sensitivity of 1 and specificity of 0. After ROSE sampling of the training dataset (Figure 6.2), we were able to get a more balanced distribution in our training set of samples with approximately 16,000 samples in both groups.

Figure 6.1 (left) This graph shows original imbalanced class distribution between no stroke and stroke. Figure 6.2 (right) This graph shows the balanced class distribution in had stroke and no stroke after ROSE sampling.

Logistic Regression and k-Nearest Neighbors

Classification with all the candidate predictor variables was done using both logistic regression and k-nearest neighbors (knn) analysis. Both models resulted in a p-value of less than $2*10^{-16}$, which was below our alpha cutoff value of 0.05, and therefore, we rejected the null hypothesis in both models. As a result, we can state that there is an effect of age, hypertension, heart disease, and average glucose level on stroke incidence.

Stepwise analysis with adding age, followed by hypertension, heart disease, and average glucose level to the logistic regression model resulted in increased accuracy, sensitivity, and specificity, as shown in Figure 7.1. All values increase with the addition of each predictor in the logistic regression model to a final sensitivity of 0.7423 and specificity of 0.7734. Although specificity decreased with the addition of average glucose level, we decided to keep it in the model since this denotes a prediction for stroke with the potential for providing preventative care. It is better to be safe and include results that increase sensitivity even if there is a lower true negative rate.

Logistic Regression Results

Predictors	P-Value	Accuracy	Sensitivity	Specificity
Age	$< 2*10^{-16}$	0.7101	0.7083	0.8077
Age + Hypertension	$< 2*10^{-16}$	0.7166	0.7092	0.8084
Age + Hypertension + Heart Disease	$< 2*10^{-16}$	0.7176	0.7153	0.8141
Age + Hypertension + Heart Disease + avg. Glucose Level	$< 2*10^{-16}$	0.7220	0.7423	0.7734

Figure 7.1 This figure shows the increasing accuracy and sensitivity of the logistic regression model as variables are added. Specificity also increases when hypertension and heart disease are added but decreases when average glucose level is added. However, this regression model proved to be better overall than the knn model.

The knn analysis (Figure 8.1 in the Appendix) showed a final sensitivity of 0.7733 and specificity of 0.6346 with all predictors. However, the addition of hypertension and heart disease, which are binary numerical features resulted in high sensitivity but a specificity of 0, due to the built-in distance function in the caret package. In the future, the distance function could be customized to incorporate binary numerical features. In addition, since the test set was not resampled using ROSE, it remained imbalanced towards patients who did not have stroke which would cause more responses in the knn analysis to be 'no stroke' since knn is sensitive to the local structure of the data.

Overall, the logistic regression analysis resulted in a better predictive model for our data. The ROC curve for the logistic regression model, as shown in Figure 7.2, shows a good level of sensitivity and specificity with an area under the curve of 0.81.

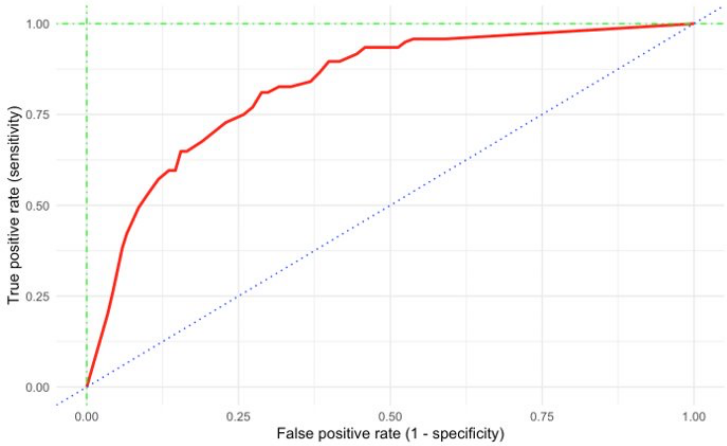

Figure 7.2 The area under the curve for the ROC plot was 0.812 which indicates that the logistic regression model with 4 predictors is a strong classifier for stroke.

Discussion and Conclusions

Given that we were able to develop a predictive model with a reasonable sensitivity and specificity, it can be concluded that older patients who have a history or combination of heart disease, hypertension, and higher glucose levels may be candidates to receive preventative care against stroke. With our analyses, increased age groups were at higher risk, and the predictive ability is stronger if other factors are also present. Despite an overall specificity of 77% which indicates that according to our model, 23% of patients predicted to have a stroke may never actually suffer from one, preventative care may still be advised in order to keep those patients alert for symptoms. This information can be used by policymakers and healthcare providers as guidelines in interactions with patients who are at higher risk, and potentially reduce the adverse health outcomes related to stroke. This would be both beneficial to patients and ease the resource load on the healthcare system.

In future work for this project, we could investigate more classification techniques to possibly improve the sensitivity and specificity of the prediction. The knn analysis could also be improved with the addition of a custom distance function for the binary numerical features to increase its accuracy and retried with a more balanced test set.

Limitations of Analysis

1. There was a significant class imbalance in our dataset with representation of stroke patients being only 1.8% of total patients recorded. This meant that many of the correlations, especially with the categorical variables that have many options, may have been weak due to lack of representation in the stroke positive subset.
2. Most of the labels in the dataset are categorical, which might affect the precision of the model due to the inherent limitations of quantitative attributes. For example, heart disease was a factor, but it was unknown whether a patient suffered from multiple heart conditions. Work type was categorized as self-employed, government job, private sector, or never employed which is not able to capture the differences in employment conditions. There are nuances in types of work which may indicate a sedentary or physically active lifestyle that may be more correlated to health outcomes. In the future, collection of daily physical effort may be collected.
3. Smoking data for 30% of the patients was unknown, and our analysis showed no significant correlations between any smoking status and stroke. In general, smoking is associated with poor health outcomes, and if that data were present, we may have seen relationships that were not visible.