# Characterizing Stroke Risk Factors

Group 4
Sean Mulherin, Robert Craig, Dan Knight,
Andy Wang, Chloe Yang, Lauren Huang

UCLA College | Physical Sciences
Master of Applied Statistics & Data Science

# Abstract

A *stroke* is a ‘brain attack’ that occurs when the blood supply to the brain is disrupted by a blockage or rupture (analogous to a heart attack).

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

The objective of this project is to see if it is possible to predict if an individual is at risk of having a stroke. We will be investigating biological factors and lifestyle factors.

After exploration, we conclude that age, hypertension, and average glucose level play the largest role in predicting the risk of stroke.

One idea for model implementation is to emphasize programs and outreach focused on preventative care for these risk factors of strokes.

UCLA College | Physical Sciences
Master of Applied Statistics & Data Science

https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke
https://www.who.int/news-room/fact-sheets/detail/hypertension
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

# Cohort

## Patient Demographics

Data authored by Fede Soriano (through Kaggle).

The dataset contains **5,110** patients in total. **202** were excluded due to missing data, resulting in a usable dataset of **4,908** patients. **80%** were then sampled randomly for training the model, with the remaining **20%** withheld for testing.

**209** patients have experienced a stroke.

**59%** of patients are male, and **41%** are female.

Ages range from **0.8** years old to **82** years old.



https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

# Research Questions

Can we predict the risk of having a stroke from:

- Biological factors (age, gender, etc.)
- Lifestyle factors (ever married, work type, etc.)

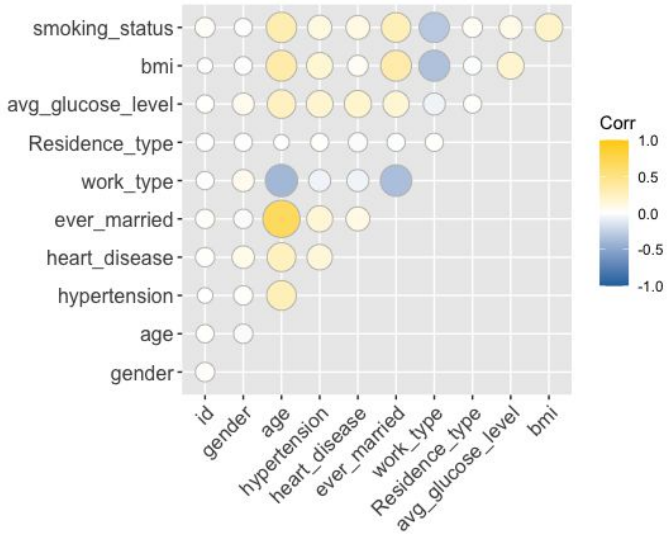
What are the most important factors contributing to the risk of having a stroke?

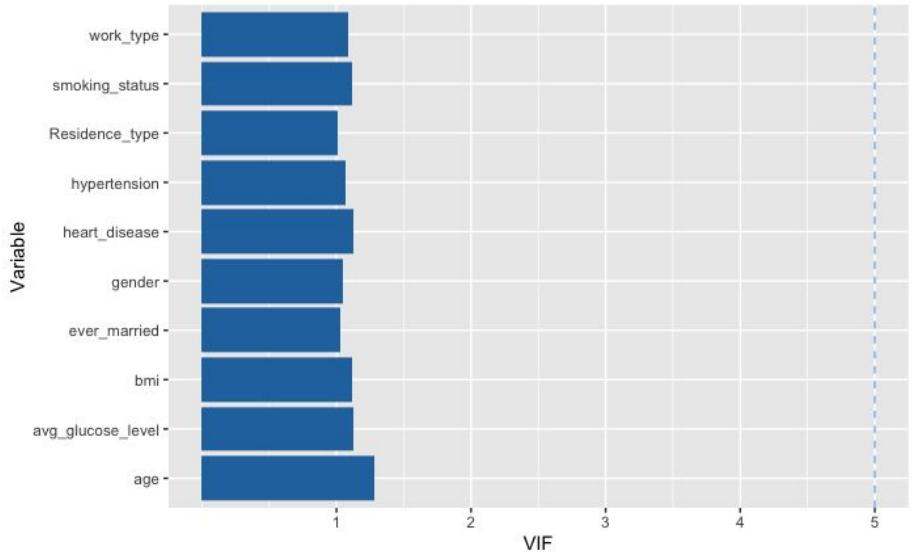
# Analysis Roadmap

| | | | | |
|---|---|---|---|---|
| Biological | **Age** | Numeric (0-82) | | |
| | Gender | Categorical (2 Levels) | | |
| | **Hypertension** | Categorical (2 Levels) | | |
| | **Avg. Glucose Level** | Numeric (55-272) | | |
| | Heart Disease | Categorical (2 Levels) | | |
| | BMI | Numeric (10-98) | → Stroke | Categorical (Binary) |
| Lifestyle | Work Type | Categorical (5 Levels) | | |
| | Residence Type | Categorical (2 Levels) | | |
| | Smoking Status | Categorical (3 Levels) | | |
| | Ever Married | Categorical (Binary) | | |

# Correlation and Multicollinearity

## No issues with important variables (VIF < 5)



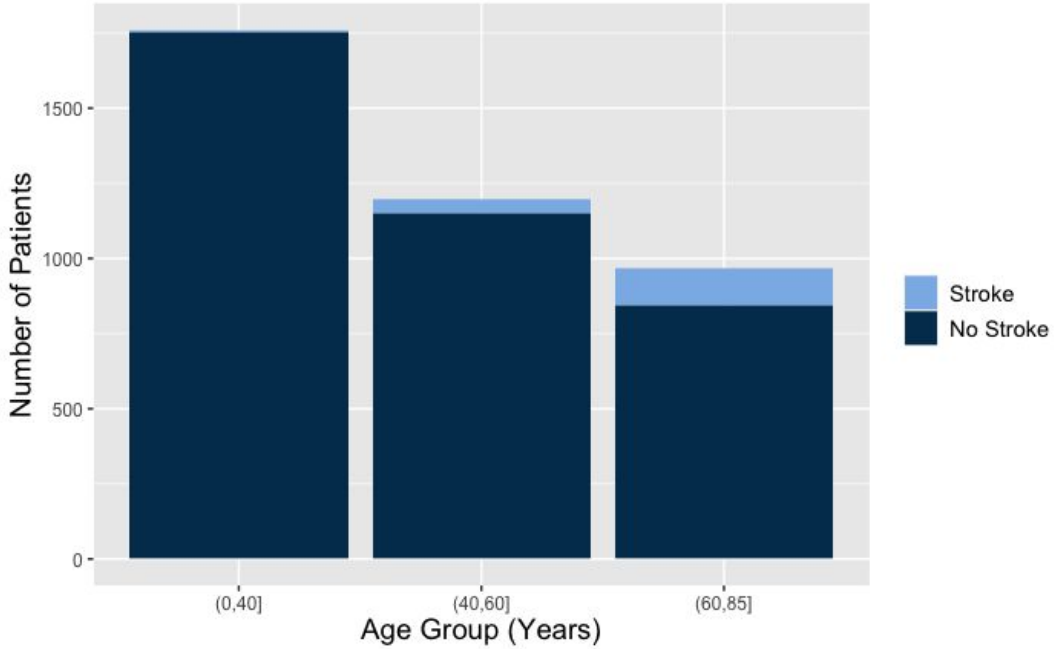



# Age Groups

To improve interpretability, we grouped patients into 3 distinct age groups:

- 0-40 (n = 1,761 with 6 strokes)
- 40-60 (n = 1,197 with 47 strokes)
- 60-85 (n = 968 with 123 strokes)

Ages were chosen for clinical relevance and statistical validity.

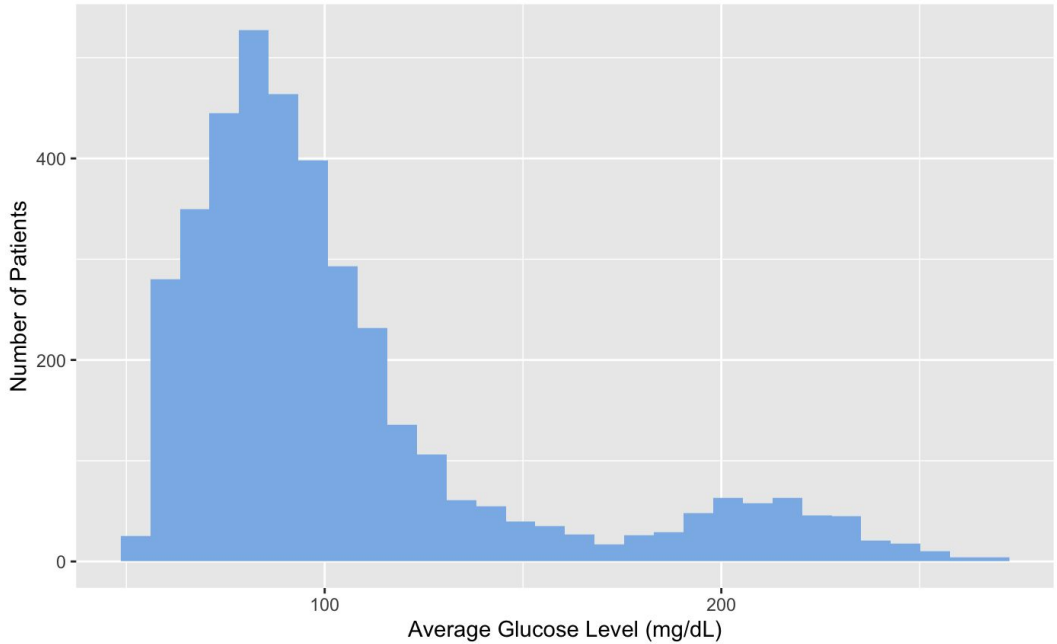


# Average Glucose Level and Diabetes

## Checking for normality

This variable is clearly not normally distributed.

It seems severely right-skewed, but it could also be bimodal.



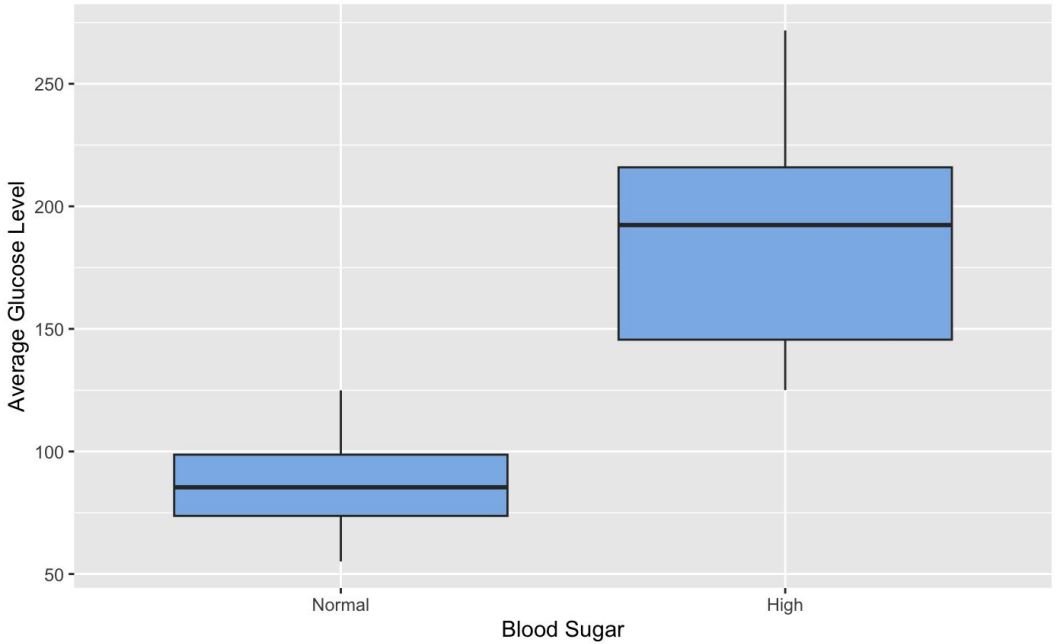
# Normal vs. High Blood Sugar

## Two groups, each normally distributed

**125 mg/dL** is a standard threshold for diagnosing diabetes (closely linked with blood sugar levels).

Splitting the patients here yielded two normally distributed groups.

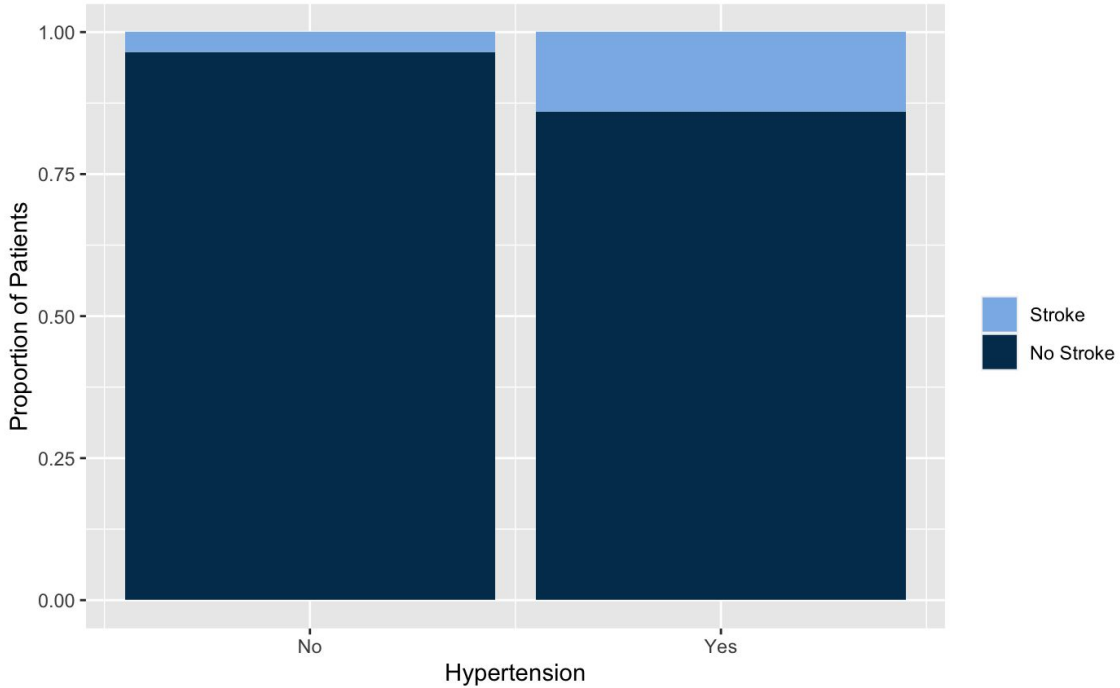
This categorical **blood sugar** variable can be used in our logistic regression model as-is.






https://www.cdc.gov/diabetes/basics/getting-tested.html#:~:text=A%20fasting%20blood%20sugar%20level,higher%20indicates%20you%20have%20diabetes.

# Other EDA (Hypertension)



# Feature Selection

## Blockwise Regression

† Indicates New Variable

| | Feature | $\beta$ (Model 1) | $\beta$ (Model 2) | $\beta$ (Full Model) |
|---|---|---|---|---|
| Biological | Age Group (40, 60]† | **2.324*** ** | | **2.008*** ** |
| | Age Group (60, 85]† | **3.333*** ** | | **3.137*** ** |
| | Gender (Male) | -0.073 | | -0.079 |
| | Hypertension | **0.587** ** | | **0.663*** ** |
| | Blood Sugar (High)† | **0.675*** ** | | **0.693*** ** |
| | Heart Disease | **0.624** ** | | **0.559*** |
| | BMI | -0.006 | | -0.004 |
| Lifestyle | Work Type (Private) | | 14.024 | 12.543 |
| | Residence Type (Urban) | | 0.023 | 0.006 |
| | Smoking Status (Smokes) | | 0.218 | 0.368 |
| | Ever Married | | **0.643*** | -0.329 |

# Final Model

| Predictors | Odds Ratio | 95% Conf. Int. | P-Value |
|---|---|---|---|
| Age(40, 60] | 7.89 | (3.5, 21) | <0.001*** |
| Age(60, 85] | 24.5 | (11.4, 64.1) | <0.001*** |
| Hypertension | 1.85 | (1.27, 2.65) | 0.001** |
| Age(0, 40] : diabetesHigh | 2.2e-6 | (9e-8, 1e-3) | 0.98 |
| Age(40, 60] : diabetesHigh | 2.28 | (1.2, 4.17) | 0.009** |
| Age(60, 85] : diabetesHigh | 2.08 | (1.4, 3.07) | <0.001*** |

# Model Interpretation

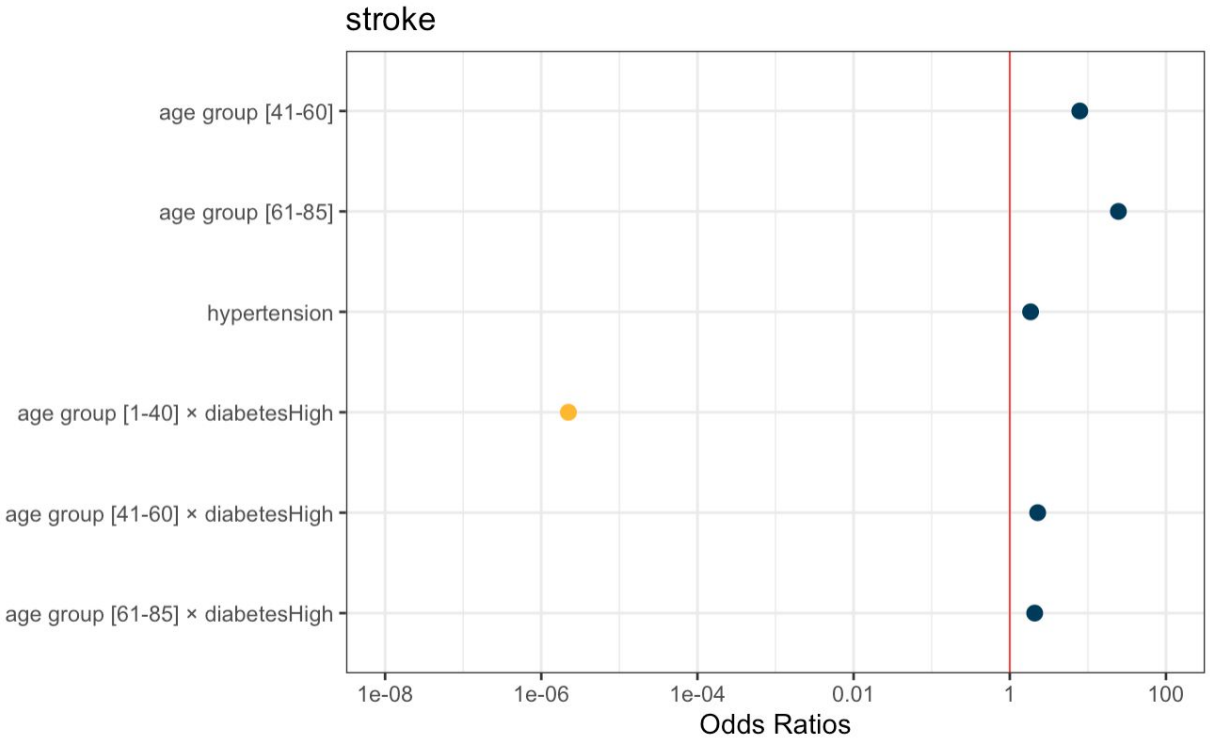
Patients who are between **40 and 60 years old** are **7.9 times more likely** to have a stroke than patients under 40.

Further, patients **over 60 years old** are **24.6 times more likely** to have a stroke than patients under 40.

Patients with **hypertension** are **1.85 times more likely** to have a stroke than patients who do not.

The effect of **high blood sugar varies with age**. Patients with an **average glucose level above 125 mg/dL** between the **ages of 40 and 60** are **2.3 times more likely** to have a stroke than patients who do not have high glucose levels.

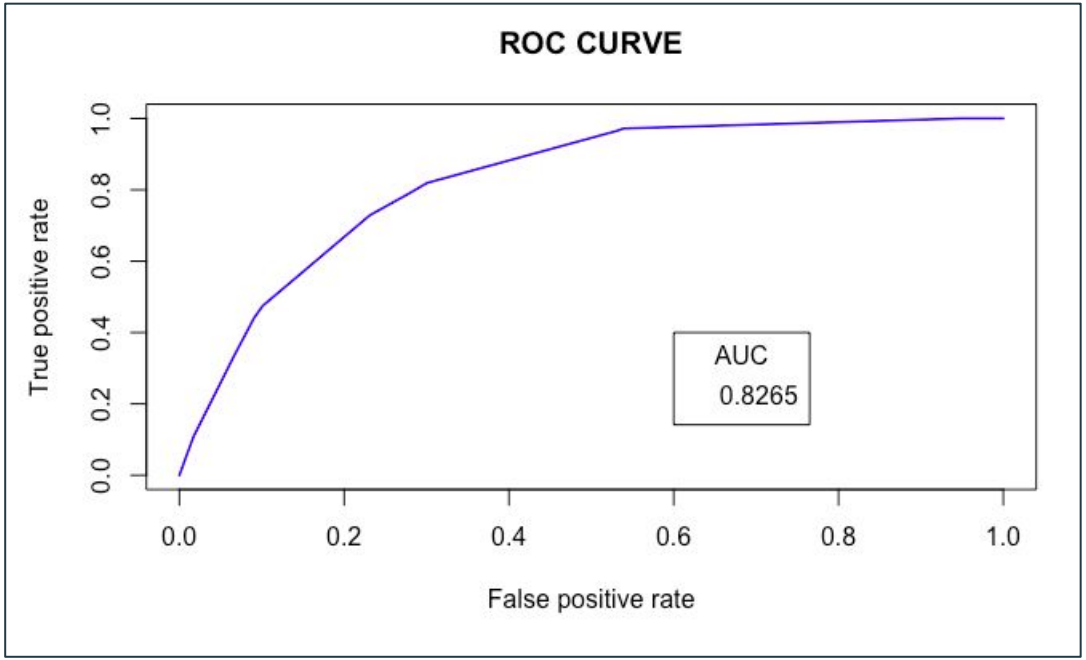
# Confidence Interval Plot



# Application of the Final Model

An **ROC curve** can be constructed to understand the model's accuracy.

This curve was created using the final model and the test dataset.



# Prediction Accuracy

In this clinical pre-screening context, we heavily prioritized **high sensitivity** over overall model accuracy.

Compared confusion matrices computed using multiple thresholds on the training data.

A **threshold of 0.05** was used in the final confusion matrix on withheld test data.

| Accuracy = 66.7%<br>Specificity = 66.5%<br>**Sensitivity = 78%** | | **Observed Outcome** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicted Outcome** | **0** | 630 | 7 |
| | **1** | 320 | 25 |

# Conclusions

**Can we predict the risk of having a stroke from biological factors (age, gender, etc.) and lifestyle factors (ever married, work type, etc.)?**

Yes. Based on the dataset, health factors such as **age, hypertension and average glucose level** are shown to have a significant relationship to stroke risk. On the contrary, lifestyle factors such as smoking, work type, and residence type are statistically insignificant. Therefore, they were not included in the final model.

Our model has **66.7% accuracy** and **78% sensitivity** with the predictors of age, hypertension, and the average glucose level.

# Conclusions

**What are the most important factors contributing to the risk of having a stroke?**

Age is the most important factor. In patients under age 40, less than 1% of the patients observed had a stroke. However, above age 40, the odds of stroke increase nearly 8-fold, and over 65 years of age, the odds increase nearly 25-fold.

Average glucose level and hypertension were also important factors. Patients with hypertension were nearly twice as likely to have a stroke, and patients above the age of 40 with average glucose levels above 125 mg/dL were also about twice as likely to have a stroke.

# Recommendations

Given the preventability of strokes, hospitals should emphasize programs and outreach focusing on the highest risk groups: older age, high blood pressure, and high blood sugar.

The groups defined in this model can be used as a free pre-screening for stroke risk using patient data which is already widely collected. Doctors and nurses will have simple, clear guidelines for categorizing patients as having high stroke risk: **40+ years old, 125(mg/dL) average glucose level and/or hypertension.** These patients can then be recommended for further scans and/or treatment.

# Shortcomings

**Dataset Limitations:** This dataset only contains post-stroke data for stroke victims. To find factors which can more accurately predict strokes before they happen, survival analysis with time-to-event data would likely be more effective.

**Predictive Power:** Our model sacrificed some prediction accuracy to improve its interpretability. This tradeoff can be justified, but it might be useful to fit another, more flexible model if more accurate predictions become necessary.

**Imbalanced Data:** Our dataset has very few stroke patients compared to non-stroke patients.