Predicting Stroke

Project 2

By Steven Phillips

Problem: Predicting Stroke

The predictability of stroke amongst individuals was explored in this project.

Data Set:

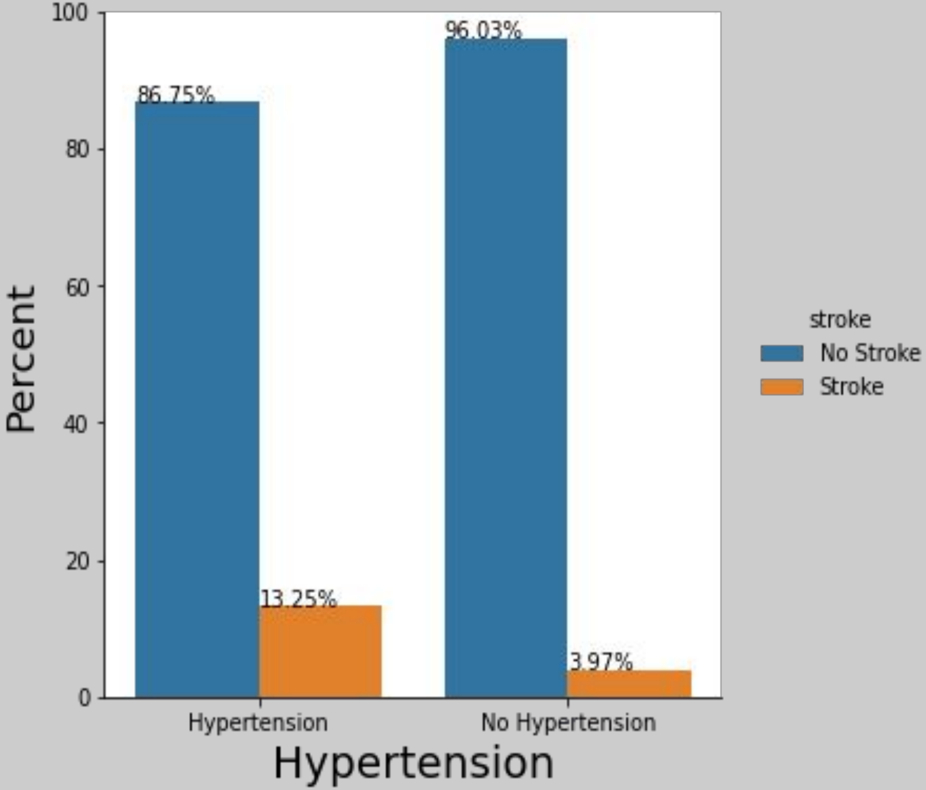
The data used for this project originates from:

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

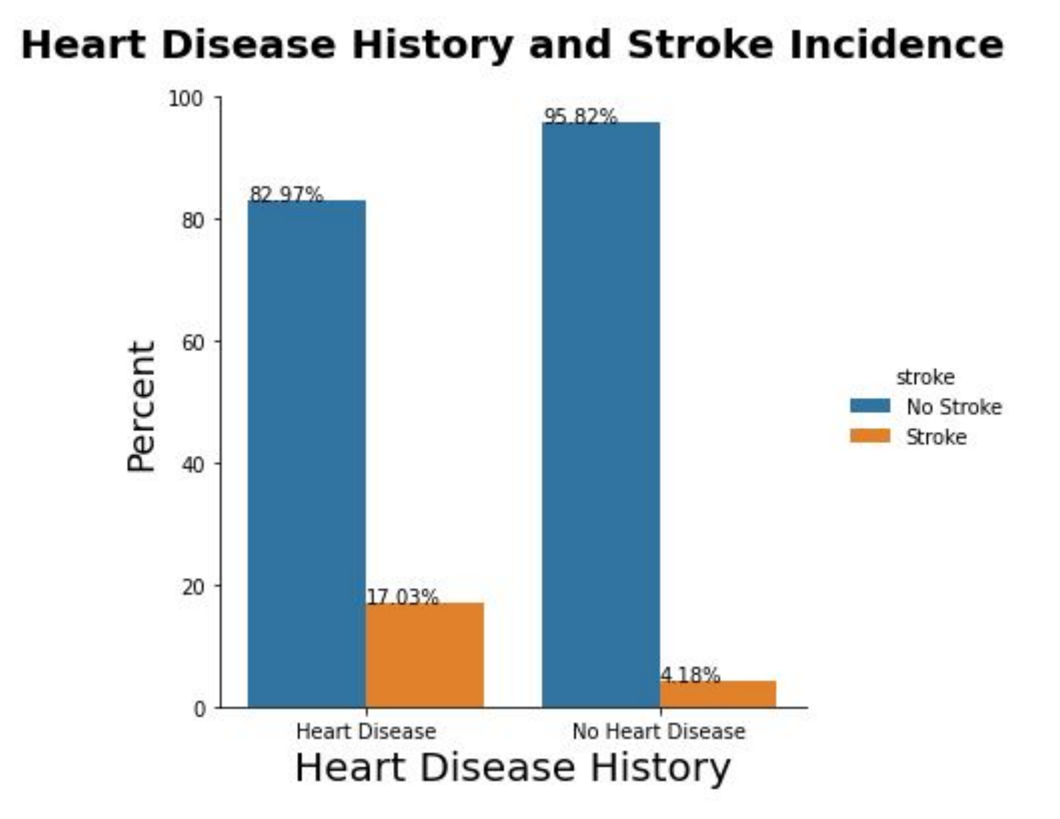
- The data set contained information from 5,110 individuals.
- Several features, including age, gender, marital history, smoking status, and type of work, were part of the data set.
- Other health indicators, including, whether or not the individual had a stroke were also available.

- More than three times the incidence of stroke for those individuals with Hypertension

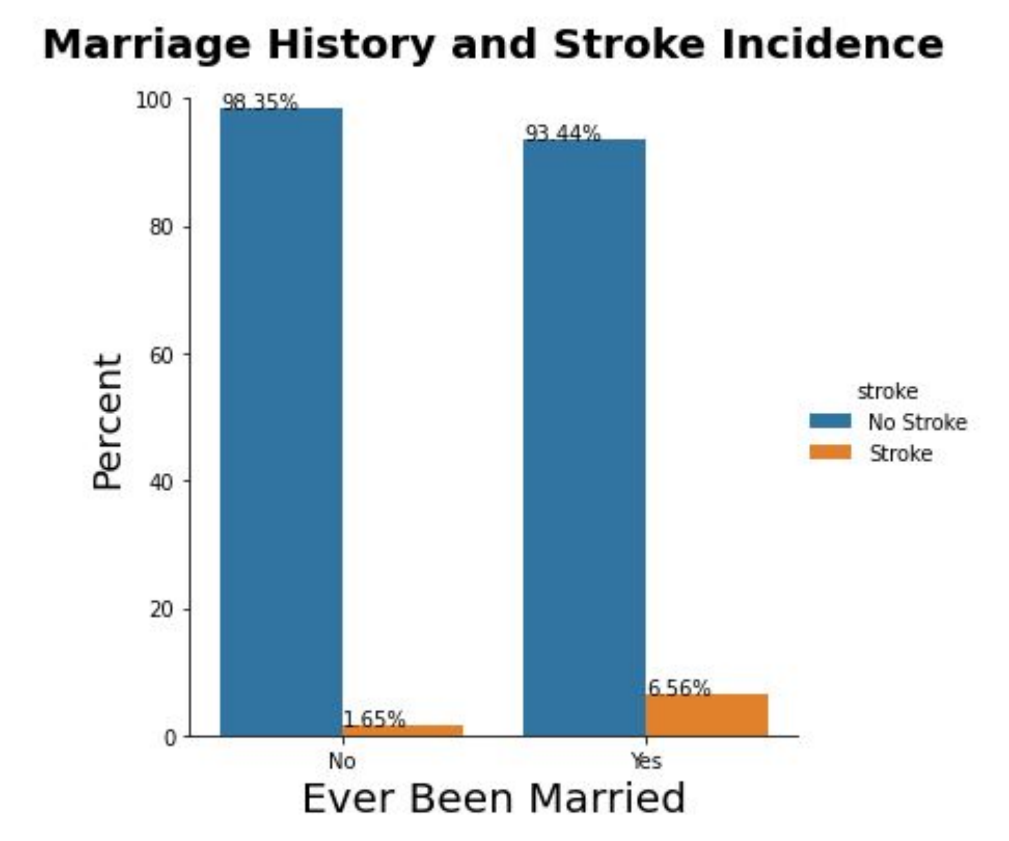
- Very low percentage of Stroke for those individuals without Hypertension



Hypertension History and Stroke Incidence



- More than four times the incidence of stroke for those individuals with Heart Disease
- Very low percentage of Stroke for those individuals without Heart Disease



- More than three times the incidence of stroke for those individuals who have ever been married

- Very low percentage of Stroke for those individuals without a Marital History

Maching Learning Predictive Models Explored:

This was a problem of classification, predicting whether or not an individual would have a stroke. These models were considered:

- **Bagged Tree Classifier**
- **Logistic Regression**
- **LightGBM Classifier**

Model 1: Bagged Tree Classifier

- Great at overall predictions

  .94 accuracy overall

- Where did it fall short?

  ~ Taking on the challenge of an imbalanced data set.

  (93.74% of the test data had the stroke condition.)

Model 2: Logistic Regression

- Great at overall predictions initially, but where did it fall short?

Also, taking on the challenge of an imbalanced data set

- How did it do after some tuning?

Accuracy suffered too much in comparison to the model ultimately chosen.

Model 3: Light GBM

- Great at overall predictions initially.

- Where did it fall short?

Initially, taking on the challenge of an imbalanced data set

- How did it improve after some adjustments?

Improved performance dealing with the imbalance and better accuracy than other comparable models

Model Chosen for Production: LightGBM

- **Reduced False Negatives: RECALL 0.85**

REDUCED PREDICTING NO STROKE FOR THOSE WITH STROKE RISK

- **Maintained decent overall predicting accuracy**

ACCURACY OF 0.68

- **Additional metrics: Precision - 0.14, F1-Score - 0.25**

Parameters of the model: Class weight - balanced, n_estimators - 50, num_leaves - 50, max_depth - 1

Final recommendations:

- The stroke rate for individuals with a marriage history is higher and also for those with hypertension and heart disease. Treatments for this population may be considered to reduce their incidence of stroke.

- Consider the LightGBM model for deployment in making decisions for the risk of stroke in individuals.