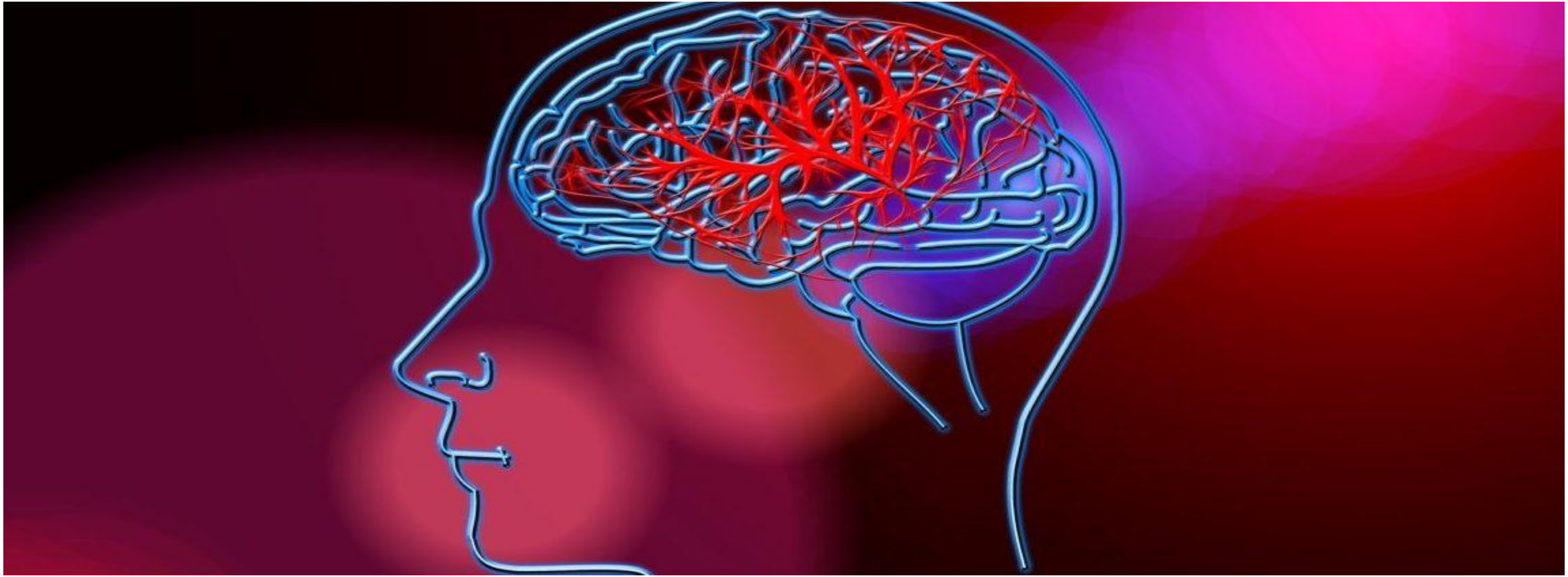


Stroke Predictions

Authors: Anisha Malhotra and Jaclyn Dwyer

Business Problem

GOAL:

Create a model for preliminary screening that can predict if a person is going to have a stroke.

WHY:

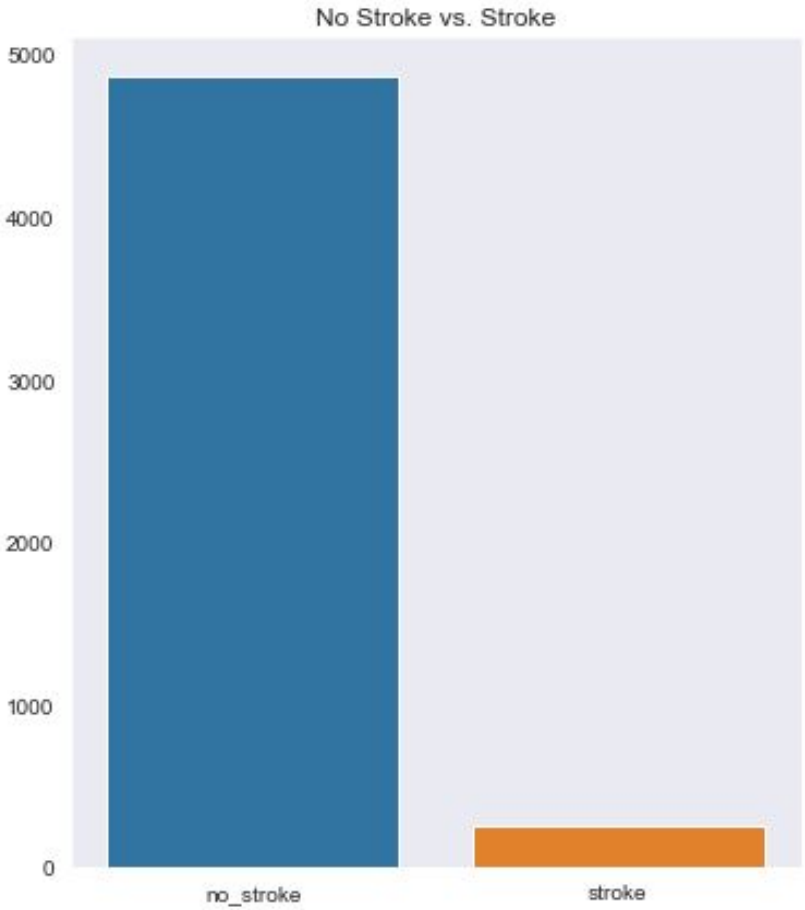
- Strokes are the 5th leading cause of death in the United States according to the CDC.
- This model would allow patients to take measures in order to prevent having a stroke

METRIC:

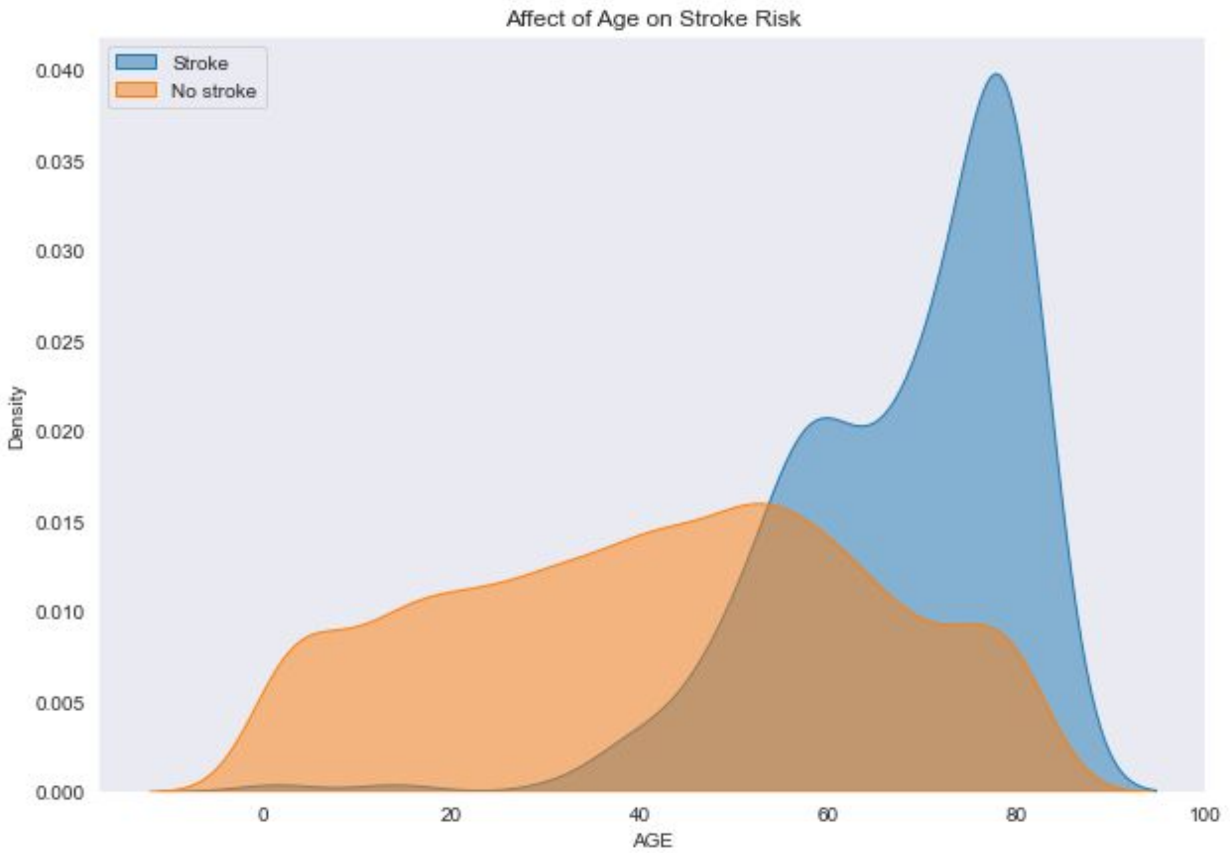
Recall Score - minimize false negatives

Data

- Obtained through Kaggle
- 5110 Rows & 12 Features
- Key Features:
  - Age
  - BMI
  - Average Glucose Level
  - Hypertension
  - Heart Disease
- Target Variable:
  - No Stroke vs Stroke
- Class Imbalance



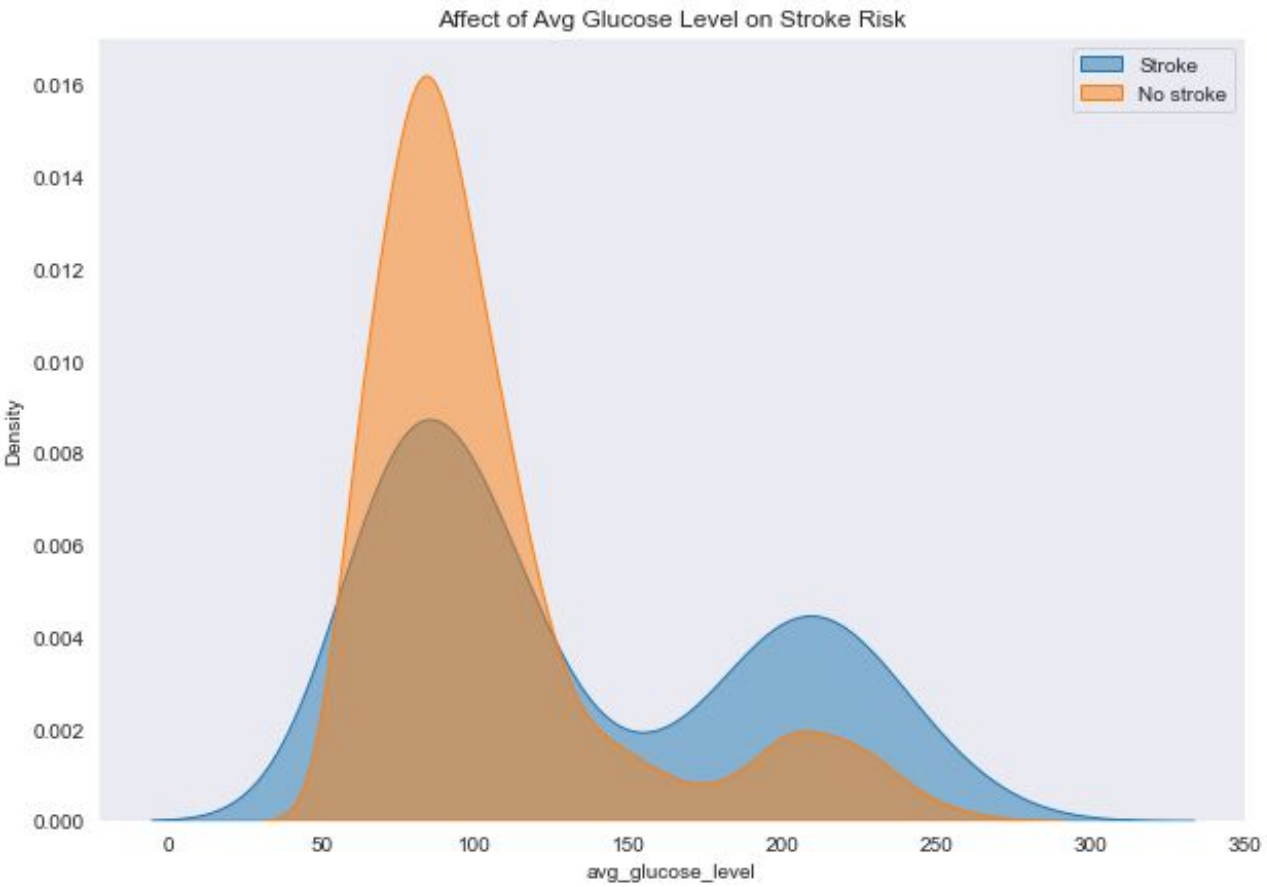
EDA: Age



Key Takeaway:

- Number of strokes increases as age increases

EDA: Average Glucose Levels



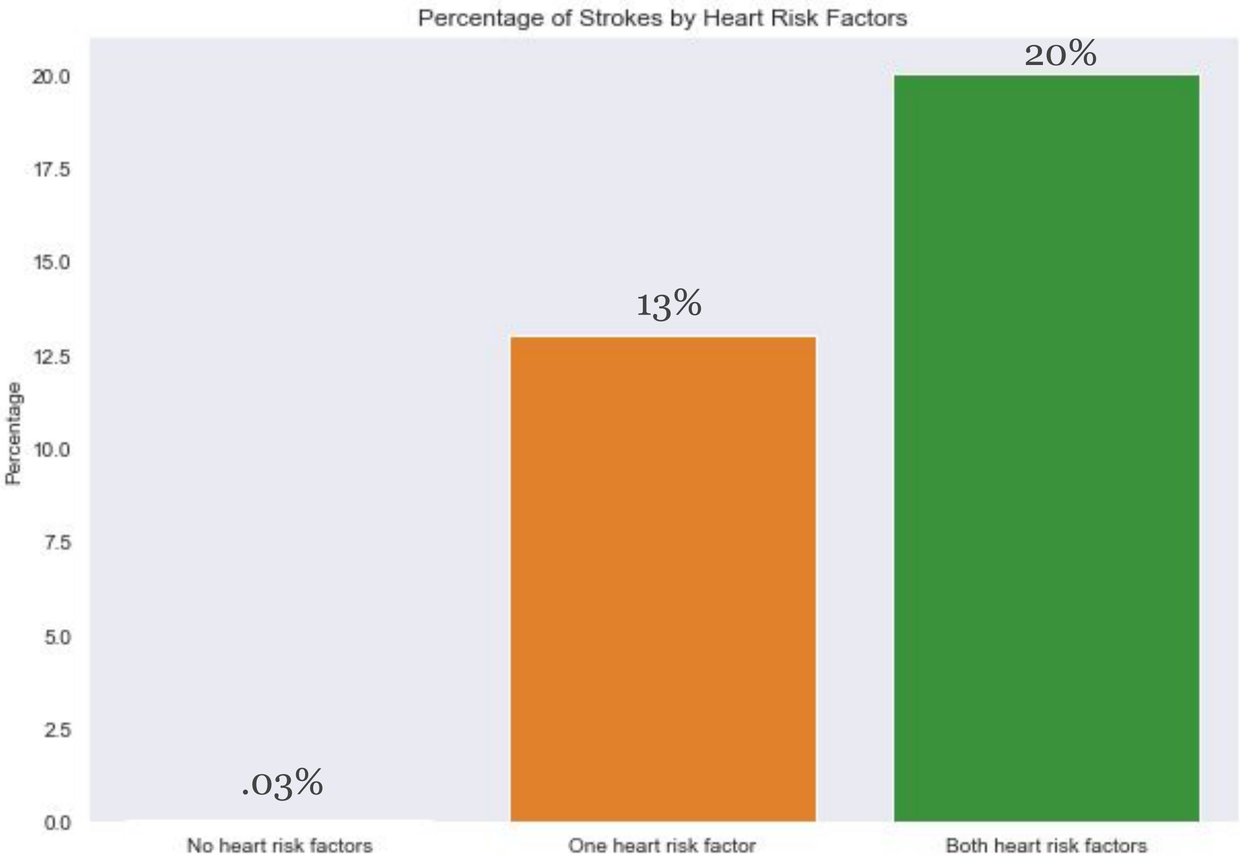
Glucose Levels:

- Normal < 140 mg/dl
- Pre Diabetic 140 - 200 mg/dl
- Diabetic > 200 mg/dl

Key Takeaway:

- At normal glucose levels, no strokes are more common
- At pre-diabetic and diabetic levels, strokes are more common

EDA: Heart Risk Factors



Heart Risk Factors:

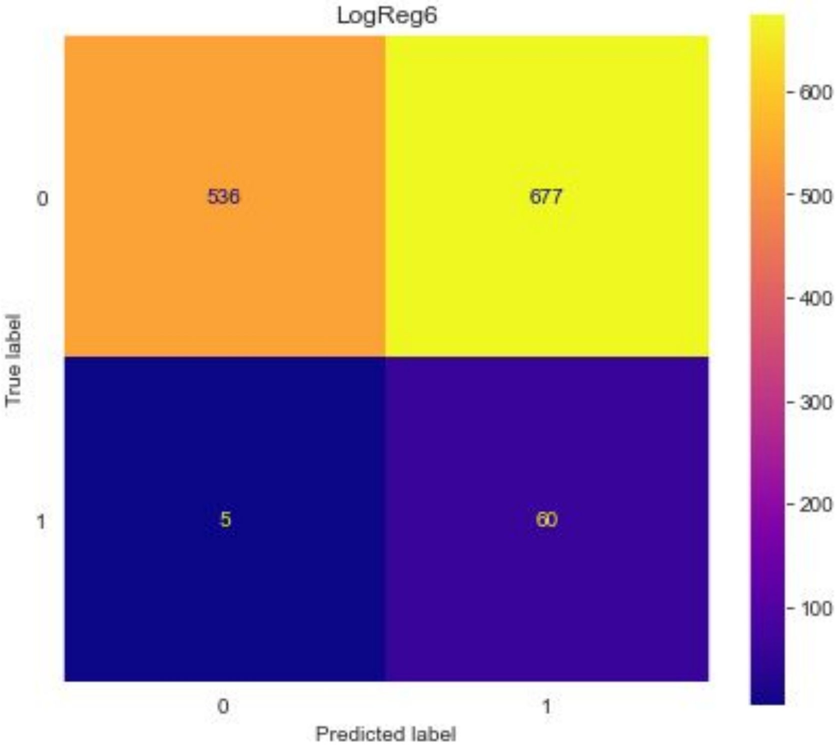
- **0** = No risk factors
- **1** = Either Hypertension or Heart Disease
- **2** = Both Hypertension and Heart Disease

Key Takeaways:

- The more heart risk factors, the higher percentage of strokes

Model Evaluations

- Logistic regression models had higher recall scores for train and test data
- Final Model:
  - Logistic Regression
  - 92% of strokes caught
- Note:
  - 56% false positive
  - Further screening would prove these patients to be healthy



Next Steps:

- Run more GridSearches on models to determine if a model with an even higher recall score exists
- Feature engineer from outside sources to implement risk factors
- Run model on more unseen data