Stroke Case Prediction Final Report

Capstone Project 3

Adam Reichenbach

June 2024 Cohort
Springboard Data Science Career Track

Introduction

According to the World Health Organization (WHO), stroke is the 3rd leading cause of death in 2021 globally (2nd in low-income, 3rd in lower-middle-income, 1st in upper-middle income, 3rd in high-income countries).

We are going to use collected data to make a machine-learning algorithm that can classify how susceptible someone is to having a stroke based on different variables (age, sex, smoking status, glucose level, BMI, heart disease status, hypertension status, etc).

Criteria for Success and Solution Space

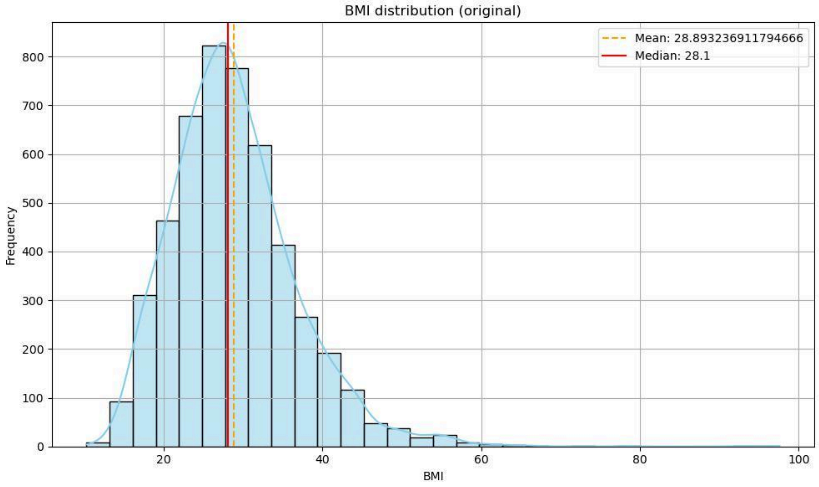
A classification model that can predict if someone will have a stroke or not with the key metric being **recall** for positive stroke cases since it is critical in a medical setting that we reduce the amount of false negatives detected. We will also look at the AUC-ROC scores. Both metrics will be compared across the models using predictors from the training and testing data.

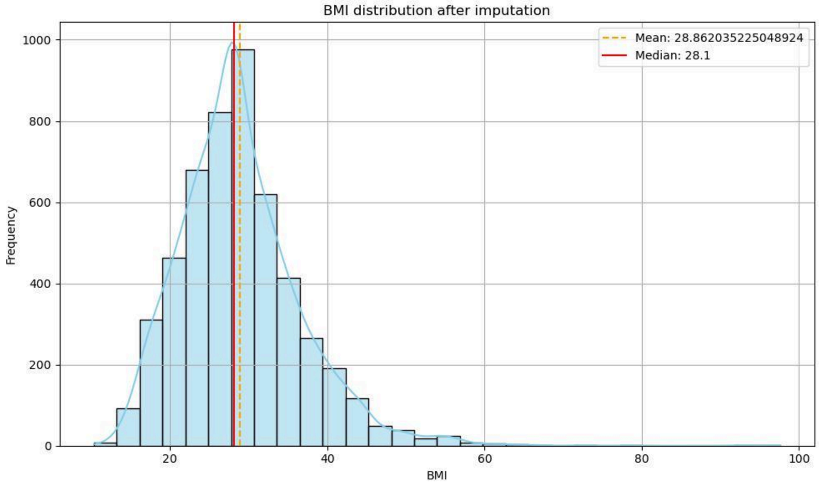
Stakeholders

- Hospitals and healthcare providers since the model can be used for early intervention and helps doctors give more correct diagnoses as well as give better preventative care for their patients.
- Med software & tech companies can use a model that they can integrate into their systems that could help HCPs and patients alike.
- Public health agencies who want to help to reduce strokes to the population as a whole and help distribute resources to more at-risk people accordingly.
- Health insurers who want to reduce the amount of expensive claims they have to pay out to stroke victims and encourage more preventative care.
- Medical researchers can use the model to help uncover other patterns not seen by anyone else.

Data Wrangling

- The data wrangling (and EDA) process can be seen here.
- The data for the project was available as a downloadable CSV file from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. The data provided was already mostly cleaned and contained 201 missing values from only the BMI feature. This was fixed by imputing the missing data with the median, it still kept the original mean and median BMI value the same.





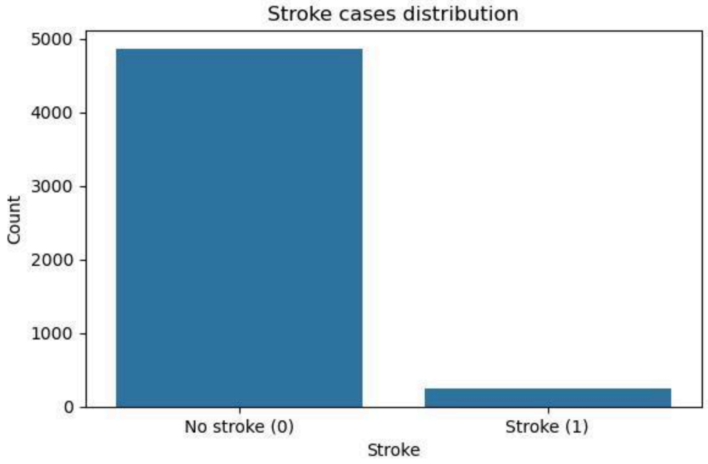
- After verifying that there weren't any more missing values and seeing that there are 5110 unique values of the patient **id** feature (same amount as the number of rows of the dataset), and double checking for any duplicate rows, it was confirmed that all observations were unique and that we could remover the **id** feature from the final dataframe for use.
- The final dataframe for use contains 5110 rows and 11 features and is ready for use in the EDA step.

Exploratory Data Analysis (EDA)

Features

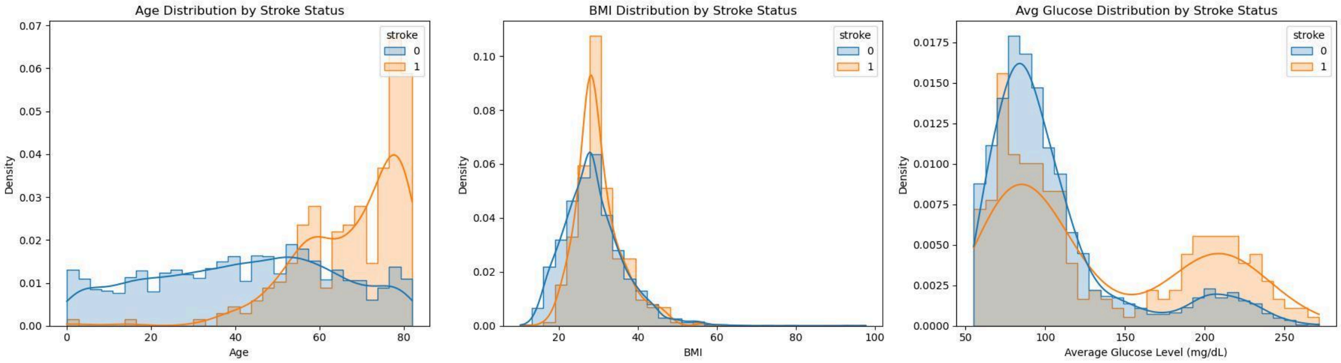
- The features of the dataset are split into 3 types:
  - **Categorical** (non-binary):
    - *Gender*: identified gender of the patient (male, female, other)
    - *Work_type*: employment type of the patient ('Children' meaning the patient is a child and therefore isn't employed)
    - *Smoking_status*: Recorded status of the patient's smoking history
  - **Categorical** (binary):
    - *Hypertension*: 0 = no hypertension, 1 = has hypertension
    - *Heart_disease*: 0 = no heart disease, 1 = has heart disease
    - *Residence_type*: patient either is from a *rural* or *urban* area.
    - *Stroke* (**target variable**): 0 = no stroke, 1 = stroke
  - **Continuous**
    - *Age*: patient's age in years (years recorded as a decimal that are less than 1 indicate the patient is an infant. Ex: 0.5 years = 6 months)
    - *Avg_glucose_level*: Average glucose level recorded from the patient. (units = mg/dL)
    - *Bmi*: Recorded Body Mass Index of the patient.

Stroke Case Distributions



- There are significantly more non-stroke cases than there are actual stroke cases. Only around 4.87% of cases recorded are positive for stroke. This would be dealt with in the preprocessing stage with SMOTE.

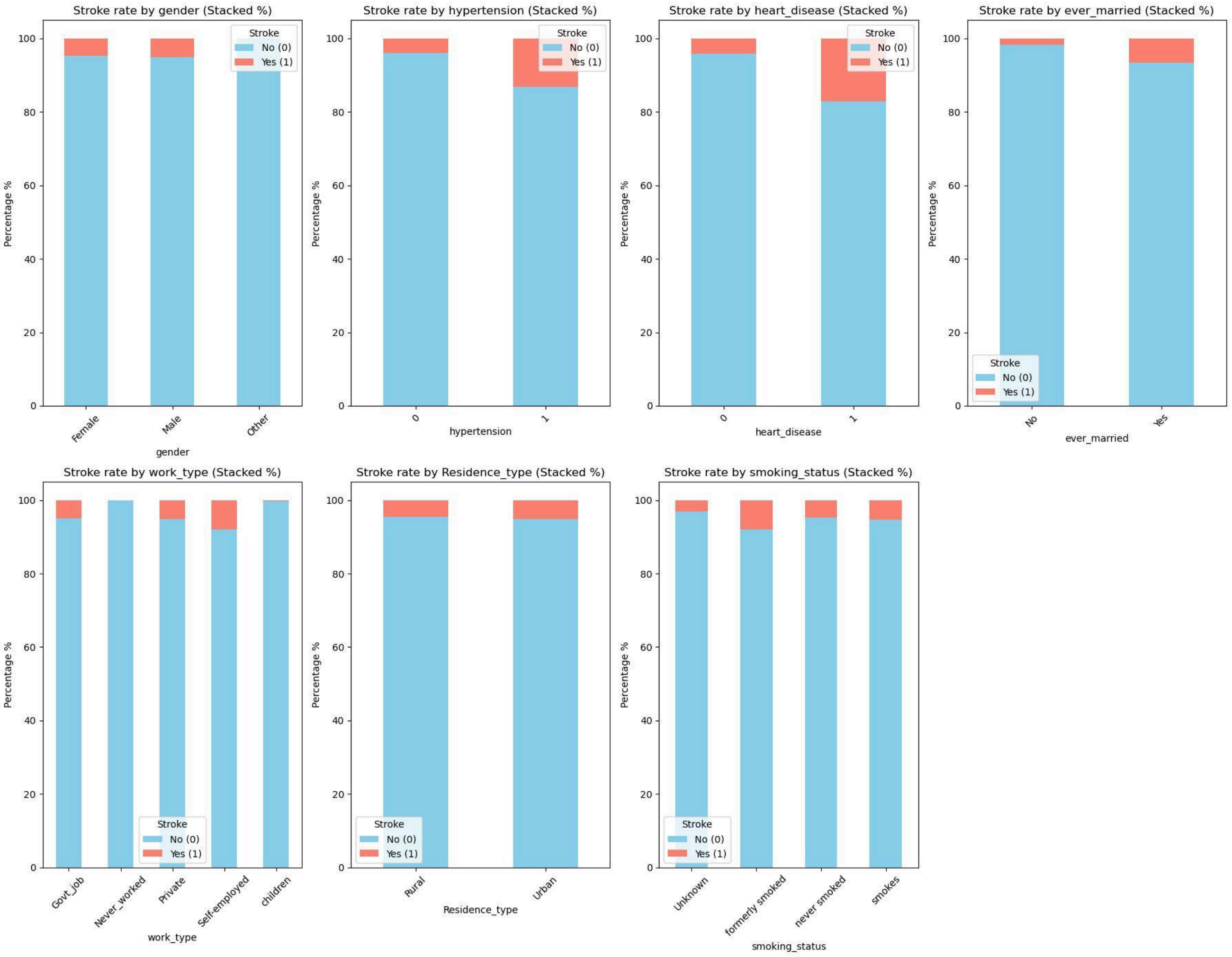
Stroke Case Distributions (continuous variables)



- Starting with the **continuous variables**, the positive stroke cases tend to skew more towards the older patients as well as patients with glucose levels above 150 mg/dL. **Age** and **glucose levels** seem to be strong candidates as driving features for stroke cases.

- Density distribution of positive and negative stroke cases are very similar with some high overlap in the **BMI** plot.

Stroke Case Distributions (categorical variables)

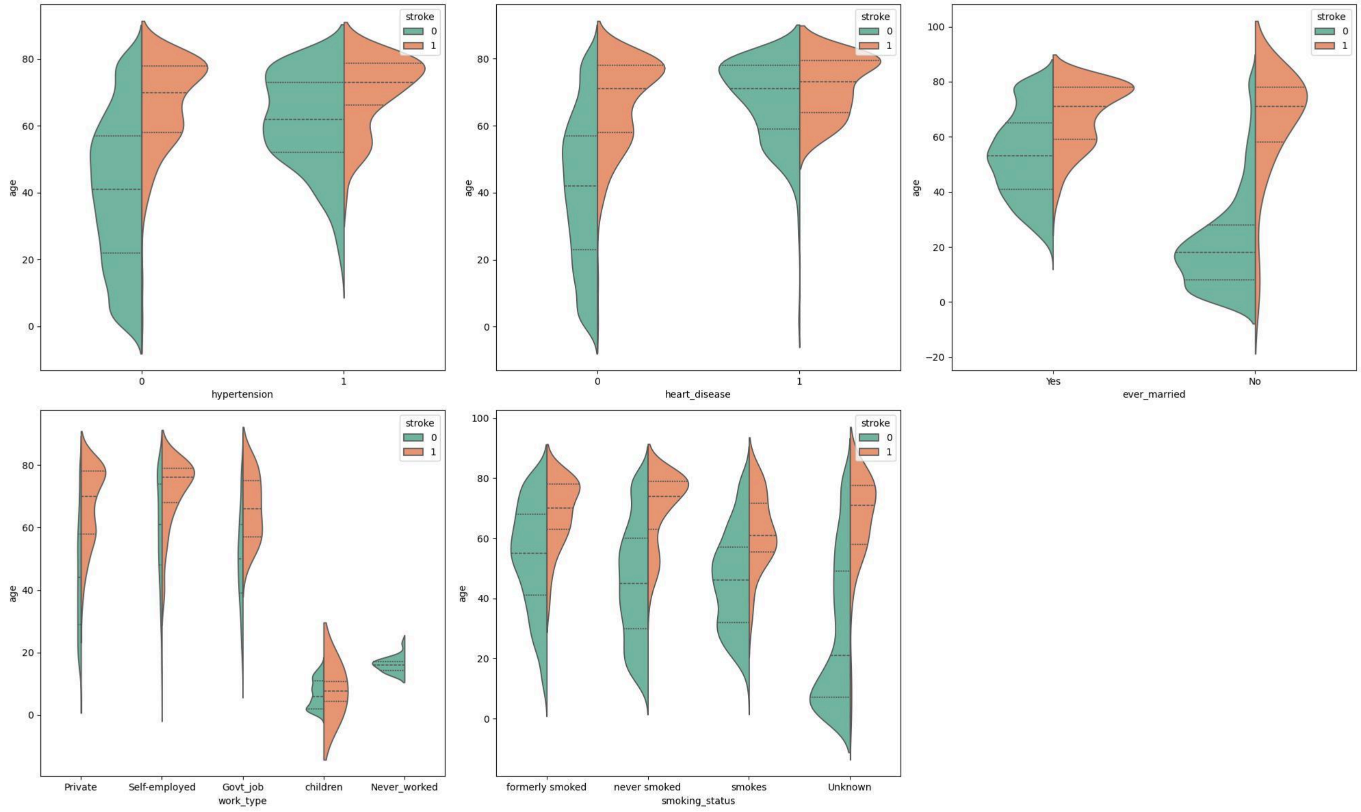


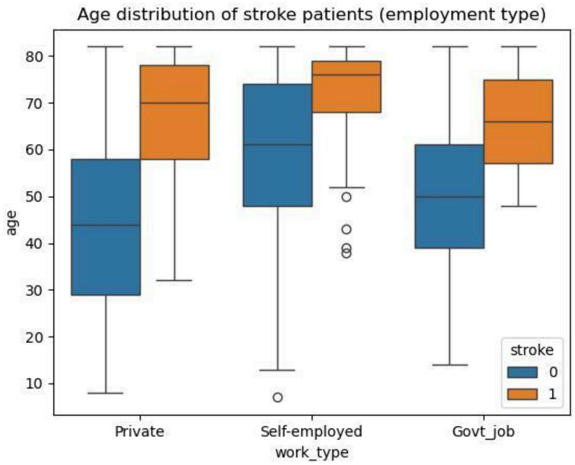
- Plotting the stroke cases percentages by categorical variables, it doesn't come as a surprise that there are higher percentages of positive stroke cases for patients who have recorded cases of **hypertension** or a **heart disease**.

- What stands out are the higher percentages of positive stroke cases for those who **have been married**, are **self employed**, and **formerly smoked**.

- It is quite possible that **age** could be a confounding variable that drives the positive stroke cases up for the previously listed.

Age as a confounding variable





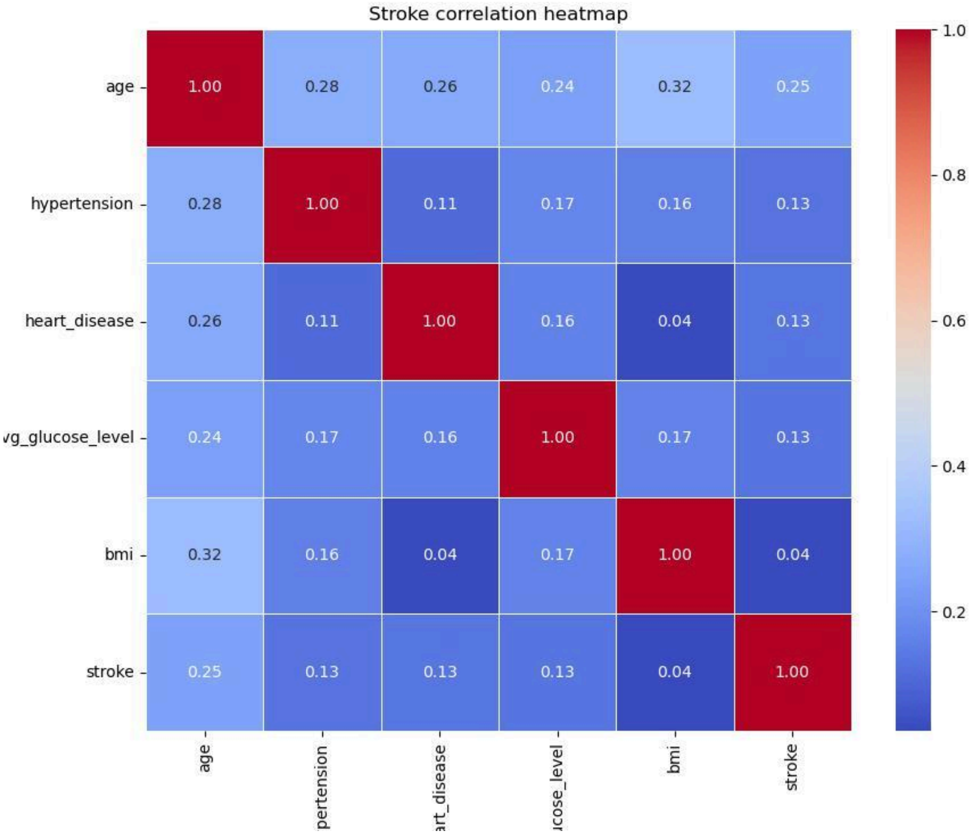
- Plotting violin plots reinforces the idea that **age** may be the confounding variable that explains the higher positive stroke case percentages in some of the groups of the categorical variables.

  - There's a higher density of older patients that have been married at one point as well as having a positive case for a heart disease or hypertension.

  - The positive stroke cases also seem to be more dense with older patients in every category.

  - I made the box plots because the density of the negative stroke cases weren't displaying properly for the violin plot for the 3 groups shown in the employment type category.

Correlation Heatmap



- Once again, **age** looks to be one of the strongest candidates as the leading variable for stroke susceptibility as it has the highest correlation with stroke. **Hypertension, glucose,** and **heart disease** look to be the next strongest with 0.13 each.
- **BMI** looks to have a weaker, near-zero correlation, but doesn't rule out non-linear effects.

EDA takeaways

- Positive stroke cases are more common among patients who are older, have higher glucose levels, and test positive for a heart disease and/or hypertension.
- BMI is less conclusive as a driving variable.
- Age also looks to influence the amount of positive strokes for features like **ever_married** and **age**.

Preprocessing the Data and Modeling

Preprocessing the data and modeling process can be seen here

Preprocessing

- The continuous variables **age**, **avg_glucose_level**, and **bmi** were scaled using StandardScaler()

- Categorical features represented by a non-numerical value were one-hot encoded.

- Binary numerical features like **hypertension**, **heart_disease**, and **stroke** were already represented as 0s and 1s and didn't need to be encoded.

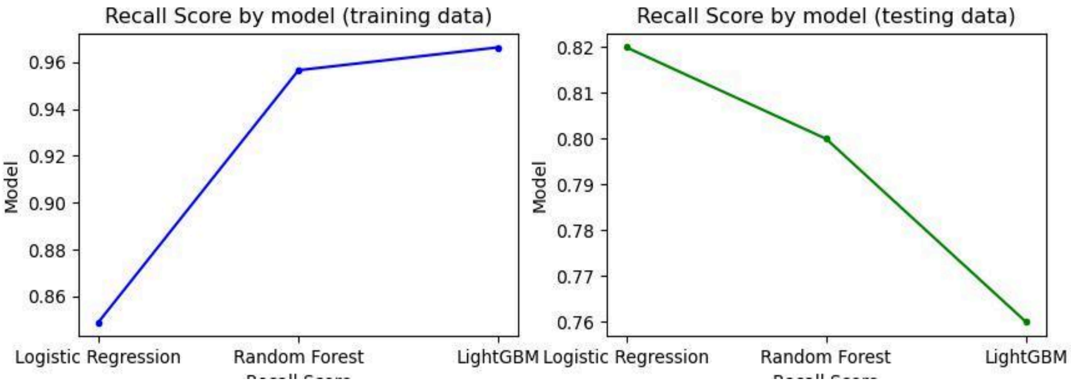
- The data was train-test-split 80/20, then SMOTE applied to handle the target variable class imbalance.

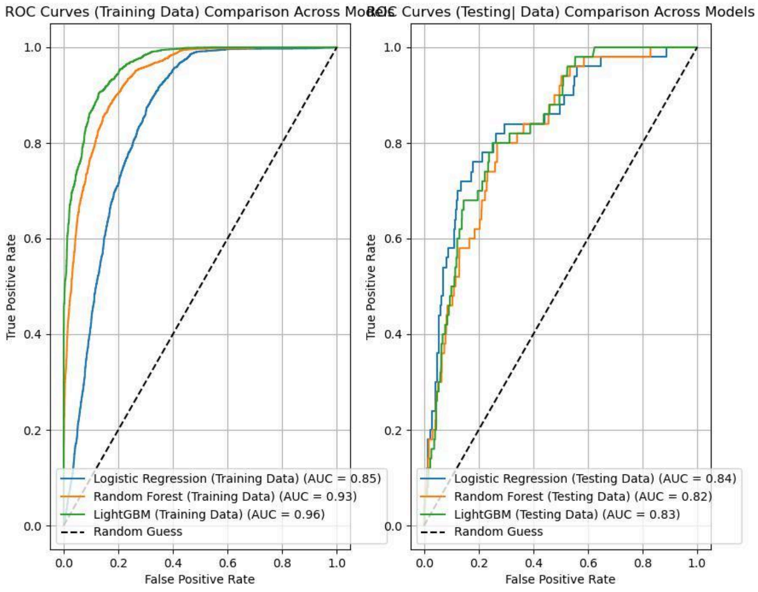
Modeling

The leading metric for model decision is **recall** as the metric is best for scoring how well a model can reduce false negatives. In a medical setting this is critical since we want to reduce the risk of false negative stroke cases predicted.

The 3 models to choose from were:

Logistic Regression, Random Forest, and **Light Gradient Boosting**





Looking at the plots of the recall and ROC-AUC, all 3 models do perform well. However, the ROC-AUC and recall scores for the Random Forest and LightGBM models from the training data are quite higher than the Logistic Regression model's. This suggests that they are more prone to overfitting.

Logistic Regression model with parameters: **{'C': 0.01, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'}** is the model of choice as it has a high/consistent recall, ROC curve, and ROC AUC score from both the training and testing data.

Final Thoughts

- The original data source for this project should have also included other possible features like family history of strokes, medications being taken, blood pressure, cholesterol levels, sleep quality, etc. These could also be other factors that could help build a more accurate model.
- There was a heavy imbalance of stroke vs. non-stroke cases with a lot more non-stroke cases which can increase the difficulty of constructing a good model. Techniques like SMOTE helped with this problem for this project, but original sampling should have more balance of the target variable.
- Age and glucose levels look to be the driving factors from the data gathered for increased risk of positive stroke cases, but other medical factors not featured that I listed above should be included in future data gathering.
- Other classification models should also be tried out in the future. Ones that are less prone to overfitting and are great with dealing with high class imbalance.
- The 201 observations that had missing BMI values should have been explored a little bit more. Maybe there was a reason those values were missing.