Project 2 - Part 5 (Core)

Erica Kitano

1. Stakeholder and Business Problem
2. Data
3. Visual Analysis
4. Model Recommendation
5. Strengths and limitations of the model

Stakeholder and Business Problem

Background:

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

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

Business problem:

Help doctors predict whether a patient is likely to get stroke based on parameters such as gender, age, various diseases, and smoking status.

Data Source

The original source of the data used is Stroke Prediction Dataset from Kaggle.

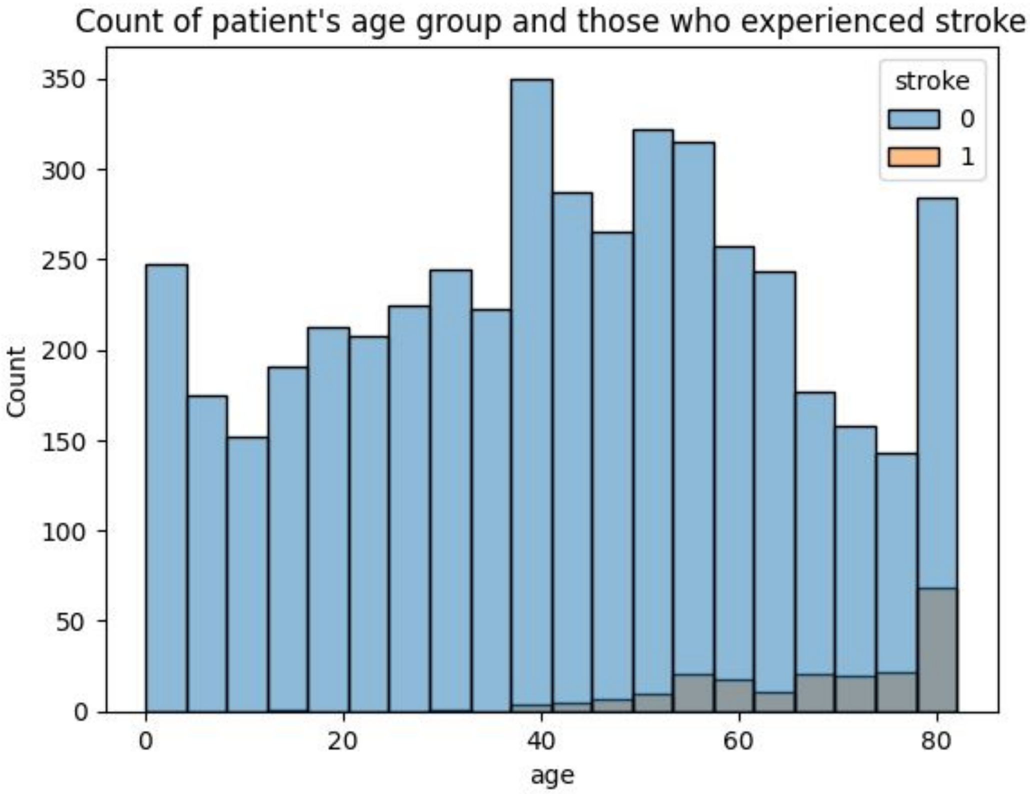
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Description of Data

Attribute Information

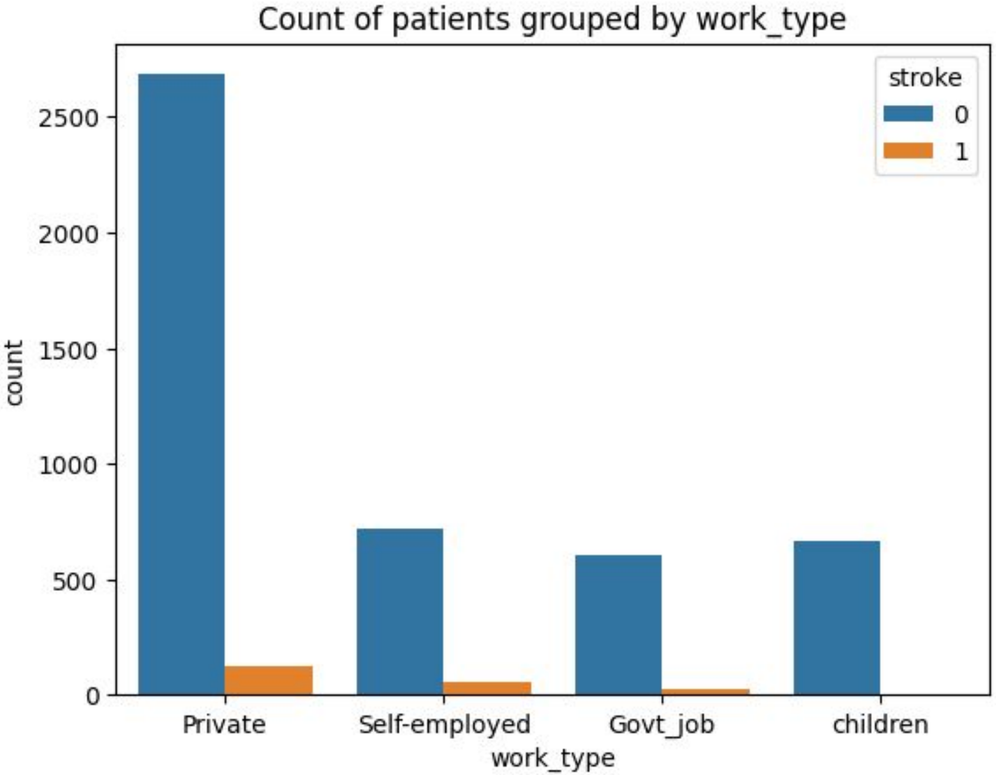
1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not
*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Visual Analysis #1



There is a trend that stroke is observed in people of older age.

Visual Analysis #2



Strokes are not common in children.

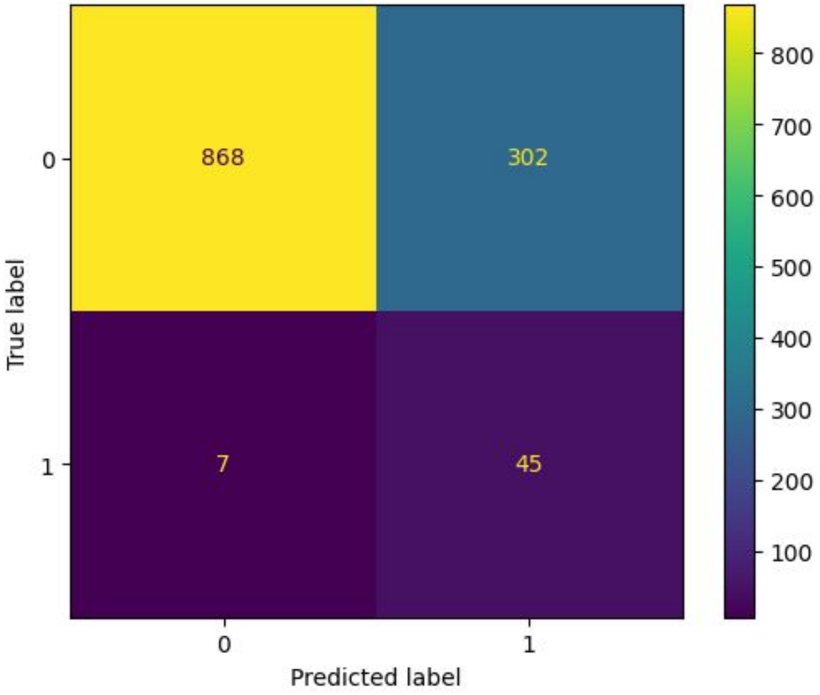
Model Recommendation

- Model Name: Tuned Logistic Regression (Balanced)
    - Parameters: C=0.01, penalty = l1, class_weight='balanced'
    - Recall score: 0.865
    - Accuracy score: 0.747

This model is recommended because it has a balance of relatively high recall score and accuracy score. Having a high recall score is important because Type 2 errors (False Negatives) are more costlier for this dataset than Type 1 errors.

Model Name	Accuracy	F1-Score	Precision	Recall
Tuned Logistic Regression (Balanced)	**0.747136**	**0.225564**	**0.129683**	**0.865385**

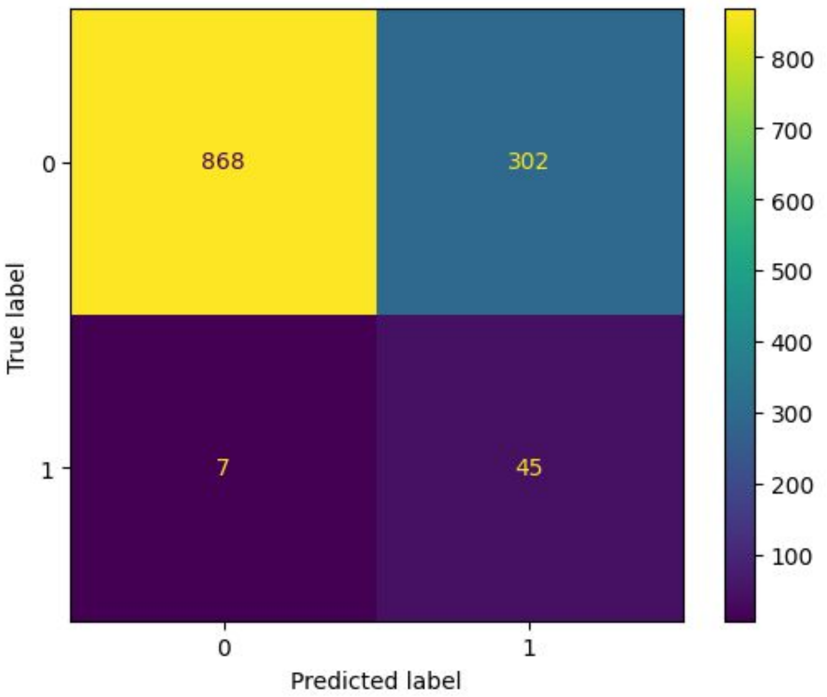
Strengths of the model



- True Negative - 868 (predicted no stroke and actually no stroke)
- False Positive - 302 (predicted stroke, but actually no stroke)
- False Negative - 7 (predicted no stroke, but actually stroke)
- True Positive - 45 (predicted stroke and actually stroke)

This model predicted 45 positive cases correctly on the test data, and only had 7 false negative cases on the test data.

Limitations of the model



- True Negative - 868 (predicted no stroke and actually no stroke)
- False Positive - 302 (predicted stroke, but actually no stroke)
- False Negative - 7 (predicted no stroke, but actually stroke)
- True Positive - 45 (predicted stroke and actually stroke)

The number of false positives is 302 and very high. For this set of data, it is difficult to find a model that has both low Type 1 and low Type 2 errors. This may be explained from the observation that the features do not correlate well with the target. However, false negatives are costlier for this dataset, therefore, having a larger false positives is considered better than having a large false negatives for this problem.