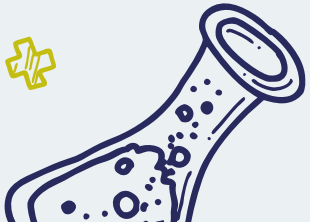
Stroke Prediction

Predicting whether a patient is likely to have a stroke or not



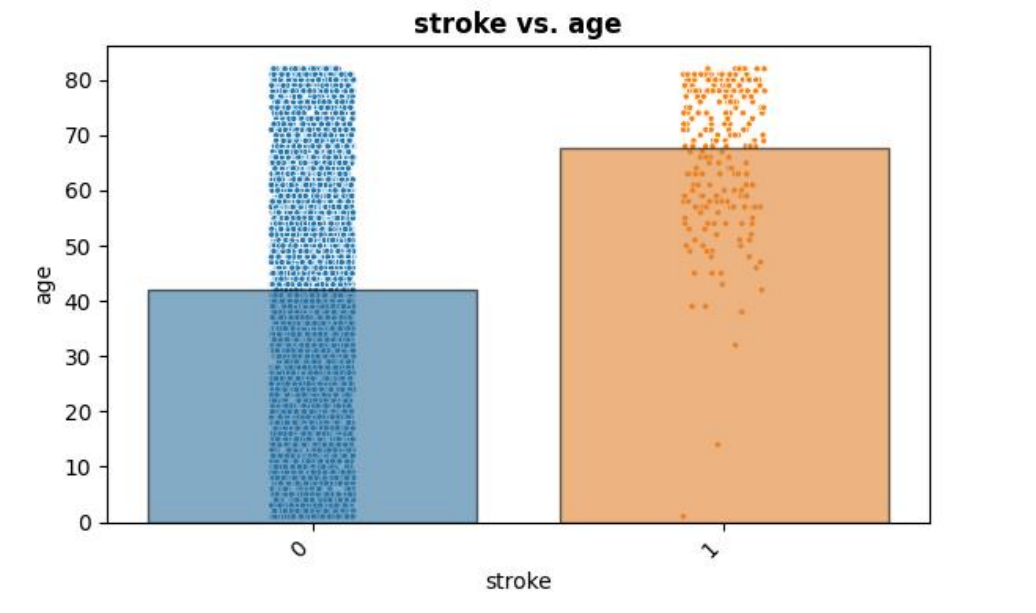
Business Problem:

A hospitals' board of directors would like to predict whether a patient is likely to have a stroke or not based on demographic and health–related data.

Dataset:

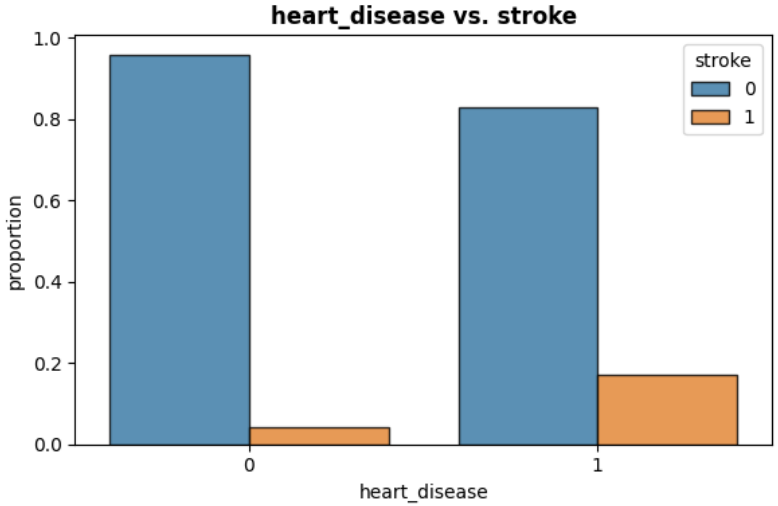
- Data sourced from WHO
- 5110 rows, 11 columns
- Includes patient demographic data such as age, gender, marital status, work type, and residence type.
- Also includes health-related data such as patient glucose levels, BMI, smoking status, heart disease and hypertension.

Stroke vs Age



- The plot above shows the average age of patients that had a stroke is higher than those that did not have a stroke
- The average age of patients that did not have a stroke is around 40 years, but there's a large number of patients older than 40 who has not had a stroke.
- There’s a few outliers where younger patients had strokes

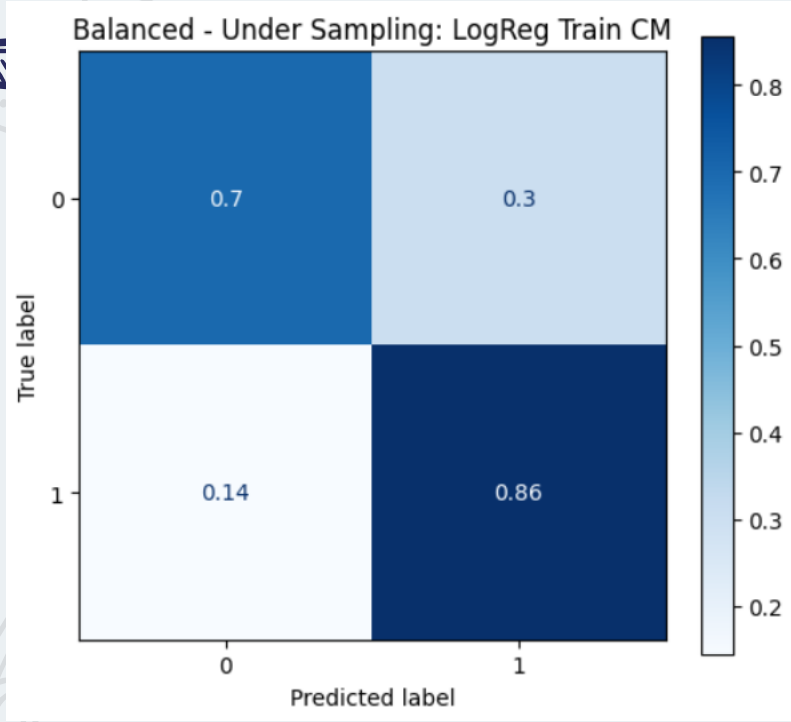
Heart Disease vs Stroke

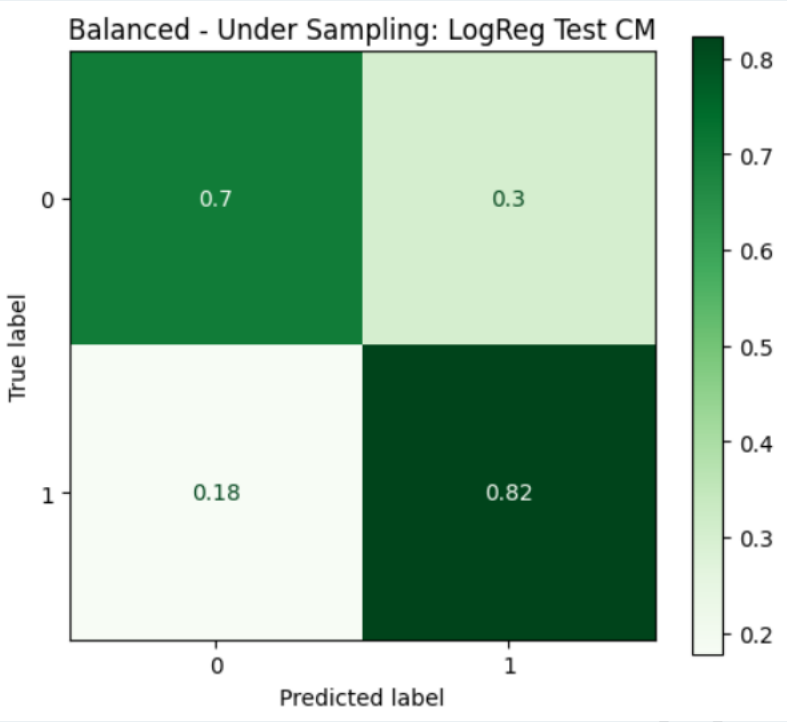


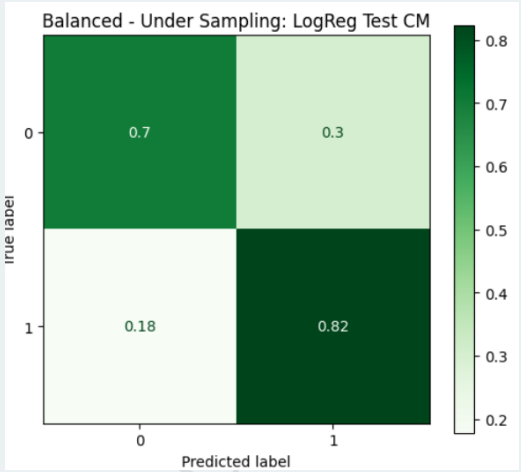
	heart_disease	stroke	proportion
0	0	0	0.958204
1	0	1	0.041796
2	1	0	0.829710
3	1	1	0.170290

- **The proportional count plot shows that 4,2% of people that do not have a heart disease are likely to have a stroke**
- **This percentage increases to 17% for people that have a heart disease**

Model Results







Model Results

- Test scores and train scores are similar indicating the model is not over fit
- The dataset was unbalanced and various techniques of balancing were attempted. Under sampling achieved the best results.
- Various models were trained an tested. The logistic regression model resulted in the best results.
- The model predicted 82% of patients that had a stroke correct, but failed to predict the remaining 18%.
- The model also incorrectly predicted 30% of patients that did not have a stroke as having a stroke.

Balanced - Under Sampling: LogReg Train

	precision	recall	f1-score	support
0	0.99	0.70	0.82	3644
1	0.13	0.86	0.22	187
accuracy			0.71	3831
macro avg	0.56	0.78	0.52	3831
weighted avg	0.95	0.71	0.79	3831

Balanced - Under Sampling: LogReg Test

	precision	recall	f1-score	support
0	0.99	0.70	0.82	1216
1	0.12	0.82	0.21	62
accuracy			0.71	1278
macro avg	0.55	0.76	0.52	1278
weighted avg	0.95	0.71	0.79	1278

Final Recommendations

For this task, the priority was to reduce the type 2 error of the model as it is more costly to predict someone not having a stroke when they actually do have a stroke

As a result, the type 1 error was sacrificed. A high type 1 error could also be harmful to the patient. This means that we could be misdiagnosing a healthy patient and giving them medication they do not need.

To further reduce type 1 and type 2 errors, features with better correlations to the target should be sourced and used for modeling.

Thank You!