Predicting Strokes from Medical and Lifestyle data

Lukas Kania

DATA 1030 Fall 2021: Andras Zsom

Brown University – December 8th, 2021

https://github.com/lkania1/data1030project

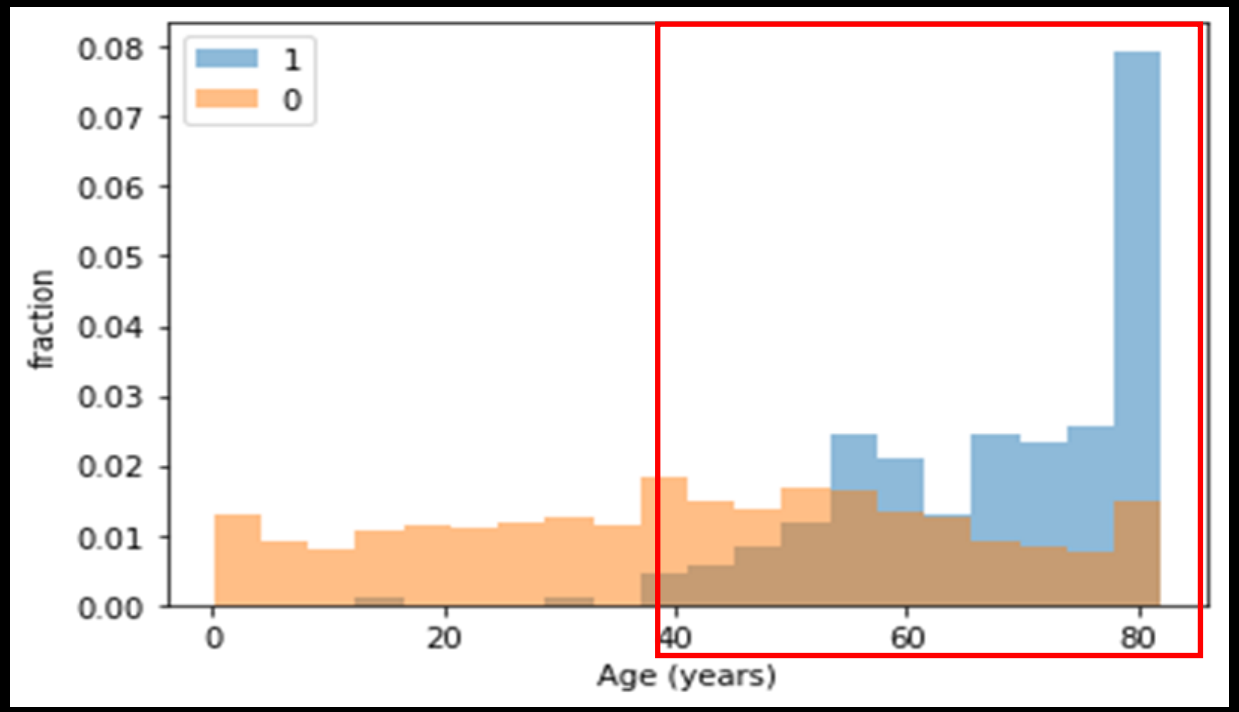
Context

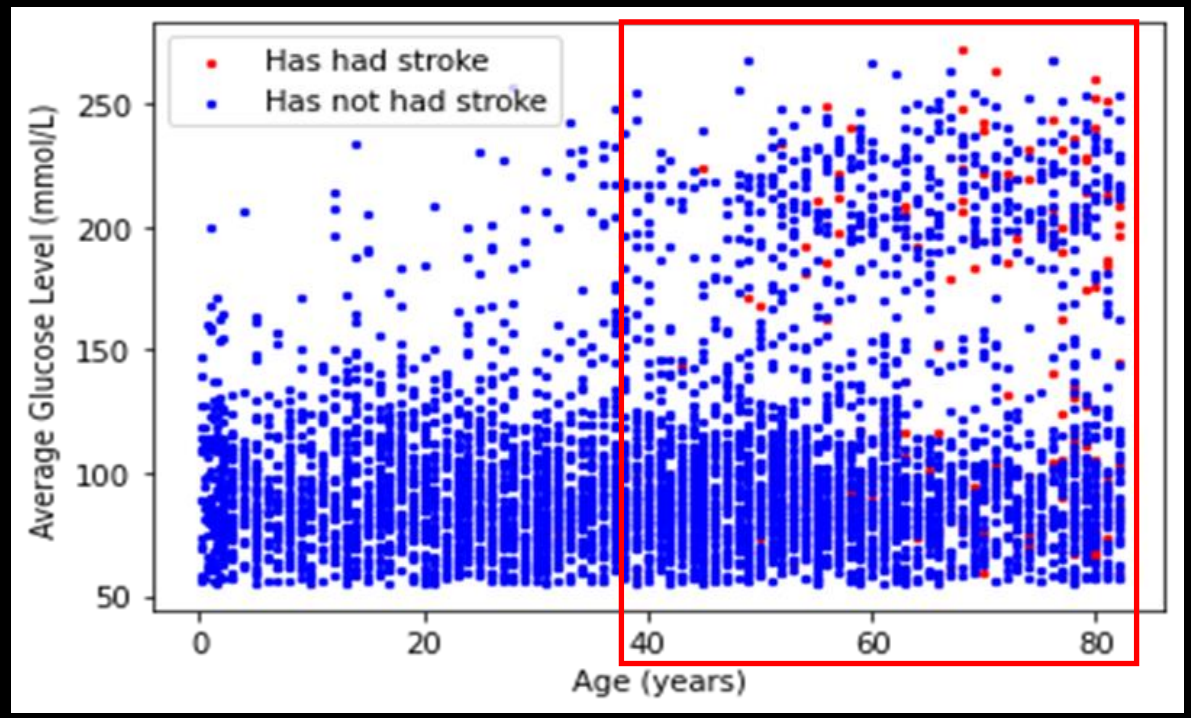
- Classification Problem: predicting “stroke” (class 1) and “no stroke” (class 0) outcomes of patients to help predict future patients’ outcomes.
- World Health Organization states that Stroke is the second highest cause of death in the world (11% of deaths).
- Dataset comprised of 5,110 patients, with 11 different feature columns

Kaggle Data set

EDA and Preprocessing:

- 201 patients had missing data for BMI. These patients were dropped as to not impute health data, which could impact a person's livelihood (4.26%).
- Dataset is heavily unbalanced (class 1: 209 patients, class 2: 4901 patients)
  - 60/20/20 training, testing, validating split utilizing Stratified K Fold with 4 folds.
- 11 features originally with 22 after preprocessing:
  - Age and average glucose level show promise as features for prediction of stroke.





Pipelines and Models:

Model	Logistic Regression	Random Forest	SVC	XGBoost
Reasoning	Processed features are binary, thus classification with Ridge regularization should work well.	Processed features are primarily binary, which work well for Random Forest Classifiers	Small number of features overall, many of which seem to be unrelated.	Powerful tree algorithm that routinely outperforms RFC and Logistic regression.
Tuning	Alpha (logspace)	Max depth (linspace) Max features (linspace)	Alpha (logspace) Gamma (logspace)	Learning rate (linspace) Column sample by tree (linspace) Sub sample (linspace)

Evaluation Metric: F1 score, with the goal of maximizing true positive and minimizing false positives.

Results: Cross Validation Scores

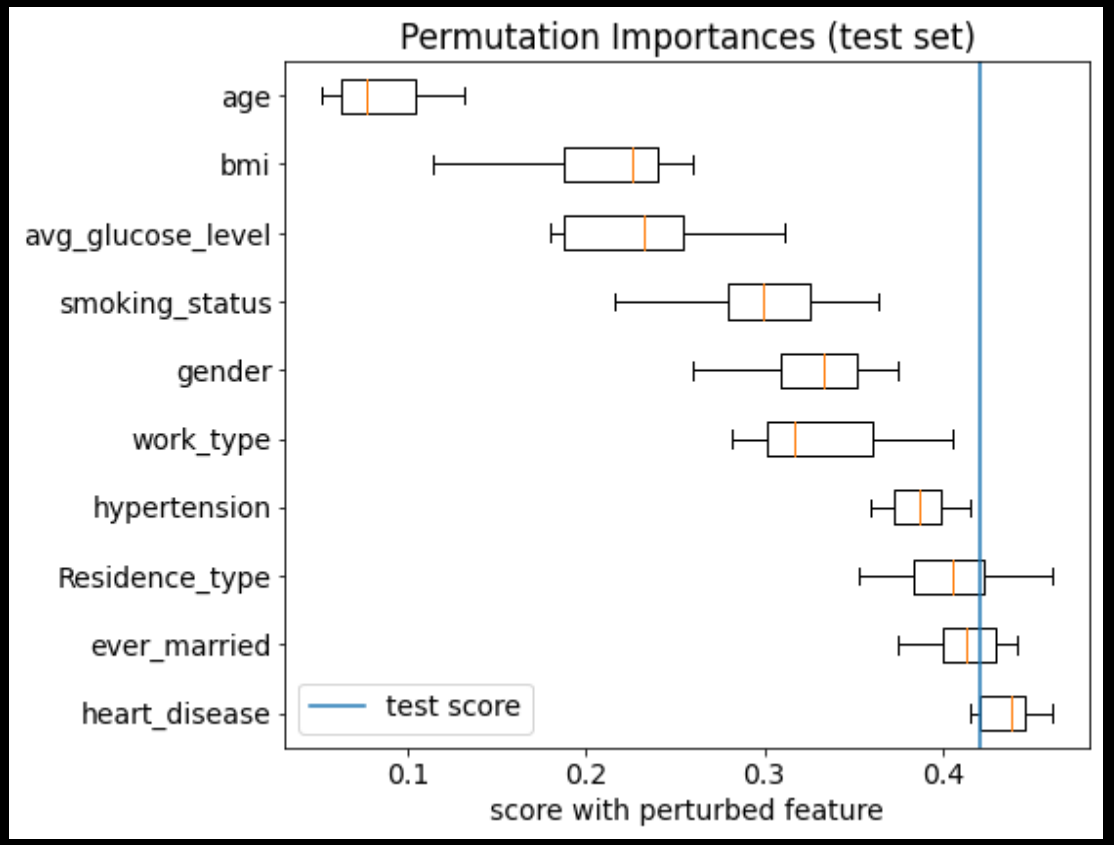
Model	Baseline	Logistic Regression	Random Forest	SVC	XGBoost
F1 Score	8.17%	1.152%	3.0303%	8.139%	12.163%
Percent Change from baseline	0.0%	-85.90%	-62.9%	-0.38%	48.87%
F1 Standard Deviation	-	0.01996	0.03329	0.03906	0.04916
Number of Std. Devs. From baseline	-	-3.5148	-0.007206	-1.5432	0.8127

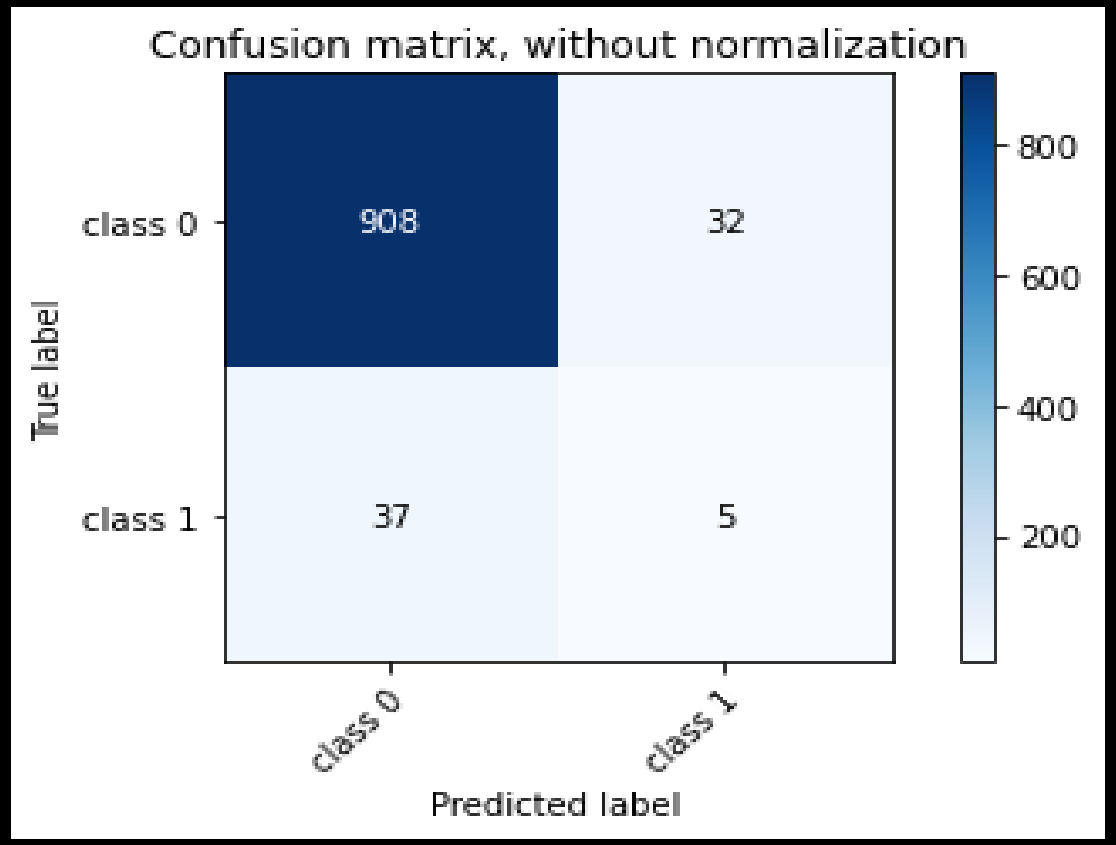
Best Model: XGBoost performs best (highest number of standard deviations And positive percent change)

Parameters:

- Learning Rate: 0.67
- Column Sample by Tree: 1.0
- Sub Sample: 0.2575

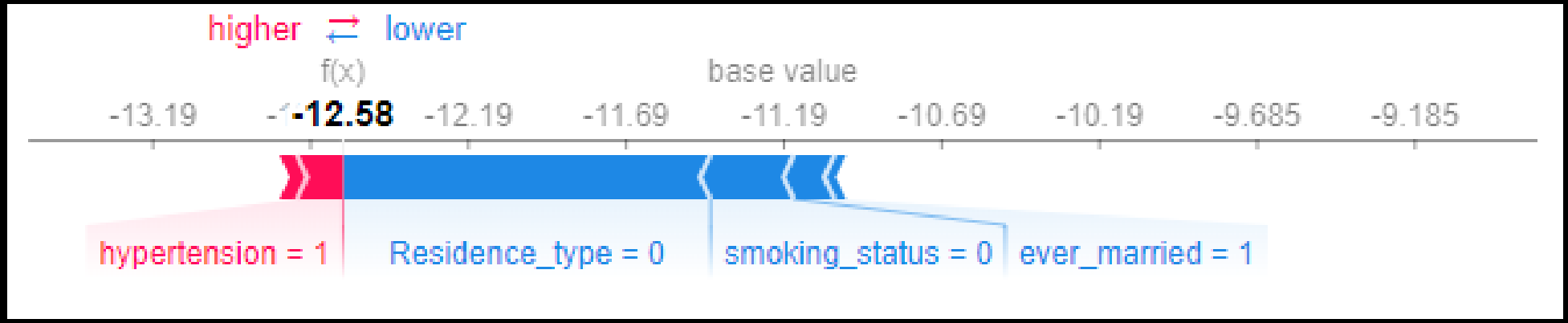
Results cont.:





Feature Importance: Age, BMI, and average glucose level all impact the outcome of a stroke the most.

Results cont.:



SHAP: Depicts a single patient's local feature importance.

- Can show patient what is causing their classification
- Residence type, smoking status and marriage push value left
- Hypertension push value right

Outlook:

- Need additional patient data to make predictions more accurate.
  - Additional stroke patients (4.26% to ~11%)
- Possible Important Features:
  - Physical activity habits
  - MRI scan data
  - Blood health metrics
- Implement KNN

Questions?

Thank you!