Cardiac Stroke Risk Stratification using Classification Model

Taesun Yoo

- June 14, 2018 -

Springboard

Cardiac Stroke Statistics: US in 2017

1 in 20 deaths

Accounts from cardiac stroke

Rank #5

Among all causes of death in US, killing 133K people a year

795K people

Experience a new or recurrent stroke

$52 Billion

Estimated indirect and direct costs for stroke

Springboard

Problem Statement

Why should you care?

- Stroke is a preventative condition
- ↑ in projected % of people having a stroke
- ↑ in cost ($) for stroke treatment

Stakeholders:

- Cardiac care unit managers and clinicians

Goal:

- Predict patients with high risks of developing a stroke

Objective:

- Help physicians to take proactive health monitoring
- Target prevention on patients with high risk for developing a stroke

Dataset Overview

Dataset contains **11** input features for predicting an "**stroke**" label:

- 8 categorical & 3 numerical features
- Lifestyle and health demographic indicators
- Sample size = 43,000 rows

Observations (rows)

ID	Gender	Age	Hypertension	Heart_Disease	Ever_Married	Work_Type	Residential_Type	Avg_Glucose_Level	BMI	Smoking_Status	Stroke
30669	Male	3	No	No	No	Children	Rural	95.1	18	NULL	0
16523	Male	58	Yes	No	Yes	Private	Urban	110.9	39.2	Never Smoked	0
56543	Female	8	No	No	No	Private	Urban	69	17.6	NULL	0
46136	Female	70	No	No	Yes	Private	Rural	161.3	35.9	Formerly Smoked	0
32257	Male	47	No	No	Yes	Private	Rural	210.1	50.1	NULL	0

Features (attributes)

Classes (label)

Challenges:

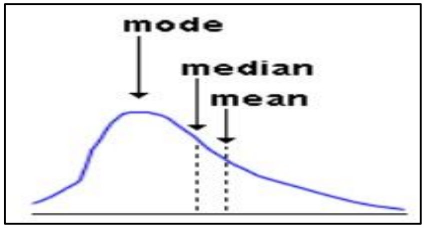
- Class imbalance (98% healthy vs. 2% stroke)
- Outliers & duplicates
- Missing values

Data Wrangling

Raw Data Loading

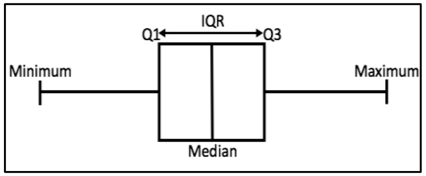
↓

Feature Drop | Feature Imputation → **Missing Value Replacement**



↓

Interquartile Range:
LB = Q1 – 1.5*IQR
UB = Q3 + 1.5*IQR
→ **Handling Outliers**



↓

Down-sampling → **Resampling**

Non-stroke (50%)	Stroke (50%)

↓

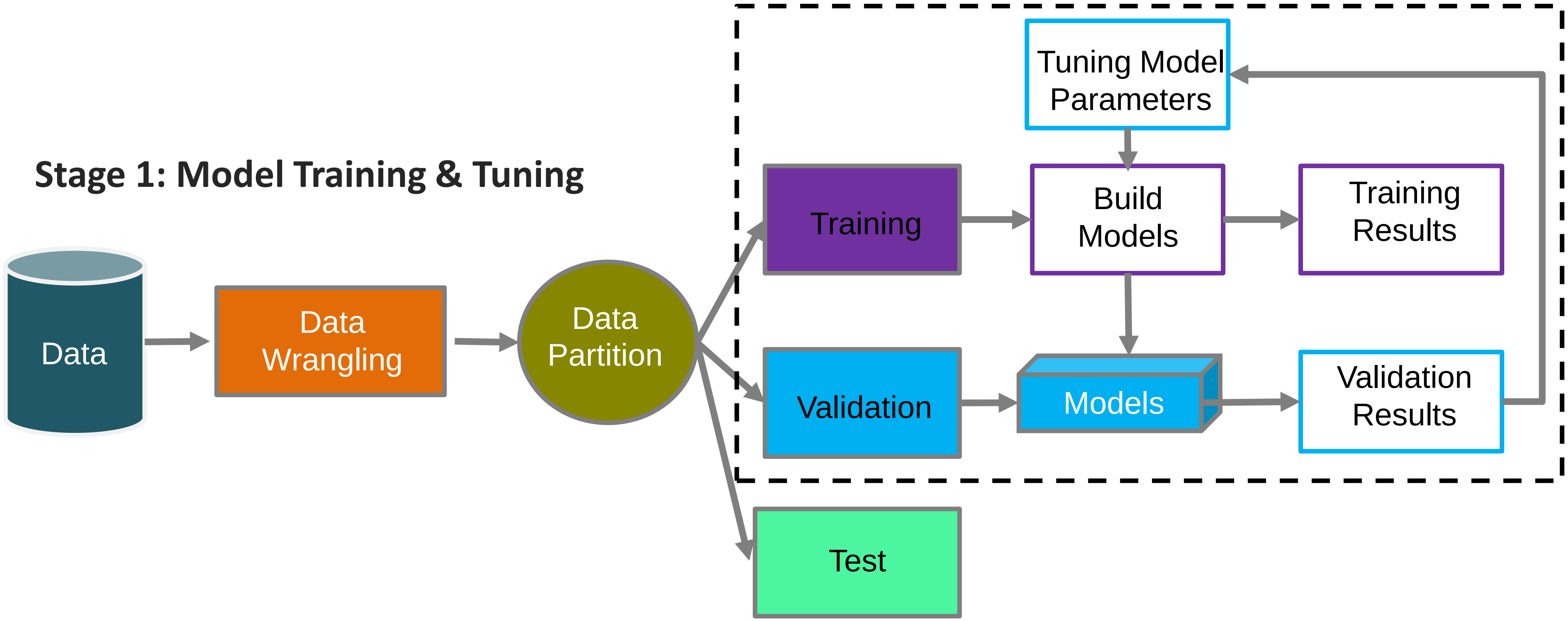
Feature Encoding | Feature Engineering | Feature Scaling → **Feature Transform**

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

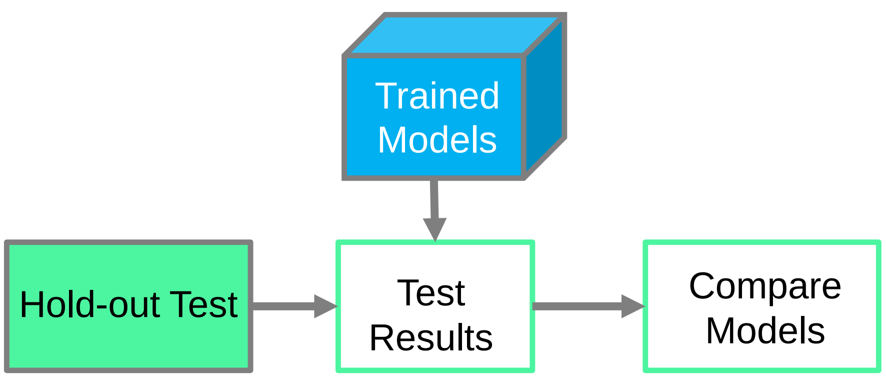
Springboard

Classification Model Workflows

Stage 1: Model Training & Tuning

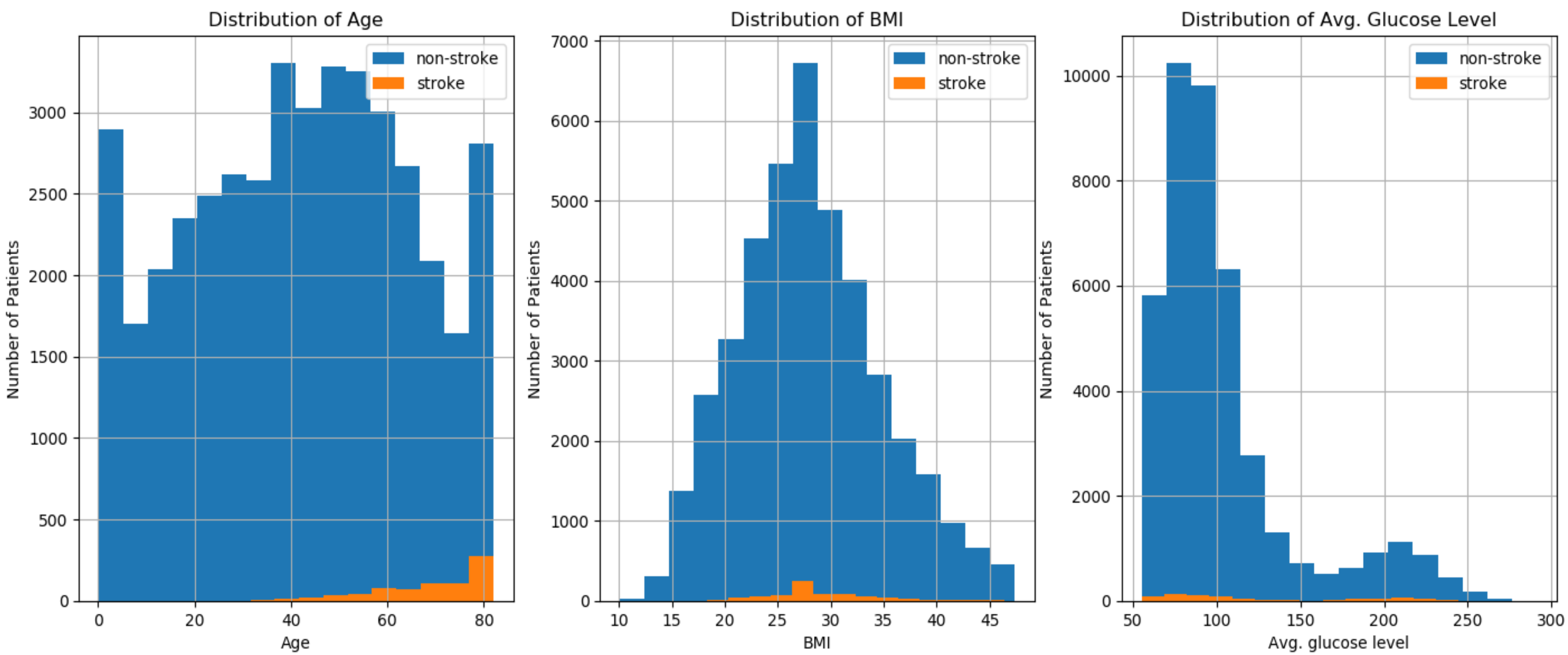


Stage 2: Model Performance Estimate



Springboard

Distributions: Healthy vs. Stroke Population



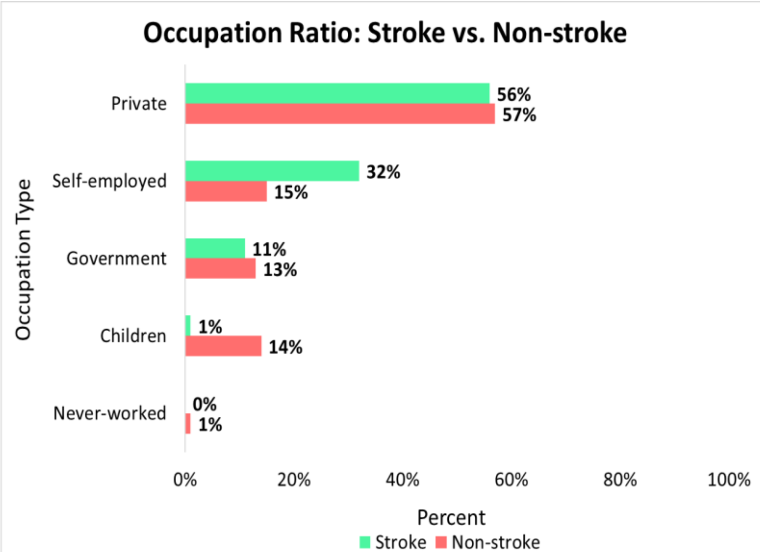
Age: majority of senior stroke patients (skewed to left)

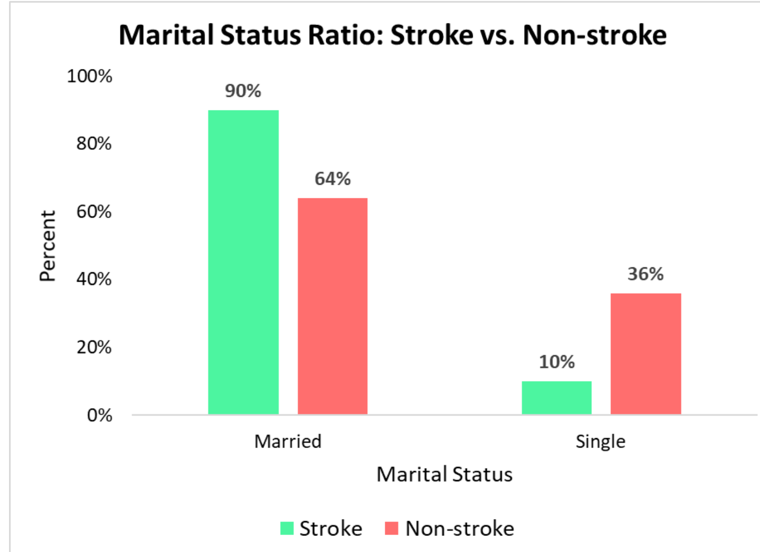
BMI: normal distribution (centralized from 25 to 30)

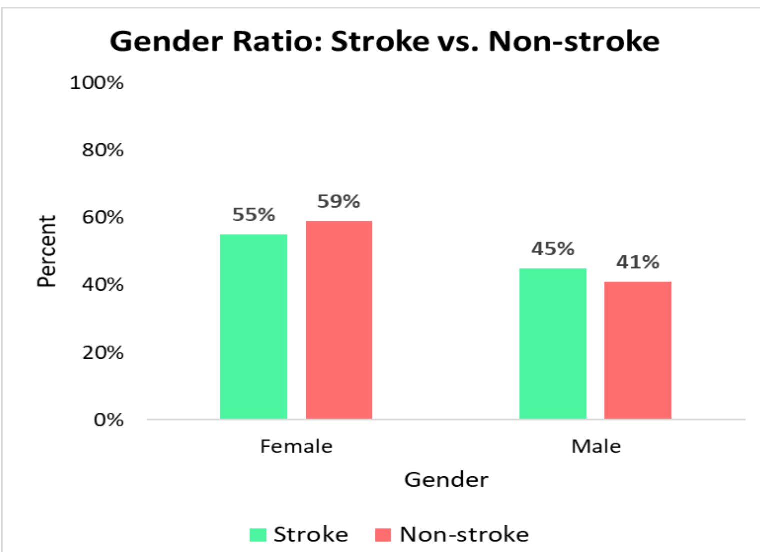
Avg. glucose level: non-normal distribution (bi-modal peaks)

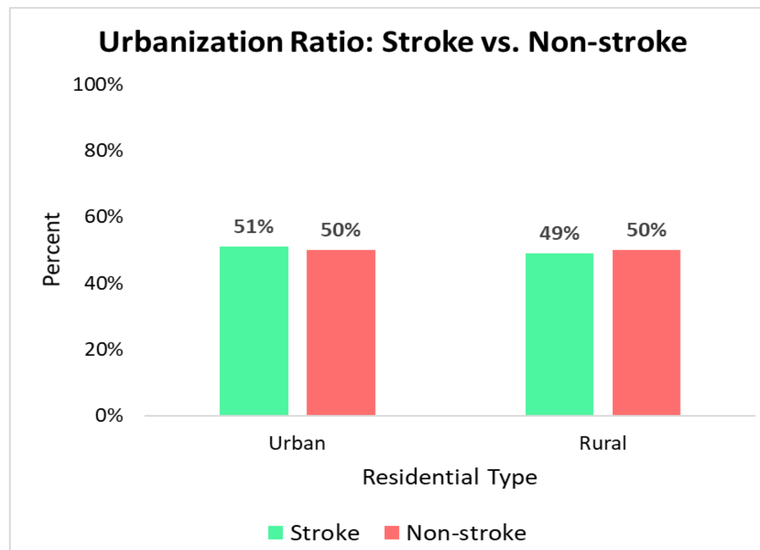
Springboard

Lifestyle Factors: Healthy vs. Stroke Population





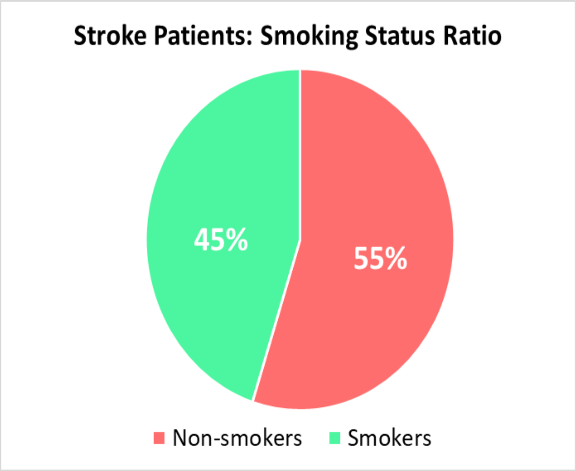


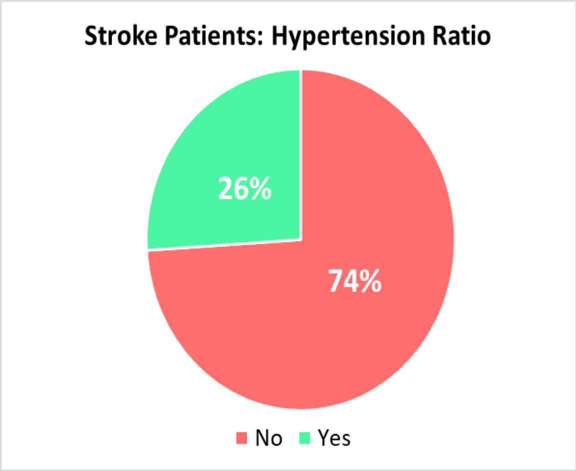


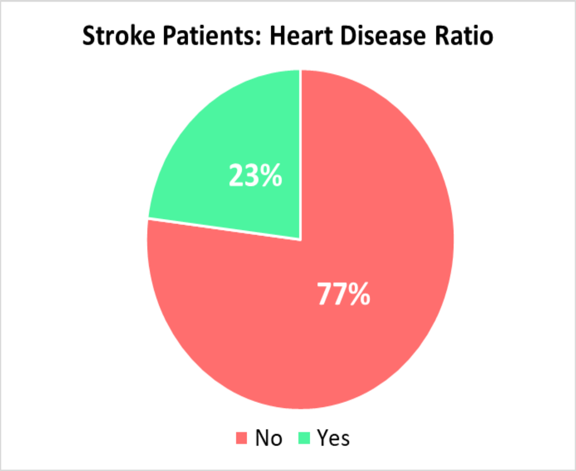
Springboard

Health Indicators: Healthy vs. Stroke Population

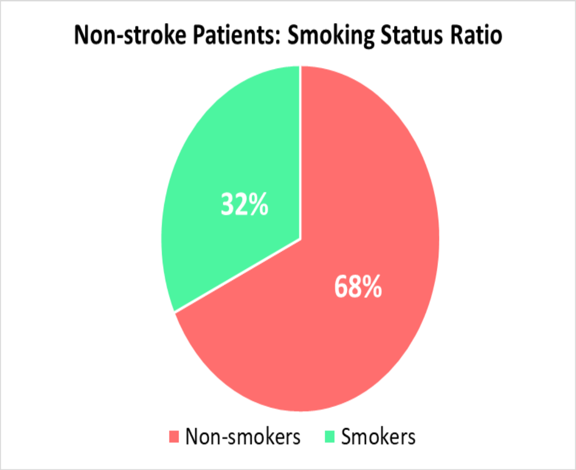
Stroke

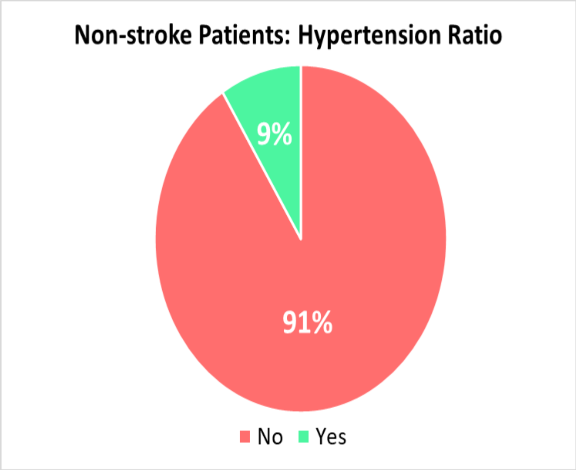


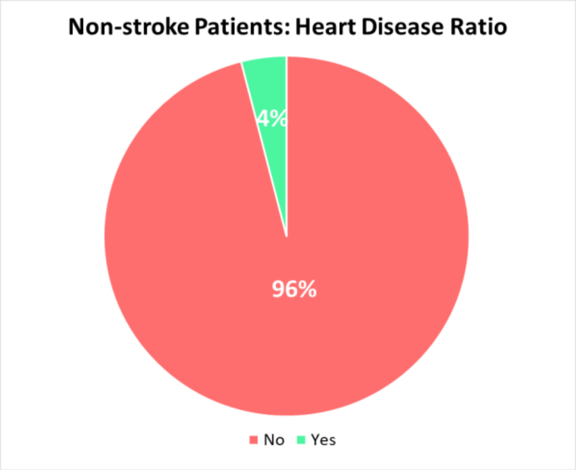




Non-stroke

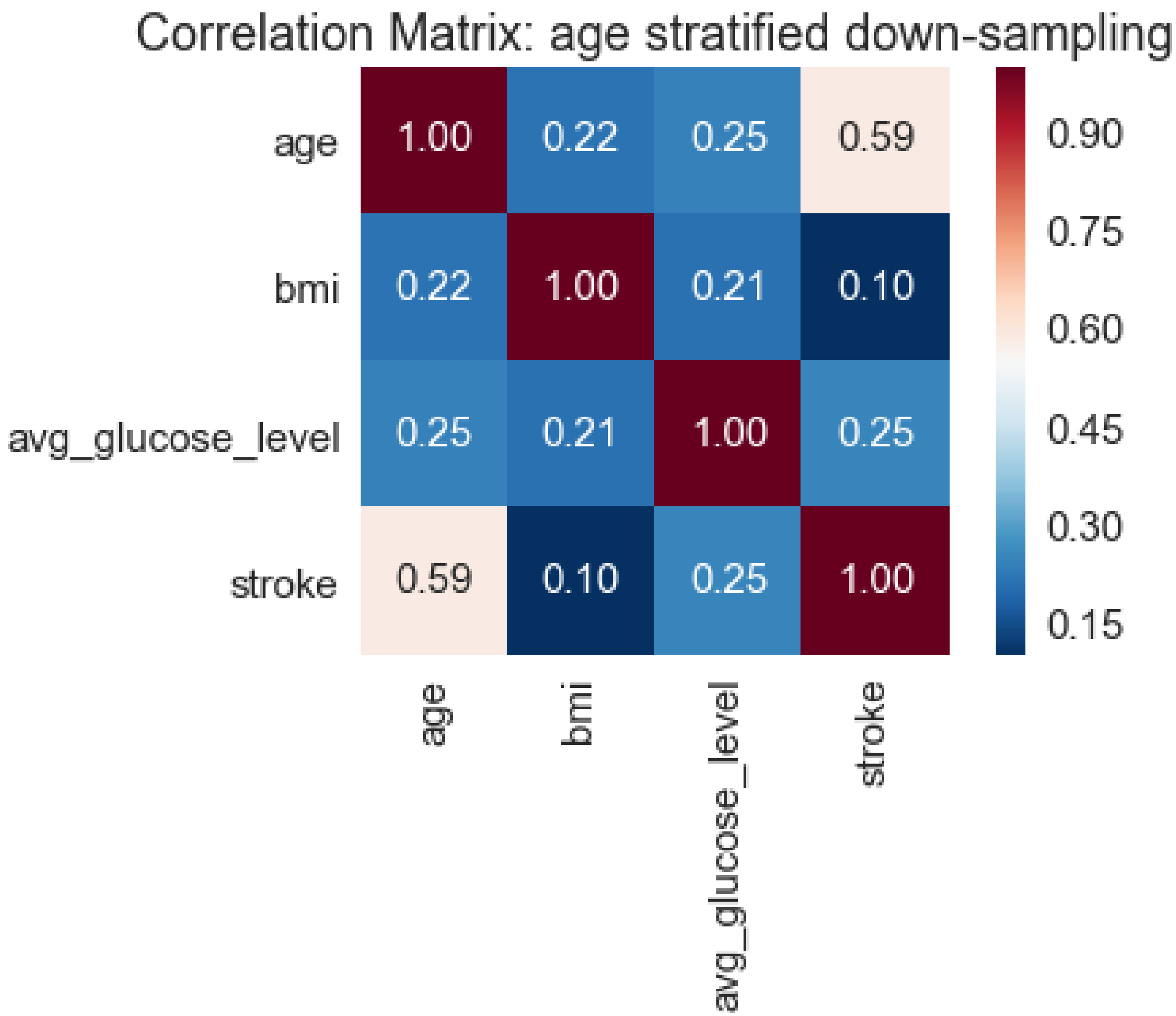






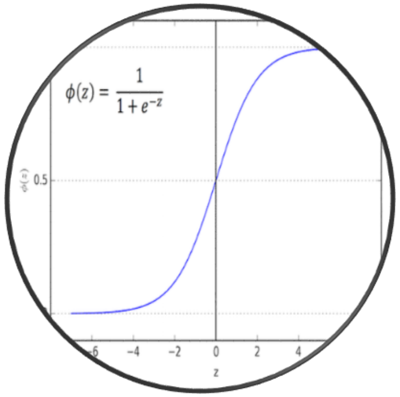
Springboard

Correlation Matrix



Springboard

Model Selections

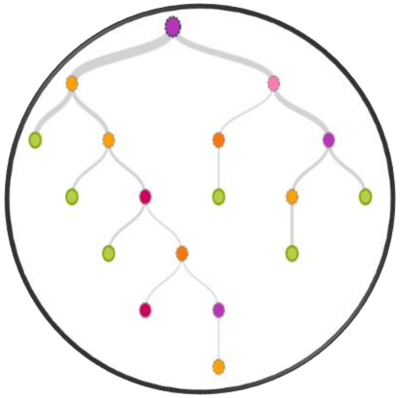


Logistic Regression

Sigmoid logit function:
log(p/(1-p))

Transforms:
Input values → estimated into prob. range (0, 1)

Works well on linearly separable classes.

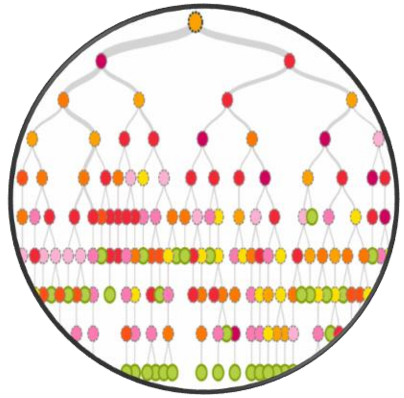


Decision Tree

Split data on features.

Repetitive splitting procedure.

Continue splitting until each node left with same class label.

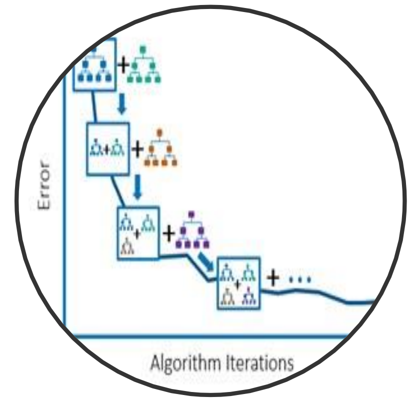


Random Forest

Ensemble learning.

Creates many decision trees.

Average performance of trees.



Gradient Boost

Sequential training.

Learn from residual errors.

Step-wise forward

$$\text{Label} = \text{mode}\ \{c_{lr}(x),\ c_{dt}(x),\ c_{rf}(x),\ c_{xgb}(x)\}$$

Majority Vote

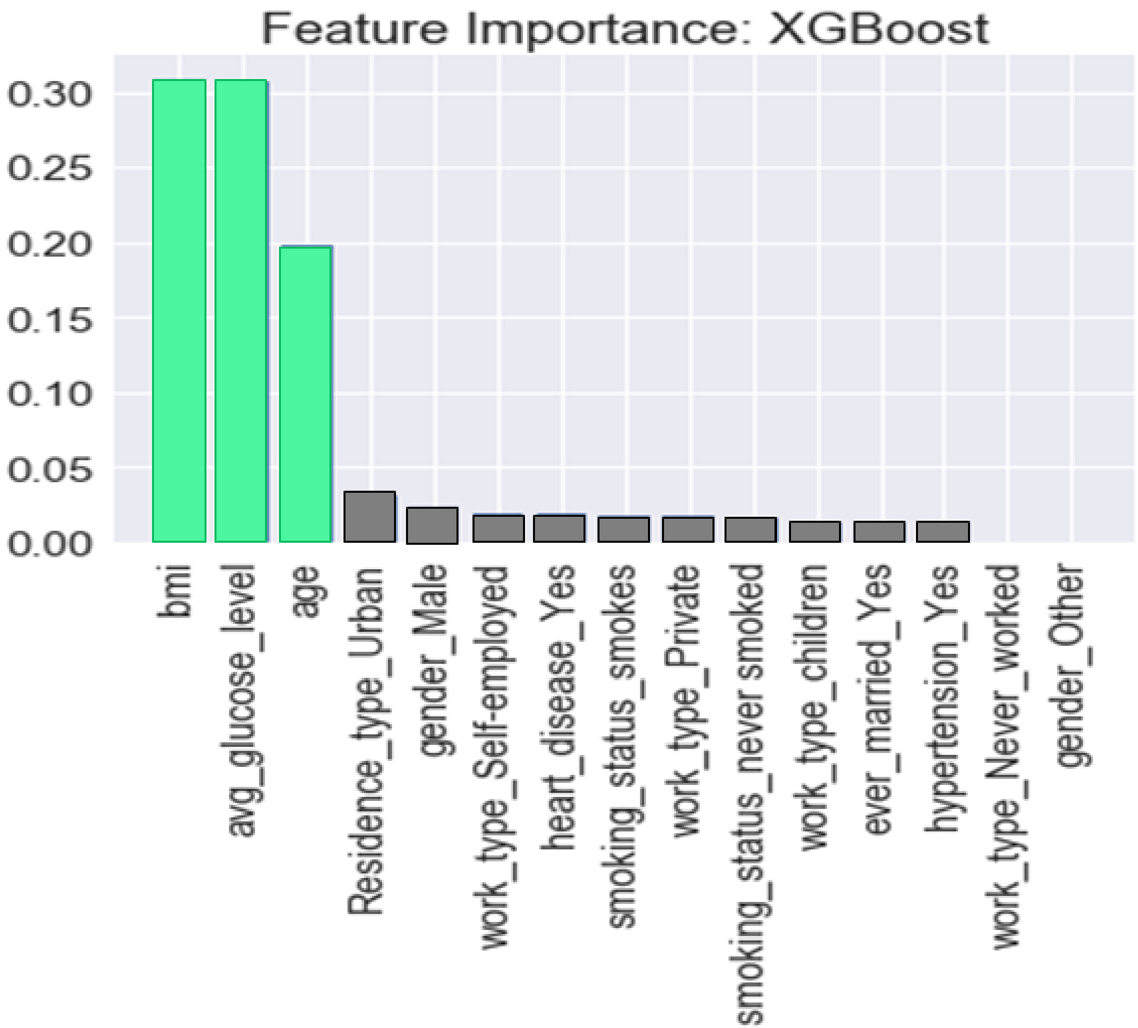
Meta-classifier

Combination of four models

Improves accuracy of model performances by majority vote

Springboard

Feature Selections



Model Comparison

	Logistic Regression	Decision Tree	Random Forest	XGBoost	Majority Vote
Accuracy	77%	75%	77%	77%	**80%**
Precision	**75%**	68%	73%	73%	78%
Recall	81%	**93%**	84%	86%	82%
ROC Score	77%	75%	77%	77%	**80%**

Overall, in terms of evaluation metrics:

- Best performing model was "Majority Vote classifier"

Springboard

Confusion Matrix

MajorityVote Classifier		
	Predicted Class	
Actual Class	Stroke	Non-stroke
Stroke	41%	9%
Non-stroke	11%	39%

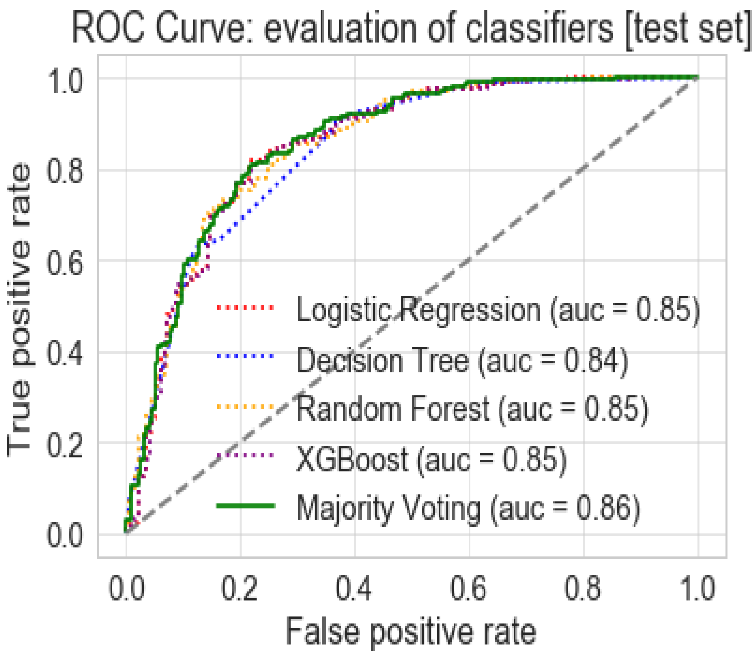
Outcome Interpretation:

- 80% of correct predictions
- 20% of mis-classification errors

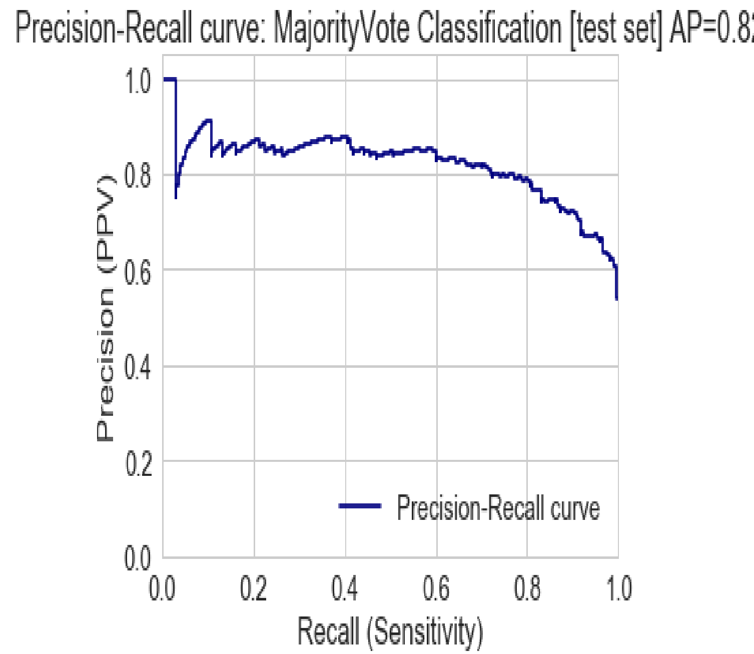
Balance between ML model and human intervention is required especially on **9% error** (*Type II error*).

Springboard

ROC and Precision-Recall Curves



ROC Curve



Precision-Recall Curve

Springboard

Summary: Stroke Classification

Goal

Predict cases at high risks of developing a stroke by classification model

Results

- Model was able to predict whether or not patients were at risk of stroke
- 80% of accurate predictions were made on test set of stroke data

Risks & Mitigation

Risks:
Model incorrectly classified with 9% error as likely patients are non-stroke but in fact had stroke

Mitigation:
Review identified cases with a group of clinicians before decision making

Next Steps

- Collection of meaningful features
- Model improvement: algorithms, resampling and designs

Springboard

Limitations & Future Work

Limitation:

- Absence of useful features/attributes
- Weak feature interaction (i.e., smoking, hypertension)

Future Work:

- Collection of features (i.e., genetic pre-disposition, physical activity, etc.)
- Model improvement: combine multiple classifiers
  - Stacking
  - Other ensemble
- Resampling strategies:
  - SMOTE
  - Oversampling (i.e., minority class: stroke cases)
- Age stratified classifiers:
  - Younger patients cohort (age < 30)
  - Senior patients cohort (age > 50)

Springboard

Recommendations

Add stroke screening test
At a recommended age (before 65)

Collect meaningful features
Stress level, physical activity,
blood pressure, genetic factors.

Conduct cohort studies
With feature engineering,
stratify patient cohorts into
diabetic and obese groups

Thank You!

Questions?