Stroke Prediction

Data Science Capstone Project

The Problem

Context:

- Stroke is the 2nd leading cause of death globally, 11% of total deaths.
- 1 in every 6 deaths from cardiovascular disease was due to stroke.*

What patterns in a patient's attributes make them more or less likely to have a stroke?

How can we track that and lower overall healthcare costs to the patient, and what measures could be taken to prevent a patient from getting a stroke?

Who are the stakeholders?

- General Public Health
- Insurance Companies tracking client health
- Healthcare Analytics
- Doctors
- Everyone who may be affected by stroke

Data Information

Source:

Data acquired on Kaggle from user: fedesriano*

Includes 11 clinical features for predicting stroke events.

Contains 5110 Observations.

File format: CSV file

Each Record: a unique patient

Data Wrangling

- ‘Bmi’ column had 201 counts of missing values, so filled in with the mean of the column. BMI is important to a patients physique.
- Categorical features such as ‘gender’ and ‘smoking_status’ were turned into numerical features.
- Dropped columns such as ‘id’, ‘Residence_type’, and ‘work_type’ because they could not be quantified.

EDA

According to the Mayo Clinic there are many factors that can increase your risk of a stroke, they include such things as:

Lifestyle risk factors are:

- Being overweight or obese
- Physical inactivity
- Heavy or binge drinking
- Use of illegal drugs such as cocaine and methamphetamine

Medical risk factors are:

- High blood pressure
- Cigarette smoking or secondhand smoke exposure
- High cholesterol
- Diabetes
- Obstructive sleep apnea
- Cardiovascular disease, including heart failure, heart defects, heart infection or abnormal heart rhythm, such as atrial fibrillation
- Personal or family history of stroke, heart attack or transient ischemic attack
- COVID-19 infection

In the kaggle Dataset we have factors of stroke such as:

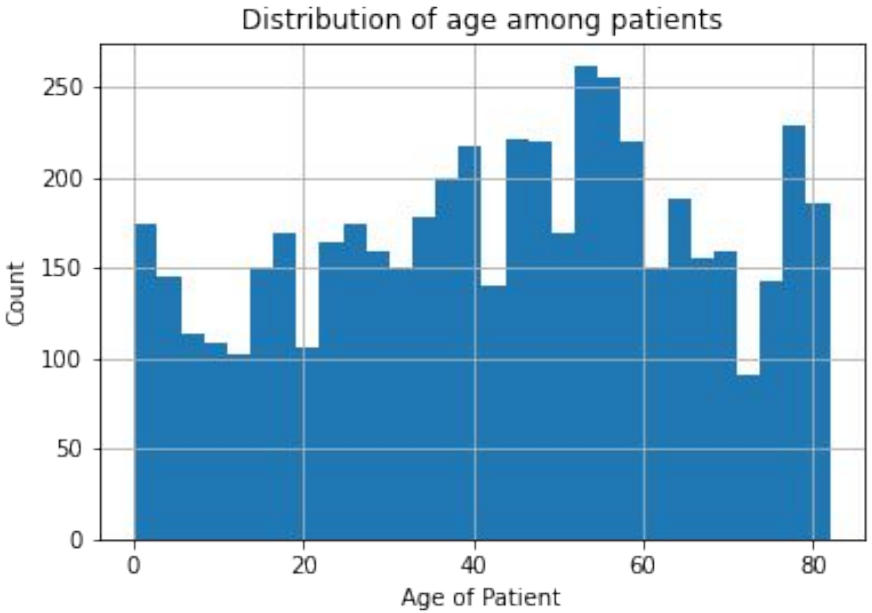
- Age
- Hypertension (which is another name for high-blood pressure)
- Heart disease
- Average glucose level (which can be an indicator of Diabetes)
- BMI Body Mass Index (which is an indicator of obesity, and a patients physique)
- Smoking Status (which tells us if they smoked)
- Stroke (if they have had a stroke before)

EDA - ‘age’ column

Age is important when we look at the likelihood of stroke in a patient. Mayo Clinic says patients above the age of 55 have higher risk of stroke.

But stroke could happen to anyone anytime!*

I plotted the Distribution of age among patients.

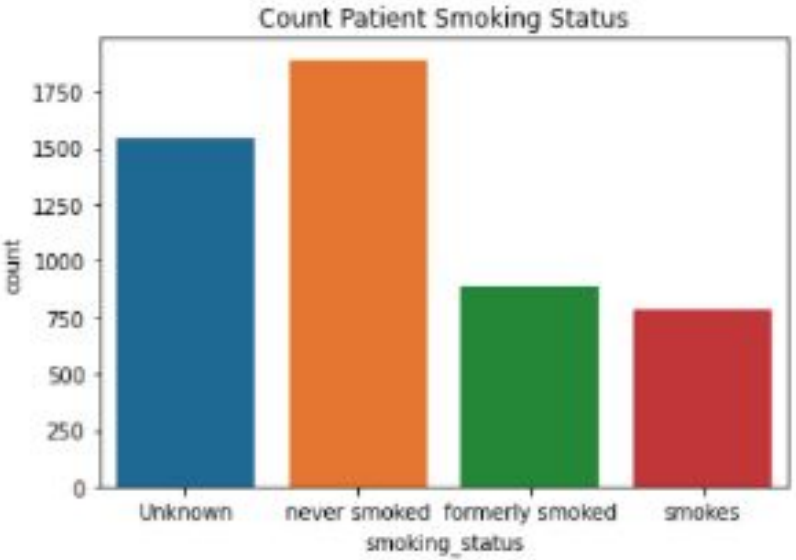


EDA - ‘smoking_status’ column

How is smoking related to heart disease and stroke? Smoking can:

- Raise triglycerides (a type of fat in your blood)
- Lower “good” cholesterol (HDL)
- Make blood sticky and more likely to clot, which can block blood flow to the heart and brain
- Damage cells that line the blood vessels
- Increase the buildup of plaque (fat, cholesterol, calcium, and other substances) in blood vessels
- Cause thickening and narrowing of blood vessels

However there are children/young adults in this dataset that would make me think they would go in the 'Unknown' and ‘never_smoked’ category.

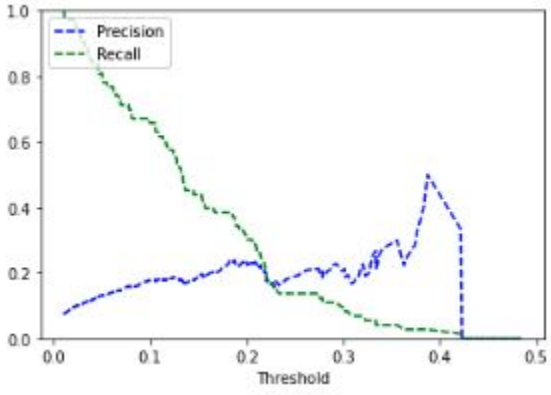


Machine Learning Modeling

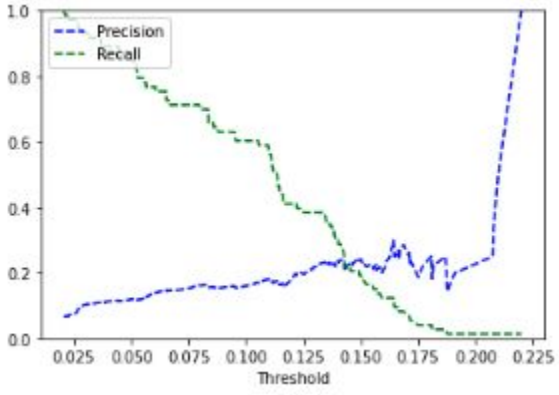
Modeling Overview

- Type:
  - Supervised Learning
- Binary Classification:
  - 1 for stroke
  - 0 for no-stroke
- Highly Imbalanced Dataset:
  - 1 249 counts
  - 0 4860 counts
- Tools: Python’s sklearn, precision_recall_plot, get_classification_report
- Modeling Steps:
  - Data Pre-processing (label, endoing, train/test split, resampling, scaling
  - Cross Validation, grid-search method
  - Training using optimal parameters, finding right threshold

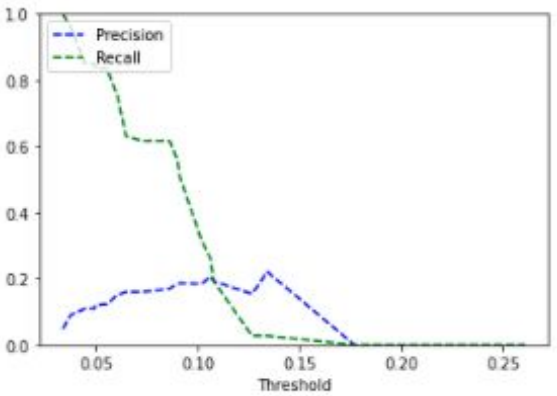
Precision-Recall Plot of Models



Logistic Regression

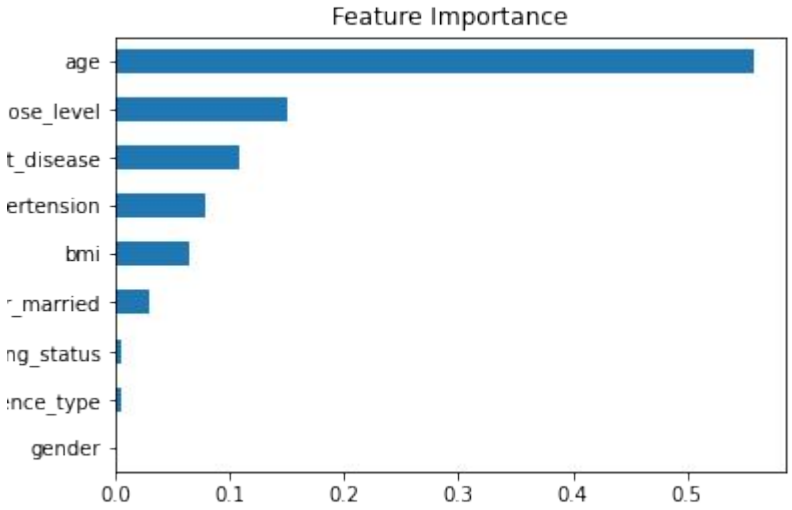


Random Forest

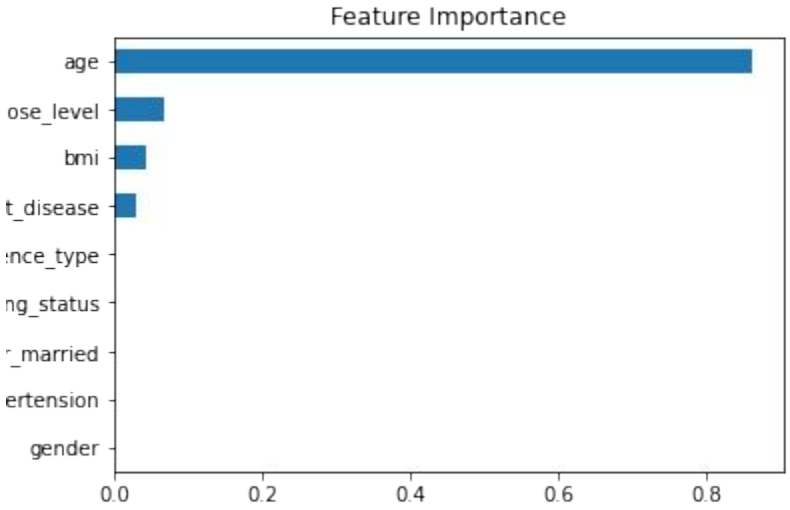


Gradient Boosting

Feature Importance



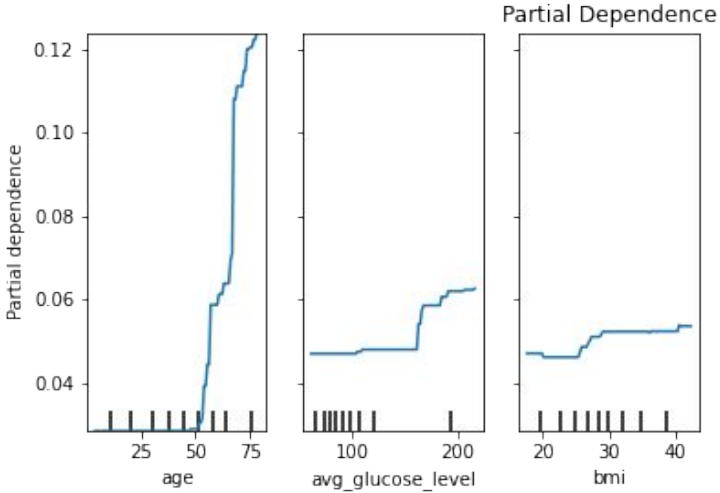
Random Forest



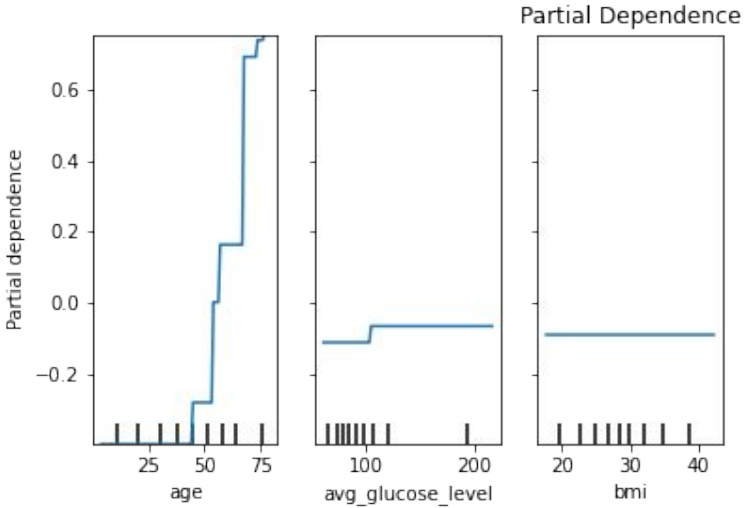
Gradient Boosting

In both Random Forest Model and Gradient Boosting Model, age was the most significant factor to determining if a patient was going to have a stroke or not.

Partial Dependence Plot of Models



Random Forest



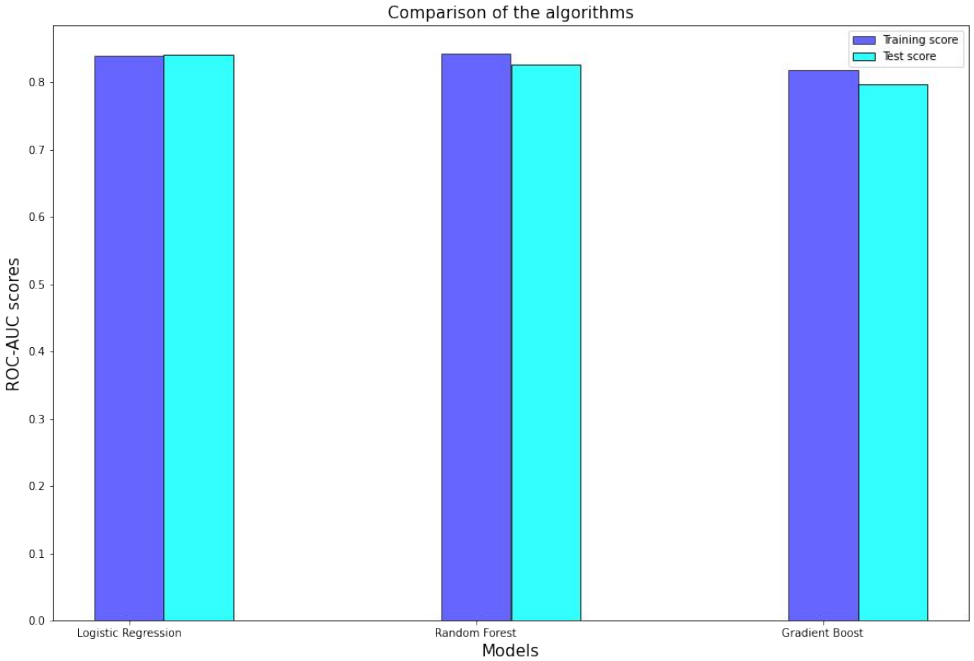
Gradient Boosting

In both Random Forest Model and Gradient Boosting Model, we can see that right after age 50 a patient's likelihood of stroke goes up. In the Random Forest Model, we can see that around 150 glucose level also has an increased risk.

Model Comparisons

Here we see that Random Forest has the best ROC_AUC scores for the training. Logistic Regression, however, has the best scores for testing.

All three performed very well.



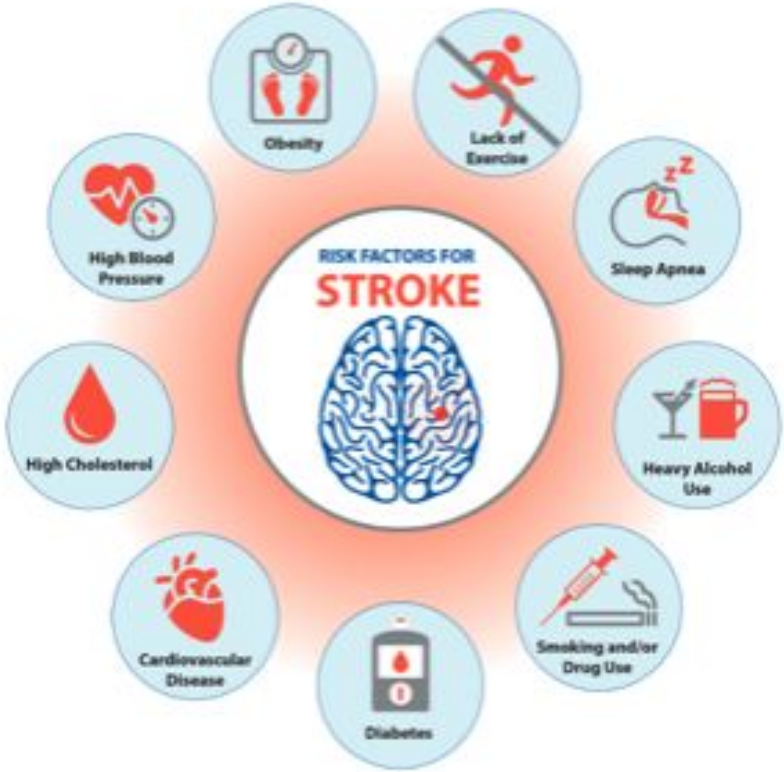
Conclusion

Logistic Regression was the model that was the best performing.

Out of 11 features, ‘age’ and ‘gluclose_levels’ were the best.

This project and this dataset could be explored much further in depth, and would be a great benefit to public health knowledge.

Even though many people know the risk factors of stroke, to see their actual different attributes of their health plotted visually would be more eye-opening.



Resources

- https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
- https://www.cdc.gov/stroke/facts.htm
- https://newsnetwork.mayoclinic.org/discussion/stroke-impacts-all-regardless-of-age-race-and-gender/
- https://www.chihealth.com/en/services/neuro/neurological-conditions/stroke/stroke-prevention.html
- My Github repo:
  - https://github.com/FayD21/Capstone-2-Stroke-Prediction

Thank You!

Fay Dennis

Fay Dennis

Email: iamfaydennis@gmail.com

GitHub: https://github.com/FayD21

Special Thanks to my Mentor!

Jeff Hevrin