November 2022

Stroke Prediction

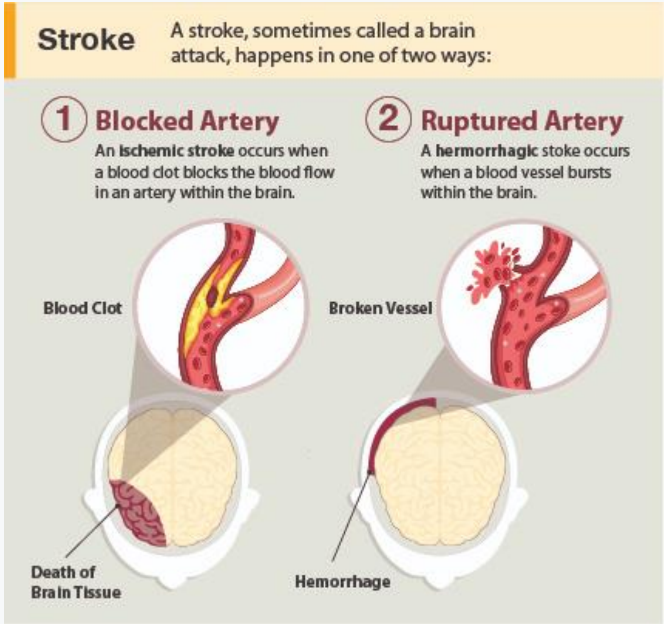
Tyler Schelling

Project Description

The purpose of this project is to predict a stroke outcome based on various data gathered from patients. Risk factors include age, gender, hypertension, heart disease, glucose levels, BMI, and more.

What is a stroke?

Per the CDC, a stroke occurs when something blocks blood supply to a part of the brain or when a blood vessel in the brain bursts. In either case, parts of the brain become damaged or die. A stroke can cause lasting brain damage, long-term disability, or even death.



Stroke A stroke, sometimes called a brain attack, happens in one of two ways:

1 **Blocked Artery**

An **ischemic stroke** occurs when a blood clot blocks the blood flow in an artery within the brain.

2 **Ruptured Artery**

A **hermorrhagic** stoke occurs when a blood vessel bursts within the brain.

Stroke Dataset

The dataset was sourced from Kaggle - Stroke Prediction Dataset from the user ‘Federsoriano’.

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

This dataset utilizes 11 features for predicting stroke events.

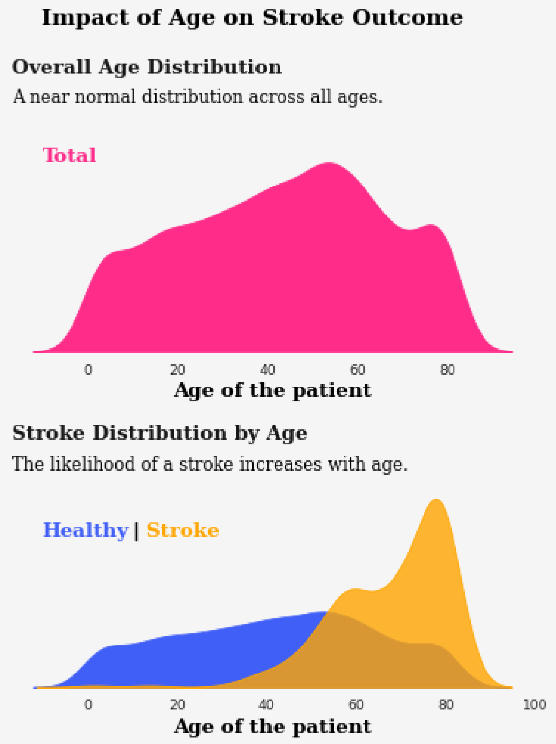
Feature Name	**Description**	**Feature Name**	**Description**	**Feature Name**	**Description**
ID	Unique Identifier	Heart_disease	0 (no) or 1 (yes)	Avg_glucose_level	Glucose level (mg/dL)
Gender	Male, Female, or Other	Ever_married	Yes or No	BMI	Body Mass Index (kg/m2)
Age	Age of the patient	Work_type	Child, Govt, Private, Self-employed, or Never	Smoking_status	Smoker, Former, Never, or Unknown
Hypertension	0 (no) or 1 (yes)	Residence_type	Rural or Urban		

Stakeholders

This analysis will be utilized by a medical insurance company in order to assist with predicting stroke outcomes to push preventative care for patients that are higher risk.

Covering the cost of a stroke can range from $20,396 to $43,652. Preventative care can improve the health of patients as well as save the insurance company money for potential claims.

Key Findings



Impact of Age on Stroke Outcome

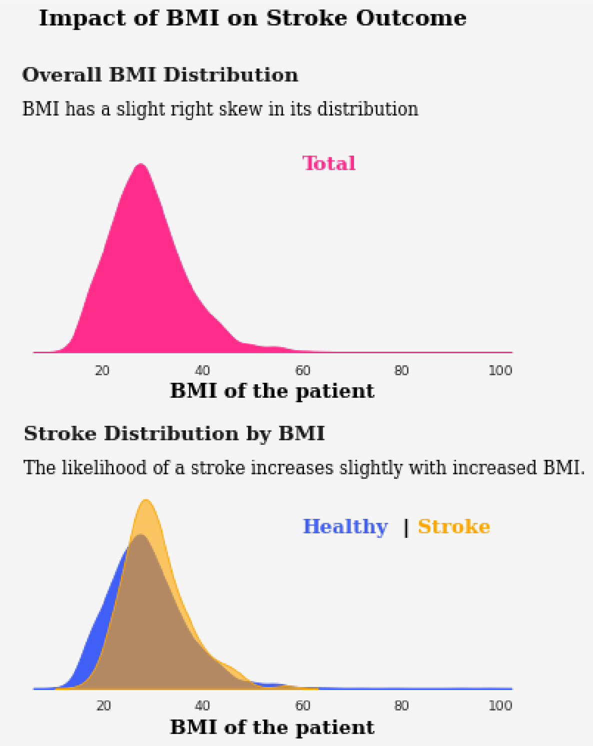
Overall Age Distribution

A near normal distribution across all ages.

Stroke Distribution by Age

The likelihood of a stroke increases with age.

Key Findings



Impact of BMI on Stroke Outcome

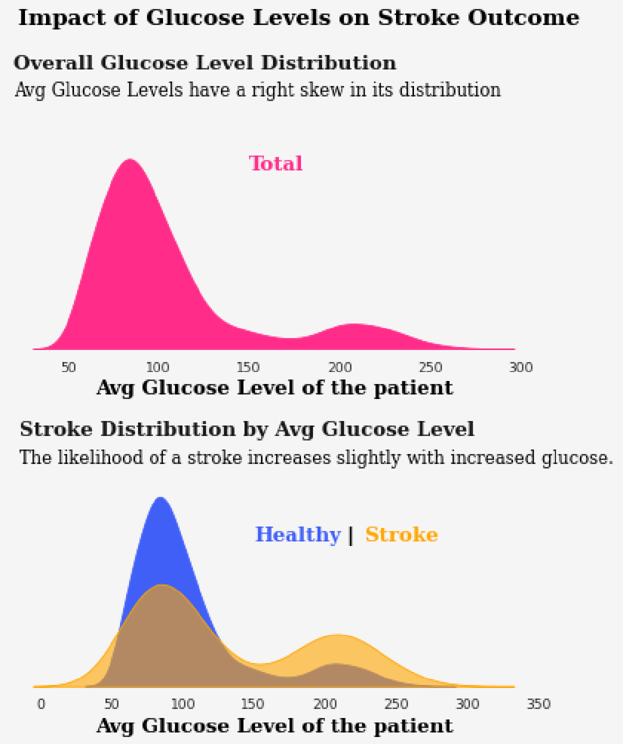
Overall BMI Distribution

BMI has a slight right skew in its distribution

Stroke Distribution by BMI

The likelihood of a stroke increases slightly with increased BMI.

Key Findings



Impact of Glucose Levels on Stroke Outcome

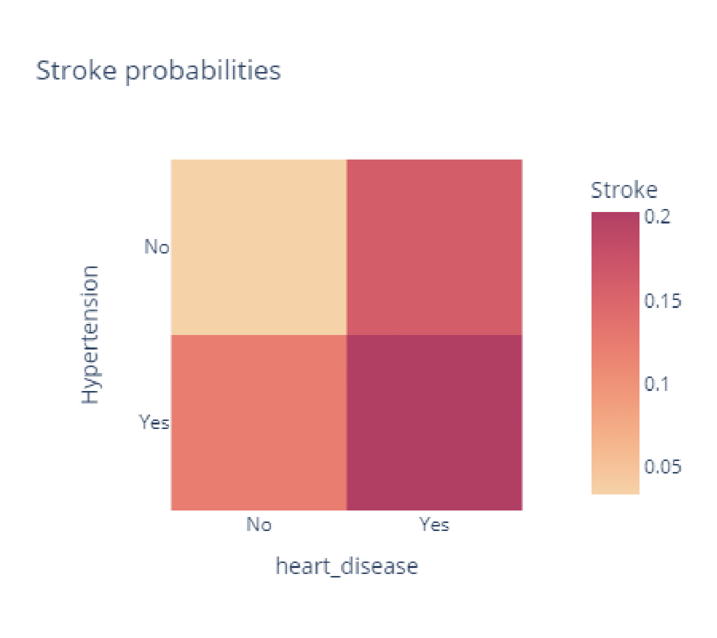
Overall Glucose Level Distribution

Avg Glucose Levels have a right skew in its distribution

Stroke Distribution by Avg Glucose Level

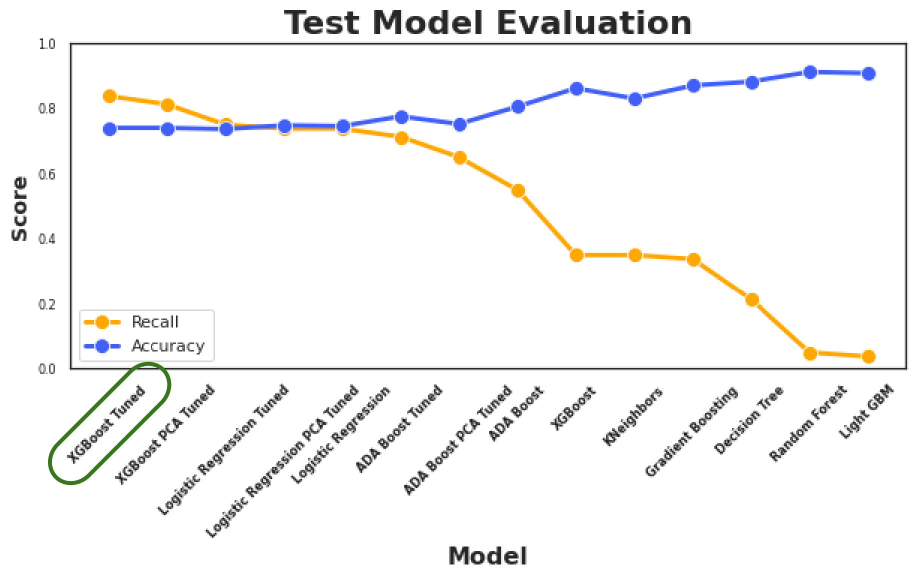
The likelihood of a stroke increases slightly with increased glucose.

Key Findings



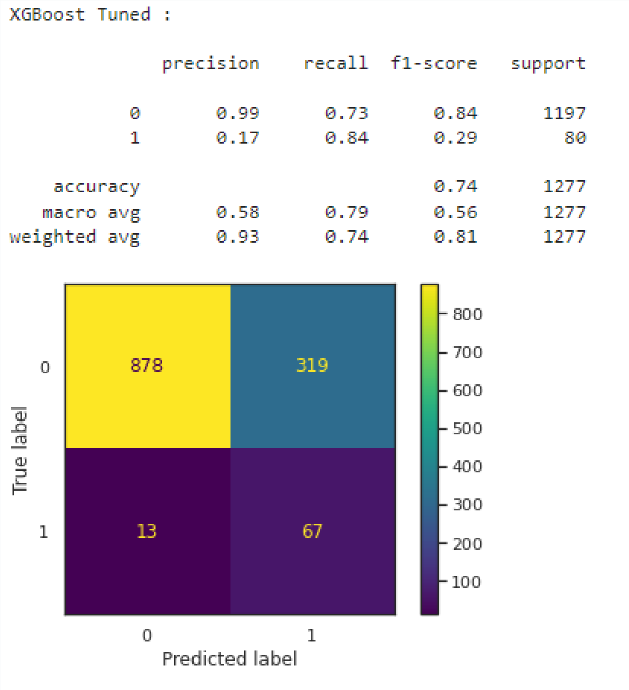
- **The likelihood of a stroke is highest for patients that have both heart disease and hypertension.**
- **Patients with either heart disease or hypertension also have an increased risk.**
- **Patients that remain healthy (no heart disease or hypertension) have the lowest risk of having a stroke.**

Model Evaluation



- **The primary metric for evaluating the models will be recall (or what proportion of actual positives were correctly identified).**
- **The best model that optimizes recall is a tuned *XGBoost model*.**
  - **Although the overall accuracy is 74.0%, the recall is the highest amongst the models at 83.8%**
  - **This means that almost 84% of the model's predicted outcome of having a stroke were correct.**

Model Evaluation



```
XGBoost Tuned :

              precision    recall  f1-score   support

           0       0.99      0.73      0.84      1197
           1       0.17      0.84      0.29        80

    accuracy                           0.74      1277
   macro avg       0.58      0.79      0.56      1277
weighted avg       0.93      0.74      0.81      1277
```

- Minimizing the false negatives will be the most beneficial for the insurance company.
- Incorrectly predicting patients will not have a stroke when they actual do (false negative) can be costly to the company and does not provide adequate resources for preventative care to patients.
- The downside to this model is the high rate of false positives. However, providing additional preventative care to patients that are likely not going to have a stroke will still be more cost effective than having a higher false negative rate.

Recommendations

- The tuned XGBoost model can lead to catching at risk patients early to provide the necessary preventative care and/or treatment.
- False negatives are still a risk in the model and some predictions may require mild manual review in order to potentially catch any concerns not captured by the predictive model.
- Aging patients, especially those with heart disease and/or hypertension, should seek medical care to get the appropriate preventative care with a medical professional.