Stoke Prediction
In the Hospital Setting

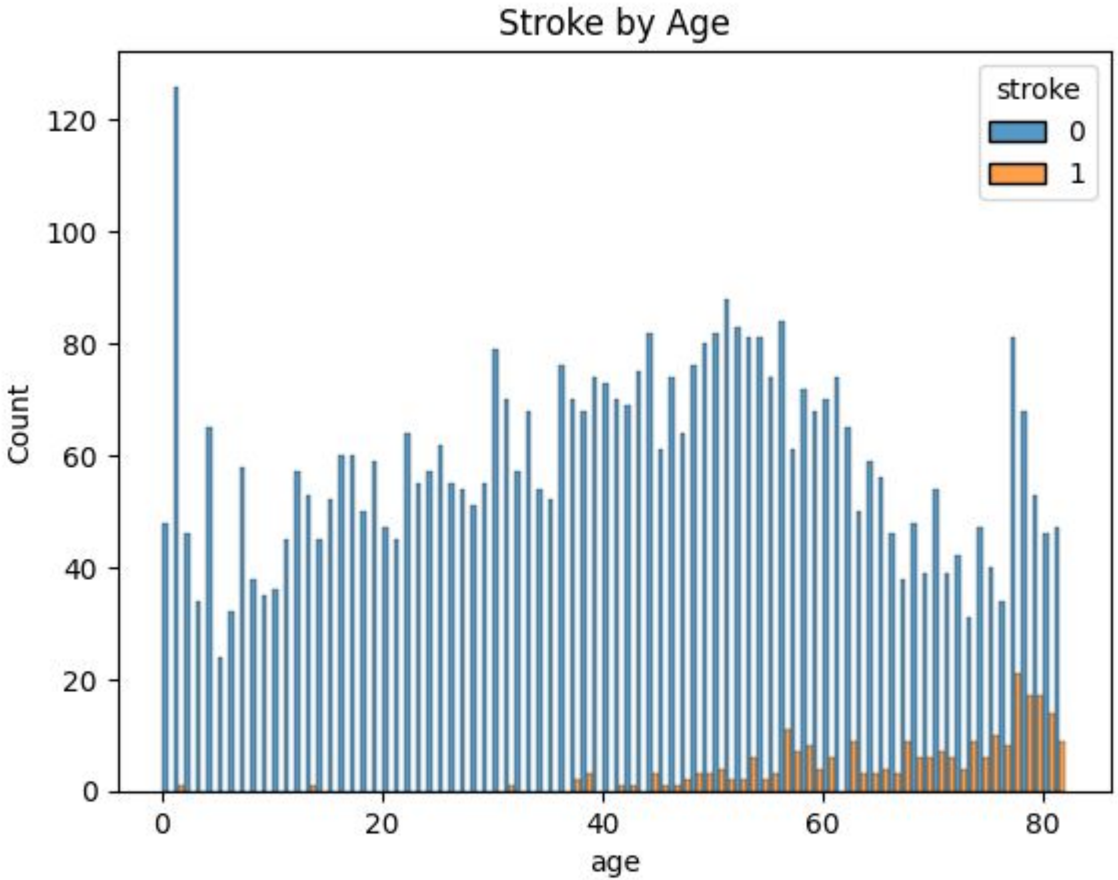
Is it Possible?

Introduction to the Data

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

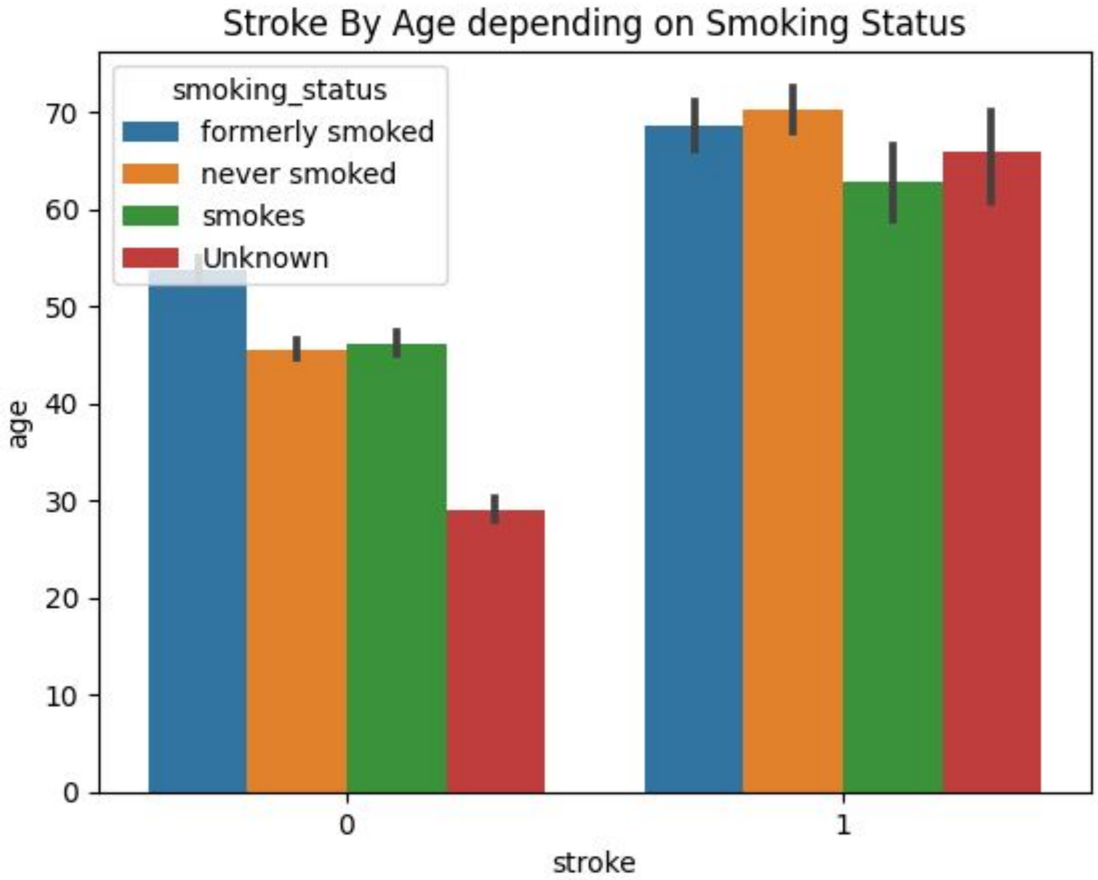
This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Stroke by Age



Previous Graph

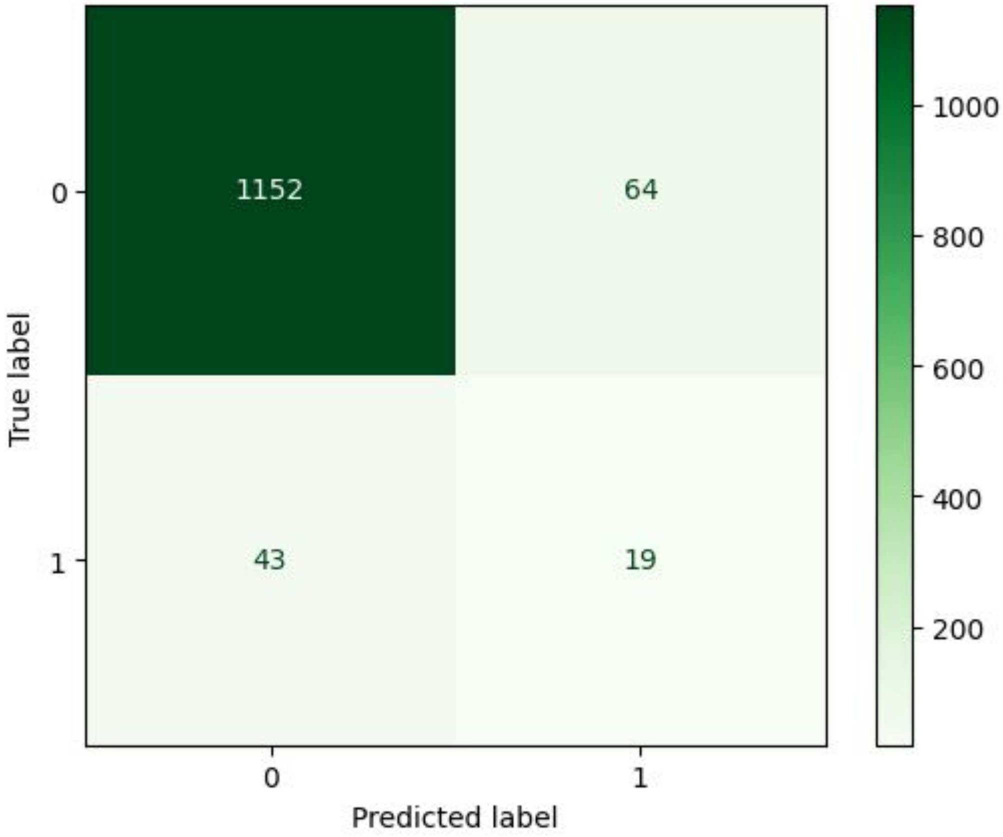
The previous graph shows the correlation of strokes with age. Most of the strokes take place after the age of 40. There are some rare strokes under the age of 40. The majority of the strokes happened around the age of 80.

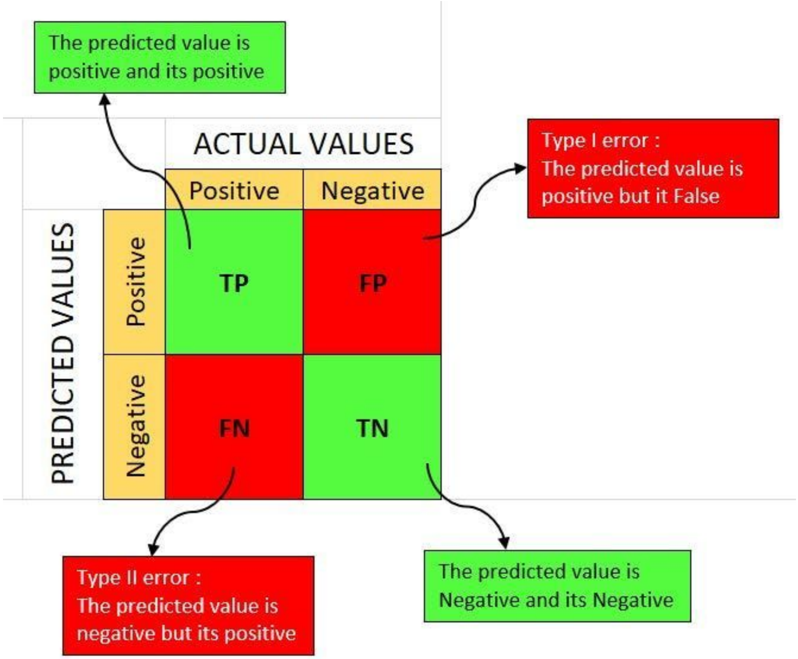

Stroke By Age depending on Smoking Status
smoking_status
formerly smoked
never smoked
smokes
Unknown
age
0
10
20
30
40
50
60
70
stroke
0
1

Previous Graph

The previous graph shows the correlation of strokes by age and smoking history. Most of the strokes take place between the age of 60 and 70. The best thing this graph shows is how not smoking puts off the onset of stroke by almost ten years.

Best Model





		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

The predicted value is positive and its positive

Type I error :
The predicted value is positive but it False

Type II error :
The predicted value is negative but its positive

The predicted value is Negative and its Negative

False Negatives and False Positives

False Negatives: When it comes to our model False Negatives are the worst outcome. This is a prediction of a patient not having a stroke but they do. This would affect the stakeholders negatively because of the cost of taking care of this stroke patient.

False Positives: When it comes to our model False Positives are second worst. This a prediction of a patient having a stroke but they do not. This would affect the stakeholders negatively just on reputation of the model alone. Just some happy patient’s after getting the worst news possible.

Final Recommendation

The balance of the data is very unbalanced. I had to use boosting techniques to increase the dataset for better modeling. Over sampling with SMOTE and logistic regression is the best model that I found. This model ended up with good accuracy of about 92%.