Stroke Prediction

Thomas Bingamon

CONTEXT:

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

Our Target

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.

Challenges

Challenge 1

Unknown Smokers

Through our analysis, a number of patients in our study had checked off **“Unknown”** in their smoking status. Are they socially smoking or are they smoking habitually?

Challenge 2

Outliers

Although our data was mostly revised and cleaned, there were a few outliers in our research.

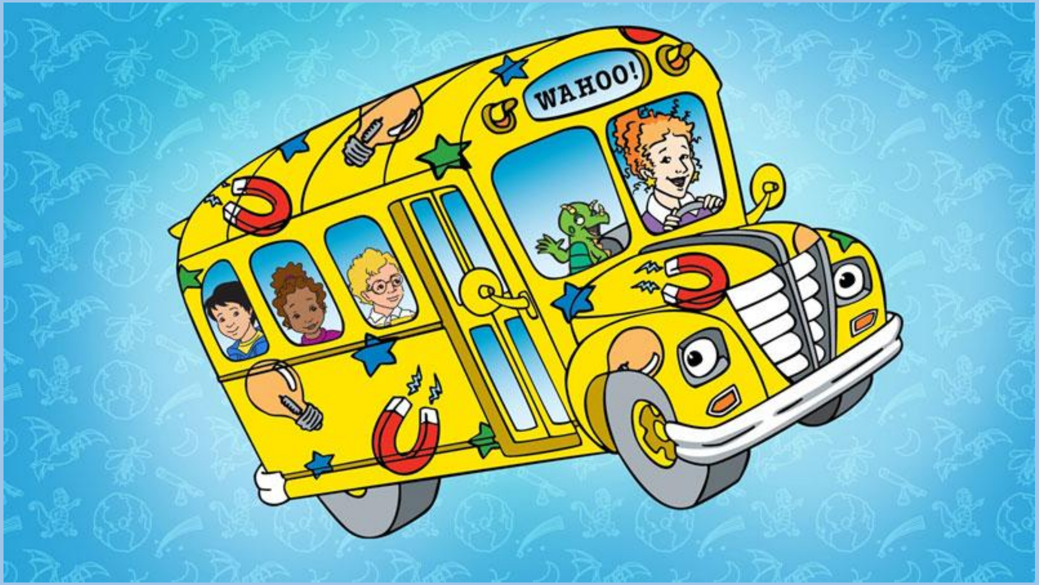
Challenge 3

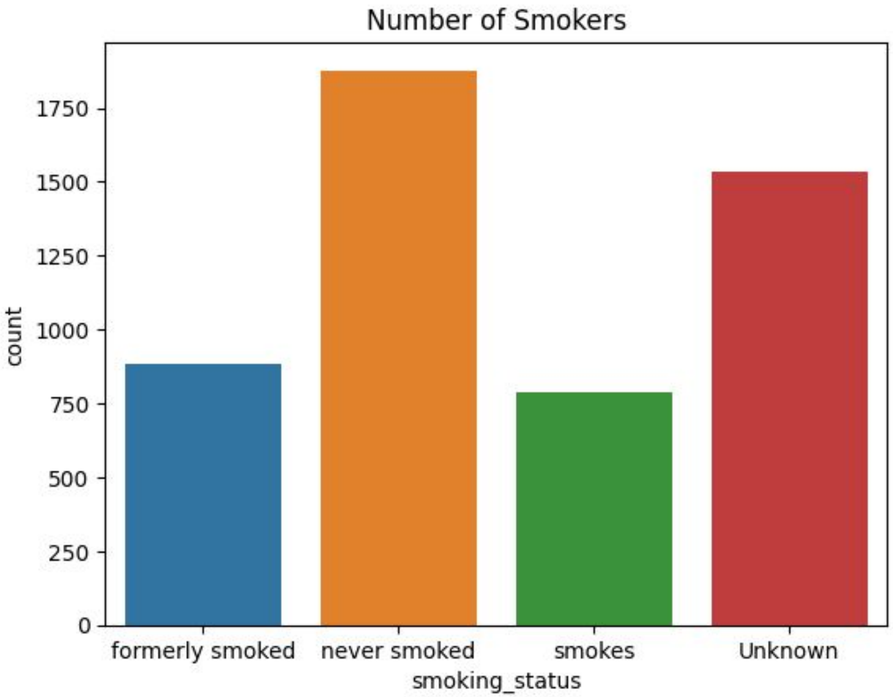
Will our predictive models work?

In various machine/predictive models, we were able to find a test model with high accuracy recall.

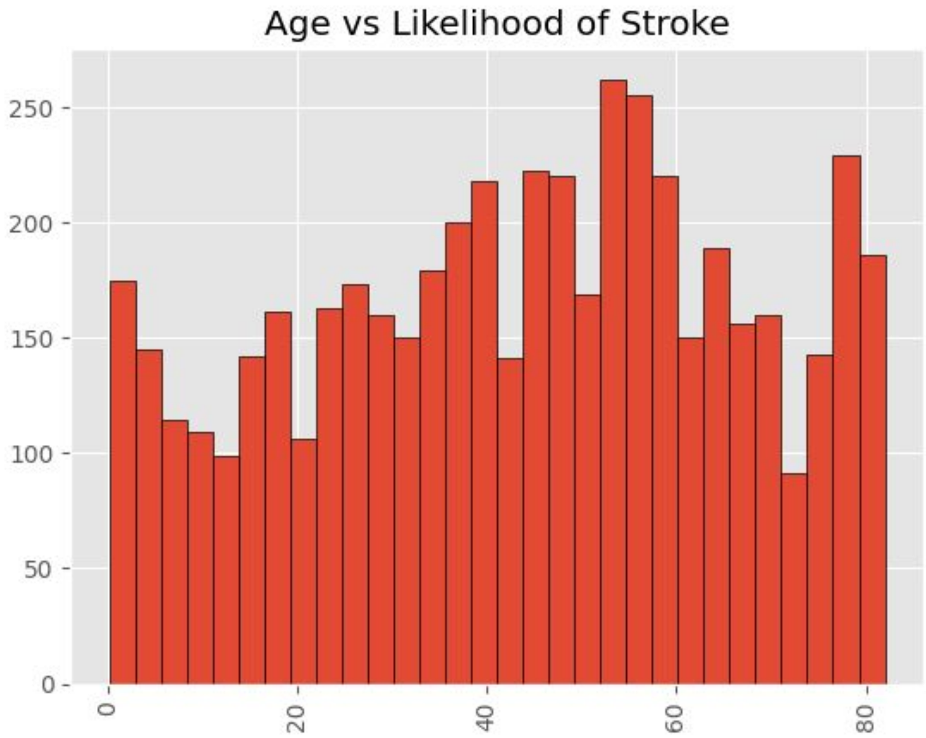
BUT FIRST...

Let us explore our findings

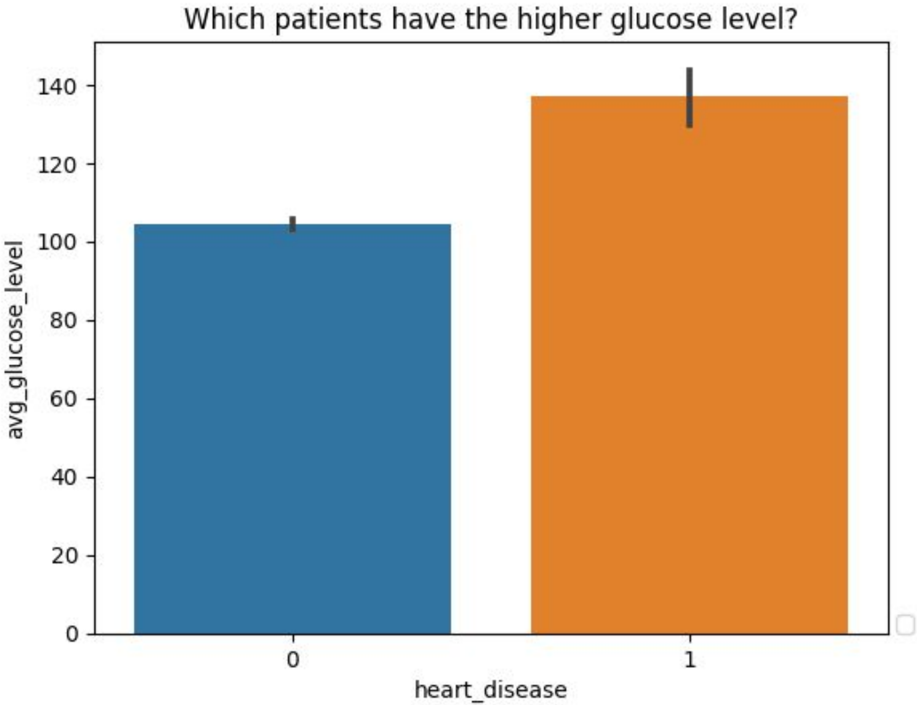




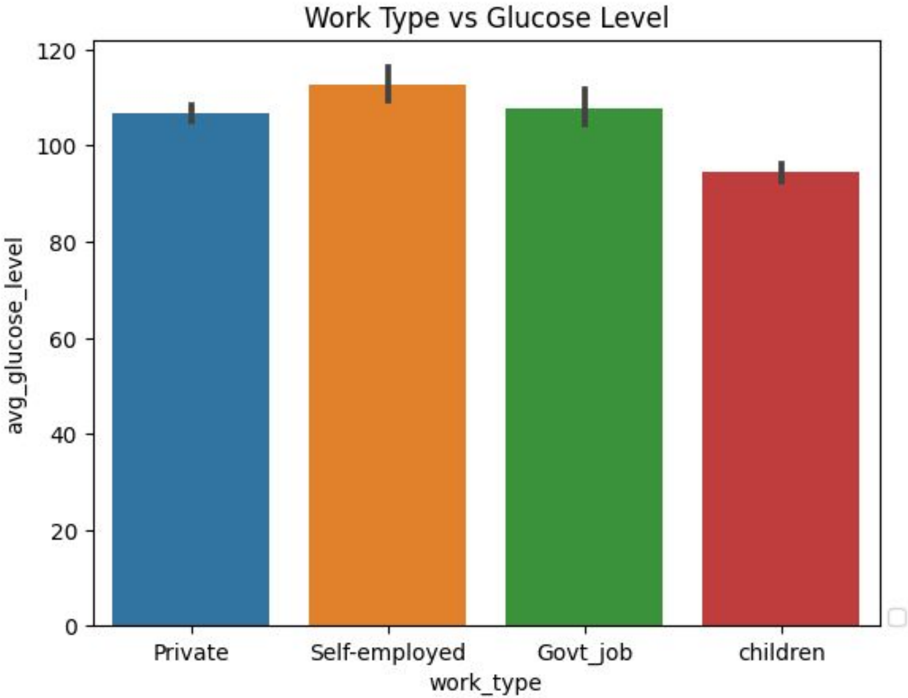
While our patients who never smoked were recorded the highest in our data, there are a significant amount of patients who marked “unknown”. We might conclude that these patients may smoke on a social level or may be not inclined to share their true smoker status.



In our findings we can see that starting from ages 50-80 there is an uptick in likelihood of getting a stroke. This could be for various reasons such as: higher cholesterol, glucose levels, unhealthy eating habits, less exercise, etc.



Our dataset has our patients labeled ‘0’ and ‘1’, we can see that the ‘1’ column has an exponential amount of higher glucose average than our patients in the ‘0’ column. We can delve deeper into this in the following slide.



In a more detailed graph, we can see how even work type may have effect on one’s glucose level. Children have the lowest glucose level, while those who are Self-employed have the highest.

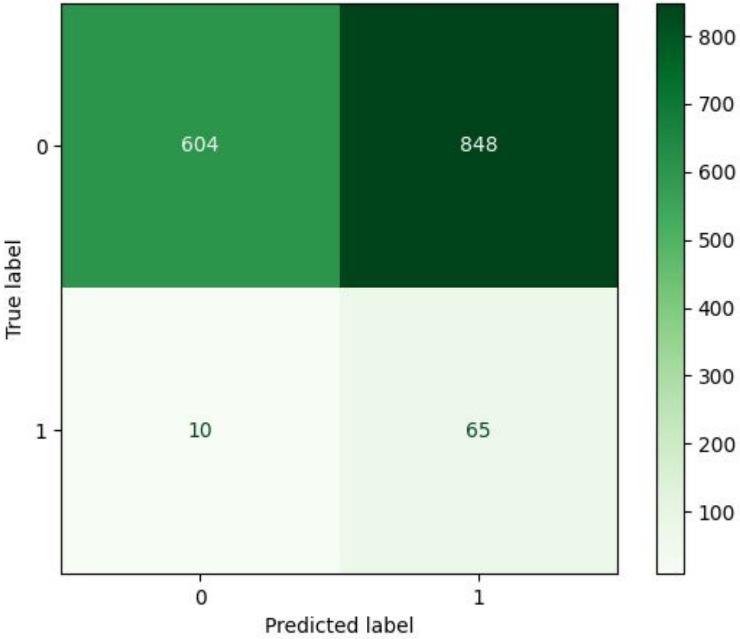
Implementation

Impact

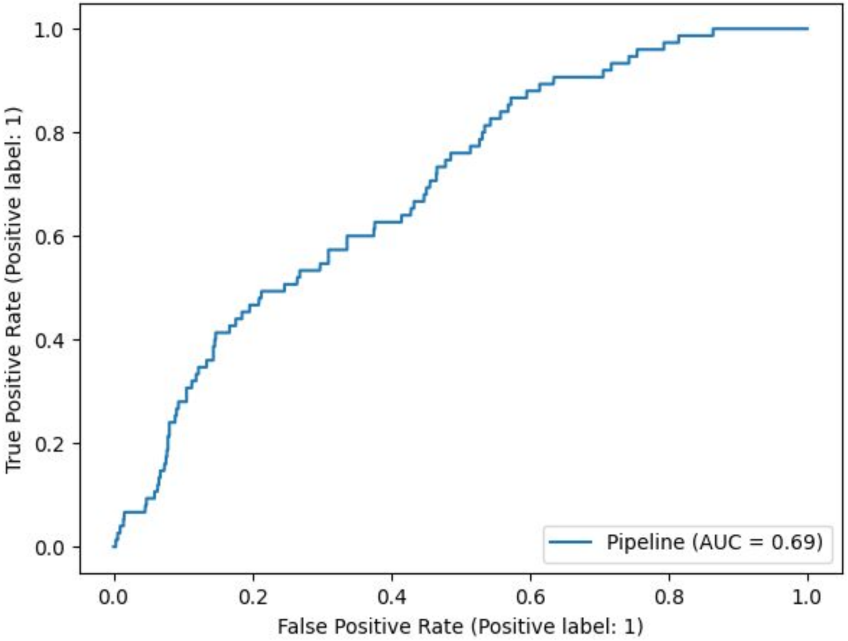
A false negative can have serious consequences, especially when positive outcomes represent critical situations(i.e. Having a stroke). Failure to detect positive instances result in missed opportunities, delays, or threats to safety or well-being. Relying on the model's predictions reduce trust and confidence if false negatives are frequent.

Impact

Minimizing the false negatives in our dataset is crucial. Stakeholders should prioritize recall or sensitivity to ensure the detection of positive cases, even if we get a higher false positive rate.



True label
Predicted label
0
1
0
1
604
848
10
65
800
700
600
500
400
300
200
100



True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Pipeline (AUC = 0.69)

Impact

Through various predictive models and machine learning tools, we were able to find a suitable predictive model that gave us a very high recall percentage of 87%, but most importantly a very low false negative.

In Conclusion:

With a low false negative rate, positive instances are being reliably detected, reducing the need for manual review or reassessment of cases that could be false negatives.

By decreasing missed opportunities, we can make more informed decisions based on a much deeper view of the data, leading to better outcomes and providing resources where needed.