Prediction of Stroke

Jonathan Nunez

For Use by Medical Professionals

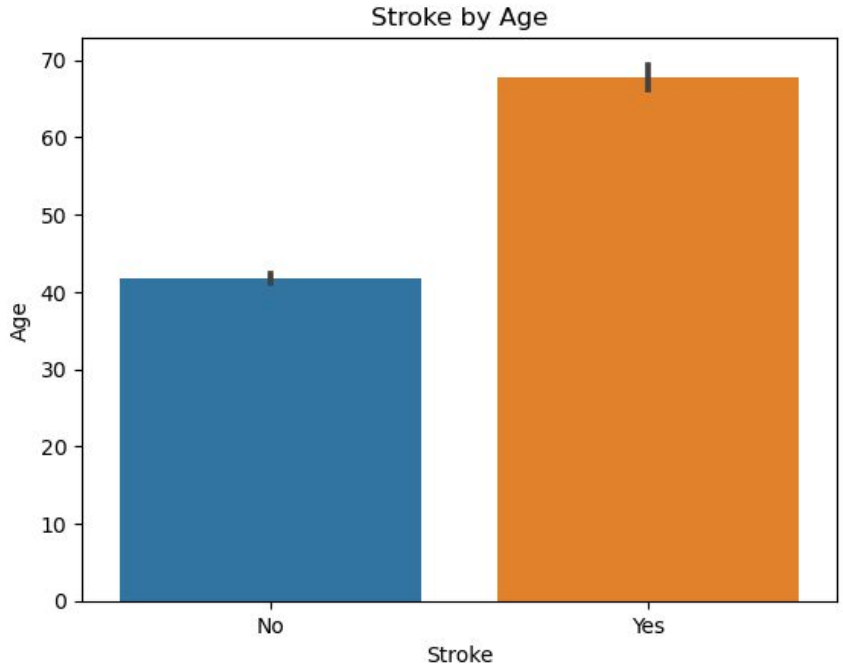
- This model can be used by various medical professionals that deal with diagnosing Strokes in patients
- Works as a supplement to normal diagnostic procedures
- Diagnosing a patient can be difficult and ambiguous
    - This model works to alleviate some of that ambiguity by pointing the doctor in the right direction

The Data

- This is a smaller dataset with 5110 entries
- The dataset is very unbalanced
  - Which is to say, there are much more instances of Negative Stroke results than Positive Stroke results
- The model will be updated as more data becomes available
- Covers features such as:
  - Gender
  - Age
  - Hypertertension (yes/no)
  - Heart Disease (yes/no)
  - BMI, etc
  - As well as the target feature: Stroke (yes/no)

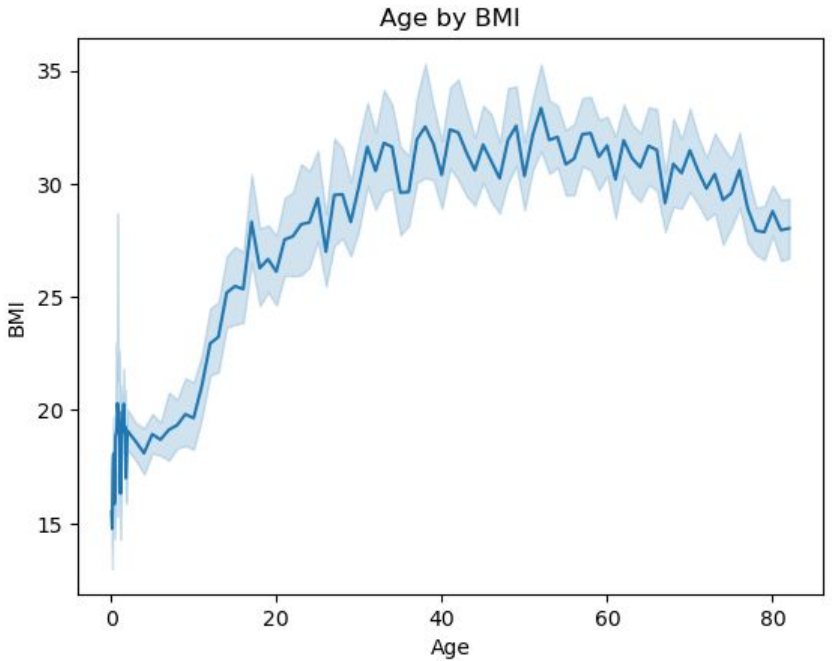
Visual: Stroke by Age

- People that get strokes are, on average, older; around 65-70 years old
- People that don't get strokes are, on average, younger at around 40 years old
- Stroke is most strongly correlated with Age



Visual: Age by BMI

- Here we can see that as Age increases so does BMI
- Though, this does seem to plateau a bit at around 30 BMI
- Age was most strongly correlated with BMI
- Older people are more likely to get a Stroke and older people are also more likely to have a high BMI



Strengths and Limitations

- The best model overall is *Tuned Under Sampling Logistic Regression*

Strengths:

  - Has an accuracy score of 72%
  - Predicts Positive samples correctly at a rate of 75%
  - I believe I struck a good balance between Accuracy of the entire model while at the same time increasing True Positives

Limitations:

  - Since the data is very unbalanced, it's inherently going to give us misleading results
    i. We will get a lot of False Negatives
  - Mitigated this issue by trying to re-balance, but there are limitations to that as well

Final Recommendations

- I recommend the *Tuned Under Sampling Logistic Regression*
- It provided the best balance between Accuracy and True Positives
- Will need to continue feeding the model more entries/samples to improve these metrics
- This is a particularly challenging dataset due to such a high imbalance between Negative results (95.74%) and Positive Results (4.26%)