Can we predict whether or not a patient will have a stroke?

Using a Random Forest Classifier or XGBoost Classifier
On the Stroke Prediction Dataset

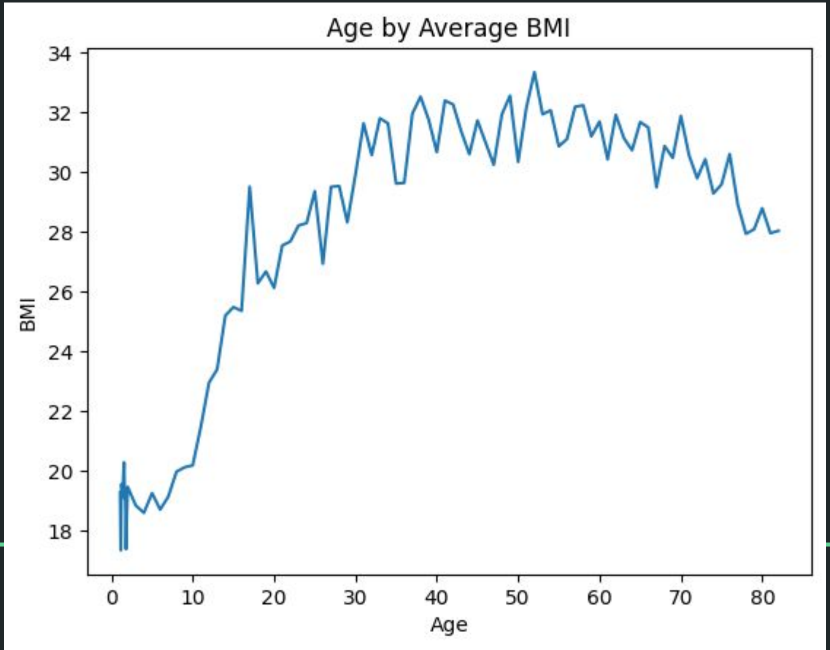
Goal:

Our goal is to use patient information like: gender, age, occupation, bmi, smoking status, etc., to predict whether or not a future patient will have a stroke.

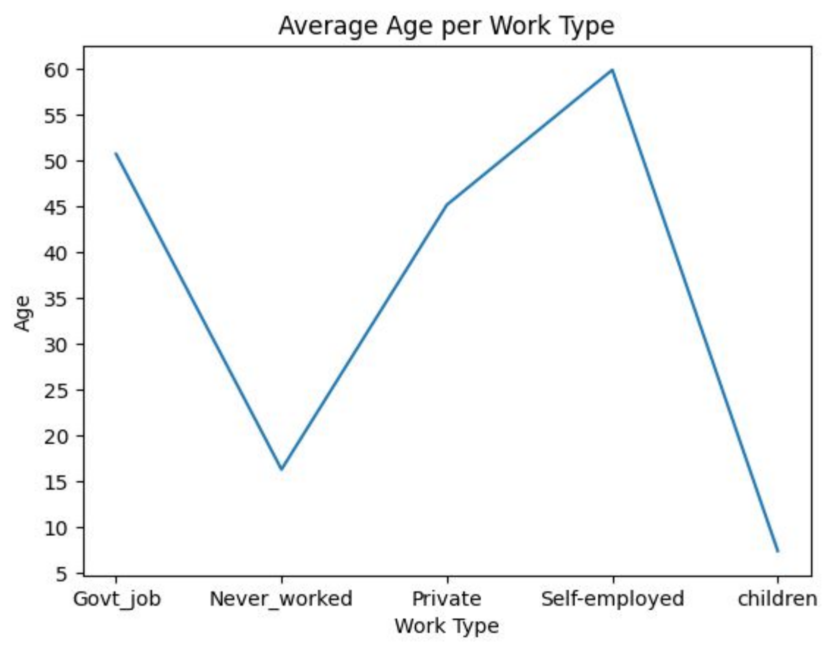
Brief Overview:

- Two models, the XGBoost Classifier, and a Random Forest Classifier were turned used to try and predict whether or not a patient will have a stroke.
- Our models were very adept at learning from the patient information, just not as well at predicting stroke.
- Both of our models underwent tuning before settling on a choice model.
- All patient information (features) collected in the data have very low correlation to each other and to the target, ‘stroke.’ Despite a lack of correlation across the board we are able to get an idea of direct and indirect correlation using the feature most highly correlated with our ‘stroke’ target: Age

Patient Age grouped by average BMI:



Insights:
This plot shows a positive correlation between age and average bmi. The data shows that on **average bmi increases as age does.** Of all features, 'age' has the strongest correlation with our target, 'stroke' (shown in heatmap). In a way, this demonstrates that having a higher BMI increases the risk of having a stroke.



Patient Average Age Grouped by Patient Occupation

Insight:

This plot shows that patients with presumably more stressful occupations are also older on average. This further shows why there is as strong of a connection between age and stroke. **In addition to high bmi being associated with high age, high age is generally accompanied by a more stressful occupation**.

Models:

XGBoost Classifier:

Limitations - This model doesn't excel in any particular area. It is likely to produce both many Type 1 and Type 2 errors. Patients would be incorrectly labeled 'likely' and 'unlikely to have a stroke.'

Random Forest Classifier :

Strengths - Less Type 1 Errors - Will predict true positives much more often than the XGBoost model. This means the model will more often correctly predict when an patient in fact WILL get a stroke.

Limitations - Even after tuning, model has many issues correctly predicting when a patient will NOT get a stroke in patients. Meaning patients may be falsely labeled 'not likely to have a stroke' (Type 2 Errors are as common as in XGBoost model).

Model Recommendations:

- If given the choice to implement either the XGBoost Classifier or the Random Forest Classifier, the most logical choice would be the Random Forest Classifier.

- This is simply due to the fact that, while both models will struggle equally in correctly predicting when a patient WILL NOT have a stroke, the Random Forest Classifier will have much less trouble in predicting when a patient WILL have a stroke. In this area, it has no competition from the XGBoost Classifier.

Final recommendations:

It would seem that after looking at a correlation heatmap, that there is not a very strong correlation between features at all. The features we have been able to draw connections from have a very small correlation to each other so it’s honestly a bit of a stretch to assume that the relationship between patient information (features) and stroke is strong enough for a model to learn from to be useful in predicting stroke in the future.

I would strongly recommend that more *quality* data be retrieved. Data with patient information that at the very least has a moderate correlation to our ‘stroke’ target. Generally, when two more-or-less powerful machine learning models can’t make predictions very well after tuning, there might be a lack of predictive power in the actual dataset.