Building a Stroke Classifier Model with Machine Learning

-Ryan Bernstein

What do we know about strokes?

- Typically caused by either insufficient blood flow to the brain or by excessive bleeding within the brain
- Best indicator of stroke risk is high blood pressure
- More than 795,000 Americans have a stroke every year
- Individuals who suffer a stroke are highly likely to develop a serious long-term disability
- A good place to find data on stroke victims is Kaggle.com!

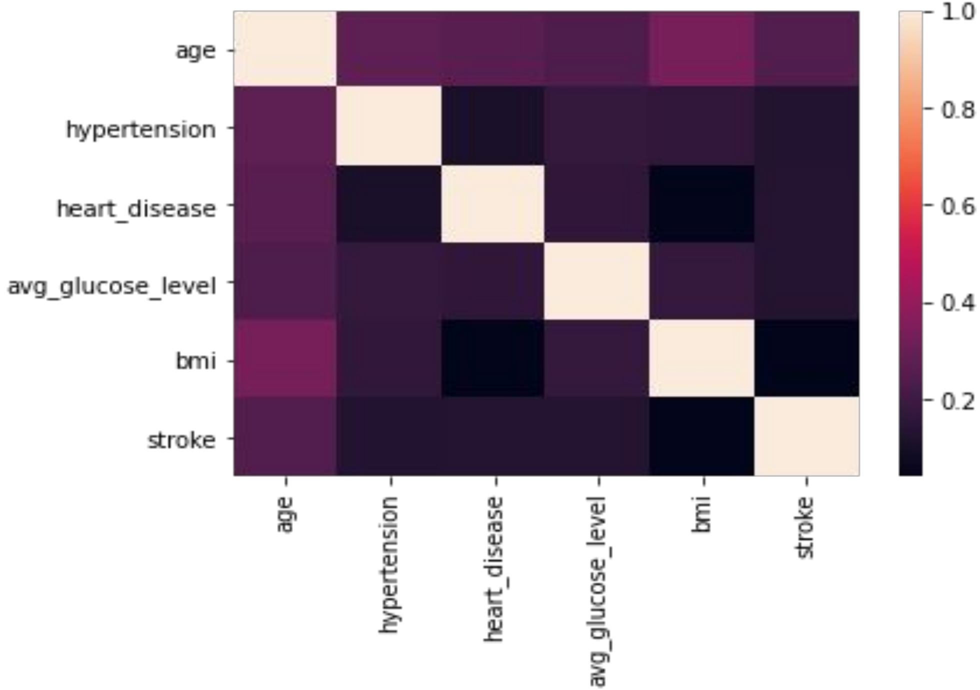
First Look At the Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

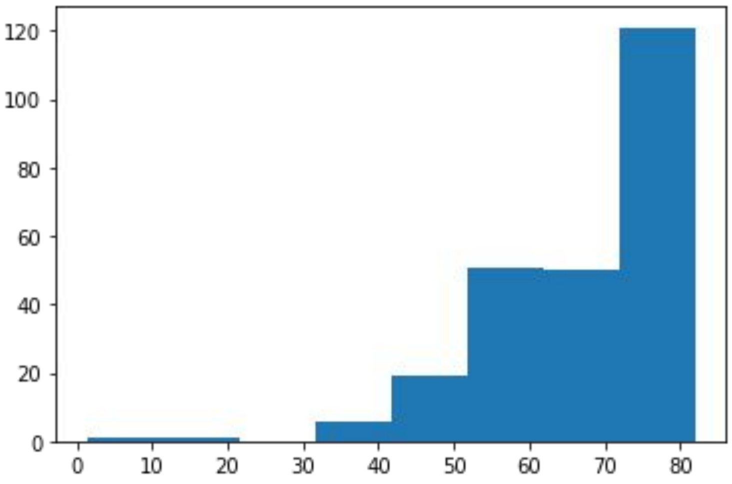
- 3 Float Variables
- 3 Indicator Variables
- 5 Categorical Variables
- 1 Variable with Null Values

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

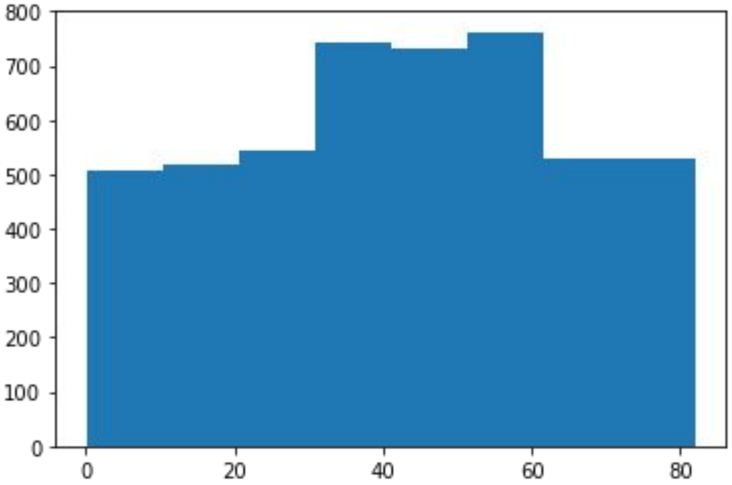
Looking For Correlations



The Impact of Age on Stroke Incidence:



Age and Frequency of those who Had a Stroke



Age and Frequency of those who Did Not

Preprocessing

- Transform categorical variables into dummy variables. This makes it easier to find where our correlations lie.
- Split data into a training set and a test set with a 70/30 split. This allows us to avoid overfitting the model on the original data
- Standardize the features with MinMaxScaler. This helps us more accurately measure the impact of our numerical features
- Impute missing values for BMI with the median BMI so that we don't have to get rid of those observations

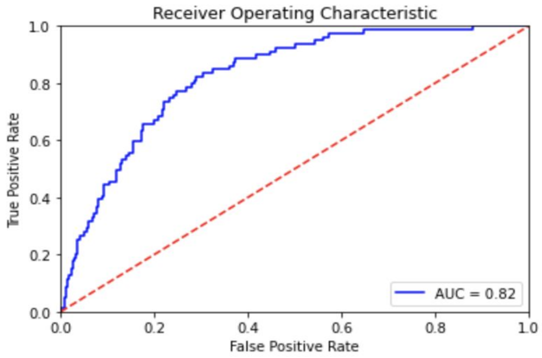
Modeling Process

- Compared two different modeling methods: Logistic Regression and Random Forest
- First ran both models without using any hyperparameters
- Next ran both models with the class_weight hyperparameter
- Finally, tuned both models by using GridSearchCV on a hyperparameter grid
- Displayed 6 metrics: accuracy, F1 Score, a confusion matrix, a classification report, average precision score, and the Matthews Correlation Coefficient
- Generated a Beeswarm Plot

Classification + ROC for Original Models

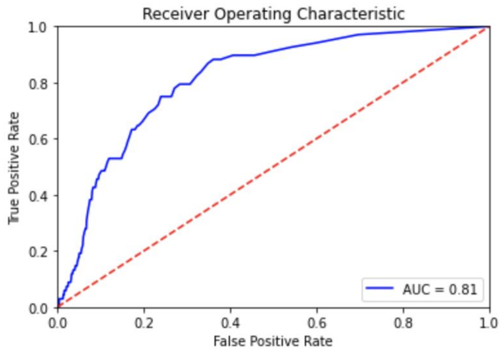
Logistic Regression

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1454
1	0.00	0.00	0.00	79
accuracy			0.95	1533
macro avg	0.47	0.50	0.49	1533
weighted avg	0.90	0.95	0.92	1533



Random Forest

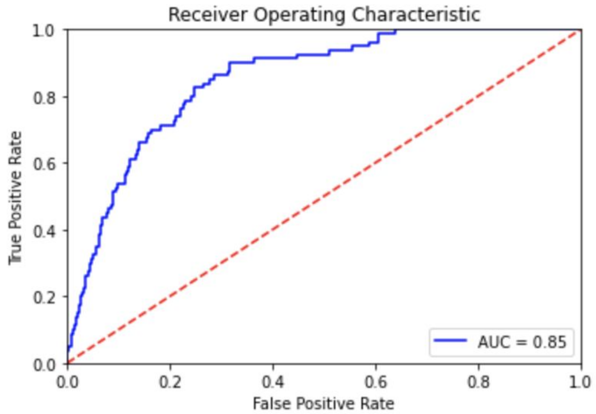
	precision	recall	f1-score	support
0	0.96	1.00	0.98	1465
1	0.00	0.00	0.00	68
accuracy			0.95	1533
macro avg	0.48	0.50	0.49	1533
weighted avg	0.91	0.95	0.93	1533



Classification + ROC for class_weight 1:99

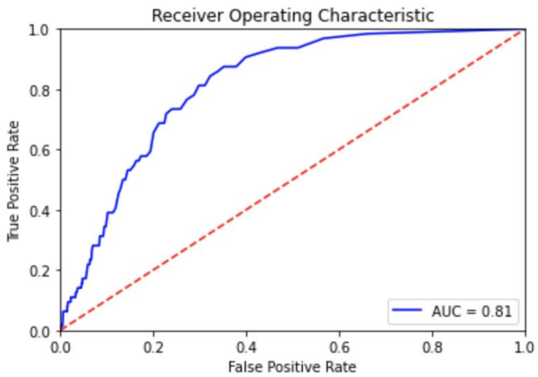
Logistic Regression

	precision	recall	f1-score	support
0	0.99	0.46	0.63	1453
1	0.09	0.94	0.16	80
accuracy			0.49	1533
macro avg	0.54	0.70	0.39	1533
weighted avg	0.95	0.49	0.60	1533



Random Forest

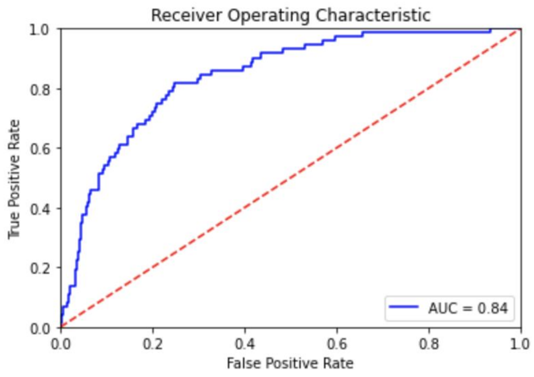
	precision	recall	f1-score	support
0	0.95	1.00	0.97	1450
1	0.00	0.00	0.00	83
accuracy			0.95	1533
macro avg	0.47	0.50	0.49	1533
weighted avg	0.89	0.95	0.92	1533



Classification + ROC for tuned class_weight

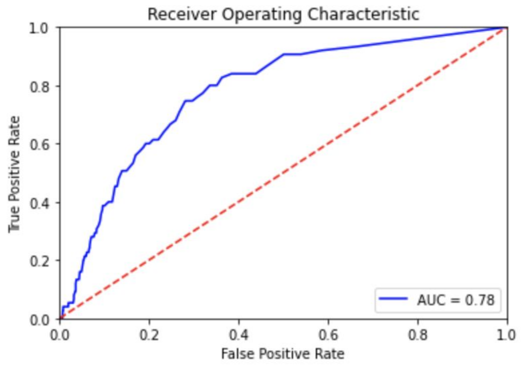
Logistic Regression

	precision	recall	f1-score	support
0	0.99	0.67	0.80	1461
1	0.11	0.85	0.20	72
accuracy			0.68	1533
macro avg	0.55	0.76	0.50	1533
weighted avg	0.95	0.68	0.77	1533

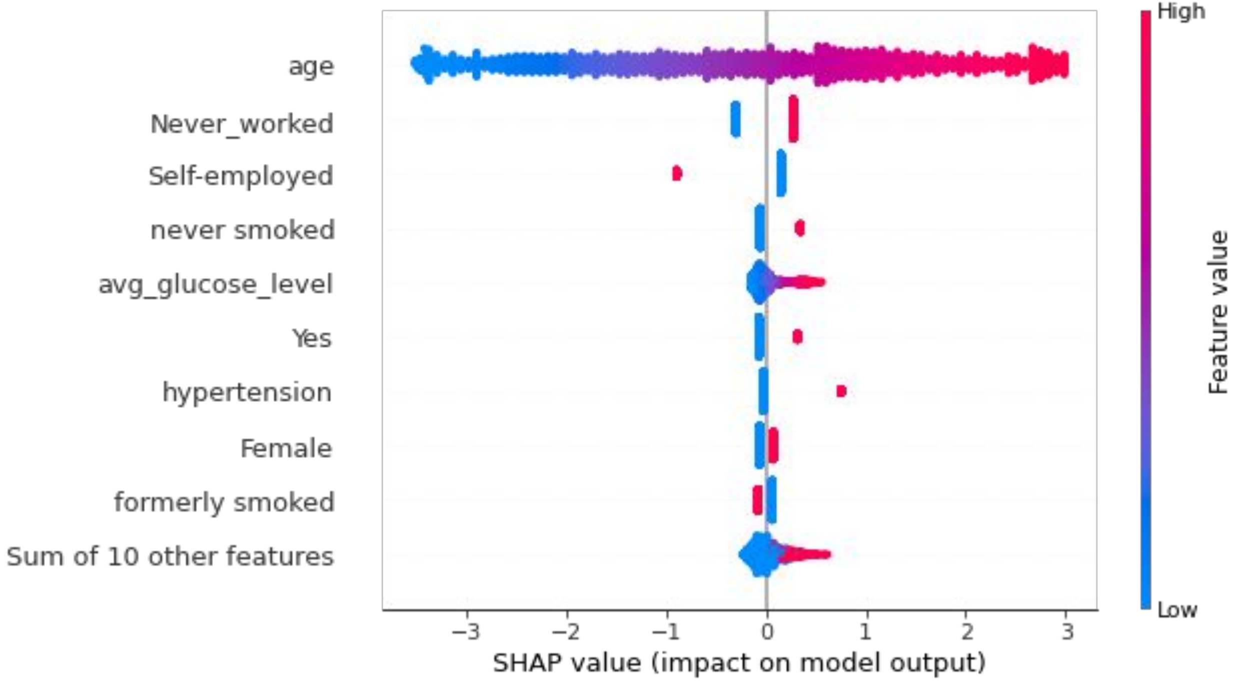


Random Forest

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1458
1	0.00	0.00	0.00	75
accuracy			0.95	1533
macro avg	0.48	0.50	0.49	1533
weighted avg	0.90	0.95	0.93	1533



Beeswarm Plot



Conclusions

- The best performing model was Logistic Regression with a 1:25 class weight. This model had a recall of .85 for positives, which means that it accurately identified 85% of positive cases (while misidentifying 15% of positive cases as negative cases).
- Not many features on their own had strong correlations with the stroke target variable. Age was the best performing feature by far.
- It seemed that the tangible health metrics provided had a similar impact on stroke risk as the provided categorical features that focused on a patient’s personal life had on stroke risk.

Future Research

- Add more features? Look for more data? Consider diversity of location or ethnicity?
- What would happen if we took preemptive action? Would the patients who underwent preemptive action have a lower incidence of stroke?