Predicting Strokes

Rhoads MacGuire

Problem

- According to the CDC over 800,000 people in the United States have a stroke each year.
- Strokes cost the US economy an estimated 34 billion dollars so there is a large monetary incentive to prevent strokes before they occur.

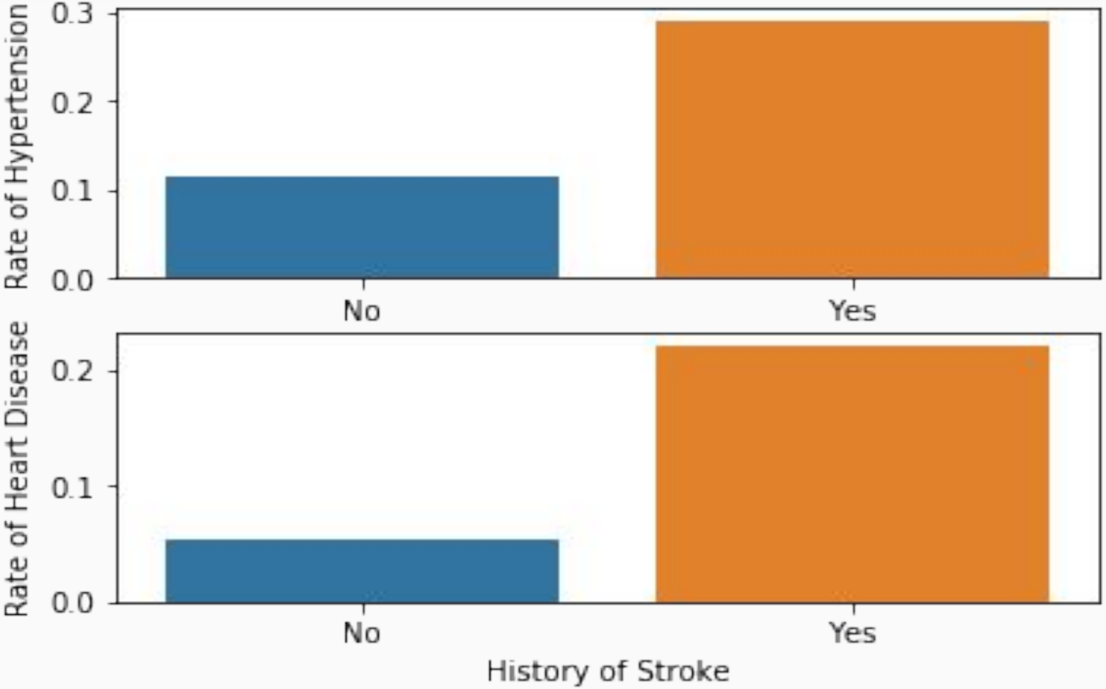
Proposal

- If we could predict the occurrence of stroke in individuals based on common health indicators we could save the money and lives of Americans
- Institutions such as the government or life insurance companies would also pay for this information

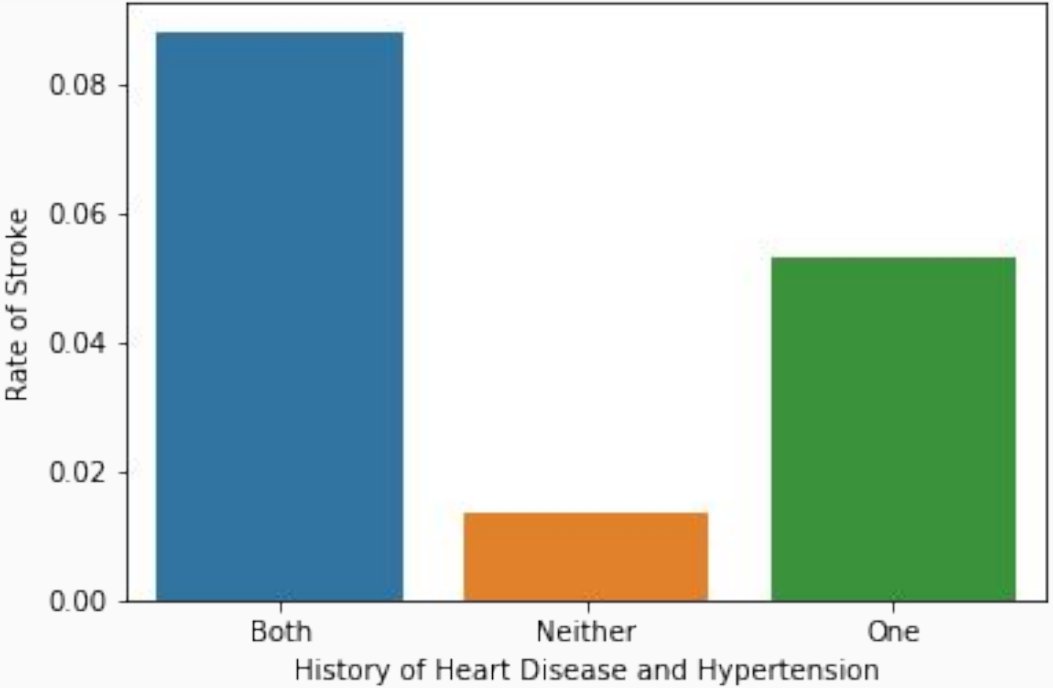
Data Wrangling

- Data was acquired from Kagle so it did not require much cleaning
- Eliminated subjects under 19 because the occurence of stroke was so low and smoking data was inconsistent.
- Replaced string variable with categorical ones

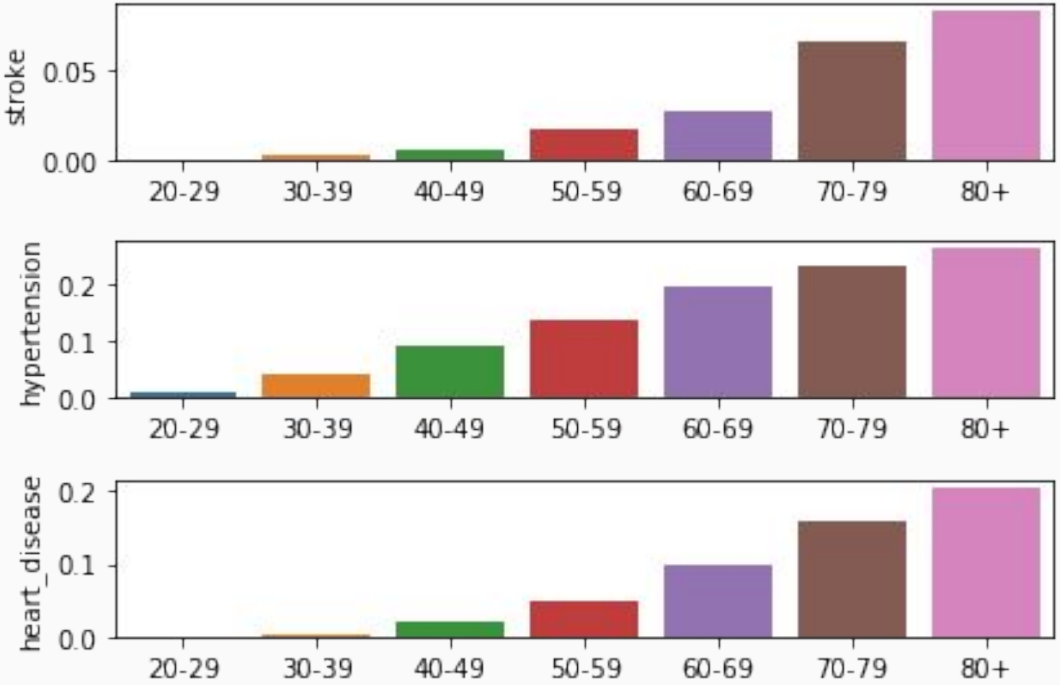
EDA- Heart Disease and Hypertension



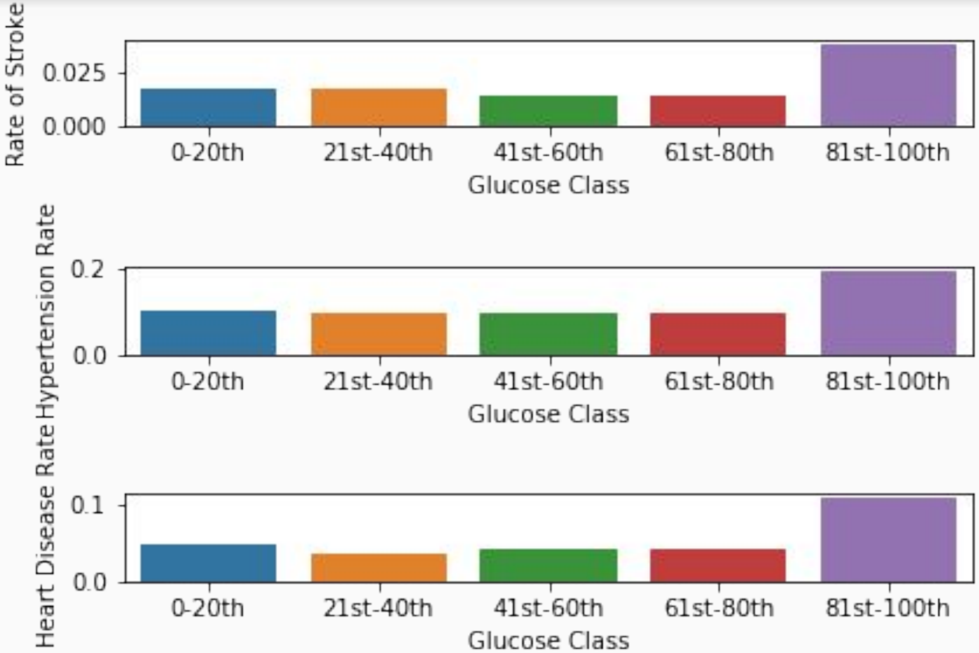
EDA- Heart Disease and Hypertension



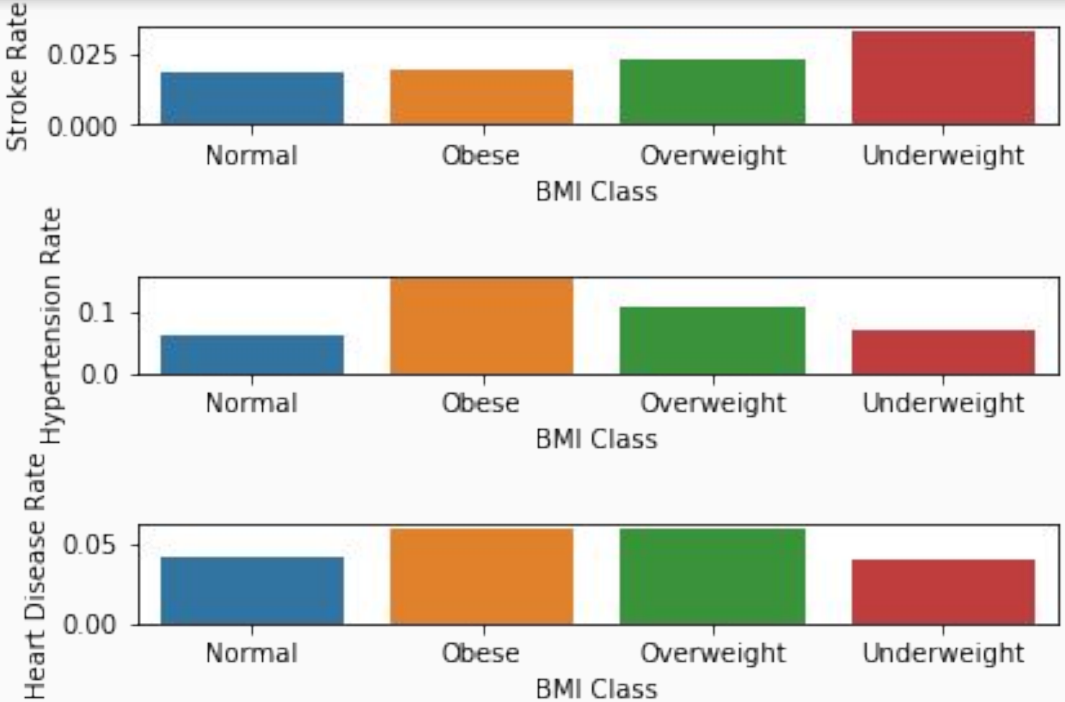
EDA- Age


stroke
0.05
0.00
20-29
30-39
40-49
50-59
60-69
70-79
80+
hypertension
0.2
0.1
0.0
20-29
30-39
40-49
50-59
60-69
70-79
80+
heart_disease
0.2
0.1
0.0
20-29
30-39
40-49
50-59
60-69
70-79
80+

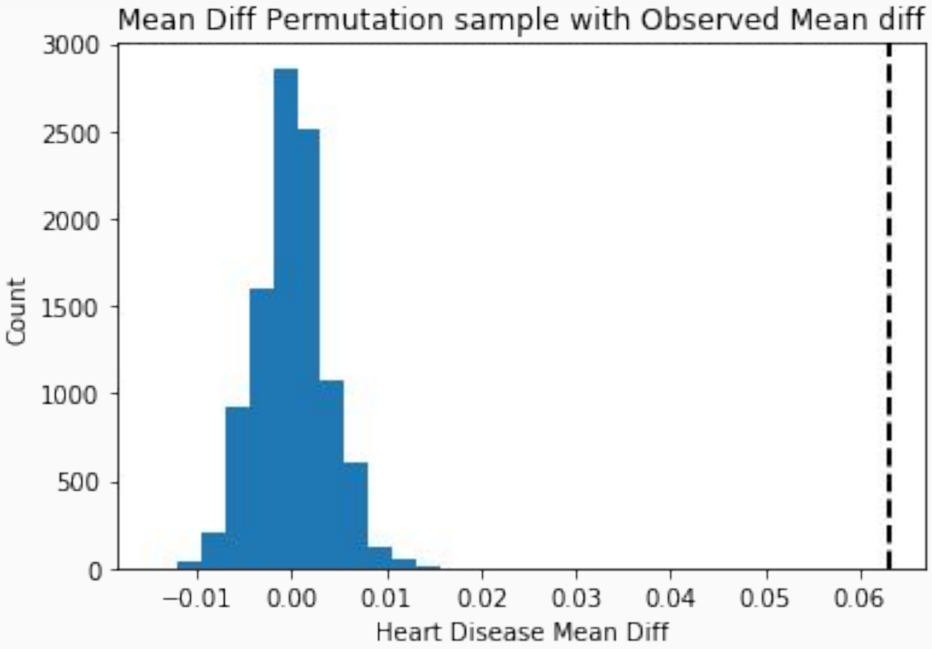
EDA- Glucose Level

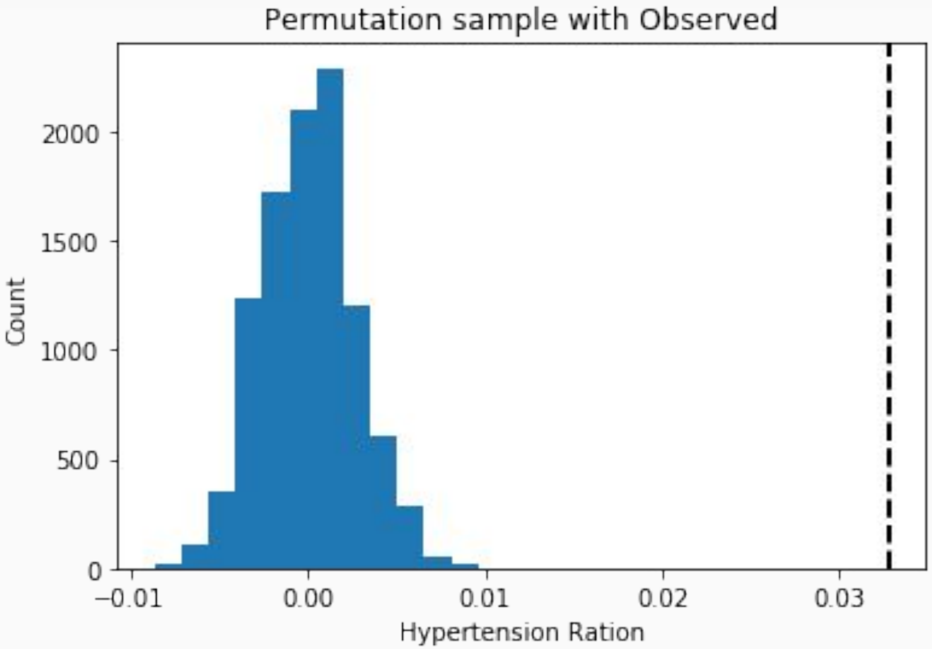


EDA- BMI


Stroke Rate
0.025
0.000
Normal
Obese
Overweight
Underweight
BMI Class
Hypertension Rate
0.1
0.0
Normal
Obese
Overweight
Underweight
BMI Class
Heart Disease Rate
0.05
0.00
Normal
Obese
Overweight
Underweight
BMI Class

Inferential Stats



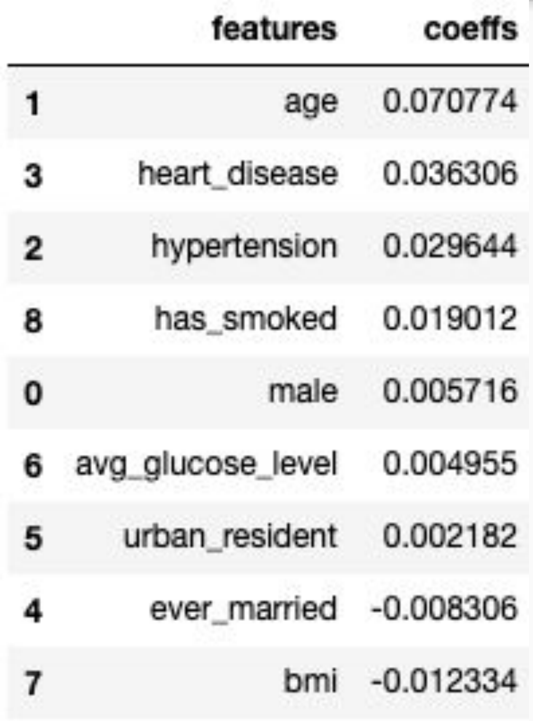


Model

- Tried a variety of models including, Logistic Regression, Random Forest and Gradient Boosted Trees
- None effective at predicting stroke
- While the accuracy was just under 98% the recall wa 0% for all models, oce tuned.

Model Insights

- While the model had little predictive power, it does confirm age, heart disease and hypertension were our most impactful features.



	features	coeffs
1	age	0.070774
3	heart_disease	0.036306
2	hypertension	0.029644
8	has_smoked	0.019012
0	male	0.005716
6	avg_glucose_level	0.004955
5	urban_resident	0.002182
4	ever_married	-0.008306
7	bmi	-0.012334

Conclusions

- While the model is not predictive, the analysis sheds useful insights such as, anyone with hypertension or heart disease is at increased risk for stroke
- BMI and glucose level should be monitored in comparison to others because obese people and those in the top 20th percentile of glucose level are at increased risk of stroke

Improvements

- Turning glucose level groups and BMI groups into categoricals to replace number values
- Possibly finding more medical features rather than personal ones. It appears the models suffered from a lack of features.