Heart Disease: Using Prediction for Prevention

January 30, 2023

Background

“Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year...Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths” (WHO, 2022). Great Lakes Hospital is starting a new campaign to prevent heart disease by identifying those with the greatest risk and intervening before a CVD occurs.

Understanding the problem

The Question

What can be developed that informs Great Lakes Hospital whether a person will have a cardiovascular disease based on their characteristics, habits, or other factors?

The Data

Dataset from the CDC taken from a nationwide telephone survey in 2020 about US residents' health status

The Deliverables

Jupyter notebooks of data wrangling through modeling, as well as this final report and presentation

The Model

Binary supervised classification

- heart disease = 1
- no heart disease = 0

Naive Bayes

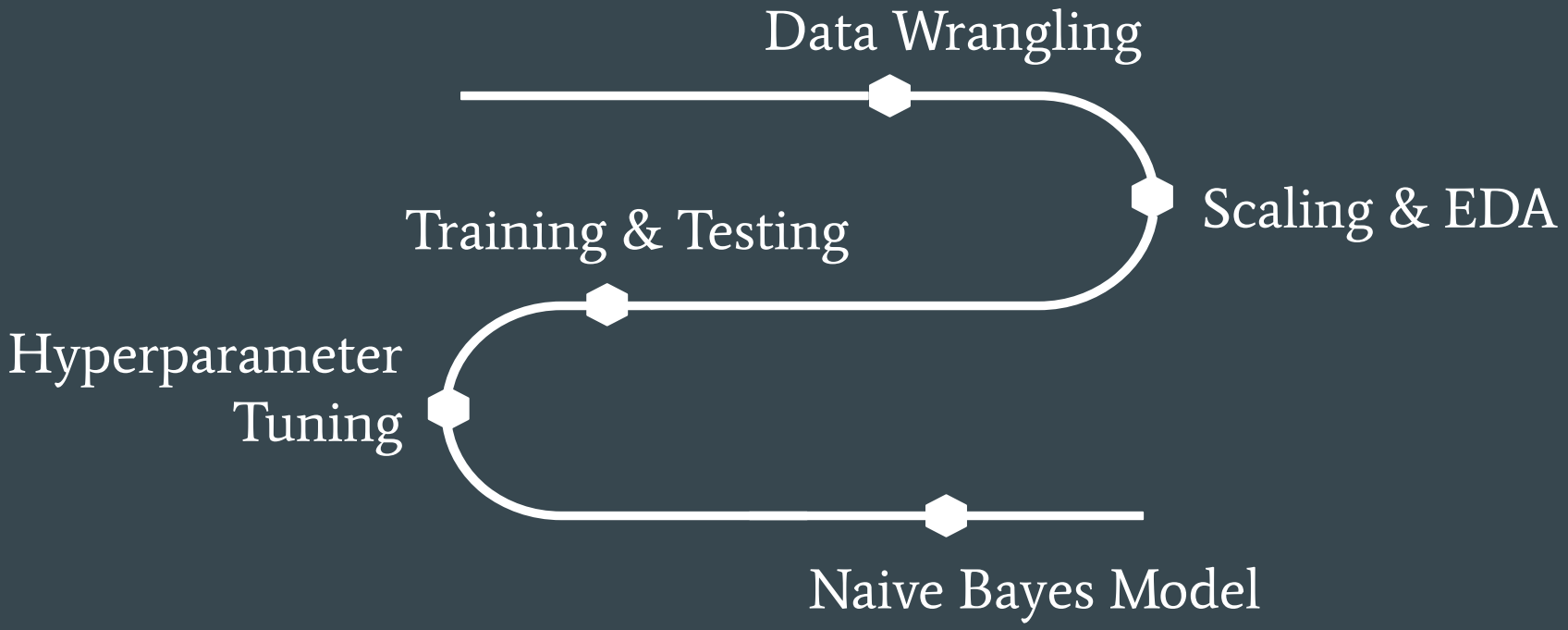
- Recall score = 70%
- ROC AUC score = 82%

The Model

Training vs Test Data Comparison

	Data Used	Recall Score	ROC AUC Score
0	Training Data	0.697689	0.816630
1	Test Data	0.699726	0.816876

The Process

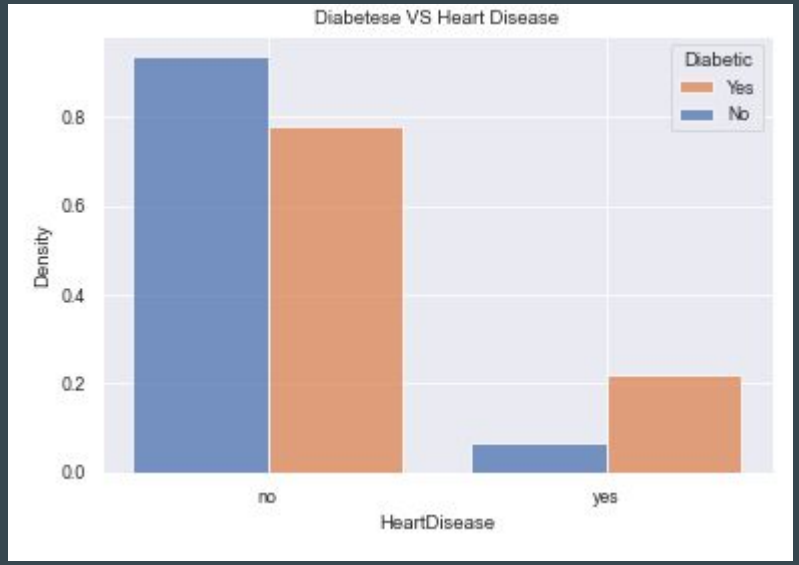

Data Wrangling
Scaling & EDA
Training & Testing
Hyperparameter Tuning
Naive Bayes Model

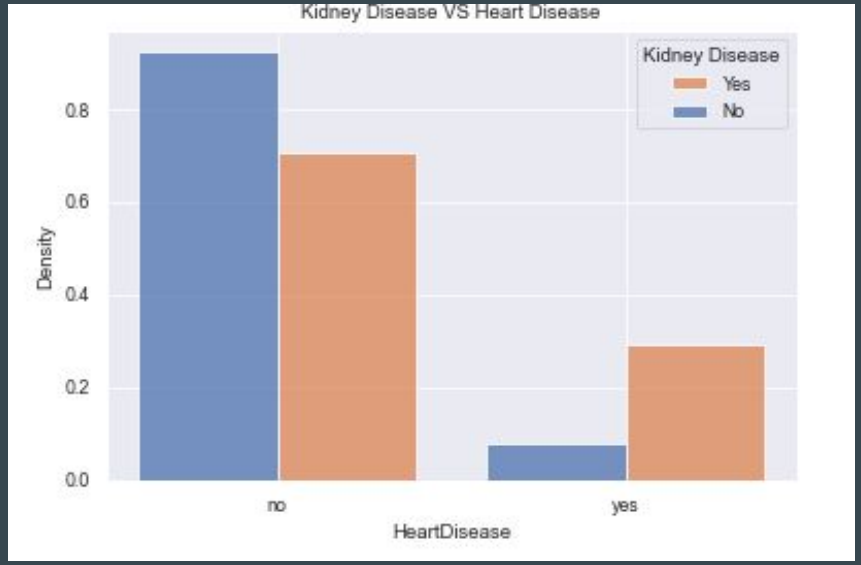
EDA Results

People with heart disease tend to

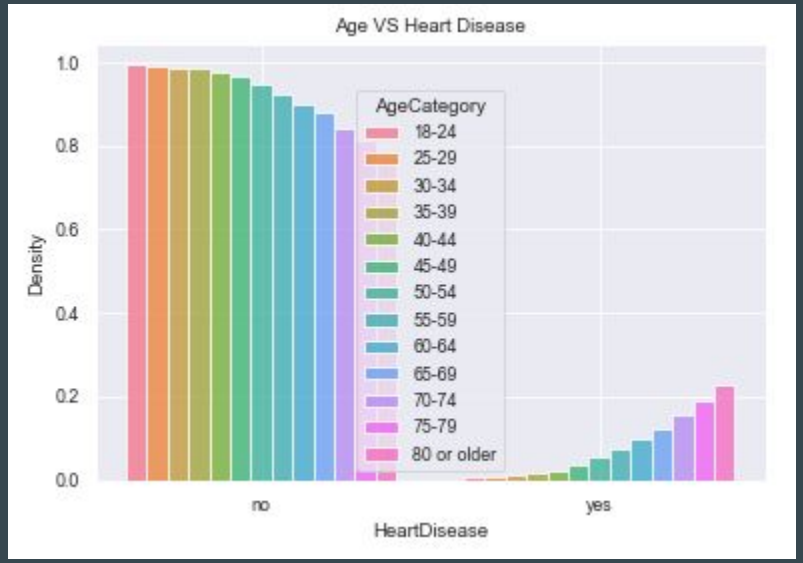
- Have higher rates of stroke, diabetes, skin cancer, kidney disease, and asthma
- Be older
- Smoke more
- Be male
- Have more bad physical health dats a month
- Have high rate of poor or fair health
- Drink less alcohol
- Have more difficulty walking
- Have a higher BMI

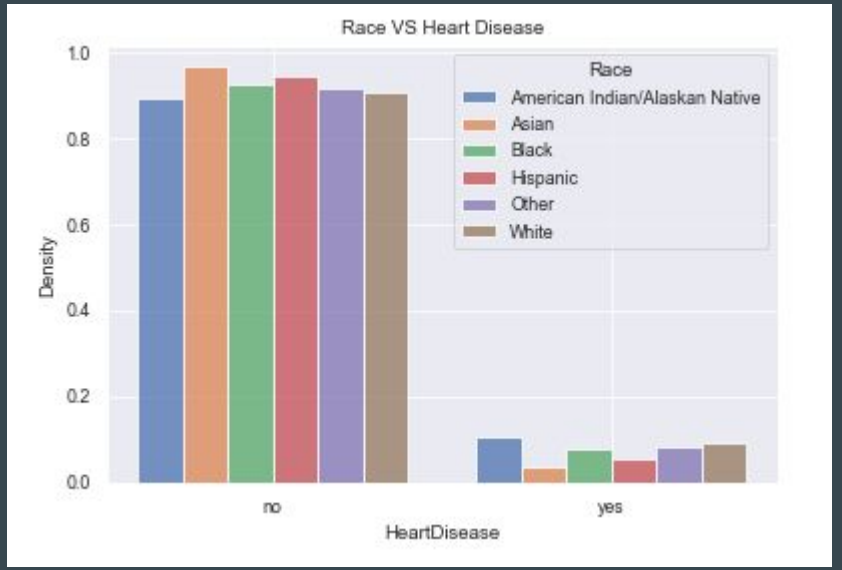
EDA Results


Diabetese VS Heart Disease
Diabetic
Yes
No
Density
0.0
0.2
0.4
0.6
0.8
no
yes
HeartDisease


Kidney Disease VS Heart Disease
Kidney Disease
Yes
No
Density
0.0
0.2
0.4
0.6
0.8
no
yes
HeartDisease

EDA Results

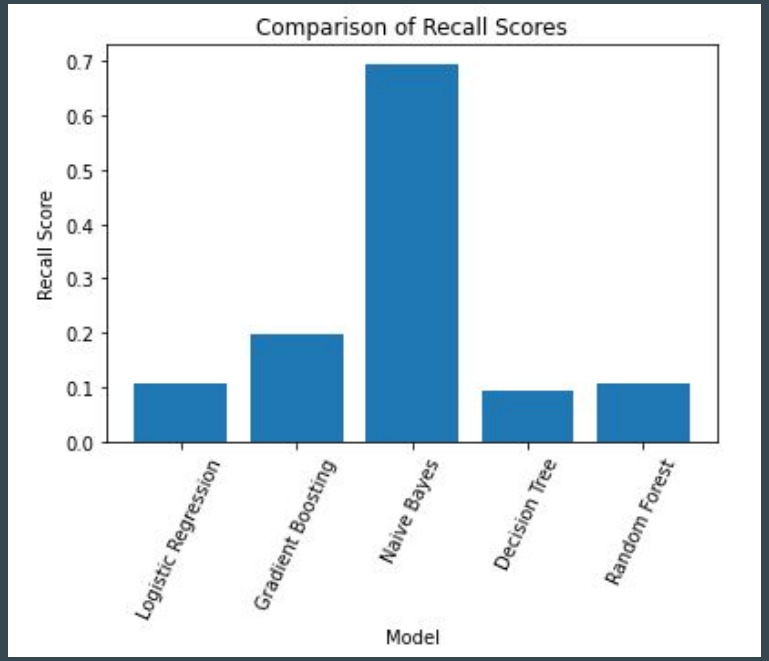

Age VS Heart Disease
Density
1.0
0.8
0.6
0.4
0.2
0.0
AgeCategory
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
no
yes
HeartDisease

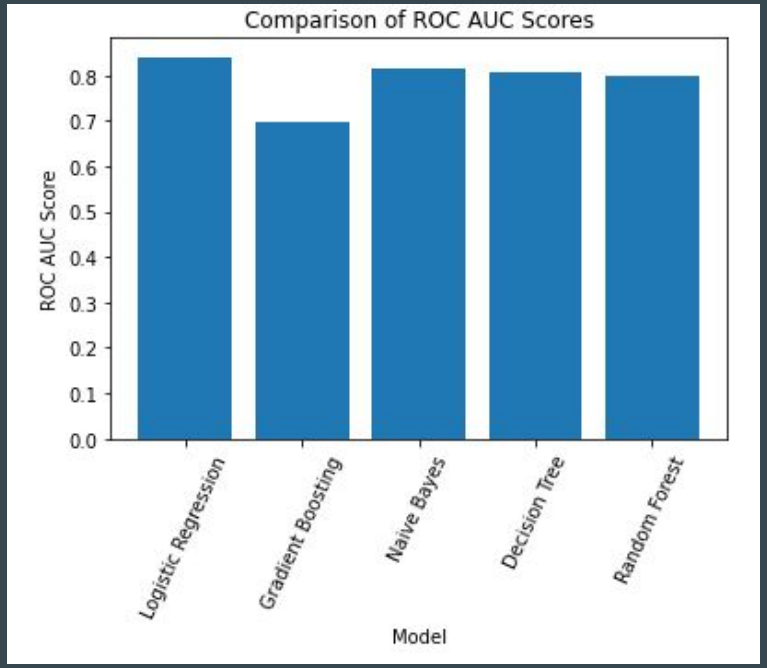

Race VS Heart Disease
Density
1.0
0.8
0.6
0.4
0.2
0.0
Race
American Indian/Alaskan Native
Asian
Black
Hispanic
Other
White
no
yes
HeartDisease

Other Models Evaluated

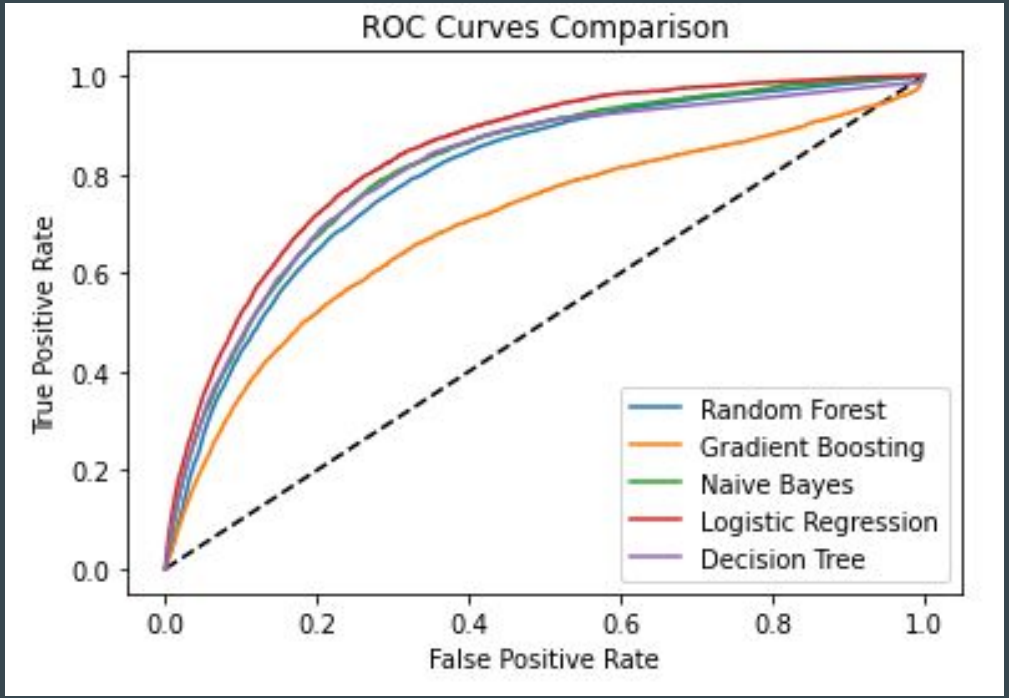
Logistic Regression	• Recall score: 11% • ROC AUC score: 84%
Gradient Boosting	• Recall score: 20% • ROC AUC score: 70%
Decision Tree	• Recall score: 9% • ROC AUC score: 81%
Random Forest	• Recall score: 11% • ROC AUC score: 80%

Model Metrics





Model Metrics



Next Steps

Collect More Data

Gather data from the surrounding community

Further Unbiased Data:

- Blood pressure
- Pulse
- Cholesterol

Integrate the Model

Allow for input of patient data to determine the chance of heart disease

Potential Rollouts:

- Internal hospital software
- Patient portal

Thanks

Andrea Ricci

Data Scientist

andreaeddy10@gmail.com

All slides, notebooks, functions, and models are available for future use.