Catching Heart Disease Early

A classification model to predict instances of heart disease

Opportunity

⅓ of all deaths in the United States are caused by Heart Disease

877,500 deaths per year

$216 Billion in healthcare costs per year from heart disease

Reduce healthcare costs by predicting instances of heart disease with easy-to-obtain information

Data from the Center for Disease Control

Data

Center For Disease Control Behavioral Risk Factor Surveillance System

Survey Data from over 300,000 American respondents in 2020

Contains Demographic Data, History of Certain Medical Conditions and Behavioral Data

Use Heart Disease Respondents as the target of a classification model

Model Selection

Logistic Regression

Random Forest

Extra Trees

Stacking Classification

Extreme Gradient Boosting

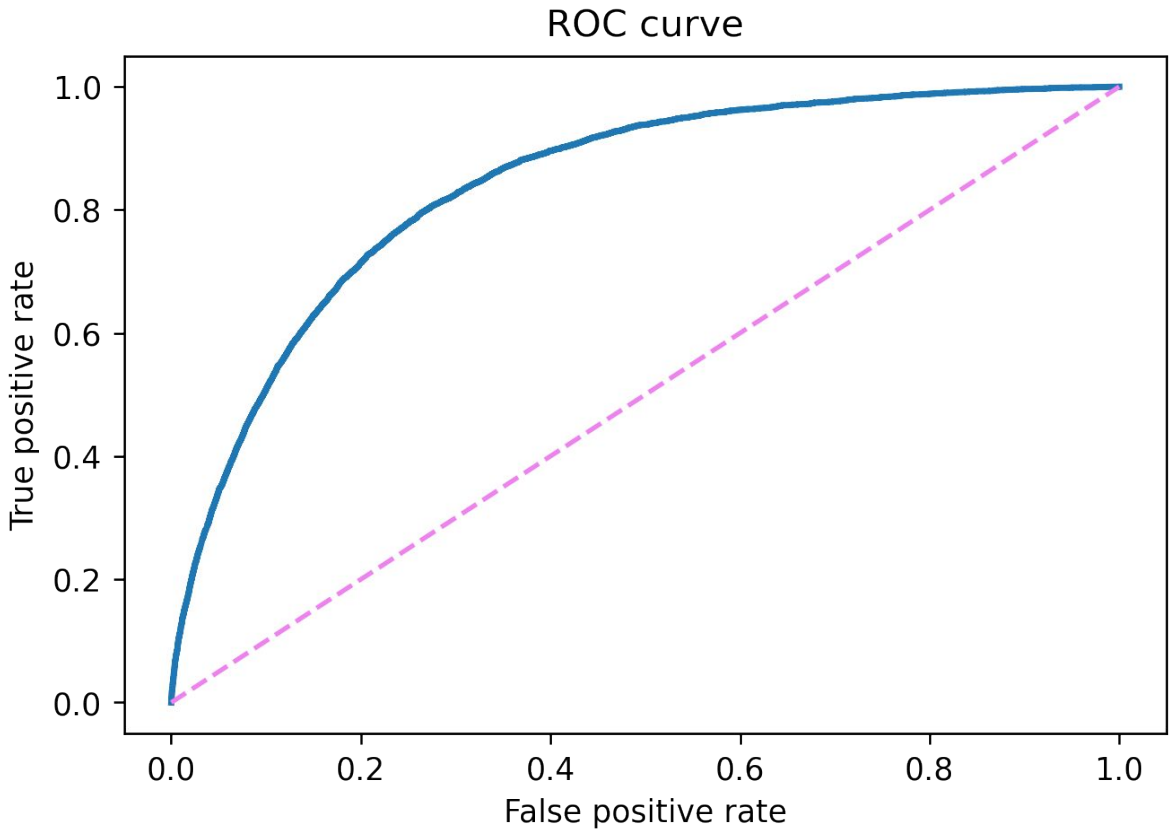
Final Model

Extreme Gradient Boosting

400 Estimators

Maximum Depth of 4

Learning Rate 0.05



Feature Relevance

Important

1. Difficulty walking upstairs
2. Diabetes
3. Poor general health
4. Age 80 or greater
5. Prior stroke

Less important

BMI

Physical activity

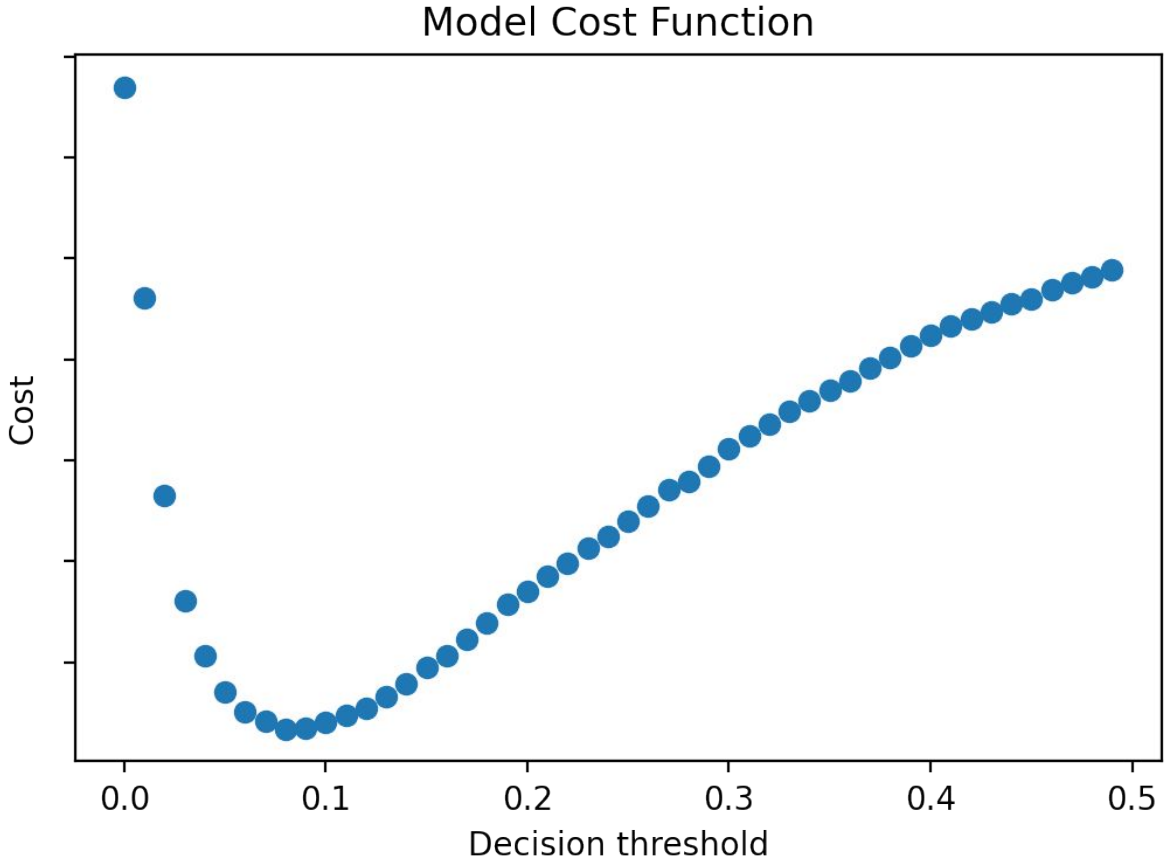
Sleep time

Decision Threshold

Assume cost of False Negative is 10x the cost of False Positive

Total Cost is minimized at p=0.08

Model will return a positive result if it estimates an 8% or more probability of Heart Disease



Model Performance

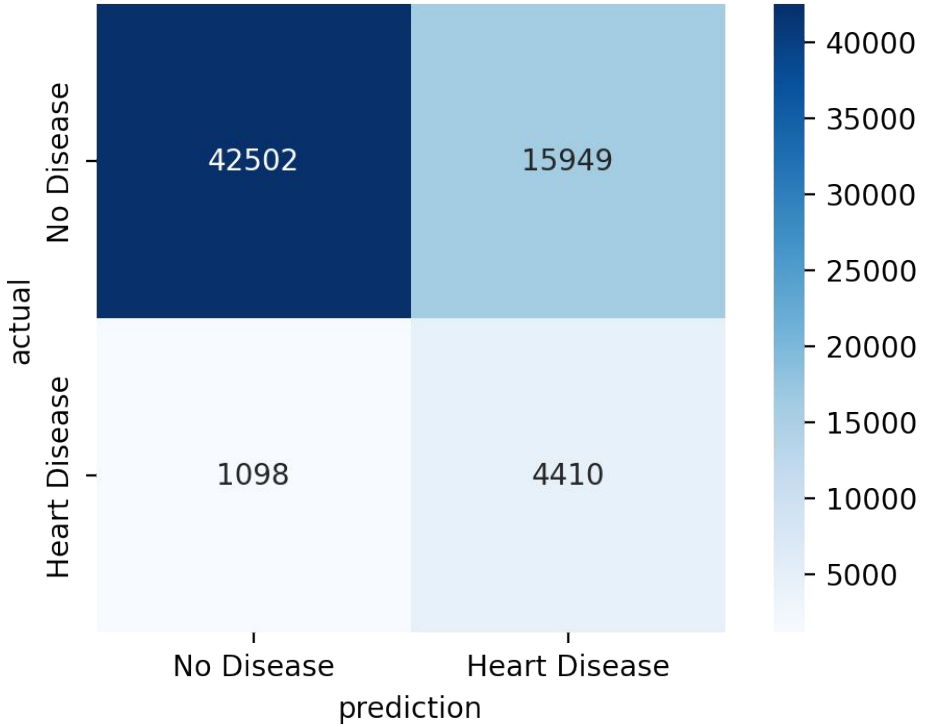
Accuracy: 73%

Recall: 0.80

Precision 0.22

F1 Score: 0.34

ROC AUC Score: 0.84



Conclusion

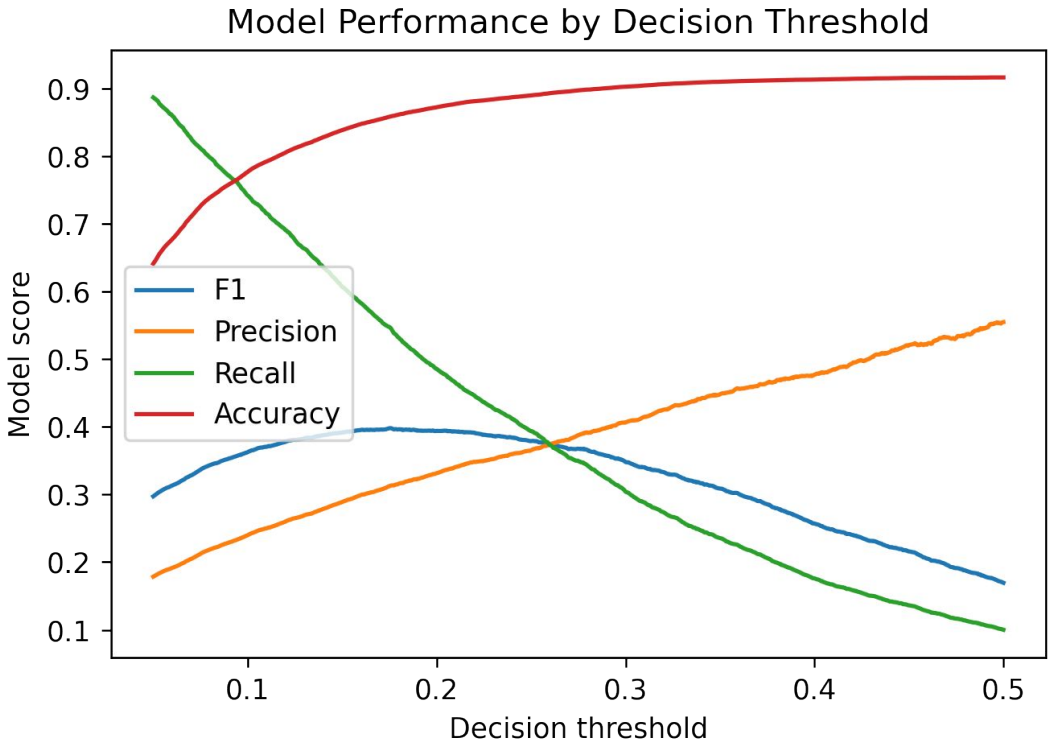
Test needs to be very sensitive in order to pick up instances of heart disease

Confirm positive results through additional testing

Model is easy to employ, and only requires easily available information

Improve model performance by incorporating additional features

Appendix 1


Model Performance by Decision Threshold
Model score
Decision threshold
F1
Precision
Recall
Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5

Appendix 2

	features	importance
6	DiffWalking	0.107653
7	Diabetic	0.073328
33	GenHealth_Poor	0.062990
25	AgeCategory_80 or older	0.060297
3	Stroke	0.056104
15	AgeCategory_30-34	0.055505
23	AgeCategory_70-74	0.052555
31	GenHealth_Fair	0.049897
24	AgeCategory_75-79	0.047021
1	Smoking	0.046568
32	GenHealth_Good	0.046492
14	AgeCategory_25-29	0.045915
16	AgeCategory_35-39	0.040979
11	KidneyDisease	0.034700
17	AgeCategory_40-44	0.031587
13	Sex_Male	0.029203
18	AgeCategory_45-49	0.026922
12	SkinCancer	0.020953
22	AgeCategory_65-69	0.020500
4	PhysicalHealth	0.012247
34	GenHealth_Very good	0.010842
19	AgeCategory_50-54	0.007167
21	AgeCategory_60-64	0.007010
2	AlcoholDrinking	0.006415
10	Asthma	0.005218
27	Race_Black	0.004180
5	MentalHealth	0.003614
0	BMI	0.003383
26	Race_Asian	0.003010
9	SleepTime	0.002931
29	Race_Other	0.002824
28	Race_Hispanic	0.002752
8	PhysicalActivity	0.002442
20	AgeCategory_55-59	0.001954