Early Detection of Heart Disease

Metis Classification Project

Learning the heart way

- Heart disease is the leading cause of death in the United States
- ~659,000 people in the United States die from heart disease each year
- Every 40 seconds an American will have a heart attack

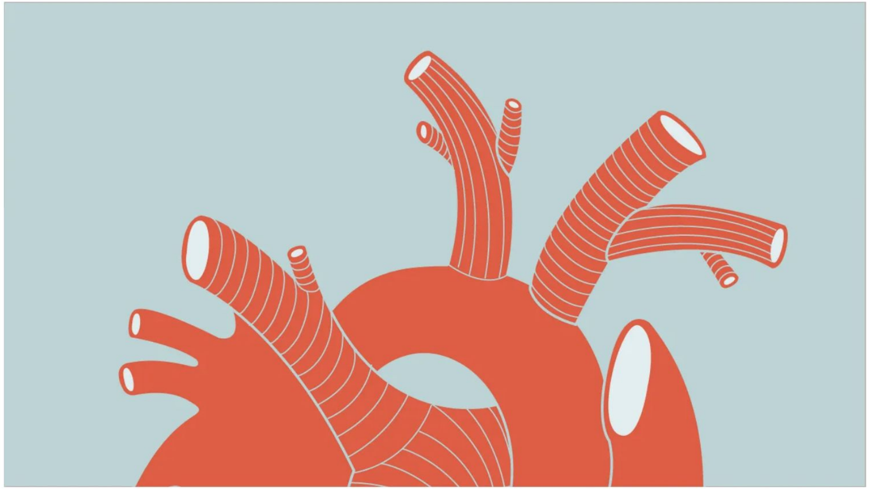


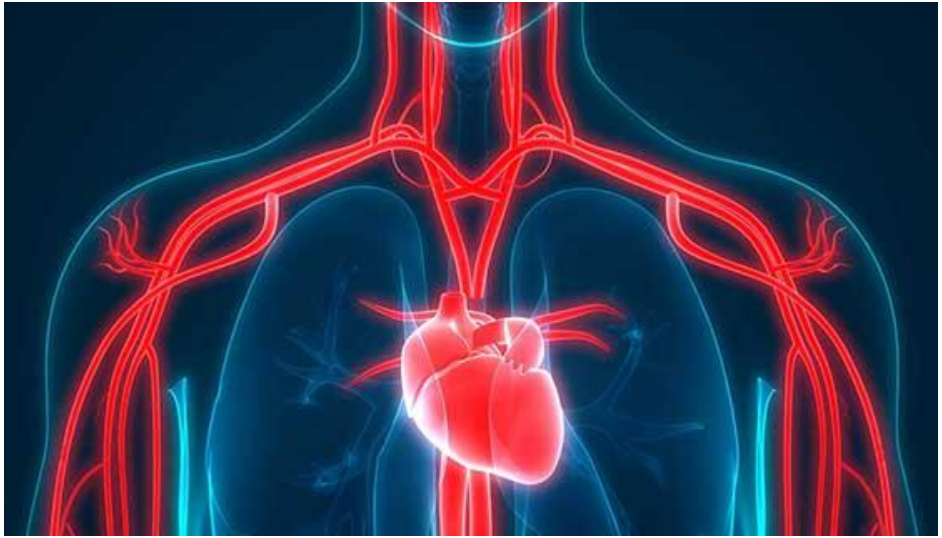
Illustration by Ruth Basagoitia and Maya Chastain

https://www.cdc.gov/heartdisease/facts.htm

Background

GOALS

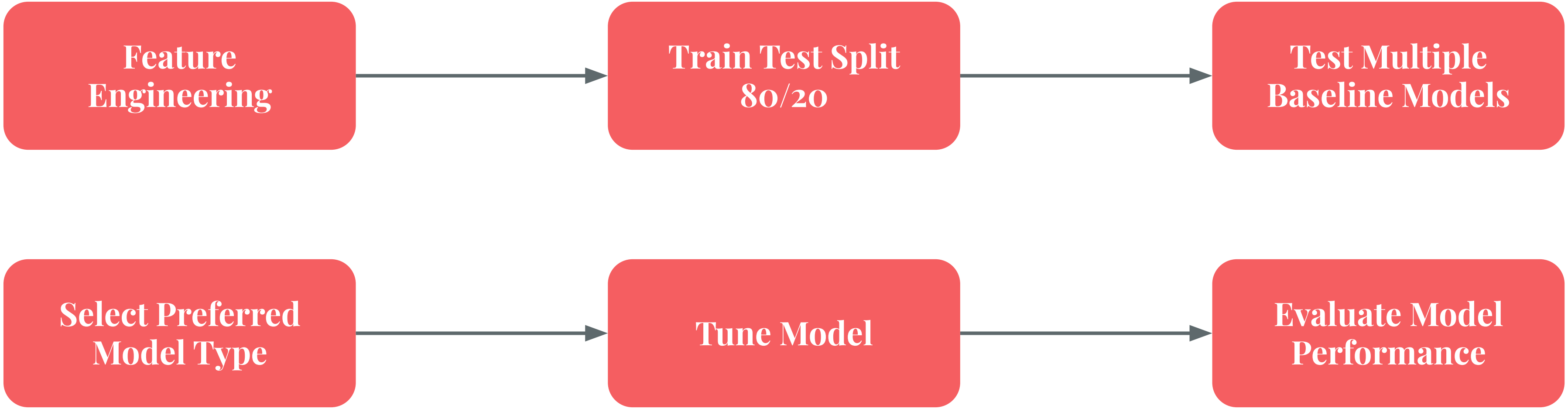
- Implement classification model to predict likelihood of heart disease or a heart attack based on other health and lifestyle characteristics
- Identify who to target for early detection and intervention
- Understand feature importance to inform screening



DATA

- Obtained from Kaggle
- Comes originally from the CDC
- 22 different attributes for 253,680 individuals

Modeling Approach



Feature Engineering

Train Test Split 80/20

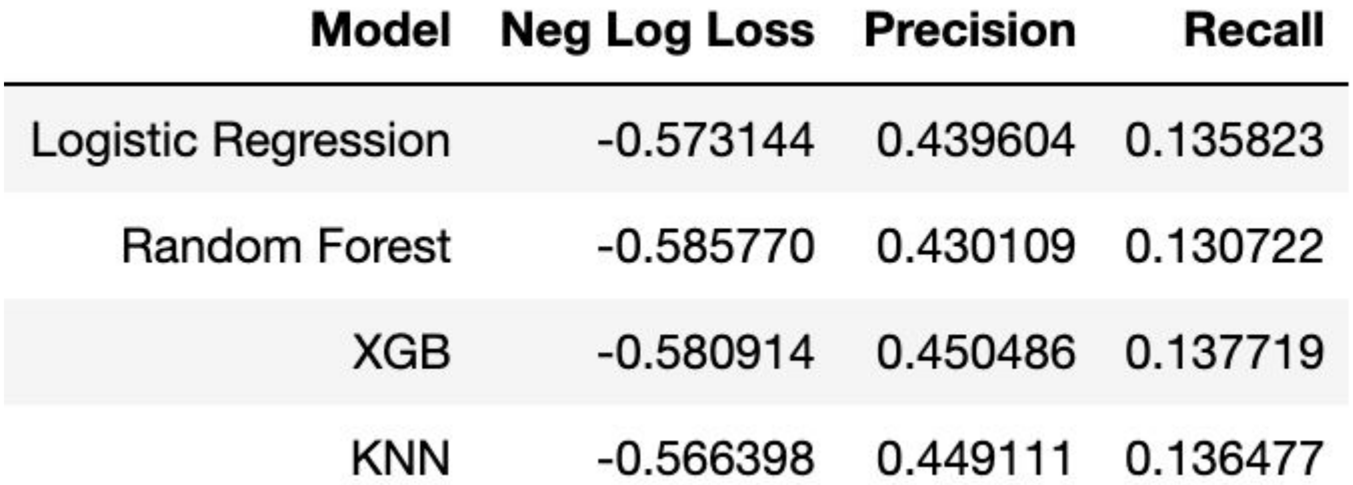
Test Multiple Baseline Models

Select Preferred Model Type

Tune Model

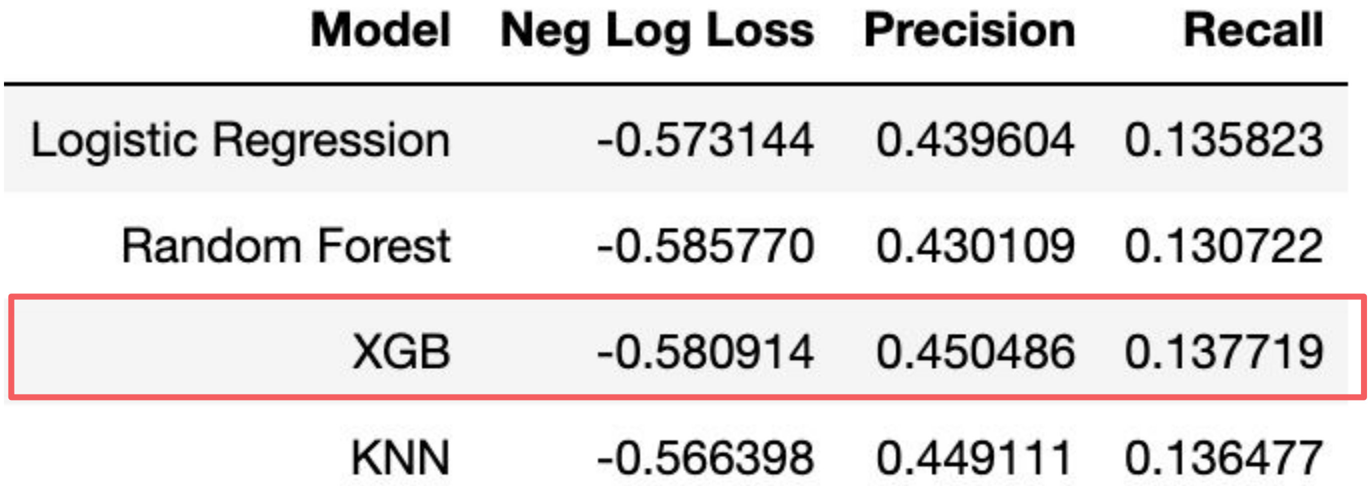
Evaluate Model Performance

Baseline Model Comparison

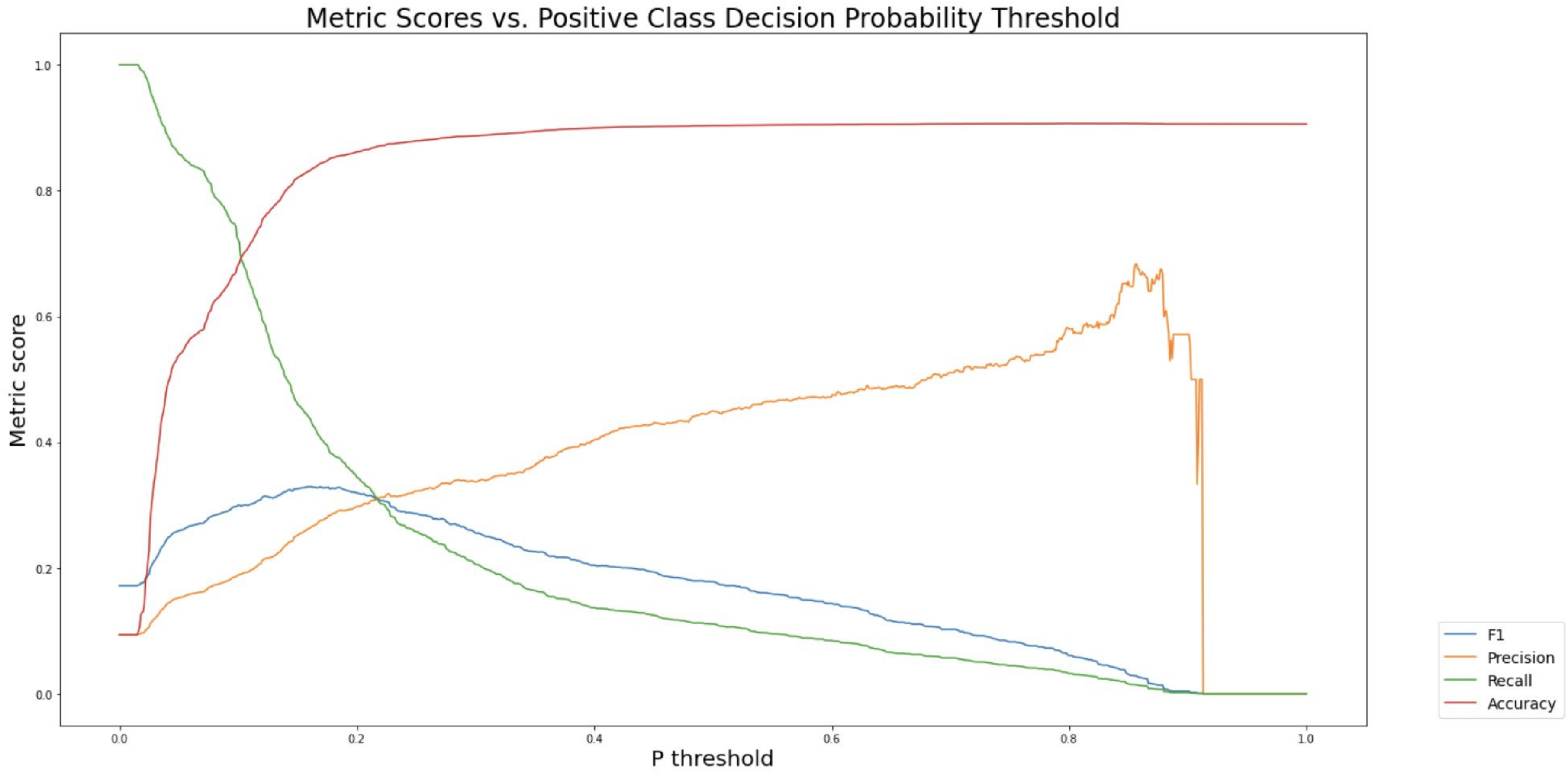


Model	Neg Log Loss	Precision	Recall
Logistic Regression	-0.573144	0.439604	0.135823
Random Forest	-0.585770	0.430109	0.130722
XGB	-0.580914	0.450486	0.137719
KNN	-0.566398	0.449111	0.136477

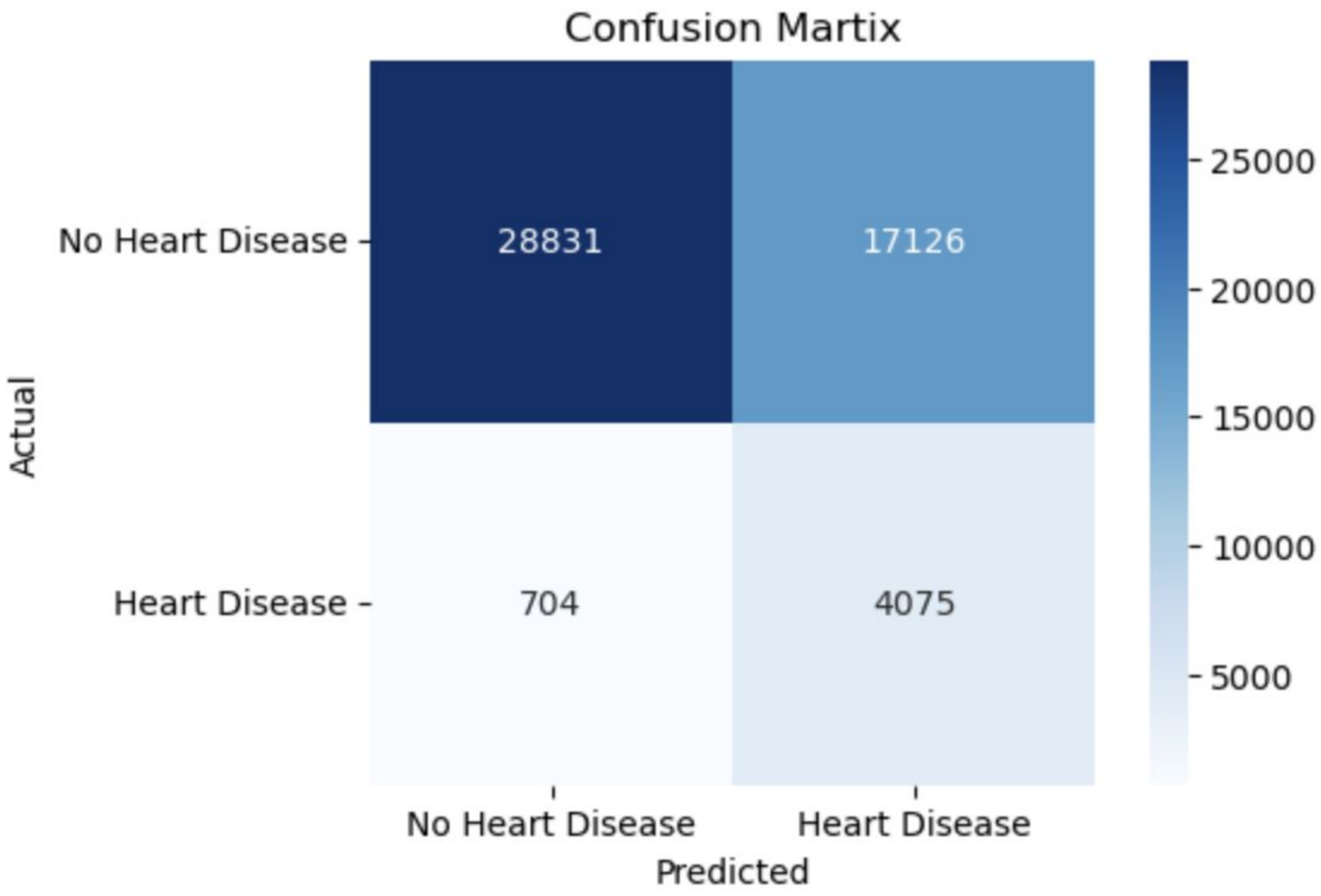
Baseline Model Comparison



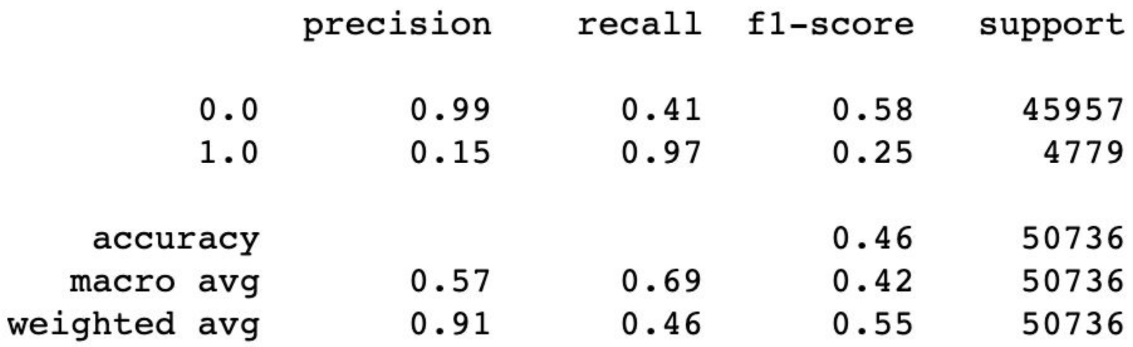
Model	Neg Log Loss	Precision	Recall
Logistic Regression	-0.573144	0.439604	0.135823
Random Forest	-0.585770	0.430109	0.130722
XGB	-0.580914	0.450486	0.137719
KNN	-0.566398	0.449111	0.136477



Best F1 score 0.330 at decision threshold >= 0.161

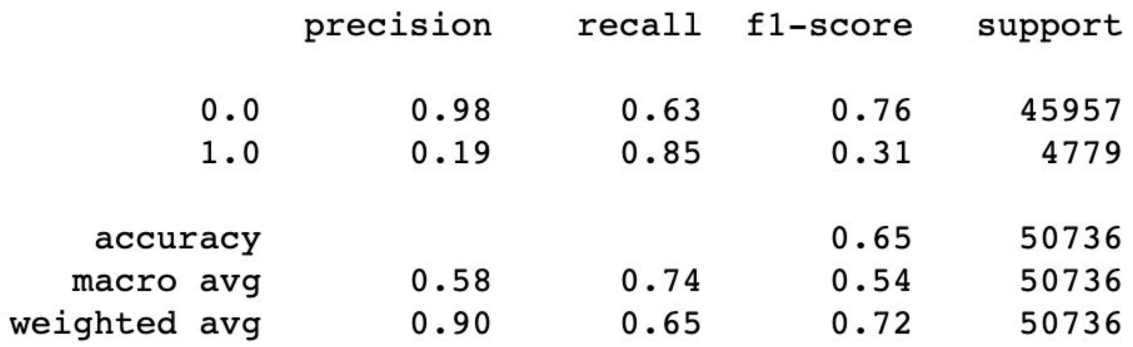

Confusion Martix
Actual
No Heart Disease
Heart Disease
Predicted
No Heart Disease
Heart Disease
28831
17126
704
4075
25000
20000
15000
10000
5000

Baseline Model

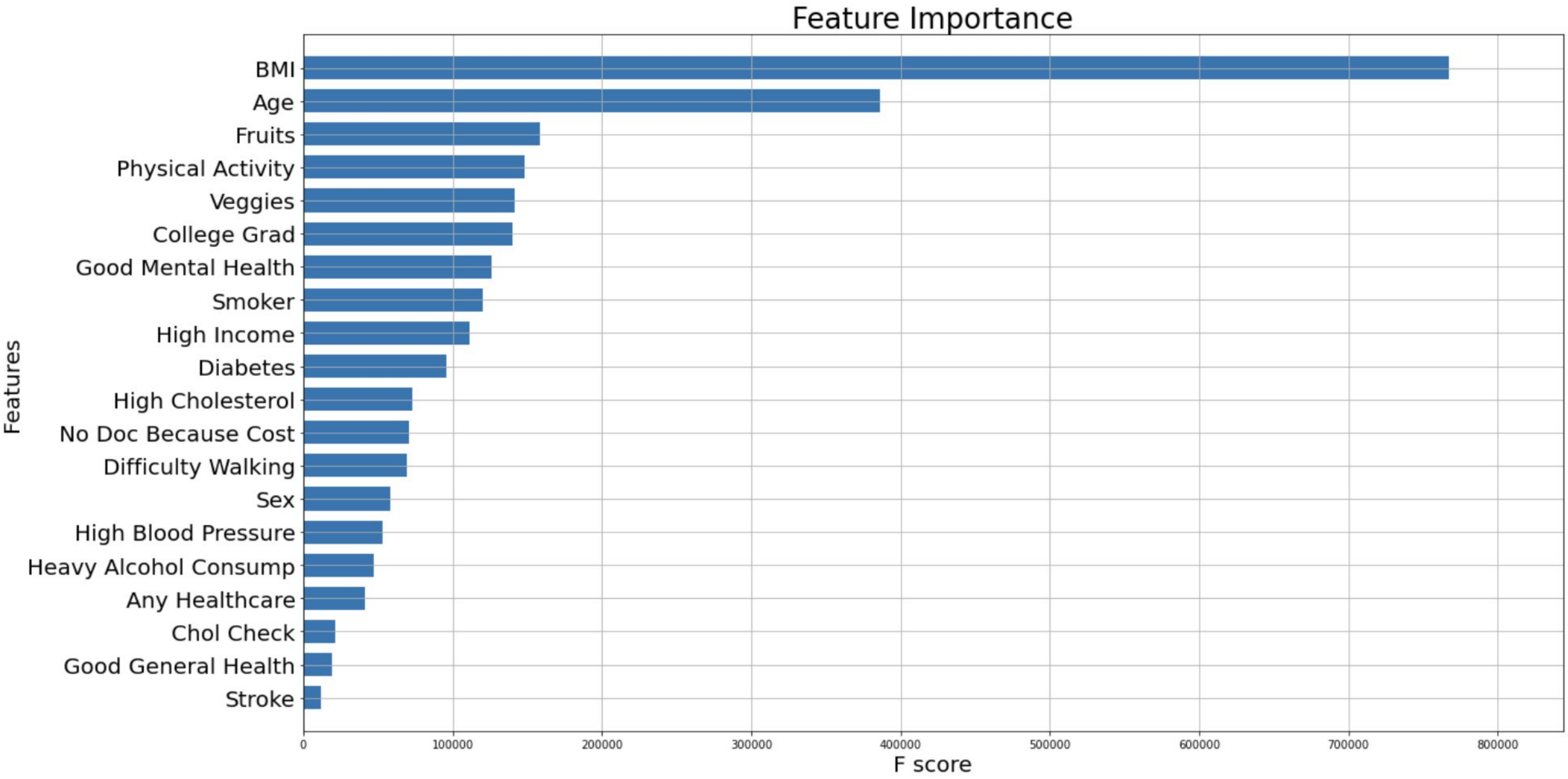


```
              precision    recall  f1-score   support

         0.0       0.99      0.41      0.58     45957
         1.0       0.15      0.97      0.25      4779

    accuracy                           0.46     50736
   macro avg       0.57      0.69      0.42     50736
weighted avg       0.91      0.46      0.55     50736
```

Tuned Model



```
              precision    recall  f1-score   support

         0.0       0.98      0.63      0.76     45957
         1.0       0.19      0.85      0.31      4779

    accuracy                           0.65     50736
   macro avg       0.58      0.74      0.54     50736
weighted avg       0.90      0.65      0.72     50736
```


Feature Importance
BMI
Age
Fruits
Physical Activity
Veggies
College Grad
Good Mental Health
Smoker
High Income
Diabetes
High Cholesterol
No Doc Because Cost
Difficulty Walking
Sex
High Blood Pressure
Heavy Alcohol Consump
Any Healthcare
Chol Check
Good General Health
Stroke
Features
0
100000
200000
300000
400000
500000
600000
700000
800000
F score

Conclusion

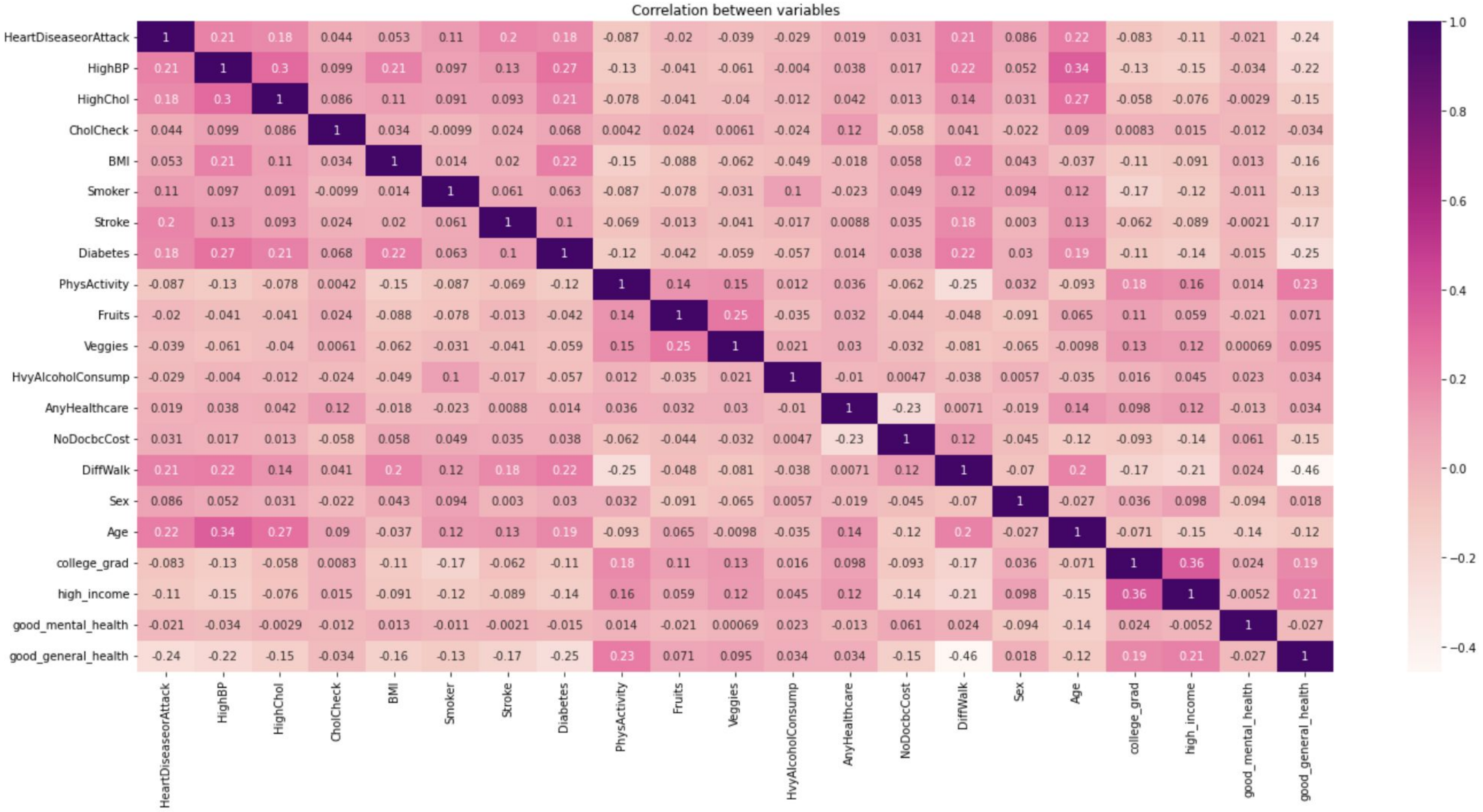
KEY TAKEAWAYS

- BMI and age are key features
- Minimizing false negatives comes with the cost of increasing false positives

NEXT STEPS

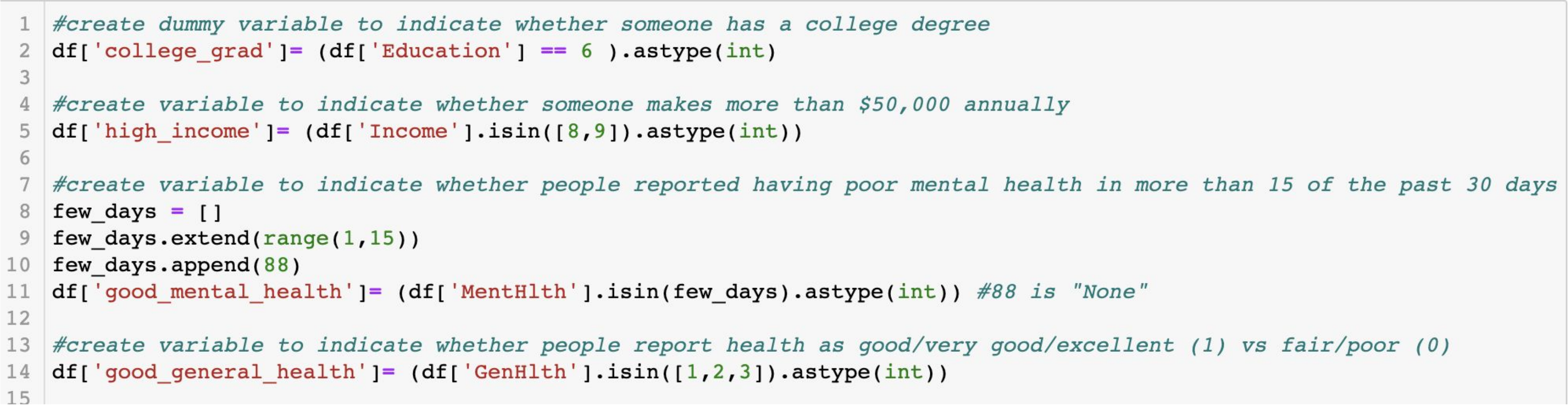
- What age should screening begin?
- What is the cost of a false positive and a false negative?
- Could ensembling improve prediction performance?

Appendix

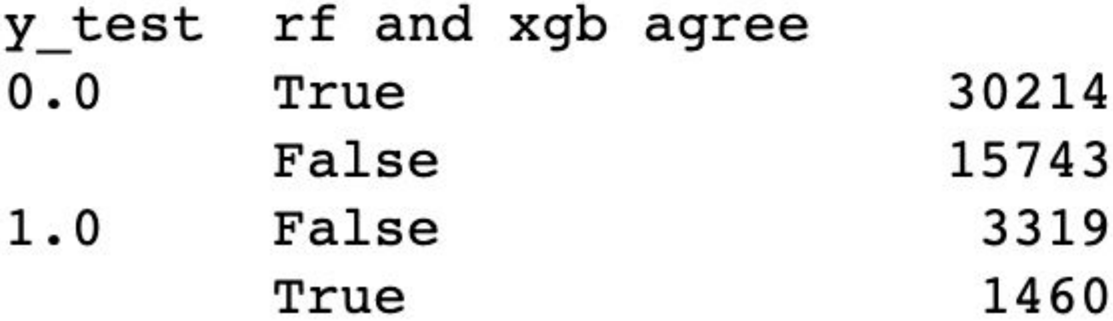


Correlation between variables

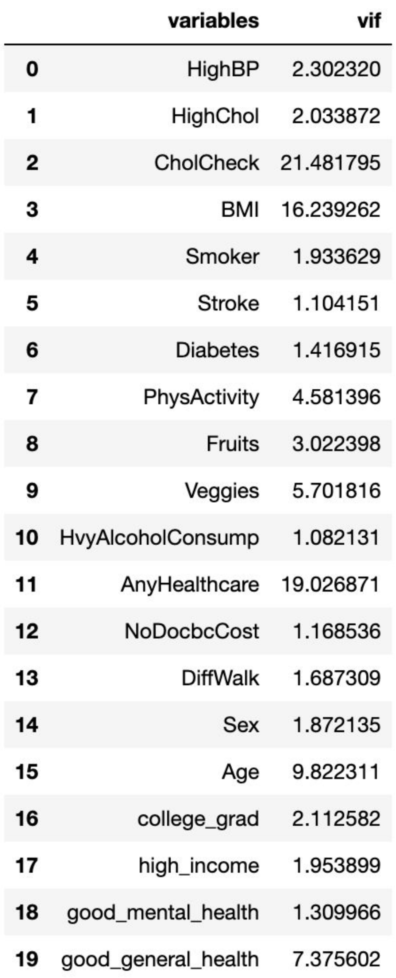
	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	DiffWalk	Sex	Age	college_grad	high_income	good_mental_health	good_general_health
HeartDiseaseorAttack	1	0.21	0.18	0.044	0.053	0.11	0.2	0.18	-0.087	-0.02	-0.039	-0.029	0.019	0.031	0.21	0.086	0.22	-0.083	-0.11	-0.021	-0.24
HighBP	0.21	1	0.3	0.099	0.21	0.097	0.13	0.27	-0.13	-0.041	-0.061	-0.004	0.038	0.017	0.22	0.052	0.34	-0.13	-0.15	-0.034	-0.22
HighChol	0.18	0.3	1	0.086	0.11	0.091	0.093	0.21	-0.078	-0.041	-0.04	-0.012	0.042	0.013	0.14	0.031	0.27	-0.058	-0.076	-0.0029	-0.15
CholCheck	0.044	0.099	0.086	1	0.034	-0.0099	0.024	0.068	0.0042	0.024	0.0061	-0.024	0.12	-0.058	0.041	-0.022	0.09	0.0083	0.015	-0.012	-0.034
BMI	0.053	0.21	0.11	0.034	1	0.014	0.02	0.22	-0.15	-0.088	-0.062	-0.049	-0.018	0.058	0.2	0.043	-0.037	-0.11	-0.091	0.013	-0.16
Smoker	0.11	0.097	0.091	-0.0099	0.014	1	0.061	0.063	-0.087	-0.078	-0.031	0.1	-0.023	0.049	0.12	0.094	0.12	-0.17	-0.12	-0.011	-0.13
Stroke	0.2	0.13	0.093	0.024	0.02	0.061	1	0.1	-0.069	-0.013	-0.041	-0.017	0.0088	0.035	0.18	0.003	0.13	-0.062	-0.089	-0.0021	-0.17
Diabetes	0.18	0.27	0.21	0.068	0.22	0.063	0.1	1	-0.12	-0.042	-0.059	-0.057	0.014	0.038	0.22	0.03	0.19	-0.11	-0.14	-0.015	-0.25
PhysActivity	-0.087	-0.13	-0.078	0.0042	-0.15	-0.087	-0.069	-0.12	1	0.14	0.15	0.012	0.036	-0.062	-0.25	0.032	-0.093	0.18	0.16	0.014	0.23
Fruits	-0.02	-0.041	-0.041	0.024	-0.088	-0.078	-0.013	-0.042	0.14	1	0.25	-0.035	0.032	-0.044	-0.048	-0.091	0.065	0.11	0.059	-0.021	0.071
Veggies	-0.039	-0.061	-0.04	0.0061	-0.062	-0.031	-0.041	-0.059	0.15	0.25	1	0.021	0.03	-0.032	-0.081	-0.065	-0.0098	0.13	0.12	0.00069	0.095
HvyAlcoholConsump	-0.029	-0.004	-0.012	-0.024	-0.049	0.1	-0.017	-0.057	0.012	-0.035	0.021	1	-0.01	0.0047	-0.038	0.0057	-0.035	0.016	0.045	0.023	0.034
AnyHealthcare	0.019	0.038	0.042	0.12	-0.018	-0.023	0.0088	0.014	0.036	0.032	0.03	-0.01	1	-0.23	0.0071	-0.019	0.14	0.098	0.12	-0.013	0.034
NoDocbcCost	0.031	0.017	0.013	-0.058	0.058	0.049	0.035	0.038	-0.062	-0.044	-0.032	0.0047	-0.23	1	0.12	-0.045	-0.12	-0.093	-0.14	0.061	-0.15
DiffWalk	0.21	0.22	0.14	0.041	0.2	0.12	0.18	0.22	-0.25	-0.048	-0.081	-0.038	0.0071	0.12	1	-0.07	0.2	-0.17	-0.21	0.024	-0.46
Sex	0.086	0.052	0.031	-0.022	0.043	0.094	0.003	0.03	0.032	-0.091	-0.065	0.0057	-0.019	-0.045	-0.07	1	-0.027	0.036	0.098	-0.094	0.018
Age	0.22	0.34	0.27	0.09	-0.037	0.12	0.13	0.19	-0.093	0.065	-0.0098	-0.035	0.14	-0.12	0.2	-0.027	1	-0.071	-0.15	-0.14	-0.12
college_grad	-0.083	-0.13	-0.058	0.0083	-0.11	-0.17	-0.062	-0.11	0.18	0.11	0.13	0.016	0.098	-0.093	-0.17	0.036	-0.071	1	0.36	0.024	0.19
high_income	-0.11	-0.15	-0.076	0.015	-0.091	-0.12	-0.089	-0.14	0.16	0.059	0.12	0.045	0.12	-0.14	-0.21	0.098	-0.15	0.36	1	-0.0052	0.21
good_mental_health	-0.021	-0.034	-0.0029	-0.012	0.013	-0.011	-0.0021	-0.015	0.014	-0.021	0.00069	0.023	-0.013	0.061	0.024	-0.094	-0.14	0.024	-0.0052	1	-0.027
good_general_health	-0.24	-0.22	-0.15	-0.034	-0.16	-0.13	-0.17	-0.25	0.23	0.071	0.095	0.034	0.034	-0.15	-0.46	0.018	-0.12	0.19	0.21	-0.027	1



```python
#create dummy variable to indicate whether someone has a college degree
df['college_grad']= (df['Education'] == 6 ).astype(int)

#create variable to indicate whether someone makes more than $50,000 annually
df['high_income']= (df['Income'].isin([8,9]).astype(int))

#create variable to indicate whether people reported having poor mental health in more than 15 of the past 30 days
few_days = []
few_days.extend(range(1,15))
few_days.append(88)
df['good_mental_health']= (df['MentHlth'].isin(few_days).astype(int)) #88 is "None"

#create variable to indicate whether people report health as good/very good/excellent (1) vs fair/poor (0)
df['good_general_health']= (df['GenHlth'].isin([1,2,3]).astype(int))

```



```
y_test  rf and xgb agree
0.0     True                 30214
        False                15743
1.0     False                 3319
        True                  1460
```



	variables	vif
0	HighBP	2.302320
1	HighChol	2.033872
2	CholCheck	21.481795
3	BMI	16.239262
4	Smoker	1.933629
5	Stroke	1.104151
6	Diabetes	1.416915
7	PhysActivity	4.581396
8	Fruits	3.022398
9	Veggies	5.701816
10	HvyAlcoholConsump	1.082131
11	AnyHealthcare	19.026871
12	NoDocbcCost	1.168536
13	DiffWalk	1.687309
14	Sex	1.872135
15	Age	9.822311
16	college_grad	2.112582
17	high_income	1.953899
18	good_mental_health	1.309966
19	good_general_health	7.375602