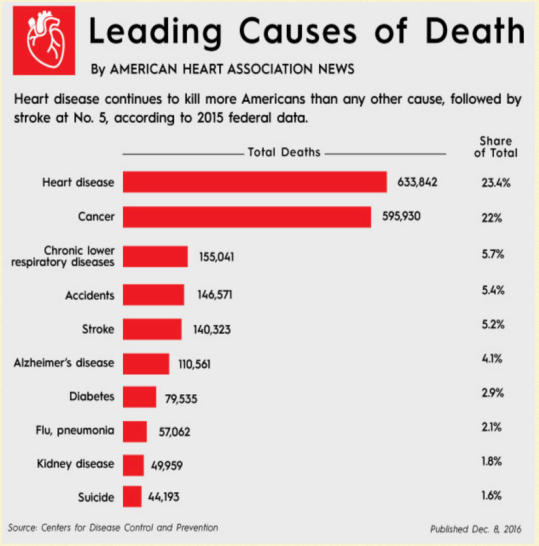
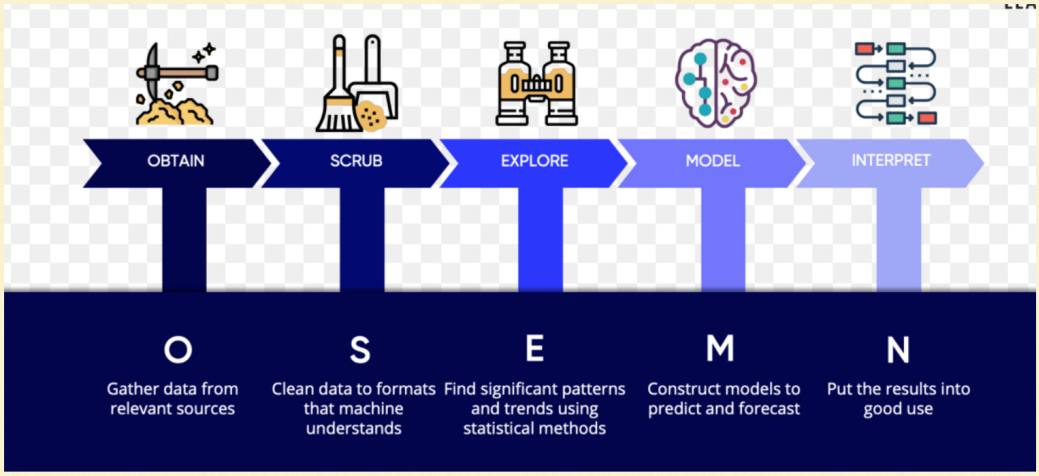
UCI HEART DISEASE DATASET MODELING

GOAL

1. Build a model that will predict if a patient has cardiovascular disease.

2. Make recommendations that will improve the diagnostic precision and improve life expectancy in people with heart disease.





METHODOLOGY USED

MODELS USED

Classifier	Runtime	Train Precision	Test Precision	Train Recall	Test Recall	Train F1	Test F1	Train Accuracy	Test Accuracy	Train ROC AUC	Test ROC AUC
Bagging Tree	0.01	1.00	0.83	0.99	0.83	1.00	0.83	1.00	0.83	1.00	0.83
Random Forest	0.01	1.00	0.81	1.00	0.70	1.00	0.75	1.00	0.77	1.00	0.77
XGBoost	0.04	0.99	0.85	0.96	0.77	0.98	0.81	0.98	0.82	0.98	0.82
Logistic Regression	0.00	0.81	0.85	0.82	0.93	0.82	0.89	0.83	0.88	0.83	0.88
KNeighbors	0.00	0.90	0.83	0.80	0.80	0.85	0.81	0.87	0.82	0.86	0.82

- Logistic Regression Model performed the best with a run time of 0.001seconds , 0.93 – Recall Score and 0.88 – Accuracy Score.

My goal is to have the least false negatives possible, therefore higher recall score. If a test comes back positive but it's actually negative, the doctor will likely order more tests and discover the truth. On the other hand, if a test comes back negative, it is less likely that the doctor will order more tests. It is more expensive to treat an advanced heart disease, because of

the irreversible heart damage. The patient will have a lower life quality and expectancy.

FINAL MODEL

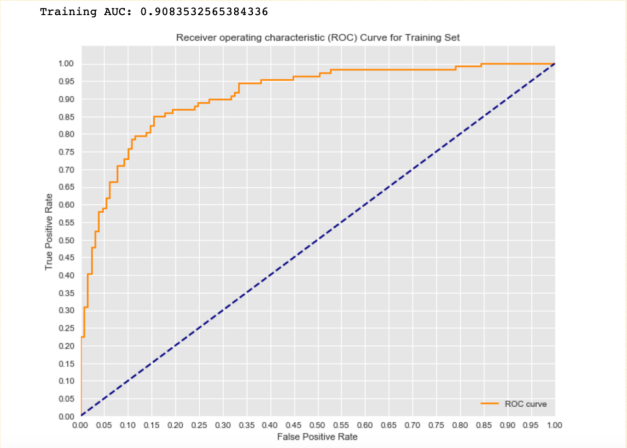
```
Training Precision:  0.8148148148148148
Testing Precision:  0.8484848484848485


Training Recall:  0.822429906542056
Testing Recall:  0.9333333333333333


Training Accuracy:  0.8347457627118644
Testing Accuracy:  0.8833333333333333


Training F1-Score:  0.8186046511627906
Testing F1-Score:  0.888888888888889
```

```
Training AUC: 0.9083532565384336
```



We trained a Logistic Regression Model with the hyper parameters found with search grid and the final Recall score was 0.93.

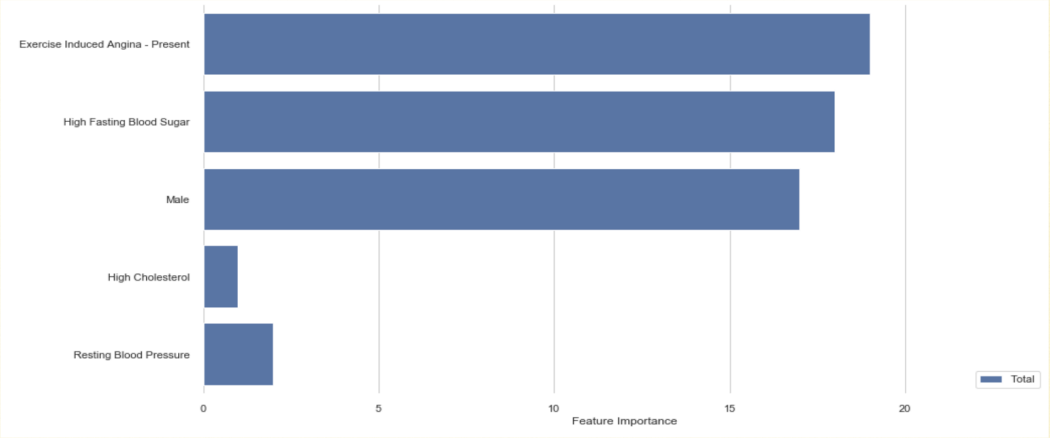
More info on scores:

Precision - Number of true positives divided by number of predicted positives(true positives + false positives). Higher false positives results in lower precision.

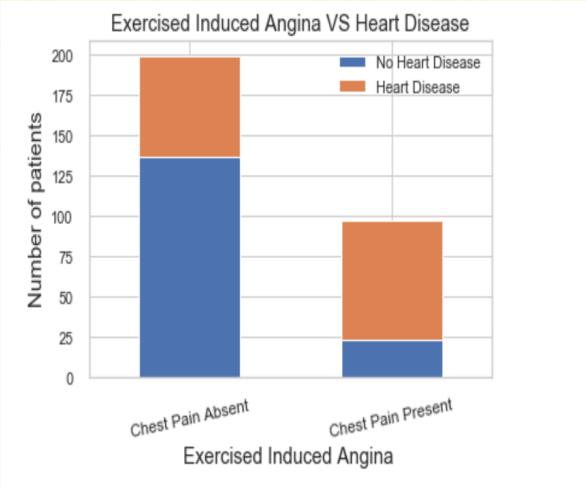
Recall - Number of true positives divided by total of actual positives (true positives + false negatives).
Higher false negatives results in lower recall.
Accuracy - True positives + true negatives divided by total number of observations (how many correct results we have out of all the observations)
F1 - Harmonic mean of precision and recall. A higher F1 will usually indicate a better performing model.(F1 = 2accuracy_score*recall_score/accuracy_score +recall_score)
AUC - stands for "Area under the ROC Curve." AUC measures the entire two-dimensional area underneath the entire ROC curve. The probability that the model ranks a random positive example more

highly than a random negative example.

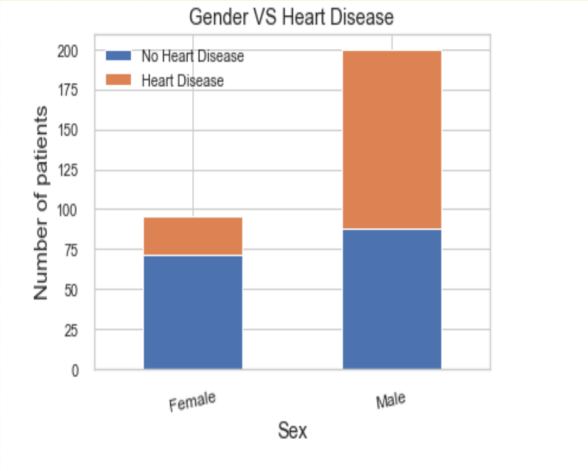
FEATURE IMPORTANCE



EXERCISED INDUCED ANGINA = HIGHER RISK FOR HEART DISEASE

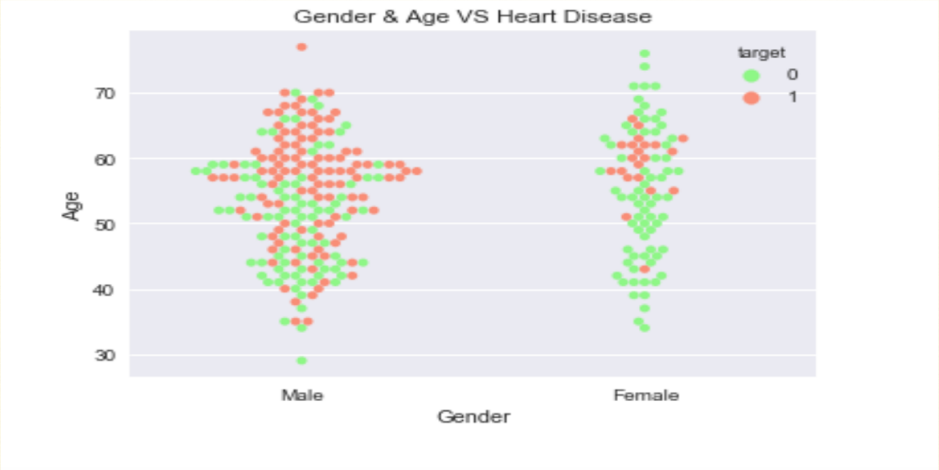


HOW GENDER INFLUENCES HEART DISEASE



Young males are more likely to develop heart disease than young females. Higher levels of estrogen in women provide some sort of protection.

HOW GENDER AND AGE INFLUENCE HEART DISEASE

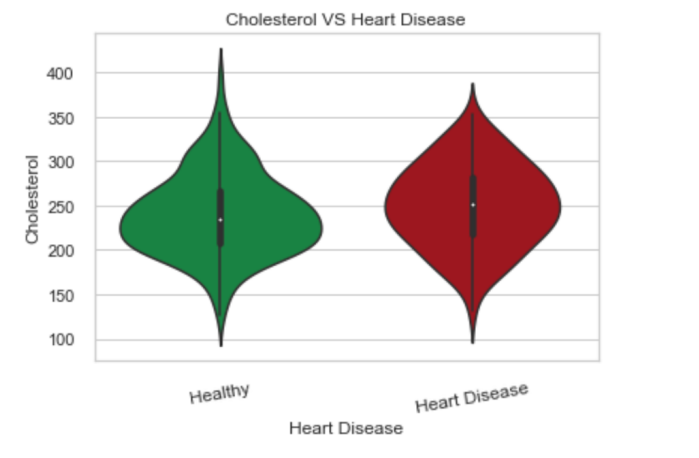


Later in life a woman is as likely to develop heart disease as a man but is less likely to be correctly diagnosed because of the stereotype that women are protected by high estrogen levels. (not necessarily true after menopause) Women are also less likely to be part of heart disease studies, because of the same stereotype. Our dataset is a good example: it consists of roughly

35% women and 65% men.

CHOLESTEROL VS HEART DISEASE

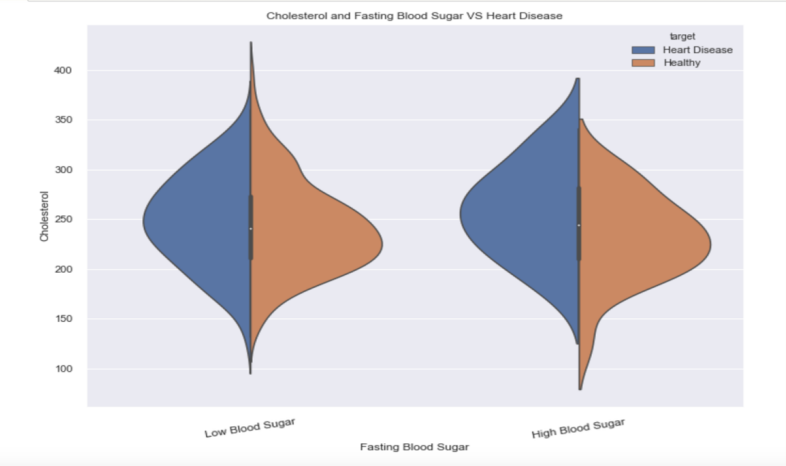
Higher cholesterol seem to increase the probability of having heart disease but not with much.



There is a common knowledge that high cholesterol increases your chance of having heart disease but our dataset doesn't seem to reflect that.

CHOLESTEROL

If you take cholesterol and blood sugar into consideration the difference is even less visible.



FINDINGS

Doctor:

A good model will improve the diagnostic precision and the time a doctor spends on it.

Patient:

An early fast diagnostic will fasten the recovery and save money.

RECOMMENDATIONS

For studies organizers:

Have a better representation for women and minorities.

For doctors:

a. Educate women about their heart disease risks.

b. Order a few cheap blood tests first like blood sugar and cholesterol. If they come back high, recommend stress tests with electrocardiographic results.

For patients:

a. Start protecting your heart health early in life through: diet (low glycemic index foods for low blood sugar) and exercise (for keeping max heart rate high later in life)

FUTURE WORK

-Improve the model:

1. Add a BMI and race columns.
2. Find a better dataset for heart disease. This one is 32 years old; many things have changed since then.

-Improve the XGBoost model with the hyperparameters found with SearchGrid.

THANK YOU