Diagnosing Heart Disease

University Circle, Inc (UCI) Cleveland Hospital

By Isabella Lindgren
Part Time Data Science Student

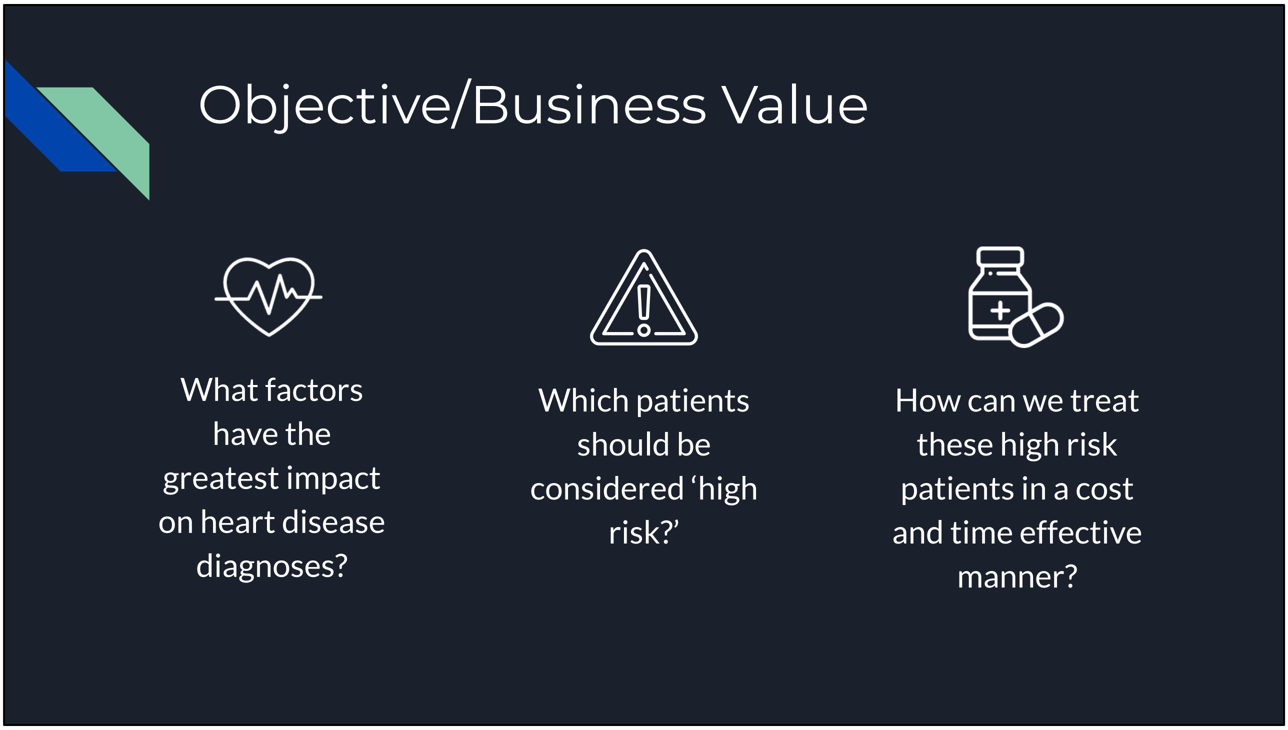
Hello everyone,
Today we are covering a topic I find extremely interesting and I hope you do too. We will be covering the topic of Heart Disease and how we can diagnose patients using Machine Learning.

Background

- 647,000 Americans die from heart disease each year
- The leading cause of death for men, women, and most racial and ethnic groups in the US
- Costs the US ~$219 billion each year

Data from the CDC -

- About 647,000 Americans die from heart disease each year—that’s 1 in every 4 deaths
- Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States.
- Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death.



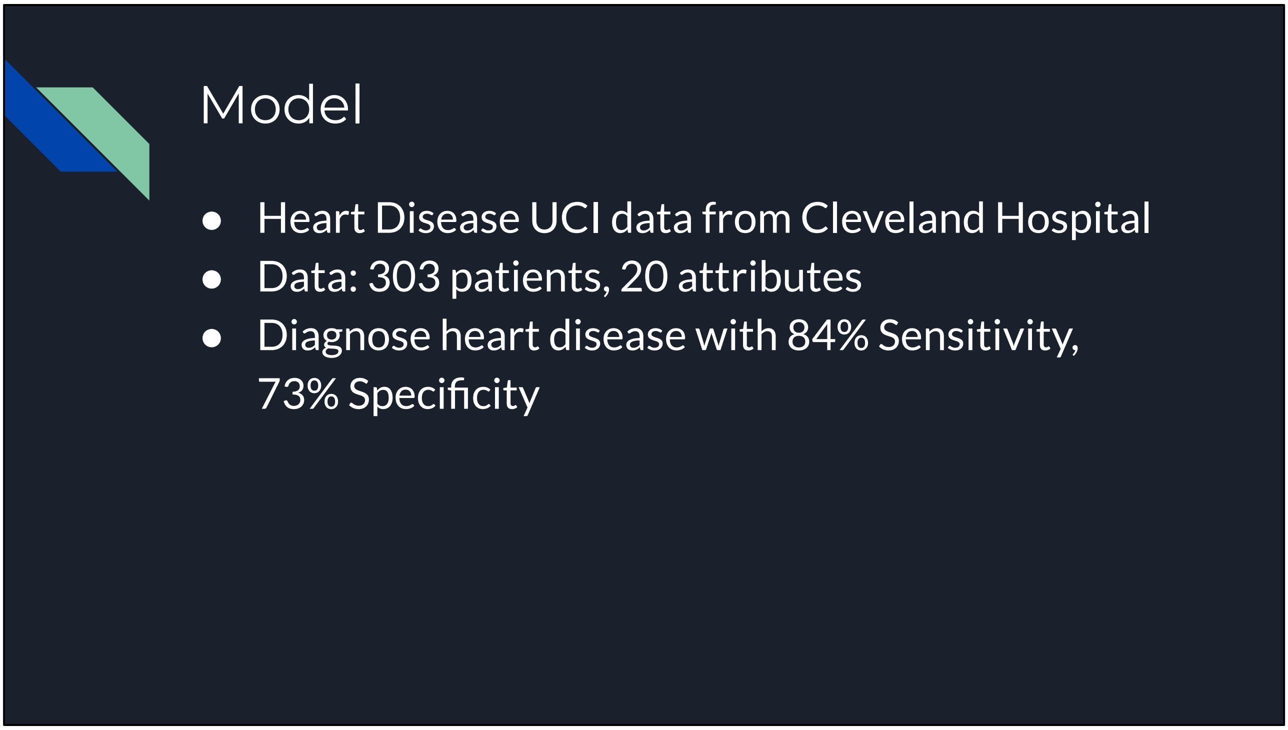
Objective/Business Value

What factors have the greatest impact on heart disease diagnoses?

Which patients should be considered ‘high risk?’

How can we treat these high risk patients in a cost and time effective manner?

- Our objective is to determine which factors have the greatest impact on determining a heart disease diagnoses and use these factors to create a model to classify patients as ‘high risk’ or ‘low risk’
- The findings from this model can help determine a more efficient way to allocate funding and to improve the health care system in regards to heart disease diagnosis and treatment.

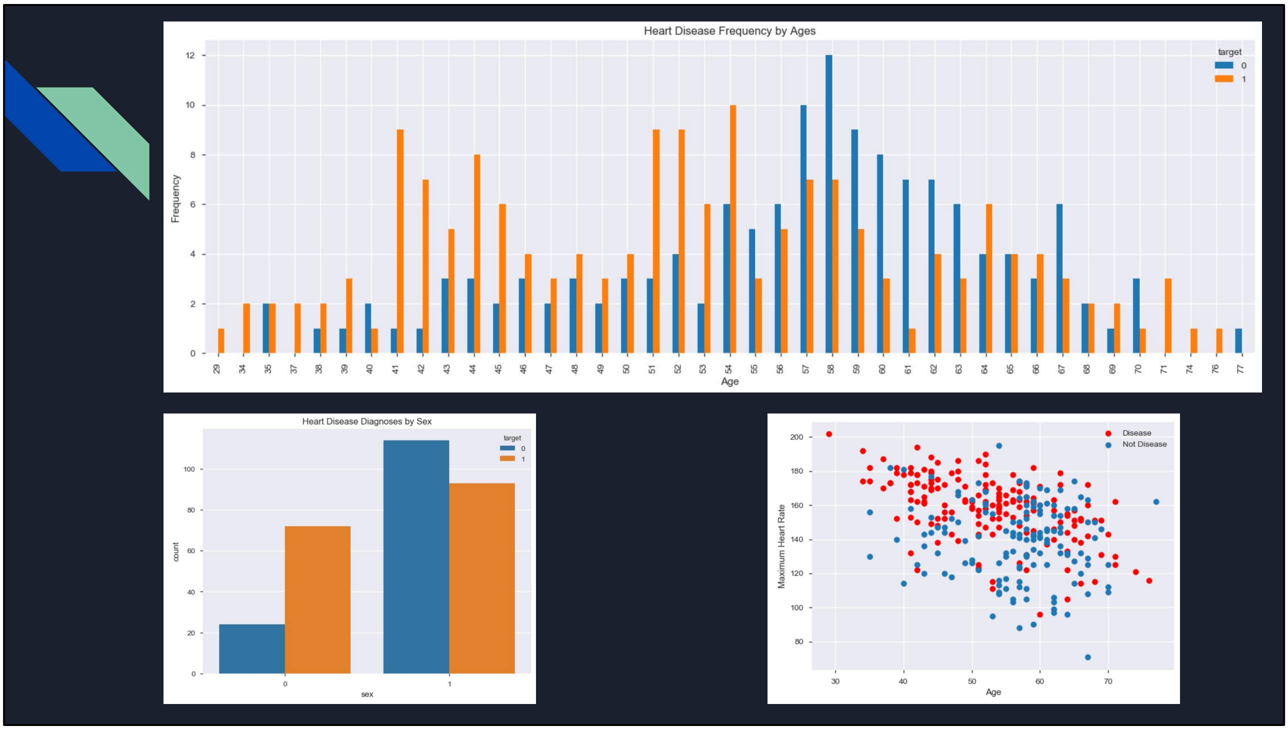


Model

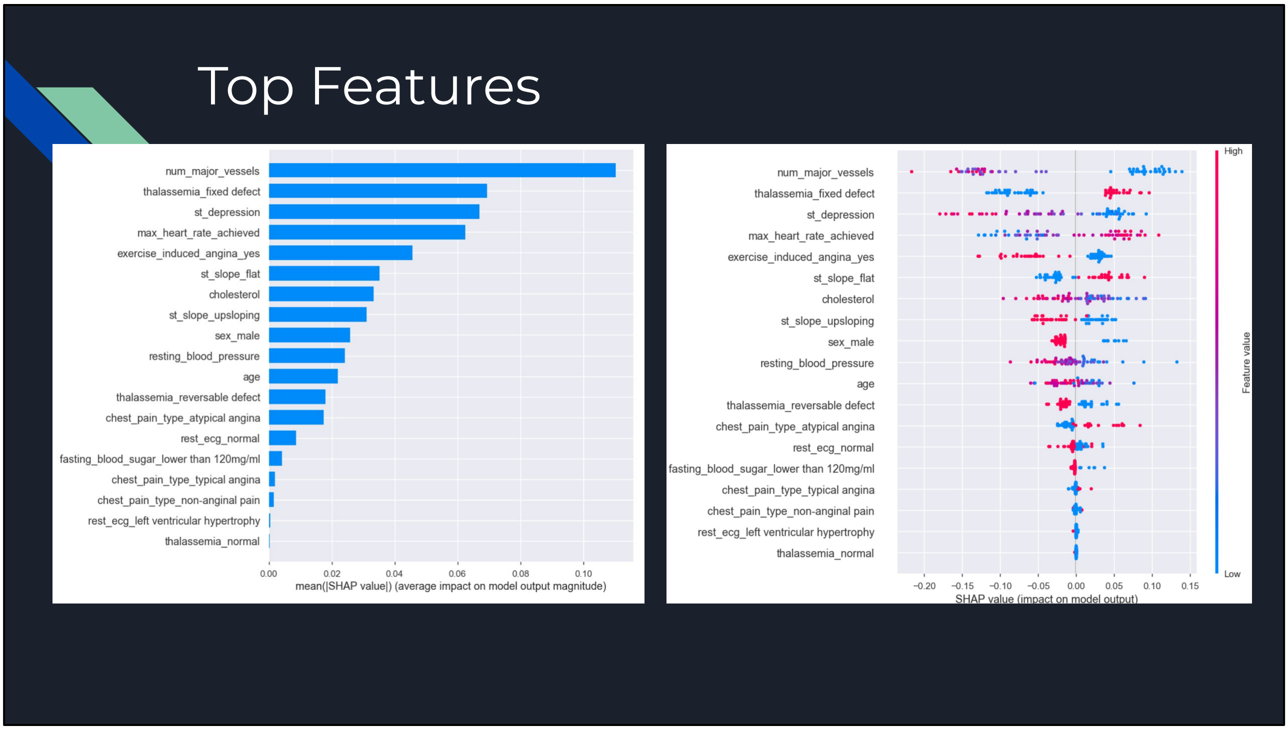
- Heart Disease UCI data from Cleveland Hospital
- Data: 303 patients, 20 attributes
- Diagnose heart disease with 84% Sensitivity, 73% Specificity

- Heart Disease UCI data from Cleveland Hospital
  Included data from 303 patients and 20 attributes including age, sex, chest pain type, resting blood pressure, cholesterol, blood sugar, testing results, etc.

- Our final model could determine with 84% Sensitivity and 73% Specificity whether a patient should be diagnosed with heart disease.



- In this dataset, we can see a trend of increasing frequency of heart disease with age
- Women were more likely to be diagnosed with heart disease compared to men in this dataset which is contradictory to the overall data of the united states. The CDC states that men are more likely to develop heart disease and at earlier ages than women.
- We can also see that the maximum resting heart rate is generally higher in patients who have heart disease, which makes sense since many heart conditions are caused by decreased blood flow, thus the heart works harder to distribute enough oxygen throughout the body.

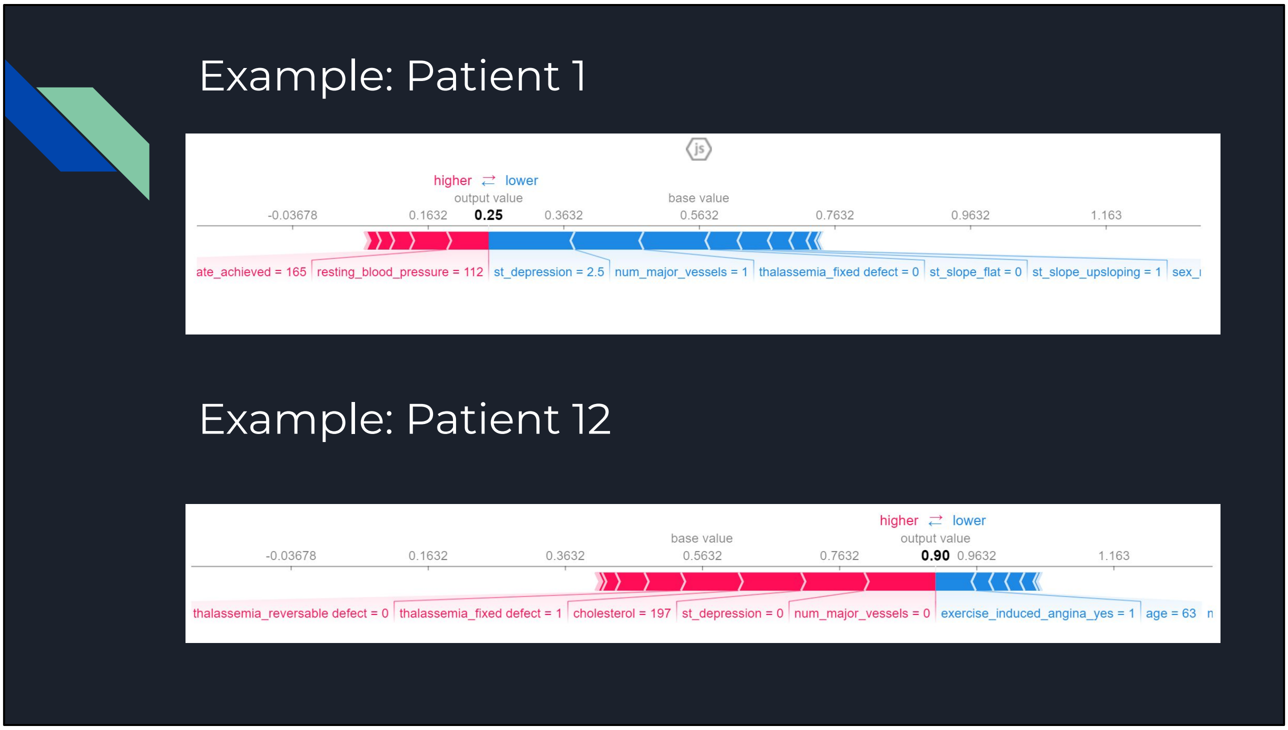


Top Features

- The features with the highest impact in our model were: number of major blood vessels, fixed Thalassemia defect, a depression in an ECG test, maximum heart rate, excercise induced angina,
- Number of major vessels had the most significant impact on our model

On the right

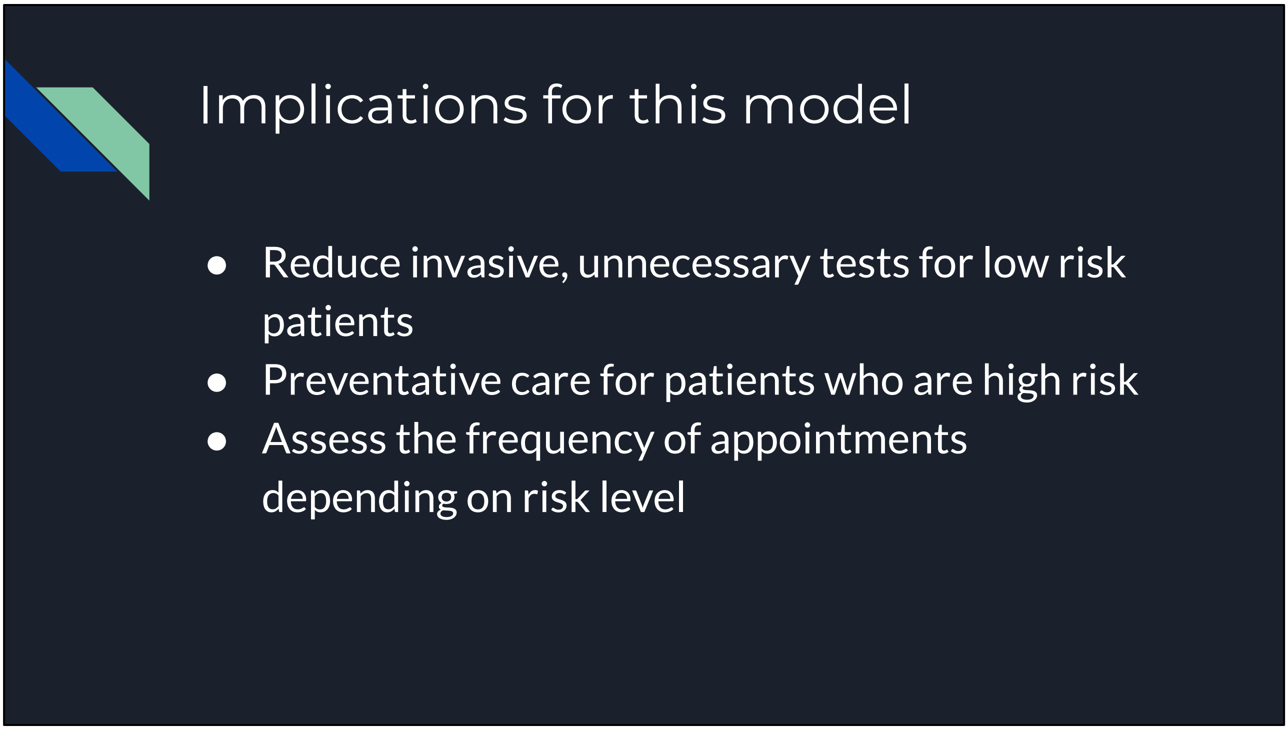
- The features with a clear divide were strong classifying features



Example: Patient 1

Example: Patient 12

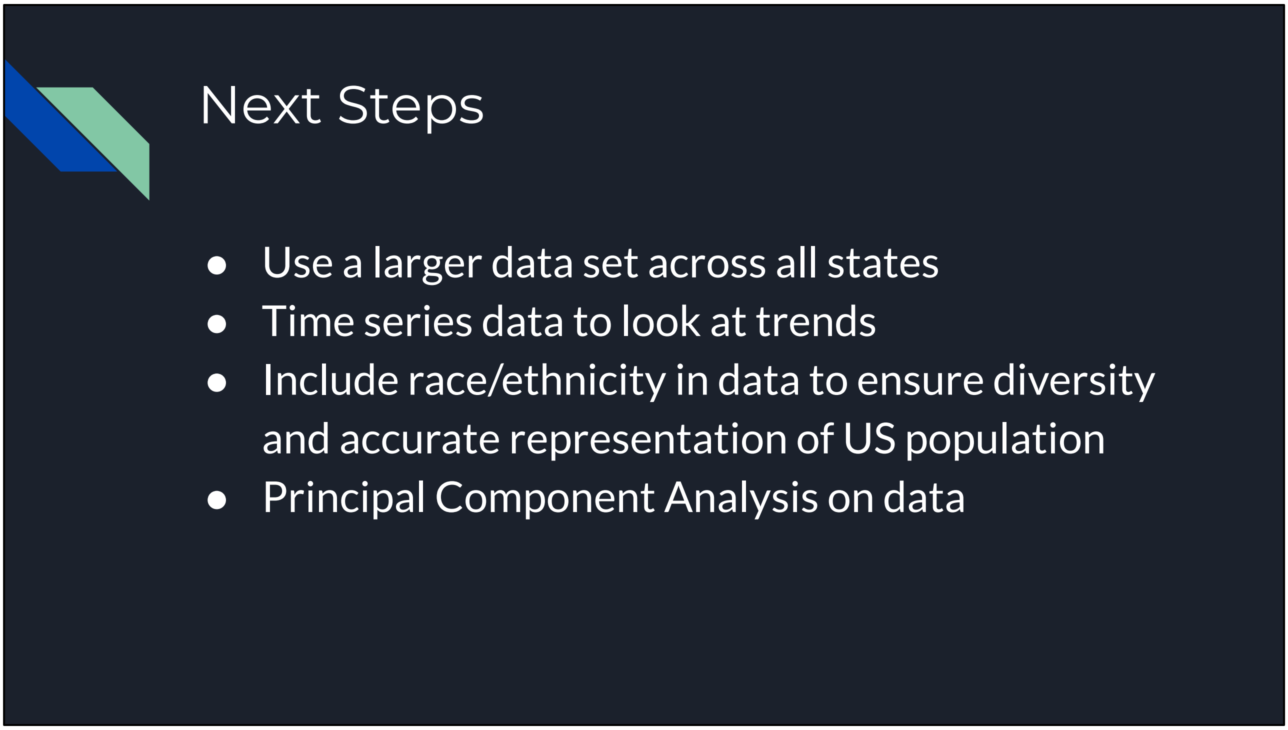
- Ex1: Patient 1: ST- depression on ECG test, higher number of vessels, No Thalassemia worked in their favor. Their output value was below baseline, classifying them as 'low risk'
- **Thalassemia** is a blood disorder passed down through families (inherited) in which the body makes an abnormal form or inadequate amount of hemoglobin
- Ex2: Patient 12: Low # major blood vessels, no ST-depression on ECG, high cholesterol and a Thalassemia fixed defect classified them as 'high risk'



Implications for this model

- Reduce invasive, unnecessary tests for low risk patients
- Preventative care for patients who are high risk
- Assess the frequency of appointments depending on risk level

-

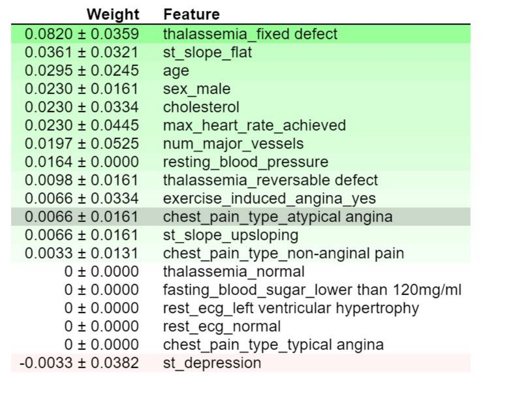


Next Steps

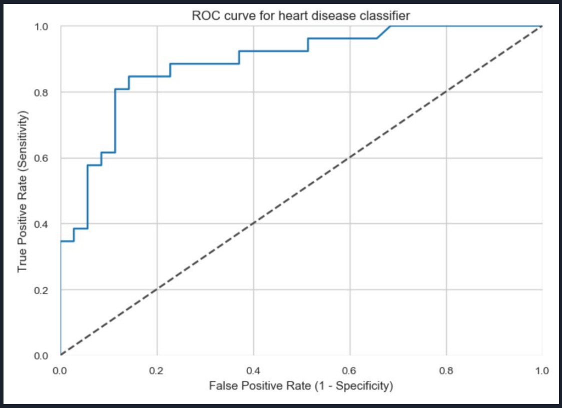
- Use a larger data set across all states
- Time series data to look at trends
- Include race/ethnicity in data to ensure diversity and accurate representation of US population
- Principal Component Analysis on data

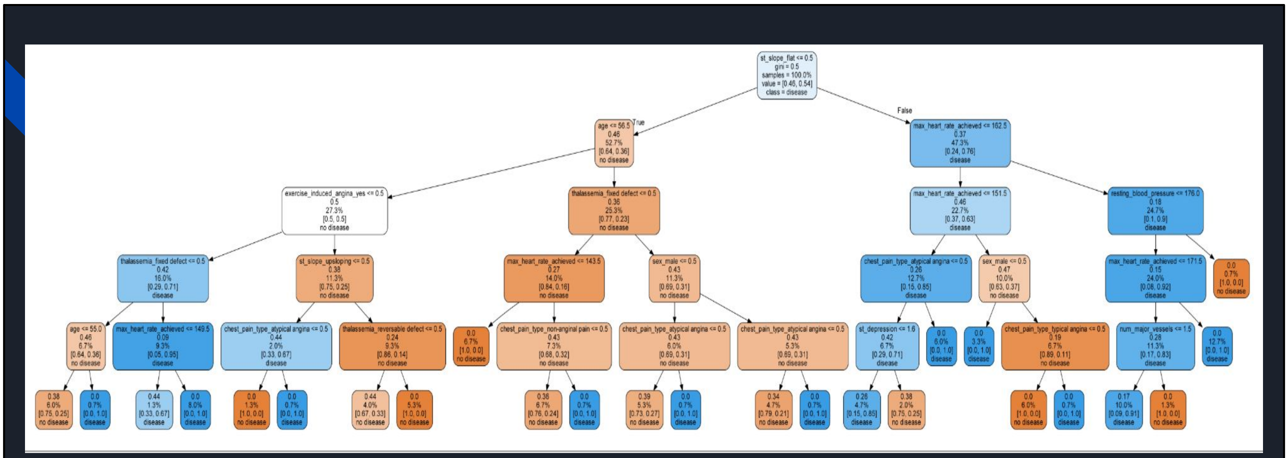
Thank you

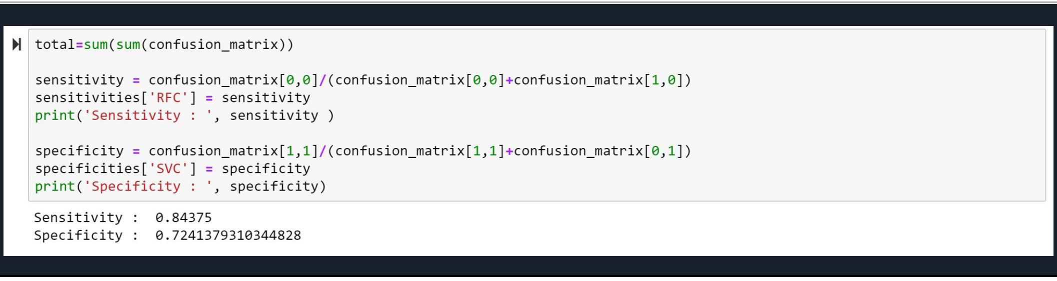
Appendix



Weight	Feature
0.0820 ± 0.0359	thalassemia_fixed defect
0.0361 ± 0.0321	st_slope_flat
0.0295 ± 0.0245	age
0.0230 ± 0.0161	sex_male
0.0230 ± 0.0334	cholesterol
0.0230 ± 0.0445	max_heart_rate_achieved
0.0197 ± 0.0525	num_major_vessels
0.0164 ± 0.0000	resting_blood_pressure
0.0098 ± 0.0161	thalassemia_reversable defect
0.0066 ± 0.0334	exercise_induced_angina_yes
0.0066 ± 0.0161	chest_pain_type_atypical angina
0.0066 ± 0.0161	st_slope_upsloping
0.0033 ± 0.0131	chest_pain_type_non-anginal pain
0 ± 0.0000	thalassemia_normal
0 ± 0.0000	fasting_blood_sugar_lower than 120mg/ml
0 ± 0.0000	rest_ecg_left ventricular hypertrophy
0 ± 0.0000	rest_ecg_normal
0 ± 0.0000	chest_pain_type_typical angina
-0.0033 ± 0.0382	st_depression

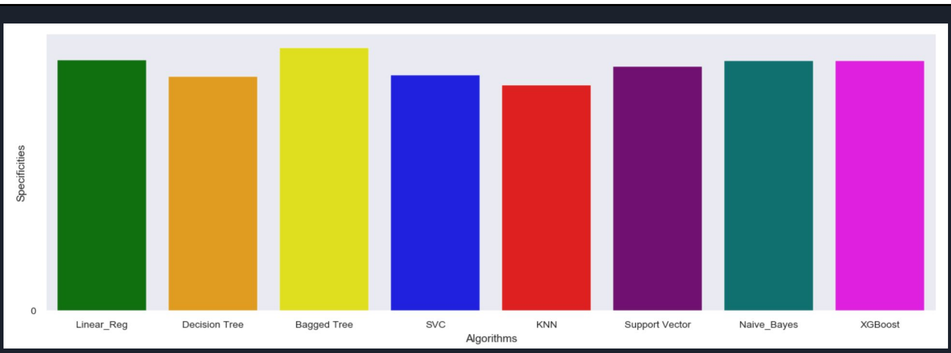



st_slope_flat <= 0.5 gini = 0.5 samples = 100.0% value = [0.46, 0.54] class = disease
True
False
age <= 56.5 0.46 52.7% [0.64, 0.36] no disease
max_heart_rate_achieved <= 162.5 0.37 47.3% [0.24, 0.76] disease
exercise_induced_angina_yes <= 0.5 0.5 27.3% [0.5, 0.5] no disease
thalassemia_fixed defect <= 0.5 0.36 25.3% [0.77, 0.23] no disease
max_heart_rate_achieved <= 151.5 0.46 22.7% [0.37, 0.63] disease
resting_blood_pressure <= 176.0 0.18 24.7% [0.1, 0.9] disease
thalassemia_fixed defect <= 0.5 0.42 16.0% [0.29, 0.71] disease
st_slope_upsloping <= 0.5 0.38 11.3% [0.75, 0.25] no disease
max_heart_rate_achieved <= 143.5 0.27 14.0% [0.84, 0.16] no disease
sex_male <= 0.5 0.43 11.3% [0.69, 0.31] no disease
chest_pain_type_atypical angina <= 0.5 0.26 12.7% [0.15, 0.85] disease
sex_male <= 0.5 0.47 10.0% [0.63, 0.37] no disease
max_heart_rate_achieved <= 171.5 0.15 24.0% [0.08, 0.92] disease
0.0 0.7% [1.0, 0.0] no disease
age <= 55.0 0.46 6.7% [0.64, 0.36] no disease
max_heart_rate_achieved <= 149.5 0.09 9.3% [0.05, 0.95] disease
chest_pain_type_atypical angina <= 0.5 0.44 2.0% [0.33, 0.67] disease
thalassemia_reversable defect <= 0.5 0.24 9.3% [0.86, 0.14] no disease
0.0 6.7% [1.0, 0.0] no disease
chest_pain_type_non-anginal pain <= 0.5 0.43 7.3% [0.68, 0.32] no disease
chest_pain_type_atypical angina <= 0.5 0.43 6.0% [0.69, 0.31] no disease
chest_pain_type_atypical angina <= 0.5 0.43 5.3% [0.69, 0.31] no disease
st_depression <= 1.6 0.42 6.7% [0.29, 0.71] disease
0.0 6.0% [0.0, 1.0] disease
0.0 3.3% [0.0, 1.0] disease
chest_pain_type_typical angina <= 0.5 0.19 6.7% [0.89, 0.11] no disease
num_major_vessels <= 1.5 0.28 11.3% [0.17, 0.83] disease
0.0 12.7% [0.0, 1.0] disease
0.38 6.0% [0.75, 0.25] no disease
0.0 0.7% [0.0, 1.0] disease
0.44 1.3% [0.33, 0.67] disease
0.0 8.0% [0.0, 1.0] disease
0.0 1.3% [1.0, 0.0] no disease
0.0 0.7% [0.0, 1.0] disease
0.44 4.0% [0.67, 0.33] no disease
0.0 5.3% [1.0, 0.0] no disease
0.36 6.7% [0.76, 0.24] no disease
0.0 0.7% [0.0, 1.0] disease
0.39 5.3% [0.73, 0.27] no disease
0.0 0.7% [0.0, 1.0] disease
0.34 4.7% [0.79, 0.21] no disease
0.0 0.7% [0.0, 1.0] disease
0.26 4.7% [0.15, 0.85] disease
0.38 2.0% [0.75, 0.25] no disease
0.0 6.0% [1.0, 0.0] no disease
0.0 0.7% [0.0, 1.0] disease
0.17 10.0% [0.09, 0.91] disease
0.0 1.3% [1.0, 0.0] no disease

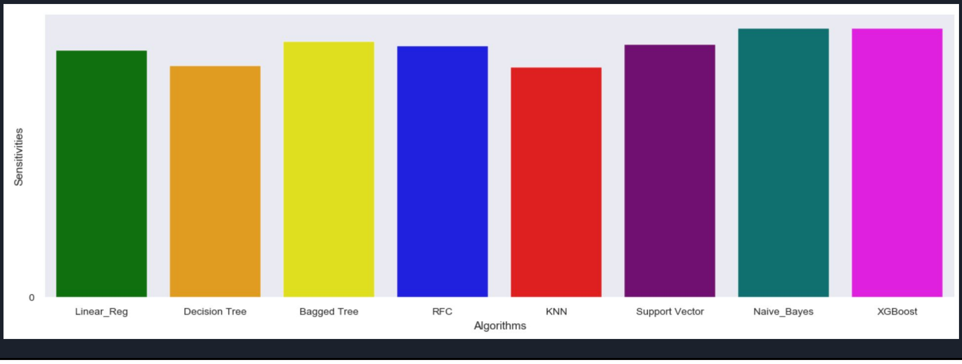


```
total=sum(sum(confusion_matrix))

sensitivity = confusion_matrix[0,0]/(confusion_matrix[0,0]+confusion_matrix[1,0])
sensitivities['RFC'] = sensitivity
print('Sensitivity : ', sensitivity )

specificity = confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[0,1])
specificities['SVC'] = specificity
print('Specificity : ', specificity)
```

```
Sensitivity :  0.84375
Specificity :  0.7241379310344828
```



Specificities
0
Linear_Reg
Decision Tree
Bagged Tree
SVC
KNN
Support Vector
Naive_Bayes
XGBoost
Algorithms



Sensitivities
0
Linear_Reg
Decision Tree
Bagged Tree
RFC
KNN
Support Vector
Naive_Bayes
XGBoost
Algorithms