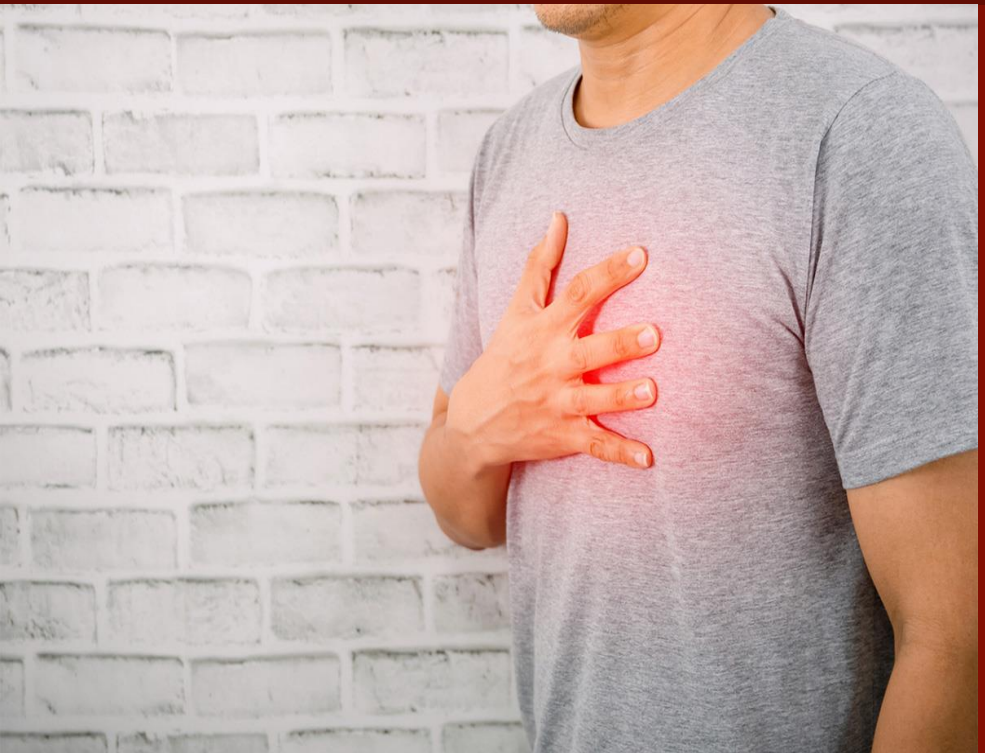
Heart Disease Prediction



Heart disease is a silent killer, striking without warning and leaving a lasting impact

Presented by James Warsing

Business Understanding

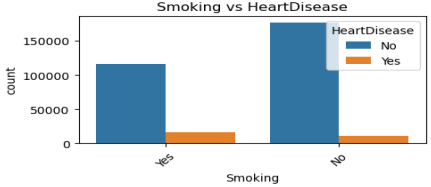
- Heart disease is one of the leading causes of death globally, posing a significant challenge for healthcare providers.
- The goal is to develop a model for early prediction and and prevention, enabling healthcare providers to implement implement targeted prevention strategies and optimize resource allocation.

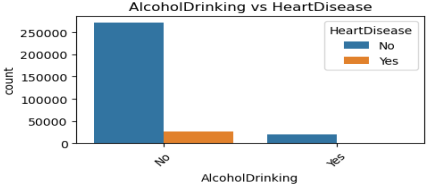
Key Questions:

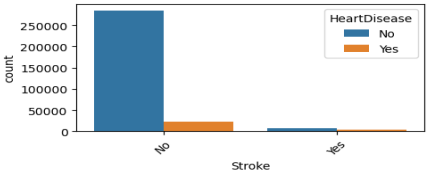
- How can healthcare providers use predictive analytics to prioritize patients for early intervention and tailored treatment plans?
- How can existing health data be leveraged to accurately predict the risk of heart disease in individuals?

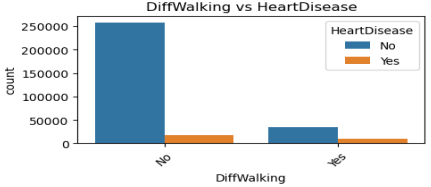
Data Overview

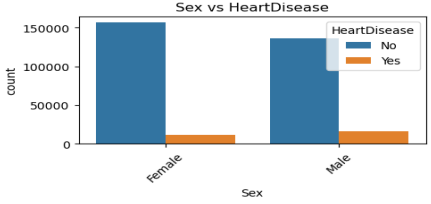
- Source: Kaggle heart disease dataset with 319,795 individuals.
- 18 health-related attributes. Some with a focus on demographics, while others were highlighted physical health and past health history.

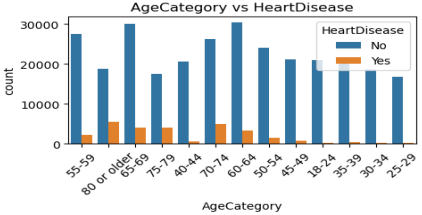

Smoking vs HeartDisease
HeartDisease
No
Yes
count
Yes
No
Smoking

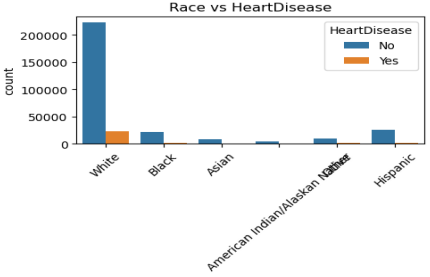

AlcoholDrinking vs HeartDisease
HeartDisease
No
Yes
count
No
Yes
AlcoholDrinking

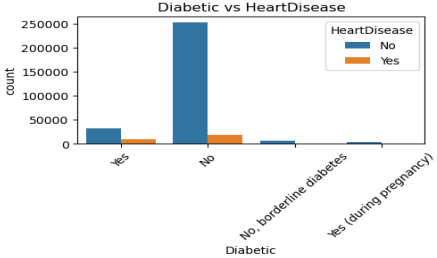

Stroke vs HeartDisease
HeartDisease
No
Yes
count
No
Yes
Stroke

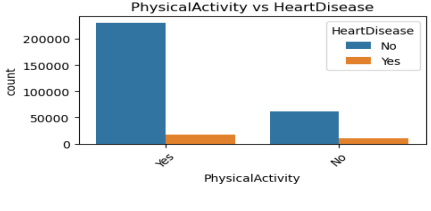

DiffWalking vs HeartDisease
HeartDisease
No
Yes
count
No
Yes
DiffWalking

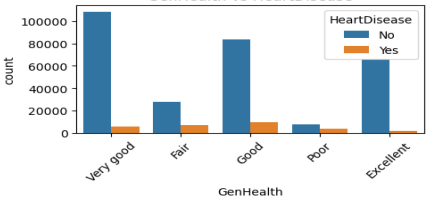

Sex vs HeartDisease
HeartDisease
No
Yes
count
Female
Male
Sex

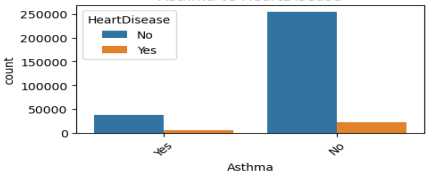

AgeCategory vs HeartDisease
HeartDisease
No
Yes
count
55-59
80 or older
65-69
75-79
40-44
70-74
60-64
50-54
45-49
18-24
35-39
30-34
25-29
AgeCategory

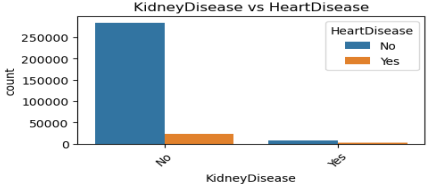

Race vs HeartDisease
HeartDisease
No
Yes
count
White
Black
Asian
Hispanic
Race

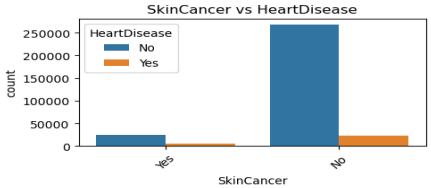

Diabetic vs HeartDisease
HeartDisease
No
Yes
count
Yes
No
No, borderline diabetes
Yes (during pregnancy)
Diabetic


PhysicalActivity vs HeartDisease
HeartDisease
No
Yes
count
Yes
No
PhysicalActivity


GenHealth vs HeartDisease
HeartDisease
No
Yes
count
Very good
Fair
Good
Poor
Excellent
GenHealth


Asthma vs HeartDisease
HeartDisease
No
Yes
count
Yes
No
Asthma


KidneyDisease vs HeartDisease
HeartDisease
No
Yes
count
No
Yes
KidneyDisease

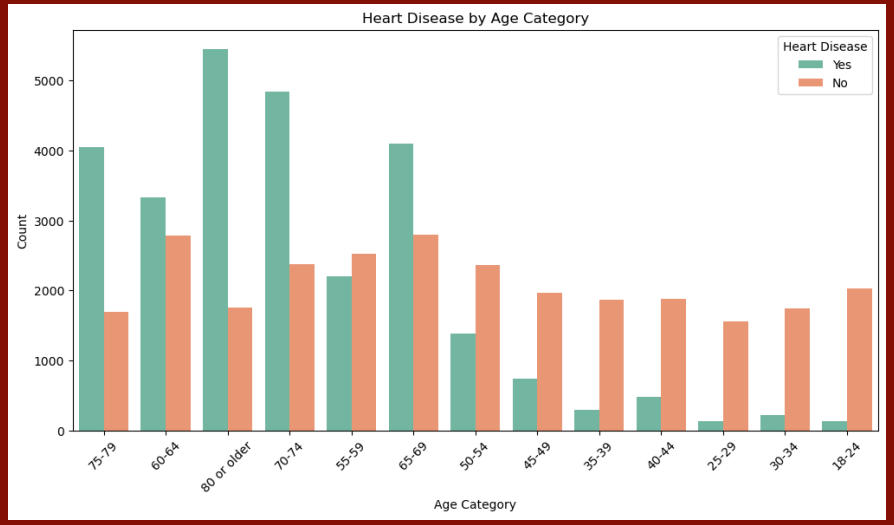

SkinCancer vs HeartDisease
HeartDisease
No
Yes
count
Yes
No
SkinCancer

Highly imbalanced dataset

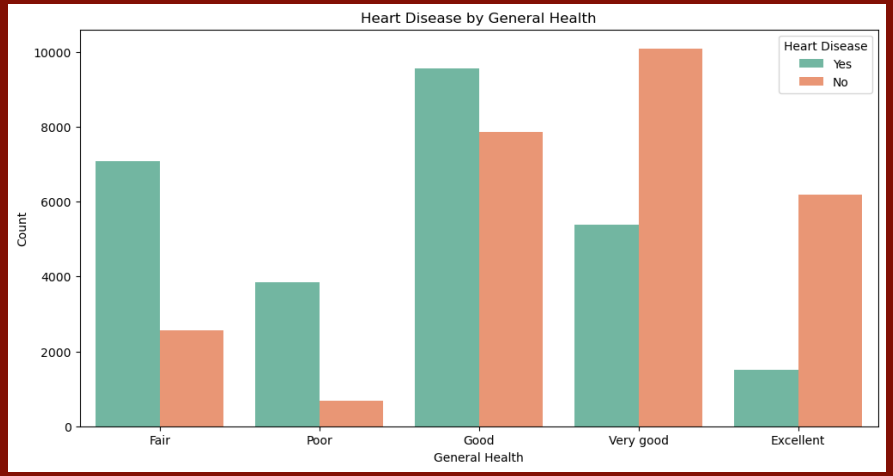
Prior health conditions can heavily impact the risk for heart disease

Higher risk for individuals over 50.

General Health is also a good indicator



A few of the key features from the data were age and and general health



Data Processing Steps

A mixture of ordinal and nominal variables required appropriate encoding

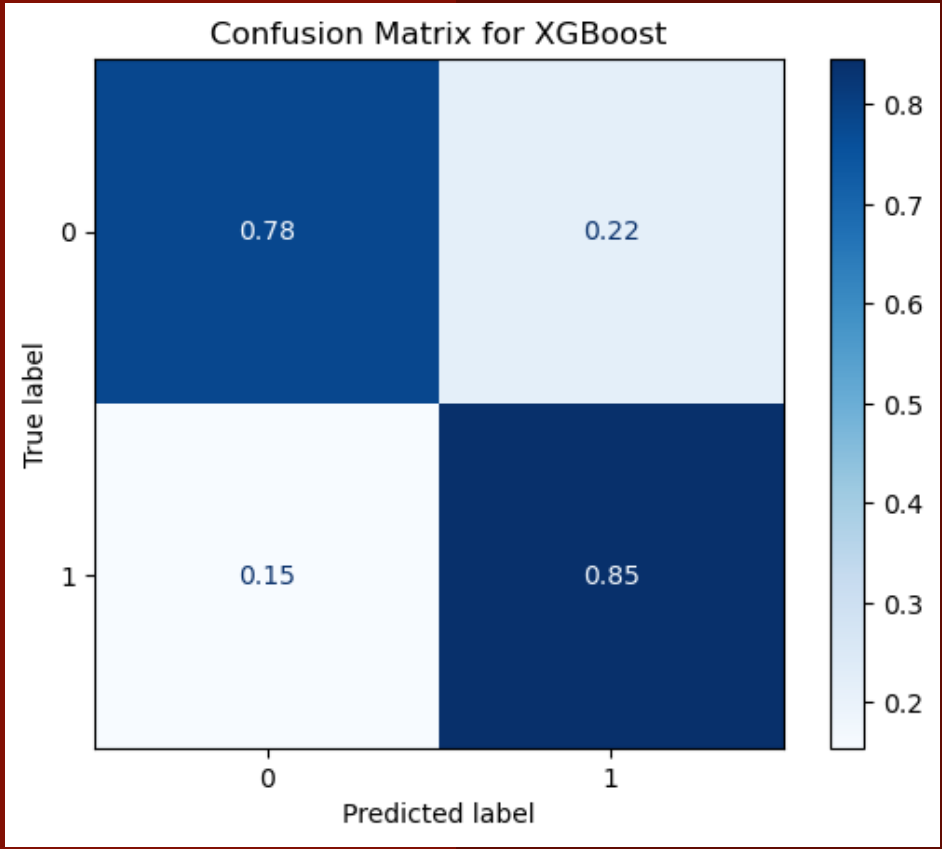
The use of balancing techniques was necessary for the models: Downsampling/SMOTE

Models

- **Logistic Regression:** Baseline model for comparison.
- **Random Forest:** Captures complex interactions between features.
- **XGBoost:** Handles imbalanced datasets effectively.
- **SVM:** Robust classifier for non-linearly separable data.

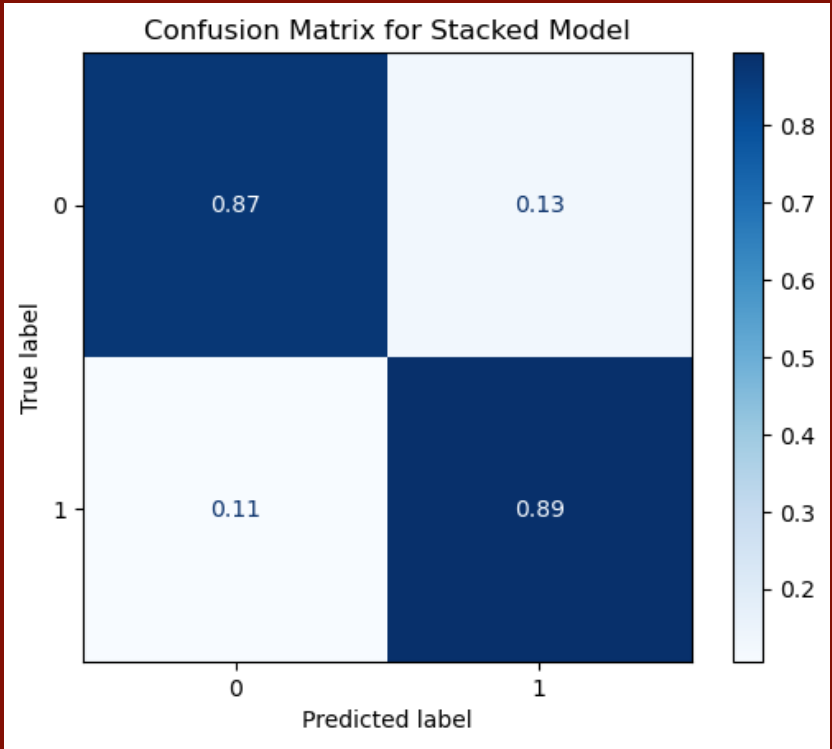
Model Evaluation Metrics

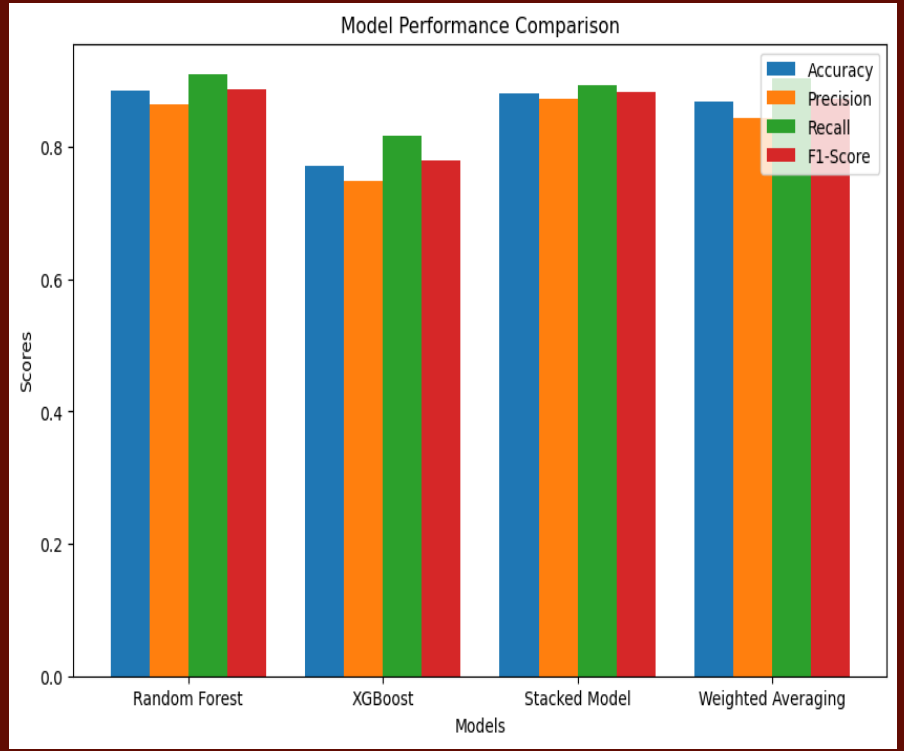
- **Random Forest: Accuracy:** 0.75, Precision: 0.74, Recall: 0.77, F!-Score 0.75
- **XGBoost: Accuracy:** 0.75, Precision: 0.73, Recall: 0.80, F1-Score 0.76
- **Logistic Regression:** Accuracy:0.75, Precision:0.74, Recall: 0.78, F1-Score: 0.76
- **SVM:** Accuracy: 0.76, Precision: 0.74, Recall: 0.81, F1 Score: 0.77


Confusion Matrix for XGBoost
0.78
0.22
0.15
0.85
True label
Predicted label
0
1
0
1
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Ensemble Techniques

- **Stacking & Voting:** Combined predictions from multiple models to use the strengths of the top two models
- Outperformed individual models overall


Confusion Matrix for Stacked Model
True label
Predicted label
0
1
0.87
0.13
0.11
0.89
0.8
0.7
0.6
0.5
0.4
0.3
0.2


Model Performance Comparison
Scores
Models
Accuracy
Precision
Recall
F1-Score
0.0
0.2
0.4
0.6
0.8
Random Forest
XGBoost
Stacked Model
Weighted Averaging

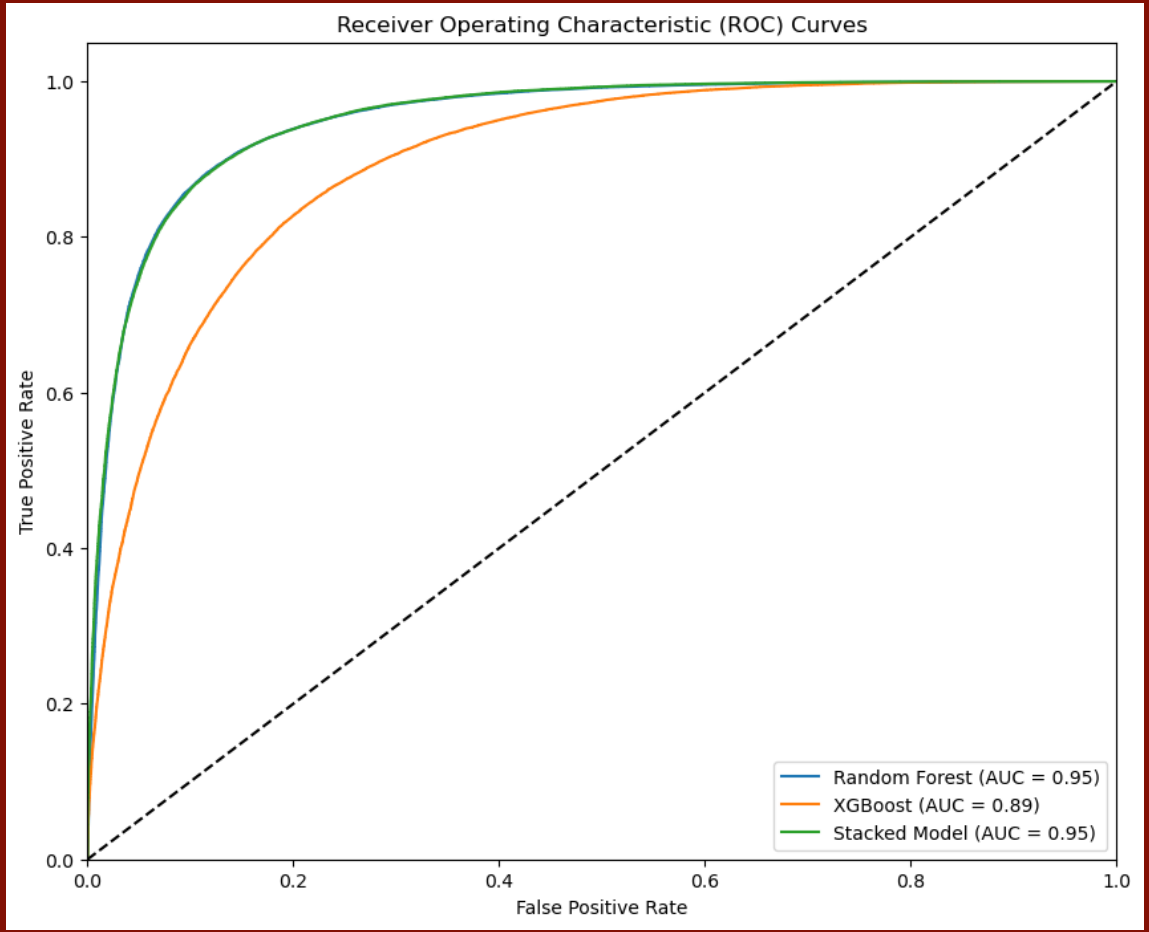
Stacked Model Performance

Accuracy: 88.19%

Recall: 89.38%

Precision: 87.30%

F1-Score: 88.33%


Receiver Operating Characteristic (ROC) Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 0.95)
XGBoost (AUC = 0.89)
Stacked Model (AUC = 0.95)

Model Results

The stacked model gave an overall improvement to the accuracy of the models

It also had the least number of false negatives which is crucial when predicting health related information

Conclusions

- Targeted Interventions: By identifying high-risk individuals based on key predictors, healthcare providers can implement personalized interventions, such as exercise programs and regular health check-ups.
- Resource Allocation: Predictive models can help allocate healthcare resources more effectively, focusing on high-risk patients and potentially reducing hospital admissions and treatment costs
- Community Outreach: Advocate for widespread lifestyle changes, public health initiatives, and early screening programs driven by predictive analytics. By encouraging healthier behaviors and early detection on a community level, we can significantly lower the long-term burden of heart disease on the healthcare system.

Future Directions

•**Broader Application**: The approach can be generalized to other health-related predictions, such as predicting other chronic diseases such as Kidney Disease and Stroke.

• **Incorporating Additional Data**: Including more data, such as detailed dietary habits or genetic information could further enhance the predictive power of the models

Questions?

Thank you for your time, I will now take any questions

View the entire project at Heart Disease Prediction Github