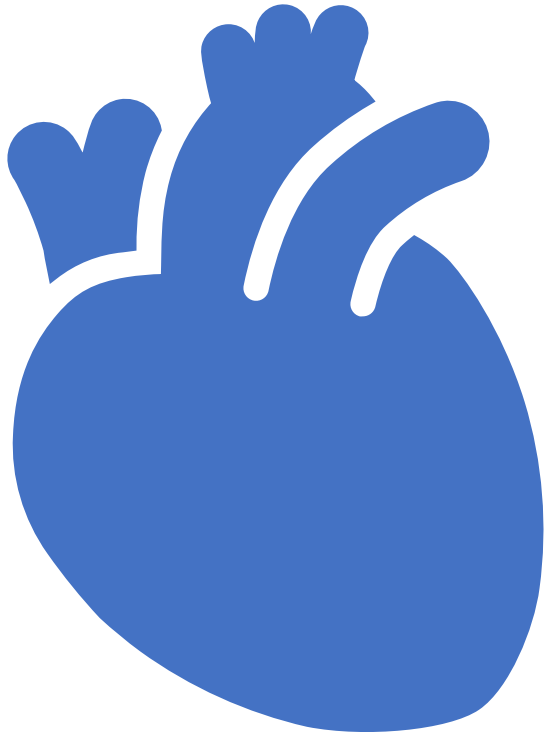


DETECTING HEART DISEASE

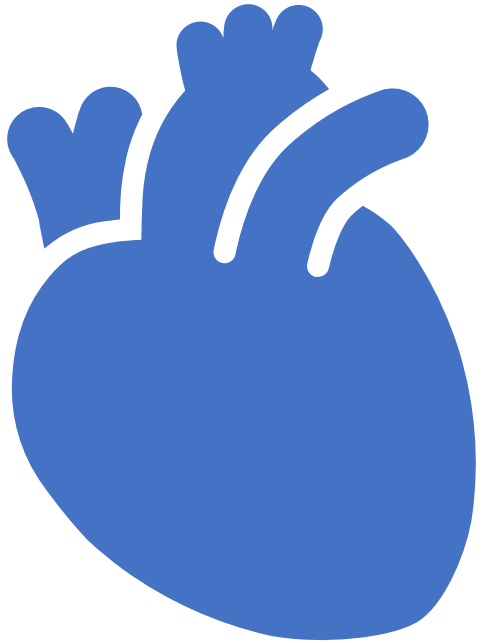
A DATA INTELLIGENCE FUSION

Data Science Capstone Introduction

By: Roland Tetteh

Problem Statement

- Globally it's estimated that 1 in 13 people are living with a heart or circulatory disease.
- In the United States, one person dies every 33 seconds from heart disease.
- According to the CDC about 5% of adults, age 20 and older have Coronary Artery Disease.
- In the United States, someone has a heart attack every 40 seconds.



Proposed Data Science Solution

In healthcare, identifying and preventing the factors that have the greatest impact on heart disease is very important.

Machine learning algorithms can detect "patterns" in data, using these risk factors, to predict a patient's condition or propensity to develop a heart disease.

This project seeks to apply machine learning techniques on a 2022 annual CDC survey data, of 400k+ adults related to their health status, to predict an individual's propensity to a heart attack.

Potential Impact

Early detection and risk assessment of heart disease.

Personalized risk assessment allowing for targeted interventions and prevention strategies.

Potentially be used in drug discovery.

Data Quality

- The dataset comes from the CDC, which conducts annual telephone surveys to collect data on the health status of U.S. residents.
- The original dataset of nearly 300 variables was reduced by the author to 40 most relevant variables.
- There were 157 duplicate rows out of the 445,132 rows and missing values in 38 out of the 40 columns.
- Some of the attributes in the data are State, Sex, MentalHealthDays, and HadDiabetes.
- The dataset is very imbalanced.

Data Preprocessing

Categorical to Numerical Conversion.

Detecting and removing multicollinearity via the Variance Inflation Factors.

Upsampling, to deal with class imbalance.

Important Findings

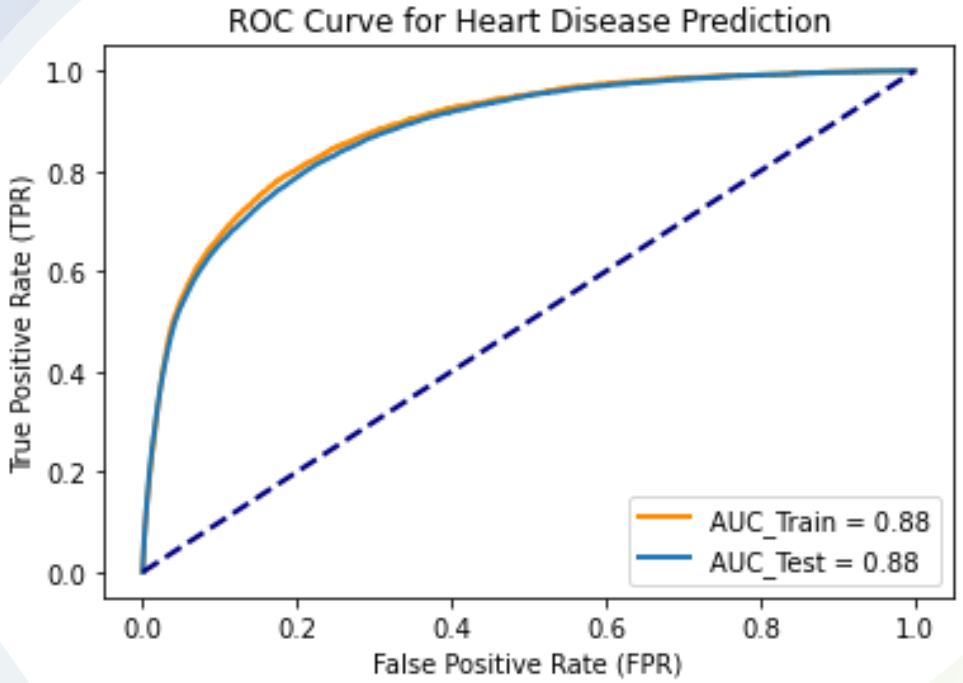
Notably the below are significant in determining an individual’s odds in having a heart attack:

- The sex
- Number of days where an individual's physical health was not good
- The number of days where an individual's mental health was not good
- The average number of sleep hours in a day
- An individual's height and weight

1st Model – Logistic Regression

Using the upsampled training data, scaling the data and applying GridsearchCV over a range of C values,

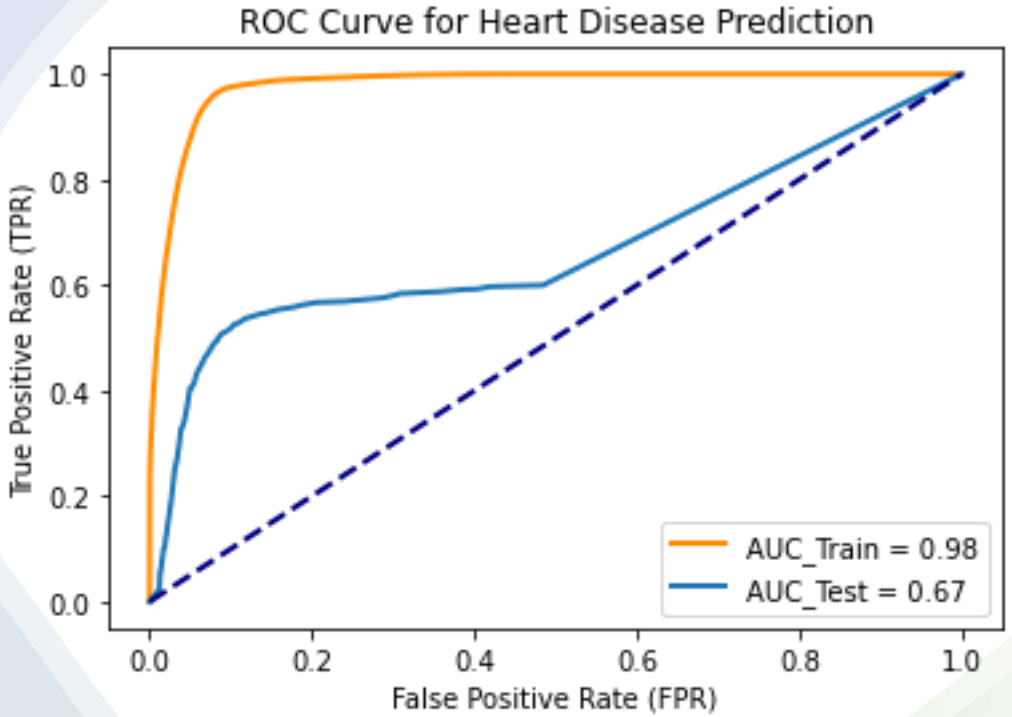
- Best C Value was 0.01
- AUC Test Score of 0.88
- Allowing for a FPR of 0.2, a TPR of 0.79 can be achieved



2nd Model – Decision Tree Classifier

Using the upsampled training data and applying GridsearchCV over a range of maximum depth and minimum samples split values,

- Best Max Depth was 20
- Best minimum samples split was 2
- AUC Test Score of 0.67
- Allowing for a FPR of 0.2, a TPR of 0.57 can be achieved



3rd Model – Support Vector Machines

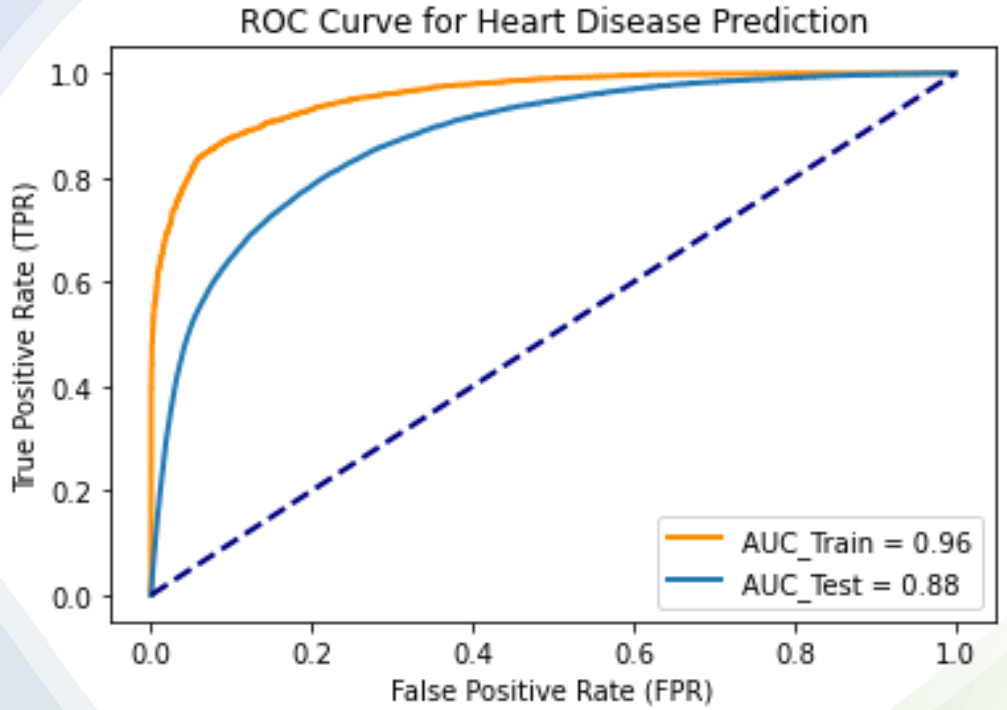
Using the upsampled training data, scaling the data and applying GridsearchCV over a range of C values,

- Best C value was 1
- Best Train AUC was 0.88

4th Model – Random Forest Classifer

Using the upsampled training data, scaling the data and applying GridsearchCV over a range of hyperparameters values,

- Best Max depth value was 30
- Minimum Samples Split of 2
- Number of Estimators was 30
- Best test AUC was 0.85
- For a FPR of 0.2, a TPR of 0.75 can be achieved



Next Steps...

Applying advanced ensemble learning methods based on the best algorithms.

Applying Unsupervised Learning