MODEL TO PREDICT POSSIBILITY OF

HEART DISEASE

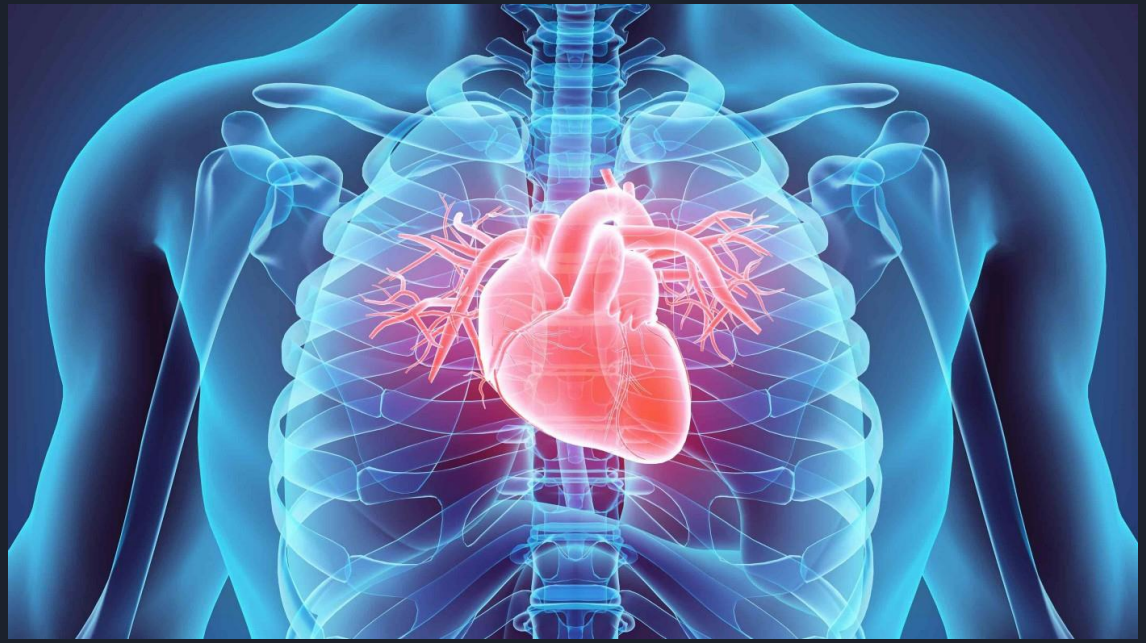
Minh Tran

Motivation

Leading causes of death in the U.S

Predict whether an individual has heart disease

Receive the necessary care and treatment to reduce the burden of their disease



Data Overview

Data Source: Personal Key Indicators of Heart Disease (Kaggle & CDC)

Before Data Processing

Item	Stats
Total variables	18
Multi-categorical variables	03
Numerical variables	04
Binary variables	11
Total records (rows)	319,795
% Heart disease (Yes)	9%

After Data Processing

Item	Stats
Total variables	23
Multi-categorical variables	00
Numerical variables	06
Binary variables	17
Total records (rows)	319,795
% Heart disease (Yes)	9%

Data Processing

Step 1: Check unique values and missing values of all variables

Step 2: Handle categorical/missing data

- Replace all binary values (Yes/No) with (1/0)
- De-binning of Age with mean values of each age category (ordinal variable)
- One hot encoding of Race (nominal variable)
- Replace GenHealth category with WOE values (weight of evidence)
- No missing data

Step 3: Check correlation among variables to remove one that have strong correlation to others

Data Processing

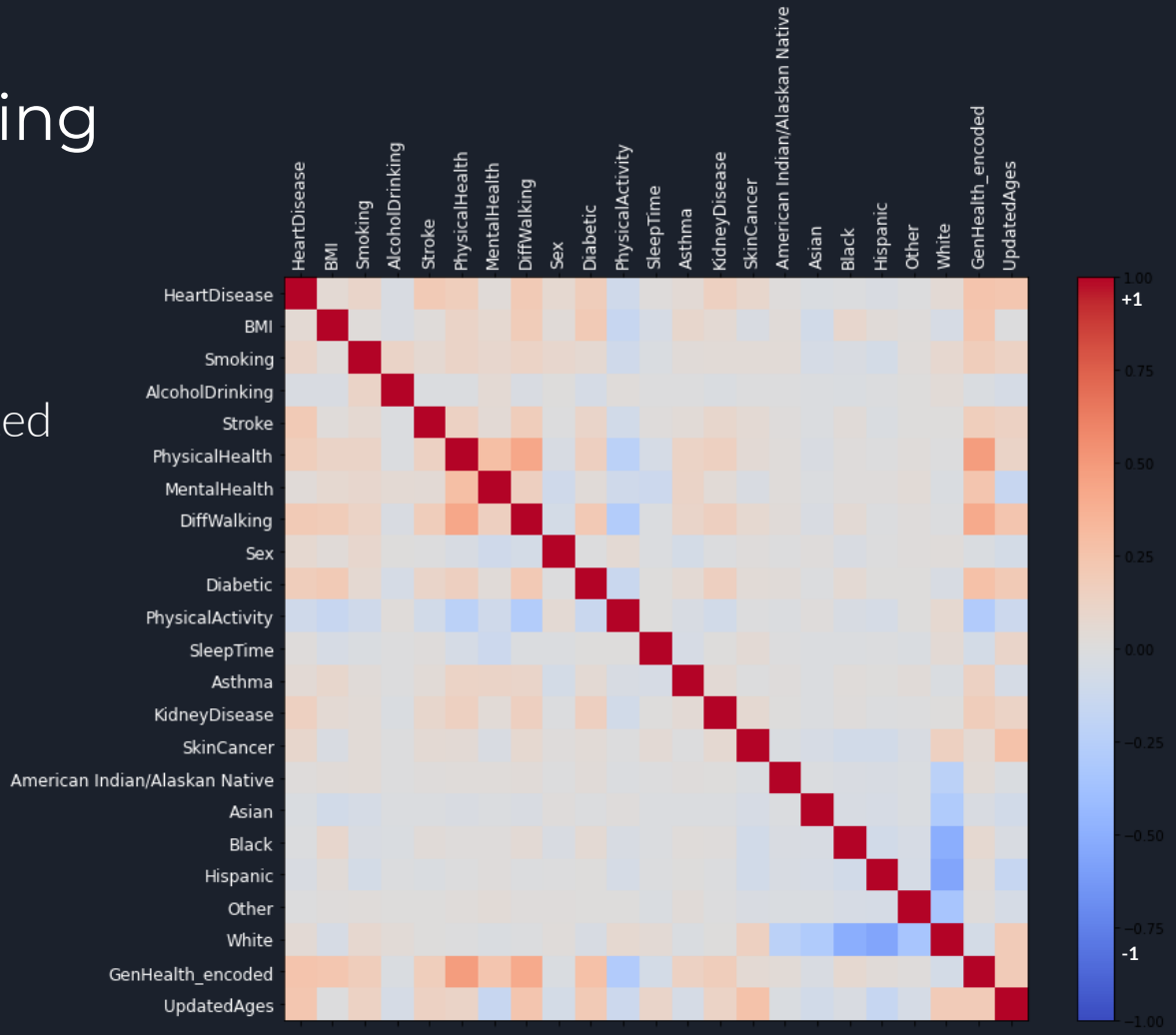
Step 3:

No strong correlation detected

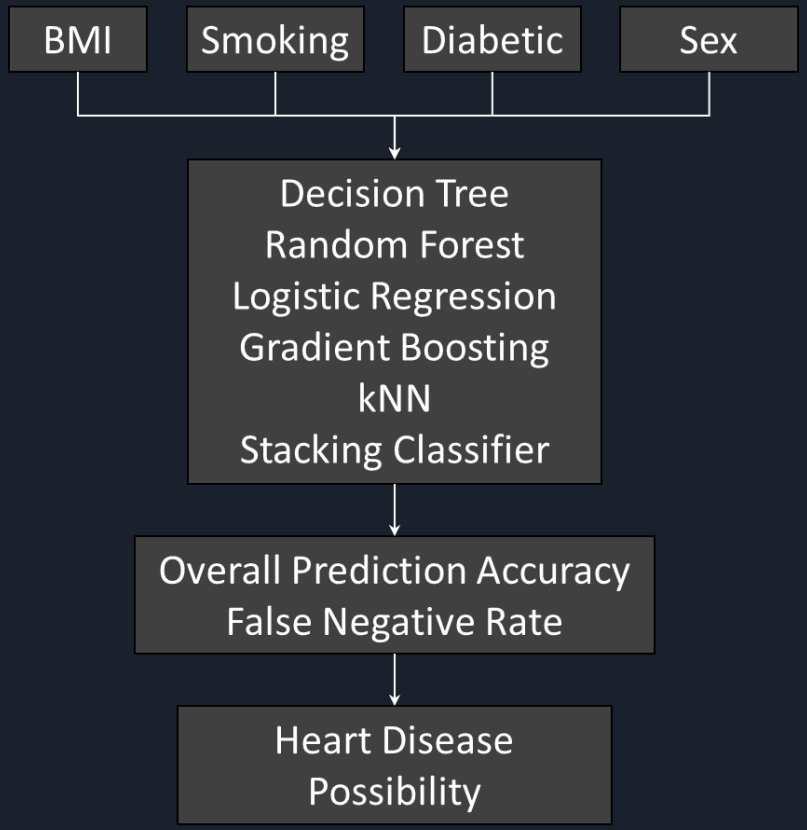
Step 4:

Split data set into:

- 80% train
- 10% validation
- 10% test



Modeling Process



1. Example Features

2. Prediction Models

(leveraging GridSearchCV)

3. Model Evaluation

4. Impact Prediction

Model Selection

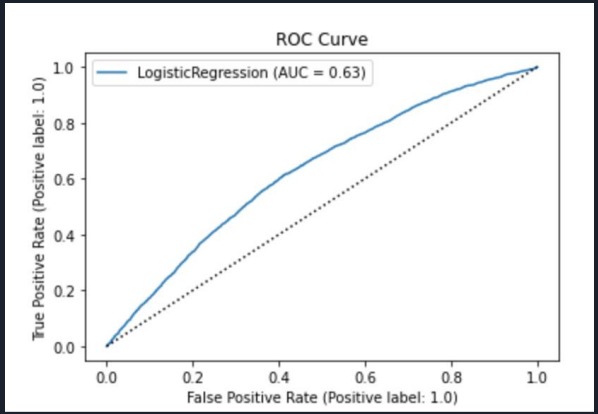
Rank	Models	Accuracy Score (Validation Set)
1	Logistic Regression	0.917
2	Decision Tree Classifier	0.914
3	Gradient Boosting Classifier	0.914
4	Stacking: 1st layer estimator = Decision Tree Classifier Final layer estimator = Logistic Regression	0.914
5	k-NN Classifier	0.910
6	Stacking: 1st layer estimator = Random Forest Classifier Final layer estimator = Logistic Regression	0.909
7	Random Forest Classifier	0.905
8	Stacking: 1st layer estimator = Logistic Regression Final layer estimator = Random Forest Classifier	0.873

Model Evaluation

Logistic Regression - ROC Curves & FNR

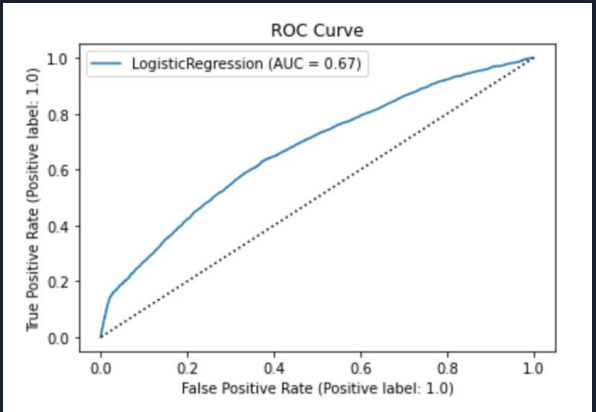
Two Features:

False Negative = 1.0



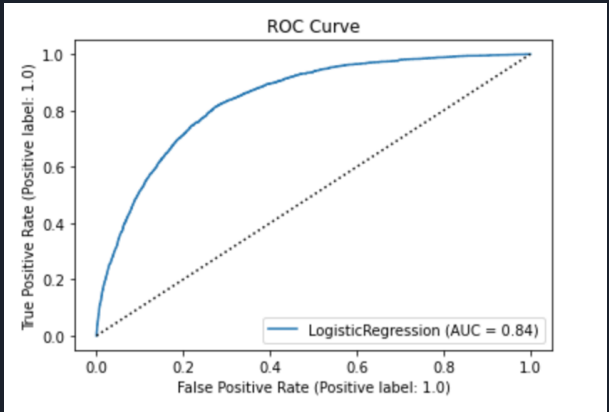
Four Features:

False Negative = .989



All Features:

False Negative = .885

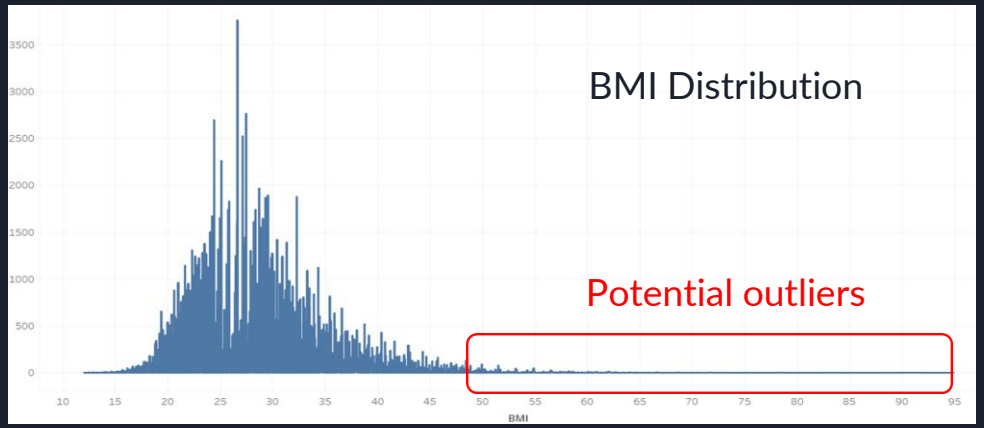


Outcomes

Testing Data Accuracy = 0.915

Areas for Improvement

- Reduce data skewness by adding more data from
  - Other races
  - Individuals with heart disease
- Increase accuracy by removing potential outliers



THANK YOU!