Heart Disease Classification

Brice Kramer
Derek Hu
Josh Fitts
John Medina
Andrew Falcone

Problem Statement

- Heart Disease is the leading cause of death in America is heart disease, taking more than 800,000 lives every year.
- Goal is to create a model to classify the risk of heart disease for individuals through a free and accessible web application.
- Help users understand which risk factors have the greatest impact on heart disease and raise awareness.

Data Overview

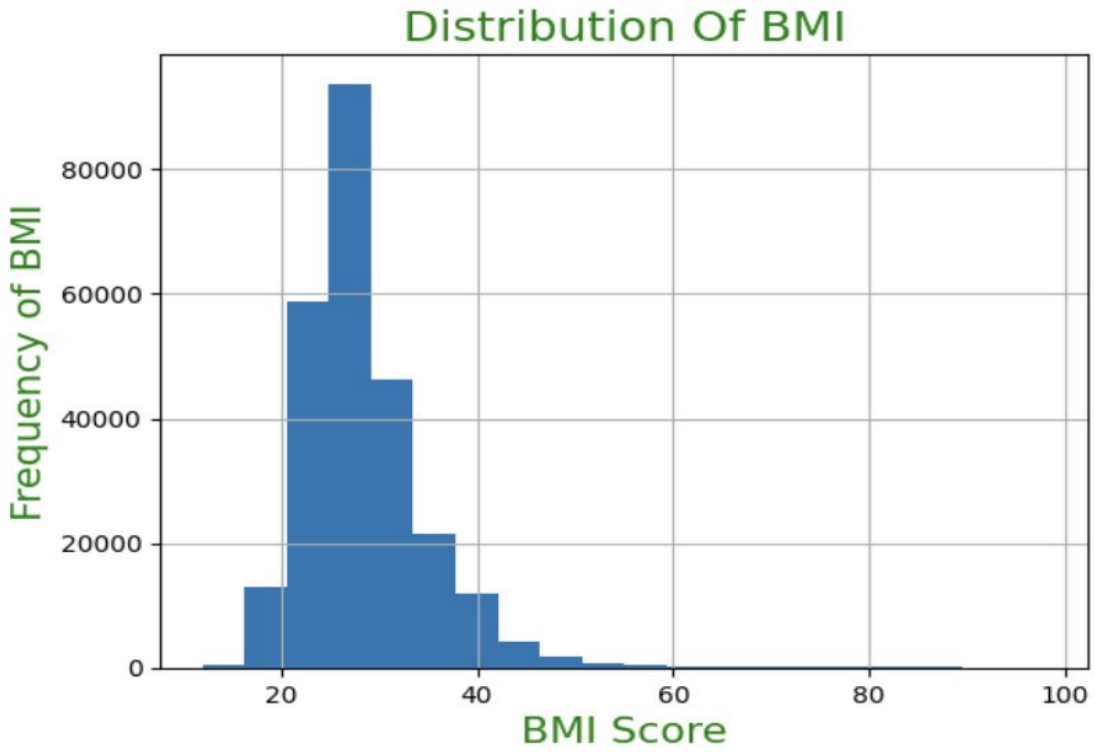
- Kaggle
- Original Data source: CDC: 2015 Behavioral Risk Factor Surveillance System
- 253,680 survey responses
- Features include:
    - Mental Health, exercise, BMI, high cholesterol, smoker, alcohol consumption, ect.

EDA (Exploratory Data Analysis)

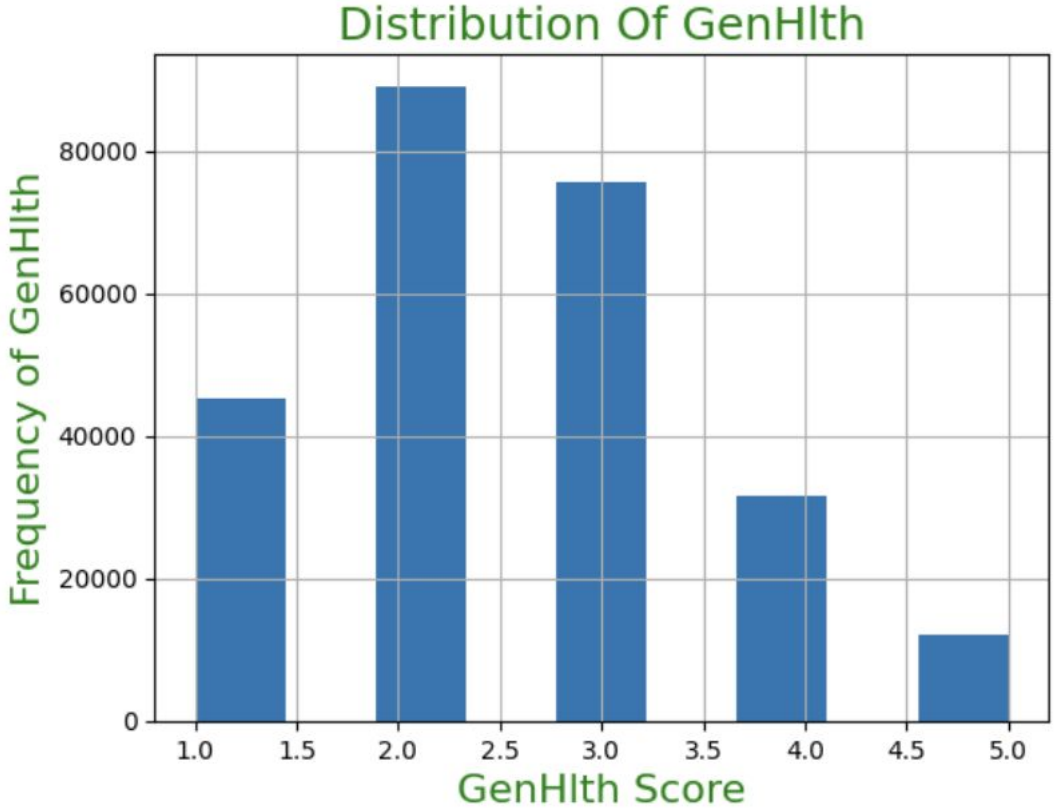
Data Cleaning

- Data was already cleaned
- No outliers, errors, or missing values
- Dependent variable ‘HeartDiseaseAttack’ was converted from boolean values to binary
- Use correlation for feature selection to explore relationship between variables

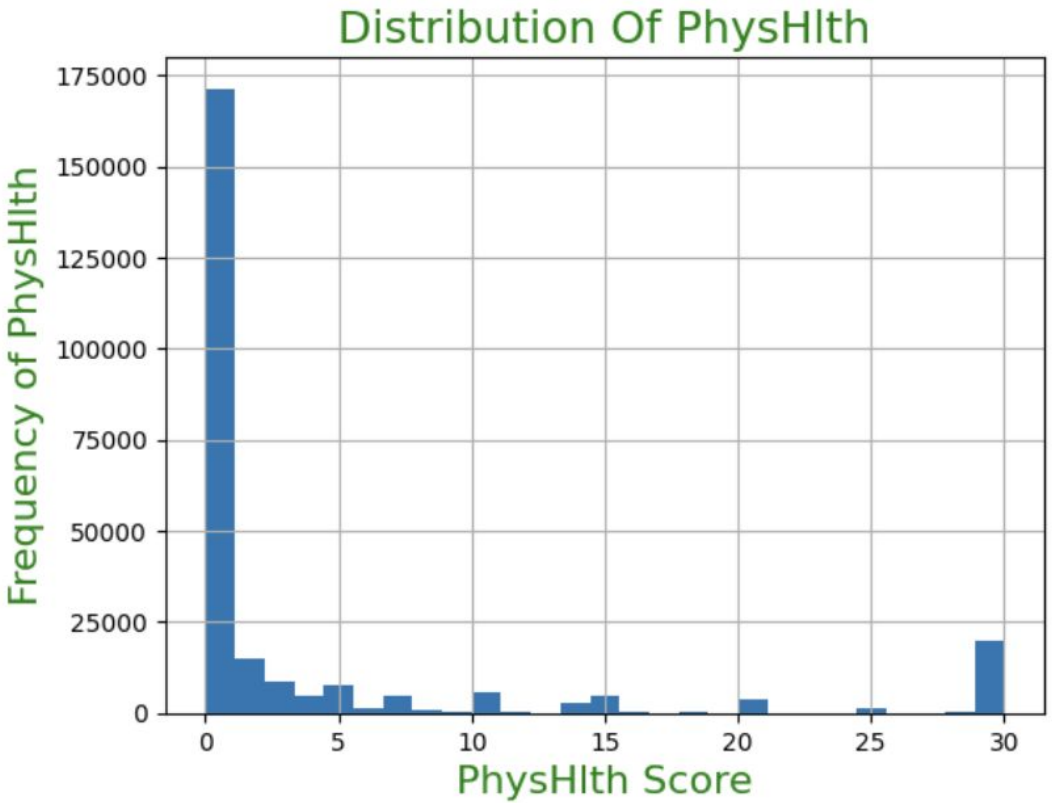
Data Visualizations



Data Visualizations

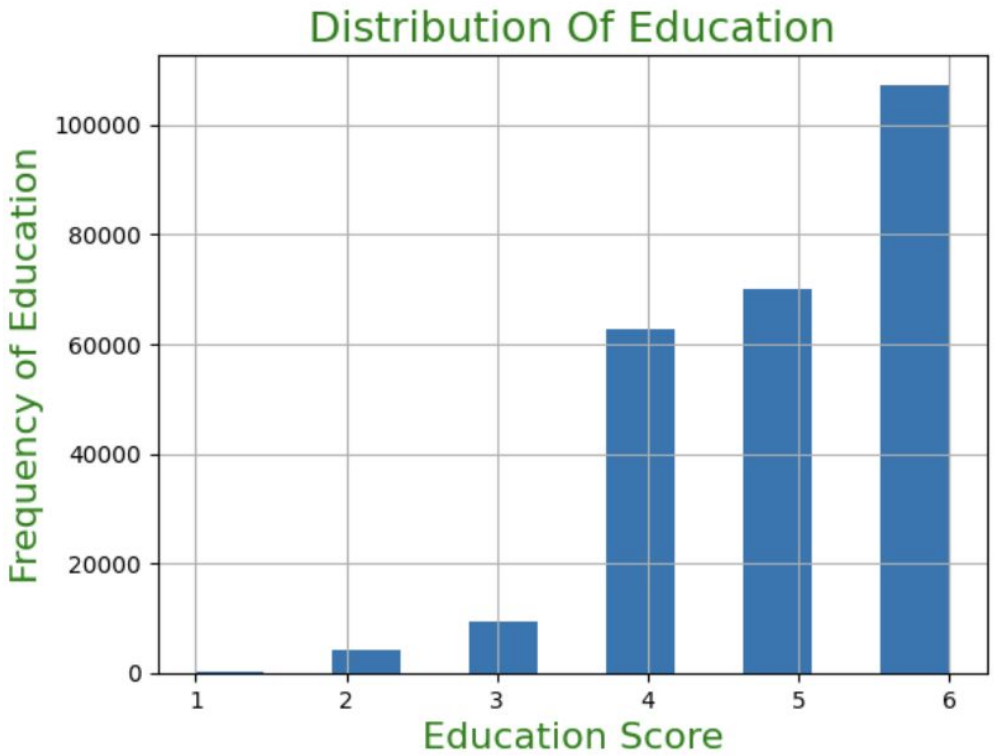


Data Visualizations - Continued



Data Visualizations - Continued

- 1 - 'Never attended school or only kindergarten'
- 2 - 'Grades 1 through 8 (elementary)'
- 3 - 'Grades 9 through 11 (Some high school)':
- 4 - 'Grades 12 or GED (High school graduate)'
- 5 - 'College 1 year to 3 years (Some college or technical school)'
- 6 - 'College 4 years or more (College graduate)'



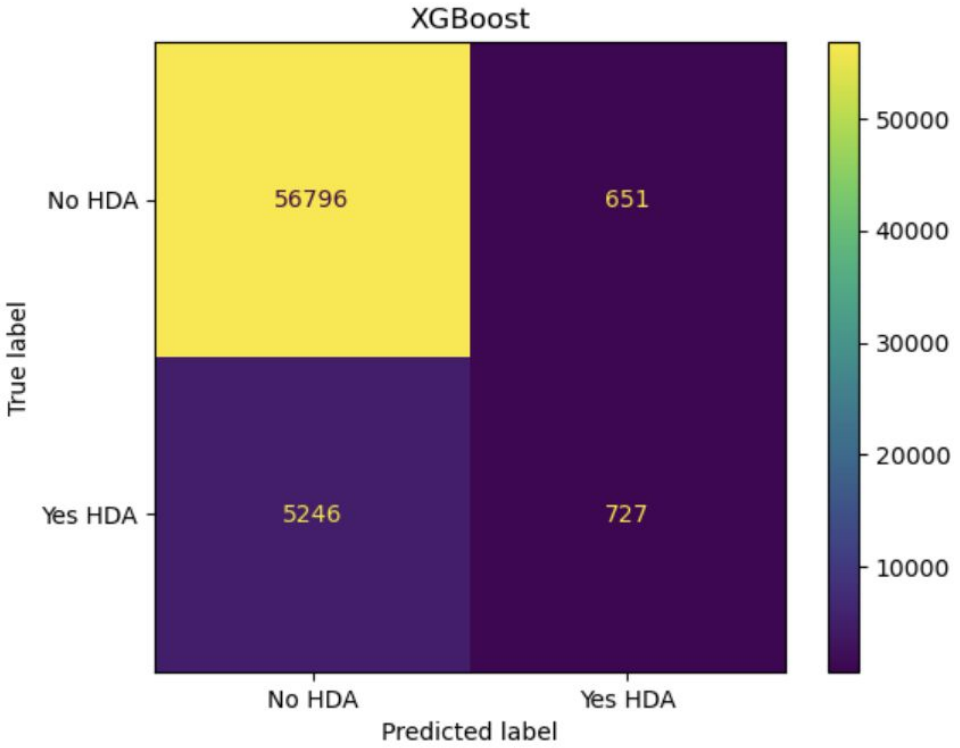
Data Modeling

- Standardized and created models using Pipeline
- Used several types of models - ADABoost, GradientBoost, and XGBoost, and a feed forward neural network
- Accuracy Results:
  - ADABoost (Training Score - 90.7%, Testing Score - 90.7%)
  - GradientBoost (Training Score - 90.9%, Testing Score - 90.8%)
  - XGBoost (Training Score - 91.5% Testing Score - 90.7%)

Confusion Matrix - XGBoost

```
True positives:  727
False positives:  651
True negatives:  56796
False negatives:  5246
```

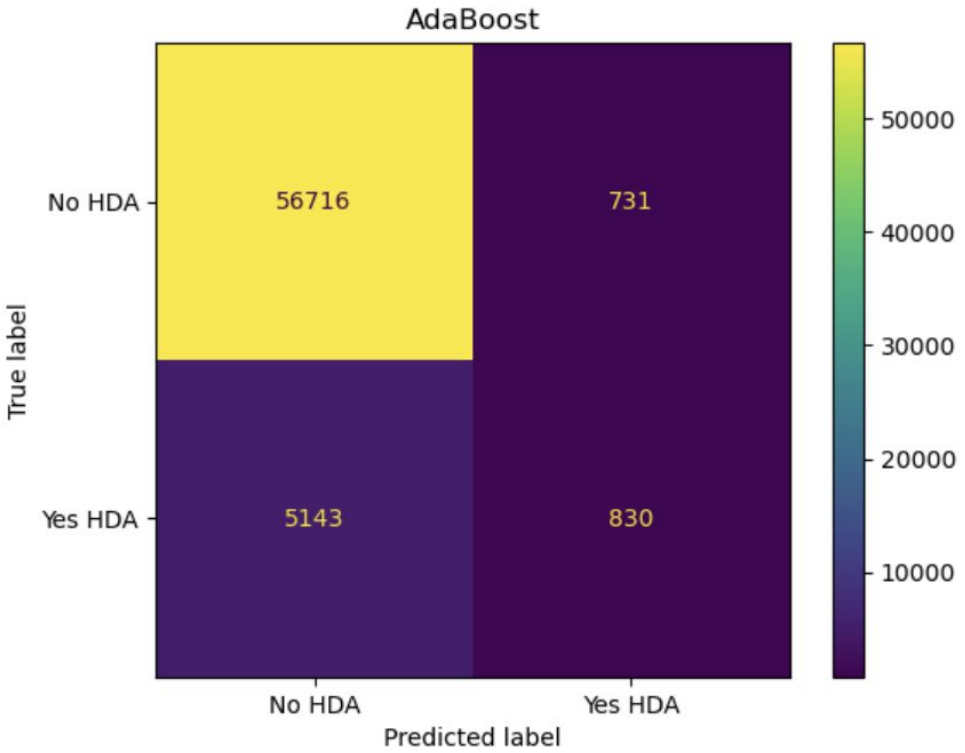
Recall Score: 12.2%
Accuracy: 90.7%



Confusion Matrix - AdaBoost

True positives: 830
False positives: 731
True negatives: 56716
False negatives: 5143

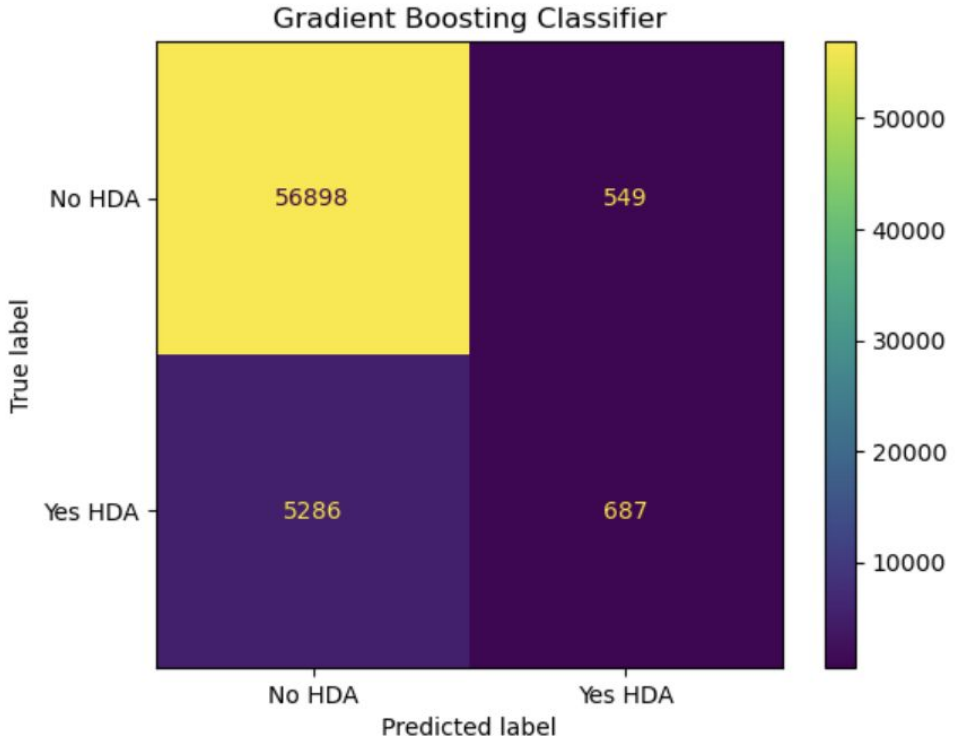
Recall Score: 13.9%
Accuracy: 90.7%



Confusion Matrix - BoostingClassifier

```
True positives:  687
False positives:  549
True negatives:  56898
False negatives:  5286
```

Recall Score: 11.5%
Accuracy: 90.8%

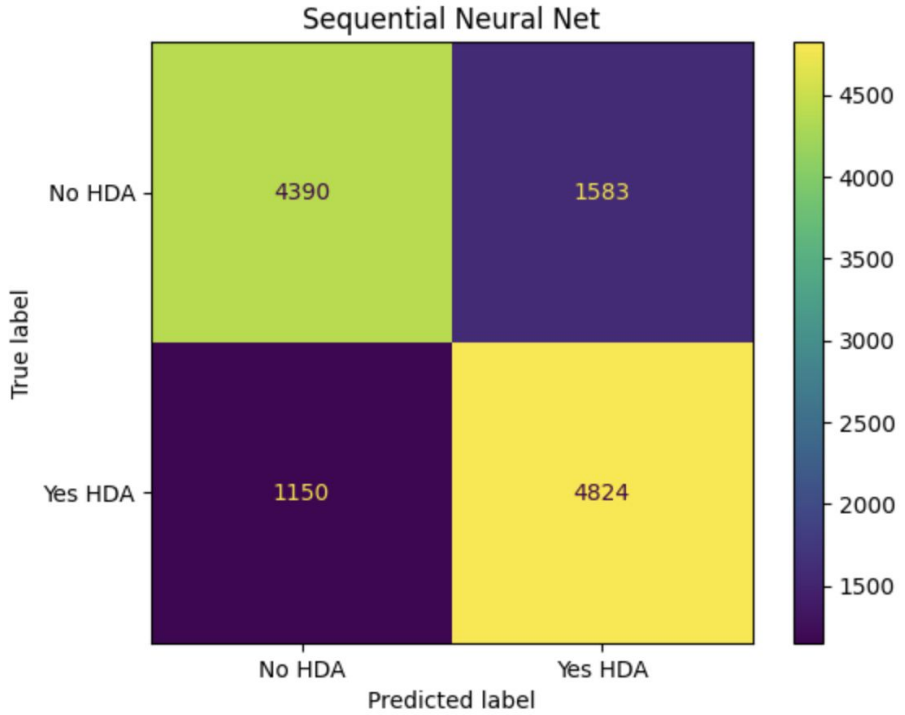


Imbalance Problem

- Our models consistently tested at 90%
- Misclassifying people with heart disease
- Solution:
  - Under sample the majority class
  - Fit the model on balanced subset with all of the minority class and a random sample of the majority class

Feed Forward Neural Network

- Accuracy = 77.1%
- Recall = 80.7%
- Precision = 75.3%



Streamlit

- Replicated the survey with the features we used
- Users can identify their risk of having heart disease
- Uploaded the neural network model and used it to make heart disease risk predictions
- https://typikal1-dsi-project-4-heart-disease-app-xtqo6y.streamlitapp.com

Conclusions

- We were able to predict heart disease risk with an accuracy of 77% with a balanced model
- Improved recall score
- BMI doesn’t have a major impact
- Age has a major impact