Deloitte.

FIND YOUR EXTRAORDINARY AI

Detection of Coronary Heart Disease in Adults Using a Random Forest Classifier

A quick and non-invasive way to assess the possible presence of Coronary Heart Disease

By Chris Christensen

MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

What is Heart Disease?

- Heart disease is a general term that refers to various conditions that affect the heart.

- The most common type of heart disease in the United States is coronary artery disease (CAD)

- CAD is caused by plaque buildup in the walls of the coronary arteries that supply blood to the heart

- This plaque build up causes arteries to narrow over time, which can partially or totally block the blood flow to the heart.

Deloitte.

Why it Matters

- Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States.

- About 697,000 people in the United States died from heart disease in 2020—that’s 1 in every 5 deaths.

- Heart disease cost the United States about $229 billion each year from 2017 to 2018. This includes the cost of health care services, medicines, and lost productivity due to death.

MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

Coronary Artery Disease

- Coronary Artery Disease (CAD), the most common type of heart disease, killed 382,820 people in 2020
- About 20.1 million adults aged 20 and older have CAD (about 7.2%).
- In 2020, about 2 in 10 deaths from CAD happen in adults less than 65 years old.
- Unfortunately, the first sign of CAD is a often heart attack.

MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

Purpose of this Project

- To address this growing health crisis by offering a solution that will first and foremost save lives but also reduce the cost and strain on the healthcare system caused by it.

- Create a quick, easy and non-invasive way to detect the possible presence of CAD that will lead to early detection, treatments, and discussions with a doctor.

- Make detection of CAD a greater possibility for people belonging to underserved and minority communities.

MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

Solution

Using a Random Forest Machine Learning Model to detect the possible presence of Coronary Artery Disease in Adults

- Takes in simple, easy to obtain metrics
- Quick and non-invasive
- Can be performed by a medical assistant or even at home.

MAKING AN IMPACT THAT MATTERS since 1845

Data and Methods

- Data was acquired from the National Health and Nutrition Examination Survey (NHANES), specifically data from the 2013-2014 Survey.

- “NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.” (1)

- “The survey is unique in that it combines interviews and physical examinations.” (1)

Deloitte.

Data and Methods

The features used to train the model were related to certain risk factors for CAD in 4 main areas

- These include demographic Data: Age, Gender, and Race
- Examination Data: BMI, Pulse, Blood Pressure
- Questionnaire Data: Time Spent Sedentary and Smoker vs. Non-Smoker
- Laboratory Data: Blood Glucose Levels and Non-HDL Cholesterol levels

MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

Data and Methods

The data was prepared fore modeling following an ELT process

- After Data extraction from a NHANES source, the data was loaded into a SQLite database for easier querying and manipulation.

- A Python library was used to connect connects to the SQLite server to query and read the necessary Data.

- Python was then used to clean and prepare the data for modeling.

MAKING AN
IMPACT THAT
MATTERS
since 1845

Data and Methods

- One challenge to creating an effective model was the imbalanced nature of the Data, out of a total of 5,328 rows there were only 215 observations with a confirmed CAD diagnosis.

- To help address the issue of imbalanced data a technique called SMOTE was used, which stands for "Synthetic Minority Oversampling Technique"

- In simple terms, it takes the minority class and creates synthetic data similar to the original data.

Data and Methods

- Another way to address the issue of imbalanced data was to evaluate the model’s performance on the right metric.

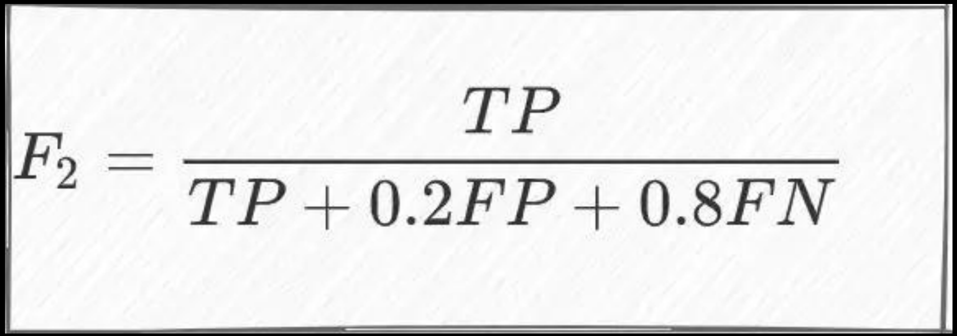
- For example, if all 215 diagnoses of CAD were predicted negative resulting in 215 false positives, the model accuracy would still come out to be 95%.

- It was determined that minimizing the rate of false negatives is most crucial in this scenario.

Data and Methods

- Therefore, the F2 score was chosen to evaluate model performance rather than overall accuracy.

- The F2 score takes in account both False Positives and False Negatives but puts more weight on false negatives as shown.



$$F_2 = \frac{TP}{TP + 0.2FP + 0.8FN}$$

Data and Methods

The Data was then trained using 3 different classification algorithms in order to determine which would produce the lowest number of false negatives:

1. A Decision Tree Classifier

2. Random Forest Classifier

3. XGBoost Classifer

Deloitte.

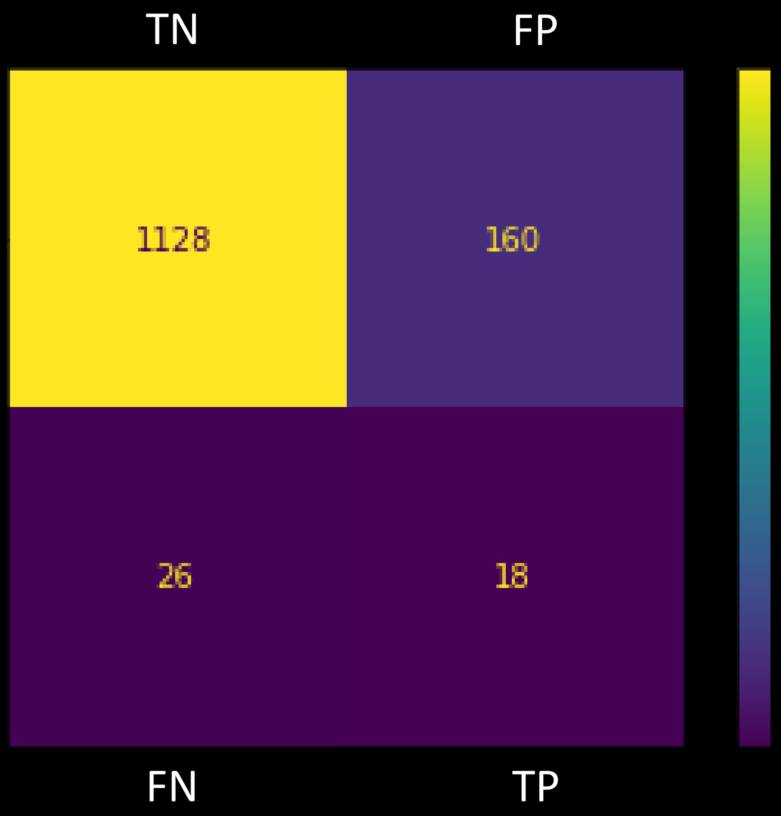
Results

Decision Tree Classifier:

Accuracy: 86.04

F2 Score: 25.42

True Negatives: 1128
False Negatives: 26
False Positives: 160
True Positives: 18



MAKING AN IMPACT THAT MATTERS since 1845

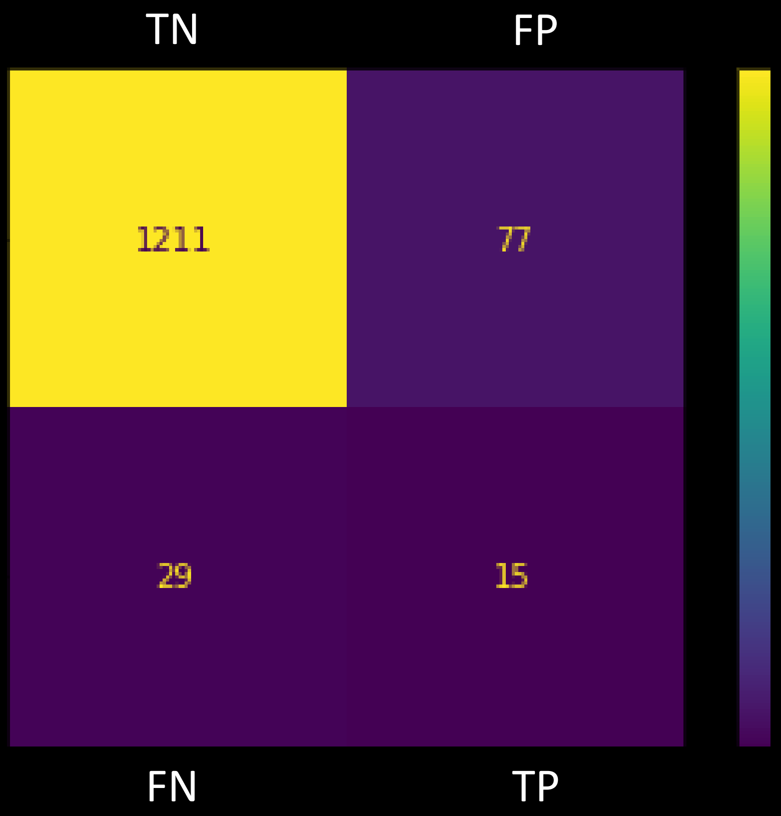
Results

Random Forest Classifier:

Accuracy: 92.04

F2 Score: 27.99

True Negatives: 1211
False Negatives: 29
False Positives: 77
True Positives: 15



Deloitte.

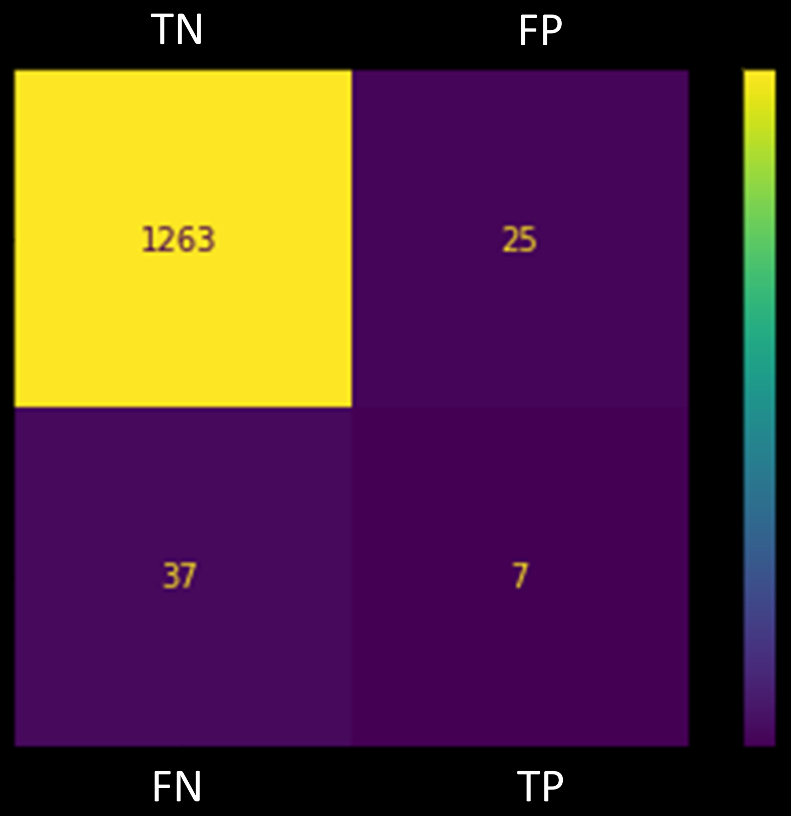
Results

XGBoost Classifier:

Accuracy: 95.35

F2 Score: 16.83

True Negatives: 1263
False Negatives: 37
False Positives: 25
True Positives: 7



MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

Results

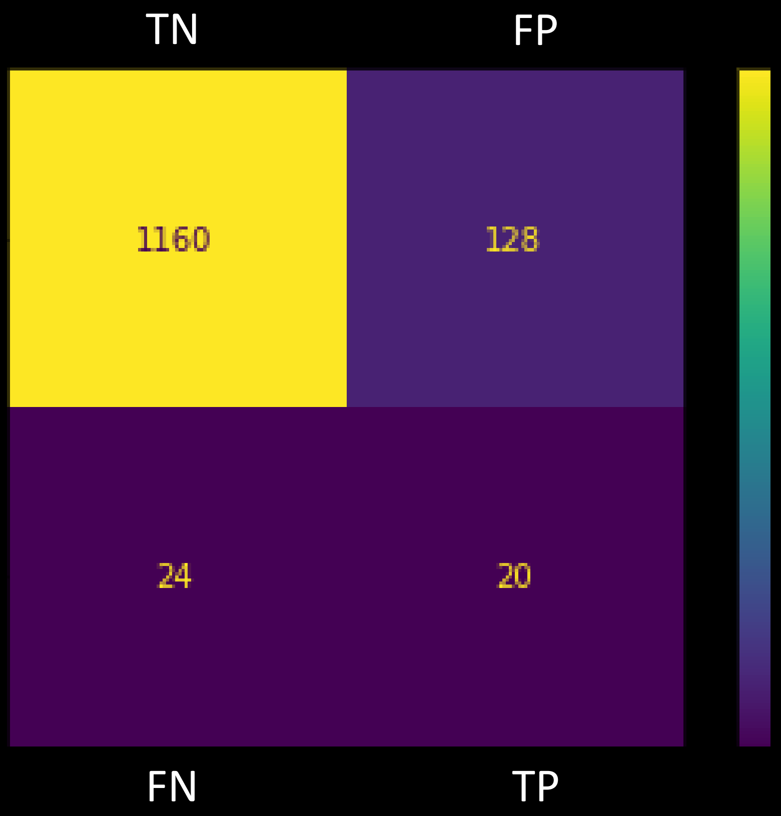
Random Forest Classifier:

(After Hyperparameter Optimization)

Accuracy: 88.59

F2 Score: 30.86

True Negatives: 1160
False Negatives: 24
False Positives: 128
True Positives: 20



MAKING AN IMPACT THAT MATTERS since 1845

Conclusion

A Random Forest Machine Classifier is the recommended model in detecting the possible presence of coronary artery disease while reducing the rate of false negatives.

This model would assist:

1. Saving lives through early detection, especially in the medically underserved and minority areas.
2. Reducing the Strain on the Healthcare system.
3. And relieving the financial burden associated with the cost of treating heart disease.

MAKING AN IMPACT THAT MATTERS since 1845

Deloitte.

References:

1. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm
2. https://www.cdc.gov/heartdisease/about.htm
3. https://www.cdc.gov/heartdisease/coronary_ad.htm
4. https://www.cdc.gov/heartdisease/facts.htm
5. https://www.nhlbi.nih.gov/health/coronary-heart-disease

MAKING AN
IMPACT THAT
MATTERS
since 1845

Deloitte.