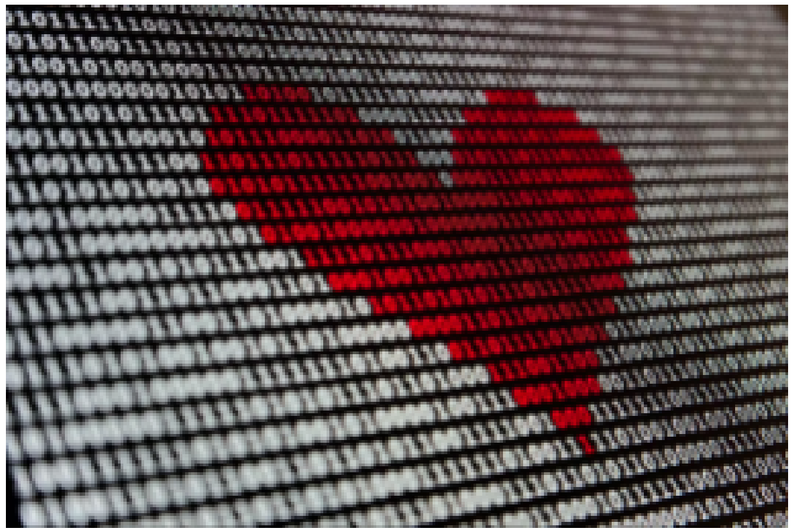
HEART DISEASE:
MACHINE LEARNING CLASSIFICATION

TRAVIS ROYCE

6/20/22

PROJECT OVERVIEW



HEART DISEASE

This project utilized a CDC dataset to create and iterate on machine learning models to predict if an individual has heart disease.

BUSINESS OVERVIEW

HEART DISEASE

The business goal is to create a final model that predicts if a person has heart disease, based on questionnaire answers regarding the individuals lifestyle.

THE FINAL PRODUCT:

WEB/PHONE APLICATION

The deliverable of this project is a web and phone application which can predict if an individual has heart disease from a few simple questions, to be used by individuals and/or their health care team.

BUSINESS OVERVIEW

THE GOAL METRIC:

Because the dataset is heavily skewed -- with only 8% of observations affirmatively having heart disease -- accuracy (the percentage of correct predictions) is not a helpful indicator of success, as a model that guessed "no heart disease" for **every** observation would achieve a 92% accuracy.

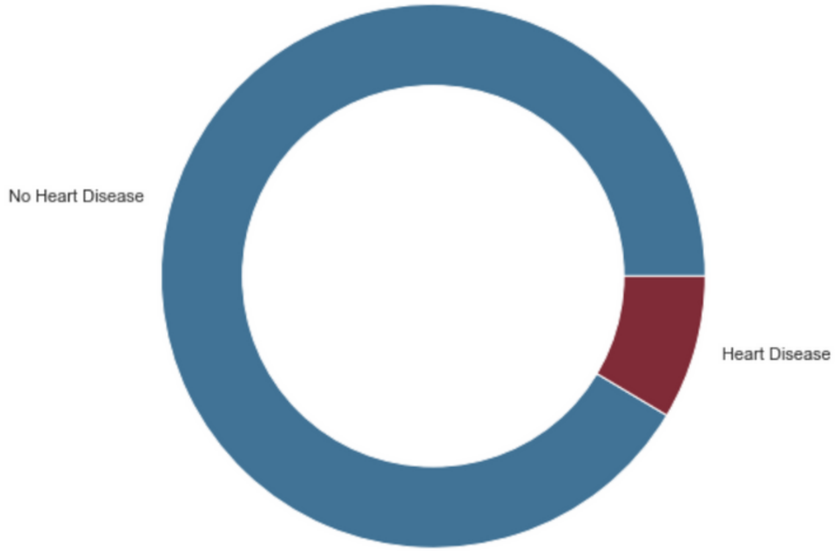
- **False negatives** could cause patients *with* heart disease to forgo further testing. This would be the worst possibility, out of the options.

- **False positives** would cost more due to testing people who did not actually have heart disease, or could cause people without heart disease to needlessly worry about their health. This is also costly, but not as costly as missing an individual with heart disease. Further, false positives may indicate an individual is at higher risk of heart disease, which can be vital information.

The **F1 Score** is the geometric mean of recall and precision.

For my goal metric, I created a **recall-weighted F score** by adding a 2x weight to the recall (i.e., false negatives) in this equation.

TARGET VARIABLE CLASS IMBALANCE



DATA OVERVIEW

THE DATASET

This dataset comes from the CDC and is a part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents.

According to the CDC, the BFRSS "collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world."

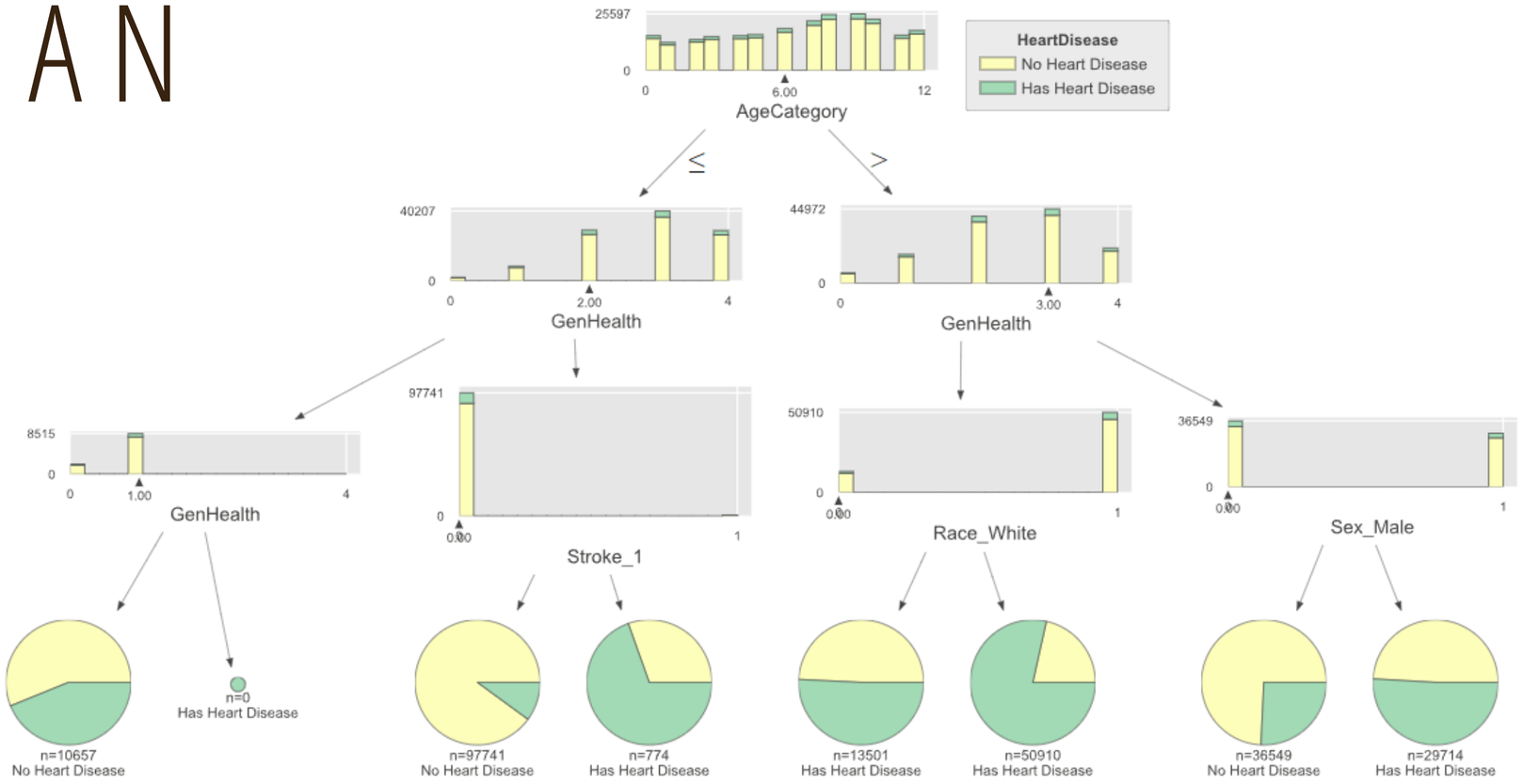
The data contains **18 variables** and approximately **320,000 observations**.

The variables in the dataset concerned:

- Heart Disease (the Target variable)
- Smoking
- Alcohol Drinking
- Having a Stroke
- Physical Health
- Mental Health
- Difficulty Walking
- Sex
- Age
- Race
- Diabetes
- Physical Activity
- General Health
- Asthma
- Kidney Disease
- Skin Cancer
- Average Sleep
- BMI

MODELS RAN

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **XGBoost**
- **Bagged Trees**
- **Extra Trees**
- **KNN**

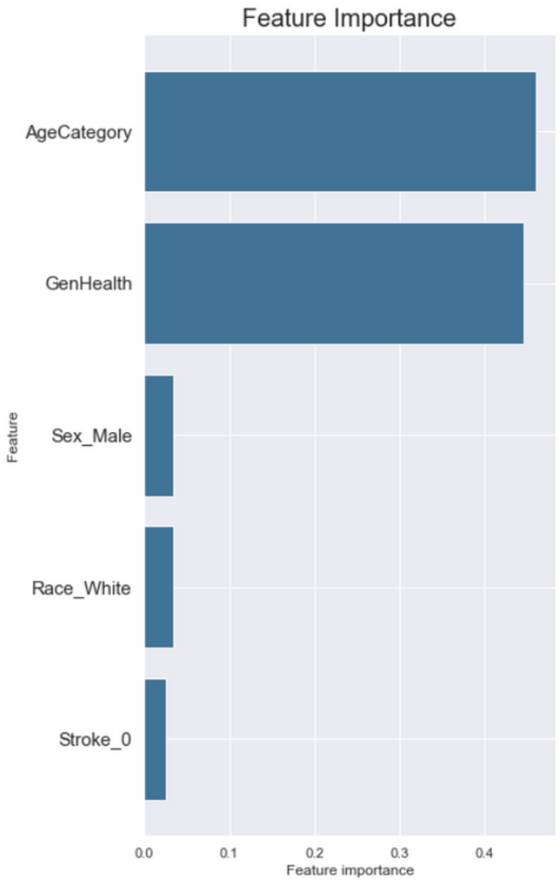


For illustration purposes.
Decision Tree with a Maximum Depth of three and a recall-weighted F1 score of 0.31

INITIAL MODEL: DECISION TREE

MOST IMPORTANT FEATURES

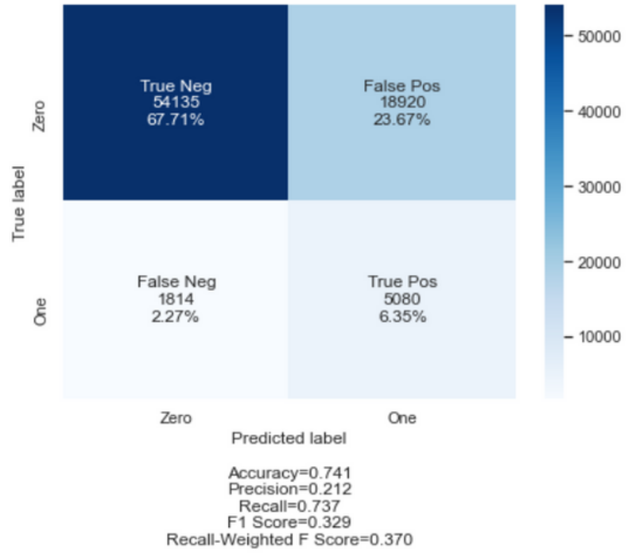
- The **age range** was the most important feature.
- **General health**, on a scale from 1-3, was the second.
- **Sex** was the next most important feature. Specifically, it is much more likely for men to have heart disease.
- The final two features were **race**. Specifically, if the observation was white, and if the observation had previously had a stroke, they were more likely to have heart disease.



RESULTS

The **recall-weighted F score** was **.395**.
While the model only predicted 1,500 false negatives (1.9% of total), it predicted over 18,000 false positives (or 23.1% of total).

I look to improve the results by checking other models, and optimizing the most promising.



INITIAL MODELS: **TESTING RESULTS**

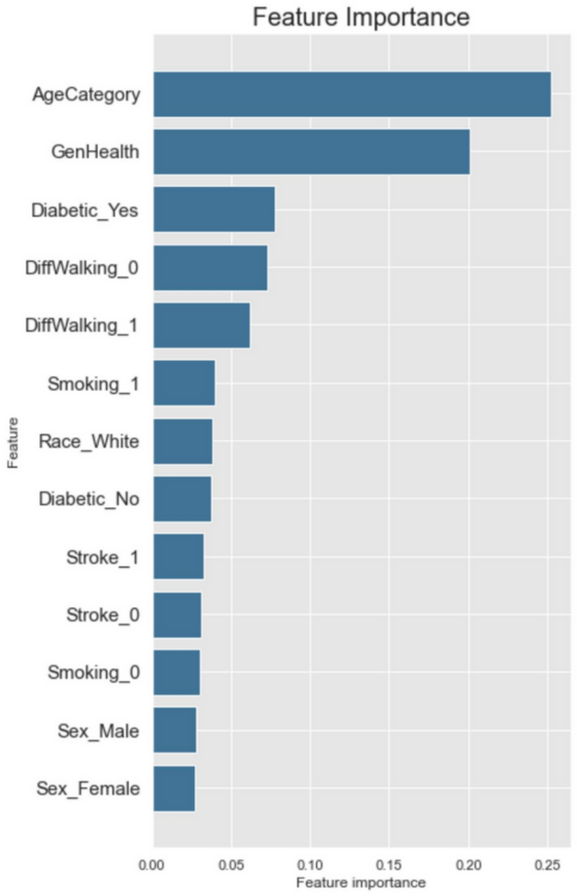
Model	RWF Score
Random Forest	*.404*
Logistic Regression	*.395*
Extra Trees	*.392*
XGBoost	*.38*
Bagged Trees	*.37*
Decision Tree	*.37*
KNN	*.346*

TOP MODELS: **ITERATION TESTING**

FINAL MODEL RESULTS

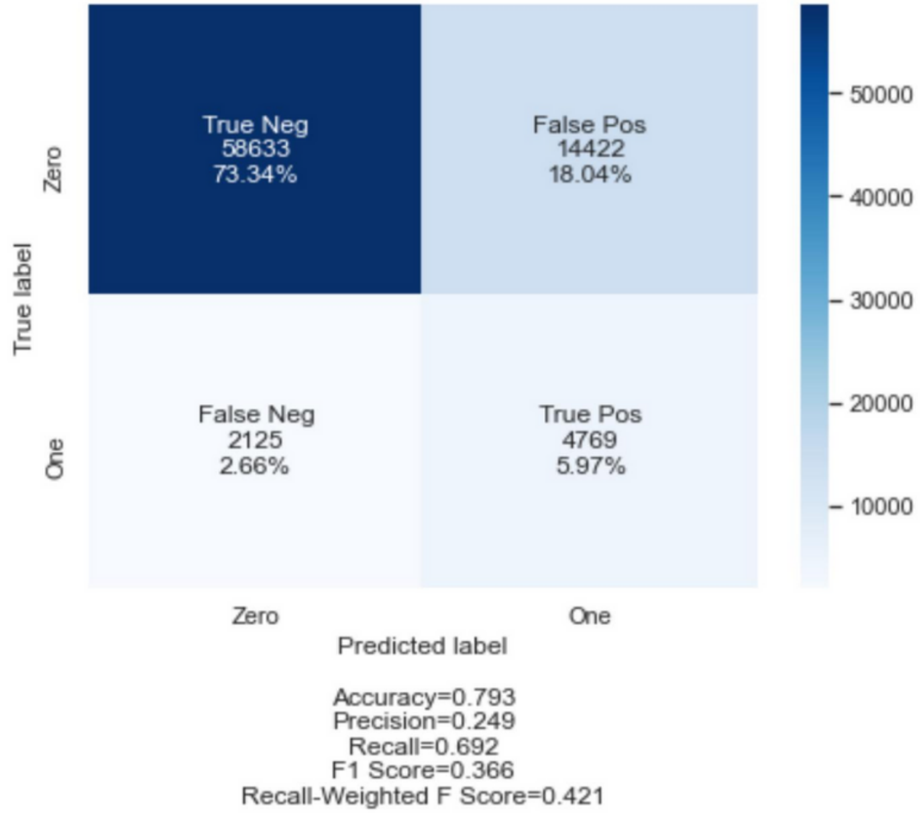
Model	Optimization	RWF Score
Random Forest	RandomizedSearch	*.421*
Random Forest	GridSearch	*.413*
Extra Trees	RandomizedSearch	*.397*
Logistic Regression	RandomizedSearch	*.394*
XBoost	GridSearch	*.379*

FINAL MODEL: **CONCLUSION**



FINAL MODEL

The Final Model is a **Random Forest model**, optimized to achieve the highest recall-weighted F score.



THANK YOU

TRAVIS ROYCE

- TravisCRoyce@gmail.com
- Github.com/tmcroyce
- https://www.linkedin.com/in/travis-royce/