HEART DISEASE:
MACHINE LEARNING CLASSIFICATION

TRAVIS ROYCE

JUNE 2022

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.

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 APPLICATION

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.

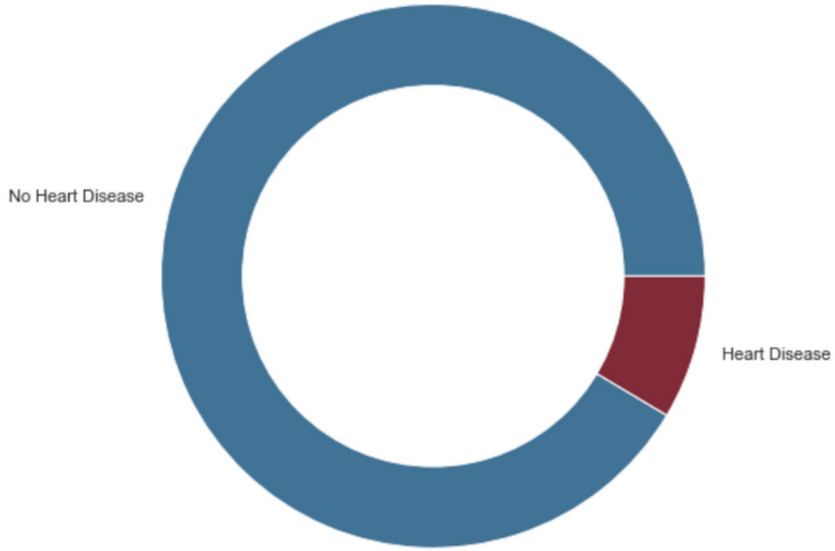
- **False negatives** could cause patients *with* heart disease to forgo further testing. This would be the worst possibility, out of the options.

- **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.

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. This was achieved using the F-beta method.

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**

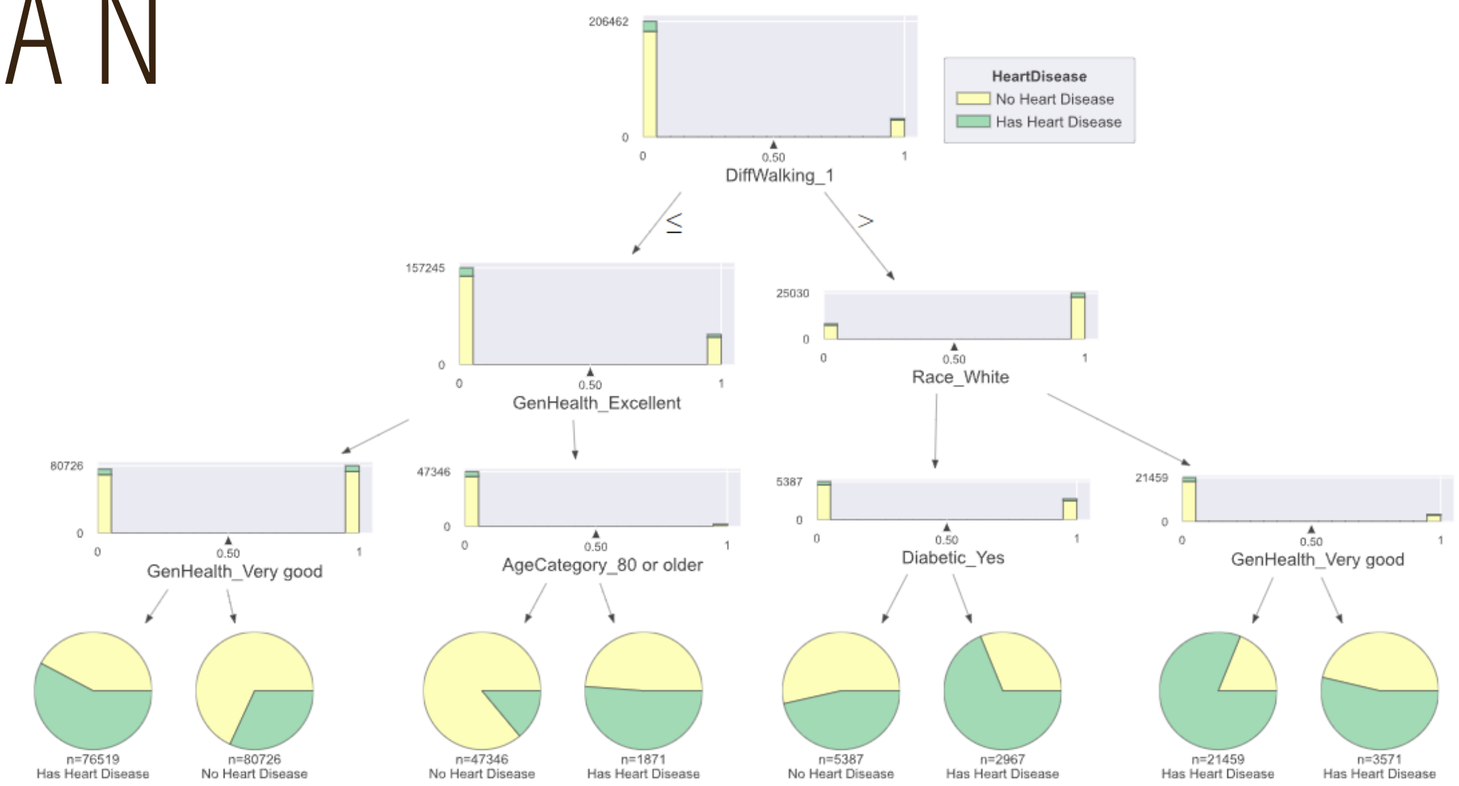
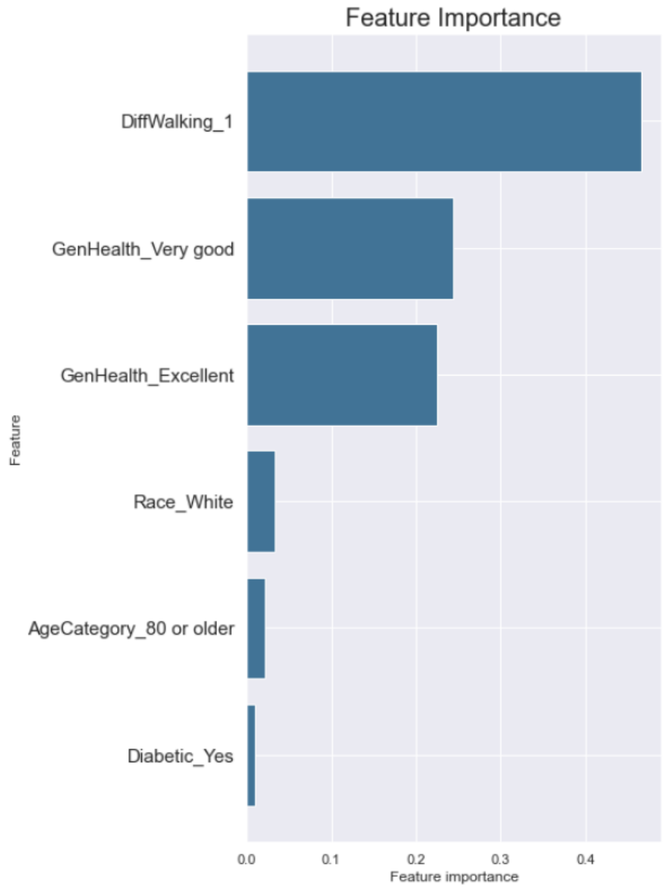


Illustration of Initial Decision Tree Model with a Maximum Depth of three and a recall-weighted F1 score of 0.265

INITIAL MODEL: DECISION TREE

MOST IMPORTANT FEATURES

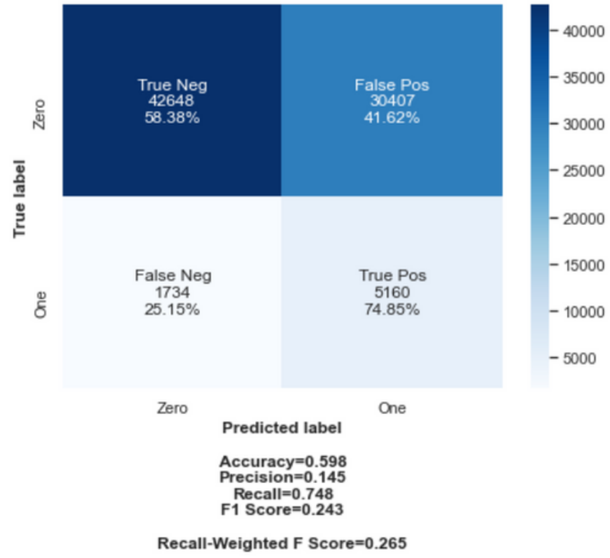
- **Difficulty walking** was the most important feature.
- **General health**, if very good or excellent, was second.
- **Race** - specifically, if the individual was white, was third.
- The final **two** features were if the individual was **80 years old** or older, and if the individual was **diabetic**.



RESULTS

The **recall-weighted F score** was **.265**
While the model only predicted 1,700 false negatives , it predicted over **30,000 false positives** (or 41.6% of top row).

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



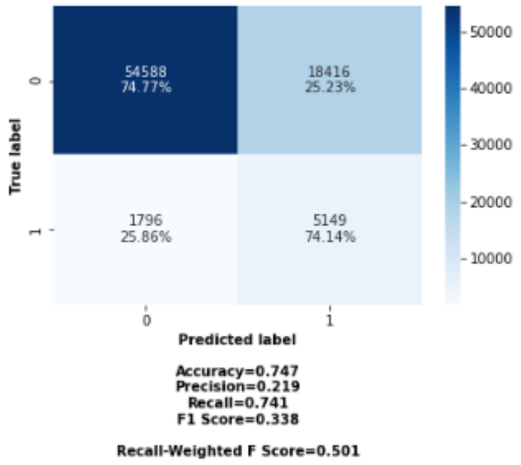
VANILLA MODELS: **TESTING RESULTS**

Model	rwF Score
Logistic Regression	*.502*
Decision Tree	*.441*
XGBoost	*.371*
Random Forest	*.305*
Extra Trees	*.3003*
Bagged Trees	*.256*

TOP MODEL: **ITERATION RESULT**

FINAL MODEL

The **Final Model** is a **Logistic Regression Model**, optimized to achieve the highest recall-weighted F score.



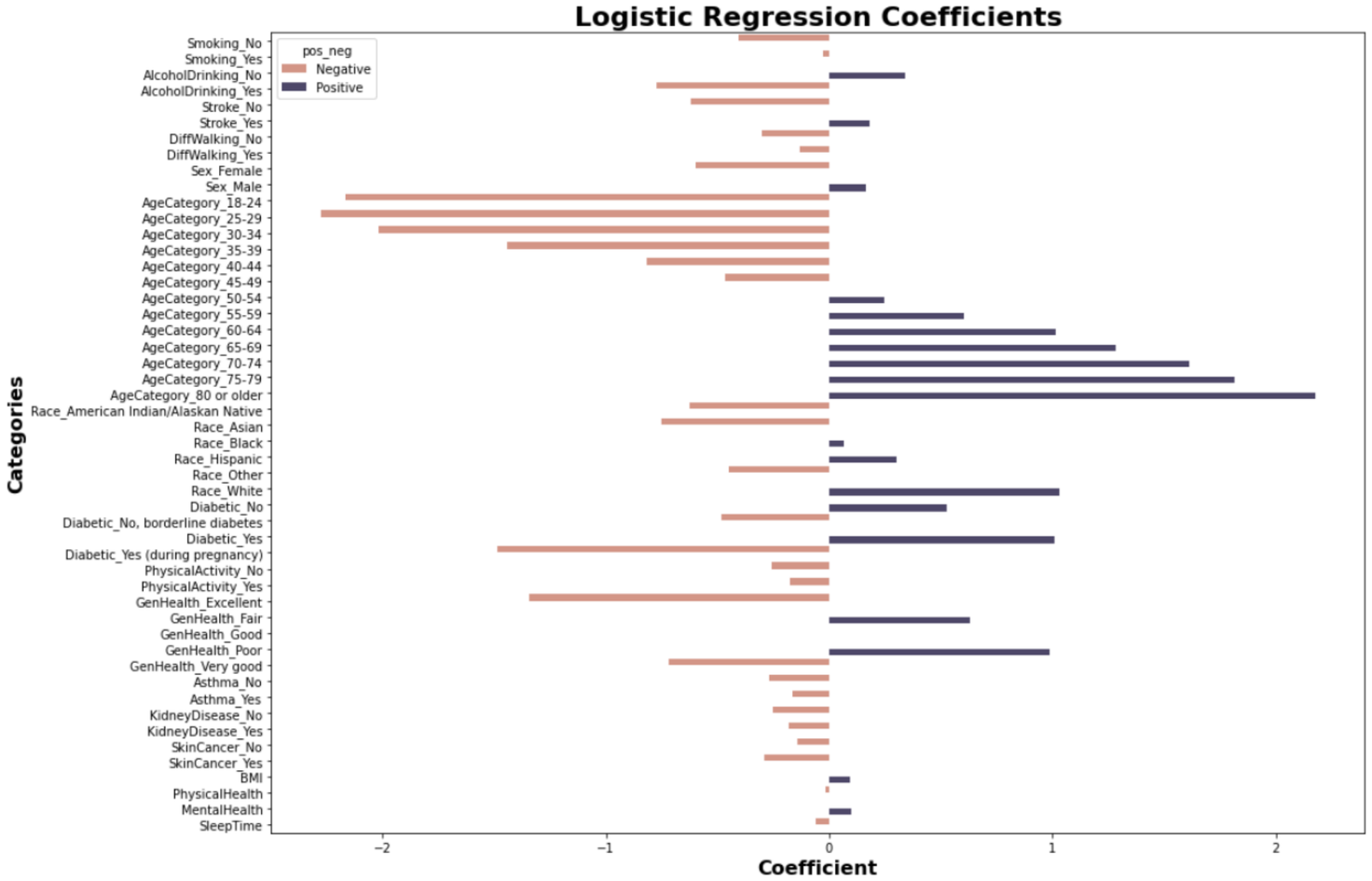
MODEL ANALYSIS

The **initial** model achieved a recall-weighted F score (rwF Score) of **.265**. The **final** model achieved a rwF Score of **.426**, an improvement of **over 60%**.

Other metric improvements are as follows:

	Initial Model	**Final Model**
Accuracy	**.60**	***.75***
Precision	**.15**	***.22***
Recall	**.75**	***.74***
F1 Score	**.24**	***.34***

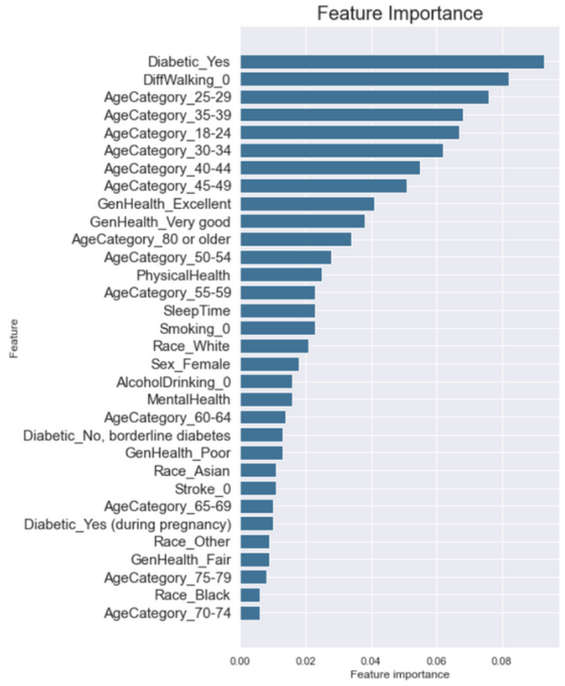
FINAL MODEL: FEATURES

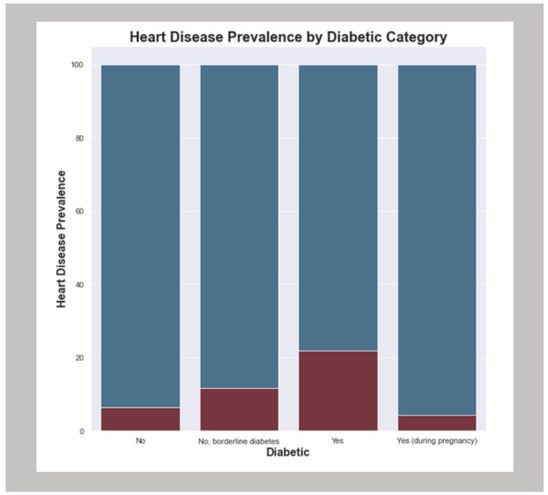


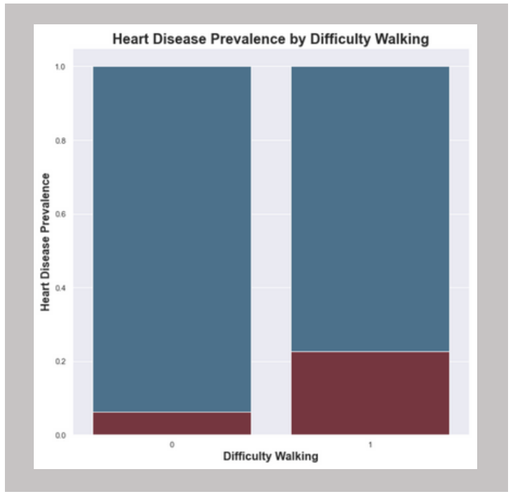
FINAL MODEL: **FEATURE IMPORTANCE**

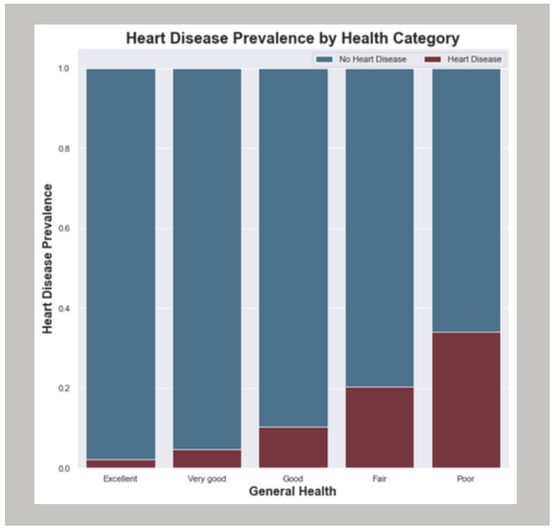
The **primary features** of the final model are:

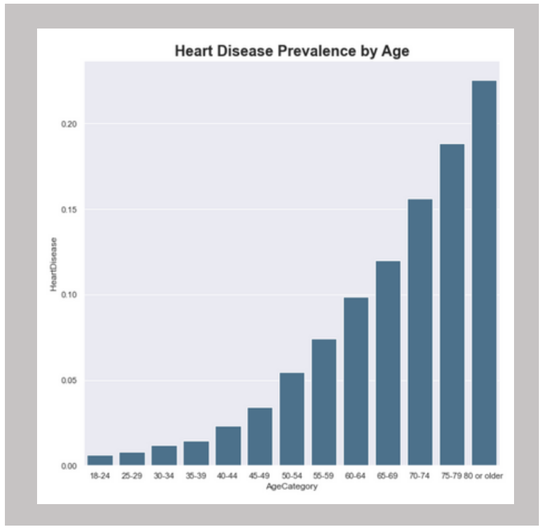
- Presence of **diabetes**
- Difficulty **walking**
- **Age** Category
- General **Health**











PRODUCT **CONCLUSION**

FINAL PRODUCT

- The **final product** can predict if an individual has heart disease based on the answers to a few simple questions with an **86 percent accuracy**.

- The algorithm is optimized to **penalize false negatives** more than false positives, because the goal is to **maximize** finding sick individuals.

- As mentioned, this product can be in the form of a **web application**, **phone application**, or **both**. Further, it can be used to inform **individuals and/or their doctors** about their heart disease risks.

THANK YOU

TRAVIS ROYCE

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