Heart Disease Final Report:
Using Prediction for Prevention

Table of Contents

[Background]

[Problem Identification]

[Data Wrangling and Preprocessing]

[Model Description]

[Model Performance]

[Model Findings]

[Next Steps]

Background

"Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year…Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths" (WHO, 2022). Great Lakes Hospital is starting a new campaign to prevent heart disease by identifying those with the greatest risk and intervening before a CVD occurs.

Problem Identification

The Question: What can be developed that informs Great Lakes Hospital whether a person will have a cardiovascular disease based on their characteristics, habits, or other factors?

The Audience: The results will be presented to the Marketing and Analytics team for use and distribution to the medical teams and the surrounding public.

The Data: Dataset from the CDC taken from a nationwide telephone survey in 2020 about US residents' health status

The Time Frame: August 2022-January 2023

The Modeling Response: heart disease = 1, no heart disease = 0

The Model: Binary supervised classification using Naive Bayes

The Deliverables: Jupyter notebooks of data wrangling through modeling, as well as this final report and presentation

Data Wrangling and Preprocessing

The data included 401,958 rows and 279 columns, but was reduced to 18. The data can be found on Kaggle.

The columns were as follows:

- HeartDisease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
- BMI: Body Mass Index
- Smoking: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
- AlcoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
- Stroke: Have you ever had a stroke?
- PhysicalHealth: Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 was your physical health not good?
- MentalHealth: Thinking about your mental health, for how many days during the past 30 days was your mental health not good?
- DiffWalking: Do you have serious difficulty walking or climbing stairs?
- Sex: Male or female
- AgeCategory: Fourteen age categories
- Race: White, Hispanic, Black, Asian, American Indian/Alaskan Native, Other
- Diabetic: Do you have diabetes?
- PhysicalActivity: Are you physically active?
- GenHealth: Excellent, Very good, Good, Fair, Poor
- SleepTime: On average, how much do you sleep in 24 hours?
- Asthma: Do you have asthma?
- KidneyDisease: Have you ever had kidney disease?
- SkinCancer: Have you ever had skin cancer?

I began the data wrangling process by checking for missing values, which there were none. Then I checked the data type for each column. I converted all binary (Smoking, Asthma, etc) columns to categorical, as well as mapped Yes=1 and No=0. I checked the summary statistics of the numerical columns (BMI, PhysicalHealth, MentalHealth, and SleepTime). I noticed a couple irregularities: a BMI of 94.85 and sleep time of 24 hours.

Then I spent time understanding each column by looking at the value counts of each category and creating histograms, box plots, and bar graphs where appropriate.

I created a new column SleepGroups by grouping sleep time together (1-6, 7-12, 13-18, 19-24 hours). I did the same for BMIGroups based on the CDC labels: 'Severe Thinness', 'Thinness', 'Normal', 'Overweight', 'Obese', 'Severely Obese', 'Morbidly Obese'.

During Exploratory Data Analysis (EDA), I found that people in this dataset with Heart Disease appear to have a higher rate of stroke and diabetes, to be older, to smoke more, to be more often male, to have more bad physical health days a month, to have a higher rate of poor or fair health, to drink less alcohol, to have more difficulty walking, and to have a higher BMI. They also appear to have higher rates of skin cancer, kidney disease, and asthma. Asians have the least amount of Heart Disease at 3%.

For preprocessing, I created indicator features for the categorical variables by using pandas get_dummies, and I removed the original columns. Y was whether a person had heart disease, and X was all the remaining features. Then I split the data into an 80/20 train/test. Lastly, I used StandardScaler to standardize the magnitudes of the numeric features (BMI, PhysicalHealth, MentalHealth, and SleepTime).

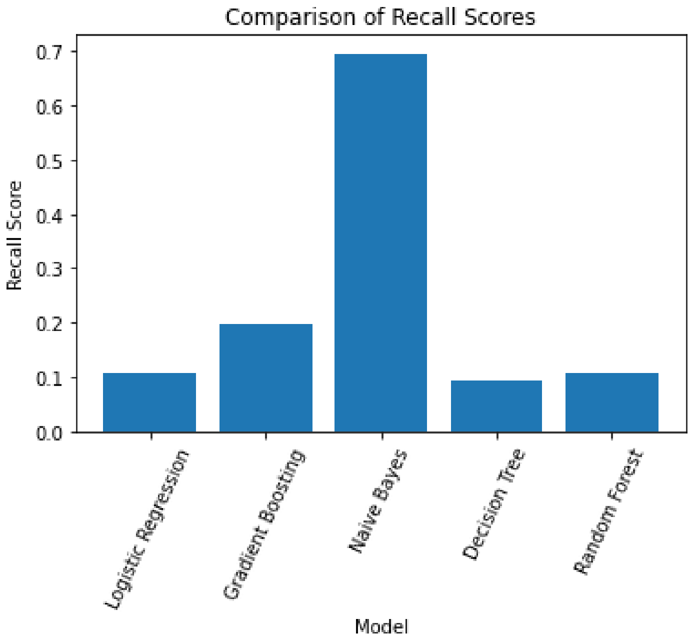
Model Description

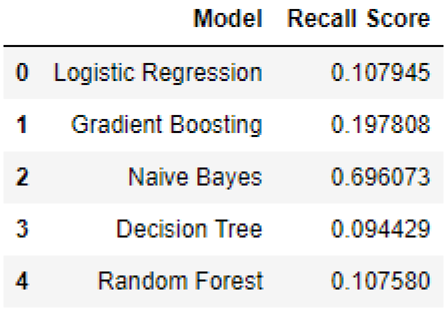
I trained five models: Logistic Regression, Gradient Boosting, Naive Bayes, Decision Tree, and Random Forest. I tried KNN and SVM, but they did not work on this dataset.

For Logistic Regression and Naive Bayes, I used the default parameters. For Gradient Boosting I used 100 estimators, 1.0 learning rate, and a max depth of 10. For the Decision Tree, I used a max depth of 10. For the Random Forest, I used 300 estimators.

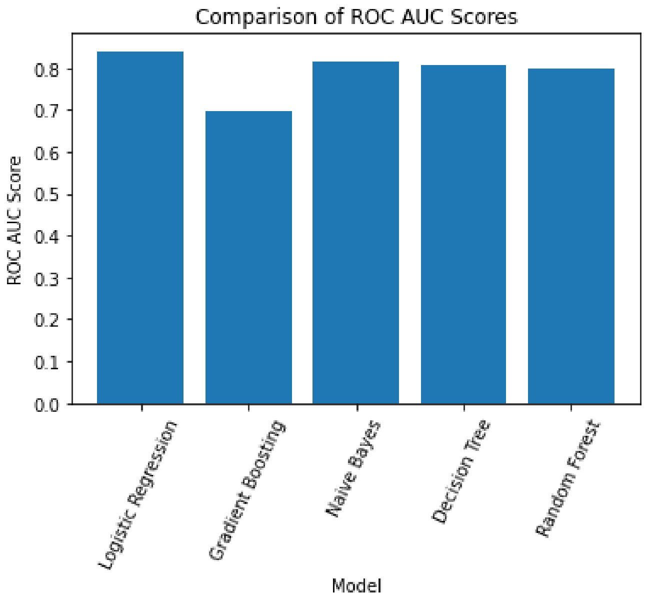
Model Performance

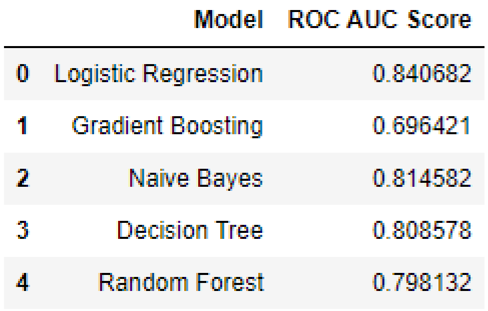
I cared most about the Recall score because we want to minimize the number of false negatives. In other words, I did not want to say a person is healthy (no heart disease) when they actually have a high chance of having heart disease. I also checked the ROC AUC score and graphed the ROC Curve for each model.



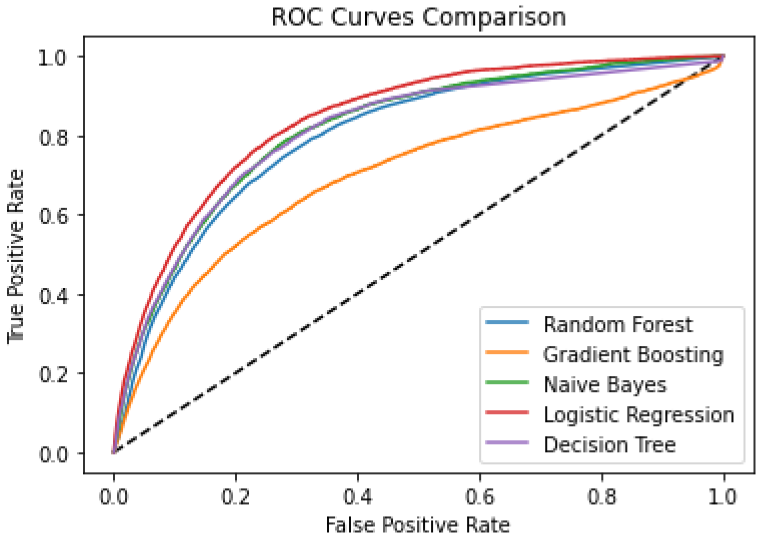


	Model	Recall Score
0	Logistic Regression	0.107945
1	Gradient Boosting	0.197808
2	Naive Bayes	0.696073
3	Decision Tree	0.094429
4	Random Forest	0.107580

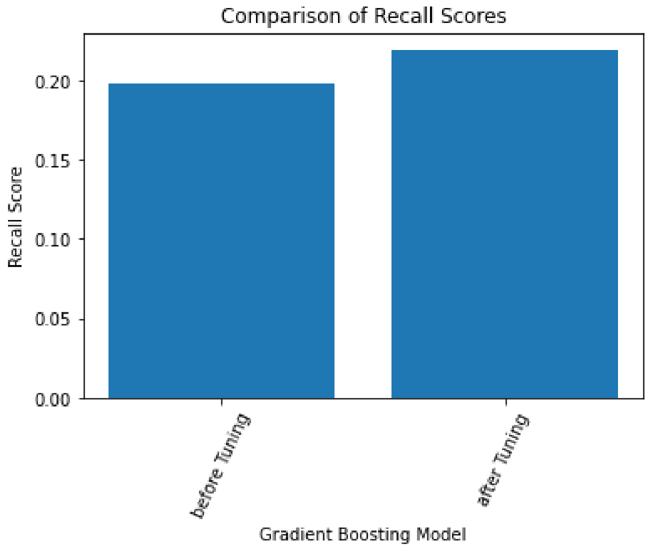


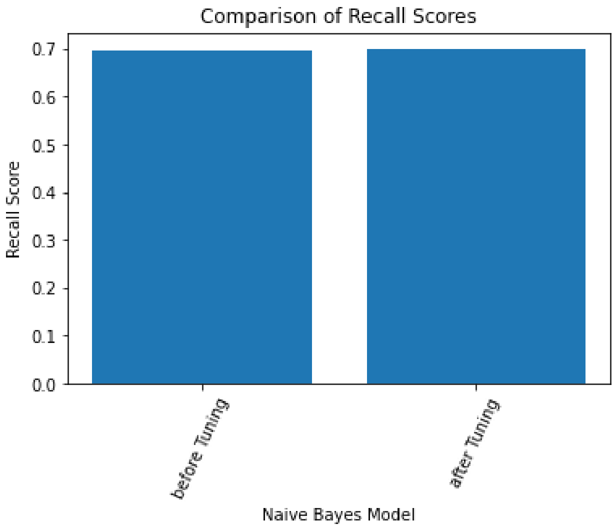


	Model	ROC AUC Score
0	Logistic Regression	0.840682
1	Gradient Boosting	0.696421
2	Naive Bayes	0.814582
3	Decision Tree	0.808578
4	Random Forest	0.798132

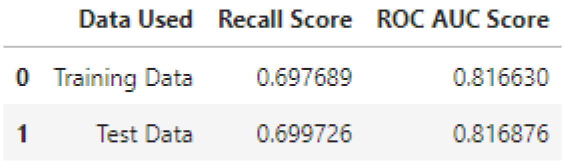


Because Gradient Boosting and Naive Bayes had the best recall scores, I tuned their hyperparameters to further improve the results using RandomizedSearchCV. Naive Bayes best parameter was 0.0029 for var_smoothing. Gradient Boosting's best parameters were n_estimators: 100, min_samples_leaf: 4, max_depth: 12, and learning_rate: 1. After tuning, the recall score for Gradient Boosting slightly increased from 19.8% to 21.9% while Naive Bayes only increased by .3%.





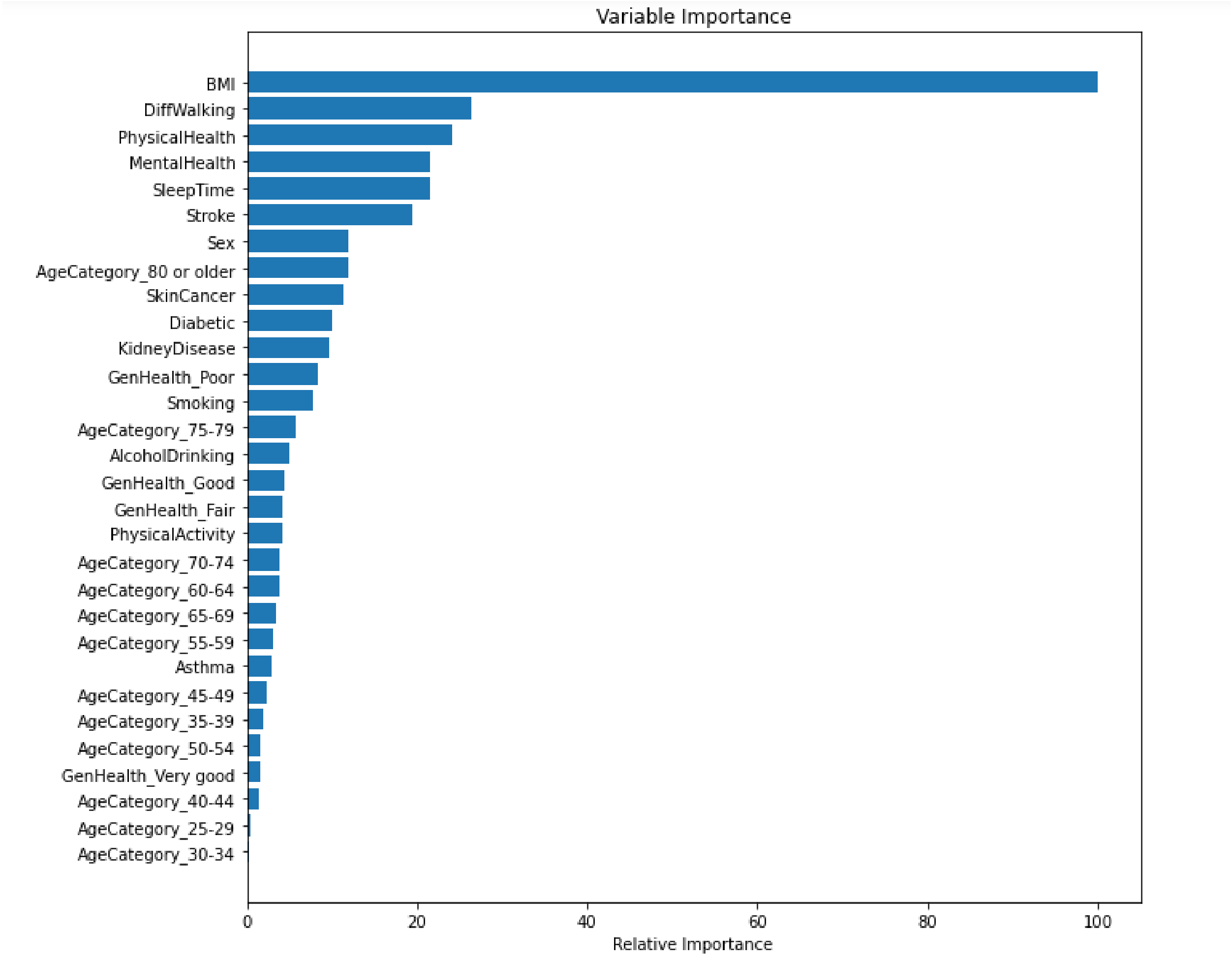
To ensure Naive Bayes was not overfitted, I compared the training recall to the test recall, as well as the roc auc score, and found very little differences.



	Data Used	Recall Score	ROC AUC Score
0	Training Data	0.697689	0.816630
1	Test Data	0.699726	0.816876

Model Findings

The Gradient Boosting showed the relative importance of the features, and it gave encouraging results. BMI was of the highest importance, which is great news because patients can change their BMI (which is not true of age, race, or sex). They can also take actionable steps to improve their walking, mental health, and sleep time.



Next Steps

I would like to have data from the surrounding community, instead of national data, to ensure there is nothing unique about the climate, culture, and geography around Great Lakes Hospital.

More unbiased data would be helpful as well. The current data includes many questions where respondents had to estimate: for example, how many days out of last month did you feel unwell? Data that can be measured, such as blood pressure, cholesterol, and pulse, could be informative.

To use the model to its fullest, the model will need to be integrated into the hospital software where medical practitioners will be able to input patient data to determine the chance of heart disease. A rollout to the community would be a future goal as well, where the application also gives specific personalized suggestions to review with their doctor to decrease their chances of heart disease.