Course Project 1: Milestone 3 – Draft White Paper

College of Science and Technology, Bellevue University

DSC680-T301: Applied Data Science (2237-1)

Alyssa Weber

July 30, 2023

Business Problem

Healthcare providers are teaming up to determine if a new panel of test results can help predict heart disease. My clients want to know if the variables, or a combination of the variables, can be used to create an accurate prediction model for heart disease.

Background/History

Heart disease is a broad term that describes a spectrum of conditions affecting the heart including blood vessel disease, arrhythmias, congenital heart defects, heart muscle diseases, and heart valve diseases (Mayo Clinic Staff, 2022). Heart disease has a range of causes commonly consisting of birth defects, heart valve problems, heart muscle complications, and arrhythmias. Since the causes have such great variance, symptoms have wide variety as well, making diagnosis of heart disease challenging. "Sometimes heart disease may be "silent" and not diagnosed until a person experiences signs or symptoms of a heart attack, heart failure, or an arrhythmia" (CDC, 2023a).

Heart Disease is the leading cause of death in the United States and killed about 695,000 people in 2021 (CDC, 2023b). A 2020 study published in Circulation called Cardiovascular Quality and Outcomes, studied patients with Type 2 heart attacks. The study found that 3 out of 5 of the subjects had undiagnosed signs of coronary artery disease (Williamson, 2023). Many of these patients were only diagnosed with heart disease after the event of a heart attack which prompted advanced heart imaging.

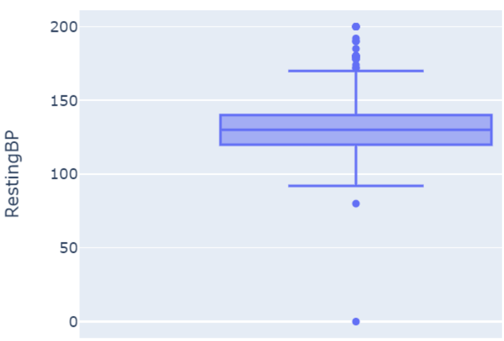
The current process for diagnosing heart disease typically begins with collecting personal and family medical history. Doctors can run laboratory tests including lipid profiles which include cholesterol levels and triglycerides. Blood tests also check for blood counts, sodium and potassium levels, kidney function, glucose, liver inflammation based on ALT and AST, and thyroid function based on TSH (UCSF Health, 2022). Testing of the heart can also be performed with the use of electrocardiogram, echocardiogram, or ultrasound.

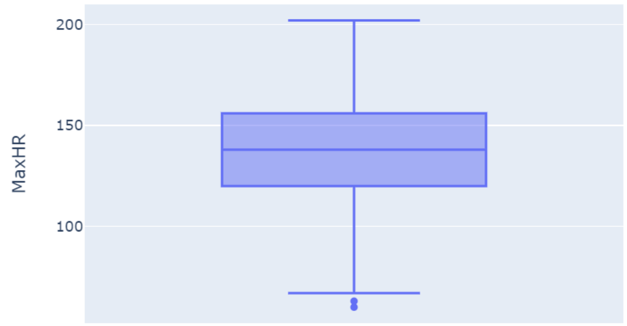
Data Explanation (Data Prep/Data Dictionary Etc)

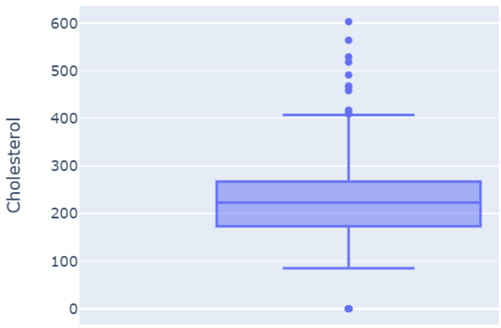
The data set used for the subsequent models is called “Heart Failure Prediction Dataset” and was found on the Kaggle website (Fedesoriano, 2021). The target variable included in the data is called HeartDisease and will be used for prediction. Each instance is labeled as either a 1 (heart disease) or 0 (no heart disease). Also included are 11 feature variables which includes:

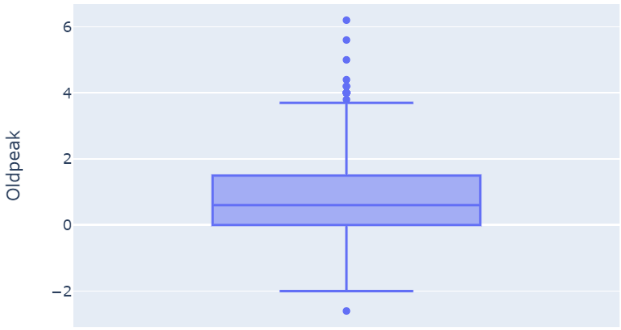
1. Age (range 28-77 years old)
2. Sex (M or F)
3. ChestPainType (ATA, NAP, SAY, TA)
4. RestingBP (0-200)
5. Cholesterol (0-603)
6. FastingBS (0 or 1)
7. RestingECG (Normal, ST, or LVH)
8. MaxHR (60-202)
9. ExerciseAngina (N or Y)
10. OldPeak (-2.6 – 6.2)
11. STSlope (Up, Flat, or Down)

To prepare the data, I began by checking for missing data, in which there was none. Next, I examined the boxplots of the numerical data, that were also not binary, to check for outliers. Age was the only variable in which no outliers were present. RestingBP and Cholesterol included at least one value of zero. When checking these instances, it was found that there was one instance in which a zero was recorded for both. This patient was removed. There were 171 other Cholesterol entries of zero. For now, those values will not be adjusted due to the high number of zero entries. Cholesterol also had many high outliers, which should be kept for the heart disease model. MaxHR had two lower outliers that will be kept as is. Oldpeak had a range of values that will also not be adjusted.









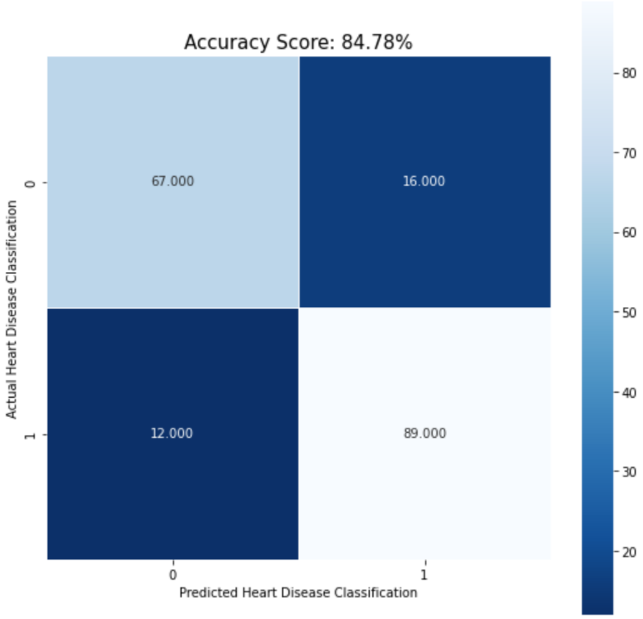
Methods

Since the outcomes are binary (0 = no heart disease, 1= heart disease) and the adjusted data set has 917 entries I plan on using a simple model. I will explore both logistic regressions and decision trees to present the best model to my clients. To measure my results, I will calculate the accuracy of both models while taking false positives and negatives into effect. Accuracy will tell me how often the model can correctly predict a patient’s heart disease status. False positives occur when the model predicts that a patient has heart disease, when in fact they do not. False negatives occur when the model predicts that a patient does not have heart disease, when in fact they do. Both errors are concerning. False positives may result in treatment that is unnecessary. False negatives will let heart disease continue to go undiagnosed which could lead to serious medical problems in the future. Ideally, the model will have a lower percentage of false negatives than false positives.

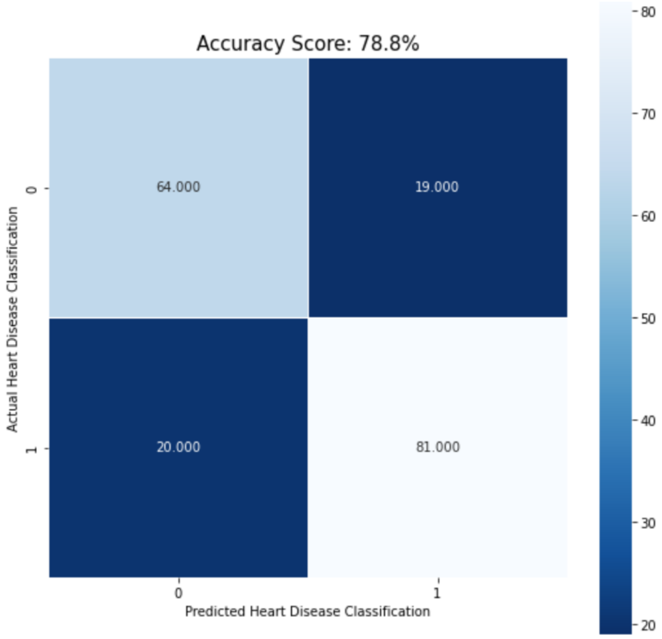
Analysis

Both the logistic model and descion tree model were run. The logisic model had an accuracy score of 84.78% compared to the decision tree model with an accuracy score of 78.8%. The logistic model resulted in 16 false poisitves and 12 false negatives. Both of these results were lower than when using the decision tree.

Logistic Recgression Confusion Matrix:



Decision Tree Confusion Matrix:



Conclusion

There is potential with the panel of new test results to predict heart disease with high accuracy. Based on my analysis, I would present a logistic regression model to my clients. The model is ready for deployment assuming that monitoring and adjusting would take place. It is also recommended that patients predicted to have heart disease have additional testing/imaging to determine the severity of their health risk.

Assumptions, Limitations, and Challenges

The dataset was built using observations from 5 different regions. Further monitoring will need to be completed when generalizing the model to the public. I am also unsure how difficult and expensive it will be to get the panel of feature variables used for the model. It will take much more time to explore the accuracies for the model if variables are unavailable from the original panel. Medical professionals will want to determine if it cost efficient and ethical compared to modern techniques of diagnosis.

Ethical Assessment

False positives and negatives, which are inevitable, are undoubtedly harmful. However, being able to predict heart disease at a level of 85% will hopefully present more lifesaving results than harm in the medical field. After the model is built, it will need continual testing when it is used with new patients.

Data collected always needs to be scrutinized before it can be assumed to generalize to the public. Because the data was collected from a variety of geographic regions, I am confident that there was no intended bias in the initial phase of modeling. Additional testing and observation will need to take place to make sure that the launched model is adapting to many different regions and populations that may not have been included in the original sampling.

Lastly, once results have been constructed, will the results be used to target people who may not know they have heart disease? Most likely, patients will be aware that they are screening for heart disease based on the panel of information being collected. However, the medical professionals will want to explore the implications of reaching out to people who are not looking for heart disease information.

References

CDC. (2023a, May 15). *About heart disease*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/about.htm

CDC. (2023b, May 15). *Heart disease facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

Fedesoriano. (2021, September 10). *Heart failure prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Mayo Clinic Staff. (2022, August 25). *Heart disease*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

UCSF Health. (2022, June 24). *Diagnosing heart disease*. ucsfhealth.org. https://www.ucsfhealth.org/education/diagnosing-heart-disease

Williamson, L. (2023, January 24). *Undiagnosed heart disease may be common in people with heart attacks not caused by clots*. www.heart.org. https://www.heart.org/en/news/2022/03/28/undiagnosed-heart-disease-may-be-common-in-people-with-heart-attacks-not-caused-by-clots