Final Report
MIS 64060
Asher John

What ails the heart?

Introduction

Cardiovascular diseases aka heart diseases are the single largest cause of death in the world. According to the World Health Organization 17.9 million people die of heart diseases every year. This shocking number of deaths represent 32% of all global deaths[i]. According to the Center for Disease Control and Prevention (CDC) almost 700,000 people in the US die of heart diseases - this is almost 25% of all deaths in the US ever year. Heart Diseases cost US a staggering $363 billion each year[ii]. The enormity of this amount can be discerned from the fact that only 40 countries in the world have a higher annual GDP than this. Heart diseases do not discriminate, people from all races, religions, and socioeconomic backgrounds suffer from them. High blood pressure, high blood cholesterol, and smoking are the leading causes of heart diseases. Angina, diabetes, and physical inactivity are other factors that cause and lead to cardiovascular diseases. A good prediction model for heart diseases can save billions of dollars and thousands of lives. This project attempts to determine a robust prediction model for heart disease based on Heart Failure dataset from Kaggle[iii]. An exploratory analysis, logistic regression, and naïve bayes analysis reveal on the data that age, sex, chest pain type, cholesterol, fasting blood sugar, exercise angina, and old peak are good predictors of heart disease.

The Problem

The project strives to find the best predictive model for heart disease from the heart failure dataset. The question addressed is; whether it is possible to come up with an accurate and good predictive model for heart disease based on all or a certain subset of the predictor variables from the Heart failure dataset?

The Data

Data used for this project was taken from Kaggle. The data consists of 918 observations with 12 attributes. The 12 attributes consist of age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG, maximum heart rate, exercise angina, old peak = ST (depression induced by exercise relative to rest), ST slope, and the target variable of heart disease. The data was selected because it is well organized and clean and comes from a reliable source. The data is representative as it contains both male and female and people with and without heart disease. An exploratory analysis, to identify any quality issues, was performed on the data. There were no gaps, missing values, or discrepancies found in the data, and it did not need any cleaning. The only changes made to the data were changing the categorical variables into factor variables.

Approach

To build a predictive model and identify a relationship between heart disease (response variable) and predictor variables (age, sex, etc.) logistic regression analysis was performed on the dataset in R. Logistic regression has been used because as it is not only easy to use and train but it also can deal with both categorical and numerical variables. In cases where the response variable is categorical as in our data it is advised and recommended to use logistic regression. Logistic regression can handle both categorical and numerical variables in the independent category of variables. It can perform both simple and multiple regression analysis. It not only provides the strength of correlation between response and predictor variables but also the direction of this association.

To find the best predictive model for heart disease a logistic regression analysis was run on the data. The data were divided into training and testing sets. The first analysis included all the predictor variables in the training set. A backward stepwise approach was adopted to find the best fit model. After running a few iterations of the model, it was decided to keep 8 predictor variables (age, sex, chest pain type, cholesterol, fasting blood sugar, exercise angina, and old peak) in the final model. The fit of the model was estimated based on the residual deviance and AIC from the R output of the logistic regression analysis. The final model has the smallest possible AIC value and the residual deviance. To check and cross validate the prediction power of the model it was applied to the test data. Table 1 shows the confusion matrix for the logistic regression prediction model. The results show that model is a good predictor of heart disease with 86% accuracy rate. After validating the model, the best subset of predictor variables was applied to entire dataset for estimating the final regression model.

Table 1. Heart Disease Prediction from logistic regression

	Predicted Value	
Actual Value	No Heart Disease	Heart Disease
No Heart Disease	132	32
Heart Disease	18	185

Naïve bayes was the second predictive algorithm used to verify and validate the predictions gained form logistic regression. It was used because it is good at solving multi class prediction issues. It is good for predicting the conditional probability of an event if the assumption of independence holds true. It is also suited for both continuous and categorical predictor variables. Naïve bayes was also conducted on the training data and then the results from the training data were used to predict the heart disease in the test data. The goodness of the model and its predictability is also reflected by the naïve bayes prediction model results. Confusion matrix of the naïve bayes prediction model is given in table 2. The results show an 84.7% accuracy rate. This result is not only good, but it also validates and verifies the results gained from our chosen model that was based on logistic regression analysis.

Table 2: Heart Disease Prediction from naïve bayes

	Predicted Value	
Actual Value	No Heart Disease	Heart Disease
No Heart Disease	129	35
Heart Disease	21	182

Results of the logistic regression analysis were transformed into their exponential form for easier understanding and description (for detailed results please see the html output of the analysis, link is given at the end).

Analysis

The results of the logistic regression analysis show a strong association between predictor variables (age, sex, chest pain type, cholesterol, fasting blood sugar, exercise angina, and old peak) and response variable (heart disease). One important caveat to keep in mid while reading and interpreting these results is that whenever the odds of having heart disease are predicted it is assumed that all other predictor variables are held constant. The accuracy of the prediction model can be observed from the predicted probability plot in figure 1. The plot shows that most of the heart disease patients are shown in high probability of having heart disease. Higher probability of heart disease is shown in the right upper corner of the plot and patients with heart disease are shown in green and the top of the S curve is very green.

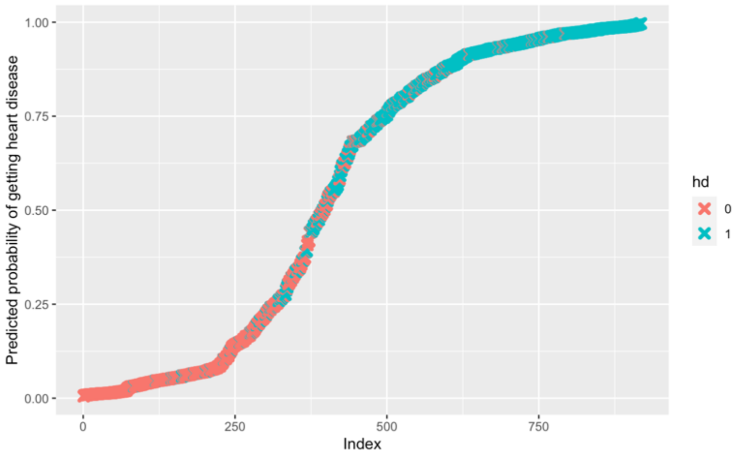

Figure 1. Prediction probability plot of getting heart disease

An important predictor of heart disease is one's biological sex and it plays an important role in contracting heart disease. Being male increases the odds of getting a heart disease by 4 times as compared to females if all other variables are held constant. This can be observed from figure 1. Although males are almost 80% of the data when it comes to heart disease males represent almost 92% of the cases.

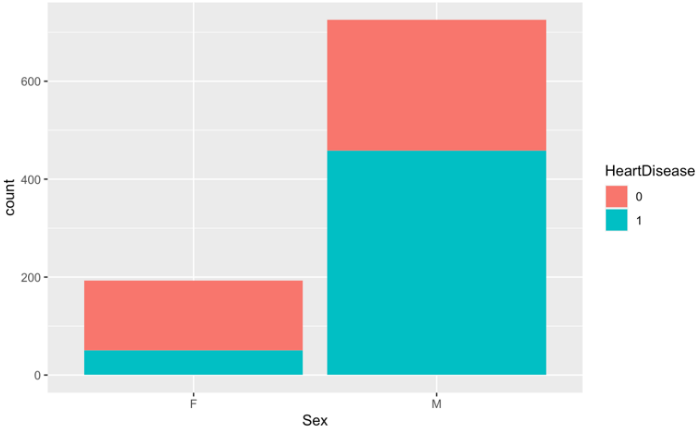

Figure 1: Sex and Heart Disease

A flat ST slope is one of the strongest predictors of heart disease form this data. Having a flat ST slope increases one's odds of having heart disease by more than 4 times as compared to having up or down ST slope. This is evident from the data as well, 76% of people who had flat ST slope became heart patients (see figure 2).

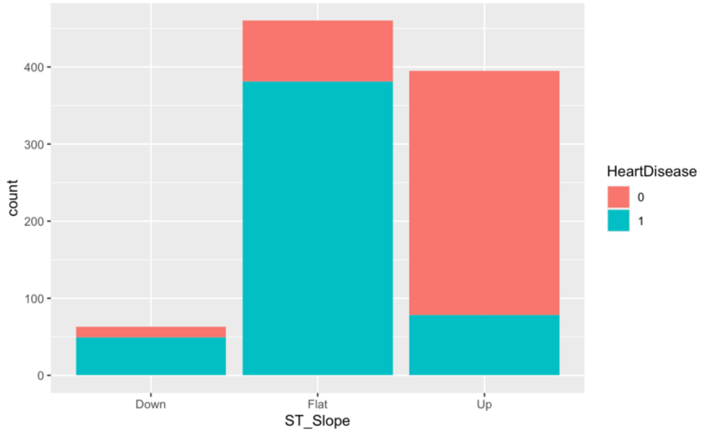

Figure 2: ST Slope and heart Disease

Having an episode of angina while exercising is also a strong indicator of heart disease as it increases the odds by more than 2.5 times. The data shows that almost 83% of people who complained of exercise angina developed a heart disease (see figure 3).

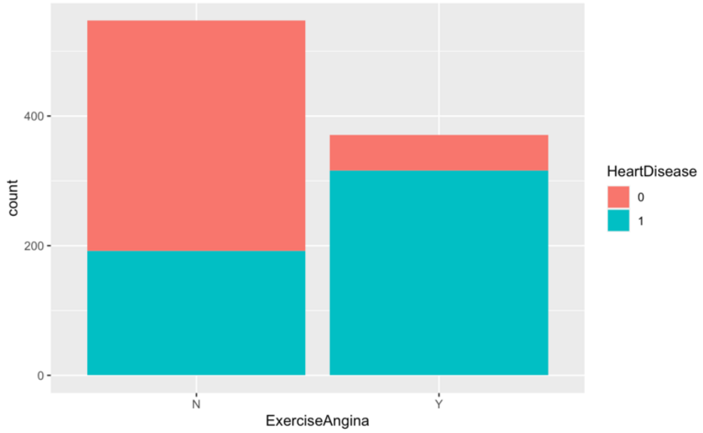

Figure 3: Exercise Angina and heart Disease

Another strong predictor of heart disease is chest pain type. Chest pain type ASY is strong predicter of heart disease and anyone suffering from this kind of chest pain has a high probability of getting heart disease. Figure 4 shows that almost 80% of people who had chest pain type ASY were diagnosed with heart disease. The predicting power of chest pain type is apparent from the fact that 80% of chest pain type ASY make almost 50% of all the subjects in the data and 80% of all heart patients! It is pertinent to observe that other chest pain types are not that indicative of heart disease.

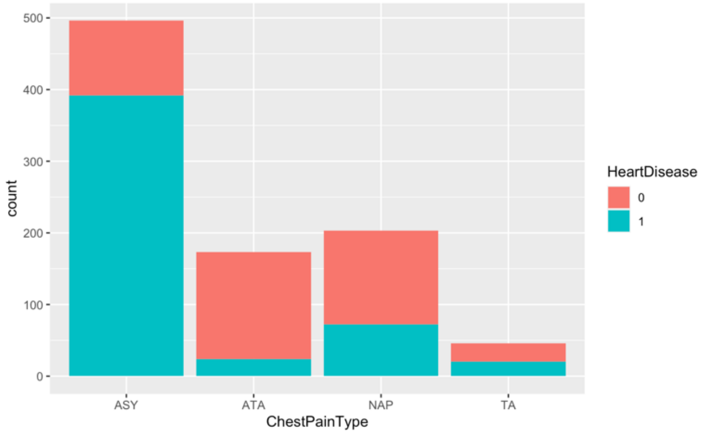

Figure 4: Chest Pain Type and Heart Disease

Conclusion

The model built in this analysis is a model of good fit and has good predictive power. The model built through logistic regression analysis had an accuracy of 86% in predicting the heart disease correctly. This was also verified by naïve bayes model and the accuracy in that model was 84%. The analysis suggests that there is a strong association between heart disease and the predictor variables that consist of age, sex, chest pain type, cholesterol, fasting blood sugar, exercise angina, and old peak. High level of fasting blood sugar is a strong predictor of having heart diseases, high blood sugar increases the odds of getting heart disease by more than 3 times as compared to not having high fasting blood sugar. Being male increases the odds of having heart disease by many times as compared to females. Chest pain type ASY is also a major predictor of the disease. ST slope (flat) increases the odds of having heart disease by 4 times as compared to other types of ST slopes. Another significant harbinger of cardiovascular diseases is exercise angina that increases the odds of having a heart disease by 2.5 times as compared to not having exercise angina.

References

Link for the R file and html output of the project:
https://github.com/asherjohn75/ajohn6_MIS64060.git

World Health organization. [i] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Center for Disease Control and Prevention. [ii] https://www.cdc.gov/heartdisease/facts.htm

Heart Failure prediction Dataset, Kaggle. [iii] https://www.kaggle.com/fedesoriano/heart-failure-prediction