Chronic Health Condition and Behavioral Risk Factors for Heart Disease

Anil Anderson
BIOS 611

December 5, 2022

Abstract: In this paper I analyze the Behavioral Risk Factor Surveillance System (BRFSS) data from 2015. The BRFSS interviews roughly 450,000 US adults by telephone each year, collecting data about health risk behaviors, chronic health conditions, and dietary, smoking, and exercise habits. The data also includes basic demographic information. Using logistic regression and gradient boosting models, I identify the main risk factors for heart disease, the leading cause of death in the United States. I also assess the predictive capabilities of the models using various accuracy metrics. I conclude that age, sex, high blood pressure, high blood cholesterol, and kidney disease are most strongly associated with heart disease, but I fail to show that healthier diet and exercise habits are associated with lower risk of heart disease.

Keywords: heart disease, logistic regression, prediction, telephone survey, behavioral risk factors

Introduction

Because heart disease is the leading cause of death in the United States, it is imporant to identify the factors that increase the risk of heart disease so that we may prevent outcomes like heart attack and death. According to the Centers for Disease Control and Prevention (CDC), the main risk factors for heart disease are health conditions like high blood pressure, high blood cholesterol levels, diabetes, and obesity. In addition, behaviors such as lack of physical activity, improper diet, excessive alcohol consumption, and tobacco use increase the risk of heart disease. In my analysis of the BRFSS data, I examine the strength of association between these variables and heart disease.

The dataset is quite large, containing several hundred variables and over 400,000 observations. I limit my analysis to roughly 30 variables and the complete cases in the dataset, and information about these variables can be found in the codebook linked on GitHub. I begin with exploratory data analysis, producing univariate and bivariate figures that help summarize information about the dataset. I then attempt somewhat unsuccessfully to perform dimensionality reduction to help visualize the data in two dimensions. Lastly, I create several predictive models that help quantify the association between heart disease and its potential risk factors. Given the large size of the dataset, I split into training and testing sets.

Preprocessing of the Dataset

Perhaps the most arduous and time-consuming step in the analysis of the dataset was the preprocessing. I first selected the variables I was most interested in–primarily variables containing information about chronic health conditions, diet and exercise habits, and demographics. Every variable in the data downloaded from Kaggle is encoded numerically, regardless of whether the variable is quantitative. I therefore manually decoded each variable by reading through the codebook and modifying the variables appropriately. Certain variables could be left as is, while others needed to be turned into factors. Some variables, for instance fruit and vegetable consumption, were encoded as 100 times the actual value (likely to avoid the use of decimal points), so I needed to divide those variables by 100 to obtain the correct values. Other variables such as strength training and cardio frequency were collected in two ways in the phone interviews. Certain people responded with a monthly frequency, whereas others responded with a weekly frequency. The codebook indicates how these different responses were recorded, and the file "setup.R" shows how I turned all of these frequencies into monthly counts. Though cleaned datasets exist on Kaggle, it was valuable for me to perform these preprocessing steps myself because they helped me understand the variables and ensure that I knew what the different levels of the variables corresponded to. With the variables appropriately transformed, I ensured that the subsequent interpretations of the regression coefficients made sense.

Exploratory Data Analysis

I first began analyzing the data by producing univariate and bivariate visualizations to help me understand the characteristics of the people in the dataset. Looking at the distribution of ages in the histogram below, we see that there are plenty of observations in all age ranges, with a median age of 58. Note that all ages above 80 are recorded as 80, which likely explains the relatively high count of people in this bin. We should also ask how representative of the population this dataset is, given that the distribution of ages is skewed left. In the general population, the distribution is likely flatter or even skewed right. Furthermore, according to the US Census Bureau, the median age in the US is roughly 40, meaning that the median in the dataset is higher than the US overall average. This makes sense, as older people who may be retired are more likely to have the time to respond to a telephone survey than younger and thus potentially busier people. I draw attention to this not because it is particularly relevant in the analysis of the data, but because it could affect how generalizable the results of the analysis are.

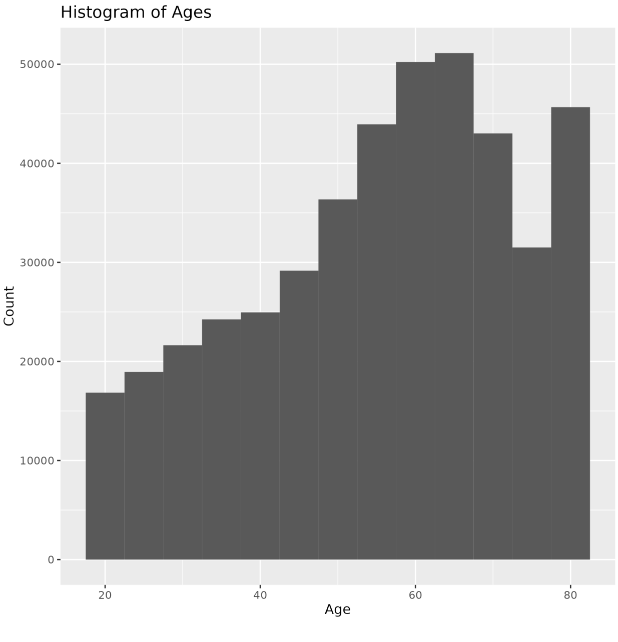

Figure 1: Histogram of ages in the dataset

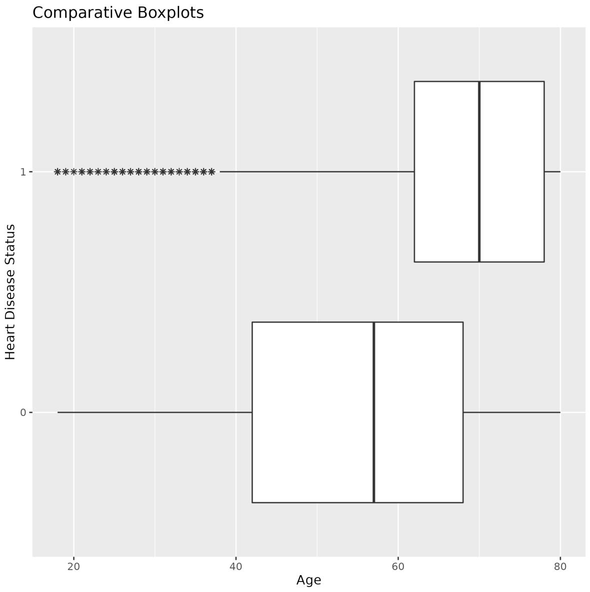

Figure 2: Boxplots of age distribution by heart disease diagnosis

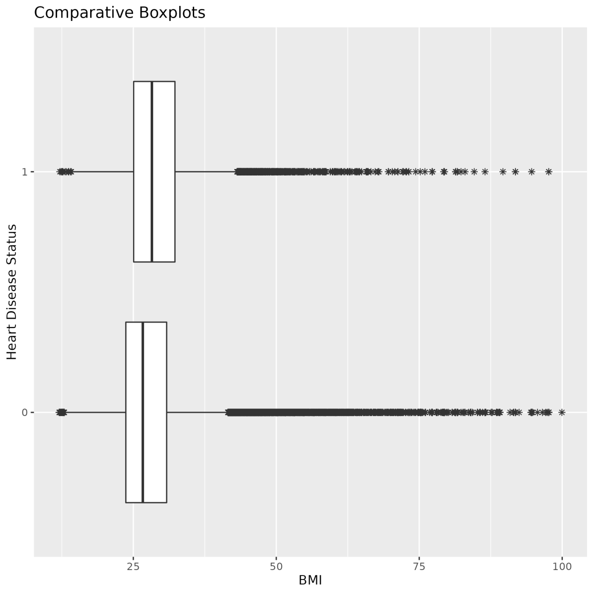

Figure 3: Boxplots of BMI distribution by heart disease diagnosis

I next began to examine the associations between heart disease and the other variables in the dataset. As seen in Figure 2, showing the side-by-side boxplots comparing age distribution by heart disease status, there is a clear positive association between age and probability of heart disease, as we would expect. As one becomes older, the risk of health complications naturally increases. This plot shows that age is an important variable we will want to include in our models. The side-by-side boxplot comparing BMI distribution by heart disease diagnosis shows there may be a positive association between BMI and probability of heart disease. This association may not be significant, however, since the distributions are somewhat similar. Fruit consumption is an example of a variable that does not appear to be significantly associated with heart disease.

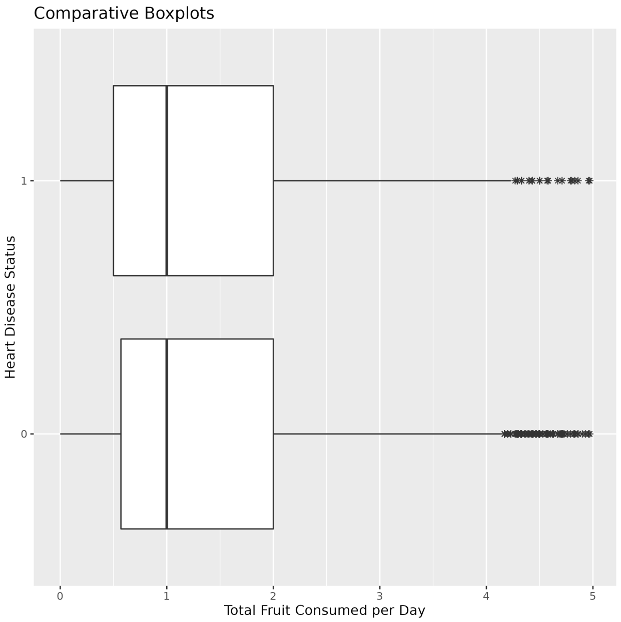

Figure 4: Boxplot comparing fruit consumed per day among those with and without heart disease diagnosis

So far, all the associations we have seen have been ones we would expect to see. One surprising relationship I found in the dataset was between smoking and heart disease. Though smoking most directly affects the lungs, we would expect more frequent smoking to be associated with higher risk of heart disease. This is not the case in our dataset, though, as seen in the Figure 5. The mosaic plot shows that former smokers are most likely to have heart disease, whereas people who smoke some days or every day are at roughly equal risk of heart disease as those who have never smoked.

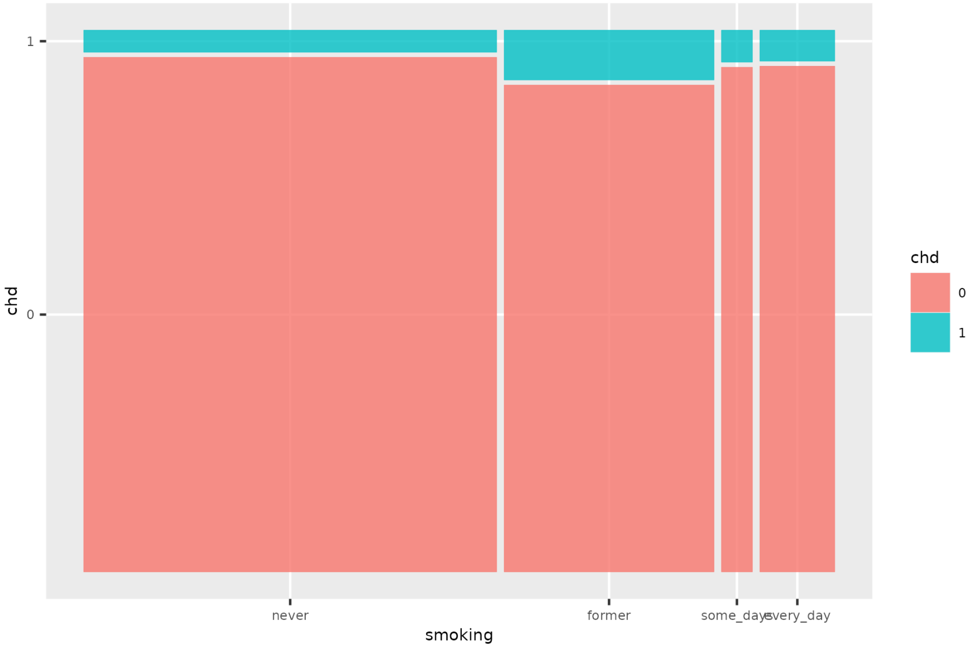

Figure 5: Mosaic plot showing association between smoking and CHD

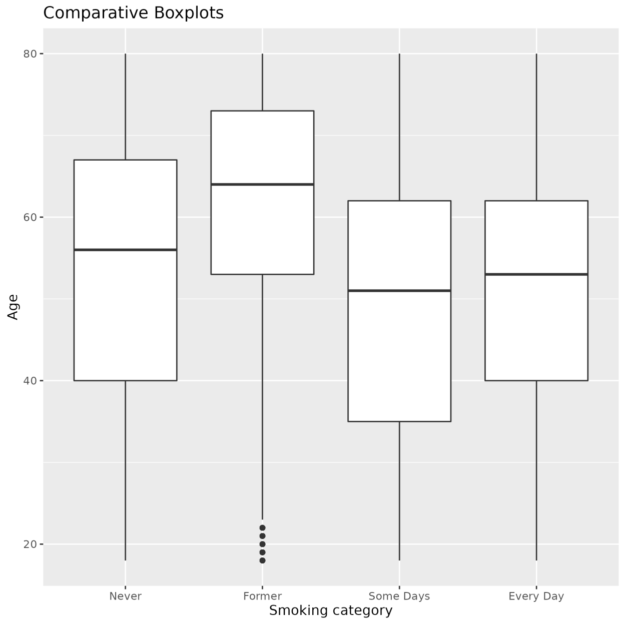

Figure 6: Comparative boxplots showing age distribution by smoking status

How could this be? The reason we see this unexpected relationship between smoking and heart disease is that age is acting as a confounding variable here. As seen in the comparative boxplots to the left, the smoking group with the highest average age is former smokers, while the other smoking groups have roughly the same distribution of ages. Because age is so strongly associated with heart disease, it makes sense that the risk of heart disease would be highest in the former smoking group and roughly the same in the other smoking groups.

The last bivariate plot I looked at was between heart disease and occurrence of heart attack. With heart disease known to put people at greater risk of heart attack, I wanted to examine the association between the two in this dataset. As seen in Figure 7, the proportion of people who have had a heart attack among those with heart disease is substantially higher than among those who do not have heart disease. This suggests that heart attack occurrence will be help distinguish between those with heart disease and those without in our models.

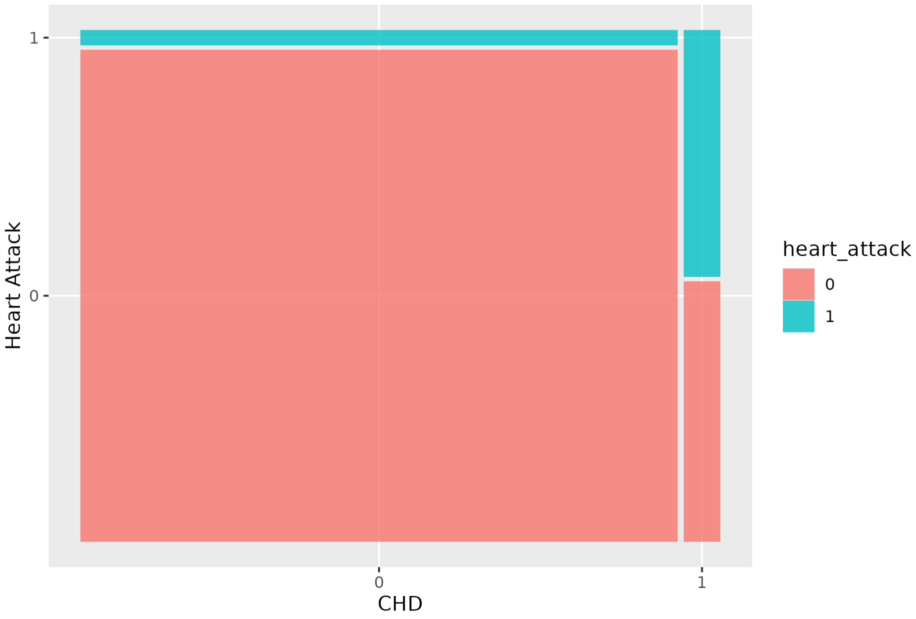

Figure 7: Mosaic plot showing association between heart attack and CHD

While these univariate and bivariate plots are helpful in determining which variables are most strongly associated with heart disease, it would really be helpful to see which features together put someone at greater risk of heart disease. I used principal component analysis (PCA) on the quantitative variables in the dataset and multiple correspondence analysis on the categorical variables to perform dimensionality reduction and help visualize all the variables in the dataset together. I first took a sample of 30,000 observations because using the whole dataset simply results in an uninterpretable blob of points. Unfortunately, no clear patterns emerged after plotting the first two principal components (PCs). Ideally, PCA would have revealed separate

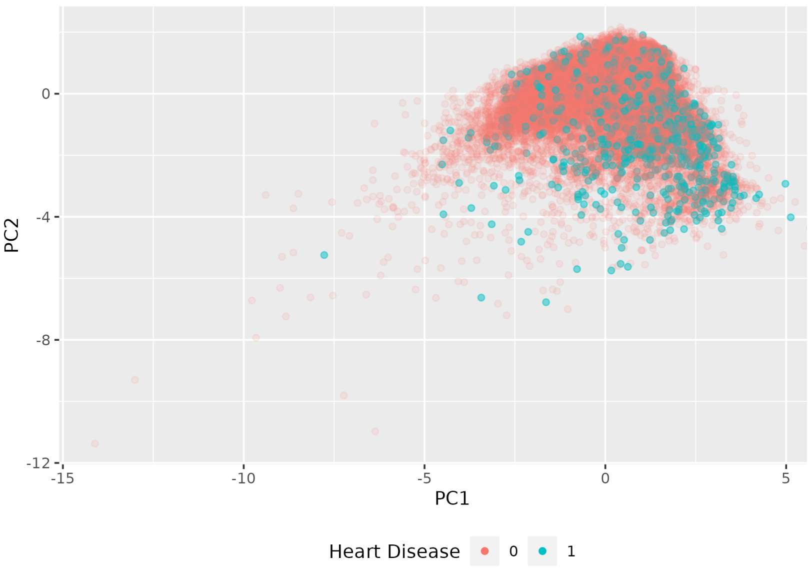

Figure 8: First two principal components for the quantitative variables

clusters corresponding to those with heart disease and those without, but we instead see a lot of overlap here. This is not at all a surprise, since the first two principal components only account for 34.22% of the variability in the data. Despite no clear separation between groups, we can observe that observations corresponding to those with heart disease seem to have slighly higher values for PC1 and slightly lower values for PC2 on average. Indeed, the median PC1 value is 0.59 among those with heart disease and 0.06 among those without heart disease. Similarly, the median PC2 value is 0.01 among those with heart disease and 0.32 among those without heart disease.

Variable	PC1	PC2
days_poor_phys_health	0.315	-0.581
days_poor_ment_health	0.272	-0.56
age	0.037	0.017
bmi	0.29	-0.179
alcohol_freq	-0.204	0.151
fruit	-0.426	-0.37
vegetable	-0.459	-0.344
cardio_freq	-0.397	-0.081
strength_freq	-0.387	-0.181

Table 1: First two columns of rotation matrix

Above are the first two columns of the rotation matrix. Since PC1 is most positively associated with days of poor physical health and most negatively associated with vegetable and fruit consumption, we might expect those with more days of poor physical health and lower vegetable and fruit consumption to be at greater risk of heart disease. Similarly, since PC2 is most positively associated with alcohol frequency and most negatively associated with days of poor physical and mental health, we might expect those with lower alcohol consumption and more days of poor physical and mental health to be at greater risk of heart disease. However, we cannot be sure of the significance of these associations without fitting a model.

Figure 9 shows the plot of the first two dimensions for multiple correspondence analysis. This method is somewhat analogous to principal component analysis and is used for categorical variables instead of quantitative variables. This plot is rather difficult to parse, but one takeaway is that because heart attack, kidney disease, stroke, diabetes, and high age categories are located close to chd1 in the bottom right corner of the plot, we can expect these variables to be strongly associated with heart disease.

Overall, my attempts at dimensionality reduction did not reveal too much information, but it was at least worth it to generate these plots in case there had been clusters we could have observed. Ideally, I would have used a method in which I could have performed dimensionality reduction on both the quantitative and categorical predictor variables, but I was not familiar with such a method.

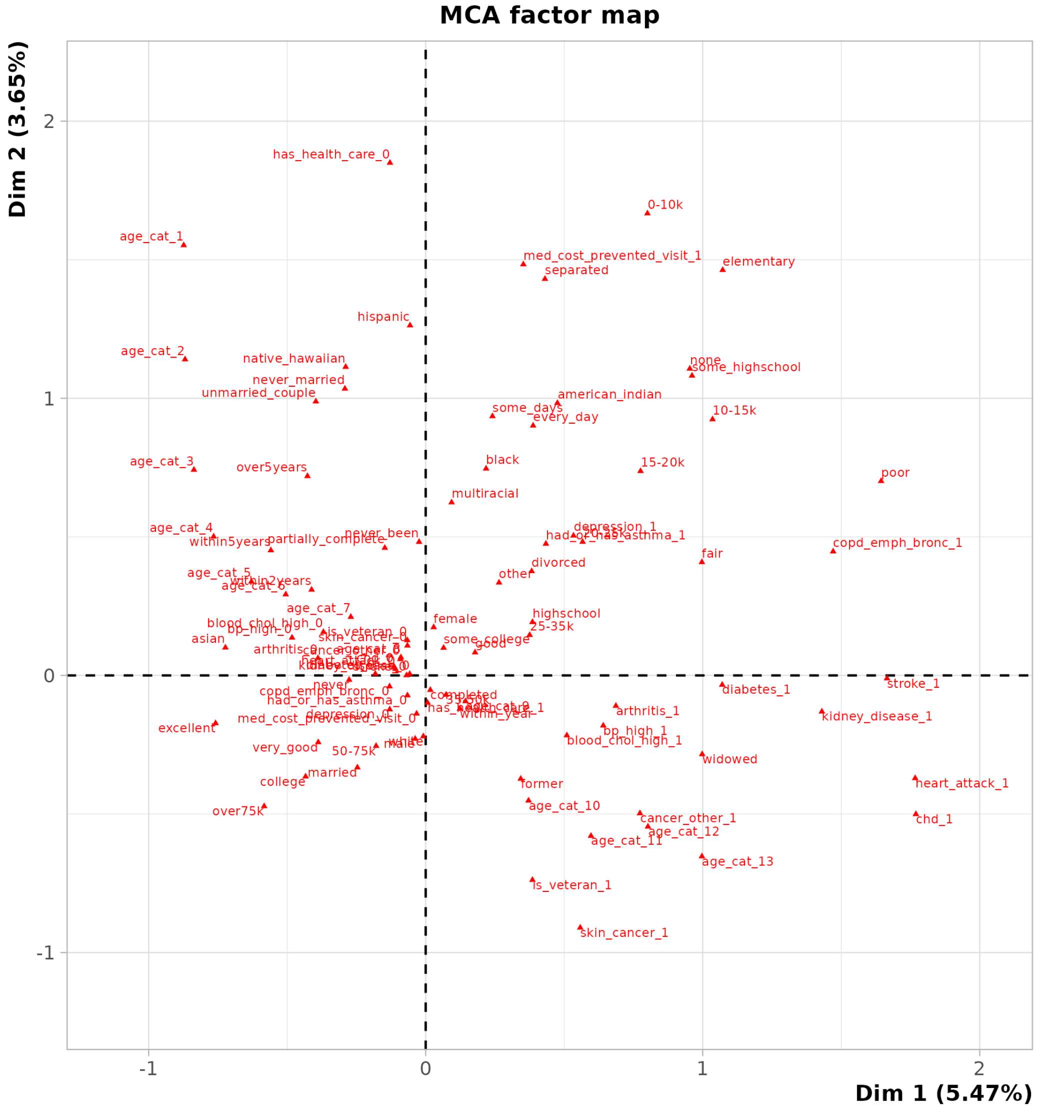

Figure 9: First two dimensions for multiple correspondence analysis

Logistic Regression Model

Given the binary outcome variable, a logistic regression model makes sense to use here. The logistic model has the added benefit over models like k-nearest neighbors or a gradient boosting machine of giving us regression coefficients that allow us to quantify the association between heart disease and the different predictor variables. Initially, I ran univariate regressions to help determine which predictors would be significantly associated with heart disease and thus valuable to include in the model. I then naively fitted a model with all the predictor variables and examined the significance of the regression coefficients and the variance inflation factors. I also implemented the stepAIC backward selection algorithm to help narrow down the set of predictors. In the end, I landed on a model with 19 predictor variables, listed below in the table of estimated regression coefficients.

The variables most strongly associated with heart disease are heart attack, high blood

Predictor	Estimate	Std. Error	z value	$P(>\|z\|)$
(Intercept)	-7.478	0.097	-77.42	0
bp_high1	0.646	0.025	25.891	0
blood_chol_high1	0.628	0.023	27.2	0
heart_attack1	2.6	0.023	110.888	0
stroke1	0.376	0.034	11.075	0
had_or_has_asthma1	0.121	0.03	4.057	0
skin_cancer1	0.124	0.028	4.369	0
cancer_other1	0.088	0.028	3.097	0.002
copd_emph_bronc1	0.646	0.029	22.215	0
arthritis1	0.302	0.022	13.494	0
depression1	0.353	0.025	13.93	0
kidney_disease1	0.563	0.037	15.41	0
diabetes1	0.304	0.025	12.4	0
sexmale	0.514	0.022	23.05	0
age	0.041	0.001	37.518	0
bmi	0.006	0.002	3.482	0
smokingformer	0.125	0.023	5.336	0
smokingsome_days	0.074	0.056	1.313	0.189
smokingevery_day	-0.045	0.039	-1.154	0.249
alcohol_freq	-0.002	0	-3.98	0
vegetable	0.013	0.006	2.327	0.02
cardio_freq	-0.002	0.001	-1.427	0.154

Table 2: Coefficient estimates for logistic regression model

pressure, high blood cholesterol, COPD/Emphysema/Bronchitis, kidney disease, sex, and age. While most of these associations are expected, I was surprised that being a male increased the risk of heart disease so much. The variables most negatively associated with heart disease were alcohol frequency and cardio frequency. Smoking every day, while negatively associated with heart disease, did not have a significant coefficient. The negative coefficient for alcohol frequency was somewhat unexpected, and while the coefficient may be significant, its small value indicates there is effectively little association. Cardio frequency also does not have a significant coefficient. This is unfortunate, as one of my objectives was to identify habits that may help decrease risk of heart disease. Unfortunately, the data do not support the hypothesis that healthy nutrition and exercise habits help reduce the risk of heart disease. We know from other studies that this is undeniably true, but this dataset does not support this conclusion.

Interpretation of selected regression coefficients:

- The estimated coefficient associated with high blood pressure is 0.646, and the 95% confidence interval is (0.597, 0.695). Holding all other variables constant, the odds of heart disease are $e^{0.646} = 1.91$ times higher among people with high blood pressure than among people without high blood pressure, on average.

- The estimated coefficient associated with age is 0.041, and the 95% confidence interval is (0.0387, 0.0430). Holding all other variables constant, the model predicts that for a 10-year increase in age, the odds of heart disease multiply by $e^{(10)(0.041)} = 1.51$, on average.
- The estimated coefficient associated with being a former smoker vs. never having smoked is 0.125, and the 95% confidence interval is (0.079, 0.171). Holding all other variables constant, the odds of heart disease are $e^{0.125} = 1.13$ times higher among former smokers than among people who have never smoked, on average. Because we have adjusted for age, we are more confident that this effect is actually due to smoking and not a result of the age confounding.

In terms of prediction, I refit the model, this time leaving heart attack out. If we are trying to predict risk of heart disease, it seems unfair to include a symptom of heart disease since the goal is to identify and lower risk before one reaches the stage of having a heart attack. Removing this predictor no doubt lowers the predictive capabilities of the model, but it makes more sense this way. Also, this prediction stage of course uses the testing set, whereas the fitting of the model used the training set. The ROC curve is shown below, and we observe an AUC of 0.8318. This is by no means a perfectly predictive model, but an AUC of 0.8313 indicates that it does a decent job at identifying those at risk of heart disease.

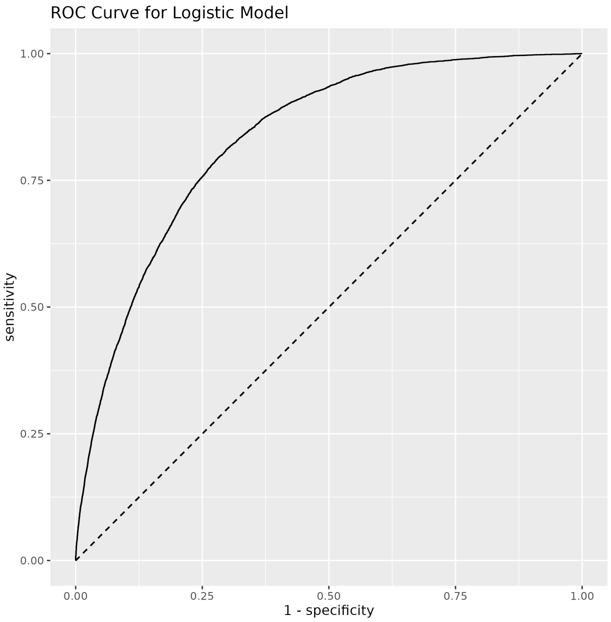

Figure 10: ROC curve for logistic regression model

Using a threshold probability of 0.073, the model achieves a sensitivity of 0.75, a specificity of 0.76, a positive predictive value of 0.18, a negative predictive value of 0.98, and an F1 score of 0.28. This means the model performs moderately well using this threshold. While the positive predictive value is low–only 18% of those predicted to have heart disease do have it–we can compare this number to the overall prevalence in the dataset of 5.8% and recognize the threefold increase in risk of heart disease given the model's prediction.

Gradient Boosting Model

I also ran a gradient boosting machine to compare its predictive capabilities to the logistic model. As seen in table 3, the gradient boosting machine identifies age, high blood pressure, and COPD/Emphysema/Bronchitis as the most influential variables. The variables identified by the GBM model as important agree with the logistic model.

Variable	Relative Influence
age	19.52
bp_high	16.70
copd_emph_bronc	15.59
blood_chol_high	14.92
stroke	10.68
diabetes	7.16
kidney_disease	5.89
sex	3.73
arthritis	3.62

Table 3: Top 9 most influential variables in GBM model

The ROC curve for the GBM model is quite similar to that of the logistic model, and we obtain a similar AUC of 0.8435. Using a threshold probability of 0.07, we observe a sensitivity of 0.75, a specificity of 0.77, a positive predictive value of 0.17, a negative predictive value of 0.98, and an F1 score of 0.28–almost exactly the same as those of the logistic model.

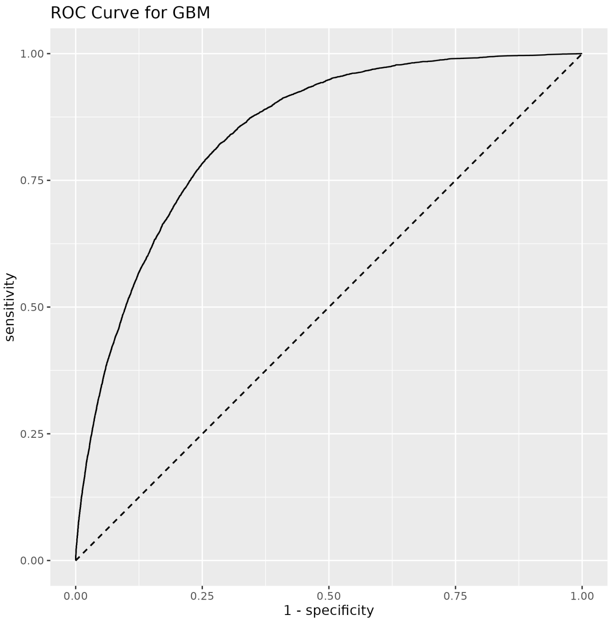

Figure 11: ROC curve for gradient boosting model

Discussion

Overall, I was pleased with the performance of the logistic model, since it had significant regression coefficients and was successful for the most part in identifying those at risk of heart disease. The GBM model had similar success with regards to prediction accuracy. The models showed that greater age, a sex of male, and common chronic health conditions like high blood pressure and kidney disease were most strongly associated with heart disease. Though I sought to show that healthy lifestle habits such as frequent exercise and good nutrition can help lower risk, the data did not support such claims. A more in-depth study of exercise and nutrition habits would be necessary to show an association with heart disease.

There are several ways I can improve this project in the future. Perhaps the largest issue in the data I ignored was the high degree of missingness. For the modeling, I simply used the complete cases, but imputation of missing data may have helped obtain more unbiased regression coefficient estimates. I would also try to perform dimensionality reduction on both the quantitative and categorical predictor variables at once. It appears this dataset is not a great candidate for dimensionality reduction, but working with all the variables together may have helped reveal more patterns I was not able to find. Lastly, I would put a greater emphasis on model selection. For both the logistic model and the GBM model, I selected variables with the assistance of stepAIC and by including variables of interest to me, but I could have put greater care into which combination of variables led to the highest predictive accuracy.