Predicting Heart Disease and Attacks

Alexander Poletaev
SID: 915803952
STA160 Section A02
adpoletaev@ucdavis.edu

Richard Ge
SID: 916466262
STA160 Section A02
rwge@ucdavis.edu

Daniel Momeni
SID: 918389787
STA160 Section A02
dmomeni@ucdavis.edu

Kunteng Miao
SID: 916878494
STA160 Section A01
ktmiao@ucdavis.edu

Jonathan Jiang
SID: 916619792
STA160 Section A01
jsjjiang@ucdavis.edu

Abstract

Heart Disease is one of the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. The purpose of this project is to use data analysis techniques to identify groups with the highest risk of heart disease and use the shared characteristics of these groups to predict whether someone is at risk of heart disease. We will first use contingency tables, conditional entropy, and odds ratios to identify at-risk groups and key predictor variables; more detail regarding at-risk groups will come from crossing multiple variables, and seeing which variables recur in the most-significantly-different groupings. Then, we will create a set of decision rules to determine whether a new respondent is at risk of heart disease. We will then compare this classifier's performance with that of a gradient boosted decision tree algorithm. We will also be using a general linear model to determine which predictor variables play the most importance when it comes to heart problems within patients. Tests will be conducted with a focus on AIC, logarithmic odds, summary statistics, and plots to aid in our determination of a best general linear model.

1 About the Data

1.1 General Description

The data set used for this project contains 253,680 processed responses from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey collected annually by the Centers for Disease Control and Prevention (CDC). The data was cleaned prior to being made available for download, and there are no missing values. The data set contained 22 variables in total. The variable names and types are as follows:

- HeartDiseaseorAttack: whether the respondent had heart disease or a heart attack. This binary categorical variable was the dependent variable of interest.
- HighChol: whether or not the respondent had high blood cholesterol diagnosed by a doctor, nurse, or other health professional (binary categorical).
- CholCheck: whether or not the respondent had a cholesterol check performed within the past five years (binary categorical).
- BMI: the respondent's Body Mass Index (continuous numeric).
- Smoker: whether the respondent has smoked at least 100 cigarettes (binary categorical).

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

- Stroke: whether the respondent has ever told they had a stroke (binary categorical).
- Diabetes: ever told they have diabetes (categorical 0-2).
- PhysActivity: whether or not the respondent did physical activity or exercise during the past 30 days other than their regular job (binary categorical).
- Fruits: whether or not the respondent consumes fruit 1 or more times per day (binary categorical).
- Veggies: whether or not the respondent consumes vegetables 1 or more times per day (binary categorical).
- HvyAlcoholConsump: whether or not the respondent is a heavy drinker. Heavy drinking is considered having more than 14 drinks per week for adult men and having more than 7 drinks per week for adult women (binary categorical).
- AnyHealthcare: whether or not the respondent has any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service (binary categorical).
- NoDocbcCost: whether or not there was a time in the past 12 months when the respondent needed to see a doctor but could not because of cost (binary categorical).
- GenHlth: the respondents' assessment of their general health, from best to worst (categorical 1-5).
- MentHlth: the number of days during the past 30 days when the respondent's mental health was not good (discrete numeric).
- PhysHlth: the number of days during the past 30 days when the respondent's physical health was not good (discrete numeric).
- DiffWalk: whether or not the respondent has serious difficulty walking or climbing stairs (binary categorical).
- Sex: the sex of respondent (binary categorical).
- Age: the age group of respondent, from youngest to oldest (categorical 1-13).
- Education: the highest grade or year of school completed by the respondent (categorical 1-6).

1.2 Class Imbalance

The main characteristic of the data was the large class imbalance between survey respondents who did not have heart disease or a heart attack and those who did. Out of the 253,680 total respondents, 229,787 did not have heart disease or a heart attack, and only 23,893 did. That is roughly a 90.5% to 9.5% class imbalance. Intuitively, this imbalance makes sense, as heart disease and heart attacks are not phenomena that occur at a 50% rate across the population. This imbalance made any count based data analysis not very helpful. As a result, we relied almost entirely on relative frequency based data analysis techniques when identifying at-risk groups.

2 Data Analysis Techniques

In our analysis, we primarily relied on three main numeric techniques: relative frequency contingency tables, conditional entropy, and odds ratios. After analyzing the resulting tables, we made visualizations for only the most significant trends, as visualizing associations that are not helpful in identifying at-risk groups was redundant and did not add meaningful value to our analysis.

2.1 Contingency Tables

A contingency table displays the frequency distribution of variables within a population. To account for the class imbalance in my data and create easily interpretable results, I displayed the percentage-wise frequencies of my variables rather than their counts. I isolated two, and later three, variables in each table. One of those variables was always my dependent variable of interest (whether or not the respondent had heart disease). Then, I computed the probabilities for both outcomes of my

dependent variable across all possible values of the independent variable(s) in question. Here is a sample table:

Table 1: Sample Table

High Blood Pressure Value	No Heart Disease/Attack	Heart Disease/Attack	Proportion of Overall Population	Conditional Entropy
0	0.958820	0.041180	0.570999	0.247678
1	0.835264	0.164736	0.429001	0.645519

2.2 Differences in Crossed Variables

As a simpler prelude to the conditional entropy analysis, we also attempted to cross all predictors multiple times and detect any association or synergistic effect between specific predictors. The process employed is fairly simple in comparison - we took two different predictors and crossed them with the heart disease outcome, and created a contingency table of the resulting counts. Dividing by row sums results in the proportions of each crossed group having, or not having, heart disease; the sum of the row is now 1 (i.e. there is a 100% chance of having or not having heart disease). Our statistic for each cross between variables is the greatest difference between rates in the column with heart disease, as this shows that different values for those variables cause a large change in the outcome. We also chose this statistic because it can easily be visualized and understood, and does not rely on any statistical tests or underlying distribution to

One of the concerns with this method is the reduced sample size after crossing multiple variables, especially when one or more of the variables has very few observations in some bins. This was mitigated by increasing the size of the bins to include more test subjects at a time, and by setting a threshold (here, $n_i \geq 30$) for the minimum amount of observations in each cell.

As an example, one such contingency table (after thresholding) is shown below. Note that the difference between the highest and lowest rates of heart attack is between women who eat fruits and have no difficulty walking (0.041638) and men who eat fruits and have difficulty walking (0.299154), which comes out to 0.257516.

Table 2: Sample Crossed Table

			No H.D or Attack	H.D. or Attack
No Diff. Walking	<1 fruit daily	Female	0.954901	0.045099
No Diff. Walking	<1 fruit daily	Male	0.907009	0.092991
No Diff. Walking	≥ 1 fruit daily	Female	0.958362	**0.041638**
No Diff. Walking	≥ 1 fruit daily	Male	0.904785	0.095215
Diff. Walking	<1 fruit daily	Female	0.806844	0.193156
Diff. Walking	<1 fruit daily	Male	0.703329	0.296671
Diff. Walking	≥ 1 fruit daily	Female	0.803883	0.196117
Diff. Walking	≥ 1 fruit daily	Male	0.700846	**0.299154**

2.3 Conditional Entropy

For each row of my contingency tables, or subgroup within the population, I also calculated the conditional entropy. Entropy is a measure of uncertainty, and the groups with the greatest conditional entropies would become the groups I would focus on when developing my set of decision rules. I used the following equation to calculate conditional entropy over a subgroup of the population:

$$H(x) = -\sum_{x \in X} p(x) log_2(p(x))$$

Where $H(x)$ is the conditional entropy of a particular subgroup and $p(x)$ is the probability of the given subgroup within the data. While it is believed that the specific logarithm does not matter,

using a base-two logarithm proved very useful in this case. I used a base-two logarithm because for each subgroup there were two possible outcomes: the respondent either had heart disease or they did not. The base-two logarithm also made interpreting my results easier, as the function for the conditional entropy of any subgroup has a maximum value of one. The function and the derivation of its maximum are displayed below.

$$H(x) = -xlog_2(x) - (1-x)log_2(1-x)$$

$$\frac{d}{dx}H(x) = -log_2(x) + log_2(1-x)$$

$$-log_2(x) + log_2(1-x) = 0$$

$$log_2(x) - log_2(1-x) = 0$$

$$log_2(\frac{x}{1-x}) = 0 \text{ at } x = 0.5$$

$$\text{At } x = 0.5,\ H(x) = -xlog_2(x) - (1-x)log_2(1-x) = 1$$

2.4 Odds Ratio

An odds ratio is a statistical measure of the strength of the association between two events. In my case, the two events are the presence of an indicator variable, such as high blood pressure, type two diabetes, etc., and presence of heart disease. I used the following formula to compute odds ratios for the variables:

$$OddsRatio = \frac{\frac{P(HeartDisease \cap A)}{P(A)}}{\frac{P(HeartDisease \cap A^c)}{P(A^c)}}$$

For multilevel categorical variables, I computed an odds ratio for each value of the variable, the only difference being that the value in the denominator corresponded to the percentage of respondents with heart disease in the compliment of the group in the numerator. So, for example, in the case of type 2 diabetes, the compliment was all respondents with no diabetes or type one diabetes. For some of the variables, such as whether or not the respondent is physically active, the inverse of the odds ratio above made sense, and I made sure to take that into account when assessing my results.

2.5 Variable Conversion

In order to use contingency tables and conditional entropy effectively for all of the variables included in the data set, I converted the numeric variables to categorical ones. Rather than keeping BMI as a continuous numeric variable, I transformed it into a six-value categorical variable based on the CDC's definition of adult overweight and obesity. The BMI categories are defined as follows:

- 1: BMI < 18.5, underweight
- 2: 18.5 $\leq$ BMI $<$ 25, healthy
- 3: 25 $\leq$ BMI $<$ 30, overweight
- 4: 30 $\leq$ BMI $<$ 35, class 1 obesity
- 5: 35 $\leq$ BMI $<$ 40, class 2 obesity
- 6: 40 $\leq$ BMI, class 3 "severe" obesity

This not only reduced the size of my tables, but also made the results of those tables more easily interpretable. Moreover, this transformation aligns with the overweight and obesity definitions of the organization that produced the data I am using. I also converted both MentHlth (number of bad mental health days out of the last 30) and PhysHlth (number of bad physical health days out of the last 30) from discrete numeric variables to five-group categorical variables. I used five groups in order to mirror the GenHlth (the respondents' assessment of their general health) variable structure. The new groups for both variables are as follows:

- 1: 5 or fewer bad days
- 2: 6-11 bad days
- 3: 12-17 bad days
- 4: 18-23 bad days
- 5: 24 or more bad days

For crossing three variables at a time, we instead split the MentHlth and PhysHlth variables into fewer bins than before. The number of bins was reduced to prevent group sizes from becoming too small after crossing multiple times. The new thresholds are as follows:

- 1: 0-10 bad days
- 2: 11-20 bad days
- 3: 21-30 bad days

3 Results

3.1 One-Way Conditioning

3.1.1 Binary Categorical Variables

Of the binary categorical variables, the most significant were HighBP, HighChol, Smoker, Stroke, and DiffWalk. The names and significant values for these variables are summarized in the table below:

Table 3: Significant One-Way Binary Results

Variable Name	Variable Value	Odds Ratio	Conditional Entropy
High Blood Pressue	1	4.592098602559366	0.645519
High Cholesterol	1	3.5890725604845954	0.623947
Has Smoked 100 Cigarettes	1	2.2039431659792883	0.561958
Stroke	1	6.936202083608327	0.959809
Has Difficulty Walking	1	4.266085291276664	0.782064

Most of these intuitively make sense, as we would expect people with high blood pressure, high cholesterol, a history of smoking, and difficulty walking to be at a higher risk of heart disease. It is also important to note that, while some of the other variables associated with poor health practices, such as not eating fruits and vegetables and not being physically active, had odds ratios over 1.5, their conditional entropies were insignificant. It is also important to note that stroke produced a conditional entropy of roughly 0.95, which is very close to the maximum possible conditional entropy.

3.1.2 Multi-level Categorical Variables

Of the multi-level categorical variables, the most significant were Diabetes, GenHlth, Age, Education, Income, and PhysHlth. More specifically, there were certain values for each variable that stood out. These results also intuitively make sense, as one would expect older, poorer, and less educated people with diabetes, poor health to be a greater risk of heart disease. It is interesting to note that, out of these subgroups, it is the people who assessed their own general health to be the worst (GenHlth = 5) that are most likely to have heart disease. This is interesting because this assessment is done by the respondent, and not by a medical professional. The names and significant values for these variables are summarized in the table below.

Table 4: Significant One-Way Multi-Level Results

Variable Name	Variable Value	Odds Ratio	Conditional Entropy
Diabetes Type	2	3.038576	0.765396
General Health	4	2.757628	0.747392
General Health	5	4.151059	0.924776
Age	11	1.935260	0.652467
Age	12	2.211918	0.708848
Age	13	2.868428	0.794262
Education	2	2.078217	0.706530
Income	2	2.078859	0.693997
Bad Physical Health Days	4	2.137758	0.717086
Bad Physical Health Days	5	2.949996	0.791877

3.2 Two-Way Conditioning

In two-way conditioning, I narrowed down the subgroups for which I was computing the conditional entropy. Now, rather than just checking heart disease outcomes across all possible values of one variable, I checked heart disease outcomes on combinations of all possible values of two variables. Here is an example of one such contingency table:

Table 5: Sample Two-Way Result

High Blood Pressure Value	Stroke Value	No Heart Disease/Attack	Heart Disease/Attack	Conditional Entropy
0	0	0.963238	0.036762	0.227245
0	1	0.723285	0.276715	0.850937
1	0	0.854462	0.145538	0.598560
1	1	0.580459	0.419541	0.981240

Some two-way combinations had no observations belonging to them, but this did not influence the result of my analysis. Almost all significant two-way groups involved the variables and values identified in one-way conditioning. Binary variables like high blood pressure, high cholesterol, stroke, difficulty walking up stairs, low levels of income and education, and high values of age, general health (so poor general health), and bad physical health days, and the presence of diabetes led to consistently high conditional entropies in the subgroups they were present in. Even when the other variable conditioned on within the subgroup was not one that was previously identified as significant, the conditional entropy of the subgroup was high enough to warrant a recheck. The only new significant subgroups that were not noted in one-way conditioning were underweight or severely obese people (the lowest and highest values of BMI). As a result, I identified the following eleven subgroups as particularly high risk groups for heart disease:

- People with high blood pressure (HighBP = 1).
- People with high cholesterol (HighChol = 1).
- People who have had a stroke (Stroke = 1).
- People who have difficulty climbing stairs (DiffWalk = 1).
- People on the low end of the income spectrum (Income = 1-3).
- People on the low end of the education spectrum (Education = 1-3).
- People on the high end of the age spectrum (Age = 11-13).
- People with a poor assessment of their general health (GenHlth = 4-5).

- People who have experienced 18-30 bad physical health days within the past 30 days (PhysHlth = 4-5).
- People with type 1 or type 2 diabetes(Diabetes = 1-2).
- People that are either underweight or severely obese (BMI = 1 or BMI = 6)

3.2.1 Three-Way Conditioning

These conclusions still hold when crossing three predictors with the outcome variable. Of the top 20 crossings with the largest differences in risk, Stroke appeared in all of them. The next most common variable was DiffWalk (which appeared in 7 crossings), then PhysActivity, HighChol, and Sex (each appeared in 6), and finally Smoker and HighBP (each appeared in 4). Interestingly, Age did not make any appearances in the top 20; this is likely due to the effects of filtering out sparse or empty cells. In the first place, filtering was necessitated by low sample sizes among certain crossed groups resulting in 100% heart disease rates, that are actually very rare in real life (e.g. 18-25 year olds who have had both ≥ 1 one stroke and Type 2 diabetes). As a result, the conclusions drawn from this analysis will be more generalizable to the majority of the population, and focused on the actual markers for heart disease in patients. Almost all of the most significant markers are associated with old age, and at younger ages, almost nonexistent. For example, the conclusion that 18-25 year olds should avoid having strokes or becoming diabetic may be somewhat pointless, as the group that this conclusion is based on may have only have 5-10 members, due to its rarity (due to the general unlikelihood that a 18-25 year old would have both of those markers). A plot of the top 20 crossings, after removing these anomalous groups, is shown below.

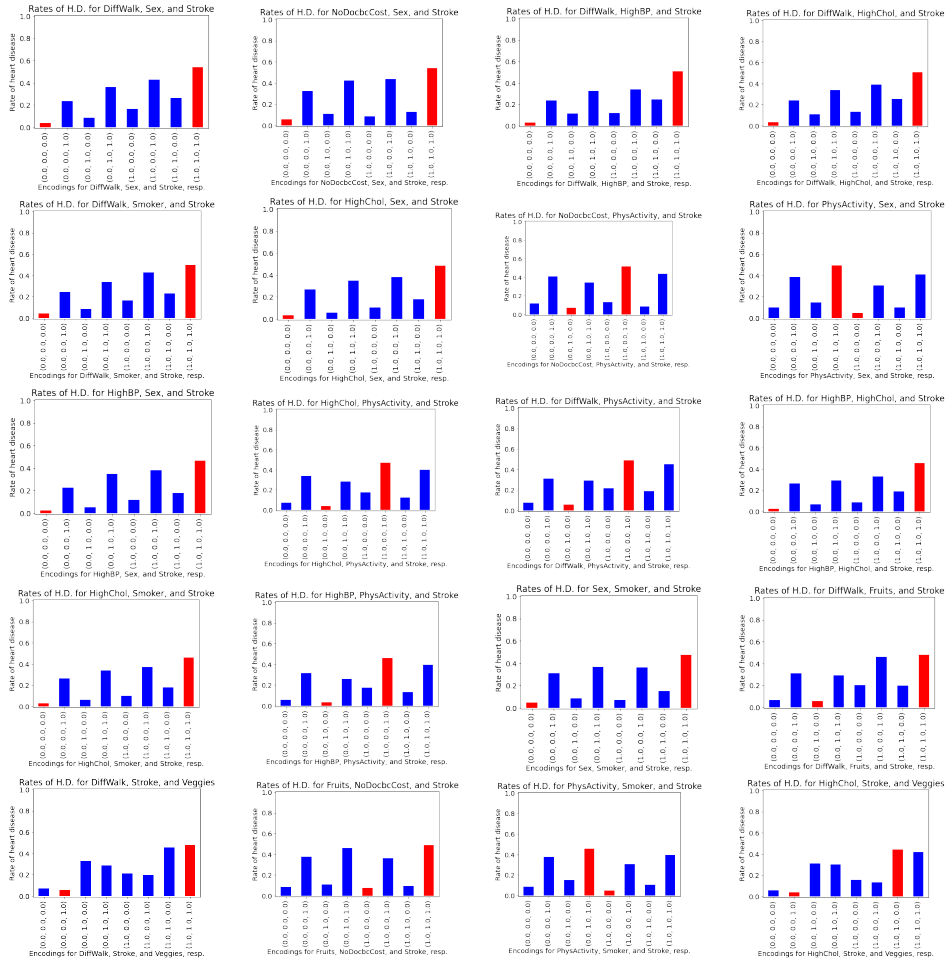

Figure 1: Top 20 Most Significant Crossings

Using these subgroups, I came up with a set of decision rules for my classifier.

4 Developing the Classifier

4.1 Goal

When constructing my classifier, my goal was to minimize the number of false negative classification, or type II errors. This is because, in the case of heart disease, a false positive is far less severe than a false positive. In the event of a false positive, the patient may take steps to improve their heart health that they will most likely need to take later down the line. In the event of a false negative, however, the patient will rest easy and ignore a potentially life threatening condition.

4.2 Structure

My classifier is structured as follows: given a respondent/observation, that respondent's 'heart disease score' is determined by the number of the at-risk groups listed above that they belong to. So their score is an integer between 1 and 11. Given a particular 'tolerance' between 1 and 11, if a given respondent's score is less than that tolerance, then they are determined to not be at risk of heart disease. If their score is greater than or equal to that tolerance, then they are determined to be at risk of heart disease.

4.3 Results

Given that the highest score any respondent got was 8, I tested my classifiers at all possible scores from 1 to 8, recording the overall accuracy and the accuracy of correctly identifying heart disease risk each time. At low levels of the score parameter, the classifier did well in correctly identifying heart disease, and poorly in terms of overall accuracy. The opposite was true for high score values. This inverse relationship is depicted in the plot below.

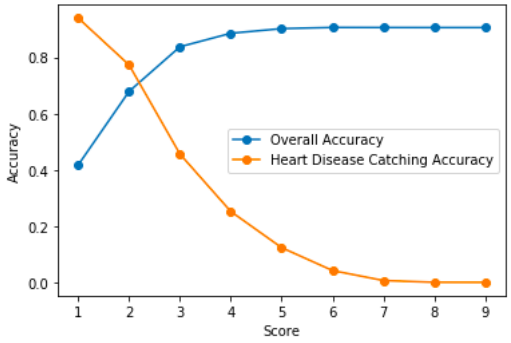

Figure 2: Classifier Accuracies

Using a score of 1, the classifier was able to accurately classify roughly 94% of heart disease cases of the ones in the data. So, while this did also result in a very poor overall accuracy of about 42%, the result is encouraging. The full table of classifier accuracies is displayed below:

5 Catboost Classifier

To put my findings into perspective, I attempted the same classification problem using the binary classifier from the CatBoost machine learning library. CatBoost is a high-performance open source library for gradient boosting on decision trees that is widely considered to provide best-in-class accuracy. I created two models, one with no specified class weights, and one where the class weights that prioritize successfully identifying heart disease risk over overall accuracy. I weighted the class with no heart disease with 1 and the class with heart disease with 10, as this roughly reflects the class imbalance proportions of my data.

Table 6: Classifier Results

Score	Heart Disease Detection Accuracy	Overall Accuracy
1	94.0%	41.7%
2	77.3%	67.9%
3	45.6%	83.7%
4	25.2%	88.6%
5	12.3%	90.197%
6	4.2%	90.622%
7	0.07%	90.596%
8	0.003%	90.581%
	0	90.581%

5.1 No Specified Class Weights

The model with no specified class weights had an overall accuracy of about 91.0% and was about 11.0% accurate in successfully identifying people at risk of heart disease. This is similar to the performance of my decision rule-based classifier using a score of 5. This result clearly reflects the prioritization of overall accuracy and the affect of the unremedied class imbalance.

5.2 With Specified Class Weights

The model with specified class weights had an overall accuracy of about 73.7% and was about 83.4% accurate in successfully identifying people at risk of heart disease. This is similar to the performance of my decision rule-based classifier using a score of 2. This result supports my previous findings, most notably the inverse relationship between successful heart disease detection and overall accuracy.

5.3 Feature Importance

Both models had the same feature importance rankings, the only notable difference being that the model with class weights had sex ahead over high cholesterol. Sex was also the only variable the classifier found important that I missed in the variable subgroups I used in my decision rules. This is very validating to my decision rules and the ideas behind my classifier.

6 General Linear Model

6.1 Goal of General Linear Model

The purpose of using a general linear model in our case is to create a compact final model that simultaneously considers several linear regression models. The data set had many predictor variables to choose from however the most important seemed to be high blood pressure, high cholesterol, sex, age, and healthcare status. It would on face value make sense for one to think both high blood pressure and high cholesterol play a role in heart disease or attacks so we wished to explore that relation to see if that is indeed the case. We also wanted to see if higher age correlated with heart disease and attacks since older populations may seem more at risk. Sex was chosen to see if there was meaningful differences between the sexes. Healthcare status was chosen to see whether or not being on healthcare made a significant role as that factor could be used to shape healthcare policy based on the results. As per the case disease or attack will be our response variable to the chosen predictor variables being age, high blood pressure, high cholesterol, sex, and healthcare.

7 GLM Data Analysis Techniques

7.1 Summary Statistics

First we wish to get a basic understanding of our data so we will look at summary statistics of our response variable compared to the individual predictor variables. When we look at Blood pressure and heart disease or attack we see that group 0 has a mean of .04 while group 1 has a mean of .16.

There does seem to be a large difference between groups that may show that high blood pressure is more likely to lead to heart problems. This will further be explored when it comes to modeling the data. The standard deviation for blood pressure and heart problems is .20 for group 0 and .37 for group 1. The standard deviations seem to be relatively high. For high cholesterol and heart problems we see a mean of .05 for group 0 and a mean of .15 for group 1. Compared to high blood pressure this could potentially be indicative that blood pressure plays a slightly more significant role in terms of heart health. The standard deviation for high cholesterol and heart problems is .22 for group 0 and .36 for group 1. There does seem to be more variation for this variable as well compared to high blood pressure. For sex we seen a mean of .07 for group 0 and .12 for group 1 with a standard deviation of .26 for group 0 and .33 for group 1. It could be the case that group 1 is more at risk. In regards to age we see that the mean increases with age along with standard deviation so older populations seem to be more at risk which seems expected. Healthcare status has a mean pf .07 for group 0 and .1 for group 1 with standard deviation of .26 for group 0 and .29 for group 1. Both groups are pretty close together so between all of our variables it could be the case that healthcare status has the lowest significance in our model.

7.2 Graphical Representation

When looking at the box plots of high blood pressure, high cholesterol, and sex we see that the counts for group 0 and larger than group 1 even though between all variables group 1 had a larger mean and standard deviation. This is interesting because the groups with 1 as the indicator are more concentrated in their results. For age we see that the count is highest in the middle which seems to be normal since the population evens out there in terms of age. For healthcare we see that most people are insured and since the means are relatively close together compared to the others it does seem to be more likely that healthcare might not play as large a role as previously thought.

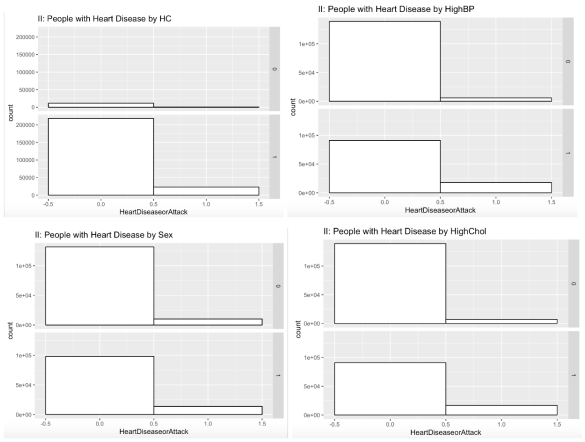

Figure 3: Histograms

7.3 Log Odds

Below are graphical representations of Log Odds for each factor that has been chosen for our general linear model analysis. Log odds will be advantageous as it will make our output easier to understand. It will brief us on the odds someone is going to get heart disease or a heart attack based on the predictor variable at play.

7.4 Individual Model AIC's

Another technique used was to create individual models between heart disease or attack with our predictor variables to see which ones individually had lower AIC's. This was done as a precursor to the final model which included all of them to see which ones would be more likely to have a greater effect on the final model. A general linear model was created for each combination, for a total of 5 results which can be seen below.

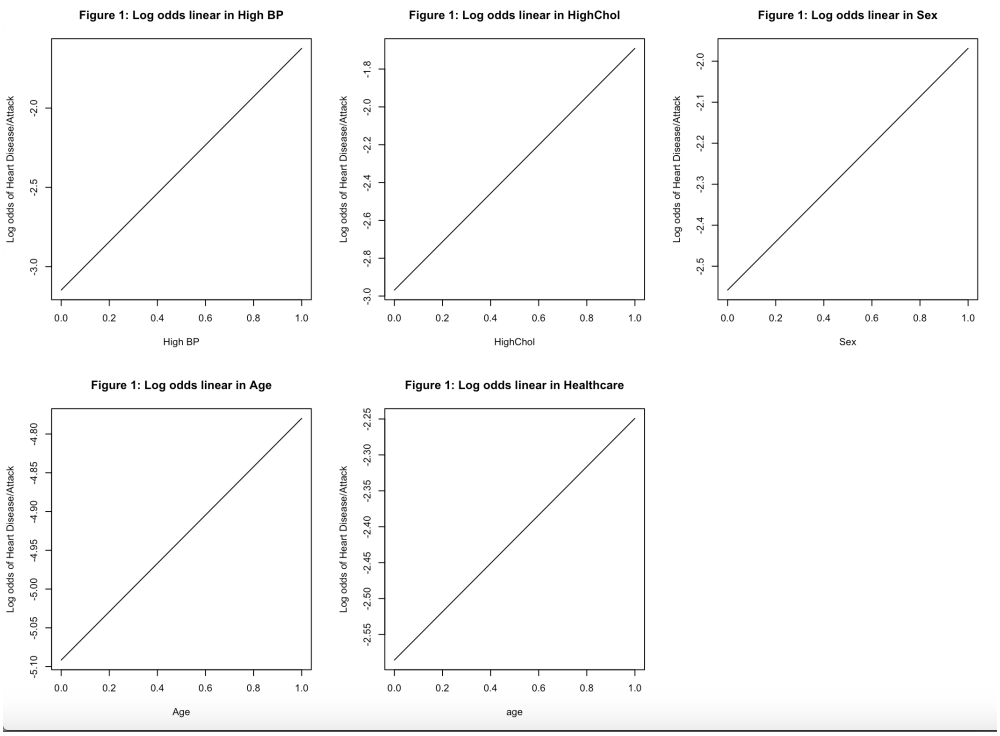

Figure 4: Log Odds

- AIC High BP - 147,100
- AIC High Chol - 150,100
- AIC Sex - 156,500
- AIC Age - 144,200
- AIC HC - 158,300

We see that the AIC of age and blood pressure are the lowest, so we would expect that those two factors are the most important in our final model that includes every factor. Our initial prediction would be that high blood pressure and age are the most likely to be associated with heart problems in one's life.

8 GLM Results

8.1 GLM Final Model Selection

Using a stepwise algorithm we wished to choose a model with the lowest AIC through a combination of our variables used in the general linear model. AIC stands for Akaike Information Criterion, and is an estimator of prediction error; thus, we would want the lowest possible AIC, as that would mean we have a lower prediction error. Our final model looked like:

HeartDiseaseorAttack Age + HighBP + HighChol + Sex + AnyHealthcare + Age:HighBP + Age:Sex + Age:HighChol + HighBP:Sex + HighBP:HighChol + Age:AnyHealthcare + Age:HighBP:HighChol + Age:HighBP:Sex

The colons between variables represent an interaction between both variables in our model. By looking at the model we see that age and high blood pressure are the most indicative of heart disease or a heart attack. Healthcare status seemed to have the least impact so we can say that whether or not someone had healthcare did not greatly affect whether someone is more likely to get a heart attack. High cholesterol and sex do seem to be moderately important factors. The binary indicator

representing 1 for sex seems more at risk for heart disease and attacks compared to the 0 indicator, which could explain why sex is not as important as age and high blood pressure. These two important factors also seem to be in line with other forms of classification done in this study, especially as seen in the early individual model AIC's and summary statistics.

9 Unsupervised Learning

9.1 General purpose of Unsupervised Learning

We attempted three Unsupervised Learning techniques: Hierarchical Clustering heat mapping, Agglomerative Hierarchical Clustering, and Principal Component Analysis, to find hidden patterns. The data set has many predictor variables possibly explaining the causes of heart disease. We will ignore pairing variables such as HighBP, HighChol,and Smoking with HeartDiseaseorAttack because those are known, not hidden patterns. We will look for correlations between possible variables with a correlation heat map and conduct Unsupervised learning.

9.2 Exploratory Data Analysis

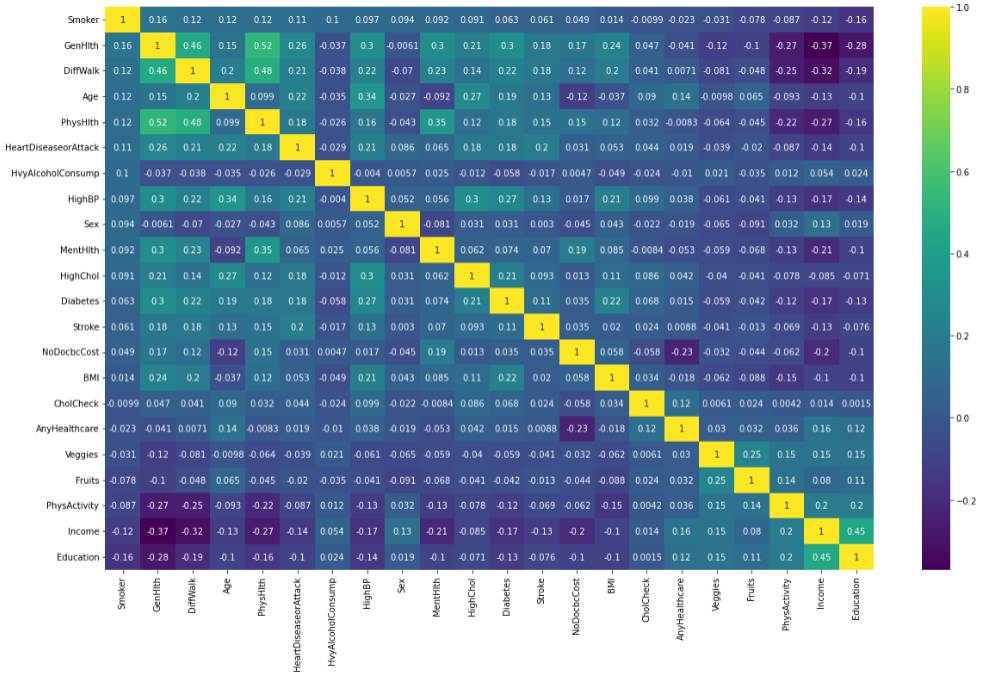

Figure 5: Correlation Heat map

We want to look at potential missed patterns between the predictor variables. Above, we constructed a correlation heat map to obtain potential patterns before using cluster analysis.

9.3 Cluster Analysis Techniques

Using the correlation heat map and information from GLM, we can identify potential patterns by observing correlations ≥ 0.25. Using hierarchical cluster analysis, we construct a dendrogram to use Agglomerative clustering on HighBP + HighChol(0.3), Age + HighBP(0.34), Age + HighChol(0.27). We notice that Sex has a very low correlation so it will not be included. The dendrograms will tell us how many clusters are necessary. Normally, linkage is a primary requisite of hierarchical cluster analysis as it measures the distance between the two clusters in various ways. However, as this dataset has many 0s and 1s, linkage will stay the same between each cluster analysis. Furthermore, we will conduct Principal Component Analysis to simplify the data as there are too many variables and the dataset is very large.

9.4 HCA and PCA results

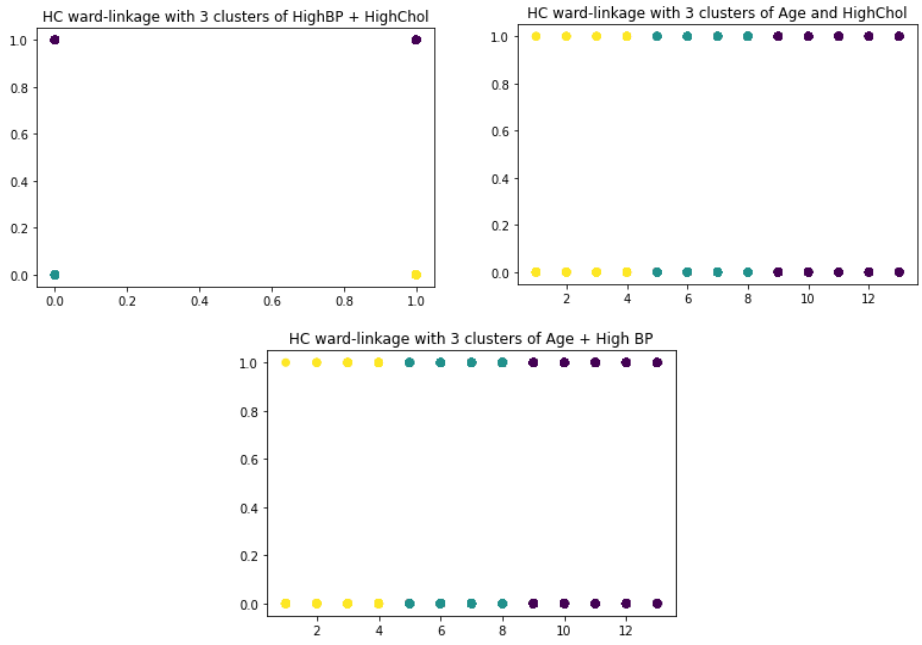

Figure 6: HCA plots

With Agglomerative Hierarchical Clustering, we observe that there are no new patterns learned from Age + HighChol and Age + HighBP. There might be potentially new patterns if the number of clusters were increased, but that would go against our dendrograms. However, HighBP + HighChol has yielded an interesting pattern. The 3 clusters shown are without either, without HighBP but with HighChol, and with HighChol with or without High BP (see figure 6). This may tell us that high blood pressure in people may be due to other factors but high blood pressure may be caused by high cholesterol.

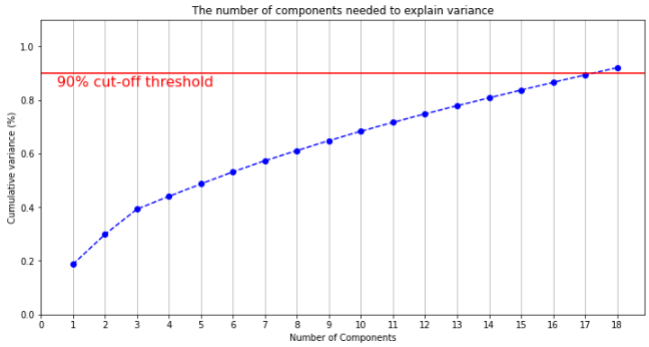

Figure 7: Explaining Variance

With PCA, we observe that we need 18 of 22 variables to achieve at least a 90% variance. Lastly, we wanted to observe which variables (if any) out of Age, HighChol, HighBP, or HeartDiseaseorAttack had separate clusters or not.

(See figure 8 below) Obviously, people with a heart disease or attack will have many different attributes than others. As people get older, they seem to share similar characteristics. Each cluster of each age looks to be it's own; however, the middle is hard to tell. People with high cholesterol seem to have varying characteristics. There's a distinct blue and yellow section, but there is a lot of overlap. On the other hand, people with or without high blood pressure have more distinct clusters. There is a noticeable part that overlaps but it is much smaller.

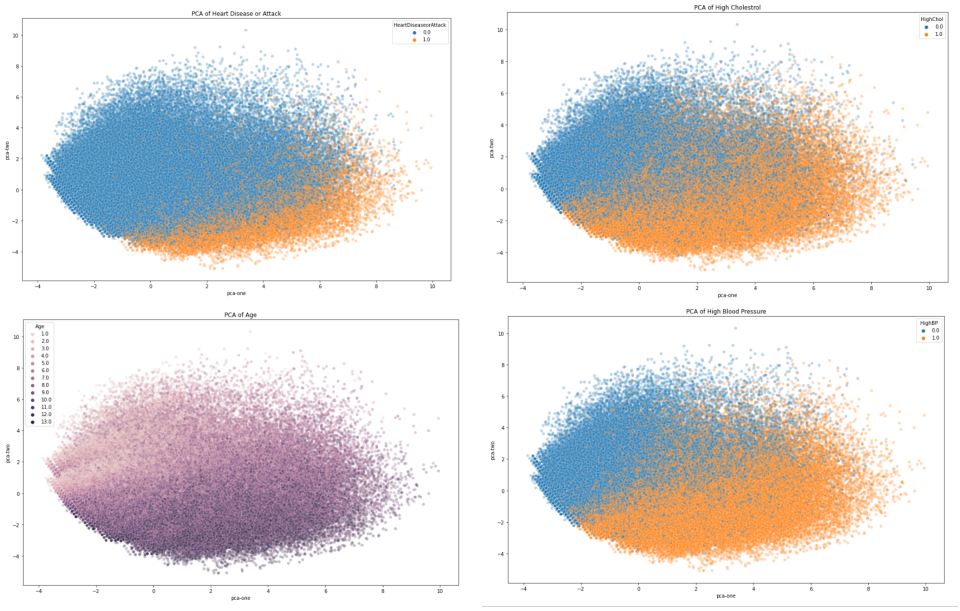

Figure 8: PCA plots

10 Feature Selection

We want to find the most important predictor variables (of features) that explain major part of variance of the response variable.To build high performing models of heart disease attack we can look at the correlation matrix heat map first to identify some important or highly related features.

10.1 Relative Importance

By using the relaimpo method we can put the relative importance of variables into linear regression models to be determined as a relative percentage. The highest three features are Genhlth(0.1601), Stroke(0.1514) and Age(0.1450). The result is consistent with the heatmap since these three features are relatively in deeper color of green compared to other features.

10.2 MARS

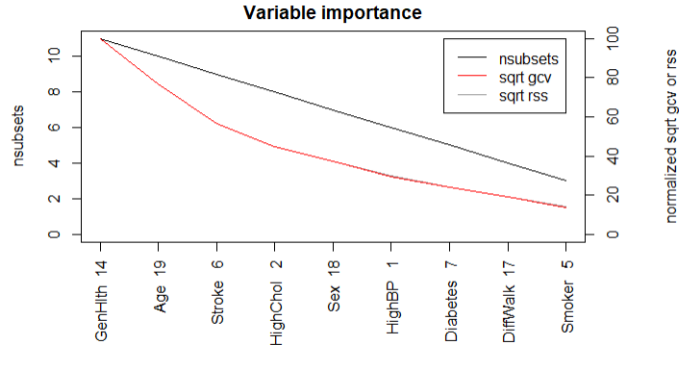

Figure 9: Variable Importance

We can implement variable importance based on Generalized cross validation (GCV), number of subset models the variable occurs (nsubsets) and residual sum of squares (RSS). The highest variable importance is still Genhlth which is consistent with the first method.

References

[1] Teboul, Alex. "Heart Disease Health Indicators Dataset." *Kaggle*, 10 Mar. 2022, https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset?resource=download.

[2] Centers for Disease Control and Prevention. *"Defining Adult Overweight & Obesity."* Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 7 June 2021, https://www.cdc.gov/obesity/basics/adult-defining.html.