BIOS 735 Project Report

Analysis of Personal Key Indicators

of Heart Disease

Group 2 - Mingwei, Di, Wanting, Eunchong, Andrew

April 27, 2022

Abstract

In 2020, heart disease was the leading cause of death in the United States with 696,962 deaths attributed (followed by cancer & COVID-19) according to the final 2020 U.S. mortality data from the CDC. Controlling for population, there were 168.2 deaths attributed to heart disease per 100,000 individuals in the United States[1]. Given the seriousness of this cause of death, it is imperative that we understand the underlying factors that contribute to increased risk of heart disease which leads to reduced quality of life and increased risk of mortality. Binary outcome statistical models were fit to estimate the relationship between heart disease occurrence and relevant clinical & demographic covariates. The greatest contributing risk factors for heart disease are old age, stroke occurrence, diabetes, and poor general health. These statistical models achieved a 99.2% specificity (true negative) rate, but only a 10.8% sensitivity (true positive) rate which corresponds to a 54.9% positive predictive value such that the chances of correctly classifying a case of heart disease are only slightly better than a coin flip. Thus, it is clear that it is challenging to predict positive cases of heart disease in part to its relatively low occurrence rate in the greater population.

1 Introduction

1.1 Background

According to the CDC, heart disease is one of the leading causes of death for people of most races in the United States (African Americans, American Indians and Alaska Natives, and white people) and it is the

leading cause of death overall with nearly 700,000 heart disease-related deaths occurring in 2020[1]. Many health status indicators are found related to heart disease such as high blood pressure, high cholesterol, smoking, diabetes, obesity (BMI > 30.0), insufficient physical activity, & excessive alcohol consumption. Understanding and detecting the factors that have the greatest impact on heart disease occurrence in populations is crucial in healthcare to improve the length and quality of life.

1.2 Aims

Considering that heart disease is the leading cause of death in the United States, we sought to investigate the relationship of the occurrence of heart disease with other relevant demographic and clinical factors available in the CDC annual survey data. The aims of this study were (a) to investigate the importance of individual demographic & clinical covariates in estimating the occurrence of heart disease, (b) estimate the relationship between heart disease and body mass index (BMI), (c) estimate the relationship between heart disease and/or stroke and risk behaviors like smoking and alcohol consumption, and (d) compare the model fitting performance between parametric methods (logistic regression) and other machine learning methods for estimating the occurrence heart disease.

1.3 Data Description

The particular data employed in this study comes from the 2020 CDC annual survey data. The dataset contains 319,795 total observations of 18 recorded variables. For our study, the outcome of interest is the occurrence of heart disease and 17 key indicators for health status are included as potential predictors. Lending to the fact that heart disease is a relatively rare ailment among the greater population, the classes of the heart disease factor in the data are significantly unbalanced. To be exact, 292,422 (91.4%) of the observations do not have heart disease and only 27,373 (8.6%) of the observations do have heart disease. In this study, the key outcomes of interest are the occurrence of heart disease and stroke. Additional covariates to be considered in analysis include: body mass index (BMI), smoking status, alcohol consumption status, days of poor physical health, days of poor mental health, difficulty walking, sex, age category, race, diabetes status, physical activity, general health classification, average sleep time, asthma, kidney disease, and skin cancer.

1.4 Exploratory Visualizations

Prior to carrying out any model building or prediction on the heart disease data, we are interested in investigating some exploratory figures to gain an understanding of some patterns that exist in the data and if there are any particularly interesting factors to consider. In Figure 1, we see that the relative proportion of heart disease occurrence among males (10.6%) is slightly larger than the proportion of heart disease occurrence in females (6.7%). It is also made clear here that the overall proportion of positive heart disease cases among all observations is quite small as well so we may be troubled by the unbalanced occurrence of heart disease with further modeling tasks, particularly with model sensitivity in prediction.

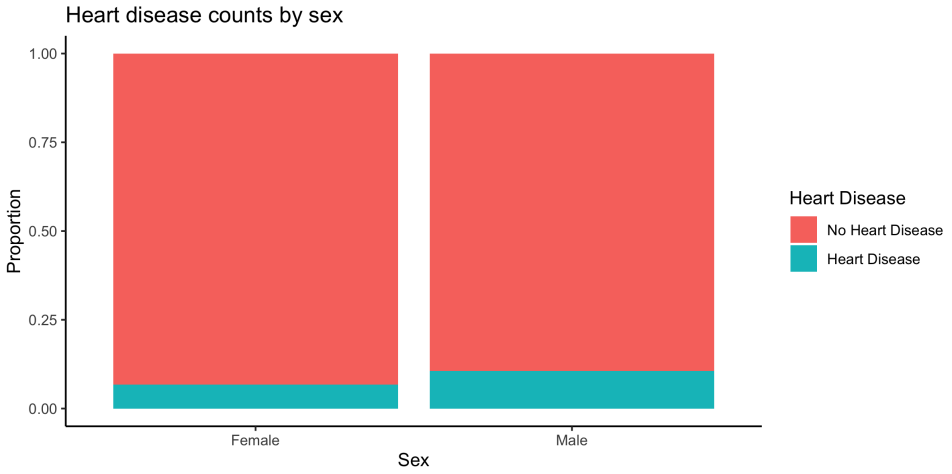

Figure 1: Proportion of male and female subjects with heart disease occurrence

In Figure 2, we visualize the relative proportions of heart disease occurrence among the different age categories recorded in the data while also controlling for sex. The age categories included range from $18-24$ up to 80 or older. Most all categories within the lower and upper bounds have age ranges of five years. We can see a clear trend here that heart disease occurrence increases with increasing age among both males and females. For example, only 0.4% of all subjects in the $18-24$ age group have heart disease but nearly 20% of subjects in the 80 of older group have heart disease. This is nearly double that of the 8.6% heart disease occurrence proportion found in the full study sample. In fact, this trend is monotonic as each older age group has a greater recorded proportion of heart disease occurrence compared to the previous age group. We also see that males have a greater rate of heart disease occurrence compared to females at all of the available age groupings. Hence, we note that old age and being male appear to be

related to a higher risk of having heart disease.

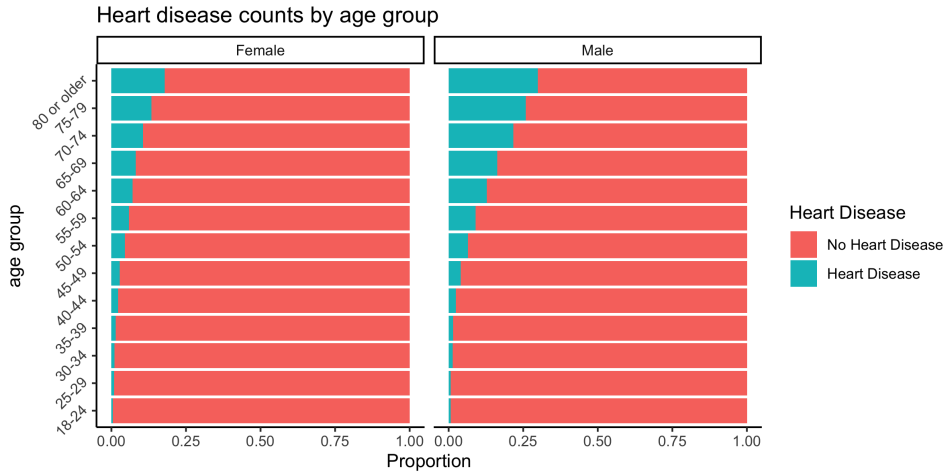

Figure 2: Proportion of subjects with heart disease in each age category, stratified by sex

The heart disease dataset also includes information on what may be considered "risky" or "unhealthy" habits. These are the consumption of all and smoking. In Figure 3, the proportion of heart disease occurrence is displayed for subjects who are smokers, individuals who have smoked at least 100 cigarettes in their lifetime, and non-smokers as well as for heavy drinkers, men who have at least 14 drinks per week and women who have at least 7 drinks per week, and those who aren't heavy drinkers. In terms of smoking, approximately 12% of individuals classified as smokers have heart disease and only 6% of non-smokers have heart disease. Hence the risk of heart disease for smokers is doubled relative to non-smokers. Furthermore, it is interesting to note that only 5.2% of heavy drinkers have heart disease but almost 9% of non-heavy drinkers have heart disease. In all cases, we see that heart disease occurrence is still relatively low in general, but it is interesting to see that smokers appear to have a greater risk yet heavy drinkers have a reduced risk of heart disease. One could suggest that this may be due to any number of confounders (ie. heavy drinkers may tend to be younger).

Figure 4 displays the distribution of subject BMI measurements according to their heart disease status. For subjects with heart disease, the average BMI is 29.4, with a median of 28.34, and standard deviation of 6.58. The minimum BMI is 12.21 and the maximum value is 83.33 with the middle 50% ranging from 25.06 to 32.69. On the other hand, for subjects without heart disease, the average BMI is 28.22, with a median of 27.26, and standard deviation of 6.33. The minimum BMI is 12.02 and the maximum value is 94.85 with the middle 50% ranging from 23.89 to 31.32. It is interesting to note that

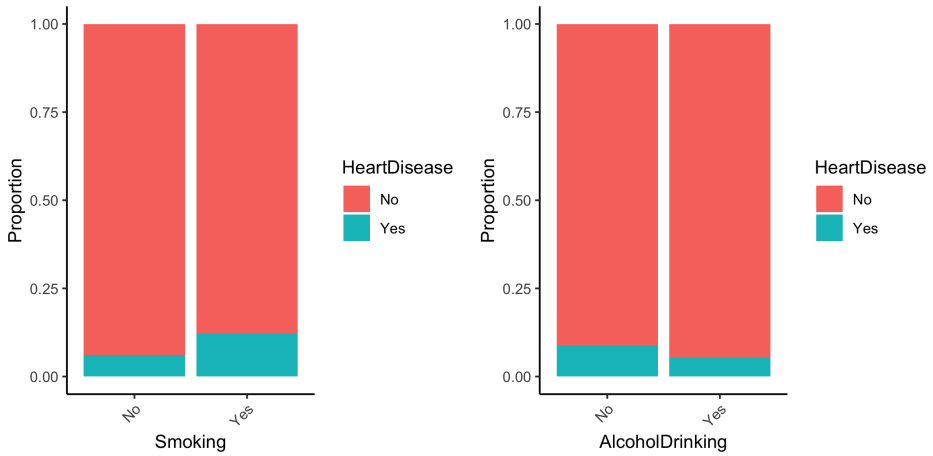

Figure 3: Proportion of subjects with heart disease based on smoking & drinking status

the mean and median BMI measurements in each heart disease classification are quite similar to each other, but they are both greater for the grouping of subjects with heart disease. Hence, a higher BMI may potentially be a risk factor for heart disease.

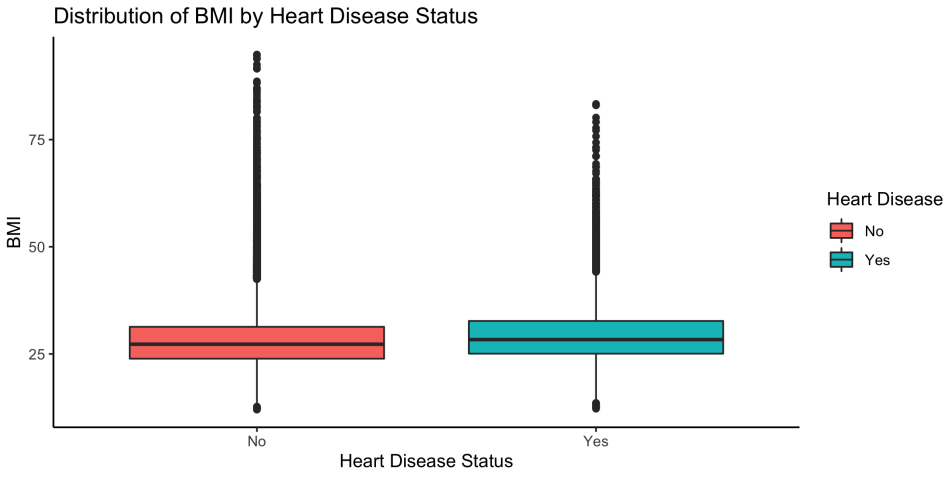

Figure 4: Distribution of BMI measurements for subjects with and without heart disease

Finally, in Figure 5 examine the relative occurrence of heart disease among diabetic classification and age groups. Previously, we noted that the proportion of subjects with heart disease increased with old age which is clearly still the case when we stratify by diabetic class. Furthermore, nearly 22% of diabetic subjects have heart disease, about 12% of subjects with borderline diabetes have heart disease, only 4% of subjects who had diabetes during pregnancy have heart disease, and about 6.5% of subjects with no diabetes whatsoever have heart disease. Hence, we note that the risk of heart disease for a patient with diabetes is over three times the risk for a patient without diabetes.

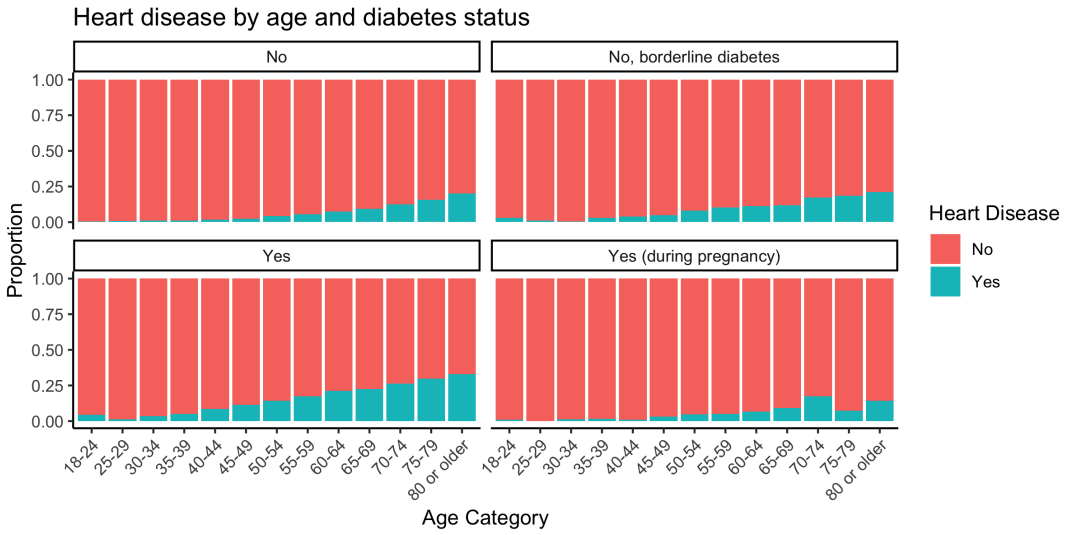

Figure 5: Proportion of subjects with heart disease in each age category, stratified by diabetic status

2 Methods

2.1 Data Preprocessing

All pre-processing and data analysis are performed with R statistical computing software. Packages utilized include: tidyverse, caret, ROCR, optimx, MASS, Rcpp, and RcppArmadillo. The "Personal key indicators of heart disease" dataset contains 319,795 observations of 18 variables as previously described in the data description. Prior to any analysis being performed on the data, a few preprocessing steps were taken. First, all binary variables or categorical/ordinal variables were converted to factors in R. Additionally, numeric variables were converted to numerics in R. Finally, balanced test & train datasets were generated with equal proportions of positive & negative heart disease cases for use in subsequent analyses.

2.2 Decision Tree

Decision trees are a non-parametric supervised machine learning method that are used for classification problems (as well as regression). They operate by learning decision rules that best separates the provided data into classes based on features of the data. Decision trees are valuable because they are quite easy to interpret and they also yield informative visualizations that depict the branching structure of the tree. At each "node" of a decision tree, the Gini impurity measure is used for splitting the data. The probability of an observation being correctly classified based on the given splitting features is provided

as well as the relative percentage of the overall data being considered at the particular node or splitting point. For example, the root node considers 100% of the data and the probability of correct classification is simply based on the proportion of positive to negative cases for a binary decision problem. Based on the heart disease data, we fit a decision tree with heart disease occurrence as the outcome to be predicted and all other covariates as predictors in the decision tree model. The decision tree was fit on the training data and predictions for heart disease (yes/no) were made on the observations in the test dataset. Due to the high degree of imbalance in the data, we increased the relative occurrence of positive heart disease cases that the model would be trained on from only about 9% in the full data to roughly 25% or a 3 to 1 ratio of negative to positive cases (75000 negative:21899 positive). This adjustment was made in order to prevent the model from simply predicting that all observations are negative cases.

2.3 Logistic Regression

To address the classification problem, one simple way is to use logistic regression. In the logistic regression model, we assume that,

$$y_i|\boldsymbol{x}_i \sim Bernoulli(\pi_i),$$

and that,

$$\text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

for $i = 1, \cdots, n$. Then the log-likelihood function for solving β can be expressed as

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i^T\boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \right\}$$

We can solve for $\boldsymbol{\beta}$ in this expression by maximizing the log-likelihood function. To achieve this, we used IRLS and BFGS methods.

2.3.1 IRLS

IRLS that is derived from Newton-Raphson (NR) is considered as the method for maximizing the log-likelihood. Its extremely fast convergence is one of the advantages. Also, the first order partial derivative of the log-likelihood and the Fisher information are known for the logistic regression. However, it is relatively more sensitive to the starting value compared to BFGS method.

The above log-likelihood can be written in the matrix form as

$$l_n(\beta) = Y^T X\beta - b(\beta)^T J_n, \tag{1}$$

where $Y = (y_1, ..., y_n)^T$, $X = (x_1^T, ..., x_n^T)^T$, $\beta = (\beta_0, ...\beta_p)^T$, $b(\beta)^T = (log(1 + exp(x_1\beta)), ..., log(1 + exp(x_n\beta))$, and $J_n = (1, ..., 1)$.

The partial derivative of the log-likelihood is $l'(\beta) = X^T(Y-\mu)$, where $\mu = (\frac{exp(x_1\beta)}{1+exp(x_1\beta)}, ..., \frac{exp(x_n\beta)}{1+exp(x_n\beta)})^T$, and the Fisher information is $I_Y(\beta) = E(-l''(\beta)) = X^T W X$, where $W = diag(\frac{exp(x_1\beta)}{(1+exp(x_1\beta))^2}, ..., \frac{exp(x_n\beta)}{(1+exp(x_n\beta))^2})$.

The NR algorithm is

$$\beta^{k+1} = \beta^k + (X^T W^k X)^{-1} X^T (Y - \mu^k). \tag{2}$$

From the NR algorithm, one can derive IRLS algorithm for β^{k+1} such that

$$\beta^{k+1} = (X^T W^k X)^{-1} X^T W^k z^T, \tag{3}$$

where a response vector is $Y = (Y_1, ..., Y_n)$, a design matrix is X having X_i as row i, $z = (z_1, ..., z_n)^T$ with $z_i = x_i\beta + e_i = x_i\beta + \frac{(1+exp(x_i\beta))^2}{exp(x_i\beta)}\left(y_i - \frac{exp(x_i\beta)}{1+exp(x_i\beta)}\right)$. The algorithm goes as follows:

Algorithm 1 IRLS

Require: Given response y_i, predictors vector x_i for $i = 1, ..., n$, starting point β_0, and convergence tolerance $\epsilon > 0$;

$k \leftarrow 0$;

while $|l(\beta^k) - l(\beta^{k+1})| > \epsilon$ **do**

for $i = 1, ..., n$ **do**

$z_i^k = x_i\beta^k + e_i = x_i\beta^k + \frac{(1+exp(x_i\beta^k))^2}{exp(x_i\beta^k)}\left(y_i - \frac{exp(x_i\beta^k)}{1+exp(x_i\beta^k)}\right)$

$w_i^k = \frac{exp(x_i\beta^k)}{(1+exp(x_i\beta^k))^2}$

end for

put $X = (x_1^T, ..., x_n^T)^T$, $W^k = diag(w_1^k, ..., w_n^k)$ and $z^k = (z_1^k, ..., z_n^k)^T$

$\beta^{k+1} = (X^T W^k X)^{-1} X^T W^k z^k$

$k \leftarrow k + 1$;

end while

2.3.2 BFGS

Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is an optimization method for multidimensional nonlinear unconstrained functions. It improves the speed by avoiding the calculation of the inverse of Hessian. Since our goal is to maximize the log-likelihood, we take the opposite log-likelihood as our objective function, i.e. $f(\beta) = -l(\beta)$ and $\nabla f(\beta) = -l'(\beta)$. The algorithm goes as follows:

Algorithm 2 BFGS

Require: Given starting point β_0, convergence tolerance $\epsilon > 0$, and inverse Hessian approximation

$H_0 = I$;

$k \leftarrow 0$

while $|f(\beta^k) - f(\beta^{k+1})| > \epsilon$ **do**

$d_k = -H_k \nabla f(\beta^k)$; ▷ Compute the search direction

$\beta^{k+1} = \beta^k + a_k d_k$ ▷ where a_k is computed from a line search procedure to satisfy the Wolfe conditions

$s_k = \beta^{k+1} - \beta^k$

$y_k = \nabla f(\beta^{k+1}) - \nabla f(\beta^k)$

$H_{k+1} = \left(I - \frac{s_k y_k^T}{sk^T y_k}\right) H_k \left(I - \frac{y_k s_k^T}{sk^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}$ ▷ Update inverse Hessian approximation

$k \leftarrow k + 1$;

end while

The line search approach first finds a descent direction along which the objective function $f(\beta)$ will be reduced and then computes a step size that determines how far β should move along that direction. Here, we consider a line search procedure to satisfy the Wolfe conditions:

$$f(\beta^k + a_k d_k) \leq f(\beta^k) + \rho a_k dk^T \nabla f(\beta^k) \tag{4}$$

$$dk^T \nabla f(\beta^k + a_k d_k) \geq \sigma dk^T \nabla f(\beta^k) \tag{5}$$

The algorithm for the line search satisfying Wolfe conditions is provided in **Algorithm 3**.

Algorithm 3 Line search satisfying Wolfe Conditions

Require: Set $\rho = 10^{-}4$, $\sigma = 0.9$,$a = 0$ $a_k = 1$ and $b = N$

while $a < b$ **do**

if $f(\beta^k + a_k d_k) > f(\beta^k) + \rho a_k d_k^T \nabla f(\beta^k)$ **then**

set $b = a_k$ and $a_k \leftarrow (a + b)/2$

else if $d_k^T \nabla f(\beta^k + a_k d_k) < \sigma d_k^T \nabla f(\beta^k)$ **then**

set $a \leftarrow a_k$

if b = N **then** $a_k \leftarrow 2a$

else $a_k \leftarrow (a + b)/2$

end if

end if

end while

2.3.3 GLM

We also fit a logistic regression model based on (2.3) with heart disease occurrence as the outcome and all covariates in the data as predictors. These models were fit via the "stock" R 'glm()' function. Based on this trained model, we made predictions of heart disease classification for the subjects in the test dataset to more directly assess model performance and viability.

Additional models were fit in consideration of the relationship between heart disease and BMI measurements, the relationship between heart disease and smoking & drinking, and the relationship between stroke occurrence and smoking & drinking. Stepwise variable selection methods were also considered to determine if any particular subset of covariates in the data lead to a more preferable model for binary classification.

2.4 R package functions & Implementation

The package named "glmLogistic" is comprised of total five functions which are "loglik", "d1.loglik", "beta.updater", "optim.IRLS", and "optim.BFGS". "loglik" function is used for calculating the value of the log-likelihood and "d1.loglik" function is used for the first derivative of the log-likelihood. Lastly, "beta.updater" computes β^{k+1} for IRLS algorithm. All functions take a design matrix with numeric values, a response vector with zero and one, and the starting value of β.

"optim.IRLS" and "optim.BFGS" are the main functions to calculate the estimates of the parameters for the logistic regression model. "optim.IRLS" executes IRLS algorithm and uses the absolute change of the log-likelihood for the convergence criterion. The function puts the estimates of the parameters, its standard error, the value of the log-likelihood, the number of iterations, and the last absolute change of the log-likelihood as a result. If the iteration does not converge, the iteration stops and warning message shows up.

2.5 Random Forest & Support Vector Machine

We fit a selection of machine learning models over the same training data set as previously described with caret package from R. In order to account for the unbalanced nature of the heart disease outcome in the data, we down-sampled the "Non-Heart Disease" cases from training data set and obtained a subset of observations with equal number of "Heart Disease" and "Non-Heart Disease" cases. For the random forest model, we used the ranger package in R. The Gini impurity measure is used for splitting the data and 5-fold cross validation is employed to tune the parameters. We tuned the parameter 'mtry' over a grid from 2 to 30. We also fit a linear kernel support vector machine model with 'tunelength = 10'. All numeric variables were centered and scaled prior to model fitting. The prediction accuracy and kappa value of the optimal fitted model based on the test data are reported. We also generated the receiver operating characteristic curve of sensitivity vs. 1-Specificity as well the AUC value ("area under curve") to evaluate the predictive performance of the machine learning model as well as its ability to correctly classify positive cases of heart disease.

3 Results

3.1 Decision Tree

The fitted decision tree used age category, general health category, and sex as variables in the final tree construction as seen in Figure 6. In general, the tree indicates that subjects who are at least 60 years old and in worse than average health are more likely to have heart disease (59%) and that subjects less than 60 years of age and in relatively good health are less likely to have heart disease ($9\% - 21\%$). Furthermore, predictions on the test data set yielded an overall accuracy of 0.866, sensitivity of 0.420, and specificity of

0.908. Hence, the model has good performance in general, but unfortunately the probability of correctly predicting positive cases is essentially a toss-up.

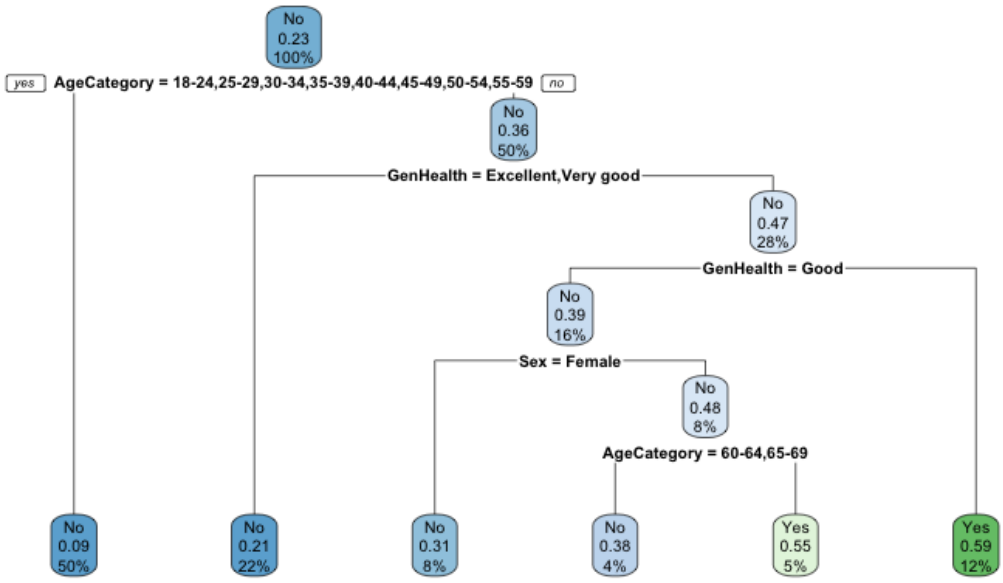

Figure 6: Decision tree classification of heart disease occurrence with age, sex, and general health

3.2 Logistic Regression (IRLS & BFGS)

The design matrix, X, was created by using a reference cell coding. The first level of the categorical variables are the reference groups, and the continuous variables are used without standardization. The response vector, Y, is comprised of zeroes and ones with an entry of 1 indicating the occurrence of heart disease. The starting value for β was the zero vector. The estimates of the parameters are shown in Table 1 by rounding up to four decimal places. The estimates of the parameters from the functions "optim.IRLS", and "optim.BFGS" agree well with the estimates from the glm function.

3.3 Logistic Regression (GLM)

For a logistic regression model fit with all covariates as predictors for the occurrence of heart disease, the trained model includes 38 total coefficient estimates (including an intercept). For each coefficient, β, a unit change in the value of a continuous covariate or the presence of a binary/categorical covariate yields an $\exp(\beta)$-fold multiplicative change in the odds of heart disease occurrence where the odds is the

Parameter	Estimate by function			Parameter	Estimate by function		
	glm	IRLS	BFGS		glm	IRLS	BFGS
Intercept	-6.3586	-6.3586	-6.3586	Race			
BMI	0.0086	0.0086	0.0086	Asian	-0.4976	-0.4976	-0.4977
Smoking: Yes	0.3546	0.3546	0.3546	Black	-0.2774	-0.2774	-0.2774
Alcohol: Yes	-0.2371	-0.2371	-0.2371	Hispanic	-0.2254	-0.2254	-0.2254
Stroke: Yes	1.0578	1.0578	1.0578	Other	-0.0125	-0.0125	-0.0124
PhysicalHealth	0.0022	0.0022	0.0022	White	-0.0221	-0.0221	-0.0221
MentalHealth	0.0051	0.0051	0.0051	Diabetic			
DiffWalking: Yes	0.2121	0.2121	0.2121	Borderline diabetes	0.1683	0.1683	0.1683
Sex: Male	0.7036	0.7036	0.7036	Yes	0.4924	0.4924	0.4924
Age Category				Yes (pregnancy)	0.1356	0.1356	0.1357
25-29	0.1483	0.1483	0.1483				
30-34	0.5313	0.5313	0.5313	PhysicalActivity: Yes	0.0203	0.0203	0.0203
35-39	0.5324	0.5324	0.5324	General Health			
40-44	1.0137	1.0137	1.0137	Fair	1.5346	1.5346	1.5346
45-49	1.3002	1.3002	1.3002	Good	1.0598	1.0598	1.0598
50-54	1.711	1.711	1.711	Poor	1.9333	1.9333	1.9333
55-59	1.9758	1.9758	1.9758	Very good	0.4788	0.4788	0.4788
60-64	2.2102	2.2102	2.2103				
65-69	2.4665	2.4665	2.4666	SleepTime	-0.0247	-0.0247	-0.0247
70-74	2.751	2.751	2.751	Asthma: Yes	0.2759	0.2759	0.2759
75-79	2.9461	2.9461	2.9461	Kidney Disease: Yes	0.5823	0.5823	0.5823
80 or older	3.1999	3.1999	3.2	Skin Cancer: Yes	0.1203	0.1203	0.1203

Table 1: Logistic regression parameter estimates for estimation of heart disease occurrence

ratio of the probability of having heart disease divided by the probability of not having heart disease. This can also be expressed as a β-fold multiplicative change in the log-odds of heart disease occurrence. Some estimated coefficients (β) of note include: BMI = 0.0086, Smoking = 0.355, Alcohol = -0.237, Age (80+) = 3.20, and Diabetes (yes) = 0.492.

The trained model had a test data prediction accuracy level of $0.916(0.9139, 0.9182)$, Cohen's kappa of 0.157, sensitivity of 0.109, specificity of 0.992, and positive predictive value of 0.545. Figure 7 depicts the model true positive prediction rate versus the false positive prediction rate. The optimal curve should "hug" the upper left corner, indicating a high true positive rate and low false positive rate. In this case, the model prediction performance yielded an AUC value of 0.838 which is excellent. For reference, an AUC value of 0.5 suggests that the classifier has no ability to correctly classify positive and negative cases among observations.

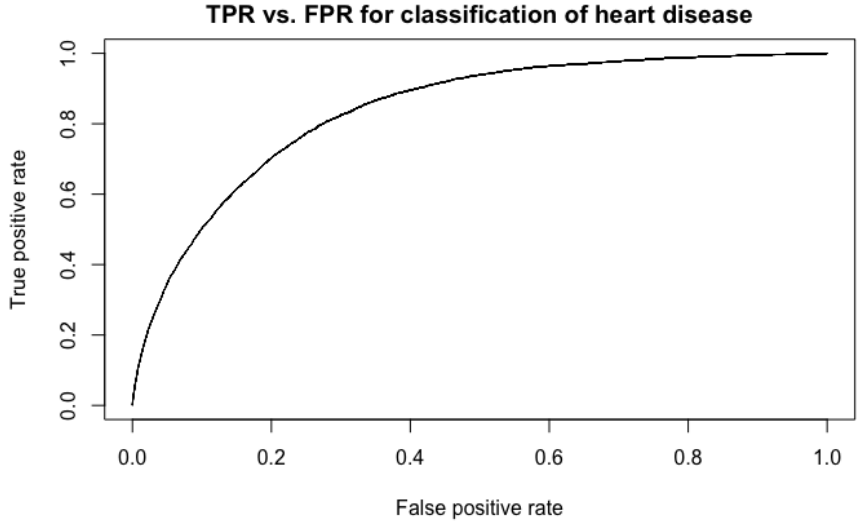

Figure 7: Receiver operating characteristic curve for classification of heart disease with GLM

Unfortunately, the models we fit with BMI and smoking & drinking as predictors yielded incredibly poor fits. Furthermore, all variable selection efforts (backward & forward) based on model AIC scores yielded the original full set of covariates as the chosen subset of predictors. Therefore, the results for these models match those for the model with all available covariates as predictors that was initially described.

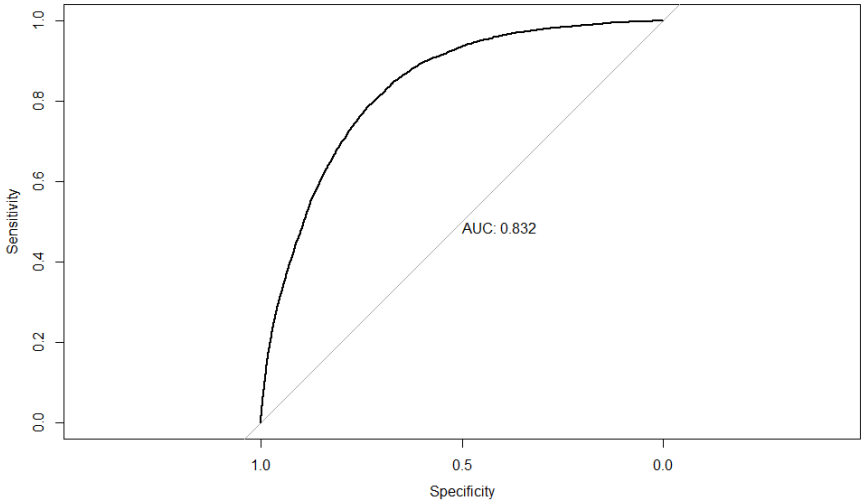

Figure 8: Receiver operating characteristic curve for classification of heart disease with random forest

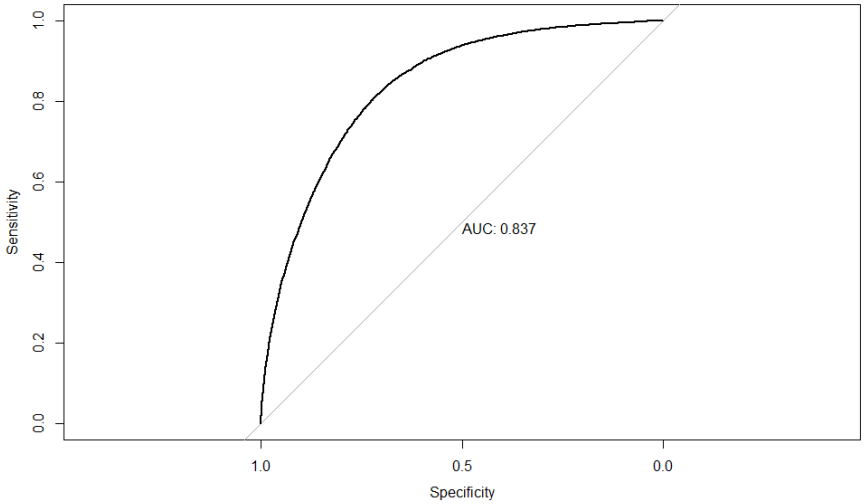

Figure 9: Receiver operating characteristic curve for classification of heart disease with support vector machine

3.4 Random Forest & Support Vector Machine

For the Random Forest model, the optimal trained model uses mtry = 6 and min.node.size = 20. The prediction accuracy of the trained model on the testing dataset is 0.727 (0.724, 0.731) with a kappa value of 0.229. The receiver operating characteristic (ROC) curve is shown in Figure 8 where the model prediction performance yielded an AUC value of 0.832, which is quite good.

For the linear kernel support vector machine model, The prediction accuracy of the trained model on the test data is 0.753 (0.750, 0.757) with a kappa value 0.248. The ROC curve for the random forest

model predictions is shown in Figure 9 where the model prediction performance yielded an AUC value of 0.837, which is excellent. A plot of the estimated top 10 most important variables from the SVM model is shown in Figure 10. Note that this includes factors such as: difficulty walking, physical health, diabetes, smoking, general health, physical activity, age, stroke occurrence, sex, and BMI.

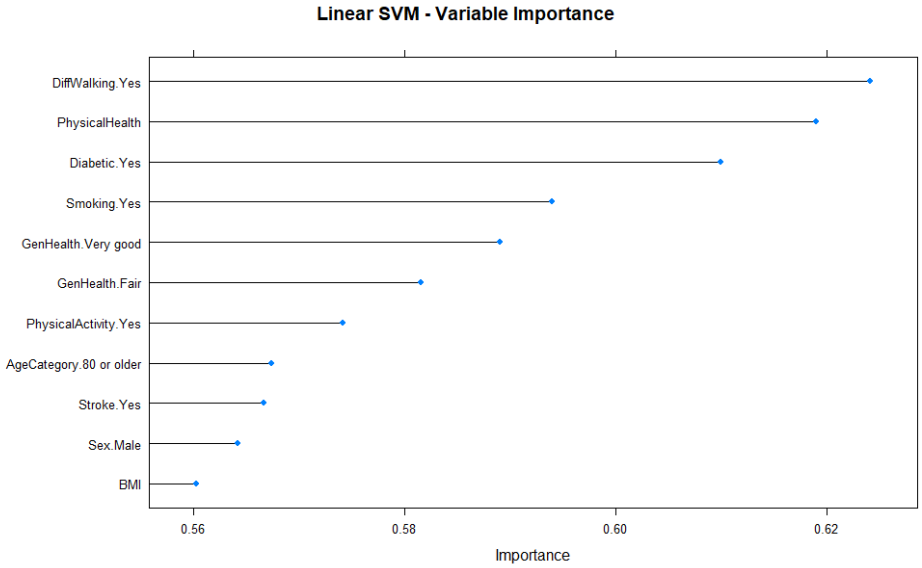

Figure 10: Variable importance (top 10) for support vector machine

4 Conclusion

In an effort to understand the factors that influence the occurrence of heart disease, we fit the following models: logistic regression, decision tree, random forest, and support vector machine. For logistic regression, we trained models with the data likelihood function as the objective function and optimized with IRLS and BFGS to obtain regression parameter estimates. We also fit logistic regression models based on existing packages in R. A comparison of the estimates for the R package we developed and estimates based on the existing 'glm()' function show that the estimates are nearly identical to each other. Hence, our "built-from-scratch" method is quite effective. For logistic regression, our model predictions obtained an accuracy level of 0.916 but a sensitivity level of only 0.109. For the other models we trained, accuracy levels were found to be as high as 0.866 with a maximum sensitivity of 0.420. We also found that some of the most influential factors in estimating the occurrence of heart disease are (among others): general health, age, smoking status, diabetic status, and sex. Referencing our stated aims, BMI has a marginal

impact on the occurrence of heart disease (1.09-fold change in odds of heart disease for each unit increase of BMI), smoking doubles the risk of heart disease but heavy drinking actually reduces the risk of heart disease (with other factors held constant), and these same relationships are found when stroke occurrence is considered as the outcome.

5 Discussion

5.1 Clinical relevance of results

Our findings from this analysis have substantial clinical relevance regarding subject characteristics that show certain groups of individuals are at greater risk of heart disease relative to others. Attaining a strong understanding of the factors that are related to heart disease can lead to targeted surveillance efforts by public health agencies at all levels that address groups of individuals with specific characteristics that we identified to be significant risk factors of the presence of heart disease. While these factors are not necessarily useful in a diagnostic approach, they may be valuable for seeking out individuals with these high risk characteristics (like old age, high BMI, diabetes, etc.) in order to ensure they continue to receive appropriate medical attention.

Furthermore, public health initiatives may also be funded and/or promoted to encourage healthy behaviors based on the evidence from our analysis that suggests habits like smoking nearly double the risk of having heart disease when compared to non-smokers. In general, we also found that individuals in worse than average general health are at a substantially higher risk of heart disease relative to subjects in good health.

5.2 Limitations & Concerns

The results of our analysis provide interesting insights into the relationship between clinical & demographic factors and the occurrence of heart disease. While we developed a number of informative models that tended to have quite good accuracy levels ($> 90\%$) in aggregate, they are not without some concerns that should be taken seriously when evaluating the results of our trained models and prediction performance on test data. For one, we mentioned previously that the outcome "heart disease" is significantly unbalanced between positive ($\approx 9\%$) and negative ($\approx 91\%$) cases. This characteristic led our models to

struggle with prediction sensitivity, where the probability of classifying positive cases as false negatives is quite high. The positive predictive value of our models is typically near or slightly greater than 50% such that the probability of correctly classifying positive cases is essentially just as good as random chance.

Furthermore, we are concerned that the presentation of the data as mostly categorical/binary predictors may reduce the granularity at which the models may "learn" features in the data in order to make accurate classifications. For example the age factor is grouped into age ranges of 5 years rather than having measurements reported on a continuous scale where age is reported as a whole number value. If more of the factors were presented on a continuous scale (quantity of smoking, quantity of drinking, etc.), we might be able to discern a more direct relationship between the magnitude of a health condition or habit and the occurrence of heart disease.

Finally, concerns in the interpretation of some estimated parameters exist. For example, we found that being a heavy drinker leads to an estimated 1.27-fold multiplicative increase in the odds of *not* having heart disease. While excessive alcohol consumption is often thought of as an unhealthy habit, it may be possible that there is a true relationship with the reduction of heart disease. However, we wonder if there may be other confounders such as age. In this case, heavy drinkers tend to fall into younger age groups which we already found to be associated with lower risk of heart disease.

5.3 GitHub Repository

The GitHub repository for this project can be accessed Here

References

[1] Sherry L Murphy, Kenneth D Kochanek, Jiaquan Xu, and Elizabeth Arias. Mortality in the united states, 2020. 2021.