Final-Report

Alex Shen, Steven Yuan, Conner Byrd

12/13/2021

I. Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. CVDs include a range of conditions that include blood vessel disease, such as coronary artery disease; heart rhythm problems (arrhythmias); heart defects at birth (congenital heart defects); heart valve disease; disease of the heart muscle; heart infections, and many more. Although many forms of CVD can be prevented or treated with healthy lifestyle choices, some can not. CVDs are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 32% of all global deaths. Over 85% of deaths from CVD were due to heart attack and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Cardiovascular diseases (CVDs) are the number one cause of death globally.

Currently, there are several different ways for physicians to diagnose patients that they believe to be at risk for Cardiovascular Diseases (CVDs). The practices vary by country, but often include the physician checking the patient's blood pressure, cholesterol level, and conducting further tests such as exercise stress tests, X-rays, etc. Currently, there are many issues with the current diagnostic methods. A [study] (https://www.sciencedaily.com/releases/2009/11/091116103435.htm) of 500 patients found a false positive reading between 77 and 82 percent in patients in patients at risk of CVD screened by ECG, and a false negative reading between 6 to 7 percent in the same patient population . People with CVDs or who are at high risk of CVDs need early detection and management wherein a machine learning model can be of great help.

Using our Cardiovascular Heart Disease data, we have two main goals. Our first goal is to create models for the purpose of prediction. These can be used to assess the likelihood of a heart disease diagnostic for potential at-risk patients based on a number of factors such as age of the and sex of the patient, blood pressure, cholesterol, heart rate, and the presence of chest pain.

Our second goal is to create models for the purpose of interpretation, which can be used to provide a greater understanding of signs that at-risk patients can analyze to check their risk for CVDs.

We chose to fit 3 different models to classify whether a patient has heart disease or not. The first model is a logistic regression model with variable selection performed by a lasso regression. We decided to use this model because of its interpretability so that we can examine the relationship our predictors have with the probability of a patient having heart disease. We also decided to use a random forest and a SVM because of their ability to perform classification and due to their predictive power despite their lack of interpretability. We will use a 10-fold CV to determine the best predictive model based on the classification error of each model. Due to the ability of ensemble models to reduce prediction error, we will also create an ensemble model where the prediction is the most common result of the 3 individual models (https://www.sciencedirect.com/topics/computer-science/ensemble-modeling). Finally, we compare the 10-fold CV errors for each individual model as well as the ensemble to find the one with lowest classification error and therefore highest predictive accuracy.

II. Data

We obtained our data from Kaggle (https://www.kaggle.com/fedesoriano/heart-failure-prediction). The dataset was originally provided by Dr. David Aha, a researcher at the US Naval Research Laboratory. It was

created by combining five heart datasets from Cleveland (303 observations), Hungary (294 observations), Switzerland (123 observations), Long Beach, Virginia (200 observations), and Stalog (270 observations), for 918 unique observations. This makes it one of the largest available heart datasets with multinational data.

The dataset has 11 predictor variables, which are listed below along with their descriptions. Out of the 11 predictor variables, five of them are categorical variables (Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope), while the rest are numeric. The categorical variables have been factored for this report. For our purposes, the dataset has a single response variable, `heartDisease`, described further below, which is whether or not the patient has been diagnosed with heart disease.

Variables	Description	Value
Age	Age of the patient in years	Numeric Value
Sex	Sex of the patient	[M/F]
ChestPainType	Chest pain type	[TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	Resting blood pressure	[mm Hg]
Cholesterol	Serum cholesterol	[mm/dl]
FastingBS	Fasting blood sugar	[1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	Resting electrocardiogram results	[Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy]
MaxHR	Maximum heart rate achieved	[Numeric value between 60 and 202]
ExerciseAngina	Exercise-induced angina	[Y: Yes, N: No]
ST_Slope	the slope of the peak exercise ST segment	[Up: upsloping, Flat: flat, Down: downsloping]
Oldpeak	The level of exercise relative to rest	Numeric value
HeartDisease	Output class denoting if patient has Heart Disease	[1: heart disease, 0: Normal]

Missing Data

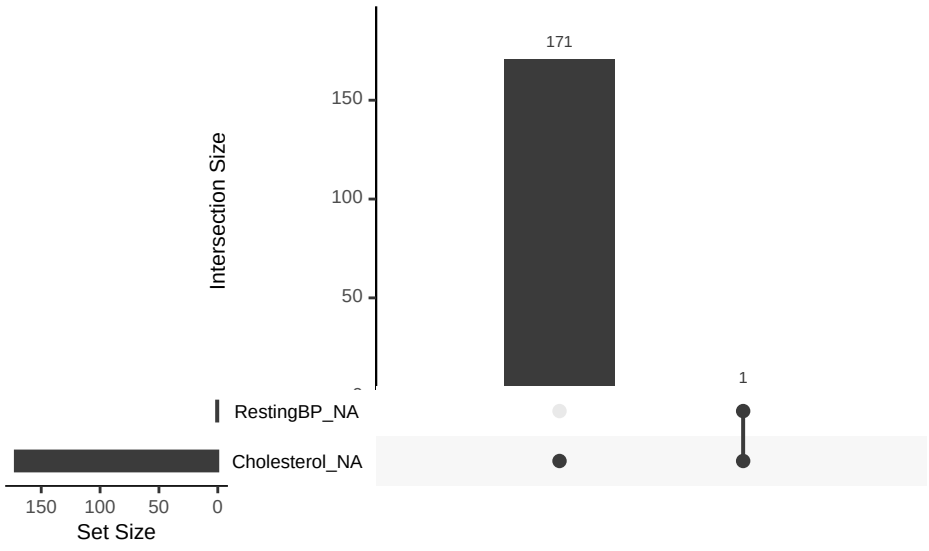


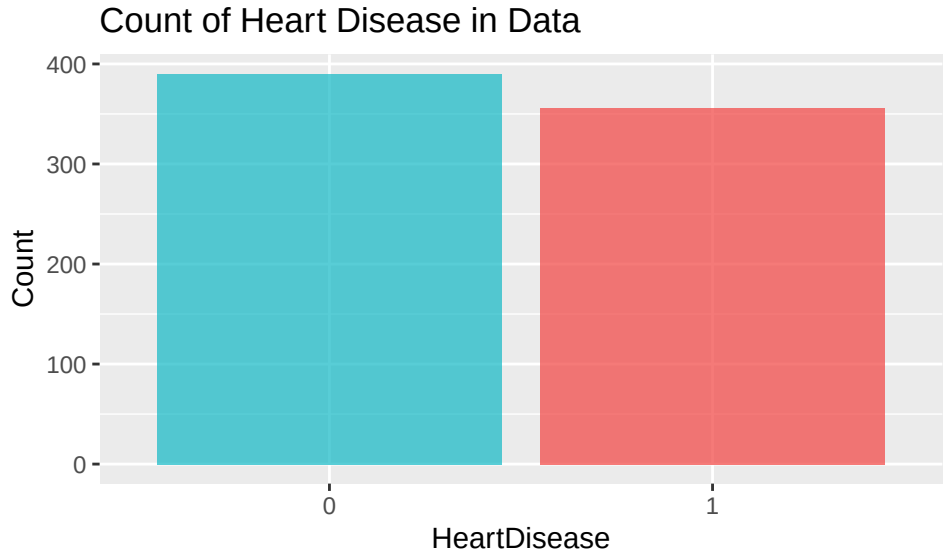
Figure 1: Missing Data Visualization

There were a total of 172 observations with missing values for `Cholesterol` and `RestingBP`. (Cholesterol

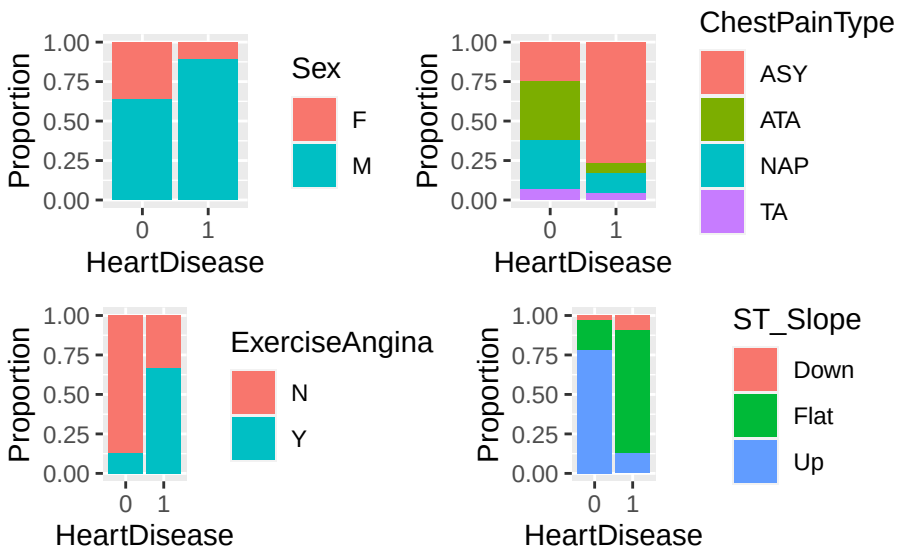
value of 0 and Resting Blood Pressure (BP) value of 0 are considered missing since these are impossible to reach in real life). The visualizations of the missing values in [Figure 1] shows that there are 171 observations with missing Cholesterol value and 1 observation with missing both Cholesterol and Resting BP value. For the one observation with missing Resting BP, it also had missing Cholesterol value. This observation was disregarded for analysis under Missing Completely at Random (MCAR) assumption. For the 171 observations with missing Cholesterol values, imputations were conducted based on Missing at Random (MAR) assumption. Multiple Imputation using Chained Equation (MICE) in R was used to impute the values. Five chains of imputations were conducted for Cholesterol with each chain using the default method of predictive mean matching (ppm) method since Cholesterol was a numerical variable.

Exploratory Data Analysis

The graph below shows that in the dataset, the frequency of a positive diagnosis for heart disease is roughly equal to the frequency of a negative diagnosis of heart disease. Thus, the data is likely not based on the general population, where the frequency of heart disease is much lower. As a result, the inference and predictions from our models do not apply to the general population, but only to the population this dataset was drawn from, which is the population of patients who are at risk of heart disease and were checked for heart disease.

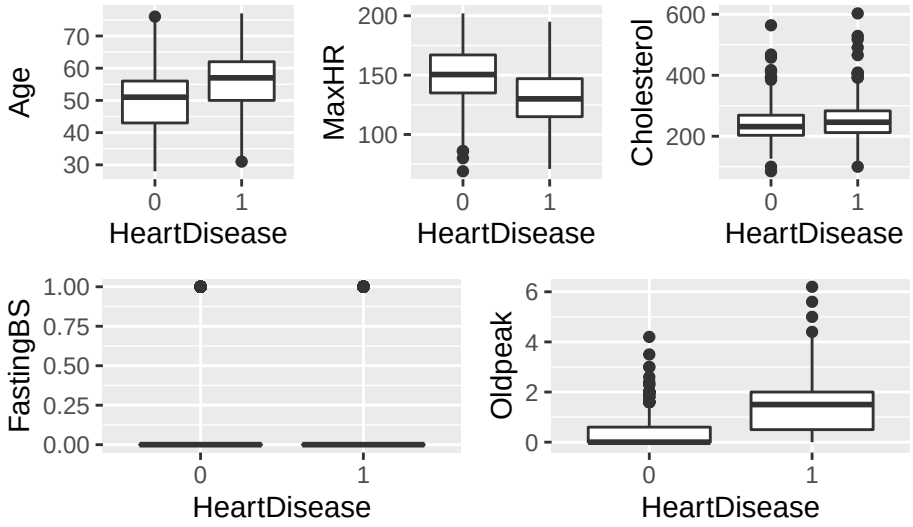


Apart from the spread of `HeartDisease`, it was important to visualize the other predictors and their relation to `HeartDisease` in the dataset. Below are plots four categorical variables (`Sex`, `ChestPainType`, `ExerciseAngina`, and `ST_Slope`)



The above 4 categorical variables, sex, chest pain types, exercise-induced angina, and ST slope seem to have some relationship with heart disease incidence, so these variables will be good to look out for in our final model. From the exploratory data analysis, it appears that a positive diagnosis for CVD tends to occur with the sex of a patient being male, asymptomatic chest pain, the presence of exercise-induced angina, and a flat slope of the peak exercise ST segment.

Next, we plotted the relationship between the incidence of heart disease and five numeric variables (`Age`, `MaxHR`, `Cholesterol`, `FastingBS`, and `Oldpeak`).



From the exploratory data analysis, it appears that the incidence of heart disease tends to occur with higher age, a lower maximum heart rate, slightly higher cholesterol, and a higher level of exercise relative to rest. These variables may be worth exploring in our models later.

III. Methodology

Logistic Regression

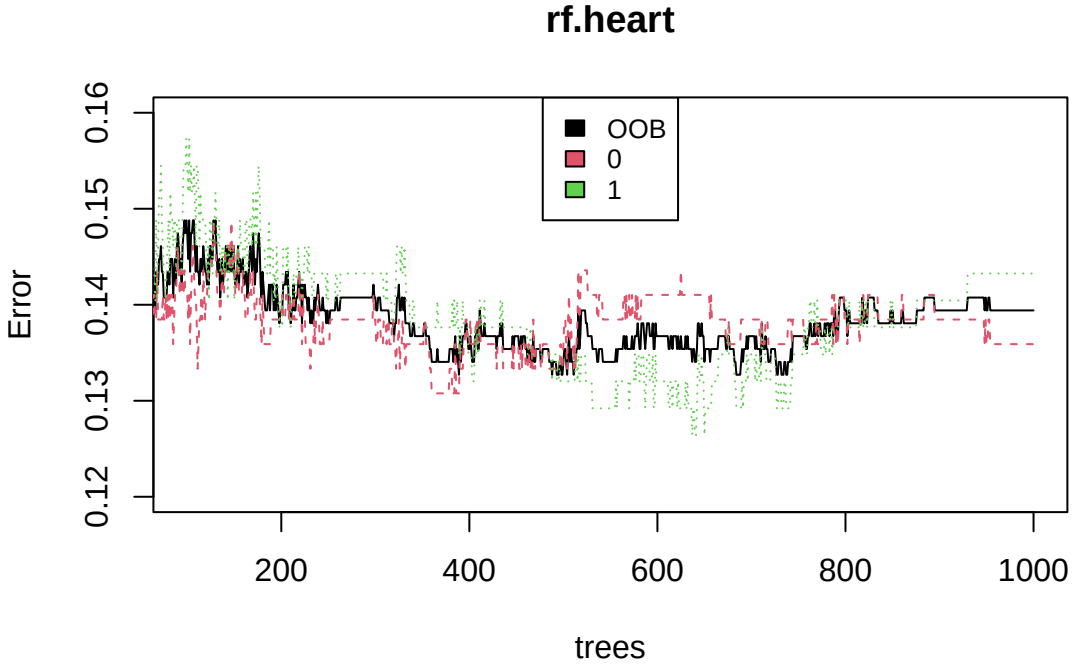
Our first model is a logistic regression model. The logistic regression model is formulated by the equation:

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Using logistic regression lends itself well to inference and classification goals, and using the Lasso method will perform model selection. Therefore, we decided to use a logistic lasso regression. This is a logistic regression with shrinkage applied to the coefficients of the logistic regression. The amount of shrinkage is controlled by a shrinkage parameter lambda, and in lasso shrinkage, the coefficients are able to be shrunk to 0, which means that lasso can perform variable selection. We used the glmnet package to fit the lasso regression model. We used cross-validation to determine the lambda value that results in the lowest CV error, and then fitted the lasso regression model with the selected lambda value. We then used any active main effects, and interaction effects with non-zero values for any level to fit the final logistic regression using the glm() function. Even though many of the main effects for the logistic regression were not active (had coefficient values of 0), we made sure to include any main effects associated with the active interaction effects due to the hierarchy principle.

Random Forest

Our second model uses a random forest. With random forests, we fit a large number of binary decision trees, each on a bagged set of data and each with a random subset of predictors, then average the trees to get a final prediction. By using random subsets of predictors, we decorrelate the trees and reduce the variance compared to other methods like bagging. The random forest method yields itself well to the project goals at hand because of the categorical and binary nature of the outcomes we're trying to predict. We built a random forest of 1000 trees built off of bootstrapped Heart Data. We specified number of predictors sampled for spitting at each node as sqrt(11) ~ 3. We chose sqrt(11) since this is the industry-standard for classification. We chose 650 trees in order to lower our false negative rate, as we can see in our error plot.



```
##
## Call:
##  randomForest(formula = HeartDisease ~ ., data = heart, mtry = sqrt(11),      ntree = 650)
##                Type of random forest: classification
##                      Number of trees: 650
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 13.4%
## Confusion matrix:
##     0   1 class.error
## 0 335  55   0.1410256
## 1  45 311   0.1264045
```

SVM

Finally, our third model uses a Support Vector Machine (SVM). An SVM uses kernels to provide a flexible classification model. It separates the domain of the data with the goal maximizing the margin distance while allowing for a small number of misclassified training points. Although SVMs are not very interpretable and therefore wouldn't fit our inference goals, they tend to yield high prediction accuracy and work well with categorical response variables, so we fit an SVM for prediction purposes to compare with our other models. The cost tuning parameters for our SVM models are determined using cross validation, and the best kernel is determined at the end based on the misclassification error from a 70-30 training-test split. The best kernel ended up being a tie between the linear and radial kernals, so we chose the linear kernel for parity. The equation for the linear kernel is:

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$$

```
##
## Call:
## svm(formula = HeartDisease ~ ., data = train, kernel = "linear",
##     cost = tune.out$best.parameters$cost)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  5.623413
##
## Number of Support Vectors:  178
##
##  ( 92 86 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
## [1] 0.1205357
```

Our final SVM model uses a linear kernal with a cost of 5.623. The error rate for this model is 12.05%.

IV. Results

Lasso:

Random Forest:

Our confusion matrix tells us that there is a 0.168 false positive rate and a 0.094 false negative rate. This means that 16.8% of the time, our model predicts someone who doesn't have heart disease to have heart disease. At the same time, 9.4% of the time, our model predicts someone who has heart disease to not have heart disease. It is better in this scenario that our false positive rate is higher than the false negative rate because it is better that we over predict people having heart disease than to under predict. If we over predict, people who didn't have heart disease but thought they did would get a second opinion and eventually realize that they don't actually have the disease. If we under predicted, then people who have actually heart disease wouldn't get the necessary treatment and may have worsened affects. Despite the false negative rate being quite high, we decided to keep the default cutoff value of 0.5 because without having done a lot more in-depth research on this medical topic, it would be fairly arbitrary of us to choose another cut off point. Intuitively, it would make sense for us to choose a higher cutoff value because there is a heavier social penalty for a high false negative value as compared to a false positive value.

SVM:

V. Conclusion

As mentioned earlier, one of the biggest limitations of our findings is the dataset itself. Although the dataset is one of the largest available datasets on heart disease, the dataset's collection methodology limits the usefulness of our models. As mentioned in the Section II of this report, the data is only collected from at-risk patients who received CVD diagnostic check-ups at hospitals. Thus, the data may be weighted disproportionately towards people who would received a positive diagnosis for CVDs, as the sample population in the data were not just at-risk, but also either self-selected for or were chosen to be checked for CVDs. Furthermore, the data was collected from only five hospitals in the United States and Europe. As a result, the data is far from representative of the entire global population that is at-risk for CVDs, and our models and results may not be accurate for predicting the likelihood of CVDs for all at-risk populations in the world, and may provide limited understanding of signs that at-risk populations who are not represented by the data can analyze to check their risk for CVDs.