# Final-Report

Conner Byrd, Eric Han, Ki Hyun, Sara Shao, Alex Shen, Mona Su, Dani Trejo, Steven Yuan

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 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 backwards selection using AIC. 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. 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**

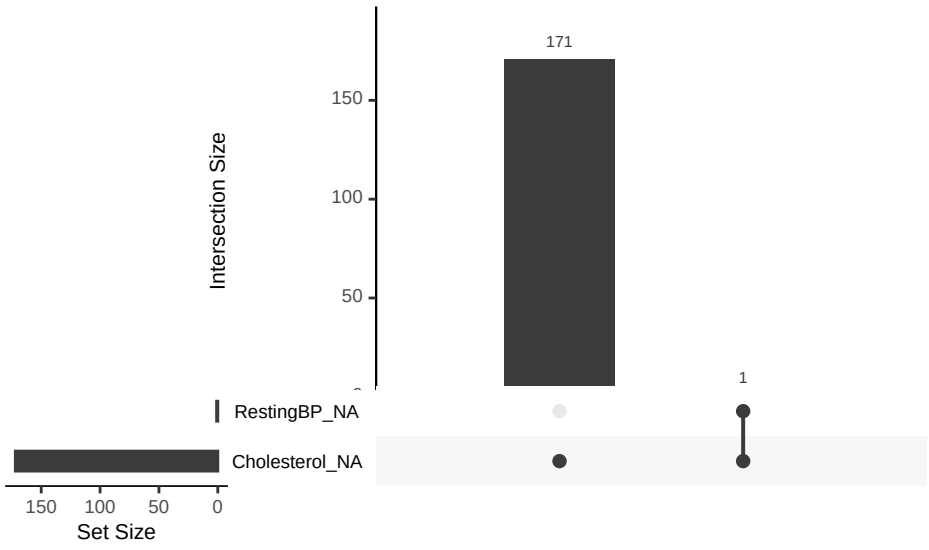



Figure 1: Missing Data Visualization

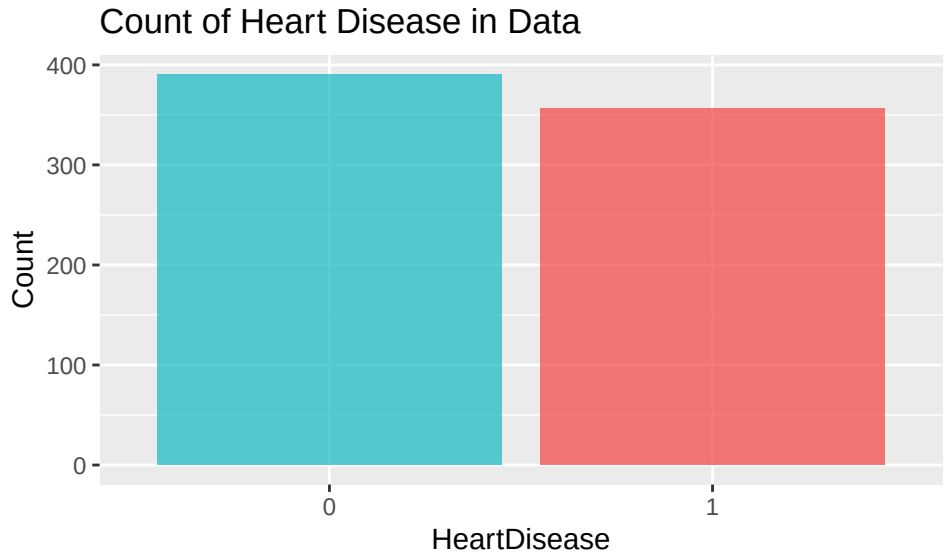
There were a total of 172 observations with missing values for `Cholesterol` and `RestingBP`. (Cholesterol value of 0 and Resting Blood Pressure (BP) value of 0 are considered missing since these are impossible to

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.

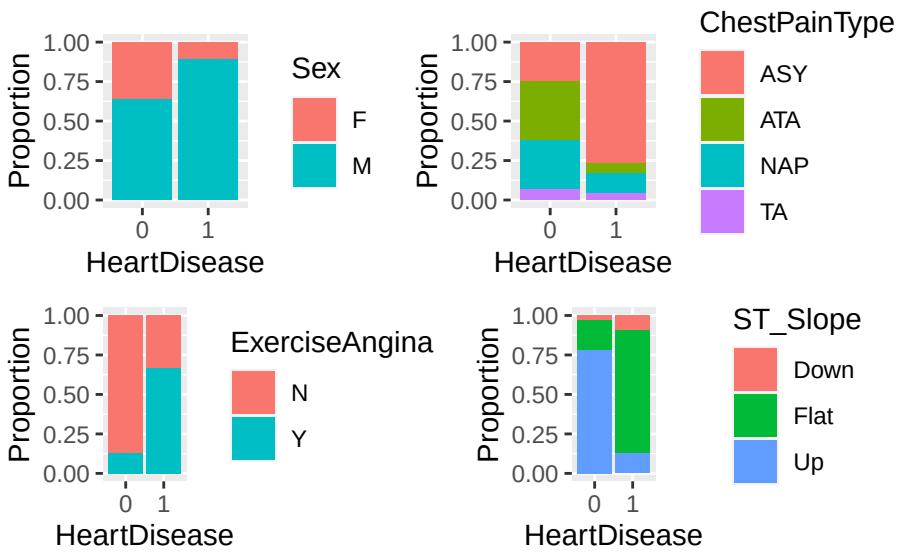
**Exploratory Data Analysis**

Note: The EDA is done excluding missing values.

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.

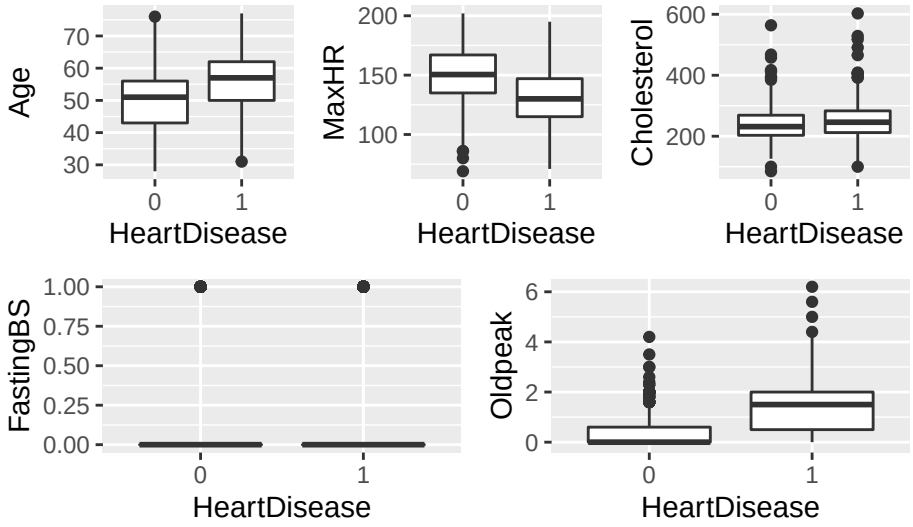



Apart from the spread of `HeartDisease`, it was important to visualize the other predictors and their relation to `HeartDisease` in the dataset. Below are plots four categorical variables (`Sex`, `ChestPainType`, `ExerciseAngina`, and `ST_Slope`)




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

Note: Results from the logistic regression and random forest are examples from fitting each model to the first of 5 chains of imputed results.

**Logistic Regression**

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

$$\log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$$

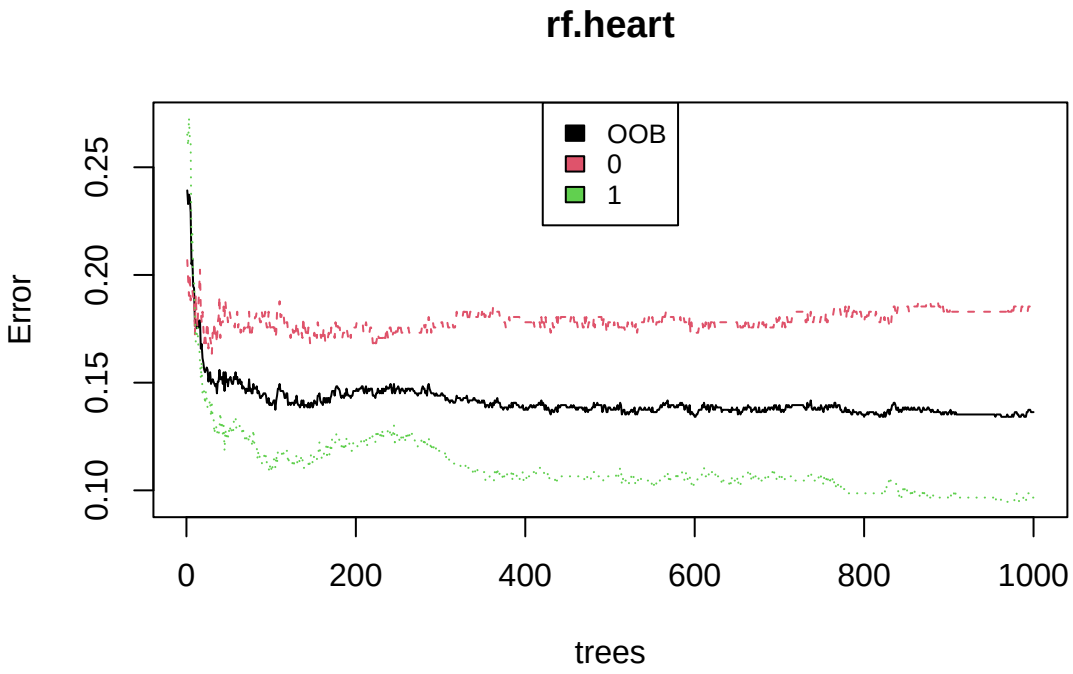
Using logistic regression lends itself well to inference and classification goals. Our original plan was to use a lasso logistic regression to perform variable selection for all main effects plus all pairwise interactions due to the ability of lasso to shrink coefficients to 0. Then, we would add the selected variables plus any associated main effects we might need to satisfy the hierarchy principle into a normal logistic regression. However, after observing the output of the normal logistic regression, we realized that many of the p values were very large, with very few p values below an alpha significance threshold of 0.05. Since our goal with the logistic model is to be able to interpret and find relationships between predictor variables and the probability of having heart disease, we decided against using this because having very few significant terms hinders our goals.

Therefore, we decided to use a logistic regression with backwards stepwise selection using AIC. Our starting full model included all main effects and all pairwise interactions. After backwards selection, we made sure that all active interaction effects had their associated main effects included in the selected model, so that the model adheres to the hierarchy principle which aids in the interpretation of interaction effects.

The backward-AIC selection method was implemented for each of the five MICE chains of imputed data. The formulas of the optimal model for each chain were saved separately as `RDS` objects and was later used for the 5-fold CV estimation of test-error rate.

**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 650 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) (square root of the number of predictors) 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.




**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 three types of kernels for SVM that we considered were linear, polynomial, and radial. For each of the kernels, we considered a range of the penalty coefficient $C$. For polynomial kernel, the degree was set to 2. The best cost tuning parameters for each of our SVM models were determined using the 'tune' function provided by the 'e1071' package.

Then, using the chosen hyperparameters for each models, we ran 5-fold cross validation to determine the estimated test error for each of the models. In each fold, we trained three svm models using 80% of the data, each for the three types of kernels with the chosen parameters. Then, on the other 20% of the data, we computed the misclassification error rate at each fold. Then, we compared the mean of the misclassification rates to select the best model.

Table 2: Estimated Test Error for the Three types of Kernels

| Type of Kernel | Estimated Test Error |
|---|---|
| **Radial** | **0.131** |
| Linear | 0.138 |
| Polynomial | 0.145 |

As shown in [Table 2], for the first chain, the radial SVM model performed best among the three types of kernels. The best hyperparameter chosen for linear SVM is $c \approx 1.78$. In our best SVM model, there are 361 support vectors.

The radial kernel was selected for the four other MICE chains as well. The corresponding best hyperparameter and kernel type was saved as a list in a `RDS` file and was referred later for the comparison of 5-fold CV estimated error rate across different models.

The summary of cost and type of kernel for each chain is shown below in [Table 3].

Table 3: SVM Model Specification for Each MICE Chain

| MICE Chain | Best Hyper-parameter | Best Kernel Type | Number of Support Vectors in the Best Model |
|---|---|---|---|
| 1 | 1.778 | radial | 361 |
| 2 | 0.562 | radial | 402 |
| 3 | 1.778 | radial | 360 |
| 4 | 1.778 | radial | 364 |
| 5 | 1.778 | radial | 364 |

**Comparision of the models through CV**

On top of this, we also considered a combination of the models chosen from each algorithm: radial SVM, random forest, and logistic regression. The combination model predicts the value that appears most among the three models' predictions. To compare the predictive accuracy of all four models (the three individual models and the combined model), cross-validation with five folds was used. The same indexes for the folds were set for the different MICE chains of imputed data. For each chain, at each fold, each of the models were trained with the hyperparameters that were selected for the specific chain. Then the 5 predictions were drawn from all the MICE chains and were averaged. For the logistic model, if the average prediction value was above the pre-specified threshold of 0.5, then a overall prediction was set as 1 at that fold. For the Random Forest and SVM models, the prediction that held majority (3 or more) was set as the overall prediction value at each fold. For the aggregate model, the prediction that held majority (2 or more) of the three models' predictions was set as the overall prediction value at each fold. Using these overall prediction values, the misclassification rate on each fold was calculated for each model, and the mean of the misclassification rates were computed in the end.

## IV. Results

**Mean misclassification rates for each model**

Table 4: Estimated Mean Test-Error Rate for Different Models

| Model | Test.Err |
|---|---|
| **Random Forest** | **0.125** |
| Aggregate | 0.129 |
| SVM | 0.129 |
| Backward-AIC Logistic | 0.141 |

As shown in [Table 4] above, the random forest model had the best predictive performance with a mean misclassification rate of $\approx 12.5\%$, outperforming even the combined model. Therefore, if we wanted a model that would assess whether an at-risk patient most likely has or does not have heart disease based on their personal and health factors, this is the model we would choose.

Note: The following analysis of the logistic regression and random forest results are based on fitting each model to the first of 5 chains of imputed results.

**Logistic Regression:**

The output of the optimal logistic model is shown below in [Table 5].

Table 5: Output of the optimal logistic model in the first MICE Chain

| Term | Estimate | Stadard Errror | P value |
|---|---|---|---|
| (Intercept) | -4.8586 | 5.8156 | 0.4035 |
| Age | 0.0425 | 0.0550 | 0.4399 |
| **SexM** | **5.3446** | **1.9689** | **0.0066** |
| **ChestPainTypeATA** | **-5.6198** | **1.9506** | **0.0040** |
| **ChestPainTypeNAP** | **-3.7440** | **1.3236** | **0.0047** |
| ChestPainTypeTA | -4.7001 | 2.5335 | 0.0636 |
| RestingBP | -0.0061 | 0.0260 | 0.8161 |
| Cholesterol | -0.0003 | 0.0028 | 0.9056 |
| **FastingBS** | **6.4489** | **2.8675** | **0.0245** |
| RestingECGNormal | 4.0192 | 2.1621 | 0.0630 |
| RestingECGST | 4.3184 | 2.7191 | 0.1123 |
| MaxHR | 0.0174 | 0.0229 | 0.4482 |
| ExerciseAnginaY | -2.9202 | 2.0178 | 0.1478 |
| Oldpeak | 0.4463 | 0.3484 | 0.2002 |
| ST_SlopeFlat | 2.3152 | 5.4322 | 0.6700 |
| ST_SlopeUp | -6.2863 | 5.5828 | 0.2602 |
| Age:ST_SlopeFlat | -0.0703 | 0.0588 | 0.2322 |
| Age:ST_SlopeUp | 0.0177 | 0.0592 | 0.7657 |
| **SexM:FastingBS** | **-2.3653** | **0.9815** | **0.0160** |
| SexM:MaxHR | -0.0216 | 0.0135 | 0.1099 |
| **ChestPainTypeATA:Cholesterol** | **0.0173** | **0.0067** | **0.0094** |
| ChestPainTypeNAP:Cholesterol | 0.0053 | 0.0047 | 0.2630 |
| ChestPainTypeTA:Cholesterol | 0.0072 | 0.0099 | 0.4664 |
| ChestPainTypeATA:FastingBS | -1.5242 | 1.0446 | 0.1445 |
| **ChestPainTypeNAP:FastingBS** | **-2.2098** | **0.8630** | **0.0105** |
| ChestPainTypeTA:FastingBS | -0.8393 | 1.0497 | 0.4239 |
| ChestPainTypeATA:RestingECGNormal | -0.8921 | 0.8441 | 0.2906 |
| ChestPainTypeNAP:RestingECGNormal | 1.2255 | 0.6729 | 0.0686 |
| **ChestPainTypeTA:RestingECGNormal** | **2.3987** | **1.0303** | **0.0199** |
| ChestPainTypeATA:RestingECGST | 0.2443 | 1.1305 | 0.8289 |
| **ChestPainTypeNAP:RestingECGST** | **2.7503** | **0.9491** | **0.0038** |
| **ChestPainTypeTA:RestingECGST** | **3.5597** | **1.4361** | **0.0132** |
| **RestingBP:FastingBS** | **-0.0390** | **0.0175** | **0.0260** |
| **RestingBP:RestingECGNormal** | **-0.0378** | **0.0166** | **0.0229** |
| **RestingBP:RestingECGST** | **-0.0450** | **0.0207** | **0.0298** |
| RestingBP:ExerciseAnginaY | 0.0295 | 0.0153 | 0.0536 |
| RestingBP:ST_SlopeFlat | 0.0484 | 0.0256 | 0.0586 |
| RestingBP:ST_SlopeUp | 0.0198 | 0.0266 | 0.4565 |
| FastingBS:RestingECGNormal | 1.3506 | 0.7595 | 0.0754 |
| **FastingBS:RestingECGST** | **2.2145** | **0.9926** | **0.0257** |
| **FastingBS:ST_SlopeFlat** | **2.7528** | **1.1883** | **0.0205** |
| FastingBS:ST_SlopeUp | 1.2494 | 1.0666 | 0.2415 |
| MaxHR:ST_SlopeFlat | -0.0250 | 0.0207 | 0.2273 |
| MaxHR:ST_SlopeUp | 0.0078 | 0.0204 | 0.7042 |
| Oldpeak:ST_SlopeFlat | -0.3837 | 0.3912 | 0.3267 |
| Oldpeak:ST_SlopeUp | 0.4351 | 0.4281 | 0.3094 |

*Note:*

The significant coefficients at alpha = 0.05 are highlighted in bold

We would like to use the results of the logistic regression in order to interpret the relationships between different variables and the response variable, despite it not being the best model for prediction.

We will be concentrating on interpreting some of the significant terms. One of the categorical terms that is significant is Sex, with an associated coefficient of 5.345. This means that being male, compared to the baseline of female, leads to an expected 5.345 increase in the log odds of having heart disease. We can also say that being male, compared to the baseline of female, leads to the odds of having heart disease to multiply by a factor of e^(5.345). ChestPainType of ATA and NAP, and FastingBS were three other main effects with significant ($\alpha = 0.05$) coefficients. For patients with ChestPainType of ATA and NAP, compared to the baseline ChestPainType of ASY, the log odds of having a heart disease by -5.620 and -3.744 respectively. For patients who have FastingBS value of 1, compared to the baseline value of 0, the log odds of having a heart disease increases by 6.449. What this means in context is that people with a fasting BS greater than 120 mg are more likely to have heart disease than those who have a fasting BS less than 120 mg. All other main effects are not considered significant at $\alpha = 0.05$.

There are a handful of significant interaction effects, most of which are interactions between 2 categorical variables. One example of this is between Sex and FastingBS. When FastingBS takes the baseline value of 0, a male compared to the baseline of female will result in an expected 5.345 increase in the log odds of having heart disease. When FastingBS takes the value 1, a male compared to the baseline of female will now result in an expected 5.345 - 2.365 = 2.980 increase in the log odds of having heart disease. One example of an interaction between a categorical and continuous variable is ChestPainType and Cholesterol. We should keep in mind that this interpretation may not be appropriate because the main effect for Cholesterol is not considered significant. For a baseline ChestPainType of ASY, an increase of 1 unit in Cholesterol will result in an expected decrease of 0.0003 in the log odds of having heart disease. For a ChestPainType of ATA, a 1 unit increase in Cholesterol now results in an expected increase in 0.0173 - 0.0003 = 0.0170 in the log odds of having heart disease.

**Random Forest:**

Table 6: Output of the optimal Random Forest model in the first MICE Chain

| | | | |
|---|---|---|---|
| **Out-of-bag Error** | **13.63%** | | |
| **Confusion Matrix** | | | |
| | Predicted as 0 | Predicted as 1 | Class Error |
| Actually 0 | 336 | 74 | 0.1805 |
| Actually 1 | 50 | 457 | 0.0986 |

As shown in [Table 6] above, the out-of-bag error from this model is 13.63%, meaning we expect this model to classify patients incorrectly around 13.63% of the time.

Our confusion matrix tells us that there is a 0.185 false positive rate and a 0.097 false negative rate. This means that 18.5% of the time, our model predicts someone who doesn't have heart disease to have heart disease. At the same time, 9.7% 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. One limitation of our random forest is that we can't tell from this model which factors are important in predicting heart disease. Another limitation is that

we can't predict heart disease as accurately with this model compared to the SVM model.

**SVM:**

Compared to the logistic regression model, the SVM performs better for prediction (12.9% misclassification rate), as we expected. However, this model does not tell us much about what factors are or aren't important in predicting heart disease, and it cannot tell us actually likelihoods of having heart disease or not.

## V. Conclusion

Although still far from perfect, our models performed quite well when it came to predicting patients with CVD and we uncovered a lot of information as a result. Our methods yielded a heart disease prediction model with an 87.5% accuracy rate for at-risk patients. We also found that patient sex and chest pain type, among others, are important predictors of CVD. More specifically, we found that men are more likely to have CVD than women, and people that have no chest pain are more likely to have CVD than people with ATA or NAP type chest pain. Expanding this to the real world opens up a myriad of medical and personal uses for our models and predictive techniques. Starting simple, if someone were to be wary of their overall health for some combination of reasons (past medical history, family medical history, some form of high blood pressure/high blood sugar/cholesterol) and have not been or are not able to be diagnosed by a professional, they could be able to input their known medical information and receive a relative prediction for their risk of obtaining a CVD. Although this does not conduce a formal diagnosis, it is a quick and simple way to obtain a relative idea of health for those who may struggle economically with visiting the doctor. Our models could also be used in a more clinical setting with some success. If a professional noted that the prediction or probability for CVD in a patient was high, it would encourage them to conduct further tests and possibly seek some form of diagnosis. Although our models themselves aren't perfect, using them in conjunction with other medical testing could certainly yield positive results.

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.