# **MAT 303 Project Two Summary Report**

Kathryn Delaney-Seamans

Kathryn.delaney-seamans@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## **1. Introduction**

The data set that I will be exploring in this problem set focuses on analyzing risk factors for heart disease. I will be analyzing different health indicators that will be used as predictor variables, such as age, fasting blood sugar, maximum heart rate, and cholesterol. The results from these analyses will be used to predict the risk of heart disease and identify high risk patients. This could be used by doctors and other medical professionals. I will be creating logistic regression models to predict the presence of heart disease, as well as using random forest models.

## **2. Data Preparation**

The important variables in this data set include the predictor variables age, resting blood pressure, exercise induced angina, maximum heart rate, chest pain experienced, sex, resting electrocardiographic measurement, and number of major blood vessels. There are 14 columns (variables) and 303 rows.

## **3. Model #1 - First Logistic Regression Model**

### **Reporting Results**

The general form of this regression model is:

In this model, represents age, represents resting blood pressure, represents exercise induced angina, and represents maximum heart rate achieved.

Then, this can be transformed into:

In the context of this model, represents the probability that someone has heart disease. represents the odds of having heart disease. is the probability of not defaulting, so the previous expression is the ratio of the probability of having heart disease to the probability of not having heart disease.

The prediction equation of this model is:

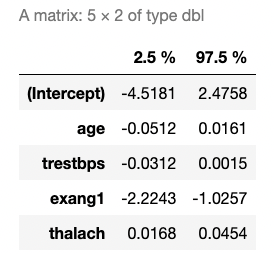
Using the outputs from the R script:

The coefficient for maximum heart rate achieved is 0.0311. This tells us that for each increase in maximum heart rate, the odds of having heart disease increase by 0.0311. Having a higher maximum heart rate during exercise is associated with having heart disease.

### **Evaluating Model Significance**

To determine if the model fits the data, we can use the Hosmer-Lemeshow goodness of fit test. The null hypothesis is that the logistic regression model fits the data well and the alternative hypothesis is that the model does not fit the data well. The p-value from this test is 0.612. Since the p value is much greater than 0.05, we fail to reject the null hypothesis and determine that the model fits the data well.

For Wald’s test, the confidence intervals are listed below:

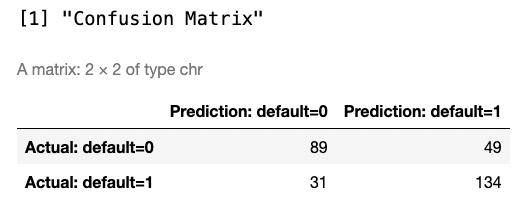


The tests for significance are listed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P value | Conclusion |
| age |  | Z=-1.024 |  | Fail to reject the null |
| trestbps |  | Z=-1.786 |  | Fail to reject the null |
| exang1 |  | Z=-5.134 |  | Reject the null |
| thalach |  | Z=4.274 |  | Reject the null |

Exercise induced angina and maximum heart exhibit a high level of significance and fall below the 5% threshold, while age and resting blood pressure do not.

The confusion matrix is listed below:



True positive is 134, true negative is 89, false positive is 49, and false negative is 31.

To obtain accuracy, we plug in the values from the confusion matrix:

The accuracy is 73.56%.

To obtain precision, we can again plug in the values from the confusion matrix:

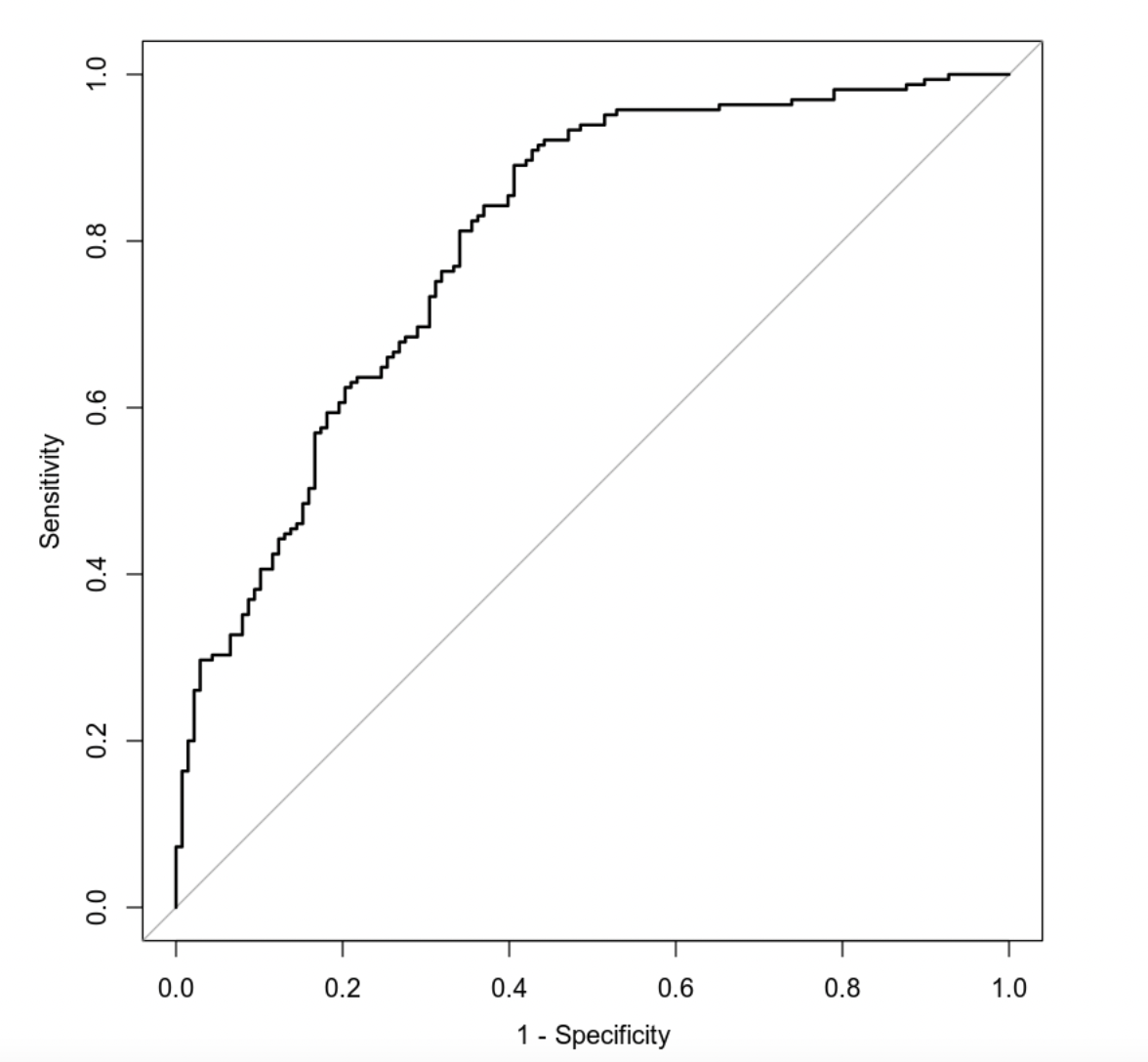
Precision is 73.22%.

To obtain recall:

8121

Recall is 81.21%.

The ROC curve is below:

****

The ROC curve tells us how successful the model is at differentiating between zero and one values. The curve is well above the diagonal line, which tells us that the model has strong discrimination ability. The AUC value is 0.8007, which suggests the model is very good at distinguishing between those who have heart disease and those who do not.

### **Making Predictions Using Model**

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 122, has exercise induced angina, and a maximum heart rate of 140 is 27.16%. The odds of this event are This tells us this person has a low probability of having heart disease. Although exercise induced angina is a significant risk factor for heart disease, it seems as though the lower maximum heart rate does lower the odds compared to the second individual with a higher heart rate.

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 130, does not have exercise induced angina, and a maximum heart rate of 165 is 78.53%. The odds of this event are These values indicate a higher risk for heart disease. The maximum heart rate likely increases the risk of heart disease substantially in the second individual.

## **4. Model #2 - Second Logistic Regression Model**

### **Reporting Results**

The general form of this regression model is:

In this model, represents age, represents resting blood pressure, represent dummy variables for type of chest pain experience, and represents maximum heart rate achieved.

Then, this can be transformed into:

In the context of this model, represents the probability that someone has heart disease. represents the odds of having heart disease. is the probability of not defaulting, so the previous expression is the ratio of the probability of having heart disease to the probability of not having heart disease.

The prediction equation of this model is:

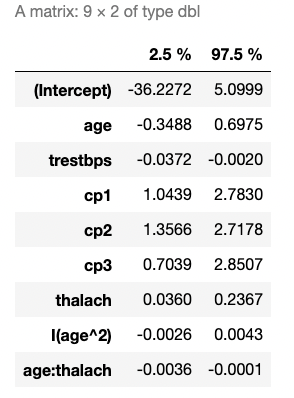
Using the outputs from the R script:

The coefficient for maximum heart rate achieved is 0.0311. This tells us that for each increase in maximum heart rate, the odds of having heart disease increase by 0.0311. Having a higher maximum heart rate during exercise is associated with having heart disease.

### **Evaluating Model Significance**

To determine if the model fits the data, we can use the Hosmer-Lemeshow goodness of fit test. The null hypothesis is that the logistic regression model fits the data well and the alternative hypothesis is that the model does not fit the data well. The p-value from this test is 0.3209. Since the p value is greater than 0.05, we fail to reject the null hypothesis and determine that the model fits the data well.

For Wald’s test, the confidence intervals are listed below:

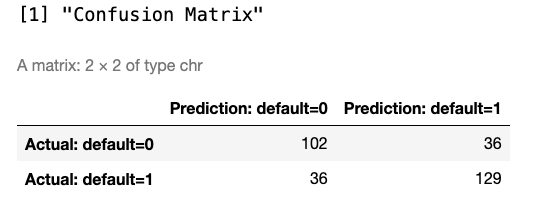


The tests for significance are listed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P value | Conclusion |
| age |  | Z=0.653 |  | Fail to reject the null |
| trestbps |  | Z=-2.181 |  | Reject the null |
| Cp1 |  | Z=4.313 |  | Reject the null |
| Cp2 |  | Z=5.867 |  | Reject the null |
| Cp3 |  | Z=3.245 |  | Reject the null |
| thalach |  | Z=2.663 |  | Reject the null |
| Age2 |  | Z=0.481 |  | Fail to reject the null |
| Age:thalach |  | Z=-2.095 |  | Reject the null |

All the predictor variables exhibit significance below the 5% threshold except for age and age2.

The confusion matrix is listed below:



True positive is 129, true negative is 102, false positive is 36, and false negative is 36.

To obtain accuracy, we plug in the values from the confusion matrix:

The accuracy is 76.24%.

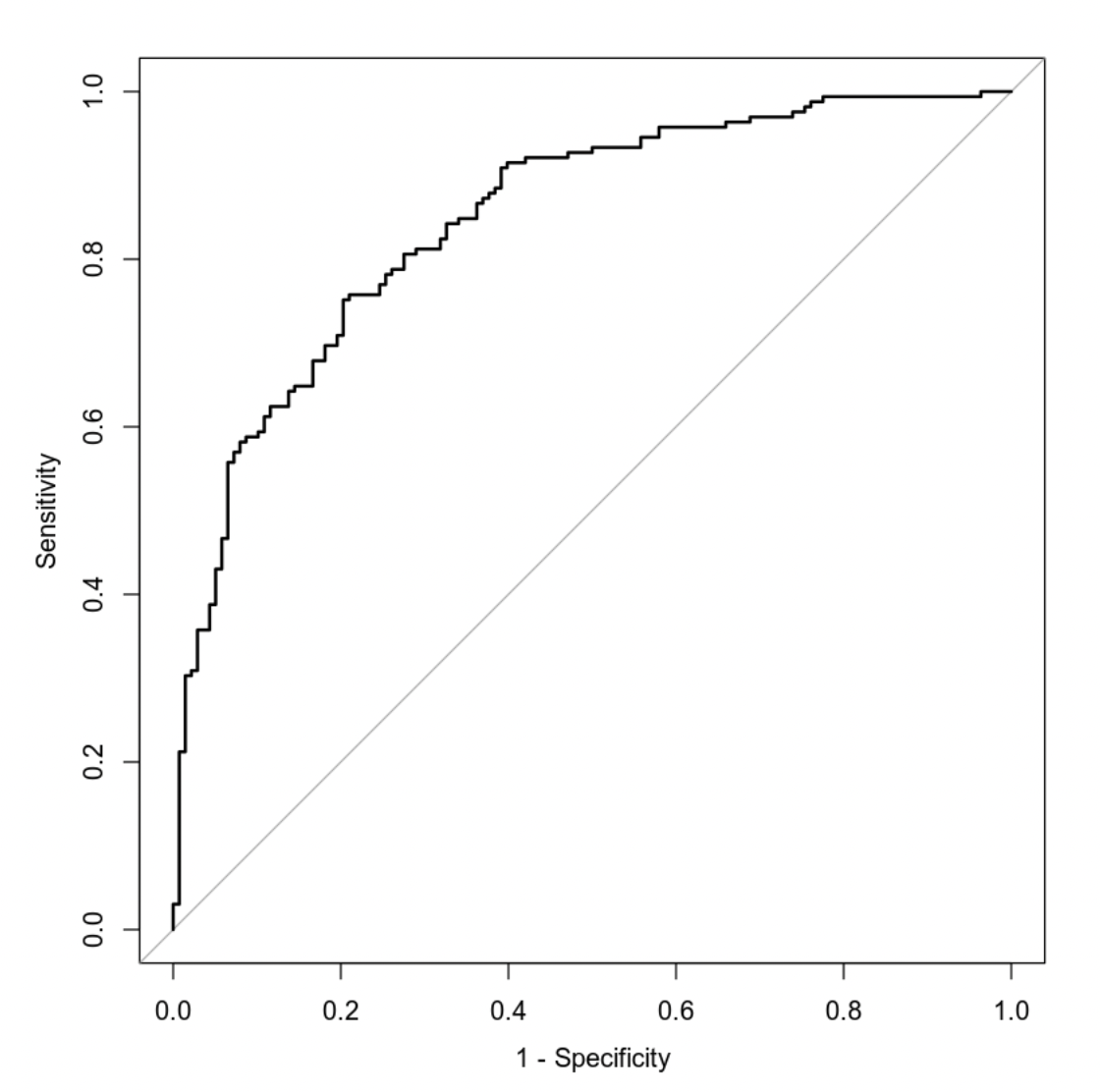
To obtain precision, we can again plug in the values from the confusion matrix:

Precision is 78.18%.

To obtain recall:

Recall is 78.18%.

The ROC curve is below:



The ROC curve tells us how successful the model is at differentiating between zero and one values. The curve is well above the diagonal line, which tells us that the model has strong discrimination ability. The AUC value is 0.8478, which suggests the model is very good at distinguishing between those who have heart disease and those who do not.

### **Making Predictions Using Model**

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 115, does not experience chest pain, and has maximum heart rate of 133is 21.88%. The odds of this event are This tells us this person has a low probability of having heart disease.

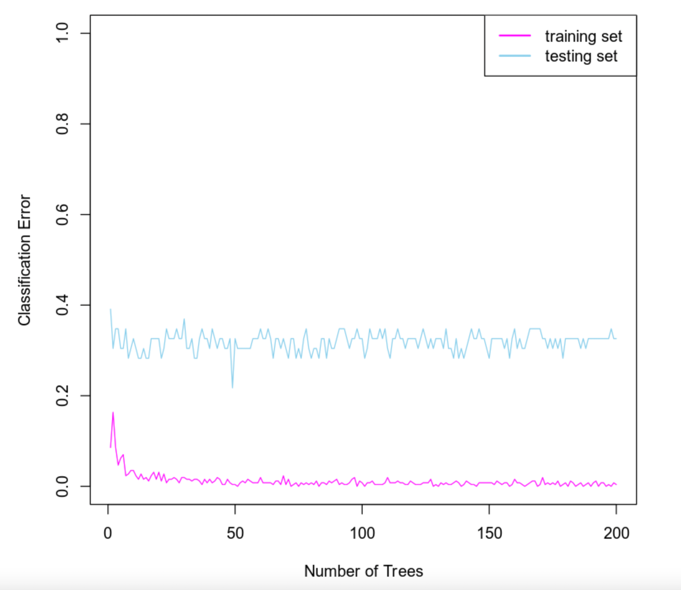
The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 125, experiences typical angina, and has maximum heart rate of 155is 80.07%. The odds of this event are These values indicate a higher risk for heart disease. The maximum heart rate and typical angina chest pain likely increases the risk of heart disease substantially compared to the first individual.

## **5. Random Forest Classification Model**

### **Reporting Results**

Based on the 85% and 15% split, there are 303 rows in the original data set, 257 in the training set, and 46 in the validation set.

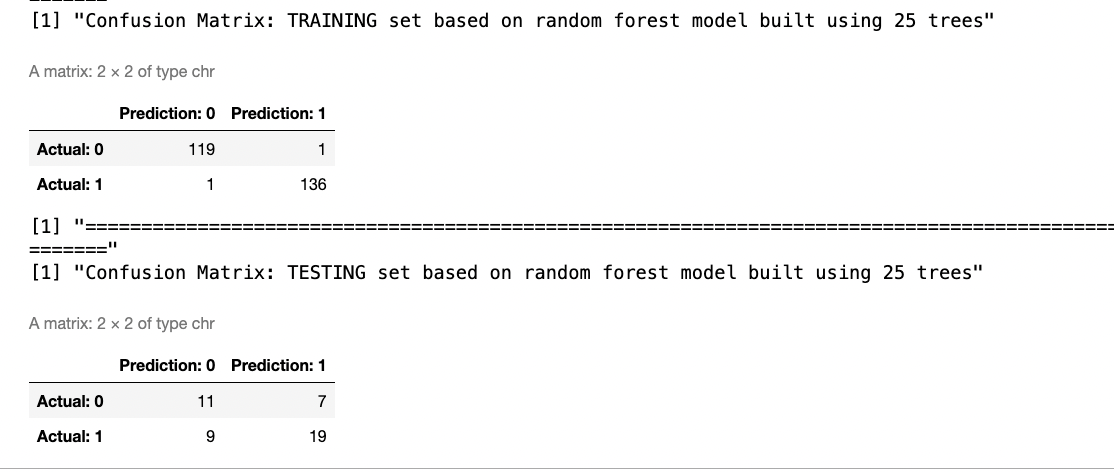
Below is the graph of the training and testing error for the predictor variables age, sex, chest pain type, resting blood pressure, cholesterol measurement, resting electrocardiographic measurement, exercise-induced angina, and number of major vessels:



Based on the graph, I can see that the training and testing errors decrease initially, until the classification error curve hits around 25 trees when it flattens. The optimal number of trees is found where the curve flattens, which in this case is around 25 trees.

### **Evaluating the Utility of the model**

Below are the confusion matrices for the training and testing sets:

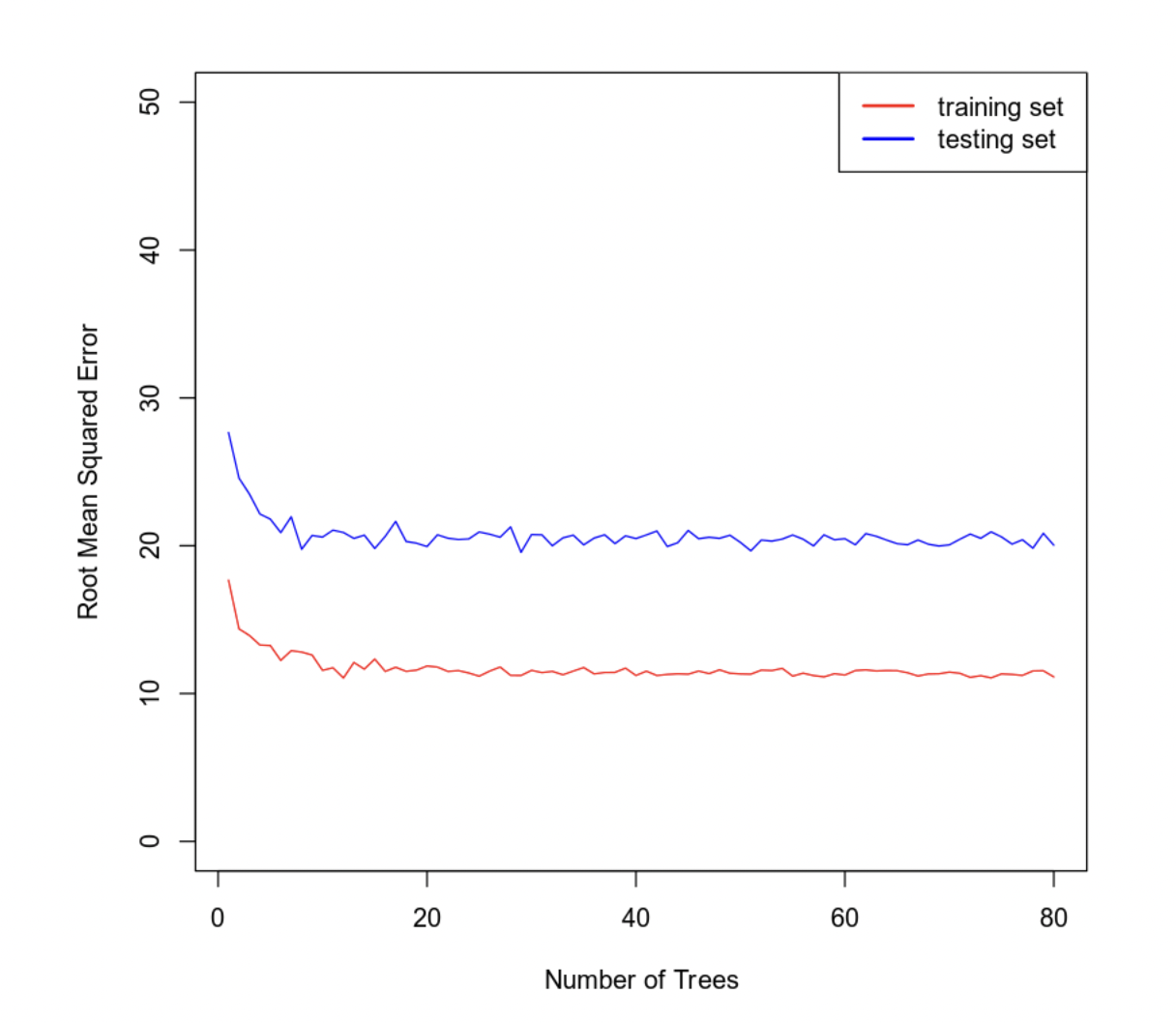


For the training set, 119 is true negative, 1 is false negative, 1 is false positive, and 136 is true positive. Based on these values, accuracy is or 99.22%, precision is or 99.27%, and recall is or 99.27%. This model shows very high values for each and indicates that the model performs well on the training set. For the testing set, 11 is true negative, 9 is false negative, 7 is false positive, and 19 is true positive. Based on these values, accuracy is or 65.22%, precision is or 73.08%, and recall is or 67.86%.

## **6. Random Forest Regression Model**

### **Reporting Results**

Using an 80% and 20% split, there are 303 rows in the original data set, 242 in the training set, and 61 in the validation set. Below is the graph for mean squared error against the number of trees using maximum heart rate achieved:



Based on the graph above, 15 trees seems to be where the curve flattens, so I will use that as the optimal number.

### **Evaluating the Utility of the Random Forest Regression Model**

Using 15 trees as the optimal number, the root mean squared error for the training set is 11.7971 and the root mean squared error for the testing set is 21.4423.

## **7. Conclusion**

Based on the statistical analyses performed, both models appear to be relatively effective for predicting heart disease. Although the second model has two predictors that were determined to be statistically insignificant (age and age2) and the first model also had two that were also statistically insignificant (age and resting blood pressure, the second model still has a better fit due to the fact that were significantly more predictor variables deemed significant compared to model one (6 compared to 2). In terms of the confusion matrix, we can compare the accuracy, precision, and recall of each model. The accuracy was 73.56% for the first model compared to 76.24% for the second model. The precision was 73.22% for the first model compared to 78.18% for the second model. The recall was 81.21% for the first model and 78.18% for the second. Based on the values, the second model overall is more accurate and precise. When comparing the values for ROC, the first model was 0.8007 compared to 0.8478 for the second model. Both models offer a good fit and we can say they will give a reasonable prediction about an individual’s risk for heart disease. However, based on the statistical insights from above, the second model is a slightly better choice.

I would recommend using the logistic regression model in this case. In the random forest classification model, there are discrepancies between accuracy, precision, and recall of the training and testing data, which would indicate potential overfitting. The logistic regression model outperforms the random forest classification model’s testing set accuracy, precision, and recall.

Both models that were used could be utilized for heart disease prevention by doctors and other medical professionals. These models may allow for early identification of individuals who are at risk for developing heart disease.