# **MAT 303 Project Two Summary Report**

Renaldo Musto

renaldo.musto@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The data set that I am going to be working with this week is a large set of historical data that contains individuals' health information related to their risk of heart disease. The data has been cleaned to make the statistical models easier to make. The data is going to be used to potentially help make better diagnosis when it comes to heart disease. If we can run a patient's chart through this model it may notice things that a human doctor did not and then we can screen them further depending on how high the predicted risk is. During this report I will be creating two logistic regression models and then evaluating them by graphing the receiver operating curve and interpreting that as well as the resulting value for area under the curve. I will also be acquiring a confusion matrix that will allow me to determine the accuracy, precision and recall of the model. Finally, I will determine significance through the hosmer-lemeshow goodness of fit tests as well as the Wald’s test for each term in the models. Next I will be making a random forest classification model and a random forest regression model. These will be evaluated using root mean square error and accuracy, precision and recall through their respective confusions matrices as well for both the training and testing sets. Lastly, I will be graphing the room mean square error against the number of trees to determine the optimal number of trees that should be used in the random forest model.

## **2. Data Preparation**

Within this data set we are going to be focusing on the variables for heart disease, maximum heart rate, chest pain, exercised induced angina, resting blood pressure, age, sex, resting electrocardiograph measurement and cholesterol. In the full data set we have 14 rows and 303 columns.

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

### **Reporting Results**

**General Form:**

Where y is heart disease, is age, is resting blood pressure, is exercised induced angina and is maximum heart rate.

**Prediction Regression Model:**

Where is heart disease, is age, is resting blood pressure, is exercised induced angina and is maximum heart rate.

**Prediction Model Equation:**

Where is heart disease, is age, is resting blood pressure, is exercised induced angina and is maximum heart rate.

**Natural Log of Odds Prediction Regression Model:**

Where is the odds of heart disease, is age, is resting blood pressure, is exercised induced angina and is maximum heart rate.

**Natural Log of Odds Prediction Model Equation:**

Where is the odds of heart disease, is age, is resting blood pressure, is exercised induced angina and is maximum heart rate.

represents the probability of having heart disease in the natural log of odds prediction equation. represents the odds of having heart disease in the natural log of odds prediction equation.The estimated coefficient for maximum heart rate is 0.0311 which means that on average the change in log odds of having heart disease is 0.0311 per unit increase in maximum heart rate. We can get more specific by putting this in terms of odds instead of log odds by calculating . This means that the odds of having heart disease increase by 3.16% per unit increase in maximum heart rate.

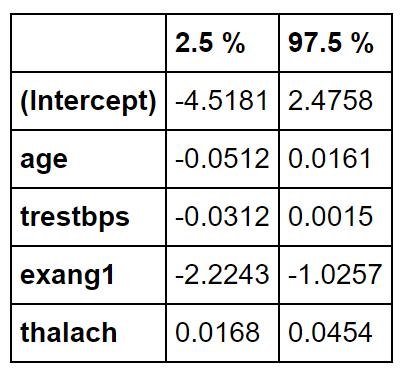
### **Evaluating Model Significance**

**GOF Test:**

| **Test Statistic** | **Null Hypothesis** | **Alternative Hypothesis** | **P-Value** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- | --- |
| X-Squared = 44.622 |  |  | 0.612 | Fail to Reject |

Given a significance level of 5% or 0.05 and a P-value that is above this level at 0.612 we must conclude that we fail to reject the null hypothesis and that the model is likely a good fit for the data.

**95% Confidence Interval**

****

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for age |  |  | Fail to reject |

The 95% confidence interval for age is between -0.0512 and 0.0161. Since 0 is within this range we must not reject the null hypothesis and conclude that there is at least a 5% chance 0 is within the interval and that therefore age is not significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for resting blood pressure |  |  | Fail to reject |

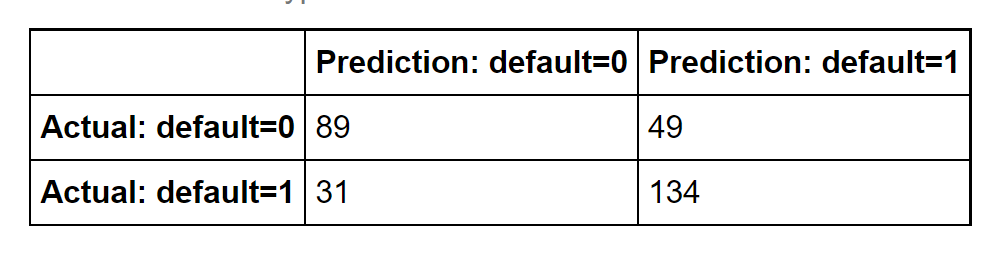
The 95% confidence interval for resting blood pressure is between -0.0312 and 0.0015. Since 0 is within this range we must fail to reject the null hypothesis and conclude that there is at least a 5% chance 0 is within the interval and that therefore the variable for resting blood pressure is not significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for exercise induced angina |  |  | Reject |

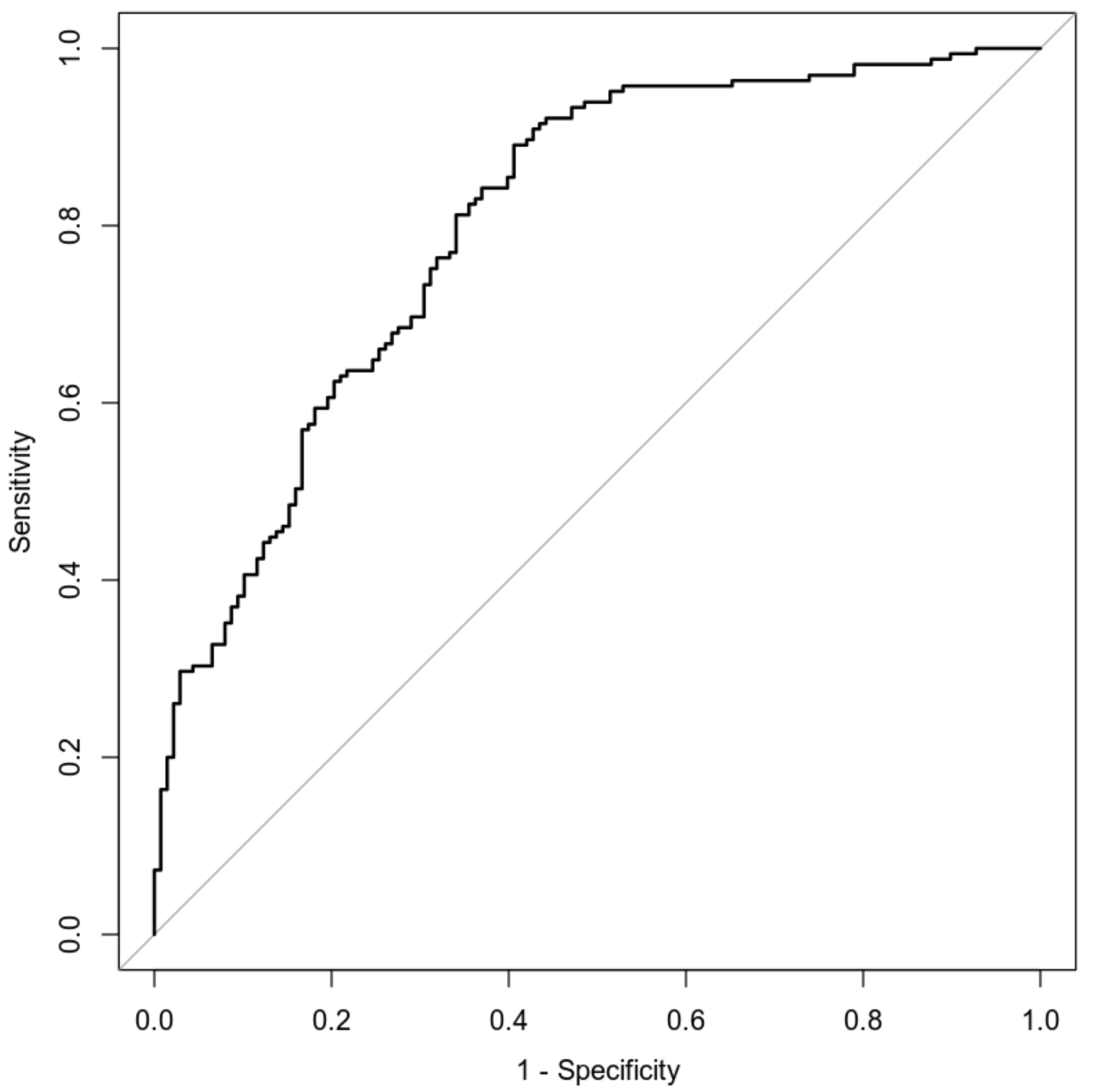
The 95% confidence interval for exercise induced angina is between -2.2243 and -1.0257. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the variable for exercise induced angina is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for maximum heart rate |  |  | Reject |

The 95% confidence interval for credit utilization is between 0.0168 and 0.0454. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the variable for maximum heart rate is significant within this model.



In the training set there are 134 true positives, 89 true negatives, 31 false positives and 49 false negatives.



This graph shows us the true positive rate on the y-axis and the true negative rate on the x-axis so ideally we would want a perfect right angle with the ends ending at zero. We have quite a steep curve and almost no straight sections even at the beginning and end points. This means that model is likely not very accurate and we can show this through the value for area under the curve or AUC, the value for AUC here is 0.8007 this means that our model has an 80.07% chance of making a correct prediction.

### **Making Predictions Using Model**

If we take a person who is 50 years old, has a resting blood pressure of 122, exercised induced angina and a maximum heart rate of 140 then run their statistics through our model we get a prediction of 27.16% for the probability of developing heart disease. To find the odds of this happening we solve for where p is the probability. Now we can get meaning that the odds of this individual developing heart disease is 679 to 1821.Considering an individual who is 50, has a resting blood pressure of 130, does not have exercise induced angina and has a maximum heart rate of 165 if we took them and ran their statistics through our model we would get a prediction of 78.53% probability of this person developing heart disease. To find the odds of this happening we solve for where p is the probability. Now we can get meaning that the odds of this individual developing heart disease is 7853 to 2147.Based on the probability and odds for the first individual I would say that they have a relatively low chance of developing heart disease while the second individual has a very high chance of developing heart disease.

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

### **Reporting Results**

**General Form:**

Where y is heart disease, is age, is resting blood pressure, , and are dummy variables for chest pain and is maximum heart rate.

**Prediction Regression Model:**

Where is heart disease, is age, is resting blood pressure, , and are dummy variables for chest pain and is maximum heart rate.

**Prediction Model Equation:**

Where is heart disease, is age, is resting blood pressure, , and are dummy variables for chest pain and is maximum heart rate.

**Natural Log of Odds Prediction Regression Model:**

Where is the odds of heart disease, is age, is resting blood pressure, , and are dummy variables for chest pain and is maximum heart rate.

**Natural Log of Odds Prediction Model Equation:**

Where is the odds of heart disease, is age, is resting blood pressure, , and are dummy variables for chest pain and is maximum heart rate.

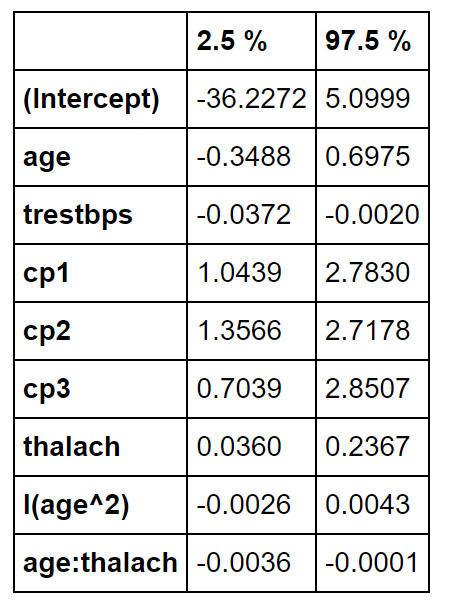
### **Evaluating Model Significance**

**Hosmer-Lemeshow GOF Test:**

| **Test Statistic** | **Null Hypothesis** | **Alternative Hypothesis** | **P-Value** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- | --- |
| X-Squared = 52 |  |  | 0.3209 | Reject |

Given a significance level of 5% or 0.05 and a P-value that is below this level at 0.3209 we must conclude that we must reject the null hypothesis and that the model is likely not a good fit for the data.

**95% Confidence Interval**

****

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for age |  |  | Fail to reject |

The 95% confidence interval for age is between -0.3488 and 0.6975. Since 0 is within this range we must fail to reject the null hypothesis and conclude that there is at least a 5% chance 0 is within the interval and that therefore age is not significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for resting blood pressure |  |  | Reject |

The 95% confidence interval for resting blood pressure is between -0.0372 and -0.0020. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that therefore the variable for resting blood pressure is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for dummy variable for chest pain |  |  | Reject |

The 95% confidence interval on the dummy variable for chest pain is between 1.0439 and 2.7830. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the dummy variable for chest pain is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for dummy variable for chest pain |  |  | Reject |

The 95% confidence interval on the dummy variable for chest pain is between 1.3566 and 2.7178. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the dummy variable for chest pain is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for dummy variable for chest pain |  |  | Reject |

The 95% confidence interval on the dummy variable for chest pain is between 0.7039 and 2.8507. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the dummy variable for chest pain is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for maximum heart rate |  |  | Reject |

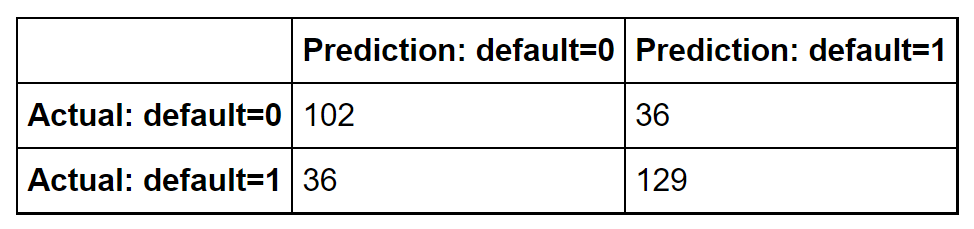
The 95% confidence interval for maximum heart rate is between 0.0360 and 0.2367. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the variable for maximum heart rate is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for interaction between age and maximum heart rate |  |  | Reject |

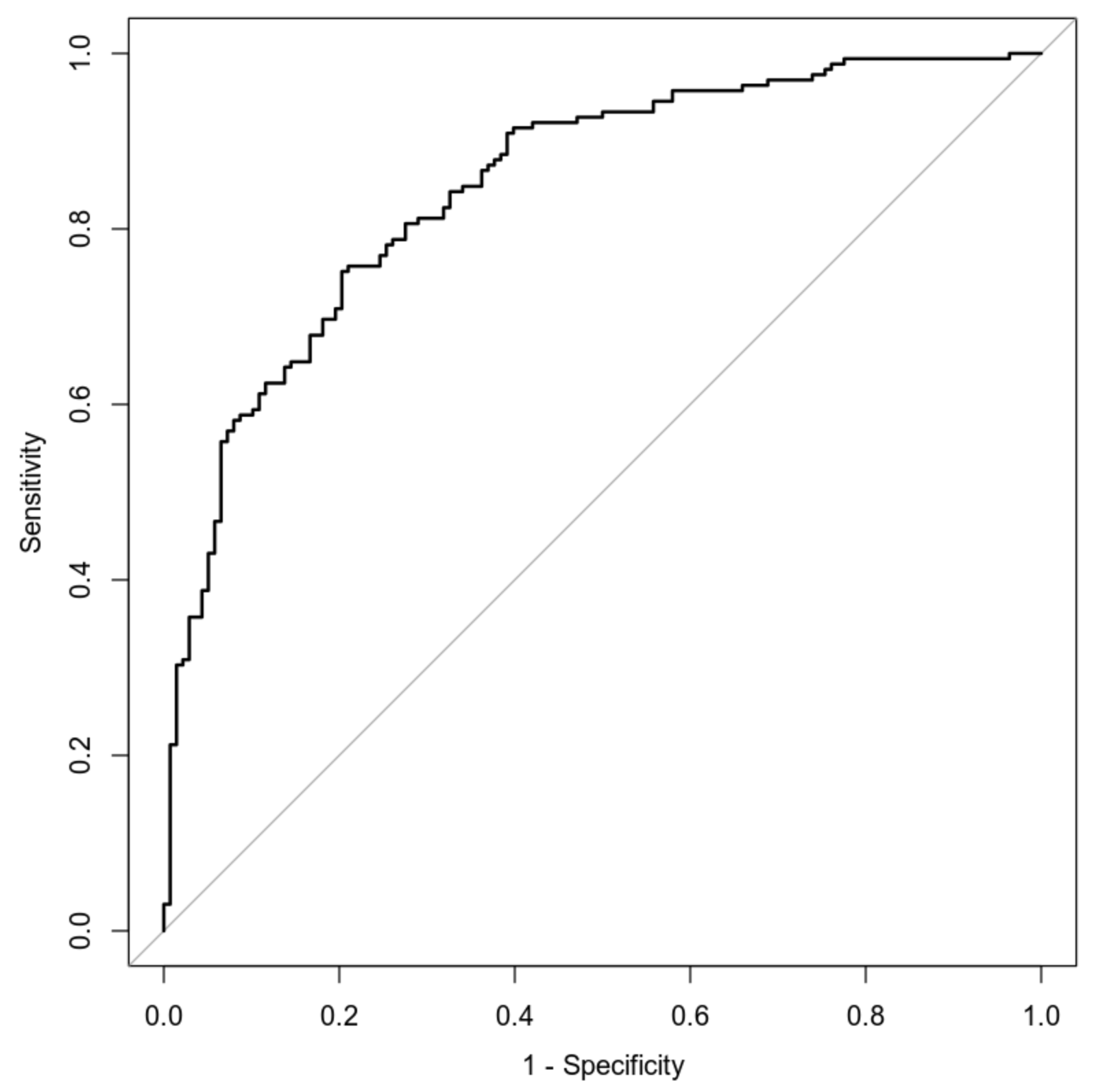
The 95% confidence interval for the interaction between age and maximum heart rate is between -0.0036 and -0.0001. Since 0 is not within this range we must reject the null hypothesis and conclude that there is less than a 5% chance 0 is within the interval and that the variable for the interaction between age and maximum heart rate is significant within this model.

| **Test Summary** | **Null Hypothesis** | **Alternative Hypothesis** | **Reject Null Hypothesis or Not** |
| --- | --- | --- | --- |
| Wald’s test for squared term for age |  |  | Fail to reject |

The 95% confidence interval for the squared term for age is between -0.0026 and 0.0043. Since 0 is within this range we must fail to reject the null hypothesis and conclude that there is at least a 5% chance 0 is within the interval and that the variable for the squared term for age is not significant within this model.

**

In the training set there are 129 true positives, 102 true negatives, 36 false positives and 36 false negatives.



This graph shows us the true positive rate on the y-axis and the true negative rate on the x-axis therefore we ideally want to see straight lines coming from zero on the x-axis and 1 on the y-axis converging at a right angle. Here we have a decent amount of curvature but it does immediately look better than the previous model. It still seems like there will be a decent amount of error as the line very quickly becomes jagged from both directions where normally you would want some straight away section at the beginning. Overall based on the look of this graph it seems like the model will be decent but still produce a significant amount of error. The value for the area under the curve or AUC here is 0.8478 and this means that the model has an 84.78% chance of making a correct prediction.

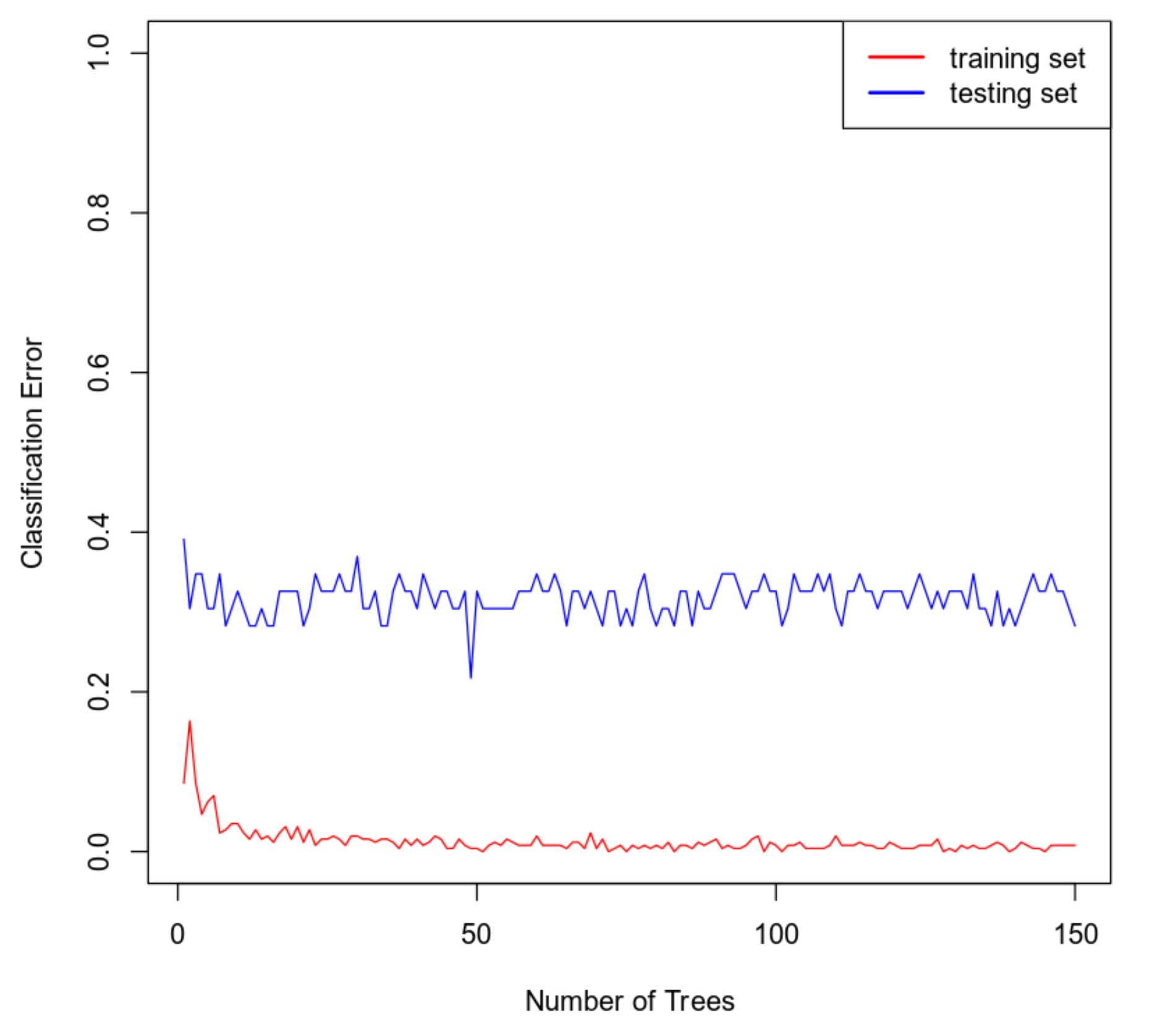
### **Making Predictions Using Model**

Using the second regression model we could plug 50 in the variable for age, 115 in the variable for resting blood pressure, no chest pain into the variable for chest pain and a 133 in the variable for maximum heart rate the model would predict that the probability of this individual having heart disease is 21.88%. To find the odds of this happening we solve for where p is the probability. Now we can get meaning that the odds of this individual developing heart disease is 547 to 1953.Based on the second regression model an individual who is 50 years old, has a resting blood pressure of 125, experiences typical angina and has a maximum heart rate of 155 would have a probability of 80.07% for developing heart disease. To find the odds of this happening we solve for where p is the probability. Now we can get meaning that the odds of this individual developing heart disease is 8007 to 1993. Given the probabilities and odds for the first individual I would say that they have a rather low chance of developing heart disease while the second individual has a rather high chance of developing heart disease

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

### **Reporting Results**

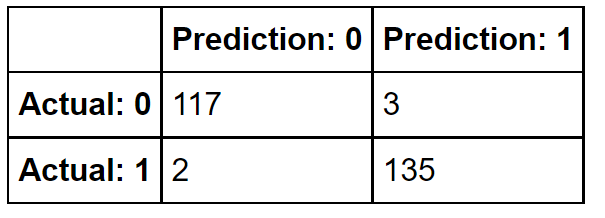
After taking my dataset and splitting it with 85% going to the training set and 15% going to the testing set we get 257 rows in the training set and 46 rows in the testing set, creating a total of 303 rows in the original dataset which matches the number we got at the start of the report.



We can tell the optimal number of trees by finding the point at which the testing set data stabilizes and looking at the graph above it appears we start to get some consistency around 60 trees. After this point the classification error seems to hit a consistent minimum and maximum point.

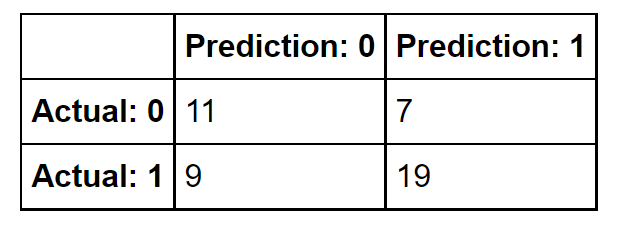
### **Evaluating the Utility of the model**

**Training Set Confusion Matrix:**



In the training set there are 135 true positives, 117 true negatives, 3 false positives and 2 false negatives.

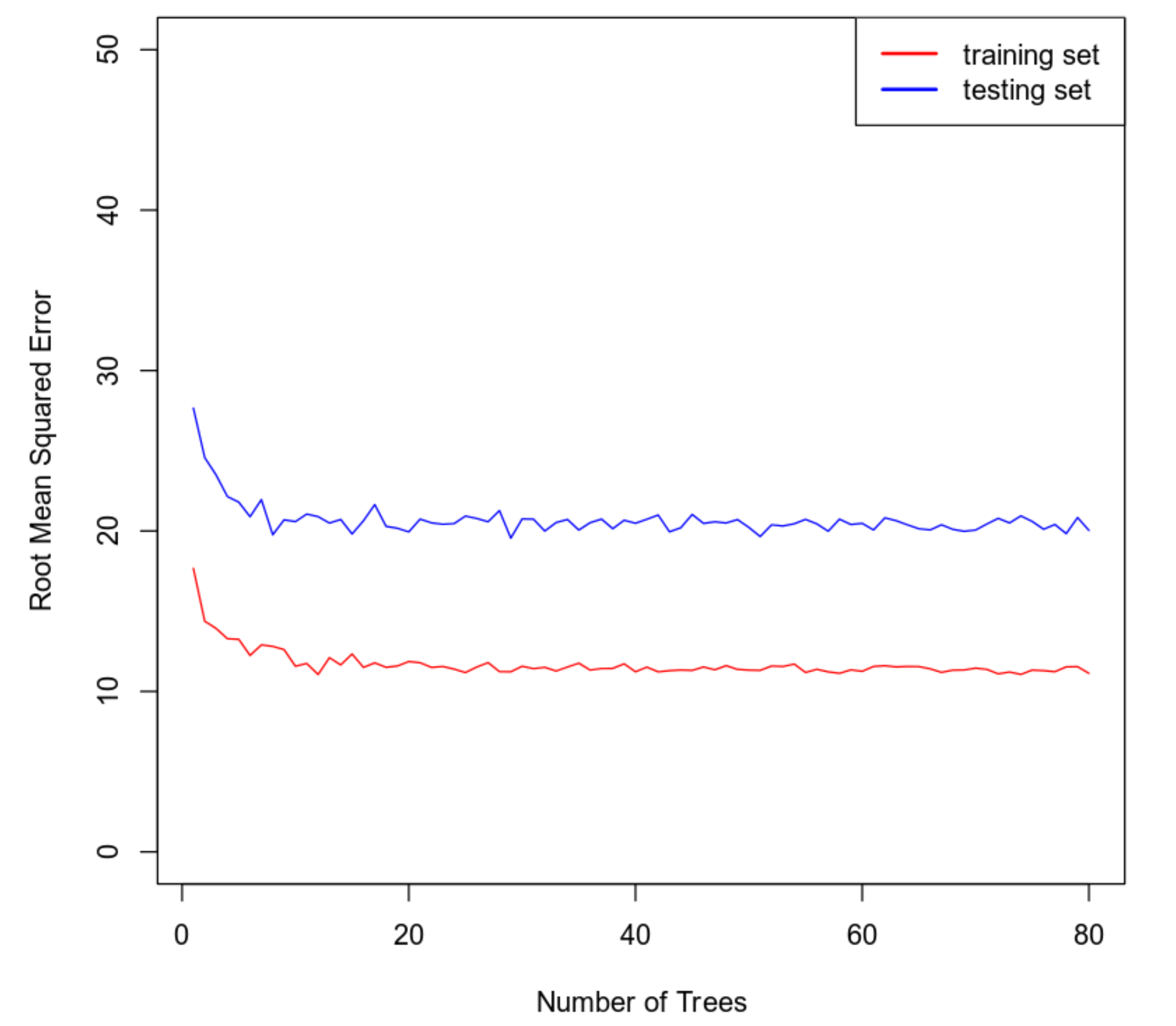
**Testing Set Confusion Matrix:**

  
In the testing set there are 11 true positives, 19 true negatives, 7 false positives and 9 false negatives.

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

### **Reporting Results**

After splitting the data set with 80% going to the training set and 20% going to the testing set we get 242 rows in the training set and 61 rows in the testing set, this makes for a total of 303 rows in the original dataset which matches the number we got at the start of the paper.



Based on the graph above the optimal number of trees to use in this random forest model would be 30. This is because we are looking for the point where the error line for the testing set starts to normalize and it seems like we pass most of the deviation at about the 30 tree mark so that seems like a good place to get our optimal tree number from.

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

The root mean squared error for the random forest regression model on the training set is 11.5619. We want our number to be as close as possible to zero so this is too far off but ideally we would like it to be lower. The root mean squared error in the testing set is 20.7872. We expect this to be higher than the error on the training data because it is not familiar with the training data. We also want this to be as close as possible to zero so while not extremely high I think that we could do better.

## **7. Conclusion**

Out of the two logistic regression models it would seem as though the second one is much better suited at predicting heart disease. The first reason for this is that the value for AUC on the second model is higher at 0.8478 meaning an 84.78% chance of making a correct prediction while the first model only had a value of 0.8007 for AUC meaning that it has a 80.07% chance of making a correct prediction. The next reason I have to choose the second model is the results of the Wald’s tests that I ran in the first model two out of the four variables were not statistically significant while in the second model only 2 out of 8 terms were deemed to be insignificant. Model 2 also has a higher degree of accuracy with the value coming out to 0.7624 while in model one we only get an accuracy of 0.7359. So while this is a rather close call it seems like the second model is slightly better at predicting heart disease than the first. Based on the results of this report I would recommend the logistic regression model over the random forest classification model. The main reason I have for this is that the accuracy, precision and recall scores are all higher on the preferred logistic regression model than in the random forest classification model. In the second logistic regression model we have a score of 0.7624 for accuracy and 0.7818 for both precision and recall. While in the random forest classification model we have scores of 0.6523 for accuracy, 0.6111 for precision and 0.55 for recall on the testing set which is what is important as it is data the model has not seen before. SInce the values are higher in all categories that evaluate utility of the model I must conclude that the logistic regression model should be preferred over the random tree classification model. The practical importance of these tests is to ensure that the model we choose is accurate enough for our purposes producing statistically significant results. This is important because if this model were to be used to help diagnose patients we would want to make sure that it is the best model available for the job and we can only confirm that through these analyses.

## **8. Citations**

*Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R. (1988). Heart disease data set [Data file and codebook]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease*