# **MAT 303 Project Two Summary Report**

Camilo Hoyos

camilo.hoyos@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The data set to be explored relates to risk factors for heart disease at a hospital. This contains patient data without any personal identifiable information. The results can be used to predict the probability of a patient having heart disease based on other health related factors. The analyses being executed will be two logistic regressions and then a random forest classification.

## **2. Data Preparation**

The important variables in this data set are age (age), resting blood pressure (trestbps), maximum heart rate (thalach), sex (sex), exercise-induced angina (exang), type of chest pain (cp), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), slope of peak exercise (slope), and number of major vessels (ca). There are a total of 14 variables and 303 rows in this dataset.

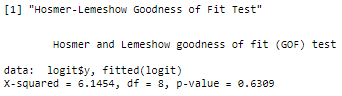
## **3. Model #1 - First Logistic Regression Model**

### **Reporting Results**

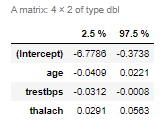
The general form of the multiple regression model is: , where X1 is age, X2 is resting blood pressure, and X3 is maximum heart rate achieved. The model in natural log of odds is: . is the odds that someone has heart disease and are the inverse odds that someone has heart disease, or someone who does not have heart disease. The logistic regression model is: and in the terms of natural log of odds it is: . The estimated coefficient of maximum heart rate achieved is 0.0427. This means that on average the change in log odds for having heart disease is 0.000427 for each percentage increase in maximum heart rate assuming all other variables are constant.

### **Evaluating Model Significance**

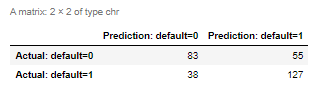
The null hypothesis (H0) is that there is no relationship between an individual having heart disease, their age, their resting blood pressure, and their maximum heart rate achieved. The alternative hypothesis (Ha) is that there is a relationship between an individual having heart disease, their age, their resting blood pressure, and their maximum heart rate achieved that a prediction could be made on whether or not they have heart disease. The Hosmer-Lemeshow goodness of fit test shows a p-value of 0.6309 which does not pass a 5% level of significance test.



Wald’s test helps to determine which terms are significant in a logistic model. As it stands the only term that could theoretically have a slope of 0 is age (-0.0409, 0.0221) and does not pass a 5% level of significance.



The confusion matrix below:

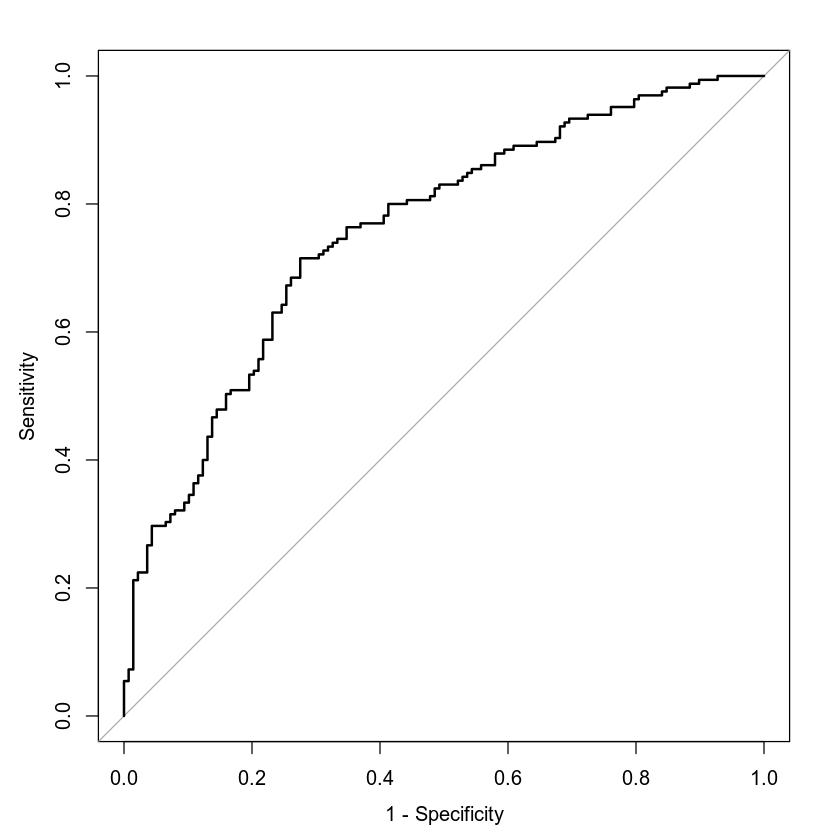


Accuracy: = =or 69.3% of correct predictions to total number of observations.

Precision: = = or 69.78% of correct positive predictions to the total predicted positives.

Recall: = = or 76.97% of correct positive predictions to the total positives examples.

The Receiver Operating Characteristic (ROC) curve is displayed below. The ROC identifies how well the model predicts the classes of 0 or 1. The larger the area under the curve, the more accurate it is. This graph has an area under the curve (AUC) of .7575 which states that 75.75% of the fitted values fall within the curve.



### **Making Predictions Using Model**

The probability of an individual who is 50 years old, has a resting blood pressure of 122, and a maximum heart rate of 140 will have heart disease is 0.4939 or 49.39%. Their odds of having heart disease is 1 to 0.98, or the patient is 0.98 times as likely to have heart disease.

The probability of an individual who is 50 years old, has a resting blood pressure of 140, and a maximum heart rate of 170 will have heart disease is 0.7248 or 72.48%. Their odds of having heart disease is 1 to 2.63, or the patient is 2.63 times as likely to have heart disease.

By reducing blood pressure from 140 to 122 and maximum heart rate from 170 to 140 can reduce the probability of having heart disease by 20%.

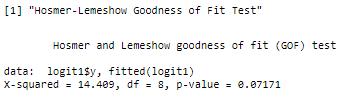
## **4. Model #2 - Second Logistic Regression Model**

### **Reporting Results**

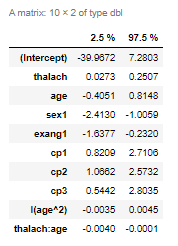
The general form of logistic regression model is: , where X1 is maximum heart rate, X2 is age, X3 is sex of the individual (1=male), X4 is exercise-induced angina, X5 is typical angina, X6 is atypical angina, X7 is non-anginal pain, and X8 is the interaction between age and maximum heart rate. The model in natural log of odds is: . The logistic regression model is: and in the terms of natural log of odds it is:

### **Evaluating Model Significance**

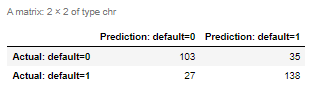
The null hypothesis (H0) is that there is no relationship between an individual having heart disease, their maximum heart rate, age, sex, exercise-induced angina, type of chest pain, and the interaction of their maximum heart rate and age. The alternative hypothesis (Ha) is that there is a relationship between an individual having heart disease, their maximum heart rate, age, sex, exercise-induced angina, type of chest pain, and the interaction of their maximum heart rate and age such that a prediction could be made on whether they have heart disease. The Hosmer-Lemeshow goodness of fit test displays a p-value of 0.0717 which does not pass a 5% level of significance test.

**

Wald’s test provides which terms are more significant than others. As it stands, the intercept and age are failing a 5% level of significance. The intercept (-39.9672, 7.2803) and age (-0.4051, 0.8148) can have a slope of 0.



Confusion Matrix:

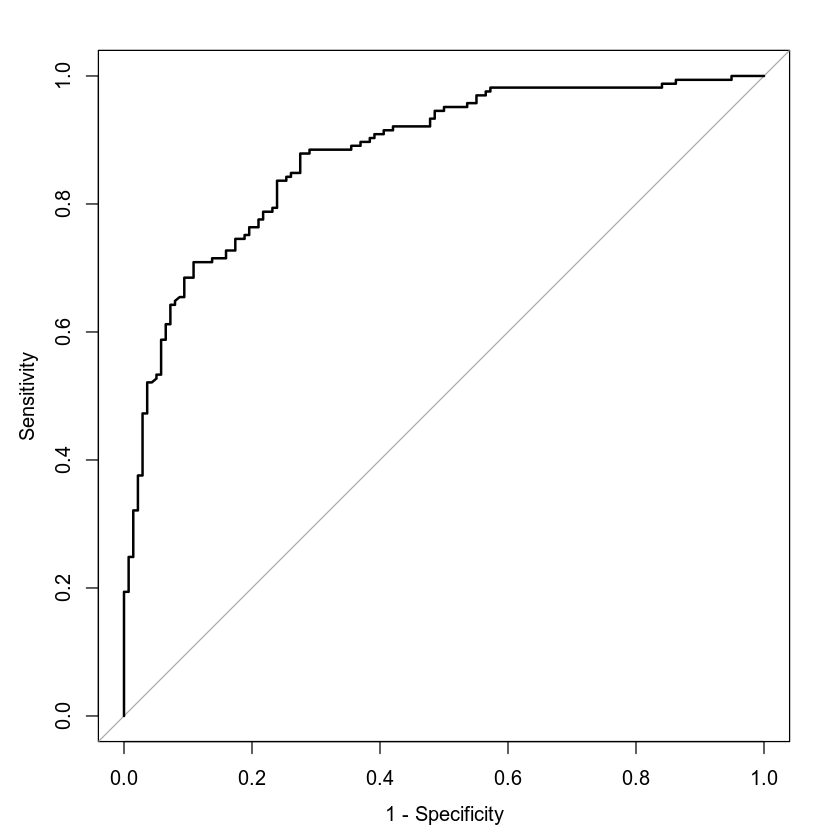


Accuracy: = =or 79.54% of correct predictions to total number of observations.

Precision: = = or 79.77% of correct positive predictions to the total predicted positives.

Recall: = = or 83.64% of correct positive predictions to the total positives examples.

The Receiver Operating Characteristic (ROC) curve is displayed below. The ROC identifies how well the model predicts the classes of 0 or 1. The larger the area under the curve, the more accurate it is. This graph has an area under the curve (AUC) of .8777 which states that 87.77% of the fitted values fall within the curve.



### **Making Predictions Using Model**

The probability of a male having heart disease who is 30 years old, has a maximum heart rate of 145, experiences exercise-induced angina, and does not experience other chest pains is 0.2654 or 26.54%. Their odds of having heart disease is 1 to 0.36 or the patient is 0.36 times as likely to have heart disease.

The probability of a male having heart disease who is 30 years old, has a maximum heart rate of 145, does not experience exercise-induced angina, and does experience typical angina is 0.8432 or 84.32%. Their odds of having heart disease is 1 to 5.38 or the patient is 5.38 times as likely to have heart disease.

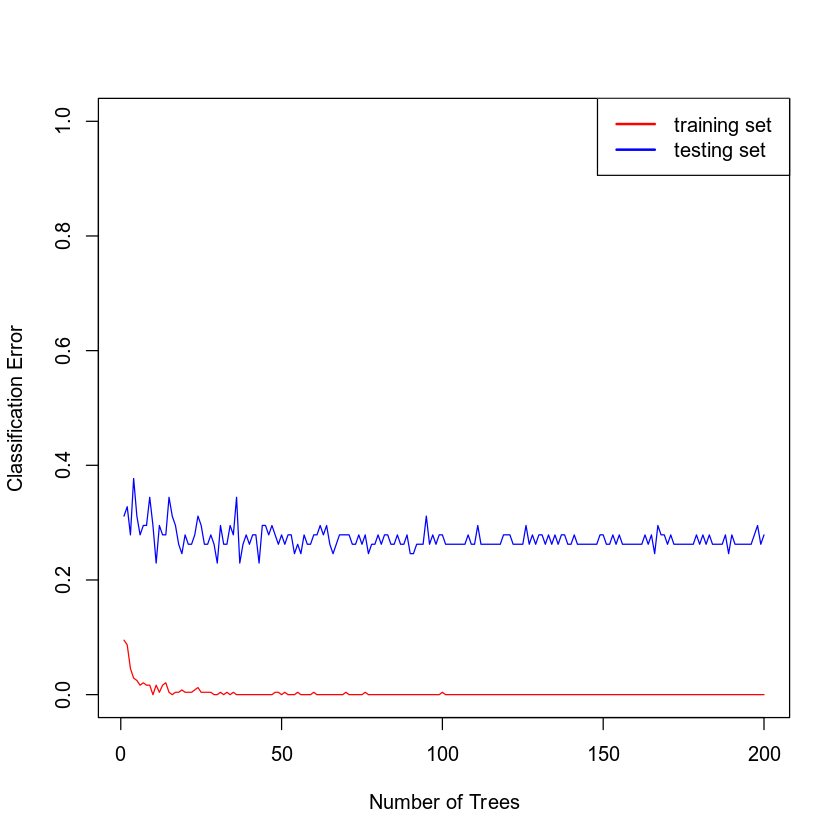
There is a massive difference in the probability of a 30 year old male having heart disease if they experience typical angina but not exercise-induced angina versus experiencing exercise-induced angina but no other chest pains. Namely, the ‘exercise-induced angina’ individual has over 50% reduced probability of having heart disease; from 84.32% to 26.54%.

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

### **Reporting Results**

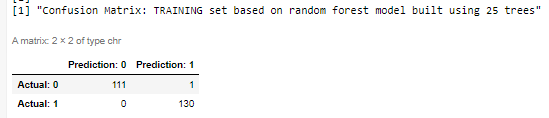
The original data set has 303 rows. The dataset has been split to have 242 records in the training set and 61 records in the validation set.

From below you can see the training and testing set graphed on 200 trees. At approximately 25 trees, the classification error flattens so we can assume 25 trees are optimal for this random forest model.



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

Confusion Matrix (training):

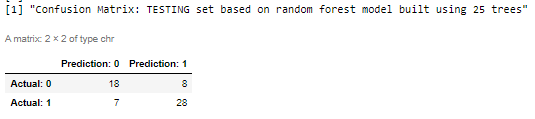


Accuracy: = =or 99.59% of correct predictions to total number of observations.

Precision: = = or 99.24% of correct positive predictions to the total predicted positives.

Recall: = = or 100% of correct positive predictions to the total positives examples.

Confusion Matrix (testing):



Accuracy: = =or 75.41% of correct predictions to total number of observations.

Precision: = = or 77.78% of correct positive predictions to the total predicted positives.

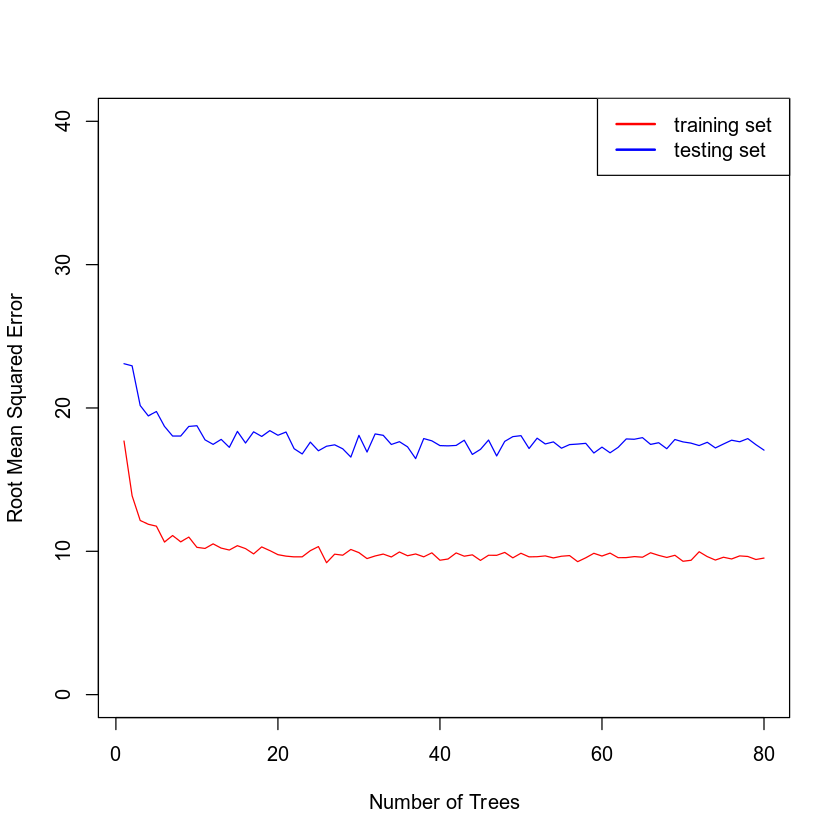
Recall: = = or 80% of correct positive predictions to the total positives examples.

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

### **Reporting Results**

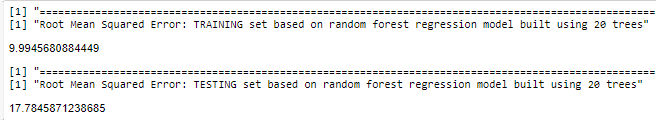
The original data set has 303 rows. The dataset has been split to have 242 records in the training set and 61 records in the validation set.

The mean squared error against the number of trees is below. The optimal number of trees is approximately 20.



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

The root mean squared error for the training set is 9.9946 and the root mean squared error for the testing set is 17.7846. This is the standard deviation of residuals, or how much they deviate from the regression line. The lower the value, the more accurate the model is.

**

## **7. Conclusion**

Of the two logistic regression models evaluated, the second logistic model has a better fit than the first. It contains a better ROC as it contains roughly 87% of the fitted values in the curve. Additionally it has a higher accuracy, precision, and recall comparatively to the first model. Lastly, it doesn’t offend the p-value of the Hosmer-Lemeshow goodness of fit as much as the first; the second model has a p-value of 0.0717 whereas the p-value of the first is 0.6309! In the end; neither of these models are near the statistical evaluation that would be preferred, such as passing the Hosmer-Lemeshow fit.

In terms of comparing the random forest classification model against the logistic regression model, the comparable tests are the confusion matrices from both models. The logistic model shows higher accuracy (79.54%), precision (79.77%), and recall (83.64%) than the random forest which has 75.41%, 77.78%, and 80%, respectively. Due to the higher statistical evaluation on the confusion matrix, I would choose the logistic regression model.

The analyses performed allowed comparison of two logistic models in terms of which predictor variables to include. We saw that the proper inclusion of more appropriate terms increased a better fit. Then we executed a random forest classification model that shared a fit test with logistic models. From this last model we were then capable of comparing which of the three were more accurate in predicting which patients would have heart disease.