# MAT 303 Project Two Summary Report

Joshua Wozny

joshua.wozny@snhu.edu

Southern New Hampshire University

## 1. Introduction

Our employer, a university hospital, has tasked us with studying risk factors for heart disease, and model how these factors can be used to predict it. We will be using a provided set of historical data that lists different health indicators (fasting blood sugar, maximum heart rate, and so on) and the presence of heart disease. We will create different logistic regression and random forest models to predict whether or not a person is at risk for heart disease.

## 2. Data Preparation

The provided heart disease dataset includes our response variable for presence of heart disease (target, 0=no, 1=yes), and the following possible quantitative predictor variables: age (age), resting blood pressure (trestbps), cholesterol measurement (chol), maximum heart rate (thalach), and ST depression (oldpeak). It also includes the following qualitative variables: sex (1=male, 0=female) cp (chest pain 0=no pain, 1=typical angina, 2=atypical angina, 3=non-anginal pain), fbs (fasting blood sugar greater than 120 1= true, 0=false), restecg (resting electrocardiogram 0=normal, 1=ST-T wave abnormality, 2=showing probable ventricular hypertrophy), exang (exercise induced angina 1=yes, 0=no), slope (of peak exercise ST 1=upsloping, 2=flat, 3=downsloping), and ca (major vessels 0, 1, 2, 3) . There are 303 rows of data in our set.

| **Variable** | **What does it represent?** |
| --- | --- |
| age | The person's age in years |
| sex | The person's sex (1 = male, 0 = female) |
| cp | The type of chest pain experienced (0=no pain, 1=typical angina, 2=atypical angina, 3=non-anginal pain) |
| trestbps | The person's resting blood pressure |
| chol | The person's cholesterol measurement in mg/dl |
| fbs | The person's fasting blood sugar is greater than 120 mg/dl (1 = true, 0 = false) |
| restecg | Resting electrocardiographic measurement (0=normal, 1=having ST-T wave abnormality, 2=showing probable or definite left ventricular hypertrophy by Estes' criteria) |
| thalach | The person's maximum heart rate achieved |
| exang | Exercise-induced angina (1=yes, 0=no) |
| oldpeak | ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot) |
| slope | The slope of the peak exercise ST segment (1=upsloping, 2=flat, 3=downsloping) |
| ca | The number of major vessels (0-3) |
| target | Heart disease (0=no, 1=yes) |

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

### Reporting Results

The general form of our regression model1 is:

where y is 1 for having heart disease and 0 for not having heart disease, x1 is age, x2 is resting blood pressure, and x3 is maximum heart rate achieved.

Where odds is the odds of having heart disease (target=1). is the expected proportional response for the model, and is the odds of having heart disease.

Running our regression yields the following model:

where y is 1 for having heart disease and 0 for not having heart disease, x1 is age, x2 is resting blood pressure, and x3 is maximum heart rate achieved.

**Interpreting the estimated coefficient of maximum heart rate achieved.**

* The estimated coefficient for the variable of maximum heart rate achieved is 0.0427. This means that on average, the change in log odds for having heart disease is increase by 0.0427 for each additional beat per second, holding all other variables constant.
* An alternative way to express this is in terms of odds (and not log odds). If we calculate:  
  - 1 = 0.0436, indicating that the odds of having heart disease increase by 4.36% for each increase in heart rate achieved, holding all other variables constant.

### Evaluating Model Significance

We begin our evaluation of the model by performing the Hosmer-Lemehow goodness of fit test, where the null hypothesis is that the model fits the data and the alternative hypothesis is that the model does not fit the data. The chi-square value for our model is 41.978, with a p-value of 0.7168. At 5% level of significancel and a p-value well above 0.05, we cannot reject the null hypothesis, and conclude that the model does fit the data.

We next review the results of Wald’s test for each term within our model. At the 5% level of significance, both resting blood pressure and maximum achieved heart rate are significant with z-scores of -2.6063 and 6.144, and corresponding p-values of 0.0392 and 8.06e-10, respectively. At a 5% level of confidence, age is not significant with a z-score of -0.186 and p-value of 0.5578.

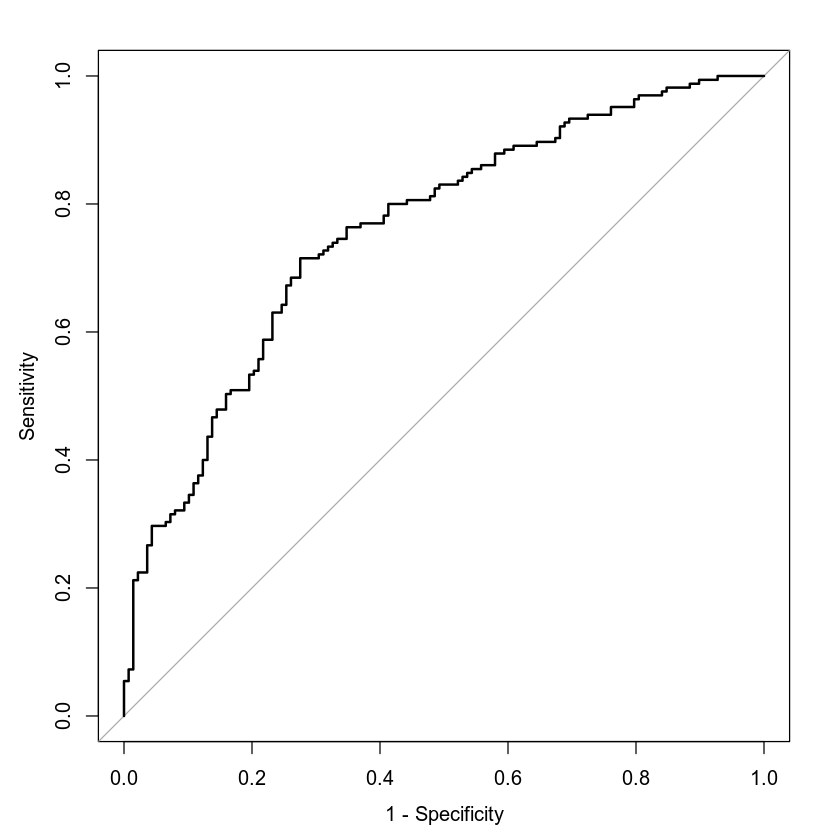
Our model produces this confusion matrix:

|  | **Prediction: target=0** | **Prediction: target=1** |
| --- | --- | --- |
| **Actual: target=0** | 83 | 55 |
| **Actual: target=1** | 38 | 127 |

We calculate the following statistics from the confusion matrix (True Positives: 127, True Negatives: 83, False Positives: 55, False Negatives: 38) from above:

* Accuracy: 0.6931
* Precision: 0.6978
* Recall: 0.7697

We continue our evaluation by reviewing the Receiver Operating Characteristic (ROC) curve. The ROC curve measures the performance of a logistic regression model at various threshold settings. This can be used to compare different models. This model has an area under the curve (AUC) of 0.7575. Generally a larger area under the curve represents a model that is better at predicting if someone has heart disease.



### Making Predictions Using Model

The probability that an individual who is 50 years old, has a resting blood pressure of 122, and has maximum heart rate of 140 having heart disease is 0.4939, with approximate odds of 1:1. A relatively high chance of having heart disease. The probability that an individual who is 50 years old, has a resting blood pressure of 140, and has maximum heart rate of 170 having heart disease is 0.7248, with approximate odds of 2.5:1. A significantly higher chance of having heart disease.

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

### Reporting Results

The general form of our regression model1 is:

where y is 1 for having heart disease and 0 for not having heart disease, x1 is age, x2, x3, and x4 are for chest pain = 1, 2, and 3 respectively, x5 is exercise induced angina = 1, x6 is sex =1 (male), and x7 is maximum heart rate achieved.

Where odds is the odds of having heart disease (target=1). is the expected proportional response for the model, and is the odds of having heart disease.

Running our regression yields the following model:

where y is 1 for having heart disease and 0 for not having heart disease, x1 is age, x2, x3, and x4 are for chest pain = 1, 2, and 3 respectively, x5 is exercise induced angina = 1, x6 is sex =1 (male), and x7 is maximum heart rate achieved.

### Evaluating Model Significance

We begin our evaluation of the model by performing the Hosmer-Lemehow goodness of fit test, where the null hypothesis is that the model fits the data and the alternative hypothesis is that the model does not fit the data. The chi-square value for our model is 60.596, with a p-value of 0.1048. At 5% level of significancel and a p-value above 0.05, we cannot reject the null hypothesis, and conclude that the model does fit the data.

We next review the results of Wald’s test for each term within our model. At the 5% level of significance, chest pain (cp=1, cp=2, cp=3), exercise induced angina, sex, maximum heart rate attained, and the interaction term between age and maximum heart rate achieved are all significant with z-scores of 3.663, 4.734, 2.904, -2.607, -4.762, 2.438, 0.240, and -2.017. And corresponding p-values of 0.000249, 2.21e-06, 0.003684, 0.009133, 1.91e-06, 0.01476, and .043666, respectively. At a 5% level of confidence, age and it’s quadratic square are not significant with z-scores of 0.658 and 0.810599. And p-value of 0.510325 and 0.810599, respectively.

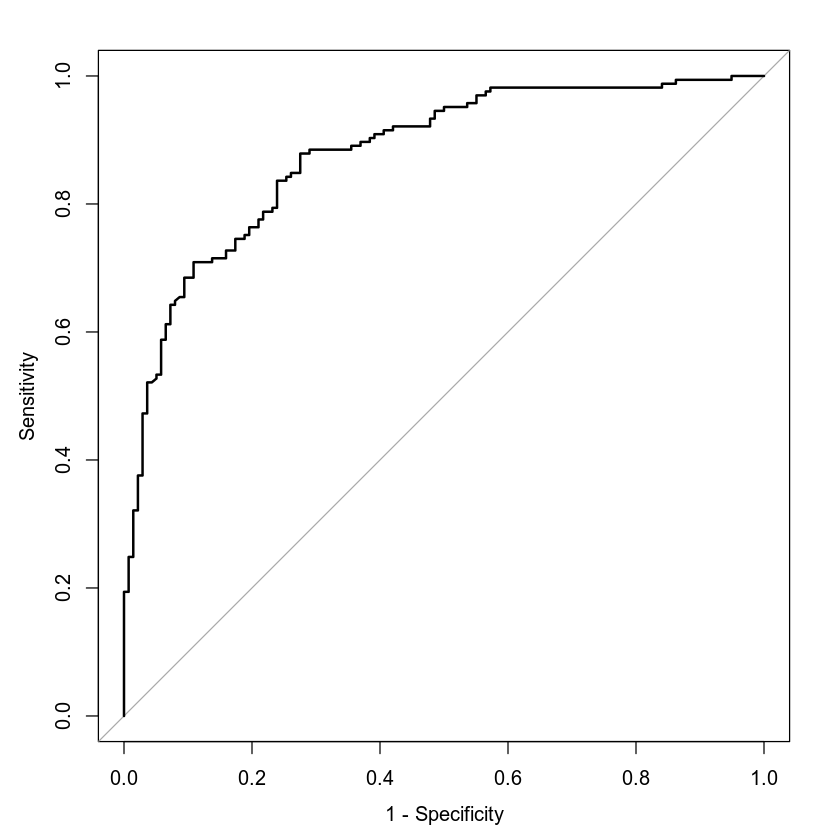
Our model produces this confusion matrix:

|  | **Prediction: target=0** | **Prediction: target=1** |
| --- | --- | --- |
| **Actual: target=0** | 103 | 35 |
| **Actual: target=1** | 27 | 138 |

We calculate the following statistics from the confusion matrix (True Positives: 138, True Negatives: 103, False Positives: 35, False Negatives: 27) from above:

* Accuracy: 0.7954
* Precision: 0.7977
* Recall: 0.8364

We continue our evaluation by reviewing the Receiver Operating Characteristic (ROC) curve. The ROC curve measures the performance of a logistic regression model at various threshold settings. This can be used to compare different models. This model has an area under the curve (AUC) of 0.8777. Generally a larger area under the curve represents a model that is better at predicting if someone has heart disease.



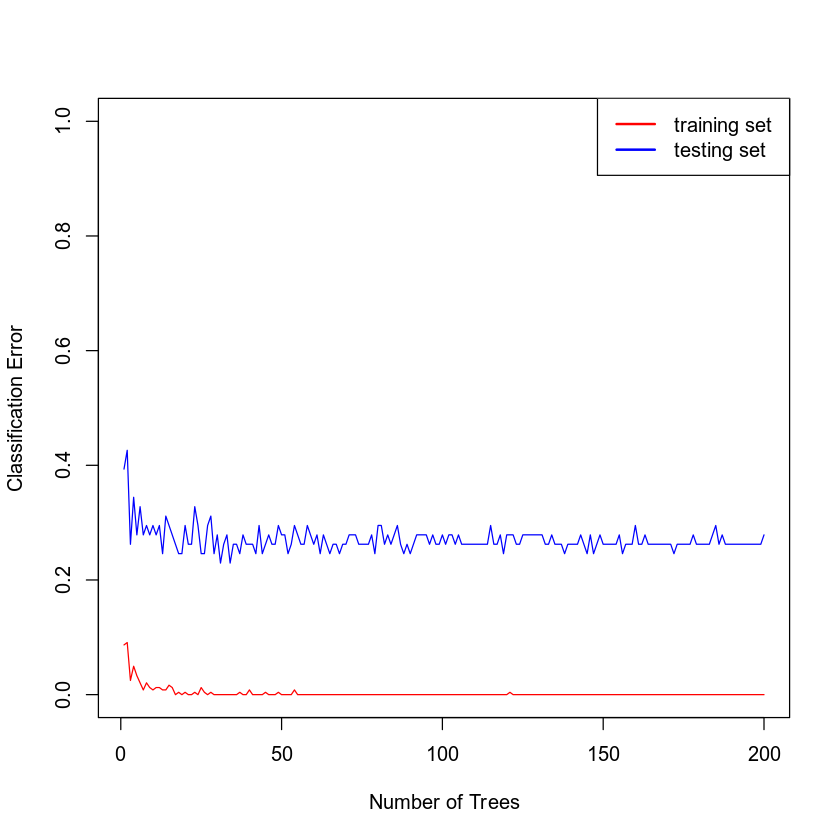
### Making Predictions Using Model

The probability that a male individual who is 30 years old; has a maximum heart rate of 145; experiences exercise-induced angina; and does not experience chest pain related to typical angina, atypical angina, or non-anginal pain heart disease is 0.2654, with approximate odds of 1:3. A relatively low chance of having heart disease. The probability that a male individual who is 30 years old, has a maximum heart rate of 145, and does not experience exercise-induced angina but experiences typical angina having heart disease is 0.8432, with approximate odds of 5.5:1. A very high chance of having heart disease.

## 5. Random Forest Classification Model

### Reporting Results

We start by splitting the dataset into training and testing sets with a 80% and 20% split, respectively. Our original dataset has 303 rows; our training set has 242 rows, and the testing set, 61. Plotting the training and testing error against the number of trees using a classification random forest model yields:



Examining the graph, it appears that it levels off somewhere before 50 trees. Beyond this, the error is roughly the same even if the number of trees increases. We will therefore use that as the optimal number of trees.

### Evaluating the Utility of the model

Using the testing data set we create the confusion matrix to calculate Accuracy, Precision, and Recall. Producing the confusion matrix yields:

|  | **Prediction: 0** | **Prediction: 1** |
| --- | --- | --- |
| **Actual: 0** | 18 | 8 |
| **Actual: 1** | 8 | 27 |

We calculate the following statistics from the confusion matrix (True Positives: 27, True Negatives: 18, False Positives: 8, False Negatives: 8) from above:

* Accuracy: 0.7377
* Precision: 0.7714
* Recall: 0.7714

Using the training data set our confusion matrix instead yields:

|  | **Prediction: 0** | **Prediction: 1** |
| --- | --- | --- |
| **Actual: 0** | 110 | 2 |
| **Actual: 1** | 0 | 130 |

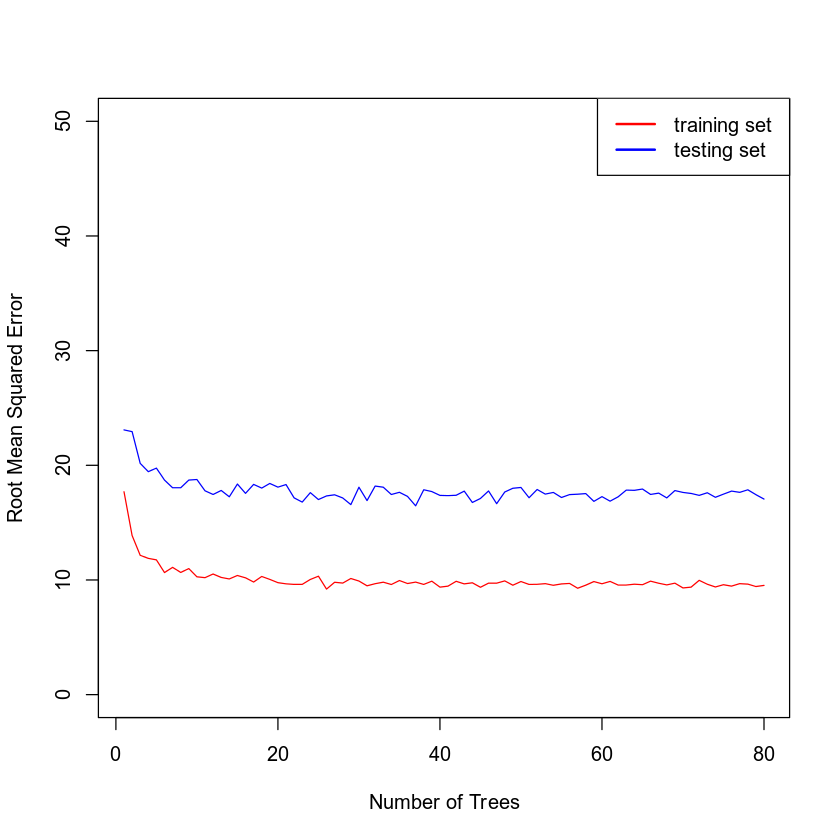
We calculate the following statistics from thisconfusion matrix (True Positives: 130, True Negatives: 110, False Positives: 2, False Negatives: 0) from above:

* Accuracy: 0.9917
* Precision: 0.9848
* Recall: 1.0000

## 6. Random Forest Regression Model

### Reporting Results

We start by splitting the dataset into training and testing sets with a 80% and 20% split, respectively. Our original dataset has 303 rows; our training set has 242 rows, and the testing set, 61. Plotting the the mean squared error against the number of trees using a classification random forest model yields:



Examining the graph, it appears that it levels off somewhere before 25 trees. Beyond this, the error is roughly the same even if the number of trees increases. We will therefore use that as the optimal number of trees.

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

Using 25 trees, we create a random forest regression model for maximum heart rate achieved using age (age), sex (sex), chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), slope of peak exercise (slope), and number of major vessels (ca). The Root Mean Squared Error: TRAINING set based on random forest model built using 25 trees is 9.5979. The Root Mean Squared Error: TESTING set based on random forest model built using 25 trees is 18.2216.

## 7. Conclusion

We created two logistic regression models to predict heart disease, each with a different set of predictor variables, the second model also included a interaction and quadratic terms. Comparing the accuracy, precision, and area under the curve between model 1 and model, we conclude that model 2 is superior. It has an accuracy of 0.7954, precision of 0.7977, and area under the curve of 0.8777, compared to model 1’s 0.6931, 0.63978, and 0.7575, respectively.

Comparing model 2 with model 3, our classification random forest model, I’d recommend model 2 based on the better accuracy and precision based on the testing set data. With an accuracy of 0.7954 and a precision of 0.7977 compared to model 3 of 0.7377 and 0.7714, respectively. I recommend collecting more data to create a larger testing set to determine if the performance of model 3 may actually perform better based on larger sample sizes.

All analyses performed were designed to tune the performance of our logistic regression and random forest models while minimizing problems with overfitting the model to our training data. The confusion matrix, Hosmer-Lemehow goodness of fit test, and the Wald’s test for individual variables allow us to judge whow well the logistic regression models predict the outcome of the response variable. Tests are performed for each type to judge performance and compare to other models: The confusion matrix for Classification Decision Trees and Root Mean Squared Error (RMSE) for Regression Decision Trees serve a similar purpose.

## 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>

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti,

Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011).

pROC: an open-source package for R and S+ to analyze and compare ROC

curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77

<http://www.biomedcentral.com/1471-2105/12/77/>