# MAT 303 Project Two Summary Report

Zachary Tankersley

[zachary.tankersley@snhu.edu](mailto:zachary.tankersley@snhu.edu)

Southern New Hampshire University

## 1. Introduction

To create statistical models that will help predict risk profiles for patient heart disease, an historical data set named “heart\_disease” (stored in a .csv format) will be analyzed. The results of these analyses will be used to spot risk factors that doctors may not easily detect. Two logistic regression models and both a decision and regression random forest tree model will be prepared for analysis.

## 2. Data Preparation

Variables that will be important for this analysis are target (heart disease), age, resting blood pressure (trestbps), maximum heart rate achieved (thalach), sex, exercise-induced angina (exang), type of chest pain (cp), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), number of major vessels (ca), and slope of peak exercise (slope). There are 303 rows of data with 14 variable columns.

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

### Reporting Results

The general form of the multiple regression model for heart disease using the predictor variables age, resting blood pressure, and max heart rate achieved is:

This model can also be expressed in linear terms via the natural log of odds form which is:

Where is the probability of an event, and represents the odds of the event happening. Using the summary analysis function of R to receive the beta parameters, the logistic regression model can be built and is represented below:

While the linear form can be represented in the following form:

The coefficient of maximum heart rate achieved (thalach) is 0.042697 which indicates that if all other variables are held constant, for every unit increase in thalach, the odds of heart disease increase by about 0.0427.

### Evaluating Model Significance

To determine whether or not this model fits the data well, a Hosmer-Lemeshow goodness of fit test is performed, however, the null and alternative hypotheses need to be set as follows:

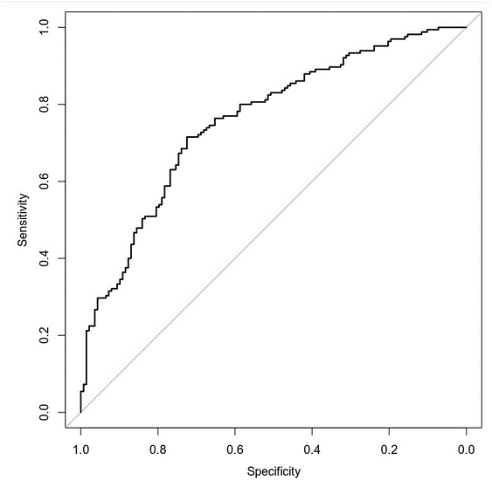
The test statistic for this test or , is 41.987 and the P-value is 0.7168. Judging against a 5% significance level or 0.05, 0.7168 is much higher, indicating in this case that the null hypothesis is confirmed and that the model *does* fit the data. To use a Wald’s test of significance on the beta parameters, null and alternative hypotheses to compare against are represented as follows:

With the P-value estimates for the beta parameters that were received from the summary analysis function, the only predictor that can be rejected is age with a p-value well above 0.05 at 0.5578. The confusion matrix which shows the true and false positive and true and false negative values is:

|  |  |  |
| --- | --- | --- |
|  | Prediction: 0 | Prediction: 1 |
| Actual: 0 | 83 (TN) | 55 (FP) |
| Actual: 1 | 38 (FN) | 127 (TP) |

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

To visually represent the model’s accuracy, an ROC curve (receiver operating characteristic) can be plotted, and an AUC (area under the curve) can be calculated which will show how many data points are accurately predicted (1 or 0 in this case):



The AUC for this model is ~0.7575 or about 75.75%. The higher the AUC the better the model is at predicting a true positive or negative value.

### Making Predictions Using Model

Using this model to make predictions, the probability of a person who is 50 years old, has a resting blood pressure of 122, and a max heart rate of 140, is about 49.39%, and the odds would be very nearly 50/50 or 1:1. For an individual of the same age, with a resting blood pressure of 140 and a max heart rate of 170, the probability is about 72.48% and the odds are about 2.6:1. These predictions mean that in two individuals of the same age, the individual with the higher resting blood pressure, and max heart rate is much more likely to have heart disease. For the first prediction, one out of two people are likely to have heart disease. For the second prediction, almost 3 of 4 people are likely.

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

### Reporting Results

The general form of the multiple regression model for heart disease using max heart rate achieved, age, sex, exercise-induced angina, and type of chest pain is shown as:

Or, in its linear form (natural log of odds):

Using beta parameter estimates from the summary function in R, the logistic regression model in its true form is:

And in linear form is:

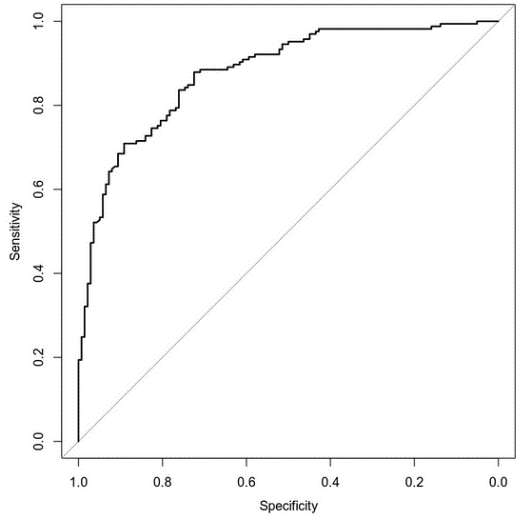
### Evaluating Model Significance

To evaluate the significance of this model, another Hosmer-Lemeshow test will be performed, comparing against the same null and alternative hypotheses as the first model. The Test statistic (chi-square) is 60.596 and the P-value is 0.1048. At a 5% significance level, the null hypothesis is favored because the P-value is higher than 0.05 and the model is confirmed to be a fit for this data. To perform a Wald’s test and determine which predictors are significant, the null and alternative hypotheses to compare against will be

When comparing the P-values for the parameters to the hypotheses, all predictors are below the significance threshold except for age, , and the age:thalach interaction term. The confusion matrix and accuracy, precision, and recall are listed below:

|  |  |  |
| --- | --- | --- |
|  | Prediction: 0 | Prediction: 1 |
| Actual: 0 | 103 (TN) | 35 (FP) |
| Actual: 1 | 27 (FN) | 138 (TP) |

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



The above ROC curve was plotted which seems to “curve higher” than the previous model, which would mean it contains more of the plot area under the curve. When calculating the AUC, the R program returns a value of ~0.8777 or about 87.77%, indicating a higher accuracy than the first model.

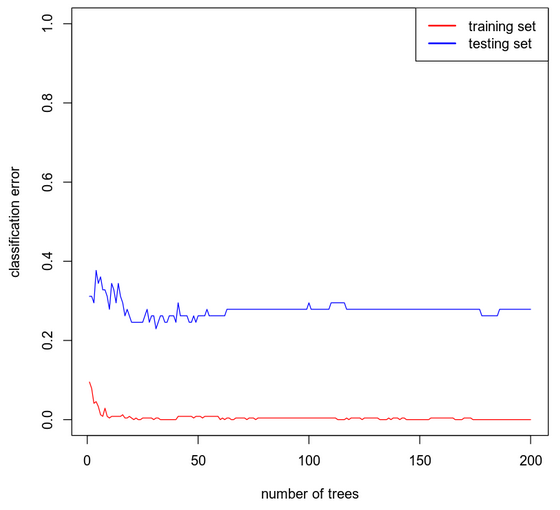
### Making Predictions Using Model

The probability of a male having heart disease, age 30, with a max heart rate of 145, exercise-induced angina, and no chest pain related to angina is about 26.54% and the odds of heart disease are about 1 out of 3 or 1:3. The probability of a male having heart disease, age 30, with a max heart rate of 145, no exercise-induced angina, who does experience typical angina is about 84.32% or about 5.3:1 or about 5 out of 6. This would indicate that people who only experience angina while exercising and have no chest pain are much less likely to have heart disease.

## 5. Random Forest Classification Model

### Reporting Results

To both train and test random forest models, training sets and testing sets must be split from the main data set. In this case, an 80/20 split will be used with 80% being the test set. In the original data set, there are 303 rows, and 242 and 61 rows in the training and testing sets respectively. This model will attempt to predict whether or not an individual has heart disease using age, sex, chest pain type, resting blood pressure, cholesterol measurement, resting electrocardiographic measurement, exercise-induced angina, the slope of peak exercise, and the number of major vessels as predictors. To determine the proper number of trees to use in the model, a graph for visual comparison must be made which will show the classification error vs the number of trees. Once the plot line for the testing set has *mostly* normalized, that is usually the increment at which increasing the tree count no longer improves the model, due to the minimization of classification errors.



In the case of this model, the proper number of trees appears to be right around the 20-25 tree mark, so 20 trees were chosen for this model.

### Evaluating the Utility of the model

To determine the usefulness of this model, the confusion matrices for both the training set and testing set, as well as their respective accuracies, precisions, and recalls will be shown below:

Training Set

|  |  |  |
| --- | --- | --- |
|  | Prediction: 0 | Prediction: 1 |
| Actual: 0 | 111 (TN) | 1 (FP) |
| Actual: 1 | 0 (FN) | 130 (TP) |

* **Accuracy**: ~0.9959
* **Precision**: ~0.9924
* **Recall**: 1

Testing Set

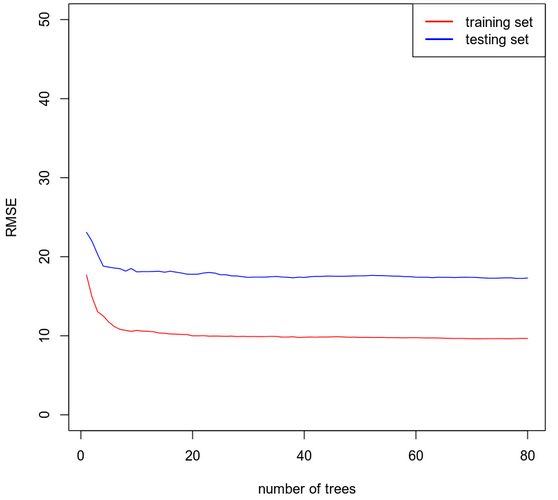
|  |  |  |
| --- | --- | --- |
|  | Prediction: 0 | Prediction: 1 |
| Actual: 0 | 18 (TN) | 8 (FP) |
| Actual: 1 | 7 (FN) | 28 (TP) |

* **Accuracy**: ~0.7541
* **Precision**: ~0.7778
* **Recall**: ~0.8

## 6. Random Forest Regression Model

### Reporting Results

In the original data set, there are 303 rows, and 242 and 61 rows in the training and testing sets respectively, due to the fact this is the same data set as used in the previous model. This model will attempt to predict the maximum heart rate achieved using age, sex, chest pain type, resting blood pressure, cholesterol measurement, resting electrocardiographic measurement, exercise-induced angina, the slope of peak exercise, and the number of major vessels. Again, to determine the proper number of trees to use in this model, a visual inspection must be made. However, this time the number of trees will be graphed against the RMSE (root mean squared error). Once again, wherever the testing set plot line seems to normalize will be the target number of trees to include, exceeding this amount serves no benefit to the model. The graph for this is shown below:



For this model, the line appears to normalize dead center between 0 and 20, indicating that 10 trees would likely be the best fit.

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

Rather than using confusion matrices to gauge the utility of a random forest regression model such as this, the RMSE is calculated for both the training and testing data sets using a model with 10 trees. The RMSE for the training set is about 12.2682 and the RMSE for the testing set is about 16.9347.

## 7. Conclusion

Regarding the logistic regression models created in this analysis, I would choose the second model. Even though the second model had several variables that the Wald’s tests would indicate do not belong in the model, the AUC was a bit higher indicating that this model is more accurate. Furthermore, the variables that did not pass the significance tests, are the same variables that failed the significance test on the first model, which would hint that age and heart disease are not as correlated as originally thought.

Recommending the random forest classification model instead of the logistic regression model is difficult. On one hand, with a very large data set, it is my opinion that a random forest model would be more accurate, however, in this particular case where the population size of the data is less than 1000 and the sample size for the model (training set) is even smaller than that, I would have to recommend the logistic regression model. The accuracy of the logistic model is several percentage points higher.

Hopefully, the insights gleaned from these analyses can be used by doctors or healthcare providers to compare variable characteristics in patients and more accurately diagnose instances of heart disease to both provide better healthcare and shift more cases of heart disease into the preventable/treatable realms.