# **MAT 303 Project Two Summary Report**

Christopher Williams

Christopher.Williams39@snhu.edu

Southern New Hampshire University

## **1. Introduction**

This data set is a large set of historical health data that can be used to analyze the relationships between various health indicators, like blood sugar, heart rate, and heart disease. I will use this data to uncover which indicators are significant in predicting heart disease. I will start with a logistic regression model to determine the liklihood of heart disease, followed by another logistic regression model to compare the effectiveness of the two models. Then, there will be a random forest classification model followed by a random forest regression model, which will be used for similar predictive purposes. These results can be used to determine if a patient is at risk of heart disease or not.

## **2. Data Preparation**

The important variables here are heart disease (target), age (age), resting blood pressure (trestbps), maximum heart rate achieved (thalach), sex of the individual (sex), exercise-induced angina (exang), chest pain type (cp), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), slope of peak exercise (slope), and number of major vessels (ca). With 14 variables measured across 303 data points, there are 14 columns and 303 rows.

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

### **Reporting Results**

The general form of the multiple regression model for heart disease (target) using variables age (age), resting blood pressure (trestbps), and maximum heart rate achieved (thalach) is

E(Y) = e^(B0 + B1X1 + B2X2 + B3X3) / 1 + e^(B0 + B1X1 + B2X2 + B3X3)

where X1 is age, X2 is resting blood pressure, and X3 is maximum heart rate achieved.

The natural log equivalent is ln(pi / 1-pi )=β0 + β1x1 + β2x2 + β3x3.

Here, pi represents the odds of heart disease while 1-pi is the odds of no heart disease. The model produces this equation: E(Y) = e^(-3.576 + -0.009X1 + -0.016X2 + 0.043X3) / 1 + e^(-3.576 + -0.009X1 + -0.016X2 + 0.043X3) where X1 is age, X2 is resting blood pressure, and X3 is maximum heart rate achieved. The estimated coefficient of the maximum heart rate achieved is much larger than the other terms. This shows that maximum heart rate was more strongly positively correlated with heart disease than other measured variables.

### **Evaluating Model Significance**

[1] "Hosmer-Lemeshow Goodness of Fit Test"

Hosmer and Lemeshow goodness of fit (GOF) test

data: logit$y, fitted(logit)

X-squared = 41.978, df = 48, p-value = 0.7168

The Hosmer-Lemeshow goodness of fit test divides the graph into sub-groups along the x-axis, like a scatterplot. The GOF is then a measure of how well observed data points within each sub-group fit the predicted value. The null hypothesis here is that the model fits the data. It follows then, that the alternative hypothesis is that the model does not fit the data. The Chi-Squared statistic is a measure of how far the observed values are from the expected values. The p-value for this GOF test is 0.7168, meaning that the likelihood of obtaining the Chi-Squared statistic, which is 41.978, given the degrees of freedom, is 71.68%, assuming the null hypothesis is true. If we are testing the null hypothesis at a 5% significance, then there is insufficient evidence to reject the null hypothesis since 5% is less than 72%. In summary, there is a 72% chance that the model does not fit the data because the Chi-squared value, 41.978, shows a significant difference between the observed and expected values for the model.

The null hypothesis for each individual term is that the variable has no effect on the likelihood of a patient having heart disease. The alternative hypothesis is that the variable does have an effect. Trestbps has a p-value of 0.0392. Thalach has a p-value of 8.06e-10 and age has a p-value of 0.5578. This means that both trestbps and thalach have a 5% significance level, since their p-values are below 5%. Their p-values can be interpreted as the likelihood that such values would be observed if the variable did not have any effect on the dependent variable. Therefore, those null hypotheses are both rejected at a 5% significance level, but age is not.

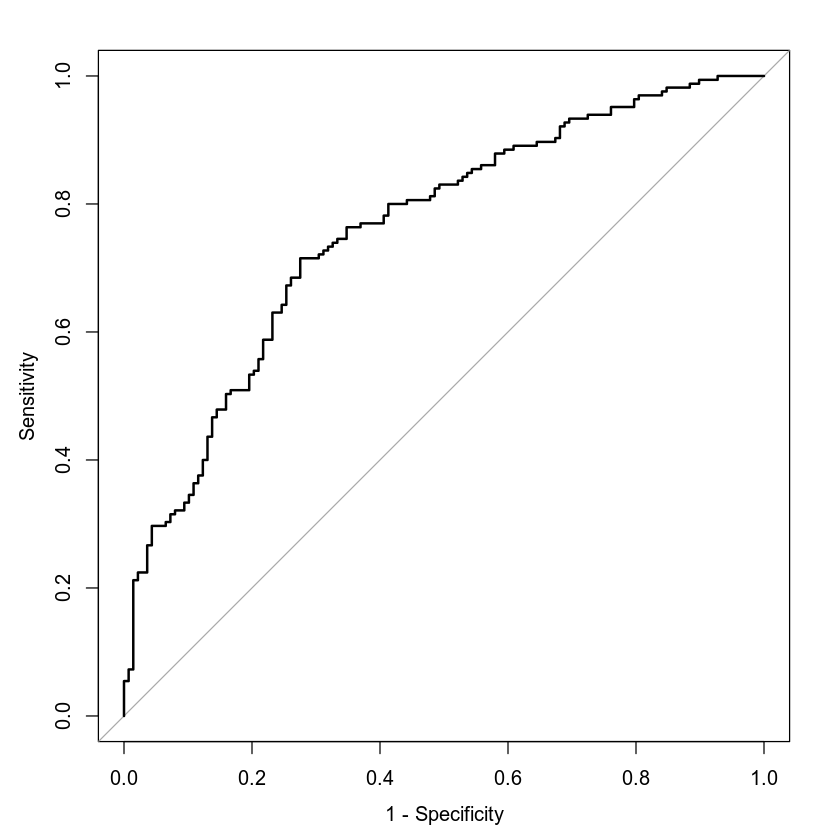
|  | Prediction: target=0 | Prediction: target=1 |
| --- | --- | --- |
| Actual: target=0 | 83 | 55 |
| Actual: target=1 | 38 | 127 |

The accuracy here is 0.6931. The precision is 0.6978. The recall is 0.7697. This all points to a model that is not very accurate or precise and only somewhat useful.

The Receiver Operating Characteristic (ROC) curve, shown below, measures the performance of a model at various thresholds. The more space there is under the graph line, the more accurate the model is interpretted to be. The AUC is the area under the ROC curve. It is a measure of how good the model is at distinguishing between whether a patient has heart disease or not. The AUC, 0.7575, indicates that the model distinguishes correctly 75.75% of the time.

*[1] "Area Under the Curve (AUC)"*

0.7575

**

### **Making Predictions Using Model**

*[1] "Prediction: patient has rbp of 122, max heart rate is 140, and age is 50"*

1: 0.4939

[1] "Prediction: patient has trestbps of 140, age of 50, and max heart rate of 170"

1: 0.7248

Based on these predictions alone, I would deduce that resting heart beat and/or maximum heart rate are great predictors of the probability that a person has heart disease.

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

### **Reporting Results**

The general form of the multiple regression model for heart disease (target) using variables age (age), resting blood pressure (trestbps), sex of the individual (sex), exercise-induced angina (exang), type of chest pain (cp), maximum heart rate achieved (thalach), the quadratic term for age and the interaction term between age and maximum heart rate achieved is

E(Y) = e^(B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6 + B7X7 + B8X8 + B9X1^2 + B10X1X6) / 1 + e^(B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6 + B7X7 + B8X8 + B9X1^2 + B10X1X6)

where X1 is age, X2 is resting blood pressure, X3 is sex, X4 is exercise induced angina, X5, X6, and X7 is type of chest pain, and X8 is maximum heart rate achieved.

The natural log equivalent is ln(pi / 1-pi )= B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6 + B7X7 + B8X8 + B9X1^2 + B10X1X6.

Here, pi represents the odds of heart disease while 1-pi is the odds of no heart disease.

Call:

glm(formula = target ~ age + trestbps + sex + exang + cp + thalach +

I(age^2) + age:thalach, family = "binomial", data = heart\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4926 -0.5907 0.1968 0.6453 2.6819

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.573e+01 1.209e+01 -1.302 0.192995

age 2.430e-01 3.126e-01 0.777 0.436959

trestbps -2.201e-02 9.310e-03 -2.364 0.018076 \*

sex1 -1.815e+00 3.710e-01 -4.894 9.91e-07 \*\*\*

exang1 -8.423e-01 3.686e-01 -2.285 0.022304 \*

cp1 1.787e+00 4.923e-01 3.630 0.000283 \*\*\*

cp2 1.865e+00 3.899e-01 4.783 1.73e-06 \*\*\*

cp3 1.942e+00 5.956e-01 3.260 0.001115 \*\*

thalach 1.474e-01 5.760e-02 2.560 0.010473 \*

I(age^2) 3.979e-04 2.076e-03 0.192 0.848053

age:thalach -2.110e-03 1.007e-03 -2.095 0.036168 \*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 417.64 on 302 degrees of freedom

Residual deviance: 257.57 on 292 degrees of freedom

AIC: 279.57

Number of Fisher Scoring iterations: 5

The model produces this equation: E(Y) = e^(-15.73 + 0.243X1 + -0.022X2 + -1.82X3 + -0.8423X4 + 1.787X5 + 1.865X6 + 1.942X7 + 0.1474X8 + 0.000398X1^2 + -0.0021X1X6) / 1 + e^(-15.73 + 0.243X1 + -0.022X2 + -1.82X3 + -0.8423X4 + 1.787X5 + 1.865X6 + 1.942X7 + 0.1474X8 + 0.000398X1^2 + -0.0021X1X6). The estimated coefficient of chest pain is much larger than the other terms. This shows that type of chest pain was more strongly correlated with heart disease than other measured variables.

### **Evaluating Model Significance**

**1] "Hosmer-Lemeshow Goodness of Fit Test"**

Hosmer and Lemeshow goodness of fit (GOF) test

data: logit2$y, fitted(logit2)

X-squared = 48.87, df = 48, p-value = 0.4379

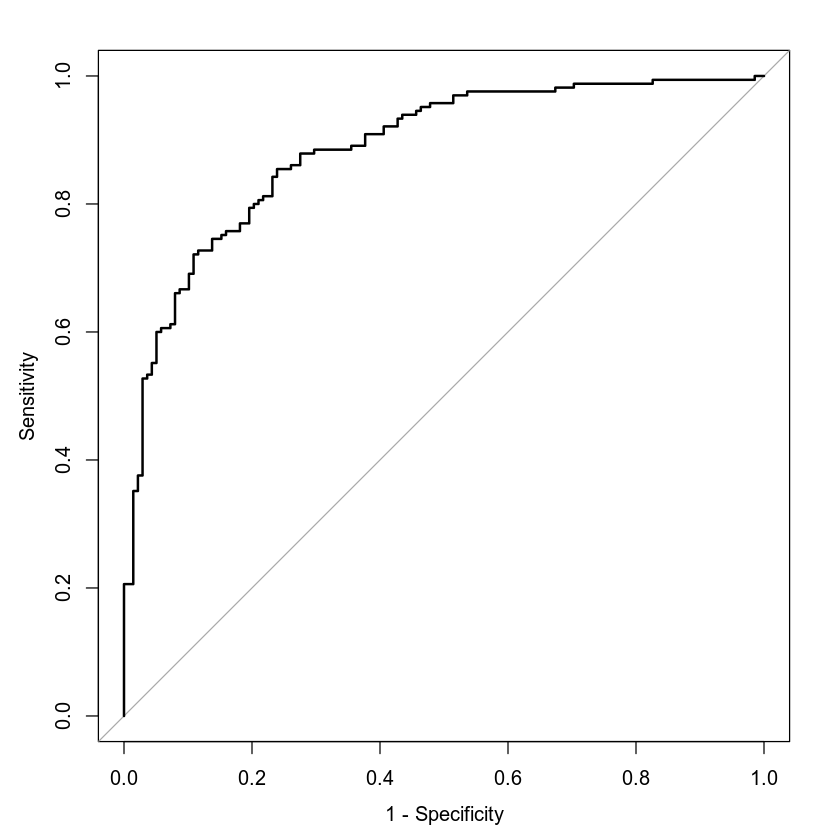
The Hosmer-Lemeshow GOF test, shown above, is to assess if the model is appropriate for the data set. The null hypothesis is that the model fits the data. This means that the p-value for this test will show the likelihood of the model fitting the data. The alternative hypothesis is then that the model does not fit. The p-value here is 0.4379. This indicates a 43.79% chance of the model being accurate. This means that the model is much more than 5% likely to be accurate. Thus, the null hypothesis is not rejected. According to the model, there are two terms with a p-value greater than 0.05. These terms are age and age^2. The null hypothesis for each term is that the term is insignificant or has no effect on the dependent variable(heart disease). The alternative hypotheses are that each term does have an effect on the presence of heart disease. Since age and age^2 are the only terms with a p-value above 0.05, they are the only ones for which we cannot reject the null hypotheses at a 5% significance.

[1] "Confusion Matrix"

A matrix: 2 × 2 of type chr

|  | Prediction: target=0 | Prediction: target=1 |
| --- | --- | --- |
| Actual: target=0 | 104 | 34 |
| Actual: target=1 | 24 | 141 |

The accuracy of this model is 80.86%. The precision is 80.57%. The recall or sensitivity is 85.45%. Shown below is the ROC. This ROC is much better than the first one. The AUC here is 88.54%, which corresponds to the chance that an outcome will be predicted correctly by the model.



*[1] "Area Under the Curve (AUC)"*

0.8854

### **Making Predictions Using Model**

*[1] "age=30, trestbps=100, sex=1, exang=1, cp=0, thalach=145"*

1: 0.3222

[1] "age=30, trestbps=100, sex=1, exang=0, cp=1, thalach=145"

1: 0.8683

The two above statistics are

1.) the probability of a male individual having heart disease who is 30 years old; has a maximum heart rate of 145; has a resting blood pressure of 100; experiences exercise-induced angina; and does not experience chest pain related to typical angina, atypical angina, or non-anginal pain, and

2.) the probability of a male individual having heart disease who is 30 years old, has a resting blood pressure of 100; has a maximum heart rate of 145, and does not experience exercise-induced angina but experiences typical angina.

These results indicate that chest pain is one of the strongest predictors of heart disease. The type of chest pain seems to be very important.

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

### **Reporting Results**

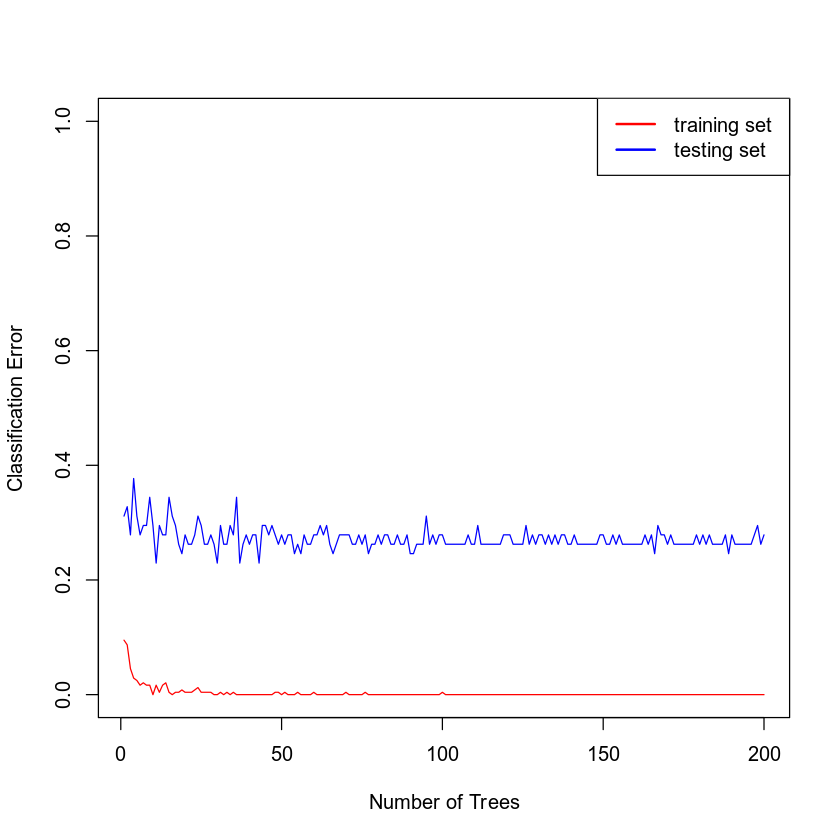
If 80% of the data is placed in a training set and 20% is placed in a testing set, then there are 242 rows in training and 61 rows in testing. Graphed below is the classification error set against the number of trees. This graph indicates that around 20 trees is all that is required for minimal errors.

[1] "Number of rows for the training set"

242

[1] "Number of rows for the testing set"

61

****

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

**[1] "Confusion Matrix: TRAINING set based on random forest model built using 20 trees"**

A matrix: 2 × 2 of type chr

|  | Prediction: 1 | Prediction: 0 |
| --- | --- | --- |
| Actual: 0 | 0 | 112 |
| Actual: 1 | 130 | 0 |

[1] "======================================================================================================================"

[1] "Confusion Matrix: TESTING set based on random forest model built using 20 trees"

A matrix: 2 × 2 of type chr

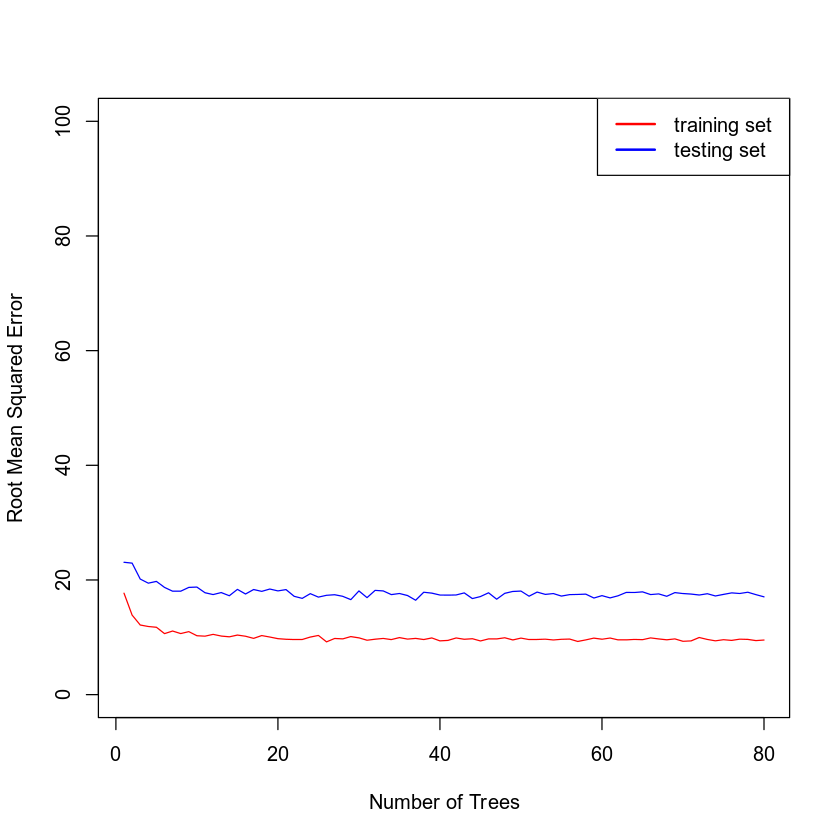
|  | Prediction: 1 | Prediction: 0 |
| --- | --- | --- |
| Actual: 0 | 11 | 15 |
| Actual: 1 | 27 | 8 |

Shown above is the confusion sets for the training and the testing sets for a 20 tree random forest that predicts the presence of heart disease (target) using variables 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 accuracy of the training set is 1. The precision of the training set is 1. The recall of the training set is 1. The accuracy of the testing set is 0.6885. The precision of the testing set is 0.7105. The recall of the testing set is 0.7714. This is a decent model, but the difference between the training and testing accuracy is significant.

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

### **Reporting Results**

The graph below shows the number of trees against the RMSE. These models are being used to predict a patients max heart rate achieved (thalach). The RMSE is lowered with each new tree until around the 10 tree mark. Past this point, it only fluctuates slightly. Therefore, it is only necessary to have around 10 trees in a random forest regression model for thalach.

**

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

**[1] "Root Mean Squared Error: TRAINING set based on random forest model built using 20 trees"**

10.547927726639

[1] "======================================================================================================================"

[1] "Root Mean Squared Error: TESTING set based on random forest model built using 20 trees"

16.9683655825757

Above is the RMSE for training and testing sets of 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 for the training set is 10.55. The root mean squared error for the testing set is 16.97. The RMSE tells the average distance from the predicted data value to the observed data point. A lower RMSE indicates a better fit for the model.

## **7. Conclusion**

For the first model, we found a GOF X^2 of 42, a p-value of 0.717, accuracy of 0.69, precision of 0.70, and AUC of 0.7575. In the second model, we found a GOF X^2 of 49, a p-value of 0.44, accuracy of 0.81, precision of 0.81, and AUC of 0.8854. These numbers indicate that the second model has a higher chance of being missplaced, but it likely has greater accuracy, precision, and fit. For these reasons, I would probably choose the second model. However, the higher p-value is cause for concern, because it could be an indicator that the model is over fit. A logical way to check this would be to gather more data, then use it as a testing set to create confusion matrices for the two models. This would be the proper way to determine if the models fit.

For the random forest classification model, the accuracy of the testing set is 0.6885 and the precision of the testing set is 0.7105. This is roughly what I would expect from the second logistic regression model, but it does not have the same danger of being overfitted. For this reason, I would choose the random forest model over the logistic regression model. The practical importance of these analyses is the ability to quantify the risk of heart disease. Doctors and nurses make mistakes and have lapses in judgment, so it is good to have a more standard system in place to identify people who are at risk. These individuals could undergo further tests that most patients would not recieve.