# **MAT 303 Project Two Summary Report**

[Luna Saccoia]

[Luna.Saccoia@snhu.edu]

Southern New Hampshire University

**1. Introduction**

*The data set I will be exploring is the heart disease data set. My results will be used to predict the likelihood of developing heart disease based on a variety of factors. The types of analyses I will be running for this project are two logistic regression models, a random forest classification model and a random forest regression model.*

## **2. Data Preparation**

*The important variables that will be used from this data set are age, sex, cp(chest pain), resting bp, cholesterol, resting ecg, exercise induced angina=true or false, slope of the peak exercise, ca (number of major vessels) and target.*

*There are 14 columns and 303 rows in the data set.*

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

### **Reporting Results**

*The general equation for the multiple regression model is:*

*The is the intercept. The are the popular regression terms for age (age), resting blood pressure (trestbps), and maximum heart rate achieved (thalach).*

*The model equation using these variables is:*

*The equation transformed to form a model that is linear in the beta terms is:*

*The model equation in terms of the natural log of odds is:*

*is the probability of the event happening, in this case the event is developing heart disease.*

*is the odds of developing heart disease.*

*The equation for this regression model is:*

*The equation for this model in terms of natural log of odds is:*

*The estimated coefficient of the maximum heart rate achieves variable is 0.042697. This means on average, the change in log odds of developing heart disease is 0.042697, holding all other variables constant.*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model predictions are close to the observed values of Y, which are either 0 or 1. In this model it is used to assess if the model fits the data or not.*

*The null and alternative hypotheses are:*

*The test statistic ( is 41.978, The P-value is 0.7168. The level of significance is 5%.*

*The P-value of 0.7168 is higher than the level of significance of 0.05. Thus the null hypothesis should not be rejected. The conclusion is that the model is appropriate for the data set.*

*The null and alternative hypotheses for calculating if age is significant based on Wald’s test with a 5% level of significance are:*

*is the age parameter.*

*The null and alternative hypotheses for calculating if resting blood pressure (trestbps) is significant based on Wald’s test with a 5% level of significance are:*

*is the resting blood pressure (trestbps) parameter.*

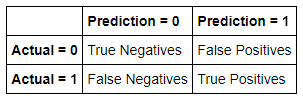
*The null and alternative hypotheses for calculating if maximum heart rate achieved (thalach)is significant based on Wald’s test with a 5% level of significance are:*

*is the maximum heart rate achieved (thalach) parameter.*

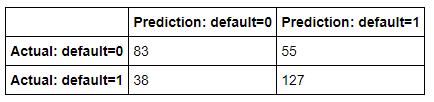
*The P-value for age is 0.5578. The P-value for resting blood pressure (trestbps) is 0.0392. The P-value for maximum heart rate achieved (thalach) is 8.06e-10.*

*The terms for resting blood pressure and maximum heart rate achieved are statistically significant at a 5% level of significance. The term for age is not significant.*

*The general form table output of a confusion matrix is:*

**

*The confusion matrix for this model is:*

**

*The confusion matrix results are:*

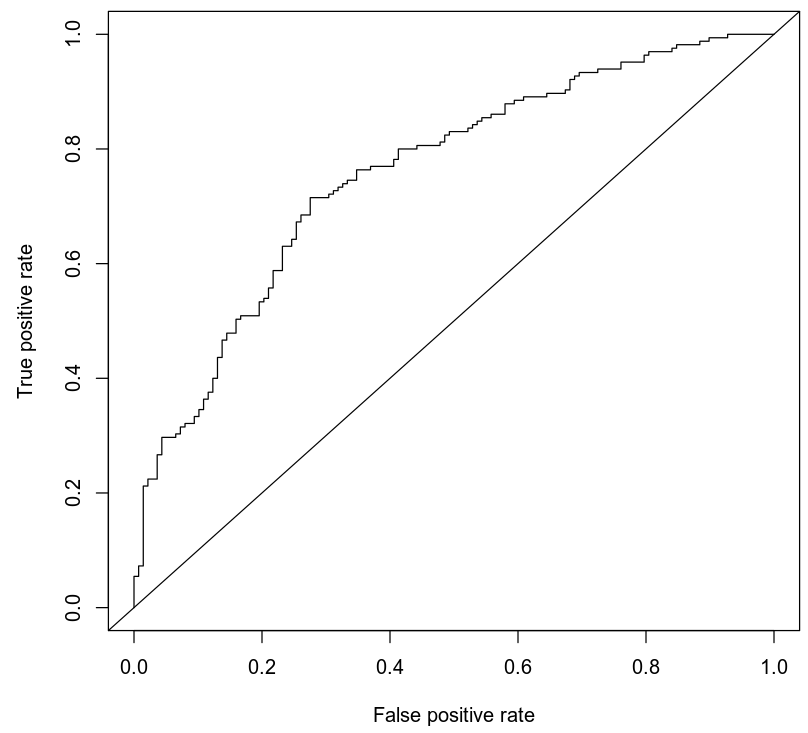
* *True positives: 127*
* *True negatives: 83*
* *False positives: 55*
* *False negatives: 38*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

*The Receiver Operating Characteristic (ROC) curve is displayed here:*

**

*The area under the curve (AUC) is 0.7575 or 75.75%. This is an indicator of how well the model distinguishes between Y = 0 and Y = 1. In general, the larger the AUC the better, because the larger the area under the curve, the better it is at predicting binary classes.*

### **Making Predictions Using Model**

***Prediction 1***

*The probability of 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. The probability of an individual, with these variables, having heart disease is 49.39%.*

*The odds are calculated by dividing the probability by 1 minus the probability. In this case . The odds in this case are 0.4939 to 0.5061 or 49.39% to 50.61%. The odds are very close to 50/50 or 1:1 of developing heart disease for an individual with these parameters.*

***Prediction 2***

*The probability of 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. The probability of an individual, with these variables, having heart disease is 72.48%.*

*The odds are calculated by dividing the probability by 1 minus the probability. In this case . The odds in this case are 0.7248 to 0.2752 or 72.48% to 27.52%. The odds are close to 75/25 or 3:1 odds for developing heart disease for an individual with these parameters.*

*The probability of these results is basically the percentage of the chance an individual with these parameters will develop heart disease. The odds are calculated to see just how likely it is. For the first prediction the odds are very close to 50/50. So if we had two people with these parameters, it could be predicted that 1 of them would have heart disease. In the second prediction, someone has an almost 75% chance of developing heart disease. So, in terms of odds, if 4 people had these parameters, it could be predicted that 3 of them would develop heart disease.*

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

### **Reporting Results**

*The general form equation of the multiple regression model is:*

*are the popular regression terms for maximum heart rate achieved (thalach), age of the individual (age), the dummy term for sex1 (male) and the dummy term for exercise-induced angina (exang1). are the dummy terms for cp1, cp2 and cp3. are the popular regression terms for age squared and the interaction of thalach against age.*

*The model for this equation is:*

*The equation for this model, in terms of log of odds is:*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model predictions are close to the observed values of Y, which are either 0 or 1. In this model it is used to assess if the model fits the data or not.*

*The null and alternative hypotheses are:*

*The test statistic ( is 60.596. The P-value is 0.1048. The level of significance is 5%.*

*The P-value of 0.1048 is higher than the level of significance of 0.05. Thus the null hypothesis should not be rejected. The conclusion is that the model is appropriate for the data set.*

*The null and alternative hypotheses for calculating if maximum heart rate is significant based on Wald’s test with a 5% level of significance are:*

*is the maximum heart rate (thalach) parameter.*

*The P-value for maximum heart rate (thalach) 0.014760. This is statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if age is significant based on Wald’s test with a 5% level of significance are:*

*is the age parameter.*

*The P-value for age is 0.510325. This is not statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if the dummy variable for sex (male) is significant based on Wald’s test with a 5% level of significance are:*

*is the dummy variable for sex (male) parameter.*

*The P-value for sex (male) is 1.91e-06. This is statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if exercise induced angina=yes (exang1) is significant based on Wald’s test with a 5% level of significance are:*

*is the exercise induced angina=yes (exang1) parameter.*

*The P-value for exercise induced angina=yes (exang1) is 0.009133. This is statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if the dummy variable for chest pain=typical angina (cp1) is significant based on Wald’s test with a 5% level of significance are:*

*is the dummy variable for chest pain=typical angina(cp1) parameter.*

*The P-value for chest pain=typical angina (cp1) is 0.000249. This is statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if the dummy variable for chest pain=atypical angina (cp2) is significant based on Wald’s test with a 5% level of significance are:*

*is the dummy variable for chest pain=atypical angina(cp2) parameter.*

*The P-value for chest pain=atypical angina (cp2) is 2.21e-06. This is statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if the dummy variable for chest pain=non-anginal pain (cp3) is significant based on Wald’s test with a 5% level of significance are:*

*is the dummy variable for chest pain=non-anginal pain(cp3) parameter.*

*The P-value for chest pain=non-anginal pain(cp3) is 0.003684. This is statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if age2 is significant based on Wald’s test with a 5% level of significance are:*

*is the age2 parameter.*

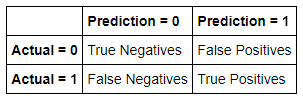
*The P-value for age2 is 0.810599. This is not statistically significant at a 5% level of significance.*

*The null and alternative hypotheses for calculating if the interaction between maximum heart rate (thalach) against age is significant based on Wald’s test with a 5% level of significance are:*

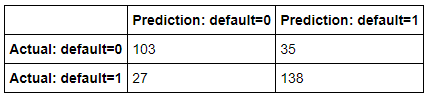
*is the interaction between maximum heart rate(thalach) against age parameter.*

*The P-value for the interaction between maximum heart rate (thalach) against age is 0.043666. This is statistically significant at a 5% level of significance.*

*The general form table output of a confusion matrix is:*

**

*The confusion matrix for this model is:*

**

*The confusion matrix results are:*

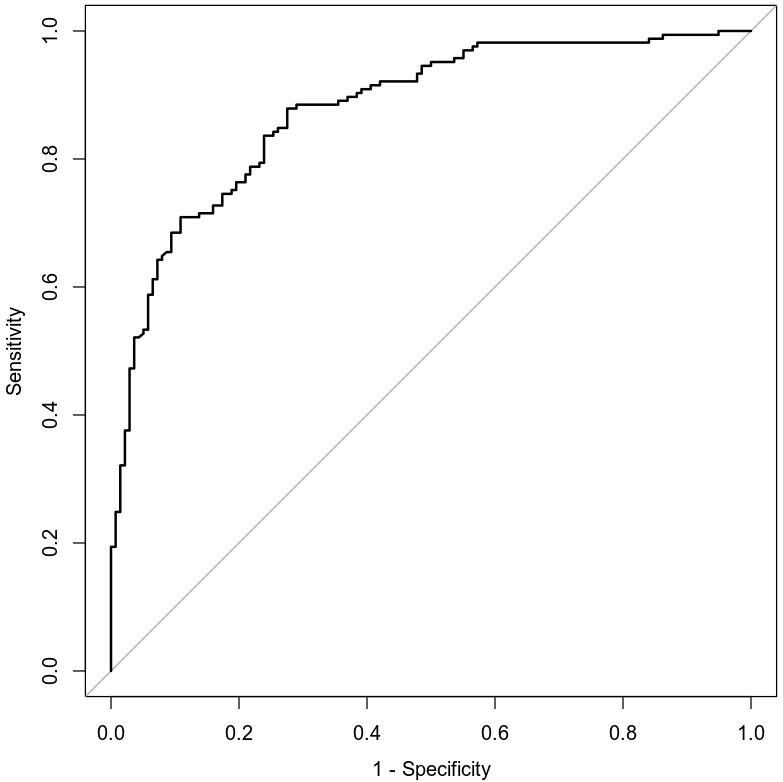
* *True positives: 138*
* *True negatives: 103*
* *False positives: 35*
* *False negatives: 27*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

*The Receiver Operating Characteristic (ROC) curve is displayed here:*

**

*The area under the curve (AUC) is 0.8777 or 87.77%. This is an indicator of how well the model distinguishes between Y = 0 and Y = 1. In general, the larger the AUC the better, because the larger the area under the curve, the better it is at predicting binary classes.*

### **Making Predictions Using Model**

***Prediction 1***

*The probability of a male individual having heart disease 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 is 0.2654 or 26.54%. The odds are a little more than 1 to 3.*

***Prediction 2***

*The probability of a male individual having heart disease who is 30 years old, has a maximum heart rate of 145, and does not experience exercise-induced angina but experiences typical angina is 0.8432 or 84.32%. The odds are a little less than 5 to 1.*

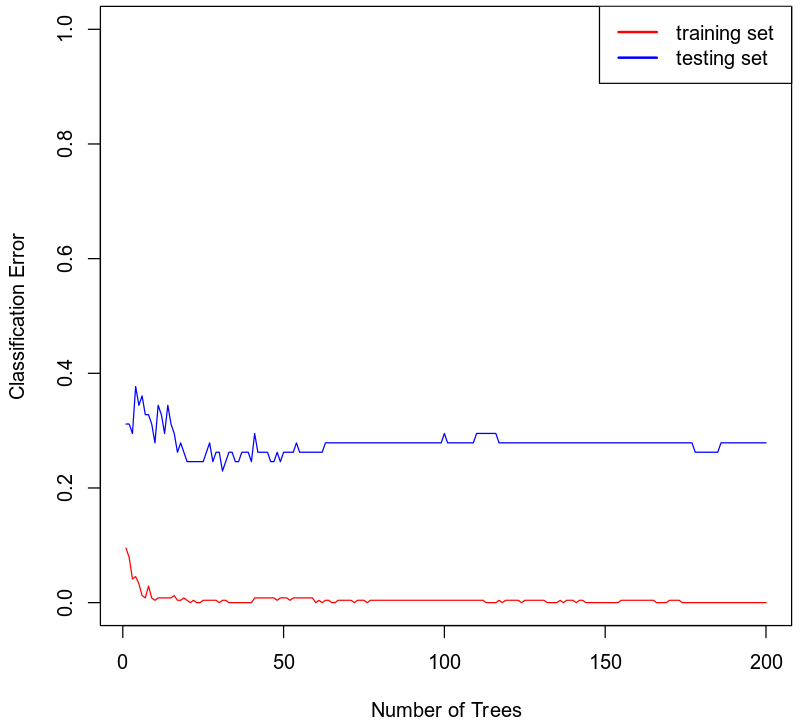
*The odds in the first prediction is approximately 1 to 3, while the second jumps to just under 5 to 1.*

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

### **Reporting Results**

*Splitting the heart disease data set into training and testing sets using 80% and 20% split, respectively. Using set.seed(511038) gives 303 rows in the original set, 242 rows in the training set and 61 rows in the testing set.*

*Graphing the training and testing error against the number of trees using a classification random forest model for 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); Using a maximum of 200 trees and using set.seed(511038) gives the following graph:*

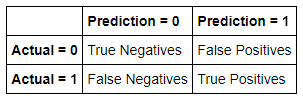
**

*The optimal number of trees is around 20 trees. This is roughly where the curve flattens out.*

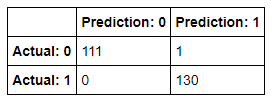
### **Evaluating the Utility of the model**

*Using the appropriate number of trees found, we can create a classification random forest model for 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). We then create a confusion matrix for this as seen below.*

*The general form table output of a confusion matrix is:*

**

*The confusion matrix for this model’s training set is:*

**

*The confusion matrix results for the training set are:*

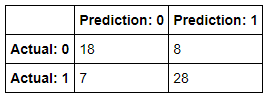
* *True positives: 130*
* *True negatives: 111*
* *False positives: 1*
* *False negatives: 0*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

*The confusion matrix for this model’s testing set is:*

**

*The confusion matrix results for the testing set are:*

* *True positives: 28*
* *True negatives: 18*
* *False positives: 8*
* *False negatives: 7*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

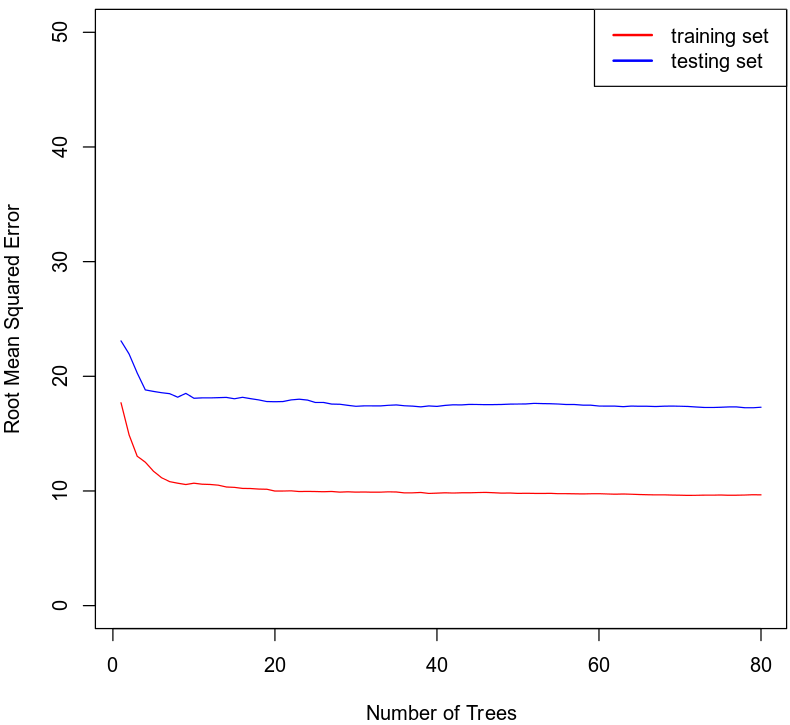
***Recall****is the ratio of correct positive predictions to the total positives examples.*

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

### **Reporting Results**

*Splitting the heart disease data set into training and testing sets using 80% and 20% split, respectively; Using set.seed(511038) gives 303 rows in the original set, 242 rows in the training set and 61 rows in the testing set.*

*Graphing the mean squared error against the number of trees for 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); using a maximum of 80 trees and using set.seed(511038) gives us the following graph:*

**

*The optimal number of trees for this random forest model is around 10 trees. This is where the curve flattens out.*

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

*Using the appropriate number of trees found, we can 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 for the training set is 9.9028.*

*The root mean squared for the testing set is 17.387.*

## **7. Conclusion**

*Of the two logistic regression models analyzed here, I would choose the second model. It includes more variables and has a larger area under the curve, which is better at predicting binary values. It also has a higher accuracy, precision and recall for the confusion matrix. I think this is a better predictor of an individual’s risk of heart disease.*

*I would recommend using the random forest classification model instead of the logistic regression model due to it having much higher accuracy, precision and recall than either of the logistic regression models.*

*The practical importance of the analyses performed are that doctors can use someone’s information and lifestyle to determine their risk of heart disease and compare that risk to different variables to see what changes the person can make to lower their risk.*

## **8. Citations**

*Zybooks MAT 303: Applied Statistics II for Science, (2016, August).*

*Retrieved April 17, 2020, from https://learn.zybooks.com/zybook/SNHUMAT303v1*