# **MAT 303 Project Two Summary Report**

Yaqub Mohamud

yaqub.mohamud@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The heart disease data set is the one that is being analyzed. Patterns between the existence of heart disease and several health markers will be examined using historical data. The findings will be utilized to forecast a person's heart disease risk depending on several variables. The analysis will include two logistic regression models, a random forest classification model, and a random forest regression model.

## **2. Data Preparation**

This data set's important variables include age (age), resting blood pressure (trestbps), exercise-induced angina (exang), maximum heart rate achieved (thalach), type of chest pain (cp), sex (sex), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), and number of major vessels (ca). The data collection has 303 rows and 14 columns.

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

### **Reporting Results**

The general form for the multiple regression model is:

E(Y )=β0+β1 X1+ β2 X2+β3 X3+ β4 X4

β0 represents the intercept. β1, β2, β3, and β4 are commonly used regression terms for age (age), resting blood pressure (trestbps), exercise-induced angina (exang), and maximal heart rate (thalach).

Using the natural log of odds, this equation takes the following linear form:

ln (odds) = β0 +β1 X1 + β2 X2 +β3 X3+ β4 X4.

The regression model's equation is given below:

The maximum heart rate achieved is calculated to have a coefficient of 0.042697. Keeping all other factors equal, this indicates that the average change in the log odds for developing heart disease is 0.042697.

### **Evaluating Model Significance**

To determine if the model's predictions closely match the observed values of Y, which can be either 0 or 1, the Hosmer-Lemeshow goodness of fit (GOF) test is used. Null Hypothesis (H0): The data and the model match each other well. Hypothesis Alternative (Hα): The data does not fit the model well. The P-value is 0.612 and the test statistic (X2) is 44.622. The null hypothesis should not be rejected because the P-value of 0.612 is greater than the significance level of 0.05. Consequently, we may say that the model fits the data set. The terms for resting blood pressure and maximum heart rate attained are statistically significant at the 5% level of significance. But the age term isn't considered significant.

The results of the confusion matrix are:

134 true positives

89 true negatives

49 false positives

31 false negatives

The True Positive (TP), when the predicted value is 1 and the actual value is 1 (default = 1). Consequently, a true positive. The True Negative (TN), when the predicted value is 0 and the actual value is 0 (default = 0). A real negative, then. The False Positive (FP), where the predicted result is 1 and the actual value is 0 (default = 0). Thus, a Type 1 error, also known as a false positive. The False Negative (FN), where the predicted value is 0 and the actual value is 1 (default = 1). Consequently, a Type 2 error, also known as a false negative.

Accuracy is determined by dividing the total number of observations by the proportion of accurate predictions. Accuracy = TP+TN/TP+TN+FP+FN = 134+89/134+89+49+31 = 73.60% is the accuracy.

Precision is defined as the proportion of accurate positive predictions among all projected positive predictions. Precision = TP/TP+FP = 134/134+49 = 73.22%

Recall is the proportion of accurate positive predictions among all the positive cases.   
Recall = TP/TP+FN = 134/134+31 = 81.21%

The following is the Receiver Operational Characteristic (ROC) curve:

A graph with a line

AI-generated content may be incorrect.

The value of the Area Under the Curve (AUC) is 0.8007, or 80.07%. This figure indicates how well the model can distinguish between Y = 0 and Y = 1. Since a bigger area under the curve signifies better prediction of binary classes, a larger AUC generally denotes a better model.

### **Making Predictions Using Model**

If a person is 50 years old, has a resting blood pressure of 122, exercise-induced angina, and a maximum heart rate of 140, their chances of developing heart disease are 0.2716, or 27.16%. The likelihood of the event occurring is divided by one minus the probability to determine the odds.

𝜋 / 1−𝜋 = 0.2716 / 1−0.2716 = 0.2716 / 0.7284, or around 1:2.66 odds.

If a person is 50 years old, has a resting blood pressure of 130, no exercise-induced angina, and a maximum heart rate of 165, their chance of developing heart disease is 0.7864, or 78.64%. The probabilities are determined in the same way:

𝜋 / 1−𝜋 = 0.7864 / 1−0.7864 = 0.7864 / 0.2136, or around 3.68:1 odds.

While the odds show how likely it is, the probability shows the likelihood that someone with these identical variable values will have heart disease. According to the probabilities of 1:2.66 for the first prediction, one in three individuals with these characteristics will have heart disease. With a 78% likelihood of getting heart disease, the probabilities were greater in prediction 2. Given that this is over 75%, the probabilities suggest that three out of four people with these characteristics will have heart disease.

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

### **Reporting Results**

The following is the multiple regression model's general form equation:   
  
E(Y) = β0 +β1 X1 +β2 X2 +β3 X3 +β4 X4 +β5 X5 +β6 X6 +β7 X7 +β8 X^2 2 +β9 X1 X2

where the beta terms for age and resting blood pressure are 𝛽1 and 𝛽2. The dummy variables for the type of chest discomfort experienced are 𝛽3, 𝛽4, and 𝛽5. We need three dummy variables since the variable might have four potential values. The beta term for the maximum heat rate is 𝛽6. The beta term for age squared is 𝛽7, while the interaction term between age and the highest heart rate attained is 𝛽8.

Using the natural log of odds, this equation takes the following linear form:

ln (π 1−π) = β 0 +β1 X1+β2 X2 +β3X3 +β4 X4 +β5 X5 +β6 X6 +β7 X7 +β8 X^2 2 +β9 X1X2

The regression model's equation is given below:

### **Evaluating Model Significance**

The Hosmer-Lemeshow goodness of fit test is used to assess if the model fits the data well. H0, which states that the model matches the data, would be the null hypothesis. The model does not fit the data, which is the alternative hypothesis. The test statistic(X2) is 60.596. With a p-value of 0.3209, above the significance level of 0.05, we are unable to rule out the null hypothesis and conclude that the model fits the data set well. According to the Wald's test, age and CP would be significant in the model.

The results of the confusion matrix are:

129 true positives

102 true negatives

36 false positives

36 false negatives

Accuracy = TP+TN/TP+TN+FP+FN = 129+102/129+102+36+36 = 76.24% is the accuracy.

Precision = TP/TP+FP = 129/129+36 = 78.18%

Recall = TP/TP+FN = 129/129+36 = 78.18%

The following is the Receiver Operational Characteristic (ROC) curve:

A graph with a line

AI-generated content may be incorrect.

The value of the Area Under the Curve (AUC) is 0.8478, or 84.8%. This figure indicates how well the model can distinguish between Y = 0 and Y = 1. Since a bigger area under the curve signifies better prediction of binary classes, a larger AUC often denotes a better model.

### **Making Predictions Using Model**

If a person is 50 years old, has a resting blood pressure of 115, does not experience chest pain, and a maximum heart rate of 133, their chances of developing heart disease are 0.2188, or 21.88%.

𝜋 / 1−𝜋 = 0. 2188 / 1−0. 2188 = 0. 2188 / 0.7812, or around 1:5 odds.

If a person is 50 years old, has a resting blood pressure of 125, experiences typicalangina, and a maximum heart rate of 155, their chance of developing heart disease is 0.8007, or 80.07%.

𝜋 / 1−𝜋 = 0.8007 / 1−0.8007 = 0.8007 / 0.1993, or around 4:1 odds.

The likelihood of a person developing heart disease at age 50 is higher if their resting blood pressure, maximum heart rate, and experience of usual angina are high, according to the odds and probability.

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

### **Reporting Results**

We separated the data set into training and testing subsets for the Random Forest classification model. 15% of the data was utilized as a testing set, while the remaining 85% was used as a training set. The training and testing sets include 257 and 46 rows, respectively, of the 303 rows in the original data set.

A graph of a number of trees

AI-generated content may be incorrect.

There should be no more than 20 trees. The curve flattens out at this approximate point.

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

The confusion matrix for the training set is given below:

137 true positives

118 true negatives

3 false positives

0 false negatives

Accuracy = TP+TN/TP+TN+FP+FN = 137+118/137+118+3+0 = 98.83% is the accuracy.

Precision = TP/TP+FP = 137/137+3 = 97.86%

Recall = TP/TP+FN = 137/137+0= 100.00%

The confusion matrix for the testing set is given below:

21 true positives

11 true negatives

7 false positives

7 false negatives

Accuracy = TP+TN/TP+TN+FP+FN = 21+11/21+11+7+7 = 69.57% is the accuracy.

Precision = TP/TP+FP = 21/21+7 = 75.00%

Recall = TP/TP+FN = 21/21+7= 75.00%

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

### **Reporting Results**

We separated the data set into training and testing subsets for the Random Forest classification model. 15% of the data was utilized as a testing set, while the remaining 85% was used as a training set. The training and testing sets include 242 and 61 rows, respectively, of the 303 rows in the original data set.

A graph of a number of trees

AI-generated content may be incorrect.

Ten trees are the optimal number for this data set as the result shows that the classification error curve flattens out at about ten trees.

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

The root mean squared error for the training set is 11.8890



The root mean squared error for the testing set is 20.5738

****

## **7. Conclusion**

I recommend using the second of the two logistic regression models we analyzed for predicting heart disease. Compared to the first model, the second model exhibits more accuracy and precision. Furthermore, the second model's AUC (area under the curve) is larger than the first logistic regression model's, suggesting that it is more relevant to the data set. Therefore, the second model would be slightly more accurate in estimating a person's risk of heart disease.

Since the Random Forest model received the greatest accuracy, precision, and recall scores of the three models, I would advise selecting it if given the choice between the Random Forest Classification Model and the Logistic Regression Model.

The practical importance of the analysis was to create a model that may be used to analyze a patient's likelihood of developing heart disease based on a variety of historical factors. To reduce the risk of heart disease in their patients, physicians may utilize these models to forecast the likelihood of heart disease and make data-driven choices regarding recommended treatments.