# **MAT 303 Project Two Summary Report**

Nicholas Kreuziger

nicholas.kreuziger@snhu.edu

Southern New Hampshire University

## **1. Introduction**

We will be researching a large set of historical heart data indicating heart disease at a university hospital. These results will be used to predict whether someone is at risk for heart disease and could be used to evaluate existing medical records to assist doctors in helping patients. The analyses used will be two Regression Models, a Random Forest Classification Model to predict the risk of heart disease and a Random Forest Regression Model to predict the maximum heart rate achieved.

## **2. Data Preparation**

In total this dataset has 14 variables and 303 rows. These variables and their descriptions are summarized in the table below:

Graphical user interface, text, application, email

Description automatically generated

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

### **Reporting Results**

We will first write the general form of a logistic regression model that estimates the likelihood of heart disease based upon the variables of age (age), resting blood pressure (trestbps) and maximum heart rate achieved (thalach). We will call this model 1.

The general form of this logistic regression model is:

E(y) = Probability of Heart Disease

β0 = Intercept Parameter; indicates when the logistic curve crosses the vertical axis.

X1 = Age (age)

β1 = Coefficient of Age

X2 = Resting Blood Pressure (trestbps)

β2 = Coefficient of Resting Blood Pressure

X3 = Maximum Heart Rate Achieved (thalach)

β3 = Coefficient of Maximum Heart Rate Achieved

The general form of this model appears:

Written in the natural log of odds to express beta terms in linear form:

In the natural log of odds, π represents the probability for heart disease, not the value of pi as it normally constitutes. represents the odds of success. In this model it represents the probability of heart disease.

Applying the General Model, we use statistical software to arrive at a summary, shown below:

Table

Description automatically generated

The general form of this model becomes:

Written in the natural log of odds to express beta terms in linear form becomes:

This can be expressed in terms of odds by,

This is to say, the odds of heart disease increase by 4.362% for every unit of Maximum Heart Rate achieved. Which makes sense, a sign of a healthy heart is more efficient heartbeats that deliver more oxygenated blood through your system. You need more maximum heart beats under exertion the less healthy you are.

### **Evaluating Model Significance**

In the interests of assessing this model’s validity we will apply a Hosmer-Lemeshow goodness of fit test. A Hosmer-Lemeshow test assesses whether a model is appropriate for the dataset by assessing whether model predictions are close to the observed values of the response variable (in this case the actual instances of credit default vs predicted instances).

The Null Hypothesis is,

H0: The model fits the data

The Alternative Hypothesis is,

Ha: The model does not fit the data

We will test at a 5% level of Significance,

α = 0.05

Text

Description automatically generated

The P-Value of 0.7168 is larger than the Significance Level 0.05, we fail to reject the Null Hypothesis. The Model fits the data.

Following a Hosmer-Lemeshow test we can further validate the model by testing its individual factors with a Wald Test. Each individual Wald Test can be summed as such,

Null Hypothesis, Coefficient of the predictor variable is 0

βi = 0

Alternative Hypothesis, Coefficient of the predictor variable does not equal 0

βi ≠ 0

The p-values of these individual tests can be found in the PR(>|z|) column of the Summary for this model:

Text

Description automatically generated

According to the Wald Test every term is significant in this model except for age, which has a p-value of 0.5578. Summarized in table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | P-Value | Significance Level | Null Hypothesis : Reject or Fail Reject |
| Age | 0.5578 | 0.05 | Fail Reject |
| Resting blood pressure (Trestbps) | 0.0392 | 0.05 | Reject |
| Maximum Heart Rate (thalach) | .000000000806 | 0.05 | Reject |

A confusion matrix is used as a diagnostic for Logistic Models. A confusion matrix is a table output as:

|  |  |  |
| --- | --- | --- |
|  | Prediction = 0 | Prediction =1 |
| Actual = 0 | True Negatives | False Positives |
| Actual = 1 | False Negatives | True Positives |

Using a confusion matrix we can assess Accuracy, Precision and Recall. Defined as,

**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.

We will use a Confusion Matrix based upon the model’s predictions of heart disease. From this we will determine the values of Accuracy, Precision and Recall.

A screenshot of a computer

Description automatically generated with medium confidence

For simplicity, True Negative = 83, False Negative = 38, True Positive = 127, False Positive = 55.

An additional test to assess this model is the ROC curve. An ROC curve measures the performance for a classifier at various threshold settings. The area under the curve is an indicator of how well the model distinguishes between Y = 0 or Y =1 (in this model, whether someone has heart disease). In the ROC curve graph below you can see there is a large area beneath the line, indicating this model does a good job of distinguishing between Y = 0 or Y = 1.

Chart, line chart

Description automatically generated

A picture containing diagram

Description automatically generated

The value of AUC (Area under curve) assessed is 0.7575. This indicates that 75.75% of the area of this plot is within the curve. There is a 75.75% likelihood this model can distinguish between heart disease and non-heart disease patients.

### **Making Predictions Using Model**

Now that we have tested the validity of this model, we can make predictions for a few scenarios.

1. What is 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?
   1. 
   2. There is a 49.39% probability of heart disease based upon these factors.
   3. The odds of heart disease occurring based upon these factors is 4939:5061. In odds this means .4939 / (1 - .4939) = .9759. There is a 9,759:10,000 (9,759 in 10,000) chance of heart disease.
2. What is 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?
   1. 
   2. There is a 72.48% probability of heart disease based upon these factors.
   3. In odds this means .7248/ (1 - .7248) = 2.6337. There is a 26,337 in 10,000 (26,337 in 10,000) chance of heart disease.

It can be concluded that a lower heart rate contributes to a lower likelihood of heart disease.

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

We will now write the general form of a logistic regression model that estimates the likelihood of heart disease based upon the variables of maximum heart rate achieved (thalach), age of the individual (age), sex of the individual (sex), exercise-induced angina (exang), and type of chest pain (cp). We will include a quadratic term for age and an interaction term between age and maximum heart rate achieved. We will call this model 2.

The general form of this logistic regression model is:

E(y) = Probability of Heart Disease

β0 = Intercept Parameter; indicates when the logistic curve crosses the vertical axis.

X1 = Max heart rate achieved (thalach)

β1 = Coefficient of Max heart rate achieved

X2 = Age (age)

β2 = Coefficient of Age

X3 = Sex 1

β3 = Coefficient of Sex

X4 = Exercise-induced Angina (exang)

β4 = Coefficient of Exercise-induced Angina

X5 = Type of Chest Pain 1(cp)

β5 = Coefficient of Type of Chest Pain 1

X5 = Type of Chest Pain 1(cp)

β6 = Coefficient of Type of Chest Pain 2

X6 = Type of Chest Pain 2 (cp)

β7 = Coefficient of Type of Chest Pain 3

X7 = Type of Chest Pain 3 (cp)

β8 = Coefficient of Interaction between Age and Max Heart Rate (age:thalach)

β9 = Coefficient of Age ^ 2

The general form of this model appears:

Written in the natural log of odds to express beta terms in linear form:

In the natural log of odds, π represents the probability for heart disease, not the value of pi as it normally constitutes. represents the odds of success. In this model it represents the probability of heart disease.

Applying the General Model, we use statistical software to arrive at a summary, shown below:

Text, table

Description automatically generated with medium confidence

The general form of this model becomes:

Written in the natural log of odds to express beta terms in linear form becomes:

### **Evaluating Model Significance**

In the interests of assessing this model’s validity we will apply a Hosmer-Lemeshow goodness of fit test. A Hosmer-Lemeshow test assesses whether a model is appropriate for the dataset by assessing whether model predictions are close to the observed values of the response variable (in this case the actual instances of credit default vs predicted instances).

The Null Hypothesis is,

H0: The model fits the data

The Alternative Hypothesis is,

Ha: The model does not fit the data

We will test at a 5% level of Significance,

α = 0.05

Text

Description automatically generated

The P-Value of 0.1048 is larger than the Significance Level 0.05, we fail to reject the Null Hypothesis. The Model fits the data.

Following a Hosmer-Lemeshow test we can further validate the model by testing its individual factors with a Wald Test. Each individual Wald Test can be summed as such,

Null Hypothesis, Coefficient of the predictor variable is 0

βi = 0

Alternative Hypothesis, Coefficient of the predictor variable does not equal 0

βi ≠ 0

The p-values of these individual tests can be found in the PR(>|z|) column of the Summary for this model:

Text

Description automatically generated

According to the Wald Test every term is significant in this model except for Age and Age ^2.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | P-Value | Significance Level | Null Hypothesis: Reject or Fail Reject |
| Maximum Heart Rate (thalach) | 0.014760 | 0.05 | Reject |
| Age (age) | 0.510325 | 0.05 | Fail Reject |
| Sex 1 (Male) | 0.00000191 | 0.05 | Reject |
| Exercise Induced Angina (exang1) | 0.009133 | 0.05 | Reject |
| Chest Pain 1 (cp1) | 0.000249 | 0.05 | Reject |
| Chest Pain 2 (cp2) | 0.00000221 | 0.05 | Reject |
| Chest Pain 3 (cp3) | 0.003684 | 0.05 | Reject |
| Age ^ 2 | 0.810599 | 0.05 | Fail Reject |
| Maximum Heart Rate: Age (thalach:age) | 0.043666 | 0.05 | Reject |

We will use a Confusion Matrix based upon the model’s prediction of heart disease. From this we will determine the values of Accuracy, Precision and Recall.

**Table

Description automatically generated**

For simplicity, True Negative = 103, False Negative = 27, True Positive = 138, False Positive = 35.

An additional test to assess this model is the ROC curve. An ROC curve measures the performance for a classifier at various threshold settings. The area under the curve is an indicator of how well the model distinguishes between Y = 0 or Y =1 (in this model, whether someone has heart disease). In the ROC curve graph below you can see there is a large area beneath the line, indicating this model does a good job of distinguishing between Y = 0 or Y = 1.

Chart

Description automatically generated

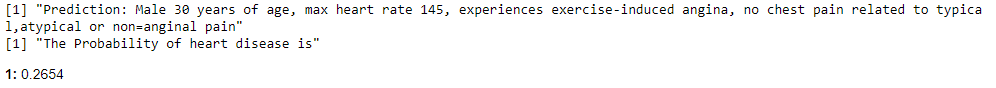
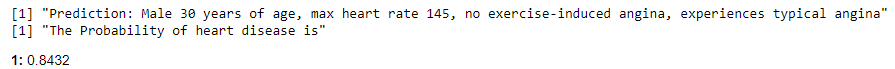
A picture containing diagram

Description automatically generated

The value of AUC (Area under curve) assessed is 0.8777. This indicates that 87.77% of the area of this plot is within the curve. There is a 87.77% likelihood this model can distinguish between heart disease and non-heart disease patients.

### **Making Predictions Using Model**

Now that we have tested the validity of this model, we can make predictions for a few scenarios.

1. What is the probability 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?
   1. 
   2. There is a 26.54% probability of heart disease based upon these factors.
   3. The odds of heart disease occurring based upon these factors is 4939:5061. In odds this means .2654 / (1 - .2654) = .3613. There is a 3,613:10,000 (3,613 in 10,000) chance of heart disease.
2. What is 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?
   1. 
   2. There is an 84.32% probability of heart disease based upon these factors.
   3. In odds this means .8432/ (1 - .8432) = 5.3794. There is a 53,794:10,000 (53,794 in 10,000) chance of heart disease.

It can be concluded that experiencing angina will significantly increase your likelihood of heart disease. This makes sense as signs an symptoms of angina are chest pain, feelings of weight and crushing on chest, pain in upper body, sickness, exhaustion, shortness of breath, lightheadedness and possibly vomiting.

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

### **Reporting Results**

Firstly, we will be splitting the data into training and validation sets. We will be using a Classification Decision Tree because the response variable, target, is a Qualitative variable (results in 1 or 0, True or False). A training set of data will be used to train our Classification Decision Tree. This fits it to the data, which in turn allows for more accurate predictions. The Validation set of data will be used to validate how well the model was trained with the training data. The total size of the data set is 303 rows. 242 rows will account for the Training Set and 61 rows will account for the Validation set. An 80/20 split of the original 303 rows.

Next, we will use a method called Random Forest to determine the best number of trees for this model. The idea is to randomly generate decision trees and up to 25 decision trees and plot the classification errors in relation to the number of trees. This way we can find the optimal number of trees to use before the returns are minimal. Below is a graph that indicates the training and test data set plotted against their classification errors. Specifically we use this graph to look for what amount of trees will yield the lowest classification error.Chart, histogram

Description automatically generated

Based upon this graph, the classification errors become steady at around 15 trees. Beyond 15 trees the error rate remains roughly the same. We will use 15 trees for this data set.

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

We will use a Confusion Matrix based upon the model’s assessment of the Training and Testing data. From this we will determine the values of Accuracy, Precision and Recall.

**Training Set:**

Text

Description automatically generated with medium confidence

For simplicity, True Negative = 111, False Negative = 1, True Positive = 129, False Positive = 1.

**Testing Set:**

**Text

Description automatically generated with low confidence**

For simplicity, True Negative = 16, False Negative = 9, True Positive = 26, False Positive = 10.

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

### **Reporting Results**

Firstly, we will be splitting the data into training and validation sets. We will use a Regression Decision tree because the response variable, Maximum Heart Rate, is reported quantitatively with a continuous integer. A training set of data will be used to train our Regression Decision Tree. This fits it to the data, which in turn allows for more accurate predictions. The Validation set of data will be used to validate how well the model was trained with the training data. The total size of the data set is 303 rows. 242 rows will account for the Training Set and 61 rows will account for the Validation set. An 80/20 split of the original 303 rows.

Next, we will use a method called Random Forest to determine the best number of trees for this model. The idea is to randomly generate decision trees and up to 80 decision trees and plot the classification errors in relation to the number of trees. This way we can find the optimal number of trees to use before the returns are minimal. Below is a graph that indicates the training and test data set plotted against their Root Mean Squared Error. Specifically, we use this graph to look for what number of trees will yield the lowest Root Mean Squared Error.

Graphical user interface, chart, histogram

Description automatically generated

Based upon this graph, RMSE levels out around a 20 Tree Model. We will use 20 trees.

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

Using Root Mean Squared Error we can assess how close, on average, the predicted value was to actual value on the Training and Testing data sets.

Table

Description automatically generated

Based upon these findings, the model could predict Maximum Heart Rate to within 18 (rounded up) heart beats.

## **7. Conclusion**

Today we summarized predicting heart disease using two logistic regression models. Below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Variables | AUC | Accuracy | Precision | Recall |
| Model 1 | Heart disease (target) using variables age (age), resting blood pressure (trestbps), and maximum heart rate achieved (thalach) | 75.75% | 69.31% | 69.78% | 76.97% |
| Model 2 | Heart disease (target) using variables maximum heart rate achieved (thalach), age of the individual (age), sex of the individual (sex), exercise-induced angina (exang), and type of chest pain (cp). Also include the quadratic term for age and the interaction term between age and maximum heart rate achieved. | 87.77% | 79.54% | 79.77% | 83.64% |

Based upon the summary table above there is consistent evidence to recommend model 2. All values assessing the model’s utility are higher. However, there are disclaimers that come with that suggestion. Model 2 has a 79% accuracy. Nearly 2/10 patients will not be correctly classified as having heart disease. Any predictions of this model should be utilized in conjunction with other patient health data. Solely relying on this model would be unwise when deciding upon a patient’s care.

This report also predicted heart disease with a Random Forest model. A logistic regression model and a random forest model both use different approaches. The logistic regression model is determining likelihoods based upon a linear equation, while a random forest model is randomly testing decision nodes in an iterative fashion with a various amount of trees and then averaging the results of multiple trees. On paper, you could reason that it’s like trying the same method 15 times and then averaging 15 results instead of only getting one attempt. Below is a summary of the key performance indicators Accuracy Precision and Recall between the Logistic Models and the Random Forest Classification Model for predicting heart disease.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variables | Accuracy | Precision | Recall |
| Logistic Model 1 | Heart disease (target) using variables age (age), resting blood pressure (trestbps), and maximum heart rate achieved (thalach) | 69.31% | 69.78% | 76.97% |
| Logistic Model 2 | Heart disease (target) using variables maximum heart rate achieved (thalach), age of the individual (age), sex of the individual (sex), exercise-induced angina (exang), and type of chest pain (cp). Also include the quadratic term for age and the interaction term between age and maximum heart rate achieved. | 79.54% | 79.77% | 83.64% |
| Random Forest Classification Model | 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) | 99.17% | 99.23% | 99.23% |

Based upon Accuracy, Precision and Recall all being almost 20% better in every instance I would be comfortable recommending the Random Forest Classification Model. Furthermore I would advise being more comfortable using this model to guide patient care.

These analyses have many practical implications. Through Logistic and Regression Modelling we’ve demonstrated one of many use-cases of Machine Learning. A Random Forest model is a powerful tool for predicting quantitative and qualitative variables in real world scenarios. These predictions can be used to help people and society make data informed decisions.