# **MAT 303 Project Two Summary Report**

Ashley De Venuto

ashley.devenuto@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## **1. Introduction**

For this project, the dataset that is being used is heart disease and its risk factors. This historical data will help with analyzing the connection between heart disease and specific factors. The dataset will be used to create a first logistic regression model, a second logistic regression model, a random forest classification model, and a random forest regression model. The models created will help predict the possibility of heart disease given certain factors of a patient’s medical history.

## **2. Data Preparation**

The dataset has 14 columns which represent the variables. The dataset has 303 rows which represent the records within the dataset. The main variables that are being used in this project are target as response variable for both the first and second logistic regression model. The first logistic regression model uses age, resting blood pressure (trestbps), exercise induced angina (exang), and maximum heart rate achieved (thalach) as the predictor variable. The second logistic regression model use age, resting blood pressure (trestbps), type of chest pain experienced (cp), and maximum heart rate achieved (thalach).

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

### **Reporting Results**

The general form of the first regression model is:

The prediction equation of a logistic regression model is:

The pi

 in the equation equals the probability of an event happening. For this specific problem/equation, it determines the probability of having heart disease. The term Equation

A fraction with pi in the numerator, and one minus pi in the denominatoris the ratio of the probability at which the event happens.

The prediction equation model in terms of natural log of odds is:

When substituting with beta terms with values obtained from the logistic regression model, the prediction model looks like this (using log of odds):

The estimated coefficient of the maximum heart rate achieved is 0.0311 (rounded to four decimal places). This means that the change in log odds on average for target is 0.00311 (rounded to four decimal places) for each increase in maximum hate rate achieved. This only happens if all other variables are constant.

### **Evaluating Model Significance**

To be able to tell if this model is a good fit, a Hosmer-Lemeshow goodness of fit test was performed. The test consists of finding the null hypothesis, the alternative hypothesis, and finding the p-value and comparing it to a 5% level of significance. The null hypothesis looks like this:

The alternative hypothesis looks like this:

The test statistic is:

44.622

The p-value is:

0.612

Since the P-value is greater than the 0.05 or 5% significance level, the null hypothesis is not rejected. This means that the model fits the data.

When using the Wald’s test, the null hypothesis is:

The alternative hypothesis is:

The p-value for age is 0.3060 which is greater than the 0.05 or 5% significance level. This means that this variable is not significant when it comes to determining heart disease. The p-value for resting blood pressure is 0.0741 which is greater than the 0.05 or 5% significance level. This means that this variable is not significant when it comes to determining heart disease. The p-value for exercised induced angina is 1.07e-07 which is less than the 0.05 or 5% significance level. This means that this variable is significant when it comes to determining heart disease. The p-value for maximum heart rate achieved is 1.92e-05 which is less than the 0.05 or 5% significance level. This means that this variable is significant when it comes to determining heart disease. The results show that two of the variables are not significant for determining heart disease (age and resting blood pressure) and two of the variables are significant for determining heart disease (exercised induced angina and maximum heart rate achieved).

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

A screenshot of a computer

Description automatically generated

The true negatives is 89, the false positives is 49, the false negatives is 31, and the true positives is 134.

Accuracy = ()

Accuracy = (

Accuracy = or 73.6%

Precision = (

Precision = (

Precision = or 73.2%

Recall = (

Recall = (

Recall = = 0.8121 or 81.2%

The Receiver Operating Characteristic (ROC) curve is used to measure the accuracy of the differential between Y = 0 and Y = 1. This model has the ability to accurately predict binary classes that are directly correlated to the size of the area under the curve. The bigger the AUC is means it is better accurate. The following image shows the ROC curve:

A graph with a line

Description automatically generated with medium confidence

The AUC is 0.8007. This means that the accuracy rate of predicting whether someone will or has heart disease is about 80.07%.

### **Making Predictions Using Model**

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 122, has exercise induced angina, and has maximum heart rate of 140 is 0.2716 or 27.16%. To find the odds, the probability is divided by 1 – the probability:

which is about a 1 to 3.73 odds

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 130, does not have exercise induced angina, and has maximum heart rate of 165 is 0.7853. to find the odds, the probability is divided by 1 – the probability:

which is a 3.66 to 1 odds

The probability represents the chance of an individual having or developing heart disease based on these variables vs the odds of which represent how likely it is. The first prediction has odds of 1:3.73. This can mean that 1 out of about 4 people with these characteristics will develop or have heart disease. Prediction 2 has a probability of 79% chance of having or developing heart disease. This would give this prediction have odds of 3 out of 4 people with these characteristics will develop or have heart disease.

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

### **Reporting Results**

The general form of the equation for a multiple regression model is:

The prediction equation of the logistic multiple regression model is:

The prediction equation model in terms of natural log of odds is:

When substituting with beta terms with values obtained from the logistic regression model, the prediction model looks like this (using log of odds):

### **Evaluating Model Significance**

To be able to tell if this model is a good fit, a Hosmer-Lemeshow goodness of fit test was performed. The test consists of finding the null hypothesis, the alternative hypothesis, and finding the p-value and comparing it to a 5% level of significance. The null hypothesis looks like this:

The alternative hypothesis looks like this:

The test statistic is:

52

The p-value is:

0.3209

Since the P-value is greater than the 0.05 or 5% significance level, the null hypothesis is not rejected. This means that the model fits the data.

The p-value for age is 0.5136 which is greater than the 0.05 or 5% significance level. This means that age has little to no significance to developing or having heart disease. The p-value for resting blood pressure is 0.0292 which is less than the 0.05 or 5% significance level. This means that resting blood pressure is significant with developing or having heart disease. The p-values for cp1, cp2, and cp3 are 1.61e-05, 4.45e-09, and 0.0012 which are all less than the 0.05 or 5% significance level. This means that cp1, cp2, and cp3 is significant with developing or having heart disease. The p-value for maximum heart rate achieved is 0.0078 which is less than the 0.05 or 5% significance level. This means that maximum heart rate achieved is significant with developing or having heart disease. The p-value for the age squared is 0.6303 which is greater than the 0.05 or 5% significance level. This means that age squared is not significant with developing or having heart disease. The p-value for age : maximum heart rate is 0.0362 which is less than the 0.05 or 5% significance level. This means that age : maximum heart rate is significant with developing or having heart disease. The results show that all the variables except for age and age squared have a significant relationship with developing or having heart disease.

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

A white rectangular object with black text

Description automatically generated

The true negatives is 102, the false positives is 36, the false negatives is 36, and the true positives is 129.

Accuracy = ()

Accuracy = (

Accuracy = or 76.2%

Precision = (

Precision = (

Precision = or 78.2%

Recall = (

Recall = (

Recall = = 0.7818 or 78.2%

The Receiver Operating Characteristic (ROC) curve is used to measure the accuracy of the differential between Y = 0 and Y = 1. This model has the ability to accurately predict binary classes that are directly correlated to the size of the area under the curve. The bigger the AUC is means it is better accurate. The following image shows the ROC curve:

A graph of a function

Description automatically generated

The AUC is 0.8478. This means that the accuracy rate of predicting whether someone will or has heart disease is about 84.78%.

### **Making Predictions Using Model**

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 115, does not experience chest pain, and has a maximum heart rate of 133 is 0.2188 or 21.88%. To find the odds, the probability is divided by 1 – the probability:

which is about a 1 out of 5 odds

The probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 125, experiences typical angina, and has a maximum heart rate of 155 is 0.8007 or 80.1%. To find the odds, the probability is divided by 1 – the probability:

which is about a 4 to 1 odds

The probability represents the chance of an individual having or developing heart disease based on these variables vs the odds of which represent how likely it is. The first prediction has odds of 1:2.80. This can mean that 1 out of about 5 people with these characteristics will develop or have heart disease. Prediction 2 has a probability of 80% chance of having or developing heart disease. This would give this prediction have odds of 4 to 1 people with these characteristics will develop or have heart disease.

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

### **Reporting Results**

While working with the random forest classification model, the data had to be split into a training subset and a testing subset. The model used 85% split for the training subset and 15% split for the testing subset. The original dataset has 303 rows, the training set has 257 rows, and the testing set has 46 rows.

When graphing the training and testing error against the number of trees with a maximum of 150 trees, the visualization of the graph looks like this:

A graph of a number of trees

Description automatically generated

The optimal number of trees for the random forest model is around 20. The training set starts to level out at 20 and the errors between the training set and testing set remain close to equal.

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

A screenshot of a graph

Description automatically generated

Training Set

Accuracy = ()

Accuracy = (

Accuracy = or 99.2%

Precision = (

Precision = (

Precision = or 98.6%

Recall = (

Recall = (

Recall = = 1 or 100%

A screenshot of a graph

Description automatically generated

Testing Set

Accuracy = ()

Accuracy = (

Accuracy = or 67.4%

Precision = (

Precision = (

Precision = or 74.1%

Recall = (

Recall = (

Recall = = 0.7143 or 71.4%

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

### **Reporting Results**

While working with the random forest regression model, the data had to be split into a training subset and a testing subset. The model used 80% split for the training subset and 20% split for the testing subset. The original dataset has 303 rows, the training set has 242 rows, and the testing set has 61 rows.

A graph of a number of trees

Description automatically generated

I believe that the optimal number of trees to use is around 18 trees.

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

A black and white text on a white background

Description automatically generated

The root mean squared error for the training set is 11.7266. The root mean squared error for the testing set is 21.205.

## **7. Conclusion**

If I were to recommend one of the two logistic regression models, I would recommend using the second logistic regression model to predict heart disease. For the second regression model, the accuracy and precision are higher than the first logistic regression model. The AUC of the second logistic regression model is also greater than that of the first logistic regression model. This indicates a better fit of the data set while using the second logistic regression model. The second logistic regression model can use additional variables, which many of these variables are statistically significant in predicting heart disease. If I were to recommend using either random forest classification model or the logistic regression model, I would recommend using the random forest model. When using the training set within the random forest classification model, it has the highest accuracy, precision, and recall values out of all the models used within this analysis. The importance of this analysis is using these models to predict the probability and odds of someone having heart disease based on certain variables from the historic dataset. Medical professionals can use these models to help predict heart disease and recommend treatments to those with a high probability. These treatments will then help gather more data and help lower the probability of heart disease for other patients.