# **MAT 303 Project Two Summary Report**

Sarah Steinbaum

sarah.steinbaum@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The statistical analyses that will be conducted in this report include a first logistic regression model, second logistic regression model, random forest classification model, and a random forest regression model. A large set of historical data will be used to find risk factors for heart disease. The logistic regression models and random forest classification model will be used to determine the likely it is for an individual to have heart disease based on variables such as age, resting blood pressure, and maximum heart rate achieved. The random forest regression model will be used to determine the maximum heart rate achieved based on variables such as age, sex, cholesterol measurement, and number of major blood vessels.

## **2. Data Preparation**

There are many important variables in the data set for heart data. Heart disease (*target*) is used as a response variable with 0 indicating no heart disease and 1 indicating heart disease. Age in number of years (*age*), resting blood pressure (*trestbps*), maximum heartrate achieved (*thalach*), whether an individual experiences exercise induced angina (*exang*) with 0 indicating no and 1 indicating yes are the important variables utilized in the first model. The type of chest pain experienced (*cp*) is presented in the second model and is a qualitative variable in which 0 indicates no chest pain, 1 indicates typical angina, 2 indicates atypical angina, and 3 indicates non-anginal pain. The third model adds variables including an individual’s sex (*sex*) with 1 indicating male and 0 indicating female, cholesterol measurement in mg/dl (*chol*), number of major vessels (*ca*) from 0 to 3, the slope of peak exercise (*slope*) with 1 indicating unsloping, 2 indicating a flat slope, and 3 indicates downsloping. Resting electrocardiographic measurement (*restecg*) in which 0 indicates normal, 1 indicates an individual has ST-T wave abnormality, 2 shows probable or definite left ventricular hypertrophy by Estes’ criteria are also introduced in this model. This dataset contains 303 rows 14 columns each representing a different variable.

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

### **Reporting Results**

The general form of a logistic regression model:

This model can be transformed to form a model that is linear in beta terms:

Since the left side of the above equation is the natural log of odds, it can be written as:

+

The prediction equation of a logistic regression model:

+

In this model, y and *odds* represent heart disease (*target*) as the response variable, represents age (*age*), and represents resting blood pressure (*trestbps*), represents exercise induced angina (*exang*), and represents maximum heartrate achieved (*thalach*). The symbol, , represents the probability of an individual to have heart disease. is the ratio of the probability an individual will have heart disease. Once the R script has ran, we can place the beta estimates into the equation.

The logistic regression model:

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

The estimated coefficient of maximum heartrate achieved is 0.031095. On average, this means that the change in log odds for developing heart disease is 0.031095 for each percentage increase in maximum heartrate achieved, holding all other variables constant.

### **Evaluating Model Significance**

We will now conduct the Hosmer-Lemeshow Goodness of Fit (GOF) test to determine whether or not the model is appropriate for the data set. The null hypothesis is that the model does fit the data set. The alternative hypothesis is that the model does not fit the data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| First Logistic Regression Model | = The model fits the data  = The model does not fit the data | *t* = 44.622 | *p* = 0.612 | Fail to reject the null hypothesis |

The P-value is greater than the 5% level of significance which indicates that there is sufficient evidence to fail to reject the null hypothesis. We can conclude that the model is appropriate for this data set. We will now conduct the Wald’s test in order to find out which terms in the model are significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Age  (*age*) |  | *t* = -1.024 | *p* = 0.3060 | Fail to reject the null hypothesis |
| Resting blood pressure  (*trestbps*) |  | *t* = -1.786 | *p* = 0.7041 | Fail to reject the null hypothesis |
| Exercise induced angina  (*exang*) |  | *t* = -5.314 | *p* = 1.07E-07 | Fail to reject the null hypothesis |
| Maximum heartrate achieved  (*thalach*) |  | *t* = 4.274 | *p* = 1.92E-05 | Reject the null hypothesis |

The variables for exercise induced angina and maximum heart rate achieved have a P-value less than the 5% level of significance which indicates that these variables are statistically significant and have a correlation to heart disease. However, the variables for age and resting blood pressure have a P-value greater than the 5% level of significance which indicates that these variables are not statistically significant and do not have a correlation to heart disease.

The values for the confusion matrix are listed below:

* True positive = 134
* True negative = 89
* False positive = 49
* False negative = 31

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 0.7359736

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 0.73224044

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 0.81212121

Chart, line chart

Description automatically generated

The graph above shows theReceiver Operating Characteristic (ROC) curve. This can be considered a probability curve and is a measurement of the performance of a classifier at different threshold settings. The value of the Area Under the Curve (AUC) is 0.8007. The AUC represents the measure of separability. Essentially, this chart tells us how accurate it is at distinguishing the binary responses, 0s as 0s and 1s as 1s. A higher AUC indicates increased accuracy.

### **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 a maximum heart rate of 140 is 0.2716. This indicates that there is a 27% likelihood that an individual with these characteristics would have heart disease. The odds of an individual having heart disease with these characteristics are 27 to 73. The probability 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 a maximum heart rate of 165 is 0.7853. This indicates that there is roughly a 79% likelihood that an individual with these characteristics would have heart disease. The odds of an individual having heart disease with these characteristics are 79 to 21.

Both scenarios have the individual’s age as 50 years old and the resting blood pressure is very similar with only a difference of 8 mmHg (millimeters of mercury) difference. However, the individual’s maximum heart rate achieved has a 25 bpm (beats per minute) difference which is a much larger difference. In the first prediction, the individual has exercise induced angina and in the second prediction they do not. Based on this analysis we can determine that an individual’s maximum heart rate achieved and whether they have exercise induced angina can have an impact on whether or not they have heart disease.

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

### **Reporting Results**

The general form of a logistic regression model:

This model can be transformed to form a model that is linear in beta terms:

+ ++

Since the left side of the above equation is the natural log of odds, it can be written as:

+ ++

The prediction equation of a logistic regression model:

+

In this model, y and *odds* represent heart disease (*target*) as the response variable, represents age (*age*), and represents resting blood pressure (*trestbps*). The type of chest pain experienced (*cp*) is a qualitative variable that can be no chest pain (*cp*=0), typical angina (*cp*=1), atypical angina (*cp*=2), or nonanginal pain (*cp*=3). This variable is represented by the dummy variables , , and . The variable, maximum heart rate achieved (*thalach*), is represented by . The quadratic term for age2 is represented by . The interaction term for age and maximum heart rate achieved is represented by . The symbol, , represents the probability of an individual to have heart disease. is the ratio of the probability an individual will have heart disease. Once the R script has ran, we can place the beta estimates into the equation.

The logistic regression model:

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

### **Evaluating Model Significance**

We will now conduct the Hosmer-Lemeshow Goodness of Fit (GOF) test to determine whether or not the model is appropriate for the data set. The null hypothesis is that the model does fit the data set. The alternative hypothesis is that the model does not fit the data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Second Logistic Regression Model | = The model fits the data  = The model does not fit the data | *t* = 52 | *p* = 0.3209 | Fail to reject the null hypothesis |

The P-value is greater than the 5% level of significance which indicates that there is sufficient evidence to fail to reject the null hypothesis. We can conclude that the model is appropriate for this data set. We will now conduct the Wald’s test in order to find out which terms in the model are significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Age  (*age*) |  | *t* = 0.653 | *p* = 0.51357 | Fail to reject the null hypothesis |
| Resting blood pressure  (*trestbps*) |  | *t* = -2.181 | *p* = 0.02916 | Reject the null hypothesis |
| Type of chest pain experienced  Typical angina  (*cp1*) |  | *t* = 4.313 | *p* = 1.61E-05 | Reject the null hypothesis |
| Type of chest pain experienced  Atypical angina  (*cp2*) |  | *t* = 5.867 | *p* = 4.45E-09 | Reject the null hypothesis |
| Type of chest pain experienced  Nonanginal  (*cp3*) |  | *t* = 3.245 | *p* = 0.00117 | Reject the null hypothesis |
| Maximum heartrate achieved  (*thalach*) |  | *t* = 2.663 | *p* = 0.00775 | Reject the null hypothesis |
| Age2  (*age2*) |  | *t* = 0.481 | *p* = 0.63025 | Fail to reject the null hypothesis |
| Age: maximum heart rate achieved  (age:thalach) |  | *t* = -2.095 | *p* = 0.03616 | Reject the null hypothesis |

The terms for resting blood pressure, chest pain characterized by typical angina, chest pain characterized by atypical angina, chest pain that is nonanginal, maximum heartrate achieved, and the interaction term *age:thalach* all have a P-value less than the 5% level of significance which indicates these terms are statistically significant and have a correlation to heart disease. The terms *age* and *age2*

Both have a P-value greater than the 5% level of significance which indicates these terms are not statistically significant and therefore do not have a correlation to heart disease according to this model.

The values for the confusion matrix are listed below:

* True positive = 129
* True negative = 102
* False positive = 36
* False negative = 36

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 0.76237624

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 0.78181818

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 0.78181818

Chart

Description automatically generated

TheReceiver Operating Characteristic (ROC) curve is presented on the graph above. This is a measurement on the performance of a classifier at different threshold settings. The value of the Area Under the Curve (AUC) is 0.8478. The AUC represents the measure of separability. The ROC tells us how accurate the model is at distinguishing the binary responses, 0s as 0s and 1s as 1s. A higher AUC indicates increased accuracy.

### **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 maximum heart rate of 133 is 0.2188. This indicates that the likelihood of someone having heart disease with these characteristics is roughly 22% and that the odds of someone having heart disease with these characteristics are 11 to 39. 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 maximum heart rate of 155 is 0.8007. This indicates that the likelihood of someone having heart disease with these characteristics is roughly 80% and that the odds of someone having heart disease with these characteristics are 4 to 1.

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

### **Reporting Results**

When using the set.seed(6522048) using a 85% and 15% split the heart disease data set has 303 rows. The training set has 257 rows and the validation set has 46. The graph below displays the training and testing error against the number of trees using a classification random forest model. The optimal number of trees for this model is around 20 as the lines for the training set and testing set tend to level off at that point. It is important to find the most optimal number of trees for the random forest model so that the model does not over fit the training data.

Graphical user interface, application

Description automatically generated

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

The values for the training confusion matrix are listed below:

* True positive = 137
* True negative = 120
* False positive = 0
* False negative = 0

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 1.0

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 1.0

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 1.0

The values for the testing confusion matrix are listed below:

* True positive = 22
* True negative = 13
* False positive = 5
* False negative = 6

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 0.76086957

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 0.81481481

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 0.78571429

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

### **Reporting Results**

When using the set.seed(6522048) using a 80% and 20% split the heart disease data set has 303 rows. The training set has 242 rows and the validation set has 61. The graph below displays the training and testing error against the number of trees using a classification random forest model. The optimal number of trees for this model is around 15 as the lines for the training set and testing set tend to level off at that point.

Chart, line chart

Description automatically generated

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

The root mean squared error for the training set is 11.7971359887268. The error for the training set is small as the model knows the data very well. The root mean squared error for the testing is 21.4423080795163. The error for the testing set is larger because this data has not been seen by the model before.

## **7. Conclusion**

The two logistic regression models assisted in determining risk factors for the presence of heart disease. We conducted a Hosmer-Lemeshow Goodness of Fit (GOF) test for the first model and determined that the model does fit the dataset. We then conducted the Wald’s test in order to find out which terms in the model were significant and found that the variables for age, resting blood pressure, and exercise induced angina had a P-value less than the 5% level of significance indicting these terms are statistically significant. We then obtained the values for the model’s confusion matrix and determined the value for accuracy was 0.7359736, precision was 0.73224044, and recall was 0.81212121. Next, we created a graph for the Receiver Operating Characteristic (ROC) curve and determined the value of the Area Under the Curve (AUC) is 0.8007.

We also conducted a Hosmer-Lemeshow Goodness of Fit (GOF) test for the second logistic regression model and determined the model does fit the data set. The Wald’s test concluded that the variables for resting blood pressure, chest pain with typical angina, atypical angina, and nonanginal pain, maximum heartrate achieved, and the interaction term for age and maximum heartrate achieved each had a P-value less than the 5% level of significance and are statistically significant. After obtaining the values for the confusion matrix we concluded the value for accuracy was 0.76237624, precision was 0.78181818, and recall was 0.78181818. After creating the ROC cure, we determined the AUC value was 0.8478. Between the two regression models, I would choose the second model as it has more statistically significant variables, a higher AUC value, and higher values for accuracy and precision.

I would recommend using the random forest classification model instead of the logistic regression model as it outperforms the other models in accuracy, precision, and recall for both the training and testing confusion matrix. The analyses that were performed have a significant practical importance. Healthcare providers can use this information in order to determine what factors can lead to heart disease and take preventative measures according to the patient’s needs and optimize patient outcomes.

## **8. Citations**

Berrier, J. (2016). MAT 303: Applied Statistics 2 for Science. Zyante Inc. (zyBooks.com)