# **MAT 303 Project Two Summary Report**

Kayla Sacks

Kayla.sacks@snhu.edu

Southern New Hampshire University

## **1. Introduction**

For this analysis I will use a data set that explores the relationship between different health indicators and the presence of heart disease. I will start my analysis by creating two logistic regression models and testing how well they can predict if an individual is at risk for heart disease. I will then create a classification random forest model that can be used to predict the risk of heart disease using the chosen variables. The final model I will create is a regression random forest which can be used to predict the maximum heart rate achieved based on the chosen variables.

The significance of logistic regression models will be evaluated using Hosmer-Lemeshow goodness of fit test, Wald’s test, creating a Receiver Operating Characteristic (ROC) curve, and the area under the ROC curve (AUC). Each of the random forest models will be split into a training set and validation set. The accuracy, precision, and sensitivity for the first three models will also be calculated. To evaluate the final model, the root mean squared error will be calculated.

The results of this analysis are important as they can be used to evaluate patient records to look for risks that can indicate the possibility of heart disease that may not be easy to find for doctors. If the individual and their doctor are aware of the risk ahead of time, they will be able to reduce it and help the patient to improve their health.

## **2. Data Preparation**

The data set that is being used for this analysis contains 13 variables and 303 rows. The data set will be split into the training set and validation set for the third and fourth models. The variables I will be using for this analysis are:

* target – If the individual has a heart disease, this will be the dependent variable for the first three models. 0 = no, 1 = yes.
* age – The age in years of the individual, a predictor variable for all models.
* trestbps – The individuals resting blood pressure, a predictor variable for all models.
* exang – Exercise-induced angina, a predictor variable for all models. 0 = no, 1 = yes.
* thalach – The individual’s maximum heart rate achieved, a predictor variable for models 1 and 2 and the dependent variable for the final model.
* cp – Type of chest pain, a predictor variable for model 2 through 4. 0 = no pain, 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain.
* sex – The individual’s sex, a predictor variable for models 3 and 4. 1 = male, 2 = female.
* chol – The individual’s cholesterol measurement in mg/dl, a predictor variable for models 3 and 4.
* restecg – Resting electrocardiographic measurement, a predictor variable for models 3 and 4. 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes’ criteria
* ca – The number of major vessels (0-3), a predictor variable for models 3 and 4.

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

### **Reporting Results**

The general form for this logistic regression model is:

The prediction regression equation is:

We can transform the prediction model into an equation that expresses the beta terms in linear form.

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

For this model, “odds” is the odds of an individual having a heart disease.

is the probability that the individual has heart disease and is the ratio of the probability, or the odds, that the individual will have a heart disease.

The logistic regression model was created using the four predictors with the summary below.

A screenshot of a computer

Description automatically generated

The prediction model equation using the outputs from the model is:

The prediction model in terms of the natural log of odds will then be:

The estimated coefficient of the maximum heart rate achieved is 0.0311. This tells us that, on average, the change in the log odds for having a heart disease is 0.0311 for each unit increase in maximum heart rate if there is no change in other variables. We can express this in terms of odds by calculating:

With this we can say that if all other variables remain constant, with each unit increase in maximum heart rate, the odds of having a heart disease increases by 0.0316.

### **Evaluating Model Significance**

The Hosmer-Lemeshow goodness of fit test will be used to see if the model is appropriate for the data set. We will start with the null and alternate hypotheses which are:

For this test I found the p-value for group sizes 10 through 50 to check if it varied significantly. Also ran the full Hosmer-Lemeshow test for the group sizes of 10 and 50 to get the test statistics (x-squared values) which are 9.192 and 44.622 respectively. Although the value varied between 0.0582 and 0.8471, they are all above the 0.05 level of significance and therefore will not reject the null hypothesis.

Next will be Wald’s test to determine with terms are statistically significant using the null and alternate hypotheses a 5% level of significance.

is the term for age and has a p-value of 0.3060 which is greater than the 0.05 level of significance and will not reject the null hypothesis.

is the term for resting blood pressure and has a p-value of 0.0741 which is greater than the 0.05 level of significance and will not reject the null hypothesis.

is the term for exercise-induced angina and has a p-value of 1.07e-07 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the term for maximum heart rate and has a p-value of 1.92e-05 which is less than the 0.05 level of significance and will reject the null hypothesis.

With the conclusion of this test, we can see that only exercise-induced angina and maximum heart rate are statistically significant in this model when using a 5% level of significance.

Now we will evaluate the confusion matrix below for the counts for true positives, true negatives, false positives, and false negatives.

*A white rectangular object with black text

Description automatically generated*

True Positives – The actual and predicted values are both 1: 134

True Negatives – The actual and predicted values are both 0: 89

False Positives – The actual value is 0 but predicted value is 1: 49

False Negatives– The actual value is 1 but the predicted value is 0: 31

Using the counts from the matrix, accuracy was calculated to find the ratio of correct predictions to the total number of observations.

Precision was calculated to find the ratio of correct positive predictions to the total number of correct predictions.

Sensitivity was calculated to find the ratio of correct positive predictions to the total number of positives.

The Receiver Operating Characteristic (ROC) curve below shows the accuracy of the model by visualizing the ratio of true positives to false positives.

A graph with a line

Description automatically generated

The area under the curve (AUC) for the ROC is 0.8007 or 80.07% and tells us how well the model distinguishes between Y = 0 and Y =1. The closer the AUC is to 1 (or 100%), the better the model’s prediction. Although the value is not excellent, an AUC of 0.8007 is still very good.

### **Making Predictions Using Model**

**Prediction 1**

The probability of an individual that is 50 years old, has a resting blood pressure of 122, has exercise induced angina, and has maximum heart rate of 140 of having a heart disease is 0.2716 or 27.16%. To find the odds, we can take the probability divided by 1 minus the probability which is . From this, we can say that the odds of an individual who meets the above criteria having a heart disease is 0.3729.

**Prediction 2**

The probability of an individual that is 50 years old, has a resting blood pressure of 130, does not have exercise induced angina, and has maximum heart rate of 165 of having a heart disease is 0.7853 or 78.53%. The odds of an individual who meets the above criteria having a heart disease is which is 3.6577.

From these two predictions, we can see that the second prediction shows much higher odds of the individual having heart disease even though the age did not change. From this, we can deduce that an increase in resting blood pressure, maximum heart rate, and/or not having exercise induces angina indicate higher odds of an individual having a heart disease.

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

### **Reporting Results**

The general form for this logistic regression model is:

The prediction regression equation is:

We can transform the prediction model into an equation that expresses the beta terms in linear form.

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

For this model, “odds” is the odds of an individual having a heart disease.

The logistic regression model was created using these predictors with the summary below.

A screenshot of a computer

Description automatically generated

The prediction model using these outputs in terms of the natural log of odds will then be:

### **Evaluating Model Significance**

The Hosmer-Lemeshow goodness of fit test will be used to see if the model is appropriate for the data set. We will start with the null and alternate hypotheses which are:

For this test I again found the p-value for group sizes 10 through 50 to check if it varied significantly. Although the value varied between 0.1413 and 0.7754, they are all above the 0.05 level of significance and therefore will not reject the null hypothesis. I also ran the full Hosmer-Lemeshow test for the group sizes of 10 and 50 to get the test statistics (x-squared values) which are 6.0481 and 52 respectively.

Next will be Wald’s test to determine with terms are statistically significant using the null and alternate hypotheses a 5% level of significance.

is the term for age and has a p-value of 0.5136 which is greater than the 0.05 level of significance and will not reject the null hypothesis.

is the term for resting blood pressure and has a p-value of 0.02916 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the first dummy term for type of chest pain and has a p-value of 1.61e-05 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the second dummy term for type of chest pain and has a p-value of 4.45e-09 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the final dummy term for type of chest pain and has a p-value of 0.0012 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the term for maximum heart rate and has a p-value of 0.0078 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the term for age² and has a p-value of 0.6303 which is greater than the 0.05 level of significance and will not reject the null hypothesis.

is the interaction term between age and maximum heart rate and has a p-value of 0.0362 which is less than the 0.05 level of significance and will reject the null hypothesis.

With the conclusion of this test, we find that the only two variables that do not meet the 5% level of significance are age and age².

Now we will evaluate the confusion matrix below for the counts for true positives, true negatives, false positives, and false negatives.

*A white rectangular object with black text

Description automatically generated*

True Positives – The actual and predicted values are both 1: 129

True Negatives – The actual and predicted values are both 0: 102

False Positives – The actual value is 0 but predicted value is 1: 36

False Negatives– The actual value is 1 but the predicted value is 0: 36

Using the counts from the matrix, accuracy was calculated to find the ratio of correct predictions to the total number of observations.

Precision was calculated to find the ratio of correct positive predictions to the total number of correct predictions.

Sensitivity was calculated to find the ratio of correct positive predictions to the total number of positives.

The Receiver Operating Characteristic (ROC) curve below shows the accuracy of the model by visualizing the ratio of true positives to false positives.

A graph with a curve

Description automatically generated

The area under the curve (AUC) for the ROC is 0.8478 or 84.78% which shows that this model fits the data slightly better than the first model with an increase of 4.71%.

### **Making Predictions Using Model**

**Prediction 1**

The probability of an individual that is 50 years old, has a resting blood pressure of 115, does not experience chest pain, and has maximum heart rate of 133 of having a heart disease is 0.2188 or 21.88%. The odds of an individual who meets the above criteria having a heart disease is 0.2801.

**Prediction 2**

The probability of an individual that is 50 years old, has a resting blood pressure of 125, experiences typical angina, and has maximum heart rate of 155 of having a heart disease is 0.8007 or 80.07%. The odds of an individual who meets the above criteria having a heart disease is 4.0176.

Comparing the two predictions, we can see that the conditions in the second prediction have a much higher probability of the individual having heart disease. We can deduce that an individual that has lower resting blood pressure, lower maximum heart rate, and/or no chest pain has lower odds of having a heart disease.

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

### **Reporting Results**

For this model, I split the original data set which contains 303 rows into a training set that uses 85% (257) of the rows and the validation set which uses 15% (46) of the rows. I have also used set.seed(6522048) for this model.

This classification random forest model is for presence of heart disease (target) using variables age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), and number of major vessels (ca). I then graphed the error against the trees from a minimum of 1 to maximum of 150 trees.

A graph of a number of trees

Description automatically generated

Since the error for the training data fluctuates slightly, I examined the point where the testing error has a major decrease to determine the optimal number of trees to be 45 for this random forest model.

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

Once the number of trees was found, a confusion matrix was created for both the training set and validation set.

The training confusion matrix is:

**A white rectangular box with black text

Description automatically generated**

|  |  |  |  |
| --- | --- | --- | --- |
| True Positives: 136 | True Negatives: 119 | False Positives: 1 | False Negatives: 1 |

The accuracy for the training set is:

The precision for the training set is:

The sensitivity for the training set is:

Next, the testing confusion matrix is:

A white rectangular object with black text

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| True Positives: 28 | True Negatives: 18 | False Positives: 0 | False Negatives: 0 |

The accuracy for the training set is:

The precision for the training set is:

The sensitivity for the training set is:

**6. Random Forest Regression Model**

### **Reporting Results**

For this model, I split the original data set which contains 303 rows into a training set that uses 80% (242) of the rows and the validation set which uses 20% (61) of the rows. I have also used set.seed(6522048) for this model as well. To find the optimal number of trees, I plotted the root mean squared (RMSE) against the number of trees with a maximum of 80 trees.

A graph of a number of trees

Description automatically generated

Based on the graph above and comparing the RMSE values different number of trees, I found the optimal number of trees to be 30.

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

*To evaluate this model, I will find the root mean square error (RMSE) which is used to find how different the predictions are from the measured values. The equation for RMSE is:*

Using 30 trees the RMSE for the training set is calculated to be 11.5619 and 20.7872 for the testing set.

## **7. Conclusion**

For the conclusion of this analysis, I’ll first compare the two logistic regression models. During the evaluation step, it was concluded that they both fail to reject the null hypothesis for the Hosmer-Lemeshow goodness of fit test which shows both models fit the data. Next during Wald’s test, we found that each model also has more then one variable that is statistically significant. We can start to consider which model would be better to choose to predict heart disease once we move to the confusion matrix which shows the first model has an accuracy and precision of 73.6% and 73.22% respectively whereas the values were 76.24% and 78.18% respectively. So, although the first model has a higher recall percentage, the second has a higher accuracy and precision. The next step that was taken to evaluate the significance for each model was the ROC curve and the area under the curve (AUC) calculation. The second model AUC is 0.8478 and the first model AUC is 0.8007 which indicates that the second models’ predictions will be more accurate. From these observations, I would choose to use the second logistic regression model to predict heart diseases as it has the higher chance to provide more accurate predictions.

When it comes to recommending using the random forest classification or logistic regression model, there are a few points that I believe need to be considered as both models can be helpful when they are used in certain situations. One aspect to consider for this is the size of the overall data set since the data needs to be split into the training and validation sets. This is important as if it is to small, there may not be enough data to train the random forest or to validate it to a decent extent. In this way, a logistic regression model might be a better choice. Another aspect to consider is outliers. Outliers can have a larger impact on logistic regression models but can be easily handled with a random forest since it splits the data, and the final value is determined by the majority result. We can see in this analysis that the random forest classification training and testing sets have better values for accuracy, precision, and recall value than the logistic regression model which indicates a better fit for the data, but we also need to consider if the data set is a significant size to ensure the best predictions for the model.

My final consideration for this analysis is that the logistic regression model is used to find the possibility and odds of an individual having heart disease whereas the random forest classification of no they do not (output of 0) or yes, they do (output of 1). So overall, I believe the decision of which model to use depends on the information that is needed. More data is always better, but I believe that there is enough for the random forest model to be useful if the question needs a yes or no response (has or does not have) whereas the logistic regression model is better for answering “ What is the probability of the patient having heart disease?”.

The practical importance of the analyses can have a high impact in areas related to healthcare. Doctors can use these analyses to predict the likelihood of their patient having or developing heart disease. By detecting any problems early, doctors can create solutions to reduce the risk for their patients. Health insurance companies can also use the analyses to assess the best coverage plan for their customer based on if they are at a higher risk for heart disease. Finally, the analyses can be used by medicine developers by being able to better understand the relationship between the different health indicators and heart disease which can help them to develop better medicine to reduce the risk or effects of heart disease.