# **MAT 303 Project Two Summary Report**

[Full Name]

[SNHU Email]

Southern New Hampshire University

## **1. Introduction**

The data set I will be exploring is heart\_disease.csv. The purpose of using the results from the models created in this project is to research the risk factors for heart disease so we can predict whether a person is at risk for heart disease or not. One of the models in this project will be used to predict the maximum heart rate achieved.

The analyses I will be conducting in this project are 2 logistic regression models, confusion matrix, Hosmer-Lemeshow Goodness of Fit (GOF) test, Wald’s test, a classification random forest model, and regression random forest model.

## **2. Data Preparation**

There are 14 columns and 303 rows. There are a lot of variables used in this project and they are:

* Age
* Sex
* Type of chest pain (cp)
* Resting blood pressure (trestbps)
* Resting electrocardiographic measurement (restecg)
* Maximum heart rate achieved (thalach)
* Exercise-induced angina (exang)
* Slope of the peak exercise ST segment (slope)
* Number of major vessels (ca)
* Heart disease (target)

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

### **Reporting Results**

**Table

Description automatically generated**

The equation above is the general form of a logistic regression model. Where y is 1 for having a heart disease and 0 for not having a heart disease, is age, is resting blood pressure, and is maximum heart rate achieved.

The equation above is another way to form a model that is linear in the beta terms (both equations are the same, the natural log of odds appear to be simpler. It will still build a logistic regression model). Where is the probability of having a heart disease and is the odds.

Logistic regression model in terms of E(y):

Logistic regression model in terms of natural log of odds:

The estimated coefficient of maximum heart rate achieved is 0.0427. Meaning the change in log odds for reaching the maximum heart rate is 0.0427 for each rate increase in maximum heart rate achieved.

### **Evaluating Model Significance**

Hosmer-Lemeshow Goodness of Fit (GOF) Test:

**Text

Description automatically generated**

Null hypothesis:

= the model fits the data

Alternative hypothesis:

= the model does not fit the data

Using 5% level of significance: the p-value of this model is 0.7168. We have insufficient evidence to reject the null hypothesis. Meaning that this model fits the data.

Wald’s Test:

Age predictor:

The z-value of this predictor is -0.586 with a p-value of 0.5578. Since the p-value is greater than the 5% significance level, we have insufficient evidence to reject the null hypothesis. Which concludes that this estimate is not significantly different from 0, suggesting that no statistically significant relationship exists between the predictor and binary response variables.

Resting blood pressure predictor:

The z-value of this predictor is -2.063 with a p-value of 0.0392. Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Maximum heart rate achieved predictor:

The z-value of this predictor is 6.144 with a p-value of 8.06 \* . Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Receiver Operating Characteristic (ROC) Curve:

**Chart, line chart

Description automatically generatedText

Description automatically generated with medium confidence**

An ROC curve measures the performance of the model and see how well it can recognize between the two different classes of a binary response variable (y = 0, y = 1). The AUC (the amount of area covered in the graph represents the value of how good the model is at predicting 0s as 0s and 1s as 1s. The AUC for this model is 0.7575. Which means there is 75.75% chance this model will be able to tell the difference between positive class and negative class.

Confusion Matrix:

**Table

Description automatically generated**

True Positive (TP): 127

True Negative (TN): 83

False Positive (FP): 55

False Negative (FN): 38

Accuracy:

This model displays 69.31% accuracy.

Precision:

This model displays 69.78% precision.

Recall:

The ratio of correct positive predictions to the total positives is 76.97%.

### **Making Predictions Using Model**

**Graphical user interface, text

Description automatically generated**

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 is 0.5339. With the probability of 53.39%, the odds for this prediction 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 is 0.7248. With the probability of 72.48%, the odds for this prediction is:

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

### **Reporting Results**

**Table

Description automatically generated**

The equation above is the general form of a logistic regression model. Where y is 1 for having a heart disease and 0 for not having a heart disease, is maximum heart rate achieved, is age, is the dummy variable for sex (1 = male or 0 = female), is the dummy variable for exercise-induced angina (1 = yes or 0 = no), , , and are the dummy variables for the type of chest pain experienced (0 = no pain, 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain). We also have the quadratic term for age , and interaction term between age and maximum heart rate achieved .

Logistic regression model in terms of E(y):

Logistic regression model in terms of natural log of odds:

### **Evaluating Model Significance**

Hosmer-Lemeshow Goodness of Fit (GOF) Test:

**Text

Description automatically generated**

Null hypothesis:

= the model fits the data

Alternative hypothesis:

= the model does not fit the data

Using 5% level of significance: the p-value of this model is 0.1048. We have insufficient evidence to reject the null hypothesis. Meaning that this model fits the data.

Wald’s Test:

Maximum heart rate achieved predictor:

The z-value of this predictor is 2.438 with a p-value of 0.0148. Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Age predictor:

The z-value of this predictor is 0.658 with a p-value of 0.5103. Since the p-value is greater than the 5% significance level, we have insufficient evidence to reject the null hypothesis. Which concludes that this estimate is not significantly different from 0, suggesting that no statistically significant relationship exists between the predictor and binary response variables.

Sex predictor:

The z-value of this predictor is -4.762 with a p-value of 1.91 \* . Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Exercise-induced angina predictor:

The z-value of this predictor is -2.607 with a p-value of 0.0091. Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Chest pain (1) predictor:

The z-value of this predictor is 3.663 with a p-value of 0.0002. Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Chest pain (2) predictor:

The z-value of this predictor is 4.734 with a p-value of 2.21 \* . Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Chest pain (3) predictor:

The z-value of this predictor is 2.904 with a p-value of 0.0037. Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Quadratic term: age predictor:

The z-value of this predictor is 0.240 with a p-value of 0.8106. Since the p-value is greater than the 5% significance level, we have insufficient evidence to reject the null hypothesis. Which concludes that this estimate is not significantly different from 0, suggesting that no statistically significant relationship exists between the predictor and binary response variables.

Interaction term: age and maximum heart rate achieved predictor:

The z-value of this predictor is -2.017 with a p-value of 0.0437. Since the p-value is less than the 5% significance level, we have sufficient evidence to reject the null hypothesis. Which concludes that this estimate is significantly different from 0, suggesting that a statistically significant relationship exists between the predictor and binary response variables.

Receiver Operating Characteristic (ROC) Curve:

**Chart

Description automatically generatedText

Description automatically generated**

An ROC curve measures the performance of the model and see how well it can recognize between the two different classes of a binary response variable (y = 0, y = 1). The AUC (the amount of area covered in the graph represents the value of how good the model is at predicting 0s as 0s and 1s as 1s. The AUC for this model is 0.8777. Which means there is 87.77% chance this model will be able to tell the difference between positive class and negative class.

Confusion Matrix:

**Table

Description automatically generated**

True Positive (TP): 138

True Negative (TN): 103

False Positive (FP): 35

False Negative (FN): 27

Accuracy:

This model displays 79.54% accuracy.

Precision:

This model displays 79.77% precision.

Recall:

The ratio of correct positive predictions to the total positives is 83.64%.

### **Making Predictions Using Model**

**Text

Description automatically generated**

The probability of 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 is 0.2654. With the probability of 26.54%, the odds for this prediction 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 is 0.8432. With the probability of 84.32%, the odds for this prediction is:

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

### **Reporting Results**

**Graphical user interface, chart, histogram

Description automatically generated**

Splitting the heart disease data set into training and testing sets using 80% and 20% split, there are 242 rows in the training set and there are 61 rows for the testing set. The curve flattens at approximately 15-20 trees so the optimal number of trees for the random forest model is approximately 15-20.

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

*Graphical user interface, text, application, email

Description automatically generated*

*Training data set:*

True Positive (TP): 129

True Negative (TN): 111

False Positive (FP): 1

False Negative (FN): 1

Accuracy:

This model displays 99.17% accuracy.

Precision:

This model displays 99.23% precision.

Recall:

The ratio of correct positive predictions to the total positives is 99.23%.

*Testing data set:*

True Positive (TP): 26

True Negative (TN): 16

False Positive (FP): 10

False Negative (FN): 9

Accuracy:

This model displays 68.85% accuracy.

Precision:

This model displays 72.22% precision.

Recall:

The ratio of correct positive predictions to the total positives is 74.29%.

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

### **Reporting Results**

**Chart

Description automatically generated**

Splitting the heart disease data set into training and testing sets using 80% and 20% split, there are 303 rows in the original data set, 242 rows in the training set and there are 61 rows for the testing set. The curve flattens at approximately 30 trees so the optimal number of trees for the random forest model is approximately 30.

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

*Table

Description automatically generated with medium confidence*

The root mean squared error (RMSE) for the training set is 9.6622. The RMSE for the testing set is 17.3059.

## **7. Conclusion**

Out of the two logistic regression models, I would choose the second model. While the Hosmer-Lemeshow Goodness of Fit (GOF) test for both states that the model fits with the data, the accuracy, precision, and recall percentages are a bit higher than the first one. The Receiver Operating Characteristic (ROC) Curve and the area under the curve (AUC) is slightly higher than the first one (87.77% vs 75.75%). The higher the AUC is, the better the model can distinguish the difference between the positive class and the negative class.

I wouldn’t recommend using the random forest classification model instead of the logistic regression model. I think it is useful to use both. But, in the logistic regression model, we can see and analyze more information comparing to random forest classification model. Also, according to the confusion matrix for random forest classification model, the testing data set’s percentages is lower than the second logistic regression model. I am also able to make predictions using logistic regression model and I think it is one of the most important statistics to use.

We want to be able to predict and obtain as much of accurate information as possible. It also would be beneficial to identify which variables is not as useful for the model and get rid of it (another reason why I would recommend logistic regression model). Predicting the risk factors would help others be aware of the risk factors and take action to prevent or reduce the risk of having a heart disease. If the predictions are not very accurate, then the model is not worth using.