# **MAT 303 Project Two Summary Report**

Jarrale Butts

Jarrale.butts@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The problems being analyzed will involve conducting statistical analyses to predict whether or not a person is at risk for heart disease. The data set being explored contains historical data that you can use to analyze patterns between different health indicators (e.g. fasting blood sugar, maximum heart rate, etc.) and the presence of heart disease. The data contains variables pertaining to an individual such as age, sex, cp (chest pain), trestbps (resting blood pressure), chol (cholesterol measurement), fbs (fasting blood sugar), restecg (Resting electrocardiographic measurement), thalach (max heart rate achieved), exang ( exercise included angina), oldpeak (ST depression induced by exercise relative to rest), slope (slope of the peak exercise), ca (number of major vessels), and target (heart disease). A model like this could eventually be used to evaluate medical records and look for risks that might not be obvious to human doctors. The type of analyses being deployed in this problem set will include different logistic regression models, a classification random forest model to predict the risk of heart disease, and a regression random forest model to predict the maximum heart rate achieved.

## **2. Data Preparation**

In this problem set, several important variables have been identified for analysis. The important variables in the data set include age, cp, trestbps, chol, restecg, thalach, exang, slope, ca, and target as the response variable. The data set containing historical data that you can use to analyze patterns between different health indicators and the presence of heart disease contains 14 columns and 303 rows.

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

### **Reporting Results**

The general form of the logistic multiple regression model for heart disease (target) using variables age (age), resting blood pressure (trestbps), exercised induced angina (exang), and maximum heart rate achieved (thalach) for model1 is: . The prediction equation for model1 is: .

The prediction model equation in terms of the natural log of odds for model1 is: .

In regard to an individual’s risk factor for heart disease I will address what and mean:

1. : this term, represents the probability that an individual is at risk for heart disease which would be represented with a value of 1. This is the response variable we are trying to predict or estimate.
2. : this term represents the odds that an individual is at risk for heart disease. Specifically, what are the odds that heart disease will be 1.

The logistical regression model for model1 is: glm(target ~ age + trestbps + exang + thalach). The prediction model equation (in terms of the natural log of odds) using outputs obtained from my R script for model1 is: .

For every one-unit increase in the maximum heart rate achieved, the odds (ln(odds)) of the outcome increase by a factor of approximately 0.0311. This means that as the maximum heart rate achieved variable increases, the odds of the outcome increase by approximately 3.11%, assuming all other predictors are held constant.

### **Evaluating Model Significance**

Evaluating model1 significance at a 5% level of significance for the regression model we’ll perform the Hosmer-Lemeshow goodness of fit test to assess whether the model is appropriate for the data set. The results of the Hosmer-Lemeshow goodness of fit test conclude that the model is not appropriate for the data set, and it is not significant at a 5% level of significance.

|  |  |  |  |
| --- | --- | --- | --- |
| Hosmer-Lemeshow Goodness of Fit (GOF) Test | | | |
| Null Hypothesis | Alternative Hypothesis | Test Statistic | P-value |
|  |  | t = 44.622 | p = 0.612 |

Based on Wald’s test both exercised induced angina (exang) and maximum heart rate achieved (thalach) in model1 are significant at a 5% level of significance.

|  |  |  |  |
| --- | --- | --- | --- |
| Individual Beta Test | | | |
| Variable | z-test | P-value | Conclusion |
| age | z = -1.024 | p = 0.3060 | Fail to reject |
| trestbps | z = -1.786 | p = 0.0741 | Fail to reject |
| exang1 | z = -5.314 | p < 0.001 | Reject the null |
| thalach | z = 4.274 | p < 0.001 | Reject the null |

Below are the confusion matrix results with reports for the counts including true positives, true negatives, false positives, and false negatives.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
|  | Prediction: target=0 | Prediction: target=1 |
| Actual: target=0 | 89 (True Negatives) | 49 (False Positives) |
| Actual: target=1 | 31 (False Negatives) | 134 (True Positives) |

Using data from the confusion matrix I can report the following:

* 1. Accuracy:
  2. Precision:
  3. Recall:

Below are the results of the Receiver Operating Characteristic (ROC) curve. The ROC curve is generated by plotting the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis at different classification thresholds between 0 and 1. The graph shows a very short vertical positive value before a steady curve, leveling off at around 0.5 and heading off on the x-axis. The ACU associated with this ROC is 0.8007 indicating that model1 has reasonably good discriminatory power. The higher the AUC, the better a model's ability to separate the positive and negative instances. In this case, an AUC of 0.8007 indicates that the model has a relatively good ability to discriminate between the classes, although there is still room for improvement.

**A picture containing diagram, line, text, plot

Description automatically generated**

### **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 a maximum heart rate of 140 is 0.2716, meaning this individual has a 27% chance of having heart disease. The odds of this event occurring are 1:2.

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 a maximum heart rate of 130 is 0.7853, meaning this individual has a 78% chance of having heart disease. The odds of this event occurring are 3:1.

The probability of 27% suggests a moderate likelihood of the individual having heart disease. The odds of 1:2 indicate that the event is less likely to occur compared to the event not occurring. The probability of 78% suggests a high likelihood of the individual having heart disease. The odds of 3:1 indicate a stronger association with the presence of heart disease. There is a higher probability and a more pronounced imbalance in favor of the event occurring.

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

### **Reporting Results**

The general form of the logistic multiple regression model for heart disease (target) using variables age (age), resting blood pressure (trestbps),type of chest pain experienced (cp), and maximum heart rate achieved (thalach), including the quadratic term for age and the interaction term between age and maximum heart rate achieved for model2 is: . The prediction equation for model2 is: .

The prediction model equation in terms of the natural log of odds expressing the beta terms in linear form for model2 is: .

The logistical regression model for model2 is: glm(target ~ age + trestbps + cp + thalach + I() + age: thalach). The prediction model equation (in terms of the natural log of odds) using outputs obtained from my R script for model1 is: 0019(age \* thalach).

### **Evaluating Model Significance**

Evaluating model2 significance at a 5% level of significance for the regression model we’ll perform the Hosmer-Lemeshow goodness of fit test to assess whether the model is appropriate for the data set. The results of the Hosmer-Lemeshow goodness of fit test conclude that the model is not appropriate for the data set, and it is not significant at a 5% level of significance.

|  |  |  |  |
| --- | --- | --- | --- |
| Hosmer-Lemeshow Goodness of Fit (GOF) Test | | | |
| Null Hypothesis | Alternative Hypothesis | Test Statistic | P-value |
|  |  | t = 52 | p = 0.3209 |

Based on Wald’s test all terms except age and in model2 are significant at a 5% level of significance.

|  |  |  |  |
| --- | --- | --- | --- |
| Individual Beta Test | | | |
| Variable | z-test | P-value | Conclusion |
| age | z = 0.653 | p = 0.5136 | Fail to reject |
| trestbps | z = -2.181 | p = 0.0292 | Reject the null |
| cp1 | z = 4.313 | p < 0.001 | Reject the null |
| cp2 | z = 5.867 | p < 0.001 | Reject the null |
| cp3 | z = 3.245 | p = 0.0012 | Reject the null |
| thalach | z = 2.663 | p = 0.0078 | Reject the null |
|  | z = 0.481 | p = 0.6303 | Fail to reject |
| age \* thalach | z = -2.095 | p = 0.0362 | Reject the null |

Below are the confusion matrix results for model2 with reports for the counts including true positives, true negatives, false positives, and false negatives.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
|  | Prediction: target=0 | Prediction: target=1 |
| Actual: target=0 | 102 (True Negatives) | 36 (False Positives) |
| Actual: target=1 | 36 (False Negatives) | 129 (True Positives) |

Using data from the confusion matrix I can report the following:

* 1. Accuracy:
  2. Precision:
  3. Recall:

Below are the results of the Receiver Operating Characteristic (ROC) curve. The graph shows a decent vertical positive value before a steady curve, leveling off at around 0.9 and heading off on the x-axis. The ACU associated with this ROC is 0.8478 indicating that model2 similarly to model1 has reasonably good discriminatory power. An AUC of 0.8478 indicates that this model also has a relatively good ability to discriminate between the classes, yet again there is still room for improvement.

A picture containing diagram, text, line, plot

Description automatically generated

### **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, meaning this individual has nearly a 22% chance of having heart disease. The odds of this event occurring are 1:3.

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, meaning this individual has an 80% chance of having heart disease. The odds of this event occurring are 4:1.

The probability of 22% suggests a low probability of the individual having heart disease. The odds of 1:3 indicate that the likelihood of not having heart disease is higher than the likelihood of having it. The probability of 80% suggests a high likelihood of the individual having heart disease. The odds of 4:1 indicate the probability of having heart disease is four times higher than the probability of not having heart disease.

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

### **Reporting Results**

Before splitting the heart disease data set into training and validation sets using 85% and 15% split, respectively there are 303 rows in the original data set. After splitting the credit card default data set there are 257 rows in the training set and 46 rows in the validation set.

Below I have attached a graph of the training and testing error against the number of trees using a classification random forest model for the presence of 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), and number of major vessels (ca), with a maximum of 150 trees. The optimal number of trees for the random forest model is 25 trees.

**A graph of a number of trees

Description automatically generated with low confidence**

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

Below are the confusion matrix TRAINING results with reports for the counts including true positives, true negatives, false positives, and false negatives.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
|  | Prediction: 0 | Prediction: 1 |
| Actual: 0 | 120 (True Negatives) | 0 (False Positives) |
| Actual: 1 | 0 (False Negatives) | 137 (True Positives) |

Using data from the TRAINING confusion matrix I can report the following:

* 1. Accuracy:
  2. Precision:
  3. Recall:

Below are the confusion matrix TESTING results with reports for the counts including true positives, true negatives, false positives, and false negatives.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
|  | Prediction: 0 | Prediction: 1 |
| Actual: 0 | 13 (True Negatives) | 5 (False Positives) |
| Actual: 1 | 6 (False Negatives) | 22 (True Positives) |

Using data from the TESTING confusion matrix I can report the following:

* 1. Accuracy:
  2. Precision:
  3. Recall:

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

### **Reporting Results**

Before splitting the heart disease set into training and validation sets using 80% and 20% split, respectively there are 303 rows in the original data set. After splitting the credit card default data set there are 242 rows in the training set and 61 rows in the validation set.

Attached is a graph of the mean squared error against the number of trees for a random forest regression model for maximum heart rate achieved using age (age), sex (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), with a maximum of 80 trees. The optimal number of trees for the random forest model is 15 trees.

*A picture containing text, screenshot, diagram, plot

Description automatically generated*

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

After creating a random forest regression model for maximum heart rate achieved using age (age), sex (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) using the appropriate number of trees of 15, I have obtained the RMSE for both the training and testing sets.Theroot mean squared error for the training set is 11.7971 and the root mean squared error for the testing set is 21.4423.

## **7. Conclusion**

I would choose Model 1 to predict heart disease as it has fewer predictors and significant variables, making it a more straightforward and reliable choice for predicting heart disease. Based on Wald’s test both exercised induced angina (exang) and maximum heart rate achieved (thalach) were significant at a 5% level of significance. Model1 had an accuracy of 73%, a precision of 73%, and a recall of 81%, as well as an ACU of 0.8007.

I would recommend using a logistic regression model over a random forest classification in certain situations. Logistic regression models have the advantage of providing interpretable coefficients and odds ratios, which can be valuable for understanding the impact of individual predictors on the outcome. Odds are something most people have herd of and can easily understand.

The practical importance of the analyses performed lies in their ability to provide insights and predictions related to heart disease. By conducting logistic regression and random forest classification models, valuable information can be obtained regarding the factors influencing heart disease and the accuracy of predicting its occurrence.