# **MAT 303 Project Two Summary Report**

Andrea Plunkett Jackson

drea.jackson11@snhu.edu

Southern New Hampshire University

## **1. Introduction**

As a data analyst researching risk factors for heart disease at a university hospital, I am exploring a large set of historical data. This dataset includes various health indicators such as fasting blood sugar levels, maximum heart rate, and other relevant metrics. The primary goal of this research is to analyze patterns between these health indicators and the presence of heart disease. The results of this analysis will be used to develop predictive models that can identify individuals at risk for heart disease. These models could be instrumental in evaluating medical records and identifying potential risks that might not be immediately apparent to human doctors.

To achieve this, I will be running several types of statistical analyses. Firstly, I will create logistic regression models to predict whether or not a person is at risk for heart disease based on the health indicators in the dataset. Logistic regression is particularly useful for this purpose as it can handle binary outcomes, such as the presence or absence of heart disease. Additionally, I will develop a classification random forest model to predict the risk of heart disease. Random forest models are advantageous because they can handle complex interactions between variables and provide robust predictions. Lastly, I will create a regression random forest model to predict the maximum heart rate achieved by individuals. This type of model is well-suited for continuous outcomes and can help in understanding the relationship between various health indicators and maximum heart rate. Overall, these analyses will provide valuable insights into the risk factors for heart disease and aid in the development of effective predictive tools.

## **2. Data Preparation**

In this dataset, several important variables are being analyzed to understand the risk factors for heart disease. These variables include:

* Age: This represents the person’s age in years and is a crucial factor as the risk of heart disease generally increases with age.
* Sex: This indicates the person’s sex, where 1 represents male and 0 represents female. Gender differences can influence the prevalence and type of heart disease.
* Chest Pain Type (cp): This variable categorizes the type of chest pain experienced by the person, with 0 indicating no pain, 1 for typical angina, 2 for atypical angina, and 3 for non-anginal pain. Different types of chest pain can be indicative of various heart conditions.
* Resting Blood Pressure (trestbps): This measures the person’s resting blood pressure, which is an important indicator of cardiovascular health.
* Cholesterol (chol): This represents the person’s cholesterol level in mg/dl. High cholesterol levels are a known risk factor for heart disease.
* Fasting Blood Sugar (fbs): This variable indicates whether the person’s fasting blood sugar is greater than 120 mg/dl, with 1 representing true and 0 representing false. Elevated fasting blood sugar levels can be a sign of diabetes, which is a risk factor for heart disease.
* Resting Electrocardiographic Measurement (restecg): This variable records the resting electrocardiographic measurement, with 0 indicating normal, 1 indicating ST-T wave abnormality, and 2 indicating probable or definite left ventricular hypertrophy by Estes’ criteria. Abnormal ECG readings can signal heart problems.
* Maximum Heart Rate Achieved (thalach): This measures the person’s maximum heart rate achieved during exercise, which can provide insights into cardiovascular fitness and potential heart issues.
* Exercise-Induced Angina (exang): This variable indicates whether the person experiences angina induced by exercise, with 1 representing yes and 0 representing no. Exercise-induced angina can be a symptom of underlying heart disease.
* ST Depression Induced by Exercise Relative to Rest (oldpeak): This measures the ST depression induced by exercise relative to rest, which relates to positions on the ECG plot. ST depression can indicate ischemia or reduced blood flow to the heart.
* Slope of the Peak Exercise ST Segment (slope): This variable describes the slope of the peak exercise ST segment, with 1 indicating upsloping, 2 indicating flat, and 3 indicating down sloping. The slope can provide information about the heart’s response to exercise.
* Number of Major Vessels (ca): This variable counts the number of major vessels (ranging from 0 to 3) that are colored by fluoroscopy. The number of affected vessels can indicate the severity of heart disease.
* Target: This is the outcome variable indicating the presence of heart disease, with 0 representing no heart disease and 1 representing heart disease. This variable is the primary focus of the predictive models being developed.

These variables are essential for building accurate predictive models to assess the risk of heart disease and understand the underlying factors contributing to it.

The dataset I am working with contains 303 rows and 14 columns. Each row represents an individual patient’s data, and each column corresponds to one of the important variables related to heart disease risk factors. This structure allows me to comprehensively analyze the relationships between various health indicators and the presence of heart disease.

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

### **Reporting Results**

The general form of a logistic multiple regression model for heart disease (target), using the variables age (age), resting blood pressure (trestbps), exercised induced angina (exang), and maximum heart rate achieved (thalach) as predictors, can be represented by the following equation:

In this equation:

E(Y) is the dependent variable, representing the probability of having heart disease ( target) .

is the intercept.

are the regression coefficients for age, resting blood pressure, exercise- induced angina, and maximum heart rate achieved.

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

The equation, when transformed to create a model that is linear in the beta terms, is as follows:

+

The left side represents the natural log of the odds of having heart disease.

x₃+

The prediction equation of the logistic multiple regression model, as well as expressed in terms of the natural log of odds (rounded to four decimal places) is:

Where ( not rounded)

is the natural log of the odds of having heart disease

= -1.021121

(age) = -0.017549

(trestbps) = -0.014888

(exang) = -1.624981

(thalach) = 0.031095

In terms of an individual having heart disease on the above model mean:

a. π : This term represents the probability of an individual having heart disease.

b. : Is the formula for the odds ratio, which is π divided by (1 - π). This is a measure of the odds of having heart disease versus not having it.

The estimated coefficient for the maximum heart rate achieved (thalach) is 0.031095. This means that for every one-unit increase in the maximum heart rate achieved, the log-odds of having heart disease increase by 0.031095, assuming all other variables remain constant. In simpler terms, a higher maximum heart rate achieved is associated with a higher chance of having heart disease, as indicated by the positive coefficient.

An alternative way to express this in terms of odds is:

e0.031095−1=0.0315

This indicates that for every one-unit increase in the maximum heart rate achieved, the odds of having heart disease increase by approximately 3.15%.

### **Evaluating Model Significance**

The Hosmer-Lemeshow Goodness of Fit Test was conducted to assess the model’s fit. The null hypothesis (H₀) posits that the model fits the data well, while the alternative hypothesis (Hₐ) suggests that it does not. The test yielded a test statistic (X-squared) of 44.622 with 48 degrees of freedom and a p-value of 0.612. At a 5% significance level, since the p-value is greater than 0.05, we fail to reject the null hypothesis. This indicates that there is no significant difference between the observed and expected event rates, indicating that the model fits the data well.

Based on Wald’s Test, we assessed the significance of the terms in the model using a 5% level of significance. Terms are considered significant if their p-values are less than or equal to 0.05. The results show that the terms in the model that are significant:

* excerise-induced angina (exang) : p-value = 1.07e-07

𝐻₀∶ 𝛽₃=0 𝐻ₐ∶ 𝛽₃≠0

* maxium heart rate achieved (thalach) : p-value = 1.92e-05
* 𝐻₀∶ 𝛽₄=0 𝐻ₐ∶ 𝛽₄≠0

Non-significant terms

* intercept : p-value = 0.5671

𝐻₀∶ 𝛽₀=0 𝐻ₐ∶ 𝛽₀≠0

* age : p-value = 0.3060

𝐻₀∶ 𝛽₁=0 𝐻ₐ∶ 𝛽₁≠0

* resting blood pressure (trestbps) : p-value = 0.0741

𝐻₀∶ 𝛽₂=0 𝐻ₐ∶ 𝛽₂≠0

Based on Wald’s test and using a 5% level of significance, the significant terms in the model are exercise-induced angina (exang) and maximum heart rate achieved (thalach). These terms have P-values less than 0.05, indicating that they are statistically significant predictors of heart disease in the model.

The output of the confusion matrix is as follows:

|  |  |  |
| --- | --- | --- |
|  | **Prediction : default = 0** | **Prediction: default = 1** |
| **Actual: default = 0** | 89 | 49 |
| **Actual: default = 1** | 31 | 134 |

From this matrix, I can identify and report the following counts:

* True Positives (TP) : 134
* True Negatives (TN) : 89
* False Positives (FP): 49
* False Negatives (FN): 31

Accuracy:

Accuracy = ≈ 0.735 or 73.5%

Precision: Precision =

Precision =

Precision = ≈ 0.732 or 73.2%

Recall : Recall =

Recall =

Recall =

The graph provided below is a Receiver Operating Characteristic (ROC) curve. This curve is used to evaluate the performance of a binary classification model by plotting the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) at various threshold settings. The curve starts at the origin (0,0) and moves towards the top right corner (1,1). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model is.

The value of the Area Under the Curve (AUC) is 0.8007. The AUC value represents the probability that the model will correctly distinguish between a randomly chosen positive instance and a randomly chosen negative instance. An AUC value of 0.8007 indicates that the model has good discriminative ability. It means that there is an 80.07% chance that the model will correctly classify a randomly chosen positive instance as having heart disease over a randomly chosen negative instance.

**A graph with a line

Description automatically generated with medium confidence**

### **Making Predictions Using Model**

Using the regression model, I can make predictions about the probability of an individual having heart disease.

**Prediction 1**

For an individual who is 50 years old, has a resting blood pressure of 122, has exercise induced angina (exang = yes(1)), and has maximum heart rate of 140 the probability of having heart disease is 0.372. This means the probability of having heart disease is approximately 27.16%, with odds of about 0.372. This indicates a lower likelihood of having heart disease. The odds of this event occurring are calculated as the probability of having divided by the probability of not having, which ≈ 0.372 or 0.372 to 1.

**Prediction 2**

For an individual who is 50 years old, has a resting blood pressure of 130, does not have exercise induced angina (exang = no(0)), and has maximum heart rate of 165 the probability of having heart disease is 3.65. This means the probability of having heart disease is approximately 78.53%, with odds of about 3.65. This indicates a higher likelihood of having heart disease. The odds of this event occurring are calculated as the probability of having divided by the probability of not having, which ≈ 3.65 or 3.65 to 1.

In summary, the second individual, who does not have exercise-induced angina but has higher resting blood pressure and maximum heart rate, shows a significantly higher probability and odds of having heart disease compared to the first individual.

Exercise-induced angina appears to be a strong predictor of heart disease, as evidenced by the higher probability and odds in the second individual who lacks this condition. These predictions highlight that factors such as exercise-induced angina, resting blood pressure, and maximum heart rate achieved are crucial in determining the likelihood of heart disease. The model indicates that individuals with higher resting blood pressure and maximum heart rate, even without exercise-induced angina, are at a higher risk of heart disease.

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

### **Reporting Results**

The general form of a logistic multiple regression model for heart disease (target), using the variables age (age), resting blood pressure (trestbps), type of chest pain experienced (cp), and maximum heart rate achieved (thalach) as predictors, including the quadratic term for age and the interaction terms between age and maximum heart rate achieved , can be represented by the following equation:

Where:

represents age

represents resting blood pressure,

represents type of chest pain experienced,

represents maximum heart rate achieved.

The prediction equation can be written as:

+

The equation, when transformed to create a model that is linear in the beta terms, is as follows:

The left side represents the natural log of the odds of having heart disease.

The prediction model equation expressed in terms of the natural log of odds (rounded to four decimal places) from outputs obtained from R script is:

Where:

represents age,

represents resting blood pressure,

represents typical angina (cp1),

represents atypical angina (cp2),

represents non-anginal pain (cp3),

represents maximum heart rate achieved.

### **Evaluating Model Significance**

The Hosmer-Lemeshow Goodness of Fit Test was conducted to assess the model’s fit. The null hypothesis (H₀) posits that the model fits the data well, while the alternative hypothesis (Hₐ) suggests that it does not. The test yielded a test statistic (X-squared) of 52 with 48 degrees of freedom and a p-value of 0.3209. At a 5% significance level, since the p-value is greater than 0.05, we fail to reject the null hypothesis. This indicates that there is no significant difference between the observed and expected event rates, indicating that the model fits the data well.

Based on Wald’s Test, we assessed the significance of the terms in the model using a 5% level of significance. Terms are considered significant if their p-values are less than or equal to 0.05. The results show that the terms in the model that are significant:

* resting blood pressure (trestbps) : p-value = 0.02916

𝐻₀∶ 𝛽₃=0 𝐻ₐ∶ 𝛽₃≠0

* typical angina (cp1) : p-value = 1.61e-05

𝐻₀∶ 𝛽₄=0 𝐻ₐ∶ 𝛽₄≠0

* atypical angina (cp2) : p-value =4.45e-09

𝐻₀∶ 𝛽₅=0 𝐻ₐ∶ 𝛽₅≠0

* non-anginal pain (cp3) : p-value = 0.00117

𝐻₀∶ 𝛽₆=0 𝐻ₐ∶ 𝛽₆≠0

* maxium heart rate achieved : p-value = 0.00775

𝐻₀∶ 𝛽₇=0 𝐻ₐ∶ 𝛽₇≠0

* age:thalach (age: maxium heart rate achieved): p-value = 0.03616

𝐻₀∶ 𝛽₈=0 𝐻ₐ∶ 𝛽₈≠0

Non-significant terms

* intercept : p-value = 0. 0.13988

𝐻₀∶ 𝛽₀=0 𝐻ₐ∶ 𝛽₀≠0

* age : p-value = 0.51357

𝐻₀∶ 𝛽₁=0 𝐻ₐ∶ 𝛽₁≠0

* age ² : p-value = 0.63025

𝐻₀∶ 𝛽₂=0 𝐻ₐ∶ 𝛽₂≠0

The output of the confusion matrix is as follows:

|  |  |  |
| --- | --- | --- |
|  | **Prediction : default = 0** | **Prediction: default = 1** |
| **Actual: default = 0** | 102 | 36 |
| **Actual: default = 1** | 36 | 129 |

From this matrix, I can identify and report the following counts:

* True Positives (TP) : 129
* True Negatives (TN) : 102
* False Positives (FP): 36
* False Negatives (FN): 36

Accuracy:

Accuracy = ≈ 0.7624 or 76.24%

Precision: Precision =

Precision =

Precision = ≈ 0.7818 or 78.18%

Recall : Recall =

Recall =

Recall =

The ROC curve below illustrates the trade-off between sensitivity and specificity, with a point closer to the top-left corner of the plot indicating better model performance. This curve aids in selecting the optimal threshold for classification by balancing the true positive rate and the false positive rate. The Area Under the Curve (AUC) value, which is 0.8478, measures the classifier’s ability to distinguish between classes. An AUC of 0.8478 means there is an approximately 84.78% chance that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. A higher AUC value indicates better model performance, and in this case, an AUC of 0.8478 suggests that the model has good classification performance.

A graph of a function

Description automatically generated

### **Making Predictions Using Model**

Using the regression model, I can make predictions about the probability of an individual having heart disease.

**Prediction 1**

For an individual who is 50 years old, has a resting blood pressure of 115, does not experience chest pain (cp=0), and has maximum heart rate of 133 the probability of having heart disease is 0.2188. This means the probability of having heart disease is approximately 21.88%, with odds of about 0.2800. This indicates a relatively low likelihood of having heart disease. The odds of this event occurring are calculated as the probability of having divided by the probability of not having, which ≈ 0.2800 or 0.2800 to 1.

**Prediction 2**

For an individual who is 50 years old, has a resting blood pressure of 125, experiences typical angina (cp=2), and has maximum heart rate of 155 the probability of having heart disease is 4.546. This means the probability of having heart disease is approximately 81.97%, with odds of about 4.5465. This indicates a high likelihood of having heart disease. The odds of this event occurring are calculated as the probability of having divided by the probability of not having, which ≈ 4.546 or 4.546 to 1.

In summary, the second individual, who experiences typical angina and has higher resting blood pressure and maximum heart rate, has a significantly greater probability and odds of having heart disease compared to the first individual, who does not experience chest pain and has lower resting blood pressure and maximum heart rate. This indicates that the type of chest pain, resting blood pressure, and maximum heart rate are crucial factors in predicting the likelihood of heart disease.

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

### **Reporting Results**

* Using set.seed(6522048) and split the heart disease data set into training and validation sets using 85% and 15% split, respectively gives 303 rows are in the original data set, 257 rows in the training set, and 46 in the testing set.
* The graph below is 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.

A graph of a training set

Description automatically generated

* Based on the graph, the optimal number of trees for the random forest model is 50 trees. This is the point where the classification error for the testing set stabilizes and does not significantly decrease with the addition of more trees. This ensures a balance between computational efficiency and predictive performance.

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

Using the optimal number of trees for the random forest model found above the output of the confusion matrix for the training set is as follows:

|  |  |  |
| --- | --- | --- |
|  | **Prediction : default = 0** | **Prediction: default = 1** |
| **Actual: default = 0** | 117 | 3 |
| **Actual: default = 1** | 0 | 137 |

Accuracy:

Accuracy = ≈ 0.9883 or 98.83%

Precision: Precision =

Precision =

Precision = ≈ 0.9786 or 97.86%

Recall : Recall =

Recall =

Recall =

Using the optimal number of trees for the random forest model found above the output of the confusion matrix for the testing set is as follows:

|  |  |  |
| --- | --- | --- |
|  | **Prediction : default = 0** | **Prediction: default = 1** |
| **Actual: default = 0** | 11 | 7 |
| **Actual: default = 1** | 9 | 19 |

Accuracy:

Accuracy = ≈ 0.6522 or 65.22%

Precision: Precision =

Precision =

Precision = ≈ 0.7308 or 73.08%

Recall : Recall =

Recall =

Recall =

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

### **Reporting Results**

* Using set.seed(6522048) and split the heart disease data set into training and validation sets using 80% and 20% split, respectively gives 303 rows are in the original data set, 242 rows in the training set, and 61 in the testing set.
* The graph below is of the mean squared error against the number of trees for a random forest regression model for maximum heart rate achieved using the same 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 80 trees.

A graph of a number of trees

Description automatically generated

* Based on the graph above, the optimal number of trees for the random forest model is 40 trees. This is the point where the root mean squared error for the testing set stabilizes and does not significantly decrease with the addition of more trees. This ensures a balance between computational efficiency and predictive performance.

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

Using the optimal number of trees found above of 40 trees, I can create 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). The Root Mean Squared Error (RMSE) for the training set is 11.5934, while the RMSE for the testing set is 20.4339. These values represent the average deviation of the predicted values from the actual values in the respective sets. The lower RMSE for the training set indicates that the model fits the training data well. In contrast, the higher RMSE for the testing set suggests how well the model generalizes to unseen data. This difference highlights the model’s performance and its ability to predict new, unseen instances accurately.

## **7. Conclusion**

After analyzing the two logistic regression models for predicting heart disease, I would choose the second model. This model includes more variables, enhancing its predictive power. The Area Under the Curve (AUC) for the second model is 0.8478, indicating a strong ability to distinguish between individuals with and without heart disease. Additionally, the Hosmer-Lemeshow goodness of fit test for the second model shows a p-value of 0.3209, indicating that the model fits the data well. Significant terms identified through Wald’s test, such as resting blood pressure, types of chest pain, maximum heart rate achieved, and the interaction between age and maximum heart rate achieved, further support the model’s robustness.

In summary, the second logistic regression model is a better predictor of an individual’s risk of heart disease due to its comprehensive variable inclusion, higher AUC, and superior performance metrics. This model provides a more accurate and reliable assessment of heart disease risk, making it the preferred choice for prediction.

Choosing between the random forest classification model and the logistic regression model depends on several factors, including the advantages and disadvantages of each. Given the performance metrics and the context of predicting heart disease, I recommend using the logistic regression model. The second logistic regression model showed strong performance with high accuracy, precision, and recall, and it is easier to interpret. This makes it a reliable and practical choice for predicting heart disease, especially when interpretability and understanding the relationship between variables are important.

The analyses performed in this study have significant practical importance, particularly in the field of healthcare and medical diagnostics. The logistic regression and random forest models developed provide valuable tools for predicting the likelihood of heart disease in individuals. Accurate prediction models can help healthcare professionals identify high-risk patients early, allowing for timely intervention and treatment. By understanding the significant factors contributing to heart disease, such as resting blood pressure, types of chest pain, and maximum heart rate achieved, healthcare providers can make more informed decisions about patient care, leading to personalized treatment plans tailored to individual risk profiles.

Accurate prediction models can also help healthcare systems allocate resources more efficiently. By identifying high-risk patients, healthcare providers can prioritize preventive measures and allocate medical resources to those who need them most, potentially reducing the overall burden on the healthcare system.

In conclusion, the analyses performed in this study have practical implications for improving patient care, enhancing decision-making, optimizing resource allocation, educating patients, advancing research, and informing public health policies. These contributions can ultimately lead to better health outcomes and a more efficient healthcare system.

## **8. Citations**

zyBooks. (2024). Zybooks.com. <https://learn.zybooks.com/zybook/MAT-303-16699.202456-1>

Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R. (1988). *Heart disease data set* [Data file and codebook]. Retrieved from <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

*‌*