# **MAT 303 Project Two Summary Report**

Daniel Gorelkin

Daniel.Gorelkin@SNHU.edu

Southern New Hampshire University

## **1. Introduction**

*This paper will analyze a Heart Disease data set, which will be read from the “heart\_disease.csv” file. This file contains 303 historical records used to research the risk factors for heart disease in relation to patients’ different health indicators and the presence of heart disease. The purpose of this report is to predict whether a person is at risk for heart disease using various logistic regression models as well as random forest classification and regression models. This report will run several Logistic Regression models, including the tests of Wald confidence intervals and the Hosmer-Lemeshow goodness-of-fit analysis, and run several random forest models to evaluate the potential existence of heart disease which will assist in evaluating medical records and pinpointing risks and factors that could lead to heart disease that might be obscured to human doctors.*

## **2. Data Preparation**

*The given dataset is considered sufficiently large and contains 303 randomly selected records, categorized into 14 variable rubrics, including the person's age in years, fasting blood sugar, resting blood pressure, maximum heart rate, cholesterol measurement, the presence of heart disease, and more. This report will focus on several important variables for creating logistic regression models, such as age (age), resting blood pressure (trestbps), exercised induced angina (exang) as binary variable, maximum heart rate achieved (thalach), and type of chest pain experienced (cp) as categorical variable to predict the potential risks of a heart disease as target binary outcome result as well as split the dataset into training and testing portions in order to train and test the random forest models.*

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

### **Reporting Results**

*The* *general form for 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) is:*

*The term π from the general equation in terms of an individual having a heart disease can be interpreted as the probability or the chance of an individual to have a heart disease, where 1-π will be identified as the probability not to have a heart disease and the term will represent the log-odds of heart disease, meaning the odds of such a heart disease outcome occurrence happening.*

*Hence, the prediction equation is:*

*Where* ***P()*** *is the probability of having heart disease, and the coefficients* *For i = 1, 2, 3, these represent age, resting blood pressure, and maximum heart rate for the quantitative variables respectively, and* *serves as a coefficient for the exercise-induced angina binary variable, with a value of 1 indicating exercise-induced angina and 0 indicating none.*

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

*By running the first logistic regression model (model\_1), the prediction model equation in terms of the natural log of odds can be written as:*

*From observing the estimated coefficient of the maximum heart rate () variable, we can conclude that due to the positive coefficient sign as the values of heart rate increase, so increases the probability of having heart disease while assuming that all the other coefficients are constant (remain unchanged), hence for every one-unit increase in maximum heart rate achieved, the odds of having heart disease increase by 0.0311 units or by approximately 3.1%.*

### **Evaluating Model Significance**

*To evaluate the significance of this logistic regression model, we will perform the Hosmer-Lemeshow goodness-of-fit test to assess whether the model is appropriate for the dataset for a 5% level of significance and set our null and the alternative hypothesis as follows:*

*The* ***Null Hypothesis*** *(H0​): The model provides a good fit to the data, with age, the person's resting blood pressure, the person's maximum heart rate achieved, and exercise-induced angina serving as effective predictors for an individual to develop heart disease.*

*The* ***Alternative******Hypothesis*** *(Ha): The logistic regression model does not provide a good fit to the data.*

*A black text on a white background

AI-generated content may be incorrect.*

*From the performed Hosmer-Lemeshow test, we can identify the Chi-Square values calculated from the differences between observed and expected frequencies to be with the P-value of 0.612.*

*From this, we can conclude that since our P-value is much greater than 0.05 (0.612>0.05), insufficient evidence exists to reject the null hypothesis, H0, in favor of the alternative hypotheses, and, therefore, the model is significant and provides a good fit to the data.*

To evaluate w*hich terms are significant in the model, we can run Wald’s test at the 5% level of significance to determine whether each coefficient is significant for the model. Thus:*

*The* ***Null Hypothesis*** *(H0​): , for i=1, 2, 3, 4. The coefficients for age, resting blood pressure, maximum heart rate, and exercise-induced angina are zero and considered not significant for the mode.*

*The* ***Alternative******Hypothesis*** *(Ha​): , for i=1, 2, 3, 4. The coefficients for age, resting blood pressure, maximum heart rate, and exercise-induced angina are not zero and are considered significant for the model.*

*A screenshot of a computer

AI-generated content may be incorrect.*

*By running the Hosmer-Lemeshow test, we can see that because the confidence intervals for (thalach), (exang)* ***does not*** *contain zero. Hence, we reject the null hypothesis (H₀) at the 5% significance level in favor of the alternative hypothesis (Ha) and conclude that all variables are statistically significant for the model.*

*From the confusion (error) matrix, we can evaluate the performance of the logistic regression model*

*and observe the counts of the accurate and false outcomes:*

A black text on a white background

AI-generated content may be incorrect.

*From this confusion matrix, we can see that the true positive(****TP****) counts, i.e., the actual and predicted values, are 136 accuracies, true negatives(****TN****) are 89, false positives(****FP****) are 49, and the false negatives(****FN****) count is 31. From these results, we can calculate the accuracy of the model*

A number with black lines

AI-generated content may be incorrect.

That shows us that despite the model being significant, its accuracy, precision, and the ratio of correct positive predictions to the total positive examples aren’t perfect and can be improved to guarantee more accurate results.

*The Receiver Operating Characteristic (ROC) curve enables us to visualize and evaluate the performance of binary classifiers at various thresholds, thereby assessing the model’s overall performance.*

*A graph of a function

AI-generated content may be incorrect.*

*From the ROC curve demonstrated above, we can identify the False Positive rate as 1-specificity and the True Positive rate as sensitivity. The curve line that climbs gradually from the origin (0, 0) towards the top-right corner indicates a good model with reasonable prediction ability, however, the model is not perfect. Hence, the model distinguishes between Y=0 (no heart disease) and Y=1(heart disease) well. The Area Under the Curve (AUC) value of 0.8007 (80%) represents a high but not perfect predictive ability for the potential development of heart disease.*

### **Making Predictions Using the 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:*

**

*The probability can be calculated by the prediction equation as:*

*Where the odds of having heart disease are*

*Where the probability of an individual having heart disease who is 50 years old, has a resting blood pressure of 130, does not have an exercise-induced angina, and has a maximum heart rate of 165 is*

**

*The probability can be calculated by the prediction equation as:*

*Where the odds of having heart disease are*

*Based on these predictions, we can deduce that a patient who has exercise-induced angina has a significantly decreased chance of having heart disease. Therefore, for a random patient with such parameters as in the first example, his probability of having a heart disease is as low as 27.2%, which means he most likely will not have a heart disease as the outcome is less than 1, the ratio can be explained as for every three person that will not develop a heart disease, one person will.*

*The second prediction, on the other hand, shows a way higher likelihood to develop heart disease due to a higher maximum heart rate while not having exercise-induced angina. For such an outcome, the likelihood of developing heart disease is as high as 78.5%, and for every single person, approximately three persons will develop heart disease.*

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

### **Reporting Results**

*Report the results of the regression model. Address the following questions in your analysis. Round all numbers to four decimal places*.

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

*The general form:*

*The prediction equation:*

*Where* ***P()*** *is the probability of having heart disease, and the coefficients* *For i = 1, 2, 3, these represent age, resting blood pressure, and maximum heart rate for the quantitative variables respectively,* *serves as a coefficient for the interaction term, is**the coefficient for age², and* *are the coefficients for the type of chest pain experienced for x6-x8 values as 1 indicates the type of pain and 0 indicates no pain.*

*The prediction equation for model\_2 in terms of the natural log of odds is:*

*Hence, the logistic regression equation in terms of the natural log of odds for model\_2 is:*

### **Evaluating Model Significance**

*To evaluate the significance of this model, we will perform the Hosmer-Lemeshow goodness-of-fit test again to assess whether the model is appropriate for the dataset for a 5% level of significance and set our null and the alternative hypothesis as follows:*

*The* ***Null Hypothesis*** *(H0​): The model provides a good fit to the data and is effective in predicting the probability for an individual to develop heart disease.*

*The* ***Alternative******Hypothesis*** *(Ha): The logistic regression model does not provide a good fit to the data.*

*A close-up of text

AI-generated content may be incorrect.*

*From the performed Hosmer-Lemeshow test, we can identify the Chi-Square values calculated from the differences between observed and expected frequencies to be with a P-value of 0.3209.*

*From this, we can conclude that since our P-value is much greater than 0.05 (0.3209>0.05), insufficient evidence exists to reject the null hypothesis, H0, in favor of the alternative hypotheses, and, therefore, the model is significant and provides a good fit to the data.*

To validate that all *terms are significant in the model, we can run Wald’s test at the 5% level of significance and determine whether each coefficient is significant for the model. Thus:*

*The* ***Null Hypothesis*** *(H0​): , for i=1 to 8. The coefficients for age, resting blood pressure, maximum heart rate, and exercise-induced angina are zero and considered not significant for the mode.*

*The* ***Alternative******Hypothesis*** *(Ha​): , for i=1 to 8. The coefficients for age, resting blood pressure, maximum heart rate, the interaction for age and heart rate, age squared, and the type of pain are not zero and are considered significant for the model.*

*A screenshot of a table with numbers

AI-generated content may be incorrect.*

*By running the Hosmer-Lemeshow test, we can see that because the confidence intervals for (thalach), (), ), (), and* ***,*** *(the type of pain)* ***does not*** *contain zero. Hence, we reject the null hypothesis (H₀) at the 5% significance level in favor of the alternative hypothesis (Ha) and conclude that all variables are statistically significant for the model.*

*From this confusion matrix, we can see that the true positive(****TP****) counts, i.e., the actual and predicted values, are 129 accuracies, true negatives(****TN****) are 102, false positives(****FP****) are 36, and the false negatives(****FN****) count is 36. From these results, we can calculate the accuracy of the model as follows*

A number and percentages on a white background

AI-generated content may be incorrect.

This indicates that, despite the model being significant, its accuracy, precision, and the ratio of correct positive predictions to the total positive examples are improved compared to the first model, and yet, there is still room for improvement to guarantee even more accurate results.

*The Receiver Operating Characteristic (ROC) curve allows us to visualize and evaluate the performance of binary classifiers at various thresholds and assess the model’s overall performance.*

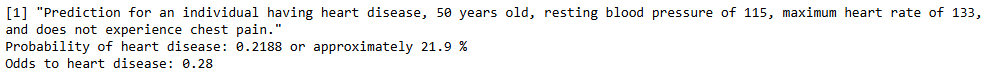
*A graph with a line

AI-generated content may be incorrect.*

*From the ROC curve demonstrated above, we can identify the False Positive rate as 1-specificity and the True Positive rate as sensitivity. The curve line that climbs gradually from the origin (0, 0) towards the top-right corner indicates a good model with reasonable prediction ability, however, the model is not perfect. Hence, the model distinguishes between Y=0 (no heart disease) and Y=1(heart disease) well. The Area Under the Curve (AUC) value of 0.8478, e.g., 85%, represents a high but not perfect predictive ability for the potential development of heart disease and performs slightly better compared to the first model.*

### **Making Predictions Using the Model**

*To test the model, we can make several predictions using the regression model, such as the*

**

*The probability of developing heart disease can be calculated by using our prediction equation:*

*The odds for such an event occurring could be calculated as*

*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 are:*

**

*And it can be calculated by using our prediction equation:*

*The odds for such an event occurring could be calculated as*

*From these predictions, we can deduce that experiencing typical angina pain can significantly ramp up the probability of developing heart disease, and a person declaring typical angina-type pain is very likely to develop heart disease as such event occurring with a ratio of four to one persons. On the other hand, a person with a lower resting blood pressure and higher maximum heart rate of the same examined age who identifies no pain at all is not likely to develop heart disease, and the odds of it happening are once in every three persons. To conclude, higher resting blood pressure and higher maximum heart rate slightly increase the chances of developing heart disease, while factors like experiencing pain can ramp up the chances of a disease.*

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

### **Reporting Results**

*To construct the random forest classification model, we will split the heart disease data set, which contains 303 rows of data, into training and testing sets using an 85% and 15% split ratio, respectively. Hence, the model will be trained on 257 rows of data and tested using 46 new, unseen data rows to ensure genuine testing.*

**

*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, sex, chest pain type, resting blood pressure, cholesterol measurement, resting electrocardiographic measurement, exercise-induced angina, and number of major vessels (ca):*

*A graph of a number of trees

AI-generated content may be incorrect.*

*From the output, the classification error curve flattens at* ***approximately 50 trees****. Beyond this, the error remains roughly the same, even as the number of trees increases up to 150. Therefore, 50 trees are the optimal choice for this model.*

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

*To evaluate the utility of the random forest classification model, we will be using the optimal number of trees found, 50, and re-create the random forest classification model for the variables mentioned above and obtain the confusion matrix for the training set as:*

*A screenshot of a graph

AI-generated content may be incorrect.A black text on a white background

AI-generated content may be incorrect.A number of numbers and percentages

AI-generated content may be incorrect.*

*And the confusion matrix for the testing set as:*

**A screenshot of a graph

AI-generated content may be incorrect.A black text on a white background

AI-generated content may be incorrect.A number with black text

AI-generated content may be incorrect.**

*From the output above, we can see that our classification model is accurately predictive around 70% of the time, despite its training statistics showing almost perfect predictive ability.*

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

### **Reporting Results**

*To construct the random forest regression model, we will split the heart disease data set, which contains 303 rows of data, into training and testing sets using an 80% and 20% split ratio, respectively. Hence, the model will be trained on 242 rows of data and tested using 61 new, unseen data rows to ensure genuine testing.*

**

*The graph the mean squared error against the number of trees for a random forest regression model for maximum heart rate achieved using age, sex, chest pain type, resting blood pressure, cholesterol measurement, resting electrocardiographic measurement, exercise-induced angina, and number of major vessels:*

*A graph of a number of trees

AI-generated content may be incorrect.*

*From the output, the regression error curve flattens at* ***approximately 20 trees****. Beyond this, the error is roughly the same, even if the number of trees increases up to 80 trees. Therefore, 20 trees are the optimal choice for this model.*

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

*By creating a random forest regression model for maximum heart rate achieved using 20 trees and the eight predictors, we can identify the root mean squared error for the training set to be 11.6282 and the root mean squared error for the testing set to be 21.1139 heartbeats.*

*A black text on a white background

AI-generated content may be incorrect.*

## **7. Conclusion**

*By comparing the two logistic regression models, we can conclude that the second model, “model\_2”, performs slightly better in predicting the potential of developing a heart disease. Since the P-value for each test is larger than the 5% significance level and all the examined variables are significant, both models are considered candidates. However, from the confusion matrices of the two, despite the first model having a slightly higher recall value, i.e., it catches slightly more true heart disease cases, the second model has a 3% accuracy difference and a 5% precision difference, which means it identifying people with and without heart disease better. The better performance of the second model is also evident in its higher Area Under the Curve (AUC) value, which provides approximately 5% higher case coverage. Additionally, the second model performs its analyses based on more predictors and interactions between the terms, which can lead to identifying cases that the first model may overlook or cannot analyze at all. Therefore, despite its complexity, the second model should be chosen.  
Comparing the random forest classification model to the logistic regression 2nd model can be done by observing their accuracy, precision, and recall values as demonstrated in the following table:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Logistic regression model\_2* | *R.F. classification model training (50 trees)* | *R.F. classification model testing (50 trees)* | *Conclusion (better performance)* |
| *Accuracy* | *76.2%* | *98.8%* | *65.2%* | *model\_2* |
| *Precision* | *78.2%* | *97.9%* | *73.1%* | *model\_2* |
| *Recall* | *78.2%* | *100%* | *67.9%* | *model\_2* |

*Comparing the performance of the 2nd model to the classification testing model, we can see a better prediction ability with the 2nd logistic model despite the very high training performance of the classification random forest model. This significant difference between the training and the testing outcomes for the random forest model can signal that the model suffers from overfitting or improper tuning and setup. Therefore, under the defined setup requirements, the performance of the random forest classification model is poor compared to the second logistic model. For this reason, we should remain loyal to the second logistic regression model, as its insights and performance are more trustworthy.*

*The practical importance of the analyses performed enables the prediction of which patients are more likely to develop heart disease and which factors may have a greater impact on its development. Thus, factors such as age, blood pressure, cholesterol levels, and the existence and type of chest pain can act as a triggering alarm and potentially save human lives if identified on time. The information from the analyses can serve as a tool for doctors and healthcare providers or even be implemented as an in-app application feature on any smart device, providing real-time recommendations and guidance for individuals at risk.*

## **8. References**

*ZyBooks. (n.d.).* [*https://learn.zybooks.com/zybook/MAT-303-12932.202516-1/chapter/5/section/1*](https://learn.zybooks.com/zybook/MAT-303-12932.202516-1/chapter/5/section/1)

*ZyBooks*. (n.d.). <https://learn.zybooks.com/zybook/MAT-303-12932.202516-1/chapter/6/section/4>

*ZyBooks*. (n.d.). <https://learn.zybooks.com/zybook/MAT-303-12932.202516-1/chapter/7/section/1>