# **MAT 303 Project Two Summary Report**

Tammy Hartline

tammy.hartline@snhu.edu

Week 7

Southern New Hampshire University

## **1. Introduction**

*In this project, we aim to address the problem of predicting heart disease using statistical analyses. We have employed logistic regression models and a random forest classification model for this purpose. The dataset under exploration, heart\_disease.csv, consists of various health-related variables such as age, resting blood pressure, chest pain type, cholesterol measurements, and more. The results of our analyses will be used to predict the presence of heart disease based on these variables. We will conduct logistic regression analyses, assess model significance, evaluate model performance using metrics like accuracy, precision, and recall, while also analyzing the Receiver Operating Characteristic (ROC) curve to gauge the discriminative power of our models.*

## **2. Data Preparation**

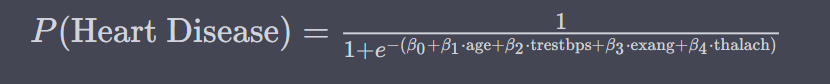
*The important variables in the dataset include target, age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), maximum heart-rate (thalach), and the number of major vessels (ca). The heart\_disease dataset we used contains 303 rows and 14 columns.*

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

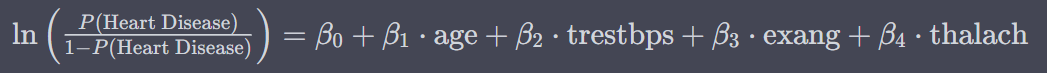
### **Reporting Results**

*The logistic regression model prediction equation examples can be seen below:*

*General Form:*



*Natural Log of Odds / Linear Form:*

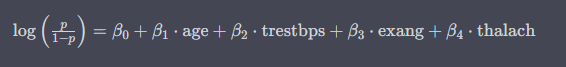
**

*The term* ***pi*** *in this scenario, is the predicted probability of someone having heart disease, based off of our logistic regression model. Where, for example, if ‘pi’ is close to 1 (i.e. 0.95 or 95%), then this suggests there is a high probability that the person, where ‘pi’ = 0.80, and the assumption that the person has the characteristics set, then there is an 80% approximated probability that the person does have heart disease.*

*The term* ***1-pi*** *in this scenario, represents the complementary probability, which is the opposite outcome, where the person is predicted to not have heart disease. Given ‘pi’ still = 80%, and then* ***1-pi****, then the 1-pi represents 0.25, or a 25% probability of not having heart disease with the same set of characteristics.*

.

*The logistic regression model equation for heart disease is:*

**

*Where p is the probability of having heart-disease is true when the target equals one, or the probability of heart-disease is false when the target is equal to zero.*

*Beta-0, the intercept, was approximately -1.0211.*

*Beta-1, the coefficient for age, was approximately -0.0175.*

*Beta-2, the coefficient for resting blood pressure (trestbps), was approximately -0.0149.*

*Beta-3, the coefficient for exercise-induced-angina (exang), was approximately -1.6250.*

*Lastly, Beta-4, the coefficient for maximum heart rate achieved (thalach), was approximately 0.0311.*

*Each of the values above are calculated estimates received as output from my logistic regression model.*

*A screenshot of a computer

Description automatically generated*

*To interpret the thalach, (maximum heart rate achieved), I needed to understand the coefficients representation value of 0.0311. I know when the target = 1, then the patient’s probability of having heart disease is true. Also, because the natural log of odds can be challenging for me to interpret directly, I broke it down a bit as follows:*

*For each 1-unit increase in thalach, the odds of the person having heart disease increases by a factor of e^0.0311, which is approximately 1.0314.*

*Breaking it down this way allowed me to determine that an increase in max heart rate achieved, is associated with a 3.14% increase in the odds of the person having heart disease, holding other variables constant.*

*Therefore, the results suggest a positive association between the coefficient for thalach with likelihood of the person having heart disease. In laymen terms, a higher thalach is associated with a slightly higher likelihood of heart disease, when other factors remain constant.*

### **Evaluating Model Significance**

*Our initial model summary, using predictor variables age, trestbps, exang, and thalach, with our response variable, target, we had a minimum deviance variable of 2.0935, and maximum deviance variable of 2.2343. Lastly, the model’s significance p-value is less than 0.001.*

*Next, I ran a Hosmer and Lemeshow Goodness of Fit Test, where the results of that fit test are as follows:*

*The alternate and null hypotheses are as follows:*

*x-squared = 44.622,*

*degree of freedom (df) = 48,*

*and p-value = 0.612.*

*Using x-squared of 44.622, we get the p-value of 0.612, which is higher than the 0.05 (5%) significance level. This means with the given data, we cannot reject the null hypothesis, and can conclude that the model is valid for the given data.*

*From here, we perform Wald’s test, to determine if each coefficient is significant.*

*A table of numbers and symbols

Description automatically generated*

*Based on the results shown above, the p-value associated with the intercept, with a 95% confidence interval between -4.5181 and 2.4758, is 0.5671, indicating the intercept is not significantly different from zero, nor statistically significant at the 0.05 significance level.*

*The coefficient for the age variable is estimated to be -0.0175, with a 95% confidence interval between -0.0512 and 0.0161. The p-value for age is 0.3060, which is greater than the 0.05 significance level, also suggesting that age is not statistically significant.*

*The coefficient for trestbps is estimated to be -0.0149, with a 95% confidence interval between -0.0312 and 0.0015. The p-value is 0.0741, which is close to, but slightly greater than 0.05. This suggests trestbps is borderline and may or may not be significant.*

*The coefficient for exang with a value of 1 (true), is estimated to be -1.6250 with a 95% confidence interval between -2.2243 and -1.0257. The p-value is very small, at (1.07e-07), indicating that it is highly significant. This means that the presence of exang (being true), has a significant impact on predicting heart disease.*

*The coefficient for thalach is estimated to be 0.0311, with a 95% confidence interval between 0.0168 and 0.0454. The p-value of for thalach is very small, at (1.92e-05), indicating that it is also highly significant. This suggests that thalach is a significant predictor of heart disease.*

*The Wald tests overall, indicate the presence of exang (where true), and thalach are significant predictors of heart disease in this logistic regression model, where age, trestbps, and the intercept for this model are not.*

*After that, we calculated the Area Under the Curve, (AUC), which was equal to 0.8007.*

*A graph with a line

Description automatically generated with medium confidence*

### **Making Predictions Using Model**

*For an individual who is 50 years old, trestbps of 122, exang is True/1, and thalach is 140, the logistic regression model predicted an approximate probability of 0.2716(27.16%) of having heart disease. The odds of this event occurring are approximately 0.2716 / (1 – 0.2716) = 0.3732.*

*For the next prediction, we have an individual who is also 50 years old, but has a trestbps of 130, exang is False/0, and thalach is 165. The logistic model predicts a higher probability of 0.7853 (78.53%), for a person with these characteristics having heart disease. The odds of this event occurring are approximately 0.7853 / (1 – 0.7853) = 3.7193.*

*A graph with a line

Description automatically generated with medium confidence*

*These predictions suggest that the second prediction, whose age is the same, but has a higher resting bp (trestbps), and a higher maximum heart rate (thalach), and whose symptoms do not include exercise-induced-angina, is associated with a significantly higher probability and odds of having heart disease when compared to the characteristics of the first person. These results indicate that people with higher blood pressure and maximum heart rate, may have a higher likelihood of developing heart disease.*

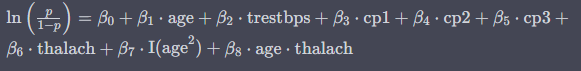
## **4. Model #2 - Second Logistic Regression Model**

### **Reporting Results**

***General Form:***

*E(y) = 1 / (1 + e^-(β0 + β1 \* age + β2 \* trestbps + β3 \* cp + β4 \* thalach + β5 \* age^2 + β6 \* age \* thalach))*

***Prediction Equation in terms of the Natural Log of Odds:***



*Where:*

*Beta0 is the y-intercept = -15.5600,*

*Beta1 is the coefficient for age = 0.1744,*

*Beta2 is the coefficient for resting blood pressure (trestbps) = -0.0196,*

*Beta3 is the coefficient for type of chest pain experienced (cp) where cp = cp1 = 1.9130,*

*Beta4 is the coefficient for type of chest pain experienced (cp) where cp = cp2 = 2.0370,*

*Beta5 is the coefficient for type of chest pain experienced (cp) where cp = cp3 = 1.7770,*

*Beta6 is the coefficient for maximum heart rate achieved (thalach) = 0.1363,*

*Beta7 is the quadratic term for age (age^2) = 0.0008,*

*and Beta8 is the interaction term between age and thalach, (age \* thalach) = -0.0019.*

*A screenshot of a computer

Description automatically generated*

### **Evaluating Model Significance**

*The results of the Hosmer-Lemeshow Goodness of Fit Test had a test statistic of 52 with 48 degrees of freedom. The p-value associated is 0.3209, which is greater than the 0.05 significance level. Therefore, we fail to reject the null hypothesis, as this suggests the logistic regression model is a good is a good fit for the data.*

*Next, I will interpret Wald’s Test for Significance of Coefficients. For the 'Age' variable, the Wald's test assesses whether the coefficient β1 (0.1744) is significantly different from zero. The z-value (0.653) is relatively small, and the associated p-value (0.5136) is greater than 0.05 (assuming a 5% significance level). Therefore, we do not have enough evidence to conclude that 'Age' is a significant predictor of heart disease in this model.*

*For 'Resting Blood Pressure' (trestbps), the test checks whether β2 (-0.0196) is significantly different from zero. The z-value (-2.181) is relatively large, and the p-value (0.0292) is less than 0.05. Thus, we can conclude that 'Resting Blood Pressure' is a significant predictor of heart disease.*

*For 'Type of Chest Pain' (cp), there are three separate coefficients (β3, β4, and β5) for each type of chest pain. All three have p-values less than 0.05, indicating that each type of chest pain is a significant predictor in its own right.*

*For 'Maximum Heart Rate Achieved' (thalach), the test examines whether β6 (0.1363) is significantly different from zero. The z-value (2.663) is relatively large, and the p-value (0.0078) is less than 0.05. Therefore, 'Maximum Heart Rate Achieved' is a significant predictor of heart disease.*

*For 'Quadratic Term for Age' (age^2) and the 'Interaction Term between Age and Maximum Heart Rate' (age \* thalach), neither coefficient is significantly different from zero as their p-values are greater than 0.05.*

*In summary, based on Wald's test results, 'Resting Blood Pressure,' 'Type of Chest Pain,' and 'Maximum Heart Rate Achieved' are significant predictors of heart disease in this logistic regression model, while 'Age,' 'Quadratic Term for Age,' and the 'Interaction Term between Age and Maximum Heart Rate' are not significant.*

*The confusion matrix shows there are 129 true positives, 102 true negatives, 36 false positives, and 36 false negatives. We needed these values to calculate the model’s accuracy, precision, recall (sensitivity).*

*Accuracy: (TP + TN) / (TP + TN + FP + FN) = (129 + 36) / (129 + 102 + 36 + 36) = 0.7576*

*Precision: (TP / (TP + FP) = 129 / (129 + 36) = 0.7812*

*Recall: (TP / (TP + FN) = 129 / (129 + 36) = 0.7812*

*The receiver operating characteristic (ROC) curve illustrates the trade-off between TP rate sensitivity and FP rate as the classification threshold changes. It shows the model’s ability to discriminate between individuals with and without heart disease.*

*A graph with a line

Description automatically generated with medium confidence*

*The AUC is approximately 0.8478, which is a measure of the model’s ability to distinguish between positive and negative cases. An AUC value close to 1 indicates good model performance, and in this case, it suggests that the model has reasonable discriminative power.*

*Overall, the logistic regression model appears to be a good fit for the data, as indicated by the Hosmer-Lemeshow test. It demonstrates reasonable discriminatory power, as shown by the AUC.*

### **Making Predictions Using Model**

*Next, we performed predictions, where the first is a person 50 years old, trestbps of 115, cp was false/0, and thalach of 133. This model with a person that has these characteristics, has a heart disease predicted probability of approximately 0.2188. (0.2188 predicted probability for having heart disease.)*

*The next prediction was a person who is also age 50, but their trestbps was 125, cp was true/1, and their thalach was 155. A person with these characteristics, have an approximate 0.8007 predicted probability of having heart disease.*

*Based on the model’s evaluation metrics and the predictions mentioned above, the model can predict the probability of heart disease for individuals based on their specific characteristics, and the evaluation metrics indicate its performance in identifying positive cases. It is worth stating, that further validation and refinement will be needed for clinical use.*

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

### **Reporting Results**

*The original dataset contains 308 rows. After splitting the data into training and validation sets, using an 85%:15% split, and setting the seed to 6522048, the training set contains approximately 257 rows, and the validation set contains approximately 46 rows.*

*Next, we trained a classification random forest model using variables age, sex, chest pain type (cp), trestbps, cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exang, and the number of major vessels (ca).*

*The next step was to evaluate the model’s performance by varying the number of trees from 1 to a maximum of 150. For each number of trees, we calculated the training error and the testing (validation) error against the number of trees.*

*Next, we were tasked with finding the optimal number of trees to use, without overfitting the model.*

*A graph of a number of trees

Description automatically generated*

*The plot above shows that the optimal number of trees is approximately 25, which I determined by examining the point at which the testing error stabilizes or starts to increase, which indicates overfitting. By visually examining and inspecting the graph, we can identify the number of trees that minimizes the testing error and, therefore, provides the best balance between bias and variance.*

*To summarize, I split the data into training and validation sets, which now consists of 257 training rows and 46 validation set rows. I then evaluated the random forest model’s performance with varying numbers of trees and determined the optimal number of trees based on the testing error curve. This analysis allows us to better understand how the model’s performance changes with the number of trees and select the most suitable number of trees for our random forest classification model. In this specific scenario, I interpreted the optimal tree count to be 25.*

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

*Upon creation of a confusion matrix, we get the following results:*

*A screenshot of a computer

Description automatically generated*

*Based on the matrices above, we are able to calculate the following values for the* ***training set****:*

*True Positives (TP) = 127,*

*False Positives (FP) = 1,*

*True Negatives (TN) = 113,*

*False Negatives (FN) = 1*

*Now that we have those* ***training set*** *values, we can calculate the accuracy, precision, and recall (sensitivity):*

*Accuracy = (TP + TN) / (TP + TN + FP + FN) = (127 + 113) / (127 + 113 + 1 + 1) = 240/242 = 0.9917*

*Precision = (TP) / (TP + FP) = 127 / (127 + 1) = 0.9922*

*Recall = TP / (TP + FN) = 127 / (127 + 1) = 0.9922*

*Next, we repeat the process for the* ***validation (testing) set:***

*TP = 30*

*FP = 11*

*TN = 13*

*FN = 7*

*Lastly, we can now calculate the accuracy, precision, and recall for the* ***validation (testing) set:***

*Accuracy = (TP + TN) / (TP + TN + FP + FN) = (30 + 13) / (30 + 13 + 11 + 7) = 43/61 = 0.7049*

*Precision = (TP) / (TP + FP) = 30 / (30 + 11) = 0.7317*

*Recall = TP / (TP + FN) = 30 / (30 + 7) = 0.8108*

*With these values, we can tell that in the case of the training set, the model achieved high accuracy, precision, and recall. In contrast to that, the testing set had slightly lower accuracy, precision, and recall values. This suggests the model’s performance may be slightly reduced when applied to new, unseen data, but it still demonstrates reasonable predictive capabilities.*

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

### **Reporting Results**

*I first started by splitting the original 303 rows into two sets once again, but this time into an 80%:20% split, resulting in 242 rows in the training set and 61 rows in the validation (testing) set.*

*A graph of a number of trees

Description automatically generated*

*Based on observation of the plot, and random testing by increments of 5 trees, I determined 25 to be the optimal number of trees for this set as well.*

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

*With the optimal number of trees (25), I then constructed a random forest regression model to predict thalach based on age, sex, cp, trestbps, chol, restecg, exang, and ca.*

*The RMSE for the training set is 11.6002, and for the testing set is 20.8735. These RMSE values indicate the model’s accuracy in predicting in predicting thalach with lower values suggesting better predictive performance on the training set and higher values on the testing set.*

*A close-up of a white background

Description automatically generated*

## **7. Conclusion**

*Based on our analyses, we have gained valuable insights into predicting heart disease using various models. The logistic regression models provided information about the significance of individual variables in predicting heart disease. The first logistic regression model indicated that exercise-induced angina (exang) and maximum heart rate achieved (thalach) are significant predictors of heart disease, while age and resting blood pressure (trestbps) are not. The second logistic regression model further supported these findings, highlighting the importance of factors like chest pain type (cp) and maximum heart rate (thalach).*

*The random forest classification model demonstrated strong predictive performance, with high accuracy, precision, and recall values on the training set. This suggests that the random forest model can effectively identify individuals at risk of heart disease based on a combination of demographic and health-related variables.*

*On the other hand, the random forest regression model aimed to predict maximum heart rate achieved. The optimal number of trees was determined to be 25, and the model exhibited a certain level of accuracy in predicting this cardiovascular metric.*

*Based on my findings, I would recommend using the first logistic regression model, as it showed that exercise-induced angina (exang) and maximum heart rate achieved (thalach) are significant predictors of heart disease. These variables provide valuable insights into a patient's likelihood of developing heart disease.*

*I would recommend considering both the logistic regression model and the random forest classification model based on the specific needs of the prediction task. The logistic regression model offers interpretability and insights into the importance of individual variables, which can be valuable for understanding the underlying factors contributing to heart disease. On the other hand, the random forest classification model demonstrates strong predictive performance, making it suitable for accurately identifying individuals at risk of heart disease.*

*The analyses conducted in this project hold practical importance in the field of healthcare. By leveraging statistical models like logistic regression and random forest classification, we can assist healthcare professionals in identifying individuals at higher risk of heart disease. This early risk assessment can lead to timely interventions and preventive measures, potentially improving patient outcomes and reducing the burden of heart disease on healthcare systems. Additionally, these analyses can serve as a foundation for further research and the development of more sophisticated predictive models in the field of cardiology.*