# **Project Summary Report**

Gregory Dionisio

## **1. Introduction**

*The data observed in this analysis is the heart disease data set. The results will be used to predict the probability of developing heart disease based on a variety of factors. To do this ill be running two logistic regression models, a random forest classification model and a random forest regression model.*

## **2. Data Preparation**

*The following variables were used in the analysis: age, sex, chest pain (CP), resting blood pressure (BP), cholesterol level, ECG pattern during rest and exercise (RPE), exercise-induced angina diagnosis (EIA), slope of peak exercise, major vessel density (Ca) and target heart rate. There are 14 columns and 303 rows in the data set.*

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

### **Reporting Results**

The second order regression model general form is . (Berrier et al., 2016) The prediction regression model equation is . (Berrier et al., 2016) After running a first order regression model we get the formula:

*The prediction model equation in terms of the natural log of odds to express the beta terms in linear form*

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

*is the probability of the event happening, in this case the event is developing heart disease.*

*is the odds of developing heart disease.*

*The equation for this regression model is:*

*The equation for this model in terms of natural log of odds is:*

*The estimated coefficient of the maximum heart rate achieves variable is . This means on average, the change in log odds of developing heart disease is , holding all other variables constant.*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model predictions are close to the observed values of Y, which are either 0 or 1. In this model it is used to assess if the model fits the data or not.*

*The null and alternative hypotheses are:*

*The test statistic ( is 44.622, The P-value is 0.612. The level of significance is 5%.*

*The P-value of 0.612 is higher than the level of significance of 0.05. Thus, the null hypothesis should not be rejected. The conclusion is that the model is appropriate for the data set.*

In our analysis of individual variables, the null hypothesis is that “no relationship exists between the response variable Y and the predictor variable” (Berrier et al., 2016) and the alternate hypothesis is “that a relationship does exist” (Berrier et al., 2016). For the variable age the pvalue is 0.3060, for trestbps 0.0741, for exang 1.07e-07, for thalac 1.92e-05. P values under 0.05 will be considered significant whereas P values above this amount we can conclude they don’t have a significant effect on heart disease. Therefore, the significant variables are exang1 and thalac. The insignificant variables are age and trestbps.

*The general form table output of a confusion matrix is:*

*Table

Description automatically generated*

*The confusion matrix for this model is:*

Table

Description automatically generated

*The confusion matrix results are:*

* *True positives: 134*
* *True negatives: 89*
* *False positives: 49*
* *False negatives: 31*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

*The Receiver Operating Characteristic (ROC) curve is displayed here:*

Chart, line chart

Description automatically generated

*The area under the curve (AUC) is 0.8007 or 80.07%. This is an indicator of how well the model distinguishes between Y = 0 and Y = 1. In general, the larger the AUC the better, because the larger the area under the curve, the better it is at predicting binary classes.*

### **Making Predictions Using Model**

If we assumed that the person is *50 years old, has a resting blood pressure of 122, has exercise induced angina, and has maximum heart rate of 140* then we can predict that the probability of having a heart attack is 27.16%. *The odds are calculated by dividing the probability by 1 minus the probability. In this case . The odds in this case are 0.2716 to 0.7284 or 27.16% to 72.84%.*

If we assumed that the person is *50 years old, , has a resting blood pressure of 130, does not have an exercise induced angina, and has maximum heart rate of 165* then we can predict that the probability of having a heart attack is 78.53%. *The odds are calculated by dividing the probability by 1 minus the probability. In this case . The odds in this case are 0.7853 to 0.2147or 78.53% to 21.47%.*

*From these tests we conclude that the people in the second prediction are more likely to get heart disease than the first prediction.*

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

### **Reporting Results**

The second order regression model general form is . (Berrier et al., 2016) The prediction regression model equation is . (Berrier et al., 2016) After running a first order regression model we get the formula:

*The prediction model equation in terms of the natural log of odds to express the beta terms in linear form*

.

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

.

*is the probability of the event happening, in this case the event is developing heart disease.*

*is the odds of developing heart disease.*

*The equation for this regression model is:*

*The equation for this model in terms of natural log of odds is:*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model predictions are close to the observed values of Y, which are either 0 or 1. In this model it is used to assess if the model fits the data or not.*

*The null and alternative hypotheses are:*

*The test statistic ( is 52, The P-value is 0.3209. The level of significance is 5%.*

*The P-value of 0.3209­ is higher than the level of significance of 0.05. Thus, the null hypothesis should not be rejected. The conclusion is that the model is appropriate for the data set.*

In our analysis of individual variables, the null hypothesis is that “no relationship exists between the response variable Y and the predictor variable” (Berrier et al., 2016) and the alternate hypothesis is “that a relationship does exist” (Berrier et al., 2016). For the variable age the pvalue is 0.5136, for trestbps 0.0292, for cp1 1.61e-05, for cp2 4.45e-09, for cp3 0.00117, for thalach 0.0078, for age squared 0.6303, and for ae with thalach 0.0362. P values under 0.05 will be considered significant whereas P values above this amount we can conclude they don’t have a significant effect on heart diesase. Therefore, the significant variables are trestbps,cp1,cp2,cp3,thalach, age and thalach. The insignificant variables are age, age squared.

*­­*

*The general form table output of a confusion matrix is:*

*Table

Description automatically generated*

*The confusion matrix for this model is:*

Table

Description automatically generated

*The confusion matrix results are:*

* *True positives: 129*
* *True negatives: 102*
* *False positives: 36*
* *False negatives: 36*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

*The Receiver Operating Characteristic (ROC) curve is displayed here:*

Chart

Description automatically generated

*The area under the curve (AUC) is 0.8478 or 84.78%. This is an indicator of how well the model distinguishes between Y = 0 and Y = 1. In general, the larger the AUC the better, because the larger the area under the curve, the better it is at predicting binary classes.*

### **Making Predictions Using Model**

If we assumed that the person is *50 years old, has a resting blood pressure of 115, does not experience chest pain, and has maximum heart rate of 133* then we can predict that the probability of having a heart attack is 21.88%. *The odds are calculated by dividing the probability by 1 minus the probability. In this case . The odds in this case are 0.2188 to 0.7812 or 21.88% to 78.12%.*

If we assumed that the person is *50 years old, , has a resting blood pressure of 130, does not have an exercise induced angina, and has maximum heart rate of 165* then we can predict that the probability of having a heart attack is 78.53%. *The odds are calculated by dividing the probability by 1 minus the probability. In this case . The odds in this case are 0.8007 to 0.1993 or 80.07% to 19.93%.*

*From these tests we conclude that the people in the second prediction are more likely to get heart disease than the first prediction.*

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

### **Reporting Results**

*Splitting the heart disease data set into training and testing sets using 85% and 15% split, respectively. Using set.seed(**6522048) gives us 257 row in the training set and 46 rows in the testing set.*

*Graphing 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). Use a maximum of 150 trees. Using set.seed(**6522048) gives us:*

A picture containing chart

Description automatically generated

*The optimum number of trees is where the red curve remains unchanged. This optimal number of trees for this model is 20.*

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

*Using the appropriate number of trees found, create 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), slope of peak exercise (slope), and number of major vessels (ca). The confusion matrix for the testing set is:*

*The general form table output of a confusion matrix is:*

Table

Description automatically generated

*The confusion matrix results for the training set are:*

* *True positives: 137*
* *True negatives: 120*
* *False positives: 0*
* *False negatives: 0*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

*The confusion matrix for this model’s testing set is:*

Table

Description automatically generated

*The confusion matrix results for the testing set are:*

* *True positives: 22*
* *True negatives: 13*
* *False positives: 5*
* *False negatives: 6*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

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

### **Reporting Results**

*Splitting the heart disease data set into training and testing sets using 80% and 20% split, respectively. Using set.seed(6522048). We receive 242 rows in the training set and 61 rows in the testing set.*

*Graphing 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). Using a maximum of 80 trees and using set.seed(6522048) we get the graph:*

Chart, line chart

Description automatically generated

*The optimal number of trees for this forest is 5 because it is at this point when the RMSE shows no change.*

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

*Using the appropriate number of trees found, 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 for the training set is 12.7008.*

*The root mean squared for the testing set is 22.4877.*

## **7. Conclusion**

*We developed two logistic regression models analyzed and after the analyses the second model is better. It includes more variables and contains a larger area under the curve, suggesting is is better at predicting binary values. As well the accuracy, precision and recall for the confusion matrix are is greater than the second one.*

*The accuracy, precision and recall that the random forest gives us is superior to the logistic regression model thus I would recommend using these in further research.*

*This data is important to determine if it may be appropriate to treat an individual a certain way with their specific variables. This will help doctors predict how to treat patients with their unique data.*

## **8. Citations**

Berrier, J., Pardoe, lain, & Watts, K. (2016). Applied Statistics Ii for Science. Zybooks, a Wiley brand. https://learn.zybooks.com/zybook/MAT-303-H3016-OL-TRAD-UG.23EW3