# **MAT 303 Project Two Summary Report**

Catherine Taft

Catherine.taft@snhu.edu

Southern New Hampshire University

## **1. Introduction**

Due to my cats’ insatiable appetite for the most premium, top-shelf, high-quality organic (but not vegetarian – I’m not a monster) cat food and their refusal of all else to the point of hunger strike, I’ve been forced to take a second job. In addition to risk analysis at the credit card company I am now researching risk factors for heart disease for a hospital. We’ll be looking at whether or not someone is likely to have heart disease, as well as assessing someone’s maximum heart rate based on a number of factors. First we’ll be doing some logistic regression models and then we’ll be doing some random forest models.

## **2. Data Preparation**

For the logistic regression models, we’ll have heart disease as the response variable (i.e. do you have it?) using age, maximum heart rate, resting blood pressure, gender, presence of angina, and type of chest pain, as well as an interaction term in the second order model. For the random forest model we’ll be looking at maximum heart rate as a function of age, gender, type of chest pain, resting blood pressure, cholesterol level, resting electrocardiograph measurement, presence of angina, slope of peak exercise, and number of major vessels. This is very outside of my wheelhouse but I can’t have my cats eating Friskies so we’ll forge ahead.

There are 14 columns and 303 rows in this dataset.

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

### **Reporting Results**

The general form of a model for heart disease using age, resting blood pressure and maximum heart rate is:

Where is age, is resting blood pressure and is maximum heart rate.

To rewrite this in linear terms we can write it in the form of the natural log of odds:

In the above equation, is the probability of having heart disease, and represents the odds of developing heart disease.

To write the model for this scenario we need to run it in R first:

Text

Description automatically generated

So based on the above, the model for our scenario here is:

And rewritten for intelligibility and linearity in the form of natural log of odds:

The estimated coefficient for maximum heart rate is 0.042967. That means that the expected increase in log odds of having heart disease is 0.042697 per unit change in the that variable.

### **Evaluating Model Significance**

Now to evaluate the model significance we’ll do the Hosmer-Lemeshow goodness of fit test. The null hypothesis is that the model fits the data, and the alternative hypothesis is that the model does not fit the data.

Text

Description automatically generated

Here we can see that the test-statistic (X-squared) is 41.978, and the p-value is 0.7168. That is above the significance level of 0.05, so we can’t reject the null hypothesis, and can say that the model does fit the data.

Now to evaluate the significance of each variable we’ll do the Wald’s test. The null hypothesis is that the coefficient is 0 and therefore the term is not significant to the model. The alternative is that it is not 0 and therefore is significant.

Based on the above output, the p-values for age, blood pressure and max heartrate are respectively 0.5578, 0.0392, and 8.06 \* 10-10. Based on this and a significance level of 0.05 we can say that the terms for blood pressure and max heart rate are relevant to the model, and age is not.

The confusion matrix is as follows:

Table

Description automatically generated

There are 83 true negatives, 38 false negatives, 55 false positives and 127 true positives.

**Accuracy**

The equation for accuracy is:

This evaluates to 0.6931 or roughly 69%.

**Precision**

The equation for precision is:

This evaluates to 0.6978 or roughly 70%.

**Recall**

The equation for recall is:

This evaluates to 0.7670 or roughly 77%.

These scores seem fairly mediocre.

Below are the AUC and ROC curve:

Chart, line chart

Description automatically generated

A picture containing text

Description automatically generated

With this graph, the more of the area under the curve (represented by the AUC value), the better the model is at classifying. Here we have a value of 0.7575 and a somewhat flattened line on the graph, so we can conclude that this model is not as great as some of the others we’ve seen previously, as also indicated by the accuracy et al. scores above.

### **Making Predictions Using Model**

Now we’re going to use the model to evaluate two scenarios. First, the probability of a 50 year old with a resting blood pressure of 122 and a max heart rate of 140. This results in 0.4939 or a 49.39% chance of having heart disease.

The second one is the probability of a 50 year old with resting blood pressure of 140 and a maximum heart rate of 170. The probability of that is 0.7248 or 72.48%.

None of that is surprising – the second scenario increased both risk factors, the second one substantially so, and so the probability of having heart disease also increased substantially in the second outcome.

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

### **Reporting Results**

Now we’re going to do a multiple regression model for heart disease using variables for maximum heart rate, age, sex, presence of angina, and type of chest pain.

The general form is:

Where the terms are, in order, : max heart rate, : age, : sex, angina, followed by the dummy terms for the 3 types of chest pain, the quadratic term for age and the interaction term for max heart rate against age.

This equation in terms of the natural log of odds is:

In order to write the equation for this scenario we need to run the model in R:

Text, table

Description automatically generated with medium confidence

I tried filling in the equation above but it literally doesn’t fit because it’s too long and I could not how to format it in a way that is readable but still fits here but the below log of odds equation is what would have been included as the exponent of e.

And in terms of the natural log of odds:

### **Evaluating Model Significance**

To evaluate the significance of the model we’ll do the Hosmer Lemeshow goodness of fit test. The null hypothesis is that the model fits the data, and the alternative hypothesis is that the model does not fit the data.

Text

Description automatically generated

The test statistic is 60.596 and the p-value here is 0.1048, so at a significance level of 0.05, we cannot reject the null hypothesis and can say that the model does fit the data.

To determine which terms are significant we’ll do the Wald’s test at the same significance level.

The null hypothesis is that the coefficient is 0, thereby making the term insignificant to the model, and the alternative hypothesis is that it is not 0, making it significant.

The p-values for each term are: heart rate: 0.014760, age: 0.510325, sex: 1.91 \* 10-6, angina: 0.009133, chest pain type = typical angina (cp1): 0.000249, chest pain type = atypical angina (cp2): 2.21 \* 10-6, chest pain type = non-anginal pain(cp3): 0.003684, age2: 0.810599, and interaction term for max heart rate:age is 0.043666*.*

Of these, all are statistically significant (we’d reject the null hypothesis) EXCEPT for age and age2 (a little bit surprising I admit).

Here’s the confusion matrix:

Table

Description automatically generated

This gives us 103 true negatives, 27 false negatives, 35 false positives and 138 true positives.

**Accuracy**

The accuracy for this model is:

This evaluates to 0.7954 or roughly 80%.

**Precision**

The precision for this model is:

This evaluates to 0.7977 or roughly 80%.

**Recall**

The recall for this model is:

This evaluates to 0.8364 or roughly 84%.

These are significantly better than the previous model.

Here are the ROC curve and the AUC value:

Chart

Description automatically generated

Text

Description automatically generated with low confidence

Here we have a plot with a curve that’s a lot less flattened, and an AUC value of 0.8777. The higher the AUC value (and the more space that’s literally under the curve in the plot) then the more effective the model is at classifying. So here from both of these things we can see that this model is more effective at classifying.

### **Making Predictions Using Model**

Now we’re going to do two predictions. The first is the probability of heart disease for a 30 year old male with a max heart rate of 145, has exercise-induced angina and does not have (I assume chronic) chest pain of any of the specified categories. That gives us a probability of 0.2654 or 25.54%.

The next scenario is for a 30 year old male with a max heart rate of 145 who does not experience exercise-induced angina but does experience typical angina. That has a probability of 0.8432 or 84.32%.

Apparently it makes a really big difference what type of angina you’re experiencing. I suppose the difference may be in having incidental chest pain with a specific trigger versus a chronic condition that could indicate a more serious underlying issue. But obviously with all other variables holding, if you have typical angina versus only exercise-induced, you are far more likely to have heart disease. Either way you should probably go to a doctor.

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

### **Reporting Results**

The original dataset has 303 rows, and upon splitting the model 80/20 into training and validation sets, we have 242 rows in the training set and 61 rows in the test set.

Here’s the graph of the training and testing error against the number of trees for a model using variables for age, sex, chest pain type, resting blood pressure, cholesterol, resting ECG, exercise-induced angina, slope of peak exercise, and number of major vessels and using a max of 200 trees.

I’d estimate the optimal number of trees (decided based on where the curve sort of flattens out) to be between 15 and 20, maybe 17.

Graphical user interface

Description automatically generated

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

Now we’ll use that number to create a model for the above-listed variables, and create a confusion matrix for the training and testing sets:

Text

Description automatically generated with low confidence

**Accuracy**

The accuracy for the training set is:

This evaluates to 0.993 or 99.3%

**Precision**

The precision is:

This evaluates to 0.9924 or 99.24%.

**Recall**

This evaluates to 1, which is 100%.... I’m actually not sure that’s a good thing? I feel like this implies overfitting.

Now for the testing set:

Text

Description automatically generated

**Accuracy**

The accuracy is:

This evaluates to 0.7377 or 73.77%.

**Precision**

The precision is:

This evaluates to 0.7568 or 75.68%.

**Recall**

The recall is:

This evaluates to 0.8 or 80%.

Not too shabby but not magnificent either.

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

### **Reporting Results**

Now we’re going to do a random forest regression model. We already split this for the previous model and I’m assuming we can just re-use that split, so that gave us 303 rows originally, 242 in the training set and 61 in the validation set.

Graphing the mean squared error against the number of trees (max 80) for a model with the same variables we’ve been working with gives us this graph:

Chart, line chart

Description automatically generated

It looks like it sort of flattens out around 9 or 10 trees (I’m not great at this eyeball estimation) and I like round numbers so I’ll go with 10.

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

Using that value of 10 we’ll do a random forest regression model for max heart rate using age, sex, chest pain type, resting blood pressure, cholesterol level, resting ECG, exercise-induced angina, slope of peak exercise and number of major vessels.

The root mean squared error for the training set is 10.6817 and for the testing set it’s 18.0886. I admit I don’t really know how to put that in context because I thought it was supposed to be a value between 0 and 1. I may have made a mistake. If this is the thing that dooms my cats to plebeian cat food I’ll never forgive myself. Fortunately their English is limited so they’ll never know.

## **7. Conclusion**

Of the two logistic regression models I’d definitely choose the second one. The AUC was much higher and the accuracy, precision and recall were all significantly higher. Not super great but definitely better.

In terms of recommending logistic regression or the random forest classification, I’d still probably recommend the logistic regression. I would be inclined initially to choose a classification model because you’re looking at a binary response variable (Heart disease: do you have it?), but between these models, the second logistic regression model was pretty effective, while the RF classification model had potential overfitting in the training set (if I was interpreting that correctly), and the accuracy, precision and recall of the validation set was pretty mediocre.

The practical importance of these of course is that you can predict whether someone is likely to have heart disease due to a variety of risk factors, and you can easily adjust the parameters of the predictor variables to see how that impacts someone’s likely outcomes.