# MAT 303 Project 2 Report

Brandon Hobbs

Brandon.Hobbs@snhu.edu

Southern New Hampshire University

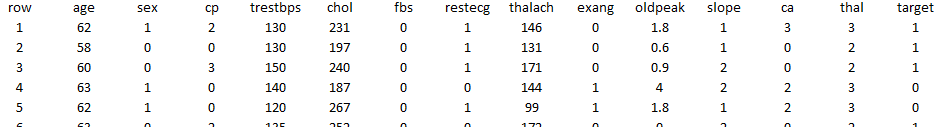
## Introduction

The data set being analyzed consists of 303 rows and 14 columns. Each row contains data about a particular individual’s heart health and their risk factors, e.g., *sex*, *age*, *chol*, etc. See Figure 1 for the first few rows for data.

The data will be used to build two logistic regression models with the purpose of predicting an individual’s likelihood to develop heat disease (*target*) from the other available data.

After the logistic models two CART analyses will be created. The first tree will be a random classification tree and the second will be a random regression tree.

First, the data in the csv-file will be ingested into a data frame so the R-language may be used for the stated purpose. Next, it will be plotted to provide a sense of the data and then the regression models, and their appropriateness, will be calculated. Finally, the models will be used to make predictions.



**Figure 1: First 5 Rows of Data used for Analysis**

## Data Preparation

To begin the analysis the data, all 303 rows and 14 columns, were imported into a data frame for consumption in the R-language. Of particular interest are the *target*, *sex, education,* *cp*, trestbps, *thalach*, *restecg*, *exang*, and *age*. Two regression models and two random trees will be created.

The first will try and predict the probability of heart disease (target) from *age, trestbps, exang, thalach.*

The second model will try and predict the probability of heart disease (target) from *age, trestbps, cp, thalach, age* squared,andan interactive term between *age* and *thalach.*

The subsequent CART analyses will build a random classification tree with a maximum of 150 tress and a random regression tree with a maximum of 80 trees.

The data for the CART analysis will be segregated into training and test sets. The trees will be built upon the larger training sets. The testing data will be fed back into these trees to calculate their ability to predict outcomes.

## First Logistic Regression Model with *age + trestbps + exang + thalach*

The first will try and predict the probability of heart disease (target) from *age, trestbps, exang, thalach.*

### Reporting Results

### The first step in building the model is to relevel the qualitative predictor, *exang*. In the data there are two possible values so only one variable is needed to represent the data – a 1 indicates that a quality exists. Table 1 shows these values:

**Table 1: Qualitative Predictors of *exang***

|  |  |
| --- | --- |
|  | ***exang*** |
| **Exercise-induced angina present** | 1 |
| **Exercise-induced angina absent** | 0 |

This first model will use four predictor variables as the modelling input for heart disease. This model will be of the form:

Where y is ‘1’ for heart disease predicted and ‘0’ for not predicted.

The final model is calculated as:

With *age* as X1, *trestbps* as X2, *exang* as X3,and *thalach* as X4.

This may be linearized into the log-odds form as:

In the above equation π represents the probability (*p*) and the odds being equal to .

The model suggests that the odds of heart disease will increase 3.2% for each unit of maximum heart rate achieved (thalach), i.e., , if all other variables are held constant.

Based on this model and the assumption that if the probability of heart disease is over 50% a person will acquire the affliction the following confusion matrix, Table 2, may be created:

**Table 2: Confusion Matrix for Model using** *age + trestbps + exang + thalach*

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

Which gives rise to the following common measures, Table 3, to help evaluate the model. The closer these metrics are to 1 the better the model is able to predict a true positive:

**Table 3: Common Measures of Logistic Models**

|  |  |
| --- | --- |
| **Accuracy** | 0.7360 |
| **Precision** | 0.7322 |
| **Recall** | 0.8121 |

To further determine if the model was relevant a Hosmer-Lemeshow Goodness of Fit (GOF) Test is conducted. A GOF is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: the model fits the data*

*Ha: the does not fit the data*

The null hypothesis states that there is a correlation between *age, trestbps, exang, thalach,* and *target*. The alternative states there is no correlation between *age, trestbps, exang, thalach,* and *target*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 4 shows the test statistic and its associated P-value:

**Table 4: Results for the Hosmer-Lemeshow Goodness of Fit (GOF) Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 44.622 |
| P-value | 0.612 |

The P-value confirms that there is not enough evidence to reject the null hypothesis, 0.612 >> 0.05; thus, the model does fit the data. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

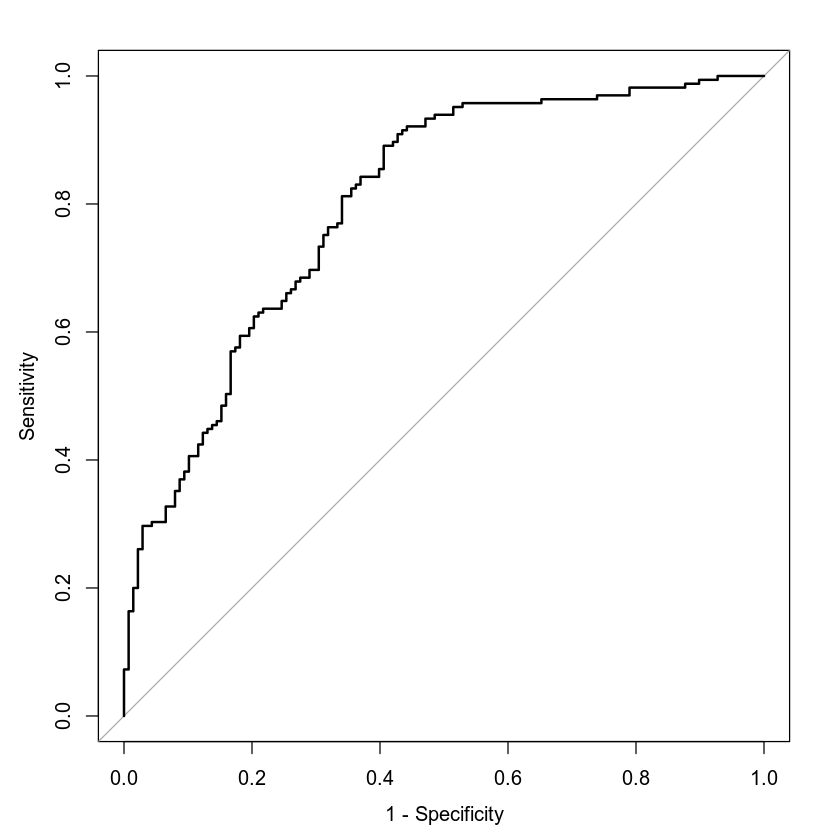
What the GOF test does not reveal is which of the predictor variables are relevant. To determine which are relevant a Wald Confidence Interval is conducted on each variable. The confidence intervals (95% confidence) can be used to determine statistical relevance, see Table 5.

**Table 5: Wald 95% Confidence Intervals**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Lower Limit** | **Upper Limit** |
| *(Intercept)* | -4.5181 | 2.4758 |
| *age* | -0.0512 | 0.0161 |
| *trestbps* | -0.0312 | 0.0015 |
| *exang1* | -2.2243 | -1.0257 |
| *thalach* | 0.0168 | 0.0454 |

Any variables’ interval that contains zero can be deemed statistically irrelevant. The intercept value, *age*, and *trestbps* could be deemed as statistically irrelevant as they all contain zero within their respective ranges.

So far, the model has shown to be underperforming based on the GOF test and the majority of terms being irrelevant from the Wald confidence intervals. One more test is to evaluate the model’s discrimination sensitivity at various thresholds. The Receiver Operating Characteristic (ROC) Curve is a plot of the true-positive rate (sensitivity) against the false-positive rate (1 - specificity) – a large area under the curve (AUC) is desired with 1 being the “perfect classifier”.



**Figure 2: ROC Plot for** *age + trestbps + exang + thalach*

The ROC curve shown in Figure 2 has an area under the curve of 0.8007. This moderately large value (~1) suggests some relevance of the generated model.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, two persons are considered for credit, Person A and Person B. Table 6 shows the relevant statistics for the two persons:

**Table 6: Relevant Statistics for Individuals Used in Predictions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Person** | **Age** | **Resting Blood Pressure** | **Maximum Heart Rate Achieved** | **Exercise Angina Present** |
| A | 50 | 122 | 140 | Yes |
| B | 50 | 130 | 165 | No |

The odds of Person A having heart disease are 27.2%:

The odds of Person B having heart disease are 78.5%:

From these two predictions the model determines that, even with equal age, the chances of heart disease are dramatically reduced in persons who have lower blood pressure and heart rate – regardless of angina.

## Second Logistic Regression Model with *age* + *trestbps* + *cp* + *thalach* + *age2* + *age*\**thalach*

The second model will try and predict the probability of heart disease (target) from *age, trestbps, cp, thalach, age* squared,andan interactive term between *age* and *thalach.*

### Reporting Results

The first step in building the model is to relevel the qualitative predictor, *cp*. In the data there are four possible values so only three variables are needed to represent the data – a 1 indicates that a quality exists. Table 7 shows these values:

**Table 7: Qualitative Predictors of *cp***

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***cp1*** | ***cp2*** | ***cp3*** |
| **No pain** | 0 | 0 | 0 |
| **Typical angina** | 1 | 0 | 0 |
| **Atypical angina** | 0 | 1 | 0 |
| **Non-anginal pain** | 0 | 0 | 1 |

This second model will use four predictor variables, including the quadratic term for age and an interaction term between age and maximum heart rate achieved as the modelling input to heart disease risk (target). This model will be of the form:

Where y is ‘1’ for heart disease predicted and ‘0’ for not predicted.

The final model is calculated as:

With *age* as X1, *trestbps* as X2, *cp1* as X3, *cp2* as X4, *cp3* as X5, and *thalach* as X6.

This may be linearized into the log-odds form as:

Based on this model and the assumption that if the probability of heart disease is over 50% a person will acquire the affliction the following confusion matrix, Table 9, may be created:

**Table 9: Confusion Matrix for Model using** *age* + *trestbps* + *cp* + *thalach* + *age2* + *age*\**thalach*

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

Which gives rise to the following common measures, Table 10, to help evaluate the model:

**Table 10: Common Measures of Logistic Models**

|  |  |
| --- | --- |
| **Accuracy** | 0.7624 |
| **Precision** | 0.7818 |
| **Recall** | 0.7818 |

To further determine if the model was relevant a Hosmer-Lemeshow Goodness of Fit (GOF) Test is conducted. A GOF is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: the model fits the data*

*Ha: the does not fit the data*

The null hypothesis states that there is a correlation between *target* and the predictor variables. The alternative states there is no correlation between the variablesand *target*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 11 shows the test statistic and its associated P-value:

**Table 11: Hypothesis Test for the Hosmer-Lemeshow Goodness of Fit (GOF) Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 52 |
| P-value | 0.3209 |

The P-value confirms that there not enough evidence to reject the null hypothesis, 0.3209 >> 0.05; thus, the model does fit the data. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

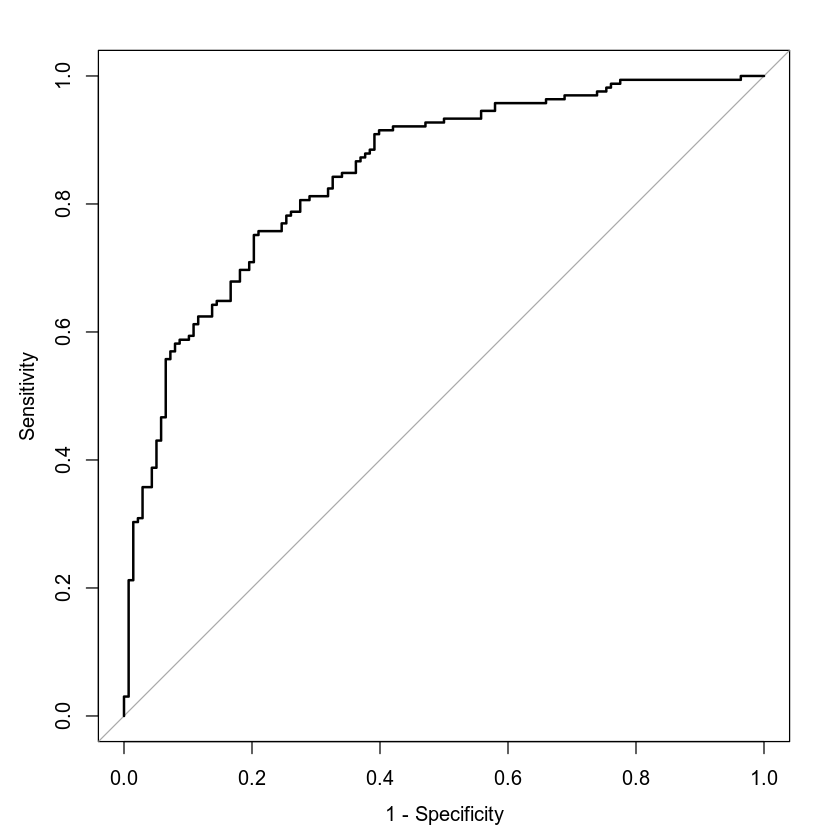
What the GOF test does not reveal is which of the predictor variables are relevant. To determine which are relevant a Wald Confidence Interval is conducted on each variable. The confidence intervals (95% confidence) can be used to determine statistical relevance, see Table 12.

**Table 12: Wald 95% Confidence Intervals**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Lower Limit** | **Upper Limit** |
| *(Intercept)* | -36.2272 | 5.0999 |
| *age* | -0.3488 | 0.6975 |
| *trestbps* | -0.0372 | -0.0020 |
| *cp1* | 1.0439 | 2.7830 |
| *cp2* | 1.3566 | 2.7178 |
| *cp3* | 0.7039 | 2.8507 |
| *thalach* | 0.0360 | 0.2367 |
| *age2* | -0.0026 | 0.0043 |
| *age\*thalach* | -0.0036 | -0.0001 |

Any variables’ interval that contains zero can be deemed statistically irrelevant. The *intercept*, *age*, *and age2* contains a zero and may be considered statistically irrelevant.

As before the ROC Curve was generated and its area calculated. The ROC is shown in Figure 3.



**Figure 3: ROC Plot for** *age* + *trestbps* + *cp* + *thalach* + *age2* + *age*\**thalach*

The ROC curve shown in Figure 3 has an area under the curve of 0.8478. This moderately large value (~1) suggests some relevance of the generated model.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, two persons are considered for risk of heart disease, Person A and Person B. Table 13 shows the relevant statistics for the two persons:

**Table 13: Relevant Statistics for Individuals Used in Predictions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Person** | **Age** | **Resting Blood Pressure** | **Chest Pain Type** | **Maximum Heart Rate Achieved** |
| A | 50 | 115 | None | 133 |
| B | 50 | 125 | Typical Angina | 155 |

The odds of Person A having heart disease are 21.9%:

The odds of Person B having heart disease are 80.1%:

From these two predictions the model determines that, even with equal age, the chances of heart disease are dramatically reduced in persons who have lower blood pressure and heart.

## Random Classification Tree

This CART analysis will build a random classification tree with a maximum of 150 tress.

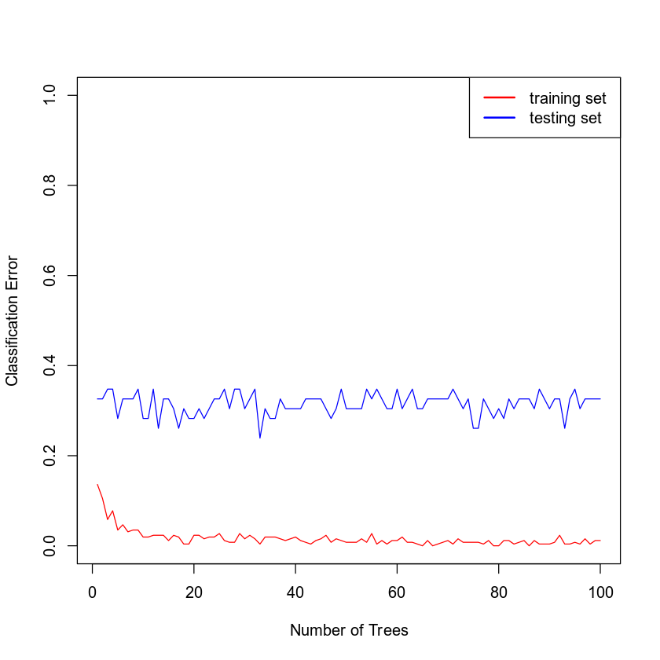
**Reporting Results**

The first step in building the model is to split the data into training and validation sets. The original data set consisted of 303 samples – this was split into 85% training and 15% validation. Table 14 shows the number of rows in each new set.

**Table 14: Training and Validation Samples used in Classification Tree**

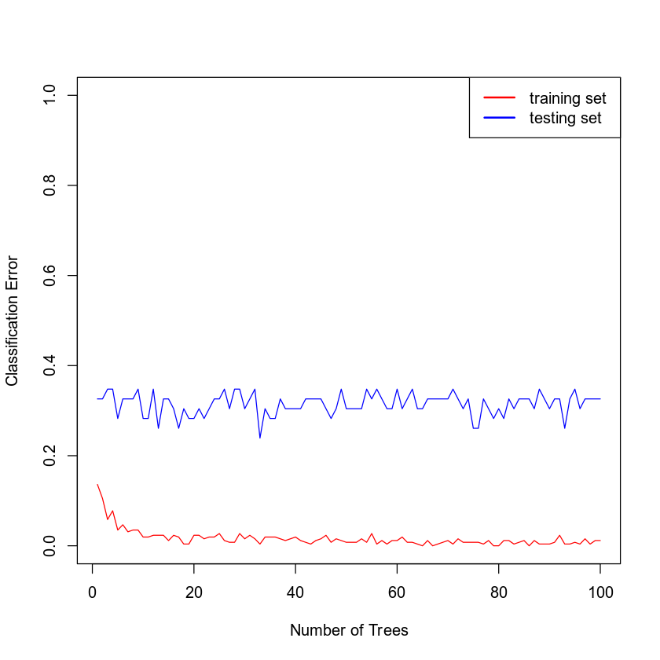
|  |  |
| --- | --- |
|  | **Samples** |
| **Training Set** | 257 |
| **Validation Set** | 46 |

Once the data was segregated between training and validation sets the tree was generated using a seed value of 6522048. Even though there was a maximum of 150 tree allowed at value of generated trees the classification error reduces to a minimum. Figure 4 displays the classification error for the training and validation sets as the number of trees increases to the maximum.



**Figure 4: Classification Error of Training and Testing Sets as the Number of Trees approaches the Maximum**

From Figure 4 it can be shown that at around 20 trees the classification error in the training set falls to a minimum, see Figure 5, and never waivers; thus, 20 is the optimum number of tress and will be used.



**Figure 5: Showing that the Minimum Error Occurs Around 20 Trees**

From these 20 random classification trees the confusion matrix and its associated performance metrics can be calculated for the training and validation sets. One would expect that the metrics for the training set would show better performance (closer to 1) than those of the validation set.

**Table 15: Confusion Matrix for Training Set**

|  |  |  |
| --- | --- | --- |
|  | **Prediction: default=0** | **Prediction: default=1** |
| **Actual: default=0** | 115 | 5 |
| **Actual: default=1** | 1 | 136 |

**Table 16: Common Measures of Classification Models for the Training Set**

|  |  |
| --- | --- |
| **Accuracy** | 0.9922 |
| **Precision** | 0.9784 |
| **Recall** | 0.9927 |

**Table 17: Confusion Matrix for Validation Set**

|  |  |  |
| --- | --- | --- |
|  | **Prediction: default=0** | **Prediction: default=1** |
| **Actual: default=0** | 11 | 7 |
| **Actual: default=1** | 8 | 20 |

**Table 18: Common Measures of Classification Models for the Validation Set**

|  |  |
| --- | --- |
| **Accuracy** | 0.6739 |
| **Precision** | 0.7407 |
| **Recall** | 0.7143 |

As Tables 15-18 show, the metrics for the training set do, in fact, show better performance.

## Random Regression Tree

This CART analysis will build a random regression tree with a maximum of 80 tress.

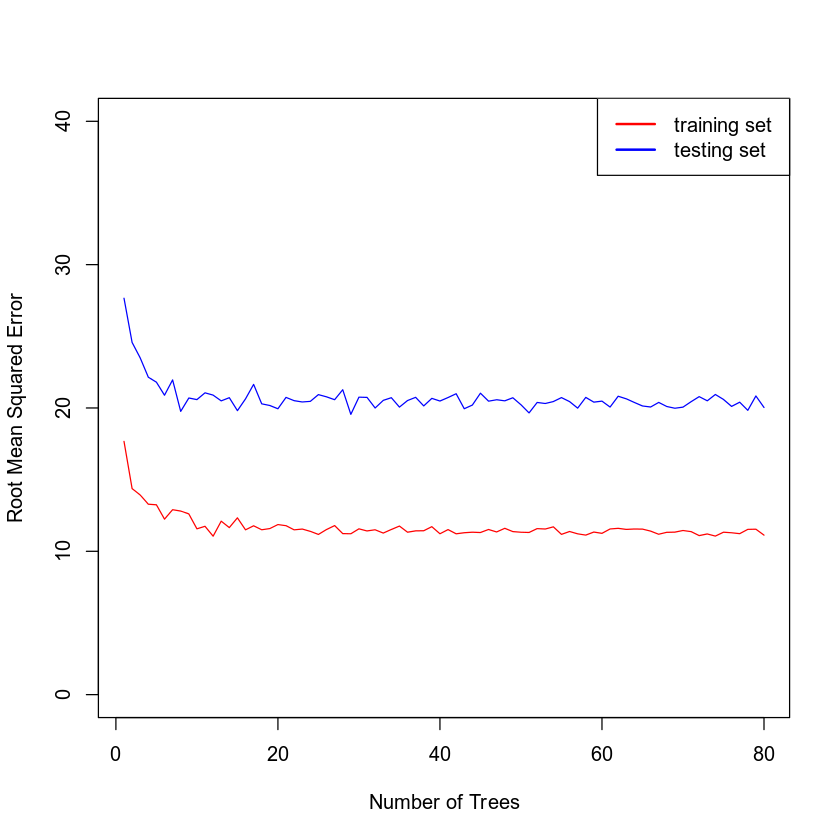
**Reporting Results**

The first step in building the model is to split the data into training and validation sets. The original data set consisted of 303 samples – this was split into 80% training and 20% validation. Table 19 shows the number of rows in each new set.

**Table 19: Training and Validation Samples used in Regression Tree**

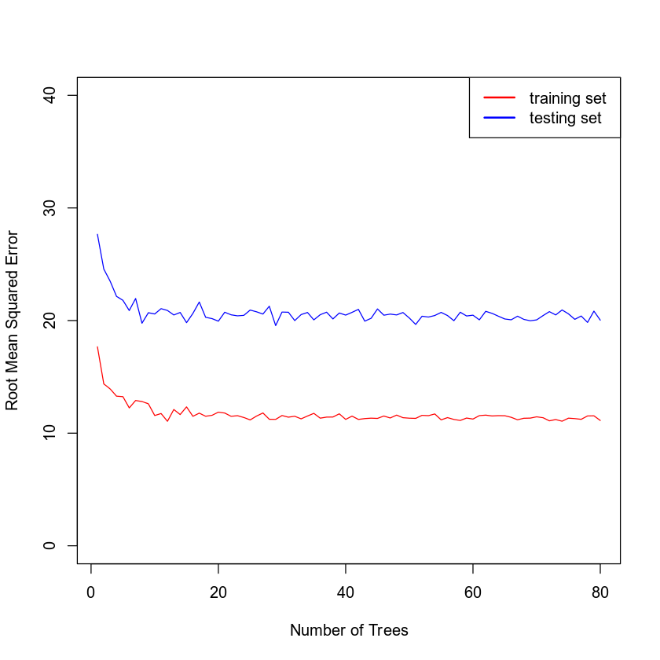
|  |  |
| --- | --- |
|  | **Samples** |
| **Training Set** | 242 |
| **Validation Set** | 61 |

Once the data was segregated between training and validation sets the tree was generated using a seed value of 6522048. Even though there was a maximum of 80 tree allowed at some value of generated trees the root mean square (RSME) error reduces to a minimum. Figure 6 displays the RSME for the training and validation sets as the number of trees increases to the maximum.



**Figure 6: RSME of Training and Testing Sets as the Number of Trees approaches the Maximum**

From Figure 6 it can be shown that at around 20 trees the RSME in the training set falls to a minimum, see Figure 7, and never waivers; thus, 20 is the optimum number of tress and will be used.



**Figure 7: Showing that the Minimum Error Occurs Around 20 Trees**

From these 20 random regression trees the RSME can be calculated for the training and validation sets. One would expect that the metrics for the training set would show better performance (closer to 0) than those of the validation set. See Table 20 for the calculated errors.

**Table 20: RSME for Training and Validation Sets**

|  |  |  |
| --- | --- | --- |
|  | **Training Set** | **Validation Set** |
| **RSME** | 11.467 | 20.88 |

## Conclusion

Two binary logistic models were built. Both models proved to be statistically relevant via their GOF tests, i.e., null hypotheses were not rejected, but which model should be used?

Table 21 shows the comparative metrics on each of the models:

**Table 21: Comparative Metrics for the Two Logistic Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **AUC** |
| **Model One** | 0.7360 | 0.7322 | 0.8121 | 0.8007 |
| **Model Two** | 0.7624 | 0.7818 | 0.7818 | 0.8478 |

As shown in the preceding table Model Two has higher accuracy (ratio of correctness), precision (ability to predict a true positive), and AUC (closer to 1 is better performance). Therefore, even though Model Two had more predictor variables not statistically relevant and lower recall, it appears to be the superior model. Therefore, in any further comparisons only Model Two will be discussed.

After the binary logistic models were created, two random CART models were generated. These were optimized by plotting their associated errors versus the number of trees allowed. Both analyses were found to be optimum at 20 trees.

The respective CART analyses may be compared to the logistic models, i.e., the classification logistic model may be compared to the random classification tree. Table 22 shows the details of these comparisons.

**Table 22: Comparative Metrics for the Random Classification Tree and Logistic Model Two**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** |
| **Model Two** | 0.7624 | 0.7818 | 0.7818 |
| **Classification Tree** | 0.9922 | 0.9784 | 0.9927 |
| **Classification Tree, Random Sample** | 0.6739 | 0.7407 | 0.7143 |

Model Two’s metrics are calculated by feeding the same training data into the model and testing – there are no segregated training and validations sets. Therefore, it is fair to compare the Classification Tree’s metrics from the training set – which are much closer to 1.0 than Model Two’s. Moreover, a random sample used with the Classification Tree compares favorably with Model Two.

Based upon the data shown in Table 22 it is recommended to use the Random Classification Tree.

The models generated in this paper would be useful for diagnostic work in a medical setting or for treatment protocols. Based upon am individuals risk factors a physician could estimate the likelihood of heart disease and a course of treatment.

## Citations

Hobbs, B. (2022). *MAT 303 project one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module five summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module six summary report*. [Unpublished report]. SNHU.