1. Introduction

I am exploring the heart_disease.csv data set. I will use this data set to analyze patterns between different health indicators such as maximum heart rate, fasting blood sugar, and the presence of heart disease. I will be running different logistic regression models that will predict whether or not a person is at risk for heart disease. I will also create a classification random forest model to predict the risk of heart disease and a regression random forest model to predict the maximum achieved heart rate. There are 303 rows and 14 columns in this data set.

2. Data Preparation

I will use important variables in this project:

- age, the person's age in years
- sex, the person's sex (1 = male, 0 = female)
- cp, the type of chest pain (0 = no pain, 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain)
- trestbps, the person's resting blood pressure
- chol, the person's cholesterol measurement in mg/dl
- fbs, the person's fasting blood sugar is greater than 120 mg/dl (1 = true, 0 = false)
- restecg, resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2=showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach, the person's maximum heart rate achieved
- exang, exercise-induced angina (1 = yes, 0 = no)
- oldpeak, ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
- slope, the slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
- ca, the number of major vessels (0-3)
- target, heart disease (0 = no, 1 = yes)

3. Model #1 - First Logistic Regression Model

Reporting Results

The general form of the multiple regression model for the heart disease (target) using variables age (age), resting blood pressure (trestbps), and maximum heart rate achieved (thalach) is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. The intercept is β_0, β_1, β_2 and β_3 are the regression terms for age, resting blood pressure and maximum heart rate achieved.

The prediction model equation with R script output is $\hat{y}$ = -3.576198 – 0.009424 x_1 - 0.016019 x_2 + 0.042697 x_3. The prediction model equation in terms of the natural log of odds to express the beta terms in linear form is:
$\ln(p /1 - p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Natural log of odds is:
$\ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
p is the probability that heart disease will occur.
p /(1 - p) is the odds of developing heart disease.

The logistic regression model is
$E(Y) = e(-3.576198 – 0.009424\ x_1 - 0.016019\ x_2 + 0.042697\ x_3)/(1 + e\ ^(-3.576198 – 0.009424\ x_1 - 0.016019\ x_2 + 0.042697\ x_3)$

The prediction model equation (in terms of the natural log of odds) is
$\ln(odds) = -3.576198 – 0.009424\ x_1 - 0.016019\ x_2 + 0.042697\ x_3$

The estimated coefficient of the variable maximum heart rate achieved (thalach) is 0.042697. This means that on average, the change in log odds for developing heart disease is 0.042697 with all variables constant.

Evaluating Model Significance

We will run the Hosmer-Lemeshow goodness of fit test to determine if this model is appropriate for this data set.
Identifying the null and alternative hypothesis:
H_0: model is significant (fits data set)
H_1: is not significant (model does not fit the data set)

The test statistic is 41.978, and p-value is 0.7168. P-value is larger than the 0.05 level of significance, so we accept the null hypothesis that the logistic regression model is significant or fits the data set.

Using Wald's tests, we will determine the significance of each term.
Checking for significant terms at .05 significance level:

β_1 = age
H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
β_2 is the resting blood pressure (trestbps), the null and alternative hypothesis:
H_0: $\beta_2 = 0$
H_1: $\beta_2 \neq 0$
β_3 is the maximum heart rate achieved (thalach), the null and alternative hypothesis:
H_0: $\beta_3 = 0$
H_1: $\beta_3 \neq 0$

The p-value for age is 0.5578, so this term is not statistically significant
The p-value for resting blood pressure is 0.0392, so this term is statistically significant
The p-value for maximum heart rate is 8.06e-10, so this term is statistically significant

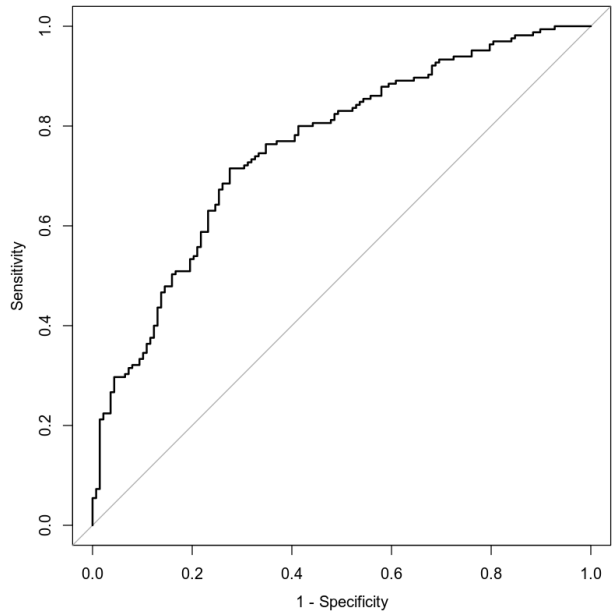
The confusion matrix for this analysis:
True positive: 127
True negative: 83
False positive: 55
False negative: 38

Accuracy is the ratio of the number of correct predictions to the total number of observations:
Accuracy = (127 + 83)/127 + 83 + 55 +38 = 0.6930 = 69.30%

Precision is the ratio of correct positive predictions to the total predicted positives:
Precision = 127/127 + 55 = 0.6978 = 69.78%

Recall is the ratio of correct positive predictions to the total positives examples:
Recall = 127/127 + 38 = 0.7690 = 76.90%

Receiver Operating Characteristic (ROC) curve graph:



The area under the curve (AUC) is 0.7575 or 75.75%. This indicates that the model strongly differentiates between Y = 0 and Y = 1. The larger the area under the curve, the model is better at predicting an individual developing heart disease.

Making Predictions Using Model

The probability of an individual having heart disease, if they are 50 years old, has a resting blood pressure of 122 and has a maximum heart rate of 140 is 0.4939 or 49.39%. This shows there is a 49.39% chance of an individual having or developing heart disease.

The probability of an individual having heart disease, if they are 50 years old, has a resting blood pressure of 130 and has a maximum heart rate of 165 is 0.7140 or 71.40%. This shows there is a 71.4% chance of an individual having or developing heart disease.

4. Model #2 - Second Logistic Regression Model

Reporting Results

The general form of the multiple regression model is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x^2_2 + \beta_9 x_1 x_2$

β_1, β_2, β_3 and β_4 are the regression terms for maximum heart rate achieved (thalach) using variables age (age), dummy terms for sex1 (male) and exercised-induced angina (exang1). β_5, β_6 and β_7 are the dummy terms for chest pain types, cp1, cp2 and cp3. The quadratic term for age and the interaction term for maximum heart rate achieved (thalach): age are a respective β_8 and β_9.

The prediction model equation in terms of the natural log of odds to express the beta terms in linear form is:
$(p/1-p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x^2_2 + \beta_9 x_1 x_2$

Natural log of odds is:
$\ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x^2_2 + \beta_9 x_1 x_2$

The logistic regression model is
$E(Y) = e(-1.634e+01 + 1.390e-01\ x_1 + 2.049e-01\ x_2 - 1.709e+00\ x_3 - 9.348e-01\ x_4 + 1.766e+00\ x_5 + 1.820e+00\ x_6 + 1.674e+00\ x_7 + 4.921e-04\ x^2_2 - 2.017e-03\ x_1 x_2)/(1 + e\ ^(-1.634e+01 + 1.390e-01\ x_1 + 2.049e-01\ x_2 - 1.709e+00\ x_3 - 9.348e-01\ x_4 + 1.766e+00\ x_5 + 1.820e+00\ x_6 + 1.674e+00\ x_7 + 4.921e-04\ x^2_2 - 2.017e-03\ x_1 x_2))$

The prediction model equation (in terms of the natural log of odds) is
$\ln(odds) = -1.634e+01 + 1.390e-01\ x_1 + 2.049e-01\ x_2 - 1.709e+00\ x_3 - 9.348e-01\ x_4 + 1.766e+00\ x_5 + 1.820e+00\ x_6 + 1.674e+00\ x_7 + 4.921e-04\ x^2_2 - 2.017e-03\ x_1 x_2$

Evaluating Model Significance

We will run the Hosmer-Lemeshow goodness of fit test to determine if this model is appropriate for this data set.
Identifying the null and alternative hypothesis:
H_0: model is significant (fits data set)
H_1: is not significant (model does not fit the data set)

The test statistic is 60.596, and p-value is 0.1048. P-value is larger than the 0.05 level of significance, so we accept the null hypothesis that the logistic regression model is significant or fits the data set.

Using Wald's tests, we will determine the significance of each term.
Checking for significant terms at .05 significance level:

H_0: $\beta_{1,2,3,4,5,6,7,8,9} = 0$
H_1: $\beta_{1,2,3,4,5,6,7,8,9} \neq 0$
β_1 = maximum heart rate achieved (thalach), the p-value is 0.014760, which is statistically significant at a 0.05 level of significance
β_2 = age, the p-value is 0.510325, which is not statistically significant at a 0.05 level of significance
β_3 = sex (sex1 = male), the p-value is 1.91e-06, which is statistically significant at a 0.05 level of significance
β_4 = exercised induced angina (exang1= yes), the p-value is 0.009133, which is statistically significant at a 0.05 level of significance
β_5 = dummy variable for chest pain (cp1= typical angina), the p-value is 0.000249, which is statistically significant at a 0.05 level of significance

β_6 = dummy variable for chest pain (cp2= typical angina), the p-value is 2.21e-06, which is statistically significant at a 0.05 level of significance
β_7 = dummy variable for chest pain (cp3= non-anginal pain), the p-value is 0.003684, which is statistically significant at a 0.05 level of significance
β_8 = quadratic term for age, the p-value is 0.810599, which is not statistically significant at a 0.05 level of significance
β_9 = interaction term for the maximum heart rate (thalach): age, the p-value is 0.043666, which is statistically significant at a 0.05 level of significance

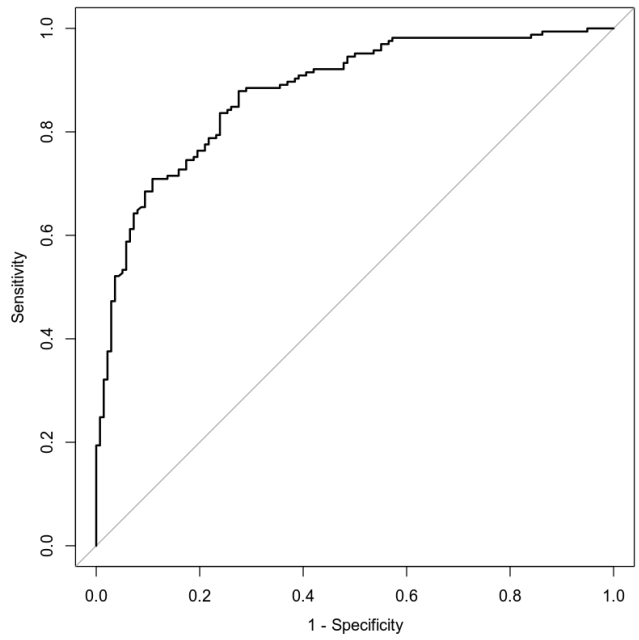
The confusion matrix for this analysis:
True positive: 129
True negative: 102
False positive: 36
False negative: 36

Accuracy is the ratio of the number of correct predictions to the total number of observations:
Accuracy = (129 + 102)/129 + 102 + 36 +36 = 0.7623 = 76.23%

Precision is the ratio of correct positive predictions to the total predicted positives:
Precision = 129/129 + 36 = 0.7818 = 78.18%

Recall is the ratio of correct positive predictions to the total positives examples:
Recall = 129/129 + 36 = 0.7818 = 78.18%

Receiver Operating Characteristic (ROC) curve graph:



The area under the curve (AUC) is 0.8478 or 84.78%. This indicates that the model strongly differentiates between Y = 0 and Y = 1. The larger the area under the curve, the model is better at predicting an individual developing heart disease.

Making Predictions Using Model

The probability of an individual having heart disease, if they are 50 years old, has a resting blood pressure (trestbps) of 115, does not experience chest pain (cp0) and has a maximum heart rate (thalach) of 133 is 0.2188 or 21.88%. This shows there is a 21.88% chance of an individual having or developing heart disease.

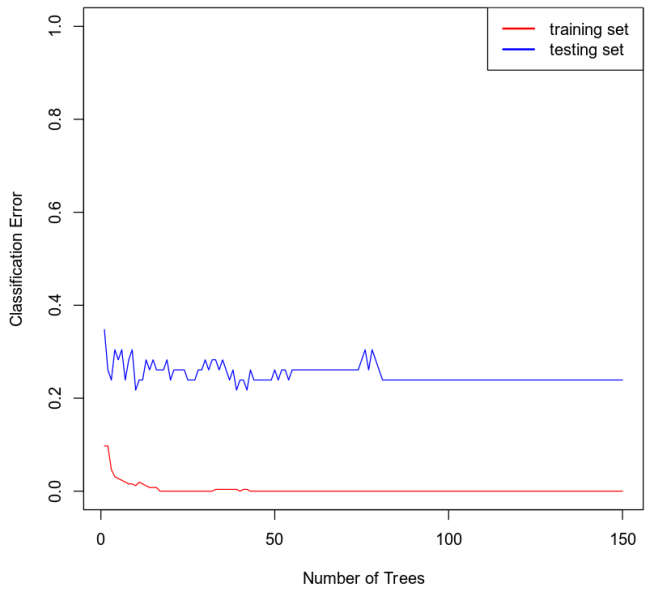
The probability of an individual having heart disease, if they are 50 years old, has a resting blood pressure of 125, does experience typical angina and has a maximum heart rate of 155 is 0.8007 or 80.07%. This shows there is an 80.07% chance of an individual having or developing heart disease.

5. Random Forest Classification Model

Reporting Results

Using set.seed(6522048) and splitting the heart disease data set into training and testing sets using 85% and 15% split, there are 257 rows for the training set and 46 rows for the testing set.

Training and Testing Error against the Number of Trees Graph



The optimal number of trees for the random forest model is 20. This is after the first drop.

Evaluating the Utility of the model

We can create a random forest classification model to determine the presence of heart disease (target), using the variables age, sex, chest pain type (cp), resting blood pressure (trestbps) cholesterol (chol), resting ecg (restecg), exercised-induced angina (exang), slope of the peak exercise ST segment (slope) and number of major vessels using the number of trees found. We can get the confusion matrix for the training set along with the accuracy, precision and recall.

```
[1] "=============================================================================
============================================="
[1] "Confusion Matrix: TRAINING set based on Random Forest model built using 20
trees"
```

A matrix: 2 × 2 of type chr

	Prediction: 0	Prediction: 1
Actual: 0	120	0
Actual: 1	0	137

The confusion matrix for the training set:
True positive: 137
True negative: 120
False positive: 0
False negative: 0

Accuracy is the ratio of the number of correct predictions to the total number of observations:
Accuracy = (137 + 120)/137 + 120 + 0 +0 = 1.0000 or 100%

Precision is the ratio of correct positive predictions to the total predicted positives:
Precision = 137/137 + 0 = 1.0000 or 100%

Recall is the ratio of correct positive predictions to the total positives examples:
Recall = 137/137 + 0 = 1.0000 or 100%

```
[1] "=============================================================================
============================================="
[1] "Confusion Matrix: TESTING set based on Random Forest model built using 20
trees"
```

A matrix: 2 × 2 of type chr

	Prediction: 0	Prediction: 1
Actual: 0	13	5
Actual: 1	6	22

The confusion matrix for the testing set:
True positive: 22
True negative: 13
False positive: 5
False negative: 6

Accuracy = (22 + 13)/22 + 13 + 5 + 6 = 0.7608 or 76.08%

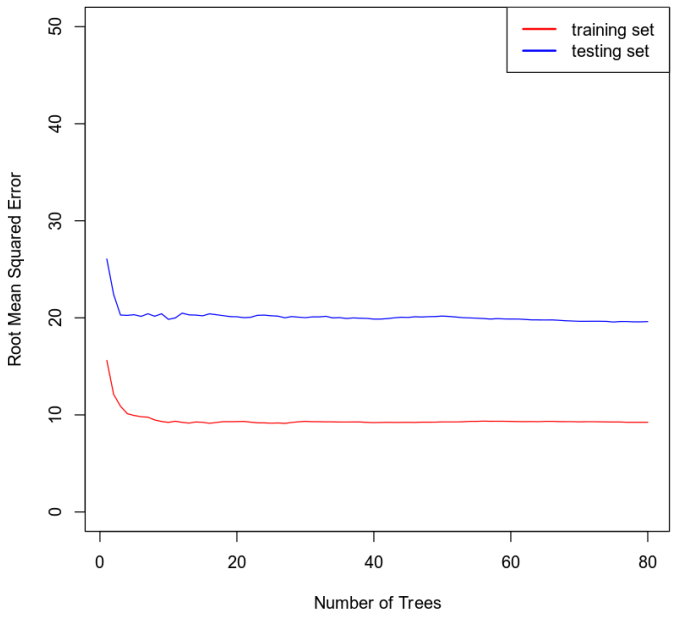
Precision = 22/22 + 5 = 0.8148 or 81.48%
Recall = 22/22 + 6 = 0.7857 or 78.57%

6. Random Forest Regression Model

Reporting Results

Using set.seed(6522048) and splitting the heart disease data set into training and testing sets using 80% and 20% split, there are 242 rows for the training set and 61 rows for the testing set.

The graph for 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). I used set.seed(6522048).



The optimal number of trees for this random forest model is 9 as this is after the first drop.

Evaluating the Utility of the Random Forest Regression Model

The root mean squared error for the training set using 80 trees is 9.2448
The root mean squared error for the testing set using 80 trees is 19.6189

7. Conclusion

I would choose to use the second of the logistic regression models to predict heart disease. The second regression model has a larger area under the ROC curve than the first model(84.78% compared to 75.75%). The second model also uses more variables, and have higher accuracy, precision and recall values (accuracy = 76.23%, precision = 78.18%, and recall = 78.18%, compared to accuracy = 69.30%, precision = 69.78%, and recall = 76.90%).

I would recommend the random forest classification model as the confusion matrix is better (accuracy, precision and recall are all at 100%). The reason for creating a model is for medical doctors to use when evaluating patients medical record data and determine heart disease risk.