1. Introduction

This project was centered on the examination of the 'heart_disease' dataset. This dataset highlights various health indicators. The goal was to craft logistic regression models to predict the likelihood of an individual developing heart disease. I will also be building two Random Forest models. The insights drawn from this investigation aim to improve diagnostic precision and offer enhanced risk evaluations through data-driven perspectives.

2. Data Preparation

The heart disease dataset features 14 columns and 303 rows.
The important variables in this dataset are listed below:

Variable	What does it represent?
age	The person's age in years
sex	The person's sex (1 = male, 0 = female)
cp	The type of chest pain experienced (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 equation of the logistic multiple regression model for heart disease (target) using variables age(age) resting blood pressure (trestbps), exercised induced angina (exang), and maximum heart rate achieved (thalach) is:

$$E(y) = \frac{e^{\beta_0+\beta_1\times\text{age}+\beta_2\times\text{trestbps}+\beta_3\times\text{exang}+\beta_4\times\text{thalach}}}{1+e^{\beta_0+\beta_1\times\text{age}+\beta_2\times\text{trestbps}+\beta_3\times\text{exang}+\beta_4\times\text{thalach}}}$$

The prediction model equation in terms of the natural log of odds that expresses the beta terms in linear form is:

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1\times\text{age}+\beta_2\times\text{trestbps}+\beta_3\times\text{exang}+\beta_4\times\text{thalach}}$$

The symbol, π represents the probability of an event occurring and 1−π represents the probability of the event not occurring.

The output from the first model is as follows:

```
Call:
glm(formula = target ~ age + trestbps + exang + thalach, family = "binomial",
    data = heart_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.021121   1.784194  -0.572   0.5671
age         -0.017549   0.017144  -1.024   0.3060
trestbps    -0.014888   0.008337  -1.786   0.0741 .
exang1      -1.624981   0.305774  -5.314 1.07e-07 ***
thalach      0.031095   0.007275   4.274 1.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 323.14  on 298  degrees of freedom
AIC: 333.14

Number of Fisher Scoring iterations: 4
```

Using the output from the model, the prediction model equation can be written as:

log(π/1−π) = -1.0211 – 0.0175 * x(age) – 0.0149 * x(trestbps) – 1.6250 * x(exang) + 0.0311 * (thalach)

The coefficient for thalach (maximum heart rate achieved) is 0.0311. In the context of logistic regression, the interpretation of coefficients is in terms of the logarithm of odds ratios. Specifically, for a one-unit

increase in thalach, the odds of having heart disease increases by 0.0311, holding all other variables constant.

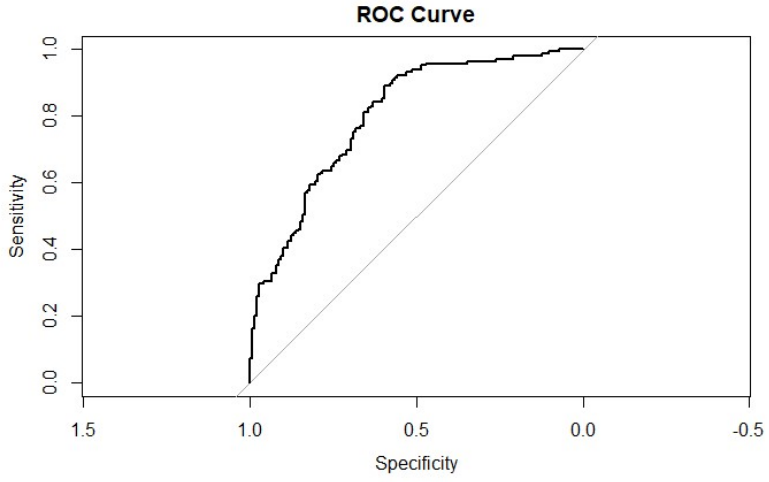
Evaluating Model Significance

The Hosmer-Lemeshow goodness of fit test evaluates if the predictions for developing heart disease are close to the observed values. The null hypothesis states that the model fits the data, and the alternative hypothesis states the model does not fit the data.

A p-value greater than 0.05 indicates that the null hypothesis should not be rejected and suggests that the model's fit is adequate. This model's p-value is 0.3264. Therefore, we are not going to reject the null hypothesis and we can conclude that the model and the data are a good fit.

To determine which terms are significant based on Wald's test, the p-values associated with each coefficient in the logistic regression output are compared to the significance level of 5%. In this model, several terms were assessed for their significance in predicting heart disease. The term "age" has a p-value of 0.3060, which is greater than the threshold of 0.05, indicating that age is not a significant predictor in this model. Similarly, both the target term of whether or not a person has heart disease as well as a person's resting blood pressure have p-values of 0.5671 and 0.0741, both exceeding the 0.05 threshold, are not significant predictors. The term that represents a person's maximum achieved heart rate, and the term that represents whether or not a person has exercise-induced angina have notably small p-values of 1.07e-07 and 1.92e-05 respectively, which are well below the 5% level of significance placed upon them. Therefore, these are the terms that would be significant predictors of heart disease in the model.

The model's accuracy is approximately 73.6% when it comes to being able to make accurate predictions about heart disease. This is a good score. The precision of the model is at 73.2%, and it tells us how often the predictions are actually correct. This ensures that unnecessary alarms are not being raised too often. The model's recall value is 81.2%. This would indicate the number of patients with heart disease that were correctly identified by the model. This number establishes that the model is doing a respectable job at identifying true cases of heart disease. If all of these values are considered out of 100, it is not difficult to see why this is a good model.



The Receiver Operating Characteristic (ROC) curve for the model shows a trajectory leaning towards the upper left corner, with its peak situated around the middle. This positioning is indicative of the model's capability to discern between the two outcomes in our study which are whether heart disease is present or not. The Area Under the Curve (AUC) is derived from the ROC curve, offering a consolidated measure of the model's proficiency in differentiating between the positive and negative cases. The AUC value can span between 0 to 1. Ideally, a value closer to 1 signifies a model with exceptional distinguishing power. For our model, the AUC is 0.8007, implying it excels at identifying individuals with and without heart disease.

Making Predictions Using Model

The equation for making predictions on the scenarios is as follows:

$\text{Log}(\pi/1-\pi) = -1.0211 - 0.0175 \times \text{age} - 0.0149 \times \text{trestbps} - 1.6250 \times \text{exang} + 0.0311 \times \text{thalach}$

π represents the probability of having heart disease and $1-\pi$ represents the odds of having heart disease. By plugging in the numbers provided and running the necessary R code we can deduce that for the first scenario which states that the predicted probability that a 50-year-old individual with a resting blood pressure of 122, who has exercise-induced angina, and has a maximum heart rate of 140, will have heart disease is approximately 0.2719 or 27.19%. The odds of this event occurring are 0.3735. This means that for every 373 individuals with these characteristics predicted to have heart disease, there would be about 1,000 who are not predicted to have heart disease.

Scenario 2 referenced the predicted probability that a 50-year-old individual with a resting blood pressure of 130, who does not have exercise-induced angina, and has a maximum heart rate of 165, will have heart disease is approximately 0.7856 or 78.56%. The odds of this event occurring are 3.6634. This means for every 3663 individuals with these characteristics predicted to have heart disease, there would be about 1,000 who are not predicted to have heart disease.

From these results we can deduce that individuals fitting the description of scenario 2 have a much higher predicted probability of having heart disease than those in scenario 1. The odds in scenario 2 are also significantly higher, indicating that for every individual not predicted to have heart disease, there are about 3.66 times more individuals predicted to have heart disease.

4. Model #2 - Second Logistic Regression Model

Reporting Results

$$E(y) = \frac{e^{\beta_0+\beta_1\times\text{age}+\beta_2\times\text{trestbps}+\beta_3\times\text{cp}+\beta_4\times\text{thalach}+\beta_5\times\text{age}^2+\beta_6\times\text{age}\times\text{thalach}}}{1+e^{\beta_0+\beta_1\times\text{age}+\beta_2\times\text{trestbps}+\beta_3\times\text{cp}+\beta_4\times\text{thalach}+\beta_5\times\text{age}^2+\beta_6\times\text{age}\times\text{thalach}}}$$

$$\log\left(\frac{E(y)}{1-E(y)}\right) = -1.762 + 0.2139 \times \text{age} - 0.02217 \times \text{trestbps} + 0.8547 \times \text{cp} + 0.1532 \times \text{thalach} + 0.0007836 \times \text{age}^2 - 0.0021 \times \text{age} \times \text{thalach}$$

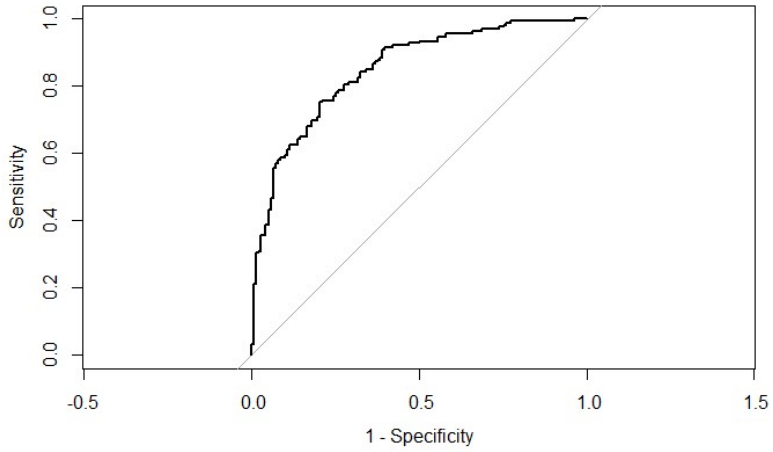
The output for the second model is as follows:

```
Call:
glm(formula = target ~ age + trestbps + cp + thalach + I(age^2) +
    age:thalach, family = "binomial", data = heart_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.556e+01  1.054e+01  -1.476  0.13988
age          1.744e-01  2.669e-01   0.653  0.51357
trestbps    -1.958e-02  8.978e-03  -2.181  0.02916 *
cp1          1.913e+00  4.437e-01   4.313 1.61e-05 ***
cp2          2.037e+00  3.473e-01   5.867 4.45e-09 ***
cp3          1.777e+00  5.477e-01   3.245  0.00117 **
thalach      1.363e-01  5.119e-02   2.663  0.00775 **
I(age^2)     8.424e-04  1.750e-03   0.481  0.63025
age:thalach -1.867e-03  8.909e-04  -2.095  0.03616 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 293.67  on 294  degrees of freedom
AIC: 311.67

Number of Fisher Scoring iterations: 5
```

Evaluating Model Significance

The Hosmer-Lemeshow goodness of fit test (GOF) was conducted on the model to see if it fits the data well. The important statistic derived from this test is the p-value, which tells us if there is a significant difference between the observed and expected values. We retrieved a p-value of 0.6418, which is greater than the threshold of 0.05 and therefore we do not reject the null hypothesis, which states that the model is a good fit for the data. The alternative hypothesis would be that the model is not a good fit.

Based on the results of the confusion matrix, the model correctly identified heart disease about 76.24% of the time, as noted by the accuracy level. When the model predicted an individual has heart disease, it was correct 78.18% of the time as per the precision level, and out of all the individuals who truly had heart disease, the model successfully detected 78.18% of them. The Receiver Operating Curve, (ROC) is pictured below, and it plots the true positive rate (sensitivity) against the false positive rate, (specificity). The Area Under the Curve (AUC) is a measure that evaluates the ability of a model to distinguish between positive and negative outcomes, and it is derived from the ROC. An AUC value of 0.8478 indicates that the model has a high discriminative power. Specifically, there's an 84.78% chance that the model will correctly differentiate between a randomly chosen positive case and a randomly chosen negative case. Values closer to 1 suggest excellent model performance, so an AUC of 0.8478 is indicative of a strong model in terms of its classification ability.



Making Predictions Using Model

The probability of an individual having heart disease who is years old, has a resting blood pressure of 115, does not experience chest pain, and has maximum heart rate of 133 is 0.2188 or 21.88%. The odds are another way to represent this likelihood. In this case our odds are 0.2800 to 1, or 28 to 100.

The second individual's case states that the probability of them having heart disease at the age of 50, with a resting blood pressure of 125, experiences typical angina, and has maximum heart rate of 155 is 0.8007 or 80.07% The odds value is 4.0176 and suggests that for every individual with the aforementioned health metrics who has heart disease approximately 4 others with similar metrics will not. In other words that odds are about 4.02 to 1.

After evaluating the results based on both of these individual cases it can be concluded that the first individual has a lower probability of having heart disease compared to the second individual. The odds of the first individual having heart disease are about 28 to 100, whereas the odds for the second individual are substantially higher at about 4.02 to 1. The risk factors of typical angina, higher resting blood pressure, and a higher maximum heart rate seem to have a significant impact on increasing the likelihood of heart disease. This is indicated by the stark difference in the probabilities and odds for the two individuals.

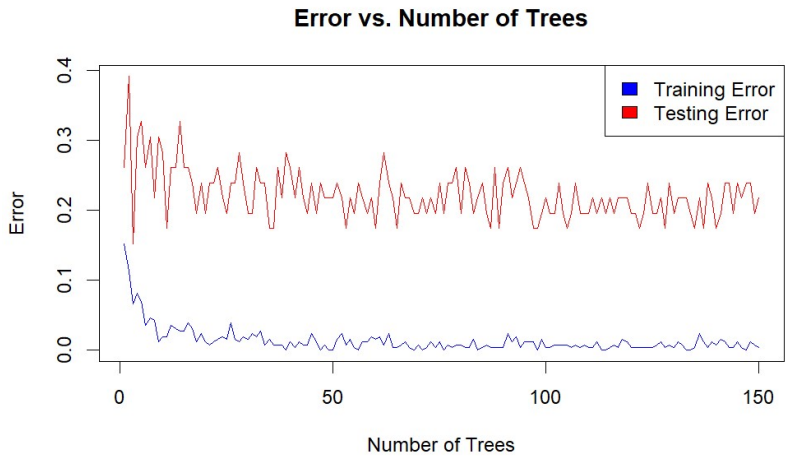
5. Random Forest Classification Model

Reporting Results

The output from the random forest classification model is as follows:

Number of rows in original data set: 303
Number of rows in training set: 257
Number of rows in testing set: 46

Optimal number of trees: 3



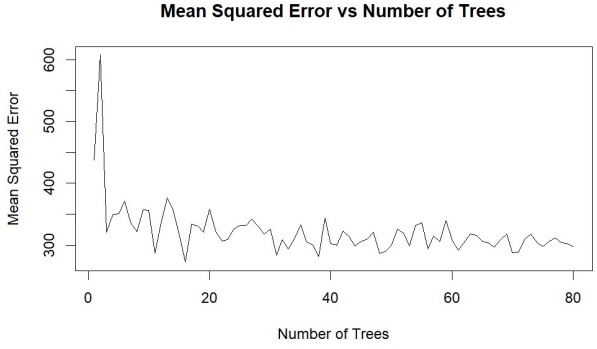
Evaluating the Utility of the model

Upon employing a random forest model for the prediction of heart disease presence, several key parameters were considered: age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), and the number of major vessels (ca). The model exhibited an accuracy rate of 65.22%, which denotes that the presence or absence of heart disease was correctly identified in about 65 out of every 100 cases. The precision stood at 68.97%, suggesting that of all cases predicted as having heart disease, approximately 69% actually had the condition. The recall rate was 74.04%, which indicated that the model successfully identified around 74% of all actual heart disease cases. These figures offer a view of the model's reliability when it comes to classifying cases correctly.

6. Random Forest Regression Model

Reporting Results

Graph of mean squared error



The optimal number of trees for this random forest regression model is 16.

Evaluating the Utility of the Random Forest Regression Model

After splitting the heart disease dataset and designating 80% of the data for training and the remaining 20% for testing purposes, the target variable was predicted based on a set of predefined predictors. Post-training, the model's performance was assessed by computing the Root Mean Square Error (RMSE) on both the training and testing sets. The training RMSE came out to 15.0095, suggesting an average deviation of about 15 units between the actual and predicted target. The testing RMSE was higher at 21.8931, indicating a mean discrepancy of about 22 units for the testing set predictions. The difference in RMSE values between the training and testing sets implies that while the model fits the training data well, it shows a reduced performance on testing data.

7. Conclusion

In conclusion, I would recommend the second logistic regression model with interactions and polynomial terms for predicting heart disease. This more sophisticated model considers interactions between factors like age and heart rate and is better equipped to predict heart disease than the simpler model. Furthermore, I would recommend the logistic regression model over the random forest classification model because the logistic regression model carries better predictive accuracy. The logistic regression model is also easier to interpret, allowing an interdisciplinary team, namely medical professionals to understand which factors are most influential in determining heart disease risk.

The analysis conducted could play a vital role in the medical field. By using the recommended model, healthcare professionals will be better equipped to diagnose and treat heart disease. Early and accurate detection of positive cases can aid in the management and possible reversal of the condition. The result would be an optimization of patient care that could also lead to potential savings in medical costs and, most importantly, better patient health outcomes.