UNIVERSITY OF CALIFORNIA

Los Angeles

Heart Disease Prediction Using Machine Learning Algorithms

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Shu Jiang

2020

ABSTRACT OF THE THESIS

Heart Disease Prediction Using Machine Learning Algorithms

by

Shu Jiang

Master of Applied Statistics

University of California, Los Angeles, 2020

Professor Yingnian Wu, Chair

This paper is focused on the possibility of having heart disease by training four machine learning algorithms. By using the data provided by the UCI Machine Learning Repository, we can analyze and compare the models of logistic regression, random forest, extreme gradient boosting and neural network to choose the most robust model and determine important features in our model.

The thesis of Shu Jiang is approved.

Frederic R Paik Schoenberg

Nicolas Christou

Yingnian Wu, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

According to the World Health Organization, Cardiovascular diseases (CVDs) are the number 1 cause of death globally: more people die annually from CVDs than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke [1]. The high mortality rate and expensive surgery cost already make heart disease become a serious threat for many families in many parts of the world, especially those poor-stricken countries. Therefore, it is crucial for people to analyze the relationship between various kinds of attributes in human and the possibility of suffering from heart disease. A robust model is helpful and meaningful in predicting which type of people is more likely to have a heart disease thus we can prepare or prevent in advance.

Machine learning is closely related to computational statistics, which focus on using mathematical optimization to deliver methods, theory and application domains to solve medical, industry, social and business problems in the real world. It can be divided into two broad categories: supervised learning and unsupervised learning. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. While in unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels. Since our objective is to predict the possibility of having heart disease based on the physical body function. And inputs and desired outputs are expected, definitely we should choose the supervised learning. In this thesis, I would use four models (logistic regression, random forest, extreme gradient boosting and neural network) in supervised learning to predict the possibility of people to have heart

disease based on people's physical function of the body.

To fulfill the objective of my thesis, the following steps may be followed.

1. Download the data set "Heart Disease" from UCI Machine Learning Repository webpage
2. Clean the data and perform exploratory data analysis
3. Apply four different models
4. Compare the results of these four models
5. Draw conclusion based on the results

Chapter 2

Exploratory data analysis

2.1 data set information

This data set is obtained from UCI Machine Learning Repository webpage. It contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The total observations are 303. And it was donated by David W. Aha.

2.2 Attribute Information

AGE: age in years

SEX: (1 = male; 0 = female)

CP (Chest Pain Type):

--Value 0: typical angina (most serious)

--Value 1: atypical angina

--Value 2: non-anginal pain

--Value 3: asymptomatic (least serious)

TRESTBPS: resting blood pressure (in mm Hg on admission to the hospital)

CHOL: serum cholesterol in mg/dl

FBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

A fasting blood sugar level less than 100 mg/dL is normal. From 100 to 120 mg is considered prediabetes. If it is 125 mg/dL or higher on two separate tests, you have diabetes.

RESTECG (Resting Electrocardiographic Results):

--Value 0: normal

--Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

--Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

THALACH: maximum heart rate achieved

EXANG: exercise induced angina (1 = yes; 0 = no)

OLDPEAK: ST depression induced by exercise relative to rest

SLOPE (the slope of the peak exercise ST segment):

--Value 0: upsloping

--Value 1: flat

--Value 2: downsloping

CA: number of major vessels (0-3) colored by fluoroscopy

THAL: 3 = normal; 6 = fixed defect; 7 = reversable defect

TARGET: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

(in any major vessel: attributes 59 through 68 are vessels)

2.3 Exploratory Data Analysis

First of all, we need to check if there exists missing value in the data set. (Figure 2.1)

Figure 2.1: Missing Values

```
#check if missing values exist
sum(is.na(data))
```

```
## [1] 0
```

Figure 2.1 indicates that our data set does not contain missing values.

Next we will do the data cleaning: transform some data to categorical variables.

The next picture (Figure 2.2) is the summary of our current data.

Figure 2.2: Summary of Our Data

```
      age              sex          cp             trestbps          chol            fbs
 Min.   :29.00   F: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
 1st Qu.:47.50   M:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
 Median :55.00           2: 87   Median :130.0   Median :240.0
 Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
 Max.   :77.00                   Max.   :200.0   Max.   :564.0
 restecg    thalach       exang       oldpeak      slope   ca       thal
 0:147   Min.   : 71.0   0:204   Min.   :0.00   0: 21   0:175   0:  2
 1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1:140   1: 65   1: 18
 2:  4   Median :153.0            Median :0.80   2:142   2: 38   2:166
         Mean   :149.6            Mean   :1.04           3: 20   3:117
         3rd Qu.:166.0            3rd Qu.:1.60           4:  5
         Max.   :202.0            Max.   :6.20
 target
 0:138
 1:165
```

From Figure 2.2 we can see that there are 5 numerical variables and 9 categorical variables. We set "target" to be the dependent variable. The mean of age is 54.37, which indicates that the elderly people tend to be more careful about heart disease than any other any groups. From the perspective of sex, there are 96 women and 207 men. It seems that the proportion of males doing the test is higher than that of women. But we cannot infer the connection of the possibility of having heart disease with sex. Most of the people who have been tested have diabetes (fasting blood sugar >= 120), which indicates that heart disease may have connection with high fasting blood sugar. About half of the tested people have ST-T wave abnormality in resting electrocardiographic results, but there are only a few have definite or probable ventricular hypertrophy by Estes's criteria.

In the next part we will see the distribution of the "Target" (Figure 2.3)

Figure 2.3 Presence and Absence of Heart Disease

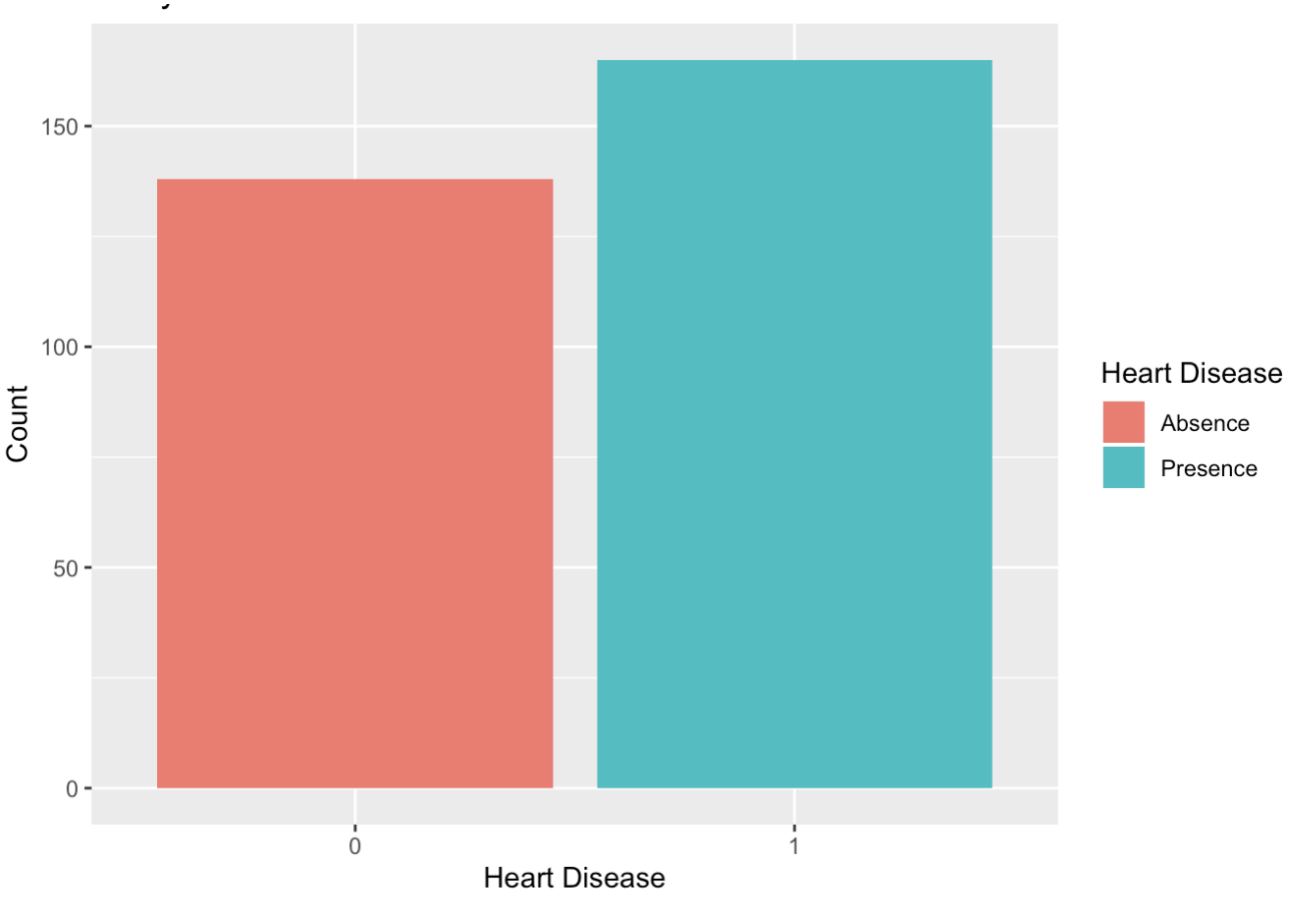


Figure 2.3 is the histogram of "target", the left red part represents the absence of heart disease and the right blue part show the presence of heart disease. The proportion of absence and presence of heart disease in our data set is .45 to .55. It indicates that there is no imbalanced value issue in the response variable.

Figure 2.4 The Distribution of Age Concerning Different Genders

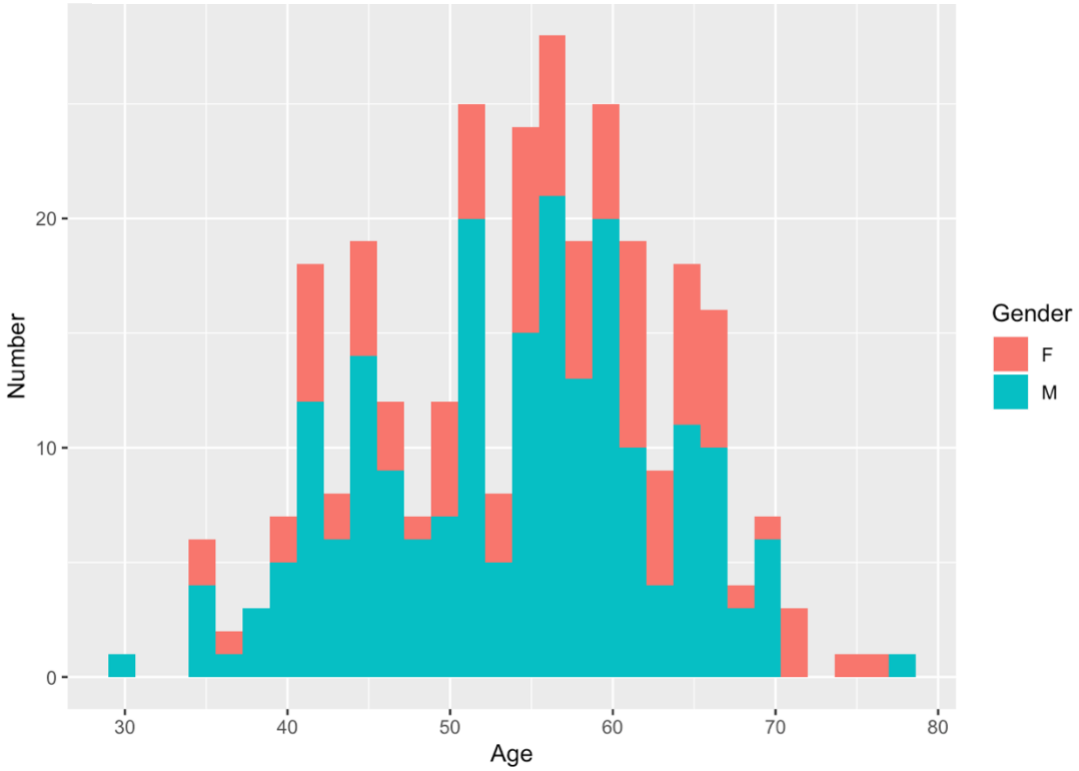


Figure 2.4 expresses that people aged from 45 to 65 are more concerned about heart disease than any other age groups. The possibility of having heart disease seems to be prevalent in people aged from 41-50 compared to that of 60 to 70. It may due to the fact that middle aged people are too concentrated on their work to ignore annual physical exam. In this increasingly competitive society, the burden of taking good care of families and heavy work pressure do not leave much space for people to consider themselves. They may come to the hospital only when they do not feel comfortable. Hence it results in the high prevalence rate in this age group. In addition, although there are more males participating in the test, from the proportion of red and blue we can infer that the possibility of having heart disease tend to be higher in females than in males.

Figure 2.5: Correlation of Five Numerical Variables

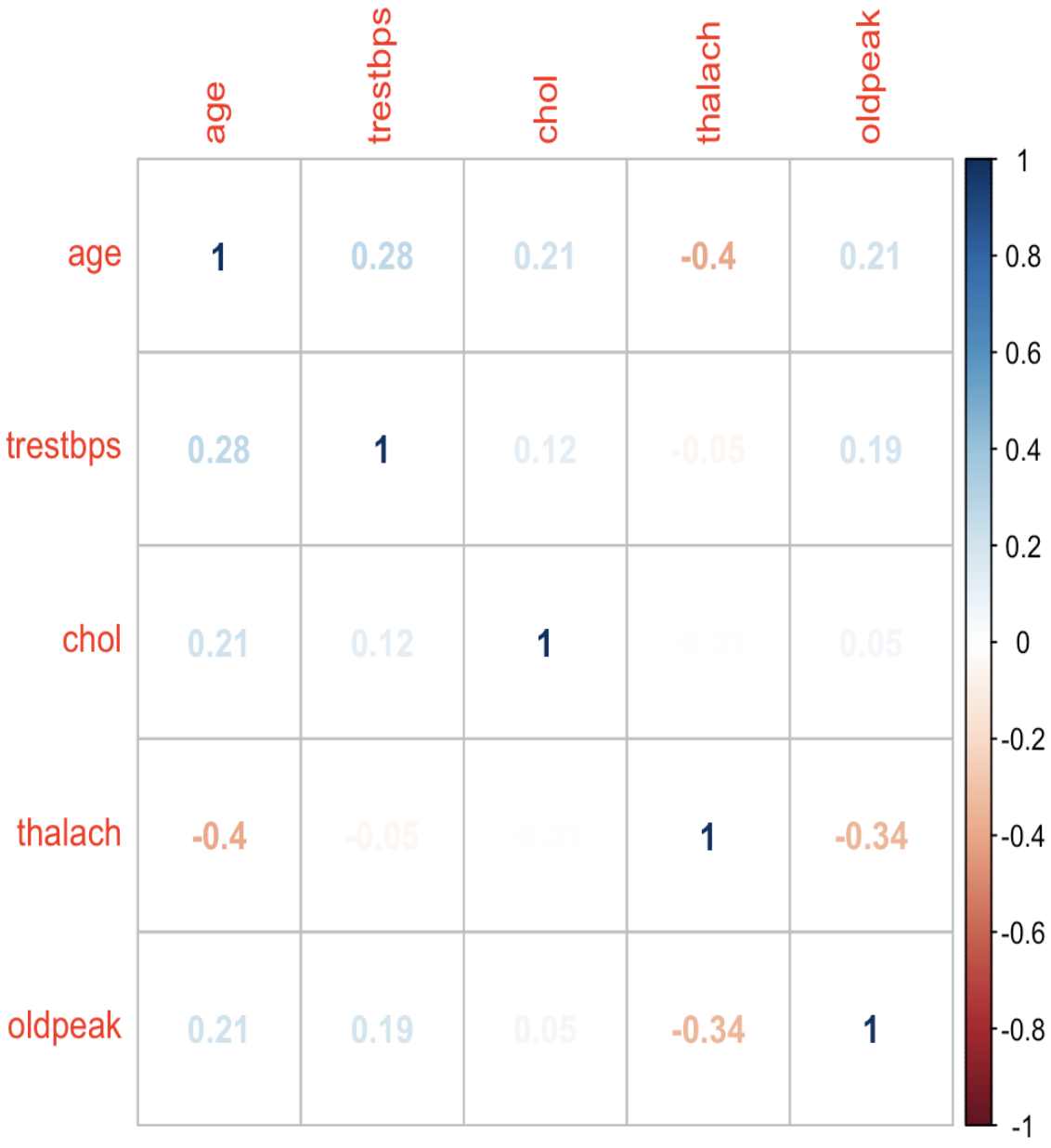


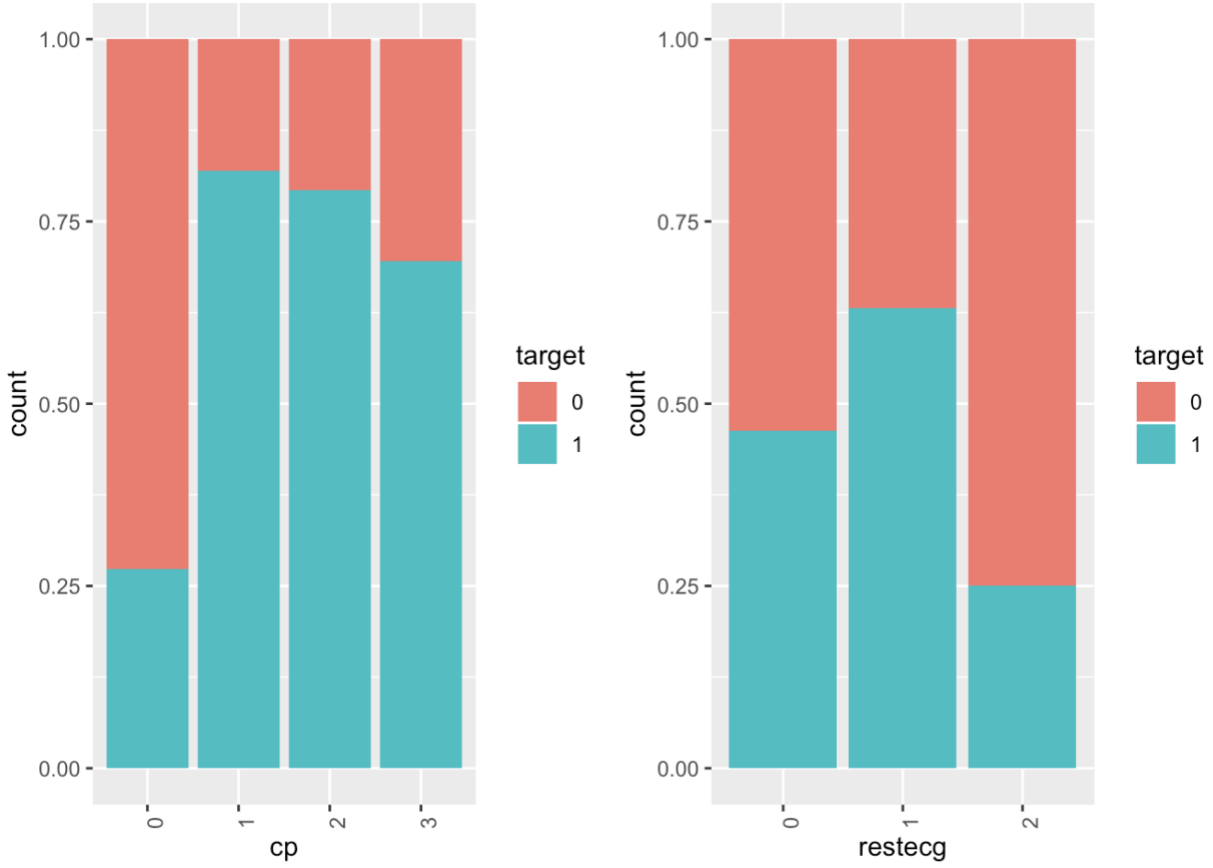
Figure 2.5 indicates the correlation of those 5 numerical variables (age, trestbps, chol, thalach and

oldpeak). We can see that age and thalach (maximum heart rate achieved) have slightly strong negative correlations. It may account for that when people get older, their physical condition may undergo some declines along with the heart rate. What's more, the correlation between thalach (maximum heart rate achieved) and oldpeak (ST depression induced by exercise relative to rest) is -0.34, it indicates that when we start to do more exercise, our heart rate will tend to increase and the ST depression will decrease.

Aside from the possible relationship, some of the variables do not seem to have much connection. For instance, in the picture we can also see that although thalach (maximum heart rate achieved) and trestbps (resting blood pressure) are all concerned with our heart, they do not seem to have any relationship since the correlation is only 0.05. At the same time, oldpeak (ST depression induced by exercise relative to rest) and chol (serum cholesterol in mg/dl) do not seem to have any connection with each other.

Last but not least, the relationship of some variables is not clear from simple visualization, we will try to evaluate their connection by using different models in the later part.

Figure 2.6: Possibility of Having Heart Disease Concerning cp and restecg



The left picture of Figure 2.6 illustrates the relationship between cp (chest pain) and target. When we suffer from typical angina, the possibility of having heart disease is much lower than those with atypical angina, non-anginal pain and asymptomatic condition. In common sense, we may take it for granted that when we come across typical angina, our heart must be responsible for that. However, many other diseases may account for chest pain, like pneumonia and other respiratory illness.

When it comes to the relationship between restecg (resting electrocardiographic results) and target, the prevalence in people who have ST-T wave abnormality seem to be much higher than those who are normal or show probable definite left ventricular hypertrophy by Estes' criteria.

Figure 2.7 Possibility of Having Heart Disease Concerning slope, ca, and thal

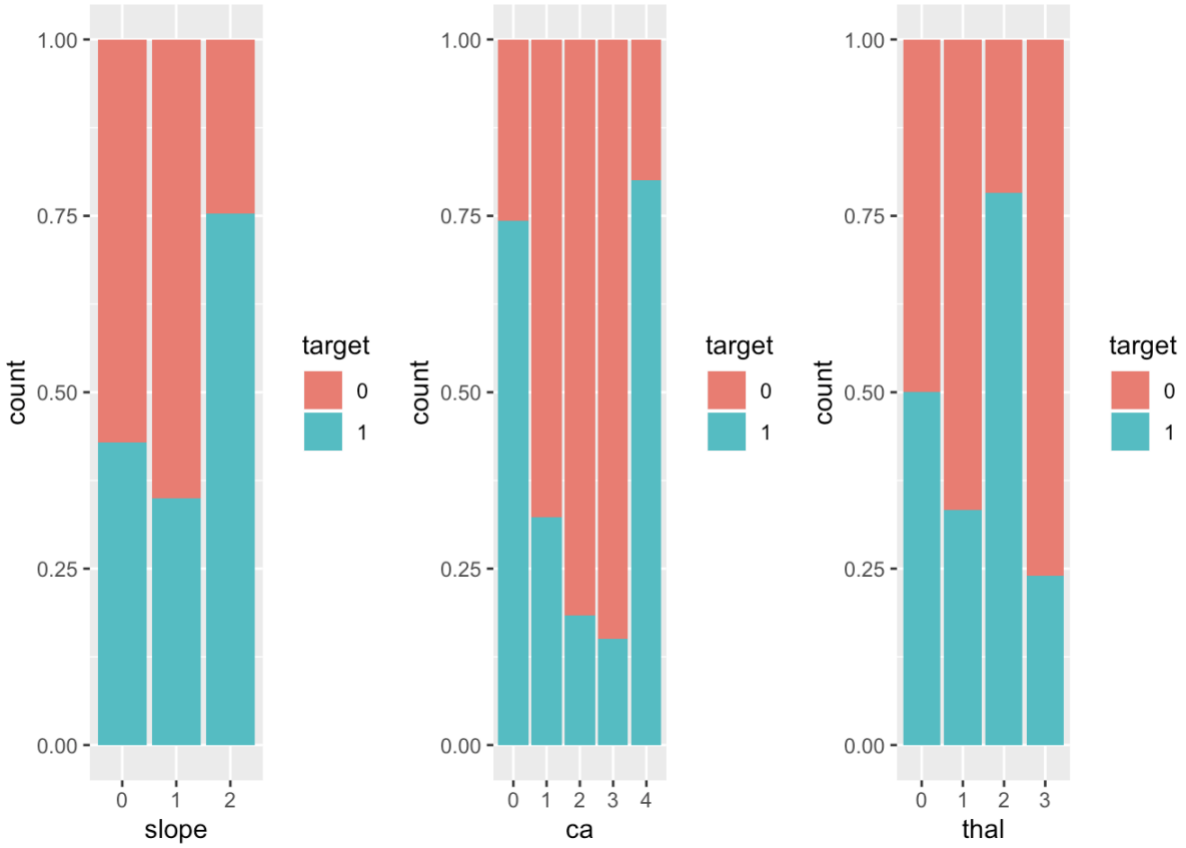


Figure 2.7 (left) indicates the corresponding relationship between slope and target. When we do exercise, the slope of ST segment may be upsloping, flat or downsloping. While the likelihood of having heart disease is normal for upsloping and flat, however the possibility get a huge climb when the slope is downward.

Figure 2.7 (middle) indicates the corresponding relationship between ca (number of major vessels (0-3) colored by fluoroscopy) and target. We may conclude that as the number of major vessels marked by fluoroscopy increases from 0 to 3, the possibility of having heart disease decreases. However, when the number of major vessels are large enough, like 4, the possibility of having heart disease has got a huge climb, even a little bit higher than the number equals to 0.

Figure 2.7 (right) indicates the corresponding relationship between thal and target. Through research we know that thal is strictly limited to the medical condition in local area. People in developed and resource-abundant countries can definitely get better treatment. However, cautious clinical management can also extend patients' life expectancy and increase living conditions. Fixed defect thal and reversable defect thal are closely connected with heart disease and diabetes.

Chapter 3

Methodology and Modeling

Before we do the modeling, we should do the cross-validation to protect against overfitting. I divided the data set into two parts with 80%-20% split. 80% is the training set, which is used to train the model, and the 20% is the testing set, which is used to validate it on data it has never seen before. Figure 3.1 is shown below to set the cross-validation.

Figure 3.1: Set Training and Testing Data Set

```
#set training and testing data set
smp_size <- floor(0.8 * nrow(data))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)
train <- data[train_ind, ]
test <- data[-train_ind, ]
```

3.1 Logistic Regression

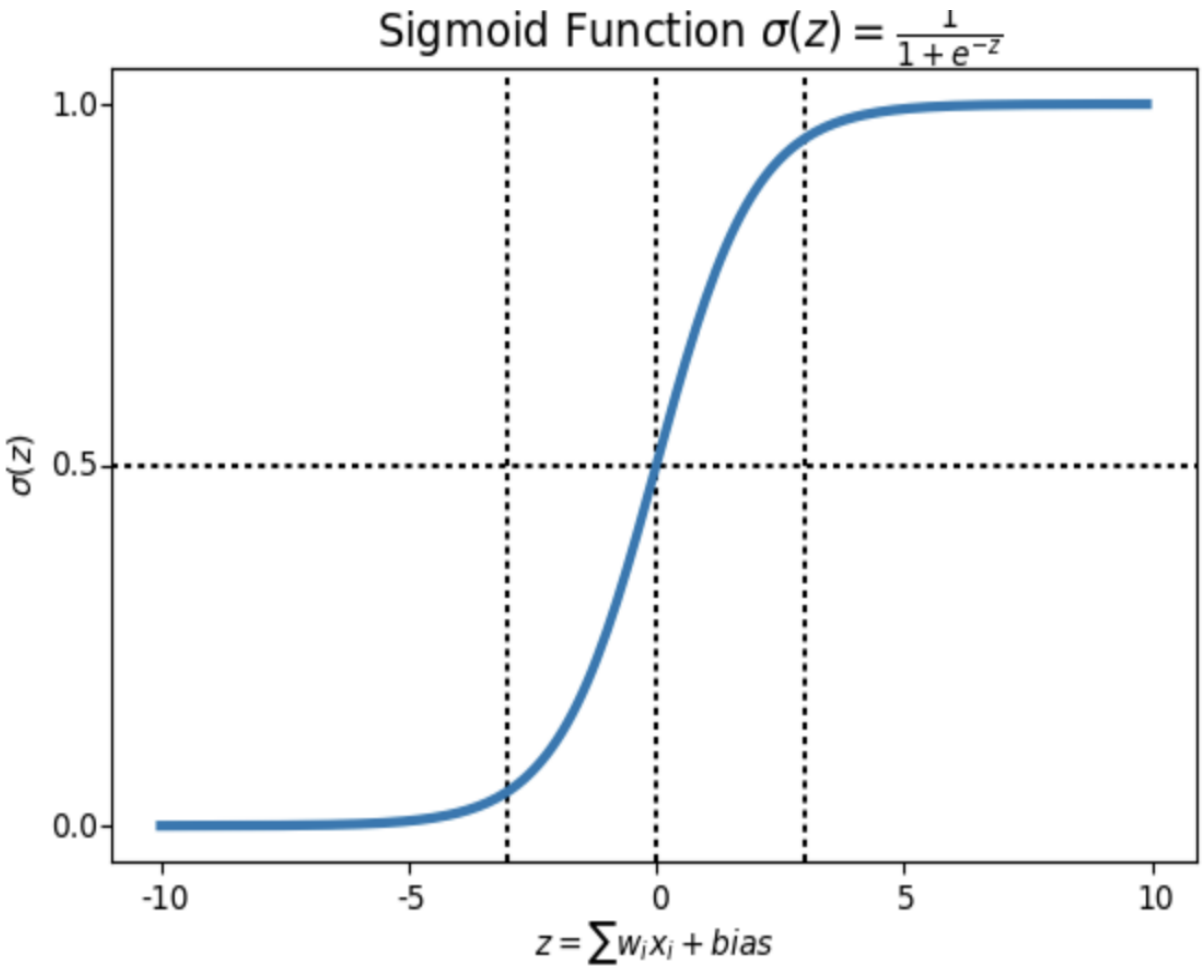
In supervised learning, categorical variables are modelled in classification. Logistic regression is an important machine learning algorithm which uses a logistic function to model a binary response variable. Since our response variable “target” is a categorical variable which has two levels 0 and 1, we can implement logistic regression algorithm. The logistic function is a sigmoid function, which takes

any real input t, ($t \in \mathbb{R}$), and outputs a value between 0 and 1; for the logit, this is interpreted as taking input log-odds and having output probability. The standard logistic function is defined as follows [2]:

$$\sigma(z) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-z}}$$

The following picture Figure 3.2 is the sigmoid function graph.

Figure 3.2: Graph of Sigmoid Function



From the perspective of our data, our response variable “Target” has two classes 0 (not having heart disease) and 1 (having heart disease). We should basically decide a threshold value 0.5, above which we classify values into having heart disease and of the value goes below the threshold then we classify it in not having heart disease.

For logistic regression, we also have our cost function. It is defined as follows:

$$J(\theta) = -\frac{1}{m}\sum[y^{(i)}\log\left(h\theta\left(x(i)\right)\right) + (1 - \log\left(1 - y^{(i)}\right)\log\left(1 - h\theta\left(x(i)\right)\right)]$$

Firstly, we put all of the variables into our model. We implement the logistic regression model and use the anova function to analyze the table of deviance.

Figure 3.3: Table of Deviance of Logistic Regression

```
#analyze the table of deviance
anova(logistic,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       241     333.83
## age       1    9.302       240     324.53  0.002289 **
## sex       1   25.242       239     299.29 5.058e-07 ***
## cp        3   67.912       236     231.37 1.195e-14 ***
## trestbps  1    4.669       235     226.70  0.030705 *
## chol      1    0.915       234     225.79  0.338760
## fbs       1    0.010       233     225.78  0.919176
## restecg   2    1.098       231     224.68  0.577452
## thalach   1   18.311       230     206.37 1.876e-05 ***
## exang     1    1.352       229     205.02  0.244882
## oldpeak   1   12.004       228     193.01  0.000531 ***
## slope     2    4.256       226     188.76  0.119091
## ca        4   31.481       222     157.28 2.442e-06 ***
## thal      3    8.393       219     148.88  0.038545 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After implementing the logistic regression, the difference between the null deviance and residual deviance keeps increasing as we add each variable one at a time. And in the end the null deviance is reduced from 333.83 to 148.88. It is clear that adding age, sex, cp trestbps, thalach, oldpeak, ca and thal significantly reduce the deviance.

Next we will briefly evaluate the fitting of the model. We would like to see how the model is doing when predicting target on the testing data set.

Figure 3.4: Accuracy of Logistic Regression for Test Data

```
#accuracy
fitted.results <- predict(logistic,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$target)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.852459016393443"
```

The accuracy of prediction is 0.852 on the testing data set, which indicates that about 85% of the test data has been predicted correctly by the logistic regression model. And this result is quite good.

Figure 3.5: Roc Plot

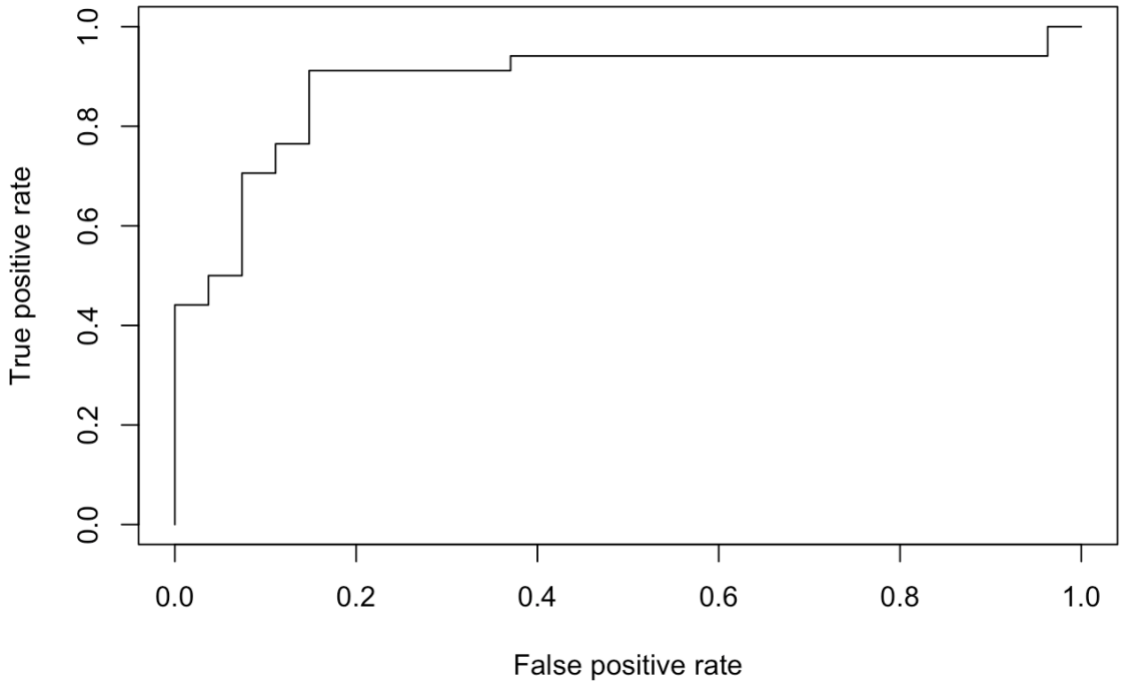


Figure 3.5 is a typical performance measurement for a binary classifier. The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. We also know that a model with good predictive ability must have an AUC closer to 1 (1 is ideal) , here the AUC is 0.89, which indicates the model is good.

After we have done the above, we sensed that some of the variables are not statistically significant, and this model may exist some irrelevant variables or cause the possibility of overfit. Therefore, we will try to exclude these variables using stepwise variable selection. In the end we have found the most important variables which have close relationship with the predicted variable target are sex, cp, trestbps, thalach and ca.

3.2 Random forest

Random forest in another important supervised machine learning algorithm. Before introducing the concept of random forest. Let's firstly try to understand the building blocks of random forest — decision trees [3].

A decision tree is a machine learning algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. Various branches of variable length are formed.

The goal of a decision tree is to encapsulate the training data in the smallest possible tree. The rationale for minimizing the tree size is the logical rule that the simplest possible explanation for a set of phenomena is preferred over other explanations. Also, small trees produce decisions faster than large trees, and they are much easier to look at and understand. [4]

Figure 3.6: Example of Decision Trees

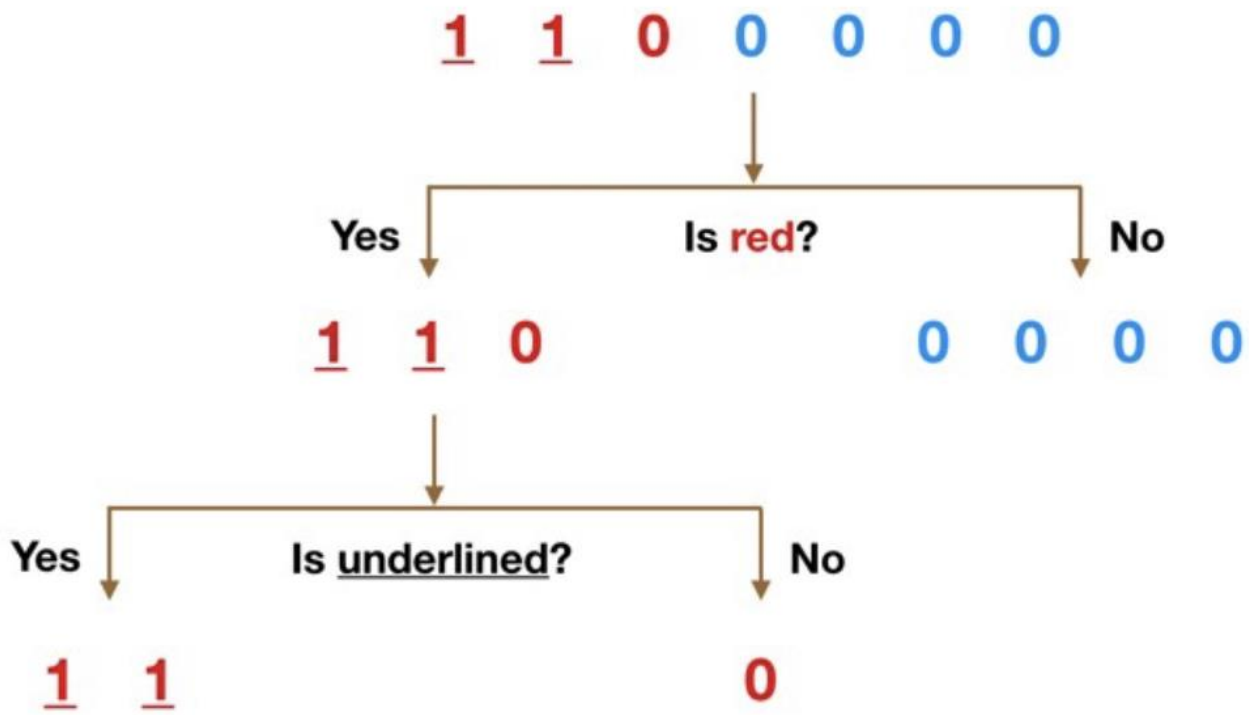


Figure 3.6 intuitively expresses that we first divide the observation by color, and then partition the observation by it is underlined or not.

Decision trees have many advantages. For instance, it has straightforward visualization, which is explicit for people to understand and interpret. However, building decision trees have some limitations simultaneously. This algorithm is too greedy, it requires algorithm to be able to determine an optimal choice at each node. It satisfies each step's optimal decision but sacrifices the global optimum at the same time. Furthermore, decision trees are prone to overfitting, especially when a tree is particularly deep.

While random forest mitigates this problem by training on different samples of the data d using a random subset of features. It consists of multitudes of individual decision trees that operates as an ensemble. Each individual tree in the random forest rules out a class prediction and the class with the most votes becomes our model's prediction. Random forest is an ensemble learning method which improves the drawbacks of decision trees' habit of overfitting to their training set.

Figure 3.7: Plot of Simplified Random Forest

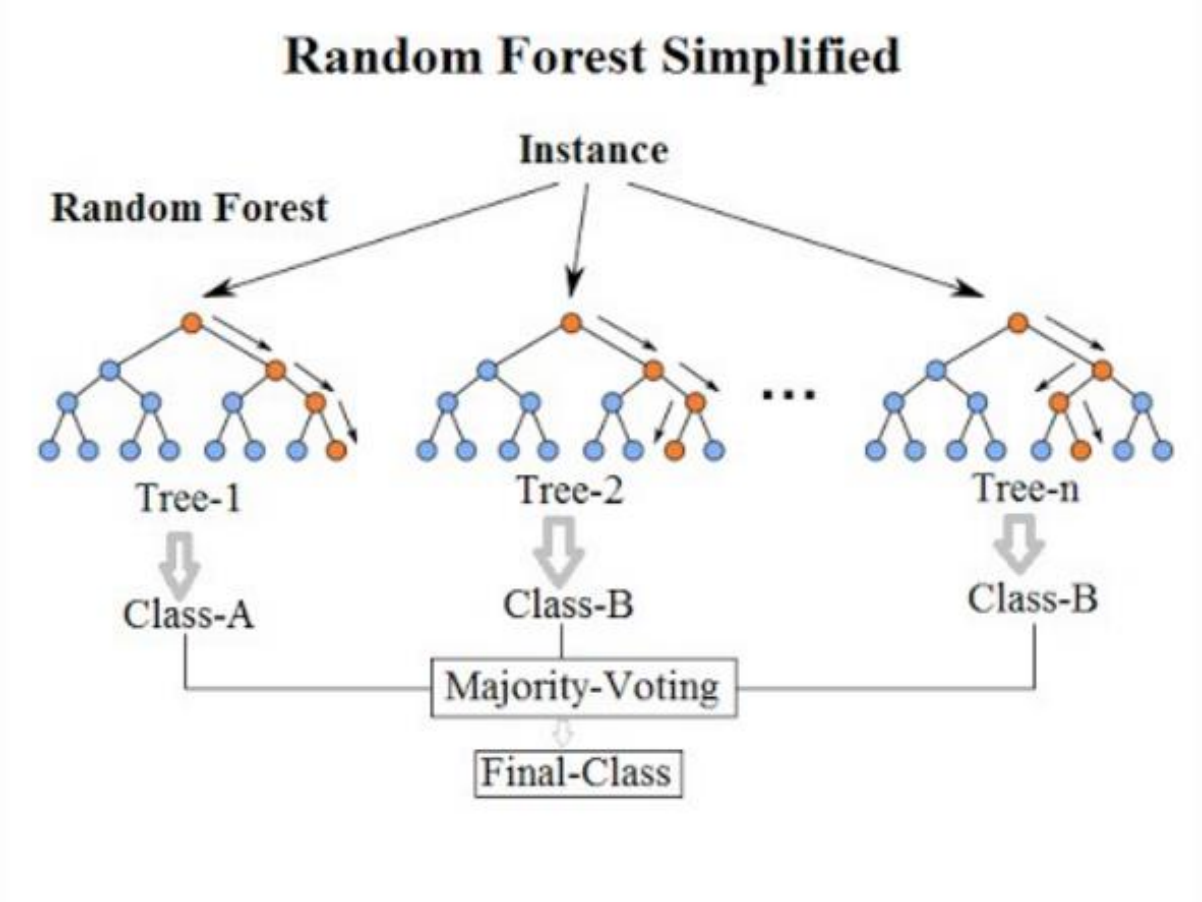


Figure 3.7 is a simplified flow chart of random forest which illustrate the process of each step.

There are two parameters in the function of random forest that we should pay attention to. First, OOB error (also called out of bag error), it calculated the prediction error using data not in bootstrap sample for each bootstrap iteration and related tree. Second, mtry is the number of variables available for splitting at each tree node. It makes random forest models differ to each other.

In our model, we should first use tuneRF function to determine the value of mtry number in order to lower the out of bag error.

Figure 3.8: OOB Error Against mtry

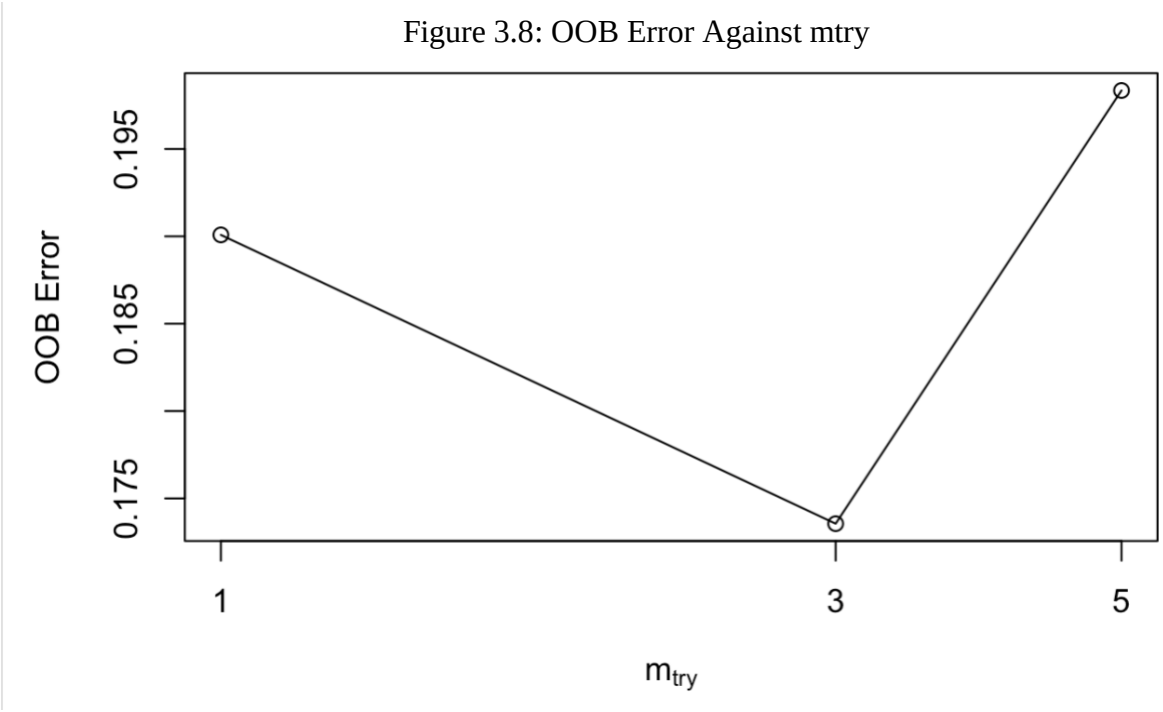


Figure 3.8 indicates that when we set mtry from 1 to 3, the OOB error decreases from 0.190 to the lowest point of about 0.15, and when we change mtry from 3 to 5, the OOB error keep increasing. Thus when we choose mtry=3 to be our parameter, we can get the lowest OOB Error.

Figure 3.9: Default OOB Error

```
randomForest(formula = target ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 18.6%
Confusion matrix:
   0   1 class.error
0 87  24   0.2162162
1 21 110   0.1603053
```

Figure 3.10: OOB Error After Optimization

```
randomForest(formula = target ~ ., data = train, mtry = 3, ntree = 200)
              Type of random forest: classification
                    Number of trees: 200
No. of variables tried at each split: 3

       OOB estimate of  error rate: 17.36%
Confusion matrix:
   0   1 class.error
0 90  21   0.1891892
1 21 110   0.1603053
```

Figure 3.9 and Figure 3.10 indicates that when we change the number of trees from 500 to 300, set the mtry value to 3, the OOB Error of the training data set decreased from 18.6% to 17.36%.

Figure 3.11: Confusion Matrix of Random Forest Model

```
confusionMatrix(predtest, test$target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 21  1
##          1  6 33
##
##                Accuracy : 0.8852
##                  95% CI : (0.7778, 0.9526)
##     No Information Rate : 0.5574
##     P-Value [Acc > NIR] : 3.37e-08
##
```

Figure 3.11 is the confusion matrix of the test data set of our random forest model. There are 33 true positive values and 21true negative values. While the total number of values is 61, the accuracy classification score is 0.88, which indicates that about 88% of the test data has been correctly predicted by our random forest model based on the training data set.

Figure 3.12: Variable Importance of the Random Forest Model

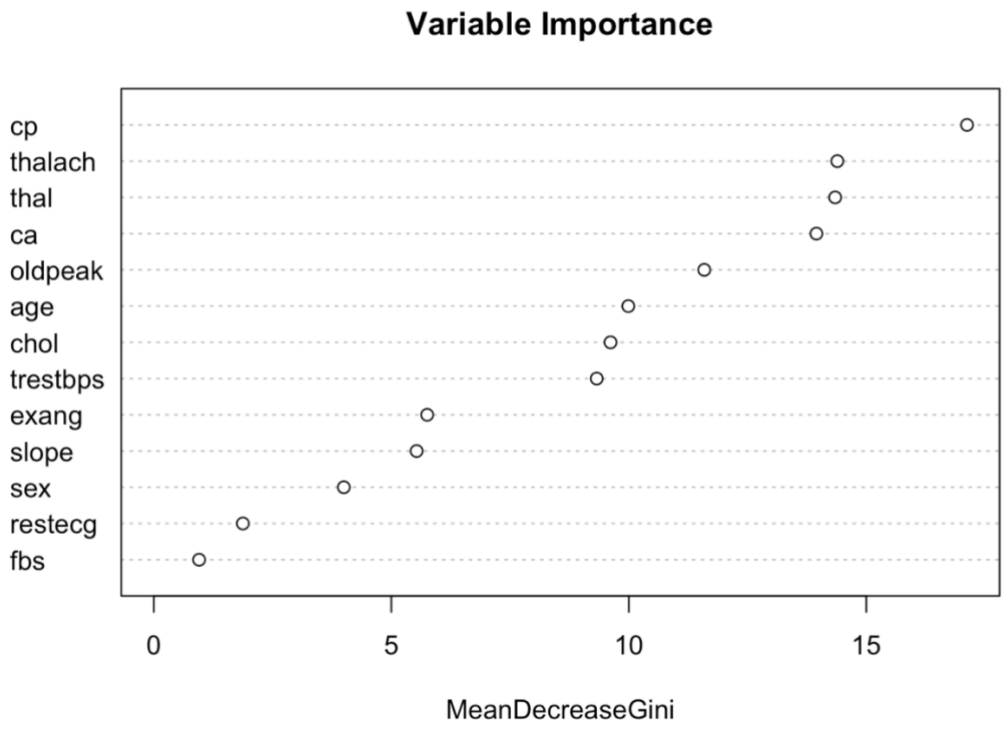


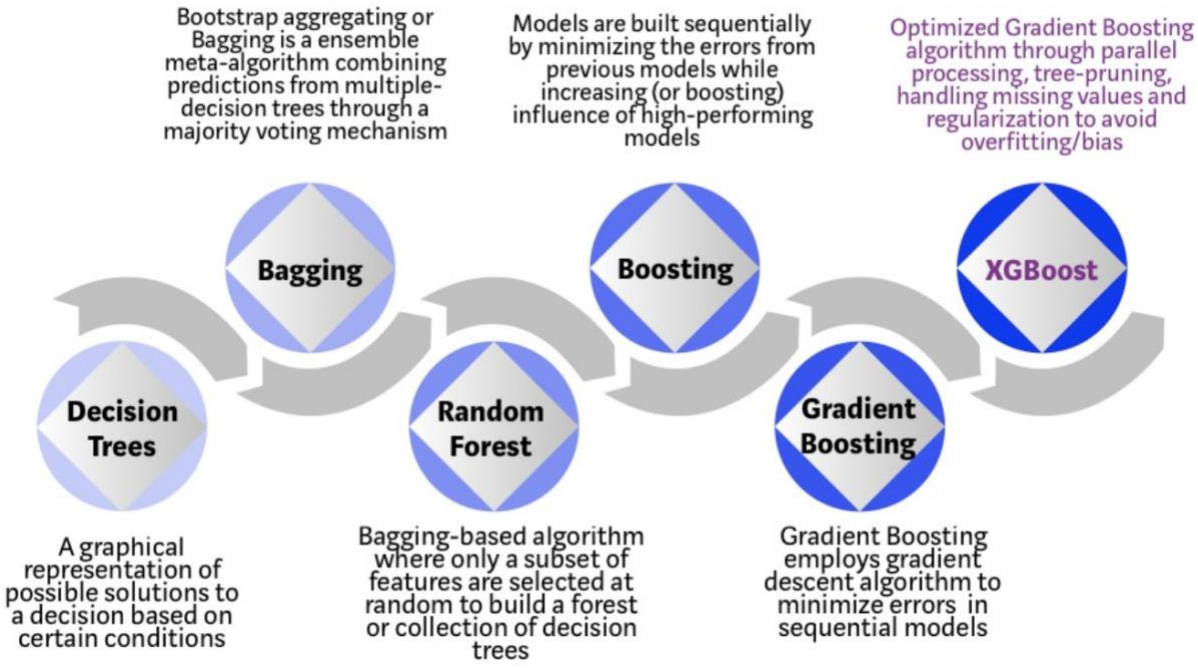
Figure 3.12 gives a clear visualization of which feature plays the most important role in predicting the response variable “Target”. We can see that cp, thalach, thal, ca and oldpeak are the top 5 crucial variables.

3.3 Extreme Gradient Boosting (XG Boost)

The extreme gradient boosting algorithm is fast, flexible, versatile and accurate among multiple machine learning algorithms. It was developed by Tianqi Chen and now is part of a wider collection of open-

source libraries developed by the Distributed Machine Learning Community. It is also a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework [5].

Figure 3.13: Evolution of Extreme Gradient Boosting Algorithm from Decision Trees



There are two parameters that we should pay attention to when using the extreme gradient boosting. The first parameter is gamma, since larger values indicates more conservative algorithm. The second parameter is subsample, since when we choose lower values, it can help us to prevent the problem of overfitting.

Figure 3.14: Error of the Extreme Gradient Boosting Model

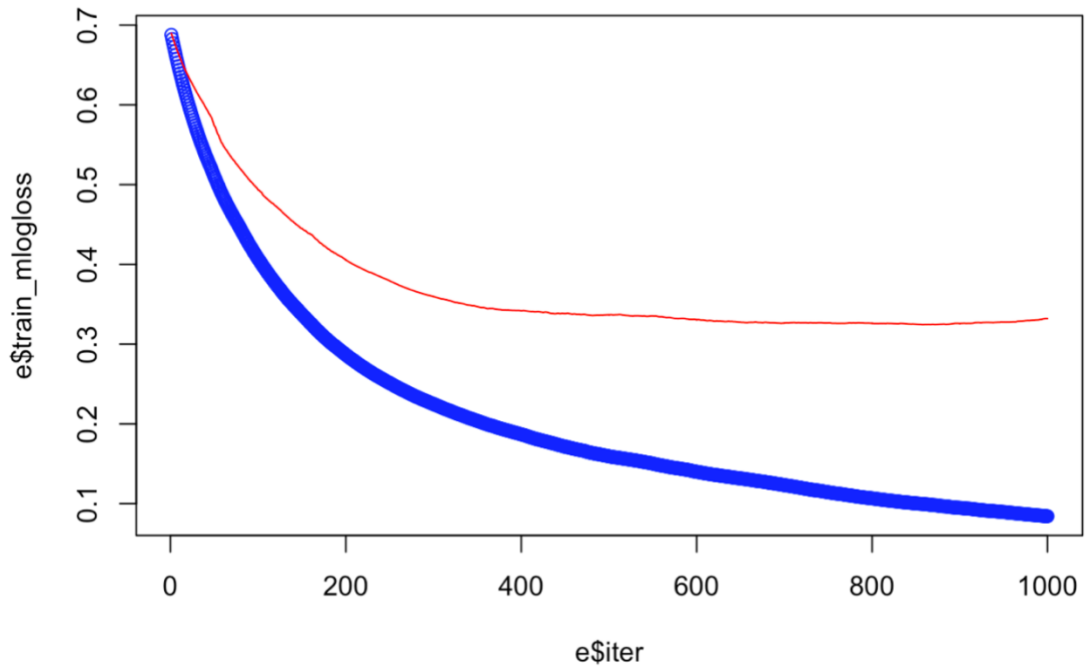


Figure 3.14 gives us a visualization of how the error of training data set and testing data set decreases as the number of iterations increase. While the blue curve indicates the error of the training data set, and the red curve shows the error of the testing data set. It is clear that when the iterations are less than 500, the error of the testing data set and training data set both go through a huge drop. And when the iterations come to 500 or more, the error of the training data set keeps decreasing while the error of testing data set keeps stable as 0.32.

Figure 3.15: Confusion Matrix of Extreme Gradient Boosting Model

```
##            Actual
## prediction  0  1
##          0 20  3
##          1  7 31
```

Figure 3.15 is the confusion matrix of the extreme gradient boosting model. The total test data is 61. There are 31 true positive values and 20 true negative values. The accuracy classification score is 51/61=0.836, which indicates that about 84% of the test data has been correctly predicted by extreme boosting model based on the training data set.

Figure 3.16 Variable Importance of Extreme Gradient Boosting Model

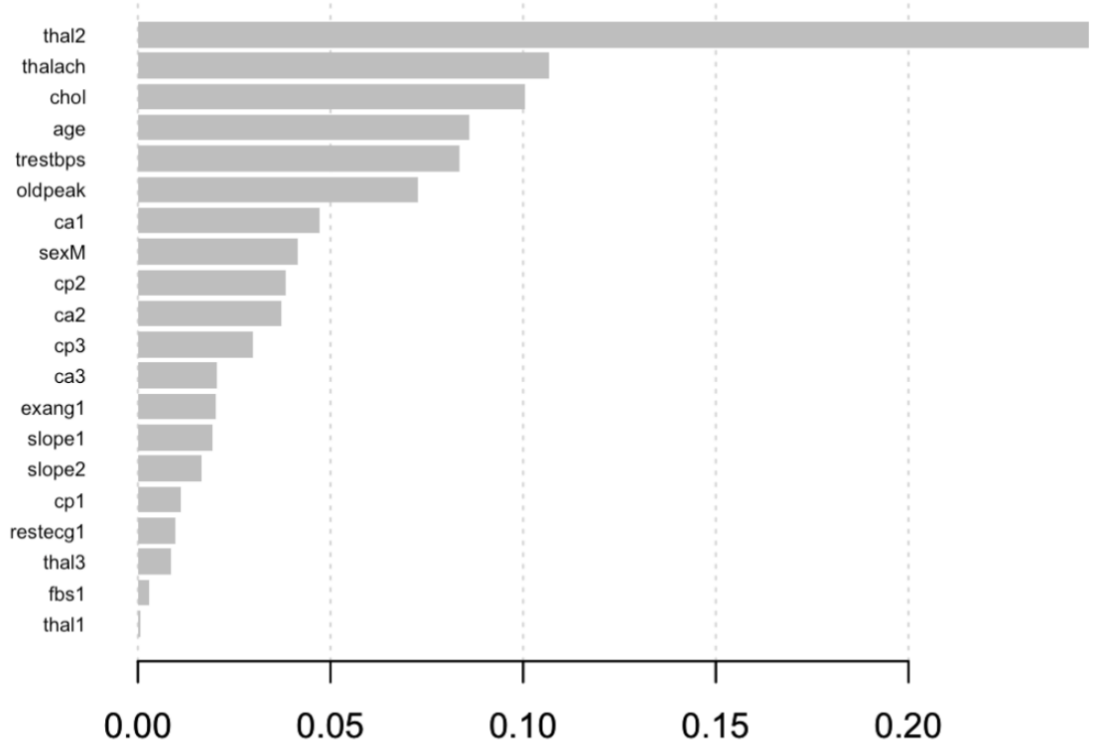


Figure 3.16 helps us to visualize the important features that play a crucial role in predicting the dependent variable “Target”. It is evident that thal2 has the most important role in predicting the output. The following comes with thalach, chol, age and trestbps.

3.4 Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. A “neuron” in a neural network

is a mathematical function that collects and classifies information according to a specific architecture. The network bears a strong resemblance to statistical methods such as curve fitting and regression analysis. A neural network contains layers of interconnected nodes. The layers can be single or multiple. In a single-layered neural network, each node is a perceptron and is similar to a multiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear. In a multi-layered perceptron (MLP), perceptrons are arranged in interconnected layers. The input layer collects input patterns. The output layer has classifications or output signals to which input patterns may map [6].

Figure 3.17 Basic Neural Network Layout

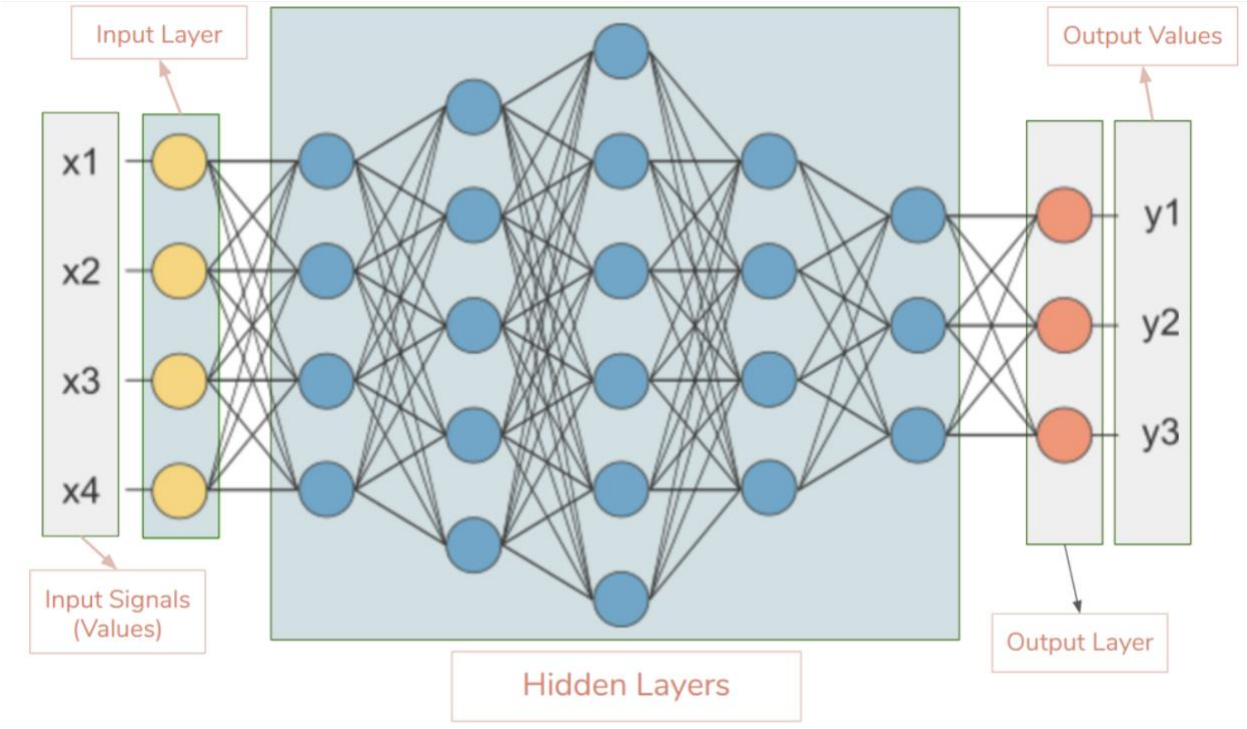


Figure 3.17 illustrate an easy and clear process of neural network for us to understand.

In order to apply the neural network model. We should firstly try to convert and normalize the variables,

making the range lie between 0 and 1. Secondly, we should also make an 80%-20% split to the data set to do cross validation and protect against overfitting.

There are several parameters in the “neuralnet” function that we should pay much attention. First of all, the number of nodes in the hidden layer, using too few neurons in the hidden layers will result in underfitting while using too many neurons in the hidden layers may result in overfitting. We may also have two hidden layers in the function, but it may cause the problem that algorithm does not converge and weights are not calculated. Secondly, “lifesign” specifies how much the function will print during the calculation of the neural network, which allows you to monitor the value of error function after each step. Thirdly, we can also use “rep” to set the number of repetitions for the neural network’s training in order to optimize the model by choosing the converged repetition that has the lowest error.

After multiple trials, I have found that when we set 5 nodes in the hidden layer, “full” in lifesign and fourth in the repetition in the model, misclassification is the lowest. Figure 3.18 is the plot of the neural network model when we set the above parameters.

Figure 3.18: Plot of Neural Networks Model

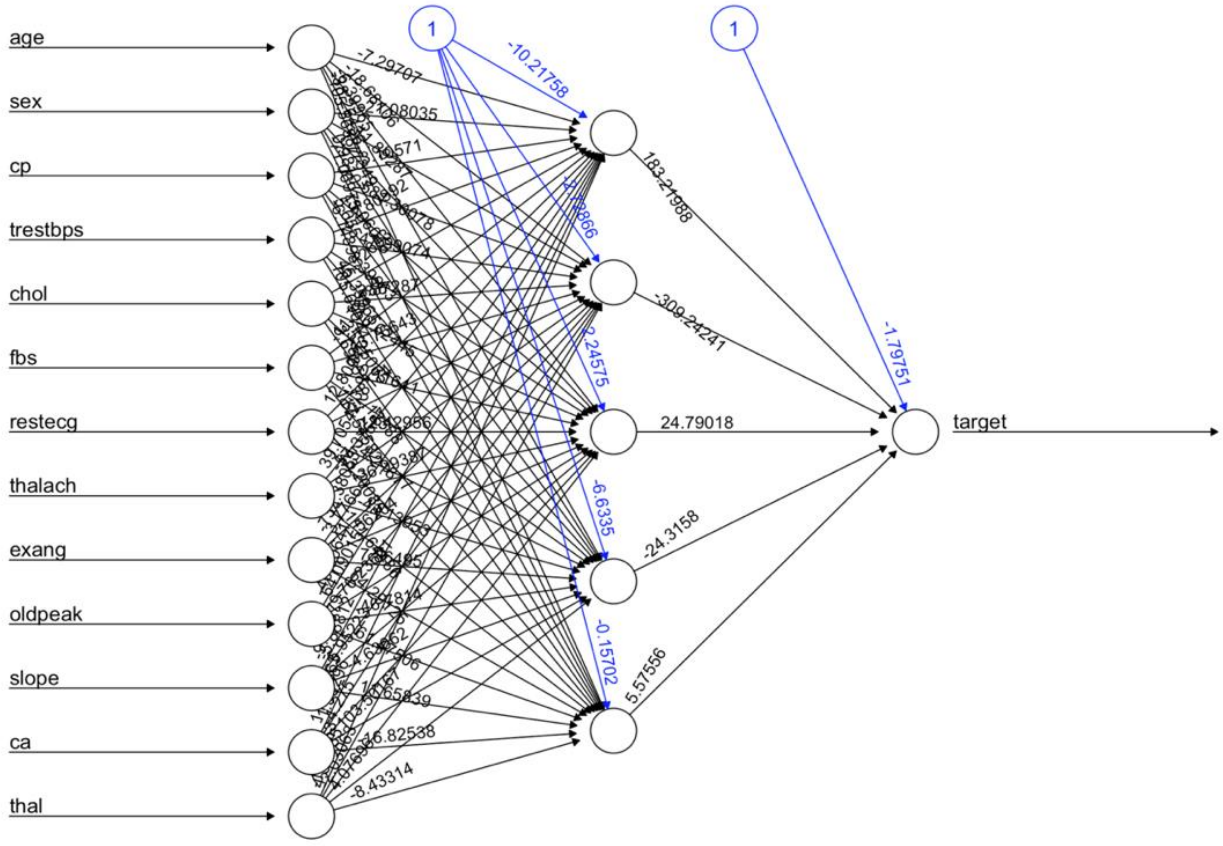


Figure 3.19: Confusion Matrix for training set

```
##
## pred1   0   1
##     0 110   4
##     1   1 127
```

Figure 3.20: Confusion Matrix for testing set

```
##
## pred2  0  1
##     0 21  3
##     1  6 31
```

Figure 3.19 and Figure 3.20 illustrate the confusion matrix for both the training data set and testing data set. For the training data set, it has 127 true positive values and 110 true negative values. The accuracy is around 98% and misclassification is 2%. On the other hand, for the testing data set, it has 31 true positive values and 21 true negative values. The accuracy is around 85% and the misclassification is

around 15%, which means about 85% of the test data set has been correctly predicted by the neural network model based on the training data set.

Chapter 4

Conclusion

4.1 Conclusion

Table 4.1: Accuracy of the four models of test data set

Models	Accuracy
logistic regression	0.852
random forest	0.885
XGBoost	0.836
neural network	0.852

If we compare the model through the perspective of accuracy, we can see that the accuracies of the four models of the testing data set are all above 80%, which is quite good.

Although the accuracy of random forest model is a little bit higher than that of the extreme gradient boosting model, extreme gradient boosting is a more advanced and robust algorithm. This is due to the fact that random forest builds each tree independently but extreme gradient boosting builds one tree at a time. Furthermore, logistic regression is the easiest model that we have used in this research paper. Hence the result of extreme gradient boosting and neural network may be more convincing.

Since difference between different models still exists, so there is so much more than the overall accuracy to investigate, more facts to consider and further research to do.

If we compare the model through the importance of variable. (from large to small)

Logistic Regression: cp, ca, sex, trestbps, thalach

Random Forest: cp, thalach, thal, ca, oldpeak

Extreme Gradient Boosting: thal, thalach, chol, age, trestbps, oldpeak

If we choose any two of the three models and compare them with each other, we can find that they share most of the variables and have little exceptions.

There are two variables playing significant roles in predicting the dependent variable “Target”. That is “trestbps” (resting blood pressure) and “thalach” (maximum heart rate achieved).

Usually we use blood pressure readings to analyze and monitor blood pressure. These tests record blood pressure using two measurements: systolic and diastolic blood pressure. According to the research, both high systolic and high diastolic blood pressure can lead to heart attack and stroke [8].

On the other hand, I have also found many reports said that a high heart rate was associated with a higher risk of all-cause mortality and cardiovascular events. It has also been proven that the relationship tends to be stronger in women than men [7]. While we do the exploratory data analysis, Figure 2.4 also conforms with this result.

Some people may tend to believe that blood pressure has nothing to do with heart rate. However, the fact is that the acceleration of heart rate is closely correlated with the increasing risk of the cardiovascular system. The heart rate for normal person is 60 to 100 times one minute. The rise of heart rate mostly results from sympathetic activation, which has reciprocal causation with high blood pressure. In a nutshell, when we have a racing heart, the risk of hypertension is higher. On the other hand, high blood pressure can also result in functional damage to our heart, and further accelerates our heart rate.

Aside from the two common variables, we should also pay much attention to cp (chest pain), thal and ca (number of major vessels colored by fluoroscopy).

Chest pain, also called angina, is caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion. Angina can also be a symptom of coronary microvascular disease. This is heart disease that affects the heart's smallest coronary arteries and is more likely to affect women than men [9]. This research confirms what we have found in exploratory data analysis in chapter 2. Hence when we feel angina, we may first take heart disease into consideration and go to the hospital as soon as possible.

While thal (thalassemia), is an inherited blood disorder that causes your body to have less hemoglobin than normal. Thalassemia can cause anemia, leaving you fatigued. Untreated anemia, in which hemoglobin levels are consistently too low, often causes tachycardia, as the heart tries to compensate for the lack of oxygen being carried through the body. Over time, the heart also becomes enlarged [10]. Hearts are unavoidably affected by thalassemia. Hence when people are diagnosed with thalassemia, they may check their body regularly to prevent emergent problems with heart.

When it comes to the X-ray fluoroscopy, we cannot deny its magical function to guide vascular and cardiac interventions, especially in real-time imaging and easy access to patients during interventions. But each coin has two sides, X-ray fluoroscopy is limited for defining soft tissue and obtaining functional information, regardless of the evidence that exposure to ionizing radiation from X-ray procedures is associated with an increased risk of cancer [11]. People may be more cautious with X-ray fluoroscopy, take advantage of its major function and keep updating the technology. And I believe in the foreseeable future, X-ray fluoroscopy will be more flexible, accurate and safe for people to use.

The good news, however, is that 80% of premature heart attacks and strokes are preventable. Healthy diet, regular physical activity, and not using tobacco products are the keys to prevention. Checking and controlling risk factors for heart disease and stroke such as high blood pressure, high cholesterol and high blood sugar or diabetes is also very important.

We can also take the following actions to prevent heart disease.

1. Eat a healthy diet.
2. Take regular physical activity.
3. Avoid tobacco use.
4. Check and control your overall cardiovascular risk [12].

4.2 Further Enhancement

Even though I spare no effort to make a complete and thorough research, there are still possible improvements for me in further analysis. For instance, in chapter 3.3, Figure 3.12 I list the evolution and development of extreme gradient boost Algorithm from Decision Trees. I may try all the other unused algorithms (decision trees, bagging, boosting, gradient boosting) to compare the results between each other.

Last but not lease, I only use 303 observations to do the analysis and I do not know if there exists some bias in the prediction. Next time I may choose other similar data set that have more observations to compare the results. What's more, since the mortality rate is much higher for people in some low-and-middle countries, I may set country as another important variable.

Reference

[1] "Cardiovascular Diseases (Cvds)". Who.Int, 2020, https://www.who.int/zh/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] "Logistic Regression". En.Wikipedia.Org, 2020, https://en.wikipedia.org/wiki/Logistic_regression.

[3] "Understanding Random Forest". Medium, 2020, https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[4] "Explanation Of The Decision Tree Model". Webfocusinfocenter.Informationbuilders.Com, 2020, https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/DecisionTree47.htm

[5] "Xgboost Algorithm: Long May She Reign!". Medium, 2020, https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

[6] "Neural Network Definition". Investopedia, 2020, https://www.investopedia.com/terms/n/neuralnetwork.asp.

[7] Perret-Guillaume C, et al. "Heart Rate As A Risk Factor For Cardiovascular Disease. - Pubmed - NCBI". Ncbi.Nlm.Nih.Gov, 2020, https://www.ncbi.nlm.nih.gov/pubmed/19615487.

[8] "Both Blood Pressure Numbers May Predict Heart Disease". Medicalnewstoday.Com, 2020, https://www.medicalnewstoday.com/articles/325861.

[9] "Angina (Chest Pain)". Www.Heart.Org, 2020, https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain.

[10] 2020, http://cooleysanemia.org/updates/Cardiac.pdf. Accessed 14 Mar 2020.

[11] Saeed, M., Hetts, S., English, J., & Wilson, M. (2012, January). MR fluoroscopy in vascular and cardiac interventions (review). Retrieved March 14, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3275732/

[12] "What Can I Do To Avoid A Heart Attack Or A Stroke?". World Health Organization, 2020, https://www.who.int/features/qa/27/en/.