Heart Disease Prediction

Executive Summary

This report describes the analyses conducted to predict whether patients have heart disease or not. Data used is the UCI Heart Disease dataset obtained from Kaggle - (https://www.kaggle.com/ronitf/heart-disease-uci).

The dataset contained 13 independent variables and one binary target variable. Among the independent variables, there were 8 categorical and 5 continuous variables.

First, exploratory analyses were conducted to get a feel for the data and visualize the differences between patients who have heart disease and those who don't on each of the independent variables.

Then, the dataset was split into train and test sets and a logistic regression model was built and improved using stepwise backward elimination. Predictions were made on the test set and performance of the model was evaluated. Performance measures used are AUC, accuracy, sensitivity and specificity. The receiver-operator characteristics curve was also plotted.

The AUC obtained with the logistic regression model was: **0.933**
Other performance measures were -
Accuracy: **0.902**
Sensitivity: **0.8929**
Specificity: **0.9091**

Analysis

The heart disease dataset contains the following independent variables -
1. age: age of the patient 2. sex: sex of the patient 3. cp: chest pain type 4. trestbps: resting blood pressure 5. chol: serum cholestoral in mg/dl 6. fbs: fasting blood sugar > 120 mg/dl 7. restecg: resting electrocardiographic results (values 0,1,2) 8. thalach: maximum heart rate achieved 9. exang: exercise induced angina 10. oldpeak: ST depression induced by exercise relative to rest 11. slope: the slope of the peak exercise ST segment 12. ca: number of major vessels (0-3) colored by flourosopy 13. thal: normal, fixed defect, reversable defect

Structure of the dataset is as follows -

```
str(data)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
##  $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
```

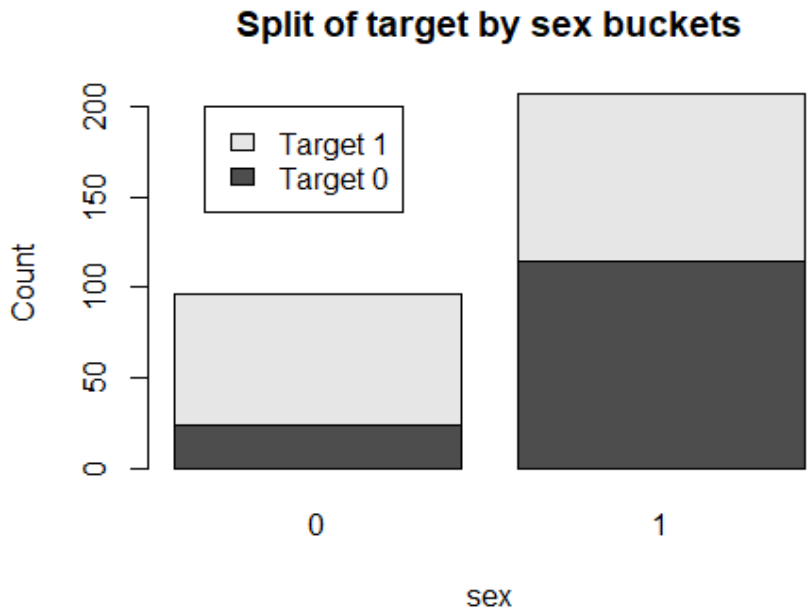
```
##  $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
##  $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
##  $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exang   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope   : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
##  $ ca      : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ thal    : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
##  $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Analysis of the heart disease dataset is split into two parts -
1. Exploratory Analyses
2. Predictive Modeling

In the exploratory analyses part, it is first ensured that there are no missing values in the data. Then, summary statistics of all the variables are analyzed. Bivariate analyses between the independent and target variables are conducted and plotted. Specifically, for categorical independent variables, a barplot showing the split of 'target' is shown, while for continuous independent variables, a frequency histogram showing the difference in distributions for the two 'target' categories is shown. Examples are shown below -

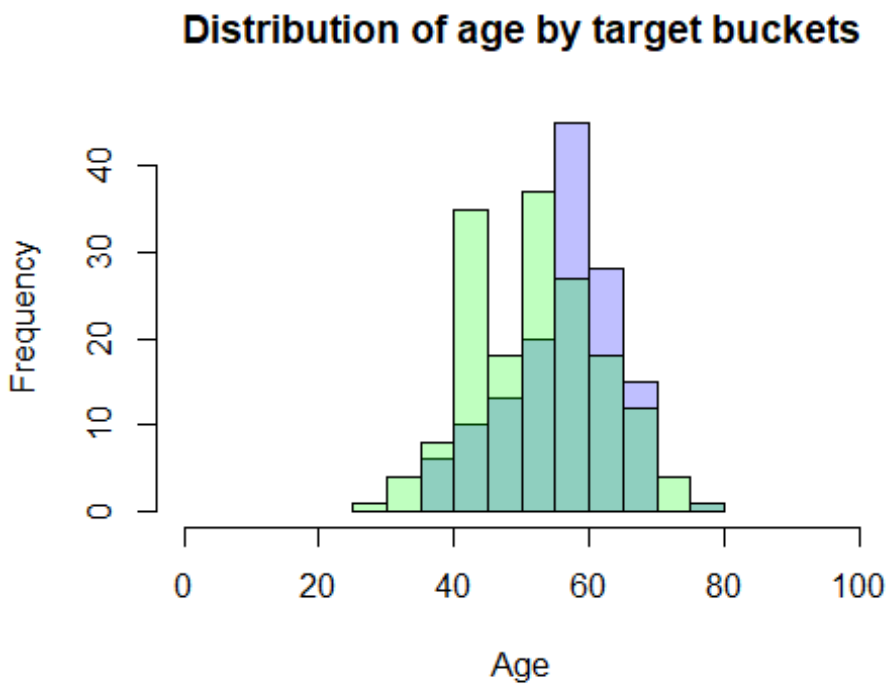
Barplot for categorical variable *sex*:

```r
barplot(table(data$target, data$sex),
        main = 'Split of target by sex buckets',
        xlab = 'sex', ylab = 'Count',
        legend.text = c('Target 0', 'Target 1'), args.legend = c(x=1,y=200))
```



Histogram for continuous variable *age*:

```
p1 <- hist(data$age[data$target==0])

p2 <- hist(data$age[data$target==1])

plot(p1, col=rgb(0,0,1,1/4), xlim=c(0,100), main='Distribution of age by
target buckets', xlab='Age')
plot(p2, col=rgb(0,1,0,1/4), xlim=c(0,100), add=T)
```



In the predictive analysis part, the dataset is first split into train and test sets such that a random 20% of the data is captured in the test set and the rest are used to train the model.

```
set.seed(1)
test_index <- createDataPartition(y = data$target, times = 1, p = 0.2, list =
FALSE)
train_data <- data[-test_index,]
test_data <- data[test_index,]

nrow(train_data)

## [1] 242

nrow(test_data)

## [1] 61
```

It is recognized that this is a binary classification problem and the logistic regression model is chosen. Stepwise backward elimination method is used to select variables. The selection criteria is AIC (Akaike Information Criteria) and p-values are used to detect insignificant variables at each step.

```
model <- glm(target~., data = train_data, family = binomial(link = 'logit'))
select_vars_model <- step(model, trace=0)
summary(select_vars_model)
```

```
##
## Call:
## glm(formula = target ~ sex + cp + trestbps + thalach + exang +
##     slope + ca + thal, family = binomial(link = "logit"), data =
train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6734  -0.3770   0.1367   0.4536   3.2581
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.40072    3.49537  -0.115 0.908729
## sex1          -1.85073    0.58086  -3.186 0.001442 **
## cp1            1.04779    0.64792   1.617 0.105843
## cp2            1.79952    0.54180   3.321 0.000896 ***
## cp3            1.58083    0.71096   2.223 0.026182 *
## trestbps      -0.01932    0.01132  -1.707 0.087819 .
## thalach        0.02115    0.01131   1.870 0.061505 .
## exang1        -0.78298    0.47465  -1.650 0.099024 .
## slope1        -0.33714    0.76959  -0.438 0.661327
## slope2         1.18561    0.78069   1.519 0.128843
## ca1           -2.18384    0.55161  -3.959 7.53e-05 ***
## ca2           -3.18865    0.74082  -4.304 1.68e-05 ***
## ca3           -2.92676    1.21591  -2.407 0.016082 *
## ca4           15.29159 1002.45099   0.015 0.987829
## thal1          2.05696    2.85311   0.721 0.470936
## thal2          1.68244    2.75612   0.610 0.541572
## thal3          0.50415    2.76643   0.182 0.855394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 333.48  on 241  degrees of freedom
## Residual deviance: 158.54  on 225  degrees of freedom
## AIC: 192.54
##
## Number of Fisher Scoring iterations: 15
```

Lastly, the trained logistic regression model is used to make predictions on the test set. ROC curve is plotted and AUC (area under the curve) is calculated to measure performance. Also, probability threshold of 0.5 is set, confusion matrix is viewed alongside key performance measures like Sensitivity and Specificity.

```
test_predicted <- predict(select_vars_model, test_data, type='response')

# ROC curve and AUC
ROCRpred = prediction(test_predicted,test_data$target)
ROCRperf = performance(ROCRpred, "tpr", "fpr")
plot(ROCRperf, main='ROC curve')

auc <- attributes(performance(ROCRpred, 'auc'))$y.values[[1]]

# Confusion Matrix for probability threshold of 0.5
predClass <- as.factor(ifelse(test_predicted>=0.5,1,0))
con_mat <- confusionMatrix(test_data$target, predClass)
```

Results

In the preprocessing stage, it was observed that there are no missing values in the data and it is useable as is.

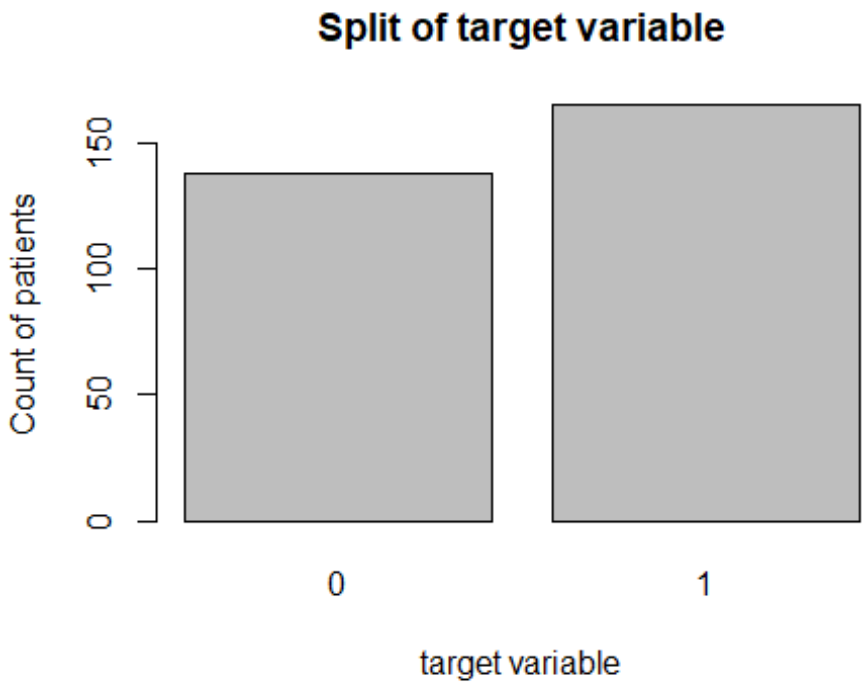
```
# Check if there are any blank values in the dataset
sapply(data, function(x) sum(is.na(x))) # No NA values

##      age      sex       cp trestbps     chol      fbs  restecg  thalach
##        0        0        0        0        0        0        0        0
##    exang  oldpeak    slope       ca     thal   target
##        0        0        0        0        0        0
```

In the exploratory analysis stage, summary statistics of the variables were obtained to see how they were distributed. It was observed that there was no major class imbalance in the target variable as 165 instances corresponded to heart disease and 138 corresponded to no heart disease.

```
# View summary of data
summary(data)

##       age        sex     cp         trestbps          chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   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
##
##
##
##
```

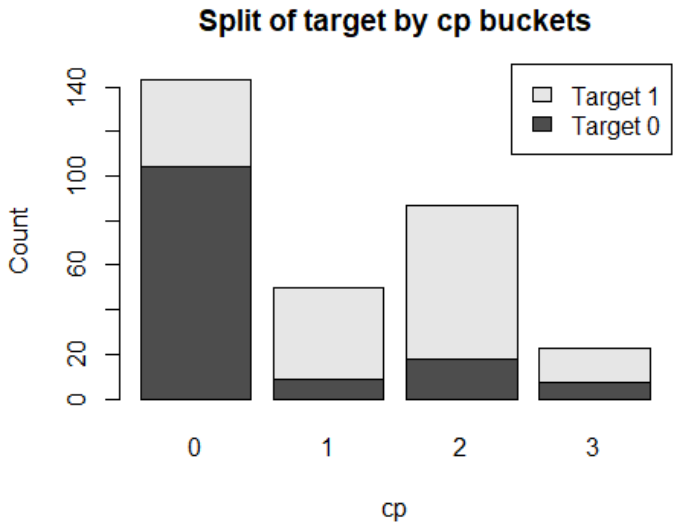
```
barplot(table(data$target), main='Split of target variable', xlab='target
variable', ylab='Count of patients')
```



Bivariate analyses showed that certain variables may highly important to predict heart disease as they showed large variation between patients with and without heart disease. These variables are -
cp, exang, slope, ca, thal, thalach

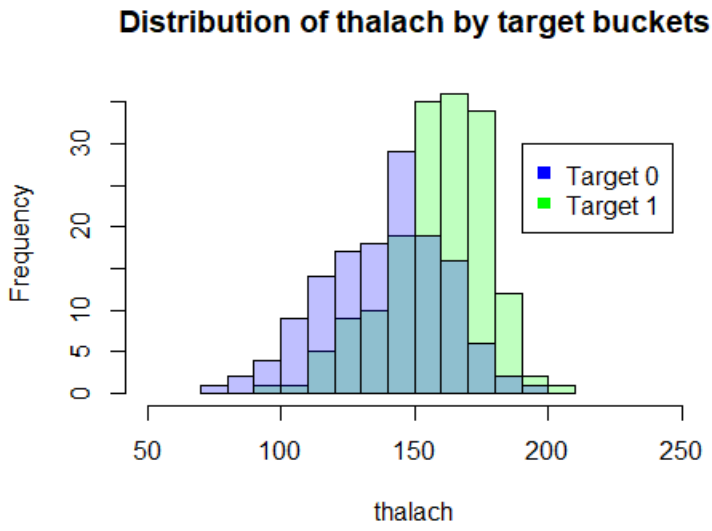
It can be seen from below plot that patients with chest pain type cp=0 are less likely to have heart disease than those with chest pain types cp=1,2 or 3

```
barplot(table(data$target, data$cp),
        main = 'Split of target by cp buckets',
        xlab = 'cp', ylab = 'Count',
        legend.text = c('Target 0', 'Target 1'), args.legend = c(x=5,y=150))
```



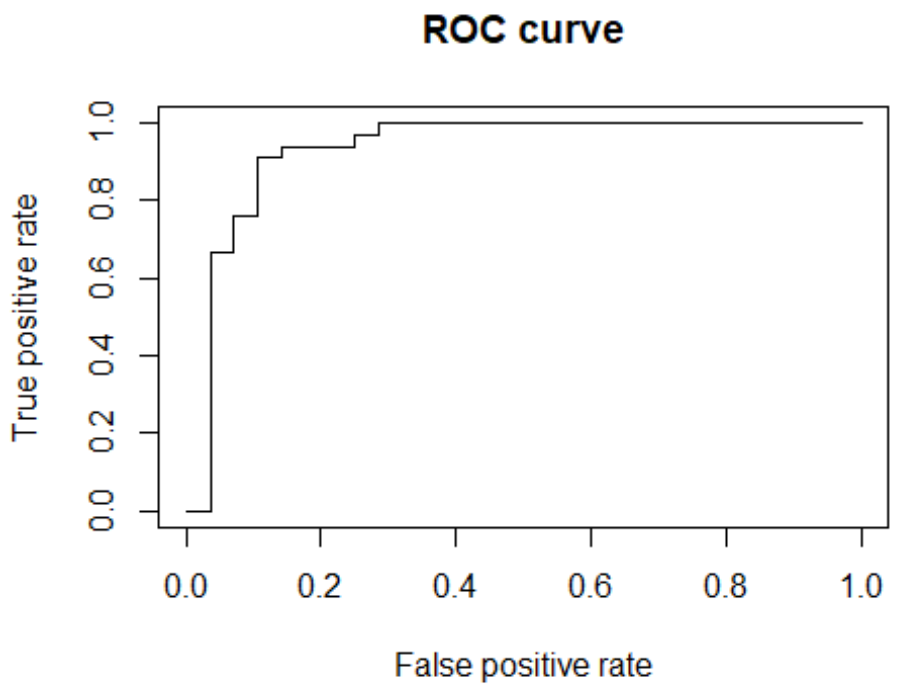
Similarly, it can be seen from below plot that patients with heart disease tend to have higher maximum heart rate than those without heart disease.

```
p1 <- hist(data$thalach[data$target==0])

p2 <- hist(data$thalach[data$target==1])

plot(p2, col=rgb(0,1,0,1/4), xlim=c(50,250), main='Distribution of thalach by
target buckets', xlab='thalach')
plot(p1, col=rgb(0,0,1,1/4), xlim=c(50,250), add=T)

legend(x=190, y=30, legend=c('Target 0', 'Target 1'), pch=15, col=c("blue",
"green"))
```



The logistic regression model fit the data well. The base model gave an AIC of **200.28**, but the best AIC obtained after variable selection was **192.54**.
The ROC curve obtained and the AUC are as follows -

```
# ROC curve and AUC
plot(ROCRperf, main='ROC curve')
```




```
auc <- attributes(performance(ROCRpred, 'auc'))$y.values[[1]]
auc
```

```
## [1] 0.9329004
```

The confusion matrix showed that 55 of the 61 instances in the test set were correctly classified at a probability threshold of 0.5. Also, sensitivity was **0.893** and specificity was **0.909**.

```
# Confusion Matrix
confusionMatrix(test_data$target, predClass)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 25  3
##          1  3 30
##
##                Accuracy : 0.9016
```

```
##                  95% CI : (0.7981, 0.963)
##     No Information Rate : 0.541
##     P-Value [Acc > NIR] : 1.252e-09
##
##                   Kappa : 0.8019
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8929
##             Specificity : 0.9091
##          Pos Pred Value : 0.8929
##          Neg Pred Value : 0.9091
##              Prevalence : 0.4590
##          Detection Rate : 0.4098
##    Detection Prevalence : 0.4590
##       Balanced Accuracy : 0.9010
##
##        'Positive' Class : 0
##
```

Conclusion

The UCI Heart Disease dataset obtained from Kaggle was used to build a logistic regression based predictive model to detect whether a patient has heart disease or not.

The best model performance was achieved after variable selection using stepwise backward elimination. It was determined that variables such as *restecg*, *fbs*, *age*, *chol* and *oldpeak* were not critical to predicting heart disease. The most significant variables were *ca*, *cp* and *sex*.

The final model had an accuracy of over 90% with a sensitivity of 89% and specificity of 91%. Sensitivity is the percentage of positive cases that are accurately captured. Any positive case that is incorrectly classified as a negative can have adverse effects on the diagnosis and thus on subsequent therapy. Therefore, there is scope to increase the sensitivity further, perhaps with the use of more advanced algorithms like Random Forest or SVM.

All in all, logistic regression is a good starting point to predict heart disease among patients.