Heart Disease Prediction

Luke Hensley

June 12, 2019

Executive Summary

This report attempts to predict which patients may have heart disease. The data used is the UCI Heart Disease dataset from Kaggle - (https://www.kaggle.com/ronitf/heart-disease-uci).

Exploratory analysis was conducted on the data to visualize differences with patients that have heart disease, and those who don't. The dataset was split into two data sets (train and test) and a regression model was built. The model was then improved using stepwise backward elimination. Predictions were made based on train set data, then performance of the model was evaluated by using the model on the test data set.

Performance measures and results:
Area under the curve (AUC): 0.933
Accuracy: 0.902
Sensitivity: 0.8929
Specificity: 0.9091

Analysis

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

Data structure:

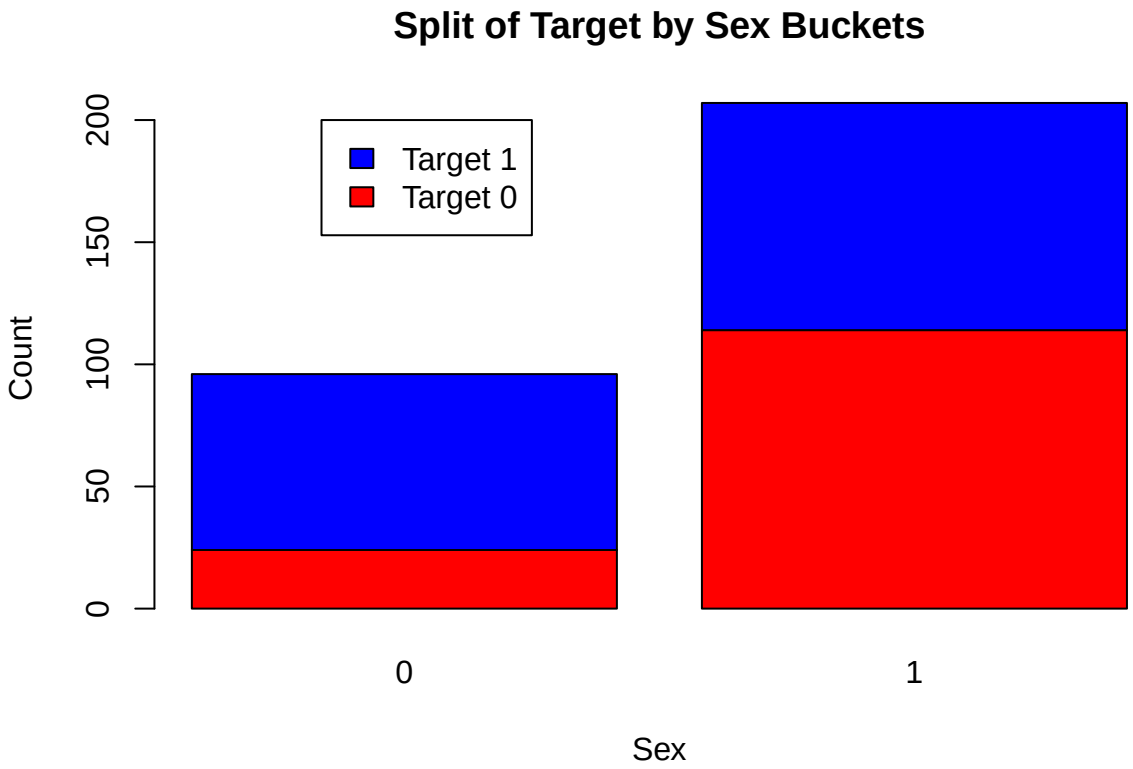
```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
##  $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
##  $ 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 ...
```

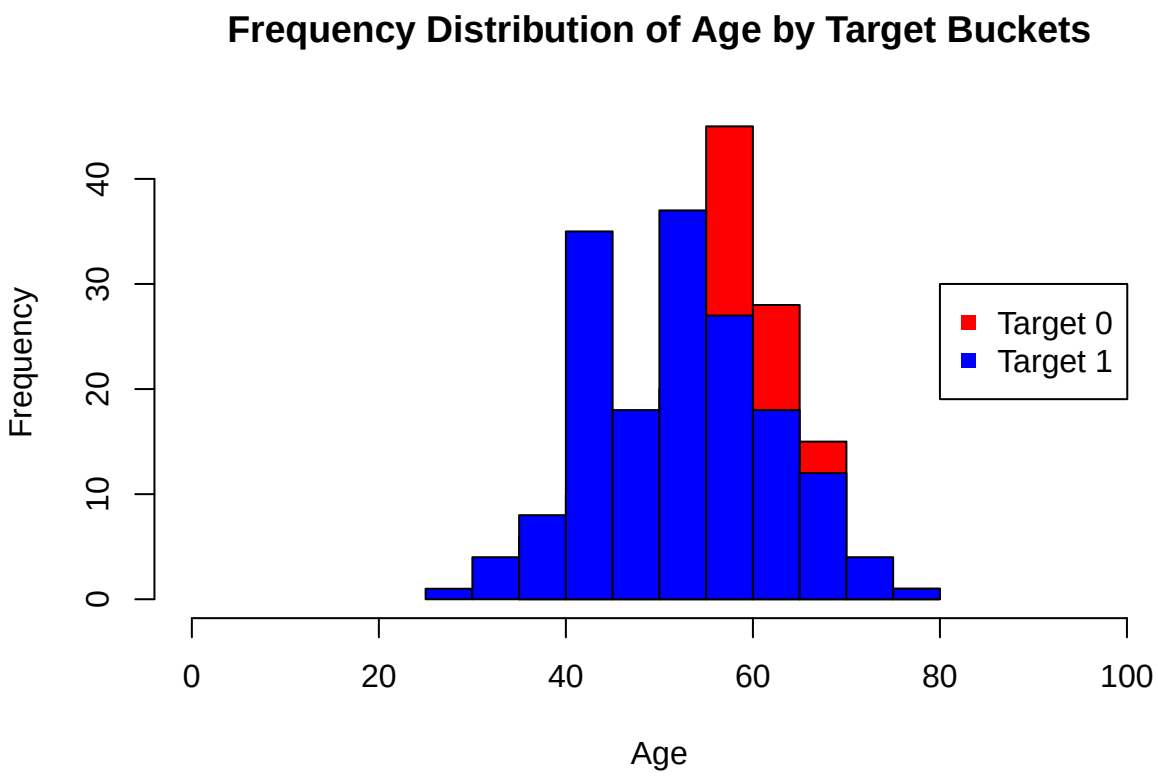
```
##  $ slope   : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
##  $ ca      : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ thal    : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
##  $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Exploratory Analyses

We first make sure that there are no missing values in the data. Then we check the summary statistics of the variables. Bivariate analyses between the independent and target variables are conducted and plotted. Categorical independent variables are plotted using a barplot to show the split of the 'target'. A frequency histogram is created to show the continuous independent variables, and the difference in distributions for the two 'target' categories is shown.

Barplot for 'Sex' variable:





Predictive Analyses

The data was split into two data sets, train and test. A random 20% of data is in the test set and the remaining 80% is used to train the model.

```
## [1] 61
```

```
## [1] 242
```

A logistic regression model is chosen, and the stepwise backward elimination method is then used to select variables. Akaike Information Criteria (AIC) is used, while p-values detect insignificant variables for each step.

```
##
## Call:
## glm(formula = target ~ sex + cp + trestbps + thalach + exang +
##     oldpeak + slope + ca + thal, family = binomial(link = "logit"),
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9437  -0.3099   0.1179   0.4076   1.9699
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  1.45733    4.01172   0.363  0.71640
## sex1        -1.58123    0.62103  -2.546  0.01089 *
## cp1          1.07415    0.67361   1.595  0.11080
## cp2          2.00572    0.57713   3.475  0.00051 ***
## cp3          2.30571    0.79618   2.896  0.00378 **
## trestbps    -0.02647    0.01258  -2.104  0.03541 *
## thalach      0.02405    0.01298   1.853  0.06382 .
## exang1      -1.03178    0.50837  -2.030  0.04240 *
## oldpeak     -0.45539    0.25429  -1.791  0.07332 .
## slope1      -1.58002    1.06194  -1.488  0.13679
## slope2      -0.01348    1.16945  -0.012  0.99080
## ca1         -2.33405    0.58129  -4.015 5.94e-05 ***
## ca2         -3.61631    0.87370  -4.139 3.49e-05 ***
## ca3         -1.29886    1.03250  -1.258  0.20840
## ca4          1.30043    1.71700   0.757  0.44882
## thal1        2.56713    3.21632   0.798  0.42478
## thal2        1.65414    3.11774   0.531  0.59572
## thal3        0.47307    3.12132   0.152  0.87953
## ---
## 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: 145.15  on 224  degrees of freedom
## AIC: 181.15
##
## Number of Fisher Scoring iterations: 6
```

The trained model is then used to make predictions on the test set. The ROC curve is plotted and the AUC is calculated for performance measurement. A probability threshold of 0.5 is set, and a confusion matrix is viewed alongside sensitivity and specificity.

Results

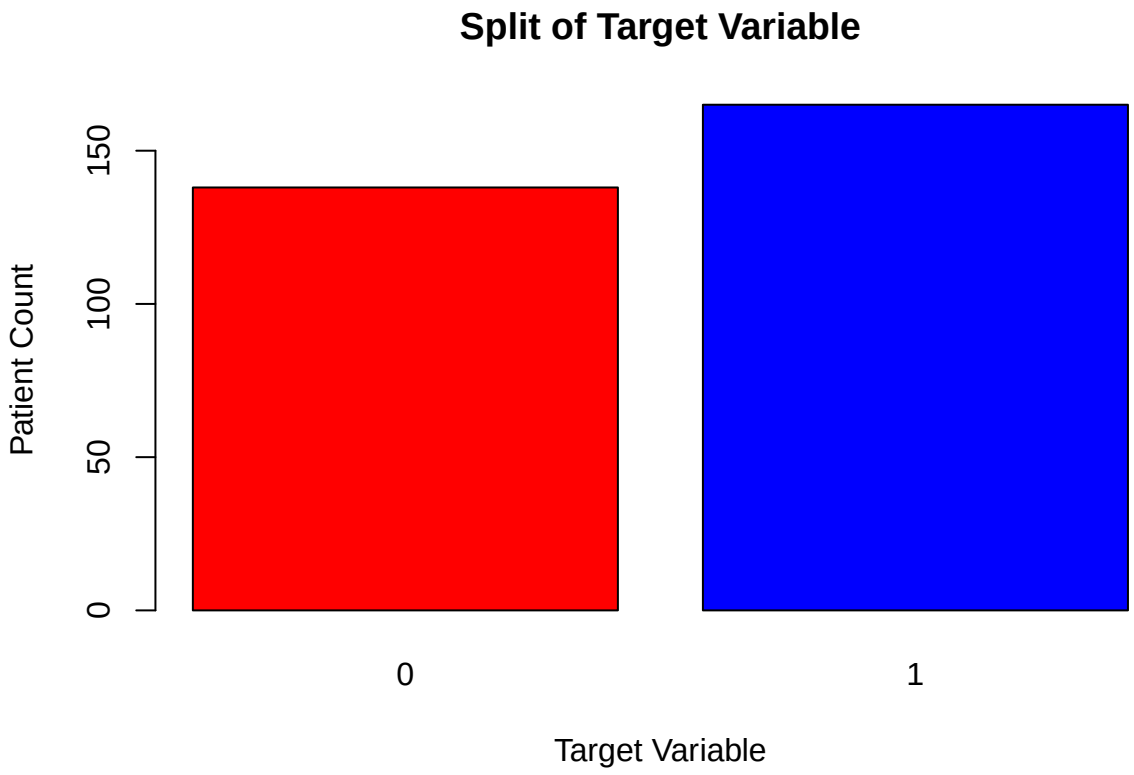
No blank or NA values are found in the data.

```
##      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
```

There is no major imbalance in the target variable.

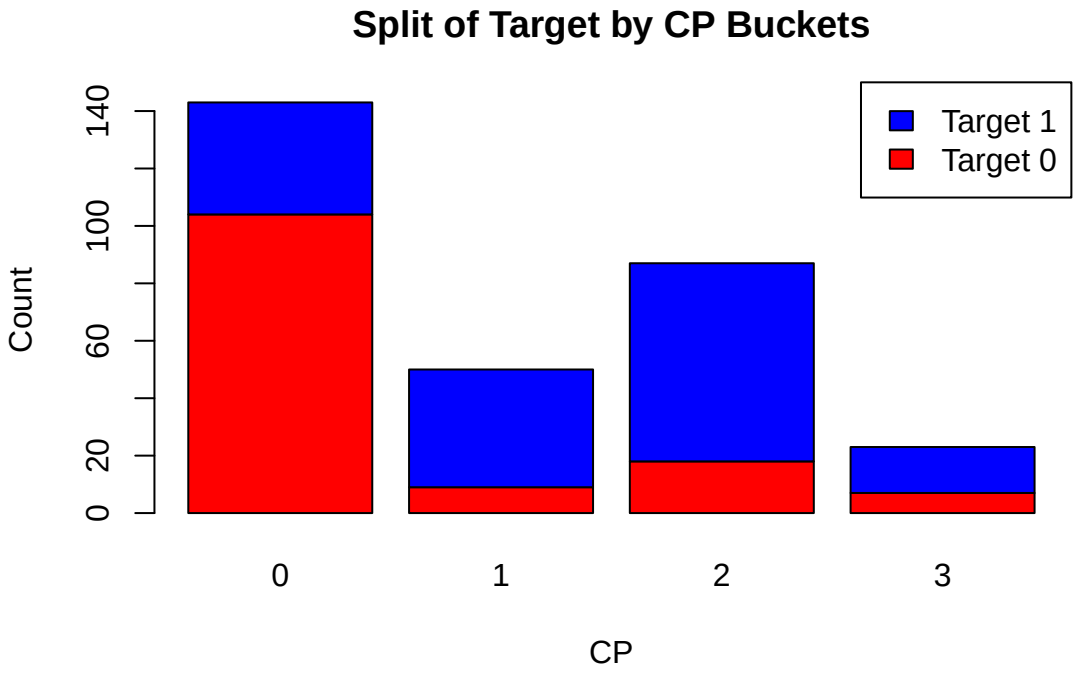
```
##       age        sex     cp         trestbps          chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :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
##
##
##
##
```



Bivariate analyses showed some variables very important to predicting heart disease (cp, exang, slope, ca, thal, thalach).

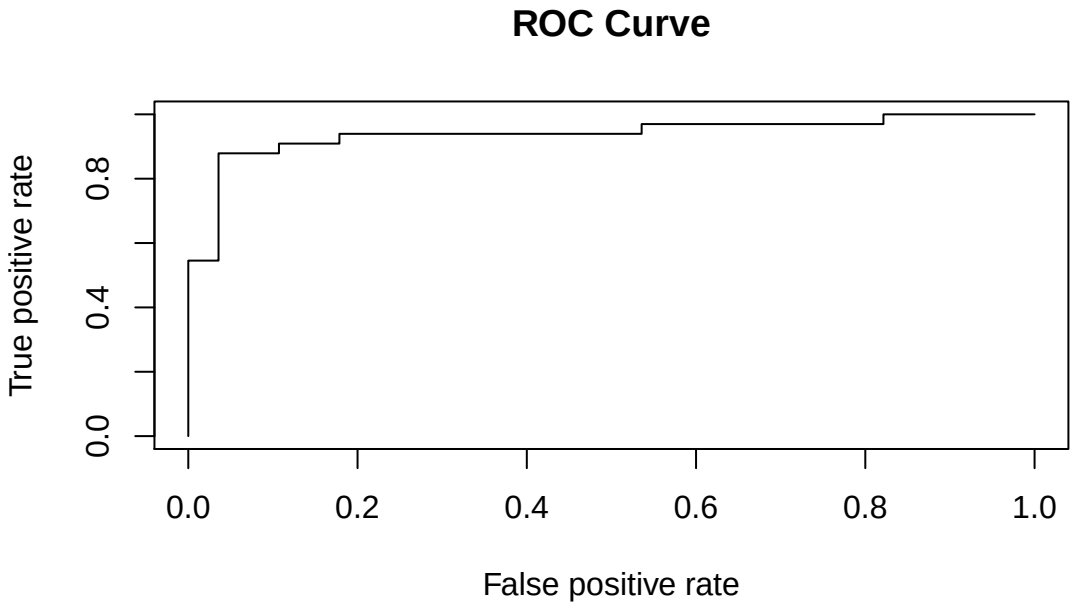
In the below plot, patients with chest pain cp=0 are less likely to have heart disease than those with chest pain cp=1,2 or 3.



The following plot shows patients with heart disease tended to have a higher maximum heart rate than those not having heart disease.

The base model gave an AIC of 200.28. The best AIC after variables were selected was 192.54.

ROC curve and AUC value:




```
## [1] 0.9383117
```

Confusion matrix showed 55 of 61 instances in the test set were correctly classified at a probability threshold of 0.5. In addition, sensitivity was 0.893 and specificity was 0.909.

```
## 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 model performed best after using stepwise backward elimination. The most significant variables were 'ca', 'cp' and 'sex'. The variables 'age', 'chol', 'fbs', 'oldpeak', and 'restecg' were not critical for heart disease prediction.

The final model had an accuracy of over 90%. Sensitivity of 89% (percentage of positive cases accurately captured), and specificity of 91%.