Capstone Project - Heart Disease UCI

Prediction System

Reza Hashemi

August 12, 2019

Abstract

This report is part of the final project capstone to obtain the 'Professional Certificate in Master of Data Science' emited by Harvard University (HarvadX), platform for education and learning. The main objective is to create a recommendation system using the Heart Disease UCI dataset, and it must be done training a machine learning algorithm using the inputs in one subset to predict in the validation set.

Contents

1 Executive Summary

The main purpose of this project is to develop a machine learning algorithm to predict wheter patients have a heart disease or not. The entire dataframe can be found at here.

The dataset contains 14 variables: 13 are independent - 8 categorical & 5 continuous variables 1 binary called `target`.

The procedure was:

1. **Exploratory Analysis**: through data and graphics, evaluate all patients who have a heart disease and those who do not, with each of the independent variables.

2. **Split Data Set**: Split the data set into train and test sets, to create and evaluate the model.

2 Introduction

The present report covers the Heart Attack UCI dataset, with aknowledgements to:

Creators: 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

The main objective for using this dataset is to build several machine learning classification models that predicts the presence of heart disease in a patient. About 165 deaths per 100.000 individuals in 2007 die of heart disease in the United States every year - that's 1 in every 4 deaths, it is the leading cause of death in US. Heart disease is the leading cause of death for both, mean and women. More than half of the deaths due to heart disease in 2009 were in men. More information can be found at Heart Disease and Stroke Statistics-2019

The machine learning models used in this report aims to create a classifier that provides a high accuracy level combined with a los rate of false-negatives (high sensitivity) .

"This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researches to this date. The 'goal' field refers to the presence of heart disease in the patient. It is integer value from 0 (no presence) to 4". [kaggle.com](https://www.kaggle.com/ronitf/heart-disease-uci)

The dataset contains 14 variables and 303 observations.

3 Data Analysis

3.1 Selected Data

This dataset contains different attributes:

Independent Variables
- Categorical **(8)**

Table 1: Attributes And Definitions

Attribute	Definition
ca	number of major vessels (0-3) colored by flourosopy
cp	pain type (0 - 3)
exang	exercise induced angina (1 = yes; 0 = no)
fbs	fbs(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
sex	sex(1 = male; 0 = female)
slope	the slope of the peak exercise ST segment
thal	thal3 = normal; 6 = fixed defect; 7 = reversable defect

- Continuos **(5)**

Table 2: Attributes And Definitions

Attribute	Definition
age	age in years
chol	cholserum cholestoral in mg/dl
oldpeak	oldpeakST depression induced by exercise relative to rest
testbps	resting blood pressure (in mm/Hg on admission to the hospital)
thalach	maximum heart rate achieved

Binary Attribute
- Binary Attribute **(1)**

Table 3: Attributes And Definitions

Attribute	Definition
target	target 1 or 0

The `target` variable represents the target feature with levels `1` or `0`, and its proportions are shown below:

```
##
##    0    1
## 0.46 0.54
```

Each attribute has been converted to `factor`:

```
## [1] 303  14
```

```
##    ï..age      sex       cp trestbps     chol      fbs  restecg  thalach
## "factor" "factor" "factor" "factor" "factor" "factor" "factor" "factor"
##     exang  oldpeak    slope       ca     thal   target
## "factor" "factor" "factor" "factor" "factor" "factor"
```

Let's see the **10 first observations** in data set:

```
##    ï..age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca
## 1      63   1  3      145  233   1       0     150     0     2.3     0  0
## 2      37   1  2      130  250   0       1     187     0     3.5     0  0
## 3      41   0  1      130  204   0       0     172     0     1.4     2  0
## 4      56   1  1      120  236   0       1     178     0     0.8     2  0
## 5      57   0  0      120  354   0       1     163     1     0.6     2  0
## 6      57   1  0      140  192   0       1     148     0     0.4     1  0
## 7      56   0  1      140  294   0       0     153     0     1.3     1  0
## 8      44   1  1      120  263   0       1     173     0       0     2  0
## 9      52   1  2      172  199   1       1     162     0     0.5     2  0
## 10     57   1  2      150  168   0       1     174     0     1.6     2  0
##    thal target
## 1     1      1
## 2     2      1
## 3     2      1
## 4     2      1
## 5     2      1
## 6     1      1
## 7     2      1
## 8     3      1
## 9     3      1
## 10    2      1
```

A **summary** of dataset:

```
##      ï..age    sex      cp         trestbps        chol      fbs     restecg
##  58     : 19   0: 96   0:143   120     : 37   197     :  6   0:258   0:147
##  57     : 17   1:207   1: 50   130     : 36   204     :  6   1: 45   1:152
##  54     : 16           2: 87   140     : 32   234     :  6           2:  4
##  59     : 14           3: 23   110     : 19   212     :  5
##  52     : 13                   150     : 17   254     :  5
##  51     : 12                   138     : 13   269     :  5
##  (Other):212                   (Other):149   (Other):270
##     thalach    exang      oldpeak     slope    ca      thal     target
##  162    : 11   0:204   0      : 99   0: 21   0:175   0:  2   0:138
##  160    :  9   1: 99   1.2    : 17   1:140   1: 65   1: 18   1:165
##  163    :  9           0.6    : 14   2:142   2: 38   2:166
##  152    :  8           1      : 14           3: 20   3:117
##  173    :  8           0.8    : 13           4:  5
##  125    :  7           1.4    : 13
##  (Other):251           (Other):133
```

The **structure** of dataset:

```
## 'data.frame':    303 obs. of  14 variables:
##  $ ï..age  : Factor w/ 41 levels "29","34","35",..: 30 4 8 23 24 24 23 11 19 24 ...
##  $ 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: Factor w/ 49 levels "94","100","101",..: 32 23 23 15 15 29 29 15 44 35 ...
##  $ chol    : Factor w/ 152 levels "126","131","141",..: 65 81 36 68 146 26 117 93 32 10 ...
##  $ 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 : Factor w/ 91 levels "71","88","90",..: 50 85 72 77 63 48 53 73 62 74 ...
##  $ exang   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
##  $ oldpeak : Factor w/ 40 levels "0","0.1","0.2",..: 23 33 15 9 7 5 14 1 6 17 ...
##  $ 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 ...
```

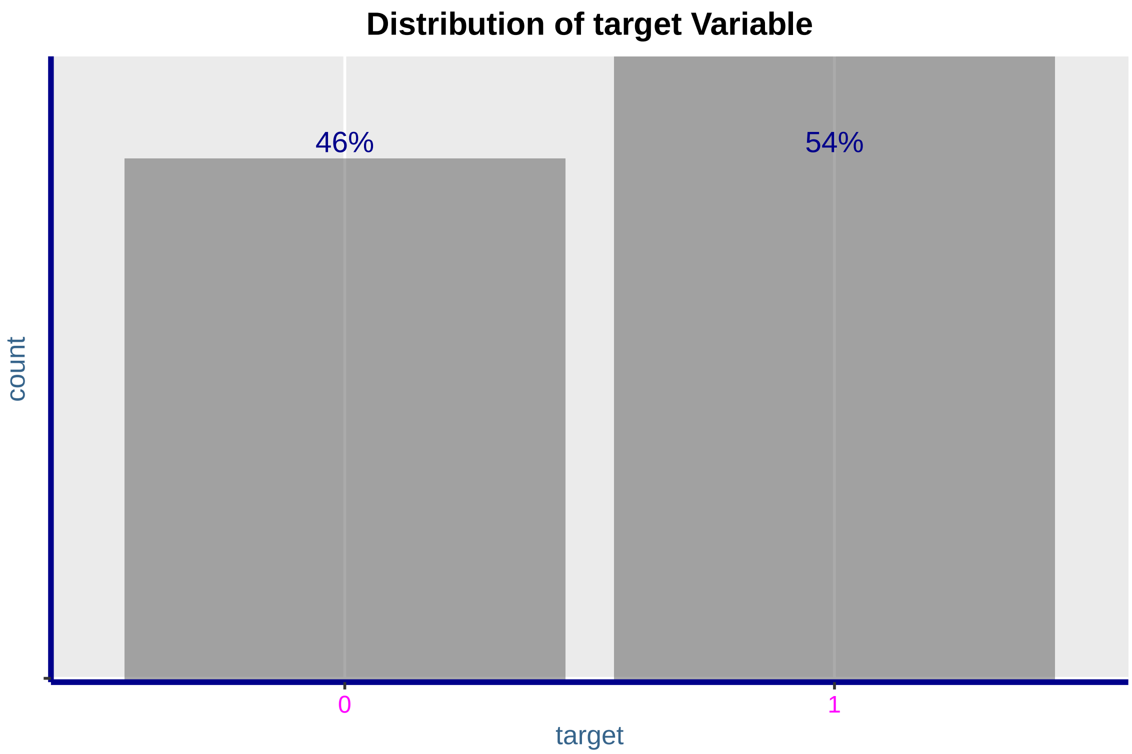
3.2 Distribution of the `target` Attribute

The `target` variable represents the target feature with levels `1` and `0`. Its proportions are shown below:

Table 4: Target Variable Distribution

Var1	Freq
0	46%
1	54%

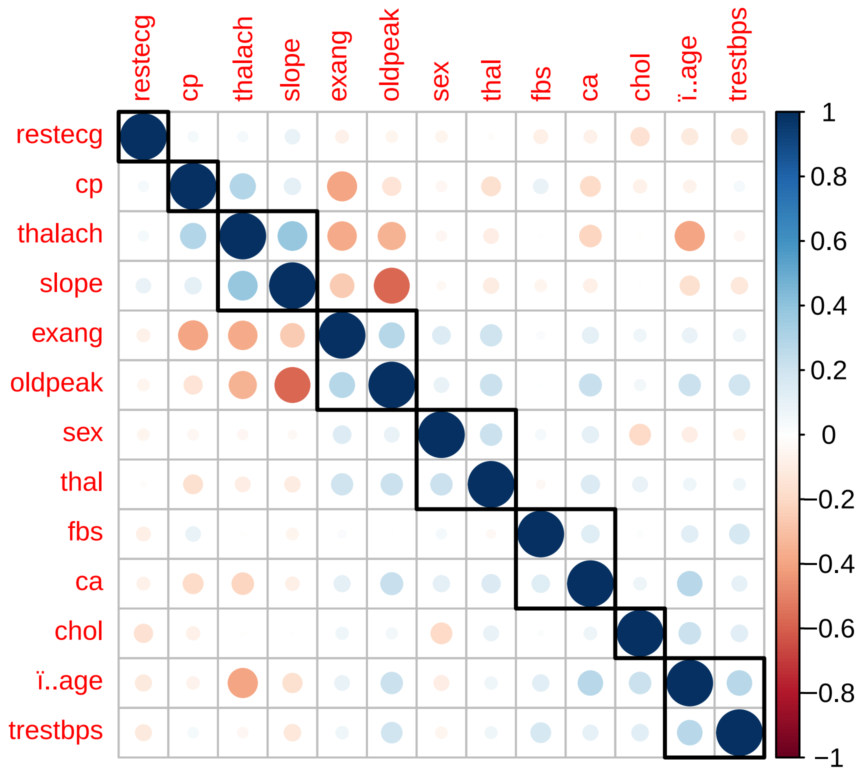
A graph that shows the previous proportions is:



3.3 Exploring the Variable's Correlation

Most machine learning algorithms assume that the predictor variables are independent from each others. This is the reason why the multiolinearity will be removed to achieve a more robust analysis.

Variables' Correlation Plot



The plot shows that none variables have a high correlation with any other, all correlations are less than 0.8.

4 Data Transformation

We will remove highly correlated predictors, based on whose correlation is above 0.9. For this purpose, we will use the `findcorrelation()` function, from caret package, which employs a heuristic algorithm to determine which variable should be removed instead of selecting bindly.

```
## [1] 14
```

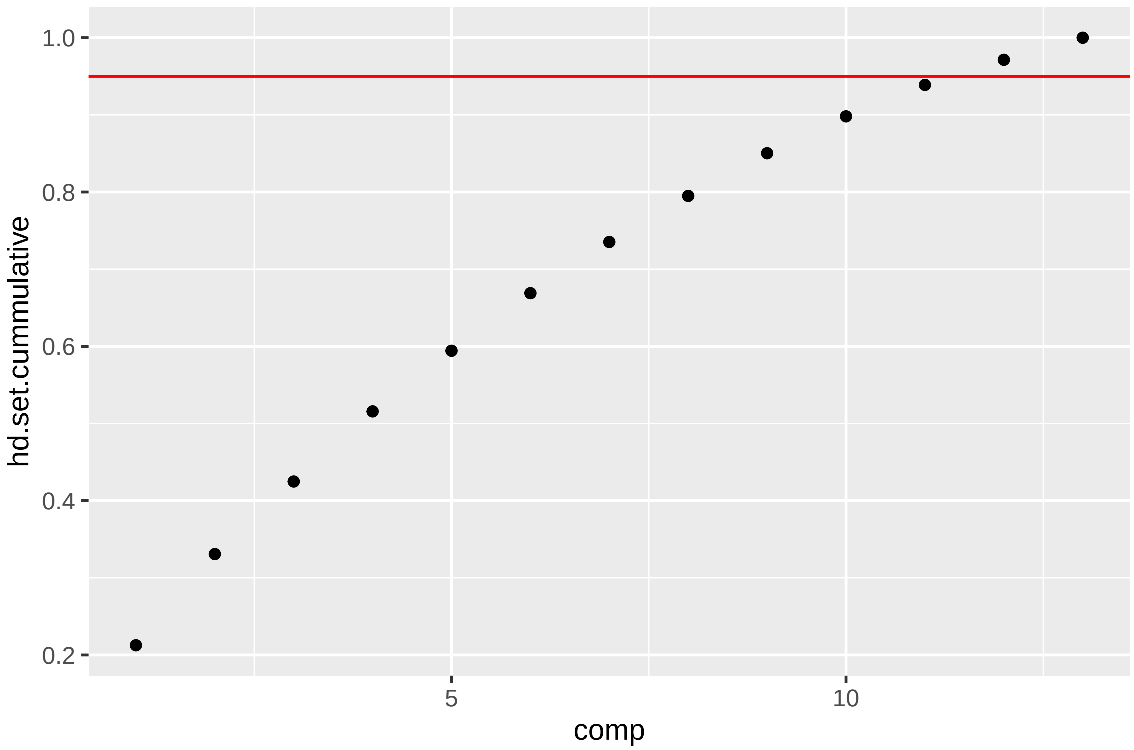
5 Data Pre-Processing

5.1 Principle Component Analysis (PCA)

The `target` variable is removed followd by scaling and centering these variables.

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     1.6622 1.2396 1.10582 1.08681 1.01092 0.98489
## Proportion of Variance 0.2125 0.1182 0.09406 0.09086 0.07861 0.07462
## Cumulative Proportion  0.2125 0.3307 0.42481 0.51567 0.59428 0.66890
##                            PC7     PC8    PC9    PC10    PC11    PC12
## Standard deviation     0.92885 0.88088 0.8479 0.78840 0.72808 0.65049
## Proportion of Variance 0.06637 0.05969 0.0553 0.04781 0.04078 0.03255
## Cumulative Proportion  0.73527 0.79495 0.8503 0.89807 0.93885 0.97140
##                          PC13
## Standard deviation     0.6098
## Proportion of Variance 0.0286
## Cumulative Proportion  1.0000
```

A plot of the compute proportion of variance is:

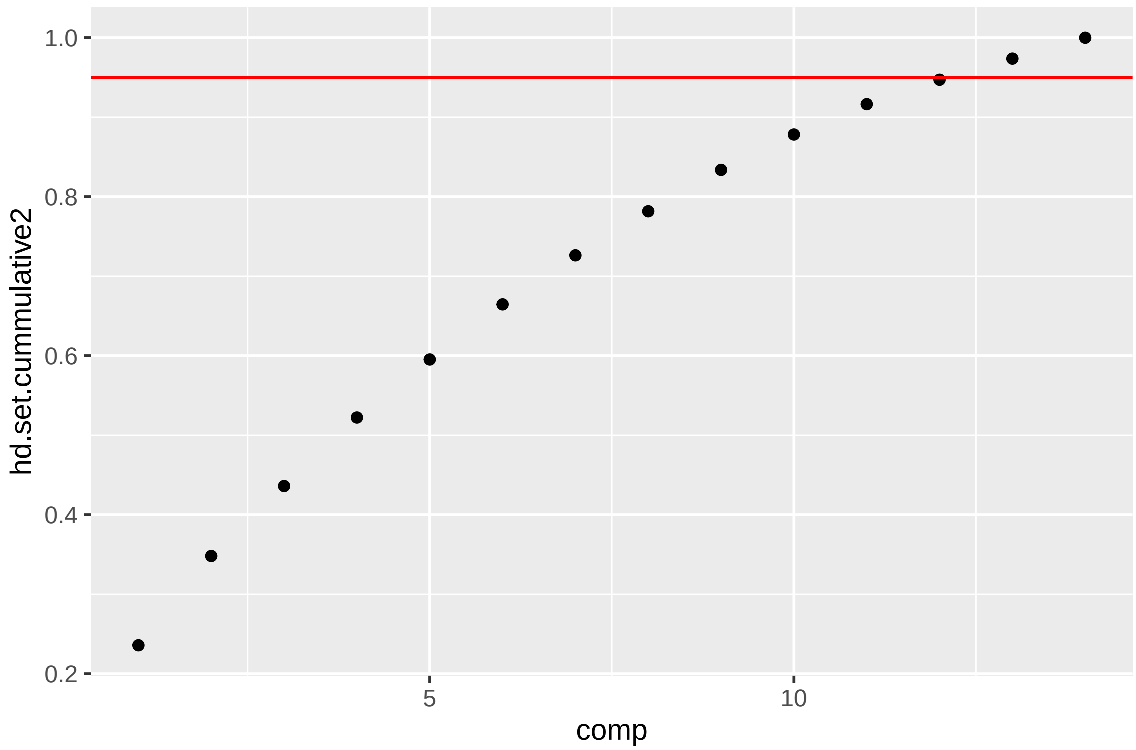


The above plot shows that 95% of the variance is explained with all PC's, working with the original dataset.

5.2 PCA Applied to the Transformed Dataset

```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5    PC6     PC7
## Standard deviation     1.8170 1.2539 1.1100 1.09847 1.0110 0.9850 0.92910
## Proportion of Variance 0.2358 0.1123 0.0880 0.08619 0.0730 0.0693 0.06166
## Cumulative Proportion  0.2358 0.3481 0.4361 0.52231 0.5953 0.6646 0.72627
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.88096 0.85393 0.78913 0.73103 0.65577 0.60982
## Proportion of Variance 0.05544 0.05209 0.04448 0.03817 0.03072 0.02656
## Cumulative Proportion  0.78170 0.83379 0.87827 0.91644 0.94716 0.97372
##                           PC14
## Standard deviation     0.60658
## Proportion of Variance 0.02628
## Cumulative Proportion  1.00000
```

A plot of the compute the proportion of variance explained is:



The above plot doesn't show any variation in comparisson with the previous plot of proportion of variance.

5.3 Linear Discriminant Analysis (LDA)

Now we will use the LDA instead of PCA, since it takes into consideration the different classes & could provide better results.

```
## Call:
## lda(target ~ ., data = hd.set.numeric, center = TRUE, scale = TRUE)
##
## Prior probabilities of groups:
##         0         1
## 0.4554455 0.5445545
##
## Group means:
##      ï..age       sex        cp trestbps     chol       fbs   restecg
## 0 56.60145 0.8260870 0.4782609 134.3986 251.0870 0.1594203 0.4492754
## 1 52.49697 0.5636364 1.3757576 129.3030 242.2303 0.1393939 0.5939394
##    thalach     exang   oldpeak    slope        ca     thal
## 0 139.1014 0.5507246 1.5855072 1.166667 1.1666667 2.543478
## 1 158.4667 0.1393939 0.5830303 1.593939 0.3636364 2.121212
##
## Coefficients of linear discriminants:
##                   LD1
## ï..age   -0.003285901
## sex      -0.784995108
## cp        0.451396013
## trestbps -0.007974151
## chol     -0.001415982
## fbs       0.069584331
## restecg   0.199649424
```

```
## thalach   0.012092920
## exang    -0.576928147
## oldpeak  -0.235458579
## slope     0.316323381
## ca       -0.402928538
## thal     -0.476771859
```

5.4 Data Analysis Between `Independent` Attributes & `target` Attribute

A Bivariate Analysis has been done, between each `independent` attribute and `target`, all these plots can be found at the end of this document.

6 Data Partition

Two sets (training & test) have been created from main dataset.

A partition has been done, into `training`(80%) & `test`(20%) datasets:

Table 5: Attributes And Definitions

Dataset	Observations
training	242
test	61

7 Model Creation

7.1 Logistic Regression Model

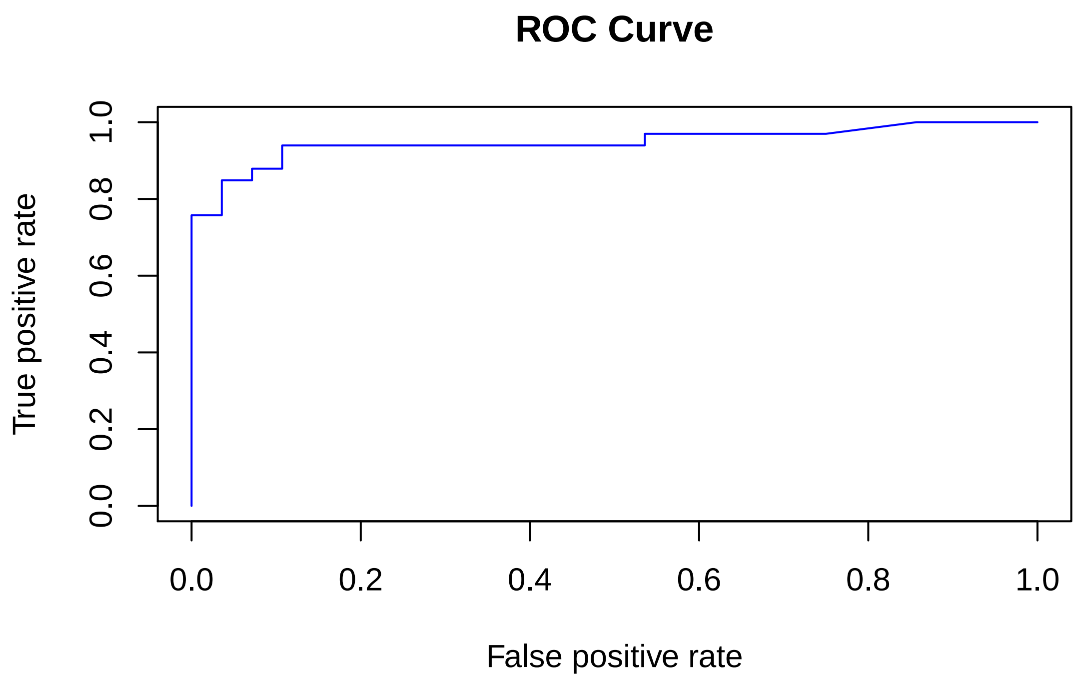
The `Regresion Model` is very useful in this analysis because this is a binary classification problem.
Then, in order to make a suitable selection of the variables, the `Stepwise Backward & Forward` elimination method was used, as well as the `AIC(Akaike Information Criteria)` for selection criteria, and, `p-values` has been used to detect the least significant variables.

`stepAIC()` function has been used to choose the best model by `AIC`. It has an option named direction, which can take the following values: i) "both" (for stepwise regression, both forward and backward selection); "backward" (for backward selection) and "forward" (for forward selection). It return the best final model. And, for out model we used the `both` option:

```
##
## Call:
## glm(formula = target ~ exang + ca + thal + slope + cp + sex,
##     family = binomial(link = "logit"), data = train.set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7607  -0.3563   0.1138   0.4380   1.8944
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1393     3.8920  -0.036 0.971447
## exang1       -1.2500     0.4773  -2.619 0.008821 **
## ca1          -2.1520     0.5360  -4.015 5.95e-05 ***
## ca2          -3.6537     0.8360  -4.371 1.24e-05 ***
## ca3          -1.9976     0.9246  -2.160 0.030735 *
## ca4           1.5660     1.7574   0.891 0.372866
## thal1         2.2362     3.8765   0.577 0.564030
## thal2         1.8122     3.7964   0.477 0.633120
## thal3         0.4881     3.8001   0.128 0.897804
## slope1       -0.5870     0.8276  -0.709 0.478178
## slope2        1.4925     0.8753   1.705 0.088192 .
## cp1           1.2461     0.6426   1.939 0.052472 .
## cp2           1.9218     0.5426   3.542 0.000397 ***
## cp3           1.8205     0.7151   2.546 0.010898 *
## sex1         -1.2987     0.5506  -2.359 0.018329 *
## ---
## 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: 157.94  on 227  degrees of freedom
## AIC: 187.94
##
## Number of Fisher Scoring iterations: 6
```

7.1.1 Prediction

We use the `Regression Model` to make predictions on the test set. If we consider all the possible threshold values and the corresponding specificity and sensitivity rate what will be the final model accuracy. ROC(Receiver operating characteristic) curve is drawn by taking False positive rate on X-axis and True positive rate on Y- axis ROC tells us, how many mistakes are we making to identify all the positives?




```
## NULL
```

The AUC (Area Under the Curve) has been calculated to measure performance, and its value is: 0.9475108.

7.1.2 Evaluating Model Performance

A value of `0.5` has been set as probability threshold. And the confusion matrix shows the key performance measures like `sensitivity (0.85)` and `specificity (0.87)`.

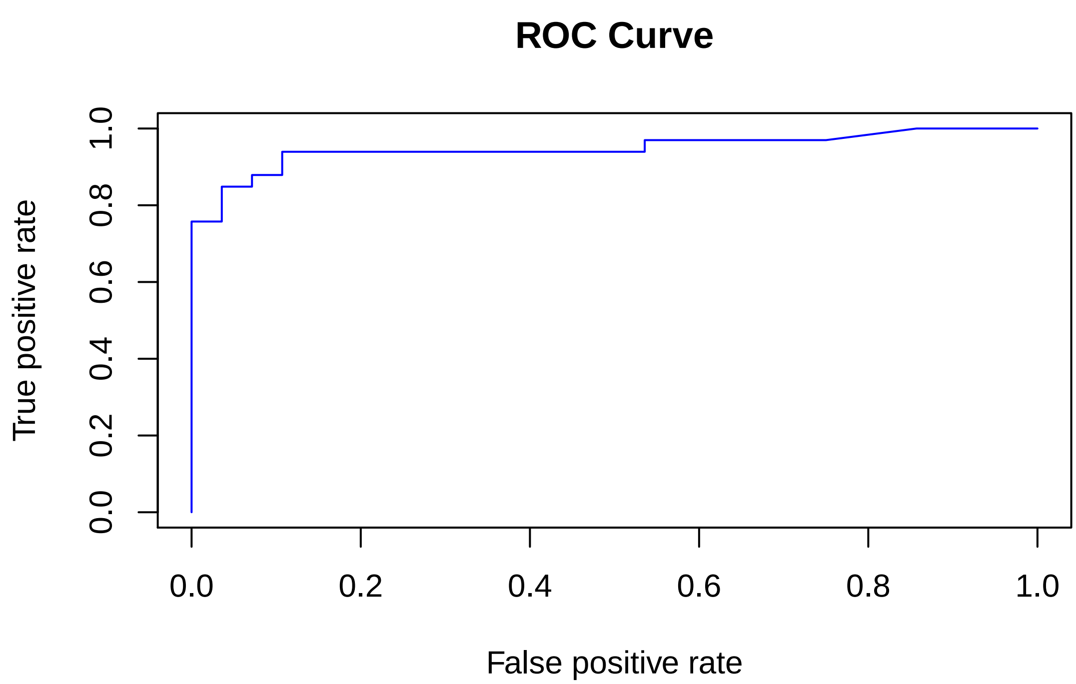
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 24  4
##          1  2 31
##
##                Accuracy : 0.9016
##                  95% CI : (0.7981, 0.963)
##     No Information Rate : 0.5738
##     P-Value [Acc > NIR] : 2.082e-08
##
##                   Kappa : 0.8009
##
##  Mcnemar's Test P-Value : 0.6831
##
##             Sensitivity : 0.9231
##             Specificity : 0.8857
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.9394
##              Prevalence : 0.4262
##          Detection Rate : 0.3934
##    Detection Prevalence : 0.4590
```

```
##        Balanced Accuracy : 0.9044
##
##         'Positive' Class : 0
##
```

7.1.3 Results

- The logistic regression model fit the data very well, tha base model gave an AIC of `194.95`.

- We want that curve to be far away from straight line. Ideally we want the area under the curve as high as possible ROC comes with a connected topic, AUC. Area Under the Curve ROC Curve Gives us an idea on the performance of the model under all possible values of threshold. We want to make almost 0% mistakes while identifying all the positives, which means we want to see AUC value near to 1. The graph that shows the AUC (Area Unde the Curve) is the following:



And, the AUC = 0.9475108
- Working with a `probability threshold` = 0.5, the confusion matrix showed that `55` of `61` instances, in test set, were correctly classified.
- The Confusion Matrix shows the key performance measures like `sensitivity (0.85)` and `specificity (0.87)`.

7.1.4 Conclusion

The dataset `Heart Disease UCI` was obtained from Kaggle. This dataset were used to construct a logistic regression based on a predictive model, in order to detect if a patient has a heart disease, or not.
The proposed model achieved the best performance after using the `stepwise` elimination process, with the option `both`, to perform a two way elimination process `backward` and `forward`. This process allowed us to identify:

Importance	Variable Name
High	ca, cp, sex
Low	age, chol, fbs, restecg

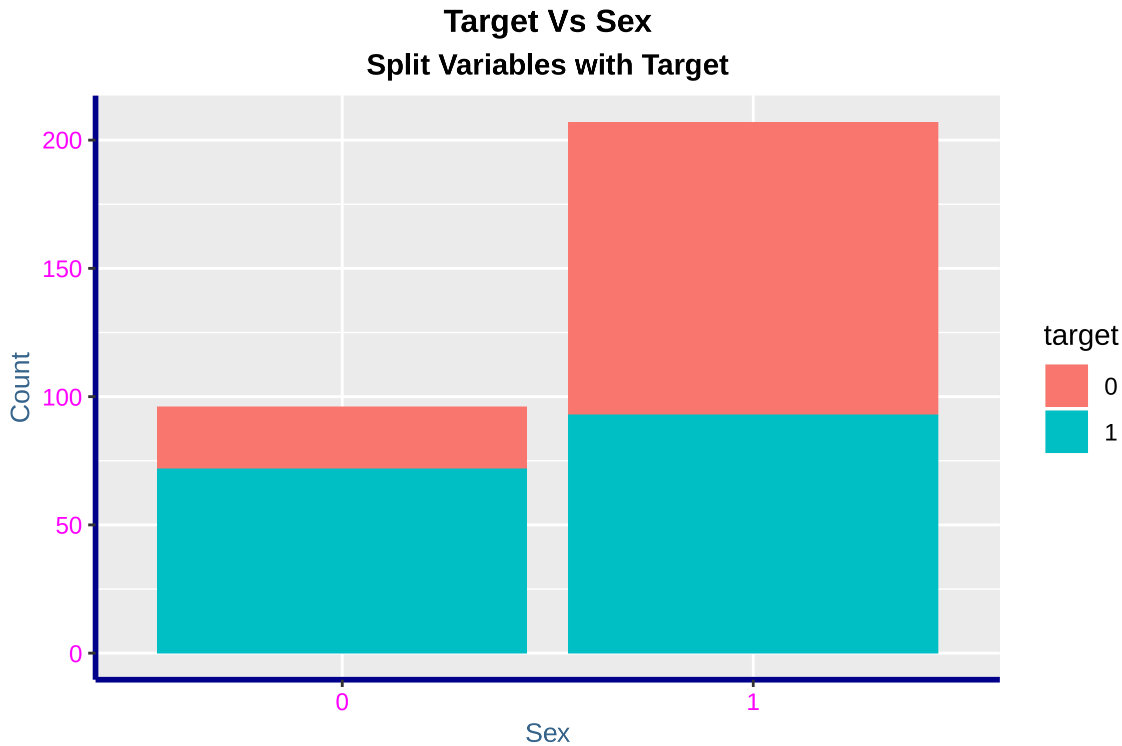
The final model results obtained, that describe the performance of the classification model, are:

Variable	Value
Accuracy	0.86
Sensitivity	0.85
Specificity	0.87

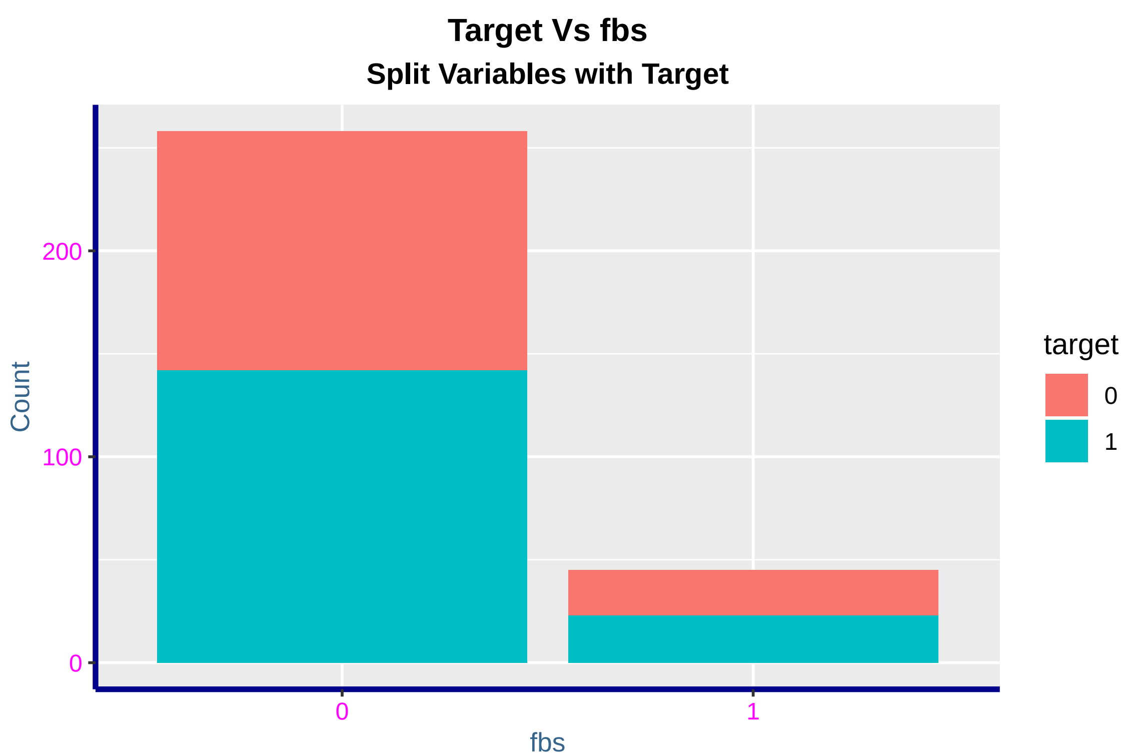
Accuracy: How often the classifier is correct **Sensitivity**: True Positive Rate Measures the proportion of actual positives that are correctly identified as such **Specificity**: True Negative Rate Measures the proportion of actual negatives that are correctly identified

8 Barplots - Bivariate Analysis

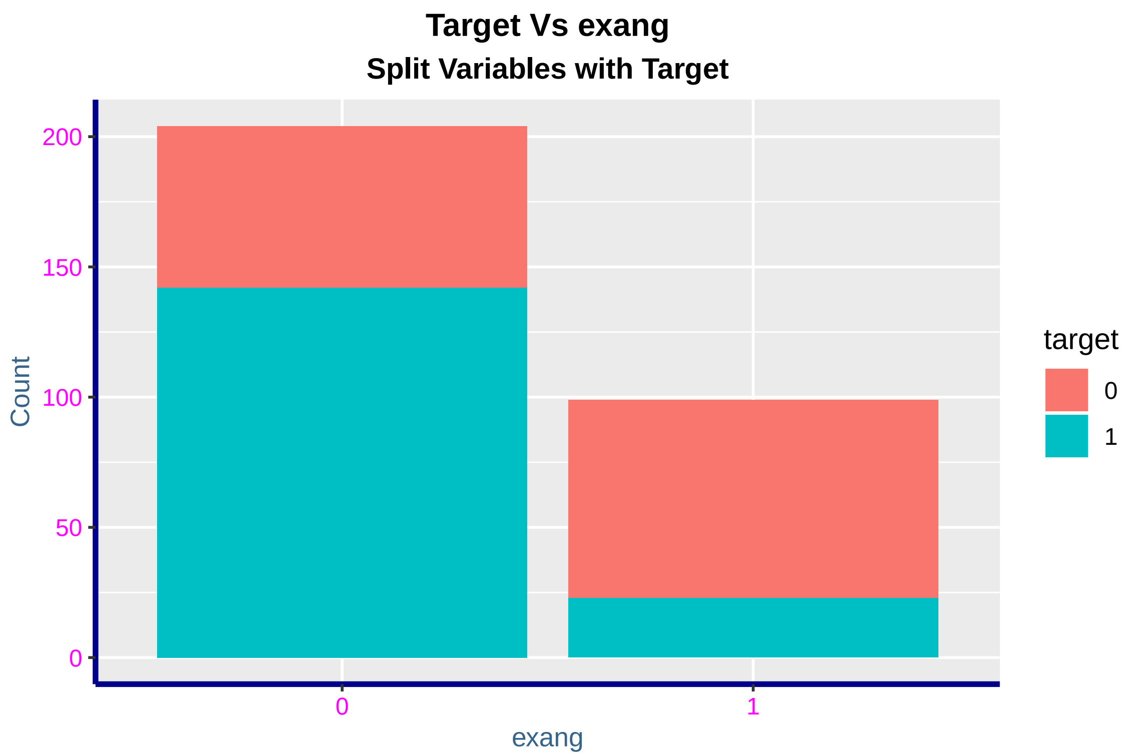
8.1 `target` Vs `sex`



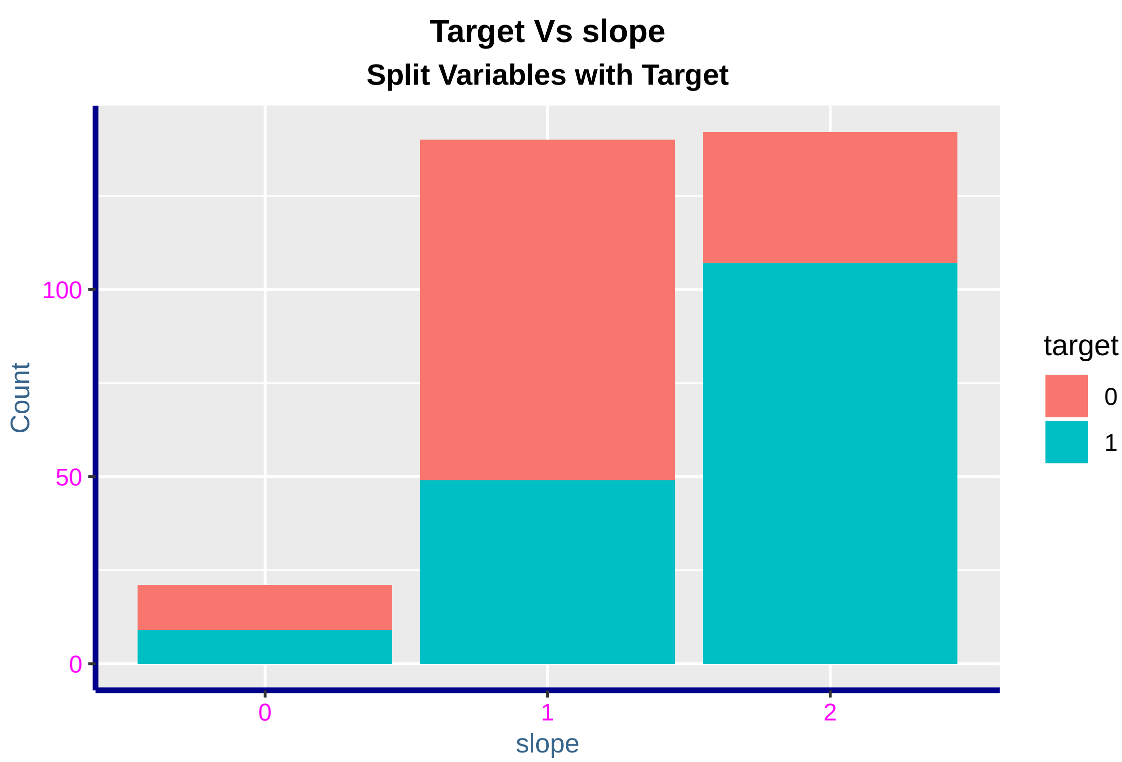
8.2 `target` Vs `fbs`



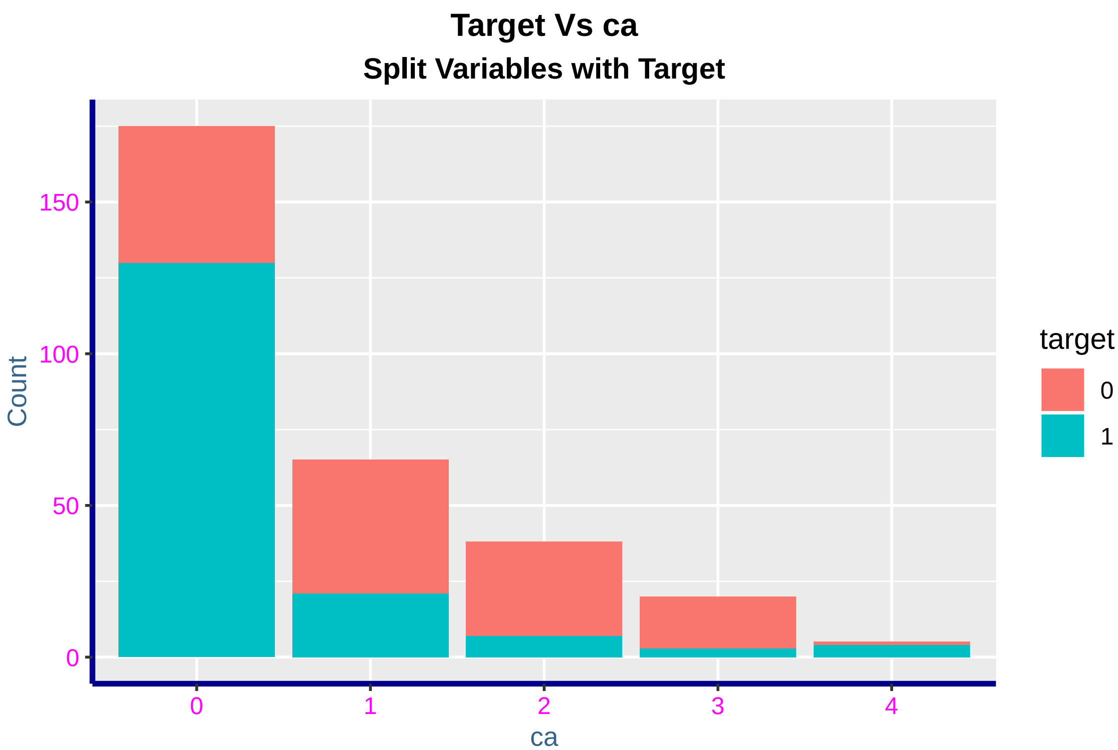
8.3 `target` Vs `exang`



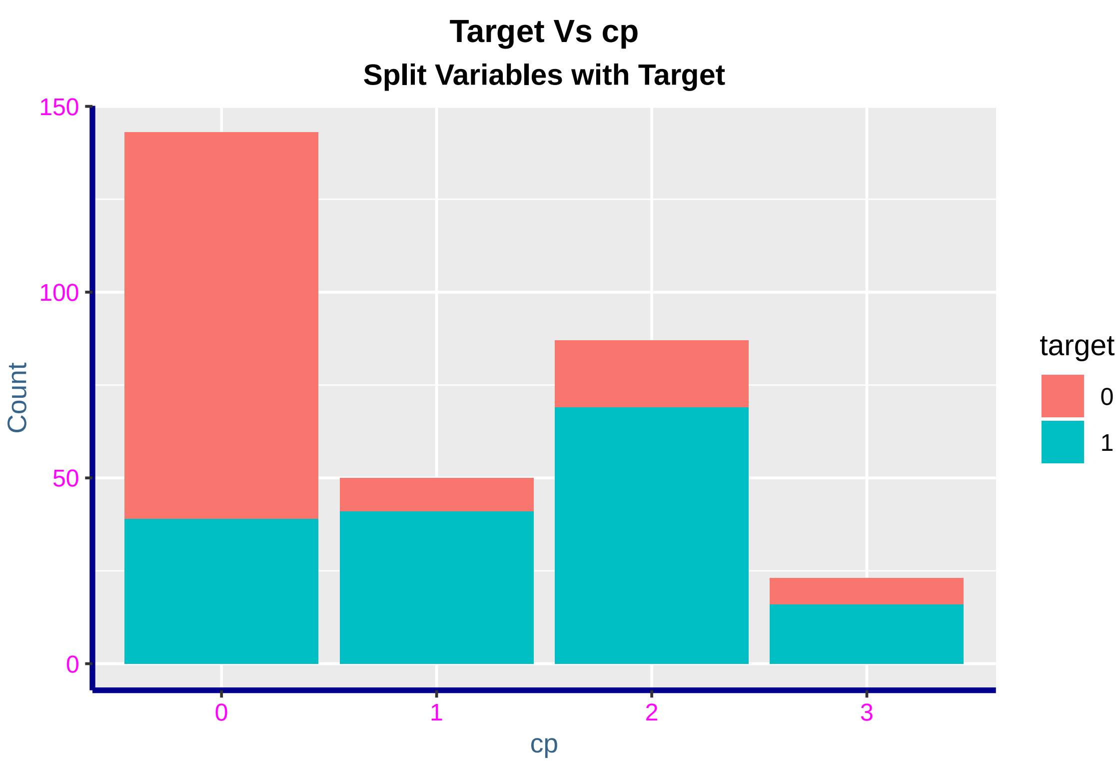
8.4 `target` Vs `slope`



8.5 `target` Vs `ca`

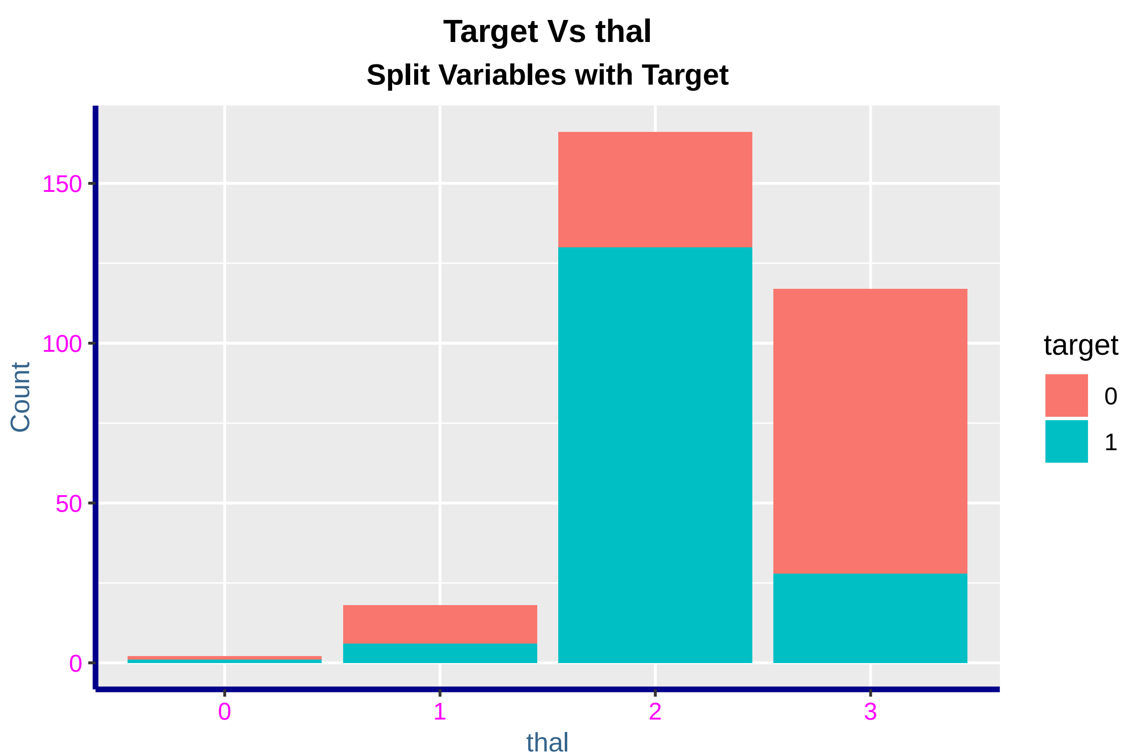


8.6 `target` Vs `cp`



8.7 `target` Vs `restecg`

8.8 `target` Vs `thal`



8.9 `target` Vs `age`

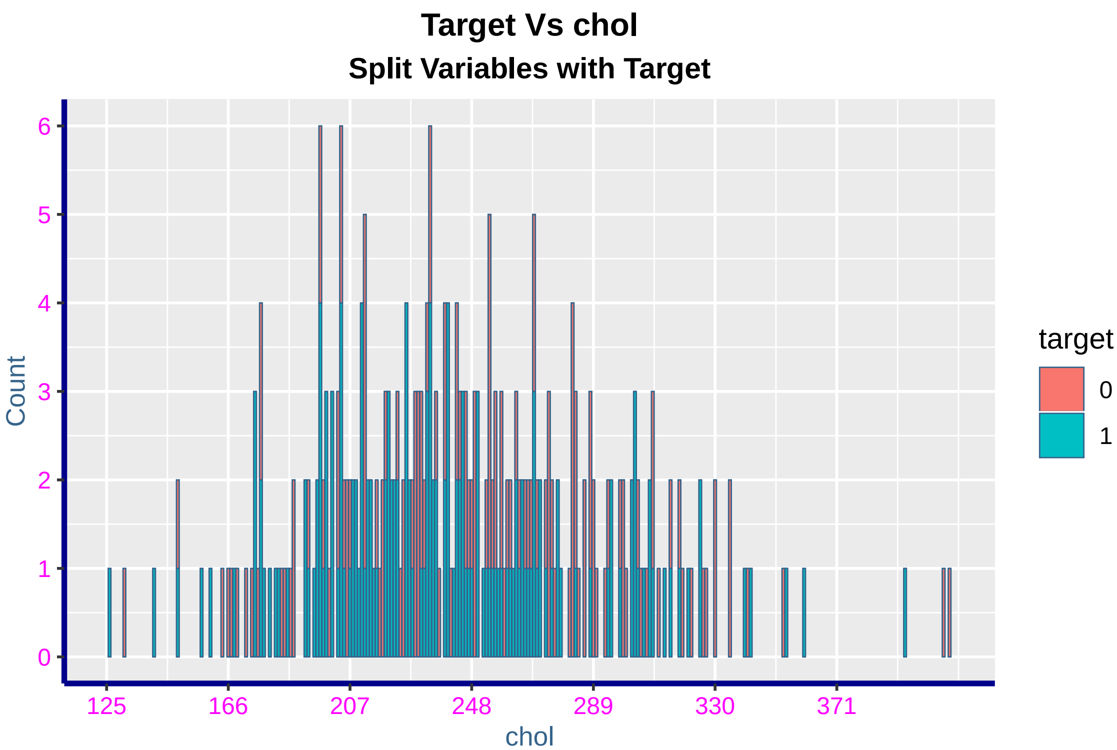
9

10 ————————————————————- Target Vs age

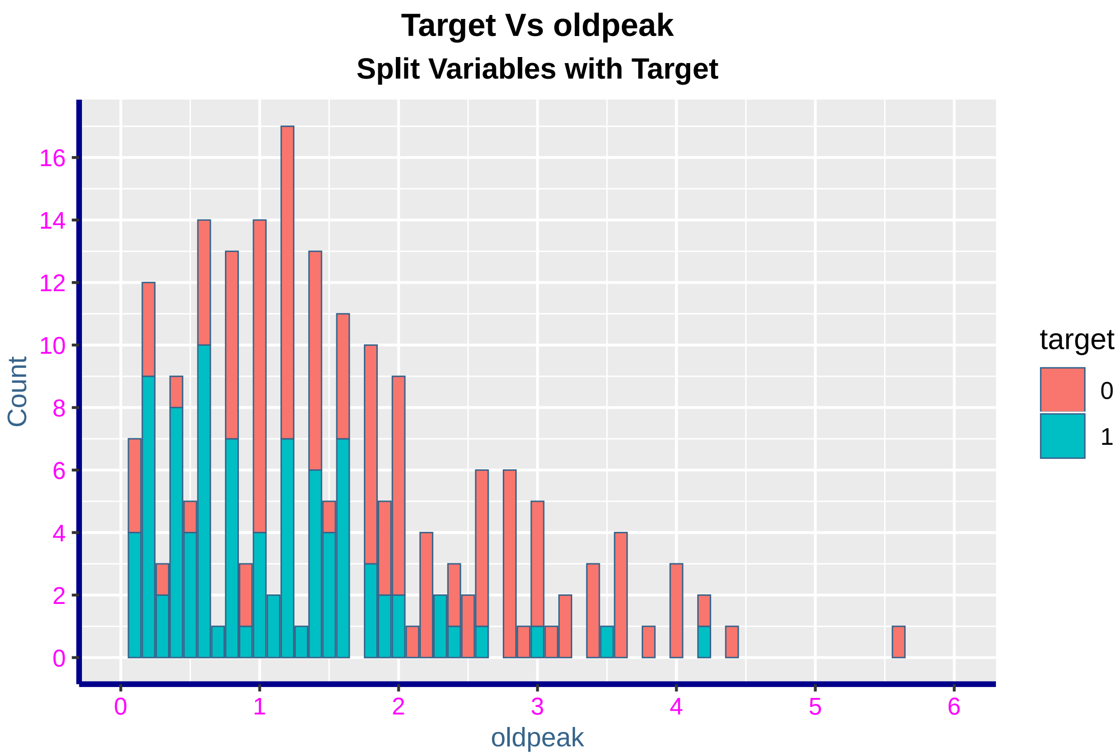
target.age <- group.target %>% dplyr::count(age) target.age$numeric < -as.numeric(as.character(target.age$age)) graph.target.geom.bar(target.age, 'age', 'Target Vs age', 'Split Variables with Target', 'age', 'Count', 'continuous', c(35, 80), c(0, 20))

"'

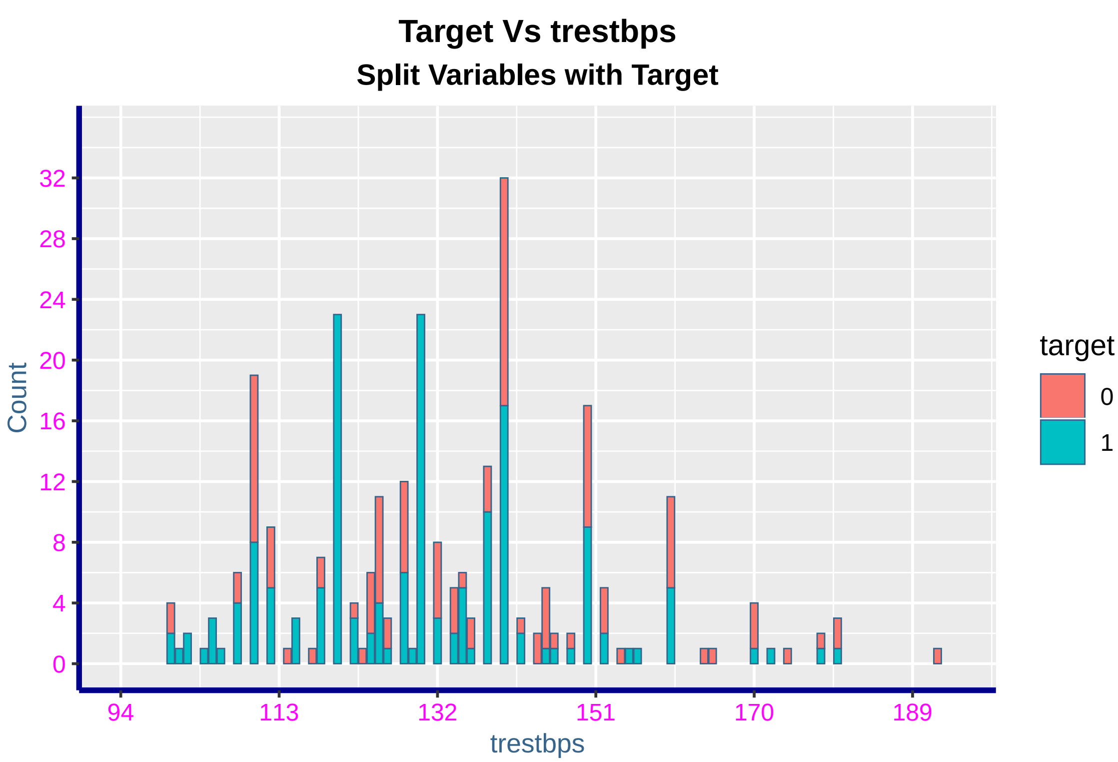
10.1 `target` Vs `chol`



10.2 `target` Vs `oldpeak`



10.3 `target` Vs `trestbps`



10.4 `target` Vs `thalach`