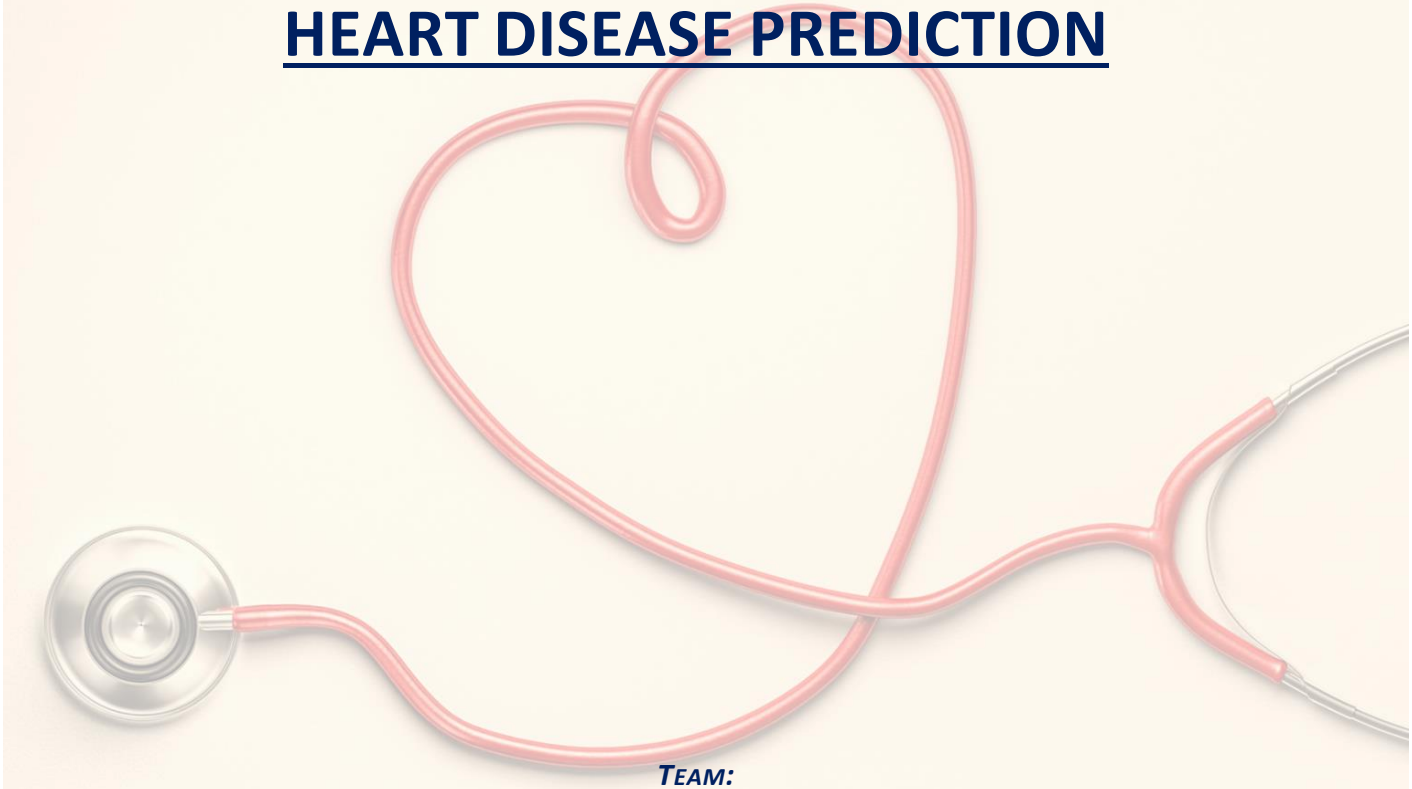
PROJECT REPORT

HEART DISEASE PREDICTION



TEAM:

SATYAM VATTS (100368408)
AVNEET KAUR (100368395)
KIRAN VIKRAM ANGA (100368125)
RADHIKA MAINI (100340257)

Data Overview

The analysis has been done on the Heart Dataset which consists of **303** observations, **13** predictors and **1** binary Response variable indicating if the person has heart disease or not. The dataset is taken from UCI Machine Learning Repository which originally constituted 76 attributes. However, all the published experiments refer to using subset of only 14 variables, so we have also used the same for our analysis. Source of the database is V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation. The variables are described below:

	Variable	*Datatype*	*Description*
1	Age	Quantitative	age of the patient [years]
2	Sex	Categorical	sex of the patient
3	cp	Categorical	chest pain type [0: Typical Angina, 1: Atypical Angina, 2: Non-Anginal Pain, 3: Asymptomatic]
4	tresbps	Quantitative	resting blood pressure [mm Hg]
5	chol	Quantitative	serum cholesterol [mm/dl]
6	fbs	Categorical	fasting blood sugar [1: if fbs > 120 mg/dl, 0: otherwise]
7	restecg	Categorical	resting electrocardiogram results [0: Normal, 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8	Thalach	Quantitative	maximum heart rate achieved
9	Exang	Categorical	exercise-induced angina [1: Yes, 0: No]
10	Oldpeak	Quantitative	ST depression induced by exercise relative to rest
11	Ca	Categorical	Number of major vessels [0 to 3] colored by fluoroscopy
12	Thal	Categorical	defect type [1: Normal, 2: Fixed Defect, 3: Reversible defect
13	Slope	Categorical	the slope of the peak exercise ST segment [0: up sloping, 1: flat, 2: down sloping]
14	target	Categorical	output class [1: heart disease, 0: Normal]

Data Pre-processing and Exploratory Data Analysis

Summary statistics of all the variables is shown below:

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

```
```{r}
sum(is.na(heart))
```

[1] 0
```

```
```{r}
dim(distinct(heart))

heart <- unique(heart)
```

[1] 302 14
```

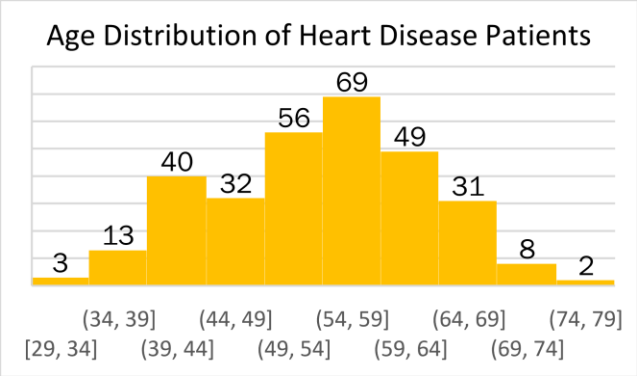
Observations:

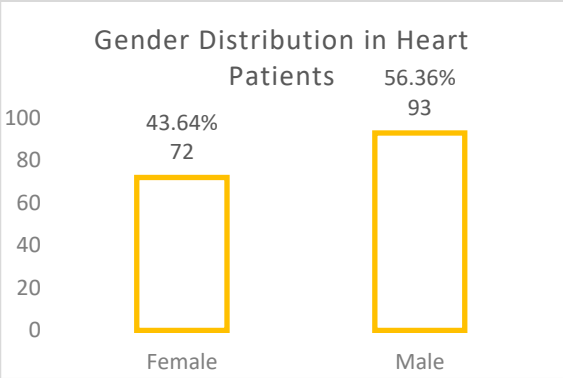
- ca : The value of 4 is invalid as data description says that the value for ca lies between 0 to 3.
- thal : 0 is an invalid value as it can either be 1 that is Normal, 1 Fixed defect, 2 Reversible Defect.
- There were no missing values
- 1 duplicate row as there are 302 distinct rows instead of 303

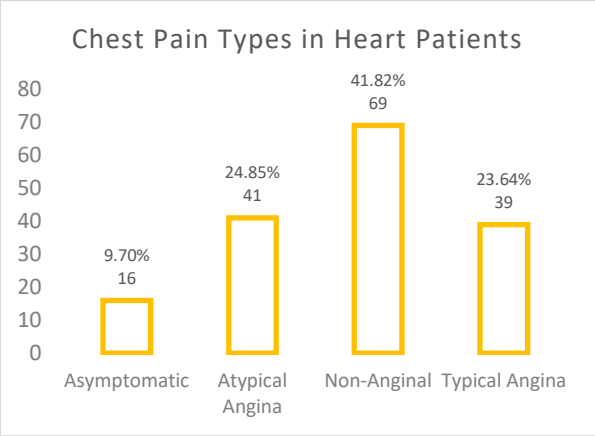
Thus, rows consisting invalid values of ca, thal and the duplicate row were dropped from the dataset.
The cleaned and final dataset now had **296** observations and **14** variables.

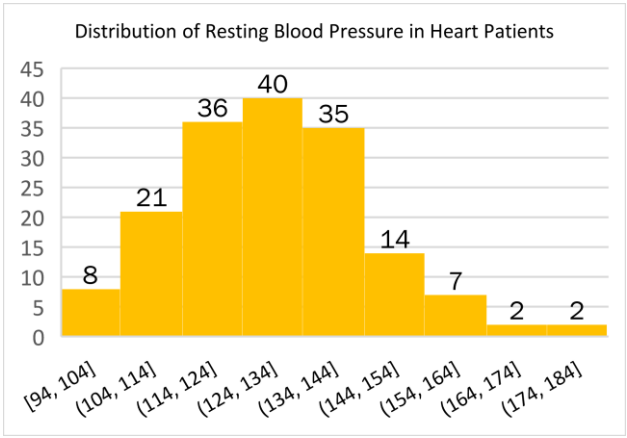
```
'data.frame':   296 obs. of  14 variables:
 $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ 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/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ thal    : Factor w/ 3 levels "1","2","3": 1 2 2 2 2 1 2 3 3 2 ...
 $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
```

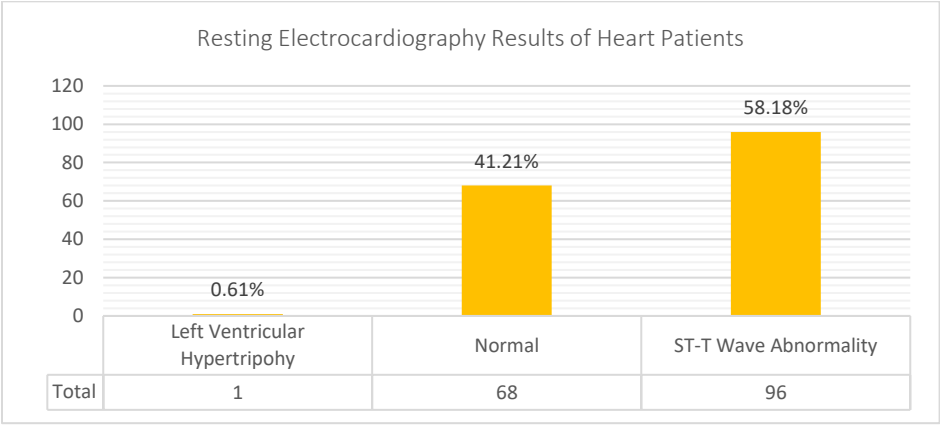
Distribution of few variables is illustrated below:











Observations:

A majority of Heart Patients

- Are in the age bracket of **50 to 60** years
- Are male
- Experience Non-Anginal Chest Pain
- Showed resting blood pressure between 114 to 144
- Showed ST-T wave abnormality in resting electrocardiography results

Research Objective

Cardiovascular Disease is the leading cause of deaths globally. Every year 17.9 million people die due to the cardiovascular diseases. Heart attack which is also known as myocardial infraction happens when the muscles in the heart don't get enough blood . The more the time passes to restore the blood flow , the more damage it can cause to the heart muscle . It requires early detection and management wherein machine learning model could be of great use. So ,the objective of the research is to **detect symptoms and clinical tests can be used to diagnose heart disease at an early stage**. We'll perform Logistic Regression as we are trying to predict the presence or absence of heart disease in patients – a binomial variable.

Inferential Analysis

T – Test for Numerical Variables

We are trying to find if there are any differences between the groups of patients based on presence or absence of heart disease. It helps us to identify the variables that show significant difference between the two groups based on mean value – for patients with and without heart disease.

$$H_0 : \textit{Mean of the two groups are equal}$$

$$H_a : \textit{Mean of the two groups are unequal}$$

Before performing T test, we looked whether the variances within groups were equal or not to perform appropriate T-Test:

```
        F test to compare two variances

data:  subset(heart, target == 0)$trestbps and subset(heart, target == 1)$trestbps
F = 1.3265, num df = 135, denom df = 159, p-value = 0.08677
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9599289 1.8412447
sample estimates:
ratio of variances
          1.326549


        F test to compare two variances

data:  subset(heart, target == 0)$chol and subset(heart, target == 1)$chol
F = 0.85308, num df = 135, denom df = 159, p-value = 0.3414
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6173151 1.1840754
sample estimates:
ratio of variances
          0.8530829


        F test to compare two variances

data:  subset(heart, target == 0)$thalach and subset(heart, target == 1)$thalach
F = 1.4229, num df = 135, denom df = 159, p-value = 0.03262
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.029668 1.975012
sample estimates:
ratio of variances
          1.422923
```

```
        F test to compare two variances

data:  subset(heart, target == 0)$oldpeak and subset(heart, target == 1)$oldpeak
F = 2.7385, num df = 135, denom df = 159, p-value = 1.558e-09
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.981635 3.800986
sample estimates:
ratio of variances
          2.738471


        F test to compare two variances

data:  subset(heart, target == 0)$age and subset(heart, target == 1)$age
F = 0.68829, num df = 135, denom df = 159, p-value = 0.02569
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4980644 0.9553400
sample estimates:
ratio of variances
         0.6882875
```

Observations:

- Variances are equal for tresbps and chol
- Variances are unequal for thalach, oldpeak and age

```
        Two Sample t-test

data:  subset(heart, target == 0)$trestbps and subset(heart, target == 1)$trestbps
t = 2.5823, df = 294, p-value = 0.0103
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.257825 9.318646
sample estimates:
mean of x mean of y
 134.4632  129.1750


        Two Sample t-test

data:  subset(heart, target == 0)$chol and subset(heart, target == 1)$chol
t = 1.3163, df = 294, p-value = 0.1891
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.946467 19.885438
sample estimates:
mean of x mean of y
 251.4632  243.4938
```

```
        Welch Two Sample t-test

data:  subset(heart, target == 0)$thalach and subset(heart, target == 1)$thalach
t = -7.9747, df = 264.36, p-value = 4.628e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -24.48009 -14.78535
sample estimates:
mean of x mean of y
 138.9485  158.5813


        Welch Two Sample t-test

data:  subset(heart, target == 0)$oldpeak and subset(heart, target == 1)$oldpeak
t = 7.8363, df = 214.28, p-value = 2.139e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.749953 1.254018
sample estimates:
mean of x mean of y
 1.600735  0.598750


        Welch Two Sample t-test

data:  subset(heart, target == 0)$age and subset(heart, target == 1)$age
t = 4.0279, df = 293.84, p-value = 7.17e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.092393 6.090695
sample estimates:
mean of x mean of y
 56.73529  52.64375
```

Observations :

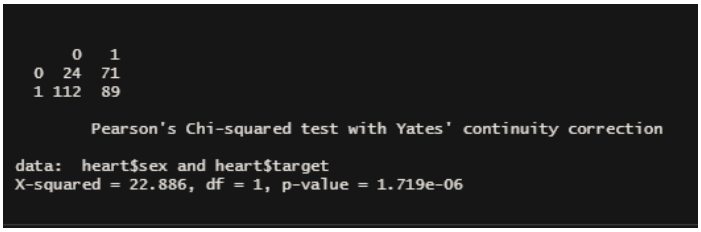
- Chol : p-value = 0.18 which is greater than 0.05. Therefore, we can say there is no difference in the means of variable chol within the groups of people with and without heart disease. Hence, chol variable is insignificant for further analysis.
- For rest of the variables, p -value are less than 0.05. Thus, they are considered significant.

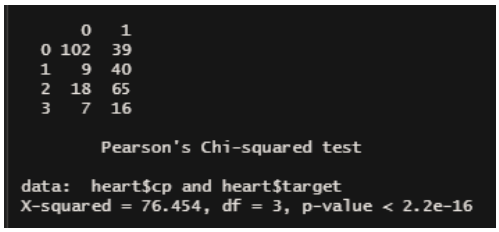
Chi-square Test for categorical variables

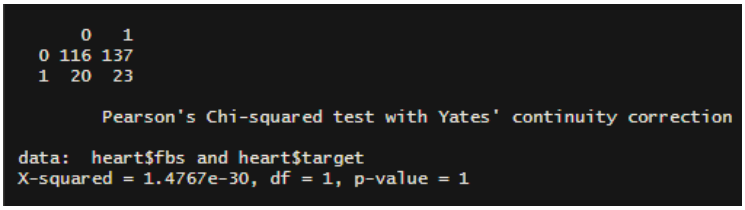
The Chi-Square Test was conducted to find whether the groups based on presence or absence of the heart diseases are independent or interdependent of various factors present in dataset that are categorical in nature . It will help us to identify variables with any association with the response variable.

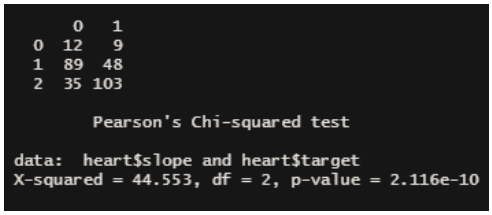
H_0 There is no association between the variables, i.e., they are independent

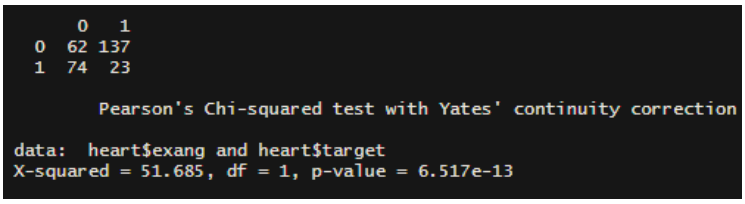
H_a :The variables are associated i.e. , they are not independent

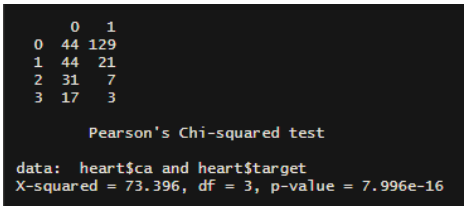


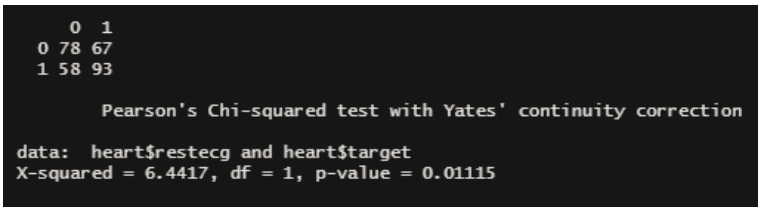


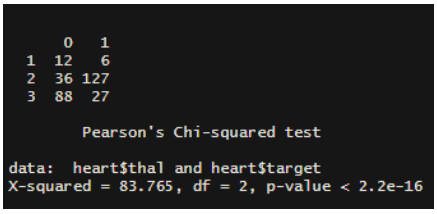












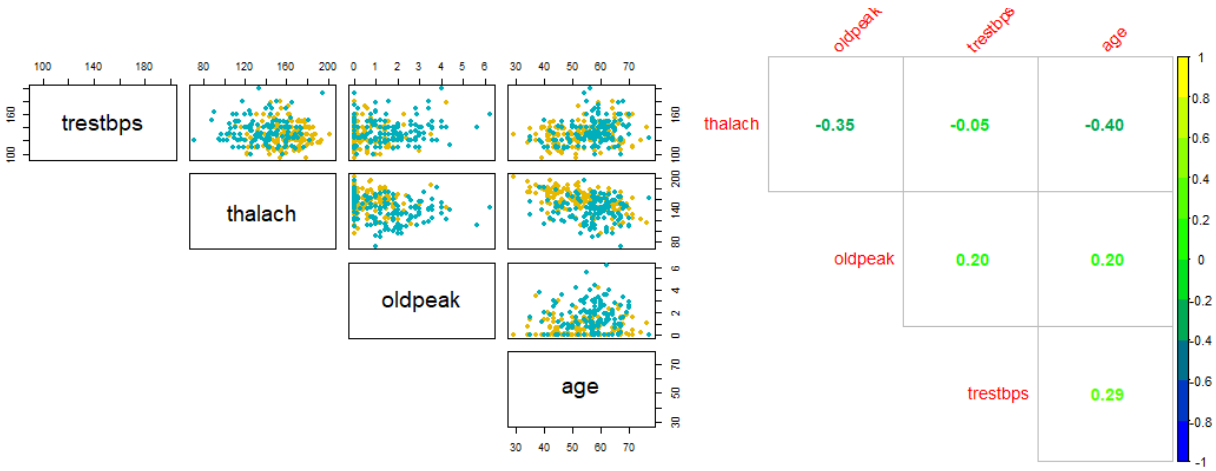
Observations:

- The variable **fbs** has a p-value of 1. This indicates that we must accept the null hypothesis and conclude that there is no significant association between the Fasting Blood Sugar Levels and person suffering for heart disease or not. Thus, it is not a significant predictor and is not considered for further analysis.
- All other categorical factors show association with the response. Thus, are important for making predictions of response.

Correlation Analysis

Correlation analysis is performed to address the issue of multicollinearity between the predictors. Multicollinearity causes unstable prediction model. Thus, we need to check and eliminate any such issue.

In order to verify whether the numerical variables are at all related, we have developed a correlation plot with correlation coefficients as shown below:



Observations:

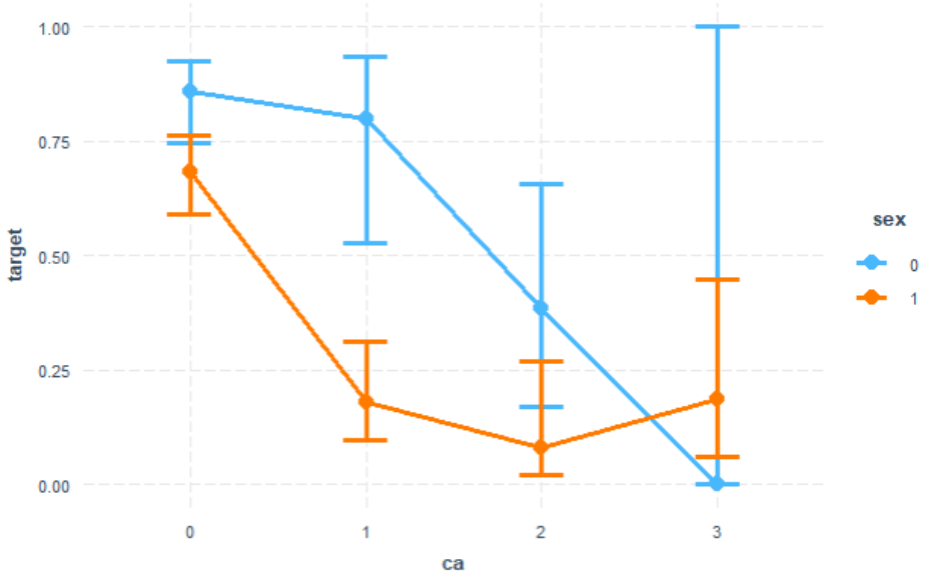
- The correlation coefficients of all the combinations is too less to be considered significant. The highest correlation coefficient is -0.4 which is still less than -0.5.
- The scatterplots verify the notion of uncorrelated predictors as there is no linear pattern for any of the variable combination.

Thus, we conclude that our dataset doesn't have multicollinearity issue.

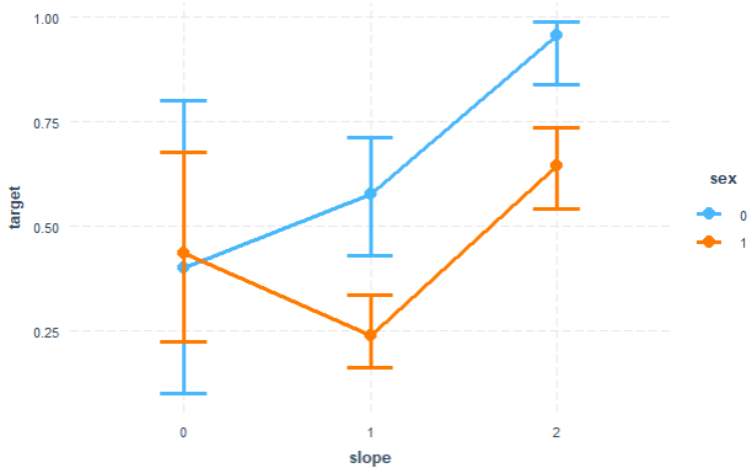
Interaction Analysis

The presence of interaction between predictors plays an important role in building prediction model. It is caused when the effect of a variable is different for different values of another variable. In this analysis, we analysed interaction between all the combination of predictors. Based on our results, three interactions were found between predictors as given below:

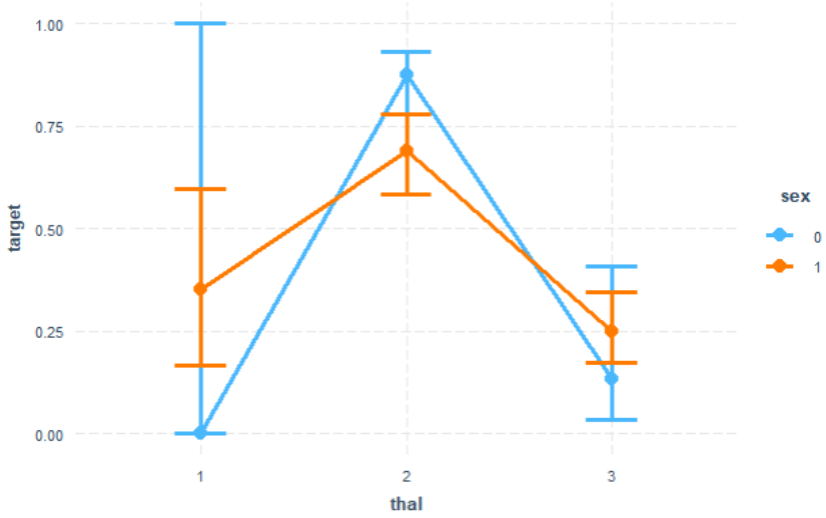
- ca and sex



- slope & sex



- thal & sex



Observations:

- The predicted probabilities are plotted against the values of categorical variables. The intersection between the lines indicated presence of interaction.
- ca:sex, slope:sex & thal:sex are the three interaction terms identified.

As the interaction between thal and sex is high(cuts the plot twice), we have taken this interaction term for the rest of the analysis & to perform comparison between models.

Regression Analysis

Ultimately, the selected set of predictors are taken to develop a Logistic Regression Model. This model aims at predicting the presence or absence of heart disease in patents based on their health status determined by different factors.

Data Splitting:

Before proceeding with model development, the dataset is split into two sets:

- Training Set: 207 Observations
- Test Set: 89 Observations

This split of 7:3 ratio is needed to validate the accuracy of the model.

Variable Selection: Stepwise regression based on AIC

Although, the inferential analysis helped us remove some of the insignificant predictors, we have performed Stepwise Logistic regression using **stepAIC** method in R. This method helped us generate a model with optimal set of predictors selected based on forward selection as well as backward elimination.

We performed the analysis for with and without the preselected interaction term, but the function resulted in the same regression model, one without any interaction term. PFB the results of the analysis:

Without Interaction

```
Start:  AIC=142.41
target ~ sex + cp + trestbps + restecg + thalach + exang + oldpeak +
    slope + ca + thal + age

          Df Deviance    AIC
- exang    1   106.42 140.42
- age      1   106.42 140.42
- oldpeak  1   107.25 141.25
- thalach  1   107.27 141.27
- trestbps 1   107.64 141.65
<none>         106.41 142.41
- restecg  1   108.95 142.95
- slope    2   111.59 143.59
- sex      1   111.01 145.01
- thal     2   122.88 154.88
- cp       3   128.25 158.25
- ca       3   130.46 160.46

Step:  AIC=140.42
target ~ sex + cp + trestbps + restecg + thalach + oldpeak +
    slope + ca + thal + age

          Df Deviance    AIC
- age      1   106.43 138.43
- thalach  1   107.27 139.27
- oldpeak  1   107.28 139.28
- trestbps 1   107.65 139.65
<none>         106.42 140.42
- restecg  1   108.95 140.95
- slope    2   111.63 141.63
+ exang    1   106.41 142.41
- sex      1   111.05 143.05
- thal     2   123.45 153.45
- ca       3   130.51 158.51
- cp       3   130.58 158.58
```

```
Step:  AIC=138.43
target ~ sex + cp + trestbps + restecg + thalach + oldpeak +
    slope + ca + thal

          Df Deviance    AIC
- oldpeak  1   107.29 137.29
- thalach  1   107.56 137.56
- trestbps 1   107.79 137.79
<none>         106.43 138.43
- restecg  1   108.95 138.95
- slope    2   111.67 139.67
+ age      1   106.42 140.42
+ exang    1   106.42 140.42
- sex      1   111.09 141.09
- thal     2   123.54 151.54
- cp       3   130.70 156.70
- ca       3   133.63 159.63
```

With Interaction: Sex:Thal

```
Start:  AIC=145.8
target ~ cp + trestbps + restecg + thalach + exang + oldpeak +
    slope + ca + age + sex + thal + sex:thal

          Df Deviance    AIC
- sex:thal  2   106.41 142.41
- age       1   105.80 143.80
- exang     1   105.82 143.82
- oldpeak   1   106.66 144.66
- thalach   1   106.68 144.68
- trestbps  1   106.91 144.91
<none>          105.80 145.80
- restecg   1   108.20 146.20
- slope     2   110.85 146.85
- cp        3   127.85 161.85
- ca        3   130.13 164.13

Step:  AIC=142.41
target ~ cp + trestbps + restecg + thalach + exang + oldpeak +
    slope + ca + age + sex + thal

          Df Deviance    AIC
- exang     1   106.42 140.42
- age       1   106.42 140.42
- oldpeak   1   107.25 141.25
- thalach   1   107.27 141.27
- trestbps  1   107.64 141.65
<none>          106.41 142.41
- restecg   1   108.95 142.95
- slope     2   111.59 143.59
- sex       1   111.01 145.01
+ sex:thal  2   105.80 145.80
- thal      2   122.88 154.88
- cp        3   128.25 158.25
- ca        3   130.46 160.46
```

```
Step:  AIC=140.42
target ~ cp + trestbps + restecg + thalach + oldpeak + slope +
    ca + age + sex + thal

          Df Deviance    AIC
- age       1   106.43 138.43
- thalach   1   107.27 139.27
- oldpeak   1   107.28 139.28
- trestbps  1   107.65 139.65
<none>          106.42 140.42
- restecg   1   108.95 140.95
- slope     2   111.63 141.63
+ exang     1   106.41 142.41
- sex       1   111.05 143.05
+ sex:thal  2   105.82 143.82
- thal      2   123.45 153.45
- ca        3   130.51 158.51
- cp        3   130.58 158.58
```

```
Step:  AIC=137.29
target ~ sex + cp + trestbps + restecg + thalach + slope + ca +
    thal

           Df Deviance    AIC
- trestbps  1   108.44 136.44
- thalach   1   108.54 136.54
<none>          107.29 137.29
- restecg   1   109.62 137.62
+ oldpeak   1   106.43 138.43
+ exang     1   107.26 139.26
+ age       1   107.28 139.28
- sex       1   112.36 140.37
- slope     2   116.40 142.40
- thal      2   125.09 151.09
- cp        3   131.59 155.59
- ca        3   138.94 162.94
```

```
Step:  AIC=138.43
target ~ cp + trestbps + restecg + thalach + oldpeak + slope +
    ca + sex + thal

           Df Deviance    AIC
- oldpeak   1   107.29 137.29
- thalach   1   107.56 137.56
- trestbps  1   107.79 137.79
<none>          106.43 138.43
- restecg   1   108.95 138.95
- slope     2   111.67 139.67
+ age       1   106.42 140.42
+ exang     1   106.42 140.42
- sex       1   111.09 141.09
+ sex:thal  2   105.83 141.83
- thal      2   123.54 151.54
- cp        3   130.70 156.70
- ca        3   133.63 159.63
```

```
Step:  AIC=136.44
target ~ sex + cp + restecg + thalach + slope + ca + thal

           Df Deviance    AIC
- thalach   1   109.78 135.78
<none>          108.44 136.44
- restecg   1   111.05 137.05
+ trestbps  1   107.29 137.29
+ oldpeak   1   107.79 137.79
+ age       1   108.33 138.33
+ exang     1   108.41 138.41
- sex       1   112.87 138.87
- slope     2   116.77 140.77
- thal      2   127.91 151.91
- cp        3   131.87 153.87
- ca        3   141.45 163.45

Step:  AIC=135.78
target ~ sex + cp + restecg + slope + ca + thal

           Df Deviance    AIC
<none>          109.78 135.78
- restecg   1   112.19 136.19
+ thalach   1   108.44 136.44
+ trestbps  1   108.54 136.54
+ oldpeak   1   108.92 136.92
+ age       1   109.18 137.18
+ exang     1   109.63 137.63
- sex       1   114.31 138.31
- slope     2   121.61 143.61
- thal      2   129.77 151.77
- cp        3   138.72 158.72
- ca        3   145.68 165.68

Call:  glm(formula = target ~ sex + cp + restecg + slope + ca + thal,
    family = binomial, data = heart_v1.train)
```

```
Coefficients:
(Intercept)         sex1          cp1          cp2          cp3     restecg1      slope1      slope2
    -0.05052     -1.45517      1.31645      3.34017      1.33195      0.81317      0.25667      2.11287
        ca1          ca2          ca3         thal2         thal3
    -2.43758     -3.31379     -3.75888      0.84313     -1.62893

Degrees of Freedom: 206 Total (i.e. Null);  194 Residual
Null Deviance:      285.6
Residual Deviance: 109.8        AIC: 135.8
```

```
Step:  AIC=137.29
target ~ cp + trestbps + restecg + thalach + slope + ca + sex +
    thal

           Df Deviance    AIC
- trestbps  1   108.44 136.44
- thalach   1   108.54 136.54
<none>          107.29 137.29
- restecg   1   109.62 137.62
+ oldpeak   1   106.43 138.43
+ exang     1   107.26 139.26
+ age       1   107.28 139.28
- sex       1   112.36 140.37
+ sex:thal  2   106.67 140.68
- slope     2   116.40 142.40
- thal      2   125.09 151.09
- cp        3   131.59 155.59
- ca        3   138.94 162.94

Step:  AIC=136.44
target ~ cp + restecg + thalach + slope + ca + sex + thal

           Df Deviance    AIC
- thalach   1   109.78 135.78
<none>          108.44 136.44
- restecg   1   111.05 137.05
+ trestbps  1   107.29 137.29
+ oldpeak   1   107.79 137.79
+ age       1   108.33 138.33
+ exang     1   108.41 138.41
- sex       1   112.87 138.87
+ sex:thal  2   107.68 139.68
- slope     2   116.77 140.77
- thal      2   127.91 151.91
- cp        3   131.87 153.87
- ca        3   141.45 163.45
```

```
Step:  AIC=135.78
target ~ cp + restecg + slope + ca + sex + thal

           Df Deviance    AIC
<none>          109.78 135.78
- restecg   1   112.19 136.19
+ thalach   1   108.44 136.44
+ trestbps  1   108.54 136.54
+ oldpeak   1   108.92 136.92
+ age       1   109.18 137.18
+ exang     1   109.63 137.63
- sex       1   114.31 138.31
+ sex:thal  2   109.05 139.05
- slope     2   121.61 143.61
- thal      2   129.77 151.77
- cp        3   138.72 158.72
- ca        3   145.68 165.68

Call:  glm(formula = target ~ cp + restecg + slope + ca + sex + thal,
    family = binomial, data = heart_v1.train)
```

```
Coefficients:
(Intercept)          cp1          cp2          cp3     restecg1      slope1      slope2          ca1
    -0.05052      1.31645      3.34017      1.33195      0.81317      0.25667      2.11287     -2.43758
        ca2          ca3         sex1         thal2         thal3
    -3.31379     -3.75888     -1.45517      0.84313     -1.62893

Degrees of Freedom: 206 Total (i.e. Null);  194 Residual
Null Deviance:      285.6
Residual Deviance: 109.8        AIC: 135.8
```

Observations:

- It can be observed that both the calls to stepAIC function i.e., with or without interaction result in same final model with 6 predictors namely sex, cp, restecg, slope, ca & thal.
- The final model AIC is 135.8 with a Null Deviance of 285.6.

Thus, it can be seen that the interaction term is not significant for the prediction model. To appreciate this result, we will consider a model with interaction term for further analysis to compare it with the model without any interaction term.

Without Interaction Model : target ~ sex + cp + slope + ca + thal

With Interaction model : target ~ sex + cp + slope + ca + thal + sex:thal

Model Comparison

The prediction models are compared to determine the significance of interaction term using two significance testing methods namely Wald Test and Log-Likelihood Ratio Test. The results of these tests are summarised below:

Model without Interaction

```
Call:
glm(formula = target ~ cp + slope + ca + sex + thal, family = binomial,
    data = heart_v1.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5683  -0.3020   0.0541   0.2942   3.1664

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.86639    1.43473   0.604 0.545931
cp1          1.41834    0.64964   2.183 0.029017 *
cp2          3.26633    0.71447   4.572 4.84e-06 ***
cp3          1.36726    0.86013   1.590 0.111928
slope1       0.09171    0.95279   0.096 0.923322
slope2       1.96605    0.95630   2.056 0.039793 *
ca1         -2.54751    0.67673  -3.764 0.000167 ***
ca2         -3.24430    0.88052  -3.685 0.000229 ***
ca3         -3.83801    1.25628  -3.055 0.002250 **
sex1        -1.47553    0.69519  -2.122 0.033798 *
thal2        0.42792    0.97674   0.438 0.661306
thal3       -1.94143    0.95112  -2.041 0.041231 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 285.57  on 206  degrees of freedom
Residual deviance: 112.19  on 195  degrees of freedom
AIC: 136.19

Number of Fisher Scoring iterations: 6
```

Model with Interaction

```
Call:
glm(formula = target ~ cp + slope + ca + sex + thal + sex:thal,
    family = binomial, data = heart_v1.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.64258  -0.30764   0.04911   0.31133   3.13659

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.43725 1455.39813  -0.009 0.993182
cp1           1.34337    0.65401   2.054 0.039971 *
cp2           3.25941    0.71091   4.585 4.54e-06 ***
cp3           1.33706    0.85470   1.564 0.117733
slope1        0.09605    0.94303   0.102 0.918877
slope2        1.95557    0.94293   2.074 0.038087 *
ca1          -2.55864    0.67966  -3.765 0.000167 ***
ca2          -3.22486    0.89913  -3.587 0.000335 ***
ca3          -3.80861    1.22937  -3.098 0.001948 **
sex1         11.89775 1455.39803   0.008 0.993477
thal2        13.94237 1455.39786   0.010 0.992357
thal3        10.33203 1455.39864   0.007 0.994336
sex1:thal2  -13.67619 1455.39811  -0.009 0.992502
sex1:thal3  -12.24171 1455.39883  -0.008 0.993289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 285.57  on 206  degrees of freedom
Residual deviance: 111.24  on 193  degrees of freedom
AIC: 139.24

Number of Fisher Scoring iterations: 14
```

- ***Wald Test:*** The test was performed to check whether the interaction term is significant for the prediction model or not.

```
Wald test:
----------

Chi-squared test:
X2 = 0.7, df = 2, P(> X2) = 0.71
```

H0 : The reduced model is appropriate | Coefficient of Interaction term is 0
Ha: Full model is appropriate | Coefficient of Interaction term is not 0

It can be observed from the result that p-value of the test is greater than 0.05, thus, we fail to reject the null hypothesis and conclude that interaction term is not significant i.e., the reduced model is appropriate.

- ***Log-Likelihood Ratio Test:*** The test was performed to verify the significance of interaction term i.e., to compare the model with & without interaction term. It is performed using the Anova function in R.
  H0: Reduced model is appropriate
  Ha: Full model is appropriate

```
[1] 0.6218851
Analysis of Deviance Table

Model 1: target ~ cp + slope + ca + sex + thal
Model 2: target ~ cp + slope + ca + sex + thal + sex:thal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       195     112.19
2       193     111.24  2  0.94794   0.6225
```

  It can be observed that p-value of chi-sq statistics obtained using the difference of residual deviances of two models is greater than 0.05 indicating that the model without interaction term is more appropriate i.e., the interaction term is not significant predictor to be considered.

The model comparison performed using Wald and Likelihood Ratio test both indicated that the interaction term is not significant for the prediction model.

Prediction Accuracy

The accuracy of model is an important metric to verify how accurate the predictions are on a testing dataset given that the model is trained on a training dataset. The accuracy is determined using two metrics, a classification report specifying confusion matrix and the other using ROC Curve. We have performed the analysis for both the without and with interaction terms model to compare. The threshold value of predicted probabilities is taken as the average of response variable in the whole data set which has a value of 0.54. Thus, probabilities more than 0.54 are considered as 1 and rest as 0.

Classification Report

The classification report is a detailed description of predictions made on test dataset and how the model performs on a new set of data.

Model without Interaction:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 31 13
         1 10 35

               Accuracy : 0.7416
                 95% CI : (0.6379, 0.8286)
    No Information Rate : 0.5393
    P-Value [Acc > NIR] : 6.965e-05

                  Kappa : 0.4827

 Mcnemar's Test P-Value : 0.6767

            Sensitivity : 0.7292
            Specificity : 0.7561
         Pos Pred Value : 0.7778
         Neg Pred Value : 0.7045
             Prevalence : 0.5393
         Detection Rate : 0.3933
   Detection Prevalence : 0.5056
      Balanced Accuracy : 0.7426

       'Positive' Class : 1
```

Observations:

- The model has an overall accuracy of 74%.
- Out of 89 test data observations, 23 were misclassified.
- The Sensitivity of detecting the positive cases i.e., people with heart disease is 72.9%
- The Specificity of detecting negative cases is 75.6%

Thus, the model is acceptable as it performs moderately good. The accuracy can be further increase if the number of training observations are more making it more robust.

Model with Interaction:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 31  9
         1 10 39

               Accuracy : 0.7865
                 95% CI : (0.6869, 0.8663)
    No Information Rate : 0.5393
    P-Value [Acc > NIR] : 1.112e-06
```

```
                  Kappa : 0.5696

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8125
            Specificity : 0.7561
         Pos Pred Value : 0.7959
         Neg Pred Value : 0.7750
             Prevalence : 0.5393
         Detection Rate : 0.4382
   Detection Prevalence : 0.5506
      Balanced Accuracy : 0.7843

       'Positive' Class : 1
```

Observations:

- The model has an overall accuracy of 78.6%.
- Out of 89 test data observations, 19 were misclassified.
- The Sensitivity of detecting the positive cases i.e., people with heart disease is 81.2%
- The Specificity of detecting negative cases is 75.6%

It can be observed that the accuracy of model with interaction is more as compared to without interaction. This result is contradictory to our earlier analysis where we discovered and verified that interaction term is not significant. However, since the dataset is small, presence of more parameters in the model allows it fit more rigidly to the test dataset. Also, the training set is small, the observations from accuracy result are contradictory.

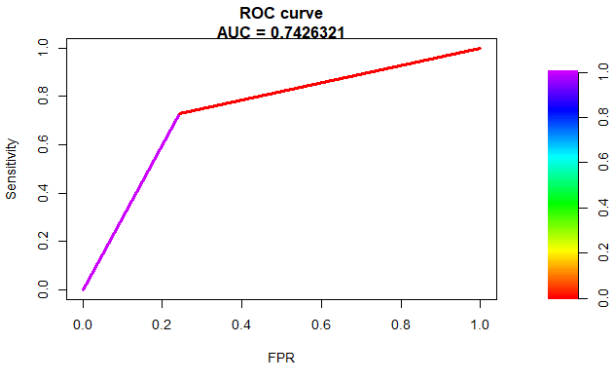
The results signifying the insignificance of interaction term can be better appreciated if we have large dataset and we can perform cross validation for different sample test dataset finally averaging out the accuracy.

ROC Curves:

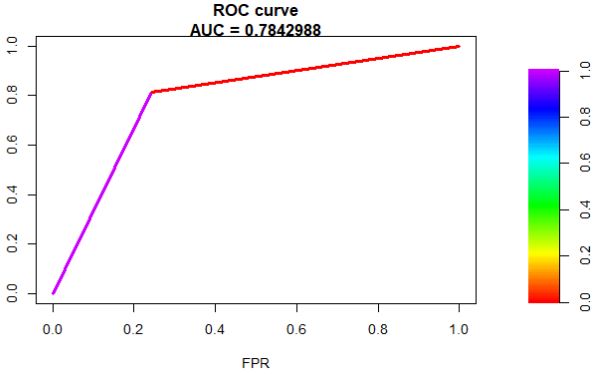
The ROC curves provide a more powerful metric to determine the accuracy of the model as they take into the account all possible values of threshold values and plot a graph between the sensitivity and specificity of the predictive model.

The plot has been made for both with and without interaction models for comparison.

Model without Interaction



Model with Interaction



Observations:

- The Area Under Curve signifying the accuracy of model is 0.7426 for model without interaction.
- The model with interaction has a higher AUC of 0.7842.

It can be observed that here again the predictive power of model with interaction is more than without interaction. As mentioned earlier, this is due to limited data available.

Fitness Analysis

Lastly, we analysed the fitness of model using Hosmer-Lemeshow Test to verify how these models compare to the saturated model and determine if they are a good fit or lacks appropriate level of fit to be considered fruitful for predictions.

H0 : The reduced model is fitted
Ha : The saturated model is fitted

Model without Interaction

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fit_wo_int$y, fitted(fit_wo_int)
X-squared = 12.658, df = 8, p-value = 0.1242
```

Model with Interaction

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fit_int$y, fitted(fit_int)
X-squared = 10.723, df = 8, p-value = 0.2179
```

The p-value of test for the models is greater than 0.05, therefore, we fail to reject the null hypothesis. The results specify that both the models are a good fit and thus can be considered for prediction of patients with or without heat disease.

Conclusion

This report detailed the analysis on heart disease dataset for 303 patients categorizing as a heart disease patient or not based on various health parameters and test results. We treated the dataset to remove invalid values from ca & thal and removed duplicated data. Inferential analysis was performed to determine the variables that distinguish the two groups of people and thus will be significant using T-test and Chi-Square tests. We found that fbs(fasting blood sugar) and chol(cholesterol) are not significant and thus removed them from dataset. We performed correlation analysis to check multicollinearity and as it was found that there is no collinearity in the dataset. In order to check whether the predictors interact to have effect on response, we performed interaction analysis and discovered 3 pairs of variables showing interaction. We chose interaction of sex and thal for further analysis as it showed high interaction based on interaction plots. We then performed the regression analysis to develop our prediction model for both with interaction and without interaction. Later ,we analysed the accuracy of these models using classification report & ROC curves. Both analyses showed model with interaction performed better although this result is due to limited data availability and may change in presence of large data. Since, the interaction term was insignificant in all the tests, the accuracy metrics are contradictory and hence require further analysis using more data which is out of scope of this research. We also determined the goodness of fit for both models and got that both are a good fit.

Such a predictive model can be used by hospitals to better identify the status of a patient and can help in early diagnosis of heart disease. This can save lives by starting medication and treatment at an early stage of disease and thus increasing the chances of recovery.

APPENDIX

Sources

UCI Machine Learning Repository : https://archive.ics.uci.edu/ml/datasets/heart+disease

Kaggle Dataset: https://www.kaggle.com/ronitf/heart-disease-uci

R-Code

Attached for reference