Heart disease prediction

Sandy Vo - August 2022

Table of Contents

I. INTRODUCTION

1. Background information

According to CDC, heart diseases or cardiovascular diseases is the leading cause of death in the world generally and in the US particularly. Everyone regardless of age, gender, social class, racial and ethnic group all has equal risks of suffering from heart diseases. Every 36 seconds, there is one person dying from cardiovascular diseases. About 6.2 million people in the US suffer from heart diseases. Apart from that, each year from 2016 to 2017, heart disease costs the US about $363 billion. This amount of money includes medicines, healthcare services and loss of productivity due to heart diseases.

In terms of the world, about 31% of the deaths in the world in 2016 was resulted from heart diseases. As reported by WHO on June 11 2021, around around 17.9 million people died from heart diseases in 2019, accounting for 32% of all the death worldwide. Moreover, 38% out of 17 million premature deaths was cause by cardiovascular diseases

2. Why is the question important?

These aforementioned data shows that heart disease has causes tremendous loss not only in societies but also in economics. Hence, taking advantage of technology in general and the advancements in machine learning in particular to provide any treatment as well as early detection in heart diseases should be seriously taken in to account. As heart diseases is hard to detect since there are a lot of risk factors such as cholesterol level, pulse rate and blood pressure. Therefore, any factors that potentially cause heart diseases should be recorded to help detect the diseases as soon as possible.

3. Questions

- What factors directly account for heart diseases?
- Given information about their heath, does that person has heart diseases?

4. Data collection

I am going to use Heart Disease Dataset to develop an effective model which helps detect Cardiovascular diseases.
(https://www.kaggle.com/datasets/cherngs/heartdisease-cleveland-uci)

It has 76 features from 303 patients; however, published studies chose only 14 features that are relevant in predicting heart disease. Hence, here we will be using the dataset consisting of 303 patients with 14 features set.

No	Name	Type		Definition
		Qualitative	Quantitative	
1	age		x	Age in years
2	sex	x		Gender
3	cp	x		Chest pain type
4	trestbps		x	Resting blood pressure
5	chol		x	Serum cholestoral in mg/dl
6	fbs	x		Fasting blood sugar
7	restecg	x		Resting electrocardiographic results
8	thalach		x	Maximum heart rate achieved
9	exang	x		Exercise induced angina
10	oldpeak		x	ST depression induced by exercise relative to rest looks at stress of heart during exercise
11	slope	x		The slope of the peak exercise ST segment
12	ca	x		Number of major vessels (0-3) colored by flourosopy
13	thal	x		Thalium stress result
14	target	x		Suffer from heart disease

Below are the description of the variables which will be used in the project.

1. age (quantitative) - age in years

As stated by National Institute on Aging, adults who are over 65 are more likely to suffer from heart disease, which is problems with the heart, blood vessels, or both, than younger people. Aging can cause changes in the heart and blood vessels that may increase a risk of developing heart disease because your heart can't beat as fast during physical activity or times of stress as it did when you were younger.

2. sex (qualitative)

It is believed that men are at greater risk of heart disease than pre-menopausal women. According to some disputes in WHO and UN, once past menopause, it has been argued that a woman and a men has equal risk of suffering from cardiovascular diseases. Moreover, it is reported that if a female has diabetes, it is more likely for her to suffer heart disease compared to a male with diabetes.

0: male

1: female

3. cp (qualitative) - chest pain type

Chest pain is a type of discomfort 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, abdomen or back. Angina pain may even feel like indigestion. In this dataset, chest pain is categorized as:

- *0: Typical angina: chest pain related decrease blood supply to the heart*
- *1: Atypical angina: chest pain not related to heart*
- *2: Non-anginal pain: typically esophageal spasms (non heart related)*
- *3: Asymptomatic: chest pain not showing signs of disease*

4. trestbps (quantitative) - resting blood pressure (in mm Hg on admission to the hospital)

Anything above 130-140 is typically cause for concern

High blood pressure (BP) is one of the most important risk factors for cardiovascular disease, which is the leading cause of mortality. Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.

5. chol (quantitative) - serum cholestoral in mg/dl

Your body needs cholesterol to build healthy cells, but high levels of cholesterol can increase your risk of heart disease. With high cholesterol, you can develop fatty deposits in your blood vessels. Eventually, these deposits grow, making it difficult for enough blood to flow through your arteries

6. fbs (qualitative) - (fasting blood sugar > 120 mg/dl)

Over time, high blood sugar can damage blood vessels and the nerves that control your heart. An amount of falsting blood sugar over 126 mg/dL signals diabetes. Having both high blood pressure and diabetes can greatly increase your risk for heart disease. Too much LDL ("bad") cholesterol in your bloodstream can form plaque on damaged artery walls.

0: false

1: true

7. restecg (qualitative) - resting electrocardiographic results

An electrocardiogram (ECG) is a medical test that detects heart problems by measuring the electrical activity generated by the heart as it contracts. A doctor may recommend an ECG for people who may be at risk of heart disease because there is a family history of heart disease, or because they smoke, are overweight, have diabetes, high cholesterol or high blood pressure. A doctor may also recommend an ECG for people who are displaying symptoms such as chest pain, breathlessness, dizziness, fainting or fast or irregular heartbeats. In this dataset, the variable is divided into 3 values:

- *0: Nothing to note*
- *1: ST-T Wave abnormality*
  - can range from mild symptoms to severe problems
  - signals non-normal heart beat
- *2: Possible or definite left ventricular hypertrophy*
  - enlarged heart's main pumping chamber

8. thalach (quantitative) - maximum heart rate achieved

The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.

9. exang (qualitative) - exercise induced angina

Everyone, including people in excellent shape, can experience pain in their chest during exercise. A heart attack occurs when the coronary arteries become blocked. The blockage causes the heart to lose oxygen. A heart attack can cause pain in the jaw, back, chest, and other parts of the upper body. The pain may go away and return, or it may last longer than a few minutes. If a person does not receive treatment, the heart muscle can die.

0: no
1: yes

10. oldpeak (quantitative) - ST depression induced by exercise relative to rest looks at stress of heart during exercise. Unhealthy heart will stress more

Exercise induced ST segment depression is considered a reliable ECG finding for the diagnosis of obstructive coronary atherosclerosis. It has also been associated with a worse prognosis for patients with a documented coronary artery disease.

11. slope (qualitative) - the slope of the peak exercise ST segment

0: Upsloping: better heart rate with exercise (uncommon)
1: Flatsloping: minimal change (typical healthy heart)
2: Downsloping: signs of unhealthy heart

12. ca (qualitative) - number of major vessels (0-3) colored by flourosopy

Arteries carry blood away from the heart while veins carry blood into the heart. Colored vessel means the doctor can see the blood passing through. The vessels colored blue indicate the transport of blood with relatively low content of oxygen and high content of carbon dioxide. The vessels colored red indicate the transport of blood with relatively high content of oxygen and low content of carbon dioxide. The more blood movement, the better because of no clots.

13. thal (qualitative) - thalium stress result

The thallium stress test is an imaging study that shows your doctor how well blood flows to your heart. It measures your blood flow during rest and after exercise. The results of this test will tell you about the flow of blood to your heart through your coronary arteries. An abnormal test result can reveal coronary blockages as well as damage from heart attacks. It can has one of the 3 values:

1 : normal
2: fixed defect: used to be defect but fine now
3: reversible defect: no proper blood movement when exercising

14. target (qualitative) - Displays whether the individual is suffering from heart disease or not

0: no
1: yes

target is the response of the project.

II. DATA PREPROCESSING

1. Duplicate values

There is 2 identical rows in the dataset. After removing the duplicate values, the dataset has **302 observations.**

2. Errors in variables ca and thal

- **ca**

ca has 4 levels corresponding to 0, 1, 2 and 3. But the dataset fails to classify categorical variables. There is 4 rows in the dataset marked with 4. So I replace 4 with the median of the variable ca.

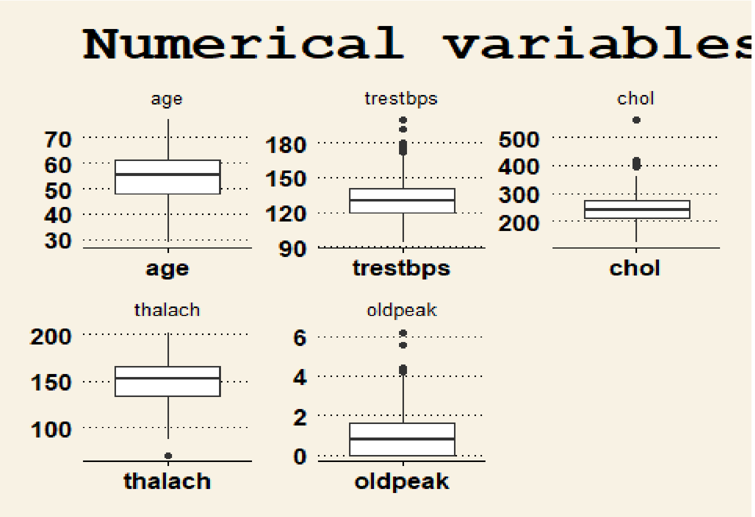
- **thal**

thal has 3 levels (1,2, and 3). But the dataset fails to classify categorical variables. There is 2 rows in the dataset marked with 0. So I replace 0 with the median of the variable **thal**.

3. Fail to classify categorical variables

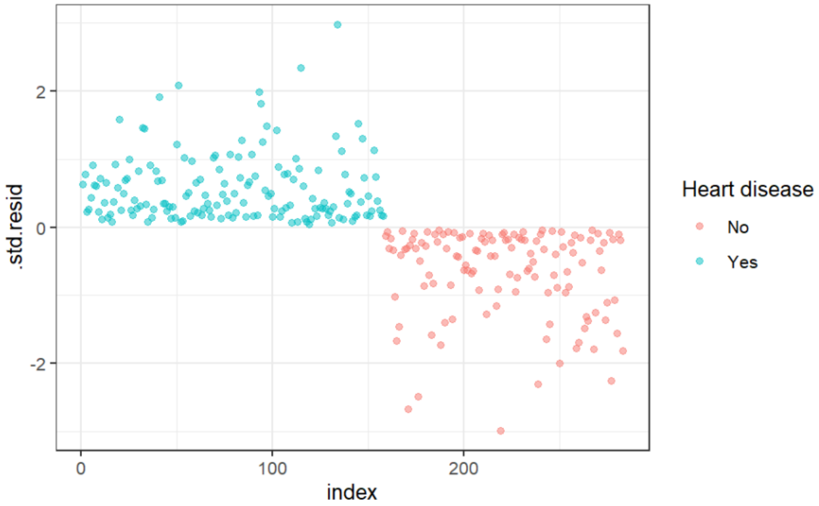
The dataset has 9 categorical variables; however, the dataset failed to classify them. They are all displayed in numeric values.

4. Outliers



Apart from **age**, the are some outliers that will potentially affect the accuracy of the prediction model. So we have to detele them.

After removing outliers, the dataset is left with 283 observations.

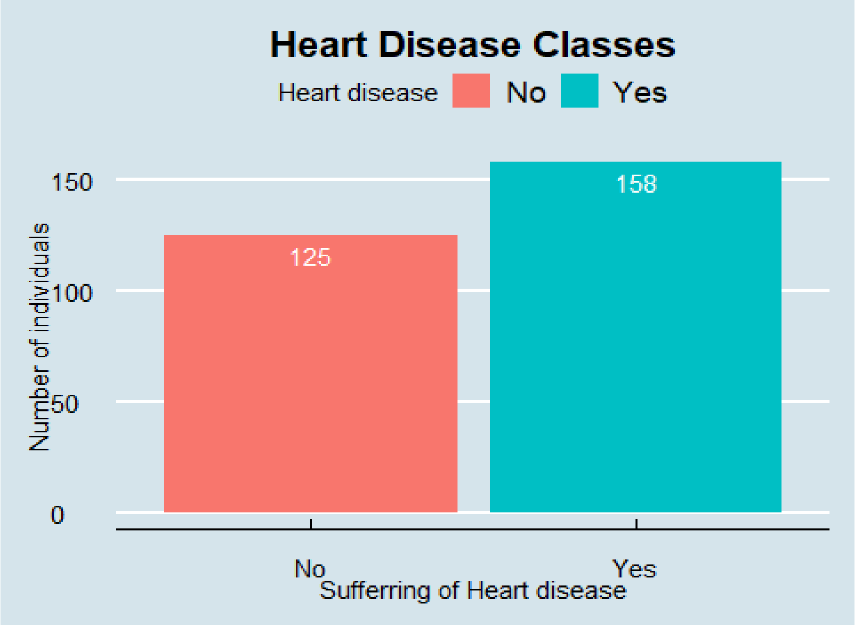


The scatter plot above shows the relationship between the studentized residuals of the logistic regression with target as a response and the other variables as predictors.

Observations whose studentized residuals are greater than 3 in absolute value are possible outliers. We can see that all the observations' studentized residuals stays within -2 and 2. So there are no outliers left in the dataset.

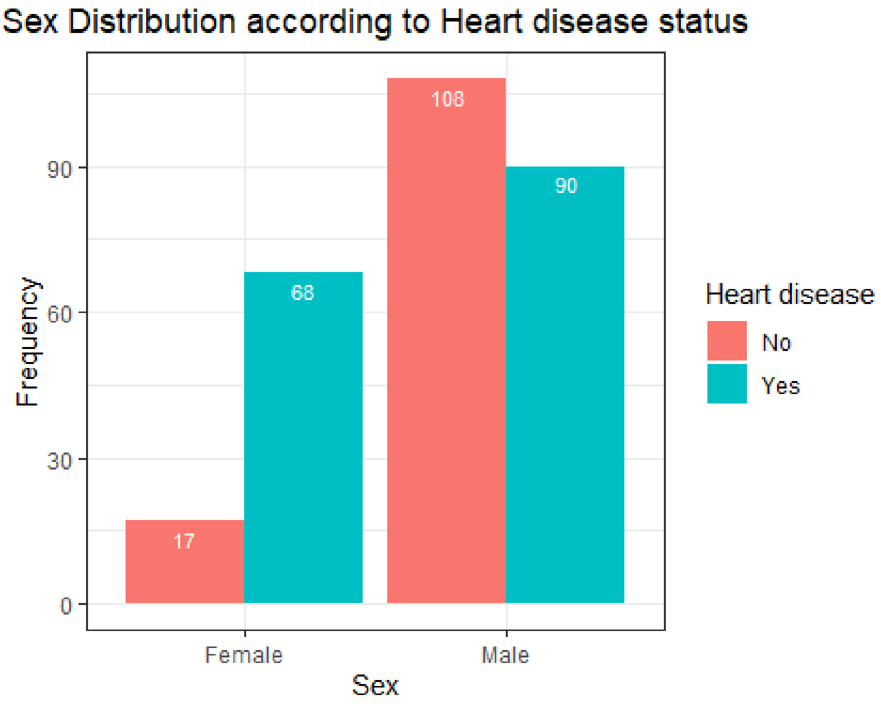
III. DESCRIPTIVE STATISTICS AND ANALYTICAL RESULTS

1. Heart disease (target)

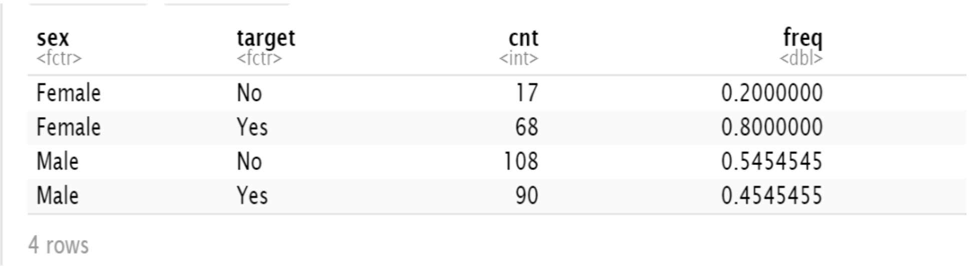


We have 158 examples where someone has heart disease based on their health parameters and 125 examples where someone doesn't have heart disease. So that's a quite balanced classification problem.

2. Sex (sex) and Heart disease (target)



The graph show that there is an imbalance in **sex** (only 85 women out of 283 individuals).



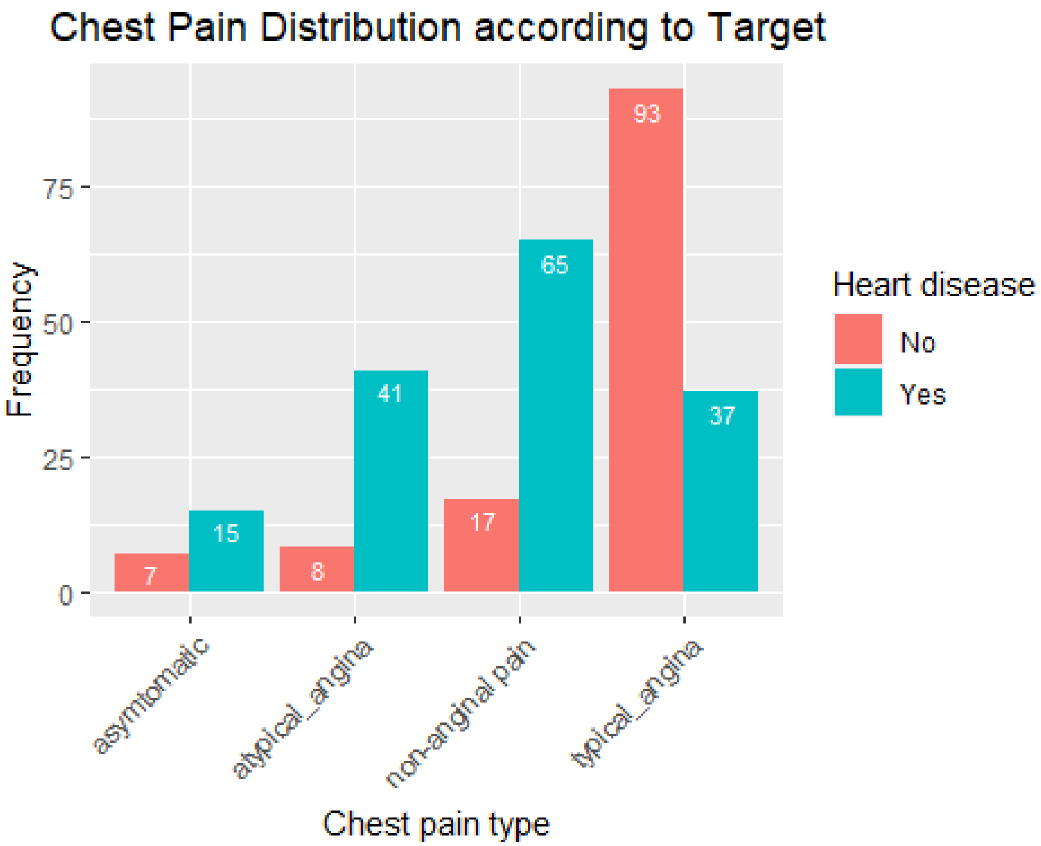
sex <fctr>	target <fctr>	cnt <int>	freq <dbl>
Female	No	17	0.2000000
Female	Yes	68	0.8000000
Male	No	108	0.5454545
Male	Yes	90	0.4545455

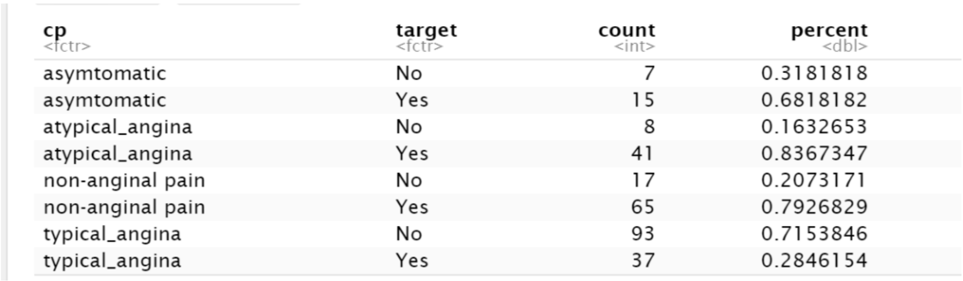
4 rows

Looking at the graph, we can observe that, in among each subgroup of sex, only 20% of **female** is patient, while it is 54.54% of patients in **male** subgroup.

So it can be concluded that **female** is more likely to suffer from heart disease than **male** (as shown in the dataset).

3. Chest pain and target





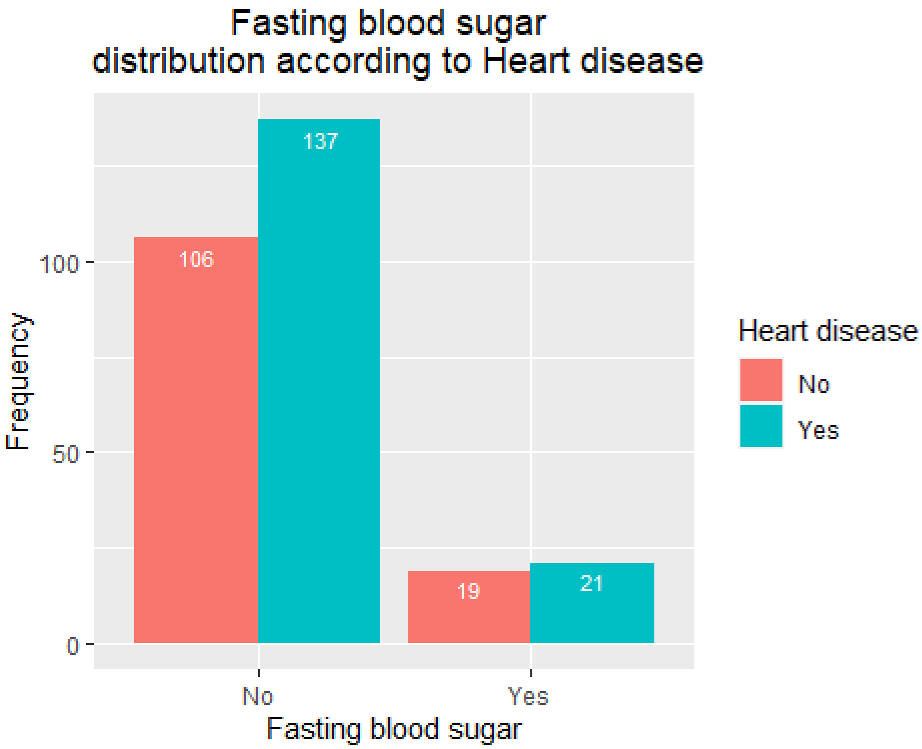
cp <fctr>	target <fctr>	count <int>	percent <dbl>
asymtomatic	No	7	0.3181818
asymtomatic	Yes	15	0.6818182
atypical_angina	No	8	0.1632653
atypical_angina	Yes	41	0.8367347
non-anginal pain	No	17	0.2073171
non-anginal pain	Yes	65	0.7926829
typical_angina	No	93	0.7153846
typical_angina	Yes	37	0.2846154

We all think that **chest pain** of all types are one of the important symptoms of heart diseases. Surprisingly, while it is true for **typical-anginal;** **asymptomatic, atypical angina** and **non-anginal pain** do not cause heart diseases (at least in the dataset). According to the table, the proportion of patient in each subgroup of chest pain type, **asymptomatic, atypical angina** and **non-anginal pain is 68.18%, 83.67% and 79.27%** respectively (much higher than

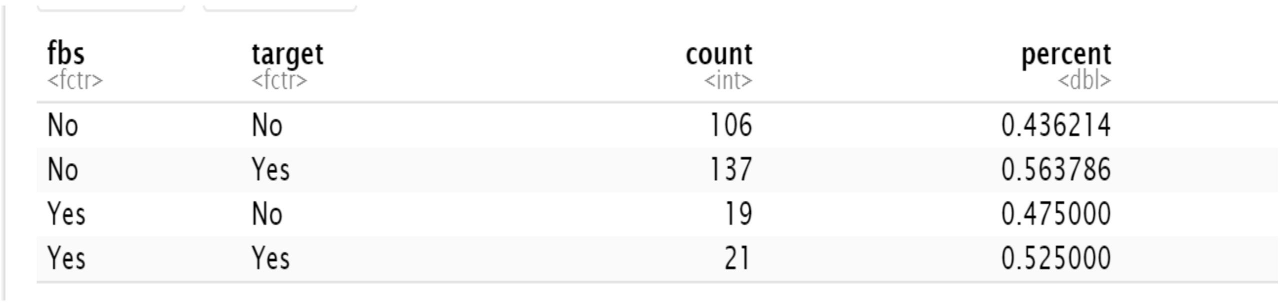
50%). However, only **28.46%** of patient is **typical-anginal** (about 1/3 of non-patient in this subgroup). So **chest pain** might be a strong predictor of heart diseases.

In addition, **chest pain** is also imbalanced, which is 130 out of 283 observations is **typical_angina** (nearly 50%). The smallest proportion is for **asymptomatic** (only 22 out of 283 observations).

4. Fasting blood sugar (fbs) and Heart disease (target)



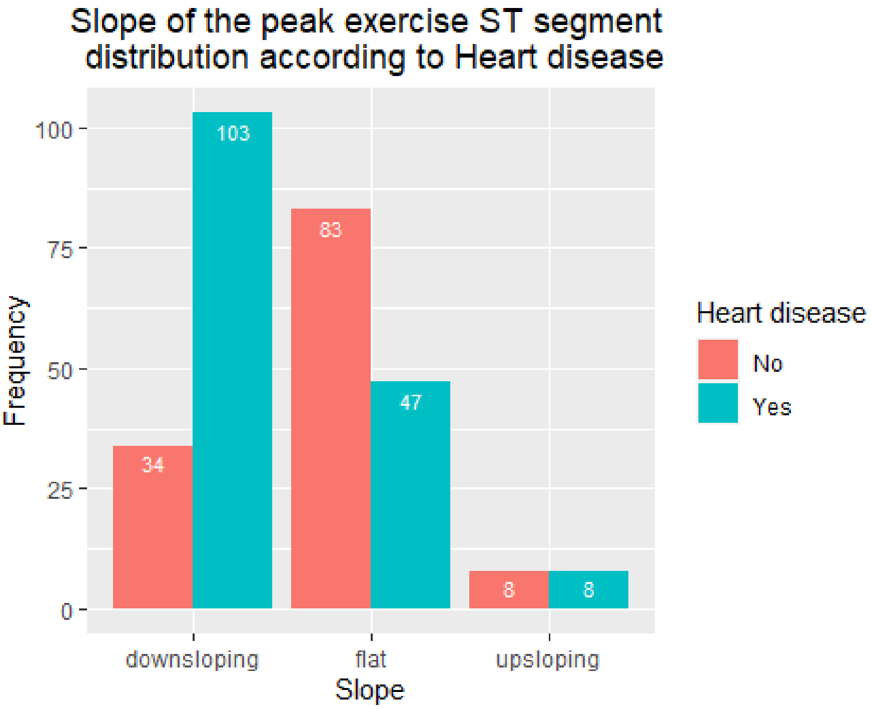
It is noticeable that the number of people with non fasting blood sugar is much higher than that number of people with fasting blood sugar.



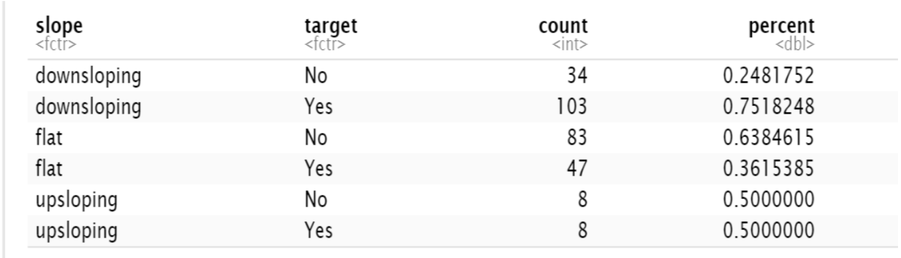
fbs <fctr>	target <fctr>	count <int>	percent <dbl>
No	No	106	0.436214
No	Yes	137	0.563786
Yes	No	19	0.475000
Yes	Yes	21	0.525000

Looking at the table, we can see that the proportion of patient in each subgroup of **fbs** is quite equal to that of non-patients, which indicates that **fbs** may not a strong predictor (at least in this dataset).

5. The slope of the peak exercise ST segment (slope) and heart disease (target)



Firstly, it can be observed that **upsloping** accounts for a small amount in **slope**, which means that **slope** is imbalanced.



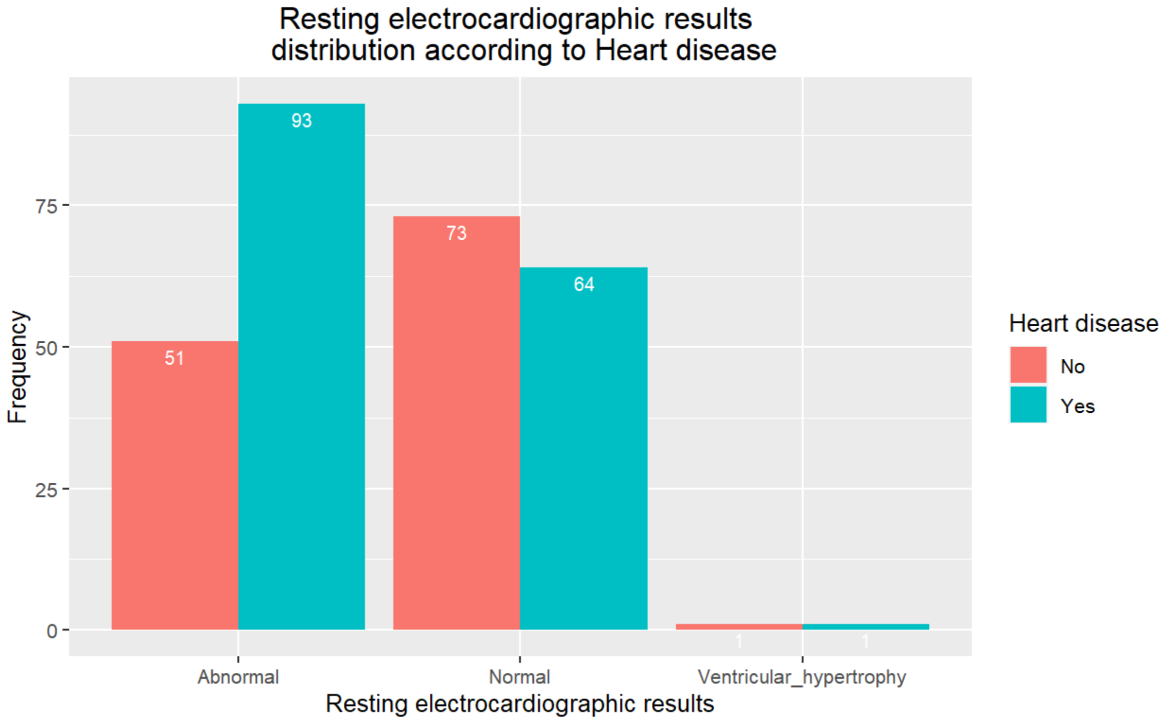
slope <fctr>	target <fctr>	count <int>	percent <dbl>
downsloping	No	34	0.2481752
downsloping	Yes	103	0.7518248
flat	No	83	0.6384615
flat	Yes	47	0.3615385
upsloping	No	8	0.5000000
upsloping	Yes	8	0.5000000

Secondly, in **downsloping,** the number of patients is nearly triple that of non-patients while in **flat** (75.18% compared to 24.82%), the former is just above half of the latter (63.85%

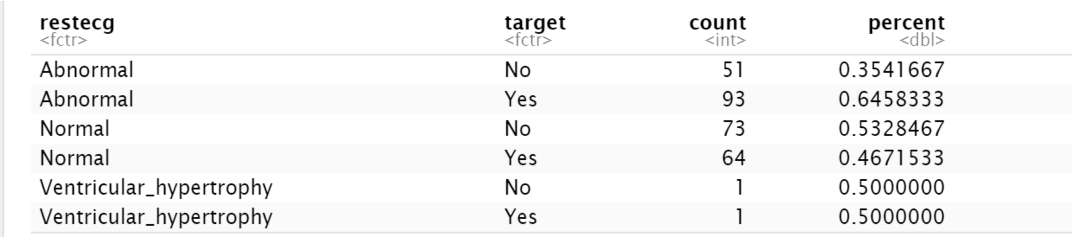
compared to 36.15%). In **upsloping**, there is no difference between the former and the latter (50% each).

So if a person has **downsloping** or **flat**, he or she is more likely to have heart disease but we can not say anything if a person has **upsloping**.

6. Resting ECG (restecg) and Heart disease (target)



We can see that **restecg** is imbalanced, only 2 out of 283 observations belongs to **ventricular_hypertrophy**, while it is it is 137 and 144 for **normal** and **abnormal** subgroup.

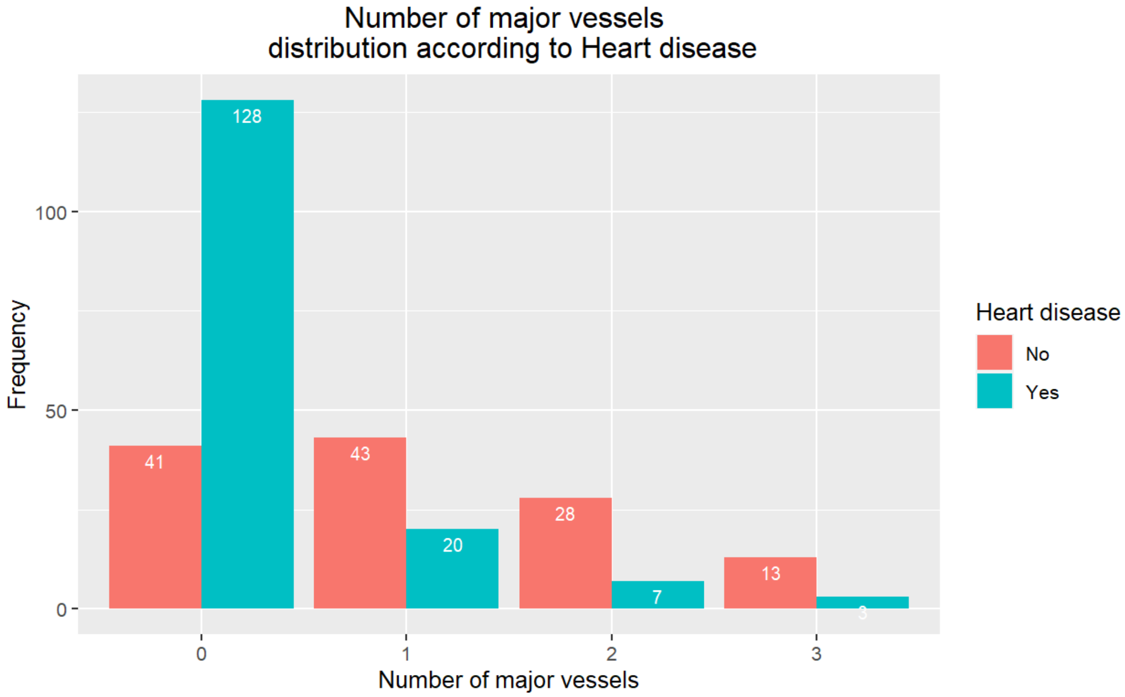


restecg <fctr>	target <fctr>	count <int>	percent <dbl>
Abnormal	No	51	0.3541667
Abnormal	Yes	93	0.6458333
Normal	No	73	0.5328467
Normal	Yes	64	0.4671533
Ventricular_hypertrophy	No	1	0.5000000
Ventricular_hypertrophy	Yes	1	0.5000000

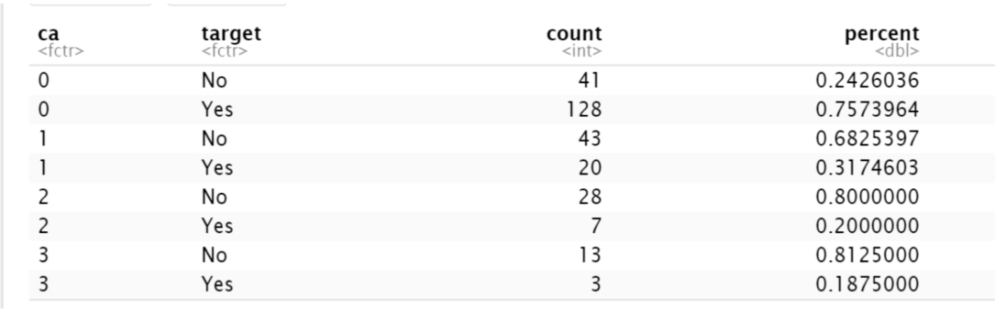
In **abnormal**, we can see that the proportion of patient 2 times that of non-patient. However, the other 2 subgroups of **restecg** experience equal proportions in patient and non-patient.

It may not a good predictor as I can not find any patterns from this variable.

7. Number of vessels colored by flourosopy (ca) and Heart disease (target)



Firstly, **ca** is imbalanced (169 out of 283 individuals has 0 major vessels colored by fluoroscopy, which is over 50%). While only 63, 35 and 16 people have **1,2 and 3** major vessels colored by fluoroscopy respectively.

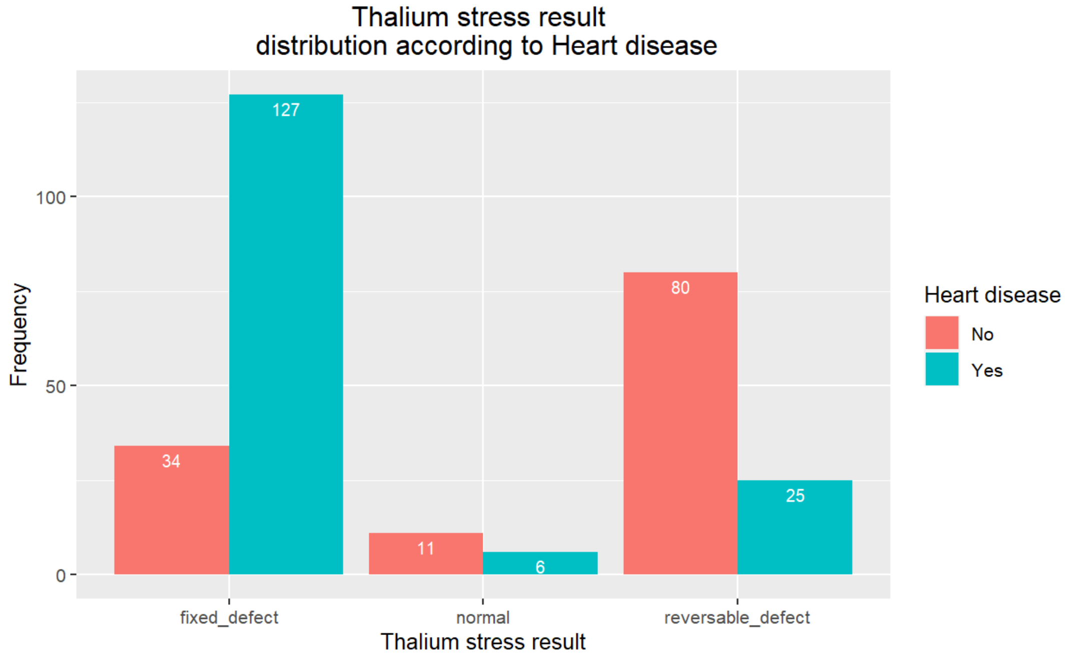


ca <fctr>	target <fctr>	count <int>	percent <dbl>
0	No	41	0.2426036
0	Yes	128	0.7573964
1	No	43	0.6825397
1	Yes	20	0.3174603
2	No	28	0.8000000
2	Yes	7	0.2000000
3	No	13	0.8125000
3	Yes	3	0.1875000

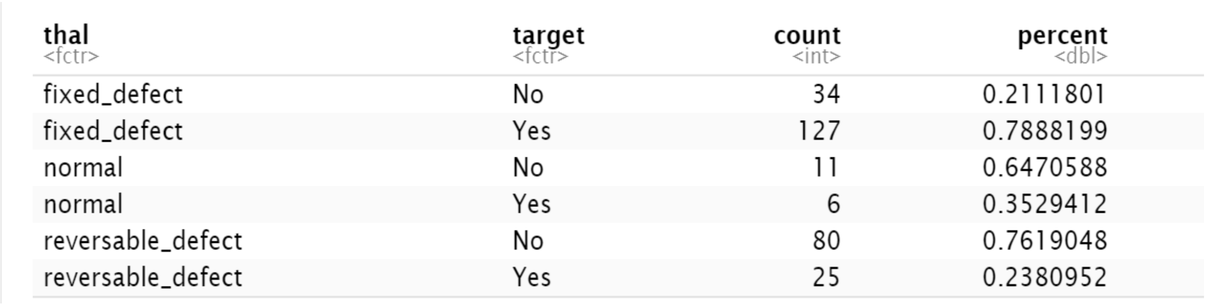
In addition, subgroup **1,2** and **3** experience the same pattern, which means that there is a much higher proportion of non-patient compared to patient. But the opposite happens in subgroup **0** (75% of patient compared to 25% of non-patient). So a person with no vessels colored by fluoroscopy is more likely to suffer from heart disease.

So **ca** might be a strong predictor of heart disease.

8. Thalium stress result (thal) and Heart disease (target)



Firstly, **thal** is imbalanced (only 17 out of 283 observations is subgroup **normal** while is it 161 and 105 for subgroup **fixed_defect** and **reversable_defect**, respectively**.**



thal	target	count	percent
<fctr>	<fctr>	<int>	<dbl>
fixed_defect	No	34	0.2111801
fixed_defect	Yes	127	0.7888199
normal	No	11	0.6470588
normal	Yes	6	0.3529412
reversable_defect	No	80	0.7619048
reversable_defect	Yes	25	0.2380952

Additionally, in subgroup **fixed_defect**, the proportion of patient is three times that of non-patient while in the other 2 subgroups, the proportion of patient is only about half of non-patient. So patients with fixed defect have a significantly higher incidence of heart disease.

So **thal** might is a strong predictor as well.

9. Logistic regression assumptions

There are some assumptions in logistic regression:

- Outcome variable is binary, where the number of outcomes is two (e.g., Yes/No). The response variable in the dataset is **target**, which is **a factor with 2 values: 1 for Yes and 0 for No**.

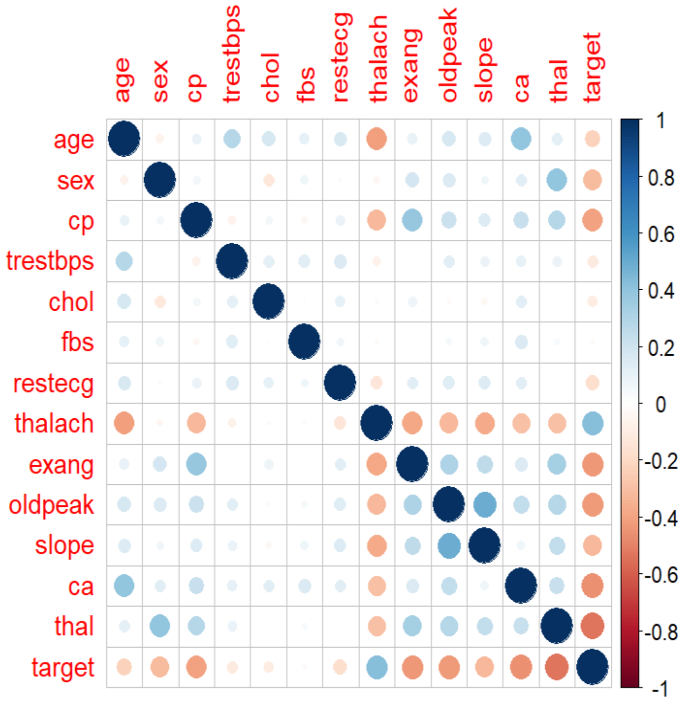
- The relationship between the log-odds of the outcome and each continuous independent variable is *linear*.

- There are no highly influential outlier data points, as they distort the outcome and accuracy of the model. This problem has been solved by removing 5 outliers from the original dataset.

- There is no high intercorrelations (i.e. multicollinearity) among the predictors.

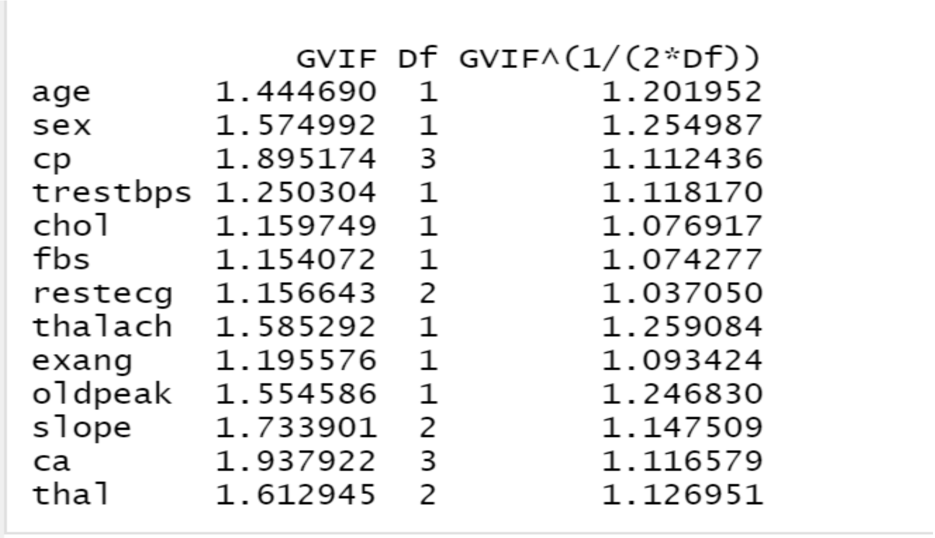
The first and third assumptions of logistic regression has been solved. We need to check if there is any colliearity in the predictors or not.

9.1 Intercorrelations between predictors



From the correlation plot, we can see that there is a relatively mild correlation between predictors such as **age and thalach**, **age and ca**, **sex and thal**, **cp and exang**, **thalach and exang**, **slope and oldpeak**. So we expect to see lasso regression shrink some less important coefficients.

Collinearity can be assessed using the R function vif() [car package], which computes the variance inflation factors.

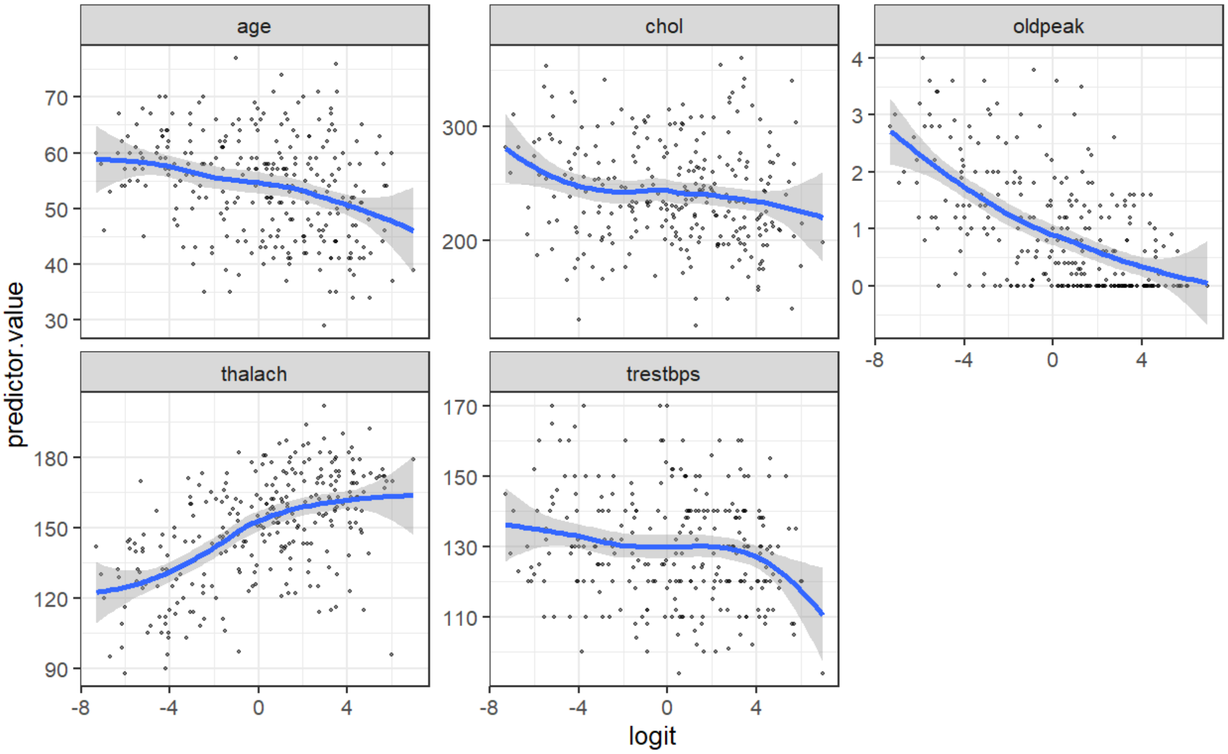


```
             GVIF Df GVIF^(1/(2*Df))
age      1.444690  1        1.201952
sex      1.574992  1        1.254987
cp       1.895174  3        1.112436
trestbps 1.250304  1        1.118170
chol     1.159749  1        1.076917
fbs      1.154072  1        1.074277
restecg  1.156643  2        1.037050
thalach  1.585292  1        1.259084
exang    1.195576  1        1.093424
oldpeak  1.554586  1        1.246830
slope    1.733901  2        1.147509
ca       1.937922  3        1.116579
thal     1.612945  2        1.126951
```

A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In this dataset, there is no collinearity since all predictors have a value of VIF well below 5.

9.2 Linearity between the response and each numerical predictors

Checking the linear relationship between numerical predictors and the logit of the response can be done by visually inspecting the scatter plot between each predictor and the logit values.

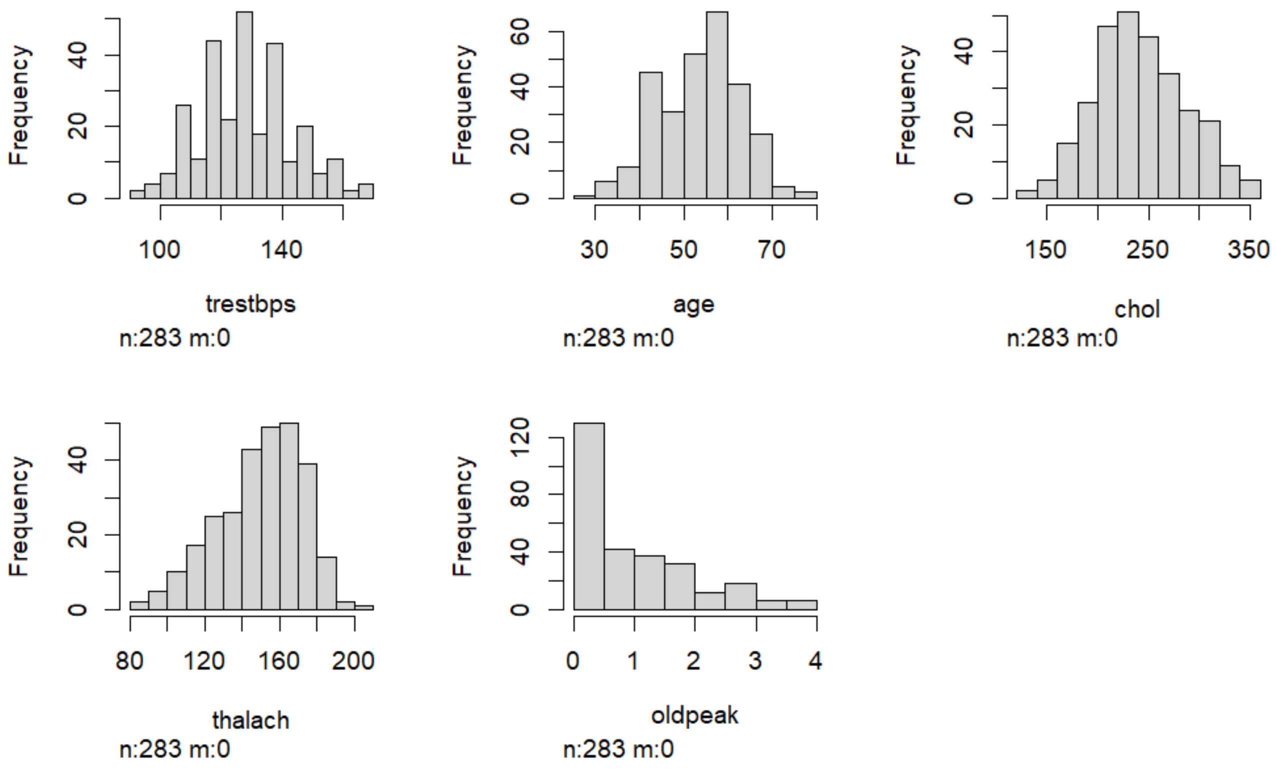


The smoothed scatter plots show that variables **age**, **chol**, **oldpeak**, **thalach** and **trestbps** are all quite linearly associated with the diabetes outcome in logit scale. However, the plot between **trestbps** and logit has a wild tail. It is because there is very few points at the tail.

9.3 Skewed data

The plot below indicates some insights: **age**, **trestbps** and **chol** are approximately normally distributed.

oldpeak is left-skewed while it is right-skewed for **thalach**.

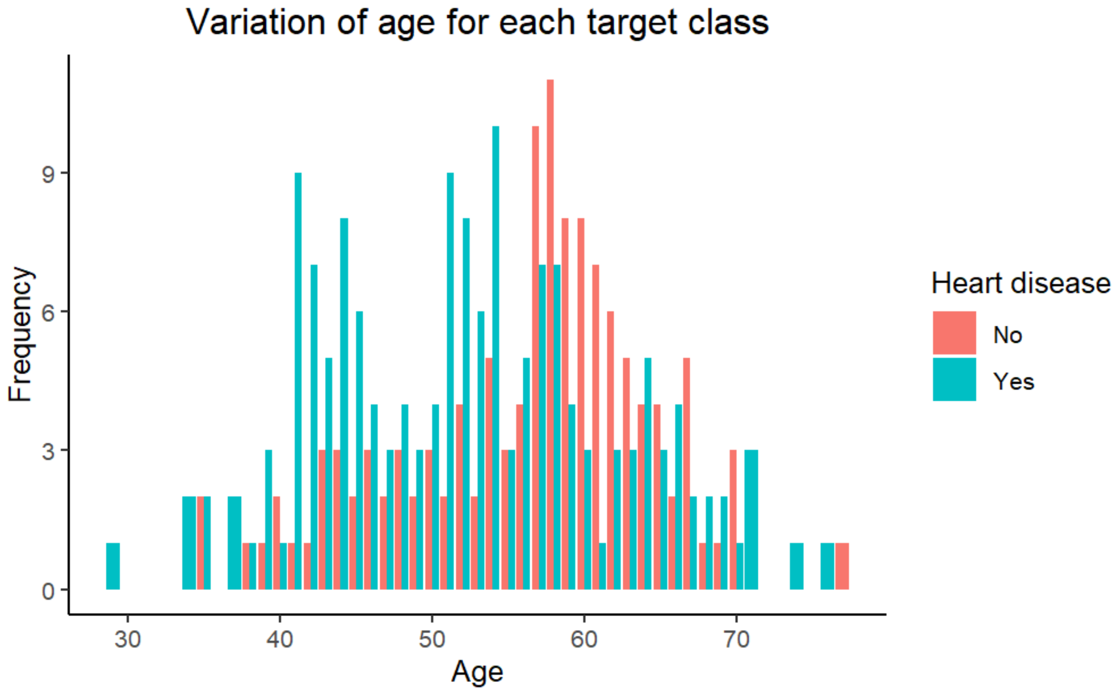


However, **there is no assumption about normality on independent variable in logistic regression**. So the skewness of the numerical predictors is not problematic.

10. Age and Target

It can be easily observed that people suffering from heart disease are of the age of 58 and 57.

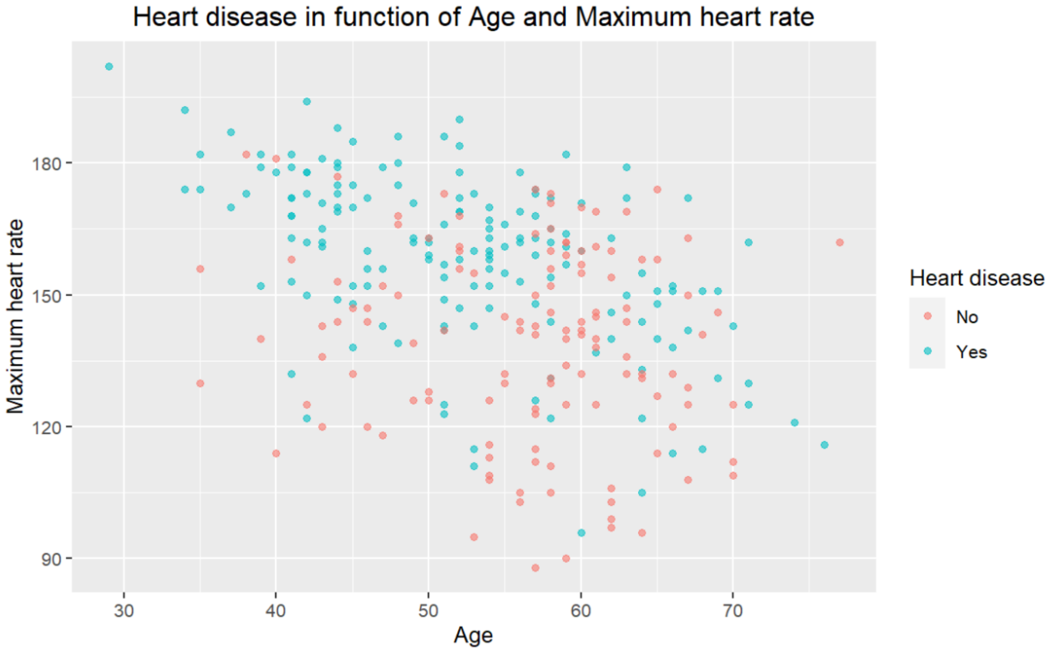
Apparently, those who are in the age group 50+ s from the disease.



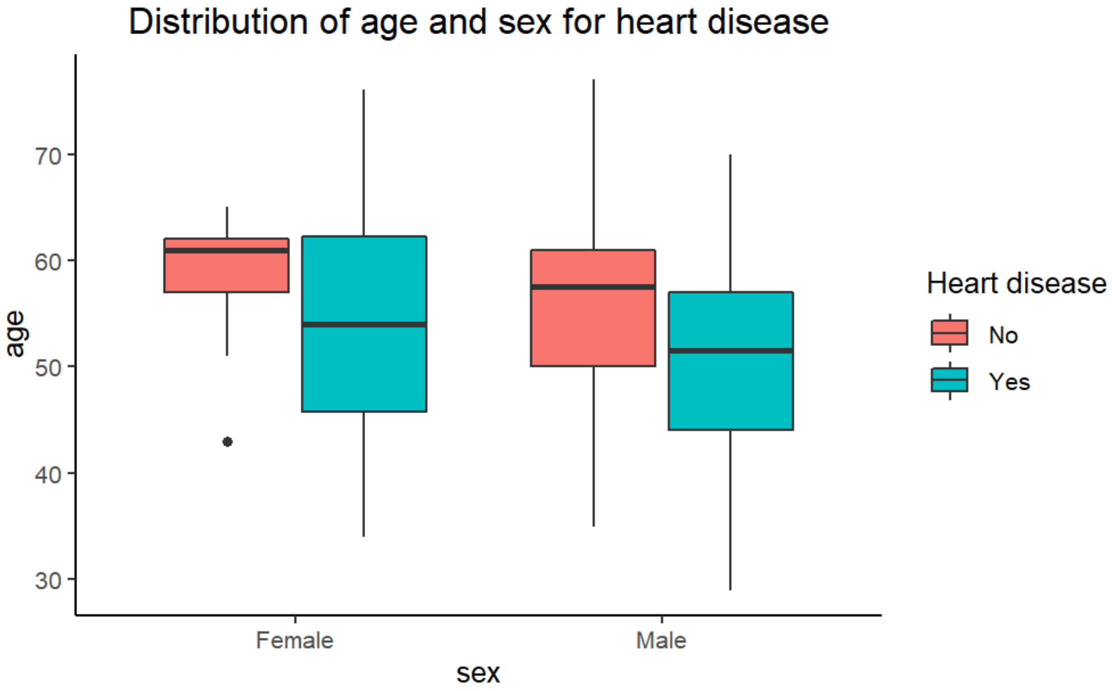
11. Age, Maximum heart rate and Target

A person gets older, their heart rate decreases. We can see a downward trend in the plot below.

It seems that maximum heart rate can be strong predictor for heart disease, regardless of age.



12. Age, Sex and Target



We see that for females who are suffering from the disease are older than males.

IV. STATISTICAL METHODS AND MODEL IMPLEMENTATION STEPS

1. Statistical methods

The model used in this project is **logistic regression, lasso logistic regression** and **cross validation** to **predict** whether a person suffers from heart disease or not.

Why do I use:

Logistic regression? Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The lasso logistic regression relies on the assumptions of logistic regression. So the plan is to use logistic regression to diagnostic potential problems against the logistic regression's assumptions.

Lasso logistic regression? Lasso regression is like linear regression, but it uses a technique "shrinkage" where the coefficients of determination are shrunk towards zero. As I am using the dataset with 13 predictors, multicollinearty is very likely to occur in the model. In lasso regression, the less important features of a dataset are penalized, which means some coefficients are made zero leading to their elimination. The dataset with high dimensions and correlation is well suited for lasso regression, which is a good choice for the heart disease prediction model. Also, lasso regression will tackle with skewed data by default setting **standardize=TRUE**.

Cross validation? When we're building a machine learning model using some data, we often split our data into training and validation/test sets. The training set is used to train the model, and the validation/test set is used to validate it on data it has never seen before. The classic approach is to do a simple 80%-20% or 70%-30% split. In cross-validation, we do more than one split. We can do 3, 5, 10 or any K number of splits. The use of cross validation in lasso regression will significantly benefit the hyperparameter tuning process, particularly **lambda** which determines the severity of the penalty.

2. Model implementation steps

1. Train – test split: split 80-20 the dataset into 2 parts: a training set and a test set.

2 Run a model

2.1 Logistic regression: run the logistic regression model using the training set

2.2 Lasso regression

a. Run the lasso regression model using the training set and 100 lambdas ranging from 0.01 to 100

b. Choose a model with lambda that has the smallest misclassification error.

c. Run that model using the training set.

3. Model evaluation

3.1 Create a ROC plot to compare 2 models.

3.2 Predict 2 models (one logistic regression model and one lasso regression model) with a cut-off of 0.5 using the test set

3.3 Create a confusion matrix and calculate accuracy rate

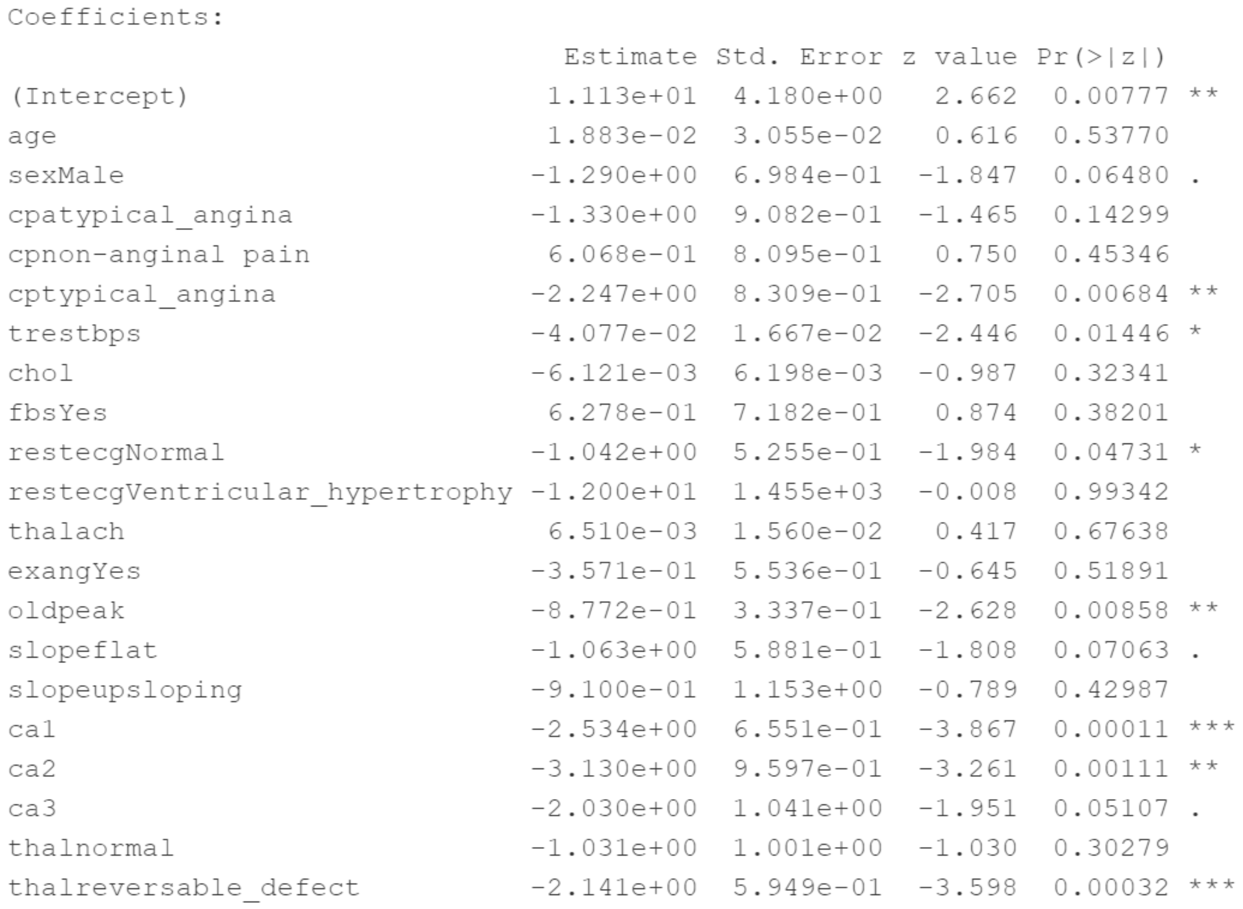
3.4 Find an optimal cut-off that weighs both sensitivity and specificity equally.

3.5 Find another optimal cut-off that weighs both sensitivity and specificity differently.

V. MODEL IMPLEMENTATION

1. Logistic regression

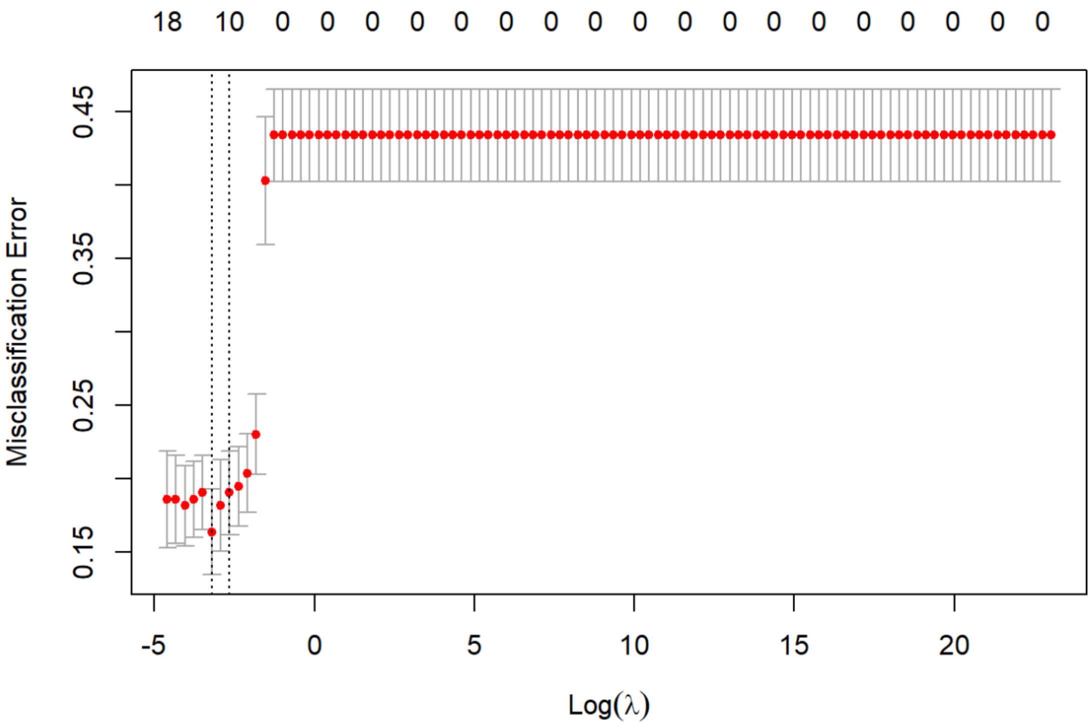
The logistic regression model summary output shows that only **4 qualitative predictors** (**cp** - Chest pain type with 2 categories (Typical agiana and the other), **restecg** - Resting electrocardiographic results with 2 categories (Normal and the other), **ca** - Number of major vessels (0-3) colored by flourosopy with 3 categories (1, 2 and the other), **thal** - Thalium stress result with 2 categories (Reversible defect and the other) and **2 quantitative predictors** (**trestbps** – Resting blood pressure and **oldpeak** - ST depression induced by exercise relative to rest looks at stress of heart during exercise) statistically significantly predict whether a person suffer from heart disease or not.



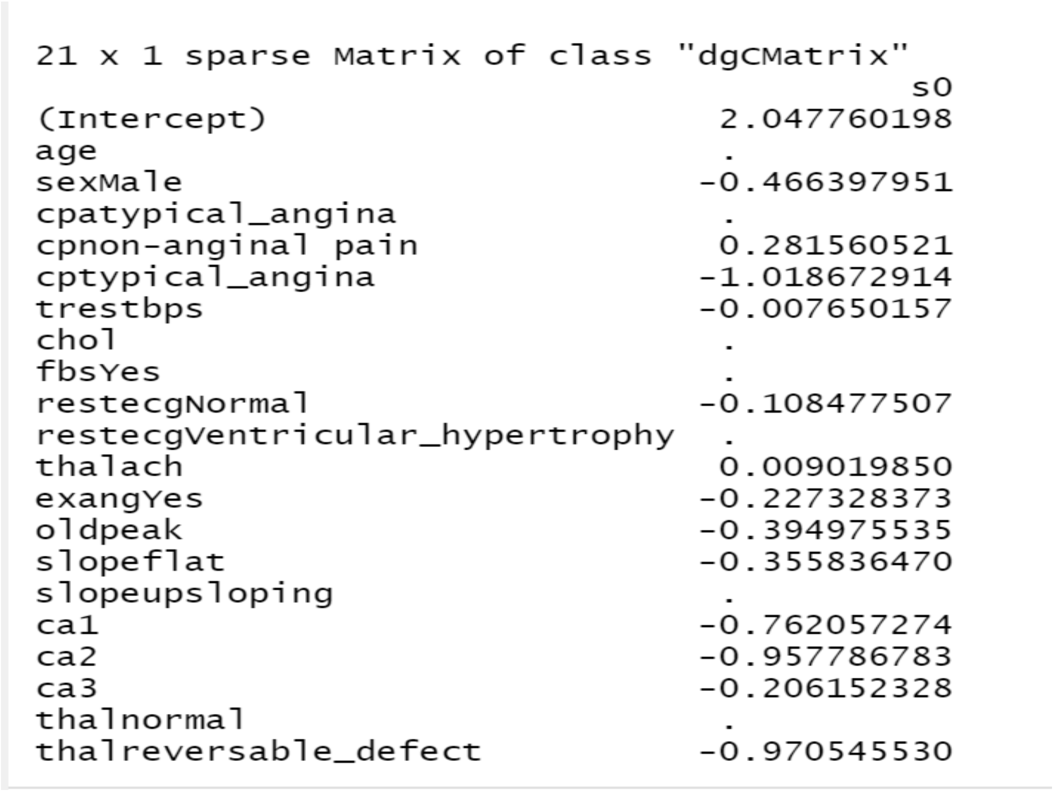
```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.113e+01  4.180e+00   2.662  0.00777 **
age                            1.883e-02  3.055e-02   0.616  0.53770
sexMale                       -1.290e+00  6.984e-01  -1.847  0.06480 .
cpatypical_angina             -1.330e+00  9.082e-01  -1.465  0.14299
cpnon-anginal pain             6.068e-01  8.095e-01   0.750  0.45346
cptypical_angina              -2.247e+00  8.309e-01  -2.705  0.00684 **
trestbps                      -4.077e-02  1.667e-02  -2.446  0.01446 *
chol                          -6.121e-03  6.198e-03  -0.987  0.32341
fbsYes                         6.278e-01  7.182e-01   0.874  0.38201
restecgNormal                 -1.042e+00  5.255e-01  -1.984  0.04731 *
restecgVentricular_hypertrophy -1.200e+01  1.455e+03  -0.008  0.99342
thalach                        6.510e-03  1.560e-02   0.417  0.67638
exangYes                      -3.571e-01  5.536e-01  -0.645  0.51891
oldpeak                       -8.772e-01  3.337e-01  -2.628  0.00858 **
slopeflat                     -1.063e+00  5.881e-01  -1.808  0.07063 .
slopeupsloping                -9.100e-01  1.153e+00  -0.789  0.42987
ca1                           -2.534e+00  6.551e-01  -3.867  0.00011 ***
ca2                           -3.130e+00  9.597e-01  -3.261  0.00111 **
ca3                           -2.030e+00  1.041e+00  -1.951  0.05107 .
thalnormal                    -1.031e+00  1.001e+00  -1.030  0.30279
thalreversable_defect         -2.141e+00  5.949e-01  -3.598  0.00032 ***
```

2. Lasso regression

Running a lasso regression model using the training set and **100 λ ranging from 0.01 to 100** results in different misclassification errors. The model with **λ = 0.04037017** brings the lowest error. So I am **using λ = 0.04037017 for the final lasso regression model**. Below is the plot between misclassification error and log of λ.



The final lasso regression model with λ = 0.04037017 gives the following regression coefficients:

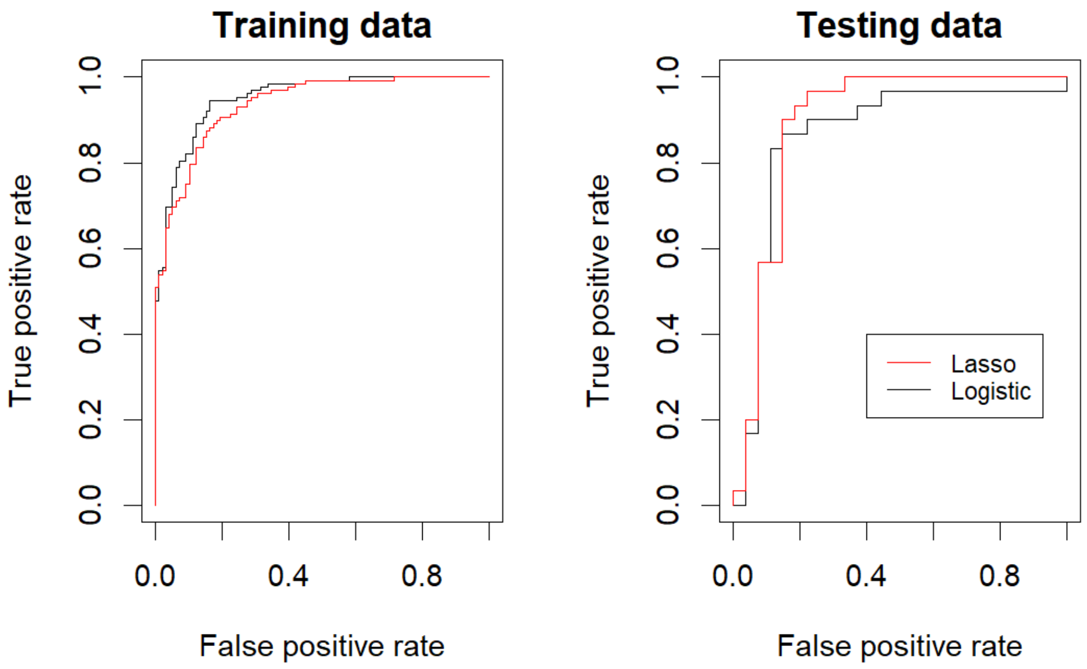


```
21 x 1 sparse Matrix of class "dgCMatrix"
                                       s0
(Intercept)                   2.047760198
age                           .
sexMale                      -0.466397951
cpatypical_angina             .
cpnon-anginal pain            0.281560521
cptypical_angina             -1.018672914
trestbps                     -0.007650157
chol                          .
fbsYes                        .
restecgNormal                -0.108477507
restecgVentricular_hypertrophy .
thalach                       0.009019850
exangYes                     -0.227328373
oldpeak                      -0.394975535
slopeflat                    -0.355836470
slopeupsloping                .
ca1                          -0.762057274
ca2                          -0.957786783
ca3                          -0.206152328
thalnormal                    .
thalreversable_defect        -0.970545530
```

From the output above, there are some variables that has coefficients exactly equal to zero such as **age**, **cpatypical_angina** (chest pain not related to heart), **chol** (serum cholestoral in mg/dl), **fbs** (fasting blood sugar), **slopeupsloping** (the upsloping slope of the peak exercise ST segment), **thalnormal** (normal thalium stress result).

In the next sections, we'll compare the model lasso regression with best λ against the full logistic regression model.

VI. MODEL ASSESSING AND COMPARISON



The ROC plots on training data shows that the best lasso model underperforms the logistic one while **in testing data, the former outperforms the latter**.

So we can say that in generally, **the best lasso model brings a better prediction compared to the logistic one**.

1. Find the best threshold

First, I am going to choose 0.5 as a cut-off for the 2 models.

Logistic			**Lasso**		
Predict	**Actual**		**Predict**	**Actual**	
	Yes	**No**		**Yes**	**No**
Yes	27	7	**Yes**	29	6
No	3	20	**No**	1	21
Accuracy	82.45 %		**Accuracy**	87.72 %	
Specificity	74.07 %		**Specificity**	77.78 %	
Recall	90 %		**Recall**	96.97 %	

Generally, **the 2 models has accuracy of over 80%**, which is by far better than random guessing. The table states that in general, **the lasso model better correctly classifies** each individual than the logistic model, with accuracy rate of 87.72% for the former and 82.45% for the latter.

However, **in medical science, sensitivity and specificity are two important metrics** that characterize the performance of classifier or screening test. The importance between sensitivity and specificity depends on the context. Generally, we are concerned with one of these metrics.

In medical diagnostic, such as in this case, we are likely to **be more concerned with minimal wrong positive diagnosis**. So, we are **more concerned about high Specificity**. Here, the model specificity is 92%, which is very good.

Note that, here I have used p > 0.5 as the probability threshold above which, declaring the concerned individuals as heart disease positive. However, if we are concerned about incorrectly predicting the heart disease-positive status for individuals who are truly positive, then we can **consider lowering the threshold**.

For example, missing someone with a heart disease based on a test may cost us $50,000 in lawsuits, but treating someone who did not have the disease may cost $10,000 in treatments. In that case, the cost of a false negative is 5 times that of a false positive, strictly in monetary measures. No cost analysis is this simple and is usually based on many factors, but most analyses do not have equal cost for a false positive versus a false negative. In this project, **assume that the cost of a false negative is 10 times that of a false positive, let's find a threshold that satisfy this assumption**.

With the cost of a false negative is 10 times that of a false positive, the optimal thresholds for the logistic model and best lasso model are 0.09283595 and 0.3717579 respectively.

Logistic			**Lasso**		
Optimal threshold	0.09283595		**Optimal threshold**	0.3717579	
Predict	**Actual**		**Predict**	**Actual**	
	Yes	**No**		**Yes**	**No**
Yes	28	12	**Yes**	30	10
No	2	15	**No**	0	17
Accuracy	75.44 %		**Accuracy**	82.46 %	
Specificity	55.56 %		**Specificity**	62.96 %	
Recall	93.33 %		**Recall**	100 %	

As we can see, in both models, using **a lower threshold brings a lower accuracy rate as well as lower specificity rate** (the proportion of identified negatives among the heart disease-negative population); however, **the recall rate** (the proportion of identified positives among the heart disease-positive population) **improves significantly, especially 100% for the lasso model**.

VI. CONCLUSION

The optimal **lasso model (λ = 0.04037017) performs better** than the full logistic regression model.

In **the logistic models**, these are statistically significant predictors:

- 4 qualitative predictors: **cp -** Chest pain type with 2 categories (Typical agiana and the other), **restecg -** Resting electrocardiographic results with 2 categories (Normal and the other), **ca -** Number of major vessels (0-3) colored by flourosopy with 3 categories (1, 2 and the other), **thal -** Thalium stress result with 2 categories (Reversible defect and the other).

- 2 quantitative predictors: **trestbps** – Resting blood pressure and **oldpeak** - ST depression induced by exercise relative to rest looks at stress of heart during exercise.

In **the lasso model**, these are statistically significant predictors:

- 3 quantitative predictors: **trestbps** (Resting blood pressure), **thalach** (Maximum heart rate achieved) , **oldpeak** (ST depression induced by exercise relative to rest looks at stress of heart during exercise)

- 7 qualitative predictors: **sex**, **cp** (chest pain type) – 3 categories (Non-anginal pain, Typical angina and the other), **restecg** (Resting electrocardiographic results) – 2 categories[1] (Normal and the other), **exang**, **slope** (The slope of the peak exercise ST segment) – 2 categories (Flat and the other), **ca** (Number of major vessels (0-3) colored by flourosopy), **thal** (thalium stress result) – 2 categories (reversible defect and the other).

There is **a trade-off** between the specificity and recall. The choice of which depends on **what is optimized in the model**. If we want to be more concerned with minimal wrong positive diagnosis, we are more concerned about high Specificity and then chose a lower threshold.

[1]