UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Heart Disease Prediction

Lorenzo Baietti	ID: 2130676
Francesco Carlesso	ID: 2125806
Matteo Mazzini	ID: 2107797

MSc Data Science

Statistical Learning

A.Y. 2023/2024

Project and Dataset Description

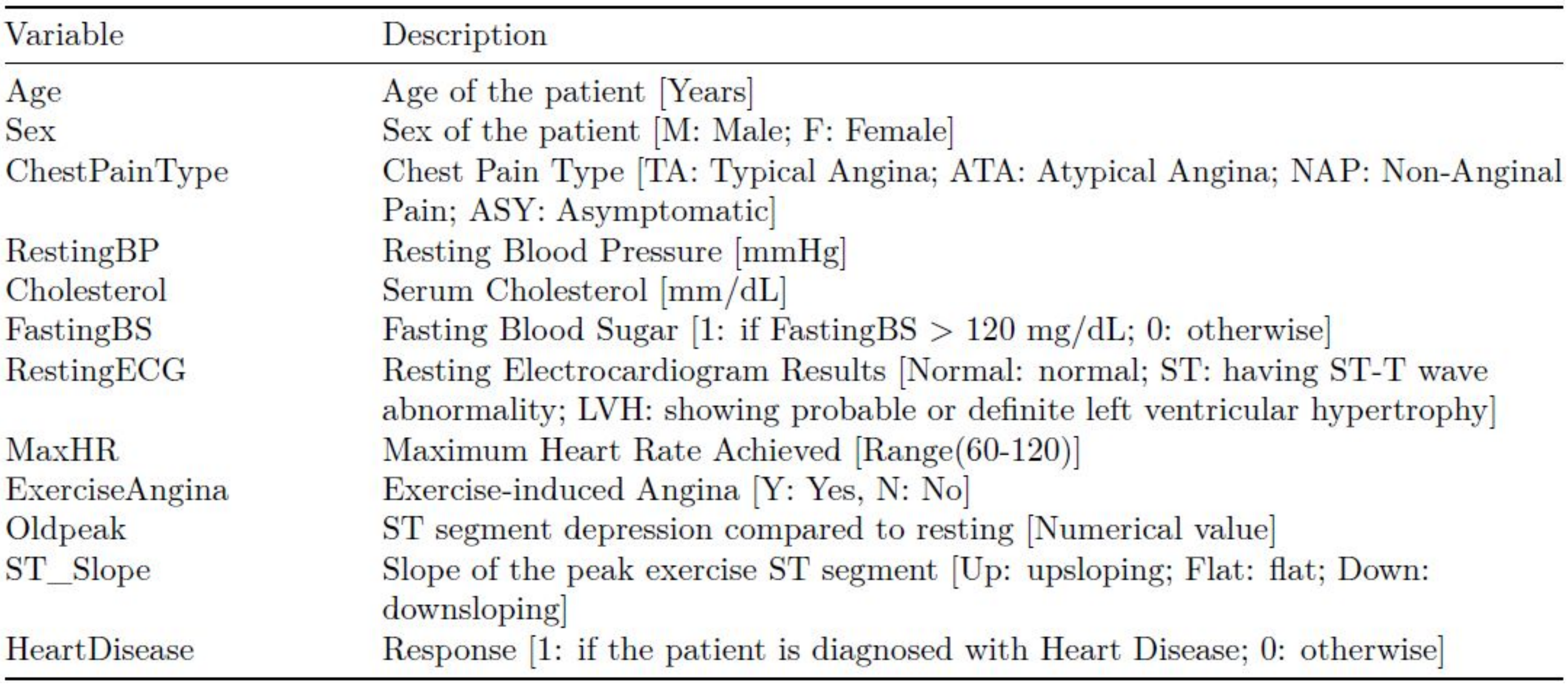
- ❖ Heart diseases can be caused by different kind of factors
- ❖ Early diagnosis is crucial for carrying out a successful treatment

Objective: Understand which are the most influential biometrics, focusing on a binary classification task which uses parameters that can be obtained simply by performing clinical tests.

Dataset: Heart Failure Prediction Dataset - combined from the UCI Machine Learning Repository

- 918 patient records
- 11 variables plus a binary target for the diagnosis
- 508 patients with a positive diagnosis

Variables Overview



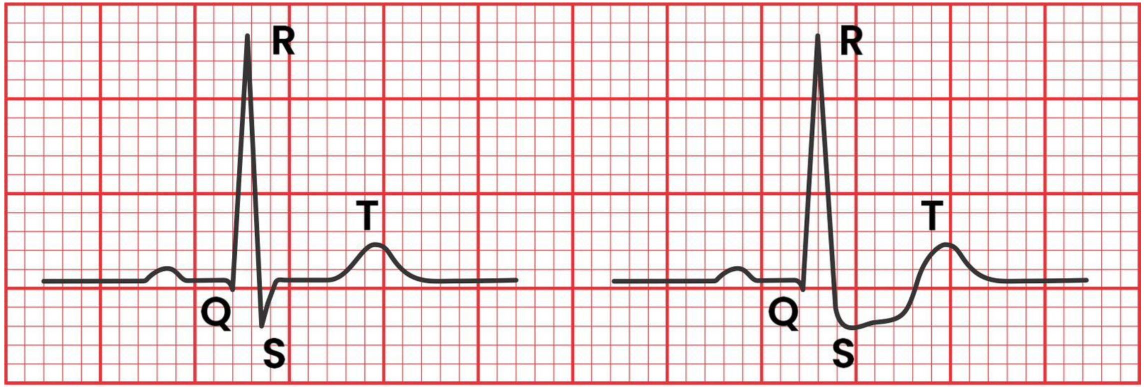
Variable	Description
Age	Age of the patient [Years]
Sex	Sex of the patient [M: Male; F: Female]
ChestPainType	Chest Pain Type [TA: Typical Angina; ATA: Atypical Angina; NAP: Non-Anginal Pain; ASY: Asymptomatic]
RestingBP	Resting Blood Pressure [mmHg]
Cholesterol	Serum Cholesterol [mm/dL]
FastingBS	Fasting Blood Sugar [1: if FastingBS $>$ 120 mg/dL; 0: otherwise]
RestingECG	Resting Electrocardiogram Results [Normal: normal; ST: having ST-T wave abnormality; LVH: showing probable or definite left ventricular hypertrophy]
MaxHR	Maximum Heart Rate Achieved [Range(60-120)]
ExerciseAngina	Exercise-induced Angina [Y: Yes, N: No]
Oldpeak	ST segment depression compared to resting [Numerical value]
ST_Slope	Slope of the peak exercise ST segment [Up: upsloping; Flat: flat; Down: downsloping]
HeartDisease	Response [1: if the patient is diagnosed with Heart Disease; 0: otherwise]

Terminology

Angina: Chest pain caused by reduced blood flow to the heart muscles

ST Segment: Electrically neutral area on the ECG, between ventricular depolarization (QRS) and repolarization (T) wave

Oldpeak: ST segment depression induced by exercise relative to rest



Normal

ST Depression

Data Preprocessing

Categorical Variables and NA Values

- Categorical variables from ‘chr’ and ‘int’ type to ‘Factor’ type
- There seems not to be any NAs at first

Issue:

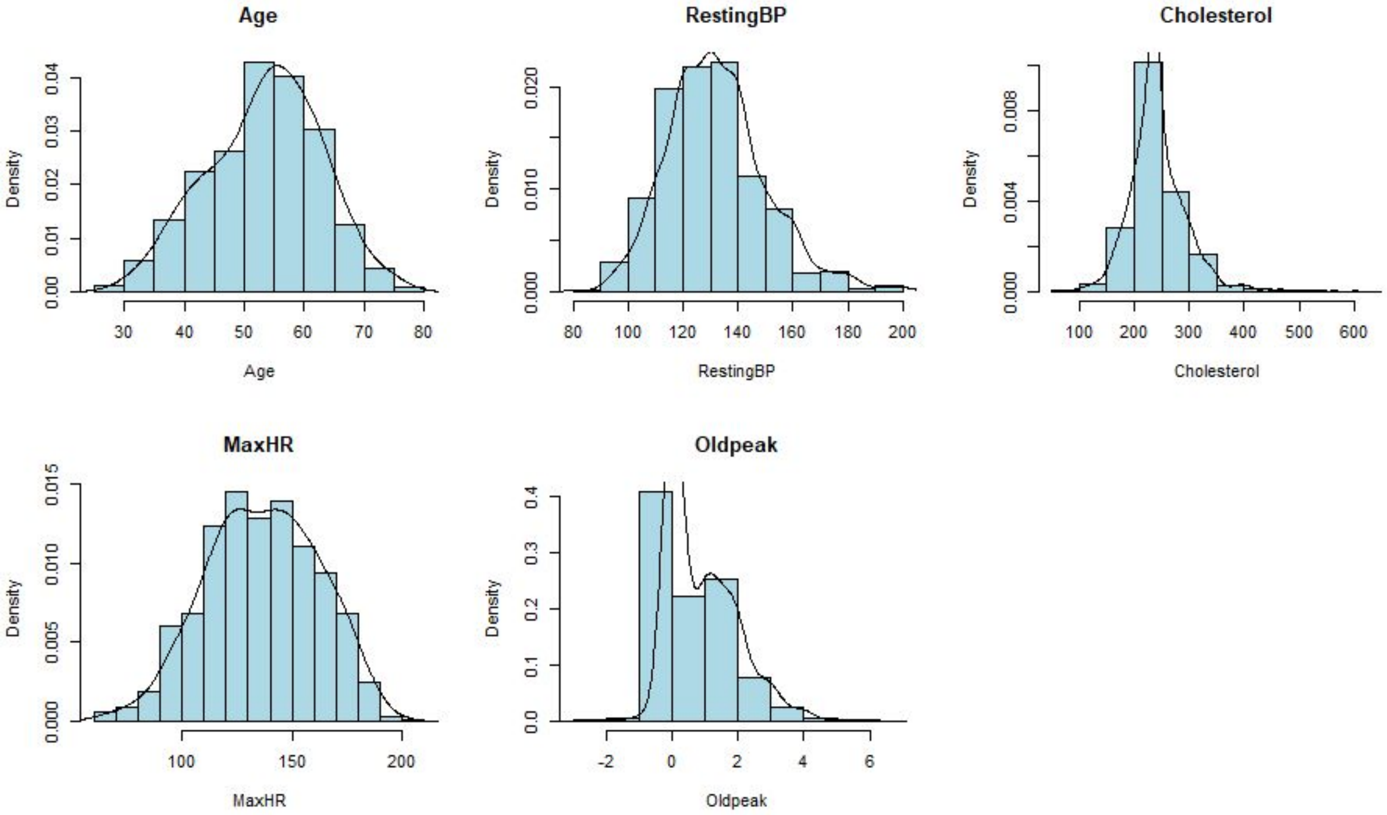
- RestingBP and Cholesterol have 0 as minimum value
- Blood pressure cannot be 0 unless the patient is dead, while 0 cholesterol is biologically impossible to observe even in deceased individuals

Solution:

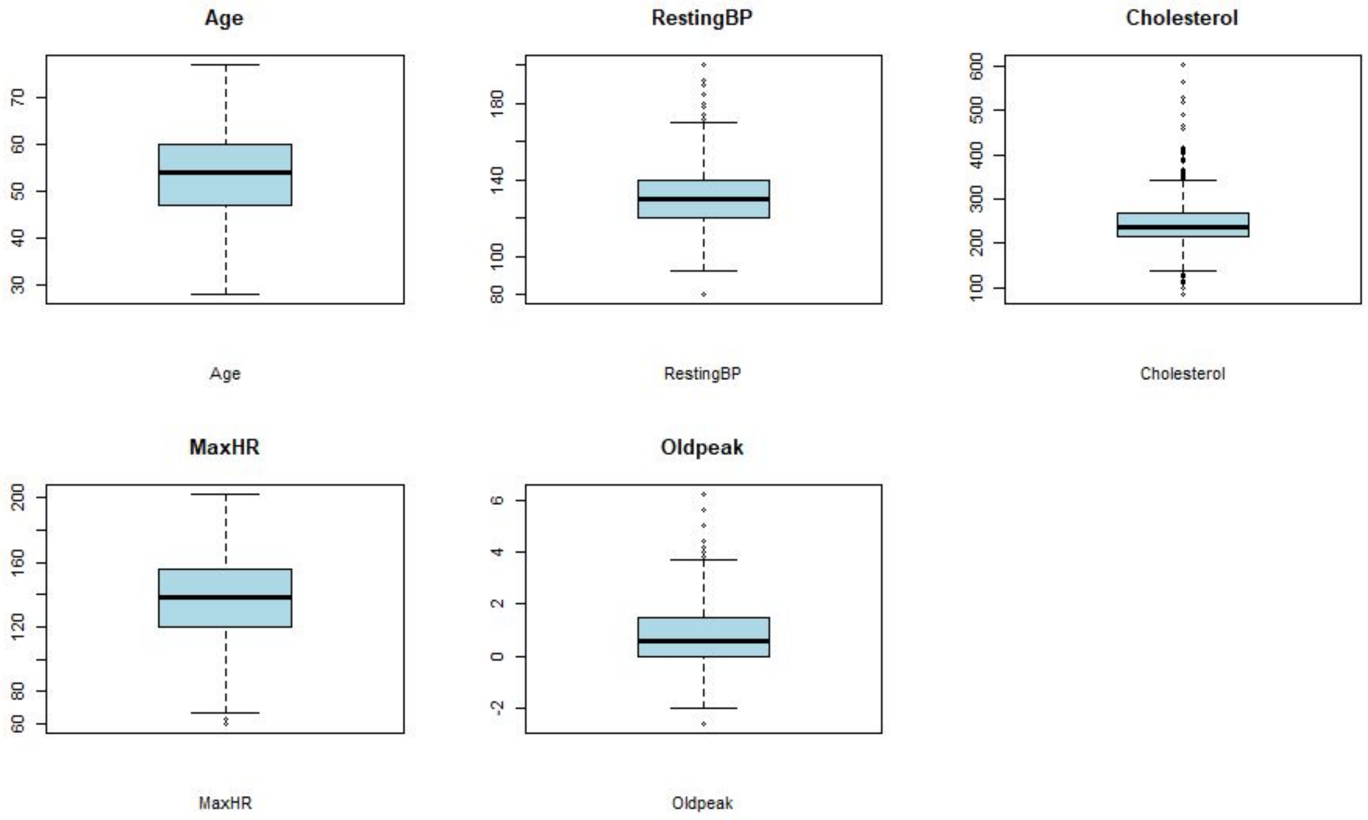
- Given that MaxHR > 0 for the RestingBP observation, we conjecture that the measurement was made on an alive patient
- We treat 0 values as NAs and substitute them with the median of the specific column

Data Exploration

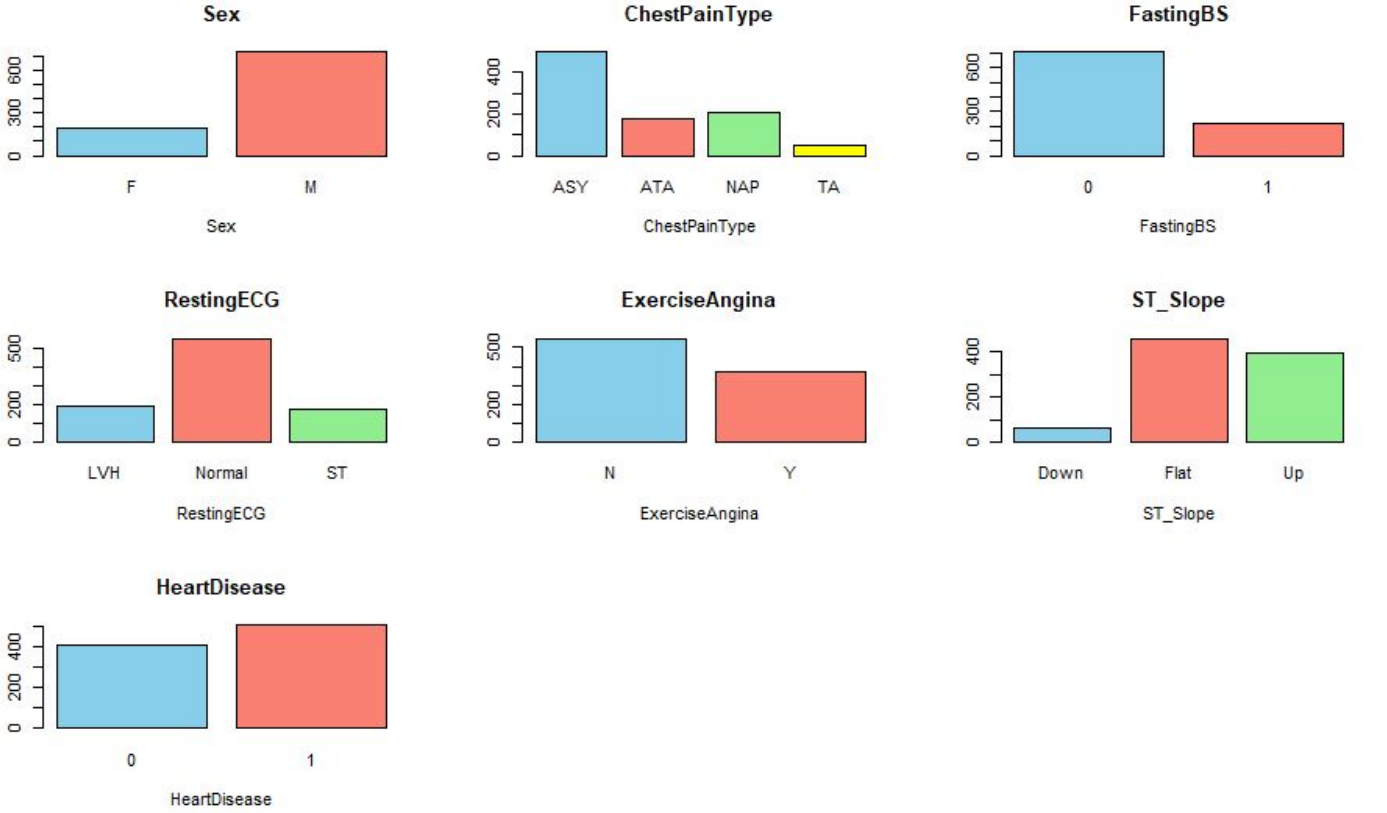
Univariate Analysis - Numerical Variables



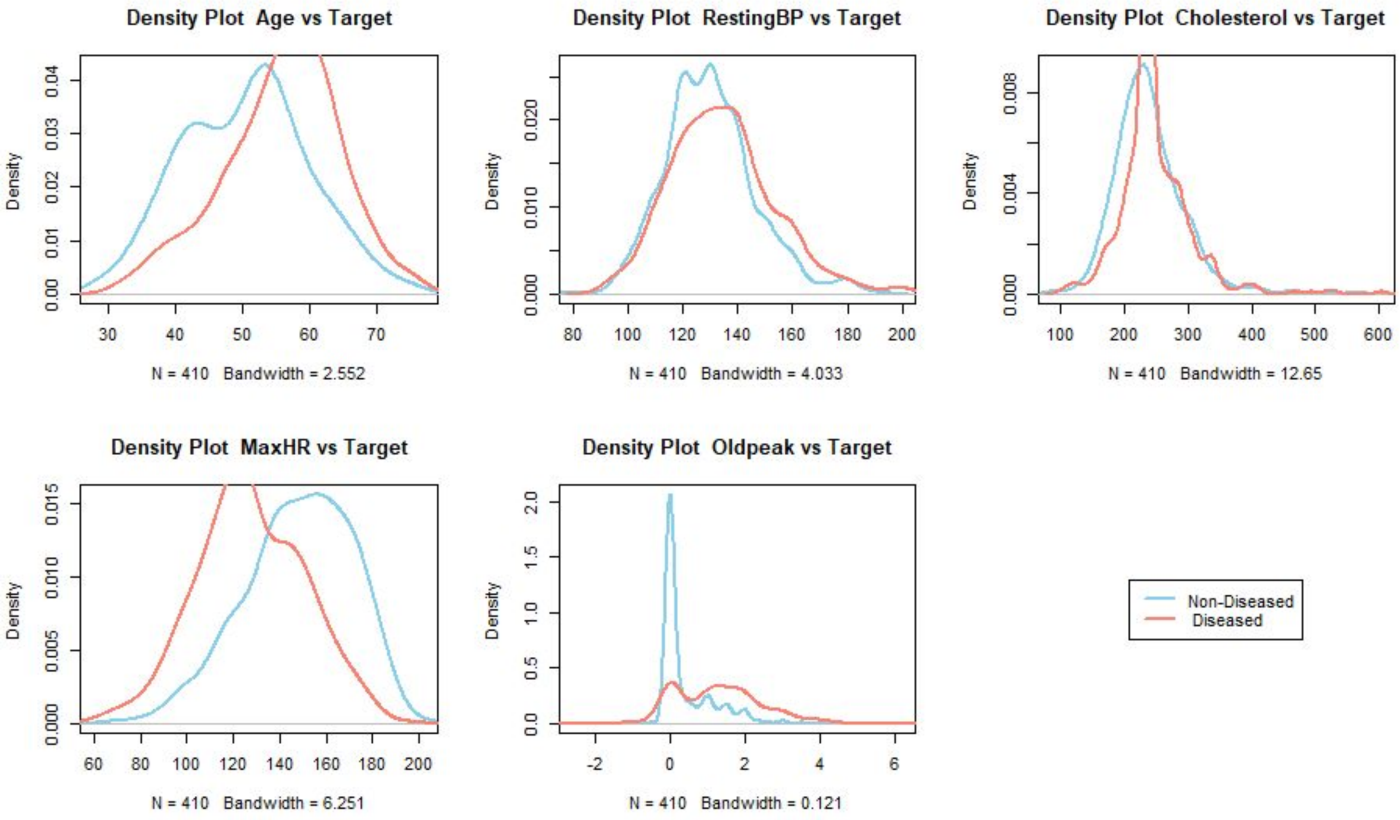
Univariate Analysis - Numerical Variables



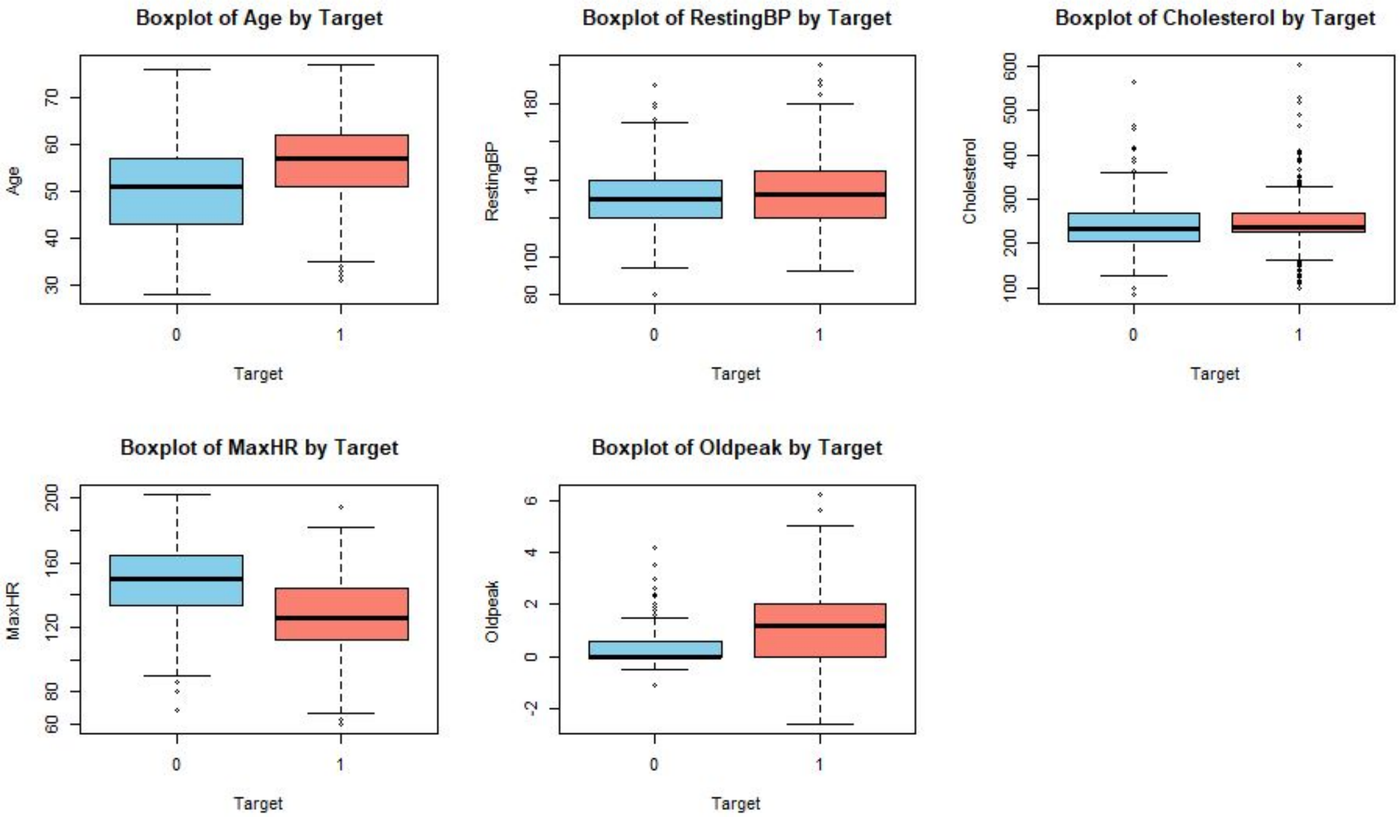
Univariate Analysis - Categorical Variables



Bivariate Analysis - Numerical Variables



Bivariate Analysis - Numerical Variables



Boxplot of Age by Target

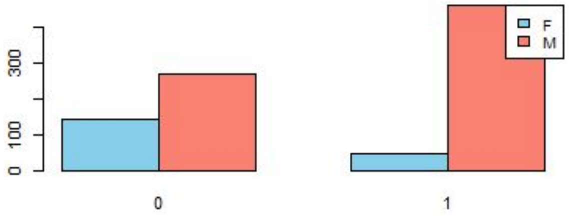
Boxplot of RestingBP by Target

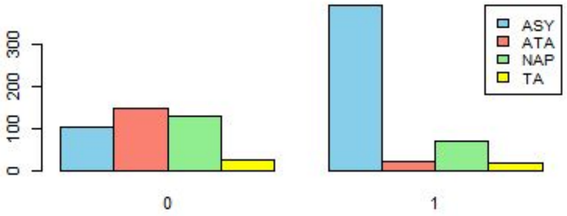
Boxplot of Cholesterol by Target

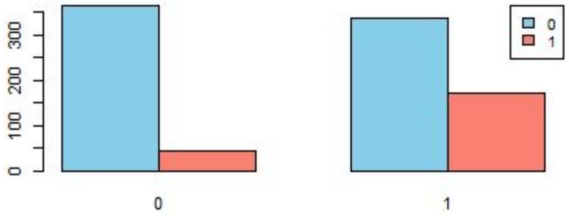
Boxplot of MaxHR by Target

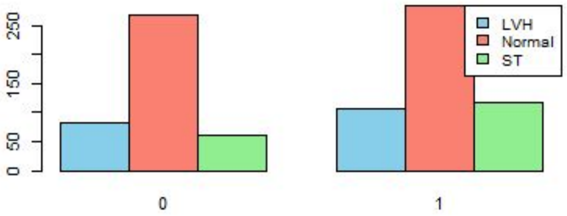
Boxplot of Oldpeak by Target

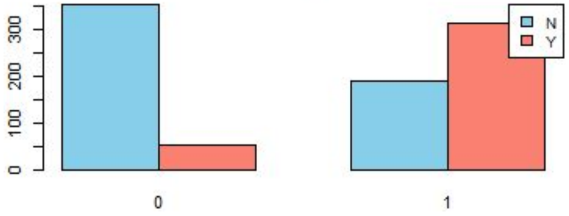
Bivariate Analysis - Categorical Variables

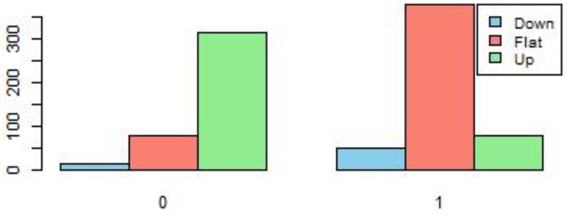

Barplot of Sex by HeartDisease


Barplot of ChestPainType by HeartDisease

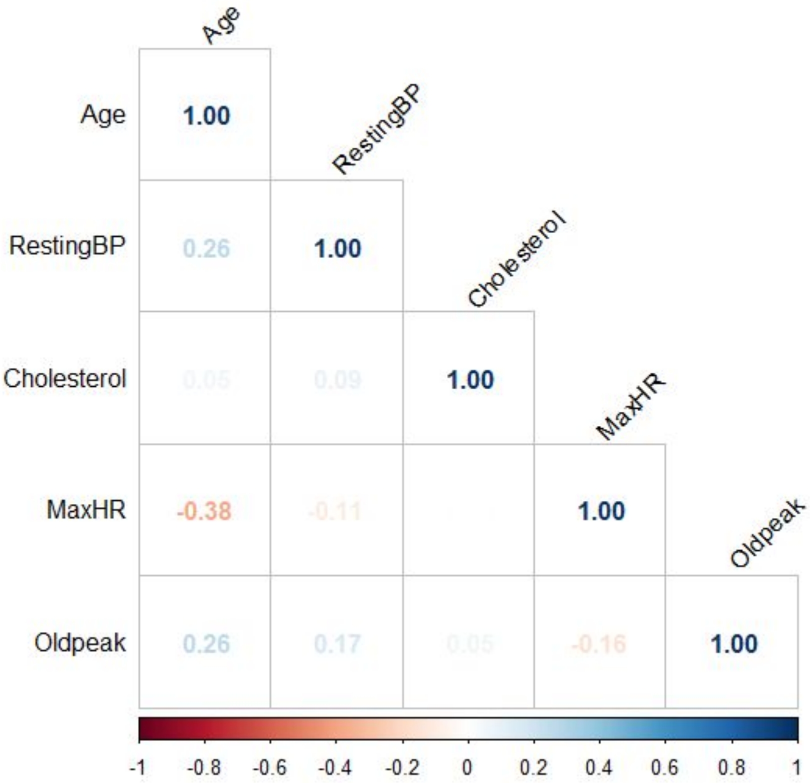

Barplot of FastingBS by HeartDisease


Barplot of RestingECG by HeartDisease


Barplot of ExerciseAngina by HeartDisease


Barplot of ST_Slope by HeartDisease

Correlation Analysis

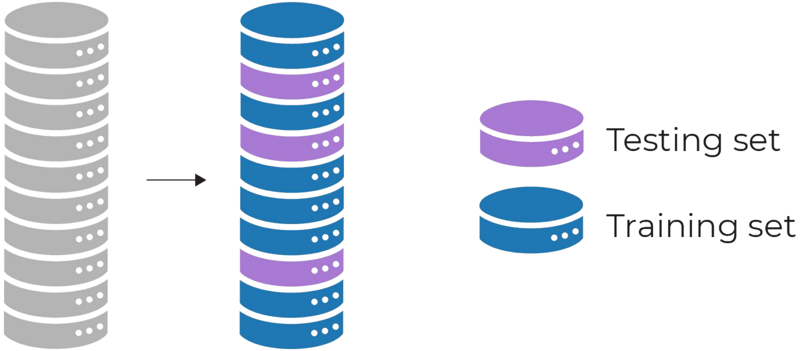


Data Modeling

Splitting and Scaling

- ❖ **Train-Test Split:** 80% - 20%

- ❖ **Standardization:** Make numerical variables follow a standard normal distribution N(0,1) to prevent features with larger scales from dominating the learning process, since the data collected has different units of measure.

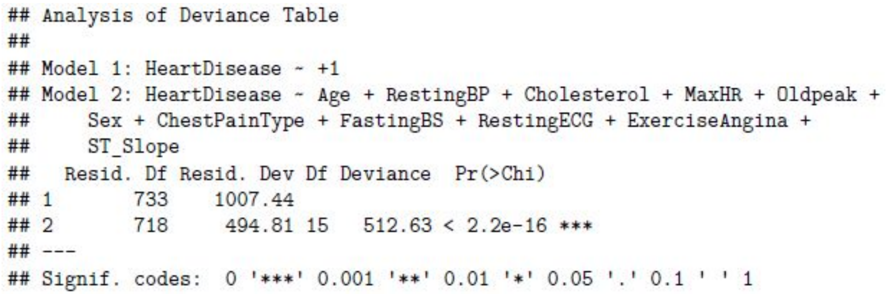


Testing set

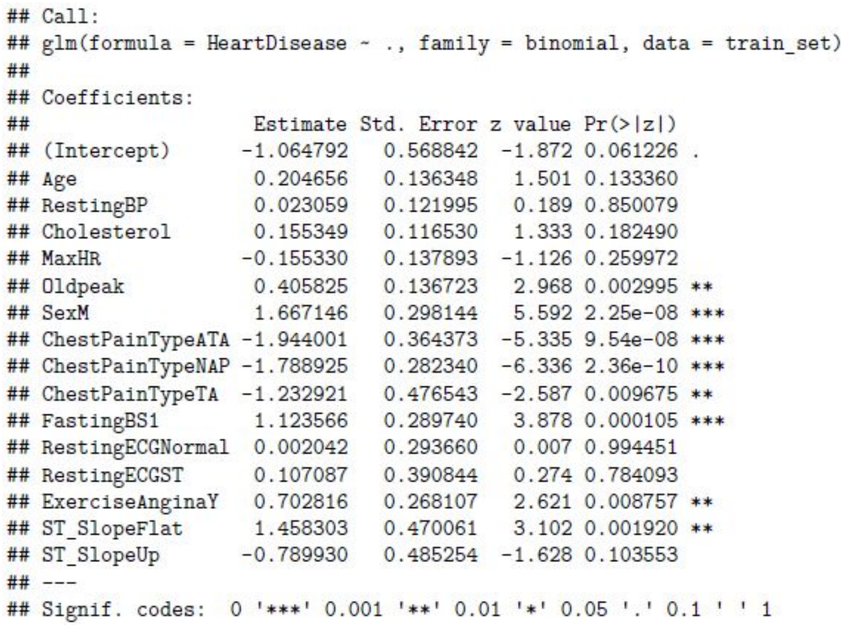
Training set

Simple Logistic Regression

- Max VIF value: 3.0
- AIC: 526.81



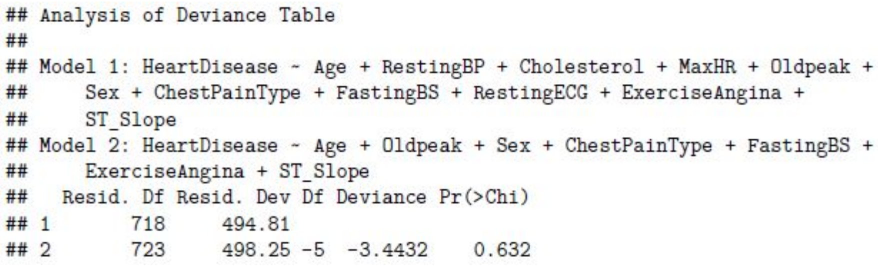
```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ +1
## Model 2: HeartDisease ~ Age + RestingBP + Cholesterol + MaxHR + Oldpeak +
##     Sex + ChestPainType + FastingBS + RestingECG + ExerciseAngina +
##     ST_Slope
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       733    1007.44
## 2       718     494.81 15   512.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



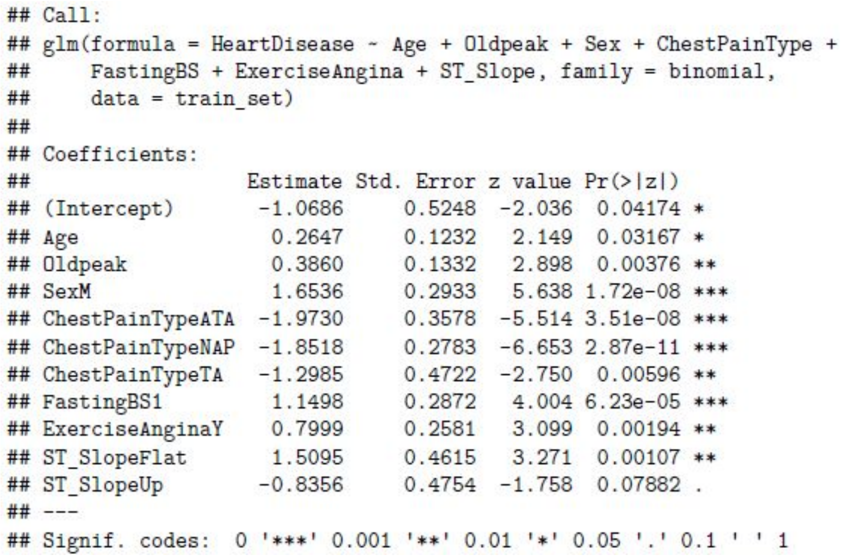
```
## Call:
## glm(formula = HeartDisease ~ ., family = binomial, data = train_set)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.064792   0.568842  -1.872 0.061226 .
## Age                0.204656   0.136348   1.501 0.133360
## RestingBP          0.023059   0.121995   0.189 0.850079
## Cholesterol        0.155349   0.116530   1.333 0.182490
## MaxHR             -0.155330   0.137893  -1.126 0.259972
## Oldpeak            0.405825   0.136723   2.968 0.002995 **
## SexM               1.667146   0.298144   5.592 2.25e-08 ***
## ChestPainTypeATA -1.944001   0.364373  -5.335 9.54e-08 ***
## ChestPainTypeNAP -1.788925   0.282340  -6.336 2.36e-10 ***
## ChestPainTypeTA  -1.232921   0.476543  -2.587 0.009675 **
## FastingBS1         1.123566   0.289740   3.878 0.000105 ***
## RestingECGNormal  0.002042   0.293660   0.007 0.994451
## RestingECGST       0.107087   0.390844   0.274 0.784093
## ExerciseAnginaY    0.702816   0.268107   2.621 0.008757 **
## ST_SlopeFlat       1.458303   0.470061   3.102 0.001920 **
## ST_SlopeUp        -0.789930   0.485254  -1.628 0.103553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stepwise Logistic Regression

- Max VIF value: 3.0
- AIC: 520.25



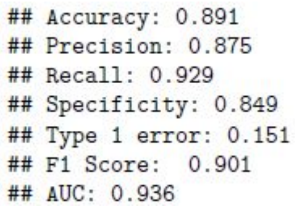
```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ Age + RestingBP + Cholesterol + MaxHR + Oldpeak +
##     Sex + ChestPainType + FastingBS + RestingECG + ExerciseAngina +
##     ST_Slope
## Model 2: HeartDisease ~ Age + Oldpeak + Sex + ChestPainType + FastingBS +
##     ExerciseAngina + ST_Slope
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       718     494.81
## 2       723     498.25 -5  -3.4432    0.632
```



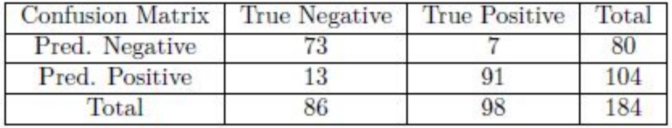
```
## Call:
## glm(formula = HeartDisease ~ Age + Oldpeak + Sex + ChestPainType +
##     FastingBS + ExerciseAngina + ST_Slope, family = binomial,
##     data = train_set)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.0686     0.5248  -2.036  0.04174 *
## Age                0.2647     0.1232   2.149  0.03167 *
## Oldpeak            0.3860     0.1332   2.898  0.00376 **
## SexM               1.6536     0.2933   5.638 1.72e-08 ***
## ChestPainTypeATA  -1.9730     0.3578  -5.514 3.51e-08 ***
## ChestPainTypeNAP  -1.8518     0.2783  -6.653 2.87e-11 ***
## ChestPainTypeTA   -1.2985     0.4722  -2.750  0.00596 **
## FastingBS1         1.1498     0.2872   4.004 6.23e-05 ***
## ExerciseAnginaY    0.7999     0.2581   3.099  0.00194 **
## ST_SlopeFlat       1.5095     0.4615   3.271  0.00107 **
## ST_SlopeUp        -0.8356     0.4754  -1.758  0.07882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stepwise Logistic Regression

❖ **Recall (True Positive Rate):** Leading performance metric in our context



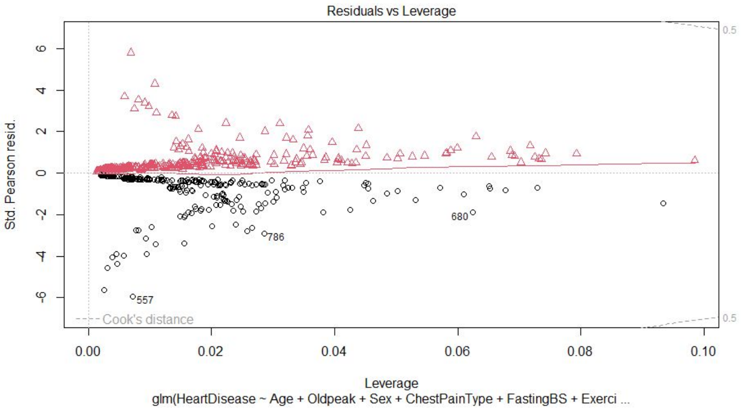
```
## Accuracy: 0.891
## Precision: 0.875
## Recall: 0.929
## Specificity: 0.849
## Type 1 error: 0.151
## F1 Score:  0.901
## AUC: 0.936
```

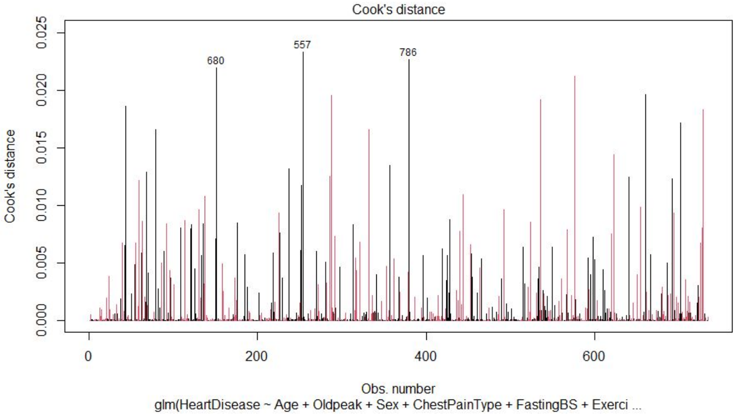


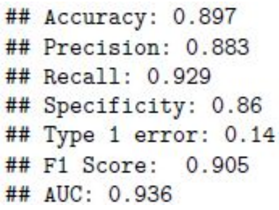
Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	73	7	80
Pred. Positive	13	91	104
Total	86	98	184

Stepwise Logistic Regression - Clean

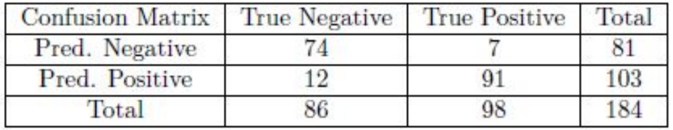
- AIC: 503.91







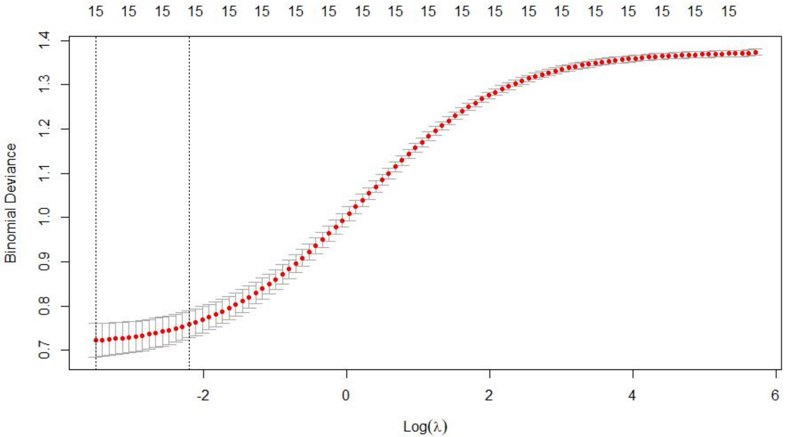
```
## Accuracy: 0.897
## Precision: 0.883
## Recall: 0.929
## Specificity: 0.86
## Type 1 error: 0.14
## F1 Score:  0.905
## AUC: 0.936
```

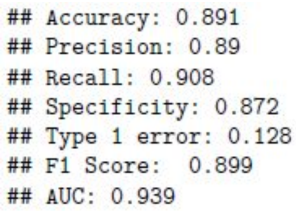


Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	74	7	81
Pred. Positive	12	91	103
Total	86	98	184

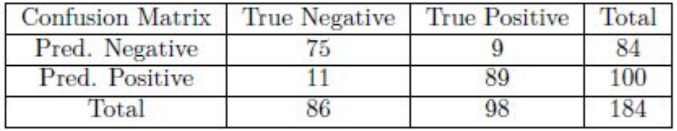
Ridge Logistic Regression

- AIC: -376.33





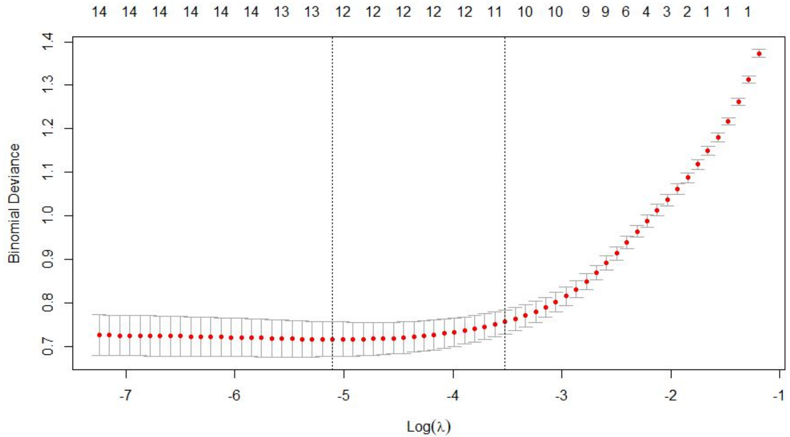
```
## Accuracy: 0.891
## Precision: 0.89
## Recall: 0.908
## Specificity: 0.872
## Type 1 error: 0.128
## F1 Score:  0.899
## AUC: 0.939
```

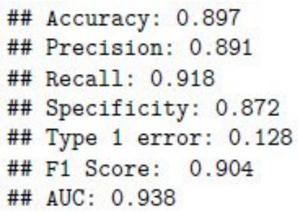


Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	75	9	84
Pred. Positive	11	89	100
Total	86	98	184

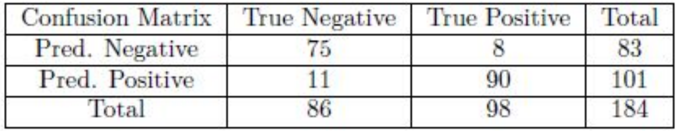
Lasso Logistic Regression

- AIC: -391.77





```
## Accuracy: 0.897
## Precision: 0.891
## Recall: 0.918
## Specificity: 0.872
## Type 1 error: 0.128
## F1 Score:  0.904
## AUC: 0.938
```

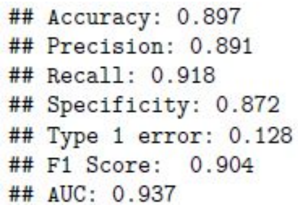


Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	75	8	83
Pred. Positive	11	90	101
Total	86	98	184

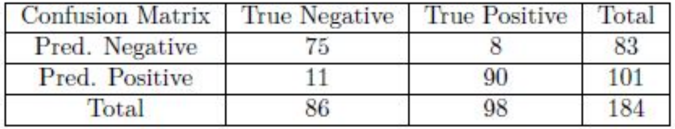
LDA and QDA

Linear Discriminant Analysis

- AIC: 677.22



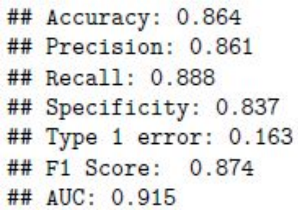
```
## Accuracy: 0.897
## Precision: 0.891
## Recall: 0.918
## Specificity: 0.872
## Type 1 error: 0.128
## F1 Score:  0.904
## AUC: 0.937
```



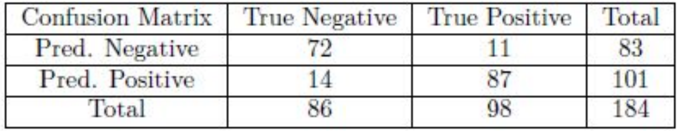
Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	75	8	83
Pred. Positive	11	90	101
Total	86	98	184

Quadratic Discriminant Analysis

- AIC: 1069.21



```
## Accuracy: 0.864
## Precision: 0.861
## Recall: 0.888
## Specificity: 0.837
## Type 1 error: 0.163
## F1 Score:  0.874
## AUC: 0.915
```



Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	72	11	83
Pred. Positive	14	87	101
Total	86	98	184

Data Interpretation

Lasso and LDA Models

❖ Same performance

❖ Lasso is more flexible

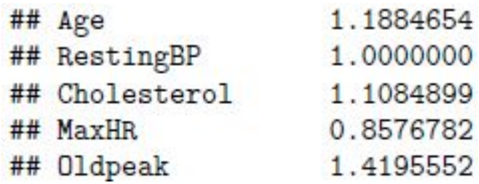
➔ Robustness: no assumptions on predictors distribution

➔ Interpretability: inherent feature selection with shrinkage

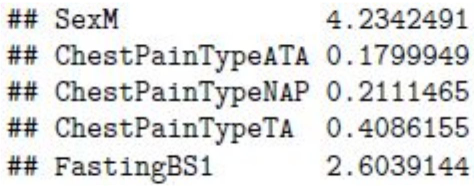
❖ **Most influential variables:**

1. Oldpeak, Sex, ChestPainType, FastingBS, ExerciseAngina, and ST_Slope
2. Age, Cholesterol, and MaxHR

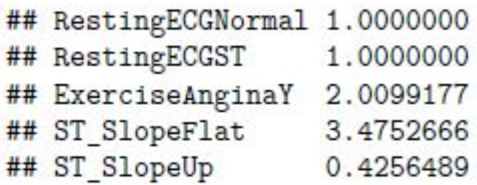
Lasso Odds Ratios:



```
## Age              1.1884654
## RestingBP        1.0000000
## Cholesterol      1.1084899
## MaxHR            0.8576782
## Oldpeak          1.4195552
```



```
## SexM              4.2342491
## ChestPainTypeATA  0.1799949
## ChestPainTypeNAP  0.2111465
## ChestPainTypeTA   0.4086155
## FastingBS1        2.6039144
```



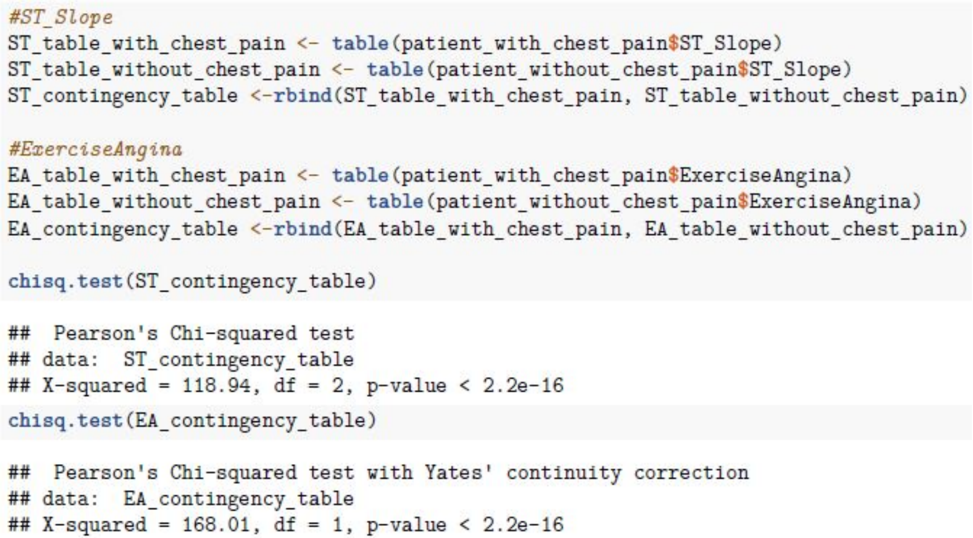
```
## RestingECGNormal 1.0000000
## RestingECGST     1.0000000
## ExerciseAnginaY  2.0099177
## ST_SlopeFlat     3.4752666
## ST_SlopeUp       0.4256489
```

Considerations on the ChestPainType Variable

❖ Asymptomatic (No-pain) as a strong predictor is counterintuitive

➔ Most of heart diseases do not bring chest pain as a symptom

Further analysis:



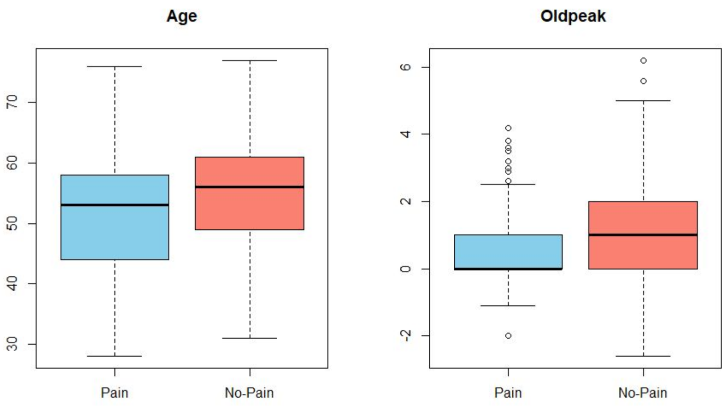
```
#ST_Slope
ST_table_with_chest_pain <- table(patient_with_chest_pain$ST_Slope)
ST_table_without_chest_pain <- table(patient_without_chest_pain$ST_Slope)
ST_contingency_table <-rbind(ST_table_with_chest_pain, ST_table_without_chest_pain)

#ExerciseAngina
EA_table_with_chest_pain <- table(patient_with_chest_pain$ExerciseAngina)
EA_table_without_chest_pain <- table(patient_without_chest_pain$ExerciseAngina)
EA_contingency_table <-rbind(EA_table_with_chest_pain, EA_table_without_chest_pain)

chisq.test(ST_contingency_table)

##  Pearson's Chi-squared test
## data:  ST_contingency_table
## X-squared = 118.94, df = 2, p-value < 2.2e-16
chisq.test(EA_contingency_table)

##  Pearson's Chi-squared test with Yates' continuity correction
## data:  EA_contingency_table
## X-squared = 168.01, df = 1, p-value < 2.2e-16
```



Considerations on the ChestPainType Variable

❖ Asymptomatic patients more connected with risk factors

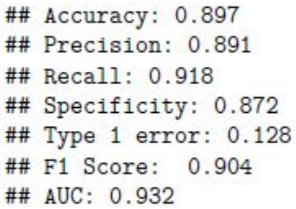
➔ Oldpeak, Age, ST_Slope, and ExerciseAngina

❖ **Confounding Effect**

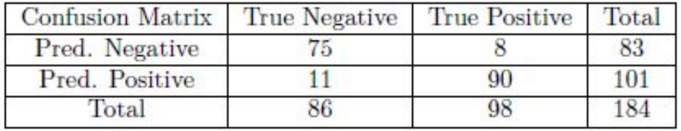
★ Remove the ChestPainType variable

Final Model: Lasso Logistic Regression without ChestPainType

❖ AIC: -397.77



```
## Accuracy: 0.897
## Precision: 0.891
## Recall: 0.918
## Specificity: 0.872
## Type 1 error: 0.128
## F1 Score:  0.904
## AUC: 0.932
```



Confusion Matrix	True Negative	True Positive	Total
Pred. Negative	75	8	83
Pred. Positive	11	90	101
Total	86	98	184

Conclusions and Potential Applications

Risk Factors

❖ **Primary Risk Factors:**

➢ Male sex

➢ High oldpeak values

➢ Fasting blood sugar higher than 120 mg/dL

➢ Exercise angina

➢ Flat ST

❖ **Secondary Risk Factors:**

➢ Old Age

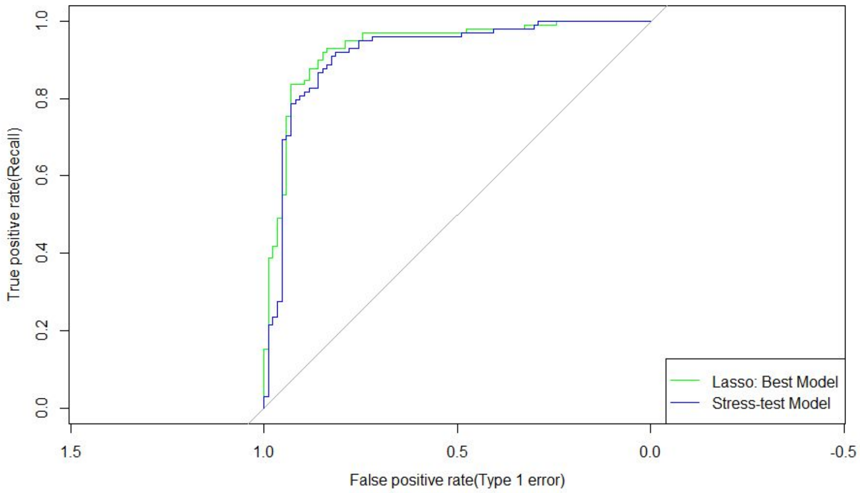
➢ High cholesterol levels

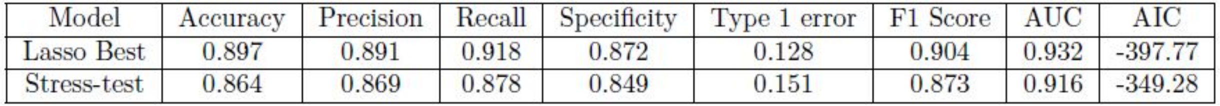
➢ Low maximum heart rate during exercise

Most of these risk factors can be evaluated by performing cardiac stress tests.

Stress Tests

Best Model without Cholesterol and FastingBS





Model	Accuracy	Precision	Recall	Specificity	Type 1 error	F1 Score	AUC	AIC
Lasso Best	0.897	0.891	0.918	0.872	0.128	0.904	0.932	-397.77
Stress-test	0.864	0.869	0.878	0.849	0.151	0.873	0.916	-349.28

Thank You!