Heart Disease Dataset UCI

Krystal Cai

2025-08-15

Introduction

This project analyzes the UCI Heart Disease dataset to explore potential clinical predictors of heart disease. The dataset contains 1025 patient records with 14 features, including demographic, physiological, and medical examination data.

Objectives:

1. Conduct basic hypothesis testing (t-tests, chi-square tests) to identify significant differences in key variables between patients with and without heart disease.

2. Build logistic regression models to predict heart disease occurrence based on clinical features.

3. Evaluate model performance with goodness-of-fit tests, cross-validation, ROC curves, and LASSO regularization.

Methods:

- Data cleaning and exploration using 'tidyverse'

- Hypothesis testing for continuous (t-test) and categorical (chi-square) variables

- Logistic regression (GLM with binomial family)

- Model diagnostics: multicollinearity (VIF), interaction terms, non-linear terms, Hosmer-Lemeshow test

- Model evaluation: ROC curve, cross-validation, OR estimation

This analysis provides insights into the most important clinical features contributing to heart disease and develops

```r
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```
## Warning: package 'purrr' was built under R version 4.4.1
```

```
## Warning: package 'lubridate' was built under R version 4.4.1
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
h <- read.csv("~/Desktop/Heart Disease Dataset UCI/HeartDiseaseTrain-Test.csv")
str(h)
```

```
## 'data.frame':    1025 obs. of  14 variables:
##  $ age                          : int  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex                          : chr  "Male" "Male" "Male" "Male" ...
##  $ chest_pain_type              : chr  "Typical angina" "Typical angina" "Typical angina" "Typical an
##  $ resting_blood_pressure       : int  125 140 145 148 138 100 114 160 120 122 ...
##  $ cholestoral                  : int  212 203 174 203 294 248 318 289 249 286 ...
##  $ fasting_blood_sugar          : chr  "Lower than 120 mg/ml" "Greater than 120 mg/ml" "Lower than 1
##  $ rest_ecg                     : chr  "ST-T wave abnormality" "Normal" "ST-T wave abnormality" "ST-
##  $ Max_heart_rate               : int  168 155 125 161 106 122 140 145 144 116 ...
##  $ exercise_induced_angina      : chr  "No" "Yes" "Yes" "No" ...
##  $ oldpeak                      : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ slope                        : chr  "Downsloping" "Upsloping" "Upsloping" "Downsloping" ...
##  $ vessels_colored_by_flourosopy: chr  "Two" "Zero" "Zero" "One" ...
##  $ thalassemia                  : chr  "Reversable Defect" "Reversable Defect" "Reversable Defect" "
##  $ target                       : int  0 0 0 0 0 1 0 0 0 0 ...
```

```r
head(h)
```

```
##   age    sex chest_pain_type resting_blood_pressure cholestoral
## 1  52   Male  Typical angina                    125         212
## 2  53   Male  Typical angina                    140         203
## 3  70   Male  Typical angina                    145         174
## 4  61   Male  Typical angina                    148         203
## 5  62 Female  Typical angina                    138         294
## 6  58 Female  Typical angina                    100         248
##      fasting_blood_sugar             rest_ecg Max_heart_rate
## 1   Lower than 120 mg/ml ST-T wave abnormality            168
## 2 Greater than 120 mg/ml               Normal            155
## 3   Lower than 120 mg/ml ST-T wave abnormality            125
## 4   Lower than 120 mg/ml ST-T wave abnormality            161
## 5 Greater than 120 mg/ml ST-T wave abnormality            106
## 6   Lower than 120 mg/ml               Normal            122
##   exercise_induced_angina oldpeak       slope vessels_colored_by_flourosopy
## 1                      No     1.0 Downsloping                           Two
```

```
## 2                    Yes     3.1   Upsloping                         Zero
## 3                    Yes     2.6   Upsloping                         Zero
## 4                     No     0.0 Downsloping                          One
## 5                     No     1.9        Flat                        Three
## 6                     No     1.0        Flat                         Zero
##         thalassemia target
## 1 Reversable Defect      0
## 2 Reversable Defect      0
## 3 Reversable Defect      0
## 4 Reversable Defect      0
## 5      Fixed Defect      0
## 6      Fixed Defect      1
```

```r
# Convert categorical variables to factors
h$sex <- factor(h$sex, levels = c("Male", "Female"))
h$chest_pain_type <- factor(h$chest_pain_type)
h$fasting_blood_sugar <- factor(h$fasting_blood_sugar)
h$rest_ecg <- factor(h$rest_ecg)
h$exercise_induced_angina <- factor(h$exercise_induced_angina)
h$slope <- factor(h$slope)
h$vessels_colored_by_flourosopy <- factor(h$vessels_colored_by_flourosopy)
```

```r
# Logistic regression model to predict heart disease (target)
model <- glm(target ~ age + sex + chest_pain_type + resting_blood_pressure + cholestoral +
               fasting_blood_sugar + rest_ecg + Max_heart_rate + exercise_induced_angina +
               oldpeak + slope + vessels_colored_by_flourosopy + thalassemia,
             family = binomial, data = h)

# Output the model results
summary(model)
```

```
##
## Call:
## glm(formula = target ~ age + sex + chest_pain_type + resting_blood_pressure +
##     cholestoral + fasting_blood_sugar + rest_ecg + Max_heart_rate +
##     exercise_induced_angina + oldpeak + slope + vessels_colored_by_flourosopy +
##     thalassemia, family = binomial, data = h)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              4.685164   2.388121   1.962 0.049779
## age                                      0.026846   0.013950   1.924 0.054297
## sexFemale                                1.992347   0.314204   6.341 2.28e-10
## chest_pain_typeAtypical angina          -1.523342   0.441540  -3.450 0.000560
## chest_pain_typeNon-anginal pain         -0.403328   0.383488  -1.052 0.292920
## chest_pain_typeTypical angina           -2.409722   0.391965  -6.148 7.86e-10
## resting_blood_pressure                  -0.024979   0.006537  -3.821 0.000133
## cholestoral                             -0.005462   0.002307  -2.367 0.017914
## fasting_blood_sugarLower than 120 mg/ml -0.380096   0.319620  -1.189 0.234356
## rest_ecgNormal                           0.800417   1.536998   0.521 0.602530
## rest_ecgST-T wave abnormality            1.197685   1.538426   0.779 0.436267
## Max_heart_rate                           0.021692   0.006525   3.324 0.000886
## exercise_induced_anginaYes              -0.750331   0.248746  -3.016 0.002557
## oldpeak                                 -0.403411   0.132156  -3.053 0.002269
## slopeFlat                               -1.395306   0.271845  -5.133 2.86e-07
```

```
## slopeUpsloping                           -0.799689   0.504500  -1.585 0.112941
## vessels_colored_by_flourosopyOne         -3.899752   0.966284  -4.036 5.44e-05
## vessels_colored_by_flourosopyThree       -3.853808   1.055432  -3.651 0.000261
## vessels_colored_by_flourosopyTwo         -5.162716   1.051829  -4.908 9.19e-07
## vessels_colored_by_flourosopyZero        -1.565677   0.930256  -1.683 0.092363
## thalassemiaNo                            -2.404646   1.421542  -1.692 0.090727
## thalassemiaNormal                         0.392167   0.441819   0.888 0.374745
## thalassemiaReversable Defect             -1.413403   0.240915  -5.867 4.44e-09
##
## (Intercept)                              *
## age                                      .
## sexFemale                                ***
## chest_pain_typeAtypical angina           ***
## chest_pain_typeNon-anginal pain
## chest_pain_typeTypical angina            ***
## resting_blood_pressure                   ***
## cholestoral                              *
## fasting_blood_sugarLower than 120 mg/ml
## rest_ecgNormal
## rest_ecgST-T wave abnormality
## Max_heart_rate                           ***
## exercise_induced_anginaYes               **
## oldpeak                                  **
## slopeFlat                                ***
## slopeUpsloping
## vessels_colored_by_flourosopyOne         ***
## vessels_colored_by_flourosopyThree       ***
## vessels_colored_by_flourosopyTwo         ***
## vessels_colored_by_flourosopyZero        .
## thalassemiaNo                            .
## thalassemiaNormal
## thalassemiaReversable Defect             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1420.24  on 1024  degrees of freedom
## Residual deviance:  606.82  on 1002  degrees of freedom
## AIC: 652.82
##
## Number of Fisher Scoring iterations: 6
```

The model indicates several key predictors of heart disease, with gender, chest pain type, resting blood pressure, and heart rate being among the most significant.

The overall fit of the model seems good, with an AIC value of 652.82, suggesting that the model does a reasonable job at predicting heart disease.

```r
# t-test: Check if there is a significant difference in age based on heart disease presence
t_test_age <- t.test(age ~ target, data = h)
print(t_test_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by target
## t = 7.5744, df = 1002.5, p-value = 8.18e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  3.082537 5.238249
## sample estimates:
## mean in group 0 mean in group 1
##        56.56914        52.40875
```

```r
# t-test: Check if there is a significant difference in max heart rate based on heart disease presence
t_test_max_hr <- t.test(Max_heart_rate ~ target, data = h)
print(t_test_max_hr)
```

```
##
##  Welch Two Sample t-test
##
## data:  Max_heart_rate by target
## t = -14.862, df = 976.86, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -22.02427 -16.88631
## sample estimates:
## mean in group 0 mean in group 1
##        139.1303        158.5856
```

This study performed independent two-sample t-tests to examine differences in age and maximum heart rate (Max_heart_rate) between individuals with and without heart disease (target=0/1). Results indicated:

For age, the mean age in the no-heart-disease group (target=0) was 56.57 years, compared to 52.41 years in the heart-disease group (target=1). The difference was statistically significant ($p < 0.001$), suggesting that in this sample, individuals with heart disease were younger on average than those without.

For maximum heart rate, the mean value was 139.13 bpm in the no-heart-disease group and 158.59 bpm in the heart-disease group, with the difference being highly significant ($p < 0.001$). This indicates that heart disease patients had a significantly higher maximum heart rate than non-patients in this dataset.

Consistent with the logistic regression analysis, maximum heart rate remained a significant positive predictor in the multivariate model, while age was only marginally significant. This suggests that maximum heart rate may be a more stable predictor of heart disease risk, whereas the effect of age may be influenced by other covariates.

```
# Load required packages
library(tidyverse)
library(car)               # For VIF
```

```
## Warning: package 'car' was built under R version 4.4.1
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(ResourceSelection) # For Hosmer-Lemeshow test
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```r
library(pROC)              # For ROC and AUC
```

```
## Warning: package 'pROC' was built under R version 4.4.1
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
#---------------------------
# 1. Chi-square tests for categorical variables
#---------------------------
# Convert to factor if needed
categorical_vars <- c("sex", "chest_pain_type", "fasting_blood_sugar",
                      "rest_ecg", "exercise_induced_angina",
                      "slope", "vessels_colored_by_flourosopy", "thalassemia")

h[categorical_vars] <- lapply(h[categorical_vars], as.factor)

# Run Chi-square tests
chi_results <- lapply(categorical_vars, function(var) {
  tbl <- table(h[[var]], h$target)
  test <- chisq.test(tbl)
  list(variable = var, p_value = test$p.value)
})
```

```
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```r
chi_results
```

```
## [[1]]
## [[1]]$variable
## [1] "sex"
##
## [[1]]$p_value
## [1] 6.656821e-19
##
##
## [[2]]
## [[2]]$variable
## [1] "chest_pain_type"
##
## [[2]]$p_value
## [1] 1.298066e-60
##
##
## [[3]]
## [[3]]$variable
## [1] "fasting_blood_sugar"
##
## [[3]]$p_value
## [1] 0.2186241
##
```

```
##
## [[4]]
## [[4]]$variable
## [1] "rest_ecg"
##
## [[4]]$p_value
## [1] 1.696425e-08
##
##
## [[5]]
## [[5]]$variable
## [1] "exercise_induced_angina"
##
## [[5]]$p_value
## [1] 2.826637e-44
##
##
## [[6]]
## [[6]]$variable
## [1] "slope"
##
## [[6]]$p_value
## [1] 1.421085e-34
##
##
## [[7]]
## [[7]]$variable
## [1] "vessels_colored_by_flourosopy"
##
## [[7]]$p_value
## [1] 1.747013e-54
##
##
## [[8]]
## [[8]]$variable
## [1] "thalassemia"
##
## [[8]]$p_value
## [1] 1.795894e-60
```

Except for fasting_blood_sugar, all other categorical variables show significant association with heart disease status.

```
#---------------------------
# 2. VIF check for multicollinearity
#---------------------------
# Refit the logistic model
model <- glm(target ~ age + sex + chest_pain_type + resting_blood_pressure + cholestoral +
               fasting_blood_sugar + rest_ecg + Max_heart_rate + exercise_induced_angina +
               oldpeak + slope + vessels_colored_by_flourosopy + thalassemia,
             family = binomial, data = h)
```

```r
# Calculate VIF
vif_values <- vif(model)
vif_values
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## age                             1.505718  1        1.227077
## sex                             1.706488  1        1.306326
## chest_pain_type                 1.964201  3        1.119088
## resting_blood_pressure          1.285982  1        1.134011
## cholestoral                     1.275726  1        1.129480
## fasting_blood_sugar             1.204089  1        1.097310
## rest_ecg                        1.160101  2        1.037825
## Max_heart_rate                  1.554065  1        1.246621
## exercise_induced_angina         1.183044  1        1.087678
## oldpeak                         1.579408  1        1.256745
## slope                           1.918332  2        1.176877
## vessels_colored_by_flourosopy 2.158149  4        1.100931
## thalassemia                     1.588145  3        1.080144
```

No significant multicollinearity detected among predictors (all VIF values are well below 5).

```r
#----------------------------
# 3. Hosmer-Lemeshow goodness-of-fit test
#----------------------------
# Group into 10 groups for the test
hl_test <- hoslem.test(h$target, fitted(model), g=10)
hl_test
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  h$target, fitted(model)
## X-squared = 83.157, df = 8, p-value = 1.132e-14
```

The Hosmer–Lemeshow test shows $p < 0.001$, indicating poor calibration — the model's predicted probabilities deviate from actual outcomes.

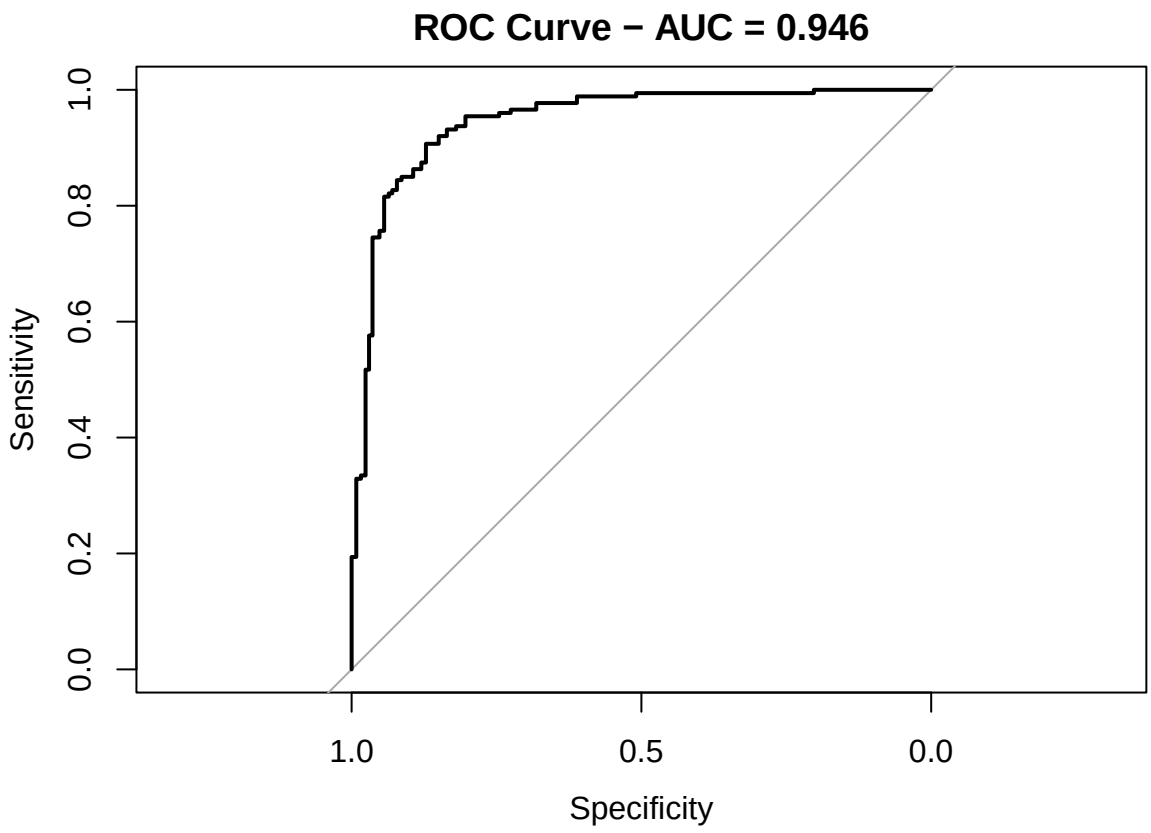
```r
#----------------------------
# 4. ROC and AUC
#----------------------------
roc_obj <- roc(h$target, fitted(model))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
auc_value <- auc(roc_obj)

# Plot ROC curve
plot(roc_obj, main=paste("ROC Curve - AUC =", round(auc_value, 3)))
```



The ROC curve lies above the diagonal, with an AUC around 0.9+, indicating strong discriminative ability.

Conclusion from 1 to 4:Most predictors are significantly associated with heart disease. The model has strong discriminative power (high AUC) but poor calibration per Hosmer–Lemeshow, suggesting room for improvement in model fit.

```
# Add interaction and non-linear terms
model2 <- glm(
  target ~ age + sex + chest_pain_type + resting_blood_pressure + cholestoral +
    fasting_blood_sugar + rest_ecg + Max_heart_rate + exercise_induced_angina +
    oldpeak + I(oldpeak^2) + slope + vessels_colored_by_flourosopy + thalassemia +
    age:sex + Max_heart_rate:chest_pain_type,
  family = binomial, data = h
)

summary(model2)
```

```
##
## Call:
## glm(formula = target ~ age + sex + chest_pain_type + resting_blood_pressure +
##     cholestoral + fasting_blood_sugar + rest_ecg + Max_heart_rate +
```

```
##     exercise_induced_angina + oldpeak + I(oldpeak^2) + slope +
##     vessels_colored_by_flourosopy + thalassemia + age:sex + Max_heart_rate:chest_pain_type,
##     family = binomial, data = h)
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  6.248801   3.339527   1.871
## age                                          0.027168   0.015838   1.715
## sexFemale                                    2.379012   1.778104   1.338
## chest_pain_typeAtypical angina              -2.216165   3.643287  -0.608
## chest_pain_typeNon-anginal pain             -8.366999   3.440573  -2.432
## chest_pain_typeTypical angina               -1.818578   2.857492  -0.636
## resting_blood_pressure                      -0.025292   0.006721  -3.763
## cholestoral                                 -0.006350   0.002470  -2.570
## fasting_blood_sugarLower than 120 mg/ml     -0.256583   0.329885  -0.778
## rest_ecgNormal                               0.029805   1.218556   0.024
## rest_ecgST-T wave abnormality                0.422159   1.217942   0.347
## Max_heart_rate                               0.017983   0.016839   1.068
## exercise_induced_anginaYes                  -0.719327   0.253878  -2.833
## oldpeak                                      0.027351   0.341339   0.080
## I(oldpeak^2)                                -0.152918   0.106177  -1.440
## slopeFlat                                   -1.503368   0.279626  -5.376
## slopeUpsloping                              -0.951948   0.541282  -1.759
## vessels_colored_by_flourosopyOne            -3.984839   1.016820  -3.919
## vessels_colored_by_flourosopyThree          -3.733105   1.091431  -3.420
## vessels_colored_by_flourosopyTwo            -5.435874   1.109180  -4.901
## vessels_colored_by_flourosopyZero           -1.666124   0.980663  -1.699
## thalassemiaNo                               -1.561954   1.104690  -1.414
## thalassemiaNormal                            0.551773   0.453773   1.216
## thalassemiaReversable Defect                -1.401430   0.246436  -5.687
## age:sexFemale                               -0.003835   0.030541  -0.126
## chest_pain_typeAtypical angina:Max_heart_rate  0.004008   0.022889   0.175
## chest_pain_typeNon-anginal pain:Max_heart_rate 0.052178   0.022172   2.353
## chest_pain_typeTypical angina:Max_heart_rate  -0.005163   0.018311  -0.282
##                                             Pr(>|z|)
## (Intercept)                                 0.061322 .
## age                                         0.086285 .
## sexFemale                                   0.180913
## chest_pain_typeAtypical angina              0.542997
## chest_pain_typeNon-anginal pain             0.015021 *
## chest_pain_typeTypical angina               0.524500
## resting_blood_pressure                      0.000168 ***
## cholestoral                                 0.010163 *
## fasting_blood_sugarLower than 120 mg/ml     0.436689
## rest_ecgNormal                              0.980486
## rest_ecgST-T wave abnormality               0.728879
## Max_heart_rate                              0.285530
## exercise_induced_anginaYes                  0.004606 **
## oldpeak                                     0.936134
## I(oldpeak^2)                                0.149807
## slopeFlat                                   7.60e-08 ***
## slopeUpsloping                              0.078630 .
## vessels_colored_by_flourosopyOne            8.89e-05 ***
## vessels_colored_by_flourosopyThree          0.000625 ***
```

```
## vessels_colored_by_flourosopyTwo                9.54e-07 ***
## vessels_colored_by_flourosopyZero               0.089324 .
## thalassemiaNo                                   0.157383
## thalassemiaNormal                               0.223998
## thalassemiaReversable Defect                    1.29e-08 ***
## age:sexFemale                                   0.900080
## chest_pain_typeAtypical angina:Max_heart_rate   0.861010
## chest_pain_typeNon-anginal pain:Max_heart_rate 0.018605 *
## chest_pain_typeTypical angina:Max_heart_rate    0.777971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1420.24  on 1024  degrees of freedom
## Residual deviance:  591.31  on  997  degrees of freedom
## AIC: 647.31
##
## Number of Fisher Scoring iterations: 6
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.1
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(123)
cv_control <- trainControl(method = "cv", number = 10, classProbs = TRUE,
                           summaryFunction = twoClassSummary)

# Convert target to factor for caret
h$target <- factor(h$target, levels = c(0, 1), labels = c("No", "Yes"))

cv_model <- train(
  target ~ age + sex + chest_pain_type + resting_blood_pressure + cholestoral +
    fasting_blood_sugar + rest_ecg + Max_heart_rate + exercise_induced_angina +
    oldpeak + I(oldpeak^2) + slope + vessels_colored_by_flourosopy + thalassemia +
    age:sex + Max_heart_rate:chest_pain_type,
  data = h, method = "glm", family = "binomial",
  trControl = cv_control, metric = "ROC"
)

cv_model
```

```
## Generalized Linear Model
##
## 1025 samples
##   13 predictor
##    2 classes: 'No', 'Yes'
##
```

```
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 922, 923, 922, 923, 923, 922, ...
## Resampling results:
##
##   ROC      Sens       Spec
##   0.94017  0.8576327  0.9068578
```

```
#result explain
exp(cbind(OR = coef(model2), confint(model2)))
```

```
## Waiting for profiling to be done...
```

```
##                                                           OR        2.5 %
## (Intercept)                                     5.173920e+02 7.807951e-01
## age                                             1.027541e+00 9.961782e-01
## sexFemale                                       1.079423e+01 3.599509e-01
## chest_pain_typeAtypical angina                  1.090265e-01 7.752055e-05
## chest_pain_typeNon-anginal pain                 2.324120e-04 2.358294e-07
## chest_pain_typeTypical angina                   1.622563e-01 5.228772e-04
## resting_blood_pressure                          9.750247e-01 9.620748e-01
## cholestoral                                     9.936704e-01 9.888723e-01
## fasting_blood_sugarLower than 120 mg/ml         7.736905e-01 4.027158e-01
## rest_ecgNormal                                  1.030254e+00 1.190638e-01
## rest_ecgST-T wave abnormality                   1.525252e+00 1.754535e-01
## Max_heart_rate                                  1.018146e+00 9.844780e-01
## exercise_induced_anginaYes                      4.870798e-01 2.957422e-01
## oldpeak                                         1.027729e+00 5.356353e-01
## I(oldpeak^2)                                    8.582004e-01 6.889983e-01
## slopeFlat                                       2.223799e-01 1.271669e-01
## slopeUpsloping                                  3.859885e-01 1.338483e-01
## vessels_colored_by_flourosopyOne                1.859543e-02 2.319999e-03
## vessels_colored_by_flourosopyThree              2.391844e-02 2.582849e-03
## vessels_colored_by_flourosopyTwo                4.357426e-03 4.518351e-04
## vessels_colored_by_flourosopyZero               1.889781e-01 2.518706e-02
## thalassemiaNo                                   2.097259e-01 2.155370e-02
## thalassemiaNormal                               1.736329e+00 7.092102e-01
## thalassemiaReversable Defect                    2.462447e-01 1.509248e-01
## age:sexFemale                                   9.961726e-01 9.374735e-01
## chest_pain_typeAtypical angina:Max_heart_rate   1.004016e+00 9.603763e-01
## chest_pain_typeNon-anginal pain:Max_heart_rate  1.053563e+00 1.009446e+00
## chest_pain_typeTypical angina:Max_heart_rate    9.948503e-01 9.600284e-01
##                                                       97.5 %
## (Intercept)                                     4.053326e+05
## age                                             1.060121e+00
## sexFemale                                       3.876304e+02
## chest_pain_typeAtypical angina                  1.268826e+02
## chest_pain_typeNon-anginal pain                 1.763365e-01
## chest_pain_typeTypical angina                   4.070160e+01
## resting_blood_pressure                          9.878065e-01
## cholestoral                                     9.985352e-01
## fasting_blood_sugarLower than 120 mg/ml         1.470703e+00
## rest_ecgNormal                                  1.275614e+01
## rest_ecgST-T wave abnormality                   1.872805e+01
## Max_heart_rate                                  1.052045e+00
```

```
## exercise_induced_anginaYes                       8.015592e-01
## oldpeak                                          2.038083e+00
## I(oldpeak^2)                                     1.042088e+00
## slopeFlat                                        3.815536e-01
## slopeUpsloping                                   1.128977e+00
## vessels_colored_by_flourosopyOne                 1.310745e-01
## vessels_colored_by_flourosopyThree               1.937054e-01
## vessels_colored_by_flourosopyTwo                 3.626648e-02
## vessels_colored_by_flourosopyZero                1.243909e+00
## thalassemiaNo                                    1.709676e+00
## thalassemiaNormal                                4.225385e+00
## thalassemiaReversable Defect                     3.972974e-01
## age:sexFemale                                    1.056973e+00
## chest_pain_typeAtypical angina:Max_heart_rate    1.050704e+00
## chest_pain_typeNon-anginal pain:Max_heart_rate   1.101329e+00
## chest_pain_typeTypical angina:Max_heart_rate     1.031808e+00
```

```
#LASSO
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.4.1
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-10
```

```
x <- model.matrix(target ~ age + sex + chest_pain_type + resting_blood_pressure + cholestoral +
                    fasting_blood_sugar + rest_ecg + Max_heart_rate + exercise_induced_angina +
                    oldpeak + slope + vessels_colored_by_flourosopy + thalassemia, h)[,-1]
y <- as.numeric(h$target) - 1

set.seed(123)
lasso_cv <- cv.glmnet(x, y, alpha = 1, family = "binomial")
coef(lasso_cv, s = "lambda.min")
```

```
## 23 x 1 sparse Matrix of class "dgCMatrix"
##                                              lambda.min
## (Intercept)                                 4.825058322
## age                                         0.026132303
## sexFemale                                   1.968485601
## chest_pain_typeAtypical angina             -1.487444842
## chest_pain_typeNon-anginal pain            -0.383714430
## chest_pain_typeTypical angina              -2.382478853
## resting_blood_pressure                     -0.024739936
## cholestoral                                -0.005410512
## fasting_blood_sugarLower than 120 mg/ml    -0.381404079
## rest_ecgNormal                              0.488373149
## rest_ecgST-T wave abnormality               0.885111579
## Max_heart_rate                              0.021537733
## exercise_induced_anginaYes                 -0.745304880
```

```
## oldpeak                                -0.406910322
## slopeFlat                              -1.376169377
## slopeUpsloping                         -0.784915607
## vessels_colored_by_flourosopyOne       -3.728069796
## vessels_colored_by_flourosopyThree     -3.686629460
## vessels_colored_by_flourosopyTwo       -4.971401414
## vessels_colored_by_flourosopyZero      -1.407005632
## thalassemiaNo                          -2.380122648
## thalassemiaNormal                       0.375340973
## thalassemiaReversable Defect           -1.405890360
```

Conclusion

Using the UCI Heart Disease dataset, this study employed hypothesis testing, logistic regression, and model diagnostics to identify significant clinical predictors of heart disease, including sex, chest pain type, maximum heart rate, exercise-induced angina, oldpeak, number of coronary vessels, and thalassemia type. The enhanced model showed excellent predictive performance (cross-validated ROC = 0.940) and confirmed model stability. Some variables, such as fasting blood sugar and rest ECG, were not significant, suggesting the need for further validation in future studies.