Predicting Heart Disease

Much research is being done on predicting diseases. In this problem, we will use a processed version of the Cleveland Dataset to predict the onset of a heart disease. In this dataset, each row represents a patient with his/her attributes and a binary indicator for whether or not they had a heart disease. We will attempt to understand the factors that can cause a heart disease.

Dataset: data.csv

Here is a detailed description of the variables:

- **Age**: Age in years
- **Sex**: 1 if male, 0 if female
- **ChestPain**: Chest pain type (discrete values 1-4)
- **RestBP**: Resting blood pressure
- **Chol**: Cholestoral
- **FBS**: 1 if fasting blood sugar > 120 mg/dl, 0 otherwise
- **RestECG**: Resting electrocardiographic (discrete values 0-2)
- **Thalach**: Maximum heart rate achieved
- **MajorVessels**: Number of major vessels (discrete values 0-3)
- **HD**:1 if there was a heart disease, 0 otherwise

In this problem, we will use various classification methods to try to predict the onset of a heart disease.

Data Source Creators:

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

Problem 1 - Exploratory Data Analysis

3.0/3.0 points (graded)

Which is the most common amount of major vessels?

- (●) 0
- ( ) 1
- ( ) 2
- ( ) 3

✔

What is the average age in the dataset?

54.5420 ✔

54.5420

What is the average cholesterol within the population of patients who had a heart disease?

251.8540 ✔

251.8540

Submit You have used 2 of 2 attempts

Problem 2.1 - Preparing the Data

2.0/3.0 points (graded)

We will now split the data into a training and testing set. To do this, we use the sample.split() function. Which variable will be used in this function?

- ○ Sex
- ○ RestBP
- ○ Thalach
- ◉ HD ✔

If running R 3.6.0, run the command:

RNGversion("3.5.3")

Set your random seed to 100 and create a training and test set using the sample.split() function in the caTools library, with 70% of the observations in the training set and 30% in the testing set.

Why do we use the sample.split() function?

- ○ It is the most convenient way to randomly split the data
- ○ It balances the independent variables between the training and testing sets
- ◉ It balances the dependent variable between the training and testing sets ✔

How many observation are there in the training set?

207 ✖

207

Submit You have used 2 of 2 attempts

Problem 2.2 - Simple Logistic Regression

0.0/2.0 points (graded)

Train a logistic regression model using Thalach as the independent variable. What is the coefficient of Thalach (the maximum heart rate achieved)?

-0.04660737 ✖

-0.04660737

Submit You have used 2 of 2 attempts

Problem 2.3 - Simple Logistic Regression

0/5 points (graded)

Using your logistic regression model, obtain predictions on the test set. Then, using a probability threshold of 0.5, create a confusion matrix for the test set. What is the (test) accuracy of your logistic regression model?

0.7444 ✖ **Answer:** 0.7303371

0.7444

Our baseline model in classification is to always predict the most frequent outcome in the test set. What is the (test) accuracy of this baseline model?

0.7444 ✖ **Answer:** 0.5393258

0.7444

What is the true positive rate of your logistic regression model?

0.7894 ✖ **Answer:** 0.6829268

0.7894

What is the false positive rate of your logistic regression model?

0.2884 ✖ **Answer:** 0.2291667

0.2884

Currently, we are predicting that there are more patients without heart disease than with. How could we change the model so that more patients are predicted to have heart disease. The motivation for such a change could be to reduce the false negative rate. Which of the following is a way to change that?

- ◯ It is impossible to predict more heart diseases with this model. To change these results, another model can be used.
- ◯ To predict more heart diseases, decrease the prediction threshold. ✔
- ◯ To predict moreheart diseases, increase the prediction threshold.
- ◉ To predict more heart diseases, create more observations with heart diseases.

✖

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 3.1 - Adding More Variables

0/2 points (graded)

Train a logistic regression model now using all of the variables provided.

Which of the following variables are significant at a level of 0.001 or less?

- [ ] Age
- [x] Sex ✔
- [ ] ChestPain ✔
- [ ] RestBP
- [ ] FBS
- [ ] RestECG
- [ ] Thalach ✔
- [x] MajorVessels ✔

✖

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 3.2 - Adding More Variables

0/5 points (graded)

Using your new logistic regression model, obtain predictions on the test set. Then, using a probability threshold of 0.5, create a confusion matrix for the test set.

What is the (test) accuracy of your logistic regression model?

0.8888 ✖ **Answer:** 0.9101124

0.8888

Which of the following is true?

- [ ] About a third of time that there is a heart disease, the model will predict ha heart disease.
- [ ] Over 90% of the times that there is a heart disease, the model will predict a heart disease.
- [x] Almost 90% of the times that there is a heart disease, the model will predict a heart disease. ✔
- [x] About 10% of the times that there is not a heart disease, the model will predict a heart disease.
- [ ] About 25% of the times that there is not a heart disease, the model will predict a heart disease.
- [ ] About 6% of the times that there is not a heart disease, the model will predict a heart disease. ✔

✖

Plot the ROC curve for your logistic regression model. Which logistic regression threshold is associated with the lower-left corner of the ROC plot (true positive rate 0 and false positive rate 0)?

- (•) 0
- ( ) 0.5
- ( ) 1 ✔

✖

At roughly which logistic regression cutoff does the model achieve a true positive rate of 90% and a false positive rate of 30%?

- ○ 0.01
- ○ 0.21
- ○ 0.41 ✔
- ◉ 0.61
- ○ 0.81
- ○ 1.01

✖

What is the AUC for your logistic regression model?

0 ✖ **Answer:** 0.9253049

0

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 4.1 - CART

2.5/5 points (graded)

If running R 3.6.0, run the command:

RNGversion("3.5.3")

Set the random seed to 100.

Then use the caret package and the train function to perform 10-fold cross validation with the training dataset to select the best cp value for a CART model that predicts the dependent variable HD using all of the possible independent variables. Select the cp value from a grid consisting of the values 0.01, 0.011,

0.012, 0.013,..., 0.05.

Remember to convert the HD column to a factor variable.

If you have called your training set train, use the following code:

train$HD = as.factor(train$HD)

Which cp value maximizes the cross-validation accuracy?

5 ✖ **Answer:** 0.034

5

Which of the following is correct?

- ◯ Patients with a chest pain value of 4 will have a heart disease.
- ◉ Women with a maximum heart rate of at least 148 and with a chest pain value of 2 will not have a heart disease.
- ◯ Patients with a chest pain value of 4 and 3 major vessels will not have a heart disease.
- ◯ All women will not have a heart disease.

✔

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 4.2 - CART

0/1 point (graded)

What is the (test) accuracy of your CART model?

0.94

✖ **Answer:** 0.8089888

0.94

Submit

You have used 2 of 2 attempts

Answers are displayed within the problem