Predicting Presence of Heart Disease with Fewer Examinations

Tiffany Lu (tlu17@illinois.edu)

Nov 8, 2020

Contents

Abstract

Cost of medical care has been a concern for many people. The more examinations that a patient need to take, the more money that the patient needs to spend. Moreover, the patient has to expend time and energy to undergo the examinations. We want to see if the patients can still get as accurate heart disease presence detection if the patient go through a fewer number of examinations. In our analysis, we build a model that only use some of the variables to predict presence of heart disease. This model has higher recall rate than that of the model using all features. We believe that this model could bring valuable insights to the prioritization and minimization of examinations for heart disease presence detection.

Introduction

A visit to a doctor costs money, and all the examination needs extra money too. Even with insurance, the cost of medical care is still a big burden to most people. If we can minimize the number of examinations in order to possibly detect for any presence of heart disease, we can help lessen the financial burden that people have to go through during the detection process. Furthermore, reducing the number of tests will reduce the physical and psychological stress that patients have to undergo when doing examinations. There will be less "wait" time before result as well, since there's less examinations that the patients needs to do.

Therefore, our goal is to see if we can build a model with minimal features that can detect presence of heart disease as good as or even better than the model that uses all features. In this case, we will use feature selection with the decision tree's feature importance feature. A good model in this case would have high recall. We want to detect as much people who really have presence of heart disease as possible. In this analysis, we will utilize the decision tree model to determine the important features and use only those features for our new model.

Our original data is obtained from UC Irvin Machine Learning Repository and the data dictionary can be found in the appendix. In the data, our response variable will be "num", which indicates the number of major heart vessels with greater than 50% diameter narrowing. V0 indicates that there's no presence of heart disease and the other values indicates otherwise.

Methods

Data

The original dataset has 15 variables, one of which is our response, num. There are a total of 920 data instances or rows. The columns are the features, and most of them are some sort of health statistic and we can view them as a result of examination. More details can be found in the appendix. Each row represents a patient. The original dataset is from the UC Irvin Machine Learning Repository and was the aggregation of the data collected from Cleveland Clinic Foundation, Hungarian Institute of Cardiology-Budapest, V.A. Medical Center, Long Beach, CA, University Hospital, Zurich, Switzerland. The data is donated on July 1, 1988, which means that it's highly likely that this dataset's result might not be as effective for heart disease detection as health machinary/examinations has been improving throughout the years. Nethertheless, there is still value in analyzing this dataset as we can get an insight to which feature has bigger impact on detection of heart disease.

We are only interested in detecting whether the person has any presence of heart disease or not. The response variable "num" indicates the number of major heart vessels with greater than 50% diameter narrowing. V0 indicates that there's no presence of heart disease and the other values indicates otherwise. We created a new variable, presence, which is 0 if num is "v0" or the person has no presence of heart disease and 1 otherwise. We then remove "num" from our dataset.

Modeling

We want to create a model that allows us to quickly determine if a patient has the possibility of heart disease of not. Every test costs money, and the high cost of examination can be a burden to patients. Therefore, we do not want to request the patient to take all of the tests possible in order to determine if he/she get the disease or not.

For the model, we use a decision tree because 1. it is simple and easy to run compared to other models, 2. it does not assume a shape of model, and 3. we can get the importance of features.

Before feeding the data into a model, we need to split our dataset into testing and training data. We did a 8:2 train-test split. To ensure that that anyone who rerun the code to have consistent result, we set a seed of 0.
***

With our training data, we use 10 fold cross validation to get an average recall for a decision tree model that predicts "presence" using all other variables. We calculate the recall for each cross validation model and take the mean. We also print out the variable importance. According to R documentation, importance value is "sum of the goodness of split measures for each split for which it was the primary variable". We only want to see the top 5variables that are the most important for our prediction.

```
##       cp  thalach location    exang      age
## 89.78678 35.05159 33.94199 24.55729 24.31630
##       cp location  thalach    exang     chol
## 89.95421 51.27225 41.65068 28.13093 26.31340
##       cp location  thalach  oldpeak    exang
## 90.25648 50.38805 33.72579 28.74280 26.40886
##       cp location  thalach    exang  oldpeak
## 89.93943 39.08359 28.94064 23.22939 17.59744
##       cp location  thalach    exang  oldpeak
## 96.68691 47.55765 34.40480 27.08740 26.59261
##       cp location  thalach    exang  oldpeak
## 93.02893 37.65602 33.76026 26.92134 24.99022
##       cp location  thalach  oldpeak    exang
## 86.43607 50.37510 30.01361 27.21368 22.97238
##       cp location  thalach    exang  oldpeak
```

```
## 97.85906 35.73525 34.63382 29.42477 28.72606
##       cp location  oldpeak  thalach    exang
## 92.62854 49.57782 41.91200 35.22265 29.90999
##       cp location  oldpeak  thalach    exang
## 94.32826 54.85405 38.79912 36.85360 28.41413
```

Our mean recall value for the model using all variables is 0.8180023.

Looking at the importance values, we can see that the variables cp, location, and thalach is the top important features for almost all of the cross validation trials. Having location has an variable is really interesting. Location variable tells which hospital is the creator of the data entry. It has nothing to do with the condition of the patient. One possibility is that some hospital admits more patients with heart-related conditions because the hospital has more resources and focus in this field. But for the analysis, we want to know more about the examinations. So we want to omit the location data from the model. We build a decision tree model that uses all features except location. We also print out the top 5 important features for each cross validation. The train recall is 0.8060976.

```
##       cp  thalach      age  oldpeak    exang
## 90.23655 43.15902 39.56375 33.20704 27.31513
##       cp      age  thalach     chol    exang
## 90.28754 33.31707 30.20614 29.28402 25.76874
##       cp      age  thalach    exang  oldpeak
## 90.25648 33.47505 30.24191 26.18274 24.08299
##       cp  thalach    exang     chol  oldpeak
## 89.93943 24.28371 23.22939 22.64219 16.28605
##       cp     chol    exang  oldpeak  thalach
## 96.68691 30.92319 27.08740 26.37087 25.25129
##       cp      age  thalach  oldpeak    exang
## 93.85437 37.95750 35.39245 27.57975 26.92134
##       cp      age  oldpeak  thalach    exang
## 90.76760 32.13787 26.95819 26.75130 23.27402
##       cp  thalach  oldpeak    exang      age
## 99.12597 37.76484 32.94945 30.73609 29.18439
##       cp  oldpeak  thalach    exang      age
## 92.64809 32.95417 31.35979 30.23432 29.55410
##       cp      age  thalach  oldpeak    exang
## 96.79082 36.02895 31.71090 28.91240 28.41413
```

```
## [1] 0.8060976
```

Now, we see that the variables cp, thalach, exang, and oldpeak is always in the top 5 important features. The variables age and chol make it to the top 5 occasionally. We use these 4 variables that is always in the top 5 for our new model to see if we can get a similar or better recall. We use 10 fold cross validation and run a decision tree model again, but this time with only the three variables: cp, thalach, exang, and oldpeak. For this model, we obtained a mean train recall of 0.8266003.

Going back to the importance values, we observed that the difference between the importance value of most important variable (cp) and the second important variable is so much bigger than the difference between the second and third. This brings up an interesting question: how good can a model be if the model only uses the most important variable (cp)? To answer this, we build a decision tree using only cp as variable to predict presence. Then, we run cross validation again to get the average train recall, which is 0.7767712.

Results

Now we have the models, let's use the test data to get the test recall value. The old model with all variables except location has a recall of 0.7525773.

```r
oldmodel = rpart(presence~.-location, data=hd_trn)
prediction = predict(oldmodel, hd_tst, type="class")
get_sensitivity(hd_tst$presence, prediction)
```

```
## [1] 0.7835052
```

The new model with only the cp, thalach, and exang, oldpeak as the variables has a recall of 0.783502.

```r
#reduced features
newmodel = rpart(presence~cp+thalach+exang+oldpeak, data=hd_trn)
prediction = predict(newmodel, hd_tst, type="class")
get_sensitivity(hd_tst$presence, prediction)
```

```
## [1] 0.8350515
```

Discussion

We build two decision tree models in this analysis: one using all variables, one using top 4 important features. The model using the top 4 important features has the higher testing recall (recall = 0.8350515) than the model using all variables. This means that we only need the details of the patient's chest pain type (cp), maximum heartrate achieved (thalach), exercise induced angina (exang), and ST depression induced by exercise relative to rest (oldpeak). This means that the patient only needs to undergo 4 examination to possibly determine if he/she has presence of heart disease. In the doctor suspect that the patient has heart disease, the doctor can prioritize examinations that examines these 4 attributes. The patient then do not have to waste so much time and energy doing all examinations.

All the four top important variables are symptoms related to heart. Chest pain clearly shows that something might be wrong with the heart. People with heart disease would most likely experience chest pain. High heart rate also can be experience when someone have heart disease. Angina is a type of chest pain, and ST segment is something that is used in electrocardiography. Therefore, it makes sense that these variables are important in determine presence of heart disease.

Throughout the process of obtaining the model, we noticed that the variable location is an important variable in determining if the person has presence of heart disease or not. But since this variable has nothing to do with the condition of the patient and is not revelant to our question, we omit it. Having such variable included in our model would cause problems. Hungarian Institute of Cardiology-Budapest is an institute specializing in cardiology. It is not surprising if most patients who go there experience cardiac problems.But the hospital should not be in a cause-effect relationship with heart disease presence.

Appendix

heart disease (hd) data dictionary can be found here: https://archive.ics.uci.edu/ml/datasets/Heart+Disease {hd: {columns: [age, sex, cp trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num, location], numrows: 920} }

age - age in years

sex - sex (1 = male; 0 = female)

cp - chest pain type – Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain – Value 4: asymptomatic

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

chol- serum cholestoral in mg/dl

fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg - resting electrocardiographic results – Value 0: normal – Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach- maximum heart rate achieved

exang - exercise induced angina (1 = yes; 0 = no)

oldpeak - ST depression induced by exercise relative to rest

slope - the slope of the peak exercise ST segment – Value 1: upsloping – Value 2: flat – Value 3: downsloping

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

thal - 3 = normal; 6 = fixed defect; 7 = reversable defect

num - diagnosis of heart disease (angiographic disease status) – Value 0: $< 50\%$ diameter narrowing – Value 1: $> 50\%$ diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)

location - which hospital's data is the patient from details obtained from https://archive.ics.uci.edu/ml/datasets/Heart+Disease