# Machine Learning on Heart Attack Issue

## Introduction

Heart attack (cardiovascular disease) is a collective term for a series of heart diseases that occurs when the flow of blood to the heart muscle suddenly becomes blocked.1 Fat, cholesterol, and other metabolic product are the main component of the blockage. That stuff will form plaque in the arteries which feed the heat (provide fresh blood with plenty of oxygen to the heart).2 Over time, these plaques build up on the lining of blood vessels, forming clots that affect blood pressure and blood flow in the blood vessels. If the heart cannot get enough oxygen supply for a long time, the heart muscle will be demanded.3 These diseases are a very heavy burden to the world. To WHO statistic results, 17.9 million people die from heart attacks every year.4 Some common symptoms include pressure and pain in the chest or arms, shortness of breath, cold sweat, and fatigue. Many reasons can increase the risk of heart attack, while, the medical study says that the human lifestyle is the main reason behind this heart problem.5 Other health-related issues such as blood sugar level and chest pain symptoms will also influence the incidence of a heart attack. Significant heart attack factors can be used to predict the probability of this disease for people with certain health condition features. If the main heart attack demographic features or health condition factors can be found, more effective public health intervention can be proposed and applied to screen people with high heart attack probability. Thus, people can control the disease at an early stage and avoid sudden death. Public health policymakers can better control the incidence of this disease and reduce the burden of a heart attack.

A heart attack-related dataset was analyzed. The data were divided into training data and test data. Different machine learning model building methods were applied to the training data, such as logistic regression, binary classification, KNN, LDA/QDA, and decision tree. The performance between models was measured by the size of the error (training error and test error). Forward/backward and AIC/BIC methods were used to select significant predictors from the best performance model.

After analysis, the binary classification model was the model with the smallest error term and 6 significant predictors were selected from 13 original predictors to predict heart attack probability.

## Related work

There are some previous works that also used machine learning methods to predict heart attack. One of the studies that was conducted in South Korea called “Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree”6was very similar to our study. This study used fuzzy logic and decision tree to make the prediction of coronary heart disease, and successfully improve the prediction accuracy and sensitivity. Our study also used decision tree like this study to make prediction for our dataset. However, we did not use fuzzy logic method because it is an approach to solve problems associated with uncertainty data in order to obtain an array of accurate conclusion. Since we don’t have that problem with our dataset, it is not necessary to use this approach to solve our problem.

**Methods**

There is rationality in assuming some health-related features can be used as predictors of a heart attack. The topic of this paper is to find the significant predictors of a heart attack. The “Health Care: Heart attack possibility” will be used for analysis.714 variables will be used to build models. During them, “target” is the dependent variable, and the rest 13 health-related and demographic features variables are the independent variables. For hypothesis, those 13 independent variables all have a significant predictive ability on a heart attack.

All statistical analyses and model building will be conducted through R and RStudio software. R is a kind of statistical analysis software that allows researchers to easily conduct quantitative and qualitative analyses. The data contained 151 observations. Firstly, different methods will be used on training data to fit a model that “targets” the outcome variable. Those methods are binary classification, LDA, QDA, KNN, decision tree, bagging, random forest, and boosting. For each method, training error and test error should be calculated to evaluate the model performance. As for the first four model approaches, LOOCV and 5-fold CV will also be calculated to compare the model fit. After the evaluation, one best model method with the smallest error will be selected (comprehensively compare the size of these errors).

So many predictors in the model may lead to the problem of overfitting and it is hard to explain a complex model. Thus, only those important predictors should be used. Thus, predictor selection should be done. Predictor selection code (subset selection) in R will be used to select important and useful independent variables between the 13 original predictors. R-square, adjusted R-square, Mallows' -statistic, and BIC values from the subset results will be plotted as the change in model size to show the results visually. We assume that with more predictors in the model, R-square and adjusted R-square will increasingly larger, while, Mallows' Cp-statistic and BIC values will continuously decrease. Forward stepwise selection and backward stepwise selection approaches will be also conducted to check the predictor selection results. Here, the forward and backward stepwise selection results are assumed as the same as subset results. Thus, the final model will be obtained

## Data and experiment setup

“Heart Disease Data Set” in 1988 is the source of the data we used in the project: “Health Care: Heart attack possibility”.8 This data is a public opening, and people can use this data to do some simple analysis. The dataset gets much information including demographic, heart attack disease-related, and health condition-related information. Those 13 independent variables we will use to analyze include: age, sex, chest pain type, resting blood pressure (in mm Hg on admission to the hospital), serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, thal and number of major vessels (0-3) colored by flourosopy.

After model selection and predictor selection.

Firstly, the data will be imported to the RStudio tool from the website. The missing value will be checked then. For building machine learning models, the data and divide into two sets equally: training data and test data. Basic descriptive statistics will be applied to check the data distribution and strange variable values, such as extreme value, median, mean and standard error.

Here is the data distribution information:

Table 1: descriptive statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable name | Variable content | Variable type | Range | Mean | Media |
| Age | Age | Continuous | 29 - 77 | 54.37 | 55.00 |
| Sex | Sex | Binary | 0 (male)  1 (female) | 0.6832 | 1.0000 |
| cp | Chest pain type | Factor | 1 (typical angina)  2 (atypical angina)  3 (non-anginal pain)  4 (asymptomatic) | 0.967 | 1.000 |
| trestbps | Resting blood pressure (in mm Hg) | Continuous | 94.0 – 200.0 | 131.6 | 130.0 |
| chol | Serum cholestoral in mg/dl | Continuous | 126.0 – 564.0 | 246.3 | 240.0 |
| fbs | Fasting blood sugar > 120 mg/dl | Binary | 1 (true)  0 (false) | 0.1485 | 0.0000 |
| restecg | Resting electrocardiographic results | Factor | 0 (normal)  1 (having ST-T wave abnormality)  2 (showing probable or definite left ventricular hypertrophy by Estes` criterial | 0.5281 | 1.0000 |
| thalach | Maximum heart rate achieved | Continuous | 71.0 – 202.0 | 149.6 | 153.0 |
| exang | Exercise induced angina | Binary | 1 (yes)  0 (no) | 0.3267 | 0.0000 |
| oldpeak | ST depression induced by exercise relative to rest | Continuous | 0.00 – 6.20 | 1.04 | 0.80 |
| slope | The slope of the peak exercise ST segment | Factor | 1 (upsloping)  2 (flat)  3 (downsloping) | 1.399 | 1.000 |
| ca | Number of major vessels (0-3) colored by flourosopy | Continuous | 0.0000 – 4.00000 | 0.7294 | 0.0000 |
| thal | NA | Factor | 3 (normal)  6 (fixed defect)  7 (reversable defect) | 2.314 | 2.000 |
| Target | Heat attack | Binary | 0 (no)  1 (yes) | 0.5446 | 1.0000 |

## Results

Here are the 8 kinds of different model fit methods and their error results:

Table 2: model selection with error results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training error | Test error | LOOCV | 5-fold CV |
| Logistic regression | 0.1513158 | 0.1523179\* | 0.1749175 | 0.1775797\* |
| LDA | 0.1644737 | 0.1589404 | 0.1683168\* | 0.1790037 |
| QDA | 0.1381579 | 0.2052980 | 0.2013201 | 0.2156914 |
| KNN(3) | 0.1710526 | 0.3841060 | 0.3828383 | 0.3942336 |
| Decision tree | 0.2171053 | 0.2715232 |  |  |
| Bagging | 0.0000000\* | 0.1986755 |  |  |
| Random forest | 0.0000000\* | 0.1655629 |  |  |
| Boosting | 0.1381579 | 0.1589404 |  |  |

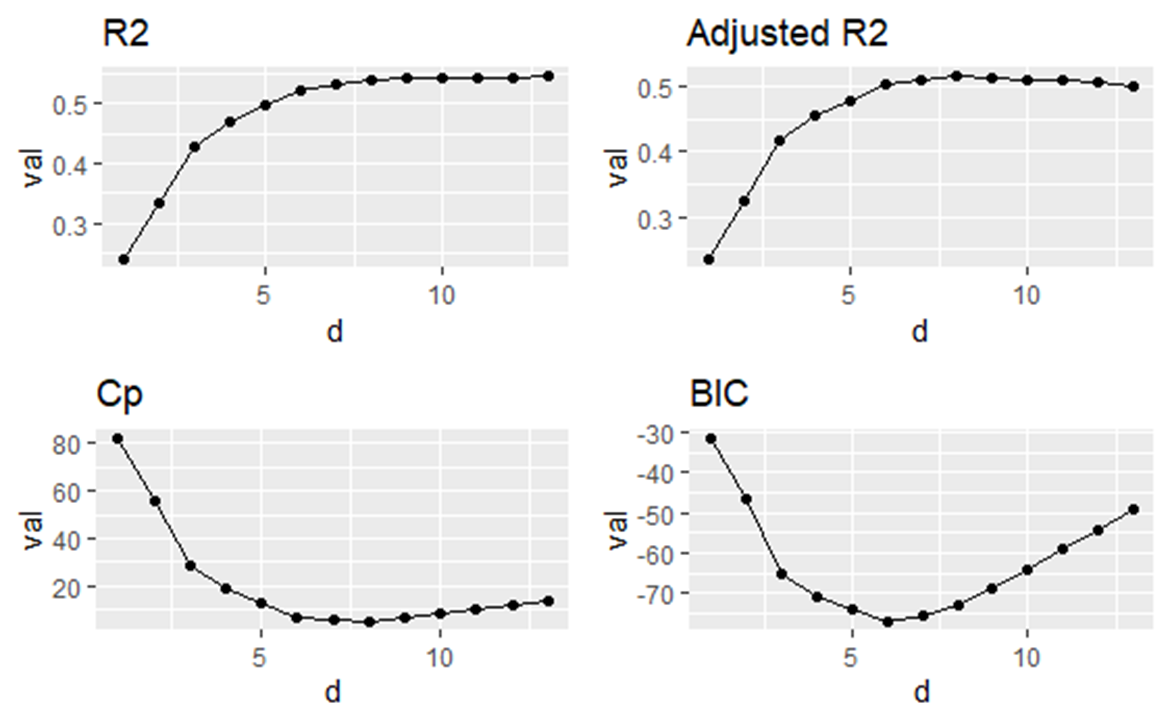
\*: the smallest error within one kind of error

Here, the binary classification method is the most suitable because it has the smallest test error and 5-fold CV. While another error for binary classification is still pretty small compared with other methods.

After selecting the most suitable model method, then, significant predictors should be selected. For predictor selection, subset, forward stepwise, and backward stepwise selection were used. R square, adjusted R square, Mallows' Cp-statistic, AIC, and BIC values are the benchmark for this section. R square and adjusted R square value should be large to explain the proportion of the variance for a dependent variable that's by independent variables. As for Mallows' Cp-statistic, AIC and BIC, they should be smaller to penalize too many parameters (overfit problem) in one model.

Here is the result for predictor selection:

Plot 1: subset result to select predictors



After comparison, 6 predictors are enough to predict the heart attack issue. From the forward stepwise and backward stepwise method, those 6 variables are sex, chest pain type, maximum heart rate achieved, exercise induced angina, number of major vessels colored by flourosopy, and thal.

Here is the final model fit result:

Table 3: final model parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Parameters | Standard error | P -value | Significance\* |
| Intercept | -3.151 | 2.142 | 0.141 | Yes |
| Sex | -2.193 | 0.633 | 0.000527 | Yes |
| Cp | 0.787 | 0.262 | 0.00264 | Yes |
| Thalach | 0.0496 | 0.0141 | 0.000436 | Yes |
| Exang | -1.555 | 0.577 | 0.00705 | Yes |
| Ca | -0.634 | 0.249 | 0.0110 | Yes |
| thal | -1.068 | 0.396 | 0.00694 | Yes |

\*: whether significant at the 0.05 level

In conclusion, the final model and the error values are:

Equation 1:

Equation 2:

Table 4: final model error terms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training error | Test error | LOOCV | 5-fold CV |
| Final model | 0.171 | 0.212 | 0.182 | 0.191 |

From the model, on average, males have 0.112 times the odds of heart attack than females holding other parameters constant. Also, people with one more unit of maximum heart rate achieved will have 1.051 times the odds of a heart attack while holding other parameters constant. For equation 1, when the parameter coefficient is larger than 0, this parameter will perform a positive effect on the heart attack issue (increase the possibility of this disease). When the coefficient is smaller than 0, this parameter will perform a negative effect on heart attack (decrease the possibility of heart attack). For equation 2, the coefficients represent the odds ratio of heart attack that the parameter can perform.

The final model contains 6 significant predictors to predict heart attack issues. While all the four error terms are larger than the former binary classification model with all the predictors. This is rational because 7 predictors are deleted from the model. Some deviance that was originally explained by those 7 predictors now cannot be explained by these remaining 6 predictors. Thus, those four error terms are all changed to be larger than before.

## Discussion

Heart attack issues can be influenced by many factors. From the dataset which was analyzed, after the model method and predictor selection, a binary classification model was conducted to predict the disease. Of the 6 significant predictors, sex, exercise-induced angina, the number of major vessels colored by flourosopy and thal has a negative effect, while, chest pain type and maximum heart rate achieved have a positive effect on a heart attack. The final binary classification model can be used to predict one’s heart attack probability when inputting the 6 parameters’ information.

The result is not completely the same as the hypothesis which supposes that 13 predictors are all significant. Here are only 6 significant predictors. While those predictors play an important role in predicting the possibility of a heart attack. For parameters with a positive coefficient, people can control and avoid that health-related action. For parameters with a negative coefficient that can reduce the odds of the disease, people can engage in more. As for demographic features, which cannot be changed by people, such as sex, a female has larger odds of heart attack than a male when do not consider other parameters, females need to put more attention to the prevention of heart attack. Using this model, people’s heart attack probability can be predicted. For the whole population, this model can be used to screen individuals with a high risk of heart attack, thus, controlling the disease as earlier as possible. If more accurate models can be built to predict disease, more disease cases can be controlled and reduce the population's heart attack burden.

The sample size is small, with only 303 observations in this data. the results cannot be generalized. At the same time, the machine learning model building methods are simple, and thus cannot be applied to complex datasets and conditions. More accurate and novelty model can be conducted in future research.

## Contribution and public link

we have uploaded our code on the Kaggle website (<https://www.kaggle.com/code/nanwang63/notebook4a16a24072>) for reviewing publicly.

For the project, we put in effort together. Lianlian Chen completed the content of the first three parts, and Nan Wan completed the content of the latter half. Throughout the process of completing the project, we communicated promptly and cooperated.

# Reference

1. American Heart Association. “Warning Signs of a Heart Attack.” *Www.heart.org*, 2016, <www.heart.org/en/health-topics/heart-attack/warning-signs-of-a-heart-attack>.

2. Centers for Disease Control and Prevention. “Heart Attack Facts & Statistics.” *CDC*, 2019, <www.cdc.gov/heartdisease/heart_attack.htm>.

3. Cleveland Clinic. “Heart Attack (Myocardial Infarction) | Cleveland Clinic.” *Cleveland Clinic*, 18 July 2019, <my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction>.

4. DeNoon, Daniel J. “Heart Attacks and Heart Disease.” *WebMD*, WebMD, Oct. 2005, <www.webmd.com/heart-disease/guide/heart-disease-heart-attacks>.

5. “Health Care: Heart Attack Possibility.” *Www.kaggle.com*, <www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility>.

6. Kim J, Lee J, Lee Y. Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthcare Informatics Research. https://synapse.koreamed.org/articles/1075743. Published July 31, 2015. Accessed May 10, 2022.

7. Mayo Clinic. “Heart Attack - Symptoms and Causes.” *Mayo Clinic*, 2018, <www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106>.

8. “UCI Machine Learning Repository: Heart Disease Data Set.” *Uci.edu*, 2019, <archive.ics.uci.edu/ml/datasets/Heart+Disease>.