Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230881

Indicators and predictions of heart disease based on machine learning scenarios

Jiaqi Fan

School of Economics, Fudan University, Shanghai 200433, China.

20307110351@fudan.edu.can

Abstract. Nowadays, various patients are suffering from heart disease and even die owing to the disease. According to common knowledge, many health problems can cause heart disease directly or indirectly, e.g., overweight, stroke, high blood pressure, and so on. This study uses Heart Disease Health Indicators Dataset from Kaggle to find out significant indicators of heart disease or heart attack, and predicts heart disease by logistic regression, random forest and LightGBM. Based on the analysis, 10 response variables, including health conditions, living habitats and age are significantly relevant to heart disease. In addition, the comparison among the model shows random forest is the most suitable model to predict heart disease with multicollinearity. This paper selects out important factors of heart disease and provides a fitting model for heart disease prediction. Based on the evaluation models, logistic regression and random forest, this paper finds random forest is the fittest model in prediction. Overall, these results shed light on guiding further exploration of indicators of heart disease.

Keywords: heart disease indicators, logistic regression, random forest, LightGBM.

1. Introduction

Contemporarily, heart disease remains a hot open topic There are plenty of available researches on this topic. Quasinowski et al. explore the globalization of cardiology and its impact on cardiovascular disease through case studies [1]. With a particular focus on heart failure, reaching a conclusion that the globalization of cardiology has a significant impact on the prevention and treatment of cardiovascular diseases worldwide. However, it also poses challenges, particularly in low- and middle-income countries, where transformations in people's lifestyle towards the Western cause an increase in cardiovascular disease but limited medical resources. Schnall et al. explore the relationship among globalization, work, and cardiovascular disease (CVD) and review the evidence from epidemiological studies and occupational health studies. suggest that globalization has led to changes in work and that the resulting factors of job insecurity, long working hours and high job demands have contributed to the development of CVD, and the implications of these findings for public health policy are discussed [2]. Cosselman et al. conduct a review, focusing on cardiovascular influences of two common and familiar environmental problems: surrounding air pollution and non-essential metals. Studies have shown that environmental factors interact with the epigenome, proteome, and metabolome, and that environmental exposure increases susceptibility to cardiovascular diseases [3].

Since the possible indicators of heart disease are widespread recently, this article is going to find out significant indicators of heart disease and find a suitable model to predict heart disease. Recently, many

DOI: 10.54254/2755-2721/16/20230881

scientists focus on exploring key indicators of heart disease. Romero et al. use PubMed literature to evaluate whether serum biomarkers are significant in predicting CVD in asymptomatic patients. The study suggests that several biomarkers may play an important role in advance the prediction of CVD risk in asymptomatic middle aged [4]. Shi et al. explore the relationship between SUA and CVD risk factors in a cross-sectional study of 11,219 adolescents aged 12 to 18 years and found that elevated SUA levels are associated with increased odds of various cardiovascular risk factors in the US adolescent population, and this is more significant in females [5]. Crea et al. conduct a review of the existing papers discussing the effect of inflammation in heart disease and heart attack, and find that inflammation can lead to the occurrence and progression of atherosclerosis, thereby increasing the risk of myocardial infarction [6]. At the same time, inflammation also directly leads to heart muscle damage, increasing the incidence and mortality of heart disease. Wu et al. conduct a systematic review and meta-analysis to explore the association between coronary calcium (CAC) scores and cardiovascular events in asymptomatic patients and find that the risk of cardiovascular events is significantly positively correlated with the CAC score [7]. By analysing tendency of risk factors of CVD in the United States, Yang et al. find that from 2011 to 2018 the prevalence of high blood pressure, high cholesterol, and diabetes decreased in cardiovascular patients in the United States, while the incidence rate of obesity and smoking remained stable, and the risk factors were correlated with age and race [8, 9].

Based on the researches above, physical condition and living habit, like physicalHealth, mentalHealth, alcoholdrinking, smoking are all indicators of heart disease. But the weight of importance and relationship between these indicators are unknown [10, 11]. This article explores the key factors of heart disease in specific data set and select suitable machine learning model to predict heart disease by given indicators. The rest part of the paper is organized as follows. The second part introduces the source and features of the chosen data set, and explains the theories of the three selected models in this paper. The third part show the processed results of the data set, including the correlations among variables and the prediction accuracy of the models, and gives explanations to the results. The last part is conclusion, bringing a brief look of the article.

Table 1. Data describes.

	count	mean	std	min	max
HeartDiseaseorAttack	253680	0.094	0.292	0	1
HighBP	253680	0.429	0.495	0	1
HighChol	253680	0.424	0.494	0	1
CholCheck	253680	0.963	0.190	0	1
BMI	253680	28.382	6.609	12	98
Smoker	253680	0.443	0.497	0	1
Stroke	253680	0.041	0.197	0	1
Diabetes	253680	0.297	0.698	0	2
PhysActivity	253680	0.757	0.429	0	1
Fruits	253680	0.634	0.482	0	1
Veggies	253680	0.811	0.391	0	1
HvyAlcoholConsump	253680	0.056	0.230	0	1
AnyHealthcare	253680	0.951	0.216	0	1
NoDocbcCost	253680	0.084	0.278	0	1
GenHlth	253680	2.511	1.068	1	5
MentHlth	253680	3.185	7.413	0	30
PhysHlth	253680	4.242	8.718	0	30
DiffWalk	253680	0.168	0.374	0	1
Sex	253680	0.440	0.496	0	1
Age	253680	8.032	3.054	1	13

DOI: 10.54254/2755-2721/16/20230881

Table 1. (continued).

Education	253680	5.050	0.986	1	6
Income	253680	6.054	2.071	1	8

2. Data & method

2.1. Data

This paper uses Heart Disease Health Indicators Dataset from Kaggle. The raw data comes from The Behavioral Risk Factor Surveillance System (BRFSS), which is a survey about health held by CDC every year. And it is cleaned from BRFSS 2015 as a binary classification of heart disease. The data set contains 1 binary target variable and 21 feature variables with a sample size of 253680. The dependent variable is HeartDisease Attack, which equals 1 if the sample suffers from heart disease attacks, while it is 0 if it doesn't. 9.4% of the samples are patients with heart disease. The feature variables contain 13 binary variables and 8 ordinal variables, and Table. 1 shows the mean values and standard errors.

2.2. Methods

Logistic regression is used to analyze and model relationships between a binary dependent variable and one or more predictor variables. This study first introduces an activation function $g(z) = 1/(1 + e^{-z})$. The logistic regression model is based on the logistic function $h_\theta(x) = g(\theta^T x)$. Here, θ is parameter vector and x is feature vector of the observed data. Through this function, the linear combination of predictors is transformed into a probability value between 0 and 1. To estimate the parameters of the logistic function to maximize the likelihood, loss function is defined as:

$$J(\theta) = \frac{1}{m}\Sigma_i \left[-y^i \log\left(h_\theta(x^i)\right) - \left(1 - y^i \log\left(1 - h_\theta(x^i)\right)\right)\right] \quad (1)$$

where m is the number of samples of training set, y^i is true value of Sample i. $x^i$is feature vector of Sample i. To minimum the loss function, gradient descent method is used to evaluate the resulting coefficients

$$\theta_j := \theta_j - \frac{\alpha \partial J(\theta)}{\partial \theta_j} \quad (2)$$

where α is learning rate. The result can be used to interpret the effect of each response variable on the dependent variable. Logistic regression is a useful regression model for its ability to handle non-linear relationships between the predictors and the dependent variable, and its ability to model the probability of a binary outcome. Meanwhile, compared with other machine learning models, it is relatively easy to implement and interpret.

Random forest is a machine learning method that classifies or makes regression with multiple decision trees. In a random forest, each decision tree is trained based on a randomly selected subset of different features and samples. In the classification task, the random forest will adopt the majority voting method to determine the final classification result.

The first step of this method is to establish random forest.

A random forest consists of many random trees. To make a decision tree, first make a SRSWR (simple random sample with replacement) from a given sample size of N, and get N samples,which is a pseudo-population. At the node of decision node, using this pseudo-population to train decision trees. At each decision tree's node to be split, if there are M attributes in each sample, then m attributes are randomly selected from the M attributes on the understanding of m << M. Next information gain or some other indexes are chosen from the m attributes to get one attribute as the split attribute of the node.

As a decision tree is forming, all nodes are split like this. If the following chosen attribute of node is used before, it becomes a leaf and stop splitting. Thus, a decision tree eventually forms. And repeating the process of make a decision tree can establishing a random forest.

DOI: 10.54254/2755-2721/16/20230881

The second step of Random Forest method is to train the model with training set. During model training, cross-validation and other methods can evaluate model performance and adjust the parameters of the model. In this process, the random forest will evaluate the importance of each feature, which can be used to further optimize the model. The last step of Random Forest is to predict with the trained model. Data is put into each tree for classification or regression, and finally output the prediction results of the random forest. In the prediction process, cross-validation and other methods can be used to evaluate the prediction of the model. In this paper, the parameters of random forest are set as the following: The number of trees in the forest is set as n_estimators = [50, 100, 500], and maximum depth of each tree is max_depths = [2, 4, 6, 8, 10]. The minimum number of samples required to split internal nodes is set as min_samples_splits = [2, 4, 6], and the minimum number of samples of leaf nodes is set as min_samples_leafs = [1, 2, 5, 10].

LightGBM is a machine learning method based on GBDT (Gradient Boosting Decision Tree). The basic decision tree construction and training is similar to those of random forest, but there are still two main differences between the two methods. First, the decision trees in LightGBM are different from those in Random Forest since they are constructed by leaf-wise tree method, while in Random Forest level- wise tree method is used. The construction method of Level-wise tree takes the number of nodes at each layer as the optimization objective, and selects the optimal nodes in the same layer for splitting each time until the preset tree depth is reached. This method is easier to control the depth of the tree and reduce the risk of overfitting, but may require more training time to achieve a better fitting effect. Meanwhile, the construction method of leaf-wise tree takes the number of Leaf nodes of each node as the optimization objective, and selects the node with the smallest sample number for splitting each time until the preset number of leaf nodes is reached. This method can achieve better fitting effect faster, but it is easy to overfit. Next, in LightGBM, a histogram-based algorithm is used to calculate the gradient. The gradient in Boosting algorithm refers to the first derivative of the objective function to the predicted value of the model. Specifically, LightGBM divides the training data into multiple histograms according to eigenvalues and calculates the gradient for each histogram. This method can greatly reduce the computational complexity and deal with sparse data and missing values effectively. During the training process, LightGBM gradually improves the forecasting ability of the model by continuously optimizing the partitioning of the histogram and calculating the gradient. In the construction process, the improvement of leaf-wise tree will give priority to the direction with the most leaf nodes for splitting, which can reduce the number of nodes and speed up the training speed. And the improvement of histogram based algorithm to calculate gradients reduces memory usage and improves training speed. In this paper, the parameters in LightGBM model are set as following. The number of trees is set as n_estimators = [50, 100, 500, 1000]; the maximum depth of the decision tree is max_depths = [2, 4, 8, 10, 20, 32]; the learning rate is set as learning_rates = [0.005, 0.01, 0.05, 0.1, 0.5]; the number of leaf nodes in the decision tree num_leaves = [2, 4, 6, 8, 10].

2.3. Evaluations

This part introduces several indicators, which are used to measure the predictive power of the model. The prediction results contains four types, TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), and the following indexes are made up of them. The first index is precision, the proportion of samples correctly predicted as positive examples by the model in all samples predicted as positive examples. The second index is recall, the proportion of the samples correctly predicted as positive examples by the model to the total samples of real positive examples. The third index is F1 value, the harmonic average of accuracy rate and recall rate, which considers accuracy rate and recall rate comprehensively. The fourth index is accuracy, the proportion of samples correctly predicted by the model to the total sample size. The first three indexes can calculate average in two ways. One is micro average, which adds up the number of true positives, false positives and false negatives for all categories and calculates accuracy, recall and F1 scores. This means that all categories are treated as equally important, so each sample is weighted equally in this case. The other is weighted average, which is a weighted average of accuracy, recall, and F1 scores for each category, where the weight is the

DOI: 10.54254/2755-2721/16/20230881

number of samples in each category. This means that the importance of each category depends on its sample size, so in this case each sample is weighted differently. The expressions are given as follows:

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 = \frac{2PrecisionRecall}{Precision+Recall} \tag{5}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

3. Results & discussion

3.1. Feature engineering

The correlation strength between dependent variable and each predictor variable are various. A predictor larger than 0.1 in absolute value is recognized to have a strong correlation with the dependent variable. This factor is a relative important indicator of heart disease. Fig. 1 shows 10 features, which are 'HighBP', 'HighChol', 'Smoker', 'Stroke', 'Diabetes', 'GenHlth', 'PhysHlth', 'DiffWalk', 'Age', 'Income', are core factors, so these factors are used in the three machine learning models to predict heart disease.

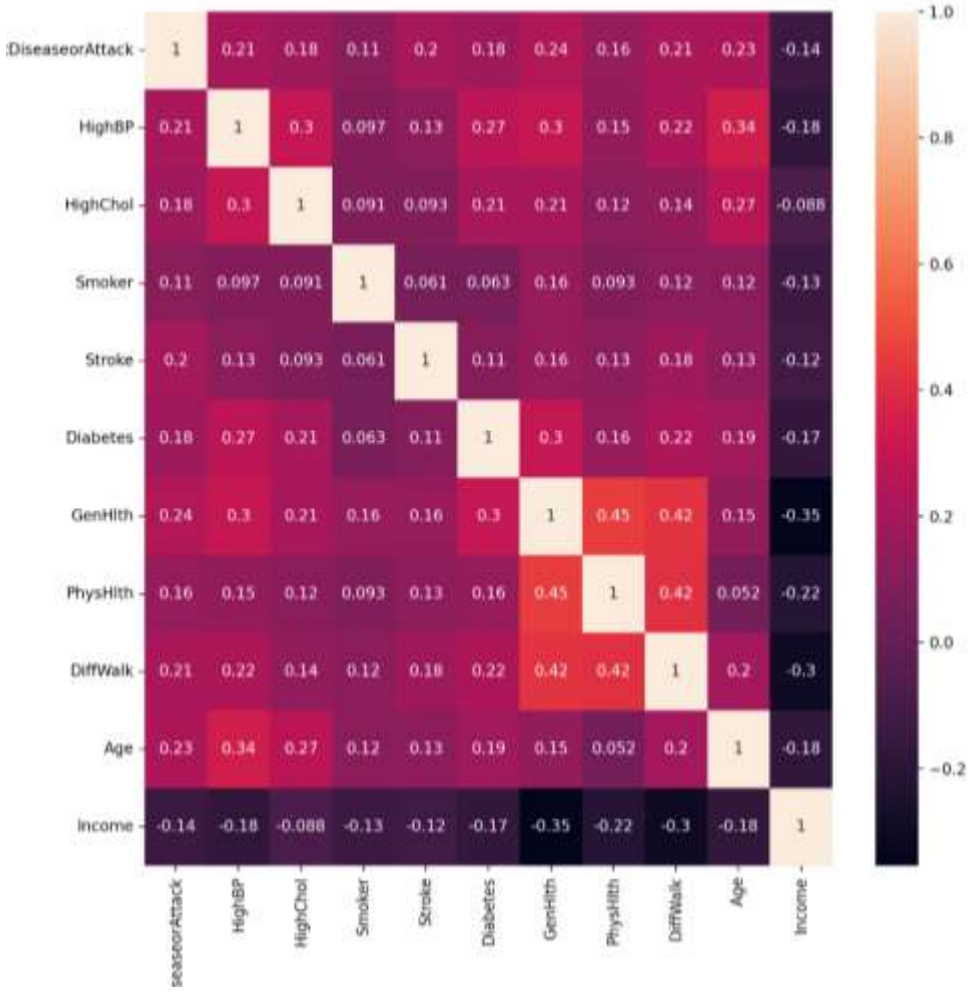

Figure 1. Correlation heat map.

DOI: 10.54254/2755-2721/16/20230881

3.2. Evaluation of models

The prediction results of the three models are shown in Table. 2, Table. 3 and Table. 4. The accuracy of the predictions is all 0.91, suggesting relatively effective. Since the numbers of samples of the two categories are of great disparity, weighted average of the indexes is more reasonable and reliable. The result shows the weighted average precision, recall and f1-score are almost the same in three models, and these three indexes are similar in the prediction of negative samples. However, the prediction of positive samples has significant in the three models. Logistic regression has the highest f1-score and the highest recall, but its precision is the lowest. Random forest has the highest precision, but its recall and f1-score are the lowest. The three indexes of lightGBM are mediate of the three. To sum up, random forest results in the best performance in both precision and recall of the three, while the logistic regression model performances the worst.

Table 2. Logistic result.

	precision	recall	f1-score	support
0.0	0.92	0.99	0.95	46042
1.0	0.53	0.11	0.18	4694
macro avg	0.73	0.55	0.56	50736
weight avg	0.88	0.91	0.88	50736
accuracy		0.91		50736

Table 3. Random forest.

	precision	recall	f1-score	support
0.0	0.91	1.00	0.95	46042
1.0	0.60	0.06	0.11	4694
macro avg	0.75	0.53	0.53	50736
weight avg	0.88	0.91	0.87	50736
accuracy		0.91		50736

Table 4. Lightgbm.

	precision	recall	f1-score	support
0.0	0.91	0.99	0.95	46042
1.0	0.57	0.09	0.16	4694
macro avg	0.74	0.54	0.56	50736
weight avg	0.88	0.91	0.88	50736
accuracy		0.91		50736

3.3. Explanations & suggestions

First, the correlation coefficients in the heat map reflects the degree of importance of the indicators to heart disease. According to the results, four categories of factors are core indicators of heart disease. The first type is about health condition, consisting of general health, high blood pressure, days with poor physical health, high cholesterol, stroke, serious difficulty of walking or climbing stairs and diabetes. It is reasonable that poor health condition increases the possibility of heart attack, since these features have a significant positive correlation coefficient with heart disease or heart attack. The second category is about living habits, including smoking and income. A significant positive correlation coefficient means smoking leads to heart disease or heart attack to some extent, while a negative correlation coefficient between income and dependent variable suggests that high income, accompanied with high pressure and heavy pressure of work, increase the rate of heart diseases. The third category has only one feature, age. A correlation coefficient of 0.23 means the possibility of suffering from heart disease increases greatly as people grow older.

Although all plus-minus signs of the correlation coefficients are consistent with common sense, the absolute value of them cannot exactly reflects the importance of the indicators to heart disease, because

DOI: 10.54254/2755-2721/16/20230881

there is multicollinearity among the features. For instance, several people have poorer health conditions as they are aging, meanwhile people are likely to have higher income as they are aging for their increasing work experience. Second, the comparison among Logistic Regression, Random Forest and Light GBM shows that Random Forest fits most in the prediction. In this problem, the correlations between response variables and each response variable are complex, rather than simple linear relations, and there is multicollinearity among response variables. Since logistic regression is based on linear regression, its prediction is not accurate enough in non-linear relationships and multicollinearity. Meanwhile, LightGBM model is over-fitting if response variables are highly related, which decreases the accuracy in prediction. By contrast, Random Forest can reduce variance and improve generalization ability by averaging or voting multiple decision trees. It can automatically process highly correlated features and select the most important features for splitting. Therefore, random forest model perform the best in the prediction of this data set.

4. Conclusion

To sum up, heart disease is a significant health concern worldwide. Heart disease indicators are important tools for assessing an individual's risk for heart disease. These indicators include blood pressure, cholesterol levels, smoking, stroke, diabetes, general health conditions, physical health conditions, difficulties in walking, age, and income. Early detection and prevention play a crucial role in avoiding these complications. Machine learning algorithms are used to predict the occurrence of heart disease based on various risk factors. These algorithms have shown promising results in accurately predicting heart disease, leading to timely interventions and improved health outcomes. However, this paper only focusses on individual indicators and have not taken environmental factors and family history into consideration. Future researches can have further researches by using more comprehensive data sets. Overall, the use of machine learning techniques in heart disease prediction can contribute significantly to reducing the burden of heart disease and improving public health.

References

[1] Quasinowski B and Liu T 2020 Int. J. Environ. Res. Public Health vol 17 p 3150.
[2] Schnall P L, Dobson M and Landsbergis P 2016 Int J Health Serv. vol 46(4) pp 656-692.
[3] Cosselman K, Navas A A and Kaufman J 2015 Nat Rev Cardiol vol 12 pp 627–642
[4] Romero C J L, Ankeny J, Fernández M A, Kales S N and Smith D L 2022 Int. J. Mol. Sci. vol 23 p 13540.
[5] Shi Q, Wang R, Zhang H, Shan Y, Ye M, and Jia B 2021 PLoS One vol 16(8) p e0254590.
[6] Henein M Y, Vancheri S, Longo G and Vancheri F 2022 Int. J. Mol. Sci. vol 23 p 12906.
[7] Tramontano L, Punzo B, Clemente A, Seitun S, Saba L, Bossone E, Maffei E, Cavaliere C, and Cademartiri F 2022 J Clin Med. Vol 11(19) p 5842.
[8] Trends in cardiovascular disease risk factors in the United States: NHANES 1999-2014.
[9] Hosmer Jr D W, Lemeshow S and Sturdivant R X 2013 Applied logistic regression vol 398 John Wiley & Sons.
[10] Agresti A 2013 Categorical data analysis vol 482 John Wiley & Sons.
[11] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W and Ye Q 2017 Advances in Neural Information Processing Systems vol 2017 pp 3149-3157.