Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230710

Predictive modelling based on statistical modeling of logistic regression for heart disease

Jincheng Wang

Shandong Experimental High School, Jinan, 250000, China

clee84177@student.napavalley.edu

Abstract. The heart is the core driving force for the continuation of human life, and the disease of this organ is bound to be fatal. There are two main types of heart disease. Congenital diseases are caused by developmental problems in unborn children. These problems, mainly in the heart, can damage various parts of the heart. Acquired sexually transmitted diseases are diseases caused by environmental factors and their own growth and development after birth. The purpose of the project model is to predict heart disease and analyse the main types of heart disease in the population. In the whole research process, the most important thing is the establishment of the model. The algorithm principle of this model is logistic regression. Logistic regression is used to make predictions and probability calculations on the data. Through such algorithms, modelling techniques can be used to predict the impact of pathogenic factors on the probability of heart disease. In addition, prevention of heart disease can be improved with accurate and convenient model predictions that can be tailored to the population that fits the predictions. This method can improve the technical level and treatment level of the hospital, and can also reduce the harm caused by heart disease.

Keywords: Heart Disease, Logistic Regression, Modelling Techniques.

1. Introduction

The heart is the most important organ in the human body, equivalent to the engine in a car. So heart problems are bound to be fatal. There are two main types of heart disease, congenital heart disease and acquired heart disease. The former can carry heart disease at birth. This is because of the abnormal development of babies before they are born. The latter is when the baby is born because of external influences that cause damage to the heart and other parts. Examples of high incidence of heart disease are: coronary atherosclerotic heart disease, hypertensive heart disease, endocrine heart disease, blood disease, nutritional metabolic heart disease, etc [1].

An electrocardiogram (ECG) is a non-invasive way to assess heart health and shows the activity and function of the heart. Scientists try to use electrocardiograms to predict heart disease. In addition, there are other ways to predict heart disease, such as blood pressure tests, Magnetic Resonance Imaging (MRIs), and heart ultrasounds [2]. Thallium stress test is an effective way to detect heart disease. Thallium is a radioactive element. Professionals can use this element for nuclear imaging tests. During the test, an isotope of thallium passes through the blood into the heart. Once radiation enters the heart, gamma cameras can detect the radiation and reveal any problems that arise in patients' heart muscle [3-4]. However, these methods require a large amount of capital investment, high cost, and large economic

DOI: 10.54254/2755-2721/41/20230710

losses. None of these methods are as good as using data models to make predictions. In 2013, the American College of Cardiology/Heart Association (ACC/AHA) was the first in the world to publish the Pooled Cohort equation (PCE) model for 10-year risk prediction of Arteriosclerotic Cardiovascular Disease (ASCVD) [5]. However, ACC/AHA also notes that the PCE model is derived from white and black US cohort data. The predictive model may not be applicable to other populations [6]. The par model in China is compared with the PCE risk prediction model. The results showed that in an externally validated cohort of 14,000 people, the number of observed cases of ASCVD overestimated the actual incidence by only 17%. The estimated 10-year incidence of ASCVD was overestimated by 54 percent by a PCE model developed based on white Americans and 54 percent by a PCE model developed based on black Americans. Overall, the China-PAR model was more accurate in predicting 10-year ASCVD risk in the Chinese population. The Chinese Population ASCVD Risk Prediction Study (China-PAR) integrated four recent Chinese prospective cohort follow-up data with a total sample size of 127,000 individuals and a maximum follow-up period of more than 23 years. The China-PAR model had good internal consistency and independent validation in two external cohorts with a total sample size of nearly 100,000 people, predicting 10 - and 5-year ASCVD risk [7]. This model used two cohorts, InterAsia and China Multicentre Collaborative (MUCA), combined with Cox proportional risk regression to establish a sex-specific 10-year risk prediction model for ASCVD [8-9].

The purpose of this project model is to predict heart disease and analyse the main types of heart disease in the population. In the whole research process, the most important thing is the establishment of the model. The algorithm principle of the model is logistic regression. Logistic regression is a statistical model. Logistic regression is used to make predictions and probability calculations on data. This algorithm can find predictive modelling techniques for the relationship between a dependent variable (the probability of having heart disease) and one or more independent variables (different causative factors). This predictive modelling technique is useful for disease analysis, because it can combine various risk factors and infer the impact of each factor on the incidence of heart disease. Moreover, prevention of heart disease can be improved through accurate and convenient model predictions, which can be tailored to the populations that match the predictions. This method can improve the level of technology and treatment in hospitals, and can also reduce the harm caused by heart disease.

2. Methodology

2.1. Dataset description and preprocessing

The dataset is publicly available on the Kaggle website [10]. Residents of Framingham, Massachusetts, conducted an analysis of cardiovascular disease. The goal of this model is to predict the effects of different factors on the incidence of heart disease. To predict the incidence of heart disease in the population. Data sets provide information about patients. It includes more than 4200 records and 15 attributes. Each attribute is a potential risk factor. There are three main types of data: demographic, behavioural, and medical. Population: Age and sex. Behaviour: Whether or not people smoke, number of cigarettes smoked per day. Medical: Historical data: whether people took blood pressure medication, whether they had a stroke, whether individuals had high blood pressure, whether they had diabetes. Current data: Total cholesterol level, systolic blood pressure, diastolic blood pressure, Body mass index, heart rate, blood sugar level.

Missing values are common in everyday data. This may be due to data corruption. However, the processing of missing values is essential in the preprocessing of data sets. Missing values are handled by deleting rows or columns with empty values. Some empty lines can be removed for processing.

2.2. Proposed approach

The main goal of this study was to develop a clear and understandable prediction model to better understand and predict whether different populations are at risk for heart disease. Follow the process in Figure 1 to initialize the logistic regression model using the data set from Kaggle. Secondly, a large

DOI: 10.54254/2755-2721/41/20230710

amount of useful information is lost in many empty values, which directly reduces the data quality, and the model effect cannot meet the target due to low quality data. The non-obvious deterministic parts of the data are significant and difficult to control. Therefore, the mining process will produce certain confusion and even unreliable data. Using the missing value filling technique, much of the real data can be reconstructed. Make use of these valuable data to improve the feasibility of the model. Deleting rows or columns (where empty refers) is an efficient way to deal with missing values. Once the missing values are processed, the accuracy of the model can be improved, thus optimizing the entire model. After model training, cross-validation, accuracy, and area under subject operating characteristic curve (AUC) were used. In this way, the stability and prediction ability of the model can be evaluated.

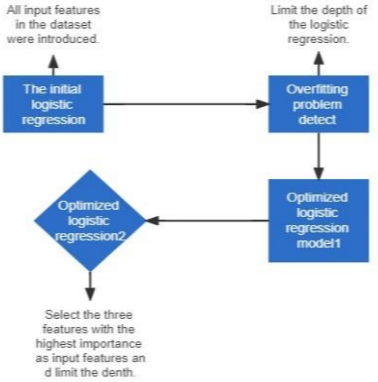

Figure 1. Flow Chart Progress.

2.2.1. Logistic regression. Classification problems are a common task in machine learning. The goal is to assign data points to different categories based on input characteristics. The model needs to train a classifier. The classifier can make predictions based on the characteristics of the input data. Logistic regression is a common classification algorithm, especially suitable for binary classification problems. The core idea of logistic regression is to map the output of linear regression to the probability space through the logarithmic probability function, thus achieving classification. The mathematical expression formula of logistic regression model is as follows:

$$\mathbf{P(Y-1|X)} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-(\boldsymbol{\omega}\cdot\boldsymbol{X}+\boldsymbol{b})}} \tag{1}$$

where P(Y=1|X) represents the probability that a data point belongs to a positive class given the input feature X. w is the weight vector. b is the offset term. e is the base of the natural logarithm. Combine the mathematical formulas of logistic regression with datasets related to heart disease:

$$P = \frac{e^{\beta_0+\beta_1 X_1}}{1 + e^{\beta_0+\beta_1 X_1}} \tag{2}$$

$$\log it(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 * Sexmale + \beta_2 * age + \beta_3 * cigsPerDay + \beta_4 * totChol + \beta_5 * sysBP + \beta_6 * glucose \tag{3}$$

This is the basic formula for logistic regression, where e is the natural logarithm. After converting the exponential form of e to the logarithm form, this is the logarithm in the following formula. Then all kinds of data in the data set are calculated for different β. p represents the probability that the result of having a heart attack is 1, and the 1-P band table shows the probability of having a heart attack is 0. beta stands for regression coefficient. If beta is negative and significant, it means that this type of data has a

DOI: 10.54254/2755-2721/41/20230710

negative impact on whether people have heart disease. If beta is positive and significant, it means that this type of data has a positive impact on whether a population has heart disease. If it is not significant, it means that this type of data has no effect on whether people have heart disease.

2.2.2. Feature selection. Feature selection is designed to remove irrelevant and redundant features. There are many benefits to feature selection that retain some of the relevant features (from the original features), such as better model performance, lower computational costs, and so on.

The goals of feature selection techniques include: Simplify models to make it easier for researchers/users to interpret them; Shorter training time; Avoid the curse of dimension; Enhance generalization by reducing overfitting.

2.2.3. Evaluation metrics and visualization tools. Evaluation of logistic regression models involves a variety of measurement and visualization techniques, including accuracy, cross-validation, AUC, and more. Such as confusion matrix, ROC curve, mathematical statistics evaluation model. Accuracy represents the correctness of the model fit or prediction. It refers to the percentage of all samples correctly identified. The calculation formula is:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

The first letter has two types: T and P. T means that the model makes a decision that is (True) correct (the decision agrees with the fact). F means that the model made a decision that was (False) wrong (contrary to the facts). The second letter is divided into P and N. P stands for making a Positive decision (predicting that the event will happen). N means that the representative makes a Negative decision (judging that the event will not happen). Confusion matrix. Confusion matrix has rows and columns. Columns represent the category to which the prediction results belong. Rows represent the real category to which the data belongs. In the image accuracy problem, it is mainly used to compare the predicted category with the actual measurement. The accuracy of classification results can be expressed in the form of confusion matrix. The Receiver Operating Characteristic (ROC) is used in psychology, medicine, machine learning, and other fields to evaluate the performance of predictive models. The prediction of how many positives there are in the real sample is the ordinate TPR. TPR is a true positive rate. The number of negatives predicted is the horizontal coordinate FPR, where FPR is the false positive rate (FPR). Sensitivity is best when the ROC curve is in the upper left corner.

2.2.4. Implementation details. This study uses Python 3.10 and Kaggle database to implement logistic regression model. Data visualization is done using the Matplotlib library. The study was conducted on a macOS device with an Apple M1 SoC. Logistic regression has the following Settings: a Gini index for the impurity measure, a minimum of 1 sample per leaf node, a minimum total sample weight of 0.0, and all features of the split are considered at each node. These Settings allow the initial logistic regression model to accurately capture patterns and relationships in the data for further optimization and improvement.

3. Result and discussion

After analysis, optimization and evaluation, the decision tree model is understood to ensure effective classification by identifying important factors. In data analysis, visualization technology is adopted to graph the data in the data set, like in Figure 2.

DOI: 10.54254/2755-2721/41/20230710

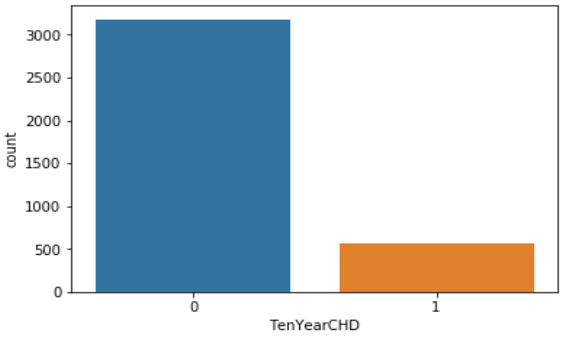

Figure 2. Visualize the data.

In addition, the importance of features was evaluated. Moreover, the feature selection method of logistic regression is used to calculate the correlation scores of each feature. The model predicts that men are more likely to develop heart disease than women. As you get older, your risk of heart disease increases. For people who smoke, the more they smoke, the more likely they are to have heart disease. However, the total cholesterol level had no significant effect on the incidence of coronary heart disease. After that, the final predictive model is trained. This simplifies the logistic regression model and then reduces the risk of overfitting. Finally, performance evaluation. The predictive ability, stability and discrimination of the trained model were evaluated. With the increase of data, the overall model can be further improved.

The accuracy of the model is 0.88. Classification models are best kept away from random classification. The model, which has an area of 0.5, performs worse than random classification. The closer the AUC is to 1, the better. The high value of AUC confirms the predictive ability, spirituality and discriminative ability of the model.

Classification accuracy of area quantization model under ROC curve; The higher the region, the greater the difference between true and false positives, and the better the model's ability to classify members of the training dataset. The optimum position for ROC curve is towards the top left corner where the specificity and sensitivity are at optimum levels. So in Figure 3, the model is more specific than sensitive.

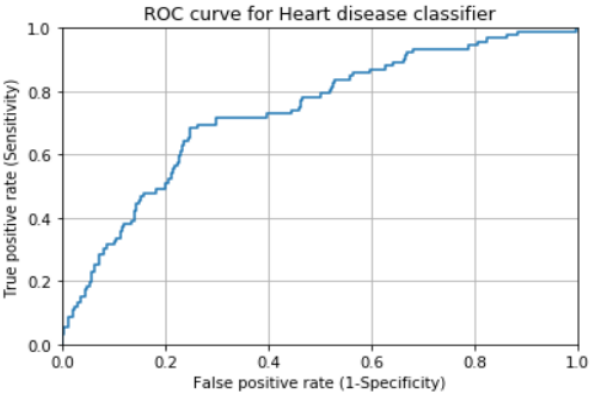

Figure 3. The AUC line of the junior classifier.

Type II errors have a big impact on models that predict heart disease. When a disease does exist, ignoring its false negatives is more unacceptable than its false positives. According to the test, lowering the threshold can effectively improve the sensitivity of the model. In Table 1, it turns out that the lower the threshold, the better.

DOI: 10.54254/2755-2721/41/20230710

Table 1. Difference in Threshold.

Threshold Consequence	Confusion Matrix	Correct Prediction	Type II Errors	Sensitivity	Specificity
0.4	652 7 86 6	658	86	0.0652173913043	0.98937784522
0.3	617 42 70 22	639	70	0.239130434783	0.93626707132
0.2	519 140 43 49	568	43	0.532608695652	0.787556904401
0.1	240 419 11 81	321	11	0.880434782609	0.364188163885

This fitting model shows that in Table 2, the odds of being diagnosed with heart disease in men (sex_male =1) compared to women (sex_male =1) are p(0.5826) = 1.788659, holding all other factors constant. In conclusion, men were estimated to be 78 percent more likely to have heart disease than women. The age coefficient model proves that p(0.0655) = 1.067574. The model predicted that each additional cigarette smoked would increase the risk of disease by 0.02. Similarly, the probability of being diagnosed with heart disease increases by 0.07 if a person is one year older. Total glucose and cholesterol levels did not change significantly.

Table 2. Prediction Consequence.

	CI 94%(2.6%)	CI 94%(7.6%)	Odds Ratio	pvalue
const	0.000042	0.000273	0.000108	Zero
totchol	0.000157+1	0.004395+1	0.002274+1	0.034
age	0.054482+1	0.080964+1	0.067646+1	Zero
cigsperday	0.011732+1	0.028127+1	0.019895+1	Zero
glucose	0.004347+1	0.010894+1	0.007612+1	Zero
sysbp	0.013291+1	0.021786+1	0.017527+1	Zero
male	0.455241+1	1.198533+1	0.788684+1	Zero

After the elimination process, P-values are useful in predicting heart disease. This is because all P-values are below 0.05. The prediction accuracy of the model is 88%. This model is more specific than sensitive. Model predictions show that women are less likely to develop heart disease than men. Not only that, heart disease is more likely to occur with age. Also, the model predicted that the risk of heart disease was also proportional to the number of cigarettes smoked. Total cholesterol level had no significant effect on SNR (Signal Noise Ratio). The number of heart attacks caused by glucose is negligible. The whole model can be improved with more data.

4. Conclusion

In this study, logistic regression is introduced for data prediction and probability calculation. This algorithm can find the predictive modeling technique that relates the dependent variable (probability of having heart disease) to one or more independent variables (different risk factors). This predictive modeling technique is useful for disease analysis as it can combine various risk factors and infer the influence of each factor on the incidence of heart disease. It is possible to predict whether a heart attack will occur if researchers have enough analysis of the heart disease data. Good predictive models can inform at-risk patients ahead of time. Then you can reduce the incidence of heart disease. Broadly speaking, this prediction can improve the health of society as a whole. This research has two important medical application. First, customized medical treatment plans: Based on existing clinical medical data and heart disease prediction data, personalized treatment plans can be optimized to provide more

DOI: 10.54254/2755-2721/41/20230710

accurate medical services for patients. Second, development of public health strategies: Analyzing large-scale data of heart disease patients can guide public health institutions in developing relevant strategies, carrying out corresponding prevention and public education activities, thereby reducing the incidence of heart disease.

References

[1] Xia Q 1985 Summary of 1601 cases of Congenital and acquired Heart disease treated by Surgery Journal of Harbin Medical University 04

[2] Li M Wen J and Han Y 1987 The value of Body surface Electrocardiogram in the diagnosis of Coronary Heart Disease Affiliated Hospital of Xi ‘an Medical University 76(6): pp 1290-1297

[3] Jabbar M A Samreen S 2016 Heart disease prediction system based on hidden naïve bayes classifier international conference on circuits controls communications and computing (I4C) IEEE pp 1-5

[4] Michael Prisant L Thomas W Dohlen Jan L Houghton Albert Can A and Martin J 1992 A Negative Thallium (± Dipyridamole) Stress Test Excludes Significant Obstructive Epicardial Coronary Artery Disease in Hypertensive Patients American Journal of Hypertension 5: pp 71-75

[5] Zhang X Pawlikowski M Olivo-Marston S et al 2021 Ten-year cardiovascular risk among cancer survivors: the National Health and Nutrition Examination Survey PloS one 2021 16(3): p e0247919

[6] Medina-Inojosa J Somers V K Hayes S et al 2020 Evaluating the sensitivity of the ACC/AHA pooled cohort risk calculator to predict atherosclerotic cardiovascular events within 10 years: how many events are we failing to predict? European Heart Journal 41(Supplement_2): p 2924

[7] Zhu J J Wang W 2020 Application of pooled cohort risk equations and China-PAR model in ASCVD risk prediction among people with physical examination Chin J Evid Based Cardiovasc Med 12(2): pp 131-134

[8] Brindle P Beswick A Fahey T et al 2006 Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review Heart 92(12): pp 1752-1759

[9] Nong Y B Lin Q Duan W H et al 2004 Constructing a Cox proportional hazard regression model of prognosis factors of acute myocardial infarction by retrospective cohort study Chinese Journal of Integrated Traditional and Western Medicine 24(9): pp 781-784

[10] Alex Teboul 2022 Heart Disease Health Indicators Dataset kaggle https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset