International Journal of Research in Engineering, Science and Management
Volume 4, Issue 7, July 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

A Modeling Approach to Predict Risk of Coronary Heart Disease using Logit and CAP Curves

Rohan Ghatpande[1*], Piyush Kendurkar[2]
[1]*Student, Institute of Management Studies, Devi Ahilya Vishwavidyalaya, Indore, India*
[2]*Assistant Professor, Institute of Management Studies, Devi Ahilya Vishwavidyalaya, Indore, India*

***Abstract*: On a global spectrum of healthcare anomalies, coronary heart disease enjoys its fair share of incidence levels and disease burden rates. The inevitability of this condition is apparent due to changing lifestyles of human beings. However, the unpredictable nature of this condition makes it a menace for management and intervention systems around the world. The study aims at generating a binary classification model sturdy enough to recognize the potent risk factors contributing towards this risk. This modeling approach could sizably reduce the area of focus to allow accurate causation insights and quantify relationships between the regressors and the outcome variables. By doing so, action plans could be created in order to deal with these red flags so obtained, ultimately aiming towards better health outcomes for people.**

***Keywords*: Coronary heart disease, Modeling, Odds ratio, Regression, Risk factors.**

1. Introduction

Coronary heart disease is one of the most common heart ailments in the world. It is also known as coronary artery disease or ischemic heart disease. It is a condition generally characterized by an insufficient supply of oxygen-rich blood to the muscles of the heart. This kind of activity is propelled by narrowing of or blockages in a coronary artery by fatty plaques. In cases of severe dearth in oxygen levels, myocardial infarction (heart attack) may be evident.

A plethora of risk factors have been associated with the condition of CHD. These include high blood pressure, elevated blood cholesterol levels, smoking, obesity, diabetes, unhealthy diet, and family history of early CHD. Individuals with predisposed hereditary conditions such as familial hypercholesterolemia (a condition in which the body's tissues are incapable of removing cholesterol from the bloodstream) are also at increased risk of contracting this condition.

Countless studies have been made to understand the epidemiologic pattern of CHD and its congruence with modern living habits of human beings. The approach of this study is inclined towards proving the universal applicability of certain risk-factors and common demographic profiles which could aid in conducting worldwide analytical regimes for the purpose of estimating risk of CHD within individuals.

Data science and machine learning have proved to be vital cogs and have articulated the wheel of healthcare by shaping up insightful ideas and patterns thereby assisting in clinical decision support and the development of clinical care guidelines. Such algorithms go the distance in intuiting information and healthcare constructs which can help alter current healthcare practices and eliminating the threadbare ideas of implementing healthcare activities according to common notion instead of smart designs. The central idea of decisive targeting instead of adopting holistic approaches can help in increasing the overall quality of such practices and moreover cater to the needs of the target groups in an efficient manner.

Scaling these analytical approaches to a global perspective can help understand the overall burden of CHD. Differences in lifestyles and day-to-day habits of people across the world have cross-disciplinary applications across the fields of medicine, psychology, sociology, and progressive dynamics of human civilizations.

Treatment plans and care practices could be made specific in nature according to the insights gained and methods with respect to the prevention of CHD could be regulated. Recommendations, hence, to deal with the incidental risks of heart disease could be channelized much efficiently and incorporated into public policy and healthcare intervention campaigns of the nation.

2. Literature Review

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014) supplied an elaborate review describing the stages leading up to the foundation of the famous Framingham Heart Study. Specific contributions from the study, which was started in 1948, were put forward along with a historical perspective of the case details paved the platform for further research and a roadmap for analytical integration studies. Shah, D., Patel, S. & Bharti, S.K. (2020) provided with the integration of data mining principles in the field of healthcare, specifically heart disease, utilising learning algorithms like Naïve Bayes, decision tree,

*Corresponding author: rrg1996@gmail.com

etc. and also discussed various attributes culminating towards heart disease. The study envisaged the probability of developing this heart condition. Gupta R, Mohan I, Narula J. (2016) established the epidemiological trend of CHD in India, thereby facilitating cross-sectional study and comparison across different demographic profiles. Epidemiologic studies have shown that there are at present over 30 million cases of CHD in the country. A study by Gajalakshmi et al during 1995–1997 showed that CVD deaths are the highest (38.6%) in urban Chennai. Similar figures are published by Joshi et al from Andhra Pradesh. These studies have insightful significance with respect to disease burden and prevention and intervention measures could be based on the projected figures of DALYs lost in India.

3. Methodology

A. Research Design

The research is primarily focused on designating the dominant characteristics that exhibit association with the said dependent variable, i.e., 10-year CHD risk among the target population under consideration. The purpose of creating a model to validate the aforementioned efforts will be to create triggers which can help understand the overall scenario and assist the introduction of appropriate methods to deal with the red flags that will be posed for interpretation at the end of the study.

B. Data Collection & Variables

The data was collected through simple random sampling of the population under study. The research approach will aim to contribute useful insights as value addition and leap over the probable specificity of the data to cover a wider ground. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor (excluding education level). There are demographic, behavioural and medical risk factors.

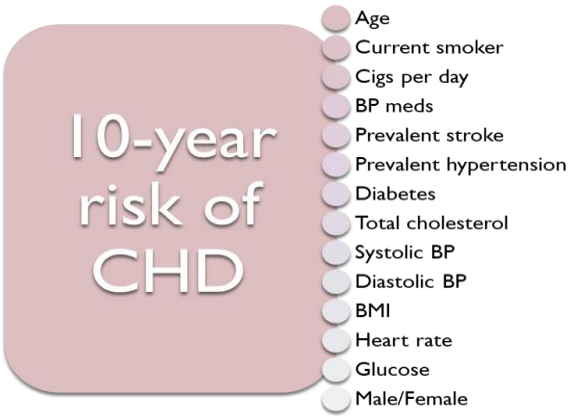

Fig. 1. List of variables

C. Data Preparation

The dataset was explored for possible construction and transformation prerequisites. To be conducive to the purpose of the research study, the original dataset was split into train and test datasets using a 70:30 random split. We trained our model on the train dataset, analysed the model performance and its validation with the test dataset using the Cumulative Accuracy Profile (CAP) Curve. The missing values for continuous variables were treated with mean computations of 'non-NA' values and those from one of the nominal variables were subjected to kNN Imputation. The data sets were split prior to any sort of transformations to avoid data leakage and overfitting.

D. Data Analysis

The method for analysis so chosen fulfils the overall motive of predicting the extent of association between the dependent and independent variables and attempts to ascribe relationships between them, if any. The intent is to create a model that fits the population sample with maximum accuracy and at the same time spare valuable insights on the way. Logistic regression was used as the tool under regression analysis to achieve the above-mentioned. The software package used for this exercise is gretl (Gnu Regression, Econometrics and Time-series Library). This is a statistical software package widely used in the field of data science.

4. Findings

After checking for consistency with assumptions of binary logistic regression (dichotomous dependent variable, standardization and multicollinearity), the train set underwent backward elimination process and after 10 iterations, the following model was obtained.

Model 10: Logit, using observations 1-2942
Dependent variable: TenYearCHD

	Coefficient	*Std. Error*	*z*	*p-value*	
const	−8.39733	0.513019	−16.37	<0.0001	***
age	0.0567469	0.00692099	8.199	<0.0001	***
cigsPerDay	0.0237059	0.00444054	5.339	<0.0001	***
prevalentStroke	1.18319	0.514207	2.301	0.0214	**
sysBP	0.0173029	0.00236610	7.313	<0.0001	***
glucose	0.00591995	0.00183353	3.229	0.0012	***
Male	0.530291	0.114988	4.612	<0.0001	***
totChol	0.00203865	0.00124043	1.643	0.1003	

Mean dependent var	0.161455	S.D. dependent var	0.368012
McFadden R-squared	0.104465	Adjusted R-squared	0.098314
Log-likelihood	−1164.717	Akaike criterion	2345.434
Schwarz criterion	2393.329	Hannan-Quinn	2362.679

Number of cases 'correctly predicted' = 2480 (84.3%)
f(beta'x) at mean of independent vars = 0.368
Likelihood ratio test: Chi-square(7) = 271.732 [0.0000]

Fig. 2. Final binary classification model

The confusion matrix for the final model achieved has been provided below:

		PREDICTED	
		1	0
ACTUAL	1	36	439
	0	23	2444

Fig. 3. Confusion Matrix

Therefore, model accuracy is computed as,

$$\frac{36+2444}{36+2444+23+439} = 0.843 \text{ or } 84.3\%$$

For the model curve, gretl can reckon the prediction values or forecasts for all 2942 observations of 10-year CHD risk. These predictions, against the actual values, could help assess the model performance and thus help us rate the quality of the model according to the index as a reference mechanism to quantify this qualitative aspect. The curve has been attached below and the projection line from the 50% mark indicates the percentage. The Cumulative Accuracy Profile Percentage comes out to be approximately 76%. This signifies that the aforementioned is a GOOD MODEL and the insights will have an acceptable level of import.

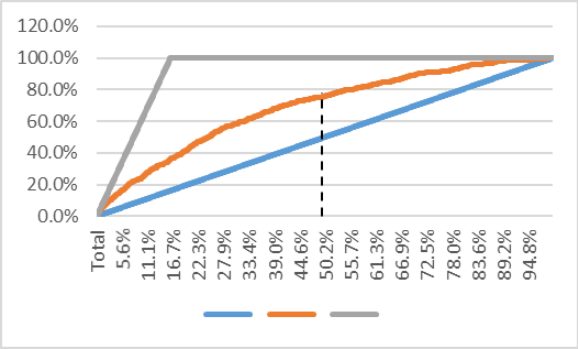

Fig. 4. CAP curve

The presence of a separate dataset to test the newly trained data holds great significance. It helps to assess overfitting and to understand if the model was biased to certain characteristics of the train dataset due to which it does not present consistent results with the test dataset. We had separated the main population randomly with a 70:30 split, and now the newly trained model will be compelled to predict values of the 10-year CHD risk incidences and after which we will adding the actual values to see the 'Goodness of Fit' of the model. Sample size of the test data post-split = 1267. In this case, the model was not subjected to any information from the test data. It was completely based on the training set. Still, the model manages to score 78% as compared to the initial value of 76%. This means that the model predicted 2% more patients with ten-year CHD risk. At the same reference line of 50%, the model was able to pinpoint 78% of such patients from the test data in a minor contrast to the 76% from the train data. To sum it up, the model seems to be performing well.

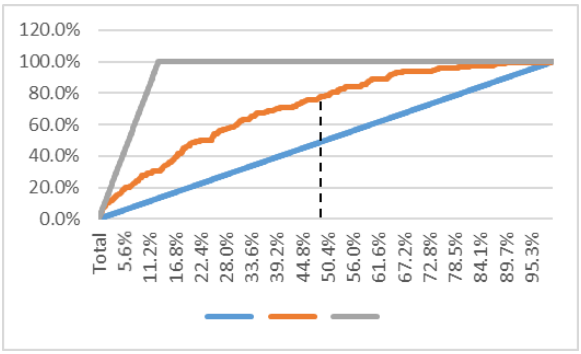

Fig. 5. CAP curve for test data

5. Interpretation

Odds Ratio (OR) is a measure of association between exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

- OR>1 indicates increased occurrence of an event
- OR<1 indicates decreased occurrence of an event (protective exposure)

One unit rise in age signifies that the odds of being associated with a potential CHD risk increases by approximately 6%. There is a 2% increase in chances of contracting CHD with every extra cigarette smoked per day. With every unit rise in the number of patients displaying a history of stroke, the risk of contracting CHD increases by a humongous factor of 3.2648. However, as the confidence interval for the same is too wide, this indicates a small sample size resulting in an inconclusive insight from only about 20 patients with a history of stroke. Glucose and total cholesterol variables display negligible impact on the odds of facing the risk of coronary heart disease. TotChol variable was only retained in the model to maintain model performance even though its p-value>0.05. Per unit increase in systolic blood pressure does not have a major influence in the probability of occurrence of CHD, i.e., only around 1.75% times the odds. It was found that whenever the records of a male subject were taken into consideration after a female subject, the odds of getting diagnosed with CHD increases by a factor of 1.6994, i.e., almost 70%.

6. Conclusion

Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic blood pressure also show increasing odds of having coronary heart disease. The attribute 'prevalent Stroke' was

Table 1
Heat map displaying the coefficients of variables under the model, p-values, odds-ratios and confidence intervals

	coefficient	p-value	Odds-ratio	95.0% conf. interval
Const.	-8.397328694	3.21E-60		
Age	0.056746937	2.42E-16	1.0584	[1.044, 1.073]
Cigs Per Day	0.023705896	9.37E-08	1.024	[1.015, 1.033]
Prev Stroke	1.183191316	0.02139146	3.2648	[1.192, 8.944]
Sys BP	0.017302881	2.62E-13	1.0175	[1.013, 1.022]
Glucose	0.005919955	0.001243448	1.0059	[1.002, 1.010]
Male	0.530291432	3.99E-06	1.6994	[1.357, 2.129]
Tot Chol	0.002038645	0.10027954	1.002	[1.000, 1.004]

deemed inconclusive even though it had the biggest impact among all the other regressors, due to the fact that the confidence interval for that particular attribute was too wide, indicating the insufficiency of the data available to draw any insights and which would have resulted in inconsistent outputs. It can, however, be called forth for future research. Along with total cholesterol, glucose too causes a very negligible change in the odds. The model predicted with 0.843 accuracy. The model was able to perform similarly on both the datasets which were split randomly and belonged to the same period of data creation. Also, the confidence interval width from the 'prevalent Stroke' attribute may not show consistency with similar background data from a different period of time, hence suggesting that the model is more specific than sensitive. Overall, the model could be improved with more data.

References

[1] Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)*, 383(9921), 999–1008.

[2] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* 1, 345 (2020).

[3] Gupta R, Mohan I, Narula J. Trends in Coronary Heart Disease Epidemiology in India. Ann Glob Health. 2016 Mar-Apr.82(2):307-15.

[4] Gajalakshmi V, Peto R., Kanaka S, Verbal autopsy of 48000 adult deaths attributable to medical causes in Chennai, India. *BMC Public Health*. 2002; 2:7.

[5] Joshi R., Cardona M, Iyengar S. Chronic diseases now a leading cause of death in rural India: mortality data from the Andhra Pradesh Rural Health Initiative. *Int J Epidemiol*. 2006; 35:1522–1529.