APPLIED STATISTICS AND CONVEX OPTIMIZATION

Predicting Presence of Heart Disease using Logistic Regression

by

ELIZABETH LIU

Candidate, MS Informatics

Department Mathematics and Computer Science

Texas Woman's University, Denton 76204

March 4, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Review of Related Literature

1.1 Background

1.1.1 Heart disease prevalence in the US

Heart disease or cardiovascular disease (CVD) is the leading cause of death for both men and women. On average, more than 2,200 Americans lose their lives to cardiovascular disease each day (Institute of Medicine, 2011). The American Heart Association (AHA) reports that approximately 82.6 million people in the United States currently have one or more forms of cardiovascular disease (Institute of Medicine, 2011). Cardiovascular disease encompasses disease types such as coronary heart disease (CHD), stroke, hypertension, and congestive heart failure (Institute of Medicine, 2011). Because of the devastating effects of heart disease, prevention efforts must be made. Prevention efforts are most effective when the person knows their risk.

1.1.2 Data mining in healthcare diagnosis

Data mining is a process of extracting and discovering patterns in large data sets. Nowadays with the ability to store copious amounts of patient health data, data mining has tremendous potential and practicality in healthcare (Ogundele, Popoola, Oyesola, & Orija, 2018). For instance, data mining has been already applied to help detect fraud and abuse, make healthcare management decisions, identify effective treatments and practices, and patients receive affordable healthcare services (Ogundele, Popoola, Oyesola, & Orija, 2018). Data mining can be used in the healthcare sector to make meaningful decisions of the diagnosis and treatment of patients, including the diagnosis of heart disease.

1.1.3 Machine Learning in healthcare diagnosis

Machine learning (ML) methods are also being used to assist healthcare decisions with rise of data mining. Some popular ML methods used in medical diagnosis include Neural Networks (NN), Bayesian Classifier (BC), Classification and Regression Tree (CART), Gradient Boosting (GB), etc. (Bhavsar, Abugabah, Singla, Alzubi, Bashir, & Nikita, 2021). In machine learning, a predictive model is built using training data. There are a wide variety of machine learning methods that can be used to build a predictive model. After the model is built, the model then undergoes model evaluation. The model is tested for its ability to preform on test data. The splitting of data into test and train data is known as the split group procedure. Currently, ML is used in the treatment, prognosis, and diagnosis of patients via their electronic health records (Bhavsar, Abugabah, Singla, Alzubi, Bashir, & Nikita, 2021). Machine learning methods have been widely used in building a predictive models for heart disease using all kinds of patient data.

1.2 Mathematical Context

1.2.1 Logistic Regression

While there are many approaches to building a predictive model, Logistic Regression will be explored.

Logistic regression has its origins in the early twentieth century. One of its earliest applications can be found in a study done by Berkson in 1944 (Berkson, 1944). To this day, it is applicable in many areas of science. In general, regression is a mathematical technique that defines the relationship between the outcome and its variables, or whether there is any relationship at all. Regression models can be built upon these relationships, where an outcome can be predicted based on the input and weight of the variables. Unlike other Regression methods, Logistic Regression is applicable when the outcome is binomial, such as in this paper, where are only two outcomes expected; the patient has heart disease and the patient does not have heart disease. In the Logistic Regression model, the probability of a particular outcome can be predicted, which in this case is the risk of heart disease (Sperandei, 2014).

1.2.2 Optimization: Cost Function & Gradient Descent

The cost function is a function that calculates the distance the between the predicted outcome and the actual outcome. The cost function can be thought of as measuring the 'error' in the model. This function is commonly used in logistic regression and machine learning to evaluate how well the model fits to the data, but is also used to get the most reliable model. The cost function can be minimized get to the most optimal model (Pant, 2019).

Gradient descent is a method that can be used to solve the minimization of the cost

function. Gradient descent is a algorithm that searches for the lowest point in the gradient, which is the point of lowest 'error' or the most minimal cost. This is where the most optimal combination of coefficients lie. Gradient Descent will be used in the optimization of the Logistic Regression model (Upadhyay, 2018).

1.3 Model Reliability

It is also pertinent to address the issue of model accuracy and reliability. The model needs to be able to provide a reliable diagnosis to influence health decisions in real life. Therefore, the model will be evaluated and compared to the base R glm() model performance. Both models will be evaluated using confusion matrix metrics. A confusion matrix is a table that is commonly used visualize and evaluate the performance of a model. From this table, useful metrics can be obtained that can be used to compare the two models (Saito & Rehmsmeier, 2018).

1.4 Project Goal

Logistic Regression will be used to build a model that can predict the risk of heart disease using a heart disease dataset. It will be built from scratch using Rstudio. The model will be optimized using the minimization of the Cost Function and the Gradient Descent algorithm. Afterward, the model will be evaluated using confusion matrix metrics. Lastly, the model performance will be compared with the base R glm() model performance to see if the model built can compete with other model performances.

Chapter 2

The Mathematics Behind the Model

2.1 Logistic Regression

Logistic regression can be thought of as a linear model with a binomial distribution and a logit link function.

Given the probability of p, the odds outcome is calculated as $\frac{p}{(1-p)}$. The logit function is the logarithm of the odds. As p is close to 1, the logit value heads towards ∞. As p is close 0, the logit value heads towards $-\infty$ (Brixius, 2019).

$$logit(p) = ln\Big(\frac{p}{1-p}\Big)$$

Figure 2.1 is logit function graphed.

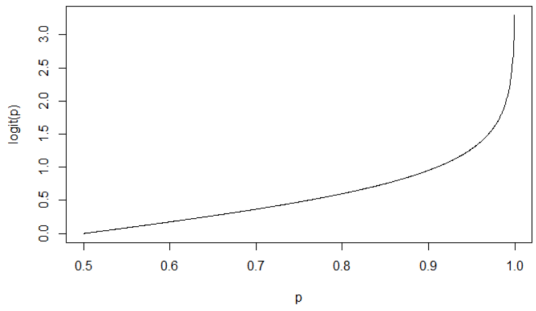


Figure 2.1: Logit Function Graph

The sigmoid function is the inverse of the logit function.

$$g(z) = \frac{1}{1 + e^{-z}}$$

Here is sigmoid is visualized in Figure 2.2.

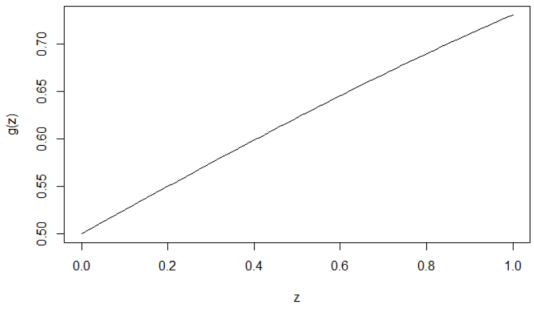


Figure 2.2: Sigmoid Function Graph

This also means that $sigmoid(logit(p)) = p$. The sigmoid function scales arbitrary real values to the range [0, 1]. Therefore, larger the value is, the closer to 1 the value will be inside the sigmoid function. The sigmoid function transforms real value variables into probabilities in this way. Using logit and sigmoid functions, variables and outcomes can be transform to fit a probabilistic scale (Brixius, 2019).

The linear relationship between x_i features and the outcome can be represented in this equation below.

$$log\Big(\frac{p^{(i)}}{1-p^{(i)}}\Big) = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ...$$

This also can be written in matrix form. $\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ...$ can be simplified to $\theta^T \cdot X$ or $\theta \cdot X$ and then placed inside a sigmoid function to get a probabilistic outcome. This equation is known as the hypothesis function for logistic regression.

$$h(x) = \frac{1}{1 + e^{-\theta \cdot X}}$$

X represents a matrix with vectors of all n data points. θ is a vector with the parameters for each feature. θ defines the parameters that determines the effect of each feature has on the outcome and how much impact. To create the model, θ must be found (Sperandei, 2014).

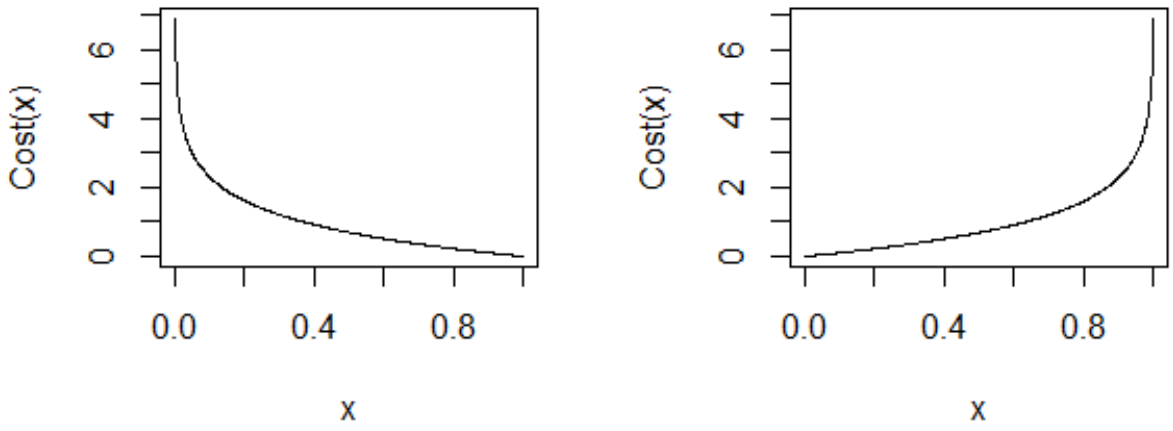
2.2 Cost Function & Gradient Descent

To get the optimal θ solution, the cost function and gradient descent will be used.

The cost function is defined as:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if y} = 1 \\ -\log(1 - h_\theta(x)) & \text{if y} = 0 \end{cases}$$

For each data point, the cost function measures each distance the actual outcome or y is from the predicted $h_\theta(x)$ value. The cost function can be used to determine how fit the model is to the data. The “cost” or error increases the further the $h_\theta(x)$ is from the actual outcome. The ”cost” decreases the closer $h_\theta(x)$ is from the actual outcome (Pant, 2019). Figure 2.3 is the cost function graphed.



(a) if $y = 1$ (b) if $y = 0$

Figure 2.3: Cost Function Graphs

The combined version of the cost function is:

$$J = -\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}log(h(x^{(i)})) + (1 - y^{(i)})log(1 - h(x^{(i)})))$$

The cost function can be minimized to solve for the optimal θ values. This optimization can be done with the gradient descent algorithm.

Gradient descent is an iterative optimization algorithm. It is used to find the minimum of a differential function, and in this case, the cost function (Pant, 2019).

The general form of gradient descent is:

Repeat until convergence {

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

}

For each feature, the algorithm searches for the minimum $J(\theta)$ value, which would yield the most optimal θ value. α is the rate at which the algorithm moves across the gradient.

The derivative is what makes the search possible and delegates the direction the algorithm should search. Positive derivative value means there is a positive slope up ahead in the gradient, which indicates a local maximum. A negative derivative value means there is a negative slope and a local minimum. In this case, the gradient descent algorithm would focus the search towards increasingly negative derivative values, and update the θ values iteratively until the most optimal solution is reached (Upadhyay, 2018).

Chapter 3

Dataset Description

The heart disease dataset comes from the heart disease database on the online UCI Machine Learning Repository. The heart disease database consists of 76 attributes. The source of the data comes from five hospitals; the Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., and V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D (Murphy & Aha, 1994). This database was used to create the heart disease dataset. The heart disease dataset used in this paper contains 270 observations and 14 attributes. Each column is a feature, and each row corresponds to data from one individual. More details about the 14 attributes in the table on the next page.

Table 3.1: Dataset Attributes in Detail

Name	Description	Value
Age	age in years	numeric
Sex	-	1 = male 0 = female
Chest	chest pain type	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic
BP	resting blood pressure in mm Hg	numeric
Cholesterol	serum cholestoral in mg/dl	numeric
FBS_over_120	fasting blood sugar >120 mg/dl	1 = true 0 = false
EKG	resting electrocardiographic results	0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular hypertrophy
Max_HR	maximum heart rate achieved	numeric
Exercise_angina	exercise induced angina	1 = yes 0 = no
ST_depression	ST depression induced by exercise relative to rest	numeric
Slope_ST	the slope of the peak exercise ST segment	1 = upsloping 2 = flat 3 = downsloping
vessels_fluro	number of major vessels colored by flourosopy	0 to 3
Thallium	test to visualize blood flow in heart	3 = normal 6 = fixed defect 7 = reversable defect
Heart_Disease, target variable	diagnosis of heart disease (angiographic disease status)	0 = <50% diameter narrowing 1 = >50% diameter narrowing

Chapter 4

Results & Discussion

4.1 Selecting of features

A correlation matrix was plotted to determine which features had strong correlations with the target variable 'Heart_Disease'. The results of the correlation matrix is visualized in Figure 4.1.

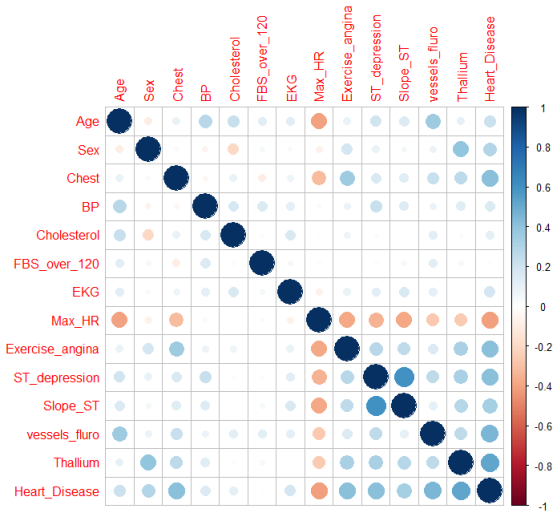


Figure 4.1: Feature Correlation plot

From this graph, it was noticed that the features 'Thallium','vessels_fluro','Chest', and 'ST_depression' had strong correlations with the target variable 'Heart_Disease'. I used the base R function glm() and the parameter 'logit' to determine if these features have enough value in predicting heart disease. The exact significance can bee seen in Figure 4.2.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.4996     1.0077  -6.450 1.12e-10 ***
Thallium       0.3991     0.1031   3.873 0.000107 ***
vessels_fluro  0.9397     0.2540   3.700 0.000216 ***
Chest          0.9370     0.2481   3.776 0.000159 ***
ST_depression  0.6242     0.1888   3.305 0.000948 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.2: Feature Significance

The result was all four features, including the intercept, were significant enough features. They all had a p-value of less than or equal to 0.001. These four features were used later on to build the Logistic Regression model. Other combinations of features were tested but these four features were the best combination of features overall.

4.2 Building from Scratch

To build this model from scratch, the Logistic Regression model and its variables had to be defined. The X variable is a n x $(i + 1)$ matrix with i number of features and n rows of datapoints. Table 4.3 shows the X matrix.

$$\mathrm{X} = \begin{array}{c} \begin{array}{ccccc} \textit{intercept} & \textit{Thallium} & \textit{vessels\_fluro} & \textit{Chest} & \textit{ST\_depression} \end{array} \\ \begin{pmatrix} 1 & 3 & 3 & 4 & 2.4 \\ 1 & 7 & 0 & 3 & 1.6 \\ 1 & 7 & 0 & 4 & 0.3 \\ 1 & 7 & 1 & 3 & 0.2 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{pmatrix} \end{array}$$

Figure 4.3: X matrix

Note: In the first column of the X matrix, an 'intercept' column was also included.

The y variable is a n x 1 matrix that holds the actual outcome for each row, with n rows of datapoints. Figure 4.4 shows the y matrix.

$$y = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ . \\ . \\ . \end{pmatrix}$$

Figure 4.4: y matrix

The θ_0 variable is a 5 x 1 matrix that holds the each of the θ values. Initially, the θ matrix will start with zeros. The zeros will eventually be that will be updated to include the optimal θ values after gradient descent. Figure 4.5 shows θ matrix.

$$\theta = \begin{matrix} intercept \\ Thallium \\ vessels\_fluro \\ Chest \\ ST\_depression \end{matrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Figure 4.5: θ matrix

After the variables where initialized, the functions in the Logistic Regression model were defined. The sigmoid function, cost function, and the gradient function were created. To perform gradient descent, the base R function *optim*() was used. The exact code used to build these function will be in the Appendix i. After running the code, the optimal θ values were calculated and obtained, as shown in Figure 4.1. This model will be referred to as 'model1'.

Table 4.1: model1 θ values

intercept	-6.4972526
Thallium	0.3988059
vessels_fluro	0.9395348
Chest	0.9369585
ST_depression	0.6239453

Using base R glm(), I created a Logistic Regression model 'model2' to compare with model1. The exact code used to build these function will be in the Appendix ii. The base R glm() model uses Fisher scoring and maximum likelihood methods for optimization of θ. The resulting model was obtained and seen in Table 4.2.

Table 4.2: model2 θ values

intercept	-6.4995820
Thallium	0.3991371
vessels_fluro	0.9397398
Chest	0.9370489
ST_depression	0.6242250

From these two figures, it is obvious that the θ values of model1 and model2 are very similar. This demonstrates that both methods are very comparable. Next, the predictive ability of the both models will be put to the test to see how much these slight θ value differences effect the performance of the models.

4.3 Model Evaluation & Comparison

I used the split group procedure, AOC-AUC, and confusion matrix metrics to evaluate the predictive ability of both model1 and model2.

The 270 observations from the heart disease dataset was split into two datasets; 'train' (with 189 observations) and 'test' (with 81 observations). The 'train' dataset was used to develop the model, and the 'test' data was used to evaluate the model. 'test' feature data was input into the model, and the model will predict an outcome. This outcome was compared with the actual 'test' data outcome. This data was used to create confusion matrix and metrics to compare the performance. The resulting comparisons can be seen in Figure 4.6.

	FALSE	TRUE
0	43	2
1	7	29

(a) model1

	FALSE	TRUE
0	43	2
1	7	29

(b) model2

Figure 4.6: Confusion Matrices of model1 & model2

Note: 0 represents the Absence of Heart Disease, and 1 represents the Presence of Heart Disease.

Looking at the confusion matrix, model1 and model2 perform the same. This demonstrates that the slight differences in the θ between model1 and model2 does not significantly make a difference in performance.

Using the confusion matrix, the accuracy, sensitivity, and specificity percentages were calculated and shown in Table 4.3.

Table 4.3: Accuracy, sensitivity, & specificity percentages

accuracy	0.89
sensitivity	0.81
specificity	0.96

The accuracy, sensitivity, and specificity determines how well both models can diagnose. Since the confusion matrix for model1 and model2 were the same, the accuracy, sensitivity, and specificity measure were also the same. Overall, accuracy, sensitivity, and specificity percentages are quite high, which means the models perform quite well and rarely give the incorrect diagnosis.

The probability both models could diagnose correctly is 89%. The sensitivity measures the probability of getting a person getting a positive result or presence of heart disease outcome in both models given that the person actually has heart disease. In this case, the sensitivity is 81%. This demonstrates the models were not too overgeneralized and still sensitive enough to detect of the presence of heart disease (Saito & Rehmsmeier, 2018). The specificity measures the probability the person would get a negative result or absence of heart disease outcome given a person does not have heart disease. The specificity was at 95%, which is pretty high. In this case, a high specificity means that the models rarely give out false-positive results (Saito & Rehmsmeier, 2018).

The ROC graph can be plotted using the results from the confusion matrix. Since the confusion matrix for model1 and model2 were the same, graph results also reflected this. ROC or Receiver Operator Characteristic graph summarizes the performance of both models by demonstrating the trade off between sensitivity and specificity (Tape, 2021). The graph is in Figure 4.7.

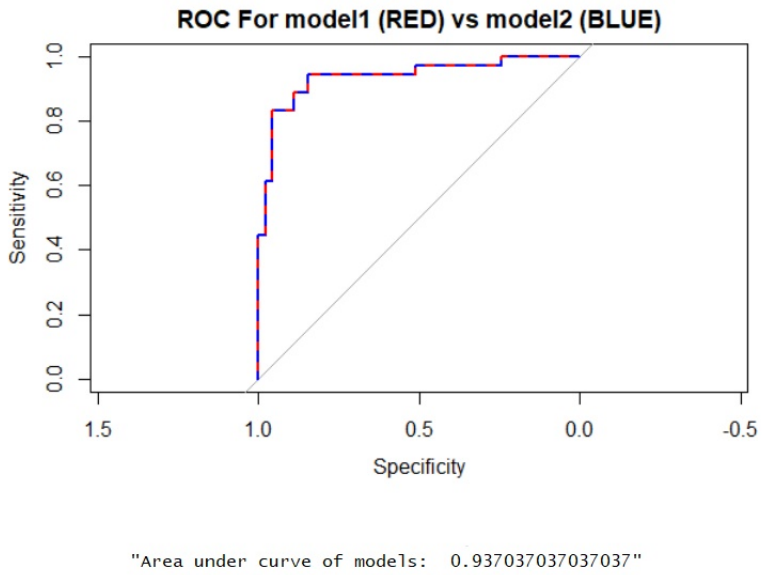


Figure 4.7: ROC graph for model1 & model 2

The gray line indicates the threshold were there is no predictive value in the model. The further the curve is from the gray line, the more predictive value there is in the model (Tape, 2021). For both of the models, the curve in the graph is far from this line, which means the model has significant predictive value. The area under the curve or AUC score can also be calculated. This score represents the probability that a randomly chosen positive instance will be achieved more often than negative instance. The highest AUC score is 1.00. In this case, the AUC score is is 0.937, which is very close to 1.00. This demonstrates that both models performed quite well.

Chapter 5

Conclusion

5.1 Summary of Project

Logistic Regression is a useful way find the relationship between variables and a binomial outcome. Using logistic regression, diagnosis models can be built with patient health data. This can be a solution for improving early detection, prevention, and treatment of diseases. In this project, a heart disease dataset was used to predict the presences of heart disease. The model were built in Rstudio. While logistic regression was the method of building the model, the optimization algorithm Gradient Descent was used to optimize the model. The model was called 'model1', and compared with another model 'model2'. model2 was built with the base R function glm() using the parameter 'logit'. The performance of the two models on test data evaluated using confusion matrix metrics.

5.2 Summary of Results

Overall, both models performed similarly. While the θ values or coefficients were slightly different between model1 and model2, the confusion matrix results for model1 and model2 were completely the same. This also meant that the other confusion matrix metrics were

the same as well. High accuracy, sensitivity, and specificity percentages indicated that the models perform quite well and rarely gives incorrect diagnoses. The high AUC score also demonstrated that both models have good predictive ability. These results indicate that logistic regression with Gradient Descent is a good method to create a predictive model. This method had almost equivalent results to the base R function glm() method in RStudio.

5.3 Limitations of Study

While the results of the project were seemingly successful, there are still ways the project can be improved, but cannot be feasibly implemented in time. One improvement is to create multiple models using various kinds of machine learning, optimization, or regression methods to compare with model1. Another improvement is using more evaluation methods, such as ANOVA or cross-validation. Because of the the slight differences in the θ values between model1 and model2, I wanted to compare how fit the two models were to the data. However, this was more strenuous to code in Rstudio because model1 was not a R model object,and thus cannot be used in any Base R functions like anova(). Therefore, I only used confusion matrix metrics to evaluate the models because it is simple to code in Rstudio. This is not to say confusion matrix is not meaningful; confusion matrix is an excellent tool to evaluate model performance and predictive ability. The heart disease dataset and feature selection can also be improved through data processing techniques such as PCA. In addition, more data and features could also be ideal. Although the models performed well on the test data, it is still difficult to determine how well will preform on real world data and whether the accuracy, sensitivity, and specificity percentages are high enough to negate the risks of a false positive or false negative result in the real world.

Appendix

i. model1 code

```
**Create Sigmoid Function**

sigmoid <- function(z){1/(1+exp(-z))}

**Create Cost Function**

cost <- function(theta, X, y){
  m <- length(y) # number of training examples
  h <- sigmoid(X %*% theta)
  J <- (t(-y)%*%log(h)-t(1-y)%*%log(1-h))/m
  J
}

**Create Gradient Function**

grad <- function(theta, X, y){
  m <- length(y)
  h <- sigmoid(X%*%theta)
```

```
  grad <- (t(X)%*%(h - y))/m
  grad
}

**Build Model & Return Coefficients**

#use the optim function to perform gradient descent
model1 <- optim(theta, fn = cost, gr = grad, X = X, y = y)

#return coefficients
model1$par
```

ii. model2 code

```
model2 <- glm(Heart_Disease~Thallium + vessels_fluro +
Chest + ST_depression, data=train, family = binomial("logit"))
```

Annotated Bibliography

Berkson, J. (1944).*Application to the logistic function to bio-assay. Journal of the American Statistical Association, 39*(227), 357. https://doi.org/10.2307/

Logistic regression has its origins in the early twentieth century. One of its earliest applications can be found in a study done by Berkson in 1944. Berkson (1944) states that the "function (referring to Logistic Regression) has been variously called the 'growth function,' the 'autocatalytic curve' and by other terms, according to the application to which it was put. It was rediscovered for the description of population growth by Pearl and Reed who, following Verhulst, called it the 'logistic' function. Since its wide statistical use stems, I feel confident, from the extensive applications made by Pearl and Reed, I shall refer to it by the general term 'logistic.'"

Bhavsar, K. A., Abugabah, A., Singla, J., Alzubi, A. A., Bashir, A. K., & Nikita. (2021).*comprehensive review on medical diagnosis using machine learning. Computers, Materials & Continua, 67*(2), 1997–2014. https://doi.org/10.32604/cmc.2021.014943

This paper offers an insight into the wide variety of studies that use ML for medical diagnosis. The paper referenced studies that designed models for all kinds diseases such as urinary tract infection, diabetes, Alzheimer's disease, etc. Not only that, the paper also summarizes the result of each study and model. The paper also offers insight to popular ML methods for medical diagnosis such as Neural Networks (NN), Bayesian Classifier (BC), Classification and Regression Tree (CART), Gradient Boosting (GB), etc. According to Bhavsar, Abugabah, Singla, Alzubi, Bashir, & Nikita (2021) the

"significant contribution of this paper includes a comprehensive review of the use of ML in medical diagnosis. Various ML algorithms used in medical diagnosis are described. The results of this study provide trends in diagnosis of various diseases using machine learning, and challenges, and future research directions."

Institute of Medicine (US) Committee on a National Surveillance System for Cardiovascular and Select Chronic Diseases. (2011).*Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung (2, Cardiovascular Disease).* Washington (DC): National Academies Press (US); Retrieved 2021, from: https://www.ncbi.nlm.nih.gov/books/NBK83160

This book gives an overview of what heart disease is, the prevalence of heart disease in the US, and the devastating effects of heart disease in the US. The book states the "Epidemiological data on heart disease, stroke, and associated risk factors are compiled and published annually in the Heart Disease and Stroke Statistical Update. This publication is a collaborative effort of the American Heart Association (AHA), the Centers for Disease Control and Prevention, the National Institutes of Health, and other government agencies. This chapter draws from the most recent edition of the report, the Heart Disease and Stroke Statistics 2011 Update, in addition to other resources to provide an overview of the burden of cardiovascular diseases in the United States."

Katz, M. H. (2003).*Multivariable Analysis: A Primer for readers of Medical Research. Annals of Internal Medicine, 138*(8), 644. https://doi.org/10.7326/0003-4819-138-8-200304150-00012.

This paper breaks down all kinds of Multivariable Analysis techniques, including Logistical Regression. The paper also give an overview of techniques for model fitting and testing reliability of the model. To determine how well the model fits the data, residuals and the Hosmer–Lemeshow goodness-of-fit test for Logistic Regression will be

used according to Katz (2003). To evaluate the reliability of the model, the model can be evaluated through the split group procedure as explained in Katz (2003).

Ogundele, I. O., Popoola, O. L., Oyesola, O. O., & Orija, K. T. (2018).*A Review on Data Mining in Healthcare. International Journal of Advanced Research in Computer Engineering & Technology, 7*(9).

This paper gives a good overview of data mining definition, techniques, and application in healthcare.Ogundele, Popoola, Oyesola, & Orija (2018) explains that the "healthcare sector requires data mining in discovery of knowledge and finding patterns for decision making. Data mining is the most advancing field of study which requires finding useful and meaningful details from a large data. Health data requires analytical methodology in identifying vital information that are used for decision making. Detection, prevention and management of diseases including fraud in the health insurance, reduce spending in the solution of medical care are some of the importance of data mining. It also help researchers to make effective healthcare policies, develop recommendation systems and health profiles for patients."

PM. Murphy, KW. Aha, *UCI Repository of machine learning databases: Heart Disease Data Set*, Irvine, CA: University of California, Department of Information and Computer Science, 1994.[https://archive.ics.uci.edu/ml/datasets/Heart+Disease]

This website provides the heart disease dataset as well as more information about the dataset such as where the data was sourced.

Sperandei S. (2014).*Understanding logistic regression analysis. Biochemia medica, 24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

This paper provides an overview of logistic regression. Sperandei (2014) states that "logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each

variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analyzing the association of all variables together."

Other Resources

Brixius, N. (2019).*The logit and sigmoid functions*, Nathan Brixius.[nathanbrixius.wordpress.com/2016/06/04/functions-i-have-known-logit-and-sigmoid]

M., J. (2019). *Logistic regression from scratch in R*, Medium.[towardsdatascience.com/logistic-regression-from-scratch-in-r-b5b122fd8e83]

Pant. A (2019).*Introduction to Logistic Regression.* [towardsdatascience.com/introduction-to-logistic-regression-66248243c148]

Tape, T. G. (2021) *Interpreting Diagnostic Tests Plotting and Intrepretating an ROC Curve*, University of Nebraska Medical Center.[http://gim.unmc.edu/dxtests/roc2.htm].

Tape, T. G. (2021) *Interpreting Diagnostic Tests The Area Under an ROC Curve*, [http://gim.unmc.edu/dxtests/roc3.htm]. University of Nebraska Medical Center.

Upadhyay, R. (2018). Gradient descent for logistic regression simplified - step by Step Visual Guide, YOUCANalytics. [http://ucanalytics.com/blogs/gradient-descent-logistic-regression-simplified-step-step-visual-guide].

Saito, T. & Rehmsmeier, M. (2016). Basic evaluation measures from the confusion matrix, Classifier evaluation with imbalanced datasets. [https://classeval.wordpress.com/introduction/basic-evaluation-measures]