Building Regularization Models to find which Model gives best RMSE for predicting Heart Disease

Instructor: Dr. Erfanul Hoque

Theoretical Machine Learning DASC5420

Thompson Rivers University

15th April 2023

Kishan Paladi Shekar

Student ID: T00710481

Abstract

In this era heart disease is a general health condition that affects people worldwide. Person is prone to heart disease by various key factors such as smoking, BMI, age, alcohol consumption, sleep time and other health conditions like physical and mental health. Main focus of this report is to compare these key factors and find which regression model is best suitable for predicting the heart disease.

Initially, data set is obtained from Kaggle that includes personal key indicators of heart disease. On this data set Exploratory Data Analysis (EDA) is performed for data visualization and later on to process the data into numerical values. After performing EDA, the new data set is generated and stored for future use. Once the EDA process is done, three different regression models are fit to calculate RMSE score to conclude which model is best fit for this data set to predict heart disease.

This paper evaluates and compares the RMSE (Root Mean Square Error) score between Lasso Regression, Ridge Regression and Elastic Regression and concludes which regression model performs better analysis to predict whether a person is prone to heart disease or not.

Key words: **Exploratory Data Analysis, Lasso Regression, Ridge Regression and Elastic Regression.**

Introduction

According to World Health Organization (WHO), cardiovascular disease is leading the number of deaths globally and its prevalence is increasing rapidly. Approximately around 17.9 million people globally die due to heart diseases [1]. Cardiovascular disease is one of the significant diseases which is growing rapidly throughout the world and is an important factor to be considered to minimize the death rate. There is an increase in number of people affected by heart disease each year, so its important to take into consideration how to minimize the death rate due to heart disease [2].

There are few important key factors which causes heart diseases. Following are the few which should be considered while concluding our results, they are BMI, age, smoking, alcohol consumption, physical activity, sleep pattern, other health conditions like physical health, mental health, general health which indicates overall health conditions, whether a person has diabetics, kidney disease or skin cancer. Based on all these parameters a researcher will be able to predict whether a person is prone to heart disease or not and based on the result person will be able to take precautions in prior. The data set which is used in this research is collected from Kaggle which has 3,19,795 tuples (number of data that is patients details for the attributes mentioned above) and has 18 attributes which includes 14 categorical variables and 4 continuous variables. On this data set to analyse the data EDA is performed and pre-processing is done to convert text data into numerical data for fitting into the model [3].

In this research paper three different regression models is performed to calculate the RMSE (root mean square error) to evaluate the performance of the model. When compared with Lasso regression model, Ridge regression model and Elastic regression model, results showed that Elastic regression model performed better compared to other two models.

Data Description

- ### Data Collection

In this generation population is growing rapidly as well as the diseases are also growing rapidly. One of the most common diseases where people are facing problems and death in large scale is cardiovascular disease. To over come this disease, we need to focus on the key factors which is affecting and improve our health so that there are less chances of getting heart disease. Hence, this paper works on data set which is gathered from Kaggle. Kaggle is one of the best data platforms for data analysts and scientists. Data set is collected from Personal key indicators of heart disease repository [3].

- ### Tools used

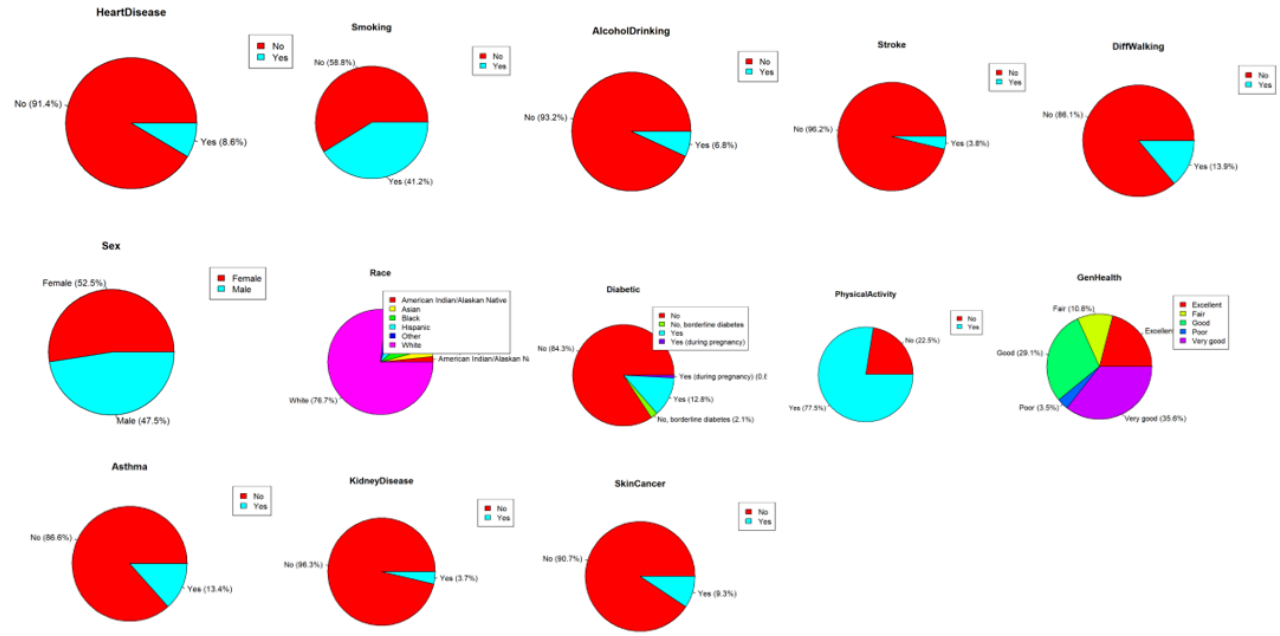
In this research R studio is used to perform all the analysis.

- ### Exploratory Data Analysis (EDA)

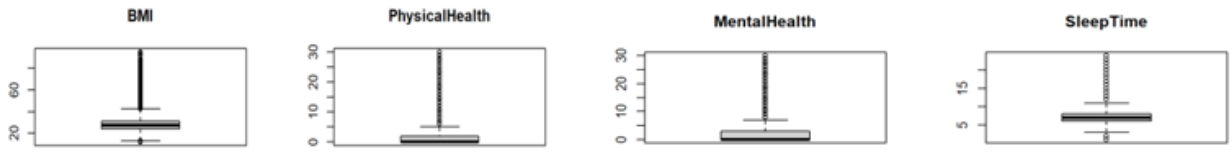
Exploratory data analysis plays a very important role in analyzing the data and build data visualization and also helps to understand better while performing pre-processing and fitting the model.

Aspects which are considered while performing EDA in this paper as follows:

1. Load the data: The data set which was collected from Kaggle is loaded into R studio.
2. Data description: In this step, number of tuples and attributes are calculated. There are 3,19,795 tuples and 18 attributes in the dataset. Later, with the help of "names" function names of attributes are printed which helps in understanding what are the attributes which are present in the data set.
3. Missing Data: This paper has used missing data function to check if there are any missing data in the data set, there were 0 missing values in this dataset.
4. Display Variables and their data types: With the help of "names" and "sapply" function variables and their data types is printed which helps while pre-processing the data and converting it into numerical form.
5. Remove Duplicates: In this paper unique function is used to check for any duplicates present, so that it can be removed for further analysis so that our model doesn't overfit and will be helpful in giving better results.
6. Describe the attributes: Here there are 4 continuous variables, with the help of describe function in this paper description of these variables are noted.
7. Plotting graph for categorical variables: In this paper, pie chart is generated to know more about our data set. Below are few pie charts which describes how our data set is split (Figure 1).
8. Check for outliers: For the continuous data we have checked for outliers using box plot (Figure 2).



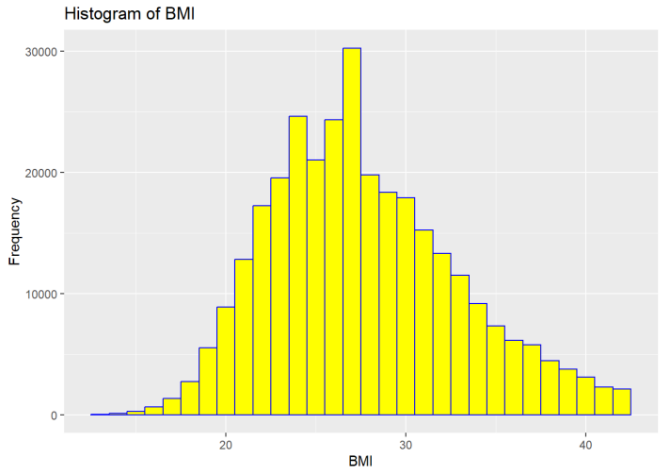
(Figure 1)



(Figure 2)

- **Pre-Processing**

This data set has numeric data as well as text data, to fit the model initially data should be converted into numerical data. Age attribute was converted into continuous variable by taking the mean of the age buckets which was given in the data set. Further, outliers were removed in the dataset by using IQR (inter quartile range) function which is one of the most common methods to deal with outliers. Later a histogram was built to check the skewness of the BMI by which we can say that BMI is normally distributed after removing the outliers (Figure 3).



(Figure 3)

Once all the variables were converted into numeric, new data set was saved for future use.

Methods

In this paper there are three regression methods used to calculate RMSE (root mean square error) value. The following methods are:

One of the common approaches for selecting between Ridge and Lasso regression is by performing k-fold cross validation with different tuning parameter lambda for each method and then selecting the value of lambda that minimizes the cross validation mean squared error (MSE).

Penalty term changes the optimization problem from minimizing the RSS to RSS + Penalty term. The strength of the penalty term is controlled by a tuning parameter (lambda), which is selected through a validation process, one of the common validation process is cross validation.

Ridge and Lasso regression are popular methods for regularized linear regression that helps to prevent overfitting of any model by adding penalty term to the objective function.

1. Ridge Regression: In ridge regression, L2-norm is added as penalty term to ordinary least square objective function. In ridge regression penalty term is proportional to square of their magnitudes.
   Ridge regression works well if the data set has many large parameters of about the same value.
   Ridge regression helps in converting the unrelated variables near to zero.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

(Equation 1)

2. Lasso Regression: In lasso regression, L1-norm is added as penalty term to ordinary least square objective function. In lasso regression penalty term is proportional to absolute value of their magnitudes.
   Lasso regression works well if the data set has small number of significant parameters and others are close to zero.
   Lasso regression helps in converting the unrelated variables to zero.

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

(Equation 2)

3. Elastic net Regression: Elastic net regression uses weighted combination of both L1-norm and L2-norm in its calculations as penalty term to ordinary least square objective function. It is a combination of both Ridge and Lasso Regression. It has two parameters lambda and alpha. In regularization it is more efficient when compared with Lasso and Ridge regression.

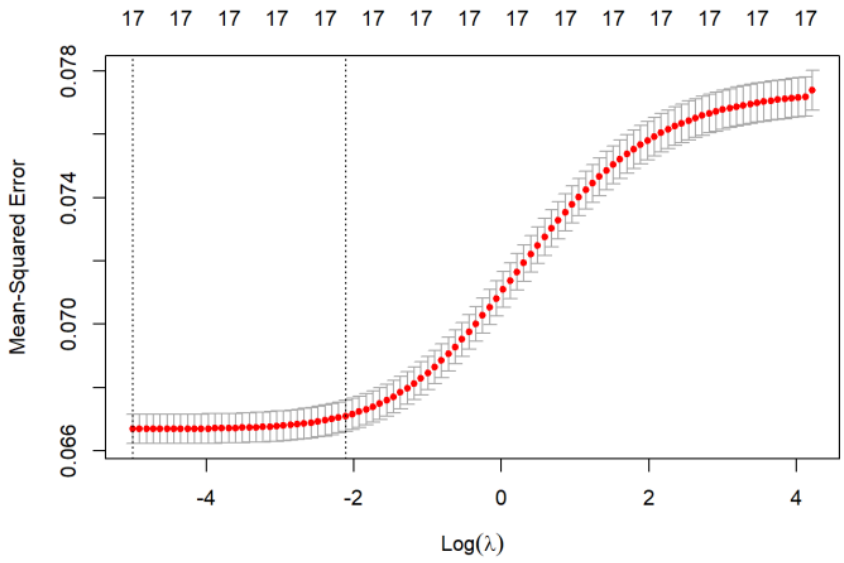
$$\sum_{i=1}^{n} \left(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left[(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]$$

(Equation 3)

Result and Analysis

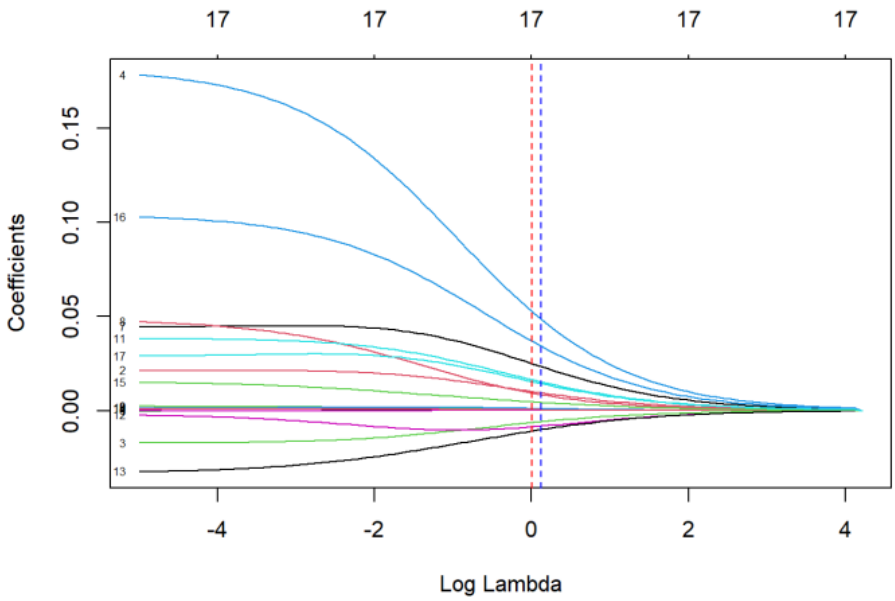
Initially the new data set is split into training and testing data. 75% of our data set is considered for training and the rest is performed for testing our model.

First, Ridge regression model is used to calculate the RMSE value to check whether our data set fits in the regression model. Initially, to get optimal lambda 10-fold cross validation is performed to minimize the error. (Figure 5) Shows the plot for optimal lambda.



(Figure 4)

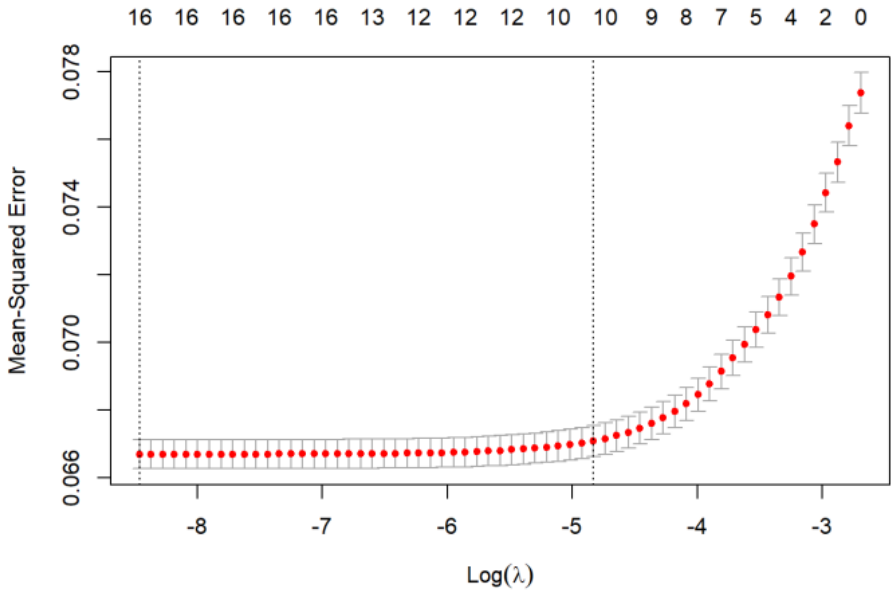
After obtaining the lambda values coefficient plot is constructed. (Figure 6) shows the ridge regression plot of coefficients vs Log lambda. From the below graph it is possible to figure out that the data set is fitting into the model as it is giving a ridge graph.



(Figure 5)

Later on, RMSE value is calculated on the test data to know whether our model is suitable for future use. By using Ridge regression model, there was a RMSE value of 0.267, by which it states that Ridge regression model is suitable for this data set.

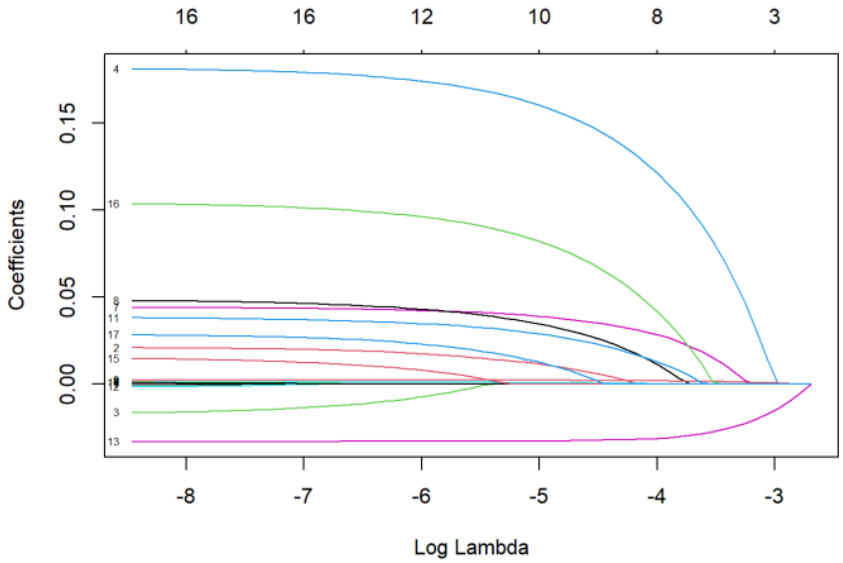
Lasso regression model is used to find the RMSE and compare with other two regression models. Initially, to get optimal lambda 10-fold cross validation is performed to minimize the error. (Figure 7) shows the plot for optimal lambda value.



(Figure 6)

Lasso regression minimizes the loss function by changing coefficient lambda to generate zero coefficients.

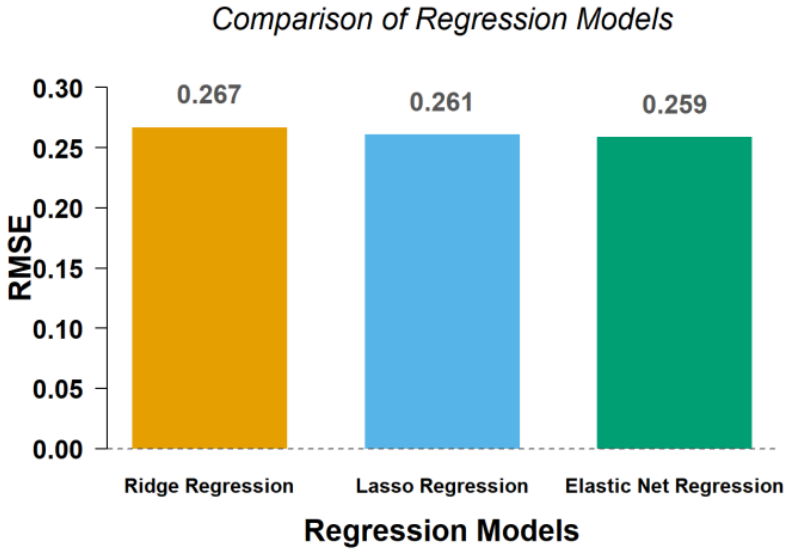
Once lambda values are obtained coefficient plot is constructed which helps in understanding better whether data set is fitting into the lasso regression model or not. (Figure 8) shows the plot of coefficients vs Log lambda values.



(Figure 7)

Later on, RMSE value is calculated on the test data to know whether our model is suitable for future use. By using Lasso regression model, there was a RMSE value of 0.261, by which still says that lasso regression model is suitable for this data set.

On the same data set elastic net regression is performed by setting the train control and providing the tune length to 10. When performed prediction on elastic net model RMSE score was 0.259 which was smaller when compared to Ridge regression and Lasso Regression.



(Figure 8)

From the above figure 9, it is evident that all the three regression model which was preformed on the heart disease data set is suitable for future analysis. When compared elastic net regression model is best suitable as it has the less RMSE value when compared with other two models.

Table to show the comparison of Regression Models

Regression models	RMSE value
Ridge Regression	0.267
Lasso Regression	0.261
Elastic net Regression	0.259

Conclusion

In conclusion, we can say that elastic net regression model is best suitable for analysis and predicting heart disease. Through this study, we can conclude that it will be helpful for healthcare professionals to take early decisions as well as prior precautions and reduce heart disease problems throughout the world.

Git hub repository link: GIT HUB REPO

References

1. World Health Organization. Cardiovascular Diseases [Internet]. World Health Organization. 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. STROKE: pathophysiology, diagnosis, and management. S.L.: Elsevier - Health Science; 2021.
3. Personal Key Indicators of Heart Disease [Internet]. www.kaggle.com. Available from: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease