**Heart Disease Dataset**

**Final Project Report, MATH 8050, Fall 2022**

Sai Vipraj Telukoti, C17202671

Vamshi Krishna Mummadi, C27032539

Yaswanth Mulakala, C16676264

**Dataset Description: -**

Heart Disease is a major health threat, that is why it is important to know risk factors so that proactive measures can be taken before it’s too late. Several health conditions such as lifestyle, age and family history can increase risk for heart disease.

This report aims to highlight factors that contribute to heart disease by analyzing a dataset obtained from Kaggle.

This dataset dates from 1988 and consists of four databases:

1. Cleveland
2. Hungary
3. Switzerland
4. Long Beach V

It contains 14 attributes and 1025 observations. While working with this dataset it was found that, it did not record the data in the most efficient way. It was discovered that the ‘target’ variable in the data was swapped. This variable provides indication as to whether the heart disease is present or not. This inconsistency in the data could be attributed to human error.

To carry out further analysis, this variable was flipped back to reflect the correct proportions.

The dataset has 14 columns. So, it is too big to take a screenshot of and show it here. Here is the name of the columns and the explanation of each variable as described in Kaggle.

1. age: The age of a person
2. sex: The person’s gender (1 = male, 0 = female)
3. cp: The types of chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. trestbps: Resting blood pressure (mm Hg on admission to the hospital)
5. chol: Cholesterol measurement in mg/dl
6. fbs: Fasting blood sugar (if > 120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality,

2 = showing probable or definite left ventricular hypertrophy by Estes’ criteria)

1. thalach: Maximum heart rate achieved
2. exang: Exercise induced angina (1 = yes; 0 = no)
3. oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot)
4. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: down sloping)
5. ca: The number of major vessels (0–3)
6. thal: A blood disorder called thalassemia (1 = normal; 2 = fixed defect; 3 = reversable defect)
7. target: heart disease (0 = no, 1 = yes)

**Modeling: -**

1. **Logistic Regression: -**

Logistic regression is a classification algorithm used to find the probability of event success and event failure. Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature. Logit function is used as a link function in a binomial distribution.

Logistic regression is also known as Binomial logistics regression. It is based on sigmoid function where output is probability and input can be from -infinity to +infinity.

Logistic regression will be used to a generate a model that can predict the heart disease in patients. For the sake of this data, logistic regression is the ideal choice for two reasons. They are: -

Firstly, the target (dependent) variable in the model is binary; secondly, it is a predictive analysis and can be used to explain the relationship between one dependent binary variable with multiple predictors.

Since resting blood pressure and cholesterol were log transformed it may be difficult to interpret it. Hence, additional feature engineering step will be carried out. Standardization of these variables will enable us to interpret the effects in terms of standard deviation.

**a) Implementation of the logistic model: -**

Table

Description automatically generated

Since the model incorporated logistic regression, the coefficients above correspond to log odds value. These coefficients directly do not state anything except that there is a positive relationship. The p-value is very small, which indicates that this relationship is statistically significant. Therefore, to understand and interpret this relationship more accurately, odds ratio will be computed.

### b) Interpreting the Coefficients through Odds Ratio: -

Graphical user interface, application, Word

Description automatically generated

Intercept will be ignored as it does not give any relevant information. However, rest of the coefficients will be interpreted as following:

(Intercept) age sex log1\_trestbpsstd ## 0.012 1.060 5.505 1.246 ## log1\_cholstd ## 1.302

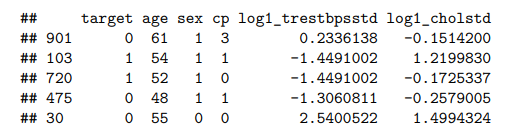
* **Age**: odds ratio of 1.060 can be interpreted as - increasing age by 1 year is associated with 1.06x increase in odds of getting a heart disease. in other words, aging 1 year is associated with a 6% [100 x (1.060 - 1)] increase in the odds of getting a heart disease.
* **Sex**: It is important to note that the odds ratio of 5.505 represents the odds of going to from female to male (since 0 = female and 1 = male). Hence, it is interpreted as, being a male is associated with 5.505x increase in odds of getting a heart disease.
* **Chest Pain**: odds ratio of 0.347 can be interpreted as increase in chest pain level by 1 lead to 65.3% [100 x (1 - 0.347)] decrease in heart disease

Log1\_trestbps: the odds of getting a heart disease are 24.6% [100 x (1.246 - 1)] higher is you are 1 standard deviation above log blood pressure.

Log1\_chol: the odds of getting a heart disease are 30.2% [100 X (1.302 - 1)] higher if you are 1 standard deviation above the average log cholesterol.

### c) Evaluating Predictive Performance: -

To evaluate predictive performance, train/test framework will be used. It is and 80/20 split to the original ‘heart data’. So, 80% training data and 20% test data.



**d) Predictions: -**

Now that the predictions are now stored for the train and test data sets, confusion Matrix function can be used to generate confusion matrix which will tabulate the correct and incorrect predictions versus the true values. Recall that since logistic regression was used in the model, classification metrics will be used.

### e) Confusion matrix for predictions on Training set and test set: -

Results between the training and the test datasets are very similar.

* **Accuracy**: is the portion of correct predictions. Accuracy of 0.6% means that, 60% of the heart rate predictions on the test set are correct.
* **Recall**: Based on the definition of recall which is the proportions of actual true that the model labelled true, it can be said that the model correctly identified 57% of the patients with heart disease
* **Precision**: is the proportion of positive predictors that are actually true. So, when the model predicted heart disease, it is true 62% of the time.
* In general, the results between the training and the test sets are very similar, point to the fact that our model can predict results on a dataset that it is very new to. In other words, overfitting was not a problem.

1. **Ridge Estimator: -**

Instead of RSS, we will calculate a version which is more penalizing:

**Diagram

Description automatically generated**

with λ ≥ 0 is called the tuning parameter.

This cost function is valid for linear regression only. \*

The larger the latter, the more our model will be simple and inverted. The penalizing term λ∑β ^ 2 is called "shrinkage penalty" and becomes small when the coefficients of our estimators are close to 0.

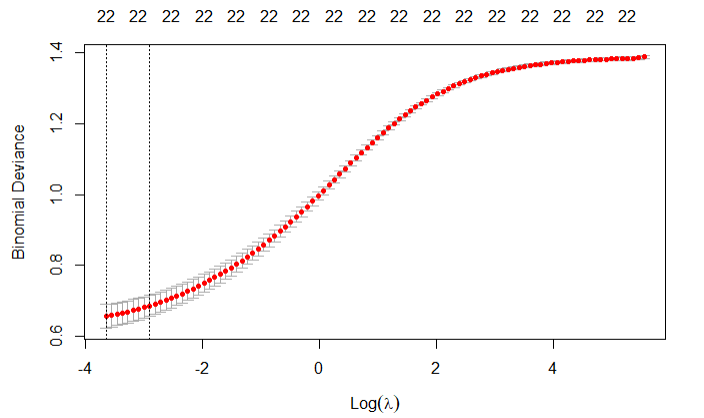
To find the optimal lambda, we must cross-validate.

Chart

Description automatically generated

We see here the evolution of the value of our coefficients as a function of lambda. What is interesting to see here is that the variables CA and CP are the most significant. However, the method does not allow us to see which variables to choose. Here, we will see how to find the optimal lambda and thus minimize the cost function of our model.

* **Cross-validation: -**

****

****

We see here that we minimize our cost function with lambda = 0.0026. Given that our model is a classification, the cost function is Deviance and not RSS as in the case of a regression.

1. **LASSO: -**

The big difference with the Ridge estimator is that here the value of the coefficients may be so penalized that it can be zero. This will therefore act as a means of selecting the variables of our model.

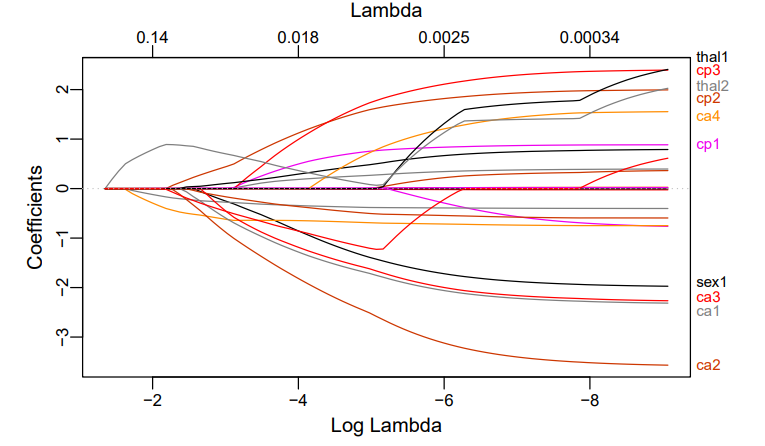
Here is the new cost function:

A picture containing text, clock

Description automatically generated

We can notice that the essential difference with the Ridge method is that we apply an absolute value and not a square for our estimator B.

For the creation of the model, the difference is that we will set alpha = 1 instead of 0.



If we compare with the graph of the ridge method, the difference is quite significant. The importance of the regressors does not necessarily change but the value of the coefficients is very different and will especially be much more penalized. Here, it would rather be necessary to select the variables Thal, CA and CP which remains a result relatively similar to the ridge method.

Chart

Description automatically generated

A picture containing logo

Description automatically generated

Obviously, the optimal lambda is no longer the same with this method!

1. **Elastic Net: -**

The main defect of the LASSO method is that in some cases it can be too penalizing. In addition, the correlation between 2 variables can have the consequence that the coefficients of the other variables are too much penalized, unlike the correlated variables. Therefore, we use the Elastic-net method which will combine the Ridge and LASSO method.

The penalizing term therefore becomes:

Text

Description automatically generated with medium confidence

The implementation of the method is no longer done with the glmnet package but with caret:

Chart

Description automatically generated

We see that the values of lambda> 0 are too penalizing! Moreover, the more our lambda will increase, the less the prediction accuracy will be good! We must therefore stay with a lambda strictly less than 0 where the precision will not change. Note also that alpha value does not matter in our case. We therefore return rather to a LASSO model, no doubt due to the correlation of our variables.

1. **Baye’s Algorithm: -**

We need to remove the extra columns we added while performing BLR before implementing Bayes algorithm.

Implementing Confusion Matrix on training set-

**Table

Description automatically generated with medium confidence**

We can say that naive bayes algorithm is 84.51% accurate with the training data. Now, Validating the model with the testing data by predicting and creating confusion matrix.

Implementing Confusion Matrix on testing set-

Table

Description automatically generated with medium confidence

we can conclude that the model generated with help of Bayes algorithm is 77.43% accurate or we can also say that the misclassification rate for Bayes algorithm is 22.57%.

# **Conclusion: -**

To conclude, we can say the feature selection methods are multiple and will mainly depend on your data and what you seek to do with it. In the case of data with little variance, a regression ridge can give very satisfactory results. If you are looking to select variables, the lasso estimator can be your ally and help you identify the variables that are important, if they are not correlated. More generally, the elastic-net method is probably the one that will give you the best results in many cases, but it is not necessarily the one that you should always consider first.

# **Summary: -**

This project aimed at understanding the factors that contribute to heart disease. A model was built to predict whether a patient has heart disease or not using a dataset obtained from Kaggle. This dataset had to be adjusted for the ‘target’ variable as the values were found to have switched. Below is a quick summary of the entire project by each variable:

**Age**

* This variable was recorded in years and displayed bimodal distribution.
* Density plot revealed that there are 3 groups. Less than 54 years (labelled as ‘adults’), greater than or equal to 54 and less than 70 years (labeled as ‘older adults’) and greater than or equal to 70 years (labelled as ‘elderly’)
* Older population in the dataset seem to have higher number of heart diseases than younger.
* Aging 1 year is associated with a 6% increase in the odds of getting a heart disease.

**Sex**

* Sex is related to heart disease, in that males have more heart disease.
* Being a male is associated with 5.5 increase in odds of getting a heart disease.

**Chest Pain**

* Side by side bar plots revealed even those that reported 0 chest pain had heart disease. This means that the level of chest pain experienced by the patient is not necessarily a result of heart disease.
* Odds ratio revealed that increase in chest pain level by 1 lead to 65.3% [100 x (1 - 0.347)] decrease in heart disease.

**Blood Pressure**

* The distribution of this variable was right skewed, so it was log transformed to make it normally distributed.
* Although those with heart disease seem to have lower blood pressure compared to those that do not, the odds of getting a heart disease are 24.6% higher is you are 1 standard deviation above log blood pressure.

**Cholesterol**

* This variable was right skewed, so it was log transformed to make it normally distributed.
* The odds of getting a heart disease is 30.2% [100 X (1.302 - 1)] higher if you are 1 standard deviation above the average log cholesterol.

# **Contributions: -**

|  |  |
| --- | --- |
| Name | Contribution |
| Sai Vipraj Telukoti | EDA, Logistic Regression, Documentation |
| Vamshi Krishna Mummadi | Data Cleaning, Ridge Estimator, Documentation |
| Yaswanth Mulakula | LASSO, Elastic-net, Documentation |