EDA and Important Predictors Identification in Heart-Analysis Dataset

Objective :

EDA and Important Predictors Identification in Heart-Analysis Dataset

What is EDA:

EDA is the process of identifying patterns, observing trends, and formulating hypothesis.

Dataset : https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Target column:

output – whether Heart-Attack is possible or not ~ 0- NO , 1= YES

Predictors/Feature Columns:

- age - Age of the patient
- sex - Sex of the patient
- cp - Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
- trtbps - Resting blood pressure (in mm Hg)
- chol - Cholestoral in mg/dl fetched via BMI sensor
- fbs - (fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
- restecg - Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy
- thalachh - Maximum heart rate achieved
- oldpeak - Previous peak
- slp - Slope
- caa - Number of major vessels
- thall - Thalium Stress Test result ~ (0,3)
- exng - Exercise induced angina ~ 1 = Yes, 0 = No

Whats done :

1. **Clean the data, Transform categorical variables to int handle missing values and outliers**
   a. Explored the distribution of the population in each column using histogram/box plot to find the outliers
   b. Transformed categorical column for the sake of analysis.
   c. Used log transformation to eliminate skewness in some columns
   d. Checked if there is null values in the dataset. No missing values

2. **Find how features are correlated with target**
    a. Visualised correlation matrix using heatmaps (triangle heatmap, exclusively with the output column)
    b. Effect of different features on other features and target variable

3. **Identify which features that are significant in predicting the target variable. (Feature Engineering)**
    a. we compare the correlation between features and remove one of two features that have a correlation higher than 0.4.

EDA Results:

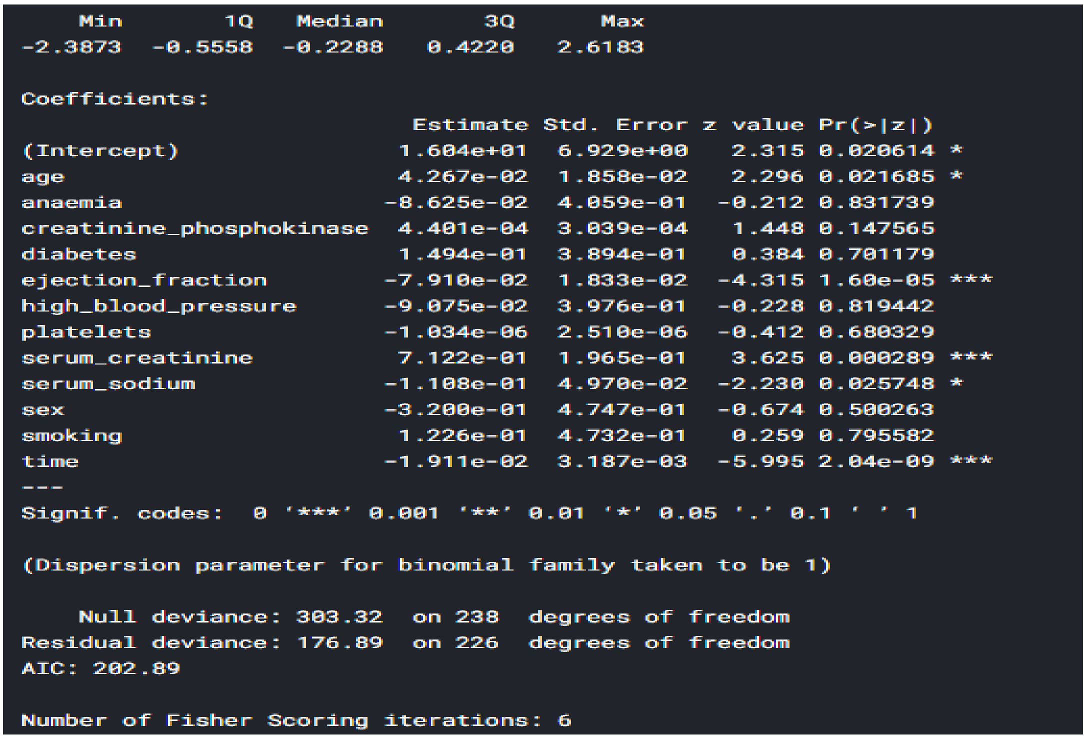
1. middle aged (45 to 60 years) persons have higher chance of heart attack
2. trtbps (Resting Blood Pressure) has week or slightly negative relation with heart attack
3. thalach (Maximum Heart Rate Achieved) has positive relation with heart attack
4. oldpeak (Previous peak) has negative relation with heart attack
5. Women are more likely to have heart problem than men (based on ratio)
6. People with High Cholestoral (more than 200) have very higher chance of heart attack
7. People with Maximum Heart Rate Achieved > 150 has higher chance of heart attack
8. Age has negative relation with thalachh (Maximum Heart Rate Achieved)
9. Chest Pain has higher chance of heart attack
10. typical angina has lower chance of heart attack than other chest pains
11. People with lower major vessels (caa) have much higher chance of heart attack
12. Age has positive relation with n major vessels (caa) (older people are more likely to have vessels)
13. People with Thall == 2 have much higher chance of heart attack
14. Women are likely to have higher levels of cholesterol compared with men
15. Cholestoral has positive relation with Age
16. Slope has positive relation with heart attack and People with Slope == 2 have much higher chance of heart attack
17. trtbps (Resting Blood Pressure) has positive relation with Age
18. oldpeak (ST depression induced by exercise relative to rest) has highly negative relation with heart attack

Logistic Regression:

- We now perform the Logistic Regression on the training set taking all of the 12 variables as predictors.
- We then study the summary of the model and then perform predictions on the test data set.
- We store the success rate on the test data into a vector for comparison later and also observe the confusion matrix to understand the numbers where the predictions worked or did not work.

Summary - Logistic Model

As can be seen from the above summary statistics that age, ejection fraction, serum creatinine, serum sodium and time (follow up time) are some of the siginifcant predictors of heart failure. We now use this model to predict the event of death due to heart failure on the test data and then calculate and store the success rate of the predicted values.



```
    Min       1Q   Median       3Q      Max  
-2.3873  -0.5558  -0.2288   0.4220   2.6183  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)               1.604e+01  6.929e+00   2.315 0.020614 *  
age                       4.267e-02  1.858e-02   2.296 0.021685 *  
anaemia                  -8.625e-02  4.059e-01  -0.212 0.831739    
creatinine_phosphokinase  4.401e-04  3.039e-04   1.448 0.147565    
diabetes                  1.494e-01  3.894e-01   0.384 0.701179    
ejection_fraction        -7.910e-02  1.833e-02  -4.315 1.60e-05 ***
high_blood_pressure      -9.075e-02  3.976e-01  -0.228 0.819442    
platelets                -1.034e-06  2.510e-06  -0.412 0.680329    
serum_creatinine          7.122e-01  1.965e-01   3.625 0.000289 ***
serum_sodium             -1.108e-01  4.970e-02  -2.230 0.025748 *  
sex                      -3.200e-01  4.747e-01  -0.674 0.500263    
smoking                   1.226e-01  4.732e-01   0.259 0.795582    
time                     -1.911e-02  3.187e-03  -5.995 2.04e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 303.32  on 238  degrees of freedom
Residual deviance: 176.89  on 226  degrees of freedom
AIC: 202.89

Number of Fisher Scoring iterations: 6
```

The confusion matrix shows the hits and misses of the model. Also we observe the **success rate of the logistic model as about 85%**. The model with 12 predictors and 239 training observations does a fairly good job in predicting the response and surely performs better than tossing a coin :)

Ridge estimator (with k = 10 Folds Cross Validation)

- we perform the Ridge Regression following the same steps as above and capture the success rates of the model and the confusion matrix.
- All of the 12 variables are taken as predictors in the model. The ridge regression values depend on the value of the lambda chosen during prediction. Each value of lambda would yield a separate set of coefficients of the predictors.

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

- with λ ≥ 0 is called the tuning parameter. This cost function is valid for linear regression only.
- To arrive at the best lambda a 10-fold Cross Validation is performed using **cv.glmnet** (cv.glmnet by default performs a 10-fold Cross Validation) on the training data set and the lambda corresponding to the minimum cross validation error is selected and used during the prediction process.

Summary of Ridge Regression

- Although the predictor space is not shrunk to remove some predictors totally but the ridge regression tries to shrink the predictor coefficients towards zero in such a way that the success rate does not get lower than using all predictors.

- **Ridge Regression has given us a success rate of 82% which is slightly lesser than the normal Logistic Regression.** The success rate rate based on the training data is quite good.

Lasso Regression (with k = 10 Folds Cross Validation)

- As a final step of the modelling process we perform a lasso regression and observe the results of the predictions and the success rates of the prediction.
- The Lasso Model also depends on the lambda value selected for the prediction.
- Just like the Ridge Regression process explained above the lambda for the Lasso Model also is selected by performing 10-folds Cross Validation using cv.glmnet (cv.glmnet by default performs a 10-fold Cross Validation) on the training data set and selecting the lambda which gives the lowest Cross Validation error.

Summary of Lasso Regression

- In Lasso Regression as we might know the predictors space is shrunk and only the significant predictors are kept in the model. Lets see which are the predictors that have non zero coefficients.
- And the predictors which have been shrunk are anaemia, diabetes, high blood pressure, platelets, sex, smoking.
- **Lasso Regression has given us a success rate of 83%**

Model Comparison & Conclusion:

- The above success rates (Ridge = 82%, Lasso = 83%, Logistic = 85%) suggest that the Normal Logistic Regression performs the best on the test data.
- The Lasso Model performs just a little below the Normal Logistic Model but has the advantage of shrinking the predictor space from 12 to 6 (half).
- If we compare the summary of the Logistic Model and the Coefficients of the LASSO model which have NOT been shrunk to zero (i.e. ‘age', 'creatinine_phosphokinase', ‘ejection_fraction', 'serum_creatinine', 'serum_sodium', 'time'), we see that they match almost nearly perfectly except for the predictor "creatinine phosphokinase" which the LASSO model thinks as significant to be included in the model whereas the summary statistics of the Logistic Model does not show as being significant.
- The Ridge model though does not shrink the coefficients of the predictors space to zero but gets them very near to the zero-value based on the best lambda used from Cross Validation. The performance of the Ridge Model is also satisfactory in as much as is it is just behind the Lasso Model.
- So, in a summary if we have the wherewithal to include all the predictors (all 12 of them) then the Logistic Model above will give a 85% accuracy in prediction i.e. the most accurate prediction. If we do not have the wherewithal to include all of the 12 predictors, then the Lasso Model with 6 predictors helps us achieve a near about same accuracy as explained above. It needs to be noted that as more and more predictors are included in the model the interpretability of the model becomes difficult and hence a model that does the job but with a lesser set of predictors is what is generally preferred. So, the Lasso Model can be used for predictions of the event of death due to heart failure.