ACTL3143 Assignment:

Heart Disease Prediction using Deep Learning

JUNE 2023

SHARON ZHOU

Z5310019

INTRODUCTION

Taking an estimated 17.9 million lives each year, cardiovascular diseases are the leading cause of death globally and are a significant consideration for health and life insurers (WHO, 2021). Through its ability to learn and adapt to intricate patterns, deep learning models have the potential to become a powerful tool in medical diagnostics. As actuaries, a predictive model that can detect early signs of heart disease can lead to a more effective risk assessment, reduced insurer costs and better health outcomes for society (Bhattacharyya and Jindal, 2022).

This assignment will use past patient variables along with a target condition of having the presence or absence of heart disease to create a predictive model for future patients. This is a binary classification type problem with a "goal" field of 1 = presence and 0 = absence. This early draft will conduct exploratory data analysis to gain insights into the dataset, before fitting a logistic regression model as a baseline. Further improvements will be made through utilising deep learning techniques.

DATA PREPARATION

The dataset to be analysed is from the University of California Irvine (UCI) data repository of 270 hospitalised Cleveland patients who were admitted to hospital in 1988. Minor data formatting was performed to prepare the data for modelling. The steps that were taken are as follows:

- Column names were shortened to the following:
    - col_names = ['age', 'sex', 'cp', 'trestbps', 'fbs', 'chol', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target']
- The *target* variable was converted from string to Boolean format.
- Categorical variables were defined and converted into string formatting.
- Numerical variables were defined and converted into integer formatting.
- Potential data imbalances were checked, with the conclusion that there is minimal dataset imbalance.

The numerical variables were also standardised using RobustScaler(), while the categorical variables were encoded using one-hot encoding. The data was then split into a 60% training, 20% validation and 20% test set.

EXPLORATORY DATA ANALYSIS (EDA)

EDA was performed on the Cleveland dataset to maximise insights of the dataset and provide a clear picture of the explanatory variables and their relationships with each other.

Firstly, Figure 1 checks the correlation between numeric variables in the dataset to ensure an absence of highly correlated variables. The correlation plot illustrates that the target variable has a high negative correlation with "thalach" and high positive correlations with "oldpeak" and "ca".

Figure 2 shows the distribution of heart disease against the chest pain variable. As can be expected, patients experiencing degree 4 chest pain tend to

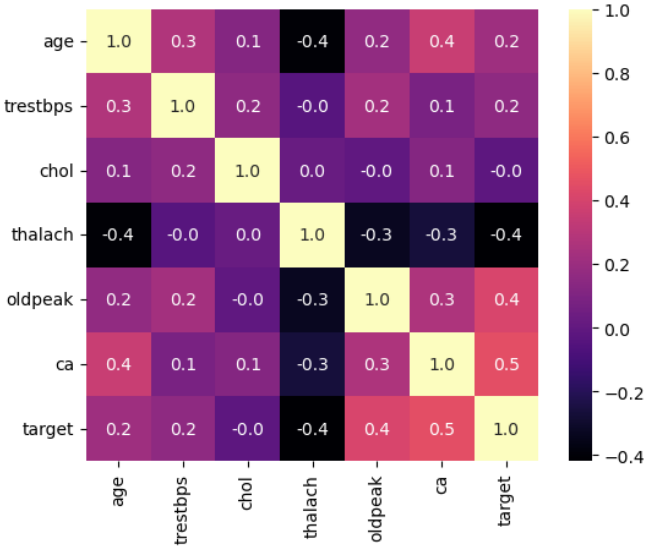


Figure 1: Correlation matrix

have the presence of heart disease. Figure 3 shows a boxplot of age against heart disease. From the plot, we can see that older age groups are at a higher risk of heart disease in the dataset. Further EDA is included in Appendix A and B, which show univariate and bivariate analysis.

Figure 2: Distribution of Heart Disease against cp

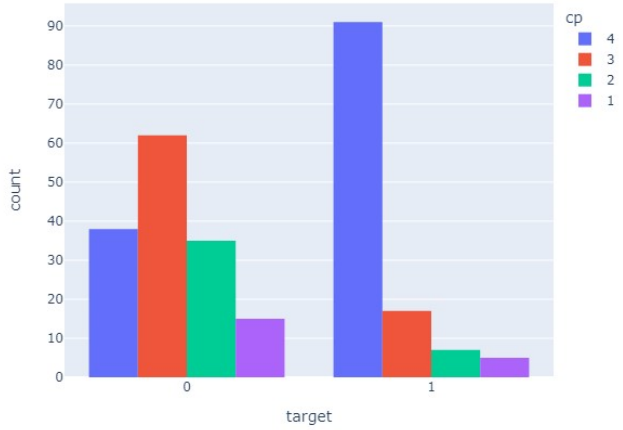
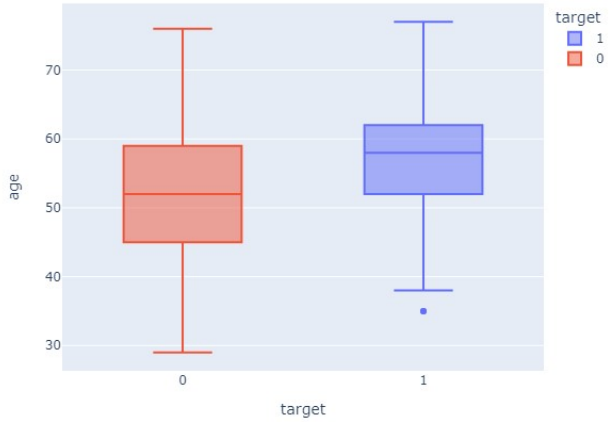


Figure 3: Boxplot of Age vs. Target



BENCHMARK MODEL: LOGISTIC REGRESSION

Our task is to solve a binary classification problem of predicting whether a patient has the presence or absence of heart disease. Logistic regression is a simple yet powerful machine learning algorithm that is a special case of linear regression. Namely, the model predicts the probability of an observation belonging to either binary class using a logit function. A linear regression equation is shown below, where y is the target variable and $X_1, \dots X_{13}$ are the explanatory variables:

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{13} X_{13}$$

A logistic regression works by fitting a sigmoid curve to the observations from the training set. A sigmoid curve is 'S' shaped and maps any real-valued number to a value between 0 and 1. The following equation shows the sigmoid function and resulting logistic regression equation:

$$\sigma(x) = \frac{1}{1 + e^{-y}} \rightarrow \sigma(x) = \frac{1}{1 + e^{-\beta_0 + \beta_1 X_1 + \cdots + \beta_{13} X_{13}}}$$ (Thorn, 2021)

The confusion matrix and ROC curves are shown in Figure 4 and 5 respectively. From the confusion matrix, the test accuracy of the model is calculated as (TP + TN) / (TP + TN + FP + FN) = 0.87. The area under the ROC curve (AUC-ROC) is 0.89, which indicates a fairly high model performance.

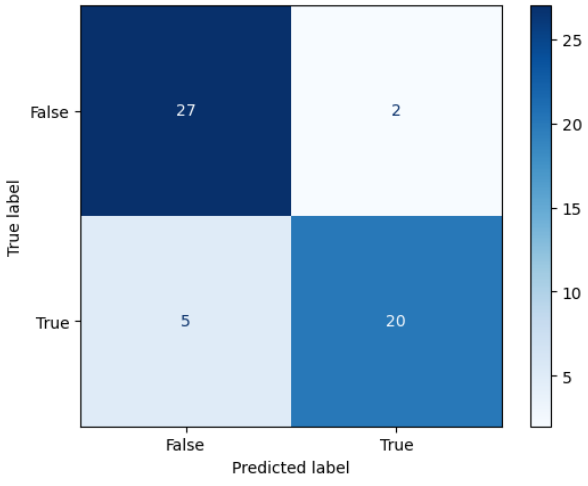


Figure 4: Confusion matrix of predictions

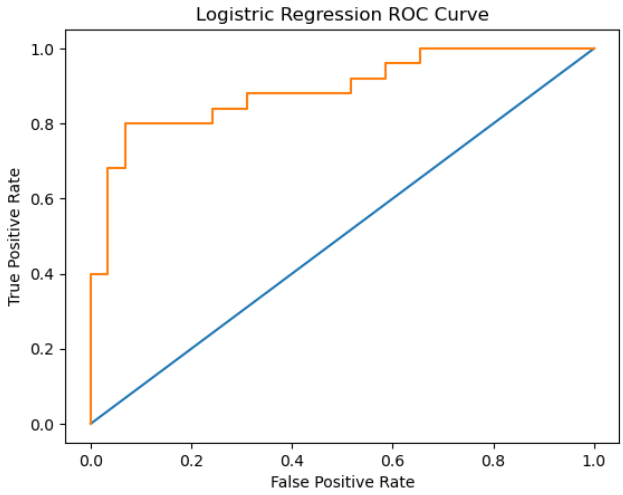


Figure 5: Logistic Regression ROC Curve

REFERENCES

Bhattacharyya, A. and Jindal, K. (2022) *Electric dreams: How machine learning could revolutionise healthcare and insurance*, *The Actuary*. Available at: https://www.theactuary.com/features/2022/01/28/electric-dreams-how-machine-learning-could-revolutionise-healthcare-and (Accessed: 27 June 2023).

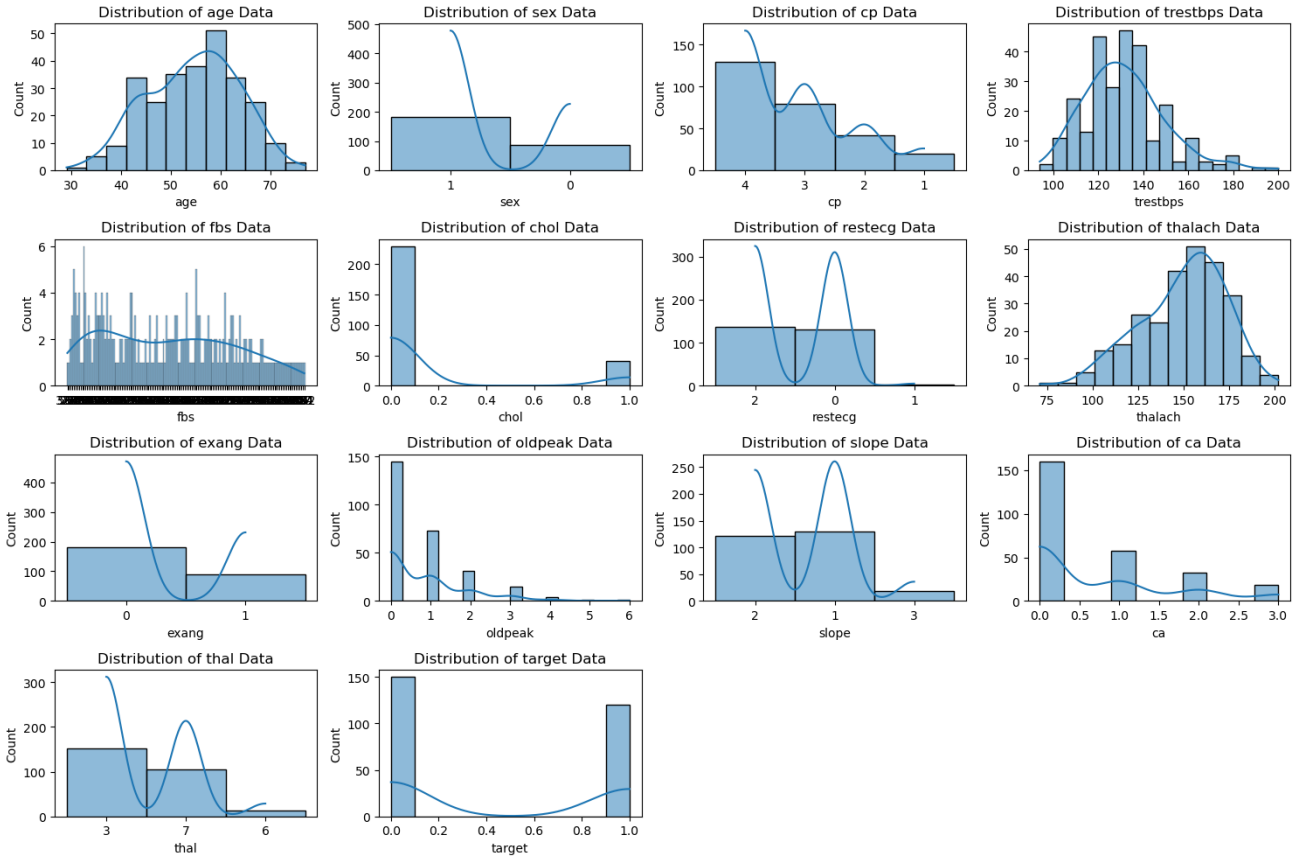
"Cardiovascular disease." *World Health Organization*, 2021, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (Accessed: 27 June 2023).

Thorn, J. (2021) *Logistic regression explained*, *Medium*. Available at: https://towardsdatascience.com/logistic-regression-explained-9ee73cede081 (Accessed: 27 June 2023).

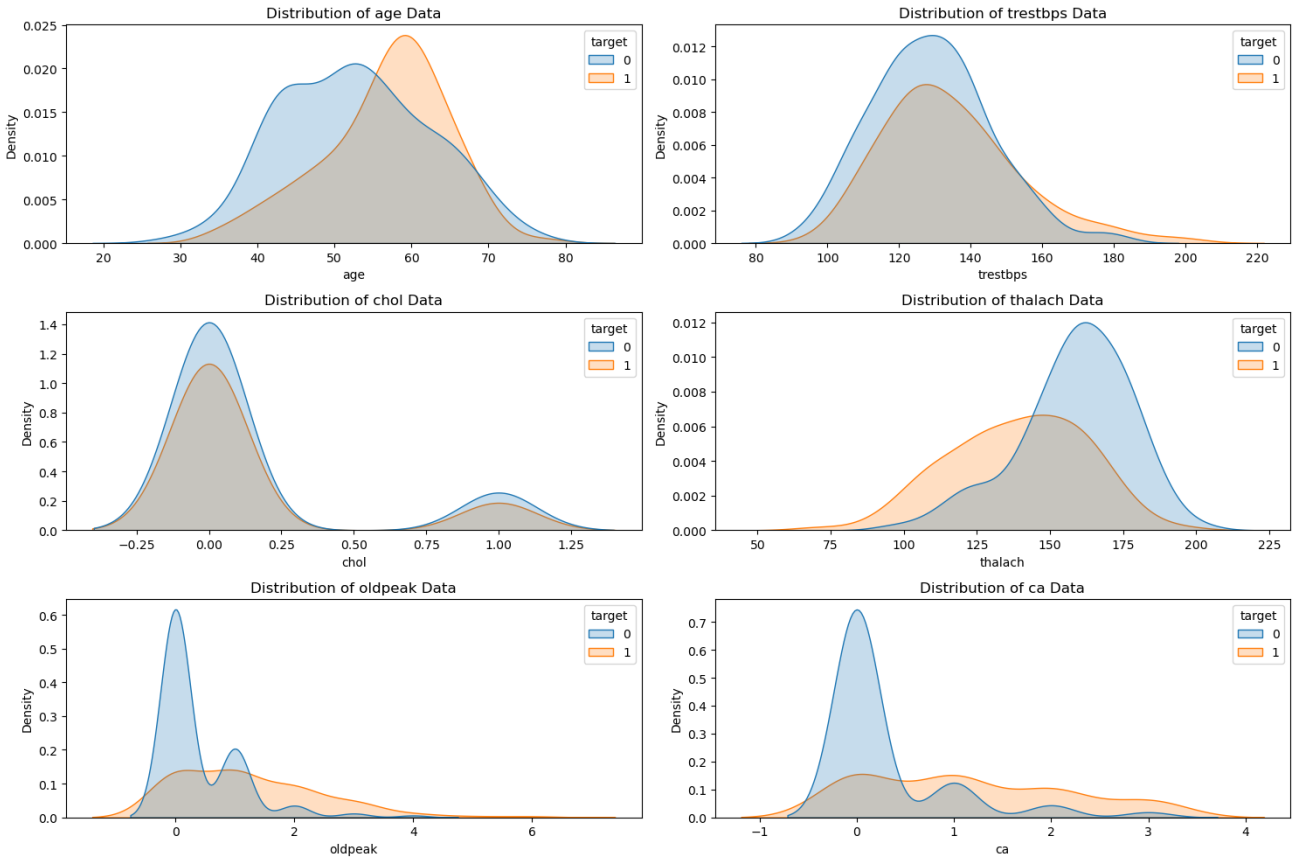
APPENDIX

APPENDIX A: UNIVARIATE ANALYSIS

The distribution graphs below have been plotted for all variables in the dataset to check for linearity of features. We can see that in general, the features are evenly distributed. However, there's an imbalance in "restecg", and "chol", "oldpeak" and "ca" are skewed towards 0.



APPENDIX B: BIVARIATE ANALYSIS



APPENDIX C: CLASSIFICATION REPORT

Appendix C shows the model's classification report, including the precision, recall and F1 score metrics.

- Precision: measures the proportion of correctly predicted positive instances out of total predicted positive instances and is calculated as $\frac{TP}{TP+FP}$.
- Recall: measures the proportion of correctly predicted positive instances out of total actual positive instances and is calculated as $\frac{TP}{TP+FN}$.
- F1 score: calculated as $2 \times \frac{Precision \times Recall}{Precision+Recall}$.

	precision	recall	f1-score	support
0	0.84	0.93	0.89	29
1	0.91	0.80	0.85	25
accuracy			0.87	54
macro avg	0.88	0.87	0.87	54
weighted avg	0.87	0.87	0.87	54