STAT 473 Project
Applications of Machine Learning To Heart Disease

May 12, 2023

Report Prepared By
Ethan Huang

Table of Contents

Introduction

Our dataset, sourced from the UCI Machine Learning Repository, is a compilation of 299 patients with heart failure. The data were all collected during a follow up period after their initial diagnosis. It contains 13 variables: 7 numeric and 6 categorical. The continuous numeric variables are age in years, creatinine phosphokinase concentration in the blood, the ejection fraction per contraction, the number of platelets, serum creatinine concentration in the blood, serum sodium, and patient follow up time in day. The six boolean variables reflect the diagnosis of anemia, high blood pressure, and diabetes in the patient. Gender, smoking, and death are also boolean variables. We will be using all variables except time as predictors throughout our study.

Questions of Interest

Our main question of interest is to see which classification model works best in predicting the patient's probability of death based on their health metrics. We utilized five machine learning models: logistic regression, linear discriminant analysis, quadratic discriminant analysis, decision trees, and random forest. To evaluate which model works best, we compare the sensitivity, accuracy, and AUC amongst the models. Second, we seek to analyze the significant predictors for our models to understand its impact. Lastly, in conducting our research, we hope to understand the real world implications of our data.

Analysis

Exploratory Data Analysis

To help understand the variables within our study, we summarized the result of our categorical variables to understand the balance within our data set. Based on our results, we see that of the 299 patients within our dataset 41.13% were diagnosed with Anemia, 41.81% were diagnosed with Diabetes, and 35.12% of our patients were diagnosed with high blood pressure. 64.88% of our patients were Male, while 32.11% smoked and 32.11% died.

Fig. 1

	Anemia	**Diabetes**	**High Blood Pressure**	**Sex**	**Smoking**	**Dead**
% Yes	43.13%	41.81%	35.12%	64.88% Male	32.11%	32.11%
% No	56.87%	58.19%	64.88%	35.12% Female	67.89%	67.89%

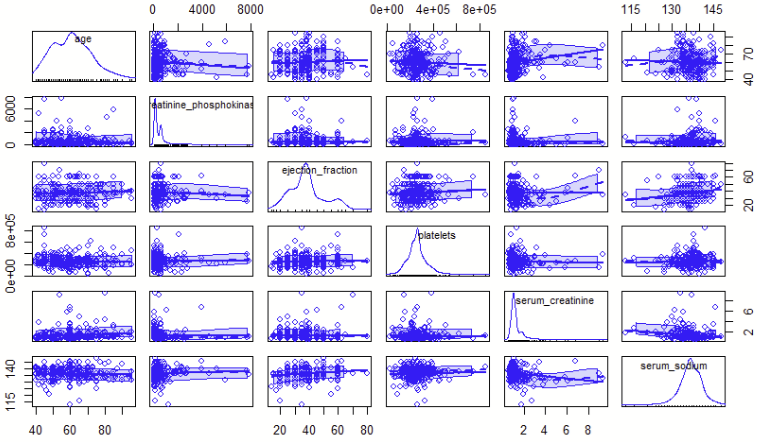
From Fig. 1 we can see that there are some unbalanced within our Dataset. The amount of Men in our sample size is 14.88 percentage points greater than the population. We also observe a large amount of smokers in our dataset when compared to the population, 22.11 percentage points greater, indicating that these two categorical variables may have a large causal relationship if one was to have heart disease, we cannot yet determine if this relationship is statistically significant or if it affects the probability of death.

The second aspect of exploratory data analysis was in regard to our six applicable continuous numerical variables.

Fig 2.

	Age	Creatinine Phosphokinase	Ejection Fraction	Platelets	Serum Creatinine	Serum Sodium
Min.	40	23 mcg/l	14%	25,100	0.5 mg/dl	113 mEq/l
Mean	60.83	581.8 mcg/l	38.08%	263,358	1.394 mg/dl	136.6 mEq/l
Max	95	7,861 mcg/l	80%	850,000	9.4 mg/dl	148 mEq/l

Fig. 3



From Fig. 2 the Mean age of our patients was 60.83 years old. For Creatinine Phosphokinase, the minimum value of 23 mcg/l is significantly closer to the mean value of 581.8 mcg/l when compared to the maximum value of 7,861 mcg/l. Ejection fraction has a mean value of 38.08%, meaning that with each heart contraction, 38.08% of the blood is ejected from the heart. A higher ejection fraction is indicative of hearth health and recovery from a heart attack.

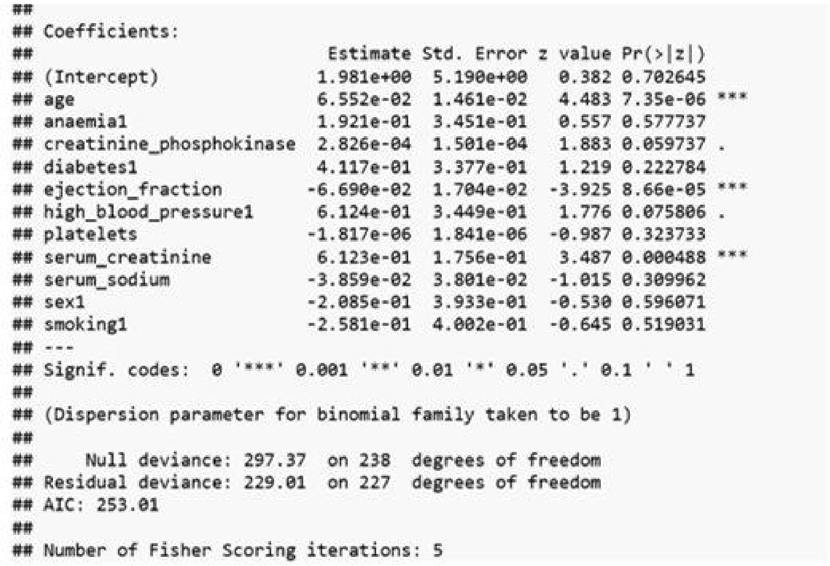
We expect that this variable would have significant predictive power for our models because it can be used as an instrumental variable for recovery from heart attacks.

With Fig. 3 we can see that the distributions of Age, Ejection Fraction, platelets, and Serum Sodium are relatively normally distributed. However we can see that there is a severe positive skew from Creatine Phosphokinase and Serum Creatinine. This can have an effect on models such as linear discriminant analysis, quadratic discriminant analysis, and logistic regressions. However for our random forest models and decision trees, the skewness would have no effect on their predictive capacities or interpretability.

Model 1: Logistic Regression

We first built a logistic regression model using our training set.

Fig. 4



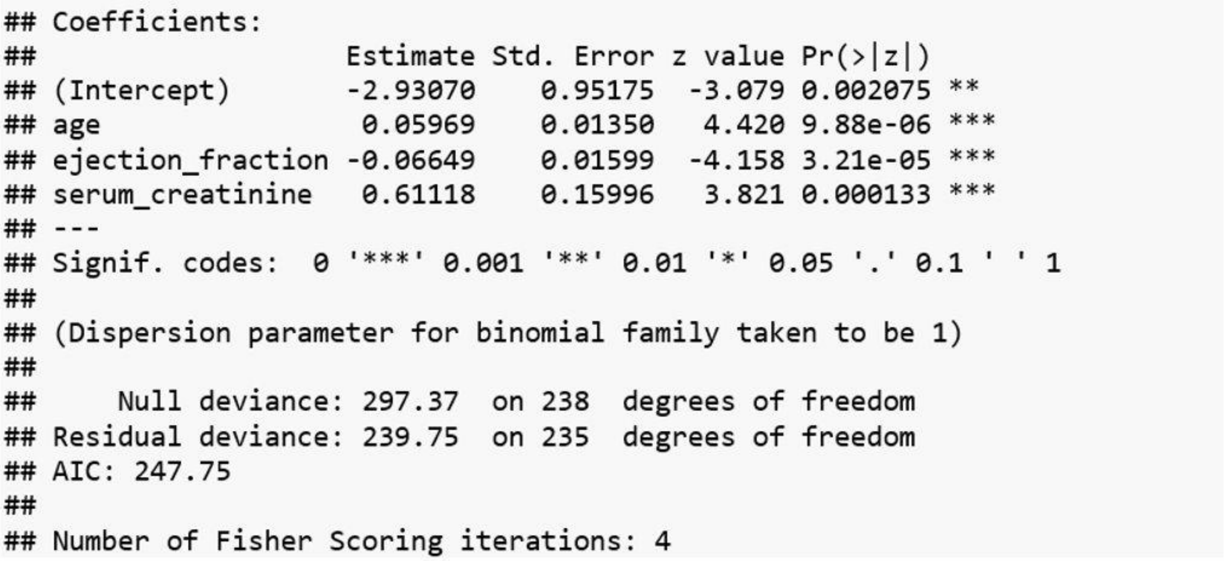
```
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.981e+00  5.190e+00   0.382 0.702645
## age                      6.552e-02  1.461e-02   4.483 7.35e-06 ***
## anaemia1                 1.921e-01  3.451e-01   0.557 0.577737
## creatinine_phosphokinase 2.826e-04  1.501e-04   1.883 0.059737 .
## diabetes1                4.117e-01  3.377e-01   1.219 0.222784
## ejection_fraction       -6.690e-02  1.704e-02  -3.925 8.66e-05 ***
## high_blood_pressure1     6.124e-01  3.449e-01   1.776 0.075806 .
## platelets               -1.817e-06  1.841e-06  -0.987 0.323733
## serum_creatinine         6.123e-01  1.756e-01   3.487 0.000488 ***
## serum_sodium            -3.859e-02  3.801e-02  -1.015 0.309962
## sex1                    -2.085e-01  3.933e-01  -0.530 0.596071
## smoking1                -2.581e-01  4.002e-01  -0.645 0.519031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 297.37  on 238  degrees of freedom
## Residual deviance: 229.01  on 227  degrees of freedom
## AIC: 253.01
##
## Number of Fisher Scoring iterations: 5
```

Fitted Model:
log (PI/1-PI)= 1.981 + 0.06552*age + 0.1921*(anemia) + 0.0002826*(creatinine phosphokinase) + 0.4117*(diabetes) - 0.06690*(ejection fraction) + .6124*(high blood pressure) - 0.000001817*(platelets) + 0.6123*(serum creatinine) - 0.93859*(serum sodium) - 0.2085*(Gender) - 0.2581*(Smoking)

We see three significant predictors: age, ejection fraction, and serum creatinine. As age increases by one year, the estimated log-odds in favor of death increase by 0.06552 holding all other predictors fixed. As the percentage of blood leaving the heart at each contraction increases by one percent, the estimated log-odds in favor of death decrease by 0.06690 percentage points holding all other predictors fixed. Last, as the level of serum creatinine in the blood increases by one milligram, the estimated log-odds in favor of death increase by 0.6123 holding all other predictors fixed.

Using our three significant predictors, we build a second logistic model.

Fig. 5



```
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.93070    0.95175  -3.079 0.002075 **
## age                0.05969    0.01350   4.420 9.88e-06 ***
## ejection_fraction -0.06649    0.01599  -4.158 3.21e-05 ***
## serum_creatinine   0.61118    0.15996   3.821 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 297.37  on 238  degrees of freedom
## Residual deviance: 239.75  on 235  degrees of freedom
## AIC: 247.75
##
## Number of Fisher Scoring iterations: 4
```

Fitted Model:
log(PI / 1- PI): -2.93070 + 0.05969*age -0.06649*(ejection fraction) + 0.61118*(serum creatinine)

As age increases by one year, the estimated log odds of death increase by 0.05969 holding all other predictors fixed. As the percentage of blood leaving the heart at each contraction increases by one percent, the estimated log odds of death decrease by 0.06449 holding all other predictors fixed.   Last, as the level of serum creatinine in the blood increases by one milligram, the estimated log-odds in favor of death increase by 0.61118 holding all other predictors fixed.

Now, we compute confusion matrices (using 0.3 as our threshold) and ROC Curves for these two models (see appendix).

The sensitivity of model one is 66.66%, and the accuracy of this model is 70%. The sensitivity of model two is 71.4%, and the accuracy is 73.3%. Hence, model two is better in terms of confusion matrices.

However, when we plot the ROC curve and compute the corresponding AUC, we see that the AUC for the first model is greater than that for the second model. The AUC for model 1 is 0.7399, and the AUC for model 2 is 0.7314. Hence, based on AUC curves, Model 1 is better than Model 2.

Model 2: LDA/QDA

The next models we chose to run were linear and quadratic discriminant analysis. For both of these models we kept all of the variables, and our response variable was the death event. If the patients did not die they were classified as “0” and if they did die they were classified as “1.” We began with the LDA model and first looked at the prior probabilities to determine whether our data was balanced. The prior probability for the group where patients did not die was 68.62% and 31.38% for the patients that did die. The data does seem to be slightly imbalanced. After running the model, we then constructed the confusion matrix to calculate the accuracy and sensitivity values. The accuracy percentage for the linear discriminant analysis was 66.7% and the sensitivity percentage was 23.8%. Since our sensitivity value is really low, the linear discriminant analysis is not good at detecting patients who actually died. In order to increase our sensitivity value, we chose to change our threshold. After implementing a threshold, our sensitivity percentage increased to 61.9%. Now, the model is better at detecting true patients that died. Next, we built an roc curve and calculated the area under the curve. From the graph we can see that it does not stretch too much towards the upper left corner, and our area under the curve was 0.7338.

We moved on to conducting a quadratic discriminant analysis and got the same quantities for prior probabilities. Then we constructed a confusion matrix and got 70% accuracy percentage and 42.86% sensitivity percentage. From the sensitivity percent we conclude that the model is not good at detecting true patients that died. Then we built an roc curve and calculated the area under the curve. From the graph we also see that it does not stretch towards the upper left corner and the area under the curve was 0.6935.

Comparing the two models we conclude that the quadratic discriminant analysis has a higher accuracy value compared to the linear discriminant analysis. This means that the QDA model is better at correctly identifying true patients that died and did not die. As for sensitivity, the LDA model has a higher value than the QDA model. This means that the LDA model is better at detecting true patients that died. Lastly, the LDA model has a slightly higher area under the curve compared to the QDA model.

Model 3: Decision Tree

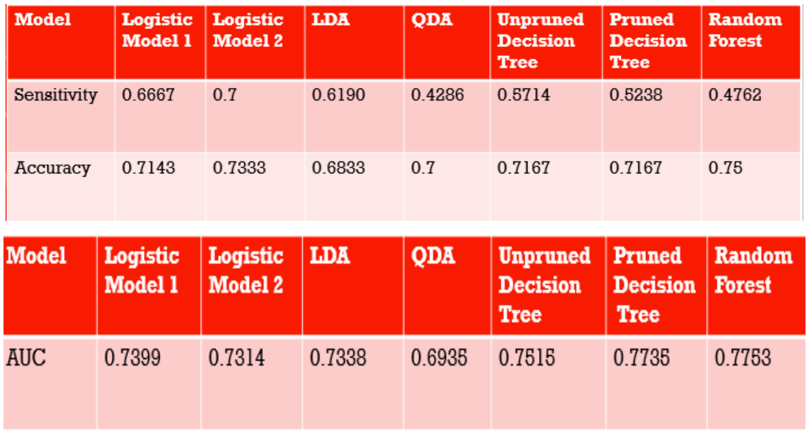
Tree method is a good choice to answer our question of interest because it is very easy to interpret as well as being a good method to apply when dealing with data that does not move linearly. We specifically need a classification tree because we are dealing with a response variable of yes death occurred “1” or no death occurred “0”. The tree model is also built on the same test set and training set proportion all the other models were built on in this project, and DEATH_EVENT is the response variable, and all other variables are predictors. We fit our model to the train set. The first tree model had 25 terminal nodes, which appeared to overfit the data. Its misclassification error rate is good though because it is small at a rate of 11.3%. The sensitivity was 57.14%, and the accuracy was 71.67%. The AUC value was 75.15% which is decently close to 100%. Pruning the tree model was the next step in the hopes of increasing interpretability and predictability. Pruning the classification tree, we now see 10 terminal nodes with a misclassification error rate of 13.39%. The sensitivity was calculated to be 52.38% and the accuracy was calculated to be 71.67%. The AUC value was 77.35% which is decently close to 100%. Comparing the unpruned and pruned tree models we observe that the pruned model does fit the data well, but not necessarily better than the unpruned original tree due to the increase in misclassification error rate.

Model 4: Random Forest

Random forest is a good approach to answer our question of interest because it will address any high variance that may have occurred in the previous decision tree model. This model was on the same test set and training set proportion all the other models were built on in this project, and DEATH_EVENT is the response variable with all other variables are predictors. The model is fit to the train set. The output revealed that serum_creatine is the most important variable followed by, age, and ejection_fraction. Applying a confusion matrix to see how well a random forest fits the model in terms of accuracy yielded results of 75% , and for sensitivity yielded results of 47.62%. The AUC value was calculated as 77.53%, which is a decently close to 100%. The accuracy of the random forest model was better than the pruned and unpruned tree models, so to analyze this data using tree methods, a random forest model should be used.

Conclusion

In this analysis of Heart Failure, five different machine learning models are built. We are most interested in answering the question of which model works best by comparing sensitivity and accuracy along with evaluating AUC of the models. Comparing the results for each model, (fig. 6) it is shown that the logistic model 2 has the best sensitivity rate, and the random forest model has the best accuracy rate. The random forest model also has the best AUC value. Overall using a logistic model would be the best choice when evaluating this data if we care most about predicting people who will die, while random forest would be best to use when we are predicting death occurring and not occuring at the same time. Locating our significant predictors can be seen in The logistic regression model and the random forest model. The top three significant predictors are the same in both models being serum_creatinine, age, and ejection_fraction. Now having an idea of the correct models to apply as well as knowing the significant predictors in the model, DEATH_EVENT may be predicted more precisely and patients' lives can be saved.



Model	Logistic Model 1	Logistic Model 2	LDA	QDA	Unpruned Decision Tree	Pruned Decision Tree	Random Forest
Sensitivity	0.6667	0.7	0.6190	0.4286	0.5714	0.5238	0.4762
Accuracy	0.7143	0.7333	0.6833	0.7	0.7167	0.7167	0.75

Model	Logistic Model 1	Logistic Model 2	LDA	QDA	Unpruned Decision Tree	Pruned Decision Tree	Random Forest
AUC	0.7399	0.7314	0.7338	0.6935	0.7515	0.7735	0.7753

Fig. 6