yg2834 Yucong Gao

Prediction of the Survivorship for Patients with Heart Failure

Introduction

Heart failure is a very serious disease globally. About 6.2 million adults in the US have heart failure and 379,800 deaths were associated with heart failure in 2018[1]. In this analysis, I use 299 records of patients who had heart failure and applied classification algorithms to predict the survivorship using 12 feature predictors. In doing so, we may understand what medical feature should be closely monitored after a patient had heart failure and what treatment can be used to reduce the probability of death.

The dataset is published in 2020 and is from the UCL Machine Learning Repository which contains 299 observations and 13 features. Among the 13 features, 7 are continuous variables and 6 are discrete (See Appendix table 1.1 for detailed variable description). This dataset is quite cleaned and do not contain missing value. So, I transformed the continuous and discrete predictors into double and factor type respectively.

Exploratory Data Analysis

For continuous variables, I used density plot to graph on which features, patients who dead and patients who were not have different density distributions (Figure 2.1). For discrete variables, I calculated odds ratio of death after heart failure whether exposed to the risk factor or not (Table 2.2). Among the density plot (Figure 2.1) of continuous variable in the two groups, patients who died had a relative high serum creatinine level, lower serum sodium level, lower ejection fraction level and are older. In terms of discrete variables, the odds of death among heart failure patients who had anaemia is 1.33 times compared heart failure patients without anaemia. And the odds of death among heart failure patients with high blood pressure is 1.42 times of that for heart failure patients without high blood pressure.

In addition, I wanted to study if the level of continuous features of patients changed over time among the patients who died and who did not. I used scatter plots to display the relationships between time and variables of serum creatinine level, serum sodium level, creatinine phosphokinase level, ejection fraction level as well as platelets level (Figure 2.3). As a result, there is no specific trend can be concluded.

Models

Methods In this section, different binary classification prediction models are built, and predictors are age, anaemia, creatinine phosphokinase level, diabetes, ejection fraction level, high blood pressure, platelets level, serum creatinine level, serum sodium level, sex, and smoking status. And I divide the whole dataset into training and test dataset (3:1). In this part, I assume that there is no interaction between follow-up time and other variables. Therefore, Logistic regression, penalized logistic regression, GAM, MARS, LDA and Naïve Bayes model are constructed to predict the survivorship of heart failure patients. In addition, a repeated cross-validation is used to choose tuning parameters and best model. On the other hand, I evaluate the performances of these models though confusion matrix and ROC curve.

Outcome Table 3.1 shows the classification models test performance and Figure 3.2 shows the cross-validation result. According to the result, MARS achieves the best cross-validation performance which has mean AUC 0.79. By comparing the AUC of cross-validation in figure 3.2, the MARS model also provides a flexible prediction as the variance of AUC is relatively small. With respect to test set performance, the MARS model also gives the best prediction performance on test set as the AUC achieves 0.925, the Kappa equals 0.511 and the accuracy equals 0.797. However, the GAM model can better predict the death as the sensitivity of 0.625 is the highest among these models. And all these models have an equal performance of predicting survivor class.

Feature Importance To further identify the most risk factors of death for heart failure patients, I visualize the feature importance for the parametric models of logistic regression, penalized logistic regression, MARS and GAM. The top three important features are level of ejection fraction, age, and level of serum creatinine. To be more specific, lower ejection fraction level, higher serum creatinine level and older patients are more associated with death.

Effect modification – Follow-up time To study whether follow-up time would modify the effect of the important risk factors, I divide patients into two groups who were followed less than 100 days, and who were followed more than 100 days. As the class imbalance for the second group, I resample and create a dataset which has 25 death cases and 25 survivor cases. Then two logistic

regressions are used to evaluate where risk factors have different impact on survivorship among different follow-up time group. The result shows that serum creatinine level is more important among patients who were followed more than 200 days which has a coefficient of 5.112 compared to 0.725 for patients followed less than 100 days.

Limitations These model are only based on a small dataset of 299 observations, a larger dataset can provide a reliable result. In addition, we have no information about geography, the same data collection should be conduct in multiple places as to provide generalizability. One more limitation is that we do not have more information to identify other risk factors of heart failure patients.

Conclusions

Through this study, we can conclude that the level of ejection fraction, patients' age as well as the level of serum creatinine are three most risk factors for survivorship of patients with heart failure. Older heart failure patients who have low ejection fraction level and high serum creatinine level are high-risk group for death. And this coincides with what we've seen in the exploratory data visualization part (Figure 2.1). Moreover, follow-up time do have a modification effect on risk factors. For patients followed up for less than 100 days, ejection fraction level, age and serum creatinine level are three most important and statistically significant risk factors. However, for patients followed for more than 100 days, serum creatinine is the only statistically significant risk factors.

Therefore, after patients' heart failure, ejection fraction level and serum creatinine level should be closely monitored, and corresponding treatment should be applied once the ejection fraction level and serum creatinine level become abnormal.

yg2834 Yucong Gao

Appendix

Table 1. Variable Description

Variable Name	**Type**	**Description**
age	continuous	age of the patient (years)
anaemia	factor	decrease of red blood cells or hemoglobin
creatinine_phosphokinase	continuous	CPK enzyme in the blood (mcg/L)
diabetes	factor	if the patient has diabetes
ejection_fraction	continuous	percentage of blood leaving the heart at each contraction
high_blood_pressure	factor	if the patient has high bp
platelets	continuous	platelets in the blood (kiloplatelets/mL)
serum_creatinine	continuous	serum creatinine in the blood (mg/dL)
serum_sodium	continuous	serum sodium in the blood (mEq/L)
sex	factor	1 for male, 0 for female
smoking	factor	if the patient smoks or not
time	continuous	follow-up days
death_event	factor	1 dead, 0 not dead

Figure 2.1 Density Plot on Continuous Predictors

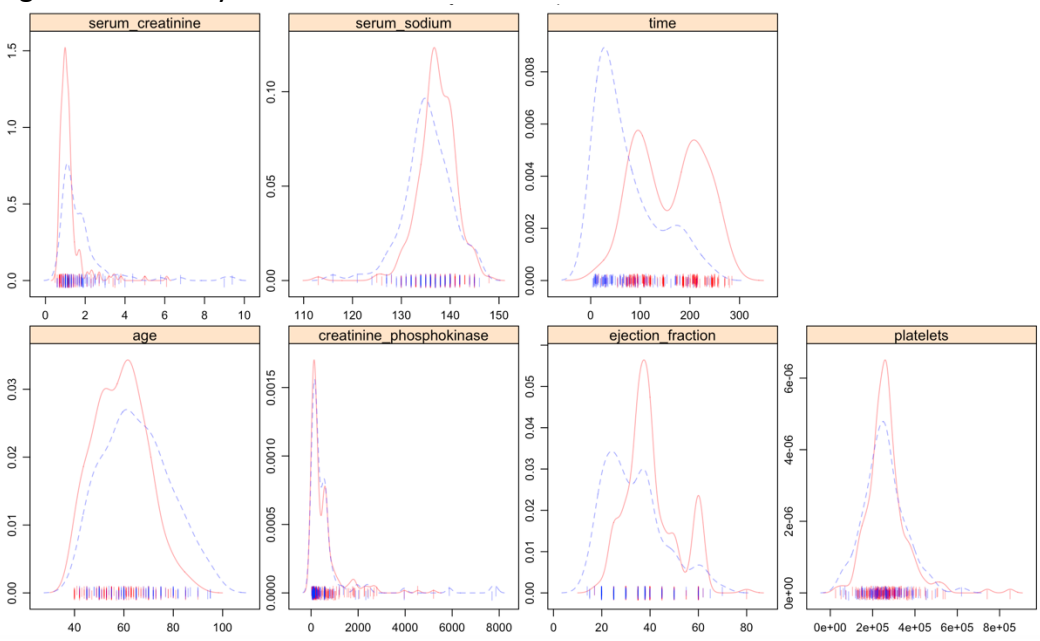

Table 2.2 Odds Ratio for Discrete Variables

Variable Name	**Odds**
anaemia	1.33
diabetes	0.992
high bp	1.42
sex	0.981
smoking	0.94

Figure 2.3 Distributions of Risk Factors by Time

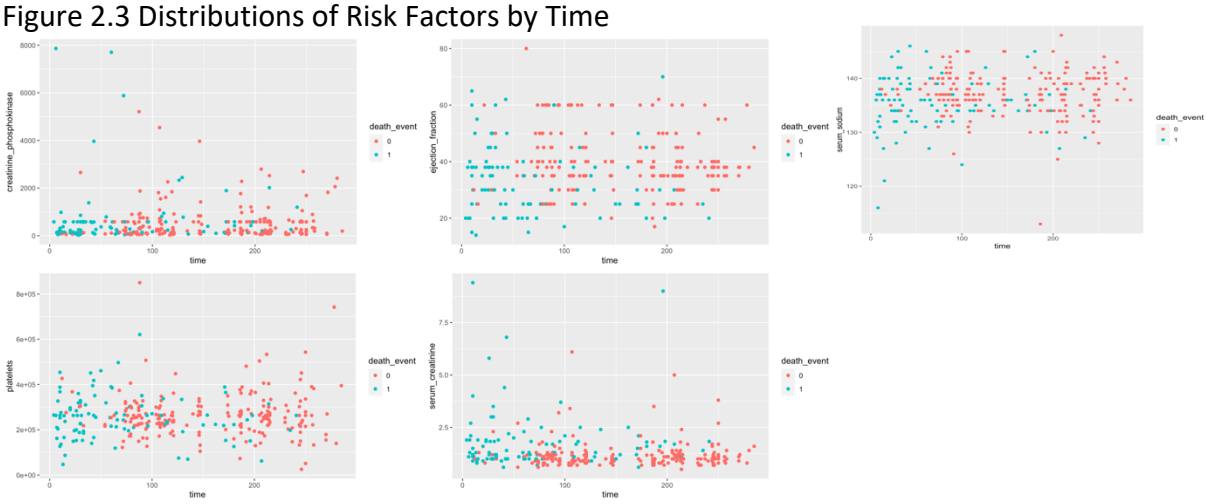

Table 3.1 Model Performances on Test set

Model	**Best Tuning Parameter**	**Accuracy**	**Kappa**	**AUC**	**Sensitivity**	**Specificity**
Logistic	/	0.784	0.46	0.883	0.5	0.92
Penalized Logistic	alpha = 0.4, lambda = 0.106	0.7432	0.3	0.892	0.29	0.96
MARS	nprune = 5, degree = 2	0.797	0.511	0.925	0.58	0.9
GAM	/	0.824	0.576	0.907	0.625	0.92
LDA	/	0.77	0.42	0.881	0.46	0.92
Naïve Bayes	usekernel = True, adjust = 1	0.743	0.335	0.83	0.38	0.92

yg2834 Yucong Gao

Figure 3.2 Cross-validation result