MATH52115 Classification Coursework

CIS Username: bgtt65

All the results in the report can be generated using the R code in following link:
https://github.com/lifewonderer/classification-coursework

Part 1: Executive Summary

Cardiovascular disease (CVD) is one of the most common causes of death claiming 17.9 million lives each year in the world, and it ranks among the top causes of death in the world except Africa. CVD is caused by atherosclerosis of the coronary arteries, and associated risk factors include age, family history, smoking, high blood glucose, high blood pressure, obesity, dyslipidemia, and so on. According to various studies, diet control and lifestyle have a great impact on CVD. By adjusting diet and lifestyle, cardiovascular-related diseases can be prevented.

Heart failure is a classic event caused by CVD, it occurs when the heart cannot pump enough blood to maintain the needs of the body. In addition, it is a common, costly, and potentially fatal disease that is difficult to be diagnosed in mild conditions. Meanwhile, it is unpredictable that most people remain stable for many years, while in some cases it may get worse quickly. Fortunately, mortality rates have declined year by year due to advances in treatment over the past three decades. Therefore, if a potential patient can be predicted or found in advance, the severe condition may be avoided.

The heart failure dataset contains 13 features that can be used to predict a possible fatal myocardial infarction. Fig. 1 shows the proportion of heart failure based on other potential features, including sex, anaemia, diabetes, high blood pressure (HBP), and smoking. Each feature may have a contribution to heart failure, the idea here is to combine all the known knowledge then have a prediction about a patient's heart condition that can help the doctor provides the early treatment.

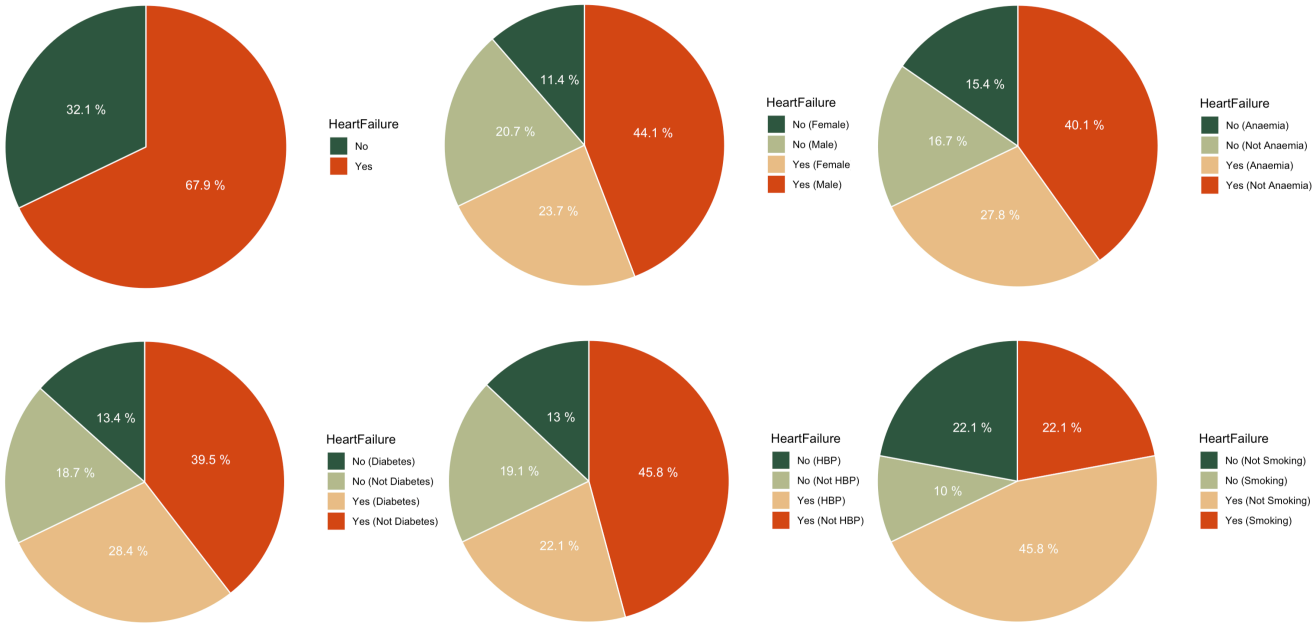

Figure 1: The relationship between heart failure and the other features

To achieve this goal, using the methods from machine learning to train and test the heart failure dataset. Then, the model can utilise new patient data to predict the possibility of fatal myocardial infarction. There are many algorithms were applied in this study that has generated a series of models. After the performance comparison, the final chosen model can reach 84% accuracy in prediction. It also can be expected that the accuracy will go up if some reliable data is added in the future.

Part 2: Technical Summary

1 Problem Description

There are some features in this dataset that we use to model the heart failure probability, including target variable and other binary and numeric variables. The information on these features is shown in Table 1. The aim of this problem is to fit the model using data of 299 patients and predict whether a patient will suffer a fatal myocardial infarction. Implementations of the fitting models are described in this report. With a series of benchmarking, the performance of the models is compared and discussed.

Target variable		**Numeric variables**	
fatal_mi	Patient suffered a fatal myocardial infarction = 1, otherwise 0	age	Age of patient
		creatinine_phosphokinase	Blood concentration of enzyme CPK (mcg/L)
Binary variables		ejection_fraction	Proportion of blood leaving the heart on each contraction (%)
anaemia	Diagnosis of conditions Yes = 1, No = 0	platelets	Blood concentration of platelets (kiloplatelets/mL)
diabetes		serum_creatinine	Blood concentration of serum creatinine (mg/dL)
high_blood_pressure		serum_sodium	Blood concentration of serum sodium (mEq/L)
sex	Male =1, Female = 0	time	Follow-up period (days)
smoking	Smoker = 1, otherwise 0		

Table 1: Features in heart failure dataset

The top left plot in Fig. 2 shows that the dataset is imbalanced, the number of patients who suffered a heart failure is only one-third of the total number of patients. Looking at other plots in Fig. 2, the proportion of the target variable is almost the same in two opposite cases of diabetes, sex, and smoking, which means these three features might not directly influence the death event. On the other hand, when a patient is diagnosed with diabetes or high blood pressure, the proportion of fatal myocardial infarction cases is more significant than those who haven't been diagnosed.

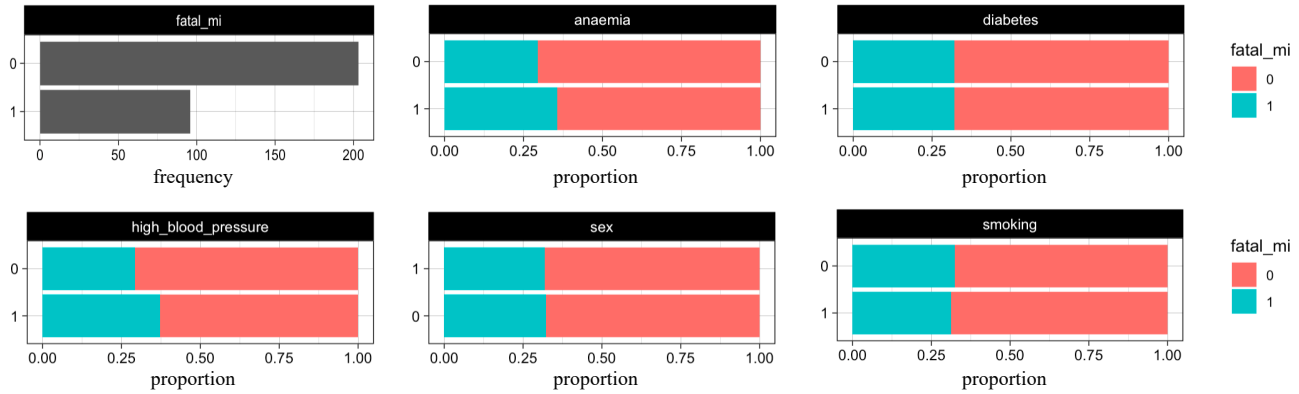

Figure 2: The bar plots for target variable and binary variables

As for Fig. 3, the box plots display the distribution of features. Take significant variables for example, the possibility of heart failure occurrence becomes much higher when patients with old age, low ejection_fraction or small time. In addition, more outliers in creatinine_phosphokinase, platelets and serum_creatinine are observed, the use of these features should be carefully considered cause extreme values may influence the prediction results.

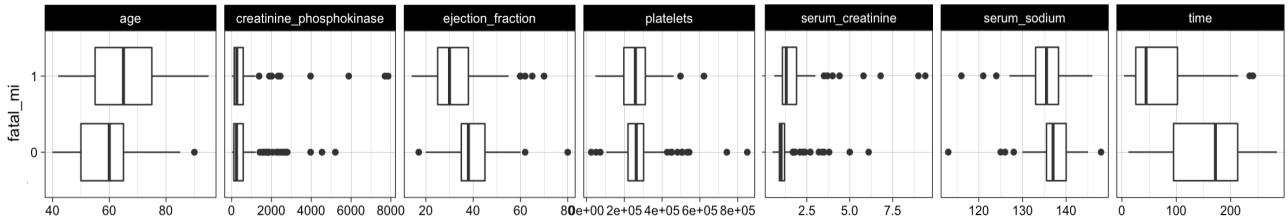

Figure 3: The box plots, numeric variables grouped by target variable

2 Model Fitting

Several algorithms were chosen to compare the performance of models in different strategies (train/test split, cross-validation, and bootstrap), including featureless classification, logistic regression, classification and regression trees (CART), and extreme gradient boosting (XGBoost). Except for featureless classification, which only analyses the target variable during train to set a baseline, the model training procedure has utilized all the features mentioned in the previous chapter. As for the loss function, the mean mis-classification error (MMCE) was employed.

2.1 Loss results

The setting for the strategies was considered. In the train/test split strategy, the dataset was split into a training set (70% data) and a test set (30% data) for model fitting and prediction. Following the usual rule-of-thumb, the number of folds for cross-validation (CV) was chosen as 5, then 5 portions of the dataset were used to train and test models on iterations. As for the bootstrap, it resamples a dataset with replacement, the repetition was set as 30 to fit the model. To observe the model performance, loss results for model fitting are presented in Table 2.

Algorithm	**Mean mis-classification error (MMCE)**		
Strategy:	Train/test split	Cross-validation	Bootstrap
Featureless classification	0.2666667	0.3210734	0.3278003
Logistic regression	0.1888889	0.1703390	0.1942986
CART	0.1777778	0.2005085	0.1993001
XGBoost	0.2222222	0.2138983	-

Table 2: Loss results for different models in model fitting

2.2 Model Comparison

In Table 2, all the loss results are less than baseline (set by featureless classification), which means the model fitting is credible. Looking at the performance of three strategies, the MMCE of featureless classification in cross-validation and bootstrap are nearly one-third, which prove the observation of the target variable in data exploration. However, the error for train/test split is much lower than expected because the number of data is small, the results are varied depending on different random sampling. Therefore, train/test split is not reliable in this case. As for the behaviour of algorithms, the powerful XGBoost implementation doesn't perform well as the small size dataset might cause the overfitting, whereas the logistic regression and CART show the good performance that can be seen in Fig. 4. After the comparison, the logistic regression using cross-validation is the best model in the model fitting step according to its low value of the loss.

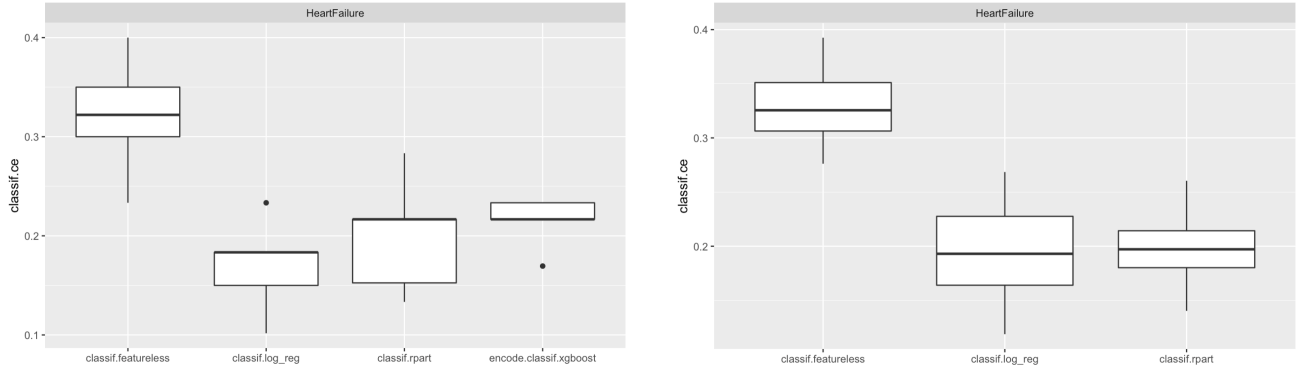

Figure 4: The box plots for CV (left) and bootstrap (right), MMCE grouped by different algorithms

3 Model Improvements

From the results of the model fitting, it can be discovered that the train/test split strategy and XGBoost algorithm is not suitable for this small size dataset. Therefore, only logistic regression classification and CART algorithms and the other two strategies (CV and bootstrap) were considered in this step.

3.1 Feature Selection

By reducing the number of variables in the dataset, feature selection decreases the cost of modelling and may improve the performance of the model. Through the wrapper methods and the observation from the data exploration, two variables (creatinine_phosphokinase and platelets) were removed. After the computation, loss area under the ROC curve (AUC) results are recorded in Table 3. Compared with Table 2, MMCE of four models are reduced, and the logistic regression using CV keep the lowest loss.

Algorithm	**MMCE**		**AUC**	
Strategy:	Cross-validation	Bootstrap	Cross-validation	Bootstrap
Logistic regression	0.1636158	0.1878416	0.8729567	0.8584077
CART	0.1871751	0.1940994	0.8219099	0.8301719

Table 3: Loss results for different models after feature selection

3.2 Tuning parameter

In this step, only CV strategy is discussed because of its good performance after feature selection. The left plot in Fig. 5 displays one of the tree models in CV. To select the cost penalty for CART algorithm, the nested cross-validation was applied. One of the results is shown in the right plot in Fig. 5. Through the comparison, the cost penalty was chosen as 0.15. Using the new parameter, Table 4 shows the loss, area under the ROC curve (AUC), false positive rate (FPR) and false negative rate (FNR) results after tuning CART model. It can be observed that the MMCE of CART is reduced and reaches the same value as that of logistic regression. However, compared with the two models, the AUC of logistic regression is much higher which means this model has better performance at distinguishing between the positive and negative classes. There is no doubt that the logistic regression using cross-validation is still the best model after model improvement, therefore, it is chosen as the final model.

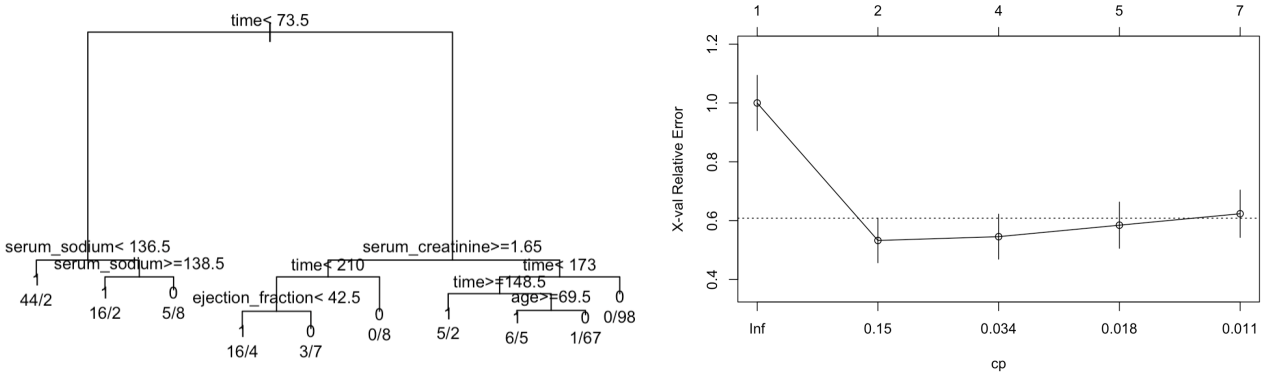

Figure 5: The tree model (left) and nested cross-validation result (right)

Algorithms	**Strategy**	**MMCE**	**AUC**	**FPR**	**FNR**
Logistic regression	Cross-validation	0.1636158	0.8729567	0.09276265	0.3141186
CART		0.1636158	0.7796564	0.06428784	0.3763993

Table 4: Computation results for different models after tuning parameter

4 Performance Report

From the model fitting and model improvements procedure, there are some significant phenomena can be observed from the performance of models. These issues are discussed in this chapter.

4.1 Model Performance Comparison

Not only the loss should be considered in model selection, but there are also some important results that may make the model more reliable. From the model fitting, feature selection to tuning parameter step, the loss results keep going down. In Fig. 6, the ROC curves in model fitting and feature selection show a similar trend which means the models still have almost the same ability to distinguish between classes without two variables However, compare two aqua curves in the mid and right plot, the latter does not perform well as the sensitivity is much lower than the former. Although the CART after tuning with quite low loss, the prediction is not clear (for example, predicting 0 classes under 0.5 but not near 0).

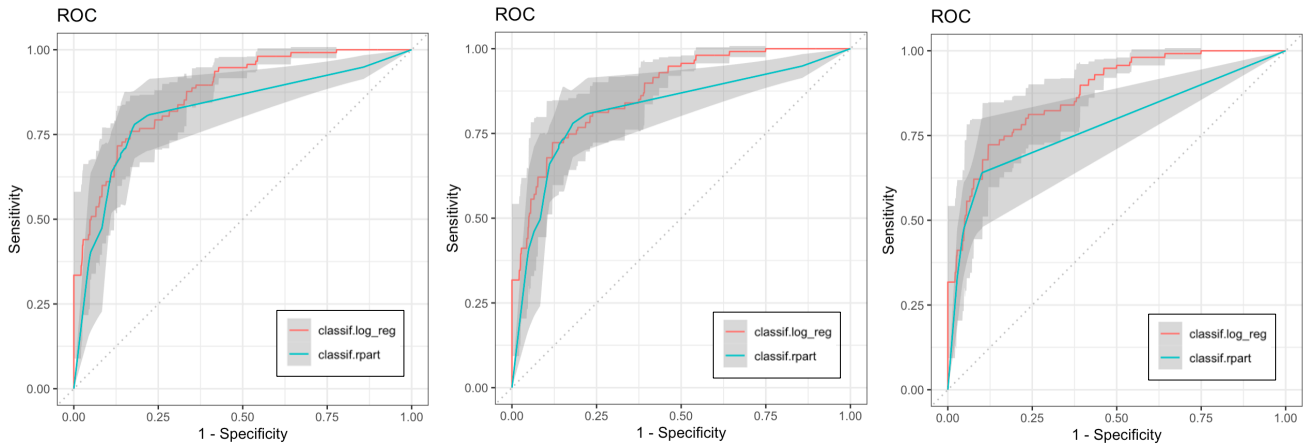

Figure 6: ROC curve in model fitting (left), feature selection (mid) and tuning CART (left)

4.2 Final Model Performance

Figure 7 shows some performance plots for final model. See the confusion matrix plot, 9.4% patients who haven't suffer a fatal myocardial infarction is predicted in positive condition, whereas 31.2% patients who have gotten heart failure is predicted in negative condition. In this medical dataset, reducing the false negative rate is more important that can find more possible patients who may experience heart failure, then the treatment can be provided in advance. Therefore, except for the loss, searching minimum FNR should also be considered in the objective function to find an optimum model.

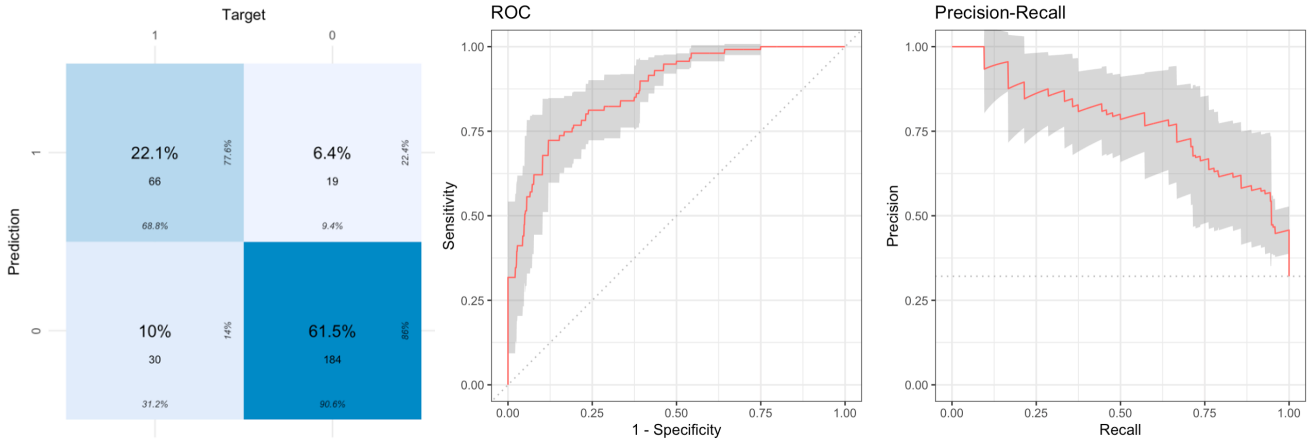

Figure 7: Confusion matrix (left), ROC curve (mid) and precision-recall curve (left) for final model