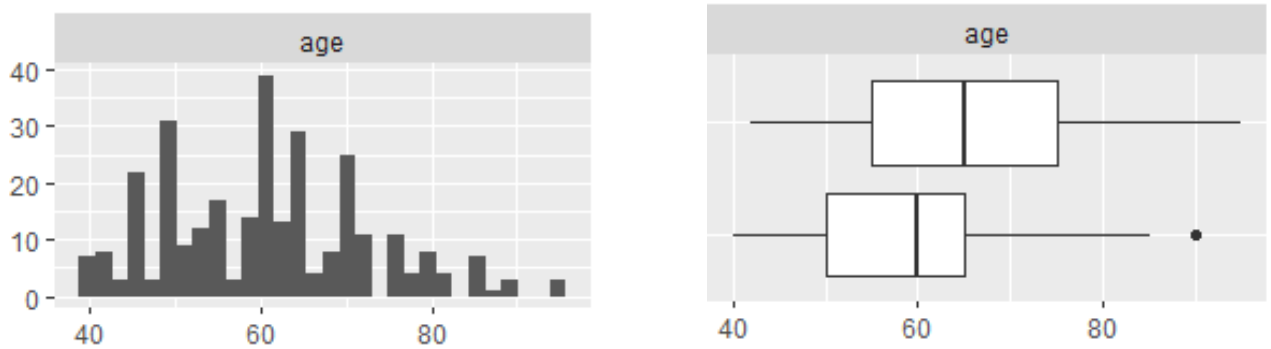
Student code nnxq47

Executive Summary

Myocardial infarction (more commonly known as a 'heart attack') is one of the leading causes of death in the United Kingdom, Europe and the United States. Fatal myocardial infarctions have direct impacts on the patient and their loved ones, as well as the wider healthcare system. By carefully monitoring patients at a high risk of suffering from a fata myocardial infarction, treatments may be offered earlier to prevent an at-risk patient from dying prematurely.

For this report, we investigated three different statistical models that were fitted using the records of 299 patients with 12 different attributes such as a patient's sex, age, and comorbidities such as diabetes and anaemia diagnoses. Unfortunately, approximately 32% of the patients ultimately suffered a fatal myocardial infarction. The average age of the patients in the dataset was around 60 years old – the upper box in the box plot below shows the distribution of the ages of patients who died as a result of cardiac arrest while the bottom box shows the distribution of surviving patients. The results show that older patients are more likely to suffer fata myocardial infarctions.

Age (years)	Anaemia (yes/no)	Creatinine phosphokinase (mcg/L)
Diabetes (yes/no)	Ejection fraction (%)	Hypertension (yes/no)
Platelets (kiloplatelets/mL)	Serum creatinine (mg/dL)	Serum sodium (mEq/L)
Sex (male/female)	Smoking status (yes/no)	Follow up period (days)



For any given patient's set of attributes in the dataset, the models were used to estimate the probability of whether the patient would suffer a fatal myocardial infarction before their next follow-up appointment, which was later used to make a prediction. To fit the models and test the accuracy of the predictions, the dataset was randomly divided into 5-subsets that were used to estimate errors by comparing the predictions to the real patient outcome.

How good are the models?

All three models made accurate predictions with overall accuracies above 80%, meaning that more than 80% of the predictions made by each model were ultimately correct. In the context of a clinical setting, a false positive result would indicate that a patient may receive follow-up appointments and treatment despite there not being a need to at the expense of healthcare services. In addition, invasive surgery may be needed to diagnose a heart condition, so false positive patients may also be at risk of surgical complications such as infections.

On the other hand, accurate predictions may lead to earlier interventions that improve a patients' outlook. As a result, false negative patients may not receive the necessary treatment, resulting in premature death. Consequently, reducing the number of false negative results in a reduction in the overall accuracy of the models. Healthcare professionals must take this into consideration selecting models to use as predictive tools for patients.

Student code nnxq47

Technical Summary

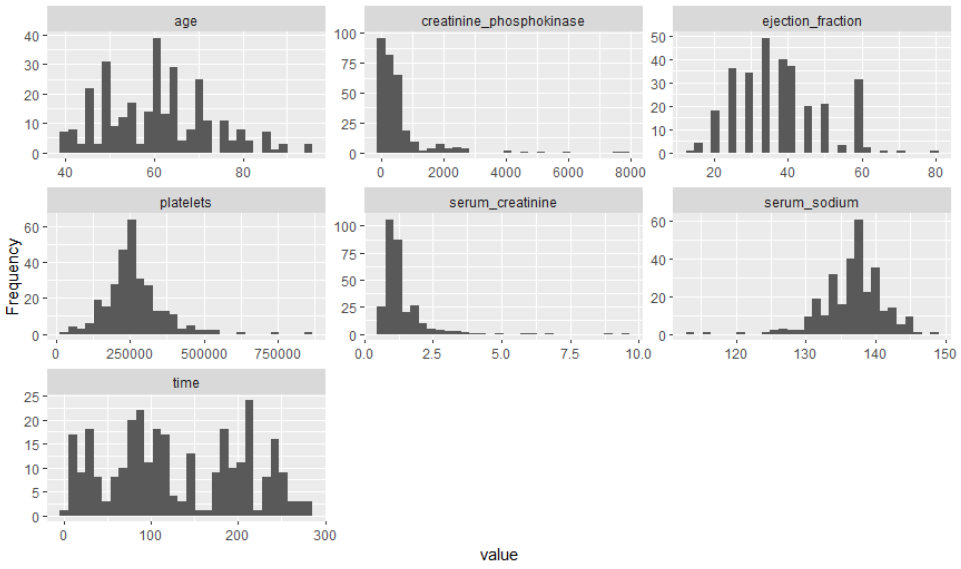
Problem Description and Data Summary

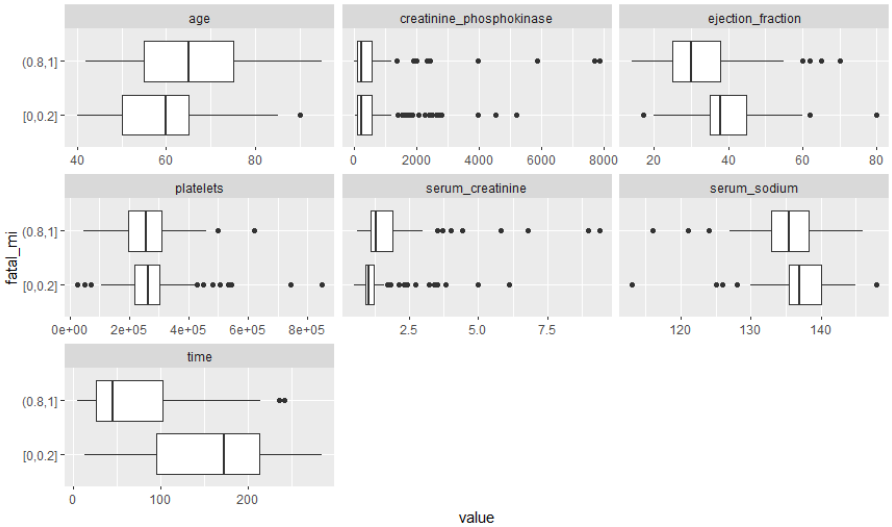
Effective decision-making processes are vital in the medical sector to help manage risks, optimise treatment options and ultimately improve the outlook of patients. Cardiac arrest is a leading cause of death in many countries within the United Kingdom, Europe, and the United States but many of those deaths could be prevented with medical intervention. The risk of any given patient suffering a fatal myocardial infarction (MI) depends on a plethora of factors, many of which cannot be directly controlled by the patient themselves. Early preventative treatments may save the lives of patients at a high risk of suffering identifying patients who are at a high risk of MI.

It is our aim to train and evaluate a set of statistical learning models to a set of anonymous patient records to predict whether any given patient will suffer from a fatal myocardial infarction. The dataset contained 299 patient records with the status of the 12 attributes below. In addition, the 13th column showed whether the patient suffered fatal MI between a follow-up appointment. Of the 299 patients, 96 patients unfortunately died because of MI making up around 32% of the sample. Patients who died during the follow-up period have been labelled as positive result – this is merely a mathematical convention and not a general statement that any individual's death is deemed as a positive result beyond the context of the machine learning models.

Age (years)	Anaemia (yes/no)	Creatinine phosphokinase (mcg/L)
Diabetes (yes/no)	Ejection fraction (%)	Hypertension (yes/no)
Platelets (kiloplatelets/mL)	Serum creatinine (mg/dL)	Serum sodium (mEq/L)
Sex (male/female)	Smoking status (yes/no)	Follow up period (days)

The histograms and box-and-whisker plots below show the distribution of each continuous variables within the dataset.





Student code nnxq47

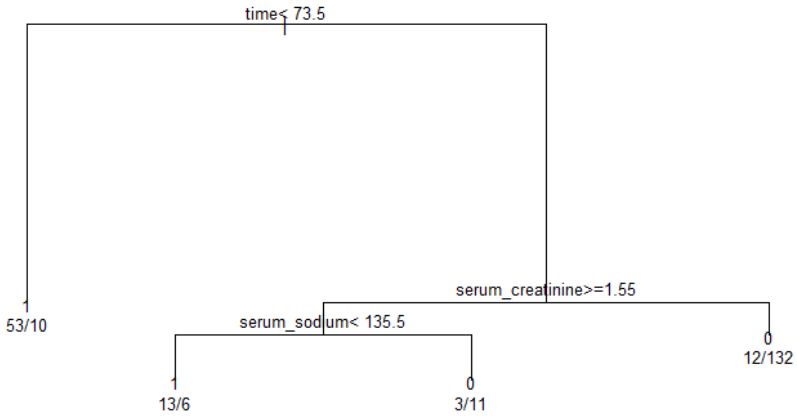
Model Fitting and Error Estimation

Three different classification models were investigated using the dataset: a logistic regression (LR) model; CP CART - a pruned classification and regression tree model with an optimised cost penalty; and a random forest (RF) model.

Cross-validation was selected for error estimation for the LR and CP CART models. A 5-fold cross-validation approach was chosen over a 10-fold cross-validation approach to balancing the respective impacts of bias and sample variance. Since the dataset used for the model contained 299 patient records with 13 variables, this approach to model selection and error estimation mitigated the effect of random false correlations between uncorrelated variables. Since the dataset was not particularly complex, all twelve features (excluding fatal myocardial infarction) were used to fit the logistic regression model.

Furthermore, we argue that higher variance associated with using nine moderately-correlated training sets and a decrease in the validity of the error estimate in each test set in each fold would negate the decrease in bias for a 10-fold approach in comparison to the 5-fold approach. Similarly, the number of samples required to train accurate models would leave too few samples remaining for precise error estimates. Ultimately, this would bias performance analysis with a sample size was too small for the true data distributions to be reflected by the empirical distribution of data.

In addition, a 5-fold nested cross-validation strategy was applied to select the parameter to prune the decision tree used in the CART model. The decision trees were pruned to reduce their complexity and variance while remaining close to the lowest mean error estimate between each fold. Since decision trees inherently suffer from large variances, further steps to reduce bias further using nested resampling. Furthermore, a random forest model was also used as an alternative to the CART model to reduce correlation between random variables and therefore reduce the variance.
The diagram below shows the CP CART decision tree, which was tuned and pruned at $\alpha = 0.02$.



Model Performance and Comparisons

The final model must be suited to aiding healthcare professionals in a clinical setting. As such the model accuracy, sensitivity, specificity and positive predictive value (PPV) after 5-fold cross validation were to choose evaluate and compare the three models.

Cardiac arrest and heart disease is a common cause of death in the UK and Europe. In comparison to other diseases that potentially result in death. for example, symptomless testing in a pandemic can waste resources. The repercussions of any given prediction being a false negative or a false positive result can vary tremendously in severity and wider impact. Ultimately, a false negative represents a preventable patient death – conversely false negative results lead to suboptimal allocations of treatment and resources and potentially even invasive surgical procedures.

The final model must be suited to aiding healthcare professionals in a clinical setting. As such the model accuracy, sensitivity, specificity, and positive predictive value (PPV) after 5-fold cross validation were to be chosen to evaluate and compare the three models. The data is presented in table 1

	Logistic Regression	CP CART	Random Forests
Accuracy	82.6%	80.9%	83.2%
Sensitivity	67.8%	71.8%	69.5%
Specificity	89.6%	85.1%	89.7%
PPV	78.4%	69.4%	76.1%
ROC AUC	87.4%	83.1%	91.9%
PR AUC	77.3%	68.7%	84.2%

Student code nnxq47

Accuracy

$$\frac{TP + TN}{TP + FN + TN + FP}$$

The logistic regression model almost matches the random forest model, whilst the CP CART model lags slightly behind in terms of overall accuracy after 5-fold cross validation.

Sensitivity

$$\frac{TP}{TP + FN}$$

A model with high sensitivities correctly predicts a high proportion of positive results out of the set of actual positive cases. As a result, sensitivity is helpful for allocating resources and treatments effectively. By contrast, the CP CART model had the highest sensitivity and thus provided fewer false negatives than the other two models, with the RF and LR models coming 2nd and 3rd respectively.

Specificity

$$\frac{TN}{TN + FP}$$

Models with high specificities are successful with correctly predicting negative results out of a set of actual positive cases. As a result, specificity is an important measure to prevent unnecessary treatments. There was little to separate the specificity of the LR and RF models, though both specificities showed a 5% improvement over the CP CART method.

Positive Predictive Value (PPV)

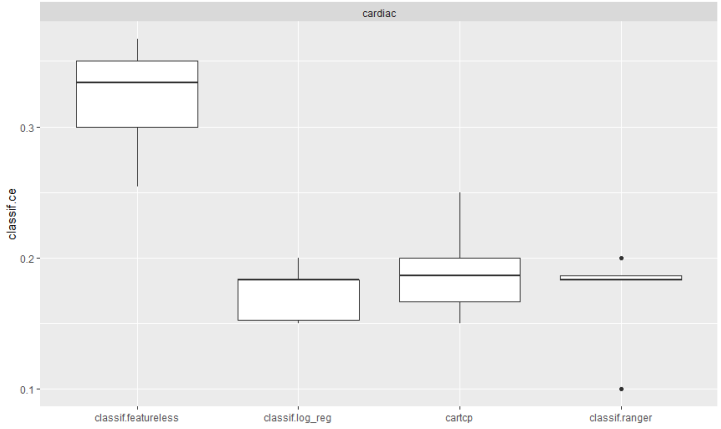
$$\frac{TP}{TP + FP}$$

PPV is helpful measure to determine how useful the classification model is in a clinical setting. A model with a high PPV will work well on large datasets with more variation in the characteristics of the patients by producing a small fraction of false positives in comparison to true positive results.

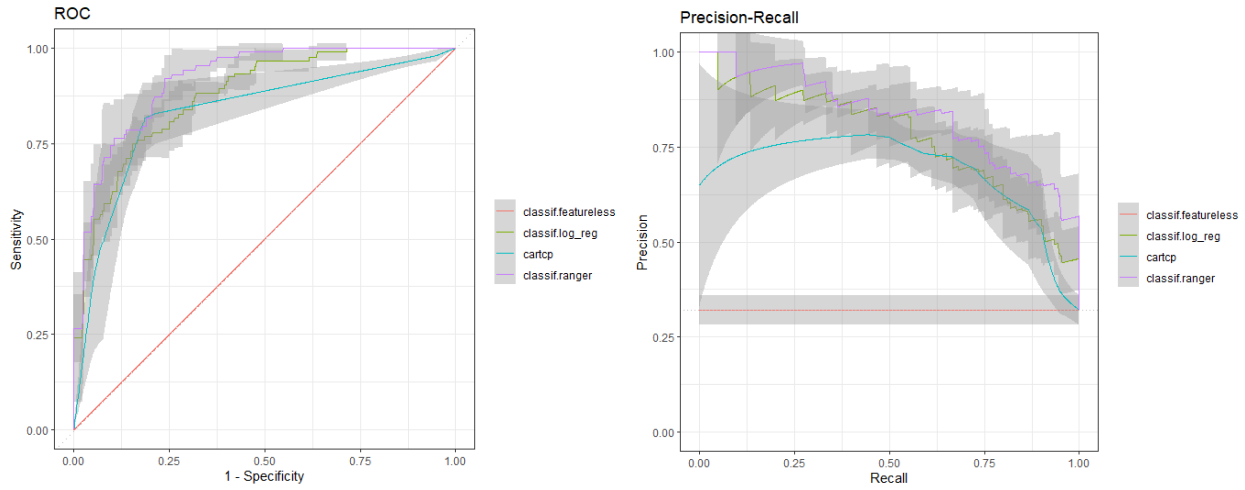
The LR model won gold in the positive predictive value, with RF winning the silver and CP CART earning bronze. PPV is a very important statistic in the context of the dataset and the classification objective since false positive results may lead to further procedures each with their own set of potential repercussions and consequences to the healthcare system and patient.

Each of the three approaches taken can be tuned further to reduce the percentage of either false negative or false positive results at a cost of decreasing the accuracy of a model. The area under the receiving operating characteristic curve is referred to as the ROC AUC whilst the area beneath the precision-recall curve is referred to as PR AUC. The ROC AUC and PR AUC values indicate the performance of each model without an explicit dependence on the false positive and false negative rates. The RF model came out top with LR and CP CART models coming second and third for each metric.

The box-and-whisker plot below shows the 5-fold cross-validation aggregate error estimate for the three models as well as the featureless baseline model that predicted the modal class. The plot demonstrates how closely matched each of the three classification models are in terms of making errors in addition to the reduction in variance between the CART CP and RF models.



Student code nnxq47



Final Model Selection and Improvements

As discussed previously, the most crucial factor to consider in the context of clinical health data is which option will minimise the negative repercussions of any decision. Given the context of the classification task and the importance of a medical professional using the model to predict and prevent deaths due to fatal myocardial infarction, the logistic regression model was selected because it is easy to alter the prediction threshold to reduce the false negative rate.

Since the dataset was moderately small in sample size and the LR model was fitted using 12 features - for further investigations, reducing the number of features used to fit the data where possible to negate the impact of potential randomly correlated features. It may also be beneficial to have the flexibility to tweak the number of features for different datasets and constrictions.

Although decision trees featured in the CART and random forest models are analogous to human decision making, it is difficult to draw meaningful conclusions from the decisions taken at each level of a tree which poses several potential problems within a clinical setting.

On one hand, non-experts may not be able to determine a suitable decision tree with relevant attributes with a true correlation to fatal myocardial infarction placed closer to the base. On the other hand, a cardiac surgeon may question whether serum creatinine level is a greater predictor of future cardiac illness than whether a patient suffers comorbidities such as anaemia or diabetes.

As discussed previously, the most important factor to consider in the context of clinical health data is which option will minimise the negative repercussions of any decision. The performance of the LR model and RF model were close and both outperformed – with LR outperforming RF for PPV. As such, the LR model is both robust, adaptable and the most likely to enhance the work of a healthcare professional.

Link to Github repository

https://github.com/dixondj1995/ASML-classification