Multiple Imputation and Cross-Validation for Classification of Survival Prediction

Robert Edwards

(2416963E)

MASTER THESIS

Biostatistics

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim of the Thesis

1.2 The Clinical Study

1.3 Study Population & Data Description

1.4 The Statistical Challenege

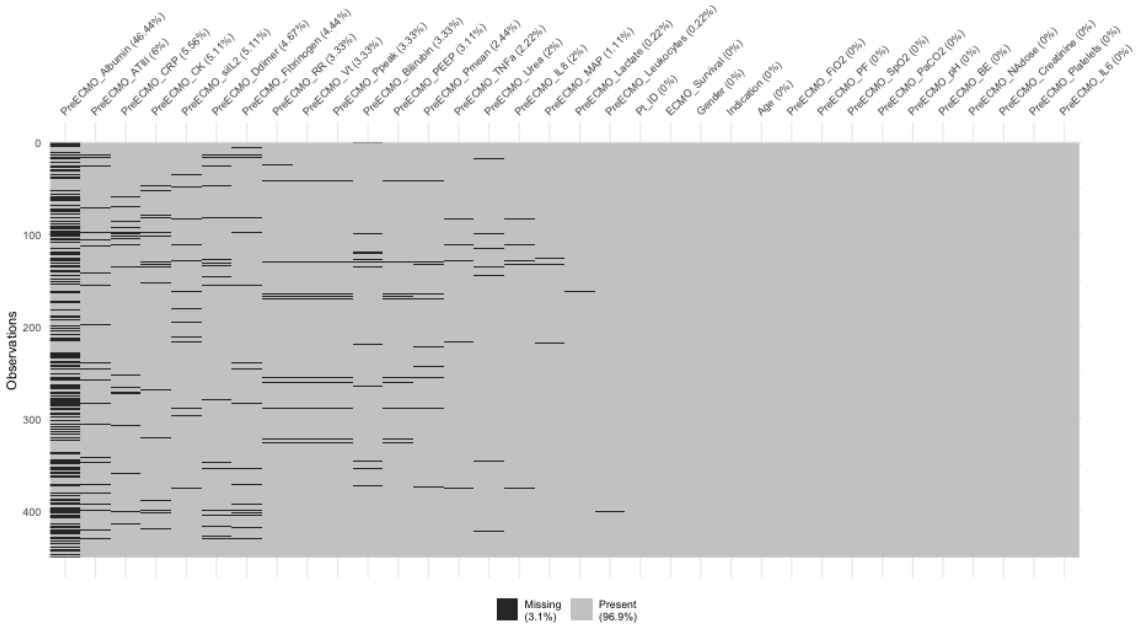

Figure 1: Visual representation of missing observations in the ARDS dataset.

2 Methodology

2.1 Basic Statistical Methods

2.1.1 Logistic Regression

Logistic regression is a widely used approach in binary classification. It is set up as a generalised linear model using a logit link that produces a probability.

2.1.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a widely used method classification method. generlization of Fisher's Linear Discriminant **(Fisher 1936)**. Ddiscriminant functions are created through a linear combination of the explanatory variables that characterize the classes.

$$\Pr(C_g \mid \mathbf{x}) = \frac{\pi_g \exp(-\frac{1}{2} d_g(\mathbf{x}))}{\sum_{i=1}^{2} \pi_g \exp(-\frac{1}{2} d_i(\mathbf{x}))} \qquad g = 1, 2$$

Assumptions of LDA:

- Explanatory variables are assumed to be normally distributed
- Homoskedasticity, equal class covariances
- No multicollinearity
- Independent observations

Drawbacks of LDA:

- Can only utilize continuous explanatory variables
- Cannot handle missing data

2.1.3 Quadratic Discriminant Analysis

(Cover 1965)

Quadratic Discriminant Analysis (QDA) is an even more generlized form of discriminant analysis than LDA. QDA has the same assumptions as LDA with the exception that the covariance of each class is not assumed to be identical.

Assumptions of QDA:

- Explanatory variables are assumed to be normally distributed
- No multicollinearity
- Independent observations

Drawbacks of QDA:

- Can only utilize continuous explanatory variables
- Cannot handle missing data

2.1.4 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a commonly used non-parametric classification method.

2.1.5 Random Forests

2.2 Missing Data

2.3 Multiple Imputation

2.4 Validation & Cross-validation

2.5 Accuracy Metrics

These are the default metrics used to evaluate algorithms on binary and multi-class classification datasets in caret.

2.5.1 Accuracy

Accuracy is the percentage of correctly classifies instances out of all instances. It is more useful on a binary classification than multi-class classification problems because it can be less clear exactly how the accuracy breaks down across those classes (e.g. you need to go deeper with a confusion matrix). Learn more about Accuracy here.

Don't use accuracy (or error rate) to evaluate your classifier! There are two significant problems with it. Accuracy applies a naive 0.50 threshold to decide between classes, and this is usually wrong when the classes are imbalanced. Second, classification accuracy is based on a simple count of the errors, and you should know more than this. You should know which classes are being confused and where (top end of scores, bottom end, throughout?)

Table 1: Confusion matrix for two classes.

	Y	N
Y	a	b
N	c	d

$$\text{accuracy} = \frac{a+d}{a+b+c+d}$$

2.5.2 ROC

2.5.3 Cohen's Kappa

Kappa or Cohen's Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes On the ARDS datasets, for example, if `ECMO_Survival` is predicted to be "Y" for all cases, then the accuracy is 75% but the prediction is no better than the baseline likelihood of the class percentages.

Let p_o be the accuracy, the relative observed agreement between observed and predicted classes and let p_e be the probability of chance agreement based on the class probabilities. Cohen's Kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

If all the observations are predicted correctly then $\kappa = 1$. **If the observations are predicted no better than expected by the class probabilities, p_e then $\kappa = 0$. If all the observations are predicted incorrectly, then $\kappa = -1$.** A positive κ indicates that the model predicts better than would be expected by chance whereas a negative κ indicates that the model predicts worse than would be expected by chance.

$$p_o = \frac{a+d}{a+b+c+d}$$

For class k, number of items N and n_{ki}, the number of times i is predicted as class k:

$$p_e = \sum_k \hat{p}_{k1}\hat{p}_{k2} = \sum_k \frac{n_{k1}}{N}\frac{n_{k2}}{N} = \frac{1}{N^2}\sum_k n_{k1}n_{k2}$$

2.5.4 Brier Score

2.5.5 F1 Score

3 Statistical Methods for the Analysis

Describe the methods step-by-step for the analysis

3.1 Complete Case Analysis

3.2 Mean Imputation

3.3 Multiple Imputation

3.3.1 Joint-Model

3.3.2 Fully Conditional Specification

3.3.3 Predictive Mean Matching

Predictive Mean Matching (PMM) is a semi-parametric imputation approach. It is similar to the regression method except that for each missing value, it fills in a value randomly from among the a observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model (Heitjan and Little 1991; Schenker and Taylor 1996). The PMM method ensures that imputed values are plausible; it might be more appropriate than the regression method (which assumes a joint multivariate normal distribution) if the normality assumption is violated (Horton and Lipsitz 2001, p. 246).

Ensemble Multiple Imputation
The steps in the ensemble approach for multiply imputed data in k-fold cross-validation are as follows:

1. Randomly partition the training data into k folds
2. Define the k^{th} as the test set and the remaining $k-1$ folds as the training set
3. Impute the training set m times, with the response variable `ECMO_Survival` included, to create m imputed training sets
4. Concatenate the m imputed training sets into one extended training set
5. A model is fitted to the extended training set
6. The test set is concatenated with the extended training set
7. Impute the combined test and extended training set, with the response variable `ECMO_Survival` excluded, to create m imputed combined test and extended training sets
8. Extract the m test sets
9. Make m predictions on the m imputed test sets
10. Take the majority vote of the m predictions as the prediction for the fitted model

11. Validate the prediction against the test set by calculating Cohen's Kappa (note there are no missing values for the response variable in the data)
12. Repeat steps 2-11 k times and validate the fitted model on each training set against the test set for each fold
13. Average the k calculated Cohen's Kappas as the estimated in-sample accuracy metric

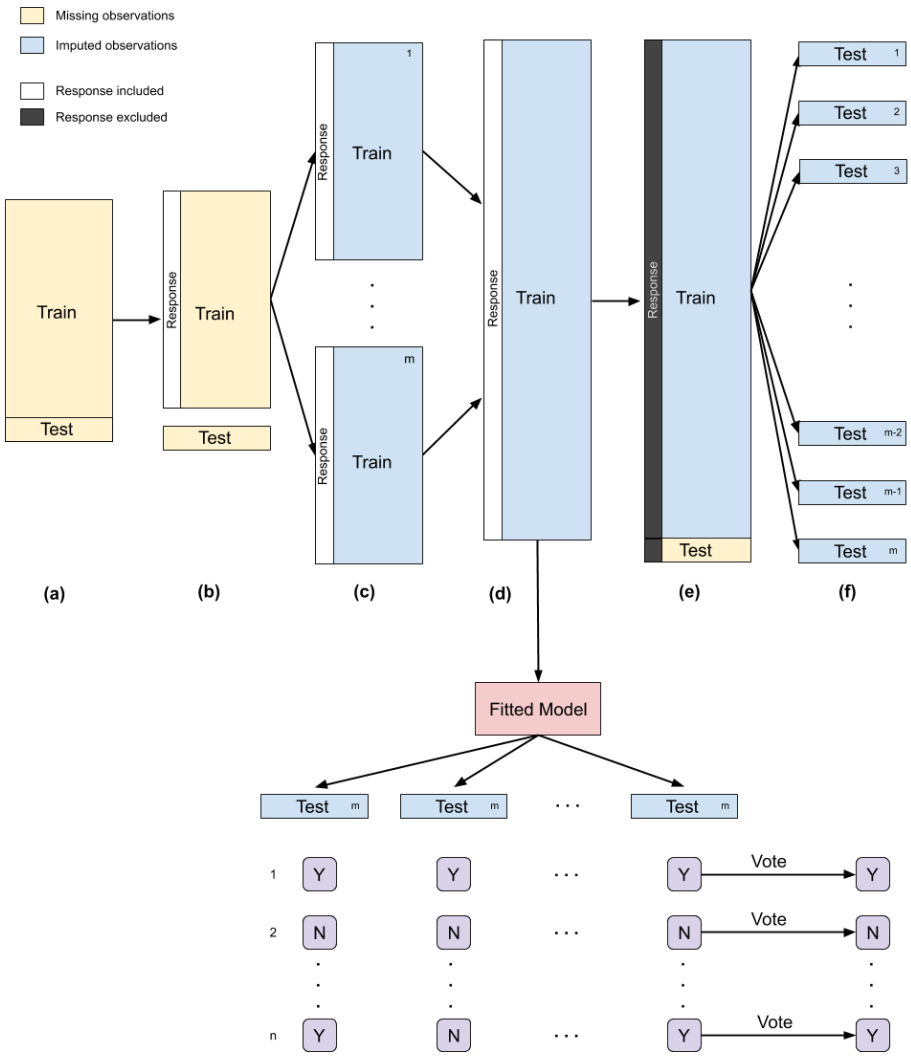

Figure 2: Outline of the algorithm used to pool predictions from multiple imputation. (a) Step 1. (b) Step 2. (c) Step 3. (d) Step 4. (e) Step 5. (f) Step 6.

4 Results

Table 2: Averaged Cohen's Kappa for each model fitted in cross-validation. The number of neighbors, K, for K-Nearest Neighbors is 15. The number of randomly selected variables at a split in Random Forests is 1.

	Logit	LDA	QDA	KNN	RF
Complete Case	0.083	0.163	NA	0.028	0.026
Mean	0.169	0.140	0.061	0.076	0.086
PMM	0.165	0.197	0.058	0.065	0.085

5 Discussion

6 Conclusion

7 Bibliography

8 Appendices

8.1 Additional Material

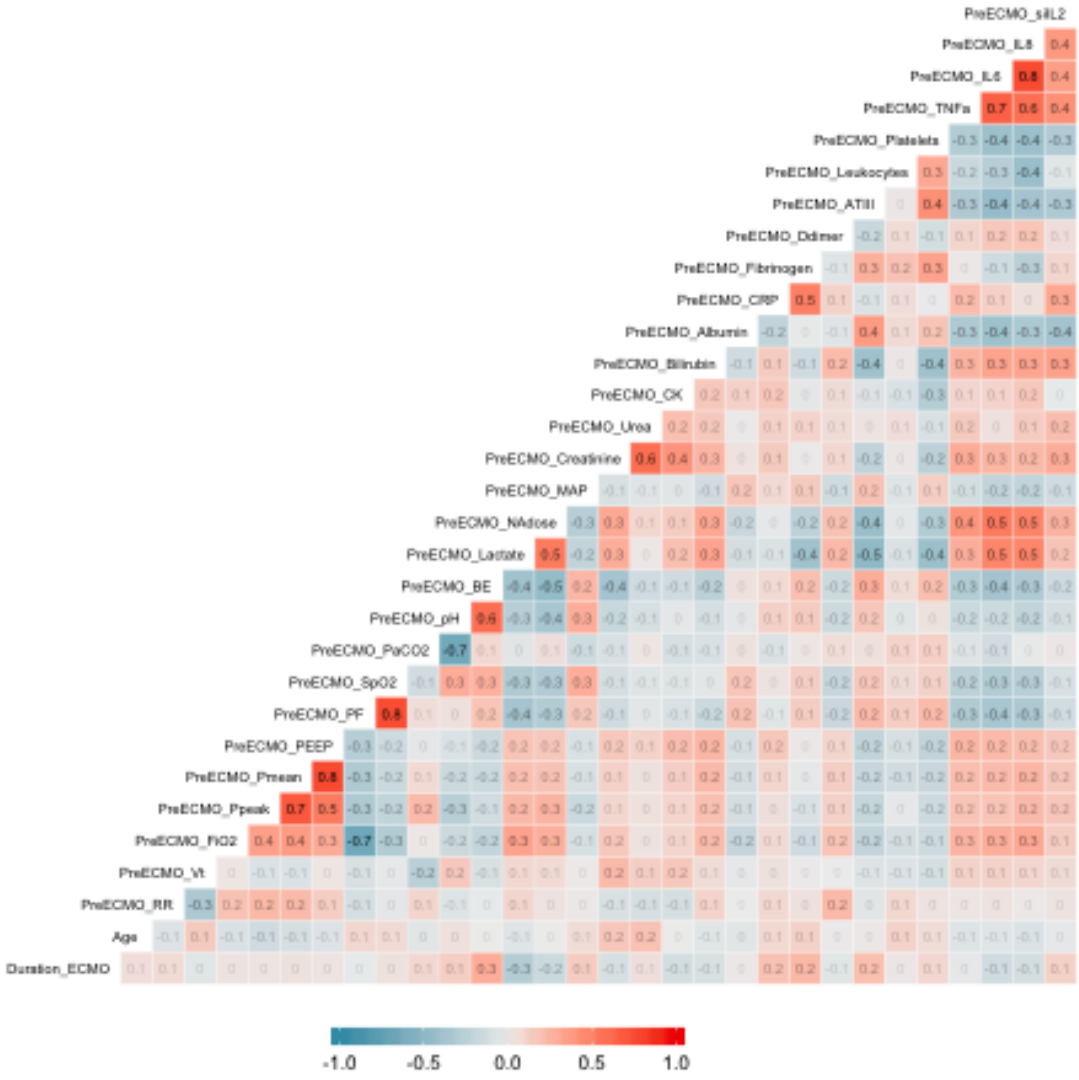

Figure 3: Heatmap of standardized and transformed variables.

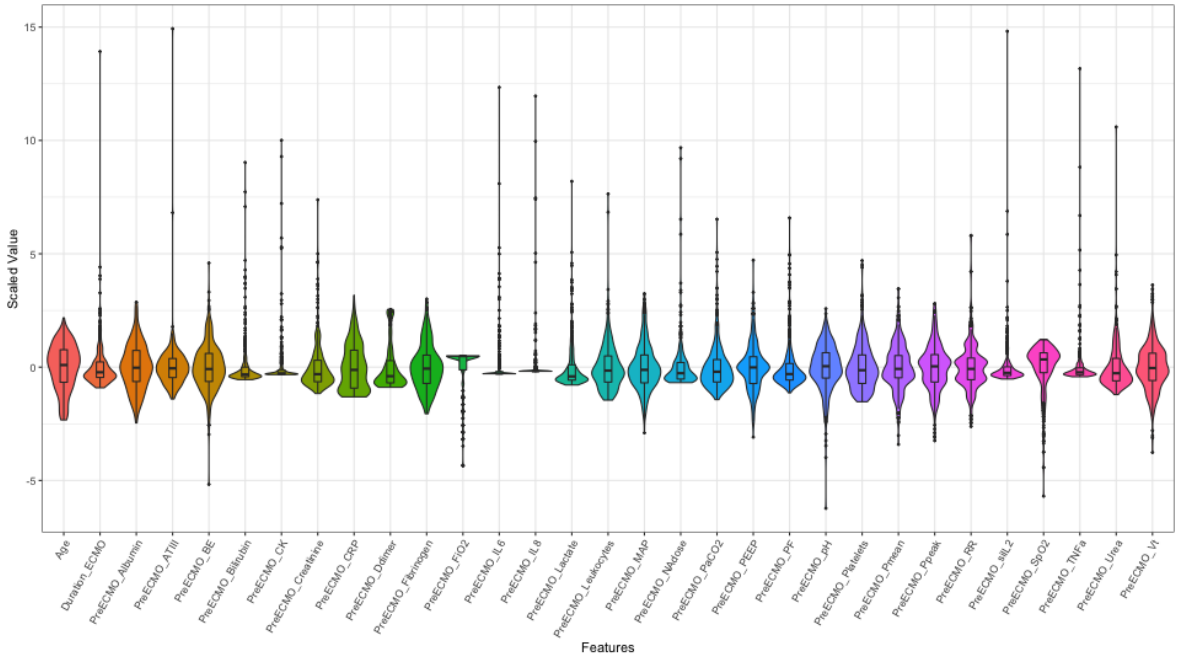

Figure 4: Violin plot of standardized variables.

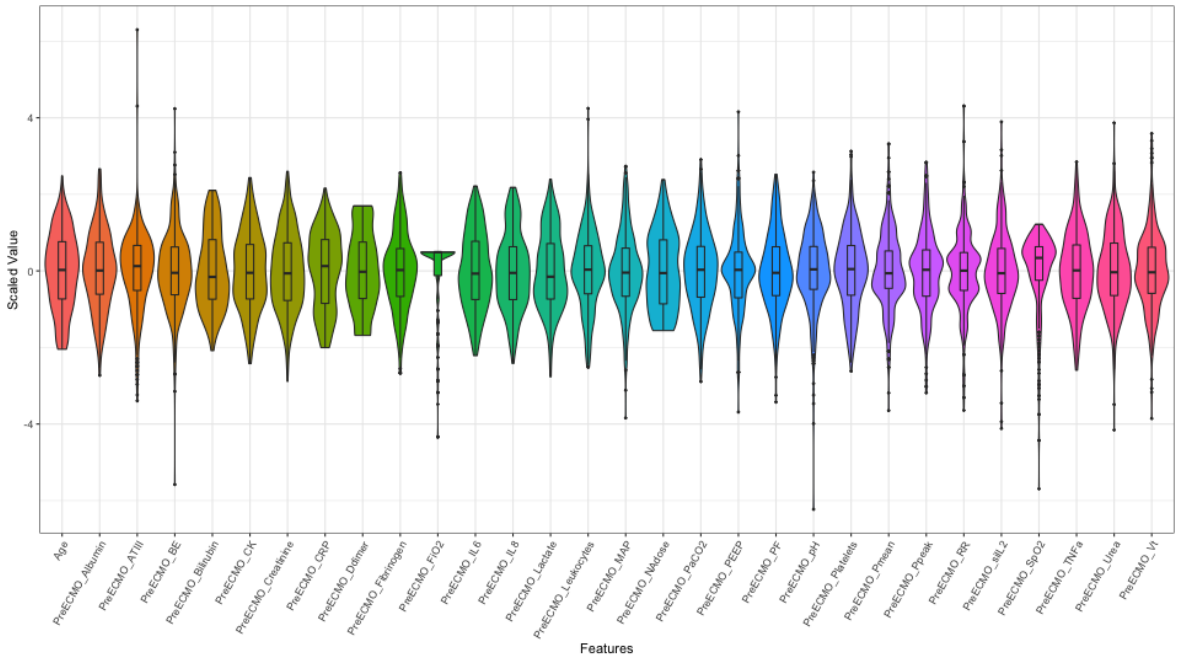

Figure 5: Violin plot of standardized and transformed variables.

8.2 R Code