Application of statistical analysis techniques learned in previous coursework.

Standardization

Table 1 shows that the predictors have vastly different scales or units. Creatinine phosphokinase has the largest mean and standard variation in the data set. Serum Creatinine has the lowest mean and standard deviationIts important to standardize the data with mean of 0 and standard deviation of 1 using z-score normalization. This process will be repeated for each predictor. This will prevent features with larger scales from dominating the learning process.

Feature Selection Techniques

We will be using several feature selection techniques to choose a subset of relevant and important features from our dataset. Feature selection will help prevent overfitting by reducing the complexity of the model. Unnecessary features might introduce noise or irrelevant information, which can degrade model performance. Irrelevant or noisy features can lead to overfitting, where the model captures patterns specific to the training data but fails to generalize to new data.

In our case we will be deploying a classification tree instead of a regression tree because our reponse variable is reported as a 0 or a 1. Feature importance in classification trees is typically calculated based on how much each feature contributes to the reduction in impurity using the Gini Index by summing up the impurity decreases for each node where a particular feature is used to split the data

A variance importance plot will be generated after deploying a Random Forest model. This plot will rank the features and show us the drop in the models accuracy as features are excluded. By examining the variable importance scores, you can decide to eliminate or retain certain features. Features with low importance scores may be candidates for removal if the goal is to simplify the model without significantly sacrificing predictive performance. The random forest will be fine tuned with different number of variables tried at each split.

We will plot the reduction in coefficients at different levels of lambda by a LASSO. You can access the coefficients to identify important features. Features with non-zero coefficients are considered selected by the Lasso. LASSO is preferred over ridge regression because LASSO forces some coefficients to 0 while ridge regression will only force them 'close' to 0. The removal of some variables will take care of any multicollinearity problems.

A z-test will be used to check the significance of the difference between the distributions. We will split the data into 2 groups one for 1 (death) and the other for 0 (survival) and calculate the marginal distribution of each feature in each group. Features with more negative Z-scores are less likely to fail and features with more positive z scores are more likely to fail.

Different methods may show that different variables are important. If I'm running a LASSO I'm not allowing interactions between the predictors but Random Forests and classification trees model interactions. Therefore, its natural for different methods to select different subsets of variables since they are allowing different behavior among the predictors. If these methods display conflicting results in terms of feature selection we will use the suggested features from each model and select the set of features that result in the lowest misclassification rate.

Predictive Techniques

We are using supervised machine learning approaches to to predict the death_event response variable. We will be utilizing a Decision Tree, Random Forest, Gradient boosting, k-nearest neighbors, negative binomial regression and GAMS.

Each technique will be k-fold cross validated to provide a more robust and reliable estimate of a model's performance compared to a single train-test split. It helps assess how well a model generalizes to new,

unseen data. Without cross-validation, the model evaluation might be highly dependent on the specific random choice of data points in the training and testing sets. Our data set is limited to only 299 rows, cross-validation allows for more efficient use of the available data. Instead of allocating a fixed portion for training and testing, cross-validation uses the entire data set for both training and testing in different folds, maximizing information extraction from the limited data. Cross validation also ensures we are not over fitting to the training data and helps generalize our model to new data. The misclassification rate for both the training and testing split will be reported. If they are very similar or identical it would mean that our model generalizes well to new data therefore we are not overfitting to the training data.

Classification trees are prone to over fitting therefore we will prune the tree using the misclassification rate with a cost complexity tuning parameter. The tree will be pruned to the respective number of terminal nodes using the lowest misclassification error generated from cross validation.

Cross validation won't be necessary for the Random Forest Model because out of bag error from bagged models can be compared to cross validated error from other models. Out of bag estimation is an efficient way for providing realistic long term error. The goal is to reduce the error rate by different the number of variables tried at each split.

A logistic regression will be deployed because our response variable is binary. This model uses the logistic (sigmoid) function to model the relationship between the predictor variables and the probability of the outcome being in the positive class. An S shaped curve will be plotted that compresses any real valued number to the range [0,1]. We can predict death if $\hat{y} > 0.5$ and survival otherwise.

Gradient boosting models are incredibly powerful and popular which is why we decided to include this technique. The distribution for this model will be binomial. The model will be built iteratively with 5000 trees but the model obtained at the iteration where cross-validation performance is optimal will be used as the final model.

In situations where the decision boundary exhibits a high degree of non-linearity, we have opted for the implementation of a non-parametric algorithm known as k-nearest neighbors (KNN). KNN will dominate logistic regression if the decision boundary is highly non-linear. Grid search will be utilized for the best value of k and the accuracy will be plotted and we will look for an 'elbow' in the plot.

GAMS

Every model will be compared using a confusion matrix, misclassification error, logloss, F1 score. F1 score will be heavily emphasized because there is an imbalance in the response variable in the data set.