Predicting Heart Failure in Medical Patients with Previous Casesof Heart Failure

Travis Froberg

Justin Park

MA 5790

October 2020

Abstract

Heart disease is the number one cause of death worldwide. Death occurs when a cardiovascular condition causes heart failure, which describes a situation where the heart cannot adequately pump blood to the rest of the body. This report uses predictive modeling on a dataset containing 12 predictors (both categorical and continuous) to predict mortality in patients who have had an instance of heart failure in the past; they had also been diagnosed with left ventricular systolic dysfunction prior to data collection. There were no degenerate, highly correlated predictors, or missing values in the data so no predictors were removed from the original dataset. Because the outcome was categorical, linear and nonlinear classification models were created. The area under the ROC curve statistic was used to determine the most effective models. The glmnet model and the FDA model were the best models on the training data. The LDA and PLSDA were the best models on the test set, with PLSDA being the best model.

Table of Contents

Background

Heart disease is the number 1 cause of death worldwide. Death occurs when a cardiovascular condition causes heart failure, which describes a situation where the heart cannot adequately pump blood to the rest of the body. Because of the prevalence and severeness of heart disease, researchers have long been trying to determine its causes. The dataset used for this study contained various predictors that are common health data collected while a patient is in a hospital. By understanding how these predictors affect a patients' likelihood of developing heart failure, clinicians may be in a better position to make informed decisions about a patient's care. Furthermore, scientists may perform more experiments to determine how important predictors are associated with heart failure.

The patients in this study all had a previous instance of heart failure prior to the study, making them a class 3 or 4 on the New York Heart Association (NYHA) classification of the stages of heart failure. Patients had also been previously diagnosed with left ventricular systolic dysfunction (LVSD) prior to the study. LSVD is a condition where the left ventricle of the heart cannot contract properly, making it unable to pump blood at normal levels into the aorta, which then carries blood to the rest of the body.

Variable Introduction and Definitions

The dataset used in this analysis contained 299 observations (i.e. medical patients); 105 were women and 194 were men. Twelve predictors were used to predict the outcome. The predictors were mostly health data that would normally be available for a patient with cardiovascular disease. A list of the predictors used and their definitions are given below:

Variable Name	Description
creatinine_phosphokinase (CPK) (continuous)	An enzyme released into the blood when muscle tissue gets damaged. Can indicate that muscles in the heart are damaged.
Ejection_fraction (continuous)	Percentage of blood the left ventricle pumps out with each contraction.
Serum_creatinine (continuous)	Waste product of creatine, which is created when muscle breaks doSerum wn.
Serum_sodium (continuous)	Amount of sodium in the blood. Low levels can be indicative of heart failure.
Diabetes (categorical)	1 is yes, 0 is no

Age (continuous)	
Anemia (categorical)	1 is yes, 0 is no
Sex (categorical)	
Smoking (categorical)	1 is yes, 0 is no
Time (continuous)	Days since initial appointment where data was collected. Range is 4-285
Platelets (continuous)	Platelet level in blood
High_blood_pressure	1 is yes, 0 is no

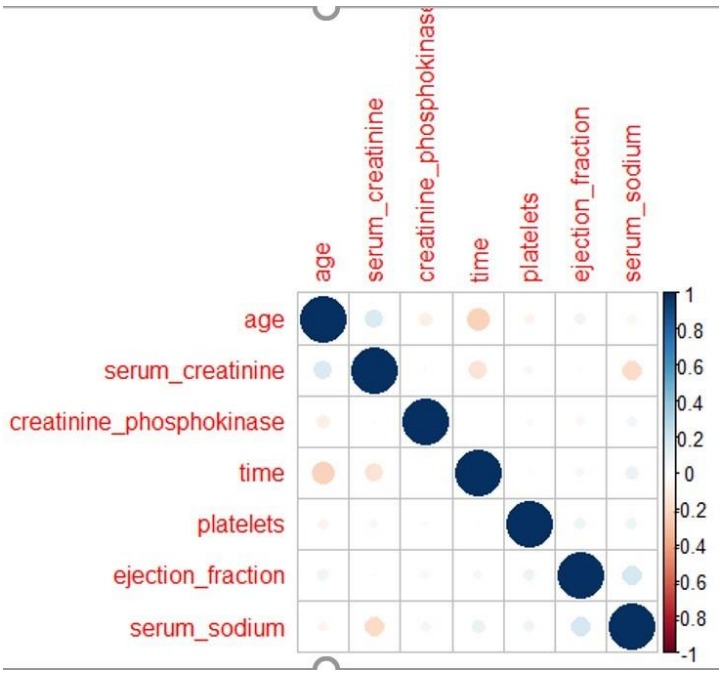
The response variable is DEATH_EVENT (i.e. mortality at the time of the follow-up); a value of 1 indicates mortality and a value of 0 indicates survival.

Preprocessing of the Predictors

Near zero variance analysis and predictors were correlated as data reduction techniques, however, there were no degenerate predictors and no highly correlated predictors (plot given below). There were also no missing values to impute. For these reasons, no predictors were removed from the data.

Dummy variables were created for all categorical predictors. Because all of them were binary, the creation of dummy variables did not increase the number of predictors in the dataset.

Correlation Plot for the 12 Predictors



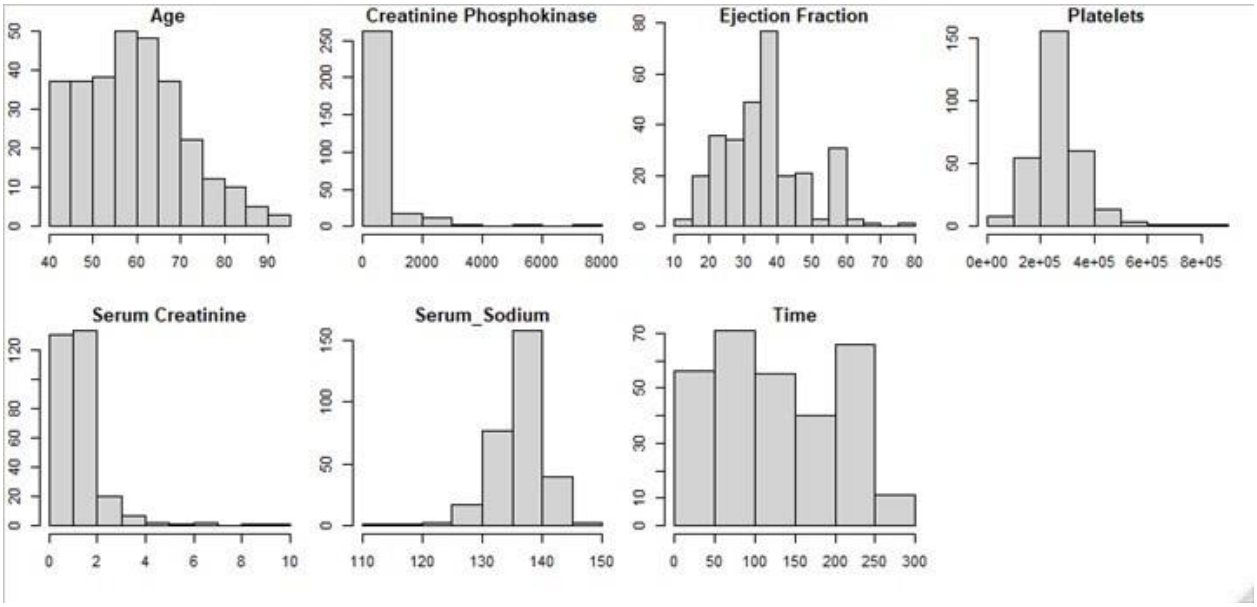
Transformations

All continuous predictors were centered and scaled prior to model building; they were also transformed with BoxCox to account for skewness.

Skewness Coefficient for Predictors Before BoxCox (Variables with High Skewness are Highlighted Yellow)

Variable	Skewness
Age	0.418827
Creatinine Phosphokinase	4.41843
Ejection Fraction	0.549823
Platelets	1.447681
Serum Creatinine	4.411387
Serum Sodium	-1.03767
Time	0.126523

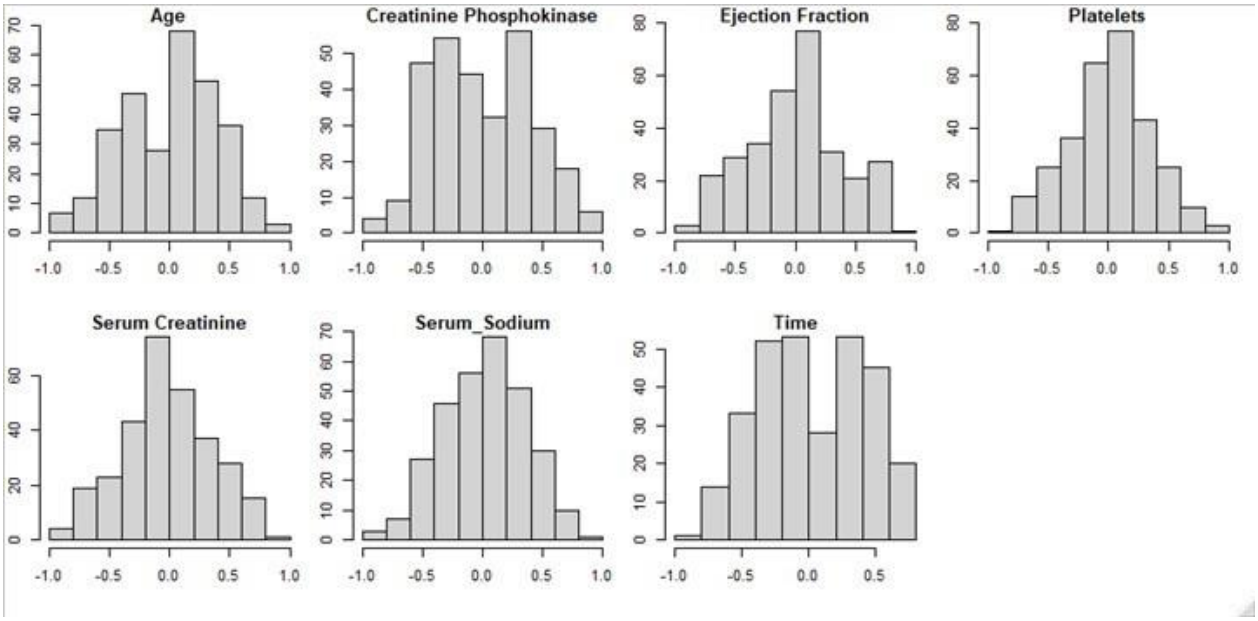
Histograms for Continuous Predictors Before BoxCox



Skewness Coefficient for Predictors After BoxCox, Centering, and Scaling

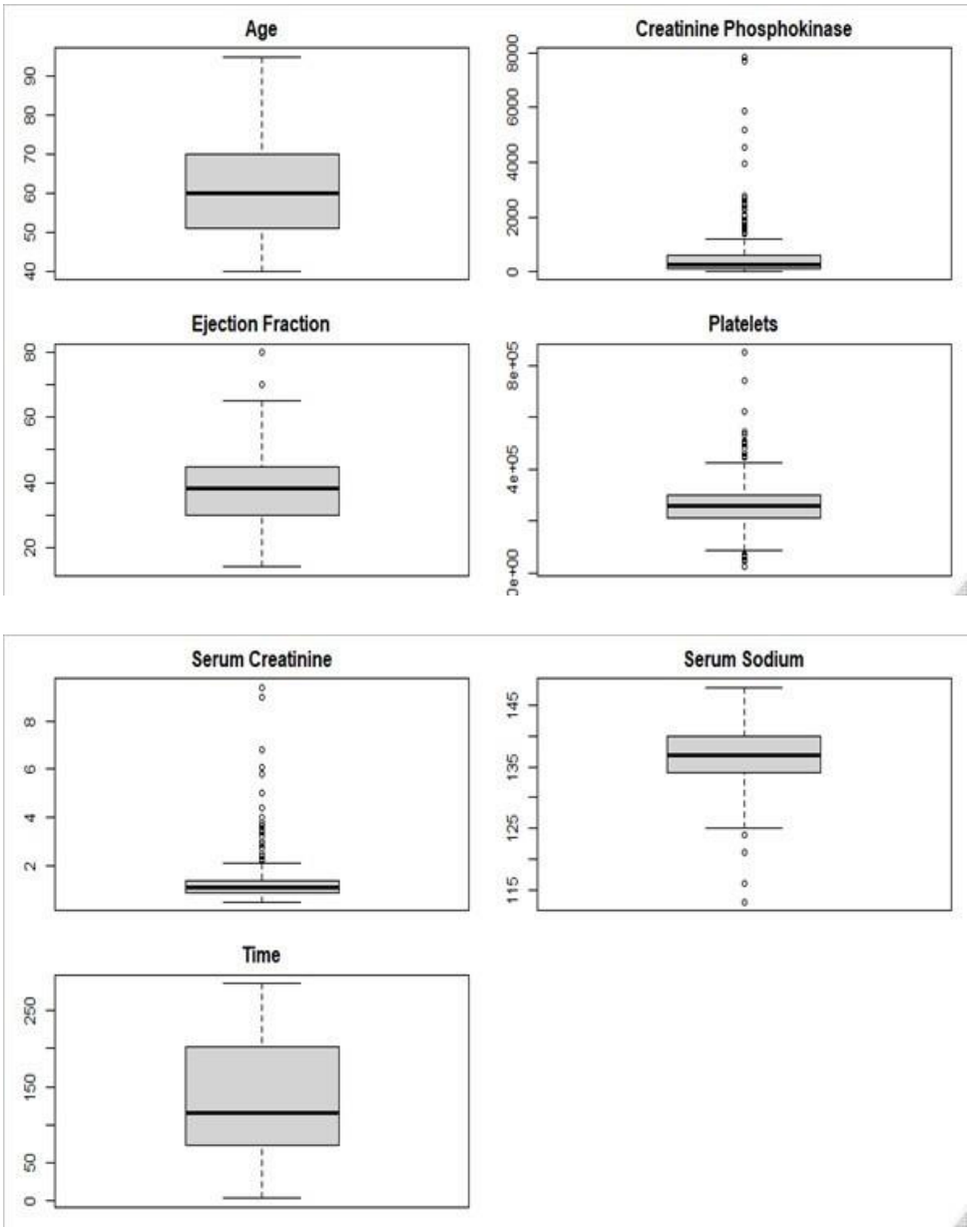
Variable	Skewness
Age	-0.11791832
Creatinine Phosphokinase	0.19009838
Ejection Fraction	-0.02320452
Platelets	0.01192747
Serum Creatinine	-0.06420098
Serum Sodium	-0.06488577
Time	-0.02705088

Histograms for Continuous Predictors After BoxCox

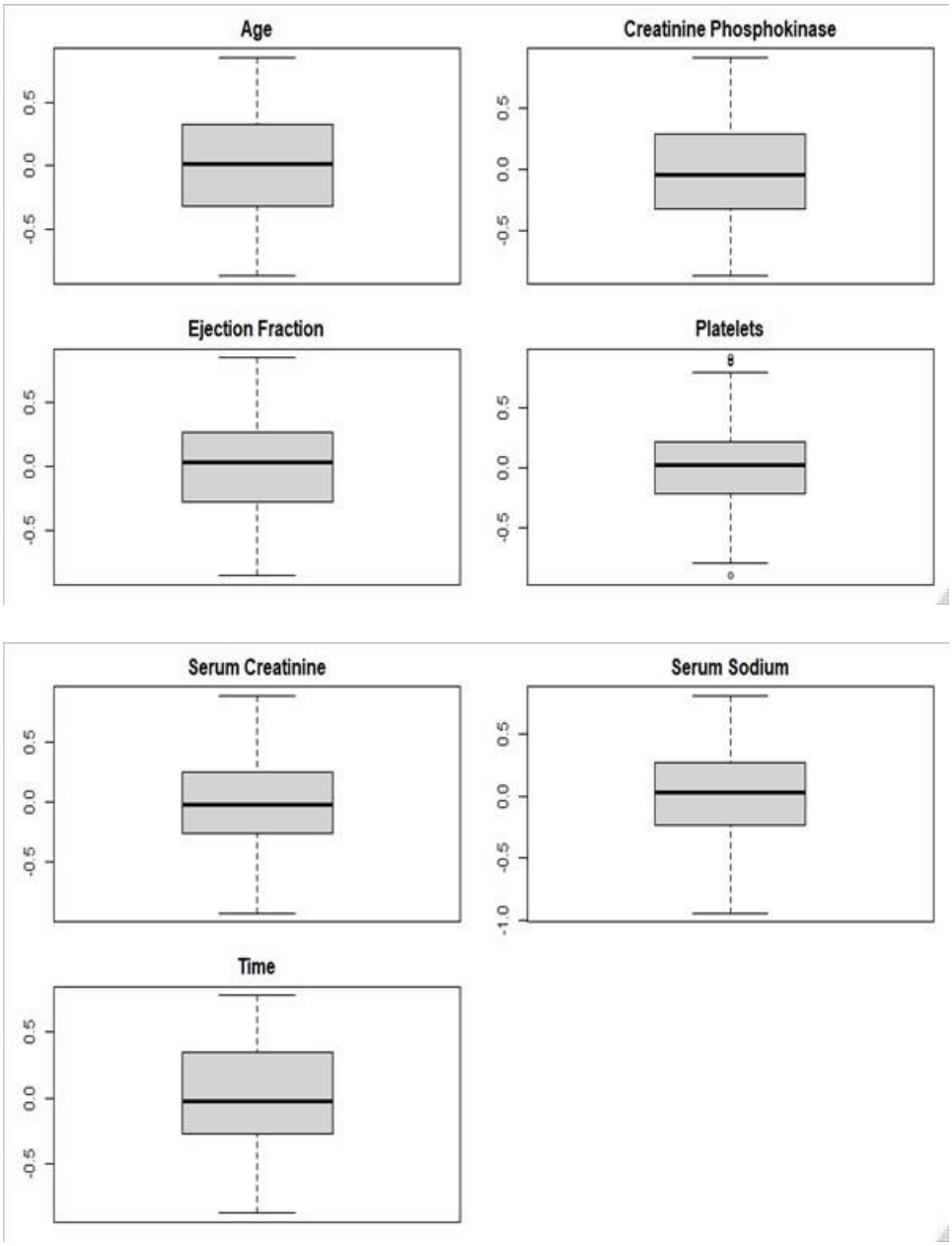


Spatial Sign transformation was done on all continuous predictors to correct for outliers (boxplots below).

BoxPlot for Continuous Predictors Before Spatial Sign Transformation

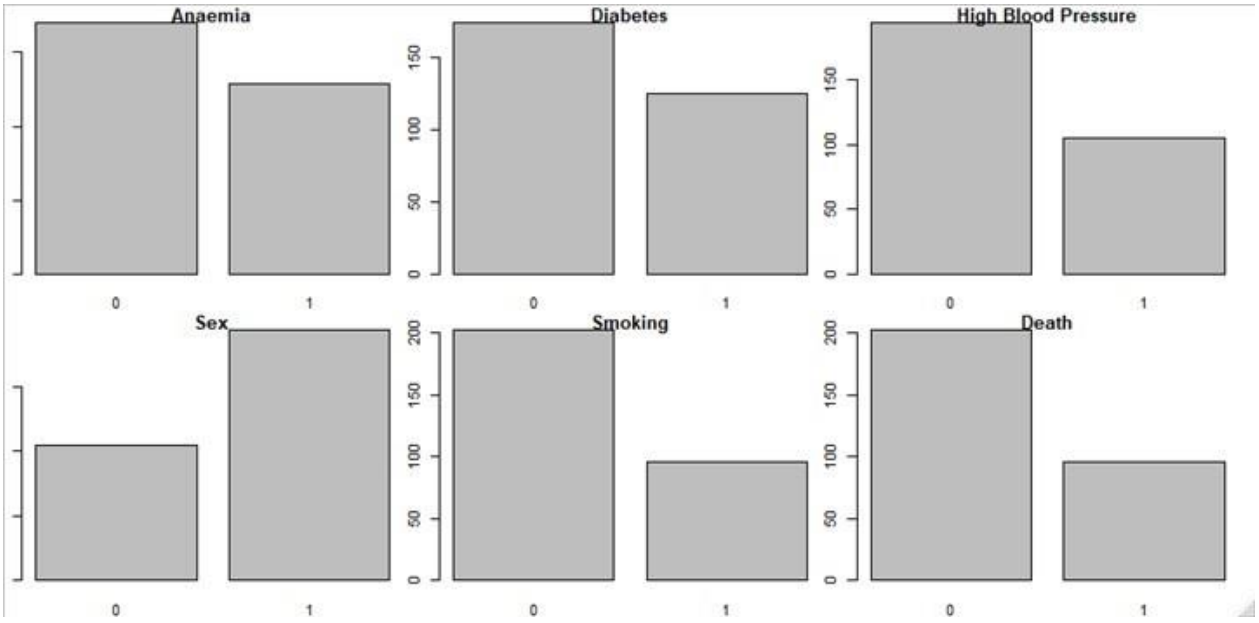


BoxPlot for Continuous Predictors After Spatial Sign Transformation



The categorical variables were all imbalanced to some degree (see graph below).

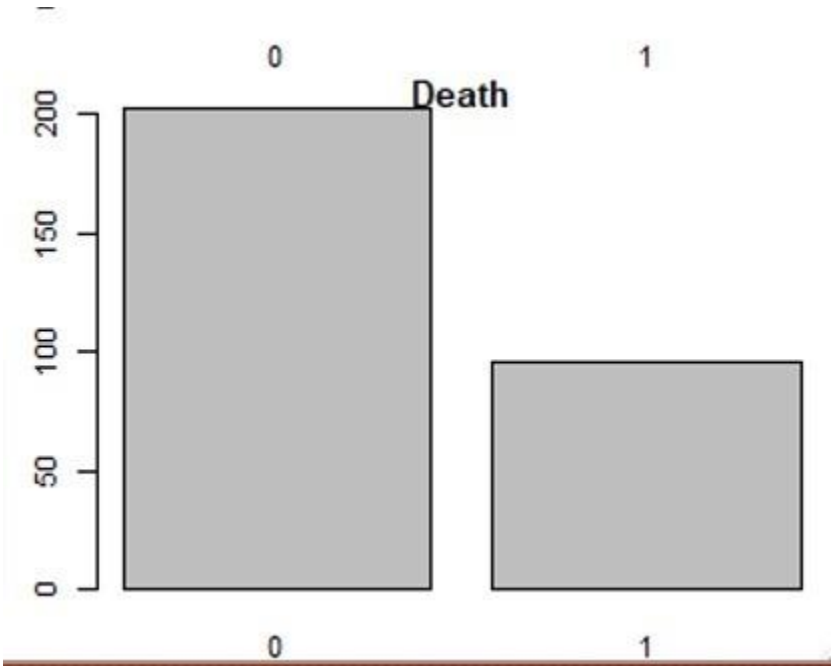
Distributions of Categorical Variables and Response Variable



Data Splitting

Because the binary categorical outcome, DEATH_EVENT, was unbalanced (see graph below), stratified random sampling was used to split the data. 80% of the data was used as the training set and 20% was used for the test set.

Distribution of Response Variable



Model Fitting

Because the outcome was categorical, linear and nonlinear classification models were created. The area under the ROC curve statistic was used to determine the most effective models. The glmnet model and the FDA model were the best models on the training data. The LDA and PLSDA were the best models on the test set, with PLSDA being the best model (see resultant confusion matrix and most important variables list below).

	Observed	
Predicted	Yes	No
Yes	14	5
No	5	35

Seeing as how the model's purpose is to predict death, a higher false positive rate would be preferable to a false negative rate. The confusion matrix shows accuracy of the model is high, with low amounts of both error types.

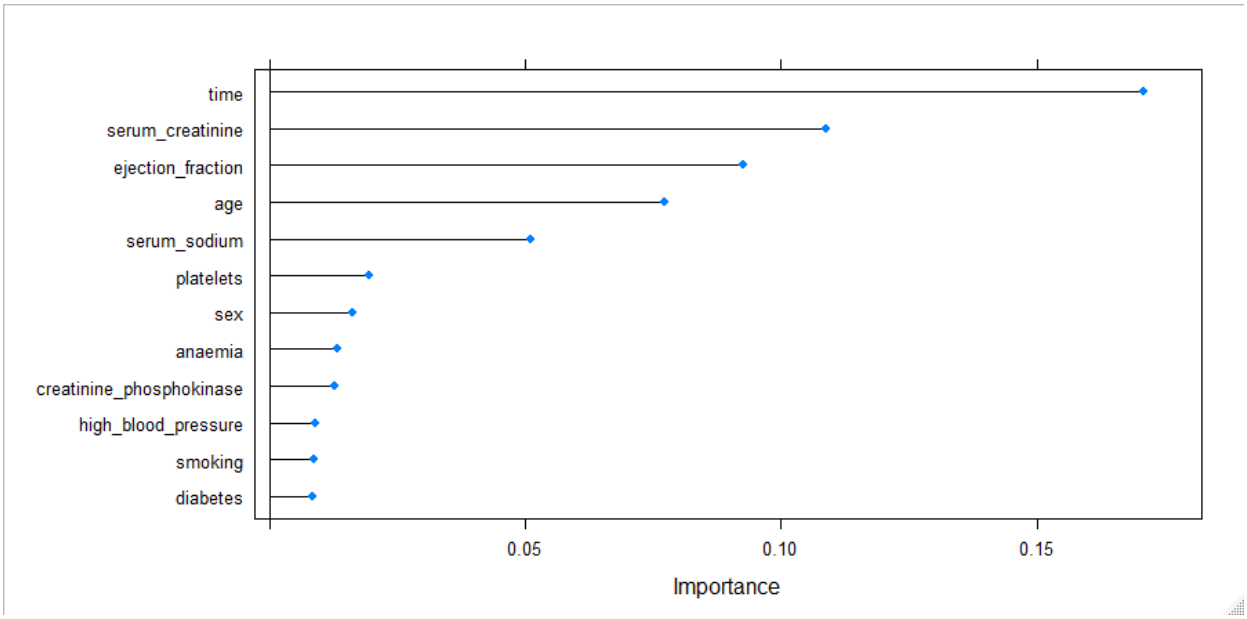
Summary

The best model that performed the best on the testing set was the PLSDA model, with an ROC AUC value of .888 (see table below).

Statistics for Various Models on the Training and Test Sets

	Training			Test		
*No tuning parameters	ROC AUC	Sensitivity	Specificity	ROC AUC	Sensitivity	Specificity
Logistic*	0.884	0.876	0.728	0.879	0.877	0.728
LDA*	0.890	0.858	0.743	0.887	0.858	0.743
PLSDA	0.890	0.868	0.737	0.888	0.865	0.735
Penalized	0.899	0.940	0.604	0.880	0.936	0.533
Neural Network	0.8966	0.889	0.714	0.837	0.876	0.673
Flexible DA	0.8971	0.895	0.701	0.862	0.878	0.692
Quadratic DA*	0.840	0.858	0.606	0.833	0.858	0.606
Regularized DA	0.877	0.884	0.613	0.859	0.872	0.651
Mixture DA	0.890	0.857	0.745	0.873	0.865	0.694
Support Vector Machine	0.887	0.857	0.724	0.858	0.864	0.674
K-nearest neighbor	0.856	0.958	0.288	0.824	0.928	0.371
Naïve Bayes*	0.889	0.91	0.606	0.885	0.910	0.606

Additionally, the most important variables for the PLSDA are shown below:

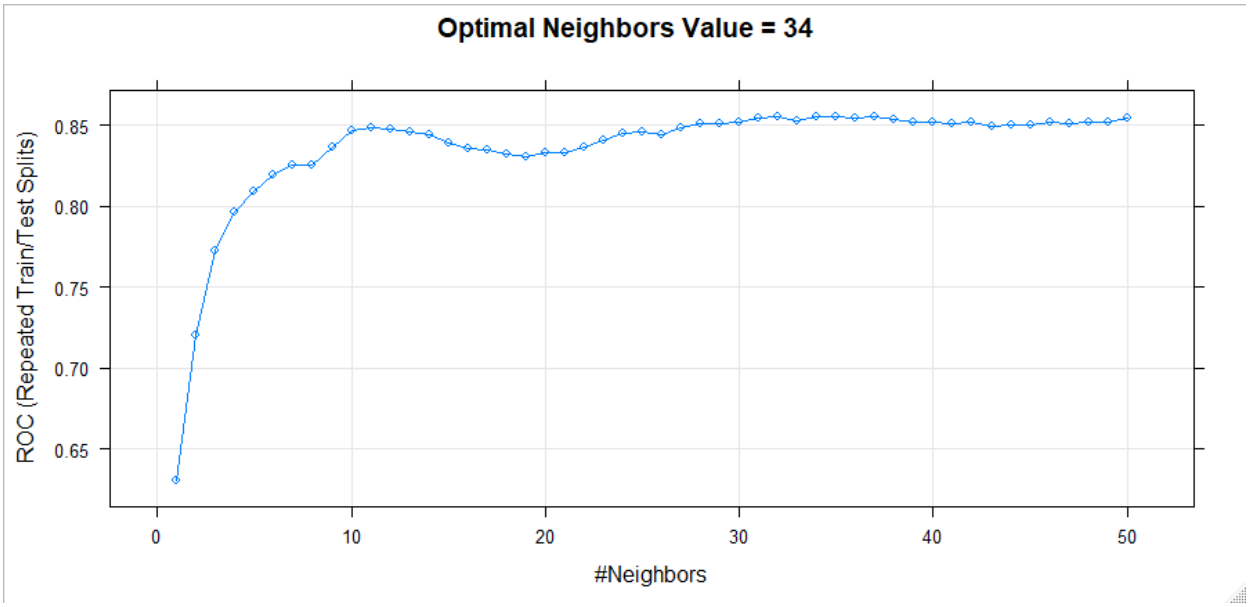


The results are somewhat surprising, as time (days since initial appointment where data was collected), a variable that could be thought of as arbitrary, was found to be the most important. There could be errors or biases that were unexpectedly affecting the data. Or on a more morbid note, it could be viewed that all these patients are very likely to die, with the only variable of import being time. It is also unexpected that smoking, diabetes, and high blood pressure are not of higher importance. There may be some errors resulting from the imbalance data.

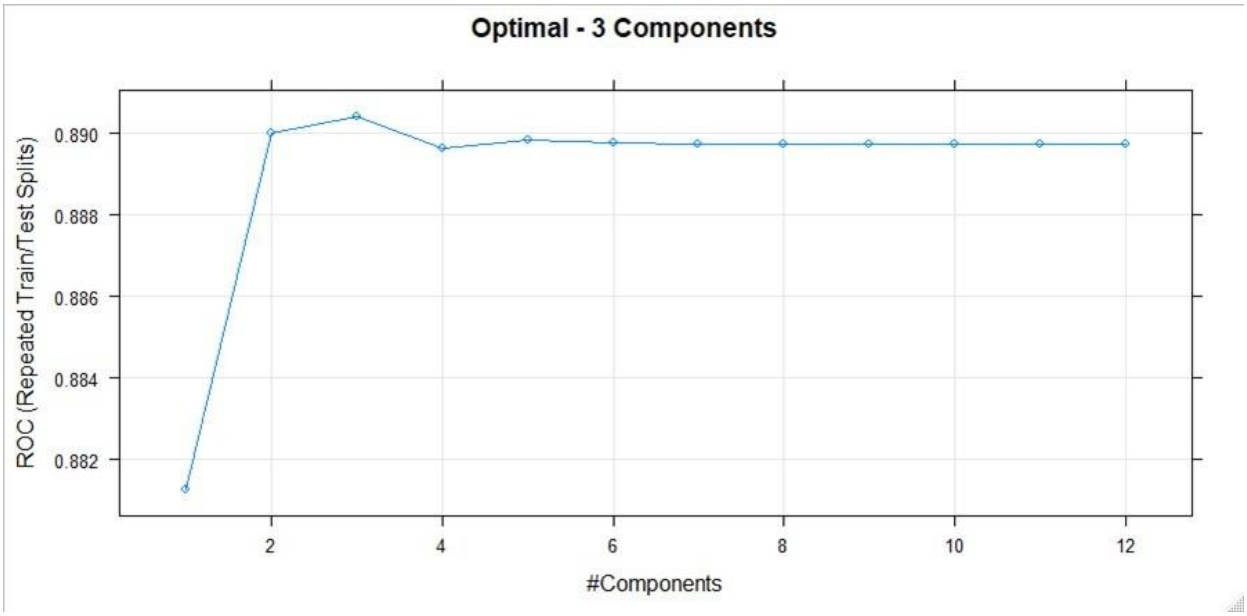
Appendix: Supplemental Material

Tuning Parameter Graphs:

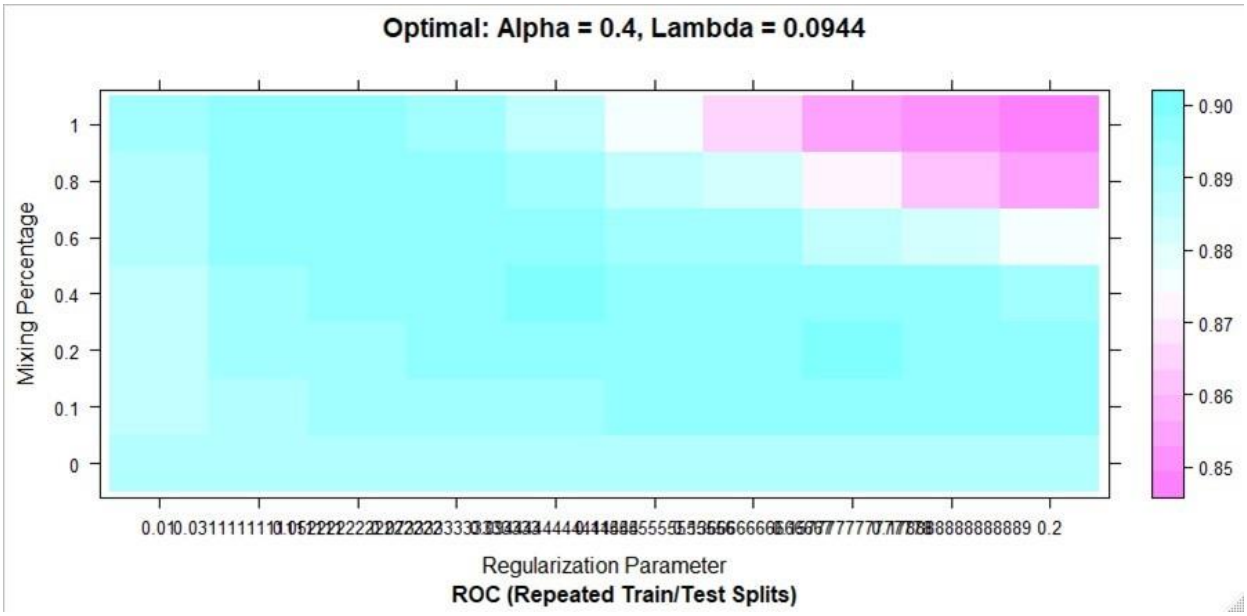
KNN



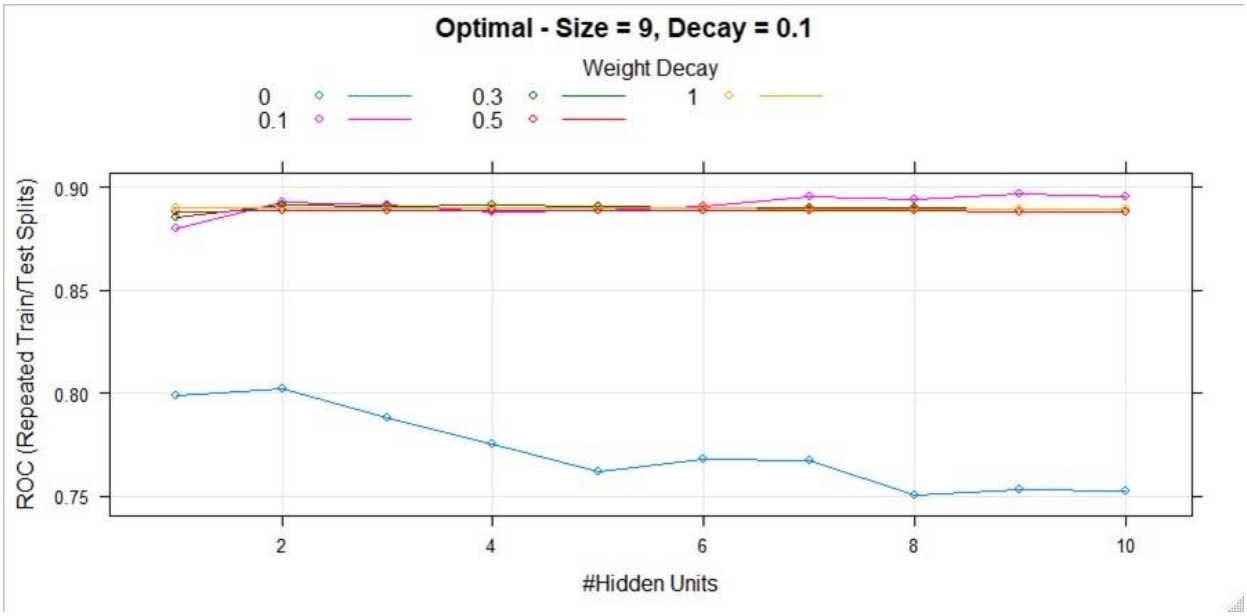
Partial Least Square Discriminant Analysis



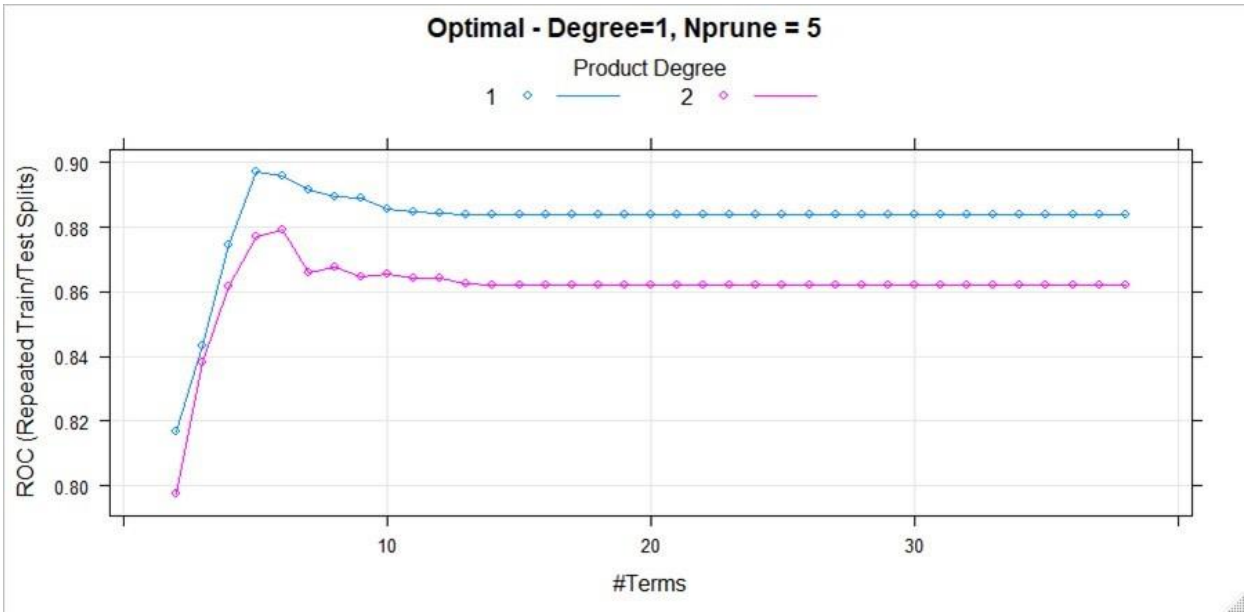
Penalized



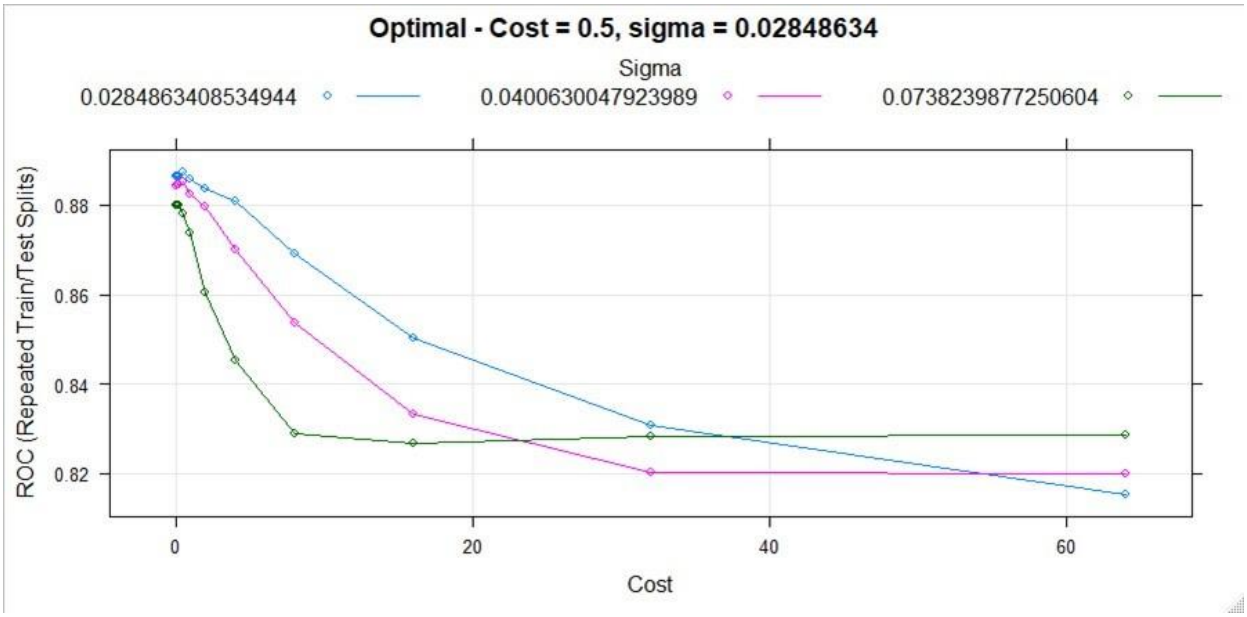
Neural Network



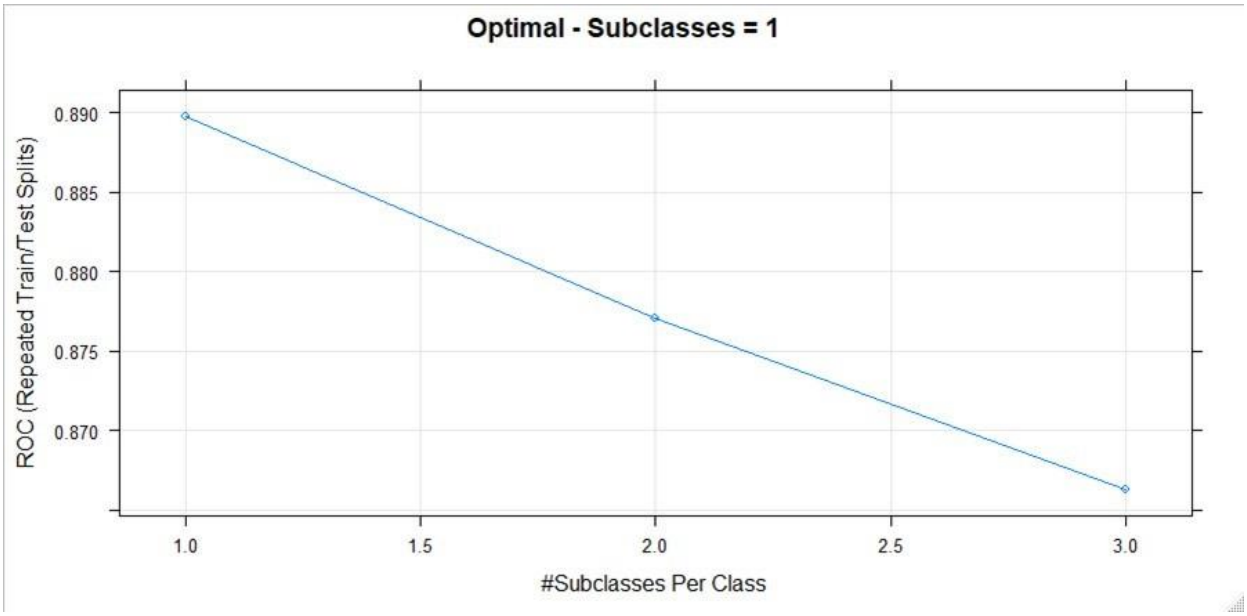
Flexible Discriminant Analysis



Support Vector Machine



Mixture Discriminant Analysis



Regularized Discriminant Analysis