Sean Moser

Statistics 4720 - Applied Statistical Models II

Final Project Report

Introduction

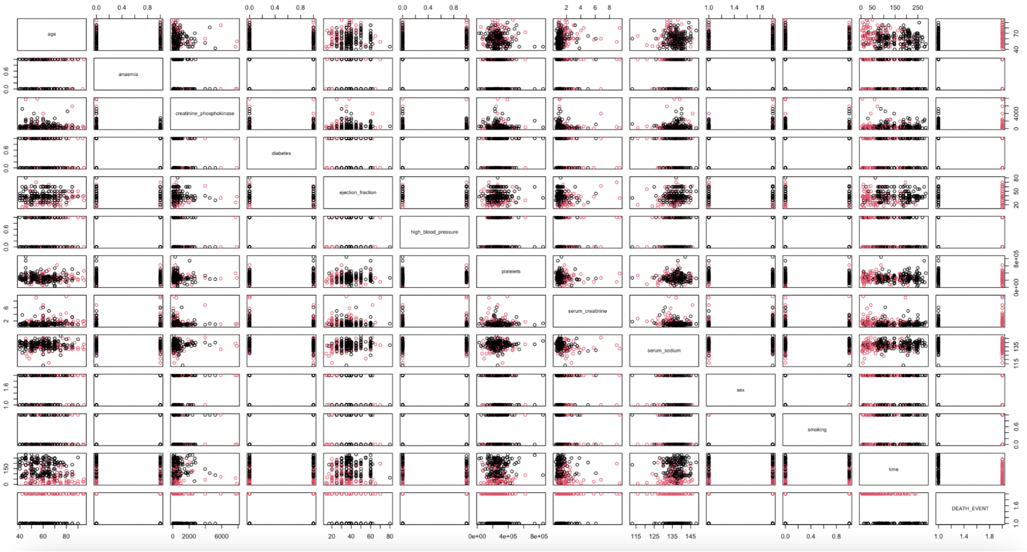
For this project, our focus is on a multivariate dataset containing 299 instances of thirteen different variables. Of these thirteen features, we will be utilizing the first twelve as predictor features that contribute towards our target variable, “death_event”, a binary response variable which tracks whether or not a particular instance, with its twelve predictors, concluded in the death of the patient or not. Some basic predictors, such as age, sex, smoking status, diabetic status, etc., will be utilized to identify trends in the dataset. More technical predictors, such as level of creatinine phosphokinase, level of serum creatinine, level of serum sodium, ejection fraction, will also be utilized as predictor variables to best identify and predict outcomes for a new dataset with the same features. Our dataset is unique because we are not tracking whether or not a person has heart disease, but rather identifies which attributes exist in patients diagnosed with heart disease that ultimately contribute to death in this population. To analyze this dataset, we will be utilizing generalized linear models to first assess our data and then make adjustments to our model based on these outcomes to better fit our dataset. It is likely we will have to utilize other classification methods to create a model that produces more accurate, precise outcomes. Possible alternative classification methods include support vector machines, linear discriminant analysis, and decision trees, which aid in a deeper, fuller understanding to answering our question: which features are most significant in predicting death in patients diagnosed with heart disease?

Objective

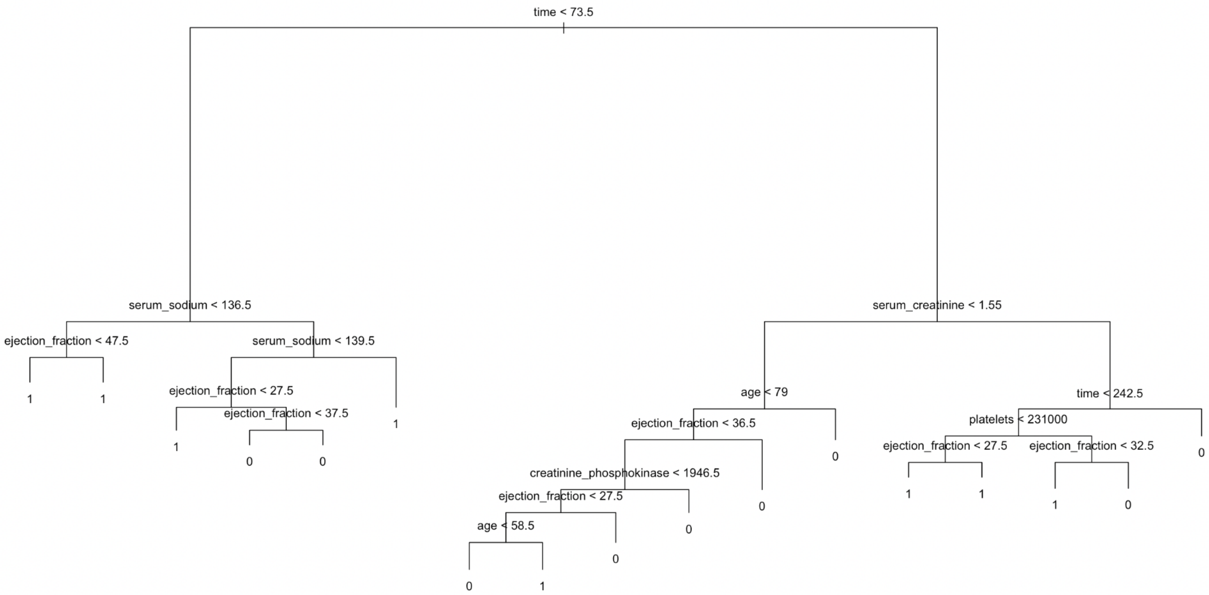
Analysis of this data is significant to cardiologists, emergency medical practitioners, public health specialists, etc., in further understanding heart disease, the leading cause of death in the United States. Furthermore, we understand what specific conditions in a diseased heart lead to the death of the patient. Our goal is to both determine which predictors are significant as well as developing a model that achieves at least 80% classification accuracy.

Initial Data Observations

Utilizing the ‘pairs()’ command, classifying responses by ‘heart.train$DEATH_EVENT”, we identify some attributes in our dataset useful for our analysis. We see relationships exist in predicting ‘DEATH_EVENT’ through predictor variables ‘ejection_fraction’, ‘time’, ‘age’, ‘serum_creatine’, and ‘serum_sodium’.

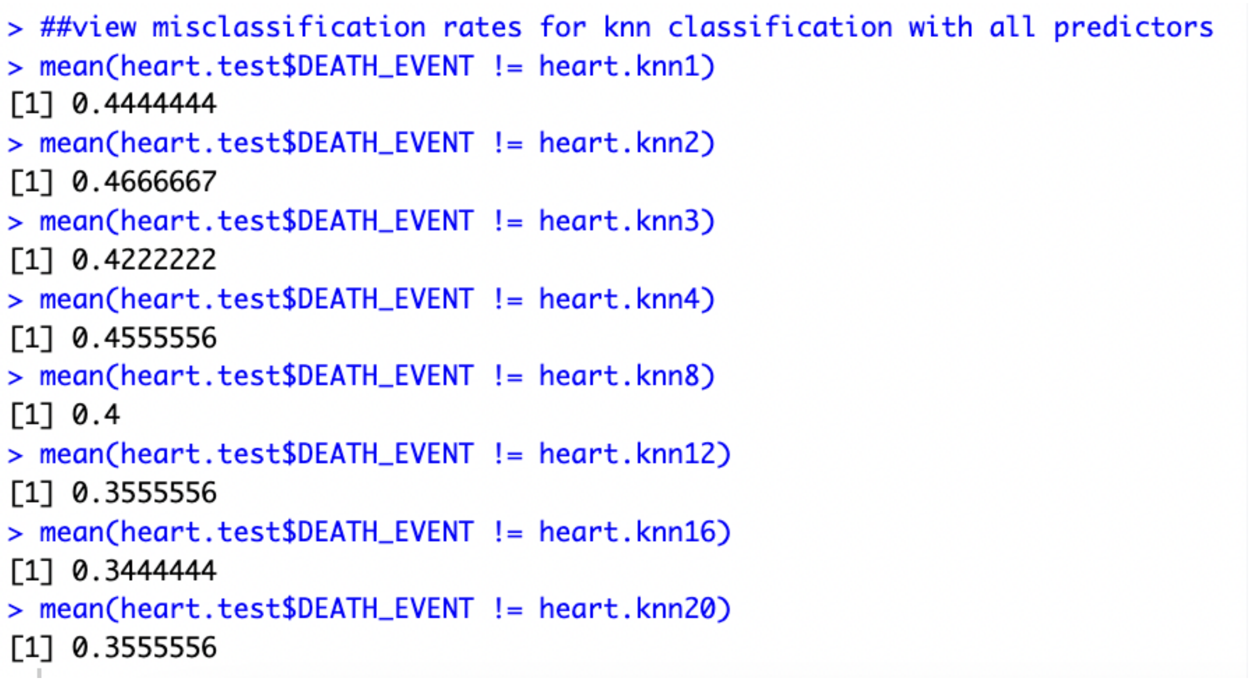


Analyzing our data through a decision tree helps us to further understand how we might better model and predict using our dataset. The resultant decision tree demonstrated the ‘time’, ‘serum_creatine’, ‘serum_sodium’, ‘age’, and ‘time’, are all important factors in predicting the outcome of ‘DEATH_EVENT’. We learn, based on the decision tree, a low time period combined with a low ejection fraction value commonly led to the binary response, 1, indicating the patient died. High levels of creatinine in the blood also leads to many outcomes of death in patients who had a sustained follow up period.



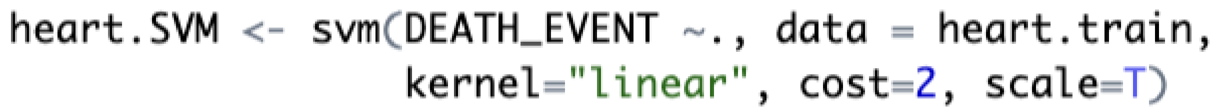
Model Fitting and Testing

This data was fitted to a variety of different models including K nearest neighbors classification, tree methodology, support vector machines, and a binomial generalized linear model. Beginning with K nearest neighbor classification, this model resulted in the least accurate predictions.



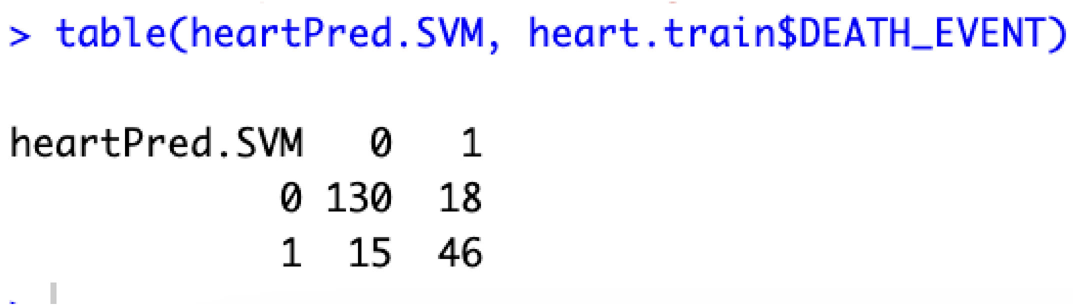
```
> ##view misclassification rates for knn classification with all predictors
> mean(heart.test$DEATH_EVENT != heart.knn1)
[1] 0.4444444
> mean(heart.test$DEATH_EVENT != heart.knn2)
[1] 0.4666667
> mean(heart.test$DEATH_EVENT != heart.knn3)
[1] 0.4222222
> mean(heart.test$DEATH_EVENT != heart.knn4)
[1] 0.4555556
> mean(heart.test$DEATH_EVENT != heart.knn8)
[1] 0.4
> mean(heart.test$DEATH_EVENT != heart.knn12)
[1] 0.3555556
> mean(heart.test$DEATH_EVENT != heart.knn16)
[1] 0.3444444
> mean(heart.test$DEATH_EVENT != heart.knn20)
[1] 0.3555556
```

Our model had the best success predicting the ‘DEATH_EVENT’ outcome when it was modeled with respect to its nearest sixteen neighbors. When using our most optimized KNN model, we can predict our binary outcome variable 65.56% of the time. Fitting, then tuning a support vector machines model found setting ‘cost = 2’ optimized the error to a minimum. The following is the syntax used in constructing the model.



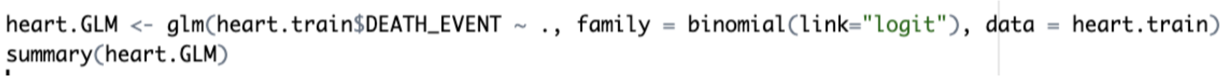
```
heart.SVM <- svm(DEATH_EVENT ~., data = heart.train,
                 kernel="linear", cost=2, scale=T)
```

Tuning our model with this cost also introduces a 10-fold cross validation on the training dataset, which improves the accuracy of its results. The resulting model correctly classified the outcome 84.21% of the time, which is a definite improvement from our K nearest neighbor binary classification. The below table demonstrates model prediction performance against training data.



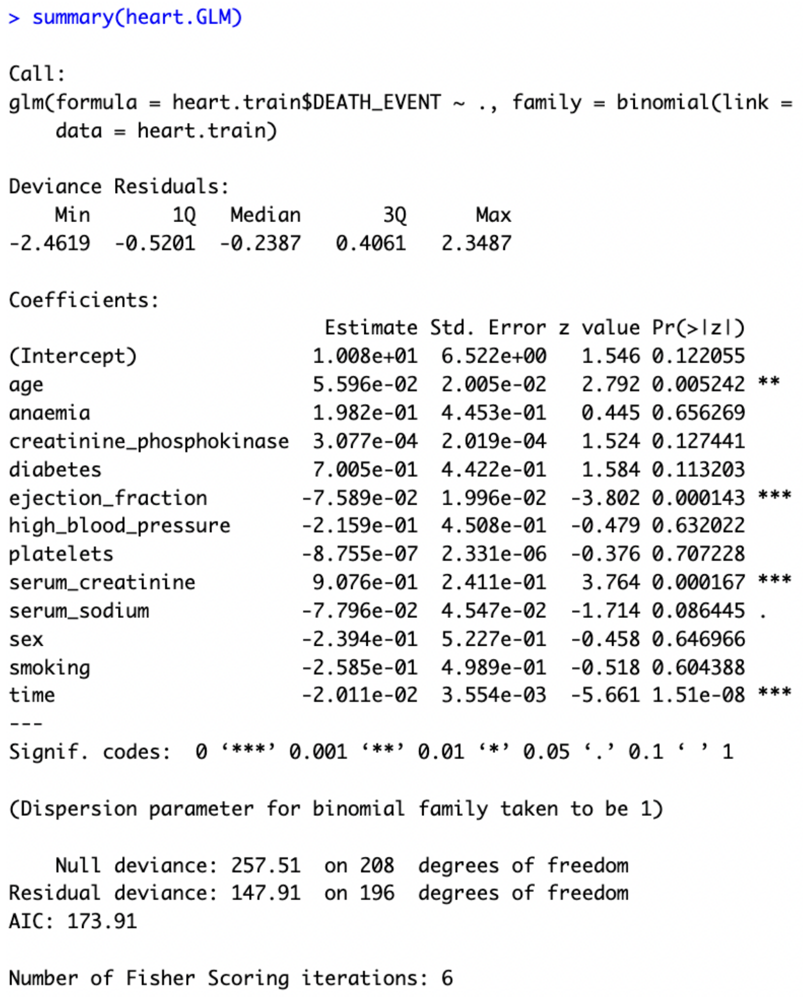
```
> table(heartPred.SVM, heart.train$DEATH_EVENT)

heartPred.SVM   0   1
            0 130  18
            1  15  46
```

The following syntax was used in constructing our generalized linear model.



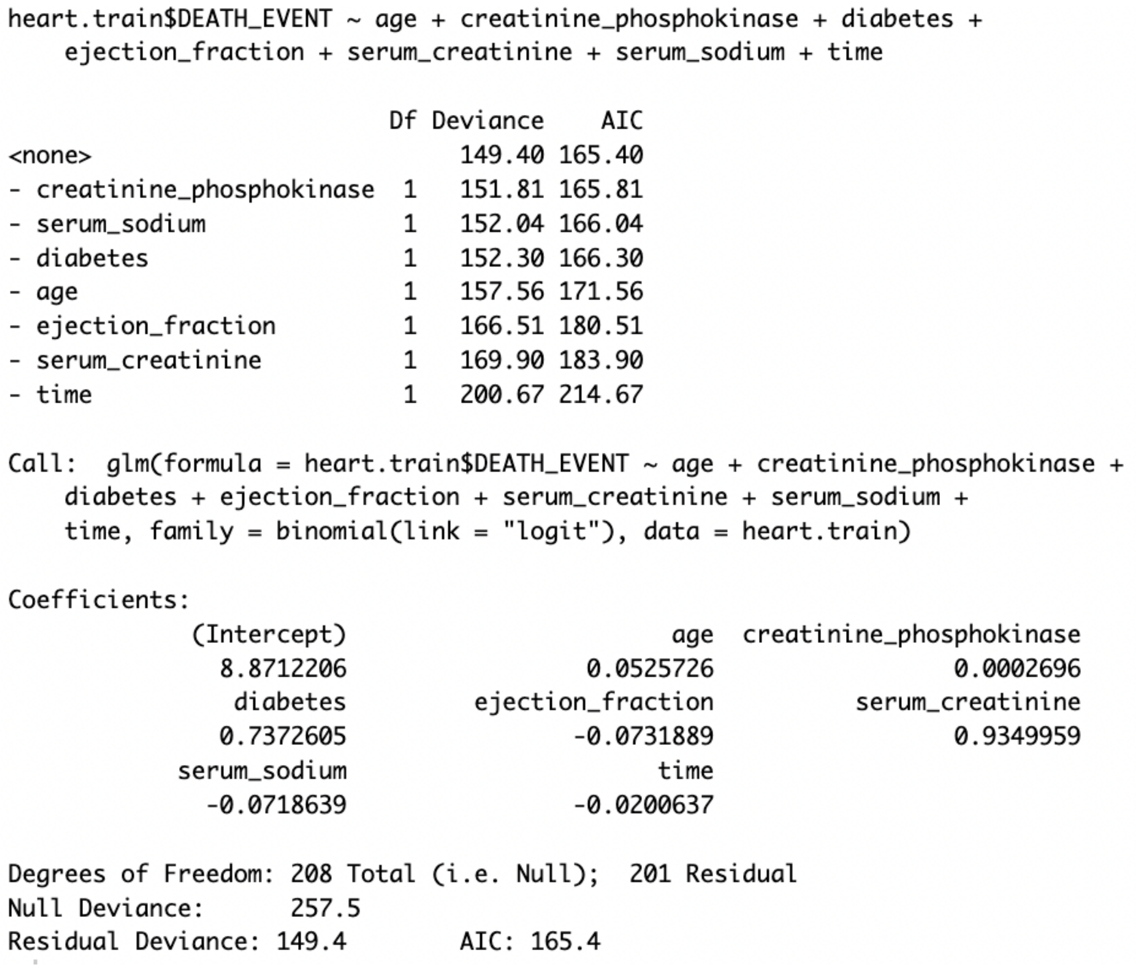
```
heart.GLM <- glm(heart.train$DEATH_EVENT ~ ., family = binomial(link="logit"), data = heart.train)
summary(heart.GLM)
```

This model utilized the ‘family = binomial(link = “logit”)’ statement to designate the link function used. Below is the syntax used in constructing this model. Our link function is the binomial link function, which helps our model map outcomes to its respecting binary outcome. Our generalized model found age, ejection_fraction, serum_creatine, and time to all be significant at a 0.01 level of significance at the least.



```
> summary(heart.GLM)

Call:
glm(formula = heart.train$DEATH_EVENT ~ ., family = binomial(link = "logit"), 
    data = heart.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4619  -0.5201  -0.2387   0.4061   2.3487  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)               1.008e+01  6.522e+00   1.546 0.122055    
age                       5.596e-02  2.005e-02   2.792 0.005242 ** 
anaemia                   1.982e-01  4.453e-01   0.445 0.656269    
creatinine_phosphokinase  3.077e-04  2.019e-04   1.524 0.127441    
diabetes                  7.005e-01  4.422e-01   1.584 0.113203    
ejection_fraction        -7.589e-02  1.996e-02  -3.802 0.000143 ***
high_blood_pressure      -2.159e-01  4.508e-01  -0.479 0.632022    
platelets                -8.755e-07  2.331e-06  -0.376 0.707228    
serum_creatinine          9.076e-01  2.411e-01   3.764 0.000167 ***
serum_sodium             -7.796e-02  4.547e-02  -1.714 0.086445 .  
sex                      -2.394e-01  5.227e-01  -0.458 0.646966    
smoking                  -2.585e-01  4.989e-01  -0.518 0.604388    
time                     -2.011e-02  3.554e-03  -5.661 1.51e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.51  on 208  degrees of freedom
Residual deviance: 147.91  on 196  degrees of freedom
AIC: 173.91

Number of Fisher Scoring iterations: 6
```

A call to the ‘step()’ function helps us in refining our model to reduce error in our model, optimizing our AIC to a minimum. The following output below is the final product of feeding our generalized linear model to the ‘step()’ function.



```
heart.train$DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes +
    ejection_fraction + serum_creatinine + serum_sodium + time

                           Df Deviance    AIC
<none>                          149.40 165.40
- creatinine_phosphokinase  1   151.81 165.81
- serum_sodium              1   152.04 166.04
- diabetes                  1   152.30 166.30
- age                       1   157.56 171.56
- ejection_fraction         1   166.51 180.51
- serum_creatinine          1   169.90 183.90
- time                      1   200.67 214.67

Call:  glm(formula = heart.train$DEATH_EVENT ~ age + creatinine_phosphokinase +
    diabetes + ejection_fraction + serum_creatinine + serum_sodium +
    time, family = binomial(link = "logit"), data = heart.train)

Coefficients:
             (Intercept)                       age  creatinine_phosphokinase
               8.8712206                 0.0525726                 0.0002696
                diabetes         ejection_fraction          serum_creatinine
               0.7372605                -0.0731889                 0.9349959
            serum_sodium                      time
              -0.0718639                -0.0200637

Degrees of Freedom: 208 Total (i.e. Null);  201 Residual
Null Deviance:      257.5
Residual Deviance: 149.4        AIC: 165.4
```

The reduced, final model decreased AIC by about 8 units, demonstrating an improvement in model performance and error reduction by removing unnecessary and irrelevant predictor variables.

Interpretation and Discussion of Results

Our initial goal when conducting this analysis was to answer the question: which features are most significant in predicting death in patients diagnosed with heart disease? We are now able to answer that question. We also found models that correctly classified outcomes with at least 80% accuracy. The age of the patient is indeed a useful predictor in determining whether a given instance of heart disease will end in death. The calculated ejection fraction, which is the amount of blood each pump of the heart disperses to the body, is also significant in determining the outcome of a patient. Abnormal levels of serum creatinine contribute to death in a diseased heart. This measure tells us kidney failure (which would cause abnormal levels of creatinine) is also likely to compound the damaging effects of heart failure, inducing death in these patients. The final variable found to be statistically significant was the ‘time’ predictor, which measured the overall length of the follow-up period in which the data was collected.

Future studies stemming from this analysis are plentiful, as we find a meaningful intersection between biophysical measurements and statistical analyses. Due to the findings of our study, how kidney issues and levels of creatinine affect the heart imply an interesting dynamic between nephrology and cardiology. Further eliminating more insignificant measurements from future datasets and collecting different potential predictor variables based on the findings of this study could further prompt more abstract discoveries and progress within the field of cardiovascular medicine and biostatistics as a whole.