

Predicting Heart Failure from Risk Factor Data

Paige Galvan, Neha Deshpande, & Witlie Leslie

2024-11-25

Introduction (Witlie)

For our project, we aimed to create a model that can predict the event of death as a result of heart failure from clinical patient data.

- ▶ Binary response variable “DEATH_EVENT”, value of 1 indicating patient deceased (0 otherwise)
- ▶ 12 risk factor variables
 - ▶ 5 binary: anemia status, diabetes status, high blood pressure status, sex, and smoking status
 - ▶ 7 numerical: age, creatine phosphokinase level, ejection fraction, platelet concentration, serum creatine level, serum sodium level, and length of follow-up period

Motivation (Neha)

- ▶ Investigating heart failure, a condition impacting millions worldwide, and exploring its complex causes and contributing factors.
- ▶ Spotting patterns between risk factors and how heart failure progresses to get a better understanding.

Linearity and Normality (Neha)

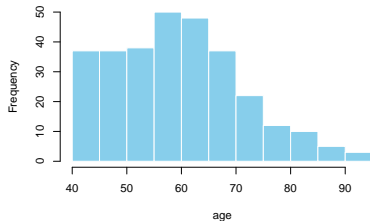
Normality

- ▶ Objective: Assess whether continuous predictor variables are normally distributed.
- ▶ Variables analyzed: Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine, Serum Sodium.
- ▶ Testing residuals ensures the model fits the data well and detects patterns that might indicate a need for model improvement.
- ▶ Tests performed:

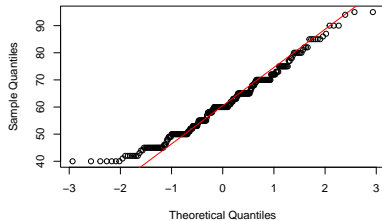
Shapiro-Wilk Test: $p\text{-values} < 0.05$ for all variables \rightarrow None follow a normal distribution.

Graphs

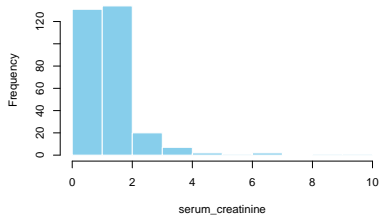
Histogram of age



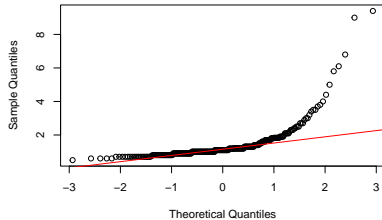
Q-Q Plot of age



Histogram of serum_creatinine



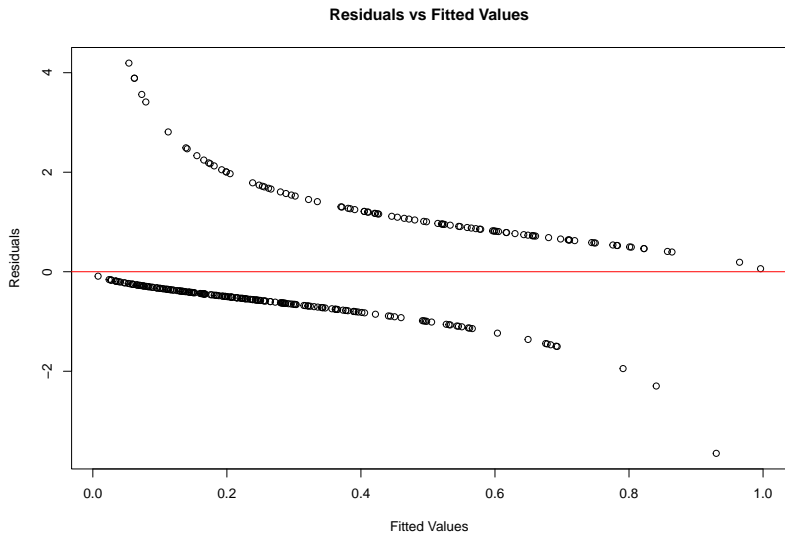
Q-Q Plot of serum_creatinine



Linearity

- ▶ Scatterplots of each variable vs. DEATH_EVENT:
- ▶ Fitted linear trendlines show non linear relationships for most variables.
- ▶ Logistic regression results: Predictor variables (e.g., Age, Serum Creatinine) significantly linked to DEATH_EVENT likelihood. Residual deviance was low, suggesting a good fit.

Graphs

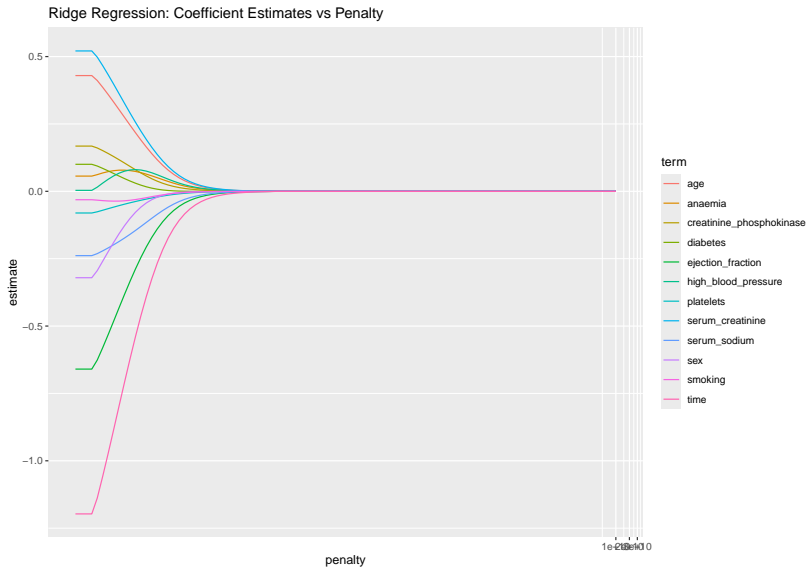


Logistic Regression (Witlie)

Logistic regression was an obvious first choice in tackling this binary classification problem. I created 3 models to assess predictive performance with different regularization methods.

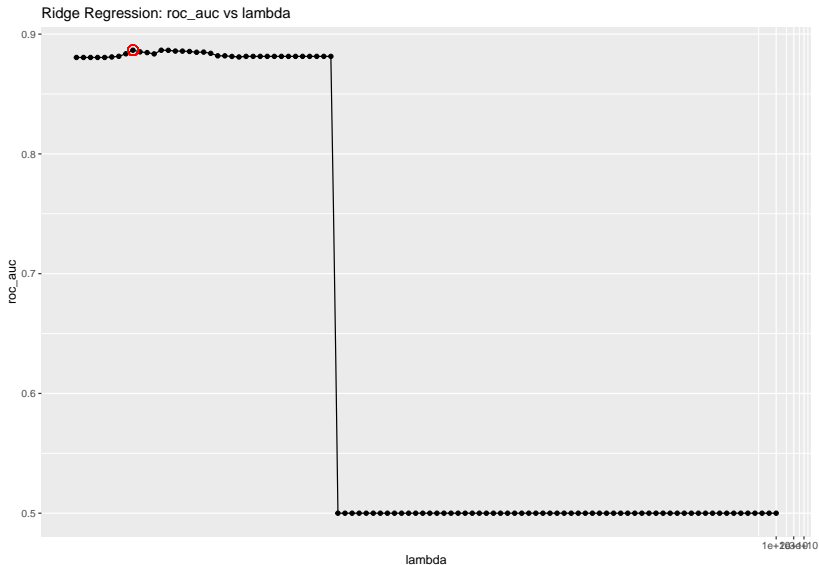
- ▶ Ridge Regression
- ▶ Lasso
- ▶ No regularization

Ridge Regression



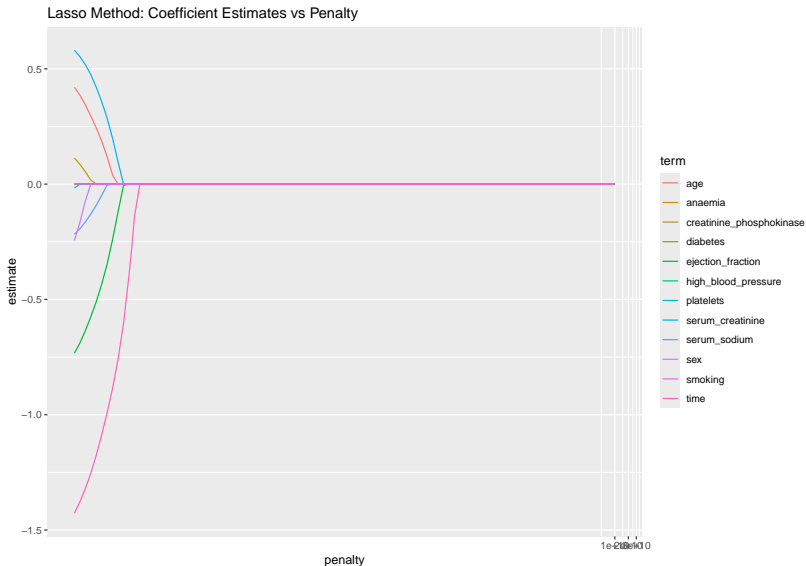
Ridge Regression

Using 10-fold cross-validation, I found the lambda with the highest ROC-AUC value of 0.887 was $\lambda = 0.0933$



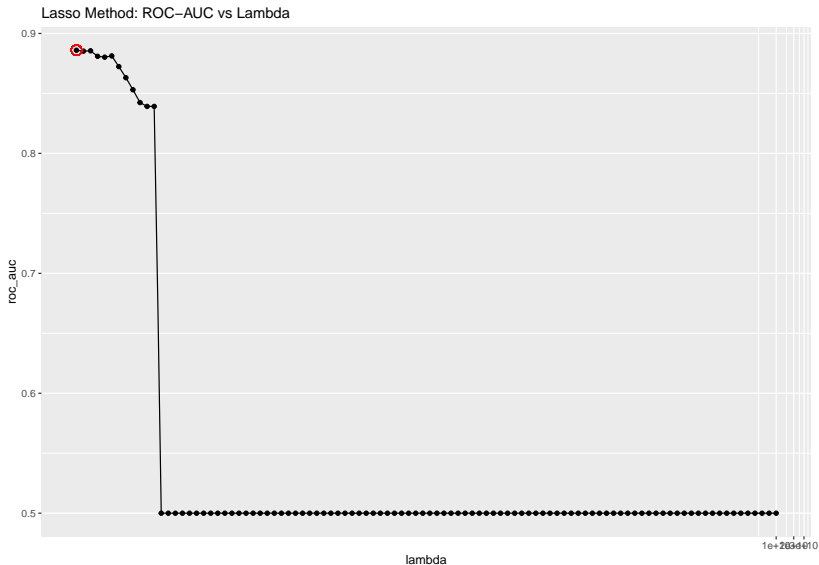
Lasso

Lasso shows a steeper dropoff of coefficient estimates as a result of feature reduction



Lasso

Using 10-fold cross validation, I found the lambda with the highest ROC-AUC value of 0.886 was $\lambda = 0.01$



Results

- ▶ No Regularization:
- ▶ Ridge Regression:
- ▶ Lasso:

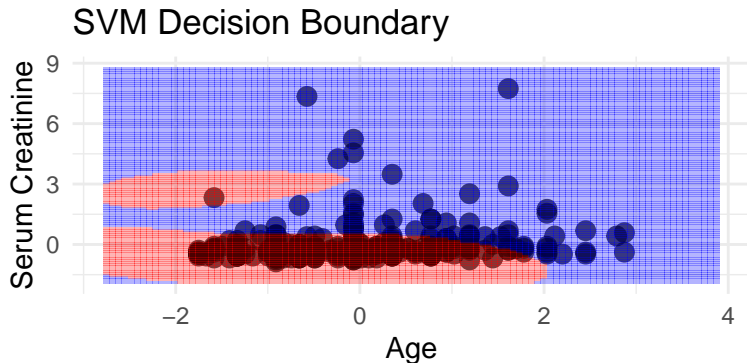
Support Vector Machine

We began by finding the accuracy rating of 81.7% and a prediction matrix

```
##           Actual
## Predicted  0   1
##           0 36   6
##           1   5 13
## [1] "Accuracy:  0.817"
```

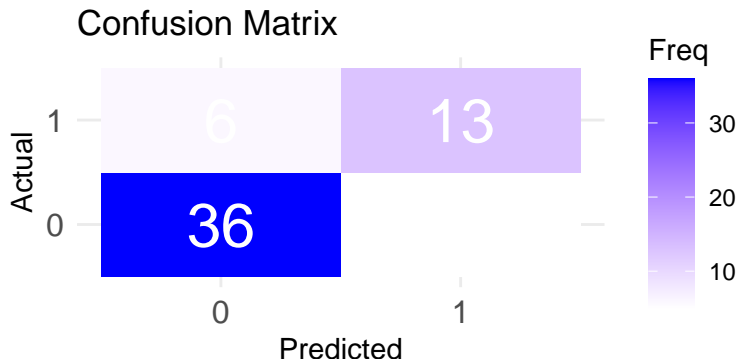
SVM Decision Boundry

The SVM Boundry analyzes the age vs serum creatinine using the death effect for males vs females. We can see that overall serum creatinine has little effect on the death effect especially in males. Meanwhile age and serum creatinine has a large response on females.



SVM Confusion Matrix

The Confusion Matrix is used to support our model's predictions, we can see the majority of true labels correctly matched with the predicted labels, indicating high accuracy.



#SVM ROC Curve

The SVM model, tuned with $C = 0.5$ and $\text{sigma} = 0.0562$, had an ROC of 0.86, a sensitivity of 86%, and a specificity of 66%.

Random Forest

We began by finding the accuracy and a prediction matrix

```
## [1] 0
```

```
##           Actual
```

```
## Predicted  0  1
```

```
##           0 39  4
```

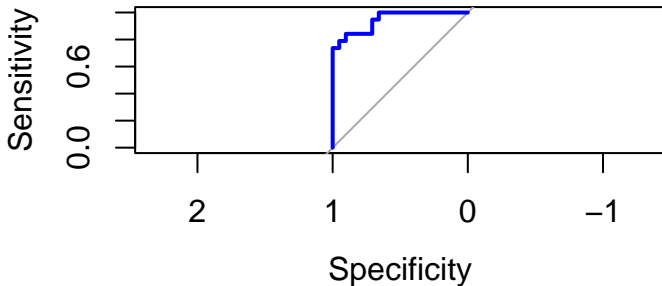
```
##           1  2 15
```

```
## Accuracy:  0.9
```

RF ROC Curve

The Random Forest model performed best with $m_{try} = 2$, giving an ROC of 0.91. It was effective at identifying true positives (91% sensitivity) but had a moderate specificity of 69%.

ROC Curve for Random Forest Model



Overall, the Random Forest model performed better, especially in terms of overall accuracy and detecting positive cases.

Results

- ▶ None of the three logistic regression models performed well. The model without regularization performed very similarly to those which implemented ridge regression and lasso, indicating that regularization is unnecessary for this data.
- ▶ The SVM model provided more insight in making predictions however still wasn't the best option.
- ▶ The Random Forest provides a sound approach to prediction as they use multiple trees and reduce risk of overfitting. This also contains the highest accuracy score with an ROC Curve that confirms it's true positive rate is the highest among it's competitors.
- ▶ Thank you!

References

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020). <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5>