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.

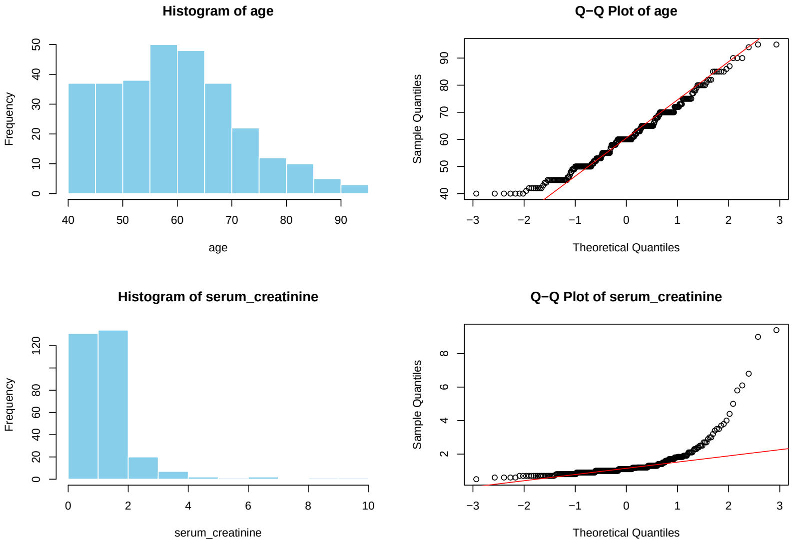
Linearity and Normality (Neha)

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-values < 0.05 for all variables $\rightarrow$ None follow a normal distribution.

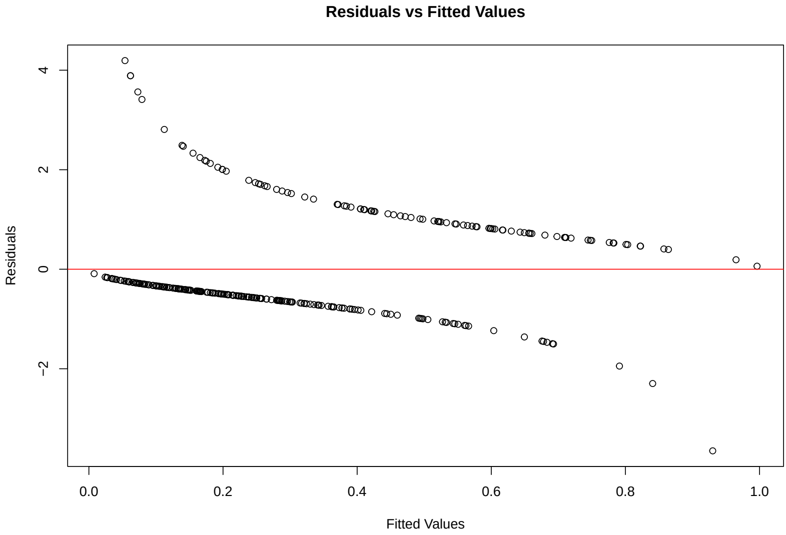
Graphs


Histogram of age
Frequency
age
Q−Q Plot of age
Sample Quantiles
Theoretical Quantiles
Histogram of serum_creatinine
Frequency
serum_creatinine
Q−Q Plot of serum_creatinine
Sample Quantiles
Theoretical Quantiles

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



Residuals vs Fitted Values
Residuals
Fitted Values
4
2
0
−2
0.0
0.2
0.4
0.6
0.8
1.0

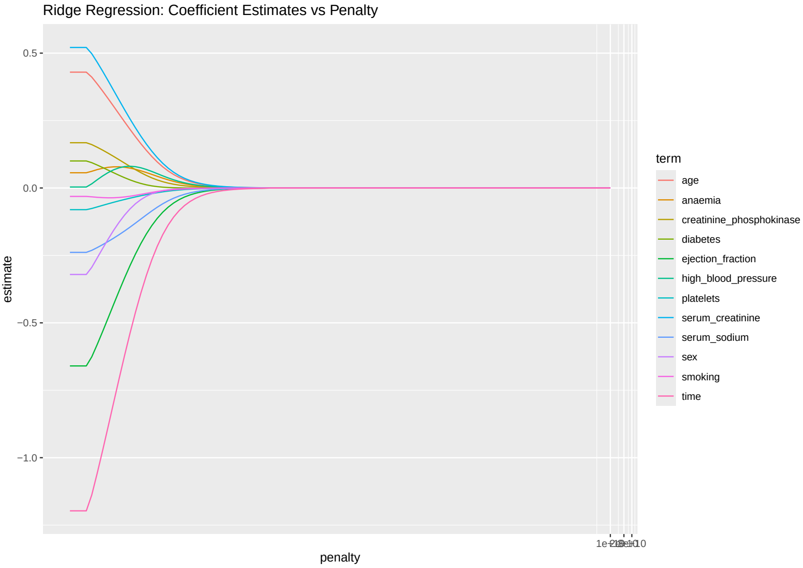
SVM & Random Forest (Paige)

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.

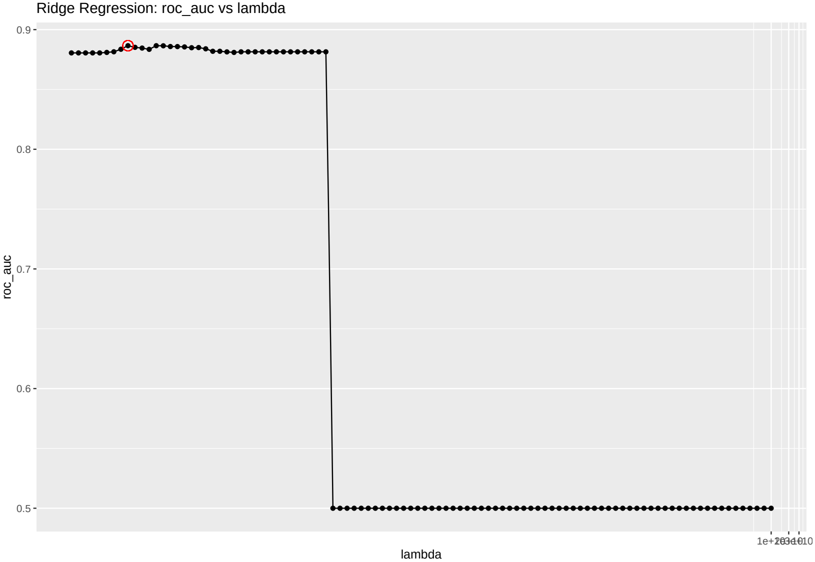
- Ridge Regression
- Lasso
- No regularization

Ridge Regression



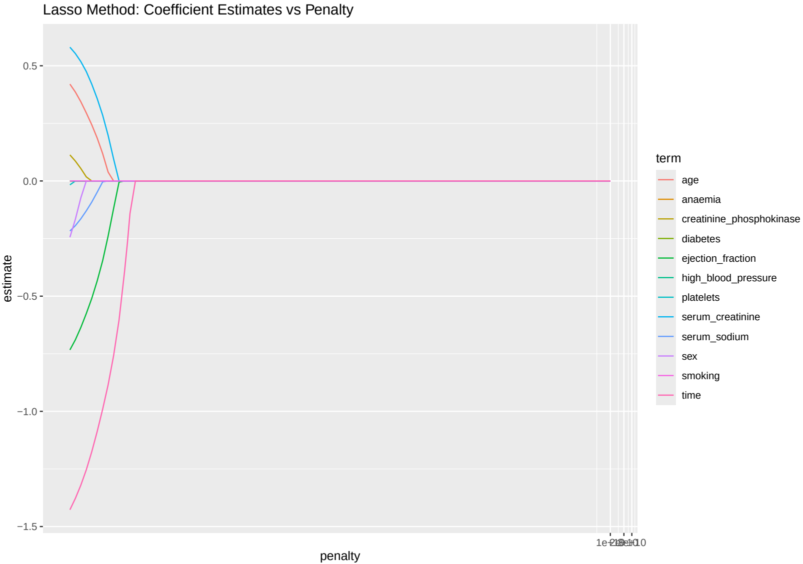
Ridge Regression

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



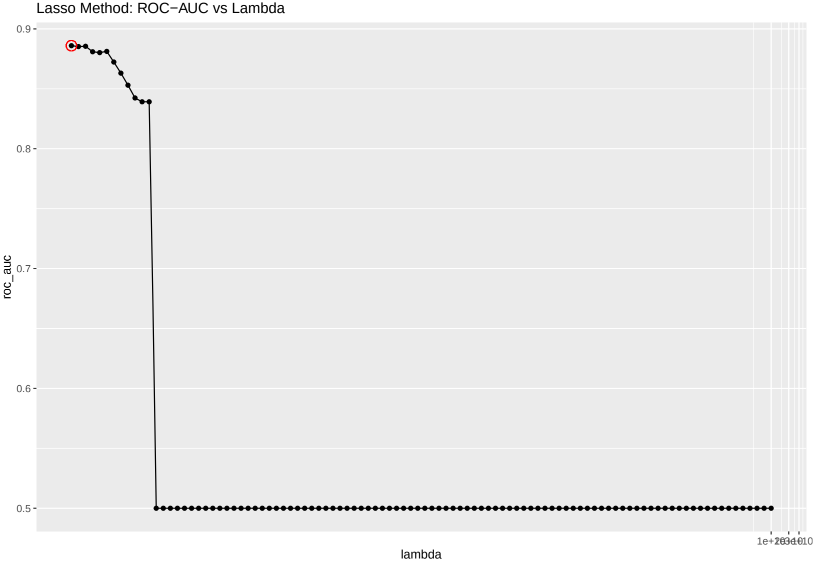
Lasso

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:

SVM

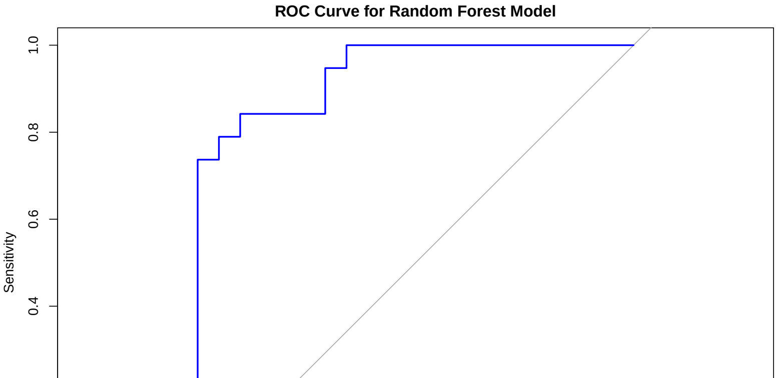
```
##
## Call:
## svm(formula = DEATH_EVENT ~ ., data = train_data, kernel
##     cost = 1, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  148
##
##  ( 69 79 )
##
##
## Number of Classes:  2
##
```

Random Forest (paige)

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

```
##           Actual
## Predicted  0  1
##          0 39  4
##          1  2 15
```

```
## Accuracy:  0.9
```



Results (Paige)

- 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://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5