Predicting Heart Failure from Risk Factor Data

Paige Galvan, Neha Deshpande, & Witlie Leslie

2024-11-25

Introduction

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

- 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.

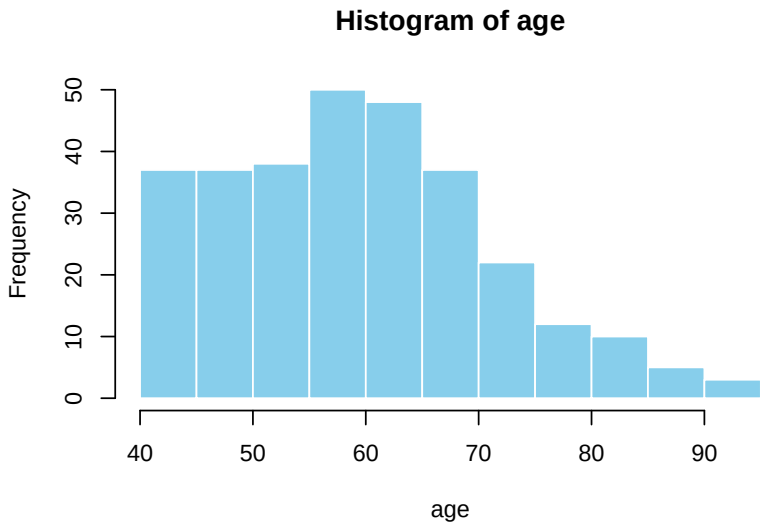
Linearity and Normality

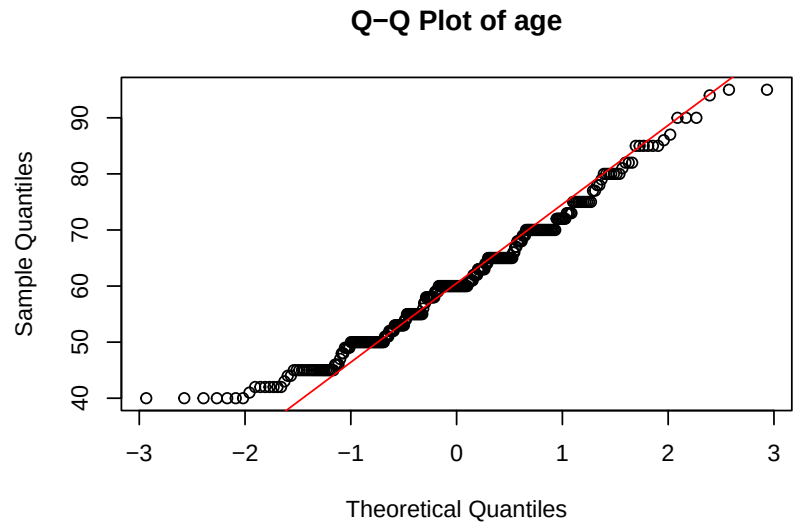
Normality

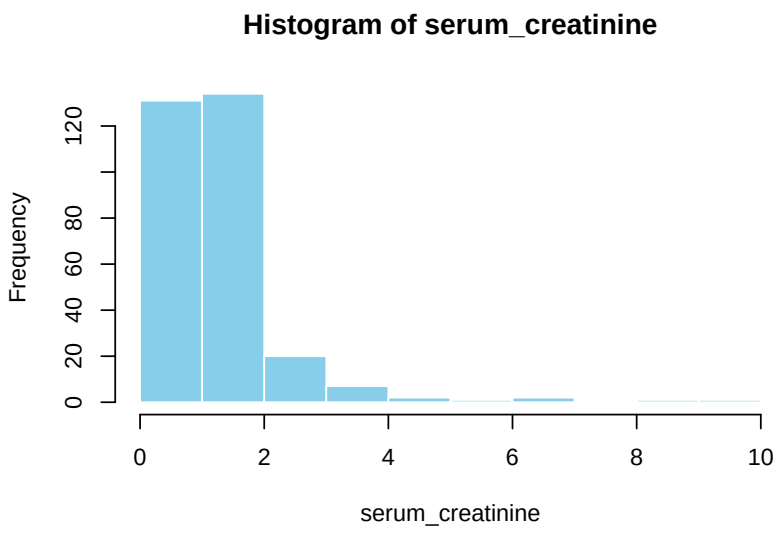
- Objective: Assess whether continuous predictor variables are normally distributed.
- Variables analyzed: Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine, Serum Sodium.
- Testing residuals ensures the model fits the data well and detects patterns that might indicate a need for model improvement.
- Most variables were either right or left skewed, showing low normality, except for platelets and serum sodium, which were closer to normal
- Tests performed:

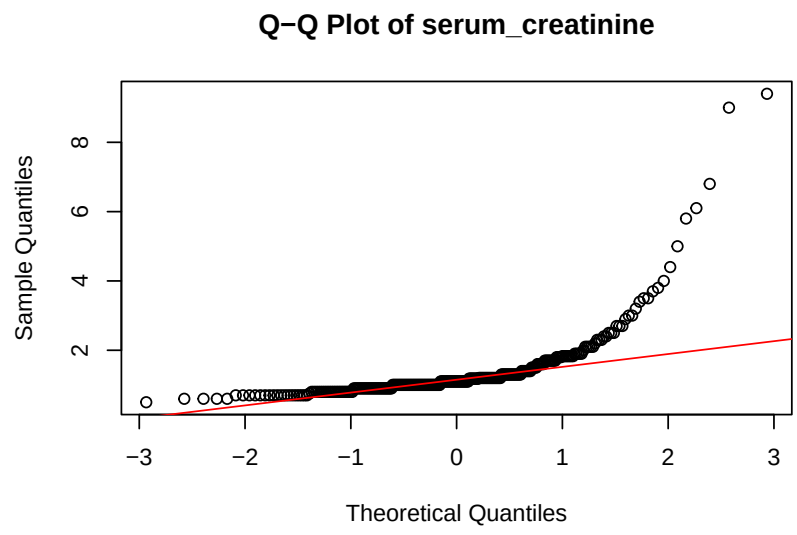
Shapiro-Wilk Test: p-values < 0.05 for all variables $\rightarrow$ None follow a normal distribution.

Graphs


Histogram of age
Frequency
0
10
20
30
40
50
40
50
60
70
80
90
age


Q−Q Plot of age
Sample Quantiles
40
50
60
70
80
90
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram of serum_creatinine
Frequency
0
20
40
60
80
120
0
2
4
6
8
10
serum_creatinine

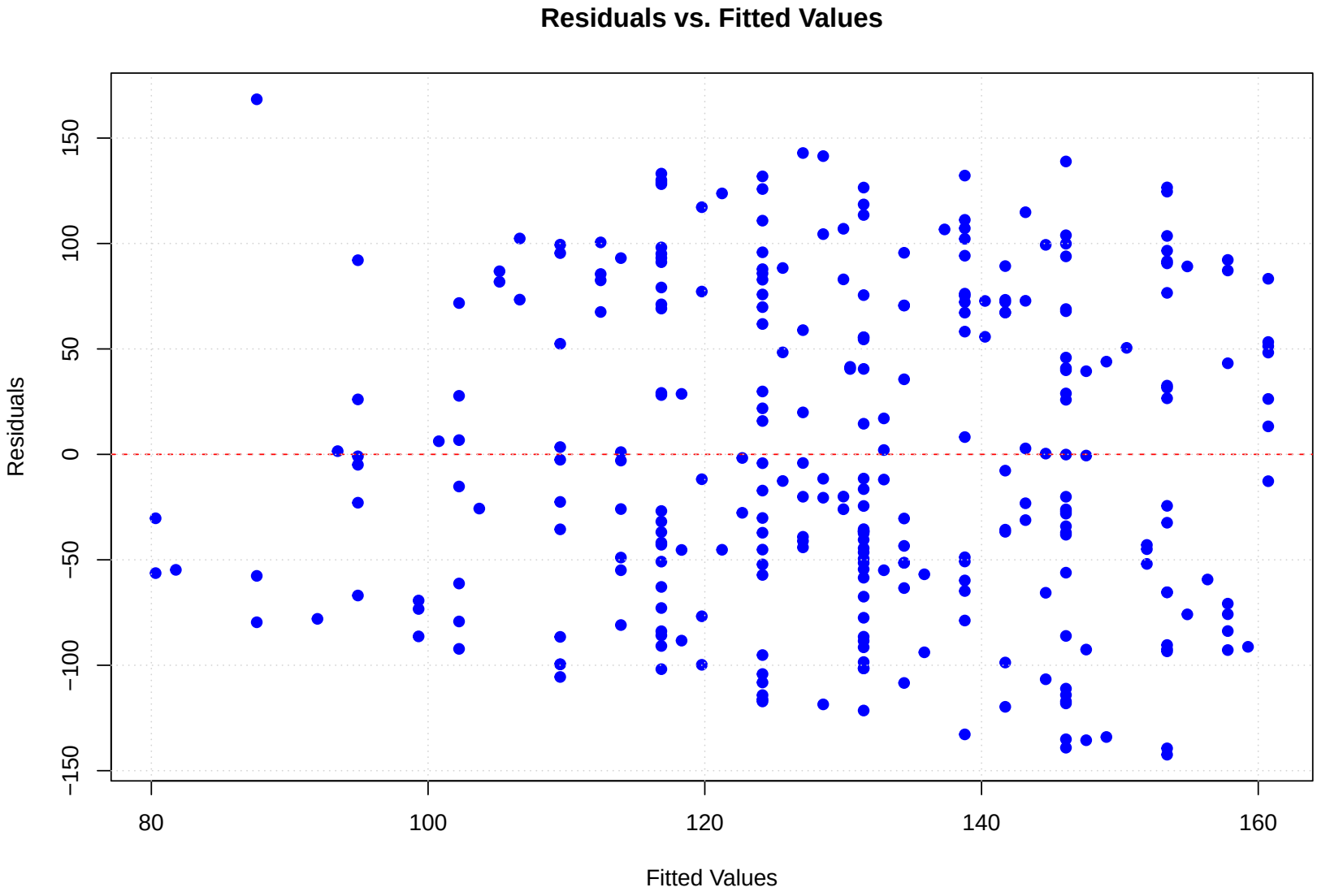

Q−Q Plot of serum_creatinine
Sample Quantiles
2
4
6
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Linearity

- Linearity assumes a straight-line relationship between the predictor variables and the log odds of the outcome in logistic regression
- Fitted linear trendlines show linear relationships for most variables.
- Logistic regression results:

The residuals exhibit no pattern, indicating that the relationship between the predictors and the response variable is pretty linear.

Graphs


Residuals vs. Fitted Values
Residuals
Fitted Values
150
100
50
0
−50
−100
−150
80
100
120
140
160

Polynomial Model to Test Linearity

- Polynomial models allow us to model nonlinear relationships between predictors and the outcome
- The polynomial model identified several significant predictors of the likelihood of a death event, including age, serum creatinine, ejection fraction, and time. However, there were other factors like anaemia, diabetes, and smoking, were found to be statistically insignificant

Logistic Regression

To begin tackling this binary classification problem, we first turned to logistic regression. We created 3 models to assess predictive performance with different regularization methods.

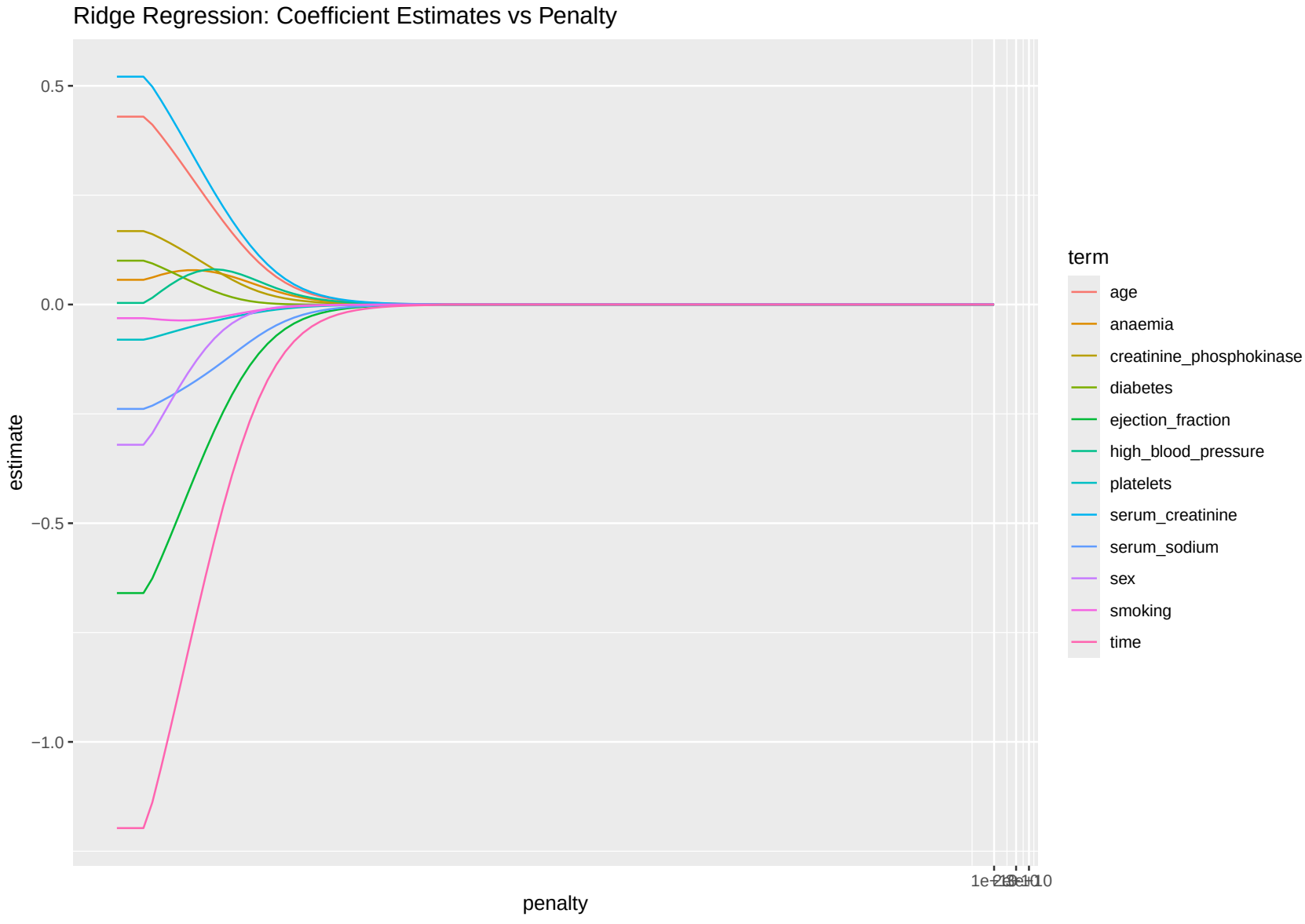
- Logistic Regression with no regularization
- Ridge Regression
- Lasso

Logistic Regression Estimates

	Estimate	Std..Error	p.value
(Intercept)	-1.034	0.487	0.034
age	0.915	0.242	0.000
anaemia	-0.059	0.414	0.886
creatinine_phosphokinase	0.429	0.230	0.062
diabetes	0.286	0.394	0.468
ejection_fraction	-0.897	0.212	0.000
high_blood_pressure	-0.253	0.418	0.544
platelets	-0.207	0.213	0.333
serum_creatinine	0.676	0.213	0.001
serum_sodium	-0.185	0.194	0.342
sex	-0.702	0.472	0.137
smoking	-0.041	0.472	0.931
time	-1.715	0.277	0.000

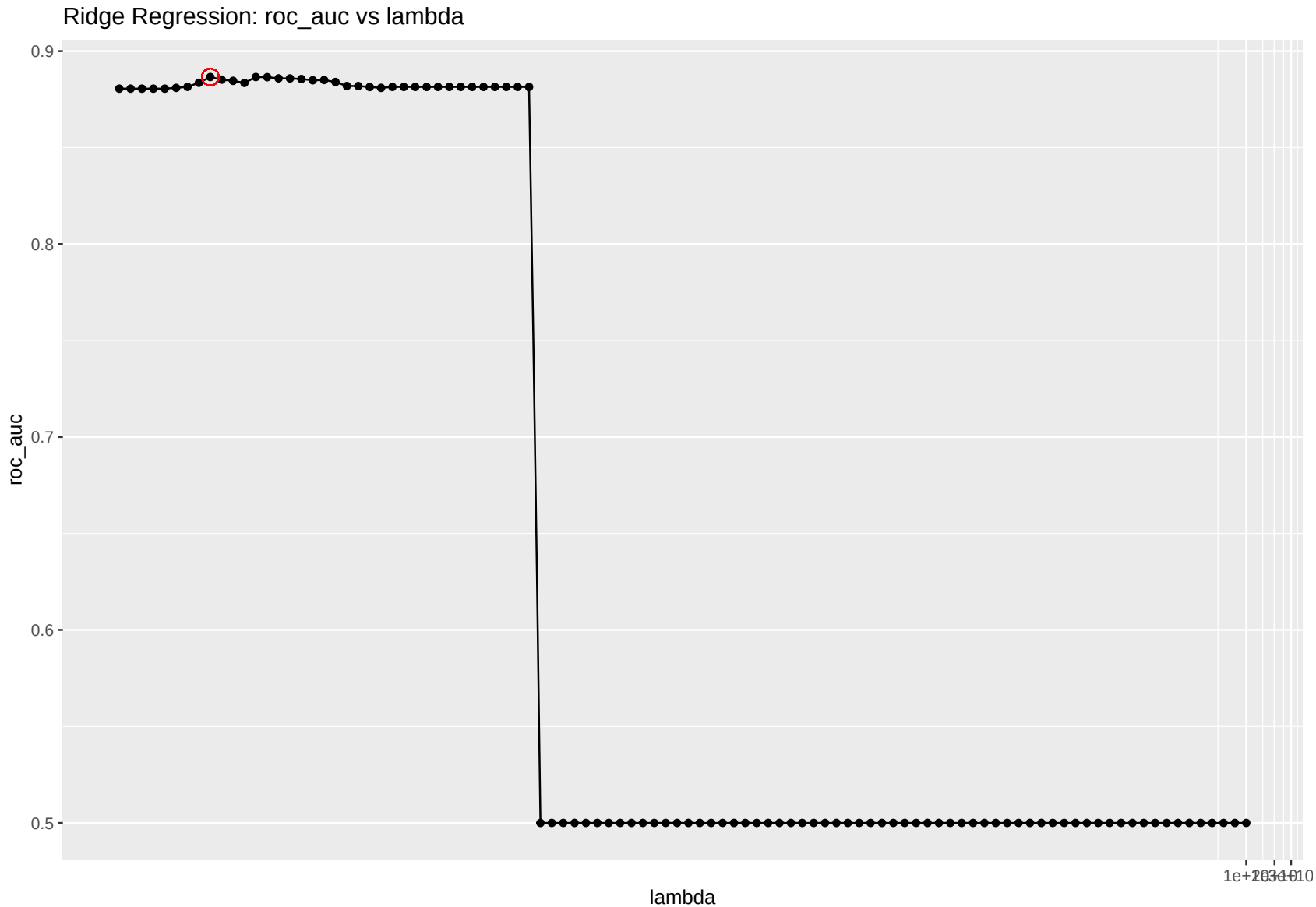
Significant Features: age, ejection fraction, serum creatine, time

Ridge Regression



Ridge Regression

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



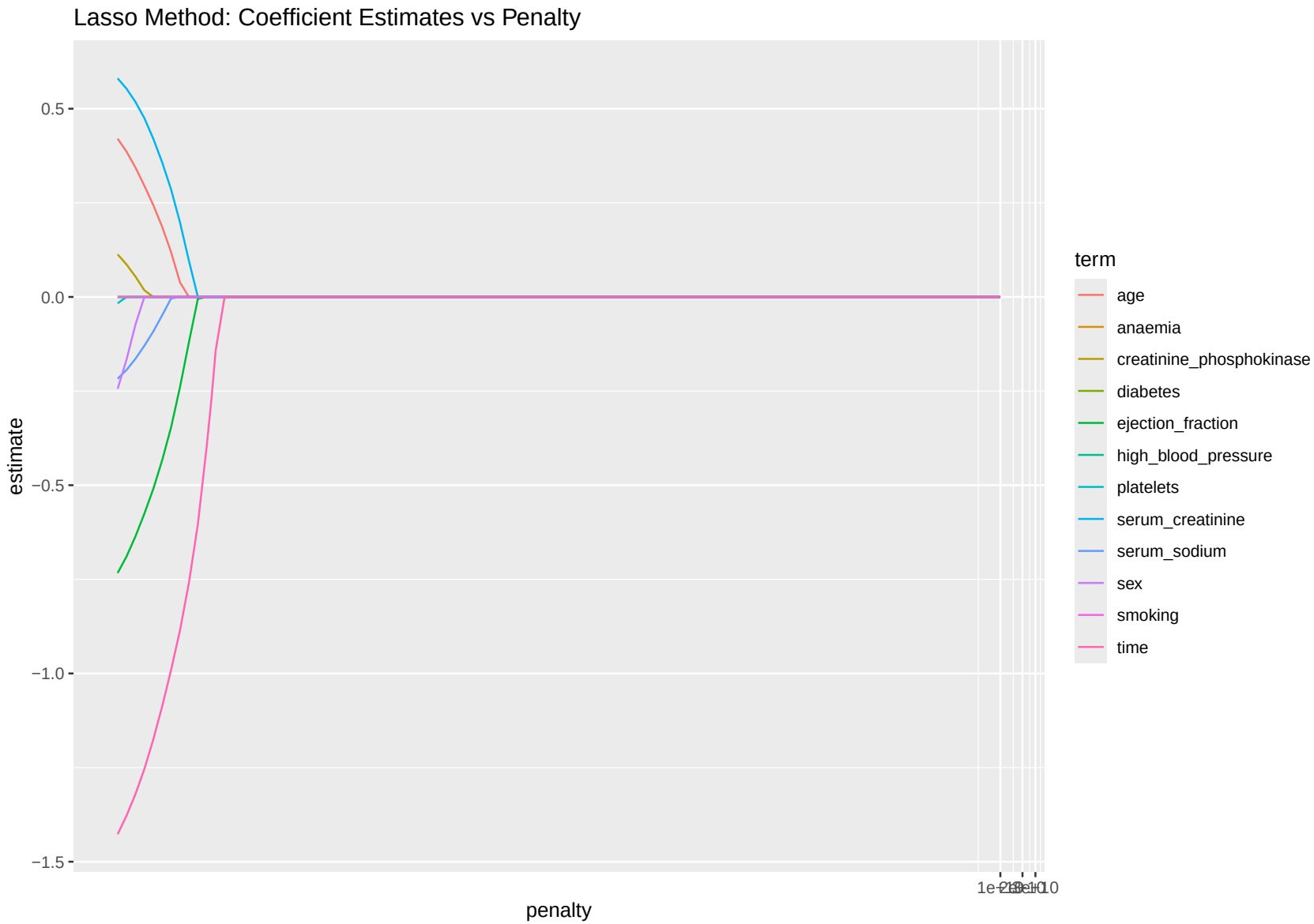
Ridge Regression Estimates

	Estimate	Std..Error	p.value
(Intercept)	-1.034	0.487	0.034
age	0.915	0.242	0.000
anaemia	-0.059	0.414	0.886
creatinine_phosphokinase	0.429	0.230	0.062
diabetes	0.286	0.394	0.468
ejection_fraction	-0.897	0.212	0.000
high_blood_pressure	-0.253	0.418	0.544
platelets	-0.207	0.213	0.333
serum_creatinine	0.676	0.213	0.001
serum_sodium	-0.185	0.194	0.342
sex	-0.702	0.472	0.137
smoking	-0.041	0.472	0.931
time	-1.715	0.277	0.000

Significant Features: age, ejection fraction, serum creatine, time

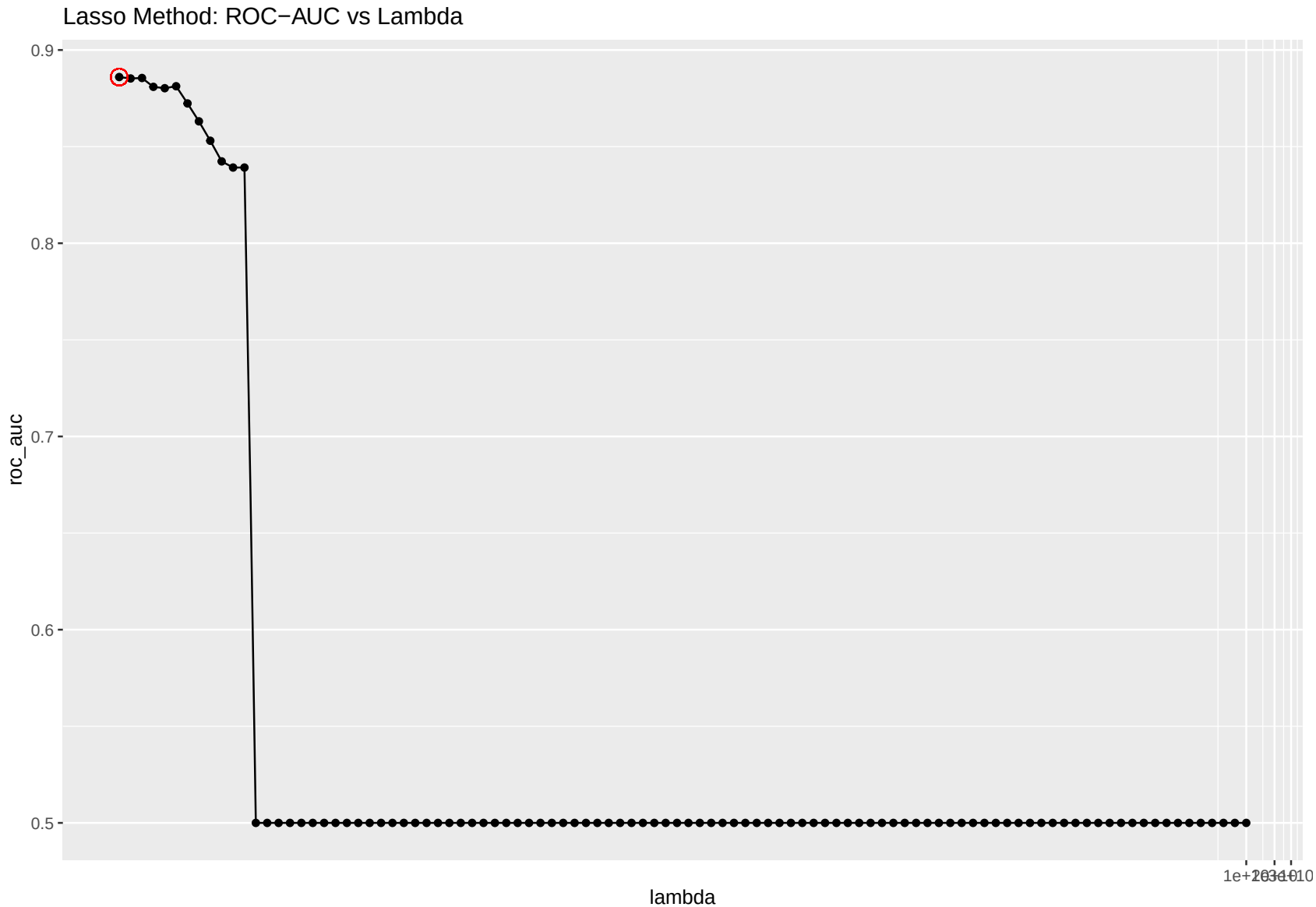
Lasso

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



Lasso

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



Lasso Estimates

	Estimate	Std..Error	p.value
(Intercept)	-1.034	0.487	0.034
age	0.915	0.242	0.000
anaemia	-0.059	0.414	0.886
creatinine_phosphokinase	0.429	0.230	0.062
diabetes	0.286	0.394	0.468
ejection_fraction	-0.897	0.212	0.000
high_blood_pressure	-0.253	0.418	0.544
platelets	-0.207	0.213	0.333
serum_creatinine	0.676	0.213	0.001
serum_sodium	-0.185	0.194	0.342
sex	-0.702	0.472	0.137
smoking	-0.041	0.472	0.931
time	-1.715	0.277	0.000

Significant Features: age, ejection fraction, serum creatine, time

Results

All Models found the same 4 predictor variables significantly correlated to the response.

No Regularization - ROC_AUC = 0.145
- Accuracy = 0.833

Ridge Regression - ROC_AUC = 0.144
- Accuracy = 0.8

Lasso - ROC_AUC = 0.134
- Accuracy = 0.833

Very similar performance metrics. Regularization appears to be unnecessary for this data set. Low ROC-AUC values indicate poor discrimination ability. It is likely that accuracy is being driven up erroneously by predictions being assigned to the majority class.

Support Vector Machine

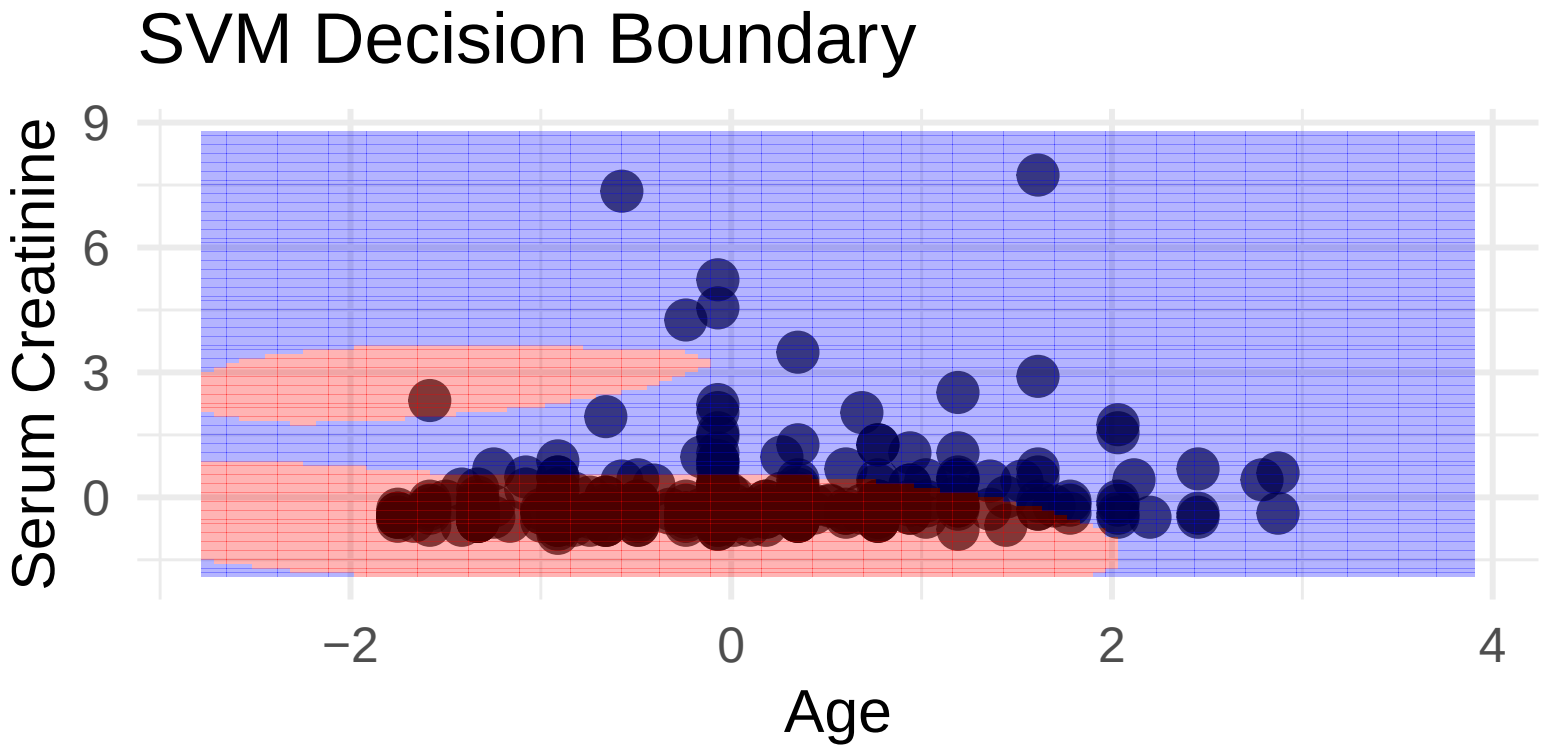
We began by finding the accuracy rating of 81.7% and a prediction matrix

```
##           Actual
## Predicted  0  1
##         0 36  6
##         1  5 13
```

```
## [1] "Accuracy:  0.817"
```

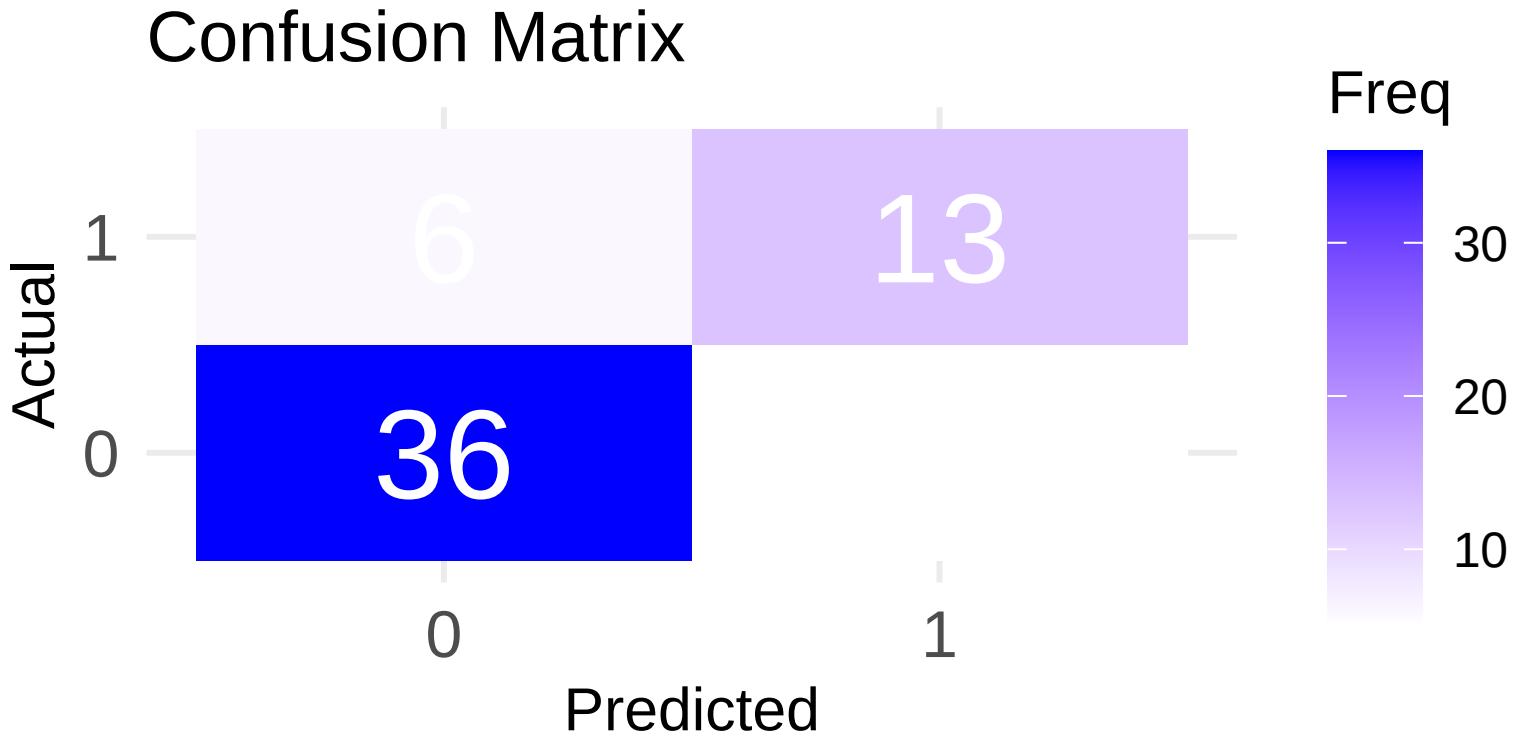
SVM Descision Boundry

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.



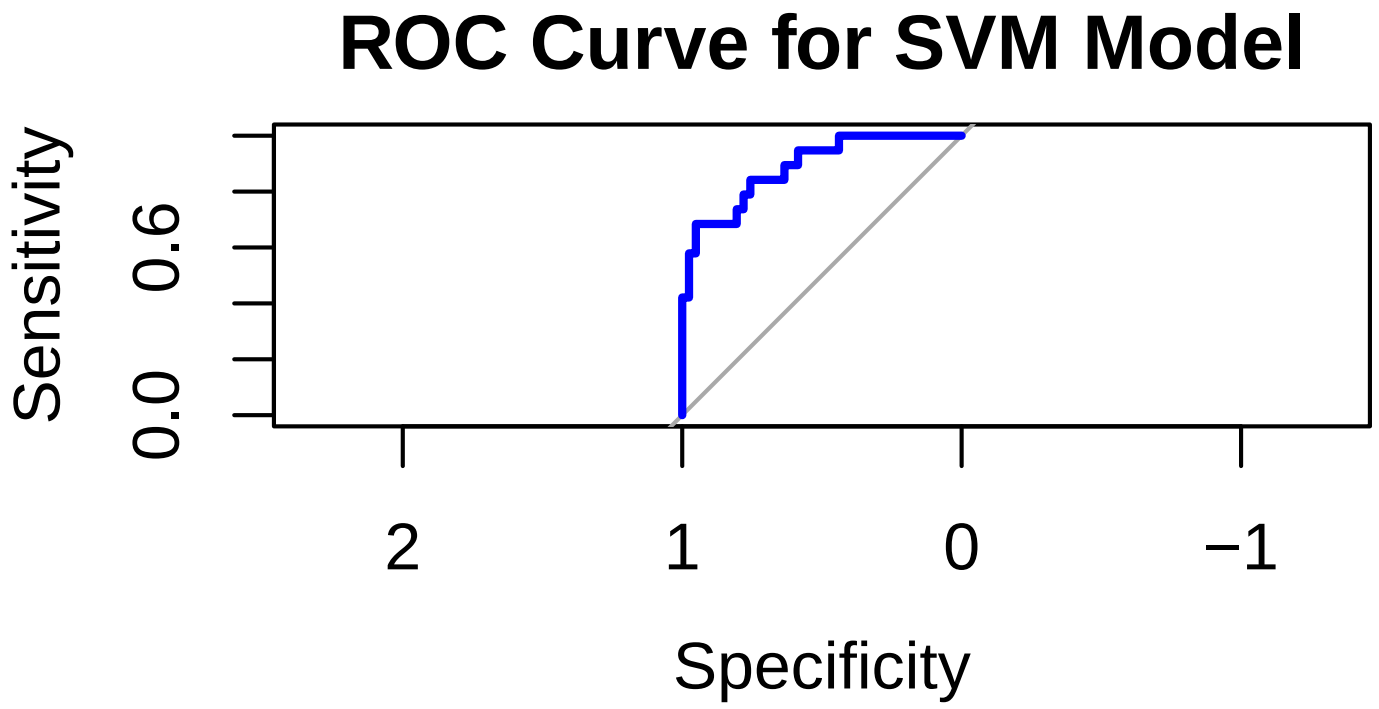
SVM Confusion Matrix

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 `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

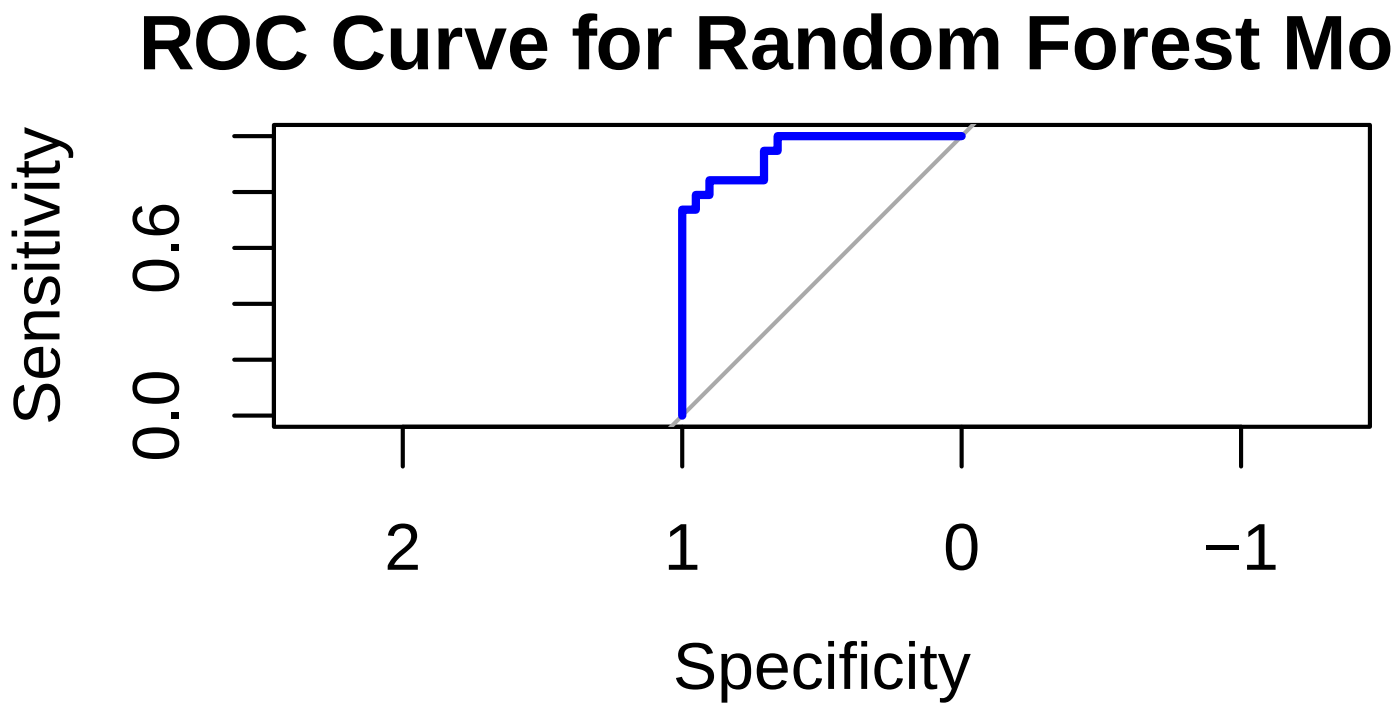
```
## [1] 0
```

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

```
## Accuracy:  0.9
```

RF ROC Curve

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



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