DSCI 445 Project Paper

Paige Galvan, Neha Deshpande, & Witlie Leslie

2024-11-25

Motivation

The goal of our project is to predict mortality from heart failure using behavioral risk factor data. Heart failure is a disease that affects millions of people yearly. Although modern medicine has improved, it can be hard to determine causes of heart failure due to how many variables can affect it. The Heart Failure Clinical Records Dataset provides a collection of medical indicators such as age, ejection fraction, serum creatinine, and co-existing conditions like diabetes and high blood pressure. By analyzing this data, researchers can uncover patterns that contribute to better understanding the progression of heart failure.

The main motivation for our group to study this dataset is to dive a little bit deeper into which factors affect heart failure. Knowing that heart failure is a leading cause of death around the world, finding meaningful patterns can inform public health strategies, such as targeted lifestyle modifications or health care campaigns. The main objective is to transform this raw data into meaningful conclusions on heart disease.

Methodology

Exploratory Analysis

Before applying machine learning models, we began by performing an exploratory analysis of the data. This included assessing the linearity and normality of the predictors, identifying any outliers, and exploring potential correlations among the variables. We visualized distributions using histograms and box plots to understand the spread of each feature, and scatter plots to check the relationships between the predictor variables and the target variable (mortality). This helped us determine whether the data required transformations before applying machine learning techniques.

Linearity and Normality

In this analysis, we begin by testing the normality of several continuous variables in the dataset, including age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, and serum sodium. First, we used histograms and Q-Q plots to visually inspect the distribution of these variables. The histograms for most variables indicated skewness, either to the right or left, suggesting that these variables do not follow a normal distribution. Platelets and serum sodium appeared to be the most normally distributed, but even they showed some deviation from normality. The Q-Q plots confirmed these observations, showing that age, ejection fraction, serum sodium, and platelets were closer to a normal distribution, while other variables exhibited greater deviations.

To check the normality of the continuous variables, we performed the Shapiro-Wilk test, and all p-values were below 0.05, indicating that none of the variables followed a normal distribution.

We then examined the relationship between these variables and the binary outcome, DEATH_EVENT, using scatter plots with linear regression lines. These plots showed that variables like age and serum creatinine were

somewhat associated with the likelihood of death, showing linear trends in most cases. Logistic regression models confirmed that many of the variables, such as age and serum creatinine, were significantly related to the risk of death.

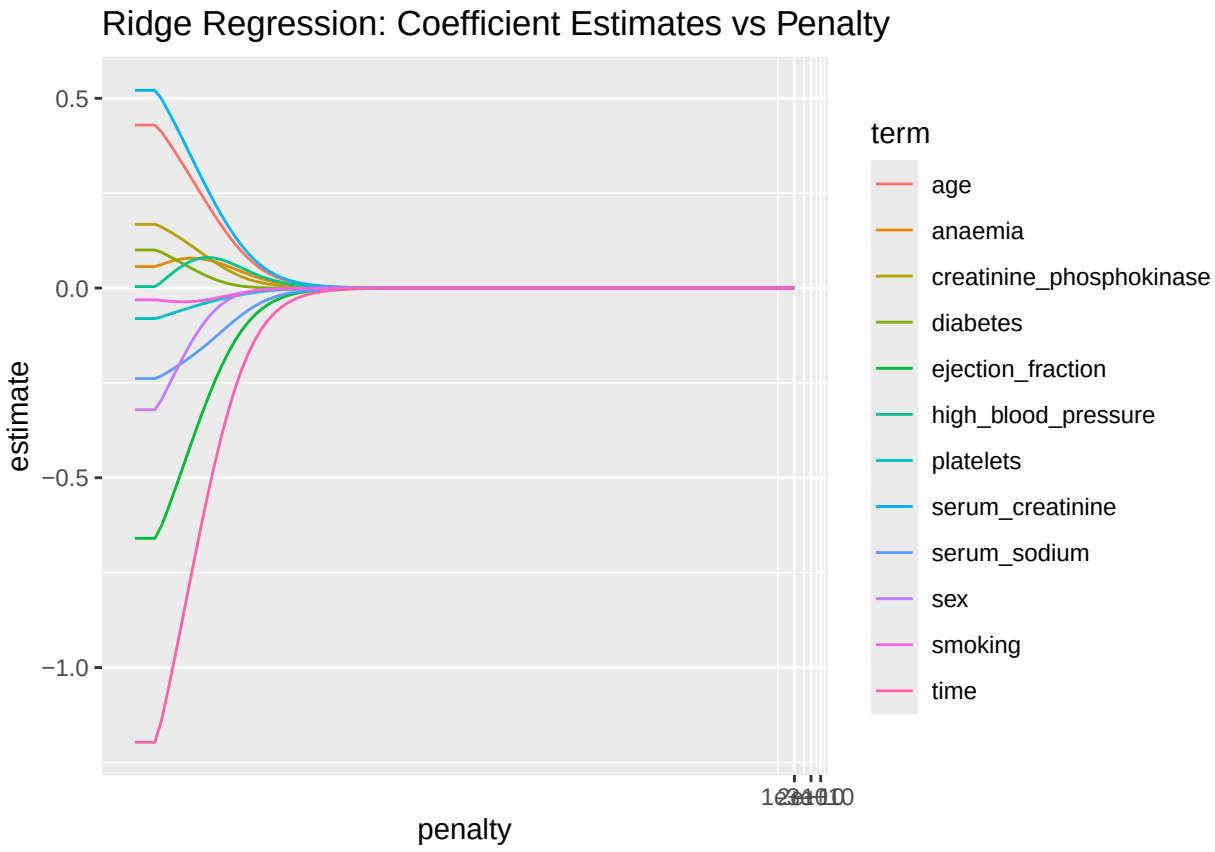
Logistic Regression with Regularization Logistic regression is a go-to method for binary classification, and we explored three versions to analyze predictive performance. First, we fit a basic logistic regression model without regularization as a baseline. While simple, it doesn't handle collinearity or irrelevant predictors. Next, we applied ridge regression regularization, which penalizes large coefficients to stabilize the model, though it doesn't eliminate predictors, making it less interpretable than Lasso. Finally, we used Lasso regularization, which not only penalizes coefficients, but also performs feature selection by shrinking some to zero, improving interpretability. Comparing their predictive power helps determine which approach balances accuracy and simplicity best.

Because the predictor variables are of varying ranges and units, we began by scaling all continuous features to prevent our regularization techniques from over-penalizing variables with larger ranges. Next, we split our data into a training set (containing 80% of the data) and a test set (containing 20%) so that we could assess the performance of our logistic regression models using 10-fold cross validation.

	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

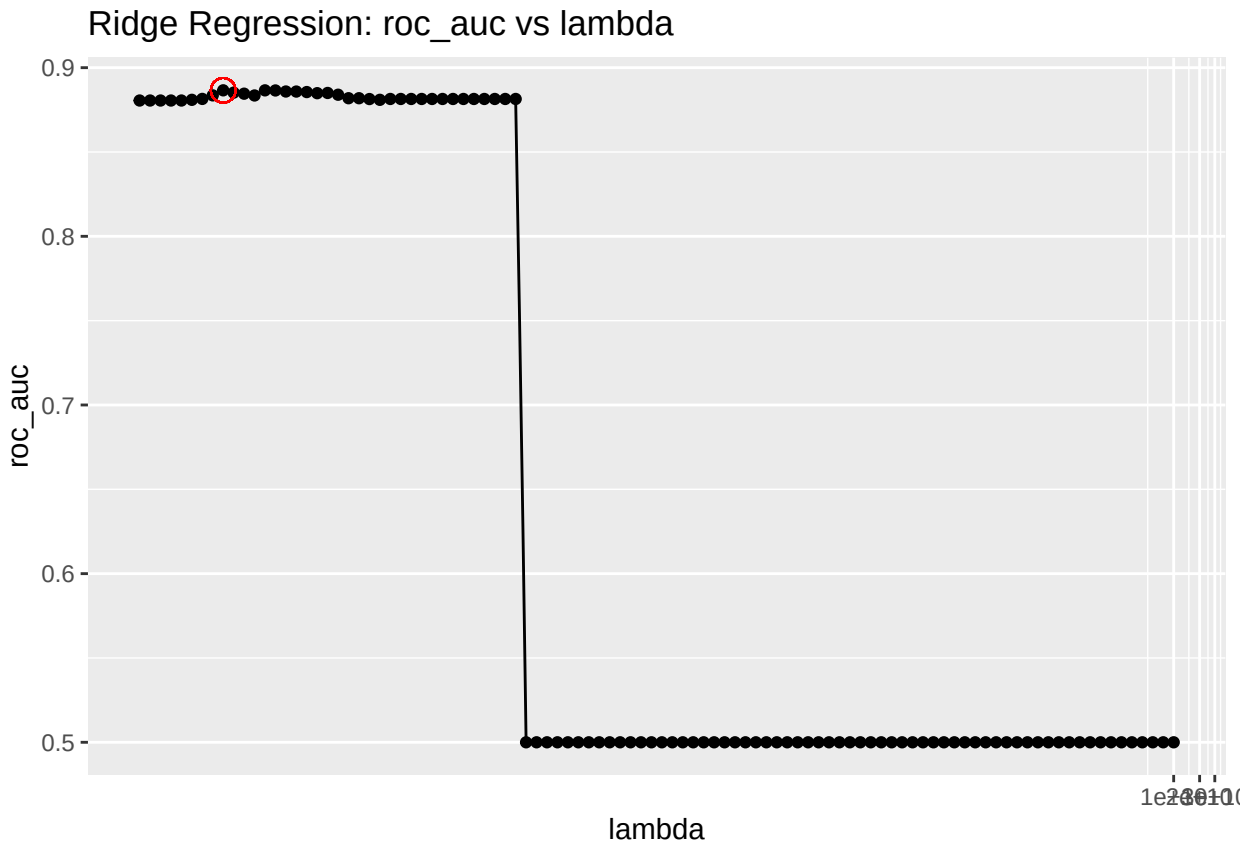
The logistic regression model found that four predictors displayed significance: age, ejection fraction, serum creatine, and time.

Next, we created a logistic regression model with ridge regression regularization. We determined the optimal lambda penalty value using 10-fold cross validation.



The graph above depicts the estimated value of each coefficient for each lambda value tested. As the lambda penalty value increases, the coefficient estimates gradually diminish towards zero without being removed entirely.

The graph below depicts the ROC-AUC value for each lambda tested. 10-fold cross validation determined the best lambda to be *INSERT LAMBDA*

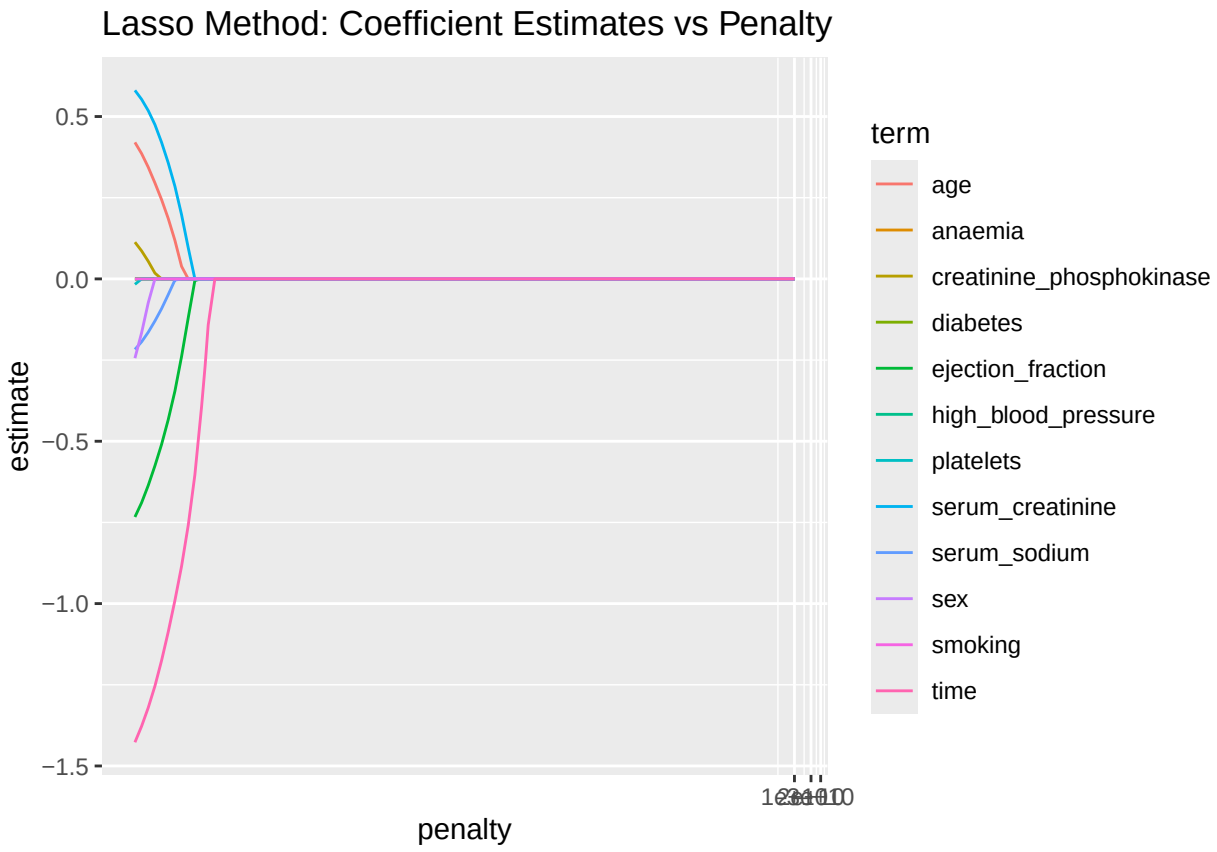


Once we have chosen the optimal penalty, we can fit the ridge regression model with this lambda value and view its estimates.

	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

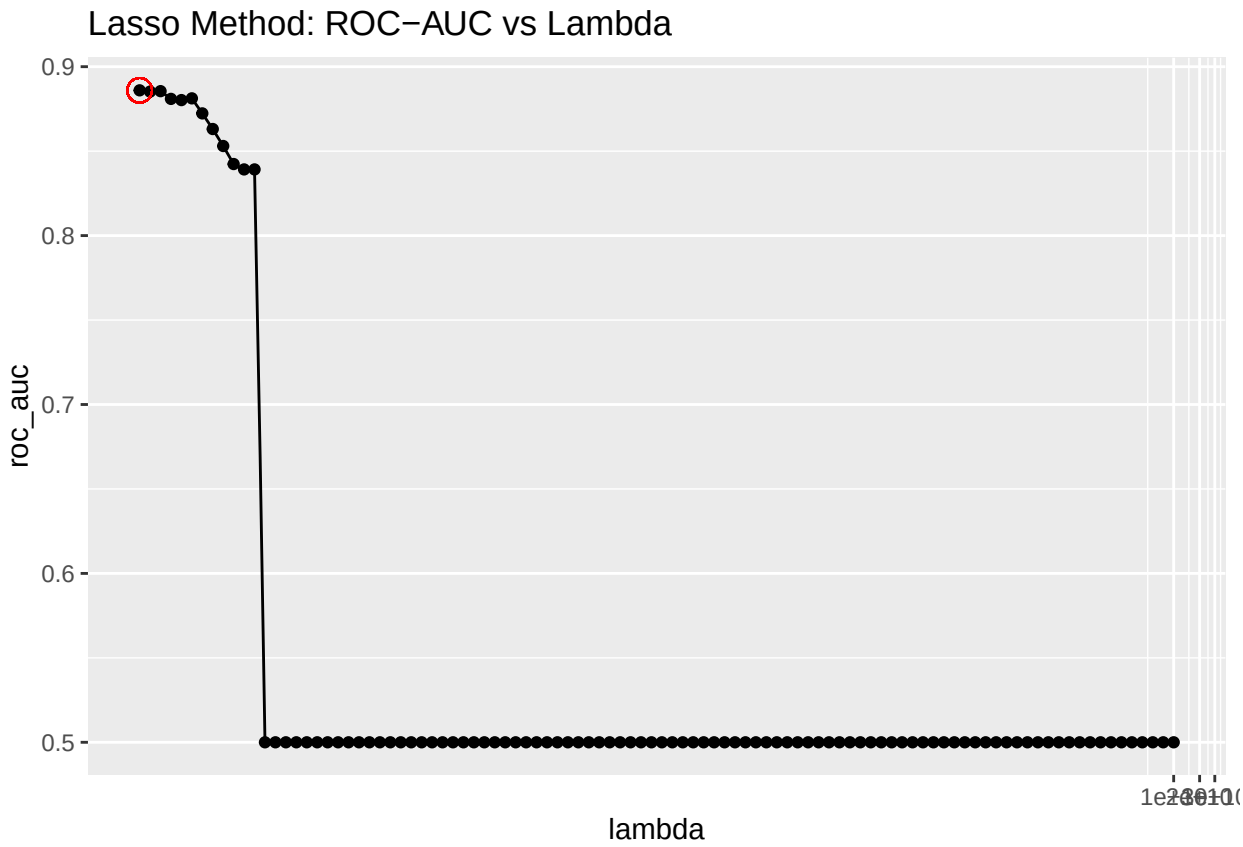
The ridge regression model found the same four predictor variables significant as the logistic regression model without regularization: age, ejection fraction, serum creatine, and time. The coefficient estimates and p-values are overall very similar to those of the first logistic regression model.

For our final logistic regression model, we will implement lasso, once again using 10-fold cross validation to determine the optimal lambda penalty value.



In the graph above, we can see the coefficient estimates diminish towards zero as lambda increases. The reduction in coefficient estimates is notably steeper for lasso than with ridge regression. Unlike ridge regression, lasso performs feature selection by driving some coefficients to equal zero, thus eliminating them from the model.

The graph below depicts the ROC-AUC value for each lambda tested. 10-fold cross validation determined the best lambda to be *INSERT LAMBDA*



Now with our chosen penalty, we can fit the lasso model with this lambda value and view its 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

The lasso model again found the same four predictor variables significant as both the logistic regression and ridge regression models: age, ejection fraction, serum creatine, and time. The coefficient estimates and p-values are again very similar to those of the logistic regression and ridge regression models.

Because the ridge regression model and the lasso model resulted in estimates very similar to those of the logistic model without regularization, this indicates that regularization was likely unnecessary for our data set.

Now we will investigate the predictive performance of our three logistic regression models. The metrics we used to assess the performance of these models are ROC-AUC—the area under the receiver-operating

characteristic (ROC) curve that represents the probability that the model will correctly rank a randomly selected positive example higher than a negative one—as well as accuracy, which is the proportion of correct predictions out of all total predictions.

REFORMAT OUTPUTS BELOW

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 roc_auc  binary         0.144
## 2 accuracy binary         0.8
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 roc_auc  binary         0.134
## 2 accuracy binary         0.833
```

DISCUSS RESULTS

Support Vector Machine Support Vector Machines (SVMs) are primarily binary classifiers. When dealing with more than two classes, SVMs can handle multi-class classification by applying techniques like "one-vs-one" or "one-vs-all," where multiple binary classifications are combined. One of the key advantages of SVMs is their ability to perform non-linear classification, which increases their flexibility and allows them to handle complex decision boundaries. This handels linear and non-linear decision boundaries. Using a linear kernel is good for approximately linear relationships, which our data is. It does not assume any specific distribution of predictors since none of our predictors are normal.

We began with our SVM model looking to the Support Vectors and the parameters of interest. The support vector are important to understanding the model's decisions. They are the informative points and make it the most critical for classification. They demonstrate the most ambiguous points of data.

From this code we conclude that there are only 13 predictive variables we should consider. We also considered the accuracy of how the SVM model this allows us to understand its predictive capabilities. We also conclude that this model will have an accuracy of 0.817 this indicates the model has a 81.7% of making correct predictions.

- The SVM Boundry analyzes the age vs serum creatinine using the death effect for males vs females. We can see that overall serium creatinine has little effect on the death effect especially in males. Meanwhile age and serum creatinine has a large response on females.
- 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 Random Forest is a powerful machine learning algorithm used for both classification and regression. It works by building multiple decision trees and aggregating their predictions to improve accuracy and reduce overfitting. Key advantages include its ability to handle complex, non-linear relationships, manage missing data, and automatically capture feature interactions. Random Forest is also robust to overfitting, particularly compared to individual decision trees, and provides a built-in estimate of model performance through out-of-bag error. Additionally, it offers valuable insights into feature importance, helping to identify which variables most influence the outcome. Overall, Random Forest is particularly effective for high-dimensional datasets, imbalanced classes, and when model interpretability is secondary to prediction accuracy.

- 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.

«««< HEAD

======= # 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. ROC-AUC values of 0.10 - 0.13 indicate that the model is performing worse than chance (0.5). A low ROC-AUC value indicates poor discrimination ability. Having a high accuracy value with a low ROC-AUC value might suggest that accuracy is being misleadingly driven up by the model assigning predictions to the majority class (in our case, death event = 0). »»»> 06e66d4c39f850846dcef0b1529f6760960600e5

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