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.

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 cross validation.

Once this setup was complete, we created our first logistic regression model with no regularization. The metrics we used to assess the performance of these models are roc_auc, which is 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. Below are the roc_auc and accuracy values of the first logistic regression model with no regularization:

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 roc_auc  binary         0.145
## 2 accuracy binary         0.833
```

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.

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.

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##
```

```
## Attaching package: 'e1071'

## The following object is masked from 'package:tune':
##
##     tune

## The following object is masked from 'package:rsample':
##
##     permutations

## The following object is masked from 'package:parsnip':
##
##     tune

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

##
## Call:
## svm(formula = DEATH_EVENT ~ ., data = train_data, kernel = "radial",
##     cost = 1, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  142
##
##  ( 74 68 )
##
##
## Number of Classes:  2
##
## Levels:
```

```
##  0 1
```

```
##          Actual
## Predicted  0  1
##         0 36  9
##         1  4 11
```

```
## [1] "Accuracy:  0.783"
```

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 inidcates the model.

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.

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