Heart Failure Risk Prediction using Logistic Regression, LASSO, Random Forest and Boosted Trees.

Norris Jaber, Rajeev Agrawal, Sha Lu, Yuxi Duan, Jen-Yu Huang

4/22/2021

1. Business Question and Case

1.1 Business Question

What factors increase the risk of death due to heart failure?

1.2 Business Case

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

2. Analytics Question

2.1 Outcome Variable of Interest

Our outcome variable of interest is whether a person died of heart failure. This is represented as a binary variable called "DEATH_EVENT" in our data set, where 0 = no death and 1 = death.

2.2 Main Predictors

The key predictors of our model include demographic information such as age and gender. This will allow us to see which demographic is more likely to die from heart failure. We will also be including health related factors such as if the patient has anemia, high blood

pressure, diabetes, as well as if the patient is a smoker. These predictors are important as it will show us which health related variables are more likely to cause heart failure in patients. Knowing these will allow caregivers and doctors know which of their patients are most at risk. We will also be including more specific predictors such as level of creatinine in the blood and level of sodium for example, which will allow us to see more a greater spectrum of patient health. Such predictors are important as they may show how close a patient is to death via heart failure and will allow caregivers to give more precise treatment.

3. Data set Description

For this study, the data set was obtained from Kaggle [1]. This is a data set of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April – December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All the patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure [2].

4. Exploratory Data Analysis

4.1 Variables

The data set contains total 299 records with the following 13 variables (Outcome variable - *DEATH_EVENT*):

- **Categorical variables**
  - *anaemia:* Decrease of red blood cells or hemoglobin; 1 = Yes, 0 = No.
  - *high_blood_pressure:* If a patient has hypertension; 1 = Yes, 0 = No.
  - *diabetes:* If the patient has diabetes; 1 = Yes, 0 = No.
  - *sex:* 1 = Male, 0 = Female.
  - *smoking:* If the patient smokes; 1 = Yes, 0 = No.
  - *DEATH_EVENT:* If the patient died during the follow-up period; 1 = Yes, 0 = No.
- **Quantitative variables**
  - *age:* Age of the patient in years.

- *creatinine_phosphokinase:* Level of the CPK enzyme in the blood in micrograms/L.
- *ejection_fraction:* Percentage of blood leaving the heart at each contraction.
- *platelets:* Platelets in the blood in kiloplatelets/mL.
- *serum_creatinine:* Level of creatinine in the blood in mg/dL.
- *serum_sodium:* Level of sodium in the blood in milliequivalents/L.
- *time:* Follow-up period in days.

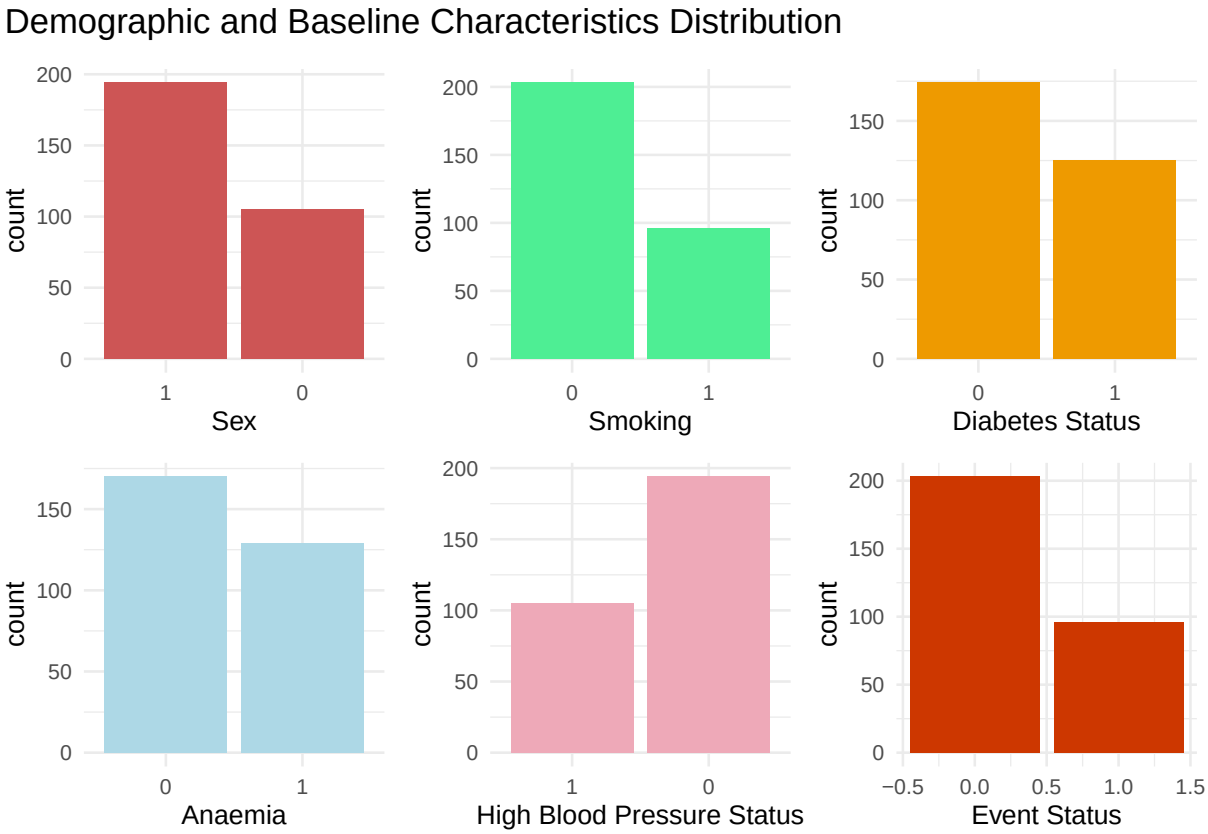
4.2 Descriptive Analytics

There are no missing values in the data set, so we do not need to do any imputation.

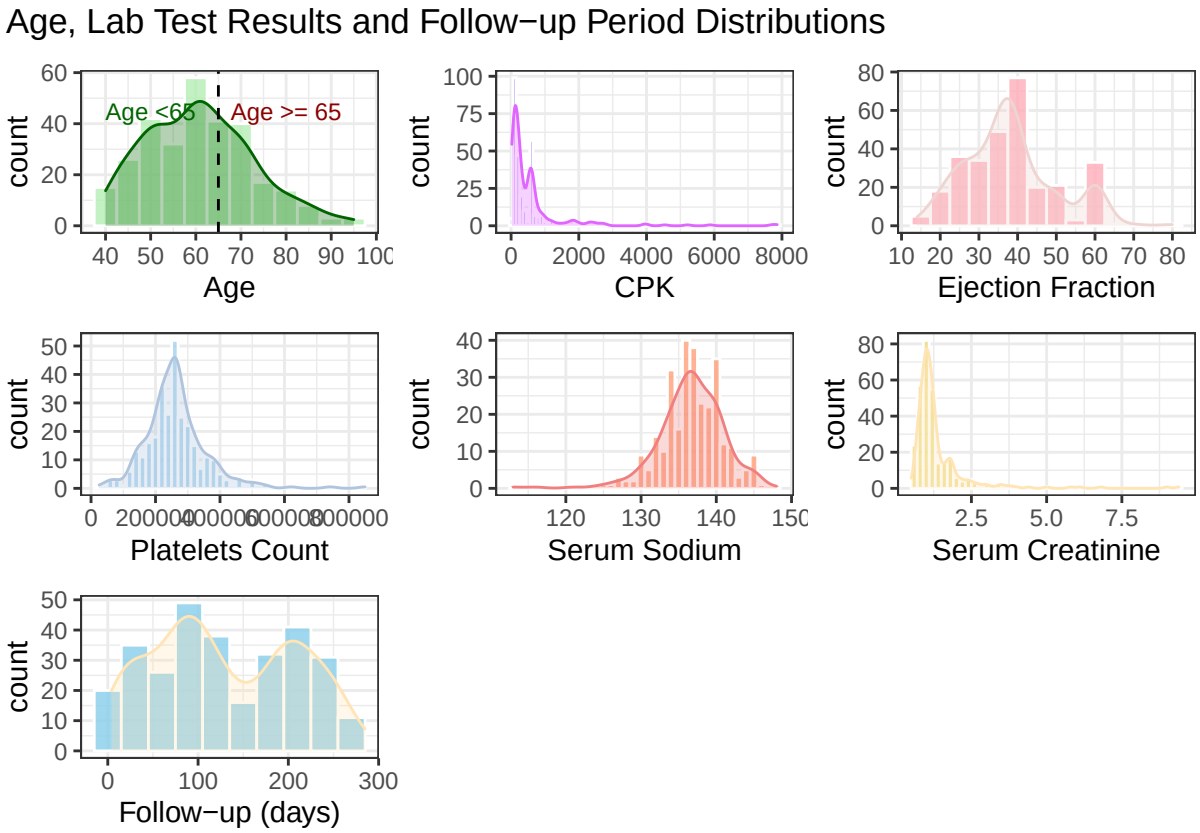
4.2.1 Quick Summary

```
##       age        anaemia creatinine_phosphokinase diabetes ejection_fraction
##  Min.   :40.00   0:170   Min.   :  23.0           0:174    Min.   :14.00
##  1st Qu.:51.00   1:129   1st Qu.: 116.5           1:125    1st Qu.:30.00
##  Median :60.00           Median : 250.0                     Median :38.00
##  Mean   :60.83           Mean   : 581.8                     Mean   :38.08
##  3rd Qu.:70.00           3rd Qu.: 582.0                     3rd Qu.:45.00
##  Max.   :95.00           Max.   :7861.0                     Max.   :80.00
##  high_blood_pressure   platelets      serum_creatinine  serum_sodium   sex
##  1:105               Min.   : 25100   Min.   :0.500    Min.   :113.0   1:194
##  0:194               1st Qu.:212500   1st Qu.:0.900    1st Qu.:134.0   0:105
##                      Median :262000   Median :1.100    Median :137.0
##                      Mean   :263358   Mean   :1.394    Mean   :136.6
##                      3rd Qu.:303500   3rd Qu.:1.400    3rd Qu.:140.0
##                      Max.   :850000   Max.   :9.400    Max.   :148.0
##  smoking      time       DEATH_EVENT
##  0:203   Min.   :  4.0   1: 96
##  1: 96   1st Qu.: 73.0   0:203
##          Median :115.0
##          Mean   :130.3
##          3rd Qu.:203.0
##          Max.   :285.0
```

4.2.2 Binary Variables Distribution



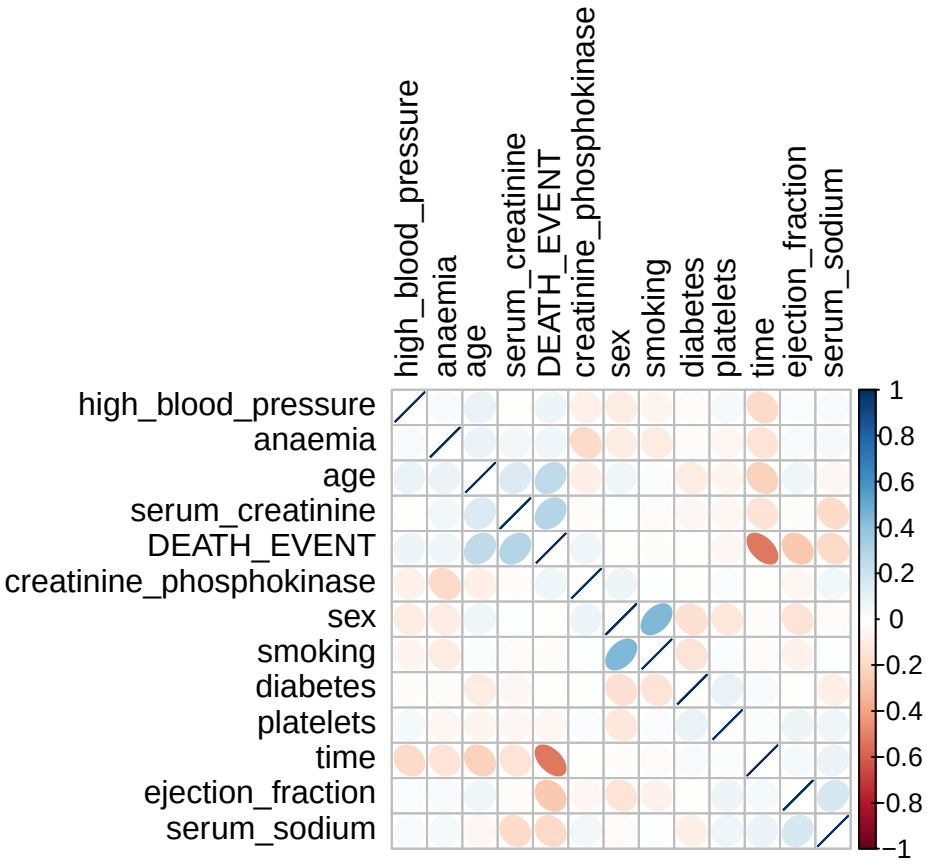
4.2.3 Continuous Variables Distribution



4.3 Correlations

4.3.1 Correlation Matrix

From the correlation matrix, we can see *Death Event* is highly correlated (point-biserial correlation) with follow-up duration, serum creatinine, age, serum sodium, and ejection fraction.


high_blood_pressure
anaemia
age
serum_creatinine
DEATH_EVENT
creatinine_phosphokinase
sex
smoking
diabetes
platelets
time
ejection_fraction
serum_sodium
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

4.4. Assumption Tests

Binomial Logistic Regression Model

Since, our outcome variable (DEATH_EVENT) is binary, the preliminary model that we will use is the logistic regression model.

4.4.1 Assumptions

1. The dependent variable is categorical (binary).
2. The observations are independent of each other.
3. There is no severe multicollinearity among the explanatory variables.
4. There are no extreme outliers.
5. The independent variables are linearly related to the log odds.
6. The sample size of the dataset is large enough to draw valid conclusions from the fitted logistic regression model.

Out of the above 6 assumptions, the 3rd assumption about multicollinearity will be tested using condition index (CI) and variance inflation factor (VIF). There is no evidence to suggest that the remaining 5 assumptions are violated.

4.4.2 Training and Evaluating the Model

We see based on the p-values in the following summary output, not all of the features in this full model are significant. Variables like age, ejection fraction, serum creatinine and time (follow-up period) are significant (p-values < 0.05), while other predictors such as anaemia, diabetes, high_blood_pressure, platelets, serum_sodium, sex and smoking are not significant (p-values > 0.05).

```
##
## Call:
## glm(formula = DEATH_EVENT ~ ., family = binomial(link = "logit"),
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1848  -0.5706  -0.2401   0.4466   2.6668
##
## Coefficients:
##                             Estimate   Std. Error z value Pr(>|z|)
## (Intercept)               9.548591580  5.613380405   1.701 0.088935 .
## age                       0.047419074  0.015800632   3.001 0.002690 **
## anaemia1                 -0.007470452  0.360489086  -0.021 0.983467
## creatinine_phosphokinase  0.000222229  0.000177933   1.249 0.211684
## diabetes1                 0.145149775  0.351188640   0.413 0.679380
## ejection_fraction        -0.076662501  0.016329130  -4.695 2.67e-06 ***
## high_blood_pressure0      0.102679427  0.358706893   0.286 0.774688
## platelets                -0.000001200  0.000001889  -0.635 0.525404
## serum_creatinine          0.666093340  0.181492576   3.670 0.000242 ***
```

```
## serum_sodium             -0.066981072  0.039735098  -1.686 0.091855 .
## sex0                      0.533658016  0.413918039   1.289 0.197299
## smoking1                 -0.013492224  0.412617798  -0.033 0.973915
## time                     -0.021044626  0.003014394  -6.981 2.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 219.55  on 286  degrees of freedom
## AIC: 245.55
##
## Number of Fisher Scoring iterations: 6
```

4.4.3 Dealing with Multi-collinearity

Multi-collinearity is a problem because it makes it difficult to separate out the impact of individual predictors on response. We evaluate the overall multi-collinearity of the model using Condition Index (CI). If the model suffers from multi-collinearity (i.e. CI > 30), we will identify which predictors contribute the most to this collinearity condition using Variance Inflation Factor (VIF). A VIF of greater than 10 indicates the presence of severe multi-collinearity and requires remediation.

```
##  [1]   1.000000   2.911576   3.350974   3.945110   4.310660   4.766095
##  [7]   4.874041   5.807561   6.946561   9.359366  11.874013  21.468950
## [13] 136.493536
```

From the output, we can see that CI, which is the square root of the ratio of largest to the smallest Eigenvalue of the correlation matrix, is 136.5, much greater than 30, implying severe multi-collinearity. Therefore, we use the VIF to estimate the variance inflation contribution of each predictor.

```
##                      age                  anaemia creatinine_phosphokinase
##                 1.104307                 1.114540                 1.085629
##                 diabetes        ejection_fraction      high_blood_pressure
##                 1.052375                 1.172842                 1.063014
##                platelets         serum_creatinine             serum_sodium
##                 1.045319                 1.102088                 1.070685
##                      sex                  smoking                     time
##                 1.380635                 1.284512                 1.151810
```

None of the VIF values are greater than 10 (or even 5). It is likely that a predictor is highly correlated with the intercept. We will try centering the data to eliminate the intercept and check again.

4.5 Data Pre-processing and Transformations

4.5.1 Centering

With centering, only the intercept changes, the β coefficients and the p-values do not change.

```
##  [1] 1.000000 1.522747 1.670010 1.808062 1.832408 1.930541 1.987264 2.155060
##  [9] 2.196331 2.626945 2.853642 3.224645 5.608133
```

From the CI value of 5.6, we can see that centering helped. The CI came down drastically from the earlier value of 136.5 to 5.6. We can also use other ways to deal with multi-collinearity such as using shrinkage methods (Ridge, LASSO) or dimension reduction methods (PCR, PLS).

4.5.2 Log Transformation

Some of the continuous predictors such as CPK and Serum creatinine are right-skewed. One of the ways to make them closer to normal distribution is to take the logarithm. However, we have a sample size of 299 (50+ data points), therefore, the predictors do not have to be normally distributed. Therefore, we leave the continuous predictors without any transformation.

5. Modeling Methods and Model Specifications

The goal is inference so we are interested in methods where in addition to the prediction accuracy, we are also able to interpret the model.

5.1 Initial Model Specification

Logistic Regression - Full Model - h_full.

5.2 Initial OLS or Logit Model Results

5.3 Assumption Tests

5.4 Model Candidates and Rationale

Rationale - Ours is a classification problem. Goal is interpretation. So we choose models accordingly and compare their performance.

- Logit Model
- LASSO Model
- Random Forest
- Boosted Trees

5.5 Model Specification Candidates and Rationale

We end up with the following two model specifications across all models.

1. **Full model with all the variables:** The full logistic model has tolerable multicollinearity after centering, so it is a good model to consider since it contains all the variables and no information is lost.

2. **Smaller nested model with only the most significant predictors:** - age, ejection_fraction, serum_creatinine, and time (follow up time). We tried the quadratic model but the coefficients were found to be not significant. Other larger models with some more predictors (control variables) was also an option. Across the different models, we found that the 4 significant variables in the logit models have relatively higher variable importance. Also, multicollinearity was not an issue in the logit model wiht these 4 variables. Hence, we used the model specification (4 significant variables listed above) across the different models and compare their performance. We used stepwise as a variable selection method here.

5.4.1 Logistic Regression

1. **Full Mode1**
2. **Smaller nested model using step-wise variable selection:** Variables selection can be based on business knowledge. It is safe to remove variables that are not statistically significant. But it is not okay to remove significant variables, unless we have sound justification or serious dimensionality issues.

Initially, $p < 0.15$, which is the default, is the criterion used for variable inclusion and removal, so as to retain marginally significant predictors as control variables.

```
## Start:  AIC=245.55
## DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
##     ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
##     serum_sodium + sex + smoking + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - anaemia                   1   219.56 243.56  0.0006 0.98115
## - smoking                   1   219.56 243.56  0.0014 0.97026
## - high_blood_pressure       1   219.64 243.64  0.1070 0.74378
## - diabetes                  1   219.72 243.72  0.2226 0.63745
## - platelets                 1   219.97 243.97  0.5359 0.46472
## - sex                       1   221.24 245.24  2.1937 0.13967
## - creatinine_phosphokinase  1   221.33 245.33  2.3120 0.12948
## <none>                          219.55 245.55
## - serum_sodium              1   222.40 246.40  3.7065 0.05519 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=243.55
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     high_blood_pressure + platelets + serum_creatinine + serum_sodium +
##     sex + smoking + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - smoking                   1   219.56 241.56  0.0012 0.97244
## - high_blood_pressure       1   219.64 241.64  0.1075 0.74327
## - diabetes                  1   219.73 241.73  0.2237 0.63662
## - platelets                 1   219.97 241.97  0.5373 0.46413
## - sex                       1   221.25 243.25  2.2102 0.13820
## - creatinine_phosphokinase  1   221.41 243.41  2.4228 0.12068
## <none>                          219.56 243.56
## - serum_sodium              1   222.43 244.43  3.7607 0.05345 .
## + anaemia                   1   219.55 245.55  0.0006 0.98115
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=241.56
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     high_blood_pressure + platelets + serum_creatinine + serum_sodium +
##     sex + time
##
##                           Df Deviance    AIC F value  Pr(>F)
## - high_blood_pressure      1   219.64 239.64  0.1073 0.74345
## - diabetes                 1   219.73 239.73  0.2258 0.63505
## - platelets                1   219.98 239.98  0.5522 0.45802
## - creatinine_phosphokinase 1   221.42 241.42  2.4405 0.11934
## <none>                         219.56 241.56
## - sex                      1   221.62 241.62  2.7078 0.10095
## - serum_sodium             1   222.43 242.43  3.7739 0.05303 .
## + smoking                  1   219.56 243.56  0.0012 0.97244
## + anaemia                  1   219.56 243.56  0.0004 0.98489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=239.64
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     platelets + serum_creatinine + serum_sodium + sex + time
##
##                           Df Deviance    AIC F value  Pr(>F)
## - diabetes                 1   219.81 237.81  0.2301 0.63185
## - platelets                1   220.08 238.08  0.5778 0.44778
## - creatinine_phosphokinase 1   221.54 239.54  2.5082 0.11435
## - sex                      1   221.62 239.62  2.6127 0.10710
## <none>                         219.64 239.64
## - serum_sodium             1   222.51 240.51  3.7802 0.05283 .
## + high_blood_pressure      1   219.56 241.56  0.1073 0.74345
## + smoking                  1   219.64 241.64  0.0007 0.97945
## + anaemia                  1   219.64 241.64  0.0005 0.98280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=237.81
## DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     platelets + serum_creatinine + serum_sodium + sex + time
##
##                           Df Deviance    AIC F value  Pr(>F)
## - platelets                1   220.22 236.22  0.5332 0.46586
## - creatinine_phosphokinase 1   221.71 237.71  2.5051 0.11457
## <none>                         219.81 237.81
```

```
## - sex                        1   221.93 237.93  2.7947 0.09565 .
## - serum_sodium               1   222.89 238.89  4.0664 0.04467 *
## + diabetes                   1   219.64 239.64  0.2301 0.63185
## + high_blood_pressure        1   219.73 239.73  0.1112 0.73901
## + smoking                    1   219.81 239.81  0.0016 0.96792
## + anaemia                    1   219.81 239.81  0.0007 0.97927
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=236.22
## DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - creatinine_phosphokinase  1   222.04 236.04  2.4157 0.12121
## - sex                       1   222.18 236.18  2.5932 0.10841
## <none>                          220.22 236.22
## - serum_sodium              1   223.46 237.46  4.2871 0.03928 *
## + platelets                 1   219.81 237.81  0.5332 0.46586
## + diabetes                  1   220.08 238.08  0.1842 0.66808
## + high_blood_pressure       1   220.12 238.12  0.1332 0.71544
## + smoking                   1   220.21 238.21  0.0144 0.90468
## + anaemia                   1   220.22 238.22  0.0000 0.99929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=236.04
## DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     sex + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - sex                       1   223.49 235.49  1.8962 0.16956
## <none>                          222.04 236.04
## + creatinine_phosphokinase  1   220.22 236.22  2.4157 0.12121
## - serum_sodium              1   225.08 237.08  3.9925 0.04663 *
## + platelets                 1   221.71 237.71  0.4376 0.50879
## + high_blood_pressure       1   221.90 237.90  0.1889 0.66419
## + diabetes                  1   221.91 237.91  0.1732 0.67755
## + anaemia                   1   221.98 237.98  0.0858 0.76982
## + smoking                   1   222.02 238.02  0.0306 0.86115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=235.49
## DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium +
```

```
##     time
##
##                            Df Deviance    AIC F value  Pr(>F)
## <none>                          223.49 235.49
## + sex                       1   222.04 236.04  1.8962 0.16956
## + creatinine_phosphokinase  1   222.18 236.18  1.7185 0.19092
## - serum_sodium              1   226.30 236.30  3.6907 0.05569 .
## + smoking                   1   223.09 237.09  0.5159 0.47315
## + diabetes                  1   223.25 237.25  0.3043 0.58162
## + platelets                 1   223.26 237.26  0.2945 0.58777
## + high_blood_pressure       1   223.46 237.46  0.0357 0.85033
## + anaemia                   1   223.48 237.48  0.0033 0.95430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     serum_sodium + time, family = binomial(link = "logit"), data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1590  -0.5888  -0.2281   0.5144   2.7959
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        9.493034   5.405768   1.756  0.07907 .
## age                0.042466   0.015030   2.825  0.00472 **
## ejection_fraction -0.073430   0.015785  -4.652 3.29e-06 ***
## serum_creatinine   0.685990   0.174044   3.941 8.10e-05 ***
## serum_sodium      -0.064557   0.038377  -1.682  0.09254 .
## time              -0.020895   0.002916  -7.166 7.74e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 223.49  on 293  degrees of freedom
## AIC: 235.49
##
## Number of Fisher Scoring iterations: 6

## [1]   1.000000   3.831667   5.243276   8.969244  15.808285 105.242967

##               age ejection_fraction  serum_creatinine      serum_sodium
```

```
##          1.053111          1.133484          1.079122          1.028355
##              time
##          1.096862
```

The h_step_15p model still has severe multi-collinearity with CI at 105 » 50 although all VIFs way below 10. So, just like before, the multicollinearity is possibly with the intercept and centering should help.

Next we try $p < 0.05$, a more restrictive criterion, to explore if the resulting model has multicollinearity issue or not.

```
## Start:  AIC=269.49
## DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
##     ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
##     serum_sodium + sex + smoking + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - anaemia                   1   219.56 265.65  0.0006 0.98115
## - smoking                   1   219.56 265.65  0.0014 0.97026
## - high_blood_pressure       1   219.64 265.73  0.1070 0.74378
## - diabetes                  1   219.72 265.82  0.2226 0.63745
## - platelets                 1   219.97 266.06  0.5359 0.46472
## - sex                       1   221.24 267.34  2.1937 0.13967
## - creatinine_phosphokinase  1   221.33 267.43  2.3120 0.12948
## - serum_sodium              1   222.40 268.50  3.7065 0.05519 .
## <none>                          219.55 269.49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=265.65
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     high_blood_pressure + platelets + serum_creatinine + serum_sodium +
##     sex + smoking + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - smoking                   1   219.56 261.81  0.0012 0.97244
## - high_blood_pressure       1   219.64 261.89  0.1075 0.74327
## - diabetes                  1   219.73 261.98  0.2237 0.63662
## - platelets                 1   219.97 262.22  0.5373 0.46413
## - sex                       1   221.25 263.50  2.2102 0.13820
## - creatinine_phosphokinase  1   221.41 263.66  2.4228 0.12068
## - serum_sodium              1   222.43 264.69  3.7607 0.05345 .
## <none>                          219.56 265.65
## + anaemia                   1   219.55 269.49  0.0006 0.98115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Step:  AIC=261.81
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     high_blood_pressure + platelets + serum_creatinine + serum_sodium +
##     sex + time
##
##                           Df Deviance    AIC F value  Pr(>F)
## - high_blood_pressure      1   219.64 258.05  0.1073 0.74345
## - diabetes                 1   219.73 258.14  0.2258 0.63505
## - platelets                1   219.98 258.39  0.5522 0.45802
## - creatinine_phosphokinase 1   221.42 259.83  2.4405 0.11934
## - sex                      1   221.62 260.03  2.7078 0.10095
## - serum_sodium             1   222.43 260.85  3.7739 0.05303 .
## <none>                         219.56 261.81
## + smoking                  1   219.56 265.65  0.0012 0.97244
## + anaemia                  1   219.56 265.65  0.0004 0.98489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=258.05
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     platelets + serum_creatinine + serum_sodium + sex + time
##
##                           Df Deviance    AIC F value  Pr(>F)
## - diabetes                 1   219.81 254.38  0.2301 0.63185
## - platelets                1   220.08 254.65  0.5778 0.44778
## - creatinine_phosphokinase 1   221.54 256.12  2.5082 0.11435
## - sex                      1   221.62 256.20  2.6127 0.10710
## - serum_sodium             1   222.51 257.08  3.7802 0.05283 .
## <none>                         219.64 258.05
## + high_blood_pressure      1   219.56 261.81  0.1073 0.74345
## + smoking                  1   219.64 261.89  0.0007 0.97945
## + anaemia                  1   219.64 261.89  0.0005 0.98280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=254.39
## DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     platelets + serum_creatinine + serum_sodium + sex + time
##
##                           Df Deviance    AIC F value  Pr(>F)
## - platelets                1   220.22 250.95  0.5332 0.46586
## - creatinine_phosphokinase 1   221.71 252.44  2.5051 0.11457
## - sex                      1   221.93 252.66  2.7947 0.09565 .
## - serum_sodium             1   222.89 253.63  4.0664 0.04467 *
```

```
## <none>                           219.81 254.38
## + diabetes                   1   219.64 258.05  0.2301 0.63185
## + high_blood_pressure        1   219.73 258.14  0.1112 0.73901
## + smoking                    1   219.81 258.23  0.0016 0.96792
## + anaemia                    1   219.81 258.23  0.0007 0.97927
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=250.95
## DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - creatinine_phosphokinase  1   222.04 248.94  2.4157 0.12121
## - sex                       1   222.18 249.07  2.5932 0.10841
## - serum_sodium              1   223.46 250.35  4.2871 0.03928 *
## <none>                          220.22 250.95
## + platelets                 1   219.81 254.38  0.5332 0.46586
## + diabetes                  1   220.08 254.65  0.1842 0.66808
## + high_blood_pressure       1   220.12 254.69  0.1332 0.71544
## + smoking                   1   220.21 254.78  0.0144 0.90468
## + anaemia                   1   220.22 254.79  0.0000 0.99929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=248.93
## DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     sex + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## - sex                       1   223.49 246.53  1.8962 0.16956
## - serum_sodium              1   225.08 248.13  3.9925 0.04663 *
## <none>                          222.04 248.94
## + creatinine_phosphokinase  1   220.22 250.95  2.4157 0.12121
## + platelets                 1   221.71 252.44  0.4376 0.50879
## + high_blood_pressure       1   221.90 252.63  0.1889 0.66419
## + diabetes                  1   221.91 252.64  0.1732 0.67755
## + anaemia                   1   221.98 252.71  0.0858 0.76982
## + smoking                   1   222.02 252.75  0.0306 0.86115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=246.54
## DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     time
```

```
##
##                            Df Deviance    AIC F value  Pr(>F)
## - serum_sodium              1   226.30 245.51  3.6907 0.05569 .
## <none>                          223.49 246.53
## + sex                       1   222.04 248.94  1.8962 0.16956
## + creatinine_phosphokinase  1   222.18 249.07  1.7185 0.19092
## + smoking                   1   223.09 249.98  0.5159 0.47315
## + diabetes                  1   223.25 250.14  0.3043 0.58162
## + platelets                 1   223.26 250.15  0.2945 0.58777
## + high_blood_pressure       1   223.46 250.35  0.0357 0.85033
## + anaemia                   1   223.48 250.37  0.0033 0.95430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=245.51
## DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + time
##
##                            Df Deviance    AIC F value  Pr(>F)
## <none>                          226.30 245.51
## + serum_sodium              1   223.49 246.53  3.6907 0.05569 .
## + sex                       1   225.08 248.13  1.5895 0.20840
## + creatinine_phosphokinase  1   225.12 248.17  1.5350 0.21635
## + diabetes                  1   225.87 248.92  0.5614 0.45430
## + platelets                 1   225.93 248.97  0.4859 0.48630
## + smoking                   1   225.95 249.00  0.4604 0.49799
## + high_blood_pressure       1   226.27 249.31  0.0453 0.83164
## + anaemia                   1   226.28 249.32  0.0334 0.85518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time, family = binomial(link = "logit"), data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1289  -0.6030  -0.2414   0.4954   2.8633
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.604473   1.036111   0.583  0.55962
## age               0.043326   0.014872   2.913  0.00358 **
## ejection_fraction -0.074804   0.015555  -4.809 1.52e-06 ***
## serum_creatinine   0.719785   0.174597   4.123 3.75e-05 ***
```

```
## time              -0.020611   0.002881  -7.153 8.48e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 226.30  on 294  degrees of freedom
## AIC: 236.3
##
## Number of Fisher Scoring iterations: 5

## [1]  1.000000  3.474370  4.862294  8.382579 17.191309
```

The h_step_05p model is much better and has a CI of 17, which is acceptable. Also, the h_step_05p model has no non-significant predictor.

3. **Non-linear Model**

We see in the h_step_05p model that age is negatively associated with the heart failure probability. To explore this further, we make use of non-linear models. We fit a quadratic polynomial model to evaluate the non-linear effect of age on DEATH_EVENT.

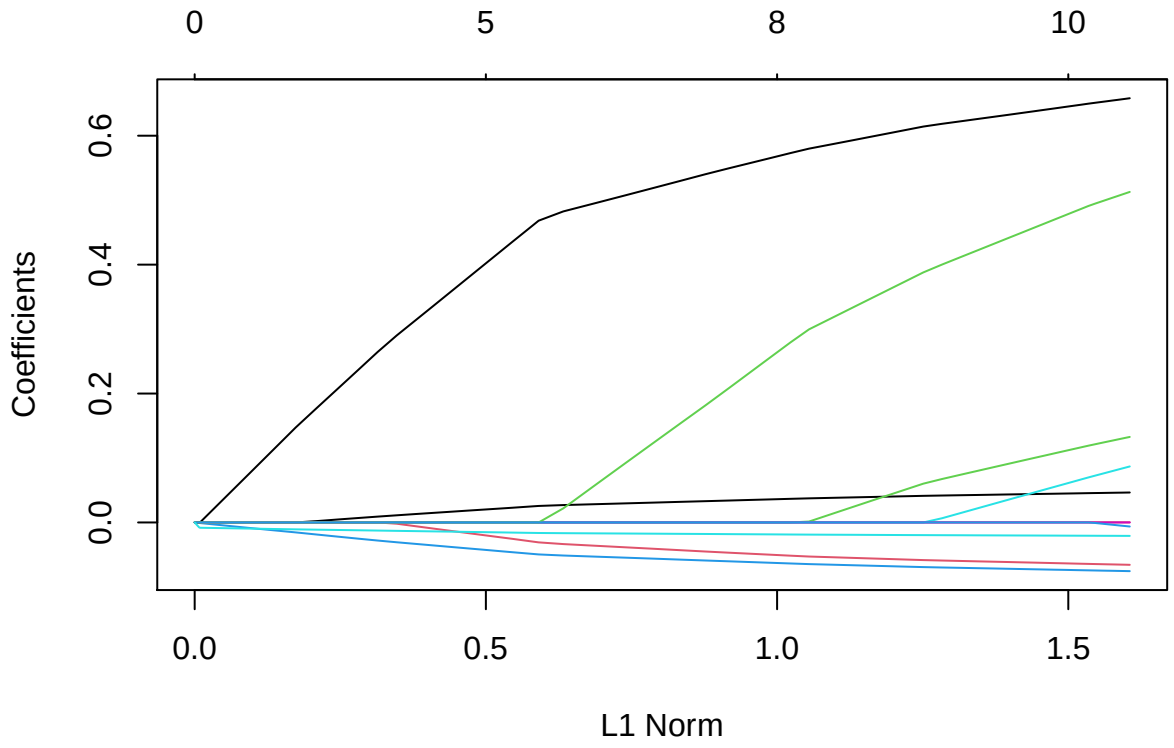
```
##
## Call:
## glm(formula = DEATH_EVENT ~ poly(age, 2, raw = T) + ejection_fraction +
##     serum_creatinine + time, family = binomial(link = "logit"),
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0340  -0.6065  -0.2521   0.4317   2.8432
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              7.076795   4.287192   1.651   0.0988 .
## poly(age, 2, raw = T)1 -0.167375   0.136202  -1.229   0.2191
## poly(age, 2, raw = T)2  0.001659   0.001075   1.543   0.1229
## ejection_fraction      -0.076153   0.015849  -4.805 1.55e-06 ***
## serum_creatinine        0.736860   0.173281   4.252 2.11e-05 ***
## time                   -0.020430   0.002886  -7.079 1.45e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 223.84  on 293  degrees of freedom
## AIC: 235.84
##
## Number of Fisher Scoring iterations: 5
```

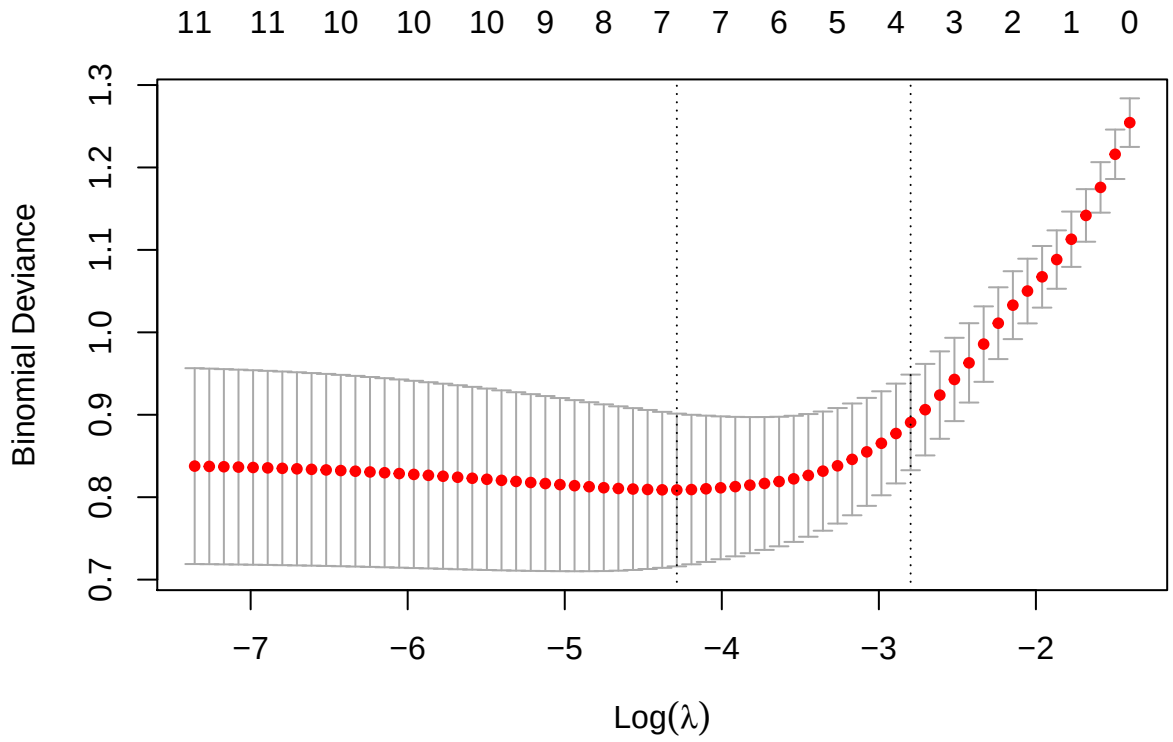
In the quadratic model h_quad, the linear effect of age is positive and the quadratic effect is negative but both the linear and the quadratic effects are not significant. So, the linear model h_step_05p is better.

5.4.2 LASSO

1. **Model Specification 1:** The full logit model has multicollinearity. One of the ways to deal with multicollinearity is to use shrinkage methods such as Ridge and LASSO. These are specially useful to reduce variance caused by dimensionality when business rationale suggests that many variables need to be retained.



In the plot, we can see how the coefficients shrink and some of them drop out as we move toward the left, that is, as the lambda goes up. Next we use 10-fold cross-validation to get the best lambda, that is, the lambda for which the deviance is the minimum.



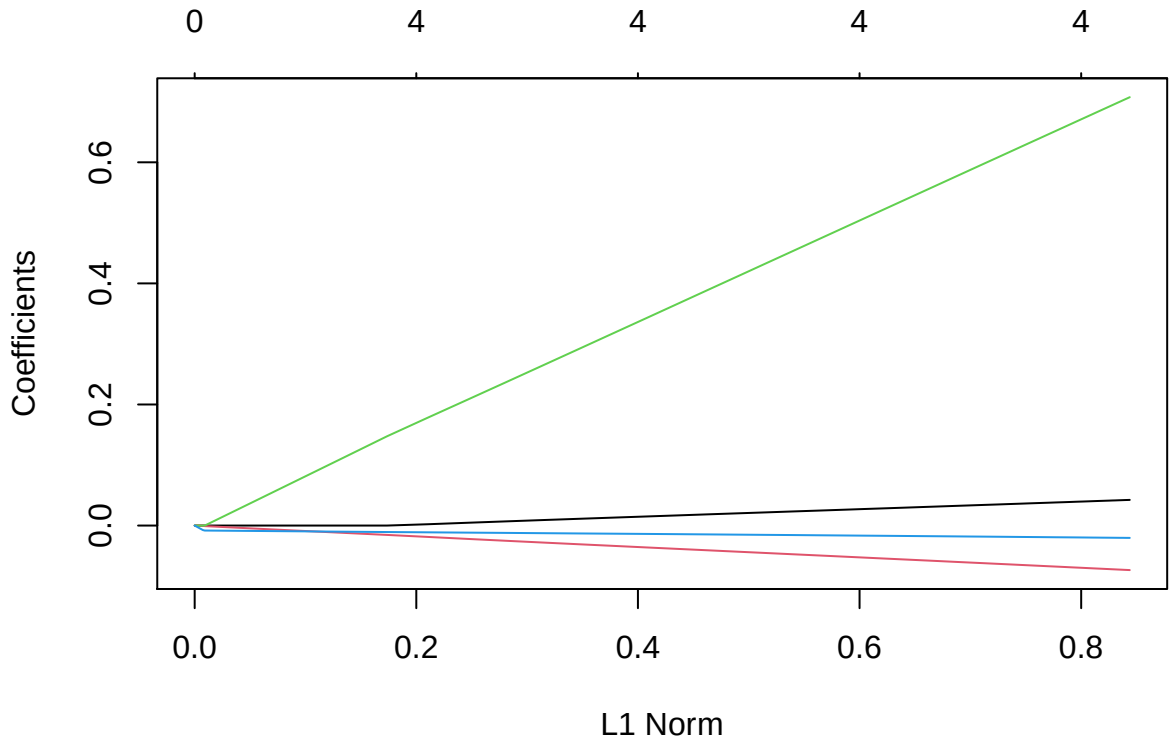

```
##      Best Lambda Best Log Lambda Best 10F-CV
## [1,]  0.01375523       -4.286336   0.8087437
```

Next we retrieve the LASSO regression coefficients from the h_lasso model object for the Best Lambda and compare them to the plain logit coefficients.

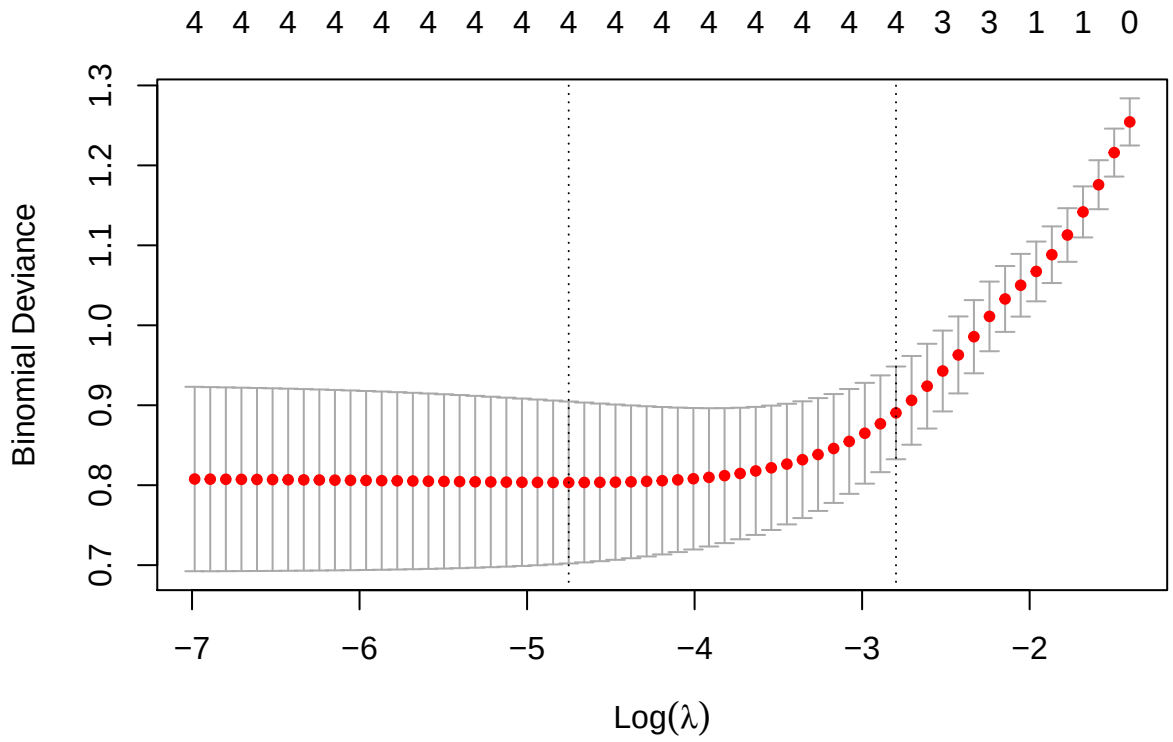
```
## 13 x 4 Matrix of class "dgeMatrix"
##                            Best LASSO    Odds 0-Lambda LASSO      Odds
## (Intercept)                     6.425 617.138          9.382 11878.569
## age                             0.032   1.032          0.046     1.048
## anaemia1                        0.000   1.000          0.000     1.000
## creatinine_phosphokinase        0.000   1.000          0.000     1.000
## diabetes1                       0.000   1.000          0.133     1.142
## ejection_fraction              -0.058   0.944         -0.076     0.927
## high_blood_pressure0            0.000   1.000          0.087     1.091
## platelets                       0.000   1.000          0.000     1.000
## serum_creatinine                0.530   1.699          0.658     1.931
## serum_sodium                   -0.043   0.958         -0.066     0.936
## sex0                            0.154   1.166          0.513     1.670
## smoking1                        0.000   1.000         -0.006     0.994
## time                           -0.018   0.983         -0.021     0.979
```

2. **Model Specification 2:**

Since, there is not much multicollinearity in the logit model with 4 variables, we expect the LASSO coefficients and the logit coefficients to be not very different as the shrinkage parameter λ is likely to quite small.



In the plot, we can see how the coefficients shrink as we move toward the left, that is, as the lambda goes up. Next we use 10-fold cross-validation to get the best lambda, that is, the lambda for which the deviance is the minimum.



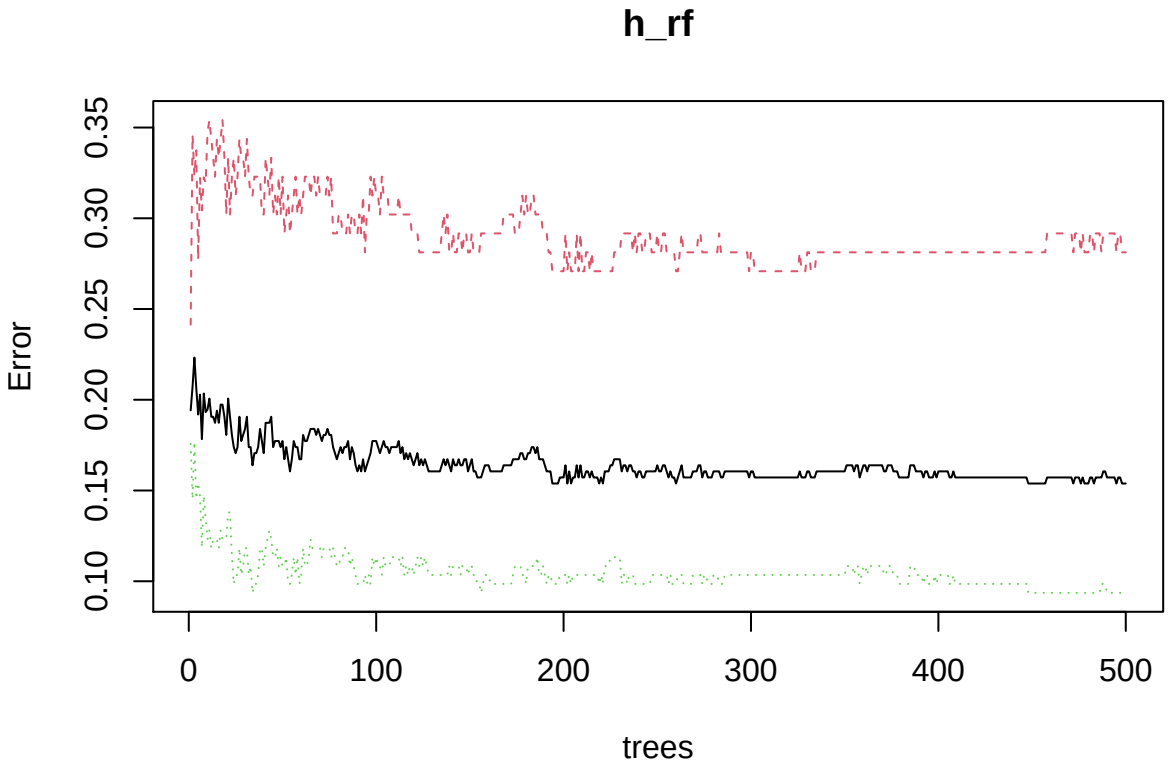

```
##      Best Lambda Best Log Lambda Best 10F-CV
## [1,] 0.008638687       -4.751505   0.8033457
```

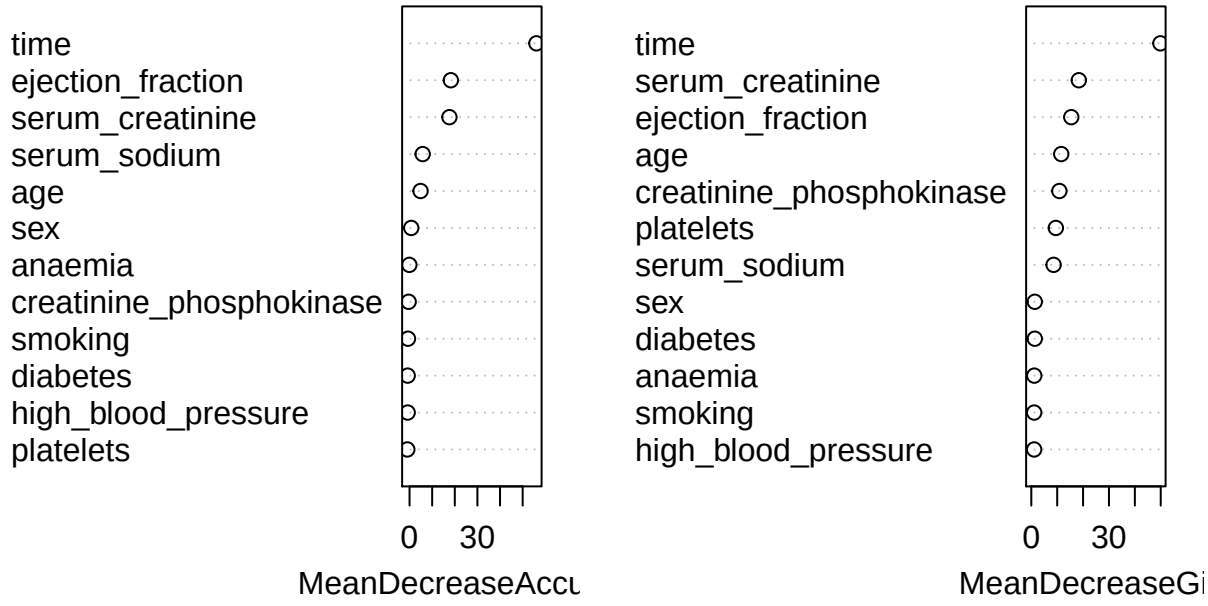
Next we retrieve the LASSO regression coefficients from the h_lasso2 model object for the Best Lambda and compare them to the plain logit coefficients.

```
## 5 x 4 Matrix of class "dgeMatrix"
##                    Best LASSO  Odds 0-Lambda LASSO  Odds
## (Intercept)             0.650 1.915          0.610 1.840
## age                     0.035 1.036          0.042 1.043
## ejection_fraction      -0.064 0.938         -0.074 0.929
## serum_creatinine        0.616 1.851          0.708 2.029
## time                   -0.019 0.982         -0.020 0.980
```

5.4.3 Random Forest

1. **Full Model Specification:** p (number of predictors) = 12. We choose M (Number of variables per tree) = sqrt(p) = 4 because it has been shown to give good performance.

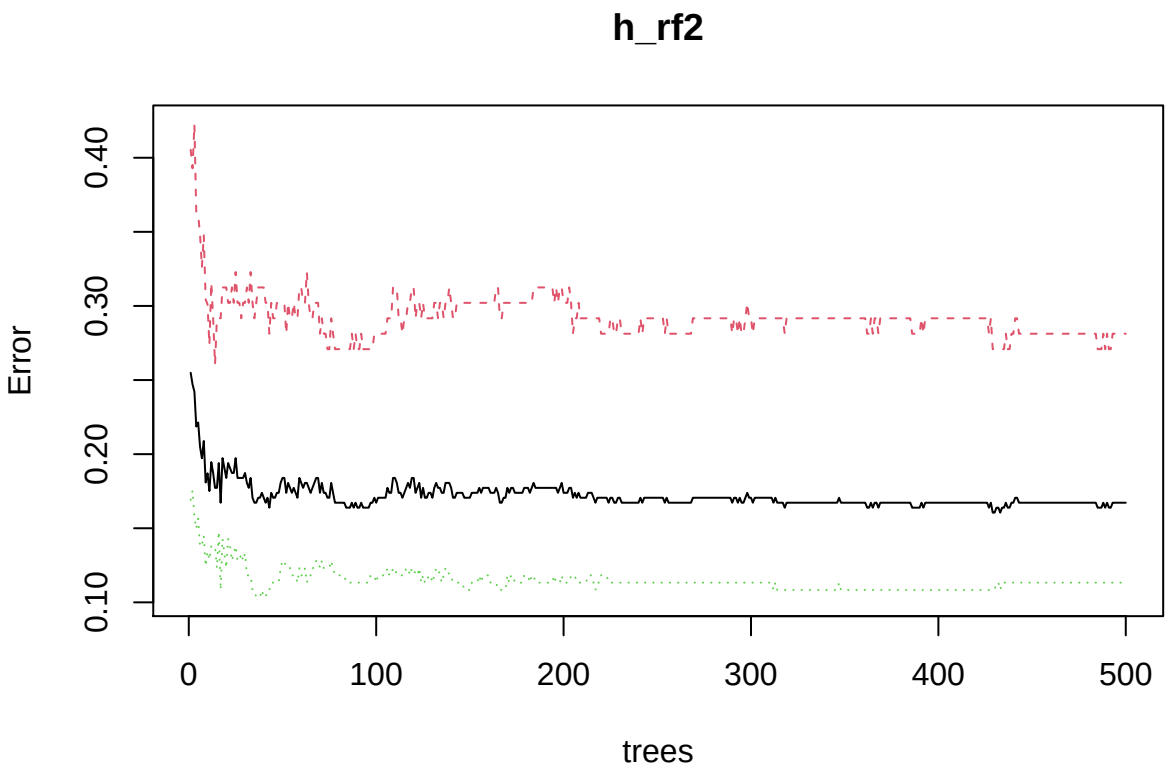


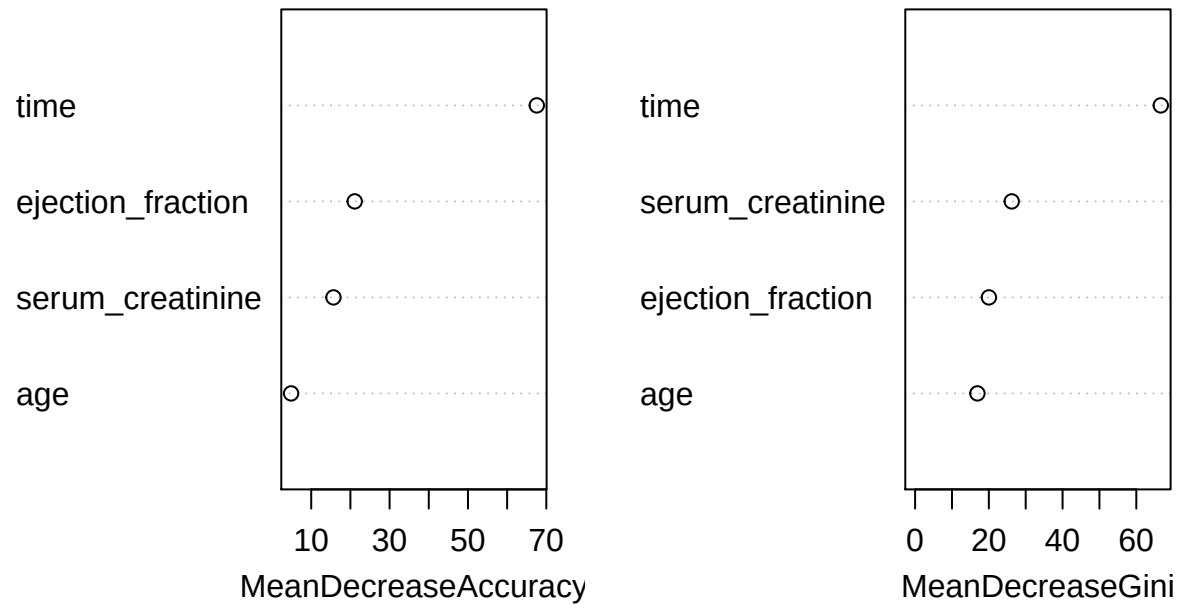



```
##                           MeanDecreaseAccuracy MeanDecreaseGini
## age                                 4.87254994        11.565924
## anaemia                            -0.08601633         1.141397
## creatinine_phosphokinase           -0.40102170        10.850111
## diabetes                           -0.82339092         1.342036
## ejection_fraction                  18.25765484        15.437616
## high_blood_pressure                -0.82713201         1.102297
## platelets                          -0.98876848         9.475315
## serum_creatinine                   17.62248854        18.360920
## serum_sodium                        5.81595279         8.613212
## sex                                 0.75614508         1.342594
## smoking                            -0.69066516         1.115784
## time                               56.08951655        49.885165

##      Mean Error Rate
## [1,]       0.1864409
```

2. **Smaller Model Specifications:** p (number of predictors) = 4. We choose M (Number of variables per tree) = sqrt(p) = 2.


h_rf2
Error
0.10
0.20
0.30
0.40
0
100
200
300
400
500
trees



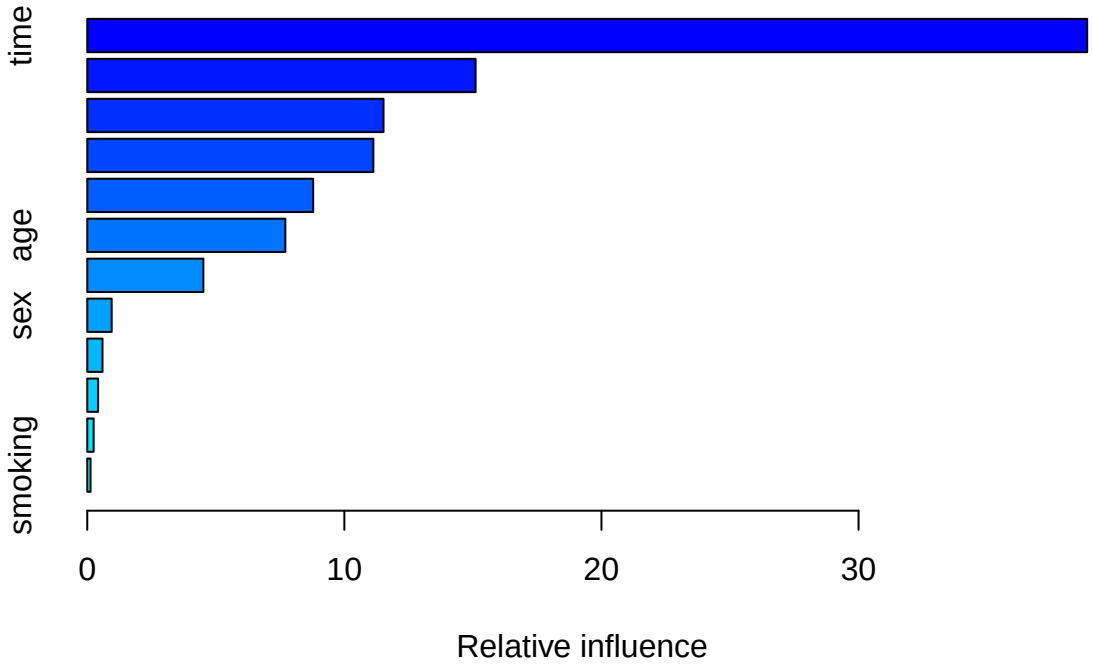

```
##                   MeanDecreaseAccuracy MeanDecreaseGini
## age                           4.870131         16.90213
## ejection_fraction            21.102261         20.02398
## serum_creatinine             15.681545         26.21005
## time                         67.565438         66.63951

##      Mean Error Rate
## [1,]       0.1929058
```

5.4.4 Boosted Trees

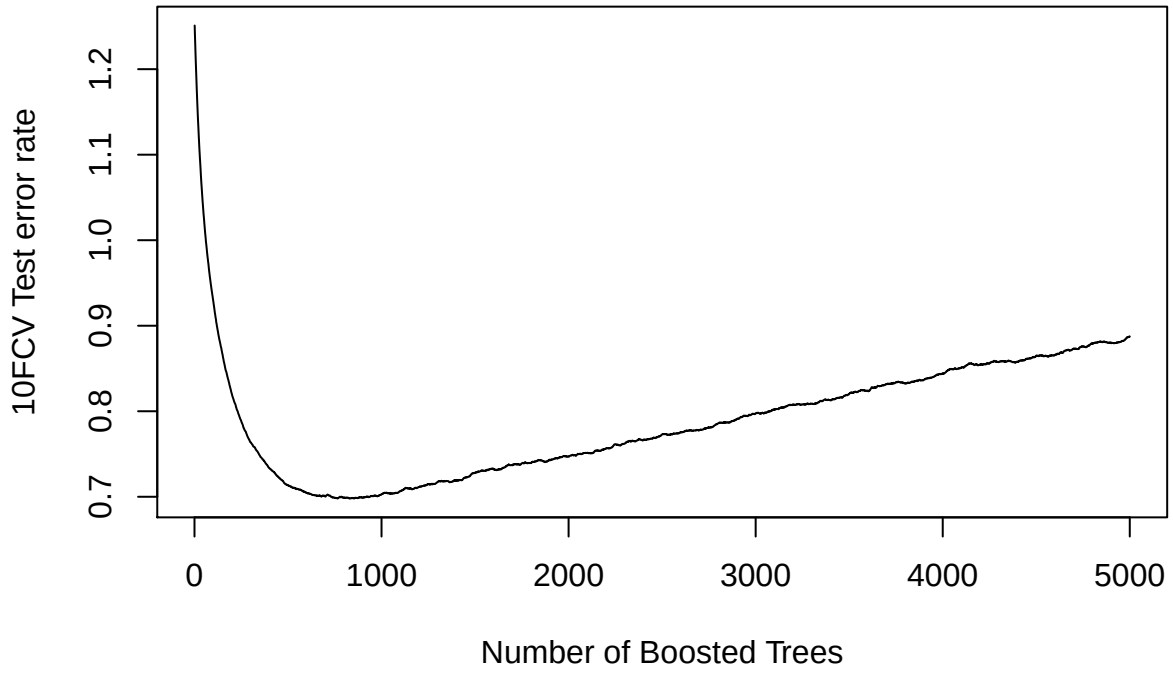
1. **Full Model Specification**

```
## gbm(formula = DEATH_EVENT ~ ., distribution = "bernoulli", data = heart,
##     n.trees = 5000, interaction.depth = 1, shrinkage = 0.01,
##     cv.folds = 10)
## A gradient boosted model with bernoulli loss function.
## 5000 iterations were performed.
## The best cross-validation iteration was 853.
## There were 12 predictors of which 11 had non-zero influence.
```

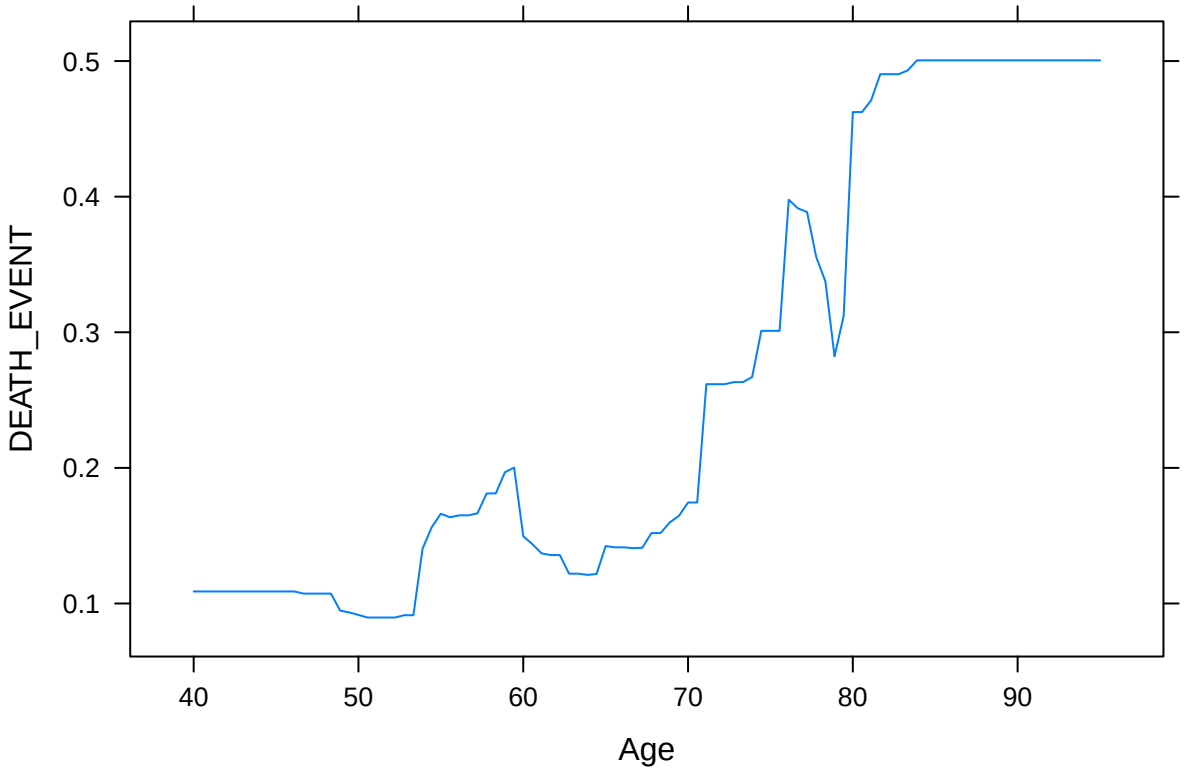


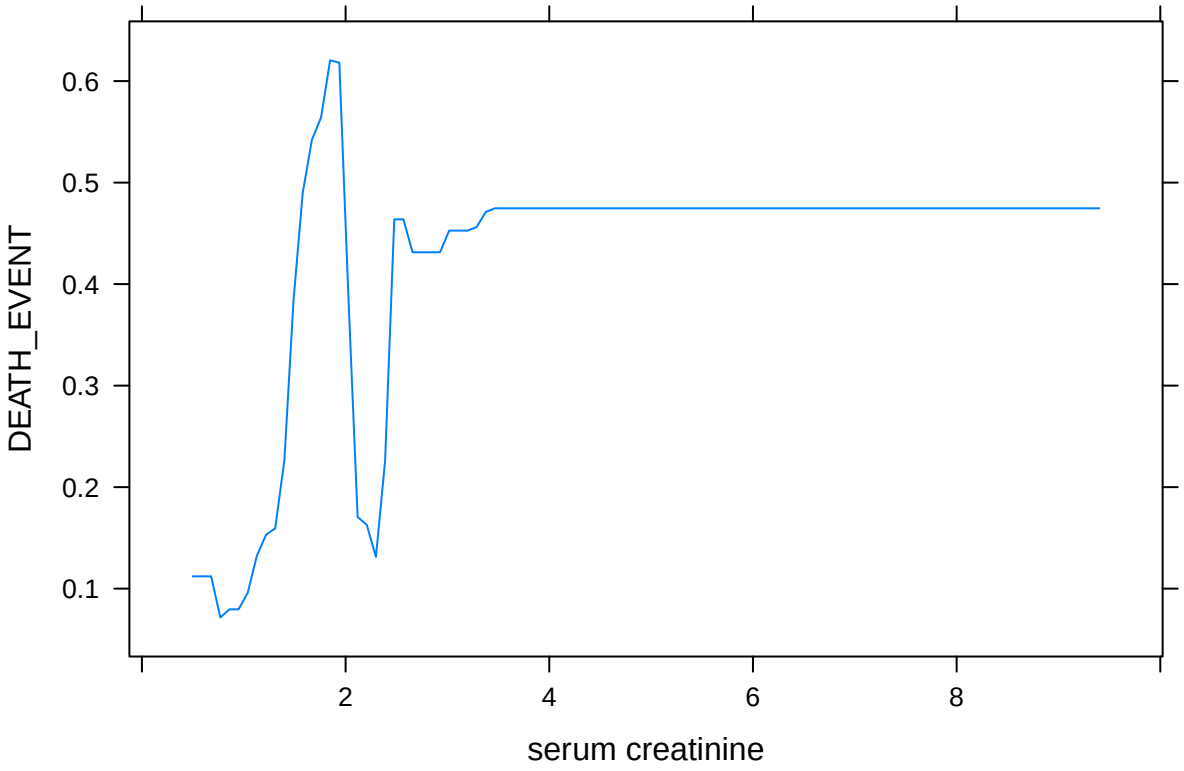

```
##                                                 var    rel.inf
## time                                           time 38.8929741
## serum_creatinine                 serum_creatinine 15.1048138
## ejection_fraction               ejection_fraction 11.5231851
## creatinine_phosphokinase creatinine_phosphokinase 11.1288614
## platelets                                platelets  8.7875055
## age                                            age  7.7044318
## serum_sodium                       serum_sodium  4.5174737
## sex                                            sex  0.9493168
## diabetes                                   diabetes  0.5929444
```

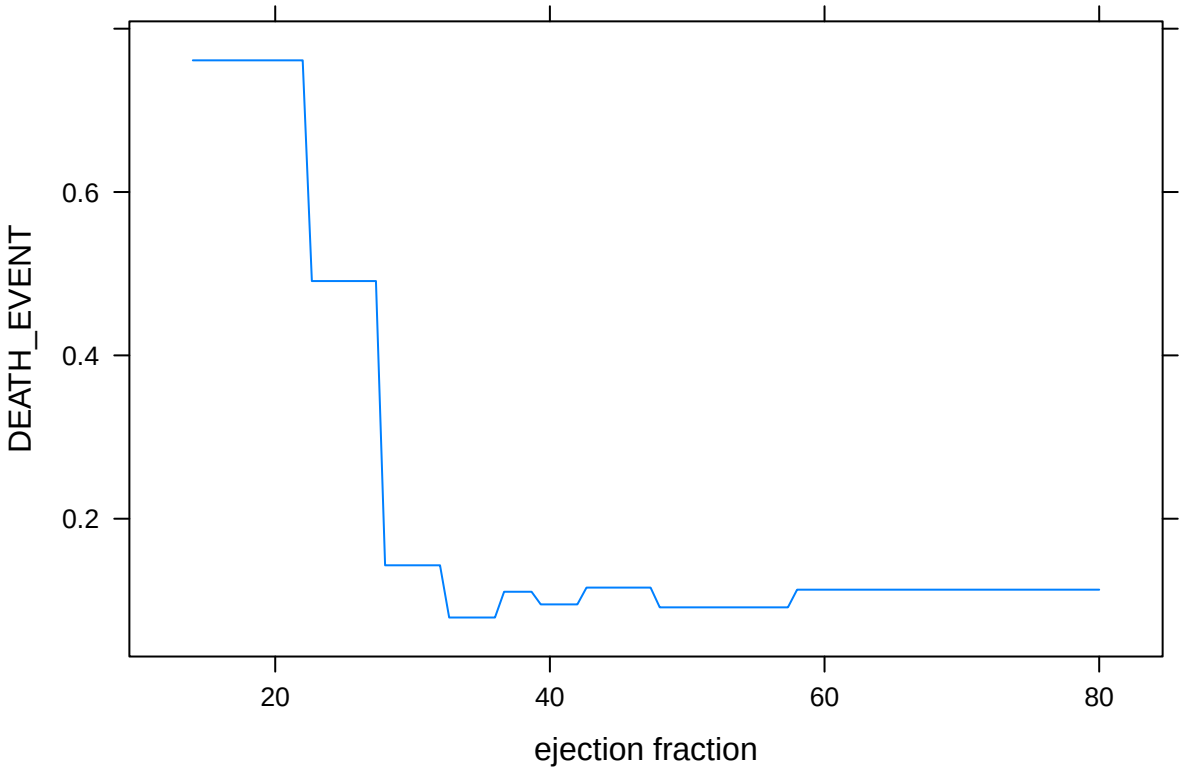
```
## anaemia                                   anaemia  0.4225037
## high_blood_pressure           high_blood_pressure  0.2488431
## smoking                                   smoking  0.1271466
```



Partial dependency graphs:
Holding everything else constant, on average,

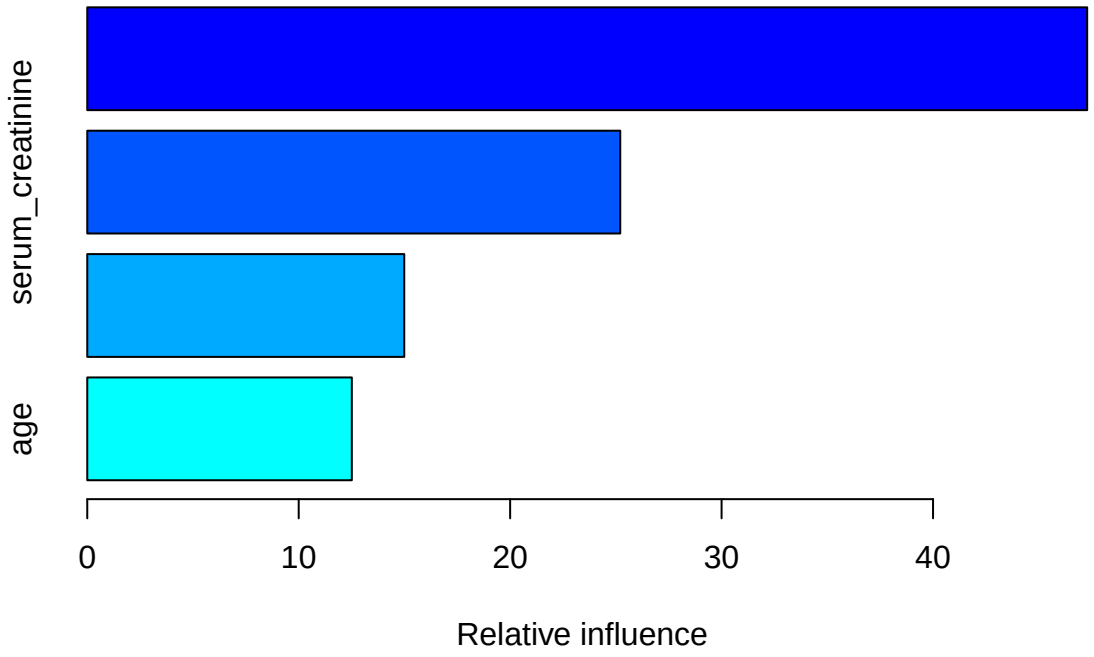

DEATH_EVENT
0.1
0.2
0.3
0.4
0.5
40
50
60
70
80
90
Age


DEATH_EVENT
0.1
0.2
0.3
0.4
0.5
0.6
2
4
6
8
serum creatinine

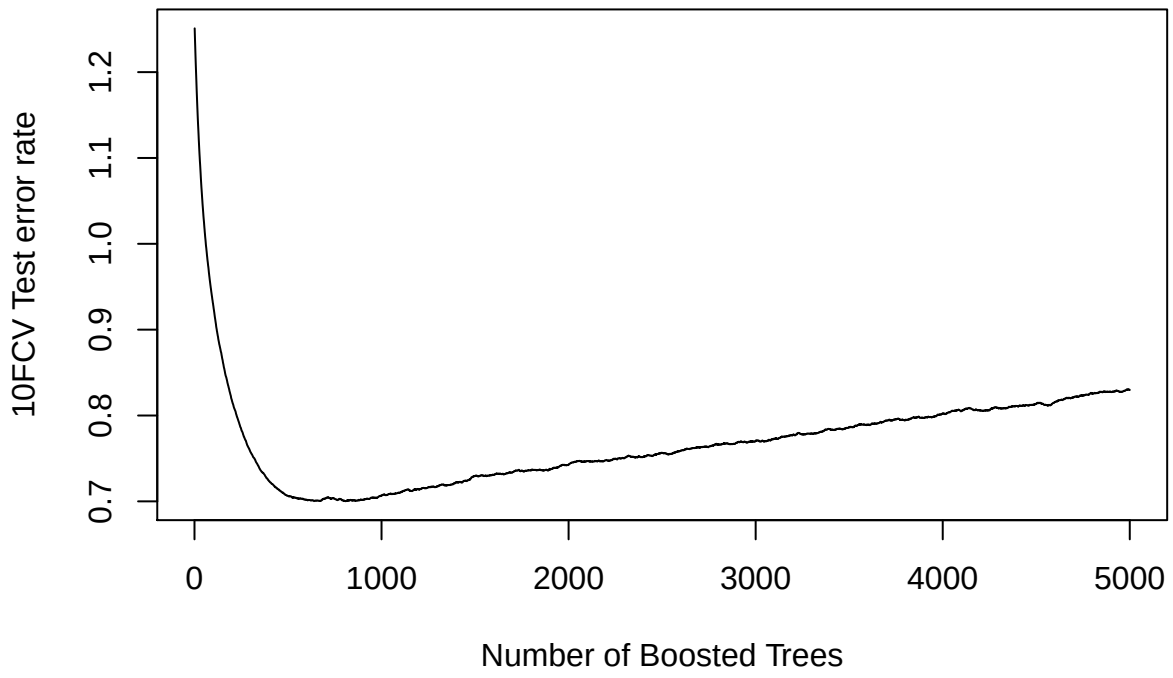


2. **Model Specification 2**

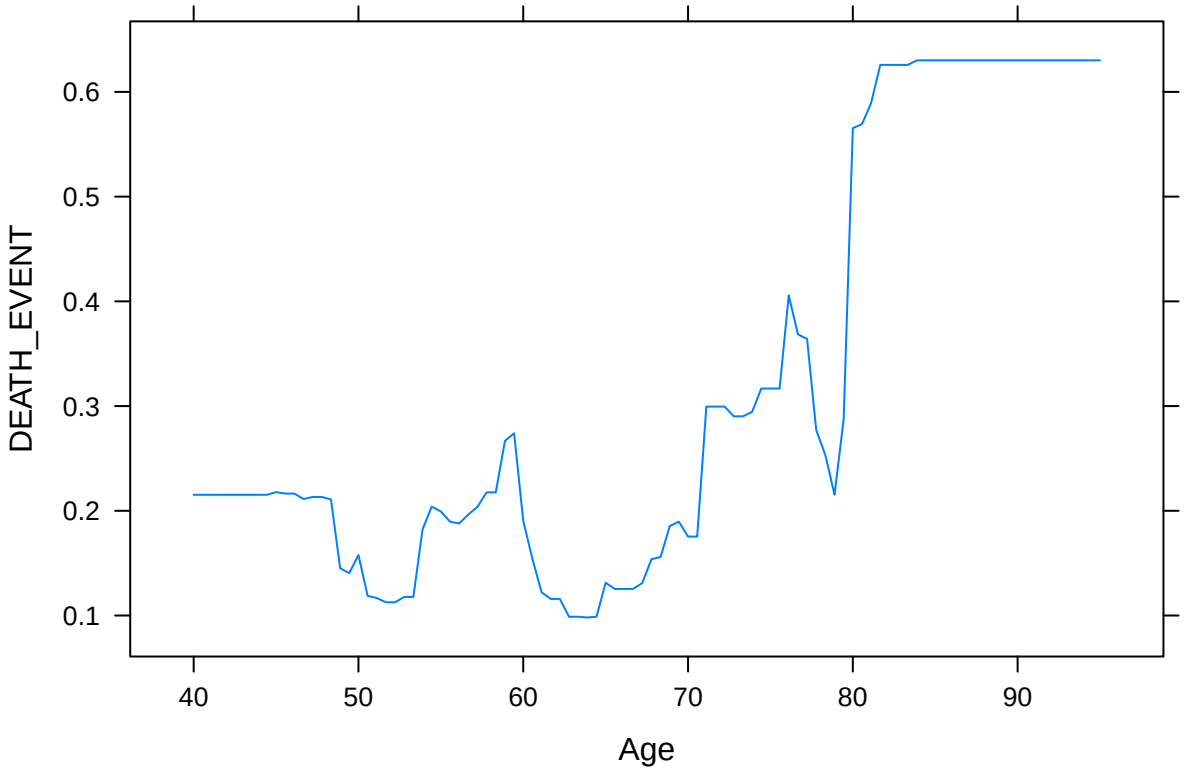
```
## gbm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time, distribution = "bernoulli", data = heart, n.trees = 5000,
##     interaction.depth = 1, shrinkage = 0.01, cv.folds = 10)
## A gradient boosted model with bernoulli loss function.
## 5000 iterations were performed.
## The best cross-validation iteration was 808.
## There were 4 predictors of which 4 had non-zero influence.
```

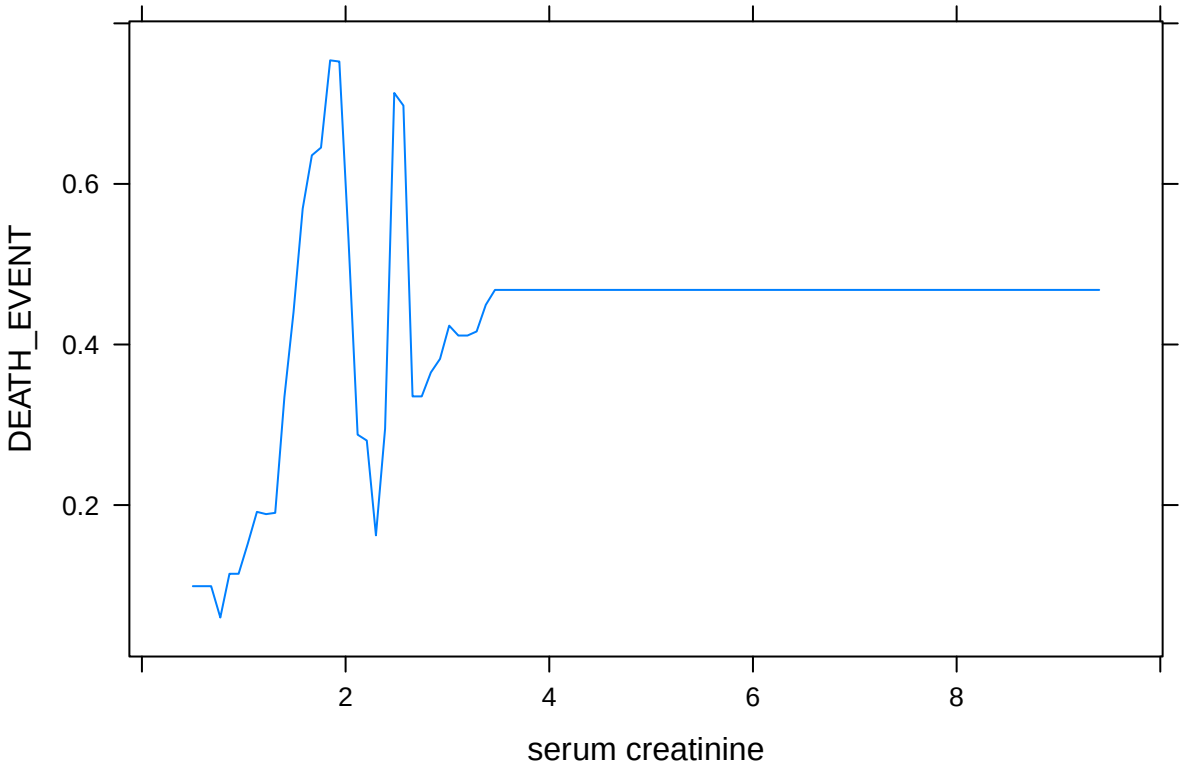


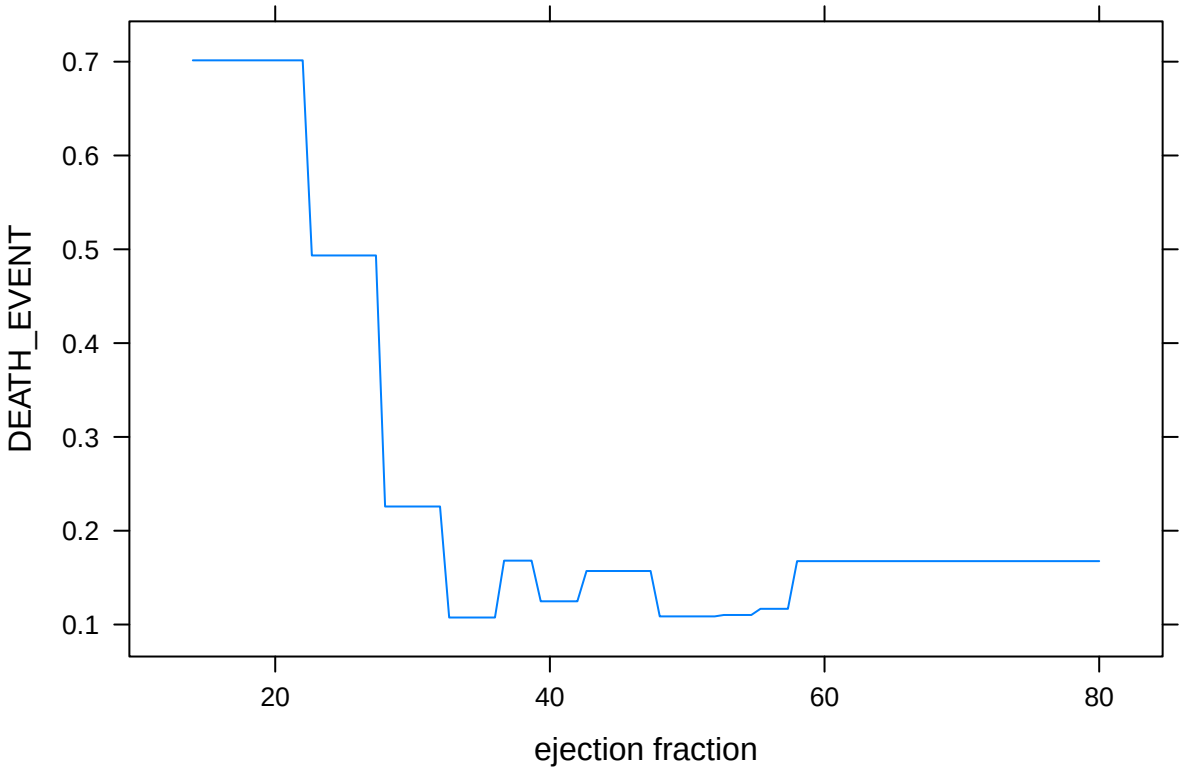

```
##                                  var  rel.inf
## time                            time 47.29000
## serum_creatinine   serum_creatinine 25.20633
## ejection_fraction ejection_fraction 14.99211
## age                              age 12.51156
```



Partial dependency graphs


DEATH_EVENT
0.1
0.2
0.3
0.4
0.5
0.6
40
50
60
70
80
90
Age


DEATH_EVENT
0.2
0.4
0.6
2
4
6
8
serum creatinine


DEATH_EVENT
0.7
0.6
0.5
0.4
0.3
0.2
0.1
20
40
60
80
ejection fraction

6. Analysis of Results

6.1 Train Models on Train Sub-Sample

To evaluate the performance of different models, we split the data set into two parts training and test. Using the training data set, we refit the models before evaluating their performance using the test data set.

Class imbalance is a common problem when working with real-world data. The performance of ML model is degraded because of it as it biases the model toward the majority class at the expense of the minority class. The class distribution of the test data set should mirror that of the original dataset because a model's performance against the test data is a proxy for its generalizability against unseen data. However, the training data should be balanced prior to the modeling process.

```
##
##     0     1
## 67.89 32.11

##
##  0  1
## 65 35

##
##  0  1
## 50 50
```

Now we fit the 4 X 2 models, 2 each of logit, LASSO, Random Forest and Boosted Tree models.

6.2 Classification Predictions

Predictions with the trained model and test subset.

6.3 Confusion Matrices

- Model 1.1

Table 1: Confusion Matrix, Prob Thres > 0.5

	No	Yes	Total
No	25	2	27
Yes	14	19	33
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]    0.733 0.267       0.905       0.641    0.359
```

- Model 1.2

Table 2: Confusion Matrix, Prob Thres > 0.5

	No	Yes	Total
No	27	3	30
Yes	12	18	30
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]     0.75  0.25       0.857       0.692    0.308
```

- Model 2.1

Table 3: Confusion Matrix, Prob Thres > 0.5

	No	Yes	Total
No	28	2	30
Yes	11	19	30
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]    0.783 0.217       0.905       0.718    0.282
```

- Model 2.2

Table 4: Confusion Matrix, Prob Thres > 0.5

	No	Yes	Total
No	27	3	30
Yes	12	18	30
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]     0.75  0.25       0.857       0.692    0.308
```

- Model 3.1

Table 5: Confusion Matrix

	No	Yes	Total
No	33	4	37
Yes	6	17	23

	No	Yes	Total
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]    0.833 0.167        0.81       0.846    0.154
```

- Model 3.2

Table 6: Confusion Matrix

	No	Yes	Total
No	33	5	38
Yes	6	16	22
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]    0.817 0.183       0.762       0.846    0.154
```

- Model 4.1

Table 7: Confusion Matrix, Prob Thres > 0.5

	No	Yes	Total
No	35	5	40
Yes	4	16	20
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]     0.85  0.15       0.762       0.897    0.103
```

- Model 4.2

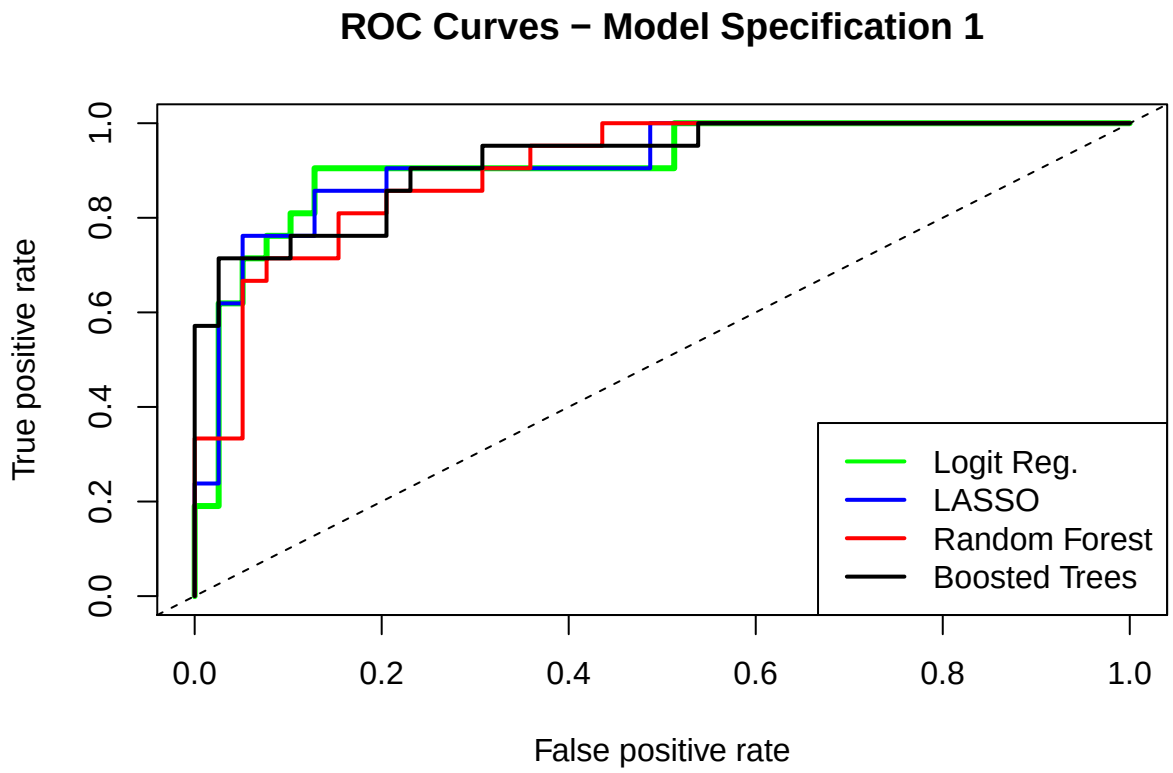
Table 8: Confusion Matrix, Prob Thres > 0.5

	No	Yes	Total
No	34	3	37
Yes	5	18	23
Total	39	21	60

```
##      Accuracy Error Sensitivity Specificity FalsePos
## [1,]    0.867 0.133       0.857       0.872    0.128
```

6.4 ROC Curves and AUC

ROC curve is commonly used to visually represent the relationship between a model's true positive rate and false positive rate for all possible cutoff values. ROC curve is summarized into a single quantity known as area under the curve (AUC), which measures the entire two-dimensional area underneath the ROC curve from (0,0) to (1,1). The higher the AUC, the better the performance of the model overall at distinguishing between the positive and negative classes.

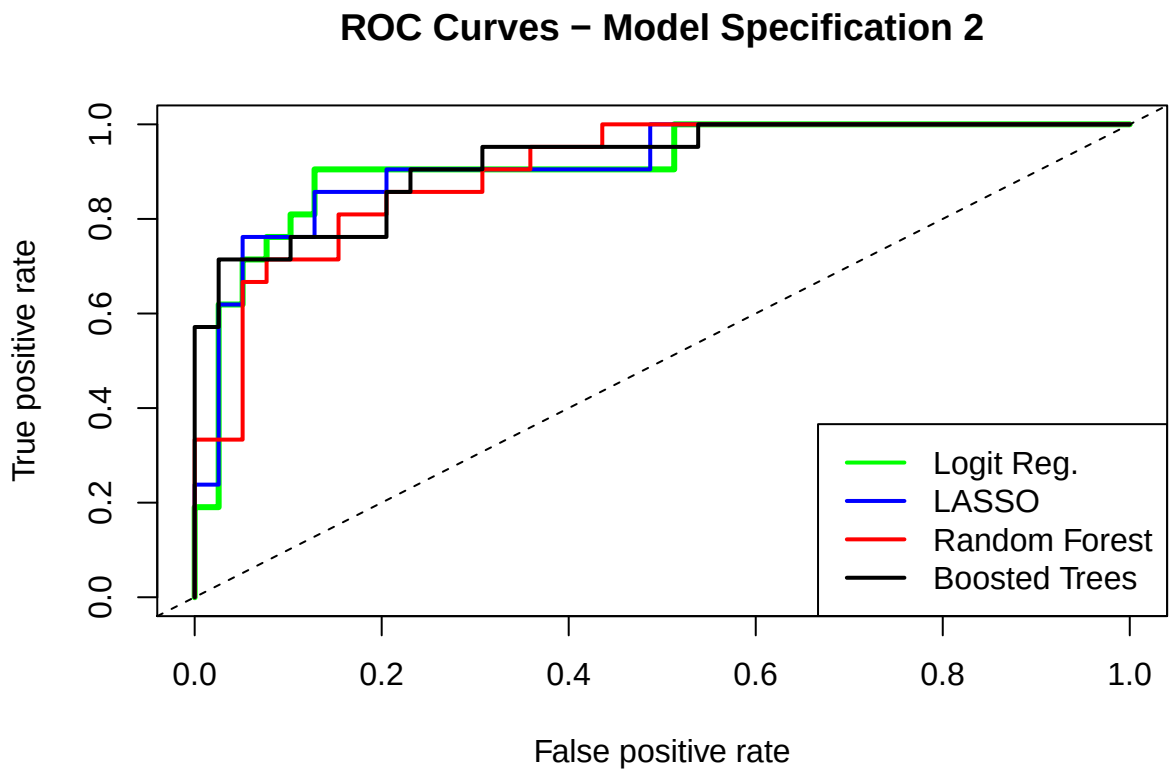



```
## [1] "Area under the ROC curve for Logit Model = 0.915"

## [1] "Area under the ROC curve for LASSO = 0.915"

## [1] "Area under the ROC curve for Random Forest = 0.902"

## [1] "Area under the ROC curve for Boosted Trees = 0.921"
```




```
## [1] "Area under the ROC curve for Logit Model = 0.907"

## [1] "Area under the ROC curve for LASSO = 0.908"

## [1] "Area under the ROC curve for Random Forest = 0.909"

## [1] "Area under the ROC curve for Boosted Trees = 0.906"
```

7. Conclusion

The capability to predict CVD early assumes a vital role for the patient's appropriate treatment procedure. Machine learning methods are valuable in this early diagnosis of CVD. In the current study, 3 machine learning techniques were applied on a training data set and validated against a test data set; both of these data sets were based on the data collected from the patients at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan). The results of our model implementations show that based on the measures of future performance - the AUC, all the models perform quite well. Since, interpretation is the goal for our case, the logistic regression classifier is the preferred model among all of them because of its high interpretatability and absence of any severe multicollinearity.

One limitation of the current study is that it may only be valid on a similar data set as was used for this study, which was sourced from a very specific location. Further research is needed to check if similar results are seen for data collected elsewhere. If an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation cohort to verify our findings.

Another limitation of the present study that we report is the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases.

Appendices

A. Data Information

B. Visuals, Graphs and Plots

C. Quantitative R Output

D. Other

E. References

1. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
2. Bredy C, Ministeri M, Kempny A, Alonso-Gonzalez R, Swan L, Uebing A, Diller G-P, Gatzoulis MA, Dimopoulos K. New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome. Eur Heart J – Qual Care Clin Outcomes. 2017; 4(1):51–8.