Final Project R Markdown

Kamaniya Chatakondu, Alberic C Kouadio, Santanu Mukherjee

12/06/2021

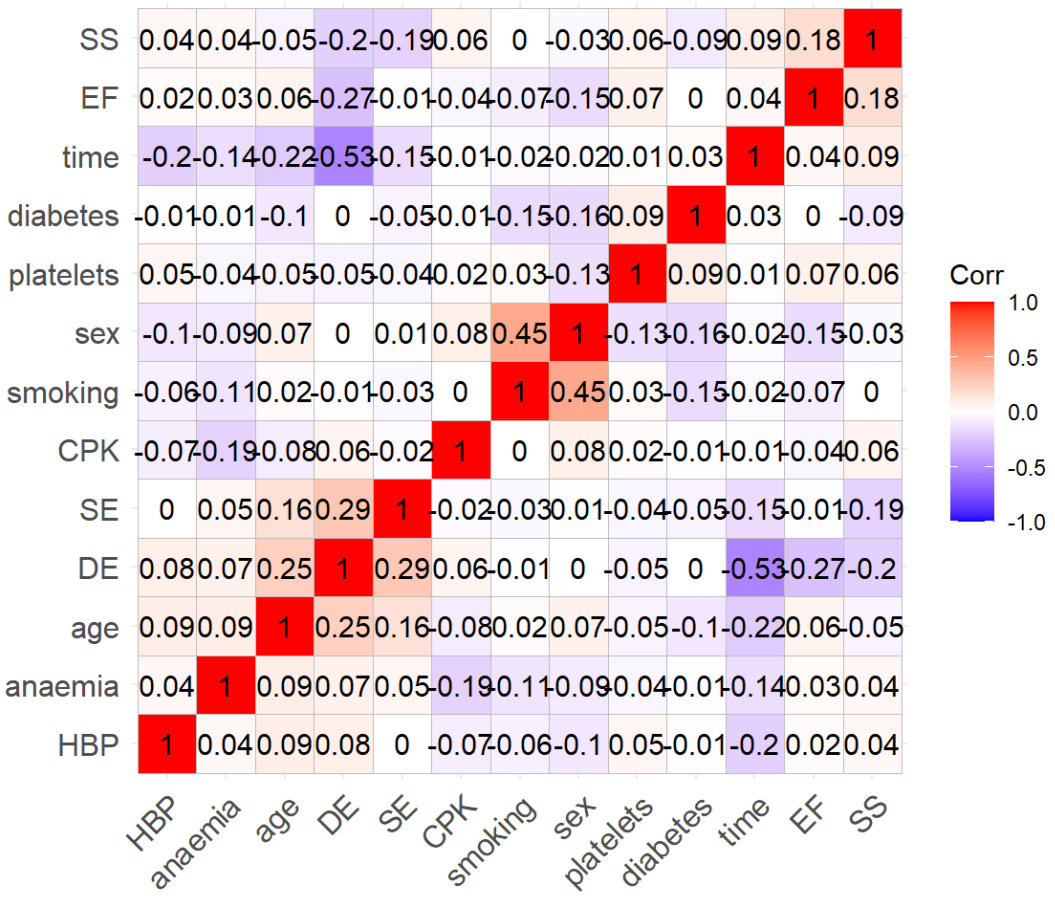
R Markdown

Final Project - Prediction of probability of death for patients having heart failures

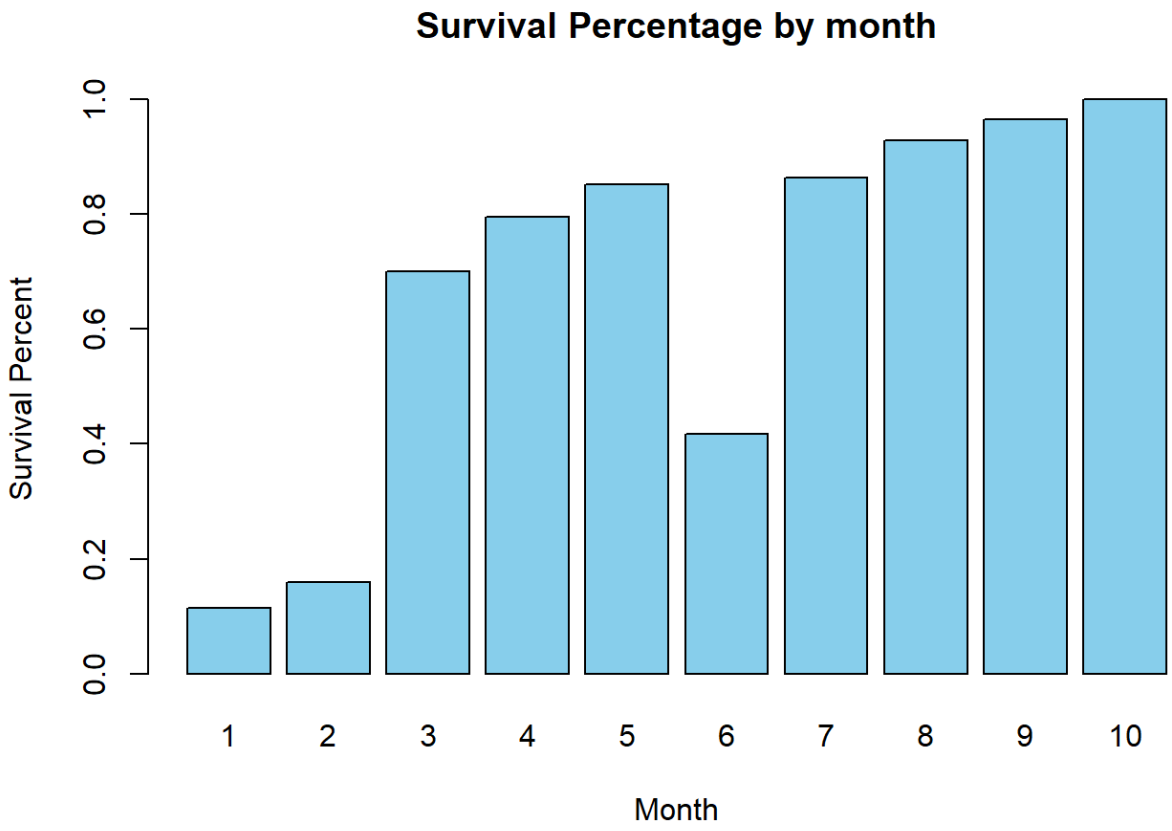
The below shows the first few records of the dataset

```
##   age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1  75       0                     582        0                20
## 2  55       0                    7861        0                38
## 3  65       0                     146        0                20
## 4  50       1                     111        0                20
## 5  65       1                     160        1                20
## 6  90       1                      47        0                40
##   high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1                   1    265000             1.9          130   1       0    4
## 2                   0    263358             1.1          136   1       0    6
## 3                   0    162000             1.3          129   1       1    7
## 4                   0    210000             1.9          137   1       0    7
## 5                   0    327000             2.7          116   0       0    8
## 6                   1    204000             2.1          132   1       1    8
##   DEATH_EVENT
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

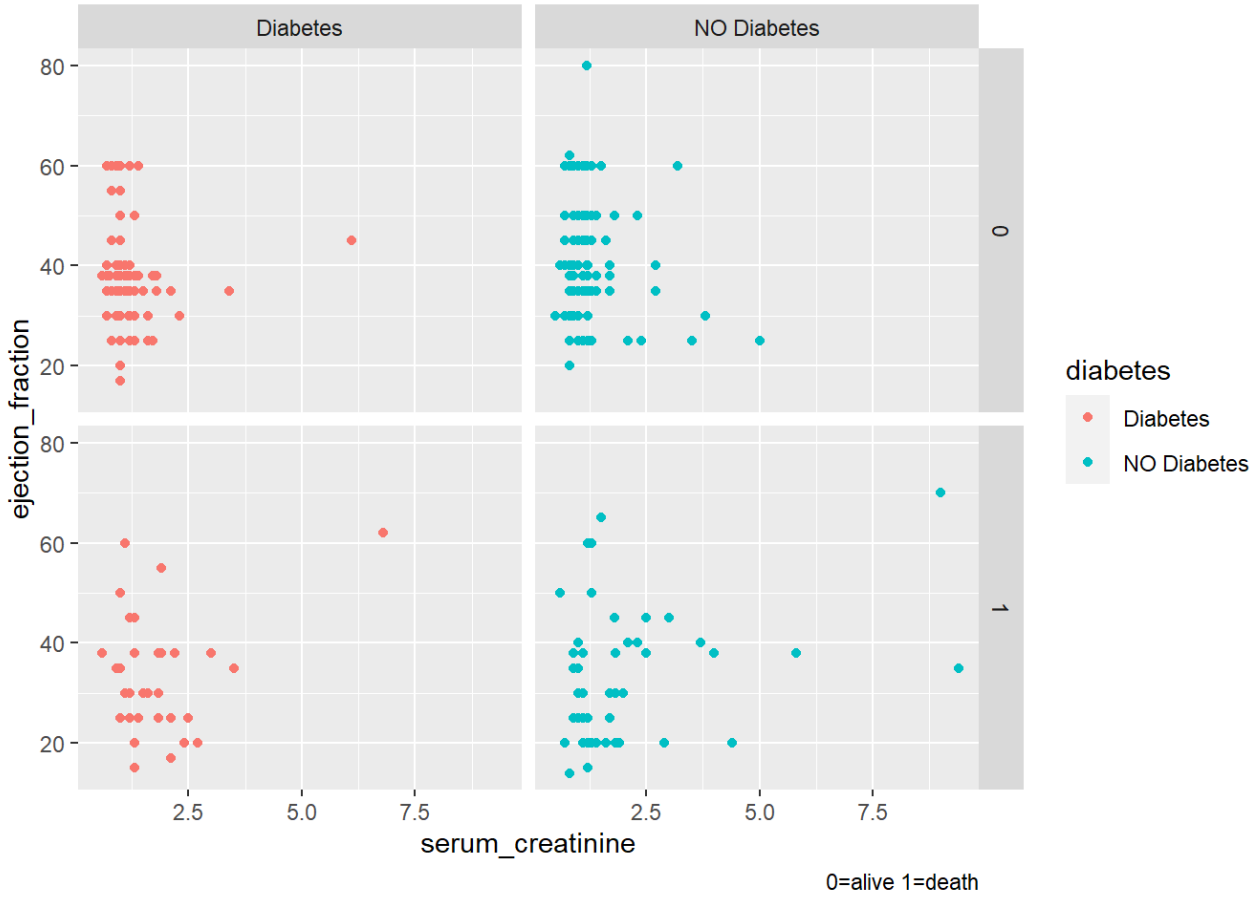
EDA (Exploratory Data Analysis - contd) - Correlation Matrix

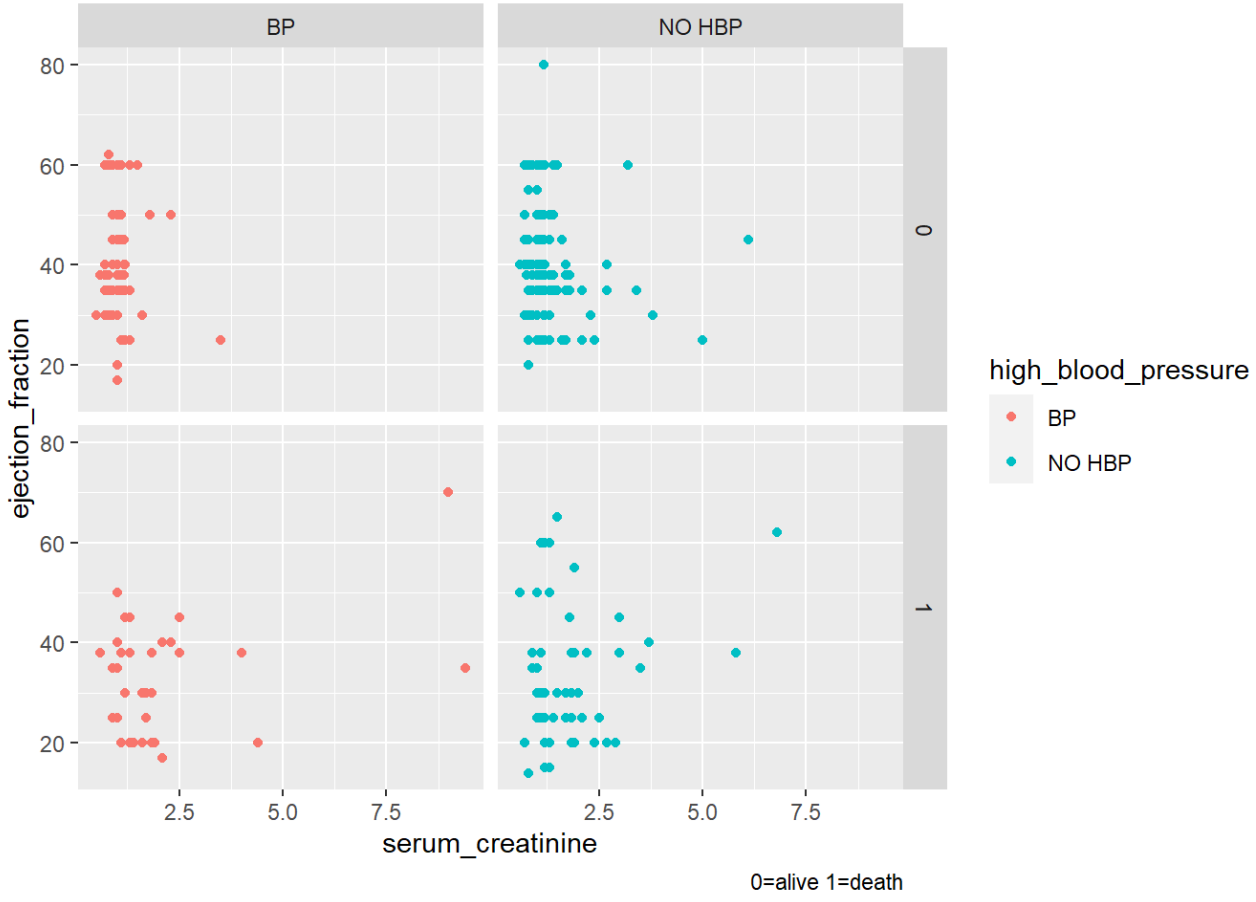


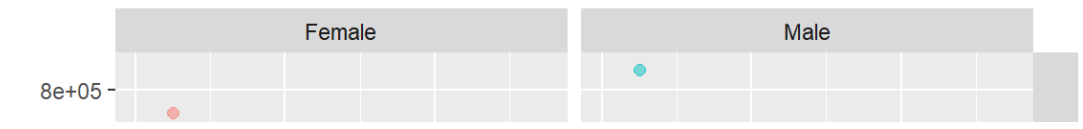
EDA (Exploratory Data Analysis - contd) - Survival percentage and Time (follow up days)



EDA (Exploratory Data Analysis - contd) - Correlation Plots


Diabetes
NO Diabetes
0
1
ejection_fraction
serum_creatinine
diabetes
Diabetes
NO Diabetes
0=alive 1=death


BP
NO HBP
0
1
ejection_fraction
serum_creatinine
high_blood_pressure
BP
NO HBP
0=alive 1=death


Female
Male
8e+05


Platelets
6e+05
4e+05
2e+05
0e+00
8e+05
6e+05
4e+05
2e+05
0e+00
40
60
80
40
60
80
Age
0
1
sex
Female
Male
0=alive 1=death

Logistic Regression model with all predictors & Prediction Result for this model

```
##
## Call:
## glm(formula = DEATH_EVENT ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2814  -0.5194  -0.1888   0.3793   2.2930
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               9.307e+00  6.280e+00   1.482 0.138370
## age                       5.647e-02  1.882e-02   3.001 0.002691 **
## anaemia                  -2.411e-01  4.168e-01  -0.578 0.563061
## creatinine_phosphokinase  4.335e-04  2.699e-04   1.606 0.108274
## diabetes                  5.718e-01  4.107e-01   1.392 0.163840
## ejection_fraction        -8.871e-02  1.855e-02  -4.783 1.73e-06 ***
## high_blood_pressure      -3.143e-01  4.196e-01  -0.749 0.453860
## platelets                -1.115e-06  2.073e-06  -0.538 0.590504
## serum_creatinine          7.233e-01  2.076e-01   3.483 0.000495 ***
## serum_sodium             -5.900e-02  4.422e-02  -1.334 0.182113
## sex                      -8.364e-01  4.784e-01  -1.748 0.080430 .
## smoking                  -1.836e-03  4.753e-01  -0.004 0.996917
## time                     -2.378e-02  3.631e-03  -6.551 5.72e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 313.97  on 246  degrees of freedom
## Residual deviance: 167.94  on 234  degrees of freedom
## AIC: 193.94
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] 0.8076923
```

The data above shows the prediction accuracy in case of Logistic Regression with the original model with all predictors is 80.7%.

To find the VIF for the original model with all predictors.

```
#VIF (Variance Inflation Factor) with original model with all predictors

vif(LogMod)
```

```
##                      age                  anaemia creatinine_phosphokinase
##                 1.171175                 1.130524                 1.130377
##                 diabetes        ejection_fraction      high_blood_pressure
##                 1.094767                 1.211300                 1.106589
##                platelets         serum_creatinine             serum_sodium
##                 1.069691                 1.173013                 1.084654
##                      sex                  smoking                     time
##                 1.414121                 1.312924                 1.295972
```

The data above shows that the VIF for the original model with all predictors is also less than 2.5

Using stepAIC, find the optimized model

```
## Start:  AIC=193.94
## DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
##     ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
##     serum_sodium + sex + smoking + time
##
##                            Df Deviance    AIC
## - smoking                   1   167.94 191.94
## - platelets                 1   168.23 192.23
## - anaemia                   1   168.27 192.27
## - high_blood_pressure       1   168.50 192.50
## - serum_sodium              1   169.71 193.71
## - diabetes                  1   169.90 193.90
## <none>                          167.94 193.94
## - creatinine_phosphokinase  1   171.03 195.03
## - sex                       1   171.08 195.08
## - age                       1   178.24 202.24
## - serum_creatinine          1   179.86 203.86
## - ejection_fraction         1   197.61 221.61
## - time                      1   239.60 263.60
##
## Step:  AIC=191.94
## DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
##     ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
##     serum_sodium + sex + time
##
##                            Df Deviance    AIC
## - platelets                 1   168.24 190.24
## - anaemia                   1   168.28 190.28
## - high_blood_pressure       1   168.50 190.50
## - serum_sodium              1   169.71 191.71
## - diabetes                  1   169.91 191.91
## <none>                          167.94 191.94
## - creatinine_phosphokinase  1   171.03 193.03
## - sex                       1   171.70 193.70
## - age                       1   178.25 200.25
## - serum_creatinine          1   180.01 202.01
## - ejection_fraction         1   197.62 219.62
## - time                      1   240.05 262.05
##
## Step:  AIC=190.24
## DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
##     ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium + sex + time
##
##                            Df Deviance    AIC
## - anaemia                   1   168.57 188.57
## - high_blood_pressure       1   168.85 188.85
## - diabetes                  1   170.11 190.11
## <none>                          168.24 190.24
## - serum_sodium              1   170.25 190.25
## - creatinine_phosphokinase  1   171.42 191.42
## - sex                       1   171.89 191.89
## - age                       1   178.52 198.52
## - serum_creatinine          1   180.35 200.35
```

```
## - ejection_fraction         1   197.91 217.91
## - time                      1   240.07 260.07
##
## Step:  AIC=188.57
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     high_blood_pressure + serum_creatinine + serum_sodium + sex +
##     time
##
##                            Df Deviance    AIC
## - high_blood_pressure       1   169.14 187.14
## - diabetes                  1   170.34 188.34
## <none>                          168.57 188.57
## - serum_sodium              1   170.71 188.71
## - sex                       1   171.97 189.97
## - creatinine_phosphokinase  1   172.07 190.07
## - age                       1   178.66 196.66
## - serum_creatinine          1   180.63 198.63
## - ejection_fraction         1   198.04 216.04
## - time                      1   241.70 259.70
##
## Step:  AIC=187.14
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time
##
##                            Df Deviance    AIC
## - diabetes                  1   170.90 186.90
## <none>                          169.14 187.14
## - serum_sodium              1   171.27 187.27
## - sex                       1   172.15 188.15
## - creatinine_phosphokinase  1   173.03 189.03
## - age                       1   178.86 194.86
## - serum_creatinine          1   181.86 197.86
## - ejection_fraction         1   198.36 214.36
## - time                      1   242.79 258.79
##
## Step:  AIC=186.9
## DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time
##
##                            Df Deviance    AIC
## <none>                          170.90 186.90
## - serum_sodium              1   173.61 187.61
## - sex                       1   174.42 188.42
## - creatinine_phosphokinase  1   174.49 188.49
## - age                       1   179.68 193.68
## - serum_creatinine          1   183.15 197.15
## - ejection_fraction         1   199.43 213.43
## - time                      1   243.31 257.31
```

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + creatinine_phosphokinase +
##     ejection_fraction + serum_creatinine + serum_sodium + sex +
##     time, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3814  -0.5030  -0.1850   0.3874   2.1936
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              10.8253126  6.1104382   1.772  0.07646 .
## age                       0.0507668  0.0181163   2.802  0.00507 **
## creatinine_phosphokinase  0.0004545  0.0002658   1.710  0.08725 .
## ejection_fraction        -0.0855337  0.0180885  -4.729 2.26e-06 ***
## serum_creatinine          0.6820178  0.1943757   3.509  0.00045 ***
## serum_sodium             -0.0712331  0.0431686  -1.650  0.09892 .
## sex                      -0.7725403  0.4157225  -1.858  0.06313 .
## time                     -0.0226869  0.0034382  -6.598 4.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 313.97  on 246  degrees of freedom
## Residual deviance: 170.90  on 239  degrees of freedom
## AIC: 186.9
##
## Number of Fisher Scoring iterations: 6
```

In R, $stepAIC$ is one of the most commonly used search method for feature selection. We try to keep on minimizing the $stepAIC$ value to come up with the final set of features.$stepAIC$ does not necessarily mean to improve the model performance, however, it is used to simplify the model without impacting much on the performance. So AIC quantifies the amount of information loss due to this simplification. AIC stands for **Akaike Information Criteria**.

AIC is only a relative measure among multiple models. AIC is similar adjusted R-squared as it also penalizes for adding more variables to the model. The absolute value of AIC does not have any significance. We only compare AIC value whether it is increasing or decreasing by adding more variables. Also in case of multiple models, the one which has lower AIC value is preferred. $stepAIC$ also removes the multicollinearity (if it exists), from the model.

```
##           1           2           3           4           6           7
## 0.987369462 0.983602767 0.962566833 0.909618202 0.955628506 0.971299162
##           8           9          10          11          12          13
## 0.332200333 0.401989038 0.999624336 0.978468775 0.829385816 0.722899327
##          15          17          18          19          20          21
## 0.796221656 0.833479882 0.929937921 0.935023287 0.761083137 0.941317875
##          22          23          24          25          27          30
## 0.928714754 0.689691131 0.126757314 0.971038689 0.929895290 0.924648407
##          31          32          33          34          35          36
## 0.949281614 0.926725107 0.804096785 0.768568452 0.303670277 0.952017975
##          37          38          41          42          43          44
## 0.696047502 0.684112974 0.959295984 0.778809557 0.546197998 0.469322849
##          45          46          47          48          49          50
## 0.248433160 0.620269321 0.835097862 0.446357850 0.994787705 0.724756116
##          51          52          53          54          55          56
## 0.818341079 0.872642578 0.974222398 0.567639113 0.860082711 0.964605346
##          57          58          59          62          63          64
## 0.807375501 0.333428638 0.731654489 0.723669241 0.351697828 0.194262490
##          65          66          67          68          69          70
## 0.020800283 0.947762358 0.819419283 0.701483620 0.820510627 0.798546604
##          71          73          74          75          76          77
## 0.100688765 0.975721777 0.209486358 0.848975090 0.581598058 0.139786226
##          78          79          80          82          83          84
## 0.116803903 0.525257224 0.246669167 0.225070330 0.910756440 0.449105680
##          85          87          88          89          91          93
## 0.663303748 0.283259534 0.048902688 0.090059470 0.236925769 0.061546121
##          94          95          96          97          98          99
## 0.671079420 0.281761320 0.034918887 0.690987665 0.169090855 0.718085483
##         100         101         102         103         104         105
## 0.367376484 0.521360769 0.342513627 0.711531303 0.673097591 0.274337370
##         106         107         108         109         110         111
## 0.801613854 0.177578347 0.265708815 0.427027051 0.148223528 0.208893805
##         112         114         115         118         119         120
## 0.276612119 0.167678831 0.449841709 0.568209869 0.079666149 0.852345099
##         121         125         126         127         128         129
## 0.079661800 0.677488988 0.115459058 0.892209962 0.054987787 0.294533297
##         130         132         133         134         136         138
## 0.372944055 0.894997591 0.211549892 0.025016343 0.336667133 0.799303428
##         139         141         143         144         145         146
## 0.413631800 0.490461887 0.439029721 0.199936701 0.676567728 0.119392241
##         147         148         150         151         152         153
## 0.172989134 0.060715463 0.365843603 0.554036474 0.020632815 0.056566629
##         155         157         159         160         161         162
## 0.358754696 0.277392032 0.289288696 0.066691101 0.188900598 0.063078848
##         163         164         165         166         167         168
## 0.136893006 0.283129434 0.232086715 0.576872421 0.012917693 0.597777702
##         170         171         172         173         174         175
## 0.388612457 0.114910004 0.168380600 0.012885082 0.149902630 0.118316007
##         176         177         178         179         180         181
## 0.011538067 0.183904340 0.027705530 0.009277515 0.046217986 0.063159570
##         182         183         184         186         188         189
## 0.194810421 0.275029514 0.297241391 0.126724777 0.466870552 0.086640463
##         190         191         193         198         200         201
## 0.014942876 0.337255956 0.028516663 0.133626528 0.404695704 0.035706264
```

```
##         202         203         204         205         206         207
## 0.002786453 0.007263530 0.364511742 0.051973514 0.025145938 0.013923461
##         208         209         211         212         213         214
## 0.108872841 0.107402922 0.264621262 0.001964903 0.032498731 0.142828311
##         215         217         218         219         221         222
## 0.048310784 0.038124174 0.427034159 0.092035852 0.383069633 0.006783900
##         223         224         225         226         228         229
## 0.011196880 0.047666650 0.064433113 0.057729731 0.055049555 0.812427757
##         230         231         232         233         235         236
## 0.271917115 0.231012901 0.054225582 0.008677715 0.010037437 0.017712685
##         237         238         239         241         242         243
## 0.007419805 0.099507458 0.060718552 0.082999391 0.071786675 0.009791197
##         244         246         247         248         250         251
## 0.068255290 0.024108411 0.090170279 0.167127977 0.042770860 0.077885714
##         252         253         254         255         256         257
## 0.025715606 0.008032357 0.115400736 0.001864389 0.015984802 0.062383598
##         258         259         260         262         263         264
## 0.009656637 0.016958093 0.002766096 0.031267316 0.096727480 0.004275297
##         265         266         267         268         269         270
## 0.011676416 0.006238288 0.110520610 0.009546262 0.008406168 0.007376483
##         271         272         273         274         275         276
## 0.023668847 0.014637599 0.041461298 0.001906093 0.018241888 0.009494175
##         278         279         280         281         282         284
## 0.015028137 0.028862757 0.018402958 0.036596230 0.044432126 0.020191490
##         285         286         287         288         289         291
## 0.002533369 0.006902783 0.013859439 0.003260934 0.025685366 0.001300395
##         292         294         295         296         297         298
## 0.010464507 0.007427349 0.003744121 0.017349633 0.001257481 0.005561190
##         299
## 0.001963510
```

```
##        Predictions
## Actuals FALSE TRUE
##       0   151   14
##       1    23   59
```

```
## [1] 0.8502024
```

```
##           5          14          16          26          28          29
## 0.997224426 0.539647104 0.734041094 0.809624897 0.603948994 0.973269259
##          39          40          60          61          72          81
## 0.946297419 0.878489225 0.883206380 0.957674404 0.347453446 0.681183135
##          86          90          92         113         116         117
## 0.052808025 0.391930756 0.278993345 0.495549325 0.320028607 0.071762619
##         122         123         124         131         135         137
## 0.449079539 0.278542836 0.425461231 0.034743943 0.864491813 0.072247539
##         140         142         149         154         156         158
## 0.267758016 0.177625862 0.667753966 0.228393892 0.384569404 0.238356753
##         169         185         187         192         194         195
## 0.204895440 0.133255534 0.024442922 0.027373747 0.109464011 0.152162207
##         196         197         199         210         216         220
## 0.053049706 0.046214006 0.237675318 0.100098455 0.113754303 0.044099804
##         227         234         240         245         249         261
## 0.092612252 0.023403497 0.009346671 0.034175319 0.007284552 0.006823063
##         277         283         290         293
## 0.037077458 0.072437406 0.040919147 0.003629188
```

```
##        Predictions
## Actuals FALSE TRUE
##       0    35    3
##       1     4   10
```

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + creatinine_phosphokinase +
##     ejection_fraction + serum_creatinine + serum_sodium + sex +
##     time, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3814  -0.5030  -0.1850   0.3874   2.1936
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              10.8253126  6.1104382   1.772  0.07646 .
## age                       0.0507668  0.0181163   2.802  0.00507 **
## creatinine_phosphokinase  0.0004545  0.0002658   1.710  0.08725 .
## ejection_fraction        -0.0855337  0.0180885  -4.729 2.26e-06 ***
## serum_creatinine          0.6820178  0.1943757   3.509  0.00045 ***
## serum_sodium             -0.0712331  0.0431686  -1.650  0.09892 .
## sex                      -0.7725403  0.4157225  -1.858  0.06313 .
## time                     -0.0226869  0.0034382  -6.598 4.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 313.97  on 246  degrees of freedom
## Residual deviance: 170.90  on 239  degrees of freedom
## AIC: 186.9
##
## Number of Fisher Scoring iterations: 6
```

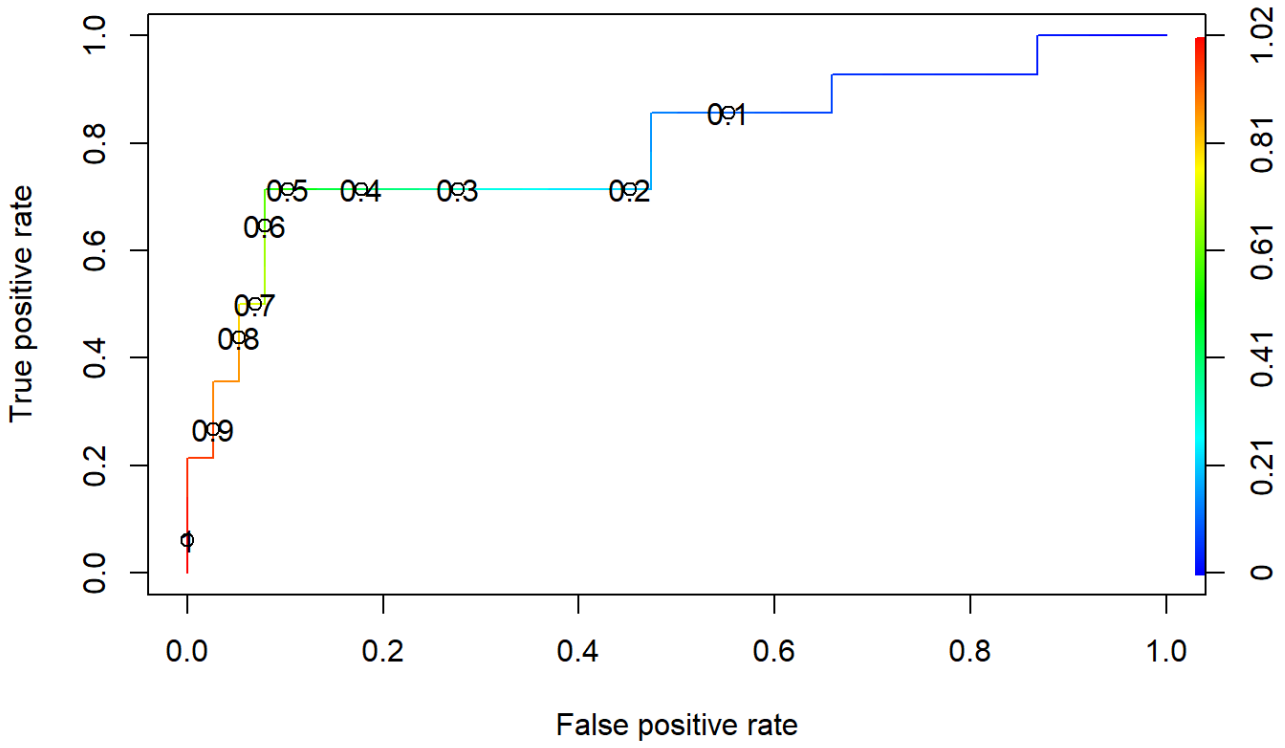
```
## [1] 0.8653846
```

The prediction accuracy in case of Logistic Regression with the selected model (res_test > 0.5) is 86.5%.

```
## [1] 0.7307692
```

When we reduce the res_test probability to res_test > 0.3, the prediction accuracy is 73%, which is lesser than than the one when the res_test probability res_test > 0.5 .

ROC plot



So, as we can see here, we want the $TPR(TruePositiveRate)$ highest as possible and $FPR(FalsePostiveRate)$ lowest as possible because the false positive rate is a misclassification. If we decrease the threshold, the FPR increases with is not desirable. so, for our model we keep the threshold to be >0.5 to have the best prediction accuracy.

Finding out AUC (Area Under the Curve)

```
## [[1]]
## [1] 0.7951128
```

What we want to find is the area under the curve that basically shows that the more area under this curve the better the accuracy of the model is. The ideal value is going to be obviously **1**. To find the area under this curve we must run the ROC performance. In our case the AUC value is **0.80** which is very good value. So the model looks good.

LDA classification

```
## Call:
## lda(DEATH_EVENT ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.6680162 0.3319838
##
## Group means:
##        age   anaemia creatinine_phosphokinase  diabetes ejection_fraction
## 0 58.91313 0.4060606                 517.8788 0.4060606          40.95758
## 1 65.53252 0.4756098                 642.5854 0.4146341          33.30488
##   high_blood_pressure platelets serum_creatinine serum_sodium       sex
## 0           0.3333333  269123.6         1.184121     137.1333 0.6666667
## 1           0.4146341  261677.8         1.826951     135.2561 0.6219512
##     smoking      time
## 0 0.3272727 160.81818
## 1 0.3170732  69.92683
##
## Coefficients of linear discriminants:
##                                   LD1
## age                       2.455275e-02
## anaemia                  -6.190849e-02
## creatinine_phosphokinase  1.967963e-04
## diabetes                  2.568402e-01
## ejection_fraction        -4.606770e-02
## high_blood_pressure      -1.188168e-01
## platelets                -3.853963e-07
## serum_creatinine          3.454956e-01
## serum_sodium             -2.996829e-02
## sex                      -4.134172e-01
## smoking                  -8.704843e-03
## time                     -1.221710e-02
```

```
##
##      0  1
##   0 31  5
##   1  7  9
```

Prediction Accuracy in LDA = $(31+9)/(31+7+5+9)$ **= 76.9% , Error =** $(7+5)/(31+7+5+9)$ **= 23.1%**

QDA classification

```
## Call:
## qda(DEATH_EVENT ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.6680162 0.3319838
##
## Group means:
##        age   anaemia creatinine_phosphokinase  diabetes ejection_fraction
## 0 58.91313 0.4060606                 517.8788 0.4060606          40.95758
## 1 65.53252 0.4756098                 642.5854 0.4146341          33.30488
##   high_blood_pressure platelets serum_creatinine serum_sodium       sex
## 0           0.3333333  269123.6         1.184121     137.1333 0.6666667
## 1           0.4146341  261677.8         1.826951     135.2561 0.6219512
##     smoking      time
## 0 0.3272727 160.81818
## 1 0.3170732  69.92683
```

```
##
##      0  1
##   0 35  8
##   1  3  6
```

Prediction Accuracy in QDA = $(35+6)/(35+8+3+6)$ **=** 78.8% **, Error =** $(8+3)/(35+8+3+6)$ **=** 21.2%

KNN classification

```
##          age    anaemia creatinine_phosphokinase   diabetes ejection_fraction
## 1  1.1909487 -0.8696469              0.000165451 -0.8461608      -1.527997920
## 2 -0.4904571 -0.8696469              7.502062717 -0.8461608      -0.007064906
## 3  0.3502458 -0.8696469             -0.449185725 -0.8461608      -1.527997920
## 4 -0.9108085  1.1460462             -0.485257493 -0.8461608      -1.527997920
## 5  0.3502458  1.1460462             -0.434757017  1.1778559      -1.527997920
## 6  2.4520030  1.1460462             -0.551217299 -0.8461608       0.161927651
##   high_blood_pressure     platelets serum_creatinine serum_sodium        sex
## 1           1.3569966  1.678834e-02       0.48923681  -1.50151891  0.7344569
## 2          -0.7344569  7.523048e-09      -0.28407611  -0.14173853  0.7344569
## 3          -0.7344569 -1.036336e+00      -0.09074788  -1.72814897  0.7344569
## 4          -0.7344569 -5.455595e-01       0.48923681   0.08489153  0.7344569
## 5          -0.7344569  6.507077e-01       1.26254973  -4.67433977 -1.3569966
## 6           1.3569966 -6.069065e-01       0.68256504  -1.04825878  0.7344569
##     smoking      time DEATH_EVENT
## 1 -0.686531 -1.626775    1.451727
## 2 -0.686531 -1.601007    1.451727
## 3  1.451727 -1.588122    1.451727
## 4 -0.686531 -1.588122    1.451727
## 5 -0.686531 -1.575238    1.451727
## 6  1.451727 -1.575238    1.451727
```

```
##
## HF_norm_KNN_pred  0  1
##                0 38  2
##                1  0 12
```

Prediction Accuracy in $KNN(k = 16)$ = $(38 + 12)/(38 + 12 + 0 + 2)$ = 96% , Error = $(2 + 0)/(38 + 12 + 0 + 2)$ = 4%