Logistic Regression Analysis

By Manisha, Rajasree, Rahul & Timothy

Heart failure prediction model

AGENDA

1. INTRODUCTION
2. EXPLORATORY DATA ANALYSIS
3. PURPOSEFUL VARIABLE SELECTION
4. STEPAIC
5. BESTGLM
6. ROC CURVES
7. CONCLUSION

Introduction

Cardiovascular diseases (CVDs) are the number 1 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 299 observations and 12 features that can be used to predict mortality by heart failure.

About the Dataset

Variable	Description	Measure
Age	The age of the person	Years (Numerical)
Anaemia	Decrease of red blood cells or hemoglobin	Has->1, Otherwise->0 (Binary/Categorical-Nominal)
Creatinine_phosphokinase	Level of the CPK enzyme in the blood	mcg/L (Numerical)
Diabetes	If the patient has diabetes	Has->1, Otherwise->0 (Binary/Categorical-Nominal)
Ejection_fraction	Percentage of blood leaving the heart at each contraction	percentage (Numerical)
High_blood_pressure	If the patient has hypertension	Has->1, Otherwise->0 (Binary/Categorical-Nominal)
Platelets	Platelets in the blood	kilo platelets/mL (Numerical)

We start off by understanding the type of data in the dataframe. We can see from the below summary that there are 12 explanatory variables associated with mortality by heart failure .The response variable is Binary variable with 1 indicating the patient died and 0 means patient did not die.

About the Dataset

Variable	Description	Measure
Serum_creatinine	Level of serum creatinine in the blood	mg/dL (Numerical)
Serum_sodium	Level of serum sodium in the blood	mEq/L (Numerical)
Sex	Gender as woman or man	Male->1, Female->0 (Binary/Categorical-Nominal)
Smoking	If the patient smokes or not	smokes->1, Otherwise->0 (Binary/Categorical-Nominal)
Time	Follow-up period	In days (Numerical)
DEATH_EVENT (Response Variable)	If the patient deceased during the follow-up period	Dead->1, Alive->0 (Binary/Categorical-Nominal)

We start off by understanding the type of data in the dataframe. We can see from the below summary that there are 12 explanatory variables associated with mortality by heart failure .The response variable is Binary variable with 1 indicating the patient died and 0 means patient did not die.

Original Dataset

```
'data.frame':   299 obs. of  13 variables:
 $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
 $ anaemia                 : int  0 0 0 1 1 1 1 1 0 1 ...
 $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
 $ diabetes                : int  0 0 0 0 1 0 0 1 0 0 ...
 $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
 $ high_blood_pressure     : int  1 0 0 0 0 1 0 0 0 1 ...
 $ platelets               : num  265000 263358 162000 210000 327000 ...
 $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
 $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
 $ sex                     : int  1 1 1 1 0 1 1 1 0 1 ...
 $ smoking                 : int  0 0 1 0 0 1 0 1 0 1 ...
 $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
 $ DEATH_EVENT             : int  1 1 1 1 1 1 1 1 1 1 ...
```

Dataset after treating binary variables as factors

```
'data.frame':   299 obs. of  13 variables:
 $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
 $ anaemia                 : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...
 $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
 $ diabetes                : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
 $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
 $ high_blood_pressure     : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...
 $ platelets               : num  265000 263358 162000 210000 327000 ...
 $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
 $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
 $ sex                     : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...
 $ smoking                 : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...
 $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
 $ DEATH_EVENT             : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

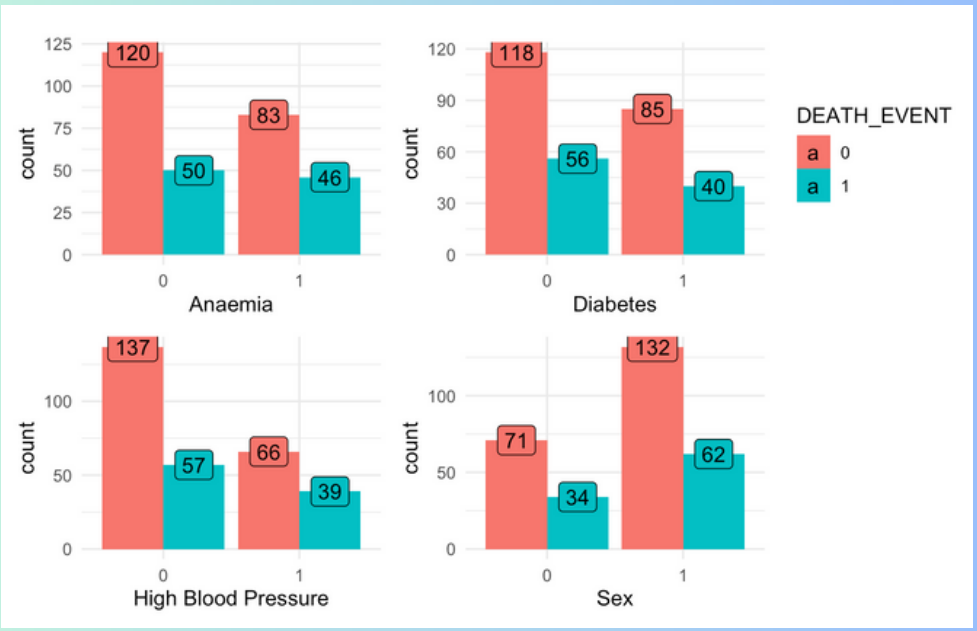
Missing Values

	Variable	Missing_Values
age	age	0
anaemia	anaemia	0
creatinine_phosphokinase	creatinine_phosphokinase	0
diabetes	diabetes	0
ejection_fraction	ejection_fraction	0
high_blood_pressure	high_blood_pressure	0
platelets	platelets	0
serum_creatinine	serum_creatinine	0
serum_sodium	serum_sodium	0
sex	sex	0
smoking	smoking	0
time	time	0
DEATH_EVENT	DEATH_EVENT	0
predicted_prob	predicted_prob	0

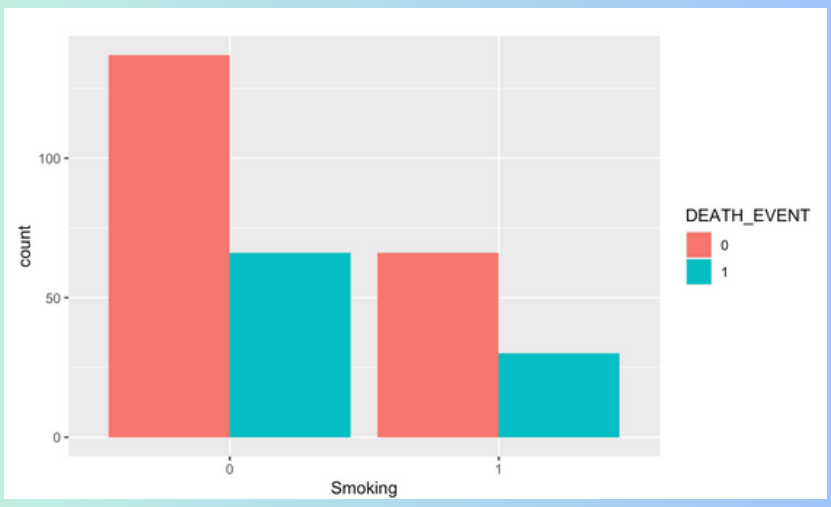
Numerical Summary

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

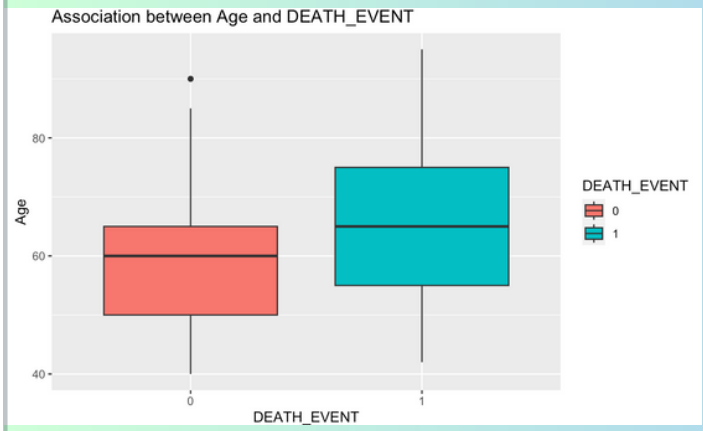
Data Visualisation



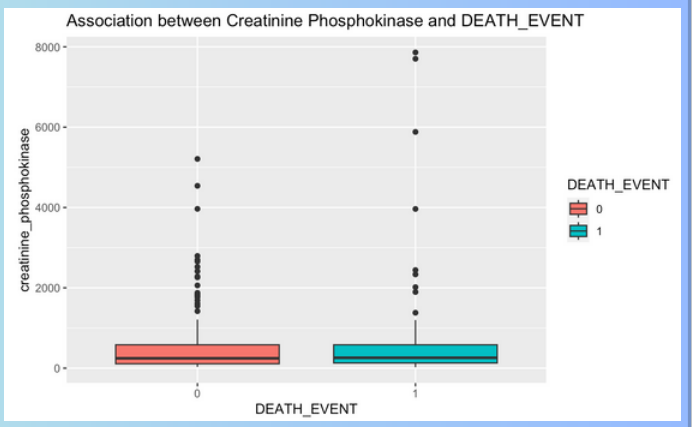
Data Visualisation



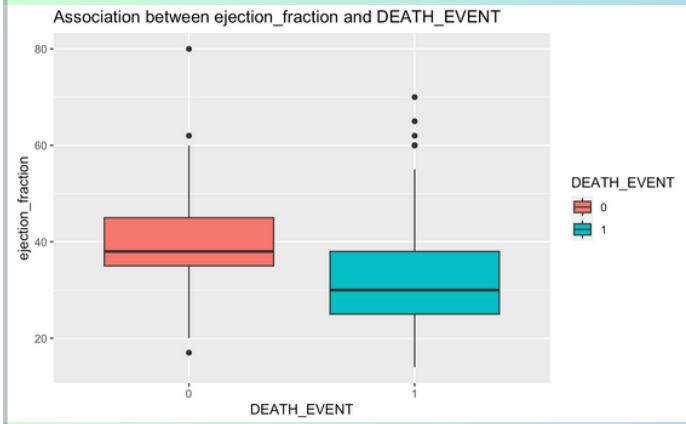
Association between Age with Death Event



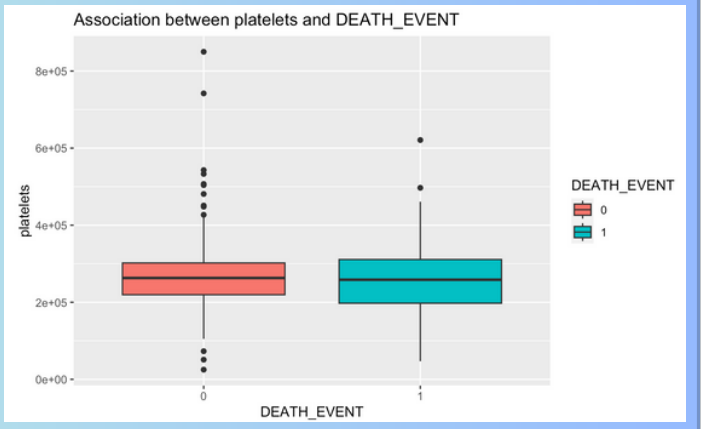
Association between Creatine Phosphokinase and Death Event



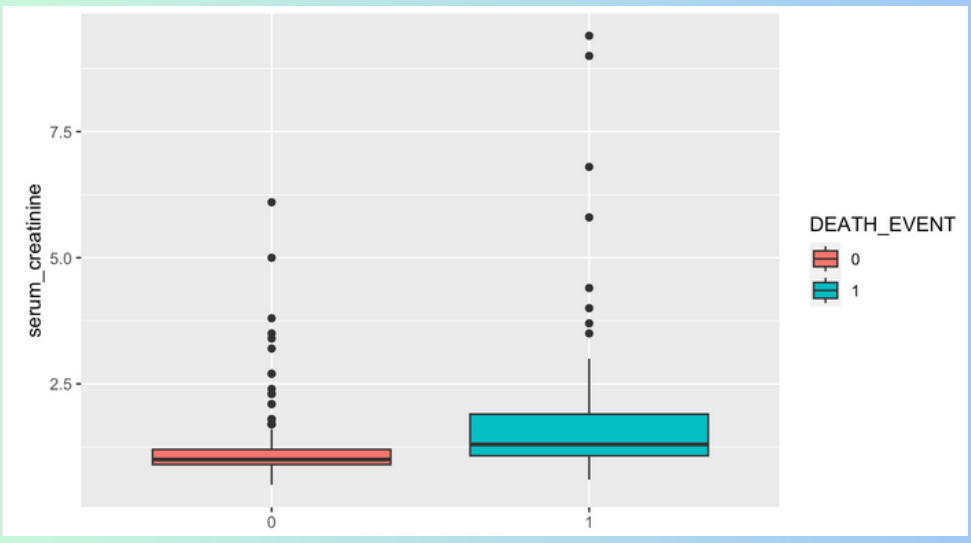
Association between Ejection_Fraction and Death_Event



Association between Plalets and Death Event

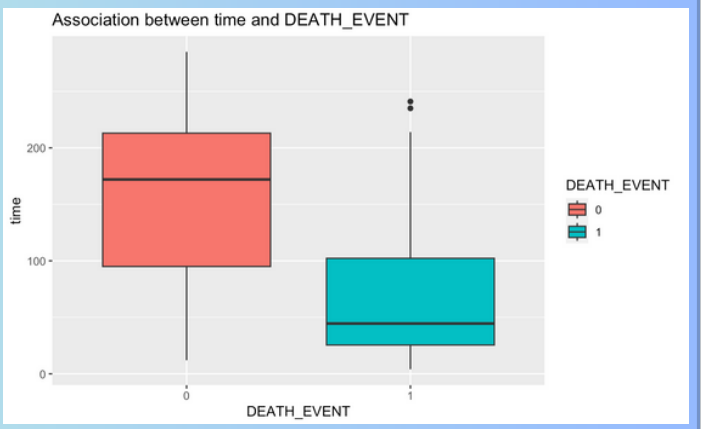


Association between Serum_creatine and Death_Event



Association between serum_sodium & Death_Event

Association between Time & Death_Event



Multicollinearity

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

Purposeful Variable Selection

STEP 1 : Construct an initial main-effects model

H0 : Reduced model is better
Ha : Saturated model is better

```
**

##                    Variable      P_Value        Decision
## 1                       age 1.084786e-05     Significant
## 2                   anaemia 2.526527e-01 Not Significant
## 3  creatinine_phosphokinase 2.902985e-01 Not Significant
## 4                  diabetes 9.731974e-01 Not Significant
## 5         ejection_fraction 1.329145e-06     Significant
## 6       high_blood_pressure 1.722801e-01     Significant
## 7                 platelets 3.903722e-01 Not Significant
## 8          serum_creatinine 1.154030e-07     Significant
## 9              serum_sodium 7.639974e-04     Significant
## 10                      sex 9.405237e-01 Not Significant
## 11                  smoking 8.269993e-01 Not Significant
## 12                     time 9.999320e-23     Significant
```

Initial Model Summary

```
Call:
glm(formula = DEATH_EVENT ~ age + ejection_fraction + high_blood_pressure +
    serum_creatinine + serum_sodium + time, family = binomial,
    data = heart)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          9.508906   5.408913   1.758  0.07875 .
age                  0.042535   0.015034   2.829  0.00466 **
ejection_fraction   -0.073461   0.015790  -4.652 3.28e-06 ***
high_blood_pressure1 -0.057587   0.348809  -0.165  0.86887
serum_creatinine     0.685565   0.174545   3.928 8.58e-05 ***
serum_sodium        -0.064480   0.038391  -1.680  0.09304 .
time                -0.020961   0.002945  -7.118 1.09e-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: 223.46  on 292  degrees of freedom
AIC: 237.46

Number of Fisher Scoring iterations: 6
```

STEP 2: Backward Elimination

H0 : Reduced model is better
Ha : Saturated model is better

	Variable	P_Value	Decision
1	time	1.403748e-18	Significant
2	serum_sodium	9.382812e-02	Not Significant
3	serum_creatinine	7.175075e-05	Significant
4	high_blood_pressure	8.687719e-01	Not Significant
5	ejection_fraction	2.844957e-07	Significant
6	age	3.448485e-03	Significant

Summary after Step 2

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time, family = binomial, data = heart)
##
## 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
```

STEP 3:

Checking Significance of variables not included after step1

H0 : Reduced model is better
Ha : Saturated model is better

```
##                        Variable   P_Value        Decision
## 1                       anaemia 0.8724652 Not Significant
## 2 creatinine_phosphokinase 0.2774741 Not Significant
## 3                      diabetes 0.5106350 Not Significant
## 4                     platelets 0.5404552 Not Significant
## 5                           sex 0.2691591 Not Significant
## 6                       smoking 0.5512882 Not Significant
```

Summary after 3

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time, family = binomial, data = heart)
##
## 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
```

STEP 4: Interaction Term

	Interaction	P_Value	Decision
1	age:ejection_fraction	0.18981653	Not Significant
2	age:serum_creatinine	0.40004436	Not Significant
3	age:time	0.21752243	Not Significant
4	ejection_fraction:serum_creatinine	0.33304352	Not Significant
5	ejection_fraction:time	0.01068685	Significant
6	serum_creatinine:time	0.59566716	Not Significant

H0 : Reduced model is better
Ha : Saturated model is better

Final Model(Model1)

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time + ejection_fraction:time, family = binomial, data = heart)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.3973042  1.2984707  -1.076   0.2819
## age                     0.0418709  0.0151455   2.765   0.0057 **
## ejection_fraction      -0.0188076  0.0265627  -0.708   0.4789
## serum_creatinine        0.8286812  0.2110598   3.926 8.63e-05 ***
## time                    0.0012112  0.0089767   0.135   0.8927
## ejection_fraction:time -0.0006536  0.0002661  -2.456   0.0141 *
## ---
## 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: 219.78  on 293  degrees of freedom
## AIC: 231.78
##
## Number of Fisher Scoring iterations: 6
```

$$\text{logit}[P(\hat{Y}=1)] = -1.397 + 0.0418 \times \text{age} - 0.0188 \times \text{ejection\_fraction} + 0.8286 \times \text{serum\_creatinine} + 0.0012 \times \text{time} - 0.000653 \times \text{ejection\_fraction:time}$$

Step AIC : start

```
Start:  AIC=242.48
DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
    ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
    serum_sodium + sex + smoking + time + predicted_prob + ejection_fraction:time

                           Df Deviance    AIC
- high_blood_pressure       1   212.48 240.48
- smoking                   1   212.48 240.48
- diabetes                  1   212.48 240.48
- anaemia                   1   212.50 240.50
- platelets                 1   213.10 241.10
- predicted_prob            1   213.20 241.20
- serum_creatinine          1   214.06 242.06
- age                       1   214.29 242.29
<none>                          212.48 242.48
- sex                       1   214.58 242.58
- serum_sodium              1   214.65 242.65
- ejection_fraction:time    1   214.84 242.84
- creatinine_phosphokinase  1   214.95 242.95
```

Final Step AIC Model

```
Call:
glm(formula = DEATH_EVENT ~ age + creatinine_phosphokinase +
    ejection_fraction + serum_creatinine + serum_sodium + sex +
    time + ejection_fraction:time, family = binomial, data = heart)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               7.1272546  5.8220089   1.224 0.220880
age                       0.0451926  0.0158289   2.855 0.004303 **
creatinine_phosphokinase  0.0002529  0.0001820   1.390 0.164619
ejection_fraction        -0.0192743  0.0274729  -0.702 0.482943
serum_creatinine          0.7865430  0.2207049   3.564 0.000366 ***
serum_sodium             -0.0615284  0.0405254  -1.518 0.128946
sex1                     -0.5874321  0.3750608  -1.566 0.117294
time                      0.0015072  0.0092733   0.163 0.870884
ejection_fraction:time   -0.0006710  0.0002769  -2.423 0.015384 *
---
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: 213.83  on 290  degrees of freedom
AIC: 231.83

Number of Fisher Scoring iterations: 6
```

Model2 :using stepAIC

```
Call:
glm(formula = DEATH_EVENT ~ age + creatinine_phosphokinase +
    serum_creatinine + serum_sodium + sex + ejection_fraction +
    time + ejection_fraction:time, family = binomial, data = heart)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              7.1272546  5.8220089   1.224 0.220880
age                      0.0451926  0.0158289   2.855 0.004303 **
creatinine_phosphokinase 0.0002529  0.0001820   1.390 0.164619
serum_creatinine         0.7865430  0.2207049   3.564 0.000366 ***
serum_sodium            -0.0615284  0.0405254  -1.518 0.128946
sex1                    -0.5874321  0.3750608  -1.566 0.117294
ejection_fraction       -0.0192743  0.0274729  -0.702 0.482943
time                     0.0015072  0.0092733   0.163 0.870884
ejection_fraction:time  -0.0006710  0.0002769  -2.423 0.015384 *
---
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: 213.83  on 290  degrees of freedom
AIC: 231.83

Number of Fisher Scoring iterations: 6
```

logit[P(Y^=1)]= 7.1272+0.04519×age+0.00025×creatinine_phosphokinase + 0.7865×serum_creatinine−0.0615×serum_sodium−0.5874×sex1−0.0192×ejection_fraction+ 0.0015×time-0.000671×(ejection_fraction×time)

Model3 :using Best subset GLM

```
Morgan-Tatar search since family is non-gaussian.
BIC
BICq equivalent for q in (0.341964872944168, 0.847734048815748)
Best Model:
                   Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)      -1.69764391 9.533605e-01 -1.780695 7.496236e-02
age               0.03784585 1.474087e-02  2.567409 1.024617e-02
serum_creatinine  0.82674999 2.002349e-01  4.128900 3.645025e-05
interaction_term -0.00065203 8.446922e-05 -7.719143 1.171148e-14
```

Model3 :using Best subset GLM

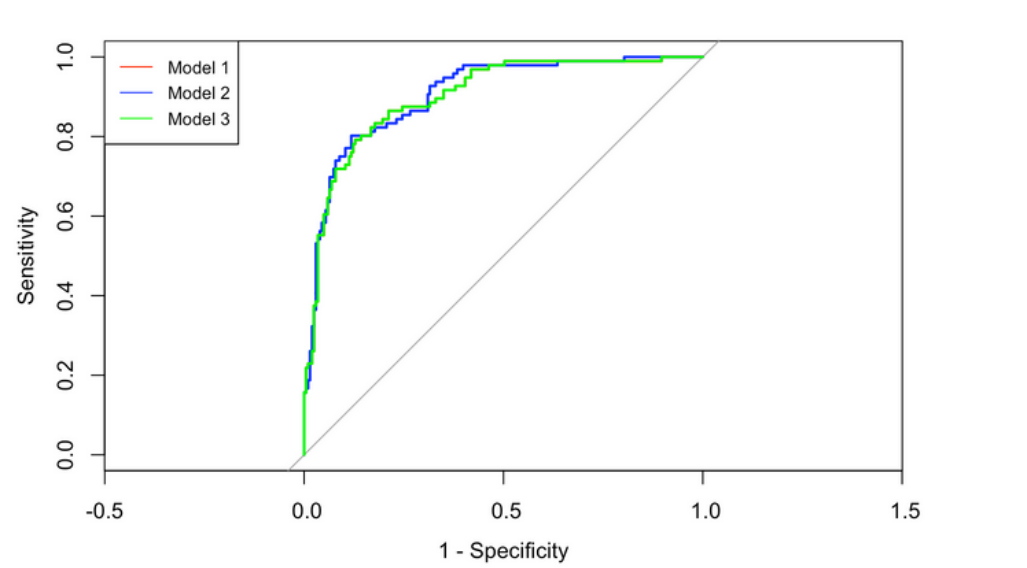
```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time + ejection_fraction:time, family = binomial, data = heart)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.3973042  1.2984707  -1.076   0.2819
## age                     0.0418709  0.0151455   2.765   0.0057 **
## ejection_fraction      -0.0188076  0.0265627  -0.708   0.4789
## serum_creatinine        0.8286812  0.2110598   3.926 8.63e-05 ***
## time                    0.0012112  0.0089767   0.135   0.8927
## ejection_fraction:time -0.0006536  0.0002661  -2.456   0.0141 *
## ---
## 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: 219.78  on 293  degrees of freedom
## AIC: 231.78
##
## Number of Fisher Scoring iterations: 6
```

```
                        exp.model3.coefficients.
(Intercept)                            0.2472626
age                                    1.0427599
serum_creatinine                       2.2902963
ejection_fraction                      0.9813682
time                                   1.0012120
ejection_fraction:time                 0.9993466
>
```

logit[P(Y^=1)]= −1.397+0.0418×age-0.0188×ejection_fraction+0.8286×serum_creatinine+0.0012×time−0.000653×ejection_fraction:time

MODELS	PREDICTORS
MODEL1	Age, Ejection_Fraction, Serum_Creatine, Time, Ejection_fraction:Time
MODEL2	Age, creatinine_phosphokinase,Serum_Creatinine, serum_sodium,sex1,Ejection_Fraction,Time,Ejection_Fraction:Time
MODEL3	Age, Ejection_Fraction, Serum_Creatine, Time, Ejection_fraction:Time

ROC Curves


Model 1
Model 2
Model 3
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
1 - Specificity

AUC Comparison

```
[1] "Area under ROC curve for Model 1 = 0.900913"
[1] "Area under ROC curve for Model 2 = 0.905378"
[1] "Area under ROC curve for Model 3 = 0.900913"
```

Conclusion

logit[P(Y^=1)]= −1.397+0.0418×age-0.0188×ejection_fraction+0.8286×serum_creatinine+0.0012×time−0.000653×ejection_fraction:time

Age(0.0418): The estimated odds of death event due to heart failure multiplied by 1.042 for each 1-year increase in age when other variables are held constant.

Ejection_fraction (-0.0188): The estimated odds of death event due to heart failure multiplied by 0.98 for each 1% increase in ejection_fraction when other variables are held constant.

Serum_creatinine (0.8286): The estimated odds of death event due to heart failure multiplied by 2.29 for each 1 unit increase in Serum_creatinine level when other variables are held constant.

Time (0.0012): The estimated odds of death event due to heart failure multiplied by 1.0012 for each 1-day increase in time when other variables are held constant.

Interaction(-0.04) [Time is held constant]: The estimated odds of death event due to heart failure multiplied by 0.95 times for each 1% increase in ejection_fraction on the fixed level of time when other variables are held constant.

Interaction(0.001) [ejection_fraction is held constant]: The estimated odds of death event due to heart failure multiplied by 1.1 times for a 1-day increase in Time on the fixed level of ejection_fraction when other variables are held constant.