**Purposeful selection of Heart failure dataset**

1. Step 1: For step 1, any predictor with p-value of less than 0.2 will be considered for the final model. The model with the sole predictor is compared with null model to check the significance, if any.
2. Age:

A black text with white text

Description automatically generated with medium confidence

As p-value <0.2, age is statistically significant for the main model.

1. Anaemia:

A close-up of a computer screen

Description automatically generated

As p-value > 0.2, anameia is not statistically significant for the main model.

1. Creatine\_phosphokinase:

A close-up of a computer screen

Description automatically generated

As p-value > 0.2, creatine\_phosphokinase is not statistically significant for the main model.

1. Diabetes:

A close-up of a number

Description automatically generated

As p-value >0.2, diabetes is not statistically significant for the main model.

1. Ejection fraction:

A screenshot of a computer code

Description automatically generated

As p-value <0.2, ejection fraction is statistically significant for the main model.

1. High blood pressure:

A screenshot of a computer code

Description automatically generated

As p-value >0.2, high blood pressure is not statistically significant for the main model.

1. Platelets:

A screenshot of a computer

Description automatically generated

As p-value >0.2, platelets is not statistically significant for the main model.

1. Serum creatine:

A computer code with black text

Description automatically generated

As p-value <0.2, serum\_creatine is statistically significant for the main model.

1. Serum sodium:

A screenshot of a computer code

Description automatically generated

As p-value <0.2, serum sodium is statistically significant for the main model.

1. Sex:

A black text on a white background

Description automatically generated

As p-value >0.2, sex is not statistically significant for the main model.

1. Smoking:

A close-up of a number

Description automatically generated

As p-value >0.2, smoking is not statistically significant for the main model.

1. Time:

A close-up of a number

Description automatically generated

As p-value <0.2, time is statistically significant for the main model.

Overall summary after step 1 is as follows:

A screenshot of a computer

Description automatically generated

Initial model with summary is presented below:

A screenshot of a computer

Description automatically generated

1. Step 2: perform backward elimination.

From here, p-value will be checked with a value less than 0.05 (5% level of significance).

1. Dropping time from model:

A computer screen shot of a table

Description automatically generated

As p-value < 0.05, time will be considered for the final model.

1. Dropping Serum\_sodium from model:

A computer screen shot of a table

Description automatically generated

As p-value > 0.05, serum\_sodium will not be considered for the final model.

1. Dropping serum\_creatine from model:

A screenshot of a computer code

Description automatically generated

As p-value < 0.05, serum\_creatine will be considered for the final model.

1. Dropping blood pressure from model:

A screenshot of a computer code

Description automatically generated

As p-value > 0.05, blood pressure will not be considered for the final model.

1. Dropping ejection\_fraction from model:

A screenshot of a computer code

Description automatically generated

As p-value < 0.05, ejection\_fraction will be considered for the final model.

1. Dropping age from model:

A white background with black text

Description automatically generated

As p-value < 0.05, age will be considered for the final model.

Overall, here is the summary after step 2.

A white background with black text

Description automatically generated

Model after step 2 has below summary:

A screenshot of a computer

Description automatically generated

1. Step 3: Now we will add variables which were not chosen initially.
2. Adding Anameia in model:

A black text on a white background

Description automatically generated

As p-value > 0.05, we will not keep anameia in model.

1. Adding creatine\_phosphokinase in model:

A computer screen shot of a math program

Description automatically generated

As p-value > 0.05, we will not keep creatine\_phosphokinase in model.

1. Adding diabetes in model:

A computer code with black text

Description automatically generated

As p-value > 0.05, we will not keep diabetes in model.

1. Adding Platelets in model:

A screenshot of a computer code

Description automatically generated

As p-value > 0.05, we will not keep platelets in model.

1. Adding sex in model:

A screenshot of a computer code

Description automatically generated

As p-value > 0.05, we will not keep sex in model.

1. Adding Smoking in model:

A computer code with black text

Description automatically generated

As p-value > 0.05, we will not keep smoking in model.

Overall, here is the summary and the model will be kept as is after step 3:

A black text on a white background

Description automatically generated

1. Step 4: Now, we will check how interactions affect the model.
2. Checking interaction between age and ejection\_fraction:

A close-up of a computer code

Description automatically generated

As p-value >0.05, we will not include the interaction term it is not significantly better (p-value=0.1898).

1. Checking interaction between serum\_creatine and age:

A computer code with black text

Description automatically generated

As p-value >0.05, we will not include the interaction term it is not significantly better (p-value=0.4).

1. Checking interaction between age & time:

A computer code with black text

Description automatically generated

As p-value >0.05, we will not include the interaction term it is not significantly better (p-value=0.2175.

1. Checking interaction between serum\_creatine and ejection\_fraction:

A screenshot of a computer code

Description automatically generated

As p-value >0.05, we will not include the interaction term it is not significantly better (p-value=0.333).

1. Checking interation between ejection\_fraction & time:

A screenshot of a computer code

Description automatically generated

As p-value < 0.05, we will include the interaction term as it is significantly better (p-value=0.01069).

1. Checking interaction between serum\_creatine and time:

A close-up of a computer code

Description automatically generated

As p-value >0.05, we will not include the interaction term it is not significantly better (p-value=0.5957).

Overall, the summary after step 4 is below:

A screenshot of a computer screen

Description automatically generated

Summary statistics for Model 1 after all the steps is below:

A screenshot of a computer code

Description automatically generated

1. BIC selection for model 2:

A close-up of a number

Description automatically generated

The best model selected based on BIC is the one with age, ejection\_fraction, serum\_creatine and time as explanatory variables and has the least AIC values with p-value < 0.05. The above model is now Model2.

Below is the summary statistics for model 2:

A screenshot of a computer

Description automatically generated

1. Checking ROC curve and AUC values for both model 1 & model 2.

A close-up of a math equation

Description automatically generated

A graph with a blue line

Description automatically generated

Comparing the ROC curves and the AUC values (area under the curve), Model1 has larger AUC compared to Model2. Based on the AUC values we can use Model1 instead of Model2, but there is not much difference between them (90% for Model1 with 5 predictors and 89% for Model2 with 4 predictors). As there is not much difference, we can say that Model2 is better model with lesser predictors and a AUC value of 0.891215.

1. Creating model 3 with all the values and plotting it against model 1 & model 2 ROC curves and checking AUC values.

Model 3 statistics:

A screenshot of a computer

Description automatically generated

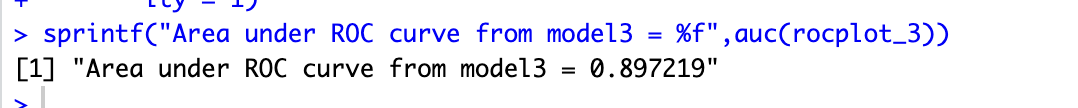
After performing AIC on model 3:

A screenshot of a computer

Description automatically generated

We can see that 5 predictors are selected after performing AIC.

ROC curve of model 1, model 2 & model 3 with AUC value for model 3:



A graph with colored lines and numbers

Description automatically generated with medium confidence

Comparing the ROC curves and the AUC values (area under the curve), As Model 2 is better than Model 1 we will compare Model 2 and Model 3. As both the models has 89% but Model 3 has 5 predictors whereas Model 2 has 4 predictors, we will choose Model 2 as it has lesser predictors and AUC value of 0.891215.