Sta303 Final Project

2024-03-26

Introduction:

The goal of this report is to explore how a comprehensive set of physiological and biochemical indicators (including age, ejection fraction, blood pressure, diabetes status, time, sex, anaemia status, serum creatinine, platelet count, and serum sodium levels) influence the risk of mortality in patients with heart failure? Based on that I aim to make a model which is more descriptive and easier to interpret. Unlike previous research, which often narrows scope of predictor factors and population, my research question encompasses a multifaceted analysis, incorporating age, ejection fraction, serum creatinine levels, and follow-up duration into a comprehensive predictive model. For instance, the study in paper *Incident Heart Failure Prediction in the Elderly* narrow down its scope within Health ABC Heart Failure Score, which was developed specifically within the Health ABC study cohort(Butleret al., 2008). Therefore, my study by potentially including a different or more diverse population, which might offer new insights into heart failure risk prediction across various demographic groups, including those not well represented in the Health ABC study. However, we also have some commons. For example, both of us acknowledge the critical role of common clinical variables such as age, serum creatinine levels, and ejection fraction in predicting heart failure risk, underscoring the importance of these factors across different models and cohorts.

Method(Variable Selection)

Literature review suggestion

At the beginning for initial variable selection, I identified a broad set of physiological and biochemical indicators known to influence heart failure outcomes based on a comprehensive review of existing literature. For example the outcome variable and most of interested predictors or confounders are selected from literature *Incident Heart Failure Prediction in the Elderly*, *High-Sensitivity Cardiac Troponin I Levels and Prediction of Heart Failure* and *Diagnosis of Nonischemic Stage B Heart Failure in Type 2 Diabetes Mellitus: Optimal Parameters for Prediction of Heart Failure.* If there are variables that appear in all those three reports then I will select them as my confounders in my project. Also if certain variables are only mentioned but not discussed in depth from those literature, then I will consider using them as main predictors in my project.

Tuition from EDA Box plot

When using box plots to inform variable selection for a model, one assesses differences in the central tendency and variability between groups for each predictor. If a box plot for a variable shows a notable difference in medians between the two DEATH_EVENT groups, this indicates the variable could be discriminative and thus useful for the model. Conversely, if the interquartile ranges and medians are similar across groups, the variable may offer less predictive value.

Stepwise variable selection

After determining model type, I filter variables so that it best fit the model by using stepwise variable selection. This method begins with an empty model and iteratively adds one variable at a time (forward

selection) while simultaneously evaluating the impact of removing variables from the model (backward elimination). At each step, the selection criteria, such as AIC and BIC, are used to determine whether adding or removing a variable improves the model fit or predictive accuracy.

Likelihood ratio test

Since from the method above,it uses 2 different selection criteria. If each criteria suggests different variable selections, then it needs to apply a likelihood ratio test. We can simply interpret the likelihood ratio test as a hypothesis test with H_0: simpler model fits better and H_a: complex model fits better. Then calculate the test statistic and P value. If P value is greater than significance level value, the null hypothesis is failed to be rejected, indicating that the simpler model provides a significantly better fit to the data and vice versa.

Method(Model Diagnostics and validation)

Dfbeta for model Diagnostics

After getting the final model, my next step is to check diagnostics and see whether there are any influential observations by using dfbetas plots. Dfbetas can be calculated using elimination of one observations. This is executing the model with and without the observation and observing the changes in a regression coefficient. If the plot shows the large absolute dfbeta values then it suggests the corresponding observation has a substantial influence on the specific regression coefficient.

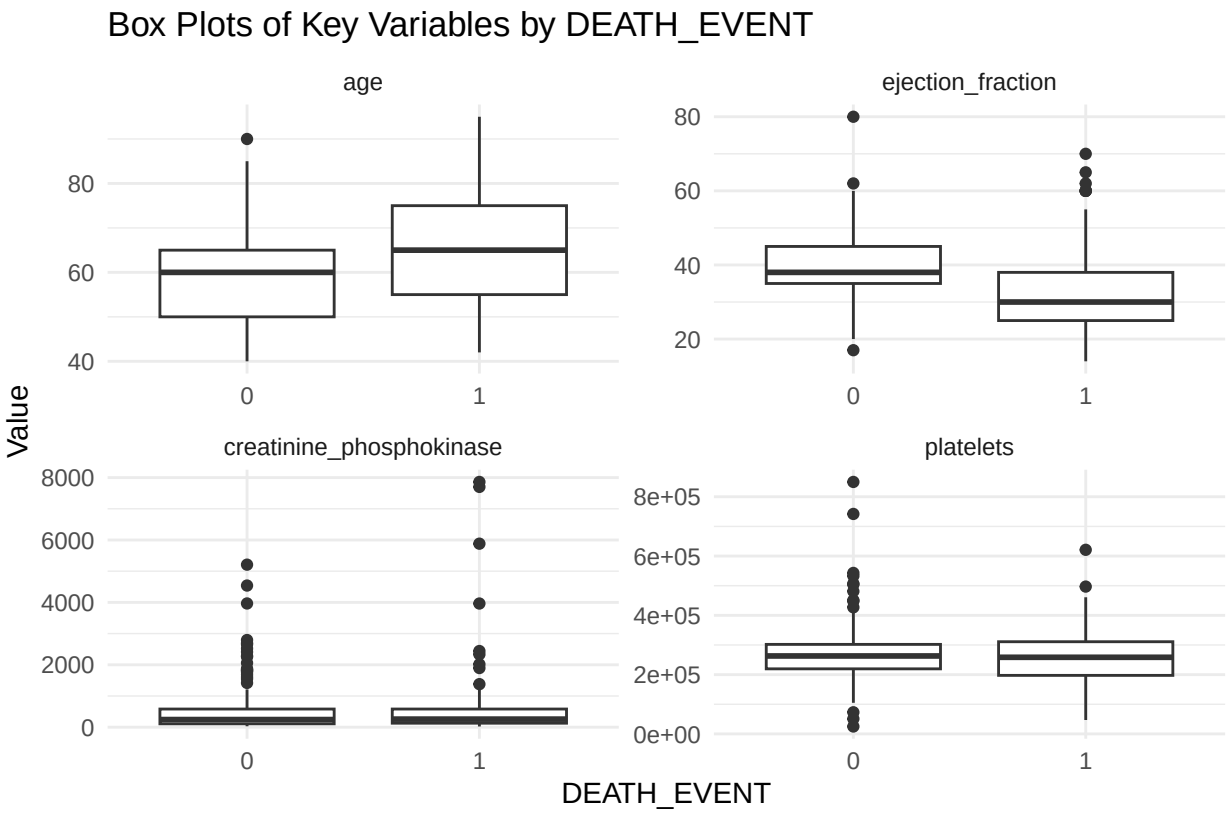
Cross validation for model validation

After checking Diagnostics, I use cross validation to avoid the case that model is overfitted. To apply CV, randomly split the data into k parts. For example when k = 5, the data is splitted into five parts. Fit the model with k - 1 parts (training) and predict the outcomes for the remaining part (test). Use all the k parts as test set. The prediction accuracy can be checked with mean absolute bias or mean squared error. The predictions can be plotted with the observed values to check the accuracy of the estimates visually. In calibration plot, there will be 3 lines apparent, bias-corrected and ideal. If the apparent and bias-corrected lines follow the pattern of ideal line then it shows the model pass the validation test.

Roc curve for model performance

After cross validation test, the final step is to draw Roc curve to explore model performance. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds, showing the trade-off between correctly identifying positive cases and falsely identifying negative cases as positive. The area under this curve (AUC) summarizes the model's ability to distinguish between the outcomes, with higher values indicating better performance.

Results Section (Descriptionof the Data)



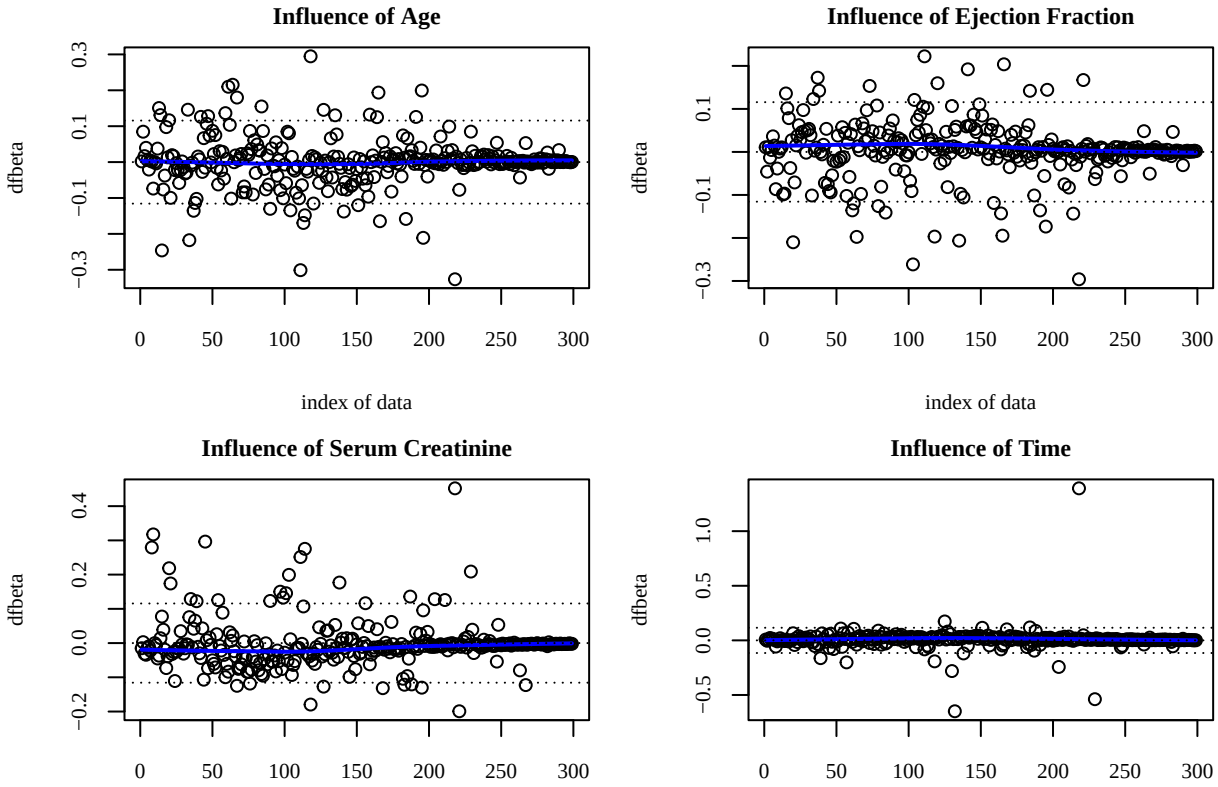
Caption: Comparative box plots of selected variables by DEATH_EVENT status.

The box plots suggest that 'age' and 'ejection_fraction' demonstrate noticeable differences in their distributions between the two DEATH_EVENT groups. These differences are discernible in their medians and interquartile ranges, hinting at their potential usefulness in predicting the outcome. Conversely, 'creatinine_phosphokinase' and 'platelets' display substantial overlap between the groups, alongside a significant presence of outliers, which could undermine their predictive power and complicate the modeling process. Therefore, 'age' and 'ejection_fraction' could be more reliable candidates for inclusion in a predictive model, while 'creatinine_phosphokinase' and 'platelets' may require further preprocessing or may be less suitable as predictors due to their variability and outlier influence.

Results Section (Presenting the Analysis Process and the Results)

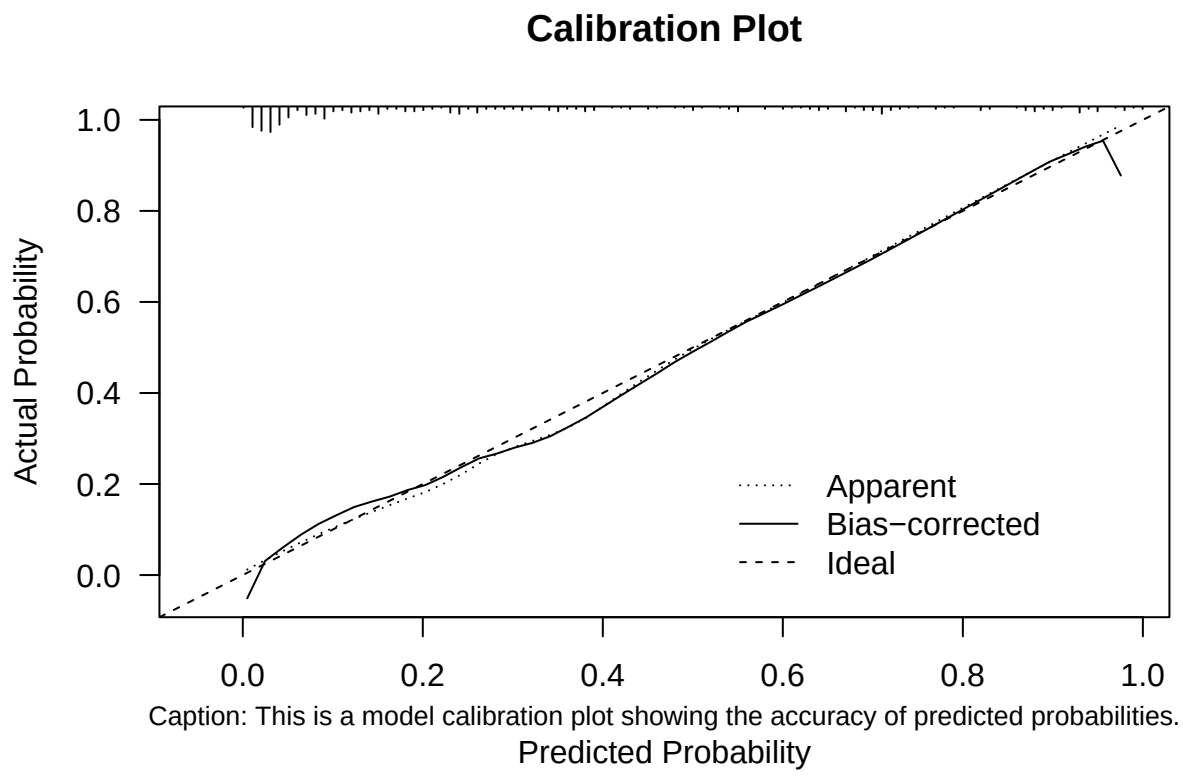
With the help from correlated literature and EDA, the initial model is a GLM model with outcome varialble as DEATH_EVENT. The predictors are age, ejection_fraction, serum_creatinine and serum_sodium. The confounders are high_blood_pressure, diabetes, sex, smoking and anaemia.

After Stepwise variable selection, different criteria provides different results. In AIC criteria the remaining variables are age, ejection_fraction, time, serum_creatinine and serum_sodium. In BIC criteria the remaining variables are age, ejection_fraction, time, serum_creatinine. To determine which model fits better, I applied the Likelihood ratio test and get P value is 0.09338078. I use 0.05 as significance level thus $0.09338078 > 0.05$ and it failed to reject H_0: simpler model fits better. Therefore, I choose BIC criteria as final model with 4 predictors age, ejection_fraction, time, and serum_creatinine.



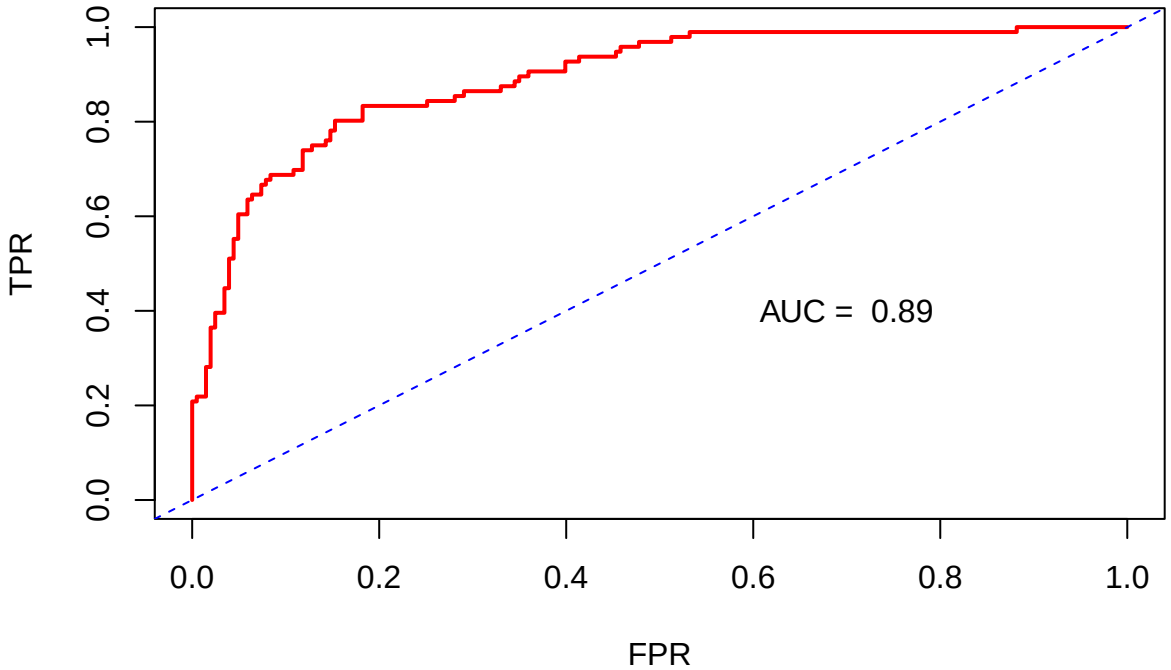
Caption: Influence of Individual Observations on Model Coefficients

Next step is to check model diagnostics. The provided image above displays four scatter plots representing the influence of individual observations on the logistic regression coefficients for different predictor variables: age, ejection_fraction, serum_creatinine, and time. Each plot shows the dfbetas values for a respective variable plotted against the index of observations. Across the plots, most dfbetas values hover around zero, with no extreme values surpassing the dotted lines at $\pm\frac{2}{\sqrt{n}}$, suggesting that there are no significantly influential observations for these predictors. The presence of LOWESS smoothing lines also suggests no strong patterns of influence across the dataset.



Caption: This is a model calibration plot showing the accuracy of predicted probabilities.

Next step is to check model validation. I use method cross validation to do that. Above is the calibration plot. It indicates that the model is reasonably well-calibrated, as shown by the closeness of the 'Bias-corrected' line to the 'Ideal'. The calibration plot also includes a measure of the mean absolute error (MAE = 0.018), which is small, and the number of observations (n=299), indicating a good fit between the model's predictions and the observed outcomes. The calibration plot, along with the validation metric, suggests that the model's estimates are reliable and it generalizes well to new data.



Caption: ROC curve plot.

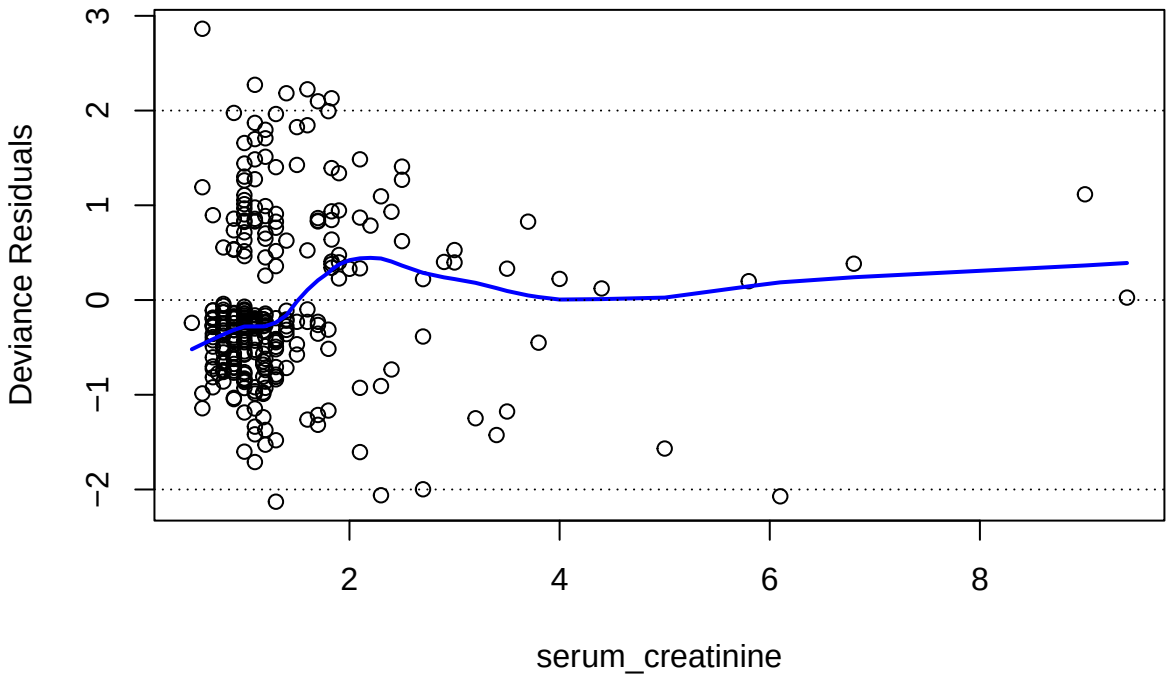
Finally, it is for performance check. The image above shows a Receiver Operating Characteristic (ROC) curve with an Area Under the Curve (AUC) of 0.89. This suggests that the model has a high ability to distinguish between the two classes being predicted. The curve stays well above the diagonal line, which represents a random chance, affirming the model's good predictive power. In summary, the ROC curve demonstrates that the model performs significantly better than random guessing in classifying the outcomes correctly.

Discussion (Final model interpretation and importance)

Based on the final model presented, which includes age, ejection_fraction, serum_creatinine, and time as predictors with a binary outcome DEATH_EVENT, we can interpret one of the coefficients to understand its impact. For instance, the ejection_fraction coefficient is -0.074804, indicating that as the ejection fraction increases, the log odds of a death event decrease, holding all other variables constant. This suggests that patients with higher ejection fraction, which typically indicates better heart function, have a lower risk of death.

In a broader sense, the model illustrates how various clinical measurements and patient characteristics relate to the likelihood of death events in heart disease patients. By incorporating significant predictors like age, heart function (ejection fraction), kidney function (serum creatinine levels), and the duration of follow-up (time), the model provides valuable insights into the factors that may increase or decrease the risk of mortality. This aligns with the research goal of identifying key factors that influence patient outcomes in heart disease, allowing for more targeted interventions and management strategies for at-risk individuals.

Discussion(Limitations of the Analysis)



Caption: Scatter plot of Deviance Residuals.

As for limitation, the scatter plot of deviance residuals against serum_creatinine appears to show some non-linearity, as indicated by the blue LOWESS curve. The residuals do not seem to be randomly distributed around the zero line but show a pattern, suggesting that the relationship between serum_creatinine and the response variable might not be perfectly captured by the model. This could potentially impact the model's effectiveness in predicting outcomes for new data points, especially if the underlying true relationship is non-linear. The presence of several points with high residuals could also indicate outliers or influential points that the model is not accounting for properly.

These issues do not have been corrected. Since the dataset is small, there is no enough data to properly model a non-linear relationship. Alternatively, the current scope of the model does not include polynomial or interaction terms that could help explain the curvature seen in the plot. If the final model is to be used for predictions, these lingering issues could reduce its predictive accuracy, and it's crucial to be aware of these limitations when interpreting its results or applying it to real-world data.

Reference

[1]: Butler, J., Kalogeropoulos, A., Georgiopoulou, V., Belue, R., Rodondi, N., Garcia, M., Bauer, D. C., Satterfield, S., Smith, A. L., Vaccarino, V., Newman, A. B., Harris, T. B., Wilson, P. W. F., & Kritchevsky, S. B. (2008). Incident heart failure prediction in the elderly. Circulation: Heart Failure, 1(2), 125–133. https://doi.org/10.1161/circheartfailure.108.768457

[2]: Takashio, S., Takahama, H., Hayashi, T., & Anzai, T. (2016). Serial measurements of high sensitivity cardiac troponin T levels in acute decompensated heart failure. Journal of Cardiac Failure, 22(9). https://doi.org/10.1016/j.cardfail.2016.07.128

[3]: Wang, Y., Yang, H., Huynh, Q., Nolan, M., Negishi, K., & Marwick, T. H. (2018). Diagnosis of nonischemic stage B heart failure in type 2 diabetes mellitus. JACC: Cardiovascular Imaging, 11(10), 1390–1400. https://doi.org/10.1016/j.jcmg.2018.03.015