Heart Failure Records

IE 500: Statistical Machine Learning

Done by:
Samir Nino, Pranav Pillai, and Sumeet Sahu

Table of Contents:

Introduction and Data Description

Heart failure is a cardiovascular condition that occurs when the heart is unable to pump enough blood to meet the requirement of the body. This can occur if the heart does not fill up with enough blood or if the heart is not strong enough to pump blood effectively. According to the Centre of Disease Control and Prevention, it is said that approximately 6 million people in the United States have suffered from heart failure. It is one of the primary causes of death, taking approximately 17.9 million lives a year. Heart failure can be acute or can progress over time. It can affect either one or both sides of the heart. Left sided and right sided heart failure might have different causes behind it. It is common for another condition to trigger heart failure for example coronary heart disease, heart inflammation and high blood pressure. Heart failure does not necessarily cause people to immediately exhibit symptoms. However, one can experience fatigue, shortness of breath and buildup of fluid around the waist or neck. Heart failure can adversely affect the liver and/or kidneys. It can also lead to other complications such as hypertension or arrhythmia. As of now heart failure has no cure but if alterations (such as regular exercise, reducing smoking) are made to lifestyle, this can lead to a better lifestyle being enjoyed.

The dataset that was obtained from kaggle.com contains 13 variable columns with exactly 299 entries collected from patients from April to December 2015 from a Pakistan-based hospital. All the patients have “left ventricle systolic dysfunction” and past heart failure. The dataset consists of different factors associated with heart failure which are : age, anaemia, creatinine phophokinase , diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time and death event. Ejection fraction provides a measure of the proportion of blood being pumped out of the left ventricle with every contraction. The serum creatinine column is a measure of serum creatinine which is a waste product of creatine which is generated due to muscle breakdown. The sex column denotes the gender of the patient with 1 indicating male and 0 indicating female. The smoking column denotes whether the patient is a smoker with 1 denoting smoker and 0 denoting non-smoker. The time column is a measure of the months of the follow up with the patient. The death event column is our independent variable, and it shows if the patient is dead or alive with 1 indicating dead and 0 as alive.

Methodology:

We will be using R to show an algorithmic approach to analyze and predict heart failure resulting in the death of the victim. We will analyze the correlation and covariance between the factors and the end result through statistical means. Logistic regression will be used to predict the “Death Event,” our independent variable using multiple variables presented in our dataset. Our logistic regression model will initially account for variables of age, anaemia, creatinine phophokinase , diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, and time to show how they affect the death event individually and as a multivariate system to determine the significance of the variables. We will split the data into training and testing data with a test ratio of 0.8 for the logistic regression model generated. The model that we eventually use will be an equation that will effectively predict whether or not the Death Event occurs to a person with their respective factors, and analyze which factors affect the Death Event the most.

Our equation will have the following form:

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

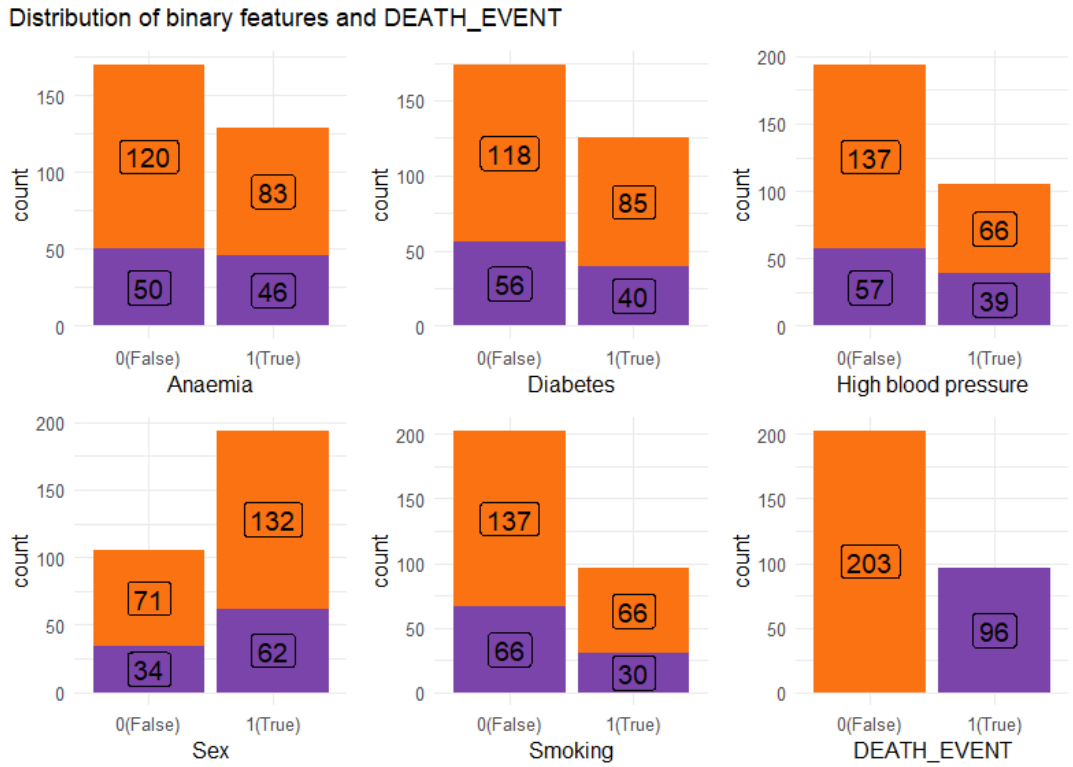
L is the predicted value of the logistic regression model

In our logistic regression model The correlation typically has a range between -1 and 1 so the correlation square has a range between 0 and 1. The higher the correlation between actual and predicted values, the higher the R square irrespective of the correlation being negative or positive. As an alternative to the tradition R-Squared and Adjusted R-Squared, we are using McFadden's. For this McFadden's R-Squared measure is used and it is expressed with the formula:

$$R^2_{\text{McFadden}} = 1 - \frac{log(L_c)}{log(L_{\text{null}})}$$

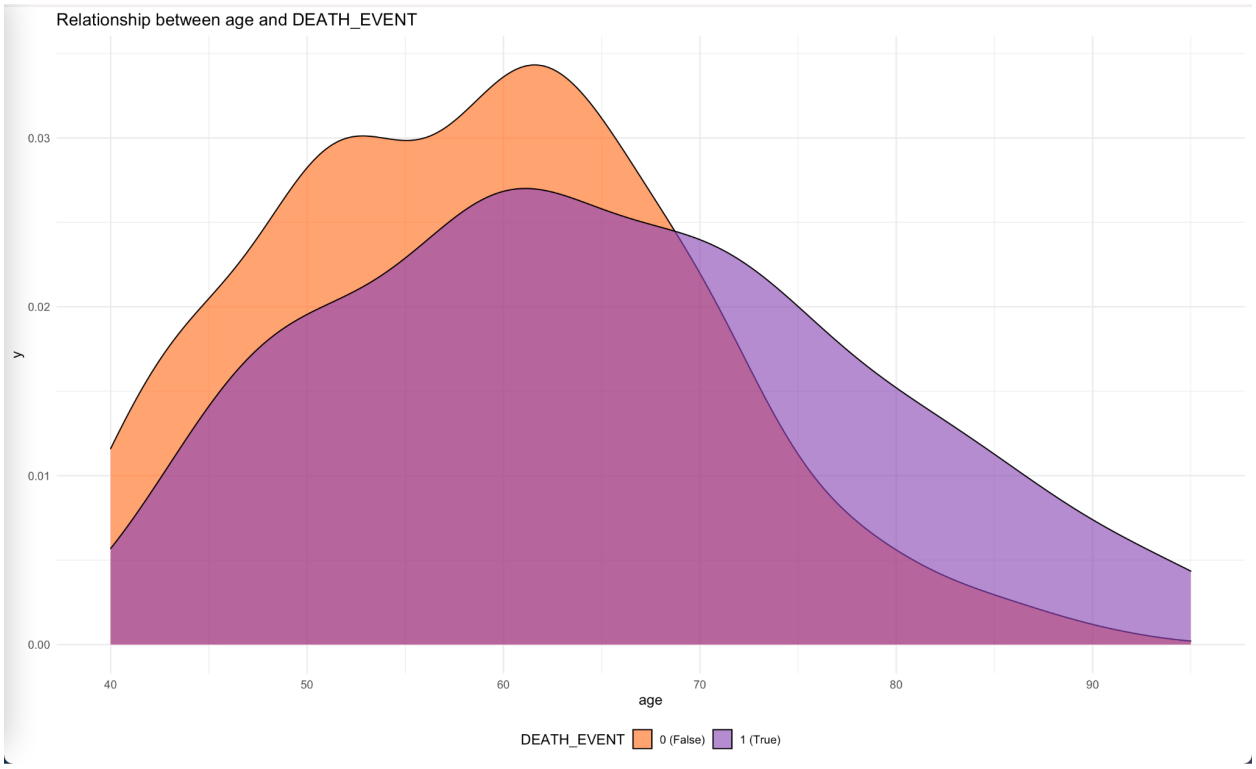
Where L_c is the maximized likelihood from the model currently fitted, (or the full model with predictors) L_{null} is the null model's corresponding value (or intercept model without predictors). This means that the model only contains an intercept and has no covariates. A likelihood is between 0 and 1 so the log likelihood is less than or equal to zero. If a model contains a low likelihood, then the log likelihood has a higher magnitude than the log of a more likely model. A low ratio of the log likelihoods denotes that the full model is a much better fit than the intercept. If two models were to be compared with the same data, the McFadden's pseudo R squared will have a higher for the model with a higher likelihood

Data Visualization



Distribution of binary features

The above distributions show the relationship between the individual binary factors and the death event. Purple shows the number of people who died with that condition and orange shows the number of people who did not die. The biggest factor in the death event is anaemia, with above a 50% chance of death if anaemia is present. Surprisingly, smoking and diabetes are the least significant factors with less than 50%.



Age vs Death Event

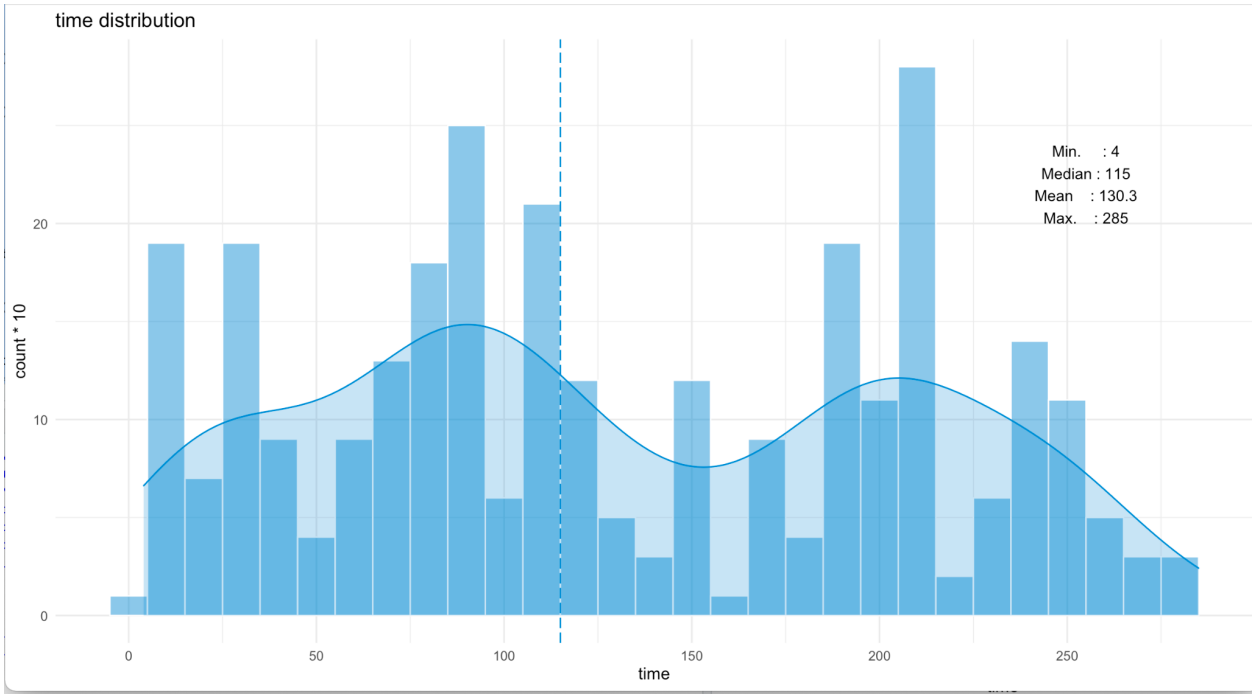
The age distribution compared to death count shows that most death events happen around the age of 60. Our data consists of ages from 40 to 100. The data is skewed to the left and the age is inversely proportional to the death event in our case. The older the person gets, the more likely that the death event will occur.

Results :

	DEATH_EVENT		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	18.023	4.864 – 32.738	**0.0101**
age	0.051	0.011 – 0.093	**0.0150**
anaemia [Anaemia]	-0.084	-0.966 – 0.775	0.8483
creatinine phosphokinase	0.001	0.000 – 0.001	0.0767
diabetes [Diabetes]	0.073	-0.797 – 0.941	0.8679
ejection fraction	-0.100	-0.147 – -0.059	**<0.001**
high blood pressure [BP]	-0.342	-1.263 – 0.543	0.4546
platelets	-0.000	-0.000 – 0.000	0.2509
serum creatinine	0.483	-0.031 – 1.087	0.0906
serum sodium	-0.110	-0.211 – -0.018	**0.0227**
sex [Male]	-1.042	-2.108 – -0.039	**0.0468**
smoking [Smoke]	0.020	-0.998 – 1.037	0.9693
time	-0.025	-0.033 – -0.017	**<0.001**
Observations	224		

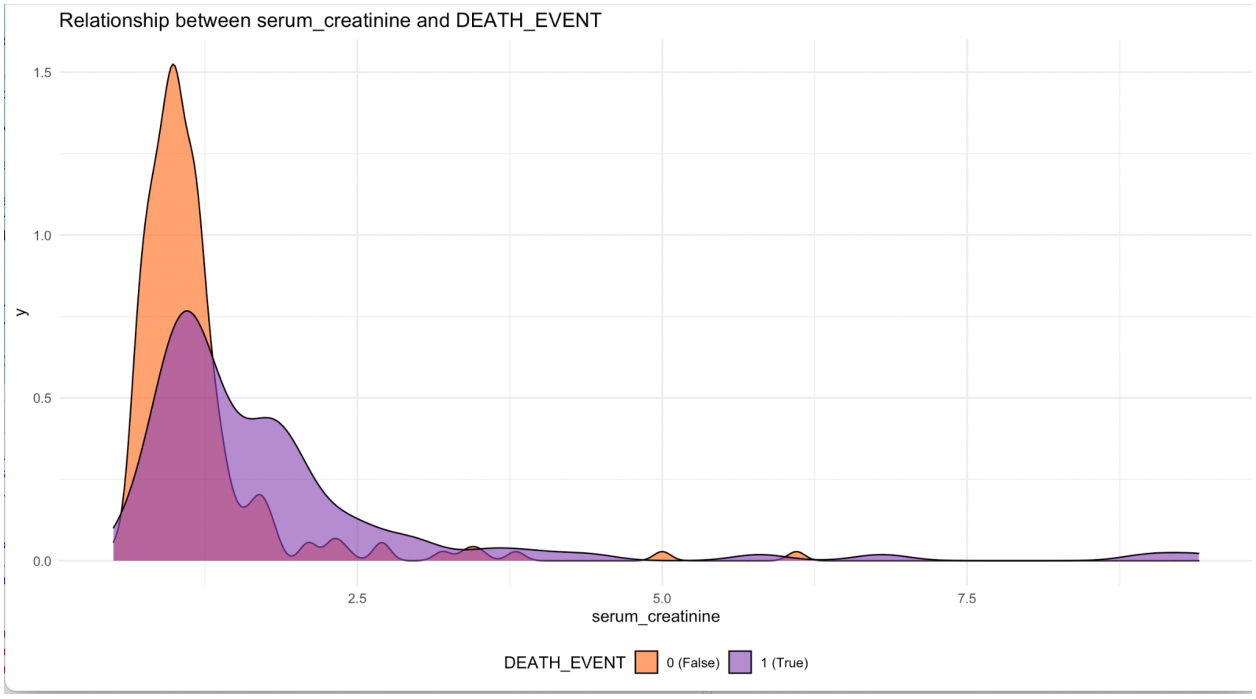
Logistic regression

The logistic regression above shows that the time and ejection fraction are the most significant factors with p-values less than 0.001. They are also inversely proportional to the death event as the coefficient estimates for time and ejection fraction are -0.025 and -0.100 respectively. Increase in these two factors decreases the probability of death event. The most significant factor that is directly proportional to the death event is serum creatinine which has a coefficient of 0.483. Increase in serum creatinine increases the probability of the death event.



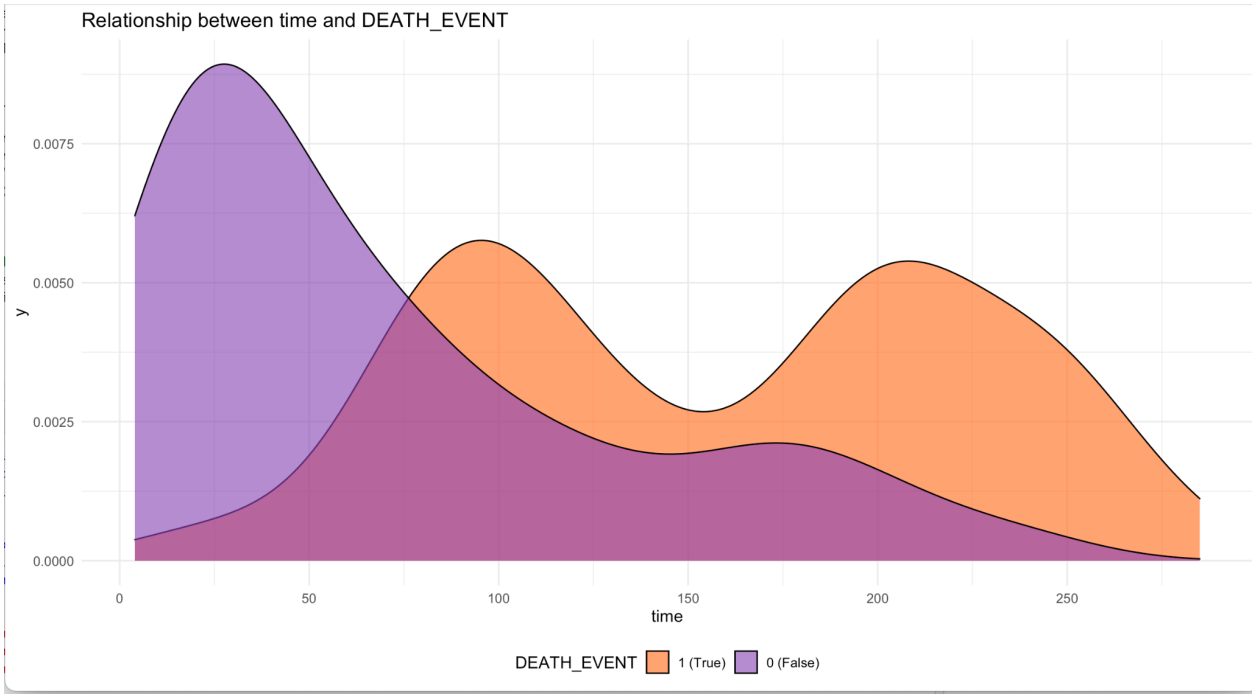
Time distribution vs Time.

The above distribution shows the time distribution vs time. Interestingly, even though the time distribution is highly correlated with the death event, its distribution appears to be random.



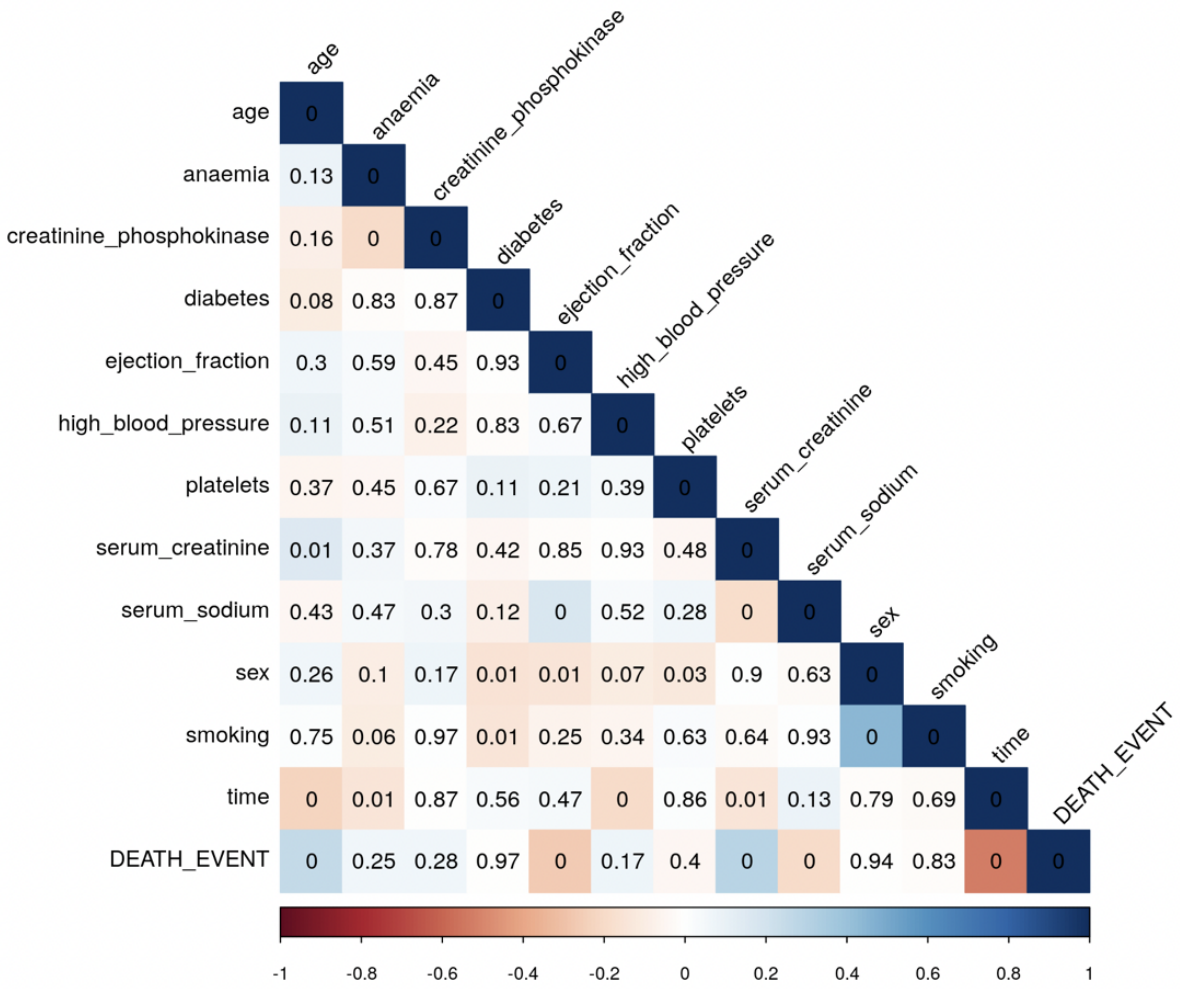
Serum Creatinine vs Death Event

The plot above shows the distribution of serum creatinine with the death event. As serum creatinine increases, the death event becomes more prevalent. This shows how important it is to keep the serum creatinine factor down.



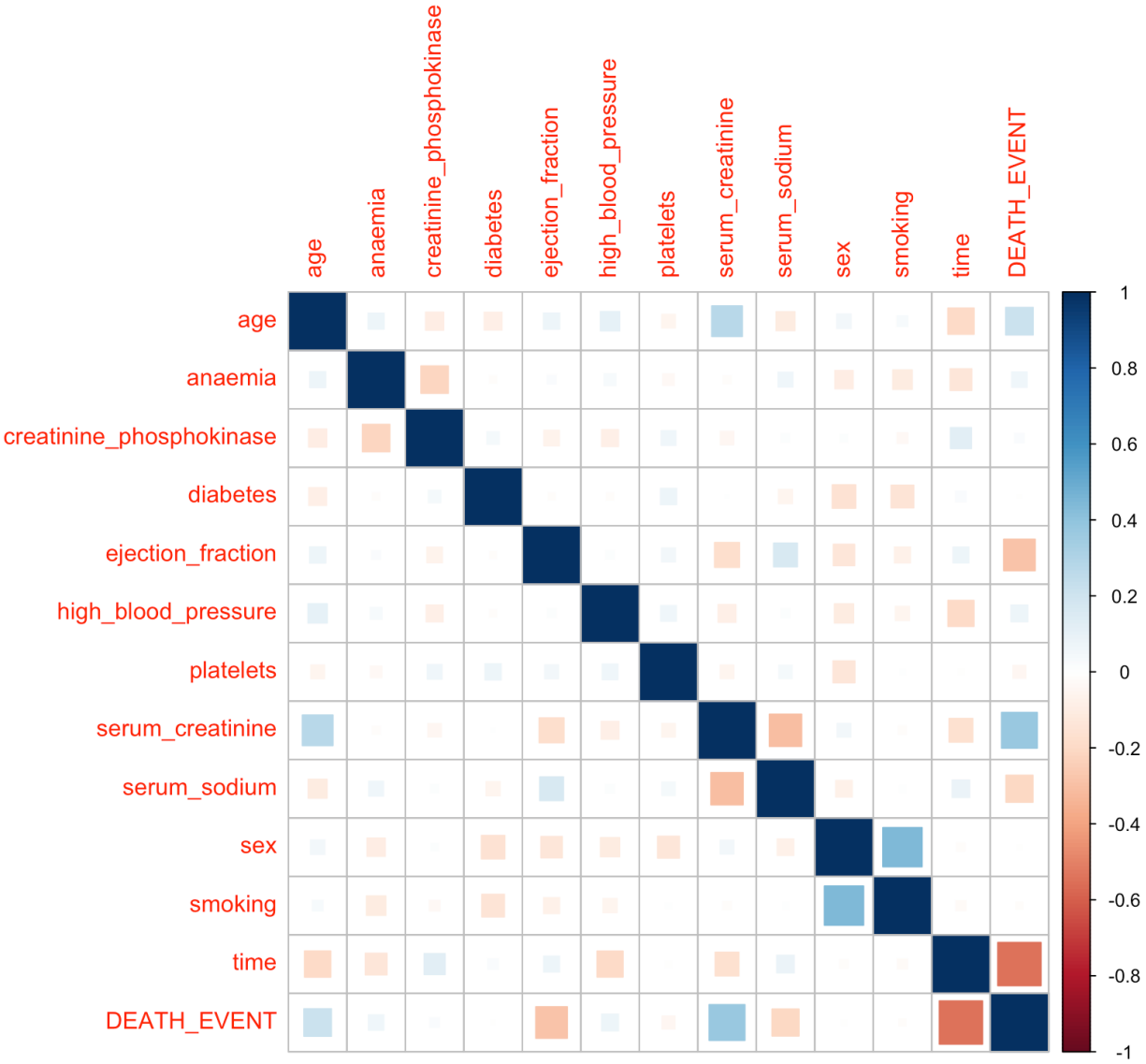
Time vs Death Event

The time spent before check up in months is inversely proportional to the death event as shown in the above plot. The more time patients spend after the initial check up to get an update, the more likely it is that the death event will become true.



Covariance

The covariance of the factors and the death event is shown above. Some factors are very closely connected with values of covariance less than 5%. Meanwhile there are some factors that are not connected to each other at all with ranging values of covariance. This shows that some of the factors do not appear at the same time in the same person and others come together. This analysis could be useful when trying to determine the best treatment for a patient and diagnosing the symptoms.



Correlation

The above plot shows the correlation between all the different factors. Time and ejection fraction are the most significant and are inversely related to the death event. Age and serum creatinine are less significant but the most important out of all the directly proportional factors. This correlation can show how much each of the factors affect the death event and how they all affect each other.

Conclusion

Out of all the factors that affect death event, time, the measure of the months of follow up with the patient, and ejection fraction, the measure of the proportion of the blood being pumped out of the left ventricle with every contraction, are the most important factors that are inversely proportional to the death event. Meaning that maintaining these two factors is the most important task of the patient. The most significant directly proportional factors are the age of the patient and serum creatinine, the measure of the waste product creatine created due to the breakdown of the muscle. Some causes for higher creatinine are dehydration and intense workout. In our dataset, the factor affect serum creatinine the most is sex (Male).

Future Directions

The limitation of the dataset was that it was small (had only 299 entries); a larger dataset would have enabled us to get a more reliable result. Having additional information such as the physical details (weight, height etc) and the occupational and lifestyle history of the patients would have been useful in gauging the supplementary risk factors of heart failure. Moreover, if a similar dataset was available for another country or region, this would have been utilized as a means of comparison to validate the findings in our project.

References

- Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20,** 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

https://www.nhlbi.nih.gov/health-topics/heart-failure

- U.S. Department of Health and Human Services. (n.d.). *Heart failure*. National Heart Lung and Blood Institute. Retrieved November 1, 2021, from https://www.nhlbi.nih.gov/health-topics/heart-failure.

https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure

- *What is heart failure?* www.heart.org. (n.d.). Retrieved November 1, 2021, from https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure.
- Bartlett, J. (2020, February 28). *R squared in logistic regression*. The Stats Geek. Retrieved December 4, 2021, from https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/.
- "FAQ: What Are Pseudo R-Squareds?" *IDRE Statsistical Counseling*, https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/.