COSC 757 Assignment 1: Analysis of Heart Failure Clinical Records

Hannah Strakna
Towson University
hstrak2@students.towson.edu

I. INTRODUCTION

Heart failure is a condition that affects people worldwide every day. In the worst scenarios, it can cause death. Therefore, it is important for hospitals to collect as much data about their patients as possible. With this data, scientists can learn how to better treat and prevent heart failure. For this data analysis, we will be examining a set of clinical records containing information about patients admitted for heart failure. We hope to discover trends within the data that can help doctors predict who is most at-risk of death.

II. DESCRIPTION OF THE DATASET

For this study, we will be using the "Heart Failure Clinical Records" data set [1] that is available on the UCI Machine Learning Repository [2]. This data set contains 299 rows of patient information and 13 different attributes as follows:

- age: age of the patient in years (from 45-95)
- anaemia: if the patient has anaemia (1=anaemia, 0=no anaemia)
- high_blood_pressure: if the patient has hypertension (1=high blood pressure, 0=no high blood pressure)
- creatinine_phosphokinase: level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (1=diabetes, 0=no diabetes)
- ejection_fraction: percentage of blood leaving the heart at each contraction
- platelets: platelets in the blood (kiloplatelets/mL)
- sex: if the patient is a woman or a man (1=man, 0=woman)
- serum_creatinine: level of serum creatinine in the blood (mg/dL)
- serum_sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (1=smokes, 0=does not smoke)
- time: follow-up period in days
- DEATH_EVENT: if the patient dies during the follow-up period (1=died, 0=lived)

III. EXPLORATORY DATA ANALYSIS

We performed an exploratory data analysis in order to find which attributes were most likely to lead to the death of a patient. For the sake of this analysis, the attribute DEATH_EVENT was used as the y, or target variable, and each of the other variables were explored with relation to y.

Because DEATH_EVENT is a categorical variable, and the other attributes are a combination of categorical and numerical variables, there were a variety of different methods we used in order to explore the data. The goal was to determine which variables were statistically significant predictors as to whether or not a patient died.

For the categorical variables (anaemia, high_blood_pressure, diabetes, sex, and smoking) we created a contingency table for each of these x variables in relation to our y variable, DEATH_EVENT. We then created a stacked bar chart to visualize the data, separating those who died from those who did not. Finally, we split the dataset into those who died vs. lived and performed a chi-square test on the two datasets to determine the p-value for each x and y pairing. We did not find any attribute among the categorical variables that was statistically significant that could be used as a predictor for DEATH_EVENT.

For the numeric variables (age, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, and serum_sodium) we created similar contingency tables and stacked bar charts to visualize how many people lived vs. died according to each variable. Although these were continuous numeric variables, their ranges were small enough that they could be easily visualized using stacked bar charts. We did not use the variable 'time' because it was not a measure of the patient's health.

Then, we performed ANOVA, a two-tail t-test and a one-tail t-test for each variable x against our y, DEATH_EVENT. From the t-tests, we determined that age,

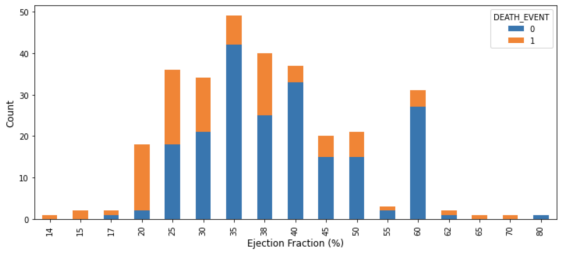
ejection_fraction, serum_creatinine, and serum_sodium were statistically significant as they yielded p-values that were less than 0.05 for both types of t-test and ANOVA, meaning that we are able to reject the null hypothesis for each of these variables and say that there is a statistically significant difference in μ of x among those who died.

***

For the rest of this exploration, we will go into more detail about the analysis of the attributes ejection_fraction (EF) and serum_creatinine (SC). According to research by Chicco, D., and Jurman, G [3], EF and SC are considered to be the two most significant factors in predicting patient death.

The "ejection fraction" in this case is the percentage of blood the heart is able to pump at each contraction. A lower ejection fraction means that the heart is pumping less blood. Additionally, during my analysis, I found that ejection_fraction had the lowest p-value, meaning that it gave me the strongest evidence to reject the null hypothesis.

Below are the results I got during my exploration of ejection_faction vs. DEATH_EVENT. The chart shows EF rates among those who lived (blue) vs. those who died (orange).



I separated the data into those who lived vs. those who died and got the following mean and standard deviation for the two data sets:

μ EF - died: 33.46875
μ EF - lived: 40.26600985221675

σ EF - died: 12.52530333701386
σ EF - lived:10.859962681586294

Next, we used a one-tail t-test to test the following hypothesis:

- H_0: μ EF died = μ EF lived
- H_A: μ EF died < μ EF lived

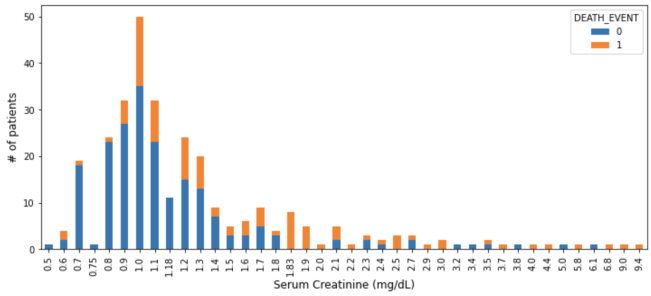
The one-tail t-test yields the following results:

statistic=-5.3171784631624845,
p-value=6.974120244970528e-07

Because the p-value is less than 0.05, we can reject the null hypothesis and say that EF is statistically significant. We can say μ EF among those who died is less than μ EF of those who lived. This means we can use EF as a predictor of patient death.

Then, we turned our analysis to something called"serum creatinine" which is a measure (in MG/dL) that shows us how much of the waste product, creatinine, is found in a patient's blood. This means that the higher the SC measure, the more waste there is in the blood.

These are the results for serum_creatinine vs. DEATH_EVENT. The chart shows SC levels among those who lived (blue) vs. those who died (orange).



I separated the data into those who lived vs. those who died and got the following mean and standard deviation for the two data sets:

μ SC - died: 1.8358333333333332
μ SC - lived: 1.184876847290641

σ SC - died: 1.4685615351275343
σ SC - lived: 0.6540826541207885

Next, we used a one-tail t-test to test the following hypothesis:

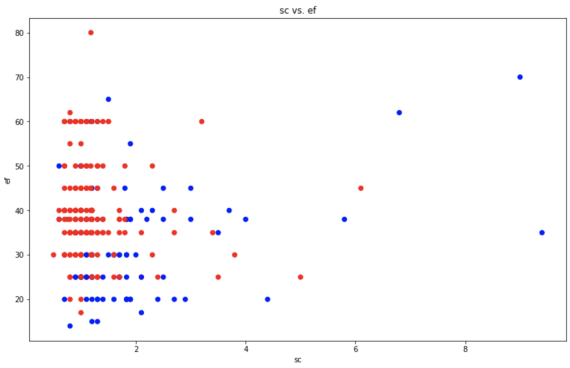
- H_0: μ SC died = μ SC lived
- H_A: μ SC died > μ SC lived

Our two-tail t-test yields the following results:

statistic=4.152639017521322,
p-value=6.398962339971595e-05

Because the p-value is less than 0.05, we can reject the null hypothesis and say that SC is statistically significant. We can say μ SC among those who died is greater than μ SC of those who lived. This means we can use SC as a predictor of patient death.

We were also curious to see if there was a relationship between SC and EF. We created the scatterplot below to examine the relationship between the variables.



Blue meant the patient died and red meant they lived. There did not seem to be much of a relationship between the two.

We predict that SC and EF can be used to predict the likelihood of death in heart failure patients.

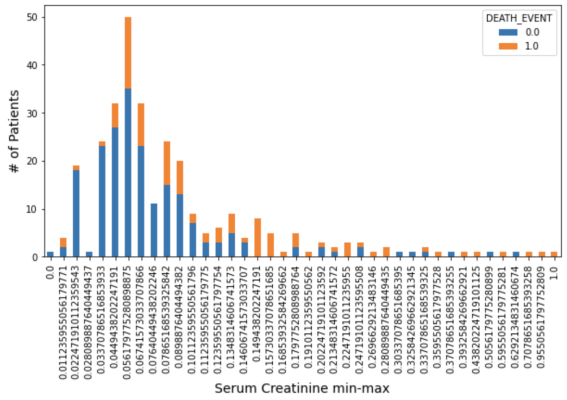
IV. DATA PREPROCESSING

To process our data for testing, we separated out the variables we intended to use for our analysis. This left us with the following variables: serum_creatinine, ejection_fraction, and DEATH_EVENT. Because our categorical variable was already in binary form, we did not need to transform it.

We found that while the data for ejection_fraction only had one slight outlier, the data for serum_creatinine had quite a few outliers very far away from our mean. We did not, however, want to directly remove outliers because many of the data points that laid above 3 standard deviations were ones where the patient died.

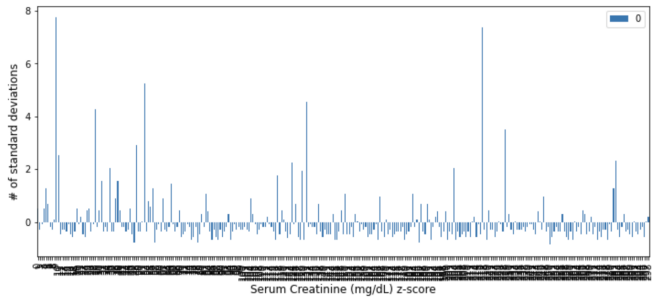
Normalization - The first step was to normalize the data. We used SC to start. The mean amount of SC was 1.39 and the standard deviation was 1.035. We had 6 outliers which included anything above 4.495, 6 points total. The range was 0.5-9.4, meaning that the highest point was quite far from the mean.

We first attempted min-max normalization, but because min-max normalization does not handle outliers well, it would not be the best choice. The mean of SC was 0.10 and the standard deviation was 0.12 when converted to a 0-1 scale.. Below is the chart for min-max SC on a scale of 0 to 1.



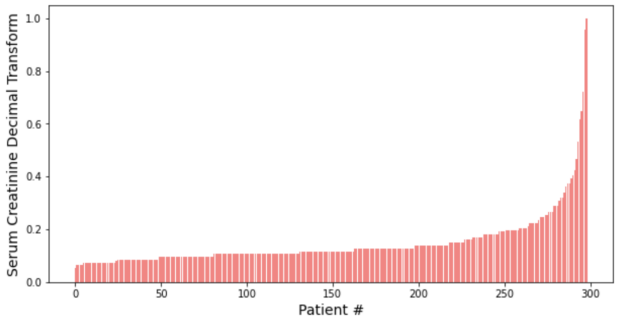
The chart ended up being almost identical to the original SC chart.

Next, we tried z-score standardization because it handles outliers better. The next plot shows the z-score distribution of the data. Each bar represents a datapoint and the numbers on the y axis tell us how many standard deviations away a point is from the mean. The mean is represented by 0 and the bars below it are data points that are less than the mean, while the bars above it are data points that are greater than the mean.



The z-score standardization helps us better visualize how the data is distributed. We can see that there is a clear positive skew.

Last, we tried to use decimal scaling to normalize the data. Our largest SC value was 9.4, so we divided each SC by 9.4. Our mean became 0.15 and standard deviation became 0.11 on a scale of 0 to 1. The sorted decimal transformation below gave us an interesting look at the data.



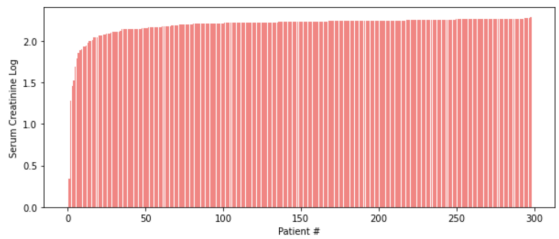
In the end, we decided to keep the outliers for SC in the dataset because these points are important to our data analysis considering the majority of patients outside of 3 standard deviations are all those who died, and high SC was a large contributor to patient death. According to doctors [4], high serum creatinine levels are alarming but not unheard of. In fact, they are a sign of medical issues such as diabetes or kidney failure. Studies have even found that those who have chronic kidney disease are more at risk for heart failure [5], therefore it is important that we not discard these outliers.

In the end, we kept the data as-is because none of the attempts at normalization had a significant enough of an effect.

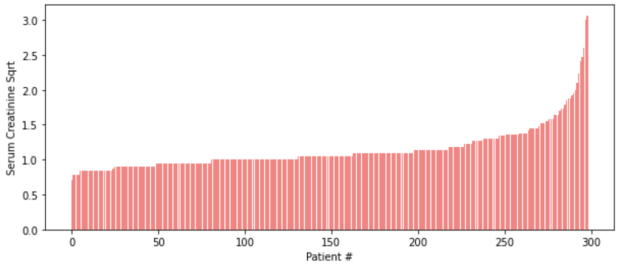
Transformation - After normalization, we took a look at improving skewness. SC was found to have a large positive skew while EF had a moderate negative skew as shown in the chart below.

	skew	kurtosis
serum_creatinine	4.455996	25.828239
ejection_fraction	0.555383	0.041409

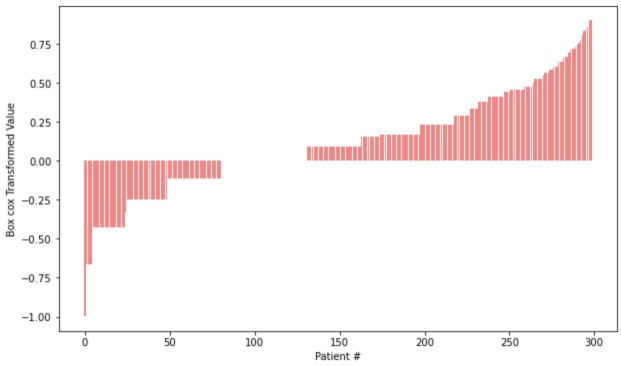
We first tried log transformation but it did not work well. It flipped the skew around and actually made it worse. The chart for log transformation is pictured below.



Next, we tried square root transformation. This worked a little better but the data was still skewed. Below is a chart of the transformation.



Finally, we tried another kind of transformation, box cox transformation, and it did a great job at transforming the data. Below is the box cox chart which helps us visualize the data a bit differently.



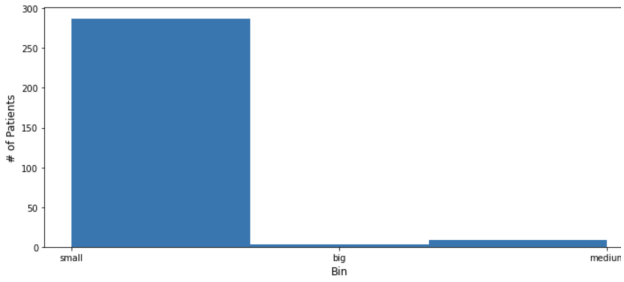
While the data using box cox are clearly still skewed, it is easy to see how there is an improvement.

The new skew measures can be seen in the chart below.

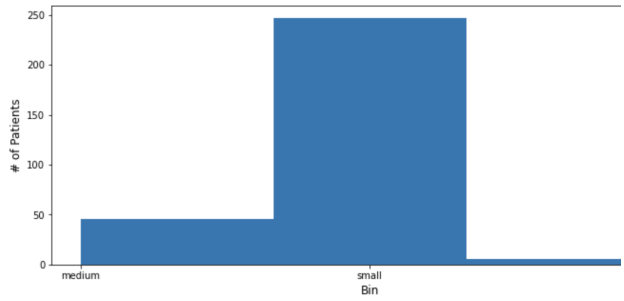
	skew	kurtosis
serum_creatinine	4.455996	25.828239
ejection_fraction	0.555383	0.041409
DEATH_EVENT	0.770349	-1.416080
sc_log	-7.767018	73.035301
sc_sqrt	2.821052	11.007459
sc_boxcox	0.021353	0.220403

Binning - Finally, we attempted to bin our SC values to see if we could better process the data. First, we tried equal width binning. Our three bins were called ‘small,’ ‘medium,’ and ‘big.’ Because our data is so skewed, this type of binning did not work well. We ended up with 287 data points in the ‘small’ bin, 9 data points in the ‘medium’

bin, and 3 data points in the ‘big’ bin. Below is a bar chart of the bins.

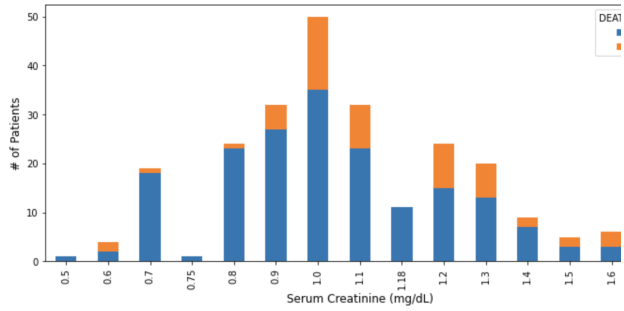


The next type of binning we tried was a bit more complex. It is called Jenks natural breaks optimization. It seeks to reduce the variance within classes and maximize the variance between classes. This binning method was a bit of an improvement, but we still had issues because the data was so skewed. We ended up with 247 data points in the ‘small’ bin, 46 data points in the ‘medium’ bin, and 6 data points in the ‘big’ bin. The chart for the Jenks binning is shown below.

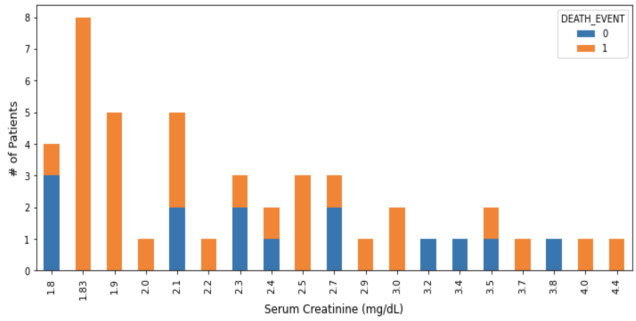


It was interesting to see how the majority of patients in the ‘medium’ and ‘big’ bins died, whereas the majority of patients in the ‘small’ bin lived.

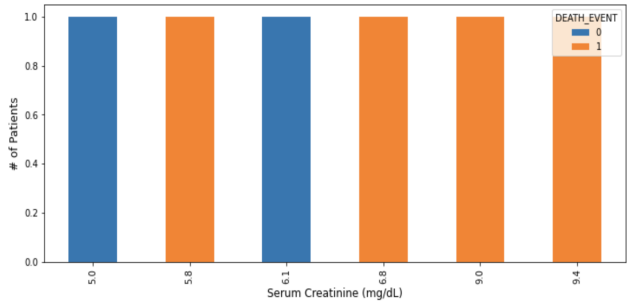
The ‘small’ bin’s range was from 0.5-1.7 and only 24% of these patients died. Below is the chart and value counts for those who lived vs. died.



The ‘medium’ bin’s range was from 1.8-4.4 and 70% of these patients died. Below is the chart and value counts for those who lived vs. died.



The ‘big’ bin’s range was from 5.0-9.4 and 60% of these patients died. Below is the chart and value counts for those who lived vs. died.



V. REGRESSION ANALYSIS

For my regression analysis, I used a logistic regression model to predict whether a heart failure patient lived or died. For this regression, we considered the following variables:

y: DEATH_EVENT
x_1, x_2: ejection_fraction, serum_creatinine

We used the following formula for logistic regression to determine the probability, *P*, that *y=1*:

$$P = \frac{1}{1+e^{-(a+b1x1+b2x2)}}$$

We then needed to find the following coefficients to determine P:

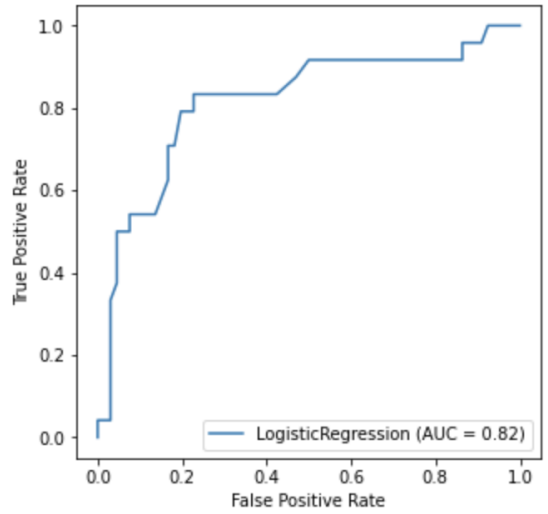
a: intercept
b1: serum_creatinine multiplier
b2: ejection_fraction multiplier

To begin the regression analysis, we first split the dataset into our training datasets x and y and our testing datasets x and y. Our training set contained 70% of the data rows and the testing set contained 30%. The data was randomized before being assigned to a set.

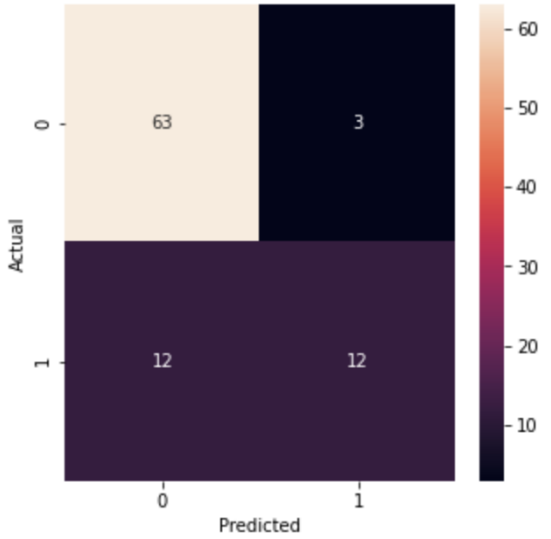
Then, we created our model. We got the following formula:

$$P = \frac{1}{1+e^{-(0.32032117+0.72719078x1+0.05309406x2)}}$$

Using this formula we got the following logistic regression curve:



We also got the following confusion matrix:



We had 63 true negatives, 12 true positives, 3 false negatives, and 12 false positives. This means that the accuracy of our model is 83.34%.

Although our formula was not accurate 100% of the time and contained a somewhat concerning amount of false positives, it was still able to accurately predict the probability of y, patient death, most of the time. To get a more accurate regression model, it might be helpful to have a larger dataset.

VI. CONCLUSION

Through much analysis of Heart Failure data, we were able to discover which factors are signs that a patient may die of heart failure.We determined that Serum Creatinine and Ejection Fraction both play a big role in predicting patient death. Although not 100% accurate, our logistic regression model is able to predict correctly 83.34% of the time. This finding is incredibly useful and can hopefully be used by doctors to determine how best to treat heart failure patients in the future.

REFERENCES

[1]http://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
[2]http://archive.ics.uci.edu/ml/index.php
[3]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
[4]https://www.healthline.com/health/high-creatinine-symptoms#outlook
[5]https://pubmed.ncbi.nlm.nih.gov/15202610/