Prof. Nathan Axvig

Ye Yint Win

Introduction to Data Analytics

April 28th, 2021

Death percentage of heart failure for patients in ICU

For my final project, I’m classifying death rate/percentage of patients in ICU with heart failure. I retrieved this data set, ‘Heart Failure Clinical Record Dataset’, from University of California, Irvine’s, Machine Learning Repository Website under the Center for Machine Learning and Intelligent Systems.

The dataset consists of 299 patients in United States in ICU who have suffered from Heart Failure, general and medical information on patients and whether they have died of heart failure or not. The general and medical information on patients consists of age, sex, smoking, levels of creatinine phosphokinase (CPK), serum Creatinine, and serum sodium, whether they have anemia, diabetes, high blood pressure or not, how many platelets are in their blood vessels, what is the percentage of blood leaving the heart at each conjunction (Ejection Fraction) and Follow-up Period on them.

Before I started doing any analysis, I try to google the terms, CPK, Ejection Fraction, Serum Creatinine, Serum Sodium, Platelets, how they are related to heart failure and what do the increase and decrease of these conditions affect heart failure of a patient. I identify my classification to be “death” and “not” (Alive-to be clearer visually in excel). My purpose is to identify whether a patient is going to die or not with the dataset.

I started my analysis with the histogram of the three health conditions of the patients, collected in the dataset: Anemia, Diabetes and High Blood Pressure in relation with death rate.

A

Chart, histogram

Description automatically generated

As you can see in the image above, the health conditions cannot be used to determine the death percentage of a patient because it is clear that it does not matter a patient have none or three of the conditions. They can be alive or dead.

Then, I tried with Serum Creatinine and Serum Sodium. Both of them show a skewed tendency to left and right respectively. However, the difference is still no vivid yet. Therefore, I made a blended variable, “Serum Feature”, which is ‘*Serum Creatinine \* Serum Sodium’*. ROC Curve of this blended variable shows a good true positive rate close to 1, but its false positive rate is too big to be considered a classifier for death percentage of these patients.

Chart

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generated

Chart, line chart

Description automatically generated

After exploring chemical variables, I came back to the most general and obvious variable and a variable that I found in the research that is a good indicator for determining heart failure patients’ survival rate. These are ‘Age’ and ‘Ejection Fraction’. I believe that age is a very obvious factor and even in its graph, the death rate starts at 45 years old significantly.

Chart, histogram

Description automatically generated

As for Ejection Fraction, it is said to be that the less ejection happens, the more it is dangerous for patients. It can also be seen in the graph at under 40-50. Chart, histogram

Description automatically generated

Even though the graphs show skewed images on facts, there are many other patients who survive from the heart failure under those conditions. They are not enough to determine the death percentages for patients. Their ROC Curves make it obvious for both of them. Even if they are able to achieve true positive rate close to 1, the false positive rate will also be close to .Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

With 4 of the variables, ‘Age’,’Ejection Fraction’, ‘Serum Creatinine’, and ‘Serum Sodium’, I was able to combine them into a IF-Function to classify death or not. However, there has to be 3 fixed variables (in which vivid values can be determined from their respective graphs), and a flexible variable for the ROC Curve. The 3 fixed variables are ‘Age’, ‘Ejection Fraction’ and ‘Serum Sodium’, which has distinctive values of ‘45’,’50’ and ’120’ respectively. I had to test to get those values to get the best true postive rates. At last, the flexible variable is ‘Serum Creatinine’. I chose this to be the flexible because the values are small in number and the range is not large and have a more distinctive ‘Death’ or ‘not’ division in the graph. The IF-formula for this is “=IF(AND(*Age>Threshold(age), Ejection Fraction<Threshold(Ejection Fraction), Serum Sodium>Threshold(Serum Sodium),Serum Creatinine>Threshold(Serum Creatinine)).*” The ROC Curve appears as follows.

Chart, line chart

Description automatically generated

The Threshold values of 1.4-2 can be the best true positive as it has 0.4 true positive rate and under 0.1 or even 0.05 false positive rate to classify as death percentage.

As for comparison, the ROC Curves appears as follow. With the lowest false positive rate of under 0.1, the combined formula of 3 fixed variables and 1 flexible variable has to be the one I would choose to classify whether a patient has a good survival rate or not. One important thing to note is that this dataset deals with human body and medical fields, both of which we are still studying as we speak. Therefore, being able to have a 40% positive rate for a classifier with under 5% false margin is a huge identifier.

Chart, line chart

Description automatically generated