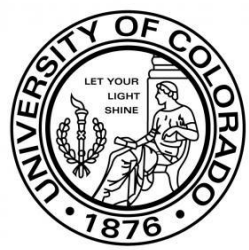
Final Project: Predicting Death of Heart Failure Patients

Jacob Pulitzer, Portia Gifford, Chidi Nzerem, Allison Bruhn

MSBC 5030, Section 006

Ksenia Polson



Heart Failure Predictions

Executive Summary:

Using the Heart Failure Clinical Records dataset, we are trying to help doctors better predict the likelihood of their patients passing away from heart failure. After running a logistic regression, we were able to determine the top variables that correlate to a patient passing away from heart failure are *age, ejection_fraction, serum_creatinine and time*. After running a stepAIC on all of the variables using death event as our dependent variable, we found that the most significant variables were *age, ejection_fraction, serum_creatinine, smoking and time.* With an AIC of 235.5, we chose to use these as our main variables for the rest of the models as they are our best predictors.

When comparing our best predictors to our dependent variable, we saw that age and serum_creatinine had a positive relationship with death event, meaning that for every one increase in these independent variables, the death event will increase by the coefficient within the model. Ejection_fraction, smoking and time all have a negative relationship with death event. This means that for every one increase in these variables, the death event will decrease by the coefficient within the model.

Next we ran a confusion matrix that showed us an error rate of .153. Our error rate could be lower but we gave up having a small error rate to allow us to make a more accurate prediction of .846 by lowering our threshold from 0.5 to 0.398.

Business Problem:

The exact business problem we are trying to solve by using this dataset is how best to predict the likelihood of a patient passing away from heart failure. This data is interesting and needed to know because as of today, nearly 5 million Americans are currently living with congestive heart failure while there are about 550,000 new cases diagnosed each year. Heart failure falls under cardiovascular diseases which is the number one cause of death in America. Since we are helping doctors at a hospital by using this data, we can more efficiently allocate our resources to people who are more prone to death due to heart failure and help those statistics go down.

Logistic Regression:

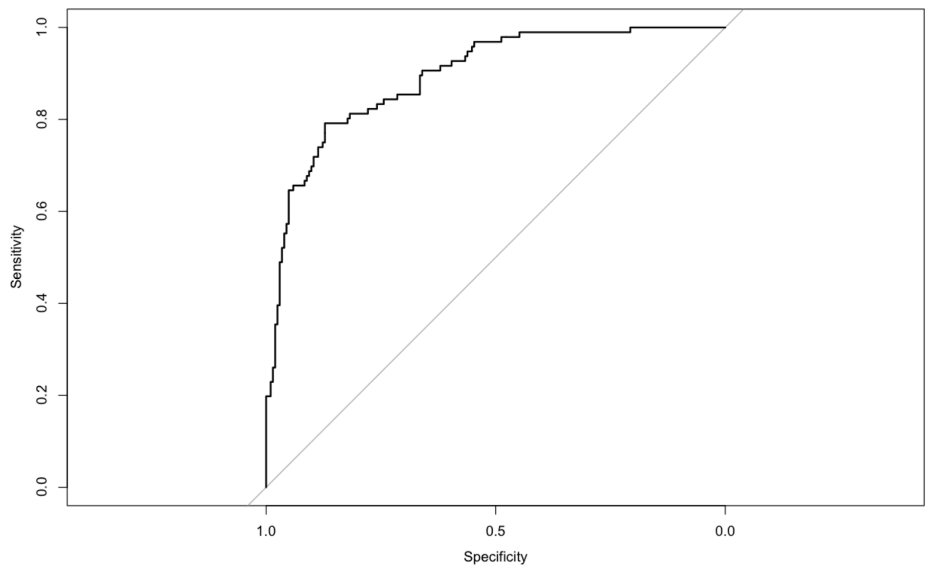
The first thing we did was create a logistic regression model. To do this, we used the variable 'DEATH_EVENT' as our dependent variable to run against all the independent variables. Meaning that the interpretation of the coefficient would be β as the unit change in a Y value form a one unit increase in X holding other variables constant.

From the regression we notice that there are only 4 significant variables, which are: *anaemia, ejection_fraction, serum_creatinine* and *time*. We can also notice that the AIC of this model is 245.55, but to find the most optimal model we run a stepAIC. The most optimal model would be the model with the lowest AIC value, after running the stepAIC we found that in our case the most optimal model was a regression containing the variables: *age, ejection_fraction, serum_creatinine* and *smoking and time*, as well as a AIC of 235.5.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1848  -0.5706  -0.2401   0.4466   2.6668

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.018e+01  5.657e+00   1.801 0.071774 .
age                       4.742e-02  1.580e-02   3.001 0.002690 **
anaemia                  -7.470e-03  3.605e-01  -0.021 0.983467
creatinine_phosphokinase  2.222e-04  1.779e-04   1.249 0.211684
diabetes                  1.451e-01  3.512e-01   0.413 0.679380
ejection_fraction        -7.666e-02  1.633e-02  -4.695 2.67e-06 ***
high_blood_pressure      -1.027e-01  3.587e-01  -0.286 0.774688
platelets                -1.200e-06  1.889e-06  -0.635 0.525404
serum_creatinine          6.661e-01  1.815e-01   3.670 0.000242 ***
serum_sodium             -6.698e-02  3.974e-02  -1.686 0.091855 .
sex                      -5.337e-01  4.139e-01  -1.289 0.197299
smoking                  -1.349e-02  4.126e-01  -0.033 0.973915
time                     -2.104e-02  3.014e-03  -6.981 2.92e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.35  on 298  degrees of freedom
Residual deviance: 219.55  on 286  degrees of freedom
AIC: 245.55
```

Some important things to note were age and serum_creatinine had a positive relationship with death_event, meaning that for every one unit increase in age and serum creatinine there was a 0.04247 and 0.68599 increase in the death_event. Ejection_fraction, serum_sodium and time had negative relationships with death_event. Meaning that every one unit increase in ejection_fraction, serum_sodium and time there was a 0.0734, 0.0645 and 0.0208 decrease in death_event.

To evaluate the model we used the receiver operating characteristic curve, a good model is one that has the output of the area under the curve close to the value of 1, our value was 0.8935 meaning that we made the correct decision in identifying the patience who survived and who did not. This can also be seen in the curve below, we are focusing on the highest point in which the two intercept.



Something to note is that we set our threshold to 0.390 instead of the standard of 0.5, because this would increase our type 2 error rate but in actuality, it makes our prediction better in terms of predicting people who actually survived. Although our overall error rate decreased our chance of making the right prediction increase, we can see this in the accuracy which is 0.846.

Confusion Matrix:

From the confusion matrix, we found our model has an error rate (ERR) of .153 and an accuracy (ACC) of .846. To determine the false positive rate and true positive rate we looked at a confusion matrix comparing our predicted values to our actual values. The true positive rate, or proportion of people who were predicted to die and did end up having a death event, is 79%. With a 79% sensitivity the model correctly identifies 79% of patients who will die, but will miss 21%. The false positive rate (FPR), or the measure of how often an actual negative will be classified as a positive, is .101. Given a FPR of 10%, and a specificity of 89%, we can conclude that our model will correctly identify 89% of patients who will not die, and falsely identify 10% of patients who will die. In general we want to have a low ERR, high ACC, a high sensitivity or

TPR, and a high specificity or a TNR. Because we are testing for a death event in our patient, allowing for a higher error rate will increase our chances of predicting someone will have a death event. This ensures we 'catch' as many death events as possible.

Conclusion:

Our results can be used/deployed in an emergency room when determining the level of care patients need who are having heart failure. After running all patient vitals, a doctor will be able to determine which of his patients are at higher risk of actually passing away, given they are having heart failure. They will then be able to focus their attention on these high risk patients, in turn decreasing the amount of patient deaths.

Risks or potential issues that can occur from deploying this model is there are many other factors that also contribute to heart failure, primarily early symptoms to look out for. By looking for early symptoms such as shortness of breath, fatigue or weakness, lack of appetite, and many more, these symptoms can help detect heart failure before it gets worse. With an ERR of .153, there is also potential for misclassification. With the dataset we looked into, yes these variables are the top signs of heart failure that lead to passing away, but doctors also need to consider the many other symptoms/factors that play into heart failure.

These other symptoms also play into the limitations of the dataset. The Heart Failure Clinical Records look into 13 different variables that take into account the chances of passing away from heart failure, but don't include whether these patients experienced other symptoms of heart failure as listed above as a few examples. These more subtle symptoms should also be considered and accounted for to help detect heart failure even earlier to prevent patients from potentially passing away.

Works Cited:

Laxel. (2020, June 20). Heart Failure Prediction. Retrieved November 15, 2020, from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data?select=heart_failure_clinical_records_dataset.csv

MFMER. (2020, May 29). Heart Failure Symptoms and Causes. Retrieved November 25, 2020, from https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142