Heart Failure Survival Prediction Using Machine Learning Techniques

Brandon May

Introduction:

Cardiovascular disease is still an important cause of morbidity and mortality worldwide. Heart failure, which reflects the heart's inability to pump blood efficiently can have high morbidity and mortality and at times diagnosis can be elusive (Choi *et al.*, 2020). While traditionally machine learning has not been extensively used in clinical medicine outside of academic research environments, there are many different applications of this. Some researchers have used machine learning to assign risk scores to help guide treatment in those with congenital heart disease (Diller *et al.*, 2019). Others have used different types of machine learning algorithms to detect abnormal radiology images and some algorithms are being trained for identifying cancer (Sarkar, 2020).

Heart failure can be a particularly devastating disease and can be caused by a wide variety of different clinical entities. Sometimes the heart failure can be due to an idiopathic or unknown cause. Some studies have noted an increased incidence of heart failure potentially as a consequence of better healthcare and the aging general population (Vasan and Wilson, 2020). In recent years, there have been many different types of therapies to minimize a person's symptoms and increase their functionality. There are significant healthcare costs associated with advanced heart failure and impact many people worldwide. Further, this is a deadly disease in that in some estimations, the 5-year mortality rate is around 50% if diagnosed early but if diagnosed during late disease manifestation, it is even higher (Hobbs, 2010).

This dataset, detailed below, includes several variables that will likely be helpful in prediction of death. The ejection fraction is a measurement of the heart's pumping ability and normally is between 50-65%. Those with systolic heart failure will have an ejection fraction <45% typically (Borlaug, 2019). There is another type of heart failure and has different manifestations and etiologies but is still a significant disease. As the heart fails, fluid balance becomes erratic. Vascular risk factors including diabetes and smoking are likely to only worsen the situation. Further, creatinine, a measure of kidney function, typically worsens as heart failure worsens.

Methods:

Data was obtained via PLOS and is based on a project in the medical journal of BMC Medical Informatics and Decision Making. The dataset can be found at the following URL: https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1. This study population was from India and had clinical risk factors as well as follow-up data on those diagnosed with heart failure. There were 299 subjects in the study population and had 12 variables describing clinical risk factors. Survival or death was considered the end point and target variable and was categorically represented as 1 in the dataset. Numeric variables included the follow-up time of the study, age, ejection fraction, sodium, creatinine level, platelets, and CPK levels. Categorical variables included gender, smoking status, diagnosis of high blood pressure, as well as presence of anemia.

Using Python, data was imported and cleaned. There were no null values in the dataset. Columns were renamed for simplicity and the categorical variables of Event, Gender, Smoking, Diabetes, BP, and Anemia were recoded. The numerical variables had some interesting descriptive statistics. The lowest age was 40 and the oldest person was 95. The lowest ejection fraction was 14% and the maximum was 80%. There was a significantly low sodium value of 113 and a high level of 148. For reference, normal sodium levels are between 135-145. The highest creatinine was 9.4 which is exceedingly high approaching renal failure range. Likewise, the lowest platelets of 25K and the highest were 850000. CPK, a muscle enzyme, had minimums of 23K with maximum of 7800K. Looking at the means, the mean follow-up time was 130 days, age 60, EF of 38%, sodium of 136, creatinine of 1.39 (slightly elevated), 263K platelets (which is normal), and CPK of 481. Based on the categorical variables, more patients survived, were male, non-smokers, non-diabetic, non-anemic, and with normal blood pressure. A correlation matrix was created to look for high absolute values of correlation >0.95, of which there were none.

Visualizing the numeric variables had sodium levels positively skewed with creatinine, platelets, EF, and CPK levels negatively skewed. Age was the only variable that appeared to be approximately normally distributed though there was some slight positive skew.

Boxplots were also created to help visualize any significant outliers. Age and follow-up time did not appear to have any significant outliers. The EF variable had a couple of ejection fractions that were 70% and 80% respectively. This is within the range of possibility, medically speaking. There were outliers with sodium levels below 125. There were significant outliers ranging from above 2 – 8 which are usually indicative of renal failure. CPK levels tended to have outliers on the higher end of the spectrum. Platelets had outliers on both the positive and negative side of the spectrum. These outliers

were deliberately kept within the dataset. The main reason for this is that when discussing survival, outliers can be predictive of mortality. Those with renal failure and heart failure are at high risk of death and could certain be predictive of mortality. Generally speaking, many of these variables that have significant outliers could have outliers for a good reason – they are portents of a poor outcome.

The numerical variables were then plotted using swarm plots to compare their distributions against the two classes of the target variable (alive vs. dead). There are several interesting findings when plotting survival vs our different numerical variables. Unsurprisingly, those who had longer follow-up in days were more likely to be alive. There was a trend that the older individuals were more likely to not have survived. It is subtle, but there appears to be an association between lower EF and survival, and this is logical; however, looking at the distributions, there was not a significant difference between the two. Likewise, there did not seem to be a significant difference between sodium level and survival. The higher the level of creatinine above 2, the more likely they were to be dead. These creatinine levels are exceedingly high and the higher the creatinine, the higher the likelihood of kidney failure which explains this trend. There did not seem to be a significant difference between survival and the platelet/CPK levels.

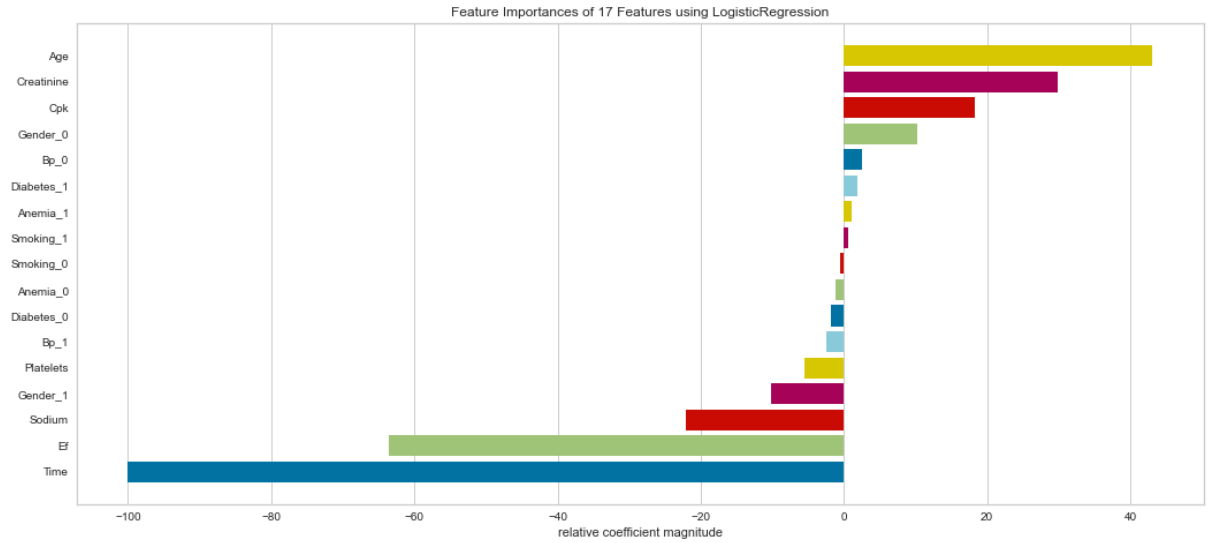
Results:

In general, when fitting the data using various models, the false negatives or predicting survival when the patient could be at risk of death should be minimized if at all possible. The risk of predicting someone who is at higher risk of death who truly is not at risk would not be as catastrophic as a false negative. The algorithms were evaluated using precision, recall, F1 scores, and ROC curves.

Logistic Regression:

The data was encoded with dummy variables to account for the categorical variables, standardized using SciKitLearn in Python. Then the data was split into 80/20 training/testing portions and fit using a logistic regression classifier and a confusion matrix, classification report, and ROC curve were obtained.

The logistic regression model performed reasonably well with an F1 score of 0.615 of predicting death and 0.815 of predicting death. There were 8 false negatives in this model. The ROC was 0.81 for both class predictions. Feature importance was done using Yellowbrick visualization. It noted that age, creatinine, CPK levels, and female gender were some of the most important features based on their relative coefficients.



KNN:

Since KNN relies on the mathematical distance between points to aid in prediction, the data was standardized, with dummy variables encoded for categorical variables, and then split into 80/20 training/testing groups. The KNN model was fit and the same metrics were calculated. This model performed worse with 11 false negatives. The F1 score for death was 0.79 and the was 0.47 for survival.

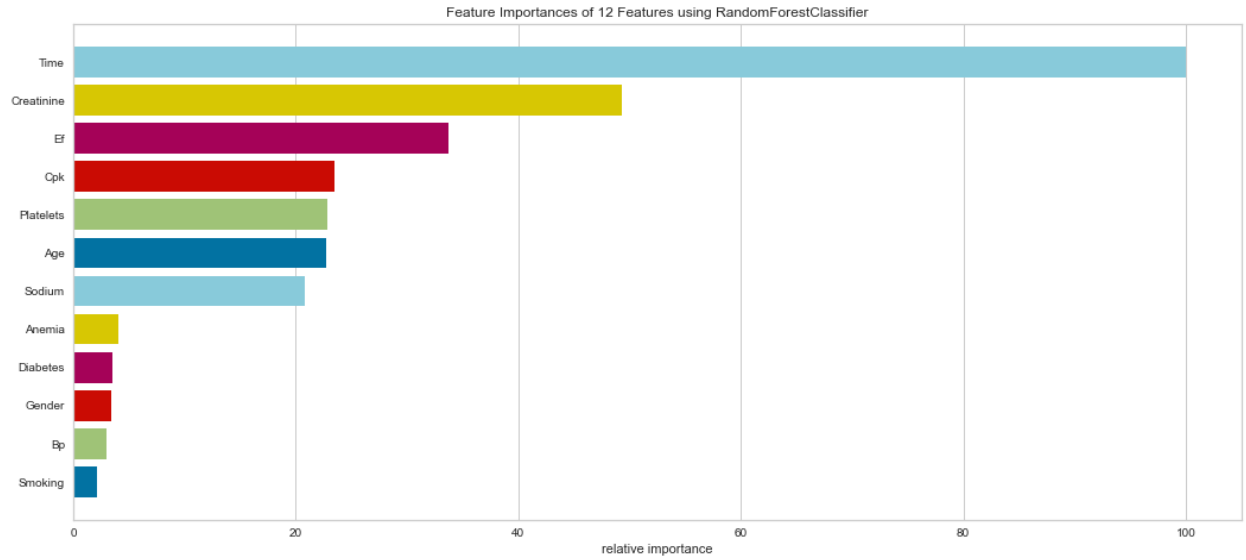
The model was then tuned using GridSearchCV and noted the best settings were using n neighbors of 5, p of 2, leaf size of 1, uniform weights, and the minkowski metric. Using these metrics the F1 score improved marginally to 0.80 for death, 0.40 for survival, but increased the false negatives to 13.

Support Vector Machine:

Support vector machine was run using the same parameters. This model had a low number of false negatives at 8 people incorrectly predicted. The F1 scores were higher than the logistic regression model at 0.847 for predicting survival and 0.629 for predicting death. The ROC was slightly better at 0.83. Overall, this appears to be marginally better. Checking feature importance, creatinine, age, cpk, and female gender were all highly important. The one difference is that the presence of anemia and current smoking also appeared to be relatively important in the model.

Random Forest Classifier:

A random forest ensemble classifier was then run. This model performed quite well. It had only 4 false negatives and an F1 score of 0.70 for predicting death, and 0.884 for predicting survival. The ROC was 0.90 for both classes. Feature importance was also computed for the random forest model as well.



It is interesting that in this model, the ensemble method put a high weight on follow-up time. Creatinine, EF, CPK, Age, Platelets, and Sodium levels were all important. Gender was less important in this model. This result is interesting in that EF is directly tied to prognosis in heart failure patients and sodium level can be indicative of fluid balance. Abnormal sodium levels can indicate worsening heart failure symptoms, so this makes medical sense.

XG Boost:

XG Boost, another ensemble method also performed quite well. There were only 4 false negatives. The F1 score for predicting death was 0.769 and predicting survival was 0.889. The ROC was 0.92. Feature importance showed a similar result to the Random Forest method showing that time of follow-up, creatinine, EF, female gender, age, and sodium level were all important among several others.

Conclusions:

Heart failure is a very deadly disease. Even 20 years ago, this diagnosis would be a death sentence. There have been significant pharmacologic advancements in recent years which have improved survival. Those with heart failure can have hospital admissions for fluid overload, trouble breathing, and even other cardiac events. However, heart failure survival is still dependent upon the severity of the symptoms and other objective measurements as well as early diagnosis and treatment (Colucci, 2019). Therefore, it is timely that we may be able to utilize machine learning algorithms to help better predict death. Perhaps if we can predict a higher risk of death based on severity of various lab

measurements and other risk factors, there could be an intensification of therapy and/or aggressive follow-up and monitoring to ensure adequate care and prolonged symptom-free survival.

It is not a significant surprise that the Random Forest and XG Boost models performed the best based on ROC and F1 scores. The Random Forest managed to edge out the XG Boost performance by a small margin. Since both of these are ensemble methods, they have multiple algorithms that are being evaluated and the prediction is then confirmed using a majority vote option. Which model to use is somewhat arbitrary given their largely similar results.

In this population of patients, there were significantly more people alive than dead at the end of the study period. This explains why many of the models predicted more to be alive than dead on a relative basis. The ensemble methods performed the best at this showing the power of these algorithm to be able to predict the harder class (death). This is the point of this project; we want to primarily identify those at high risk of mortality to either perform some sort of intervention or close monitoring. However, predicting survival can also be beneficial in that if a patient were to be predicted to survive by certain criteria, you could ensure certain factors about their health are optimized and be reassured.

Finally, it is fascinating that the machine learning algorithms all seemed to agree on certain variables that relatively more important to the prediction. The majority of them noted age, creatinine, CPK, and female gender to have more predictive power. The ensemble methods picked up on more of the traditional variables of sodium, ejection fraction, among others. Many of these variables are already medically validated as related to heart failure prognosis. The machine learning algorithms, at least based on this dataset, agreed with these factors. These factors do differ slightly based on the different algorithms used but many of the algorithms picked out similar important variables.

This dataset is smaller and has a small number of patients. This is partly due to the low numbers of heart failure overall. This is a limitation on the predictive power. Much larger datasets would be necessary among a large patient demographic to apply to the general population. This does continue to show that machine learning can aid healthcare workers and providers in clinical management.

References Cited:

Ahmad, Tanvir; Munir, Assia; Bhatti, Sajjad Haider; Aftab, Muhammad; Ali Raza, Muhammad (2017): DATA_MINIMAL. PLOS ONE. Dataset. https://doi.org/10.1371/journal.pone.0181001.s001

Borlaug, B. A., MD. (2019). Clinical manifestations and diagnosis of heart failure with preserved ejection fraction. Retrieved September 27, 2020, from https://www.uptodate.com/contents/clinical-manifestations-and-diagnosis-of-heart-failure-with-preserved-ejection-fraction?search=heart+failure

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making, 20*(1). doi:10.1186/s12911-020-1023-5

Choi, D., Park, J. J., Ali, T., & Lee, S. (2020). Artificial intelligence for the diagnosis of heart failure. *Npj Digital Medicine, 3*(1). doi:10.1038/s41746-020-0261-3

Colucci, W. S., MD. (2019). Prognosis of Heart Failure. Retrieved September 27, 2020, from https://www.uptodate.com/contents/prognosis-of-heart-failure?search=heart+failure+mortality

Diller, G., Kempny, A., Babu-Narayan, S. V., Henrichs, M., Brida, M., Uebing, A., . . . Gatzoulis, M. A. (2019). Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients. *European Heart Journal, 40*(13), 1069-1077. doi:10.1093/eurheartj/ehy915

Hobbs, F. (2009). Clinical burden and health service challenges of chronic heart failure. *European Journal of Heart Failure Supplements, 8*(Supplement 1), I1-I4. doi:10.1093/eurjhf/hfp011

Sarkar, T. (2020, April 30). AI and machine learning for healthcare. Retrieved September 27, 2020, from https://towardsdatascience.com/ai-and-machine-learning-for-healthcare-7a70fb3acb67

Vasan, R. S., MD, DM, FACC, & Wilson, P. W., MD. (2020). Epidemiology and causes of heart failure. Retrieved September 27, 2020, from https://www.uptodate.com/contents/epidemiology-and-causes-of-heart-failure?search=heart+failure

Appendix A – Variable Descriptions

Time: Integer, signifying length of follow-up in days.

Event: Categorical Target Variable (0: Alive, 1: Dead)

Gender: Categorical (0: Female, 1: Male)

Smoking: Categorical (0: Non-Smoker, 1: Smoker)

Diabetes: Categorical (0: No Diabetes, 1: Diabetes)

BP: Categorical (0: No Hypertension, 1: Hypertensive)

Anemia: Categorical (0: No Anemia, 1: Anemia)

Age: Integer (in years)

Ejection.Fraction: Integer (percentage)

Sodium: Integer (mg/dL)

Creatinine: Float (mg/dL)

Platelets: Integer (mg/dL)

CPK: Integer