Nicholas Mosca

BF550 Foundations on Programming, Data Analytics, and Machine Learning in Python
Final Report 12/10/2020

Introduction:

Cardiovascular disease is related to millions of deaths globally each year. Disease is typically linked to myocardial infractions and heart failures (HF). Heart Failure occurs when the heart cannot pump enough blood to the rest of the body. Multiple cardiovascular diseases (CVD) contribute to HF. Currently quantitative measurements and predictions of heart disease risk rely on the New York Heart Association (NYHA). The NYHA ranking system has 4 classes. Class 1 shows no symptoms and is least at risk. Class 4 risk can be demonstrated by discomfort and symptoms at rest. Unfortunately, this classification system is broad and leaves room for vast amounts of uncertainty. Poor classification for such a vital organ stresses the need for more accurate risk classification systems based on clinical data.

Electronic medical records including symptoms, body features and clinical laboratory tests are typically used to evaluate patient health. Heart failure is typically predicted by the ejection fraction value, among other features. Ejection Fraction measures the portion of blood pumped away from the heart during a single contraction [1]. Less efficient hearts will likely have an ejection fraction number under 50%. Ejection fraction is an important metric, but the importance compared to other clinical features has been unknown till recently. The raw data for this analysis consists of electronic medical records (EHR) from patients in Pakistan. The data was acquired from the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan) from April – December 2015[2]. Each patient has shown previous signs of heart failure or left ventricular dysfunction class 3 or higher according to the New York Heart Association.

The aim of the study is to rank each clinical feature and predict patient's survival rates. Biostatistics and machine learning methods are being applied to the data to unlock key features from medical record data. Identification of key clinical features can provide an importance metric to values helping providers assist risk at a high level of accuracy.

Methods:

Data:

Each medical record has 13 features collected from symptoms, clinical lab tests, and lifestyle information for each patient. Medical records for 299 patients have been aggregated into a csv file. Binary features include anemia, high blood pressure, diabetes, sex, and smoking. Anemia was considered true if hematocrit levels were lower than 36%. Creatinine phosphokinase (CPK) is feature that indicates the level the CPK enzyme in blood. When a muscle (heart) becomes damaged, CPK leaks into the blood stream. High levels of CPK in blood is another indication used for HF risk. Ejection fraction states the percentage of blood leaving the left ventricle per contraction. The serum creatine is a waste product generated by creatine, when a muscle breaks down. Currently serum creatine is a feature typically used for kidney function and its relation to HF risk is not well annotated. High levels of serum creatinine in the blood can be an indication of renal dysfunction [3]. Serum sodium is a standard feature measured in routine blood tests. Sodium is an essential mineral for the body, low levels can be a feature associated with HF. Death event was also included in the dataset as a binary feature that was measured within the 130 day follow up window.

Feature ranking:
Typical biostatistics analysis including mean, standard deviation, min, max, and percentile data was conducted on the full dataset via Python 3.76 (Numpy, Pandas). These metrics were used as a baseline look at the data before cleaning and creating a machine learning model.

Machine learning:
This abbreviated analysis focuses on the binary prediction of a patient's survival.
Feature importance ranking was conducted using Random Forests classifier. Two techniques within the Random Forest classifier provided ranking systems of the most important clinical features for predicting death. Gini importance (or mean decrease impurity), permutation-based feature importance both predicted feature rankings following the Random Forest structure. The data was trimmed and fed into a Random Forest model containing 10,000 decision trees. Overall, the model checked all the binary outcomes of each decision tree and chose ranking outcomes based on majority vote.

Feature ranking using Gini impurity which is defined as total decrease in note impurity. Within each branch the selected or targeted feature is used to make decisions on how to divide the rest of the dataset into separate sets with similar outcomes as the feature node. The separation is generated based on similarity, specifically information gain in the form of variance reduction for each subset. The features with the highest decrease (least different) are selected for additional nodes. The difference between notes is measured with a metric called Gini importance, as importance drops impurity increases. Gini impurity is measured on a scale between 0-1 with all features adding to one and the largest being most significant.

Permutation importance was also used to measure the increased prediction error for the model. This technique can also be defined as the model error rate for each feature. All machine learning models were created using Python 3.76 (Pandas, Numpy, Sklearn.ensemble, and Sklearn.model_selection.

Code structure:
The code structure for my analysis is broken into multiple files that represent the figures that have been reproduced. One file is strictly breaking down and organizing the data into subsets used for plotting. Each additional file is designed to run and generate the targeted figure. The Random Forest model, Gini Importance, and Permutation importance are all written into one file along with the figure creation.

Results:

Figure 1:

A.

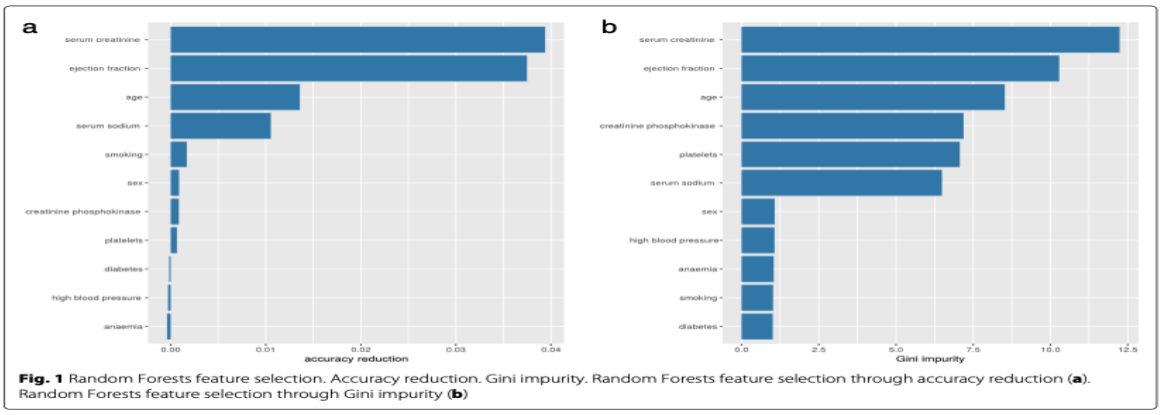

Fig. 1 Random Forests feature selection. Accuracy reduction. Gini impurity. Random Forests feature selection through accuracy reduction (**a**). Random Forests feature selection through Gini impurity (**b**)

B.

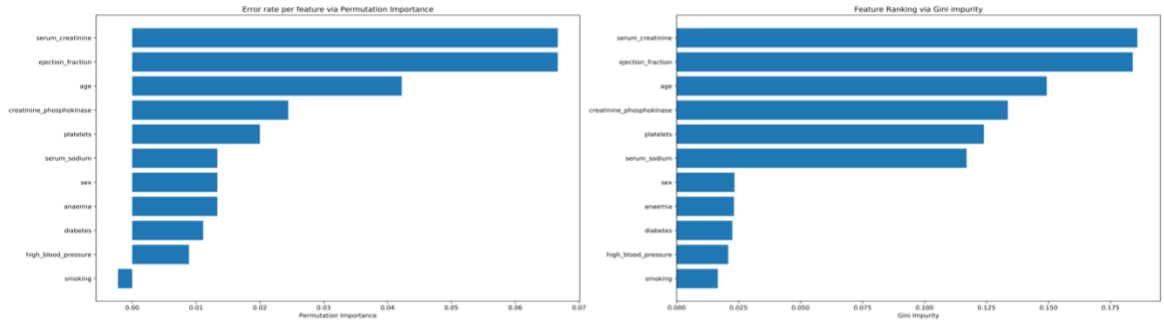

Figure 1A shows the Random Forest feature ranking from Chicco et al. Gini impurity and accuracy were used to identify the most important clinical features related to heart disease and prediction of death. The top three respectively were serum creatine, ejection fraction, and age. The accuracy reduction shows the error rate for the respective feature within the model.

Figure 1B is my replication using the same clinical data. After cleaning the data, I created a model using sklearn and conducted feature ranking using gini impurity methods. The training set for the model was 70% of the data and the remaining 30% was used for testing. I also used permutation importance to visualize the error per feature. My results mirrored the majority of the feature ranking from Chicco et al. Figure 1A and 1B show the same five clinical features with identical rankings. One difference is the permutation importance errors; this can likely be attributed by difference in model construction or randomization of the tree design.

Figure 2:

A.

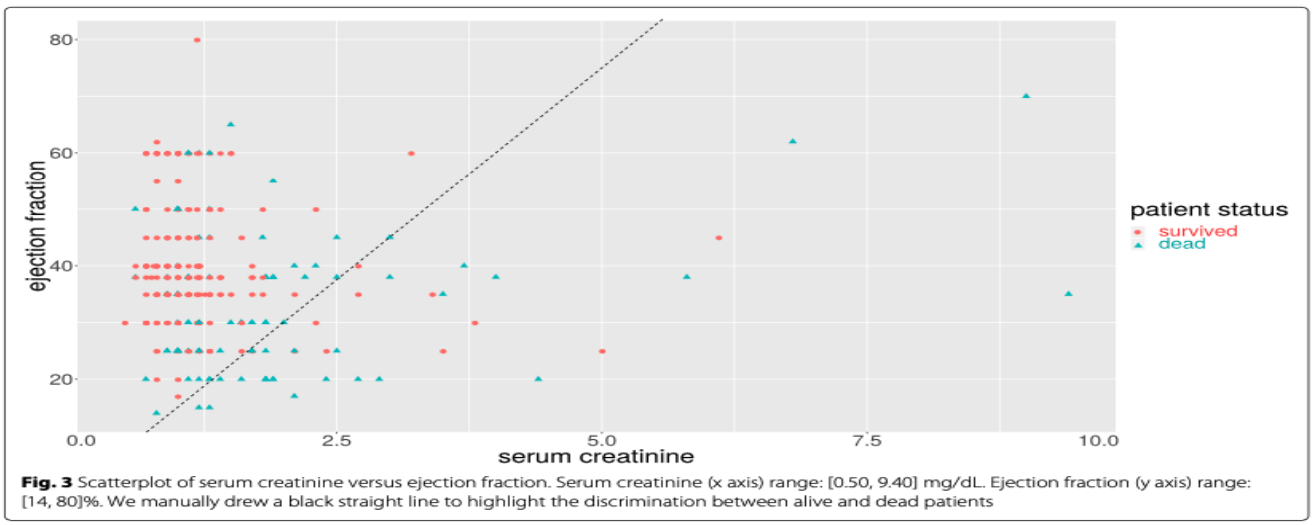

Fig. 3 Scatterplot of serum creatinine versus ejection fraction. Serum creatinine (x axis) range: [0.50, 9.40] mg/dL. Ejection fraction (y axis) range: [14, 80]%. We manually drew a black straight line to highlight the discrimination between alive and dead patients

B.

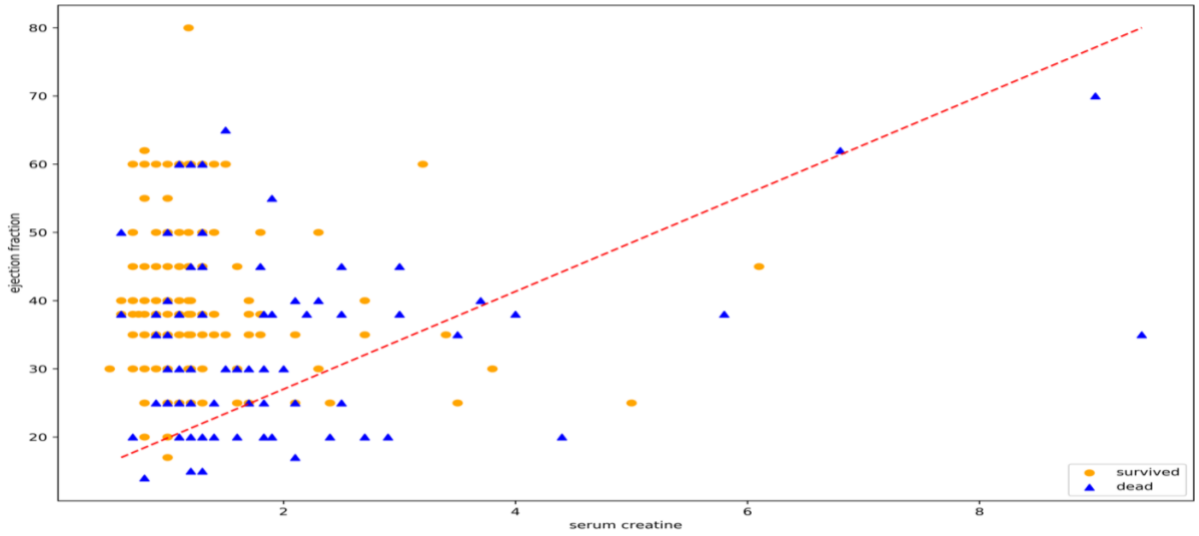

Figure 2A displays a comparison of serum creatine versus ejection fraction for both living and deceased patients. The x axis ranges from 0.50 to 9.40 mg/dL while the y axis ranges from 0 to 80 %. Figure 2B is a replication of figure 3 from Chicco et al showing similar results. All patient ejection fraction and serum creatine levels were isolated and broken into survived and dead subsets. This figure shows the relationship between the two most important clinical features according to Figure 1 A and B. A clear coloration is shown between higher serum creatine and death. This makes sense because creatine is a biproduct of damaged muscle in the heart.

Future Ideas:
Overall, I am satisfied with my replication attempts from Chicco et al. Although not shown I would focus on the correlation between serum creatine, ejection fraction, and its relationship to death. I would likely start with Peterson's correlation coefficient to get a rough estimate of R. Another possible analysis could be to try a different machine learning model and conduct future ranking again and compare. I think a drawback of this model is the small sample size (n). I would be curious to see this same analysis with more patients. I do think the feature rankings would be the same, but the permutation importance would drop with a larger n.

References:

1. 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. 2020 Feb 3;20(1):16. doi: 10.1186/s12911-020-1023-5. PMID: 32013925; PMCID: PMC6998201.

2. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, Dahlström U, O'Connor CM, Felker GM, Desai NR. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. J Am Heart Assoc. 2018 Apr 12;7(8):e008081. doi: 10.1161/JAHA.117.008081. PMID: 29650709; PMCID: PMC6015420.

3. Meng F, Zhang Z, Hou X, Qian Z, Wang Y, Chen Y, Wang Y, Zhou Y, Chen Z, Zhang X, Yang J, Zhang J, Guo J, Li K, Chen L, Zhuang R, Jiang H, Zhou W, Tang S, Wei Y, Zou J. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China. BMJ Open. 2019 May 16;9(5):e023724. doi: 10.1136/bmjopen-2018-023724. PMID: 31101692; PMCID: PMC6530409.