Predicting Survival from Heart Failure

Ethan Biegeleisen

Manhattan College
MATG 557 - Machine Learning
Spring 2021

Why this Topic?

- Had mostly done projects in other topics (usually government data)
- Wanted to work with medical data
- Wanted to test out prediction methods using various risk factors
- Wanted to compare the the accuracy of risk factors with each other

[1]

A Bit About Heart Failure [2]

- Heart muscles enlarge, restricting the pumping of blood out of the heart
- Heart chambers can lose flexibility and have trouble filling properly between heartbeats
- The heart gradually becomes unable to meet the body's requirements, which leads to trouble breathing
- Primary causes of heart failure include coronary heart disease, diabetes, and high blood pressure
- Can also be caused by HIV, alcohol abuse, cocaine, thyroid disorders, excessive Vitamin E, radiation, or chemotherapy

Project Goals

- Test feature effectiveness at predicting survival using Logistic Regression, Support Vector Machine (SVM) with linear, polynomial, Gaussian, and sigmoid kernels, Random Forest, and Dr. DeBonis' classifier
- Test these methods with repeated 10-fold cross-validation
- Check standard prediction accuracy and Area Under the Receiver Operating Characteristic Curve (ROC AUC)
- Use Least Absolute Shrinkage and Selection Operator (LASSO) to find the data set's most accurate predictors for survival
- Compare effectiveness of only using the most accurate predictors vs all predictors

How is the Data Structured? (Part 1) [3]

There are 299 patients, 12 features, and 1 target. 203 of the patients survived. The variables in the data set are the following:

- Age (Years)
- Anaemia (Binary) - Decreased red blood cell count or decreased hemoglobin
- Creatinine Phosphokinase (mcg/L) - Level of CPK enzyme in blood
- Diabetes (Binary)
- Ejection Fraction (Percentage) - Percentage of blood leaving the heart at each contraction
- High Blood Pressure (Binary)
- Platelets (kiloplatelets/mL) - Amount of platelets in the blood

How is the Data Structured? (Part 2) [3]

- Serum Creatinine - Level of creatinine in the blood
- Serum Sodium - Level of sodium in the blood
- Smoking (Binary)
- Sex (Binary)
- Time (Days) - Follow-up period
- Target: Death Event (Binary) - Whether the patient died during the follow-up period

Sample of the Data:

age	anaemia	creatinine_	diabetes	ejection_fi	high_blooc	platelets	serum_cre	serum_soc	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1

Initial Prediction Results

Modeling the entire data set with 1 iteration:

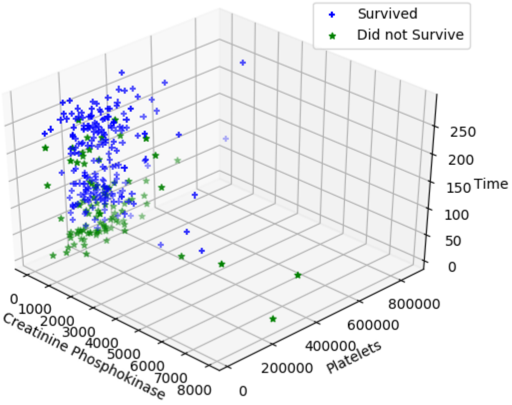
Method	Accuracy
Logistic Regression	0.896
SVM (linear)	0.906
SVM (polynomial)	0.886
SVM (Gaussian)	0.920
SVM (sigmoid)	0.679
Random Forest	0.803
Dr. DeBonis' classifier	0.829

LASSO Results

The order of importance found for the features:

1. Platelets
2. Creatinine Phosphokinase
3. Time
4. Ejection Fraction
5. Age
6. Serum Sodium
7. Serum Creatinine
8. Sex
9. Diabetes
10. High Blood Pressure
11. Smoking
12. Anaemia

Plotting the 3 Most Accurate Predictors



Logistic Regression Results

Using repeated 10-fold cross-validation with 100 iterations:

# Features	Mean Accuracy	Mean ROC AUC
12	0.825 ± 0.067	0.852 ± 0.078
7	0.825 ± 0.068	0.852 ± 0.077
6	0.826 ± 0.067	0.854 ± 0.077
5	0.829 ± 0.068	0.856 ± 0.077
4	0.808 ± 0.069	0.817 ± 0.086
3	0.820 ± 0.067	0.817 ± 0.086
2	0.674 ± 0.078	0.525 ± 0.113

Gaussian SVM Results

Using repeated 10-fold cross-validation with 100 iterations:

# Features	Mean Accuracy	Mean ROC AUC
12	0.679 ± 0.078	0.512 ± 0.032
7	0.679 ± 0.078	0.509 ± 0.029
6	0.679 ± 0.078	0.514 ± 0.023
5	0.679 ± 0.078	0.501 ± 0.010
4	0.679 ± 0.078	0.505 ± 0.030
3	0.679 ± 0.078	0.517 ± 0.034
2	0.684 ± 0.078	0.530 ± 0.052

Random Forest Results

Using repeated 10-fold cross-validation with 100 iterations:

# Features	Mean Accuracy	Mean ROC AUC
12	0.820 ± 0.066	0.884 ± 0.063
7	0.822 ± 0.066	0.886 ± 0.063
6	0.822 ± 0.068	0.872 ± 0.071
5	0.813 ± 0.067	0.870 ± 0.070
4	0.822 ± 0.068	0.862 ± 0.073
3	0.796 ± 0.069	0.788 ± 0.096
2	0.610 ± 0.086	0.527 ± 0.116

Results from Dr. DeBonis' Classifier

Using repeated 10-fold cross-validation with 10 iterations:
(Class 0 - survived; Class 1 - Did not survive)

# Features	Mean Error	Mean Accuracy	Mean ROC AUC
12	0.238 ± 0.078	0.762	0.785 ± 0.081
7	0.207 ± 0.073	0.793	0.820 ± 0.088
3	0.226 ± 0.074	0.774	0.778 ± 0.095

# Features	Class 0 Mean Error	Class 1 Mean Error
12	0.146 ± 0.089	0.432 ± 0.178
7	0.124 ± 0.083	0.385 ± 0.156
3	0.115 ± 0.082	0.464 ± 0.155

Conclusions

- Most to least effective methods for predicting survival were Random Forest, Logistic Regression, Dr. DeBonis' classifier, and Gaussian SVM
- SVM was the only method to work poorly
- Every method worked just as well or better without the binary features
- Logistic Regression worked best with 5 features, Gaussian SVM worked best with 2 features, Random Forest worked best with 7 features, and Dr. DeBonis' classifier worked best with 7 features

Future Directions

- Investigate why SVM performed noticeably worse than the other methods
- Investigate why Dr. DeBonis' classifier poorly predicted Class 1
- Check the accuracy of predicting Class 1 using the other methods
- Investigate why the binary features were seemingly unnecessary predictors
- Would a feature like Smoking be a stronger predictor if it had more than 2 categories? (no/light/moderate/heavy vs no/yes)

References

Special thanks to Dr. DeBonis

Thanks for listening! Any questions?

References

[1] Society of Cardiovascular Angiography and Interventions (2015). Diagnosing heart failure. http://www.secondscount.org/heart-condition-centers/info-detail-2/diagnosing-heart-failure

[2] Ahmad T., Munir A., Bhatti SH., Aftab M., Raza M.A. (2017) Survival analysis of heart failure patients: A case study. *PLoS ONE* 12(7): e0181001. https://doi.org/10.1371/journal.pone.0181001

[3] Chicco, D., Jurman, G. (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* 20(16). https://doi.org/10.1186/s12911-020-1023-5

[4] Brownlee, J. (2020) Repeated k-Fold cross-validation for model evaluation in Python. https://machinelearningmastery.com/repeated-k-fold-cross-validation-with-python/