Jessica Pak
PH245 Fall 2021

Final Project: Modelling Heart Failure

Abstract: Heart failure is a leading cause of death in the US, accounting for a high portion of healthcare spending. Electronic medical records of 299 patients with heart failure were analyzed and fitted with supervised machine learning models to classify patient survival and rank clinical features corresponding to the most important risk factors. Linear SVM was found be the most accurate in classifying patient's binary outcome of a death event, with prediction accuracy at 76.57%. Both logistic regression and random forest models ranked the most important clinical features to be age, CPK, ejection fraction, and serum creatinine (not in order of importance), while the random forest model also provided insight into 2 other predictors, platelets and serum sodium, that may also be important in predicting a death event from heart failure.

Background:

Cardiovascular diseases remain the leading cause of death in the US according to recent data, and mainly presents itself as myocardial infarctions and heart failure (HF). HF in the US afflicts about 6.2 million adults and was estimated to cost the nation $30.7 billion in 2012 (and likely more today)[1]. There are well-known risk factors for HF including coronary artery disease, diabetes, hypertension, obesity, other heart conditions, smoking tobacco, eating high fat, cholesterol, and/or sodium foods, physical inactivity, and/or excessive alcohol intake[1]. Electronic medical records (EMR) of patients are able to quantify symptoms, physical features, and clinical test results. Multivariate statistical analysis can use these data features to detect patterns and correlations that may otherwise be undetected by healthcare professionals[2]. Machine learning, in particular, can predict patients' survival from their clinical data and highlight the most important features. With more accurate predictions, patients and healthcare professionals could intervene on the clinical and lifestyle factors putting patients most at risk of a death event and potentially decrease healthcare spending on heart failure complications.

Methods:

For this project, I used EMR data obtained from UCI's Machine Learning Repository (link) of 299 patients with heart failure, collected during their follow-up period. Each patient profile has 13 clinical features as shown below:

- age: age of the patient (numeric, years)
- anaemia: decrease of red blood cells or hemoglobin (boolean)
- high blood pressure: if the patient has hypertension (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (numeric, mcg/L)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (numeric, percentage)
- platelets: platelets in the blood (numeric, kiloplatelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (numeric, mg/dL)
- serum sodium: level of serum sodium in the blood (numeric, mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (numeric, days)
- [target] death event: if the patient deceased during the follow-up period (boolean).

Because the variable "time" does not quantify a patient's physical or lifestyle feature, I decided to exclude it as an independent (predictor) variable in my analyses. Thus, a total of 11 independent variables (age:smoking) were included with 1 dependent (outcome) variable, death event.

I started with a simple classification method using a logistic regression model. After diagnosis of this model, I compared the prediction accuracy rates of different supervised machine learning methods[3], including random forests (RF), k-Nearest Neighbors (KNN), and support vector machines (SVM, linear and radial) using k-fold cross-validation. Based on the results of these comparisons, I then compared feature rankings from the logistic regression model to a supervised machine learning model. Given the comparable accuracy of the RF model to the logistic regression model (as shown in results), as well as its feature ranking techniques: mean accuracy reduction and Gini impurity reduction, I used the RF model to compare important clinical features to the ones detected by the logistic regression model[4]. Using different supervised machine learning methods, I aimed to predict patients' survival from heart failure and determine which predictor variables from this clinical dataset were significant.

Results:

Logistic regression model

```
Call:
glm(formula = DEATH_EVENT ~ ., family = binomial, data = hf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3184  -0.7692  -0.4436   0.8293   2.4880

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.964e+00  4.601e+00   1.079 0.280625
age                       5.569e-02  1.313e-02   4.241 2.23e-05 ***
anaemia1                  4.179e-01  3.009e-01   1.389 0.164904
creatinine_phosphokinase  2.905e-04  1.428e-04   2.034 0.041907 *
diabetes1                 1.514e-01  2.974e-01   0.509 0.610644
ejection_fraction        -7.032e-02  1.486e-02  -4.731 2.23e-06 ***
high_blood_pressure1      4.189e-01  3.061e-01   1.369 0.171092
platelets                -7.094e-07  1.617e-06  -0.439 0.660857
serum_creatinine          6.619e-01  1.734e-01   3.817 0.000135 ***
serum_sodium             -5.667e-02  3.338e-02  -1.698 0.089558 .
sex1                     -3.990e-01  3.508e-01  -1.137 0.255394
smoking1                  1.356e-01  3.486e-01   0.389 0.697300
---
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: 294.28  on 287  degrees of freedom
AIC: 318.28

Number of Fisher Scoring iterations: 5
```

Fig1: Based on the output, coefficients for age, CPK, ejection fraction, and serum creatinine seemed to be significant, $\alpha > 0.05$.

Before accepting these results, I needed to perform diagnosis of the model. Assumptions to check for in the logistic regression model were as follows[5]:

1. Outcome is a binary or dichotomous variable

   ```
     0  1
   203 96
   ```

2. Linearity between the log odds of the outcome and each continuous predictor variables

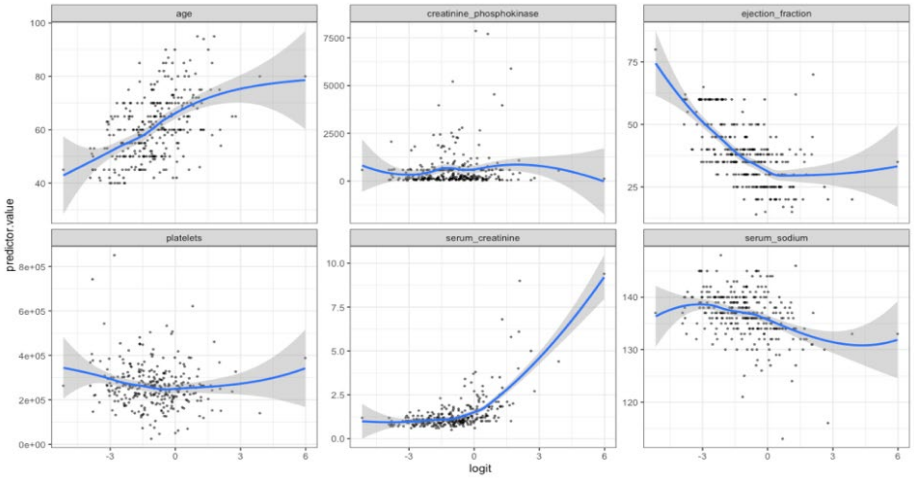

Fig 2: Continuous predictor variables in the dataset were age, CPK, ejection fraction, platelets, serum creatinine, and serum sodium. As shown from the plots, not all continuous predictor variables had a linear relationship with the log odds of the outcome variable.

3. No influential values (extreme values or outliers) in the continuous predictors

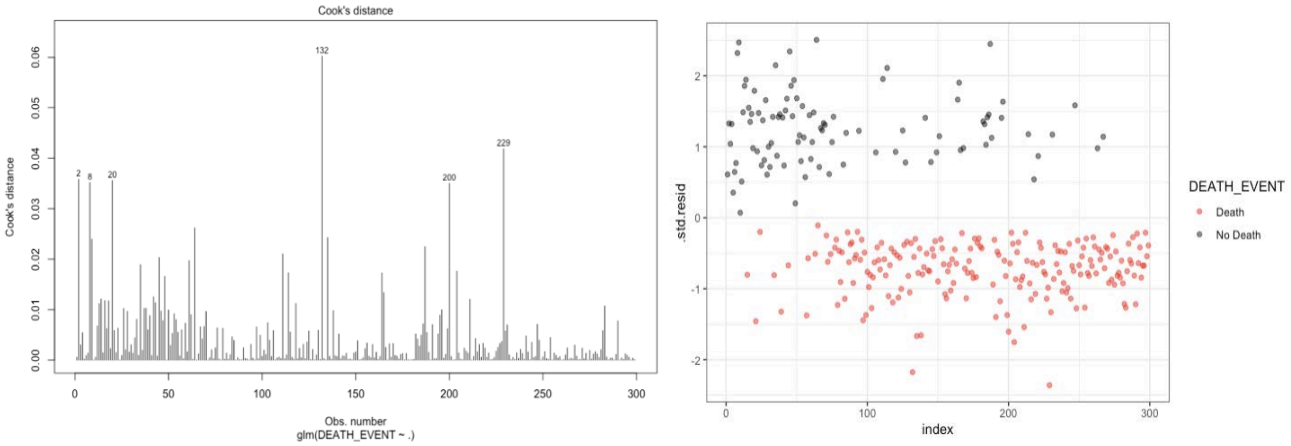

Fig 3: Top 6 outliers are highlighted in the Cook's distance plot (left). Not all outliers are influential observations. To check whether the data contains potential influential observations, the standardized residual error were inspected in the next scatterplot (right). Among these outliers, there were no standardized residuals above absolute value of 3, therefore, no influential observations in data.

4. No high intercorrelations (i.e. multicollinearity) among the predictors

```
                 age              anaemia creatinine_phosphokinase
            1.128444             1.091638                 1.112243
            diabetes    ejection_fraction  high_blood_pressure
            1.056803             1.122656                 1.059847
           platelets     serum_creatinine         serum_sodium
            1.057122             1.059148                 1.073572
                 sex              smoking
            1.377926             1.260903
```

Fig 4: No evidence of multicollinearity: all predictor variables have VIF values below 5

Prediction accuracy comparisons of supervised machine learning methods

Given that some assumptions were met and others were not, this logistic regression model may not have been the most accurate. The prediction accuracy of this model was compared to other non-parametric classification methods (KNN, RF, linear SVM, and radial SVM)[3] to assess which model can most accurately predict death outcome based on the same clinical features. To compare the accuracy of supervised machine learning methods, k-fold cross validation (k=10) was used.

```
Generalized Linear Model

299 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270, 269, 268, 269, 270, 269, ...
Resampling results:

  Accuracy   Kappa
  0.7386837  0.3416912
```

```
k-Nearest Neighbors

299 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270, 269, 268, 269, 270, 269, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.6453059  0.103823728
  7  0.6380942  0.055891920
  9  0.6385391  0.002623921

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

```
Random Forest

299 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270, 269, 268, 269, 270, 269, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7391435  0.3522128
   6    0.7321320  0.3621891
  11    0.7421394  0.3901636

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 11.
```

```
Support Vector Machines with Linear Kernel

299 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270, 269, 268, 269, 270, 269, ...
Resampling results:

  Accuracy   Kappa
  0.7656952  0.4104863

Tuning parameter 'C' was held constant at a value of 1
```

```
Support Vector Machines with Radial Basis Function Kernel

299 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270, 269, 268, 269, 270, 269, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.6856730  0.04799357
  0.50  0.7258027  0.26205241
  1.00  0.7422544  0.34253241

Tuning parameter 'sigma' was held constant at a value of 0.05858374
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.05858374 and C = 1.
```

Fig 5: Based on these cross validation results, linear SVM produced the highest accuracy at 0.7656952 (76.57%) with cost parameter held constant at 1.

Feature rankings

With the comparable accuracy of the RF model (0.7391435 for mtry = 2) to the logistic regression model (0.7386837) as well as its feature ranking techniques (mean Gini impurity reduction), the RF model was further employed to compare important clinical features to the ones that were detected by the logistic regression model earlier. Data was randomly split into 80% training (n= 239) and 20% testing (n= 60) to fit the training dataset in the RF model with 500 trees split at 2 randomly selected variables.

```
     Y.test
p.hat  0  1
    0 35 16
    1  1  8
```

```
                          MeanDecreaseGini
age                              13.986790
anaemia                           2.200725
creatinine_phosphokinase         13.006010
diabetes                          2.466730
ejection_fraction                14.821063
high_blood_pressure               2.127333
platelets                        12.464337
serum_creatinine                 17.297475
serum_sodium                     11.957433
sex                               2.281040
smoking                           2.321736
```

The confusion matrix is presented first, the RF model has a misclassification rate of 0.2833333. Overall, the feature rankings from the RF model confirmed the significant predictors from the logistic regression model's coefficients (age, CPK, ejection fraction, and serum creatinine) and provided insight into the other predictors (platelets and serum sodium) that may also be important in predicting a death event from heart failure.

Discussion

Results from this project show that it might be possible to predict the survival of patients with heart failure solely from the 11 predictor variables that describe a patient's clinical and lifestyle features. Due to different distributions and variations in these predictor variables, some assumptions of the logistic regression model were unmet.

Comparison of the model's prediction accuracy to other supervised machine learning models showed that linear SVM was the best model, at rate of 76.57% for prediction accuracy. The accuracy rate results were unexpected, as I would have expected the training sets of the more complex models (KNN, RF, and SVM) to have performed significantly better than the logistic regression model. However, KNN performed worse, and the other models performed only slightly better than the logistic regression model. This could have been due to relatively smaller sample sizes of the dataset that was randomly split into the training, testing, and validation sets in the cross-validation or perhaps the number of folds (k =10) that was chosen for the k-fold cross-validation. Further analysis of other non-parametric supervised methods as well as larger number of clinical records could provide more accurate comparisons.

In addition to classification comparisons, the feature ranking results between the logistic regression and RF model were relatively similar. In predicting a death event and the odds of a death event from heart failure, the clinical features age, CPK, ejection fraction, and serum creatinine were ranked important in both models, with the RF model including platelets and serum sodium. The limitations of the heart failure dataset should also be considered. There are many other features that were not included but nonetheless significant, such as weight, diet, exercise, occupation, etc. Further tests with random forest models of a larger (including records from other geographical regions) and expanded dataset could provide a more holistic approach and rank patients' clinical and/or lifestyle features more accurately.

Overall, the approach to modelling heart failure in this report showed that supervised machine learning can be used effectively for binary classification of electronic medical records of patients with cardiovascular hearth diseases.

All work was done in R Studio, .Rmd file code can be found at: https://github.com/jihyeonpak/ph245

References:

1. CDC. Heart Failure | cdc.gov. Centers for Disease Control and Prevention. Published September 8, 2020. Accessed December 12, 2021. https://www.cdc.gov/heartdisease/heart_failure.htm

2. Al'Aref SJ, Anchouche K, Singh G, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. 2019;40(24):1975-1986. doi:10.1093/eurheartj/ehy404

3. Li L. Introduction to Multivariate Statistics Lecture 8: Classification. Presented at: PH245: Art/Antho Practice Bldg 160; November 22, 2021; UC Berkeley.

4. Han H, Guo X, Yu H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. In: *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. ; 2016:219-224. doi:10.1109/ICSESS.2016.7883053

5. Li L. Introduction to Multivariate Statistics Lecture 5: Logistic Regression. Presented at: PH245: Art/Antho Practice Bldg 160; October 28, 2021; UC Berkeley.