Heart Failure analysis and modeling - report

Alexander Kheirallah

07/10/2020

Please see file `code.html` *to explore the code executed and outputs generated as part of data exploration, data pre-processing and model training.*

Background

This report documents my attempt to leverage **heart disease cleveland dataset** for the provision of predictive model for *angiographic disease status*, as well as identification of correlated features to aid in the understanding of disease pathogenesis and risk factors.

Methods

The challenge has been approached from two angles:

- 1. Development of optimal ML model
- 2. Statistical learning for identification of disease-correlated features

Data pre-processing

Justifications of data pre-processing can be found in `code.html`*.*

(1) ML model optimization

The following 3 steps have been utilized to train and optimize the disease prediction model:

- (I) Obtain a ranked list of important features (from most important to least important) through a random forest. For this step rely on function `xgboost()` from `XGboost` package.
- (II) To avoid over fitting, randomly split the data into train and test (I decided to give 40% of data to testing, which is a big proportion, but seems crucial to do as dataset size isn't that big and hence might be accidentally sampling 'outlier' cases).
- (III) To achieve an optimal model bias/variance trade off, do 13 iterations of logistic regression model training (`glm()` function of `stats` package) where at each loop incrementally add one feature at time, starting with most important feature and going down the list of ordered important features. After each training, test model's efficacy using ROC's *AUC* generated through predictions and actuals of **test** partition.

The number of features beyond which AUC starts to decrease suggests the optimal number of features that should be included in the model.

(2) Statistical learning

In order to determine the presence or absence of association between any feature and disease status, the probability of data given the null hypothesis that log-odds are 0 (i.e. coefficients of logistic regression) was calculated by calling `summary()` upon model object that was trained on all 13 features.

Results and Conclusions

Fig.1 below shows that the optimal number of features to include is 4. This is because adding more features to logistic regression model reduces AUC in the case of test partition that wasn't seen by the model. The AUC is constantly increasing for train partition, highlighting the random forest over fitting problem which was overcome by using test data.

Fig.2 shows the RUC curves for 4 features models, applying prediction on both train and test partitions. It was surprising to see that AUC was higher for test when using split seed 123 (which was also used throughout the loop-based model evaluation process). I hypothesised that this is due to a "lucky" random split of the data, given the small size of the dataset. Hence splitting was repeated with a different seed and which yielded an AUC lower for test relative to train validating my hypothesis.

Results of statistical analysis can be found in `code.html`. In summary, at a FDR level of 5%, features `thal`, `ca`, `oldpeak`, `slope`, `sex`, `cp4` (note the `4` here; other categories do not show the association) and `trestbps` associate with disease. In congruence with ML model development `thal`, `ca` and `cp` are among the top 4 most important features. However in disagreement, feature `age` which was among top best performing ones doesn't show evidence of association. Moreover, feature `sex` appears as strong risk factor with men having odds `exp(1.7)=5.4` times higher than women for having the disease; but `sex` showed to be among least important features.

All-in-all this analysis and training acts a starting point for possible model production. Future work would require the repeat of iterative model evaluation procedure explained above, using at least 3-4 different random train/test split (controlled by seed). We saw that this might be an issue when finding RUC for test higher than for train with seed 123. Finally results of statistical analysis, although make partial sense in light of model development, ought to be taken with caution as feature cross-correlation has not been tested and if present may have spuriously driven some of the reported results.

Fig.1:Train partition

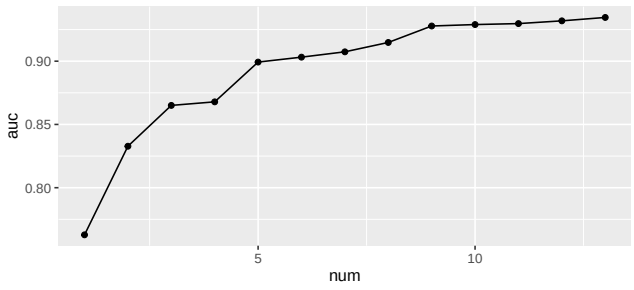

Fig.1:Test partition

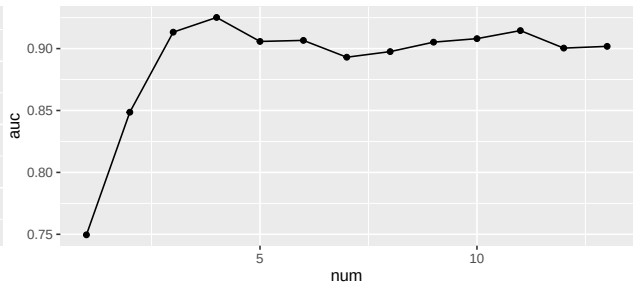

Fig.2:Seed 123/Train−blue,Test−red