REPORT 1

Maximum heart rate (HRmax) and maximum oxygen absorption (VO2max) are clinically valuable information and are known to be a great indicator for Cardioraspitory fitness (CRF). Cardiorespiratory fitness is a component of physiologic fitness and relates to the ability of the circulatory and respiratory systems to supply oxygen during sustained physical activity. A low CRF is one of the most important predictors of health outcomes. It is of high importance that we have a good estimate of what a »normal« CRF for an individual is since CRF decreases with age and is generally higher in men observants. For an individual, we would then know what the CRF for a person and his peers should be.

Today’s problem is that medical staff use simple linear models. These consist mostly of basic measurements (age, height, weight, etc.) which are not able to give accurate results.

A similar problem has already been addressed concerning the prediction of cardiovascular risk. In the study they used different machine learning models to get a significant improvement in accuracy. This resulted in and increasing number of patients who were identified early who then benefited from preventive treatment.

Machine learning has been booming for the last couple of years. But can it also help in prediction CRF? How good is it able to predict the two variables HRmax and VO2max to identify and prevent certain heart diseases?

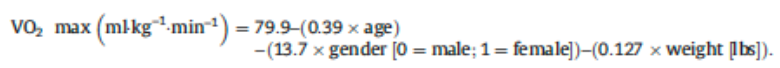
We are trying to find out just how much we can improve the prediction accuracy of our HRmax and VO2max target variables. Our plan is to use state-of-the-art Machine learning techniques to help improve the predictions. In this report I will present some techniques use to predict the two variables.

Firstly, we will have to clean the data and do extensive pre-processing of the given medical records. We will also use domain knowledge to make some feature selection among all the available variables. This will then help us create different machine learning models such as Lasso Regression, SVM, Neural Networks, Random Forests, etc.

Our expectations for machine learning predictions are high. We tend to drastically improve the prediction accuracy and this way help detect and prevent some upcoming diseases.

We are currently predicting VO2max with the following formulas:

1. Mayer’s formula from Friend registry



1. Wasserman’s formula

And HRmax with:

1. 220 – age
2. 209.2 − 0.72(age)

As we can see in the examples above, these formulas really do present a very simple linear regression solution. For our first test we also used a linear regression but with more variables and with L1 regularization to get a feel of which features are important. Before we could do that, we had to do some pre-processing.

We first got rid of current predictions that were present in the data set, so that it wouldn’t mess with our predictions later. We also deleted a few rows that did not contain any value for our target variables. We used one-hot encoding on some attributes to be able to use them as continuous variables. To get our first features we introduced domain knowledge. We checked all the value densities of all attributes and got an evaluation from the cardiology specialist on each of them.

Final attributes for building models was:

* Smoking (merged: @1rsSmoked, PackDay\_1.0, PackDay\_3.0, Smoked \_1.0 Smoked \_2.0),
* PhysInactive (merged: all PAStatus),
* Aritmia (merged: resting\_ecg, gxt\_arr),
* waist,
* bodyfat,
* fvc,
* fev1,
* AnyMeds AN do AY,
* All\_disease,
* resting\_dbp,
* resting\_sbp,
* resting\_hr,
* CR\_Code (one-hot encoded),
* ethnic\_code (dummies),
* height inch->cm,
* weight Ib->kg,
* BMI,
* ageattest,
* Gender\_Male

We still had to replace some missing values. We use a very simple approach since there were a lot of missing values in some attributes. We used mean value for true continuous variables and median value for one-hot encoded attributes. Finally, we did a minmax normalization and saved min and max values for later transformations.

Simple linear regression did not give any good results. We therefore used a L1 regularized or Lasso regression. This regression also gave a feature selection since it sets “unnecessary” attributes to 0.

**LASSO Regression for HRmax.** With a 10-fold cross validation our model found that the following attributes are important:

'Smoked', 'PhysInactive', 'AnyMeds AN do AY ', 'resting\_hr',

'weight Ib->kg', 'ageattest', 'Gender\_Male', 'CR\_Code\_1.0',

'CR\_Code\_2.0', 'CR\_Code\_3.0', 'CR\_Code\_4.0', 'ethnic\_code\_1.'

'ethnic\_code\_4.0'

With the following weights:

-3.79688263, 6.55353599, -2.4996638, 13.67687209, -4.30314275,

-41.97324887, -2.38872816, -2.95567776, 12.9042888, -3.0678952,

20.4823549, -5.54386238, -0.90631053

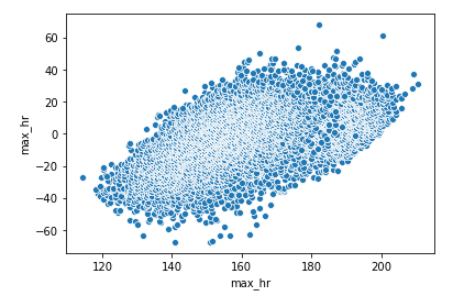
The L1 regression gave us the following results:

OVERALL RRMSE

0.497

OVERALL RMSE

14.362



A quick test was made with random forests just for a comparison. We used 250 trees and max tree

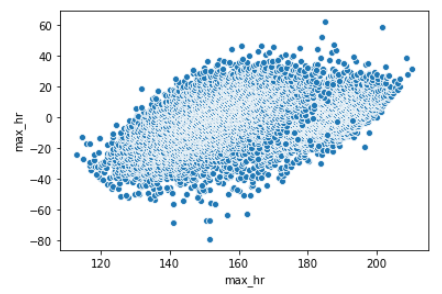
depth of 10.

OVERALL RRMSE

0.4165857036939675

OVERALL RMSE

13.151043250655091



All the test were made without any feature selection algorithms and are done on all the attributes.