REPORT 1

INTRODUCTION

Maximal heart rate (HRmax) and maximum oxygen consumption (VO2max) are clinically valuable information and are known to be a great indicator for cardiorespitory 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. CRF decreases with age and is generally higher in men observants, however it is infuenced by many other factors. It is of high importance that we have a good estimate of what a »normal« CRF for an individual is.

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 of Author different machine learning models were used to get a significant improvement in accuracy. This resulted in an increasing number of patients whose risk was identified early and who benefited then from preventive actions.

Machine learning has been booming for the last couple of years. To date, there is lack of information weather it can help in the prediction of CRF. Moreover, it is unknown how good it is able to predict the HRmax and VO2max, which is important in overall body assessment and disease prevention. Based on previous findings, o.

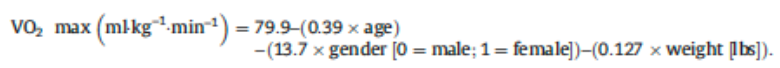
We hypothesized that state-of-the-art Machine learning techniques may help to improve the accuracy of HRmax and VO2max prediction. Accordingly, in the current study we applied several techniques to predict these two variables.

METHOD:

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

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 were:

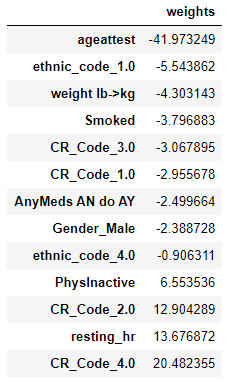
* Smoking status (merged: @1rsSmoked, PackDay\_1.0, PackDay\_3.0, Smoked \_1.0 Smoked \_2.0),
* PhysInactive (merged: all PAStatus),
* Arhythmia (merged: resting\_ecg, gxt\_arr),
* waist,
* bodyfat,
* fvc,
* fev1,
* AnyMeds AN do AY,
* All\_disease,
* resting\_dbp,
* resting\_sbp,
* resting\_hr,
* CR\_Code (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.

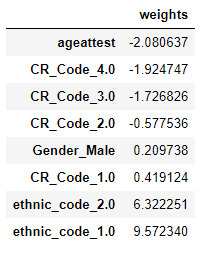
We used 2 different models to test the efficiency. The first one was Lasso, or L1 regularized Regression. Since Lasso tends to set the “unnecessary” attributes to 0 we can use it as features selection. We can see the Lasso selection on the two tables for HRmax and VO2max. The second model was a Random Forests model with 250 trees and max depth of each tree was equal to 10.

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

RESULTS

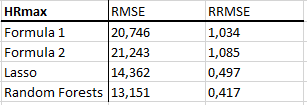
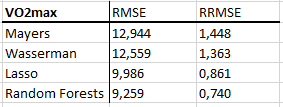


HRmax

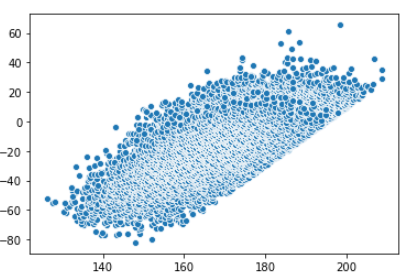
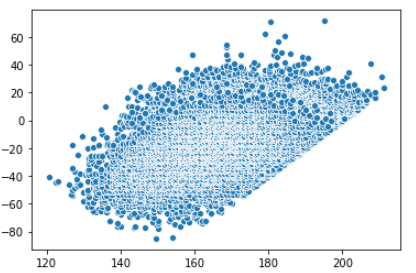


VO2max

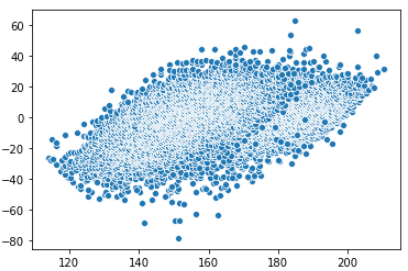
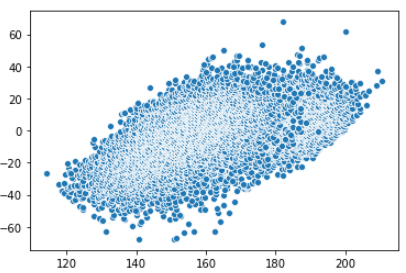
Our results compared to the baseline formulas are the following:



BLAND – ALTMAN plots for HRmax:

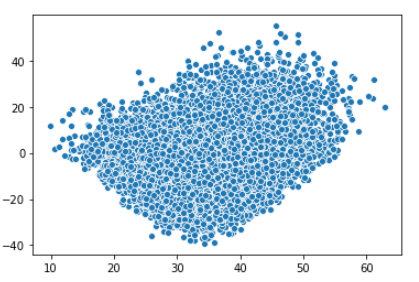
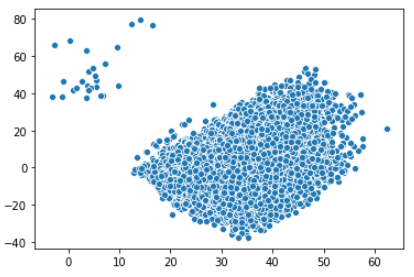


Formula1 Formula2

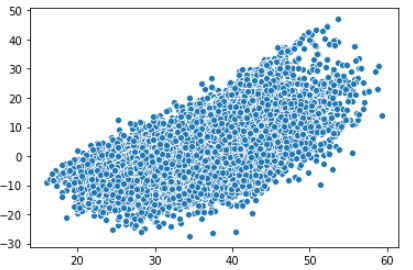
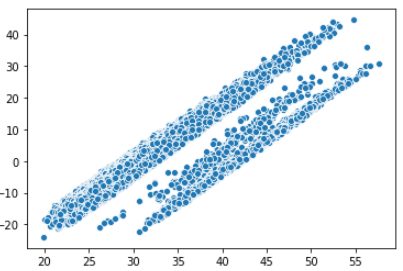


Lasso Random Forests

BLAND – ALTMAN plots for VO2max:

Mayers Wasserman



Lasso Random Forests