**Group model report**

1. Group variables:

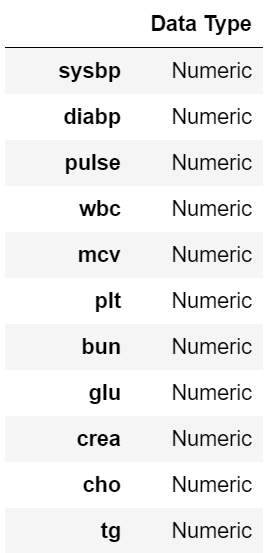
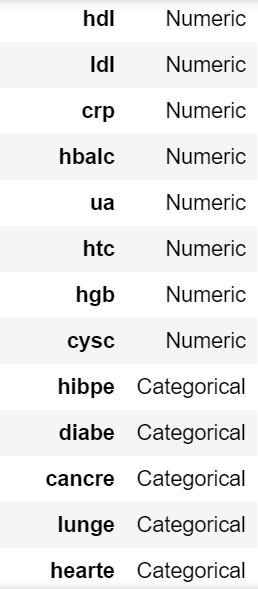
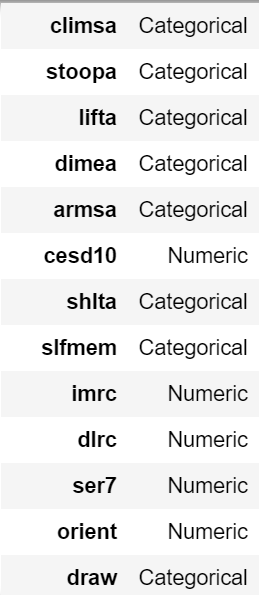
biomarkers = ['sysbp','diabp','pulse','wbc','mcv','plt','bun','glu','crea','cho','tg','hdl','ldl','crp','hbalc','ua','htc','hgb','cysc']

chronic disease = ['hibpe','diabe','cancre','lunge','hearte','stroke','psyche','arthre','dyslipe','livere','kidneye','digeste','asthmae','memrye']

self-reported functional limitation = ['dressa','batha','eata','beda','toilta','urina','moneya','medsa','shopa','mealsa','housewka','joga','walk1kma','walk100a','chaira','climsa','stoopa','lifta','dimea','armsa']

cognition/depression = ['cesd10','shlta','slfmem','imrc','dlrc','ser7','orient','draw'] #exclude tr20

2. Variable types:

With 2 ordinal variables:

'shlta' : ['Very Poor','Poor','Fair','Good','Very good'];

'slfmem' : ['Poor','Fair','Good','Very Good','Excellent']

3. Data filtering

2011+2015 with no age missing & 40<=age<=85, sample\_size=19695

Missing rate:

cysc ~13.1%

dlrc ~ 7.2%

imrc ~6.9%

After drop nan value from cysc, dlrc, imrc, sample\_size = 15834

4. Feature engineering & normalization

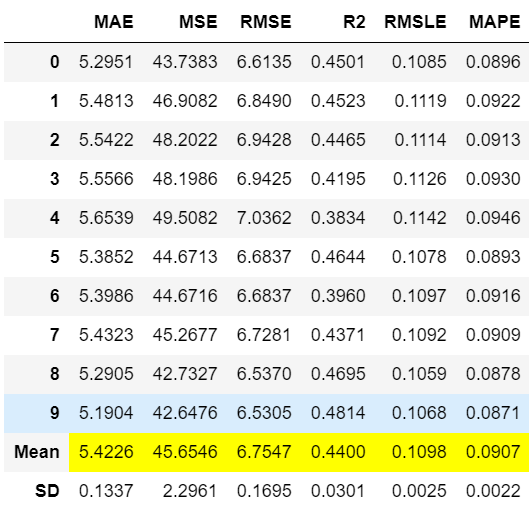
For all the numerical feature, implementing with mean value

For all the categorical feature, implementing with most frequent category

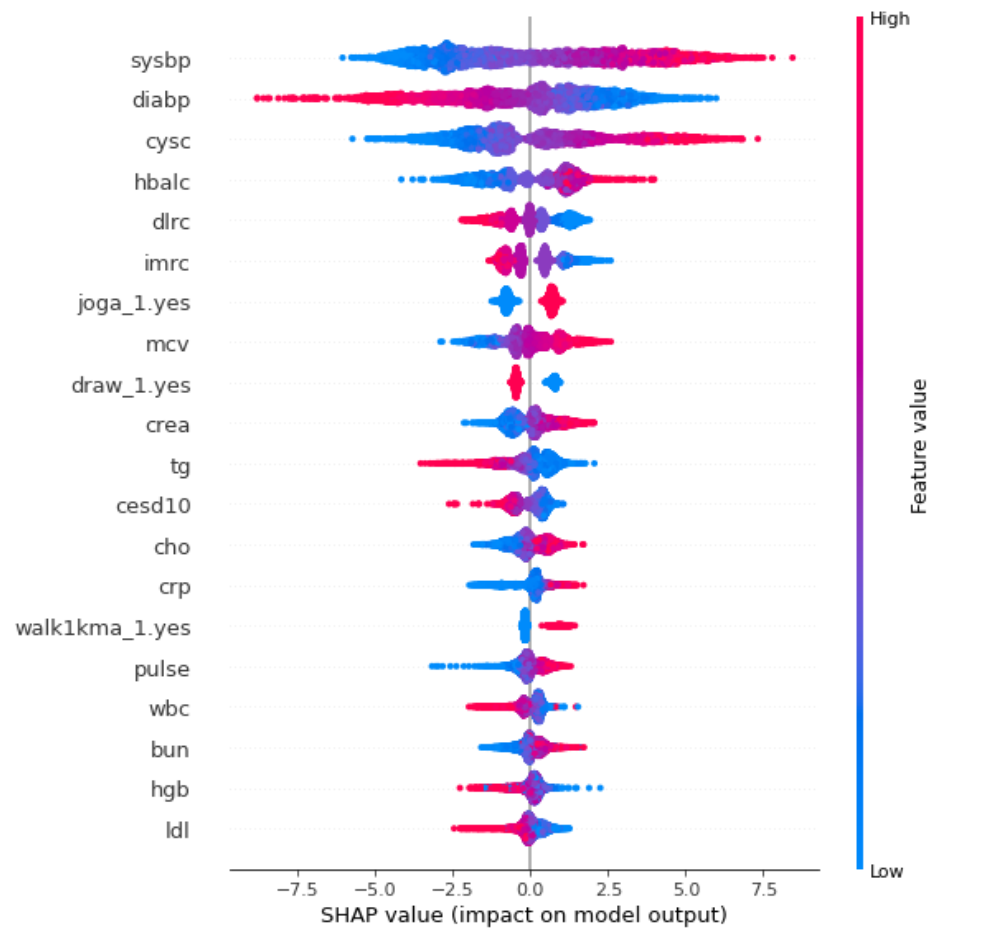
Normalization:min\_max

5. Output

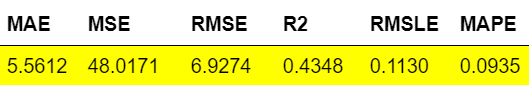
R^2: bio->0.3844 => all->0.44



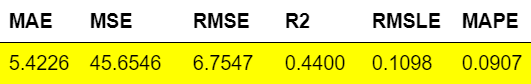
6. Feature importance



7. Model re-train with top feature-importance features



Compare to model with all variables:



Change of R^2 = 0.01, which indicates that features we choose can almost represent all the predict information of all the variables in the original model.

Features to focus: *['sysbp','diabp','cysc','hbalc','dlrc','imrc','joga','mcv','draw','crea','tg','cesd10','cho','crp','walk1kma','pulse','wbc','bun','hgb','ldl']*

Abs(pred - age) <=5, on 2011 dataset: 4179(<=5),2465(>5)