

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:

	Data Type				
		hdl	Numeric	stroke	Categorical
sysbp	Numeric	ldl	Numeric	psyche	Categorical
diabp	Numeric	crp	Numeric	arthre	Categorical
pulse	Numeric	hbalc	Numeric	dyslipe	Categorical
wbc	Numeric	ua	Numeric	livere	Categorical
mcv	Numeric	htc	Numeric	kidneye	Categorical
plt	Numeric	hgb	Numeric	digeste	Categorical
bun	Numeric	cysc	Numeric	asthmae	Categorical
glu	Numeric	hibpe	Categorical	memrye	Categorical
crea	Numeric	diabe	Categorical	dressa	Categorical
cho	Numeric	cancre	Categorical	batha	Categorical
tg	Numeric	lunge	Categorical	eata	Categorical
bede	Categorical	hearte	Categorical		
toilta	Categorical	climsa	Categorical		
urina	Categorical	stoopa	Categorical		
moneya	Categorical	lifta	Categorical		
medsa	Categorical	dimea	Categorical		
shopa	Categorical	armsa	Categorical		
mealsa	Categorical	cesd10	Numeric		
housewka	Categorical	shlta	Categorical		
joga	Categorical	slfmem	Categorical		
walk1kma	Categorical	imrc	Numeric		
walk100a	Categorical	dlrc	Numeric		
chaira	Categorical	ser7	Numeric		
		orient	Numeric		
		draw	Categorical		

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 & $40 \leq \text{age} \leq 85$, sample_size=19695

Missing rate:

cysc ~13.1%

dlrc ~ 7.2%

imrc ~6.9%

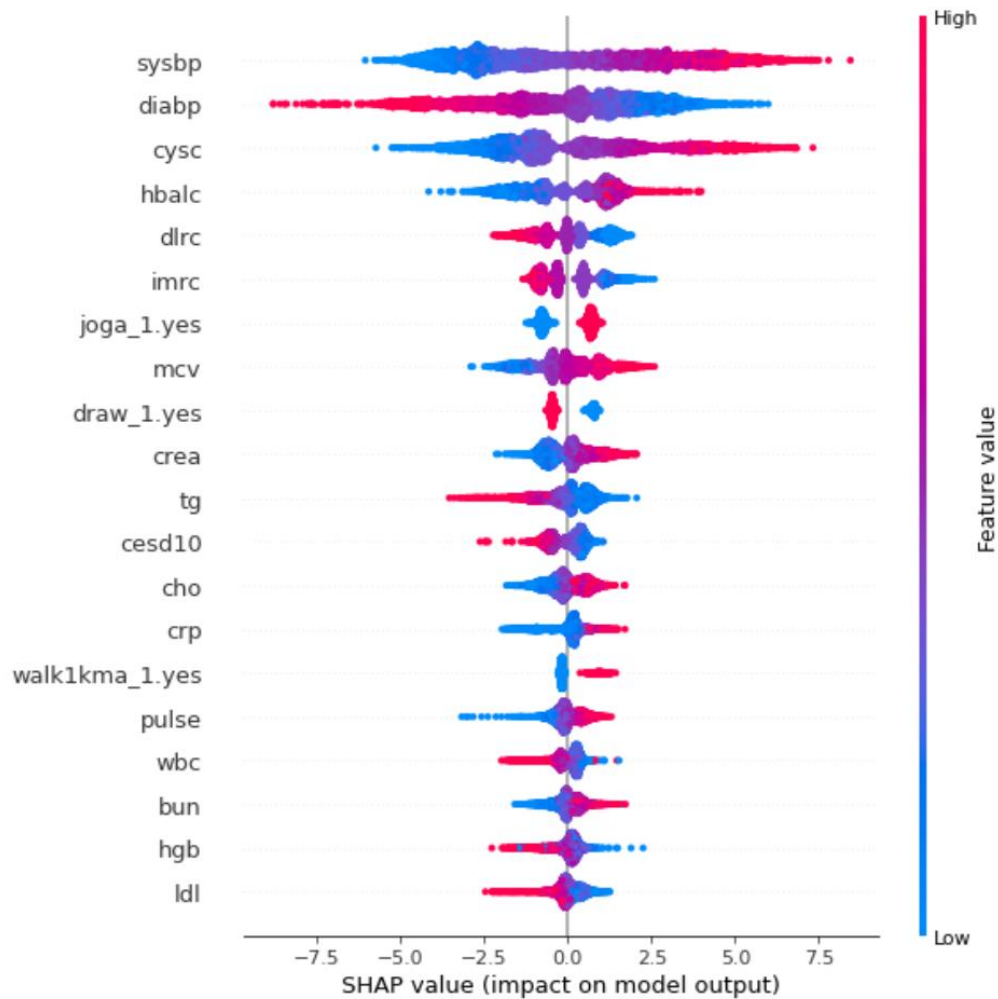
After drop nan value from cysc, dlrc, imrc, sample_size = 15834

4. Output

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

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	5.2951	43.7383	6.6135	0.4501	0.1085	0.0896
1	5.4813	46.9082	6.8490	0.4523	0.1119	0.0922
2	5.5422	48.2022	6.9428	0.4465	0.1114	0.0913
3	5.5566	48.1986	6.9425	0.4195	0.1126	0.0930
4	5.6539	49.5082	7.0362	0.3834	0.1142	0.0946
5	5.3852	44.6713	6.6837	0.4644	0.1078	0.0893
6	5.3986	44.6716	6.6837	0.3960	0.1097	0.0916
7	5.4323	45.2677	6.7281	0.4371	0.1092	0.0909
8	5.2905	42.7327	6.5370	0.4695	0.1059	0.0878
9	5.1904	42.6476	6.5305	0.4814	0.1068	0.0871
Mean	5.4226	45.6546	6.7547	0.4400	0.1098	0.0907
SD	0.1337	2.2961	0.1695	0.0301	0.0025	0.0022

5. Feature importance



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

MAE	MSE	RMSE	R2	RMSLE	MAPE
5.5612	48.0171	6.9274	0.4348	0.1130	0.0935

Compare to model with all variables:

MAE	MSE	RMSE	R2	RMSLE	MAPE
5.4226	45.6546	6.7547	0.4400	0.1098	0.0907

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','wa
lk1kma','pulse','wbc','bun','hgb','ldl']*

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