I think it makes sense to go with the blackboost results in the DTI set. They're quite significant and the variable distribution looks interesting.

If maximizing R2, then the kernelpls results would be the ones to use. The R2 is not too impressive though.

If the goal is to maximize subject instead, and go with the anatomical dataset, we could only use the conditional forest results because svmLinear isnot significant for inatt. And the R2 results, although significant, are quite pathetic.

I just left it computing the R2 for the blackboost models, just so we can report those as well if needed.

2020-03-21 07:51:58

Philip asked for more info on these models. First, what are the results for blackboost in the anatomy set?

```
> res[res$model=='blackboost' & res$nfolds==10, c(1:2, 4:6)]
       sx      model nfolds nreps meanRMSE
612 inatt blackboost     10    10 0.627078
613    hi blackboost     10    10 0.527887
614 inatt blackboost     10    10 0.562351
615    hi blackboost     10    10 0.460291
```

The top two are from the anatomy dataset. Can we select a result that works best across datasets as well?

```
params = c()
scores = c()
res =
read.csv('~/data/baseline_prediction/prs_start/residsFixed_slope_impInter.
csv', header=F)
colnames(res) = c('sx', 'model', 'fname', 'nfolds', 'nreps', 'meanRMSE',
'sdRMSE')
for (reg in unique(res$model)) {
    for (nf in unique(res$nfolds)) {
        for (nr in unique(res$nreps)) {
            idx = (res$model == reg &
                    res$nfolds == nf &
                    res$nreps == nr)
            pos = which(idx)
            if (length(pos) == 4) {
                my_str = paste(c(reg, nf, nr), collapse='_')
                params = c(params, my_str)
                scores = c(scores, mean(res[pos, 'meanRMSE']))
            }
        }
    }
}
```

```
a = sort(scores, decreasing=F, index.return=T)
print(params[a$ix[1]])
```

Here we get blassoAveraged_10_10:

```
       sx          model nfolds nreps meanRMSE
432 inatt blassoAveraged     10    10 0.623292
431    hi blassoAveraged     10    10 0.527534
533 inatt blassoAveraged     10    10 0.562877
537    hi blassoAveraged     10    10 0.459838
```

Again, top two are for anatomy. Like before, not that much difference between using blassoAveraged and blackboost. Maybe their variable importance will be different?

If we do the same thing for R2, we get:

```
params = c()
scores = c()
res =
read.csv('~/data/baseline_prediction/prs_start/residsR2_slope_impInter.csv
', header=F)
colnames(res) = c('sx', 'model', 'fname', 'nfolds', 'nreps',
'meanRsquared', 'sdRsquared')
for (reg in unique(res$model)) {
    for (nf in unique(res$nfolds)) {
        for (nr in unique(res$nreps)) {
            for (fn in unique(res$fname)) {
                idx = (res$model == reg &
                        res$nfolds == nf &
                        res$nreps == nr)
                pos = which(idx)
                if (length(pos) == 4) {
                    my_str = paste(c(reg, nf, nr), collapse='_')
                    params = c(params, my_str)
                    scores = c(scores, mean(res[pos, 'meanRsquared']))
                }
            }
        }
    }
}
a = sort(scores, decreasing=T, index.return=T)
print(params[a$ix[1]])
```

kernelpls does the best across datasets as well:

```
       sx     model nfolds nreps meanRsquared
316 inatt kernelpls     10    10     0.046883
```

```
310    hi kernelpls     10    10     0.074217
312 inatt kernelpls     10    10     0.081295
329    hi kernelpls     10    10     0.094364
```

Top two are anatomy, as usual. But they're not great, so maybe keep selecting based on RMSE?

Let's look at the updated Excel sheet and the variable importances there:

sx	model	metric	value	pval	dataset	notes
inatt	blassoAveraged	RMSE	0.562877	p<.001	DTI	counterpart to best RMSE; best across DTI+anat
hi	blassoAveraged	RMSE	0.459838	p<.001	DTI	best RMSE; best across DTI+anat
inatt	blassoAveraged	RMSE	0.623292	p<.001	anatomy	best across DTI+anat
hi	blassoAveraged	RMSE	0.527534	p<.001	anatomy	best across DTI+anat
inatt	blackboost	RMSE	0.562351	p<.001	DTI	best inatt RMSE; also best average sx results
hi	blackboost	RMSE	0.460291	p<.001	DTI	counterpart to best inatt RMSE
inatt	blackboost	RMSE	0.627078	p<.001	anatomy	best inatt RMSE; also best average sx results; added to compare to blassoAveraged
hi	blackboost	RMSE	0.527887	p<.001	anatomy	counterpart to best inatt RMSE; added to compare to blassoAveraged
inatt	svmLinear	RMSE	0.64174	p = 0.798795	anatomy	counterpart to best RMSE within anatomy
hi	svmLinear	RMSE	0.522771	p<.001	anatomy	best RMSE within anatomy
inatt	cforest	RMSE	0.619933	p<.001	anatomy	best inatt RMSE within anatomy; also best average sx result within anatomy
hi	cforest	RMSE	0.52578	p<.001	anatomy	counterpart to best inatt RMSE within anatomy
inatt	rvmLinear	R2	0.067602	p<.001	DTI	counterpart to best R2
hi	rvmLinear	R2	0.102126	p<.001	DTI	best R2
inatt	kernelpls	R2	0.081295	p<.001	DTI	best inatt R2; also best average sx results; best across DTI+anat
hi	kernelpls	R2	0.094364	p<.001	DTI	counterpart to best inatt R2; best across DTI+anat
inatt	kernelpls	R2	0.046883	p<.001	anatomy	best across DTI+anat
hi	kernelpls	R2	0.074217	p<.001	anatomy	best across DTI+anat
inatt	evtree	R2	0.067843	p<.001	anatomy	best inatt R2 within anatomy (best HI didn't work for inatt based on varimp)
hi	evtree	R2	0.063882	p<.001	anatomy	counterpart to best inatt R2 within anatomy

These are the variable importances for blassoAveraged, both datasets, both sx:

```
                       Overall
FSIQ_IR_165             100.00
PS_RAW_IR_165            71.41
cerebellumR_165          64.31
striatumR_165            61.78
unc_adR                  56.15
ADHD_PRS0.400000.origR   53.69
ADHD_PRS0.500000.origR   49.13
ADHD_PRS0.005000.origR   46.43
amygdalaR_165            46.16
OFCR_165                 44.94
lateral_PFCR_165         41.58
CC_ad_R                  41.45
slf_rdR                  41.05
ADHD_PRS0.010000.origR   39.29
cing_adR                 38.55
ilf_adR                  38.44
ADHD_PRS0.300000.origR   36.67
CC_rd_R                  35.80
unc_rdR                  33.27
SS_RAW_IR_165            32.81
Bayesian Ridge Regression (Model Averaged)

  RMSE       Rsquared    MAE
  0.5628775  0.06616144  0.4299283
```

```
[1]
"inatt,blassoAveraged,/home/sudregp/data/baseline_prediction/prs_start/gf_
impute_based_dti_165.csv,10,10,0.562877,NA"


                       Overall
striatumR_165           100.00
OFCR_165                 89.81
unc_adR                  86.02
amygdalaR_165            66.71
thalamusR_165            66.37
ADHD_PRS0.050000.origR   57.52
slf_rdR                  55.67
ilf_adR                  43.52
ADHD_PRS0.100000.origR   42.60
slf_adR                  35.79
cingulateR_165           34.90
ADHD_PRS0.200000.origR   34.64
CC_rd_R                  34.03
cing_rdR                 32.92
ADHD_PRS0.300000.origR   30.89
CST_rdR                  29.96
unc_rdR                  29.88
VMI.beery_RAW_IR         28.70
PS_RAW_IR_165            28.56
CC_ad_R                  27.51
Bayesian Ridge Regression (Model Averaged)

  RMSE       Rsquared   MAE
  0.4598384  0.0833015  0.3103396

[1]
"hi,blassoAveraged,/home/sudregp/data/baseline_prediction/prs_start/gf_imp
ute_based_dti_165.csv,10,10,0.459838,NA"


FSIQ_IR                        100.00
striatumR                       50.63
amygdalaR                       50.43
PS_RAW_IR                       44.79
OFCR                            41.07
ADHD_PRS0.000100.origR          36.09
ADHD_PRS0.000050.origR          35.40
ADHD_PRS0.500000.origR          29.34
ADHD_PRS0.001000.origR          28.15
ADHD_PRS0.005000.origR          27.78
ADHD_PRS0.400000.origR          27.50
EstimatedTotalIntraCranialVolR  26.43
VMI.beery_RAW_IR                23.84
thalamusR                       23.63
SS_RAW_IR                       23.04
lateral_PFCR                    21.53
ADHD_PRS0.300000.origR          19.91
cingulateR                      17.69
```

```
ADHD_PRS0.100000.origR           17.38
ADHD_PRS0.000500.origR           17.10
Bayesian Ridge Regression (Model Averaged)

  RMSE      Rsquared    MAE
  0.623292  0.05071923  0.4542848

[1]
"inatt,blassoAveraged,/home/sudregp/data/baseline_prediction/prs_start/gf_
impute_based_anatomy_272.csv,10,10,0.623292,NA"


OFCR                            100.00
amygdalaR                        93.35
striatumR                        91.60
ADHD_PRS0.000050.origR           74.13
VMI.beery_RAW_IR                 46.85
ADHD_PRS0.000100.origR           38.24
cingulateR                       34.17
thalamusR                        30.30
PS_RAW_IR                        30.14
DS_RAW_IR                        29.04
lateral_PFCR                     26.17
cerebellumR                      19.51
FSIQ_IR                          17.57
EstimatedTotalIntraCranialVolR   17.34
ADHD_PRS0.000500.origR           15.35
ADHD_PRS0.001000.origR           15.28
ADHD_PRS0.100000.origR           14.84
ADHD_PRS0.500000.origR           14.39
ADHD_PRS0.200000.origR           13.59
ADHD_PRS0.050000.origR           12.09
Bayesian Ridge Regression (Model Averaged)

  RMSE       Rsquared    MAE
  0.5275338  0.05700907  0.326785

[1]
"hi,blassoAveraged,/home/sudregp/data/baseline_prediction/prs_start/gf_imp
ute_based_anatomy_272.csv,10,10,0.527534,NA"
```

For comparison, these are the variable importances for blackboost, both datasets, both sx. They are both equally good models when selecting using RMSE:

```
                       Overall
FSIQ_IR_165             100.00
PS_RAW_IR_165            71.41
cerebellumR_165          64.31
striatumR_165            61.78
unc_adR                  56.15
ADHD_PRS0.400000.origR   53.69
```

```
ADHD_PRS0.500000.origR   49.13
ADHD_PRS0.005000.origR   46.43
amygdalaR_165            46.16
OFCR_165                 44.94
lateral_PFCR_165         41.58
CC_ad_R                  41.45
slf_rdR                  41.05
ADHD_PRS0.010000.origR   39.29
cing_adR                 38.55
ilf_adR                  38.44
ADHD_PRS0.300000.origR   36.67
CC_rd_R                  35.80
unc_rdR                  33.27
SS_RAW_IR_165            32.81
[1]
"inatt,blackboost,/home/sudregp/data/baseline_prediction/prs_start/gf_impu
te_based_dti_165.csv,10,10,0.562351,0.000000"

                       Overall
striatumR_165           100.00
OFCR_165                 89.81
unc_adR                  86.02
amygdalaR_165            66.71
thalamusR_165            66.37
ADHD_PRS0.050000.origR   57.52
slf_rdR                  55.67
ilf_adR                  43.52
ADHD_PRS0.100000.origR   42.60
slf_adR                  35.79
cingulateR_165           34.90
ADHD_PRS0.200000.origR   34.64
CC_rd_R                  34.03
cing_rdR                 32.92
ADHD_PRS0.300000.origR   30.89
CST_rdR                  29.96
unc_rdR                  29.88
VMI.beery_RAW_IR         28.70
PS_RAW_IR_165            28.56
CC_ad_R                  27.51
[1]
"hi,blackboost,/home/sudregp/data/baseline_prediction/prs_start/gf_impute_
based_dti_165.csv,10,10,0.460291,0.000000"

FSIQ_IR                         100.00
striatumR                        50.63
amygdalaR                        50.43
PS_RAW_IR                        44.79
OFCR                             41.07
ADHD_PRS0.000100.origR           36.09
ADHD_PRS0.000050.origR           35.40
ADHD_PRS0.500000.origR           29.34
ADHD_PRS0.001000.origR           28.15
ADHD_PRS0.005000.origR           27.78
ADHD_PRS0.400000.origR           27.50
```

```
EstimatedTotalIntraCranialVolR   26.43
VMI.beery_RAW_IR                 23.84
thalamusR                        23.63
SS_RAW_IR                        23.04
lateral_PFCR                     21.53
ADHD_PRS0.300000.origR           19.91
cingulateR                       17.69
ADHD_PRS0.100000.origR           17.38
ADHD_PRS0.000500.origR           17.10
[1]
"inatt,blackboost,/home/sudregp/data/baseline_prediction/prs_start/gf_impu
te_based_anatomy_272.csv,10,10,0.627078,0.000000"

                               Overall
OFCR                            100.00
amygdalaR                        93.35
striatumR                        91.60
ADHD_PRS0.000050.origR           74.13
VMI.beery_RAW_IR                 46.85
ADHD_PRS0.000100.origR           38.24
cingulateR                       34.17
thalamusR                        30.30
PS_RAW_IR                        30.14
DS_RAW_IR                        29.04
lateral_PFCR                     26.17
cerebellumR                      19.51
FSIQ_IR                          17.57
EstimatedTotalIntraCranialVolR   17.34
ADHD_PRS0.000500.origR           15.35
ADHD_PRS0.001000.origR           15.28
ADHD_PRS0.100000.origR           14.84
ADHD_PRS0.500000.origR           14.39
ADHD_PRS0.200000.origR           13.59
ADHD_PRS0.050000.origR           12.09
[1]
"hi,blackboost,/home/sudregp/data/baseline_prediction/prs_start/gf_impute_
based_anatomy_272.csv,10,10,0.527887,0.000000"
```

These are the variable importances for kernelpls, both datasets, both sx. This was the best model when selecting on R2:

```
striatumR               100.00
amygdalaR                99.19
SS_RAW_IR_165                       50.62
OFCR_165                            49.91
CC_ad_R                             47.37
ADHD_PRS0.050000.origR              43.59
ilf_adR                             42.87
cerebellumR_165                     42.25
FSIQ_IR_165                         41.56
CC_rd_R                             40.01
```

```
cing_adR                              36.62
ADHD_PRS0.000100.origR                36.59
ADHD_PRS0.000050.origR                31.97
ilf_rdR                               31.26
DS_RAW_IR_165                         26.80
ADHD_PRS0.010000.origR                25.71
cingulateR_165                        25.53
slf_rdR                               24.12
EstimatedTotalIntraCranialVolR_165    22.72
lateral_PFCR_165                      22.65
[1]
"inatt,kernelpls,/home/sudregp/data/baseline_prediction/prs_start/gf_imput
e_based_dti_165.csv,10,10,0.081295,0.012537"

striatumR               100.00
amygdalaR                99.19
amygdalaR_165                         87.43
VMI.beery_RAW_IR                      83.69
cingulateR_165                        79.42
slf_adR                               72.98
slf_rdR                               68.00
SES_group3_165                        55.47
ilf_adR                               44.27
ADHD_PRS0.050000.origR                44.16
OFCR_165                              41.28
ADHD_PRS0.000500.origR                40.54
cerebellumR_165                       38.57
EstimatedTotalIntraCranialVolR_165    38.20
ilf_rdR                               36.74
ADHD_PRS0.100000.origR                34.85
DS_RAW_IR_165                         34.81
ADHD_PRS0.000100.origR                33.01
ADHD_PRS0.500000.origR                29.77
thalamusR_165                         29.49
"hi,kernelpls,/home/sudregp/data/baseline_prediction/prs_start/gf_impute_b
ased_dti_165.csv,10,10,0.094364,0.011303"

striatumR               100.00
amygdalaR                99.19
FSIQ_IR                         83.67
SS_RAW_IR                       77.43
PS_RAW_IR                       69.79
striatumR                       61.05
ADHD_PRS0.000100.origR          52.23
cerebellumR                     51.80
DS_RAW_IR                       48.91
ADHD_PRS0.000050.origR          45.28
OFCR                            44.21
cingulateR                      42.08
VMI.beery_RAW_IR                41.92
EstimatedTotalIntraCranialVolR  41.20
ADHD_PRS0.400000.origR          40.66
ADHD_PRS0.005000.origR          39.54
ADHD_PRS0.001000.origR          37.23
```

```
ADHD_PRS0.300000.origR          36.29
SES_group3                      34.63
thalamusR                       31.70
"inatt,kernelpls,/home/sudregp/data/baseline_prediction/prs_start/gf_imput
e_based_anatomy_272.csv,10,10,0.046883,0.003794"

striatumR               100.00
amygdalaR                99.19
OFCR                     94.19
VMI.beery_RAW_IR         74.23
ADHD_PRS0.000050.origR   67.05
SES_group3               61.50
lateral_PFCR             48.15
cingulateR               40.98
FSIQ_IR                  36.85
SS_RAW_IR                31.16
ADHD_PRS0.000500.origR   28.87
ADHD_PRS0.010000.origR   22.05
ADHD_PRS0.200000.origR   20.17
ADHD_PRS0.300000.origR   19.48
ADHD_PRS0.400000.origR   18.69
ADHD_PRS0.005000.origR   18.00
ADHD_PRS0.500000.origR   17.52
ADHD_PRS0.100000.origR   16.95
ADHD_PRS0.050000.origR   15.37
thalamusR                15.29
[1]
"hi,kernelpls,/home/sudregp/data/baseline_prediction/prs_start/gf_impute_b
ased_anatomy_272.csv,10,10,0.074217,0.011465"
```