WE R NOT OKAY

BEN BRILL, EMMA CHI, CECILE WU

Lecture 3

INTRODUCTION

- Explore tuning parameters of random forest in order to collect data
- Goal is to find the most important tuning parameters of random forest that can best facilitate cross-validation accuracy,
- Two subsets in order to compare the two: one builds a random forest; the other evaluates the algorithm.
- Cardiovascular disease data
- Build a design for the tuning parameters using different techniques and deciding on one for cross-validation.

02 + 03

METHODOLOGY AND RESULTS

design + analysis process and
final model + evaluation

METHODOLOGY

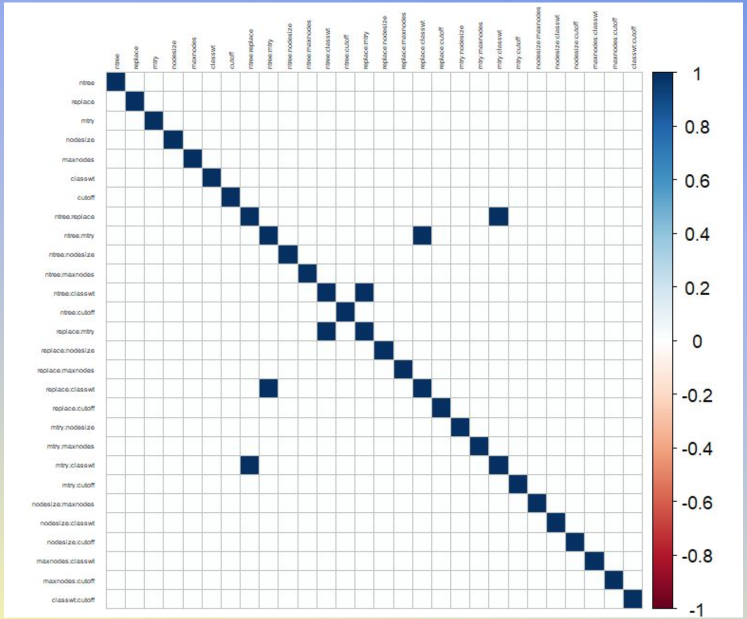
- Experimental Design
  - Fractional factorial design + experimental design using optimal design approach → 7 factors
  - 7 total factors → 2^(7-2) design with Resolution IV
  - No main effects will be aliased
  - Fractional factorial design: FrF2 function, 32 runs using the 7 factors
  - Center points to address large ranges
  - Optimal design: optFederov function, same 7 factors, 35 trials, 100 repeats

run	ntree	mtry	replace	nodesize	classwt	cutoff	maxnodes
1	−1	−1	−1	−1	−1	−1	1
2	1	−1	−1	−1	−1	1	−1
3	−1	1	−1	−1	−1	1	−1
4	1	1	−1	−1	−1	−1	1
5	−1	−1	1	−1	−1	1	1
6	1	−1	1	−1	−1	−1	−1
7	−1	1	1	−1	−1	−1	−1
8	1	1	1	−1	−1	1	1
9	−1	−1	−1	1	−1	−1	−1
10	1	−1	−1	1	−1	1	1
11	−1	1	−1	1	−1	1	1
12	1	1	−1	1	−1	−1	−1
13	−1	−1	1	1	−1	1	−1
14	1	−1	1	1	−1	−1	1
15	−1	1	1	1	−1	−1	1
16	1	1	1	1	−1	1	−1
17	−1	−1	−1	−1	1	−1	−1
18	1	−1	−1	−1	1	1	1
19	−1	1	−1	−1	1	1	1
20	1	1	−1	−1	1	−1	−1
21	−1	−1	1	−1	1	1	−1
22	1	−1	1	−1	1	−1	1
23	−1	1	1	−1	1	−1	1
24	1	1	1	−1	1	1	−1
25	−1	−1	−1	1	1	−1	1
26	1	−1	−1	1	1	1	−1
27	−1	1	−1	1	1	1	−1
28	1	1	−1	1	1	−1	1
29	−1	−1	1	1	1	1	1
30	1	−1	1	1	1	−1	−1
31	−1	1	1	1	1	−1	−1
32	1	1	1	1	1	1	1

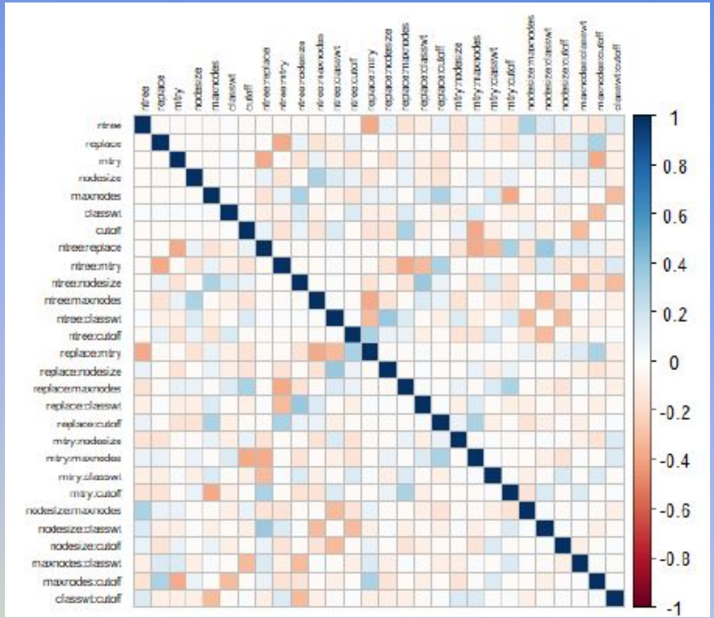
run	ntree	mtry	replace	nodesize	classwt	cutoff	maxnodes
2	1	−1	−1	−1	−1	−1	−1
4	1	1	−1	−1	−1	−1	−1
8	1	1	1	−1	−1	−1	−1
9	−1	−1	−1	1	−1	−1	−1
11	−1	1	−1	1	−1	−1	−1
15	−1	1	1	1	−1	−1	−1
17	−1	−1	−1	−1	1	−1	−1
22	1	−1	1	−1	1	−1	−1
30	1	−1	1	1	1	−1	−1
39	−1	1	1	−1	−1	1	−1
41	−1	−1	−1	1	−1	1	−1
44	1	1	−1	1	−1	1	−1
49	−1	−1	−1	−1	1	1	−1
54	1	−1	1	−1	1	1	−1
55	−1	1	1	−1	1	1	−1
60	1	1	−1	1	1	1	−1
62	1	−1	1	1	1	1	−1
63	−1	1	1	1	1	1	−1
69	−1	−1	1	−1	−1	−1	1
72	1	1	1	−1	−1	−1	1
77	−1	−1	1	1	−1	−1	1
81	−1	−1	−1	−1	1	−1	1
83	−1	1	−1	−1	1	−1	1
90	1	−1	−1	1	1	−1	1
92	1	1	−1	1	1	−1	1
95	−1	1	1	1	1	−1	1
98	1	−1	−1	−1	−1	1	1
99	−1	1	−1	−1	−1	1	1
102	1	−1	1	−1	−1	1	1
108	1	1	−1	1	−1	1	1
109	−1	−1	1	1	−1	1	1
110	1	−1	1	1	−1	1	1
116	1	1	−1	−1	1	1	1
119	−1	1	1	−1	1	1	1
121	−1	−1	−1	1	1	1	1

METHODOLOGY: COMPARISON OF FRACTIONAL FACTORIAL DESIGN AND OPTIMAL DESIGN

FRACTIONAL FACTORIAL



OPTIMAL

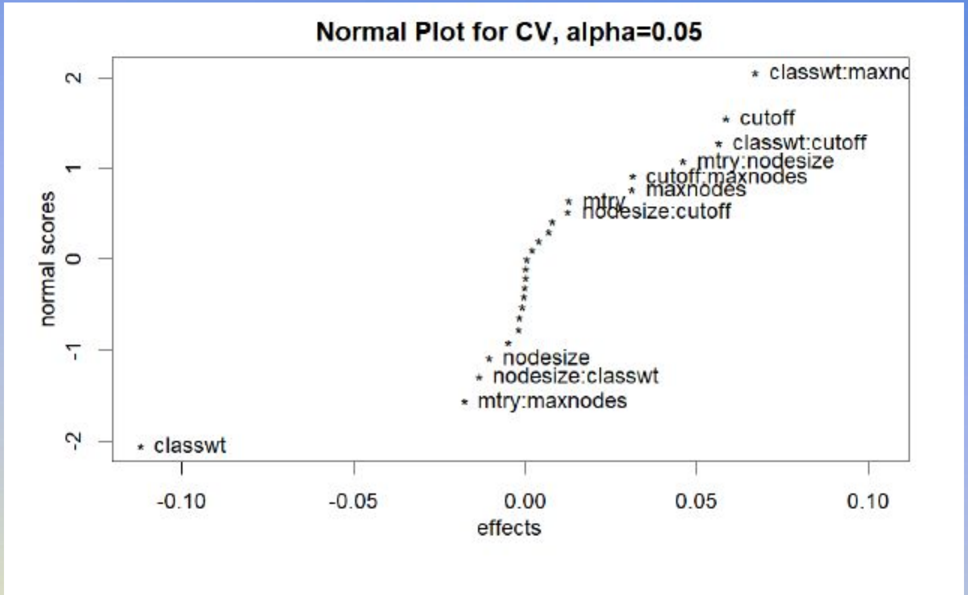


METHODOLOGY: COMPARISON OF FRACTIONAL FACTORIAL DESIGN AND OPTIMAL DESIGN

```
     (Intercept)            ntree           replace              mtry          nodesize
        1.178323         2.741298          3.037319          2.648838          1.934686
        maxnodes          classwt            cutoff     ntree:replace        ntree:mtry
        2.564909         1.660915          3.200686          3.073128          5.232742
  ntree:nodesize   ntree:maxnodes     ntree:classwt      ntree:cutoff      replace:mtry
        2.823134         4.998752          2.220494          3.261080          4.804835
replace:nodesize replace:maxnodes   replace:classwt    replace:cutoff     mtry:nodesize
        2.402817         2.317594          2.176831          3.102789          2.330797
   mtry:maxnodes     mtry:classwt       mtry:cutoff nodesize:maxnodes  nodesize:classwt
        4.170552         1.955657          2.253476          2.007650          2.264969
 nodesize:cutoff maxnodes:classwt   maxnodes:cutoff    classwt:cutoff
        2.259380         1.927194          3.905456          1.652216
```

RESULTS

- Used Half Normal plots to determine significant parameters: classwt, mtry:maxnodes, nodesize:classwt, nodesize, nodesize:cutoff, mtry, maxnodes, cutoff:maxnodes, mtry:nodesize, classwt:cutoff, cutoff, classwt:maxnodes
- We wanted to consider the more significant factors so...



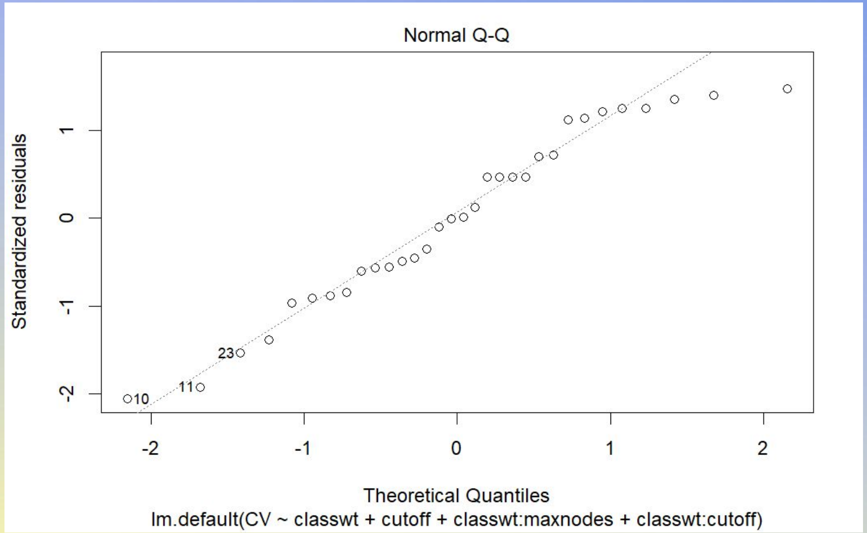
RESULTS

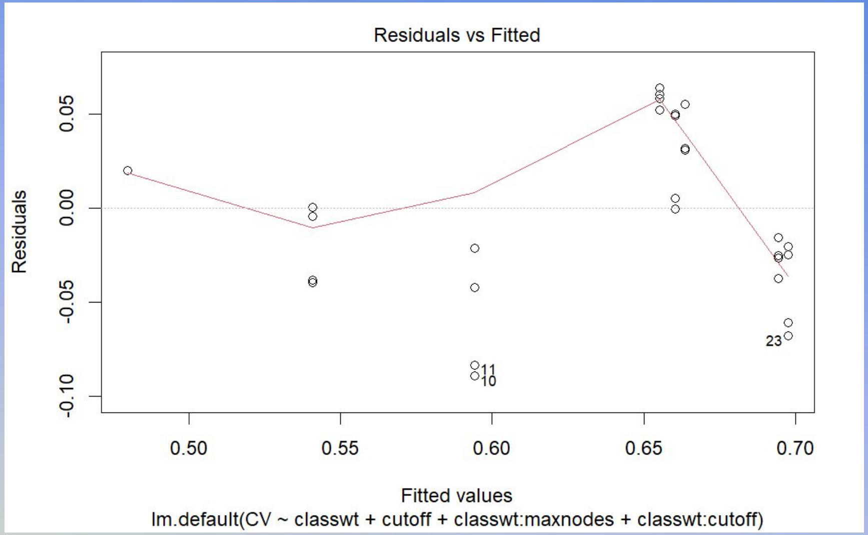
- **We chose classwt, cutoff, classwt:maxnodes, and classwt:cutoff**
- This was based off of numerous model testings with combinations of several significant factors. We ended up with these four when we eliminated those that were not significant when running the summary function.
- The final model is:

$$CV = 0.9319 - 0.5456(\text{classwt}) - 0.2270(\text{cutoff})$$
$$- 6.867\text{x}10^{-5}(\text{classwt:maxnodes}) - 0.4642(\text{classwt:cutoff})$$

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        9.319e-01  5.942e-02  15.684 4.35e-15 ***
classwt           -5.456e-01  8.246e-02  -6.616 4.25e-07 ***
cutoff            -2.270e-01  1.019e-01  -2.228  0.03444 *  
classwt:maxnodes   6.867e-05  2.330e-05   2.947  0.00654 ** 
classwt:cutoff     4.642e-01  1.400e-01   3.316  0.00261 ** 
```

RESULTS: RESIDUAL ANALYSIS





04

CONCLUSIONS

recommendations and strengths

CONCLUSIONS

STRENGTHS

- We added center points to our experiment to address some of the larger quantitative ranges
- We made sure our residual plot did not have any clumping or obvious patterns
- Our factors are significant
- Inspected multicollinearity via several methods

RECOMMENDATIONS

- Keep the classwt and the cutoff levels lower because the coefficients were negative
- Many of our factors were significant based on the half normal plot. The two-factor interactions included all main effects, however we wanted to keep the recommendation simpler. There may be more to consider in the future
- Make some more factors three levels (mixed-level design) to account for larger ranges

THANK YOU FOR WATCHING!