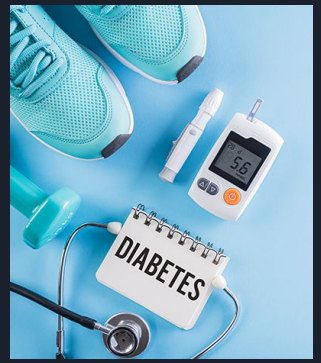
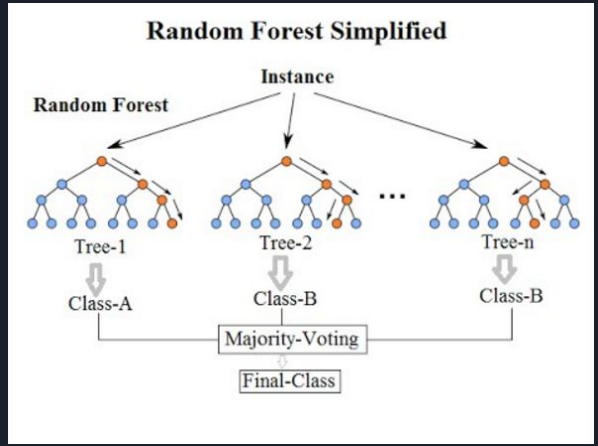
Random Forest Cross Validation Optimization

By Cara Drake, Stacy Deng, & Christopher Thornton

Introduction



- Affects over 11 % of our planet's population
- Nearly 40 % of adult US population has prediabetes
- Analyzing this data set to know which random forest tuning parameters are most essential to correctly predicting the binary data
- Random Forest:
  - Either or choice -> Decision Tree -> Random Forest
- There are seven factors involved in the Random Forest
  - Which carry the most weight in the decision making process



Methodology

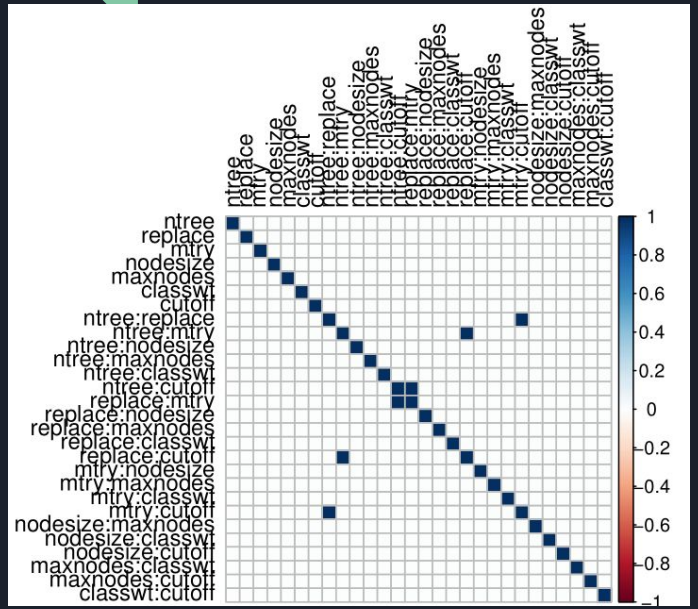
Factorial Fraction Design

- Number of runs must be a factor of 2, thus 32 trials
- Resolution IV because seven variables are used

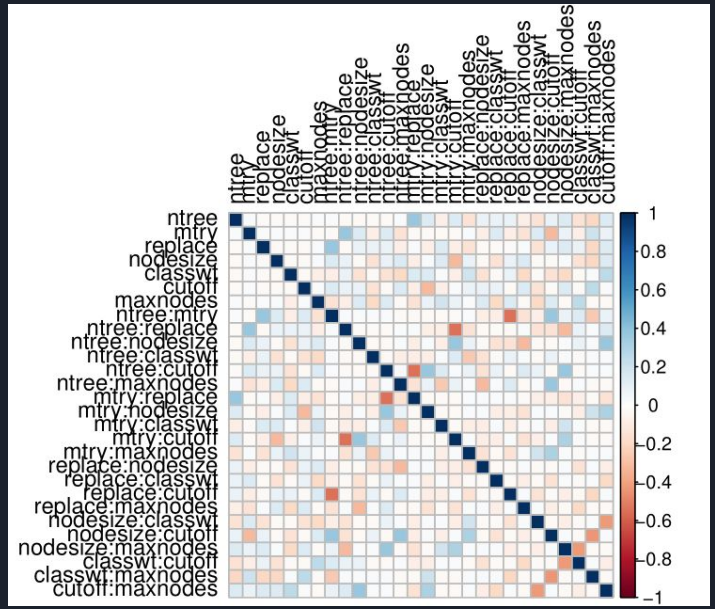
Optimal Design

- Only requires 29 trials for a resolution V model
- Can use all two parameter interactions

Design choice



Fractional Factorial Design



Optimal Design

Initial Model & Simplification

- Initial model contains all interaction terms between the seven terms of two parameter or less
- Simplifying based on p value and correlation plots

Analysis of Variance Table

Response: CV

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ntree	1	0.000003	0.000003	0.0029	0.958656
mtry	1	0.000587	0.000587	0.6465	0.452040
replace	1	0.000296	0.000296	0.3256	0.588989
nodesize	1	0.001145	0.001145	1.2605	0.304464
classwt	1	0.122507	0.122507	134.8095	2.457e-05
cutoff	1	0.031043	0.031043	34.1607	0.001106
maxnodes	1	0.015311	0.015311	16.8485	0.006324
ntree:mtry	1	0.000018	0.000018	0.0193	0.893972
ntree:replace	1	0.000032	0.000032	0.0354	0.856881
ntree:nodesize	1	0.006771	0.006771	7.4511	0.034201
ntree:classwt	1	0.001623	0.001623	1.7861	0.229845
ntree:cutoff	1	0.000401	0.000401	0.4411	0.531269
ntree:maxnodes	1	0.000048	0.000048	0.0532	0.825266
mtry:replace	1	0.000714	0.000714	0.7852	0.409655
mtry:nodesize	1	0.011081	0.011081	12.1933	0.012954
mtry:classwt	1	0.000535	0.000535	0.5885	0.472099
mtry:cutoff	1	0.000094	0.000094	0.1030	0.759090
mtry:maxnodes	1	0.002227	0.002227	2.4502	0.168546
replace:nodesize	1	0.003645	0.003645	4.0111	0.092071
replace:classwt	1	0.000000	0.000000	0.0000	0.999665
replace:cutoff	1	0.003344	0.003344	3.6794	0.103529
replace:maxnodes	1	0.000007	0.000007	0.0078	0.932671
nodesize:classwt	1	0.002260	0.002260	2.4864	0.165905
nodesize:cutoff	1	0.013343	0.013343	14.6825	0.008641
nodesize:maxnodes	1	0.000846	0.000846	0.9311	0.371843
classwt:cutoff	1	0.010608	0.010608	11.6733	0.014202
classwt:maxnodes	1	0.005914	0.005914	6.5083	0.043423
cutoff:maxnodes	1	0.004722	0.004722	5.1961	0.062844
Residuals	6	0.005452	0.000909		

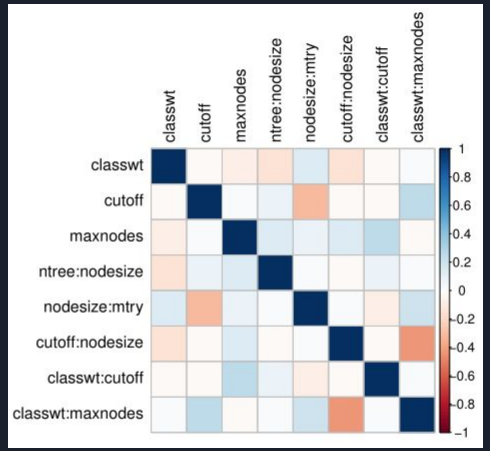
Analysis

Cross Validation Accuracy = .618271 − .057672 × classwt + .027108 × cutoff + .014352 × maxnodes + .012185 × ntree × nodesize +.011742 × nodesize × mtry − .007203 × cutoff × nodesize + .019271 × classwt × cutoff + .022981 × classwt × maxnodes

Analysis of Variance Table

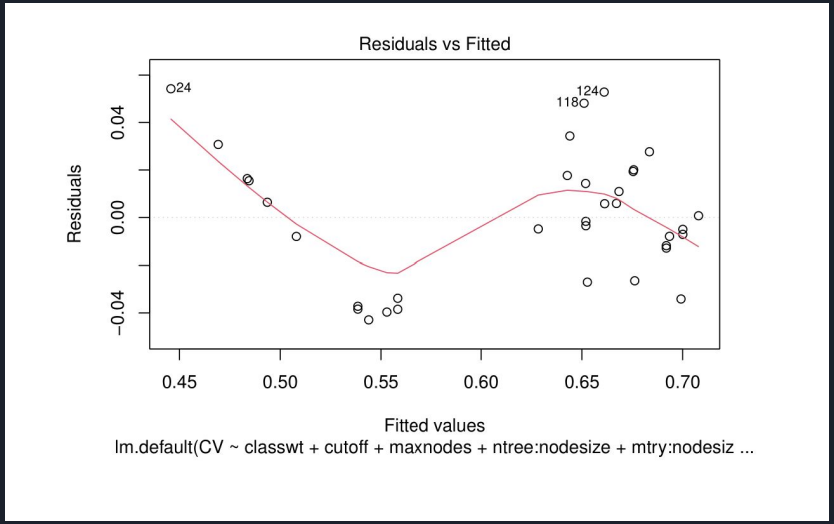
Response: CV

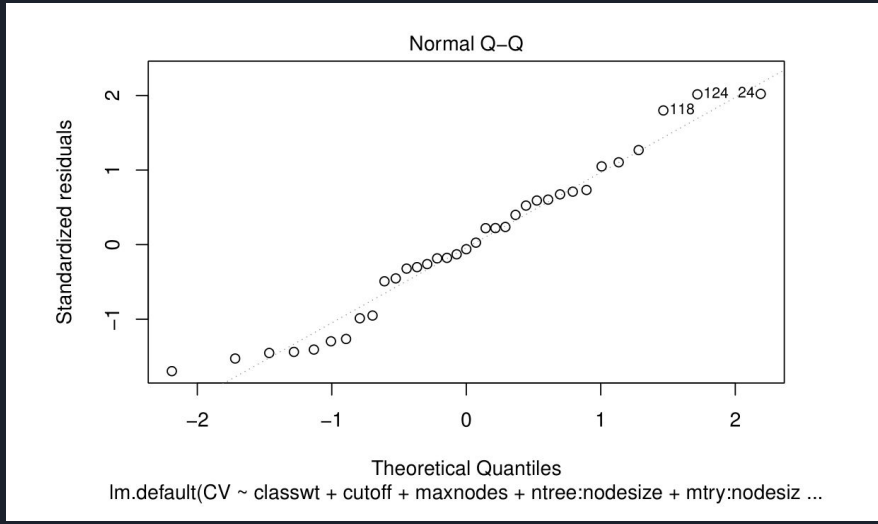
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
classwt	1	0.120431	0.120431	125.1900	1.945e-11
cutoff	1	0.032252	0.032252	33.5263	4.237e-06
maxnodes	1	0.014576	0.014576	15.1521	0.0006183
ntree:nodesize	1	0.007270	0.007270	7.5569	0.0107258
nodesize:mtry	1	0.008011	0.008011	8.3272	0.0077534
cutoff:nodesize	1	0.011599	0.011599	12.0570	0.0018196
classwt:cutoff	1	0.013521	0.013521	14.0554	0.0008964
classwt:maxnodes	1	0.011905	0.011905	12.3757	0.0016207
Residuals	26	0.025012	0.000962		



Conclusion

- Resolution V
  - Evaluates all main effects and first order interactions
- Model meets assumptions for linear regression
- Adjusted R-squared of .87





Retrospective

- Recommend full factorial design in future experimentation given sufficient funding
- Experiment with additional levels
- Analyze the effects of alternative data sets

References

Centers for Disease Control and Prevention. (2022, January 18). National Diabetes Statistics Report. Centers for Disease Control and Prevention. Retrieved June 4, 2022, from https://www.cdc.gov/diabetes/data/statistics-report/index.html

Cha, M., & Vazquez, A. (2022). Chapter 8: Two-Level Fractional Factorial Designs. Stats 101B.

MathWorks. (n.d.). Select a web site. D-Optimal Designs - MATLAB & Simulink. Retrieved June 4, 2022, from https://www.mathworks.com/help/stats/d-optimal-designs.html#:~:text=D%2Doptimal%20designs%20are%20model,estimates%20for%20a%20specified%20model

Montgomery, D. C. (2012). Design and analysis of Experiments. Wiley.

Natoli, C. (2018). Classical designs: Fractional factorial designs - 2019. afit.edu. Retrieved June 5, 2022, from https://www.afit.edu/stat/statcoe_files/Classical%20Designs-Fractional%20Fractorial%20Designs%20Rev1.pdf

Photo citation: https://www.cdc.gov/diabetes/images/library/spotlights/diabetes-stats-report-724px.png?_=42420

https://upload.wikimedia.org/wikipedia/commons/7/76/Random_forest_diagram_complete.png