UCLA STATS 101B

Final Project: Random Forest Cross Validation Optimization

Group CCS: Cara Drake, Stacy Deng, and Christopher Thornton

1. Introduction

The goal of our project was to determine which of the seven tuning parameters of random forest - ntree, replace, mtry, nodesize, maxnodes, classwt, and cutoff - affect the cross validation accuracy. We considered two types of experimental designs: a fractional factorial design and a D-optimal design. To help find the cross validation accuracy, the auxiliary dataset assigned to us is the “diabetes.RData” dataset, which was used to train a random forest.

2. Methodology

Part 1: Experimental Design

Question 1. *Propose a fractional factorial design for the problem. In addition, propose an experimental design constructed using the optimal design approach.*

The experiment is limited to only 35 trials for testing due to constraints in budget. Due to the limited number of trials, a full factorial experiment is not a viable option for this set of data since we would need n^k trials, where n is the number of levels per factor and k is the number of factors, which would be 2^7 for this experiment. Fractional factorial design is a reduced model of that full design, where some values become aliases of each other. So for the fractional factorial design we need to do a number of runs less than or equal to 35 that are a multiple of two. This gives the largest number of levels available for the fractional factorial design would be 32. There are seven factors for consideration: ntree, mtry, replace, nodesize, classwt, cutoff, and maxnodes. This gives **Figure 3** when we do a random run order for these 32 levels. To create this matrix, FrF2 was used with the parameters 32 for the nruns, for the number of levels, nfactors of 7 for the number of main factors being used, and randomize was set to true so that the order of runs were randomized. Since this experiment will use 32 trials and all the factors are being used at two levels, the exponential value will be 2^5, or more often written as 2^(7-2) since there are seven main parameters which are being computed.

When the number of trials available is limited, a different type of design which optimizes the trials for a model is optimal design. D-optimal design is an alternative method used with traditional data to find the minimal covariance between different factors in a model. The optimal design approach can use all 35 levels even with the seven factors and despite that it is not a multiple of two. **Figure 1** gives the optimal design approach. To create this matrix in R, the use of the optFederov function was employed, with 35 nTrials for the different levels, and 1000 nRepeats for how many times this process will be repeated for randomization.

Question 2. *Compare the optimal design with the fractional factorial design in practical and statistical terms. For instance, what is the performance of the designs for studying the main effects of the tuning parameters only? Can they estimate all two-parameter interactions? Why or why not? How do they compare in terms of multicollinearity?*

For the fractional factorial design, all the factors have two levels we are investigating. This means that since we have 7 factors, it will be a 2^(7 -x) = 32 experiment. 2^5 = 32 so our model with 32 trials is 2^(7-2). This is because instead of running 2^7 trials needed for the full factorial design, our experiment only uses 32. Based on table 8.30 from *Design and analysis of Experiments, Montgomery,* it is clear this model has resolution IV. This means all main tuning parameters can be included and some of the two factor interaction terms can also be included, but not all because some are aliases with each other.These values will be aliases with each other, which means that since there are only 32 trials, some of the effects influence more than one interaction. Looking at the correlation plot **Figure 4** for the factorial fraction design, it is clear the interaction terms which are aliases of each other are mtry:cutoff =ntree:replace, ntree:mtry = replace:cutoff, and ntree:cutoff = replace:mtry. In terms of multicollinearity this means these terms are completely equivalent and have a correlation of exactly one. With the limited number trials to run and seven factors, it is impossible to consider all two parameter interactions. The interactions between just the main effects for this plot is none, which means they can all be analyzed independently.

This D-optimal design has more flexibility in how many trials can be use for it, allowing for all 35 levels to be used for this design. For a level five resolution which would cover all the interaction terms of two parameters, there would need to be runs equal to or greater than 1 for the intercept, 7 for the main factors, and 7*6/2 = 21 more for each of the two parameter interactions (a 7 choose 2 probability). This would then take 29 runs to be able to achieve, and since this test has up to 35, it is possible for this model to include all the two parameter interaction terms and main factors because it is resolution five. There is some collinearity between different interactions, looking at **Figure 2** but none of them are completely correlated as was the case with the fractional factorial design. There is only some very minor interactions between the main effects on this correlation plot, but it is slightly higher than the fractional factorial design which had no interaction at all. Both models could be built out of only eight runs if the desire was to analyze only the main effects and the intercept, however for the sake of interactions with the main effects and the two parameter interactions, those can only be fully analyzed by a the optimal design with at least 29 trials or a fractional factorial design with at least 64 runs.

Question 3. *Recommend one experimental design between the two options in Question 1, motivate your decision*

We recommend using the D-optimal design over the fractional factorial design. As mentioned in Question 2, our D-optimal design is a Resolution V design; none of our main effects or two-factor interactions are aliased with any other main effect or two-factor interactions, so we can analyze all of the effects in our design. As Figure 2 shows, there is slight multicollinearity between the main effects and interactions, but none are completely correlated.

On the other hand, our fractional factorial design is a Resolution IV design and has interaction terms which are aliases of each other, as Figure 4 shows that some effects are completely correlated. The lower the resolution of the design, the more restrictions there are on our assumptions about which interactions can be used, which is why we decided on a D-optimal design.

Question 4. *Using a commercial software, the TAs and I came up with the experimental design shown in Table 2. How does your recommended design in the previous question compare with this one?*

The experimental design shown in Table 2 uses less runs (22 runs), so it might be more efficient compared to our D-optimal design (35 runs), as less runs require less time to compute. However, the Table 2 experimental design has 3 levels per factor (except for replace, which can only have two responses) while our D-optimal design has 2 levels per factor. Running an experimental design with 3 levels is more costly than running a design with 2 levels. In addition, not all of our main effects and interaction effects can be tested; it takes 29 runs for all main and interaction effects to be tested with an experimental design with 2 levels, so 22 runs is not enough to run all the main and interaction effects given by the design in Table 2 with 3 levels.

Comparing the correlation plots of the Table 2 experimental design and our D-optimal design, our D-optimal design may be a better design to choose, as there are much stronger correlations between interaction effects.

Part 2: Data Analysis

Question 5. *Collect data using your recommended design in Question 3.*

We collected data from the cross validation function using the D-optimal design. The cv.rf function takes the values of all seven parameters for each of the 35 runs in our design and outputs the cross validation value that results from each run in a new data frame to the right of the seven corresponding tuning parameter values. The cross validation value helps to measure the performance of the random forest algorithm. The data is shown in Figure 12.

Question 6. *Conduct a detailed data analysis. What are the influential tuning parameters? What is the final model that links the tuning parameters to the cross-validation accuracy? Does the final model provide a good fit to the data?*

Based on analysis of variance of our initial model (Figure 13), we find that classwt, cutoff, and maxnodes are significant main effects, and the interactions between ntree and nodesize, mtry and nodesize, cutoff and nodesize, classwt and cutoff, and classwt and maxnodes are also significant.

Minimal aliasing occurs between the effects in our new model (Figure 15). VIF is under 5 for all effects (Figure 16). This means that our model satisfies the requirement of low multicollinearity.

The residuals appear identically distributed (Figure 17), and the normal Q-Q plot is linear (Figure 17), satisfying the assumption of normality.

The final model is:

$$\mathit{Cross\ Validation\ Accuracy} = .618271 - .057672 \times \mathit{classwt} + .027108 \times \mathit{cutoff} + .014352 \times \mathit{maxnodes} + .012185 \times \mathit{ntree} \times \mathit{nodesize} + .011742 \times \mathit{nodesize} \times \mathit{mtry} - .007203 \times \mathit{cutoff} \times \mathit{nodesize} + .019271 \times \mathit{classwt} \times \mathit{cutoff} + .022981 \times \mathit{classwt} \times \mathit{maxnodes}$$

With an F statistic of 28.53 on 8 and 26 degrees of freedom we can conclude that this model is significant. We find that the final model has an adjusted R-squared of .87 (Figure 18). The model is a good fit for the data.

3. Conclusions

Using the D-optimal design works well because it is Resolution V. There is therefore not aliasing between main effects and the interaction terms, and all terms can be evaluated. The Resolution IV fractional factorial design lacks this capability. Further, the correlation plot of the D optimal design showed minimal evidence of multicollinearity between effects.

The model that we generated using the D-optimal design met the assumptions of normality, non-multicollinearity, constant variance. With an adjusted R^2 of .87, the model also does a good job of predicting the cross validation accuracy.

In future experimentation, we would recommend running a full factorial design of 128 runs as this would eliminate aliasing. We would also recommend evaluating the cross validation accuracy with different values for the tuning parameters to better assess their impact on cross validation accuracy.

4. Appendix

An Appendix showing your R code. The code should be such that I can run it and get the exact same output as you. Moreover, it should be well-documented and organized.

5. Statement of Contribution

INSTRUCTIONS, REMOVE: Clearly state the contribution of each team member to this project and report.

Christopher Thornton:

- Creation of D-optimal design
- Question 5
- Question 6
- Conclusion

Cara Drake

- Question 1
- Question 2
- Appendix
- Presentation Slides
- Code Editor

Stacy Deng

- Introduction
- Question 3
- Question 4
- Fractional Factorial Design

6. 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

R Libraries:

Barrios E (2016). _BHH2: Useful Functions for Box, Hunter and Hunter II_. R package version 2016.05.31, <https://CRAN.R-project.org/package=BHH2>.

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Taiyun Wei and Viliam Simko (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot

Ulrike Gr"omping (2014). R Package FrF2 for Creating and Analyzing Fractional Factorial 2-Level Designs. Journal of Statistical Software, 56(1), 1-56. URL https://www.jstatsoft.org/v56/i01/.

Wheeler B (2022). _AlgDesign: Algorithmic Experimental Design_. R package version 1.2.1, <https://CRAN.R-project.org/package=AlgDesign>.