

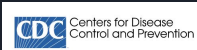


STAT 101B Final Project

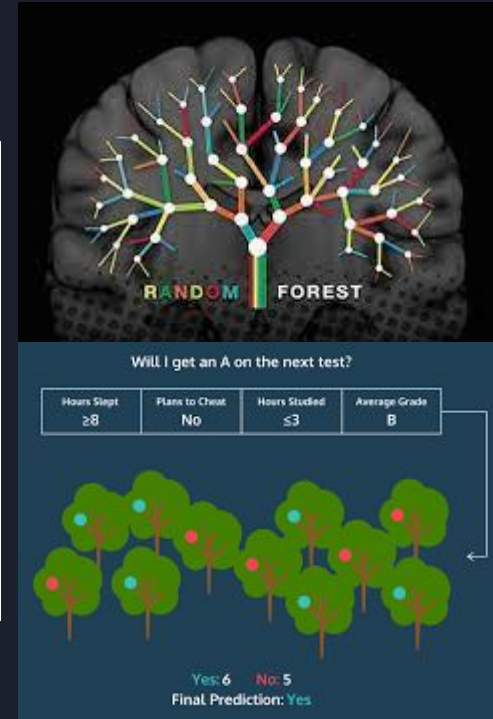
By Team 'Hey You'
Kyungchae Baek, Euijun Kim, Kaili Nguyen

Methodology and Results

- Classification by Random Forest and DataSet



Diabetes_binary	response	categorical	0 = no diabetes; 1 = prediabetes or diabetes
BMI	predictor	numeric	Body Mass Index
MentHlth	predictor	numeric	days of poor mental health scale in the past 30 days
PhysHlth	predictor	numeric	physical illness or injury days in past 30 days
HighBP	predictor	categorical	0 = no high BP; 1 = high BP
HighChol	predictor	categorical	0 = no high cholesterol; 1 = high cholesterol
CholCheck	predictor	categorical	0 = no cholesterol check in 5 years; 1 = yes cholesterol check in 5 years
Smoker	predictor	categorical	Have you smoked at least 100 cigarettes in your entire life? (0 = no; 1 = yes)
Stroke	predictor	categorical	Have you had a stroke? (0 = no; 1 = yes)
PhysActivity	predictor	categorical	physical activity in past 30 days - not including job (0 = no; 1 = yes)
Fruits	predictor	categorical	Consume Fruit 1 or more times per day (0 = no; 1 = yes)
Veggies	predictor	categorical	Consume Vegetables 1 or more times per day (0 = no; 1 = yes)
HvyAlcoholConsump	predictor	categorical	(adult men ≥14 drinks per week and adult women ≥7 drinks per week) (0 = no; 1 = yes)
GenHlth	predictor	categorical	Would you say that in general your health is: scale 1-5 (1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor)
DiffWalk	predictor	categorical	Do you have serious difficulty walking or climbing stairs? (0 = no 1 = yes)
Sex	predictor	categorical	(0 = female 1 = male)
Age	predictor	categorical	13-level age categorical



Methodology and Results

- Design and Analysis Process

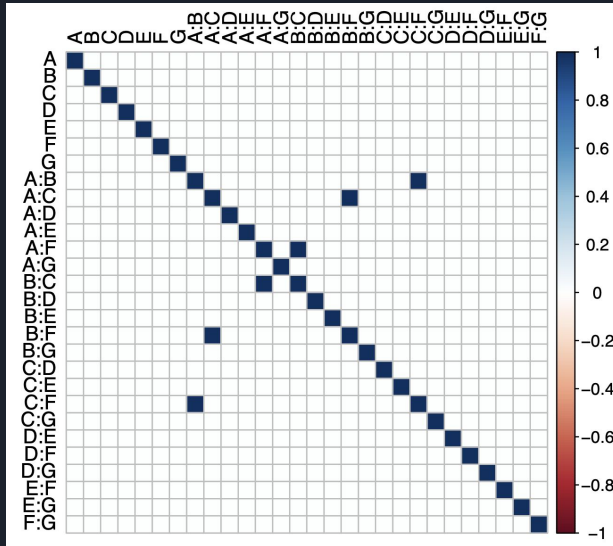


Figure 1 : Fractional Design with 32 run-size

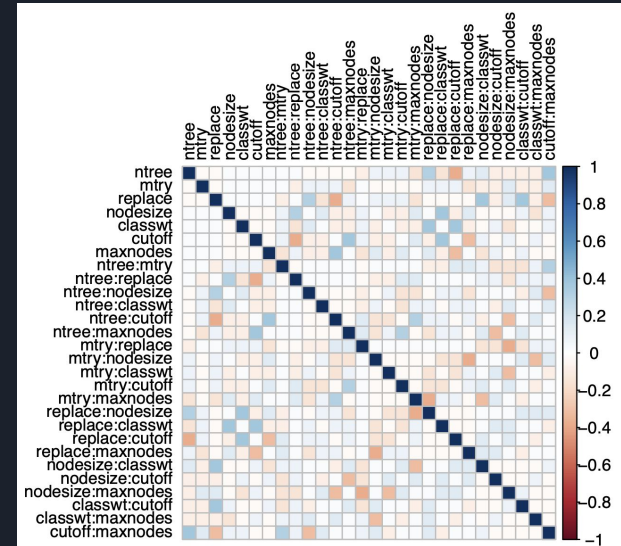


Figure 2: Optimal Design with 35 run-size

Methodology and Results

- Design and Analysis Process

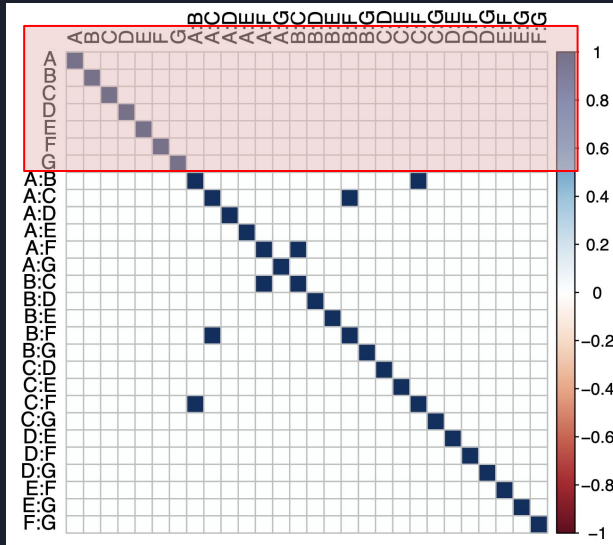


Figure 1 : Fractional Design with 32 run-size

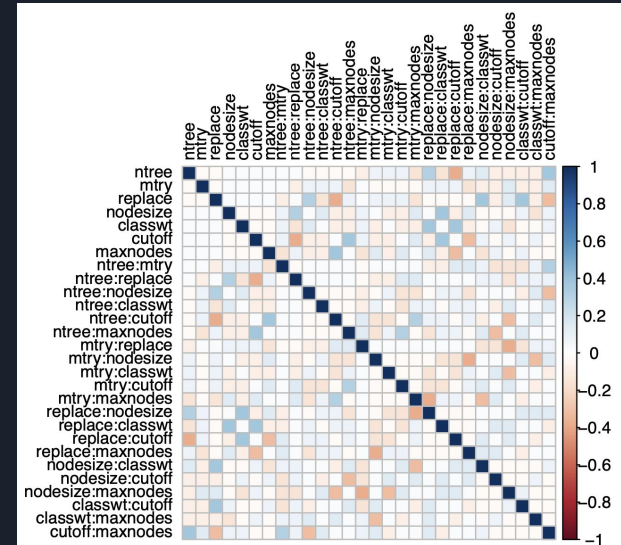


Figure 2: Optimal Design with 35 run-size

Methodology and Results

- Initial Model and How Discard Effects

```
##
## Call:
## lm.default(formula = CV ~ .^2, data = coded.frac.design)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.035267 -0.003779  0.000000  0.003779  0.035267
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6160116  0.0059075 104.276 5.24e-11 ***
## ntree         -0.0016634  0.0059075  -0.282  0.78773
## mtry          0.0064609  0.0059075   1.094  0.31605
## replace       0.0041166  0.0059075   0.697  0.51198
## nodesize      0.0001064  0.0059075   0.018  0.98621
## classwt      -0.0573193  0.0059075  -9.703 6.88e-05 ***
## cutoff        0.0211215  0.0059075   3.575  0.01171 *
## maxnodes      0.0197074  0.0059075   3.336  0.01569 *
## ntree:mtry    -0.0035539  0.0059075  -0.602  0.56946
## ntree:replace -0.0025930  0.0059075  -0.439  0.67608
## ntree:nodesize 0.0087954  0.0059075   1.489  0.18710
## ntree:classwt -0.0002265  0.0059075  -0.038  0.97066
## ntree:cutoff   0.0057255  0.0059075   0.969  0.36988
## ntree:maxnodes -0.0018116  0.0059075  -0.307  0.76947
## mtry:replace   NA         NA         NA         NA
## mtry:nodesize  -0.0004858  0.0059075  -0.082  0.93714
## mtry:classwt   0.0041618  0.0059075   0.705  0.50753
## mtry:cutoff     NA         NA         NA         NA
## mtry:maxnodes  -0.0027243  0.0059075  -0.461  0.66093
## replace:nodesize 0.0121399  0.0059075   2.055  0.08565
## replace:classwt 0.0065507  0.0059075   1.109  0.30994
## replace:cutoff   NA         NA         NA         NA
## replace:maxnodes 0.0043273  0.0059075   0.733  0.49147
```

```
## nodesize:classwt -0.0039610  0.0059075  -0.671  0.52749
## nodesize:cutoff   0.0056848  0.0059075   0.962  0.37306
## nodesize:maxnodes 0.0003546  0.0059075   0.060  0.95409
## classwt:cutoff    0.0186196  0.0059075   3.152  0.01977 *
## classwt:maxnodes  0.0341566  0.0059075   5.782  0.00117 **
## cutoff:maxnodes   0.0227729  0.0059075   3.855  0.00841 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 0.03342 on 6 degrees of freedom

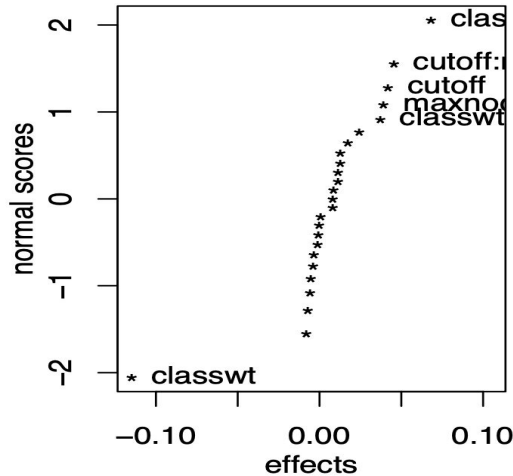
Multiple R-squared: 0.9694, Adjusted R-squared: 0.8418

F-statistic: 7.597 on 25 and 6 DF, p-value: 0.008985

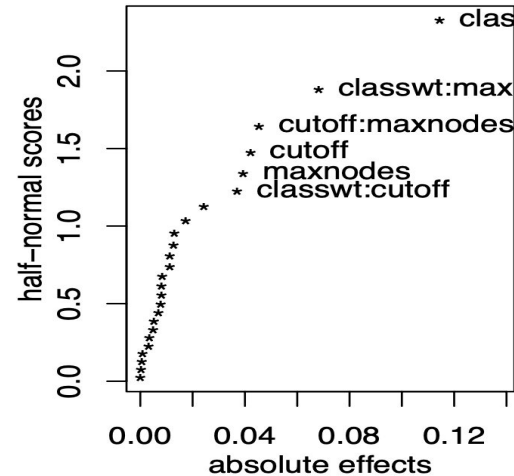
Methodology and Results

- Additional way to find Significant Effects

Normal Plot for CV, alpha=0.05



Half Normal Plot for CV, alpha=0.05



Methodology and Results

- Final model

```
##
## Call:
## lm.default(formula = CV ~ classwt + cutoff + maxnodes + classwt:maxnodes +
##           cutoff:maxnodes + classwt:cutoff, data = coded.frac.design)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.049940 -0.020995 -0.000721  0.012140  0.080060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.616012   0.005238  117.613 < 2e-16 ***
## classwt       -0.057319   0.005238  -10.944 5.04e-11 ***
## cutoff        0.021122   0.005238   4.033 0.000456 ***
## maxnodes      0.019707   0.005238   3.763 0.000909 ***
## classwt:maxnodes 0.034157   0.005238   6.521 7.86e-07 ***
## cutoff:maxnodes 0.022773   0.005238   4.348 0.000202 ***
## classwt:cutoff  0.018620   0.005238   3.555 0.001538 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02963 on 25 degrees of freedom
## Multiple R-squared:  0.8997, Adjusted R-squared:  0.8756
## F-statistic: 37.38 on 6 and 25 DF, p-value: 2.643e-11
```



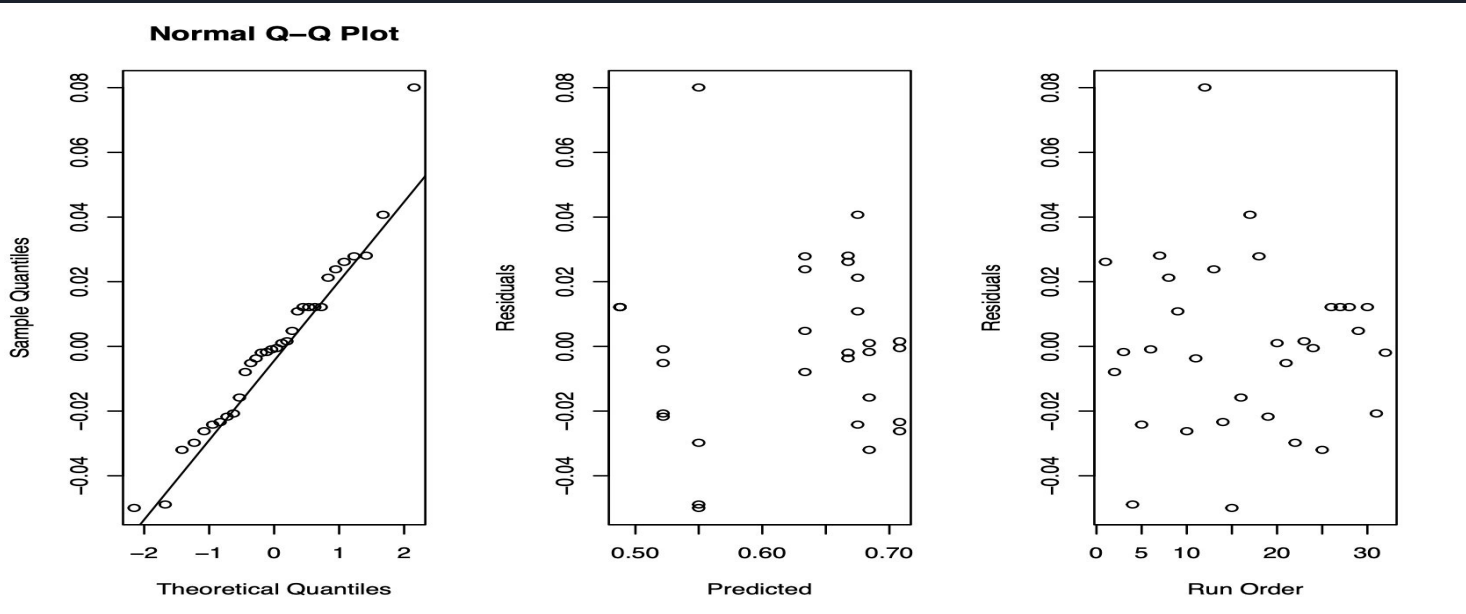
Methodology and Results


- Final model

$$\begin{aligned} \text{cross-validationaccuracy} = & 0.616012 - 0.057319 * \text{classwt} + 0.021122 * \text{cutoff} + 0.019707 * \text{maxnodes} \\ & + 0.034157 * \text{classwt} : \text{maxnodes} + 0.022773 * \text{cutoff} : \text{maxnodes} + 0.018620 * \text{classwt} : \text{cutoff} \end{aligned}$$

Methodology and Results

- Final Model and Evaluation by Residual Analysis





Methodology and Results

- multicollinearity

##	Var.32run	VIF.32run
## classwt	0.03125	1
## cutoff	0.03125	1
## maxnodes	0.03125	1
## classwt:cutoff	0.03125	1
## classwt:maxnodes	0.03125	1
## cutoff:maxnodes	0.03125	1
## classwt:cutoff:maxnodes	0.03125	1



CONCLUSIONS - Strengths of the Model

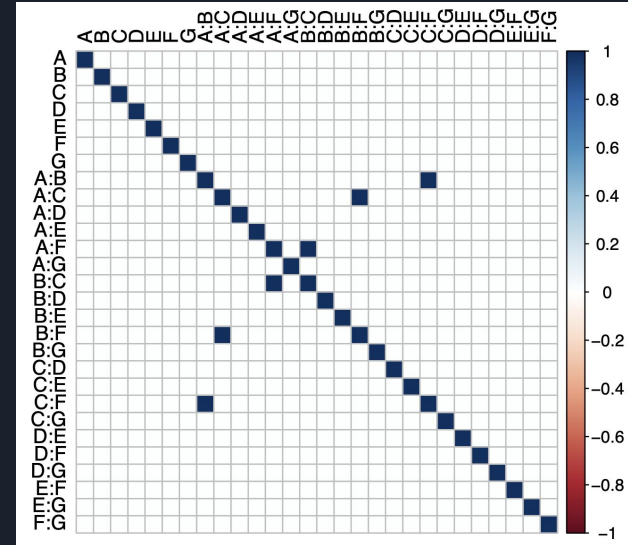
$$y_i = 0.616012 - 0.057319x_{iF} + 0.021122x_{iG} + 0.019707x_{iE} + 0.034157x_{iFiE} + 0.022773x_{iGiE} + 0.018620x_{iFiG}$$

where y_i is Random Forest's calculated Cross Validation value, x_{iE} is the i -th maximum number of nodes in a tree, x_{iF} is the i -th assigned probabilistic weight of each class being studied, x_{iG} is the i -th threshold for binary classification of observations, x_{iFiE} is the interaction between the i -th classwt(F) and i -th maxnodes(E), x_{iGiE} is the interaction between the i -th cutoff(G) and the i -th maxnodes(E), and x_{iFiG} is the interaction between the i -th classwt(G) and the i -th cutoff(F), where $i = 1, \dots, 25,000$.

- High Adjusted R-squared value
 - Explains 87.29% of variability in the model
- Interpretable
 - Only 6 parameters
 - No transformations
- Low VIF
 - Very little multicollinearity

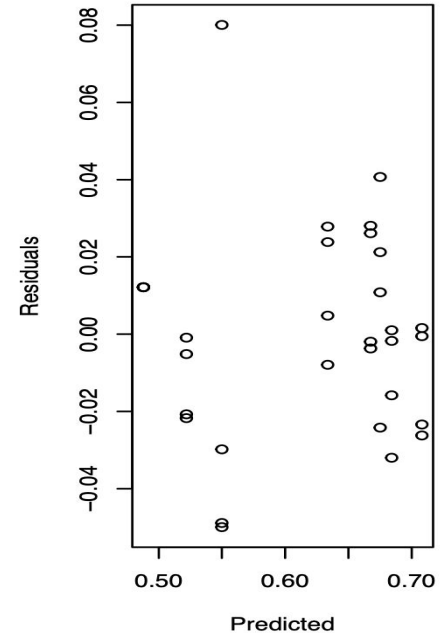
CONCLUSIONS - Strengths of the Design

- Fractional Factorial Design
 - Resolution IV
 - No aliasing
 - No multicollinearity
- 32 runs
 - Decrease experimental variance
 - Saves money



CONCLUSIONS - Recommendations

- More precision
 - Our model 87.29% of variability in the CV
 - Our design did not incorporate higher interactions
- Transformation
 - Residual vs Predicted plot
- Interpretability
 - Interactions are hard to interpret
- Experimental Variability
 - Didn't utilize all the runs





THANK YOU