Parameter Grid Search for Random Forest Classifier on Fall Detection Data

Vladislav Belov

Project Overview

This small project takes a look at the Fall Detection Data from China. The data contains 6 variables:

- **TIME** - the total time of patient's monitoring;
- **SL** - the level of sugar in the organism;
- **EEG** - electroencephalography monitoring rate;
- **BP** - blood pressure;
- **HR** - heart beat rate;
- **CIRCULATION** - blood circulation.

The response variable **ACTIVITY** classifies the type of activity patients were doing during the period of taking measurements of variables presented above:

Table 1: Types of Activity

ACTIVITY	Type of the Activity
0	Standing
1	Walking
2	Sitting
3	Falling
4	Cramps
5	Running

Here is a quick look at the head of the dataset we will be dealing with (it contains 16382 rows in total):

Table 2: Fall Detection Data from China - Sample

ACTIVITY	TIME	SL	EEG	BP	HR	CIRCLUATION
3	4722.92	4019.64	-1600.00	13	79	317
2	4059.12	2191.03	-1146.08	20	54	165
2	4773.56	2787.99	-1263.38	46	67	224
4	8271.27	9545.98	-2848.93	26	138	554
4	7102.16	14148.80	-2381.15	85	120	809

In this project we will use the Random Forest classifier to train the machine to classify activities according to basic inner body measurements. The classifier will be trained on a pre-prepaired dataset, as it is going to be cleaned and, moreover, all of the explanatory variables will be normalized.

There will be multiple training sessions depending on the initial parameters of the Random Forest model. This set of parameters will be called the grid of parameters. Each training session on the grid will be divided into K folds (K-fold cross-validation), and the prediction accuracy will be evaluated as the mean accuracy of cross-validation. Aforementioned procedures will provide us with a new dataset with the parameter grid as explanatory variables and model accuracy as the response variable. We will analyze significance of different effects and their interactions and attempt to fit model accuracy.

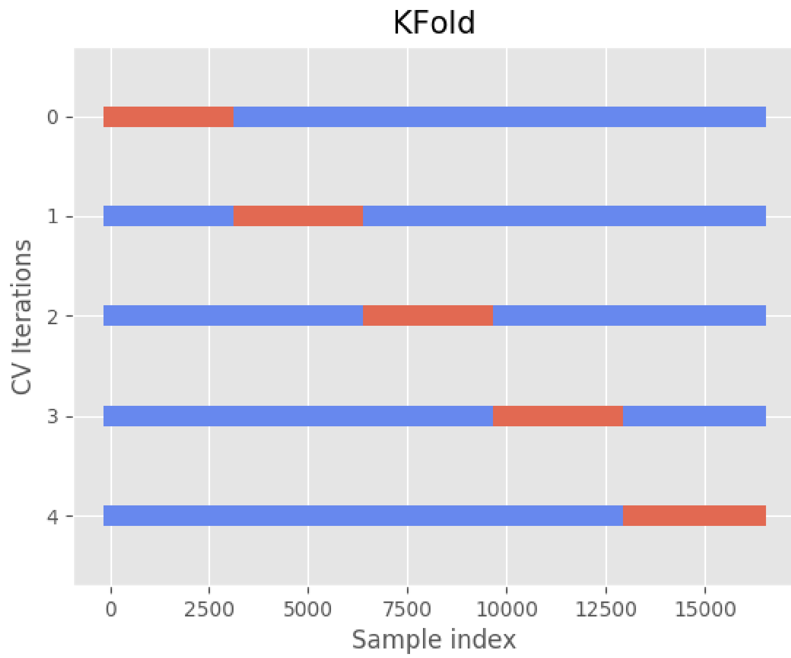


Figure 1: K-fold Cross-Validation Visualization

The grid of parameters we will be focusing on is the following one:

Table 3: Parameters of the Random Forest Classifier

Parameter Name	Description	Parameter Type	Considered Values
`bootstrap`	whether bootstrap samples are used when building trees	*factor*	{True, False}
`max_depth`	the maximum depth of the tree	*numeric*	{10, 40}
`max_features`	the number of features to consider when looking for the best split	*factor*	{'sqrt', 'log2'}
`min_samples_split`	the minimum number of samples required to split an internal node	*numeric*	{4, 20}
`criterion`	the function to measure the quality of a split	*factor*	{'gini', 'entropy'}
`n_estimators`	the number of trees in the forest	*numeric*	{10, 500}

As we are dealing with the 2^6-factorial design, we will also measure center points for numeric variables, i.e. model accuracies for 25, 12, 255 of `max_depth`, `min_samples_split` and `n_estimators`, respectively.

Results of the Random Forest Classifier Training

After all training sessions results are as follows:

Table 4: Random Forest Classifier Accuracy on the Grid

n_estimators	min_samples_split	max_features	max_depth	criterion	bootstrap	accuracy
10	4	sqrt	10	entropy	True	0.7125736
10	4	sqrt	10	entropy	False	0.7159569
10	4	sqrt	10	gini	True	0.7067552
10	4	sqrt	10	gini	False	0.7167142
10	4	sqrt	40	entropy	True	0.7455045
10	4	sqrt	40	entropy	False	0.7460128
10	4	sqrt	40	gini	True	0.7435407
10	4	sqrt	40	gini	False	0.7443576
10	4	log2	10	entropy	True	0.7154351
10	4	log2	10	entropy	False	0.7134020
10	4	log2	10	gini	True	0.7106681
10	4	log2	10	gini	False	0.7154786
10	4	log2	40	entropy	True	0.7436256
10	4	log2	40	entropy	False	0.7458334
10	4	log2	40	gini	True	0.7486115
10	4	log2	40	gini	False	0.7438735
10	20	sqrt	10	entropy	True	0.7097638
10	20	sqrt	10	entropy	False	0.7212099
10	20	sqrt	10	gini	True	0.7027979
10	20	sqrt	10	gini	False	0.7080121
10	20	sqrt	40	entropy	True	0.7430615
10	20	sqrt	40	entropy	False	0.7509147
10	20	sqrt	40	gini	True	0.7404538
10	20	sqrt	40	gini	False	0.7513531
10	20	log2	10	entropy	True	0.7128923
10	20	log2	10	entropy	False	0.7104494
10	20	log2	10	gini	True	0.7026604
10	20	log2	10	gini	False	0.7105104
10	20	log2	40	entropy	True	0.7450455
10	20	log2	40	entropy	False	0.7509433
10	20	log2	40	gini	True	0.7415975
10	20	log2	40	gini	False	0.7456683
500	4	sqrt	10	entropy	True	0.7331561
500	4	sqrt	10	entropy	False	0.7305980
500	4	sqrt	10	gini	True	0.7289745
500	4	sqrt	10	gini	False	0.7306739
500	4	sqrt	40	entropy	True	0.7686580
500	4	sqrt	40	entropy	False	0.7590120
500	4	sqrt	40	gini	True	0.7673942
500	4	sqrt	40	gini	False	0.7601011
500	4	log2	10	entropy	True	0.7342055
500	4	log2	10	entropy	False	0.7307965
500	4	log2	10	gini	True	0.7301915
500	4	log2	10	gini	False	0.7307161
500	4	log2	40	entropy	True	0.7665523
500	4	log2	40	entropy	False	0.7595934
500	4	log2	40	gini	True	0.7677508
500	4	log2	40	gini	False	0.7598164
500	20	sqrt	10	entropy	True	0.7249124
500	20	sqrt	10	entropy	False	0.7276351

n_estimators	min_samples_split	max_features	max_depth	criterion	bootstrap	accuracy
500	20	sqrt	10	gini	True	0.7232871
500	20	sqrt	10	gini	False	0.7254269
500	20	sqrt	40	entropy	True	0.7588414
500	20	sqrt	40	entropy	False	0.7623857
500	20	sqrt	40	gini	True	0.7591891
500	20	sqrt	40	gini	False	0.7621528
500	20	log2	10	entropy	True	0.7264710
500	20	log2	10	entropy	False	0.7291640
500	20	log2	10	gini	True	0.7234847
500	20	log2	10	gini	False	0.7262218
500	20	log2	40	entropy	True	0.7580831
500	20	log2	40	entropy	False	0.7617788
500	20	log2	40	gini	True	0.7583701
500	20	log2	40	gini	False	0.7622724
255	12	sqrt	25	entropy	True	0.7647718
255	12	sqrt	25	entropy	False	0.7649124
255	12	sqrt	25	gini	True	0.7648560
255	12	sqrt	25	gini	False	0.7635025
255	12	log2	25	entropy	True	0.7639772
255	12	log2	25	entropy	False	0.7624089
255	12	log2	25	gini	True	0.7632867
255	12	log2	25	gini	False	0.7639014

Summary of the generated data set:

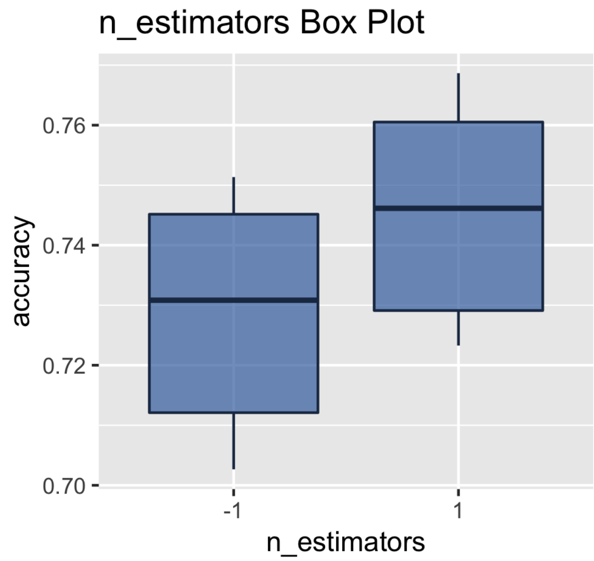
```
##   n_estimators min_samples_split max_features         max_depth
##  Min.   : 10   Min.   : 4        Length:72          Min.   :10
##  1st Qu.: 10   1st Qu.: 4        Class :character   1st Qu.:10
##  Median :255   Median :12        Mode  :character   Median :25
##  Mean   :255   Mean   :12                           Mean   :25
##  3rd Qu.:500   3rd Qu.:20                           3rd Qu.:40
##  Max.   :500   Max.   :20                           Max.   :40
##   criterion          bootstrap            accuracy
##  Length:72          Length:72          Min.   :0.7027
##  Class :character   Class :character   1st Qu.:0.7246
##  Mode  :character   Mode  :character   Median :0.7436
##                                        Mean   :0.7399
##                                        3rd Qu.:0.7596
##                                        Max.   :0.7687
```

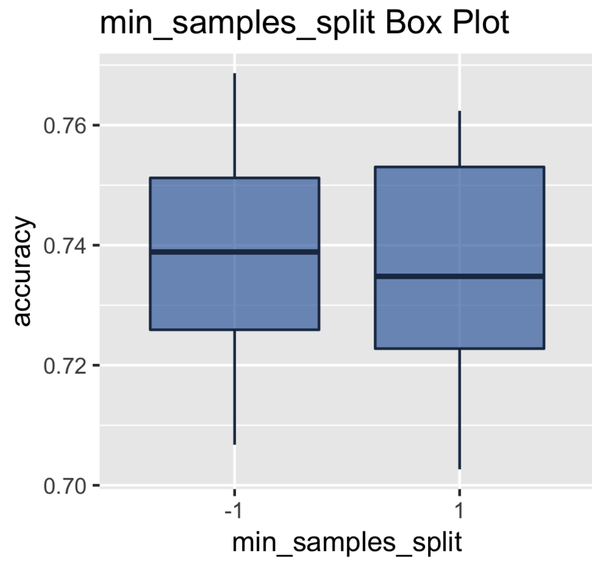
First Look into the Data

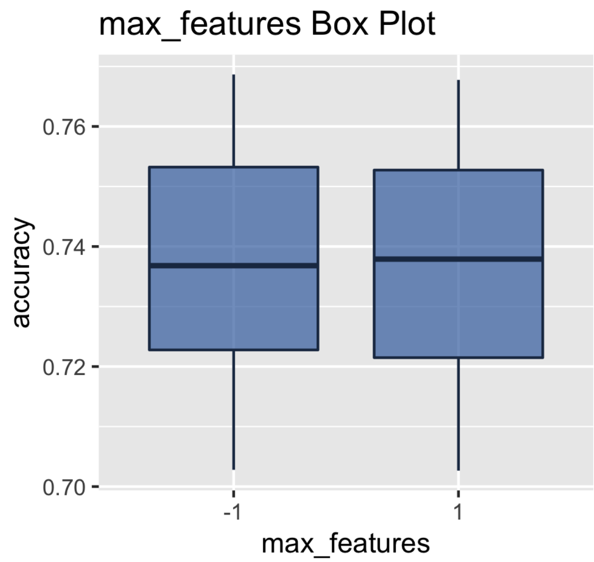
Visual Analysis

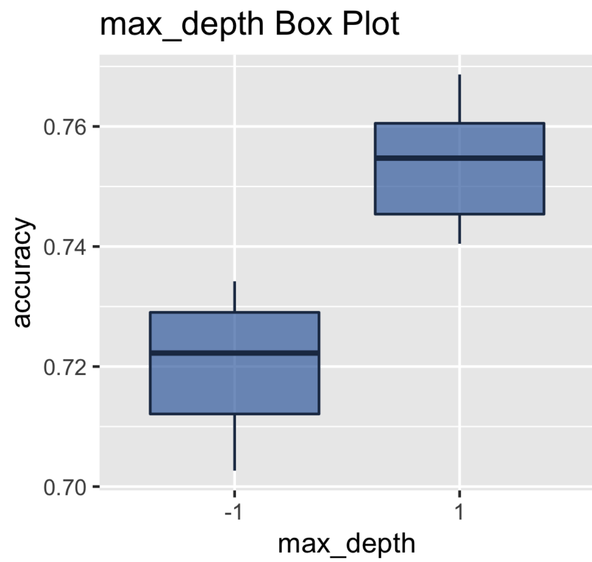
Preliminary analysis via box plot visualization indicates that not all of the variables are significant for performance of the Random Forest model.[1] For instance, only `n_estimators` and `max_depth` draw our interest. Regarding the rest of variables, none of them indicate any significance, however, more precise analysis is needed.[2]

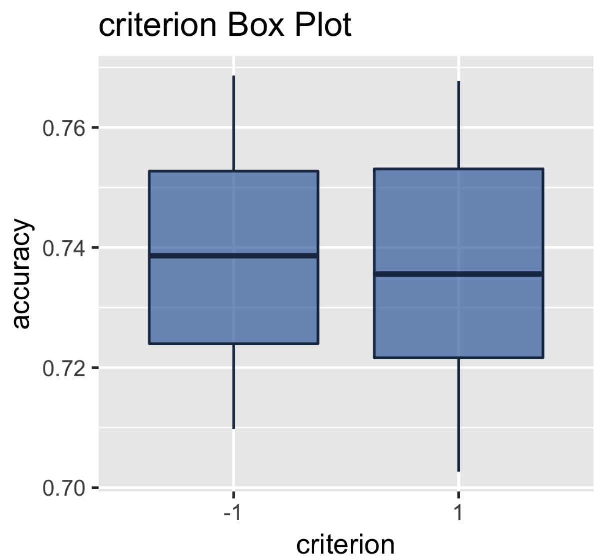
[1] All of the variables were mapped from actual ones to -1, 0 (in case of center points presence), and 1.
[2] All tests will be performed on the significance level of $\alpha = 5\%$.

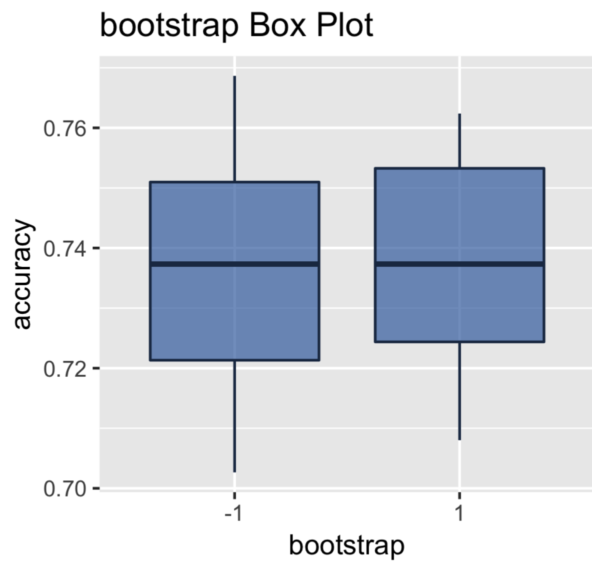

n_estimators Box Plot
accuracy
0.76
0.74
0.72
0.70
-1
1
n_estimators


min_samples_split Box Plot
accuracy
0.76
0.74
0.72
0.70
-1
1
min_samples_split


max_features Box Plot
accuracy
0.76
0.74
0.72
0.70
-1
1
max_features


max_depth Box Plot
accuracy
0.76
0.74
0.72
0.70
-1
1
max_depth


criterion Box Plot
accuracy
0.76
0.74
0.72
0.70
-1
1
criterion


bootstrap Box Plot
accuracy
0.76
0.74
0.72
0.70
-1
1
bootstrap

ANOVA without Interactions

If we take a closer look into differences between factors, then we discover that actually more variables are of interest to us. Firstly, we perform ANOVA without interactions and see that `min_samples_split` and `criterion` are also quiet important, `bootstrap` is on the margin.

```
##                   Df   Sum Sq  Mean Sq  F value  Pr(>F)
## n_estimators       1 0.004425 0.004425  384.390 < 2e-16 ***
## min_samples_split  1 0.000125 0.000125   10.877 0.00168 **
## max_features       1 0.000000 0.000000    0.001 0.97709
## max_depth          1 0.018264 0.018264 1586.404 < 2e-16 ***
## criterion          1 0.000067 0.000067    5.803 0.01925 *
## bootstrap          1 0.000046 0.000046    4.035 0.04931 *
## Residuals         57 0.000656 0.000012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey's HSD

"Honest Significant Differences" indicates the same fact, as only `max_features` confidence interval includes zero:

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov.default(formula = accuracy ~ ., data = df_mapped)
##
## $n_estimators
##            diff        lwr        upr p adj
## 1--1 0.01663093 0.01493231 0.01832954     0
##
## $min_samples_split
##              diff          lwr          upr     p adj
## 1--1 -0.002797601 -0.004496219 -0.001098983 0.0016807
##
## $max_features
##              diff          lwr         upr     p adj
## 1--1 2.446391e-05 -0.001674154 0.001723082 0.9770929
##
## $max_depth
##            diff        lwr        upr p adj
## 1--1 0.03378605 0.03208743 0.03548467     0
##
## $criterion
##              diff        lwr          upr     p adj
## 1--1 -0.002043483 -0.0037421 -0.000344865 0.0192495
##
## $bootstrap
##            diff          lwr         upr     p adj
## 1--1 0.00170396 5.341979e-06 0.003402577 0.0493076
```

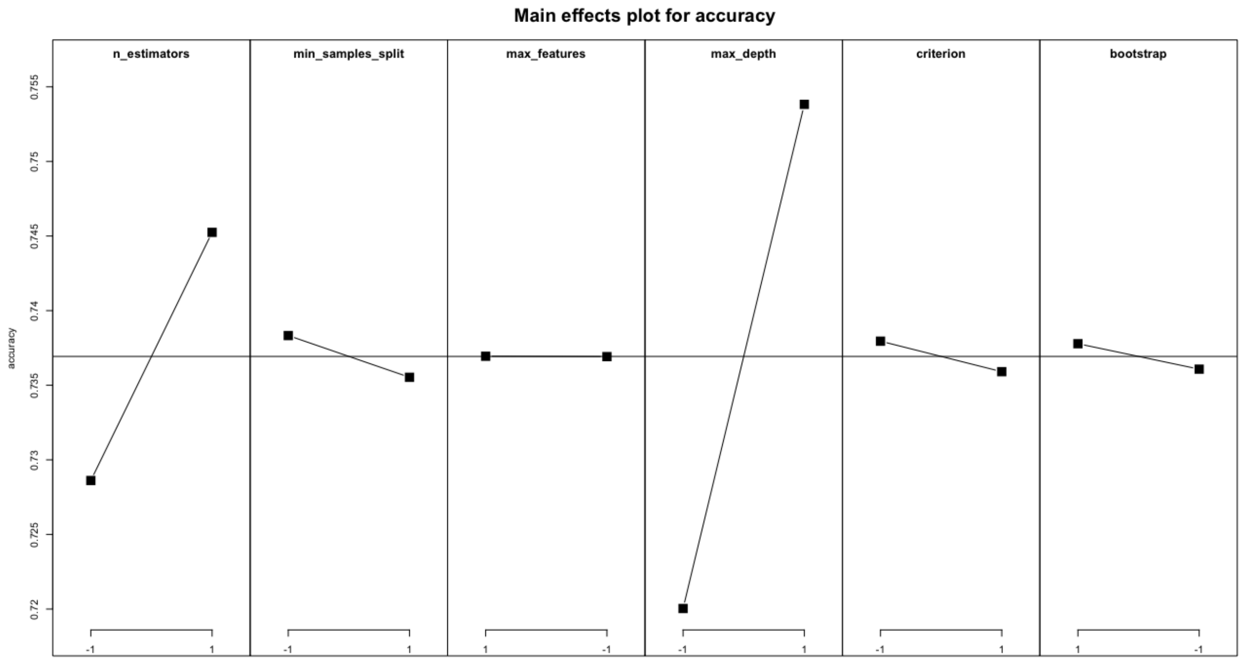


Figure 2: Main Effects Plot

Main Effects

Looking at the main effects plot and taking into account facts presented above, we can conclude, that `max_depth`, `n_estimators`, `min_samples_split`, `criterion` and `bootstrep` (presented in the order from the highest importance to the lowest) provide us with an explanation of the model accuracy behavior.

Analysis of Interactions

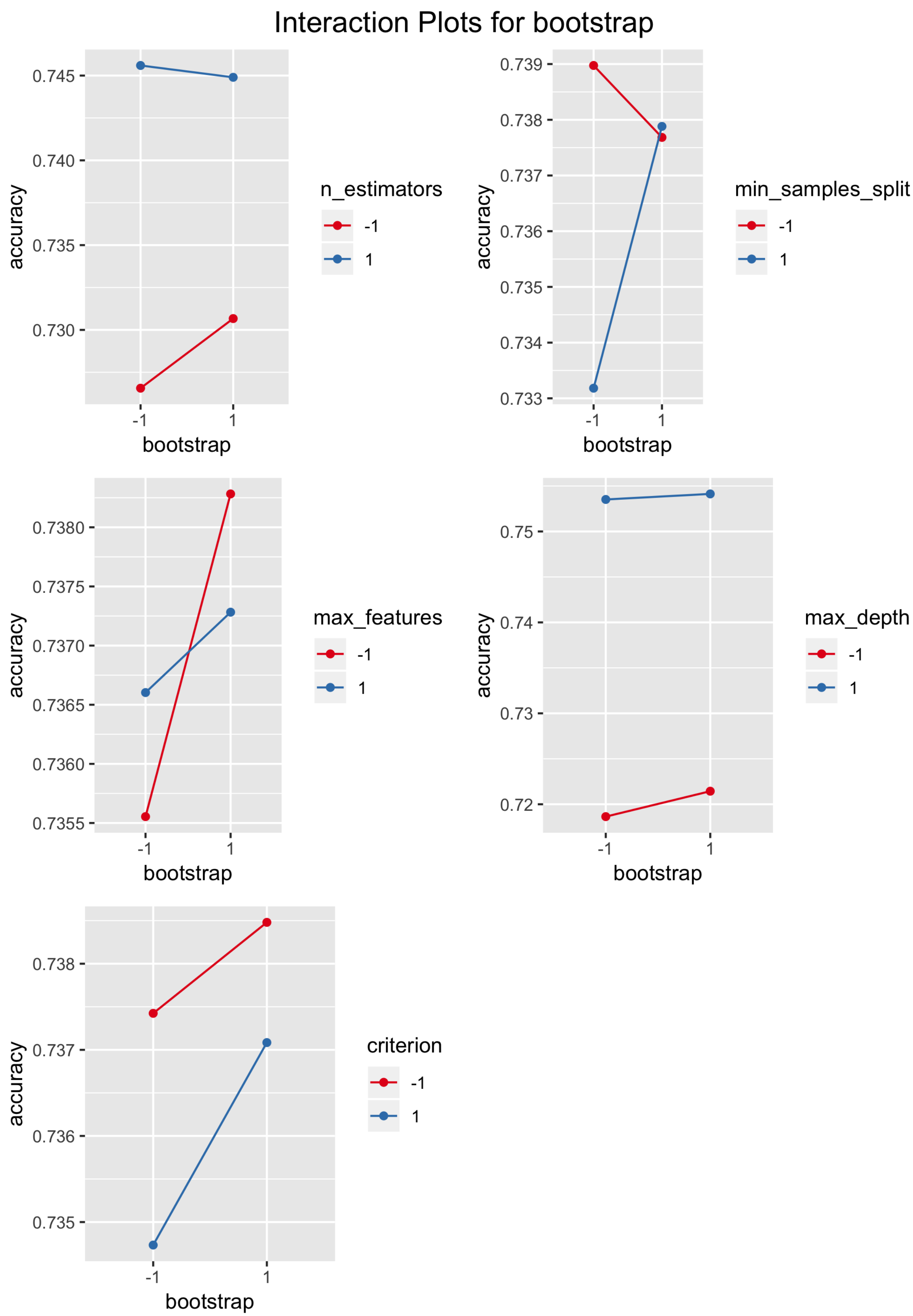
Visual Analysis of Double Interactions

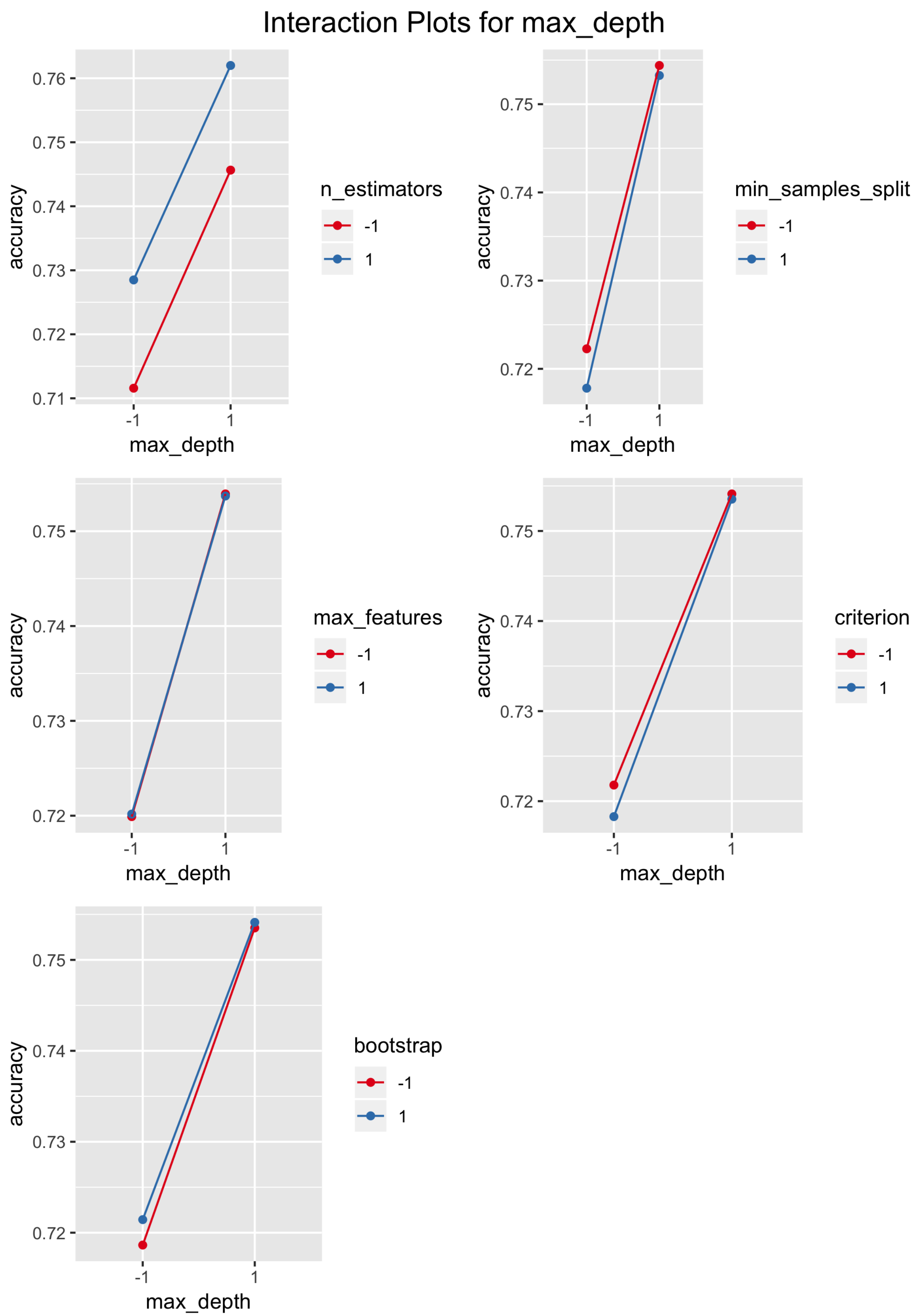
From interaction plots presented below we can empiracally assess the importance of interactions between variables ("**X**" - important, "**-**" - not important):[3]

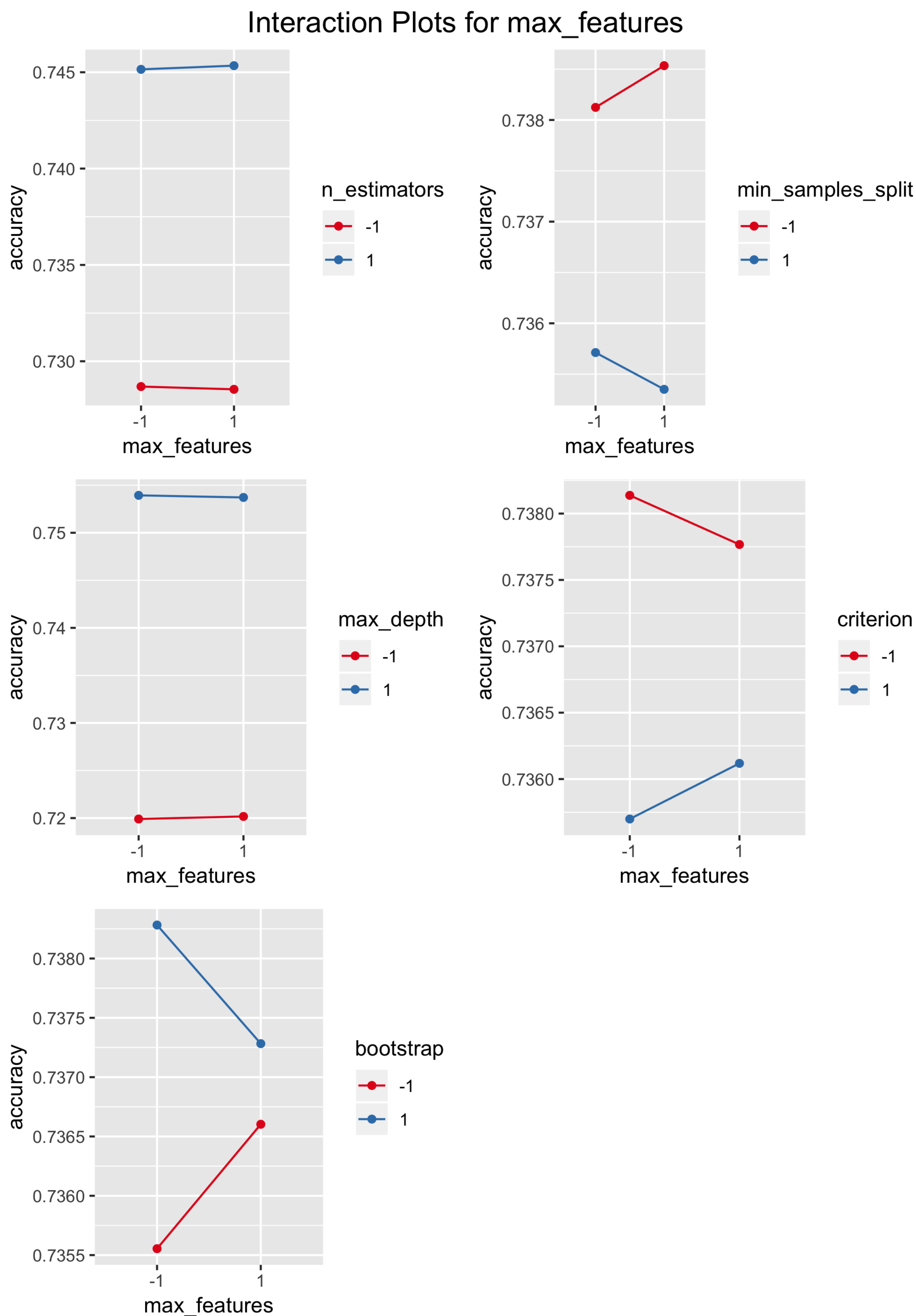
Table 5: Empirical Assessment of Pairwise Interactions

Variables	A	B	C	D	E	F
A	*NA*	**-**	**X**	**X**	**-**	**X**
B	**-**	*NA*	**-**	**-**	**-**	**-**
C	**X**	**-**	*NA*	**X**	**X**	**-**
D	**X**	**-**	**X**	*NA*	**X**	**-**
E	**-**	**-**	**X**	**X**	*NA*	**-**
F	**X**	**-**	**-**	**-**	**-**	*NA*

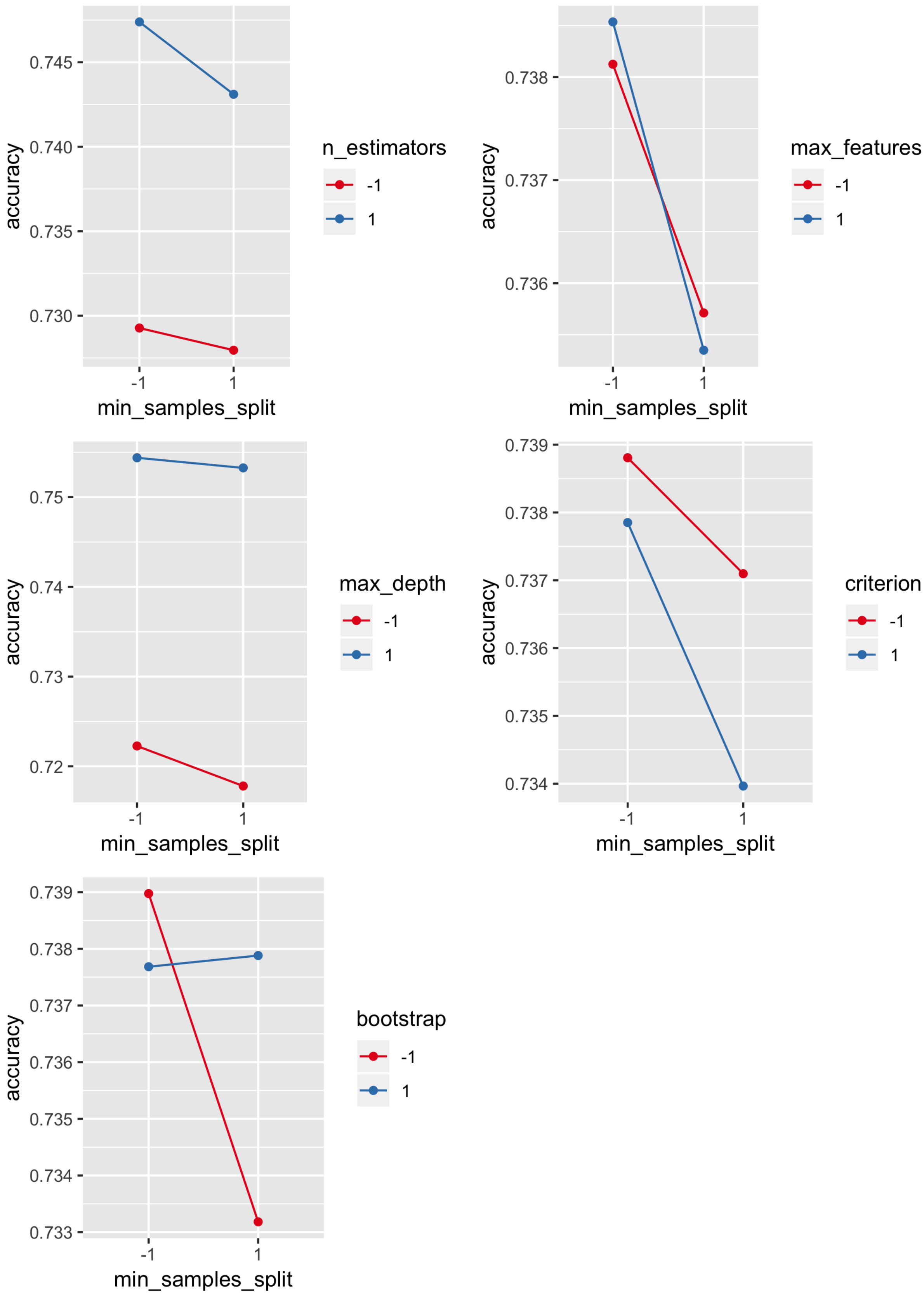
[3]For the reason of taking less space on the page here we introduce the following notation: `A` - `bootstrap`, `B` - `max_depth`, `C` - `max_features`, `D` - `min_samples_split`, `E` - `criterion`, `F` - `n_estimators`.


Interaction Plots for bootstrap
accuracy
bootstrap
n_estimators
-1
1
min_samples_split
max_features
max_depth
criterion
0.745
0.740
0.735
0.730
0.739
0.738
0.737
0.736
0.735
0.734
0.733
0.7380
0.7375
0.7370
0.7365
0.7360
0.7355
0.75
0.74
0.73
0.72
0.738
0.737
0.736
0.735

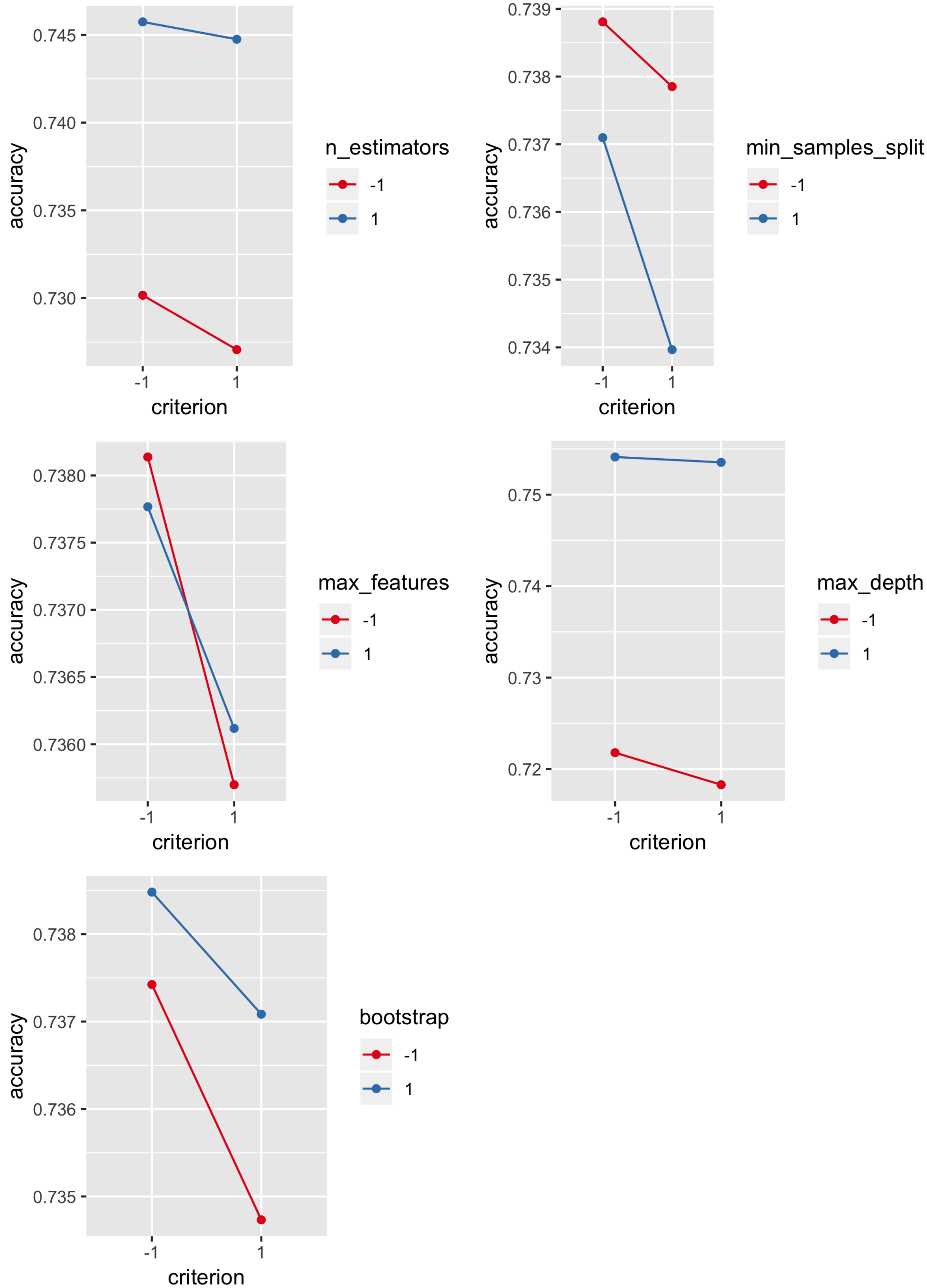

Interaction Plots for max_depth
accuracy
max_depth
n_estimators
min_samples_split
max_features
criterion
bootstrap
-1
1
0.71
0.72
0.73
0.74
0.75
0.76


Interaction Plots for max_features
accuracy
max_features
-1
1
n_estimators
min_samples_split
max_depth
criterion
bootstrap
0.745
0.740
0.735
0.730
0.738
0.737
0.736
0.75
0.74
0.73
0.72
0.7380
0.7375
0.7370
0.7365
0.7360
0.7355

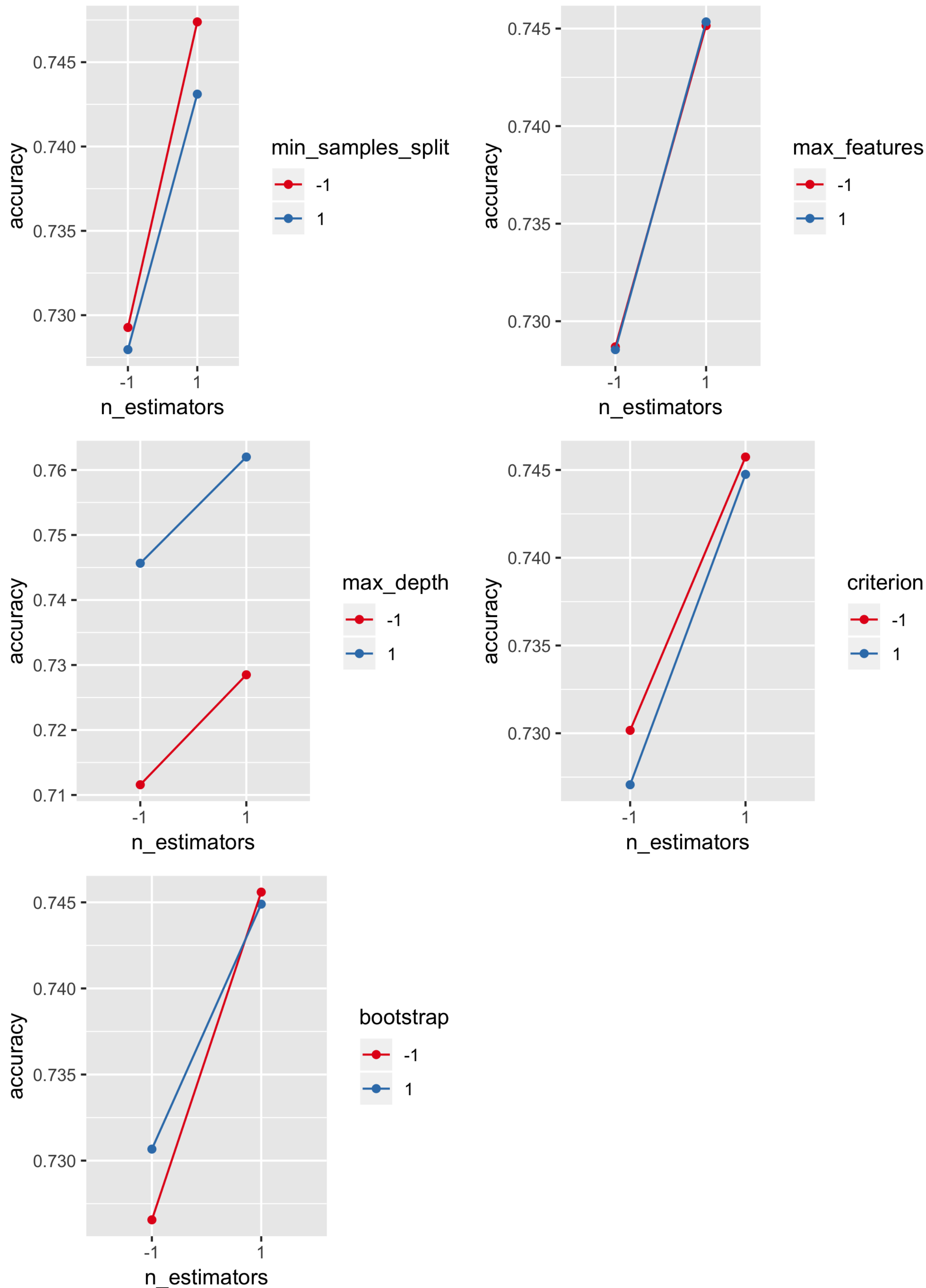
Interaction Plots for min_samples_split



Interaction Plots for criterion

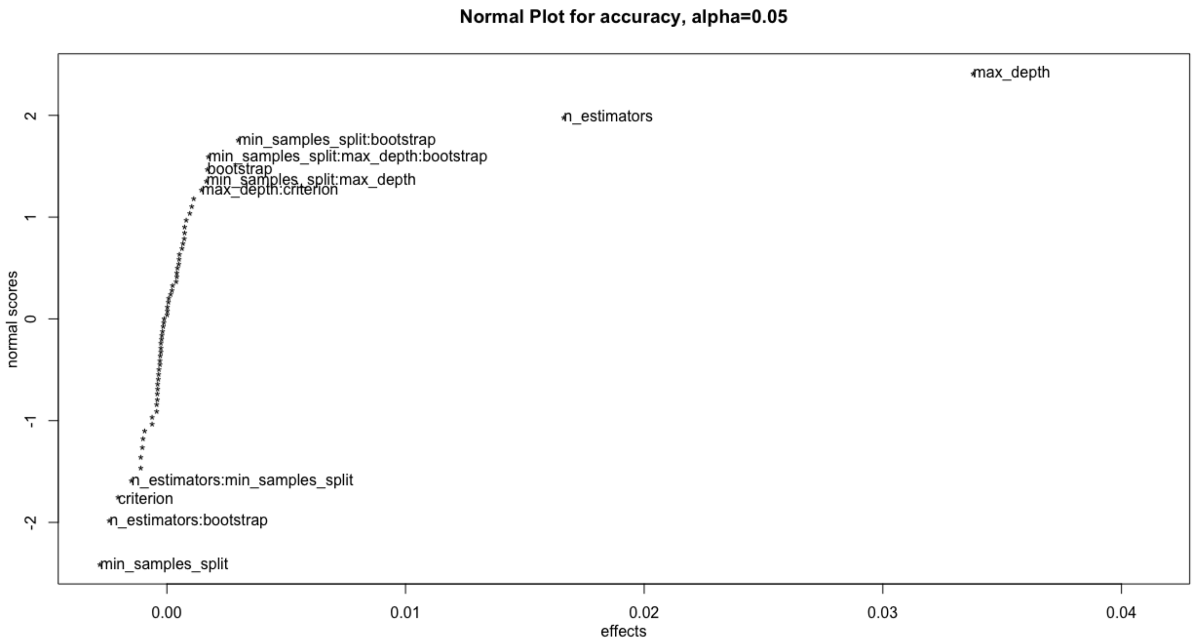


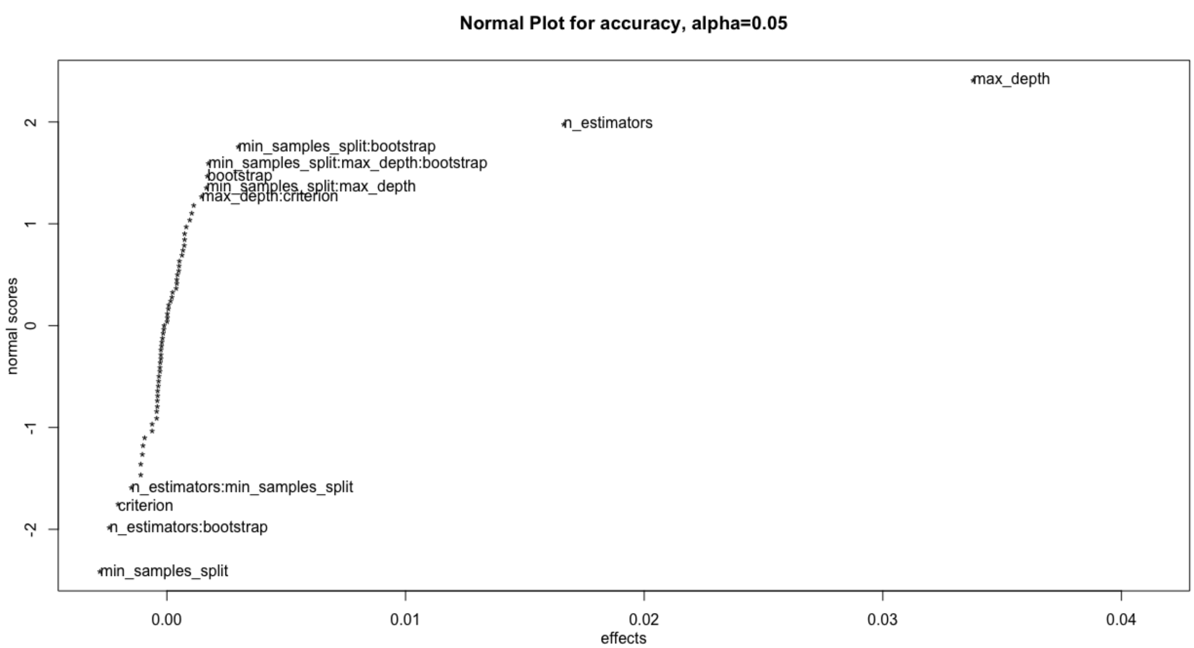
Interaction Plots for n_estimators



Daniel Plot

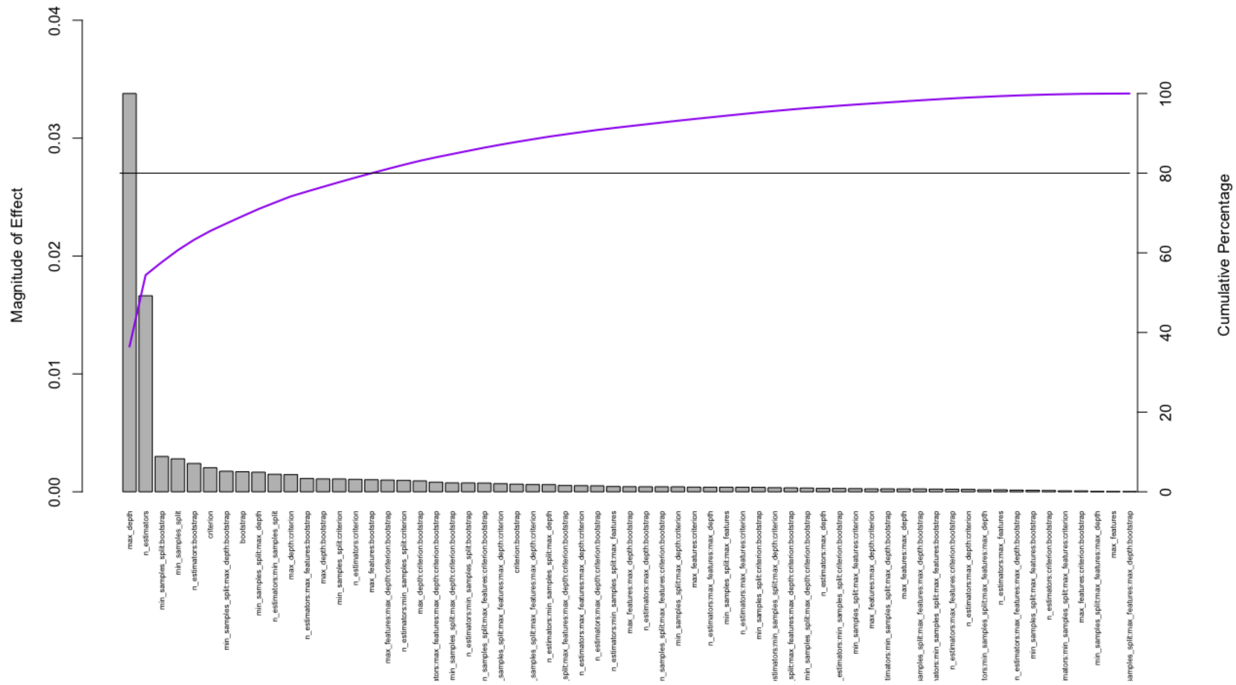
Daniel plot investigates variable signigicance with an assumption that main effects and interactions have Gaussian distribution with constant variance. Non-null effects then look like outliers on the normal plot. In figures presented below only significant effects have names near their the markers.

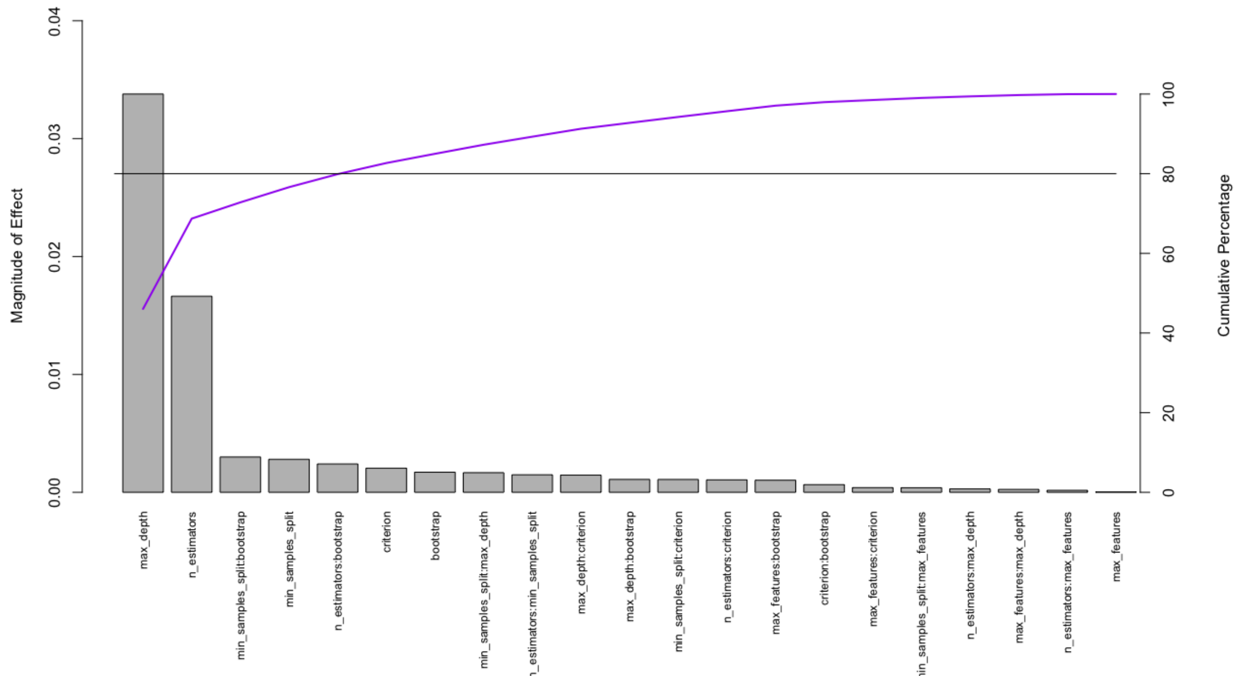




Pareto Plot

Pareto plot is another instrument which aids us in the investigation of effects. The most significant ones are obviously those with the highest magnitude.





ANOVA with Double Interactions

From the visual analysis given by three previous paragraphs it is arguably clear that none of higher interactions than double ones are significant enough for sufficient explanation of `accuracy` of the Random Forest classifier. Therefore, we also perform ANOVA with double interactions:

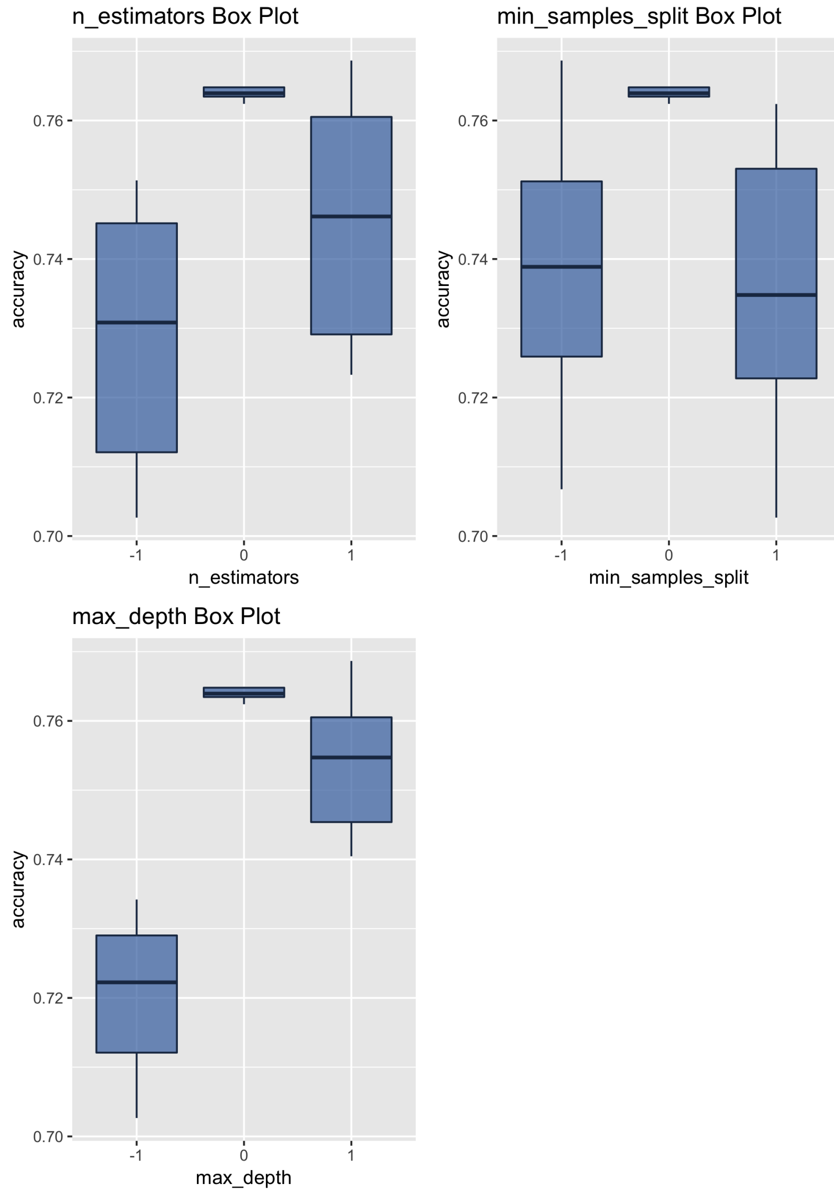
```
##                                Df   Sum Sq  Mean Sq  F value   Pr(>F)    
## n_estimators                    1 0.004425 0.004425  845.673  < 2e-16 ***
## min_samples_split               1 0.000125 0.000125   23.930 1.51e-05 ***
## max_features                    1 0.000000 0.000000    0.002 0.966082    
## max_depth                       1 0.018264 0.018264 3490.153  < 2e-16 ***
## criterion                       1 0.000067 0.000067   12.768 0.000902 ***
## bootstrap                       1 0.000046 0.000046    8.877 0.004783 ** 
## n_estimators:min_samples_split  1 0.000035 0.000035    6.738 0.012948 *  
## n_estimators:max_features       1 0.000000 0.000000    0.086 0.771145    
## n_estimators:max_depth          1 0.000001 0.000001    0.246 0.622328    
## n_estimators:criterion          1 0.000018 0.000018    3.401 0.072204 .  
## n_estimators:bootstrap          1 0.000092 0.000092   17.648 0.000135 ***
## min_samples_split:max_features  1 0.000002 0.000002    0.455 0.503434    
## min_samples_split:max_depth     1 0.000044 0.000044    8.473 0.005746 ** 
## min_samples_split:criterion     1 0.000019 0.000019    3.615 0.064130 .  
## min_samples_split:bootstrap     1 0.000144 0.000144   27.431 4.91e-06 ***
## max_features:max_depth          1 0.000001 0.000001    0.182 0.671464    
## max_features:criterion          1 0.000002 0.000002    0.477 0.493793    
## max_features:bootstrap          1 0.000017 0.000017    3.208 0.080475 .  
## max_depth:criterion             1 0.000034 0.000034    6.513 0.014434 *  
## max_depth:bootstrap             1 0.000019 0.000019    3.647 0.063026 .  
## criterion:bootstrap             1 0.000007 0.000007    1.282 0.264008    
## Residuals                      42 0.000220 0.000005                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In conclusion, interaction analysis enables us to construct ANOVA with double interactions containing only significant effects:

```
##                                Df   Sum Sq  Mean Sq  F value   Pr(>F)    
## max_depth                       1 0.018264 0.018264 3156.545  < 2e-16 ***
## n_estimators                    1 0.004425 0.004425  764.839  < 2e-16 ***
## min_samples_split               1 0.000125 0.000125   21.643 2.23e-05 ***
## criterion                       1 0.000067 0.000067   11.547 0.001295 ** 
## min_samples_split:bootstrap     2 0.000190 0.000095   16.419 2.82e-06 ***
## n_estimators:bootstrap          1 0.000092 0.000092   15.961 0.000201 ***
## max_depth:min_samples_split     1 0.000044 0.000044    7.663 0.007748 ** 
## n_estimators:min_samples_split  1 0.000035 0.000035    6.094 0.016824 *  
## max_depth:criterion             1 0.000034 0.000034    5.891 0.018653 *  
## Residuals                      53 0.000307 0.000006                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Center Points

Up to this point we used 2^6 factorial design and tried to determine significant effects on model `accuracy` using linear fits. Now center points for numeric variables are introduced to determine whether any curvature with respect to the response variable is present. Firstly we take a look at box plots:



Box plots indicate that curvature is present to some extent. Nonetheless, if a linear model without an intercept is built for the whole data set (original 2^6-factorial design with added center points), one can see that factors `n_estimators`, `max_depth`, `min_samples_split` are still essential even as linear terms.

```
##
## Call:
## lm.default(formula = accuracy ~ -1 + n_estimators + min_samples_split +
##     max_depth, data = df_all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34000 -0.03888  0.11356  0.20898  0.51403
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

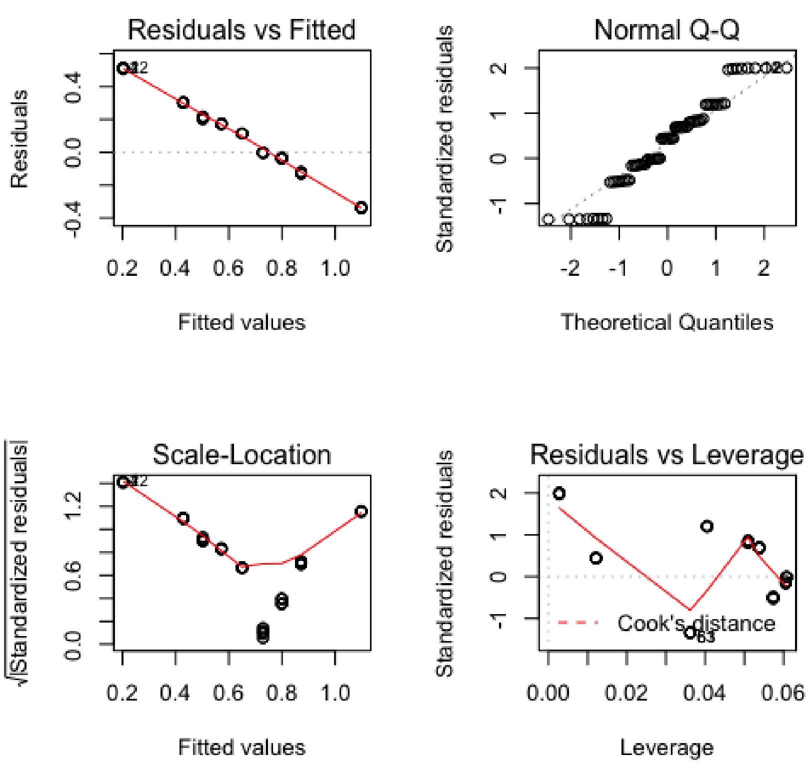
```
## n_estimators      0.0004619  0.0001205   3.834 0.000276 ***
## min_samples_split 0.0187151  0.0033063   5.660 3.19e-07 ***
## max_depth         0.0123202  0.0016625   7.411 2.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2569 on 69 degrees of freedom
## Multiple R-squared:  0.8846, Adjusted R-squared:  0.8796
## F-statistic: 176.3 on 3 and 69 DF,  p-value: < 2.2e-16
```

Linear Regression

Pure Linear Fit

Finally, we perform a linear fit for our design data with center points.

```
##
## Call:
## lm.default(formula = accuracy ~ -1 + n_estimators + min_samples_split +
##     max_depth, data = df_fit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34000 -0.03888  0.11356  0.20898  0.51403
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## n_estimators      0.0004619  0.0001205   3.834 0.000276 ***
## min_samples_split 0.0187151  0.0033063   5.660 3.19e-07 ***
## max_depth         0.0123202  0.0016625   7.411 2.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2569 on 69 degrees of freedom
## Multiple R-squared:  0.8846, Adjusted R-squared:  0.8796
## F-statistic: 176.3 on 3 and 69 DF,  p-value: < 2.2e-16
```



We also note that residuals of this linear model are normally distributed and Breusch-Pagan test against heteroskedasticity enables us to accept the null hypothesis.

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals(lm.center)
## D = 0.093391, p-value = 0.1247
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm.center)
## W = 0.94956, p-value = 0.00591
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.center
## BP = 2.4168, df = 2, p-value = 0.2987
```

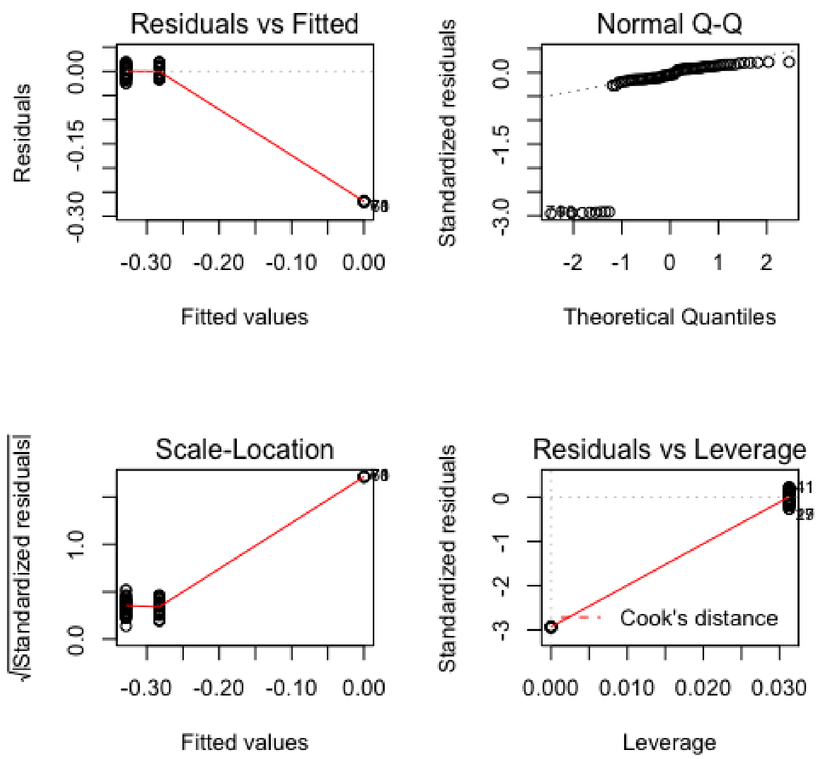
Empirical Fit with a Linear Model

Then we construct a linear model with explicit functional dependence

$$\log(Y) = \beta_1(X_1 - median(X1))^2 + \beta_2(X_2 - median(X_2))^3$$

where $Y \equiv$ `accuracy`, $X_1 \equiv$ `n_estimators`and $X_2 \equiv$ `max_depth`. The model formula is deduced empirically from box plots displaying center points. One can easily notice that the R-squared statistic has improved in comparion to the previous model. However, for this model, residuals are clearly not normally distributed and Breusch-Pagan test against heteroskedasticity rejects the null hypothesis.

```
##
## Call:
## lm.default(formula = log(accuracy) ~ -1 + I((n_estimators - median(n_estimators))^2) +
##     I((max_depth - median(max_depth))^3), data = df_fit)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.271272 -0.013378 -0.004169  0.009377  0.019571
##
## Coefficients:
##                                              Estimate Std. Error t value
## I((n_estimators - median(n_estimators))^2) -5.091e-06  1.912e-07 -26.634
## I((max_depth - median(max_depth))^3)        6.796e-06  3.400e-06   1.999
##                                            Pr(>|t|)
## I((n_estimators - median(n_estimators))^2)   <2e-16 ***
## I((max_depth - median(max_depth))^3)         0.0495 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09179 on 70 degrees of freedom
## Multiple R-squared:  0.9106, Adjusted R-squared:  0.9081
## F-statistic: 356.7 on 2 and 70 DF,  p-value: < 2.2e-16
```



Polynomial Fit

Lastly we perform a fit with orthogonal polynomials given only available data. This model surpasses the first one in terms of R-squared statistic. Moreover, it passes the test of residuals normality and accepts the null hypothesis of Breusch-Pagan test.

```
##
## Call:
## lm.default(formula = accuracy ~ poly(n_estimators, 1) + poly(max_depth,
##     2), data = df_fit)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0090615 -0.0021788 -0.0000271  0.0022641  0.0094880
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.7399328  0.0004288 1725.64   <2e-16 ***
## poly(n_estimators, 1)  0.0665237  0.0036384   18.28   <2e-16 ***
## poly(max_depth, 2)1     0.1351442  0.0036384   37.14   <2e-16 ***
## poly(max_depth, 2)2    -0.0720580  0.0036384  -19.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003638 on 68 degrees of freedom
```

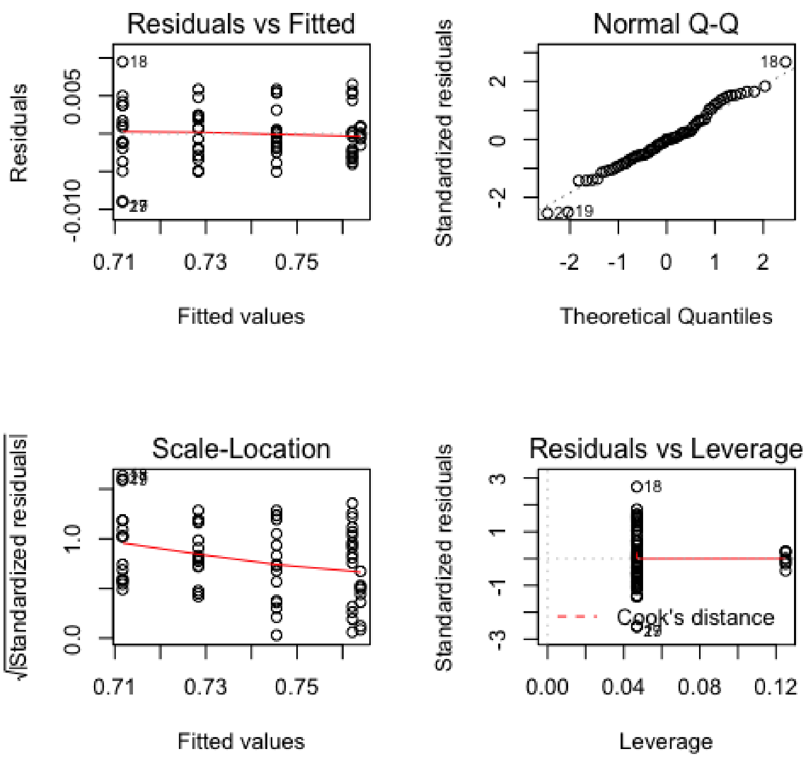
```
## Multiple R-squared:  0.9687, Adjusted R-squared:  0.9673
## F-statistic: 702.1 on 3 and 68 DF,  p-value: < 2.2e-16
```

Tests results:

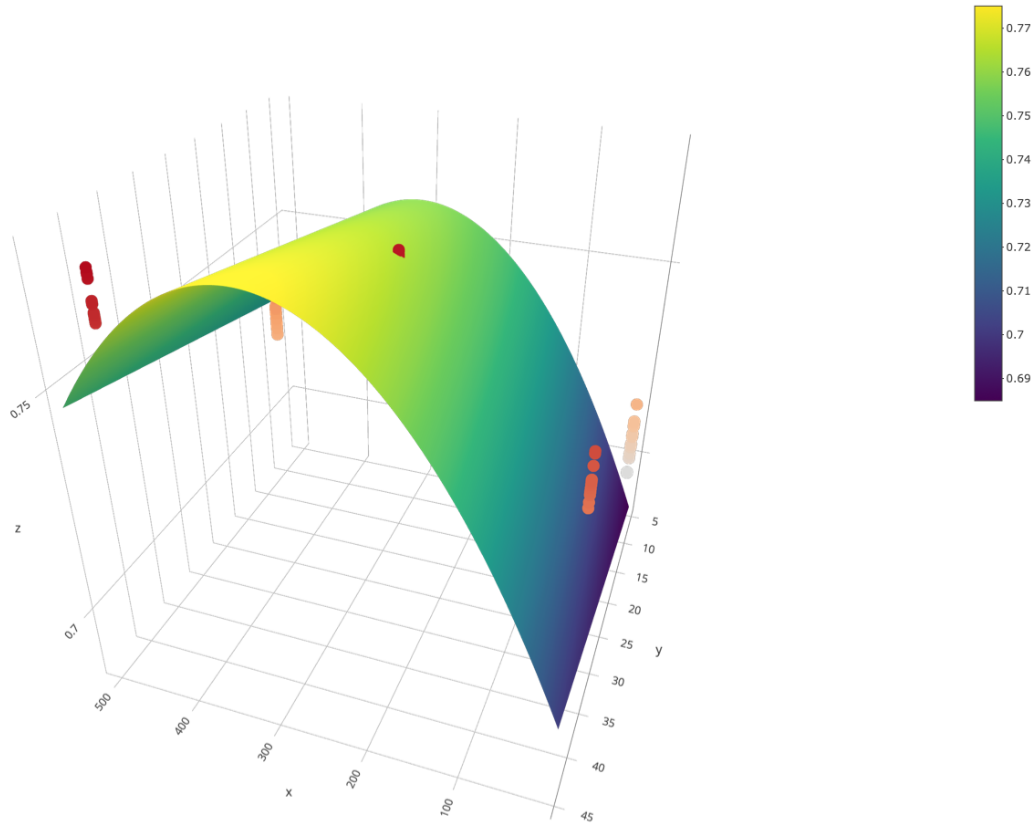
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals(lm.numeric)
## D = 0.088144, p-value = 0.1803

##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm.numeric)
## W = 0.98213, p-value = 0.3986

##
##  studentized Breusch-Pagan test
##
## data:  lm.numeric
## BP = 6.6567, df = 3, p-value = 0.08368
```



Finally we choose this model to fit `accuracy` of the Random Forest classifier. Here we also present a 3D visualization of this fit:



Conclusion

The polynomial fit we chose has provided us with arguably best parameters for `n_estimators` and `max_depth`. This report is conlcluded with results of the small grid search with `n_estimators` = 314, `max_depth` = 45: with `min_samples_split` = 4, `max_features` = "sqrt", `criterion` = "gini", `bootstrap` = True we obtain `accuracy` equal to 0.7691845. Computed accuracy is higher than that of our prior grid search, so this project was not done for nothing and has "beared its fruit".