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 parameters and their interactions and look for the most efficient model given their range which was present in the grid.

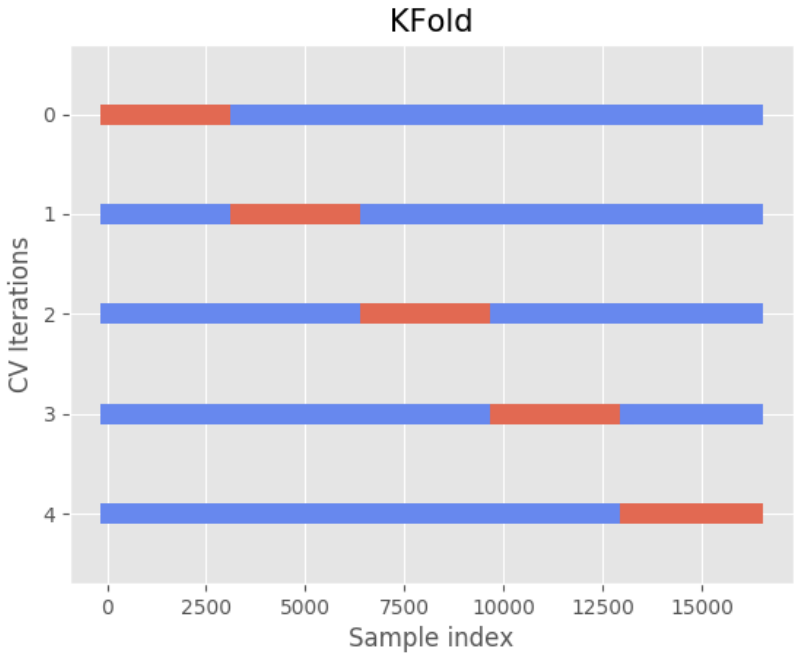


Figure 1: K-fold Cross-Validation Visualization

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
```

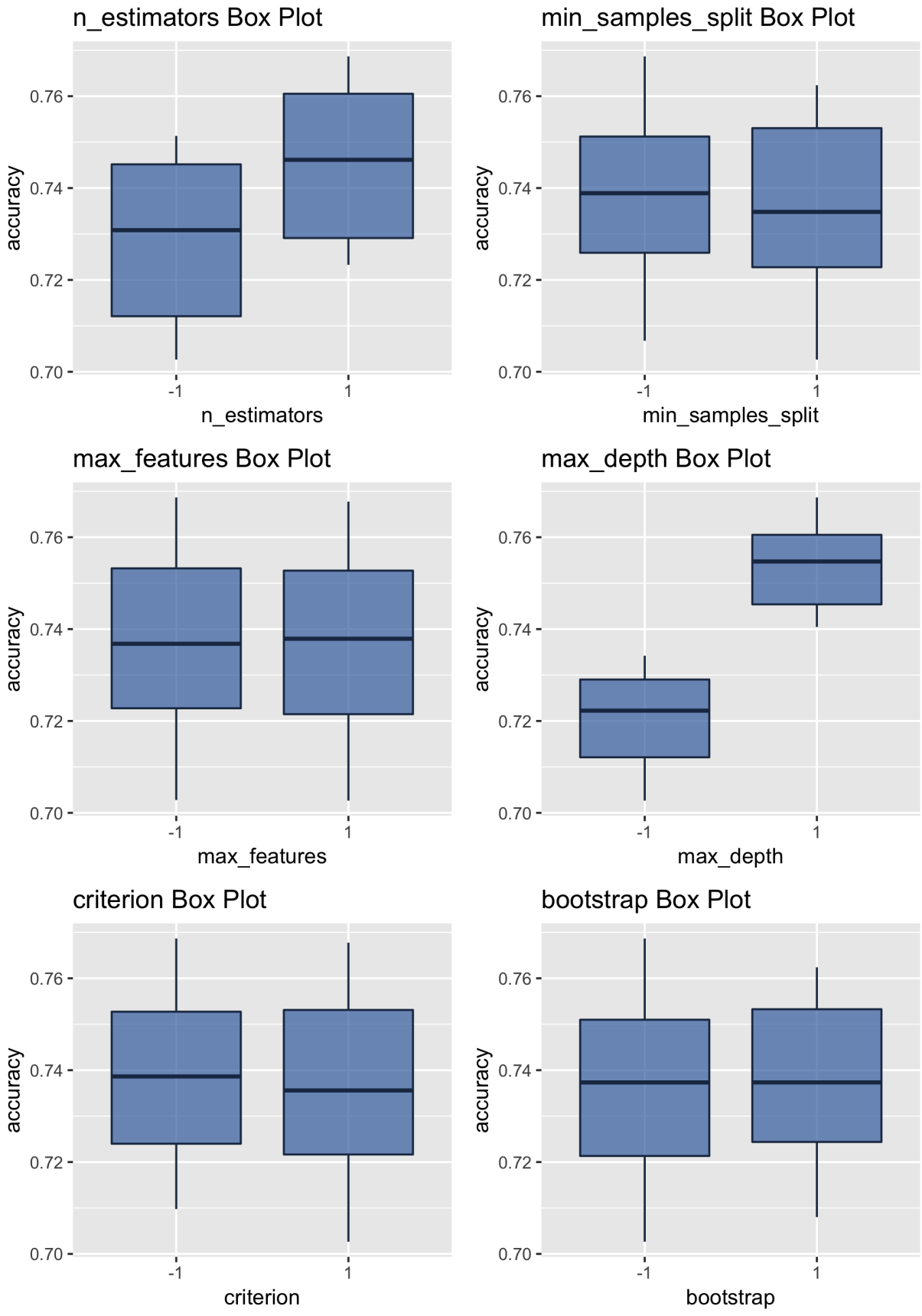
First Look into the Data

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 are 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
```

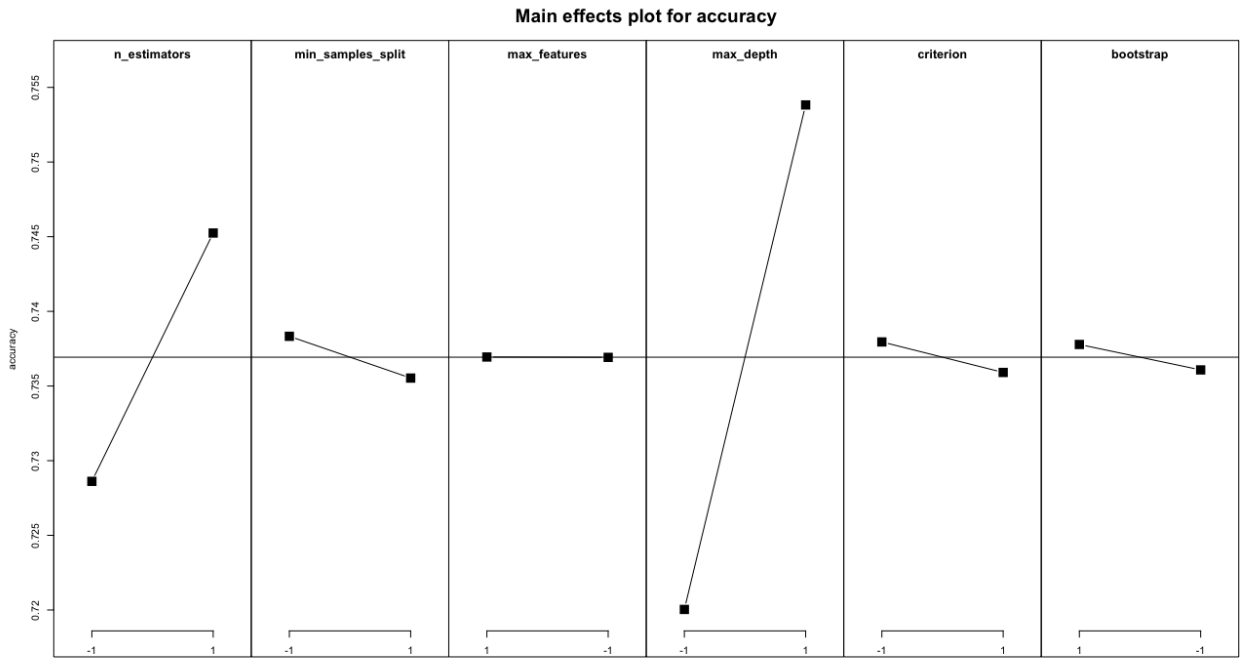


Figure 2: Main Effects Plot

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.

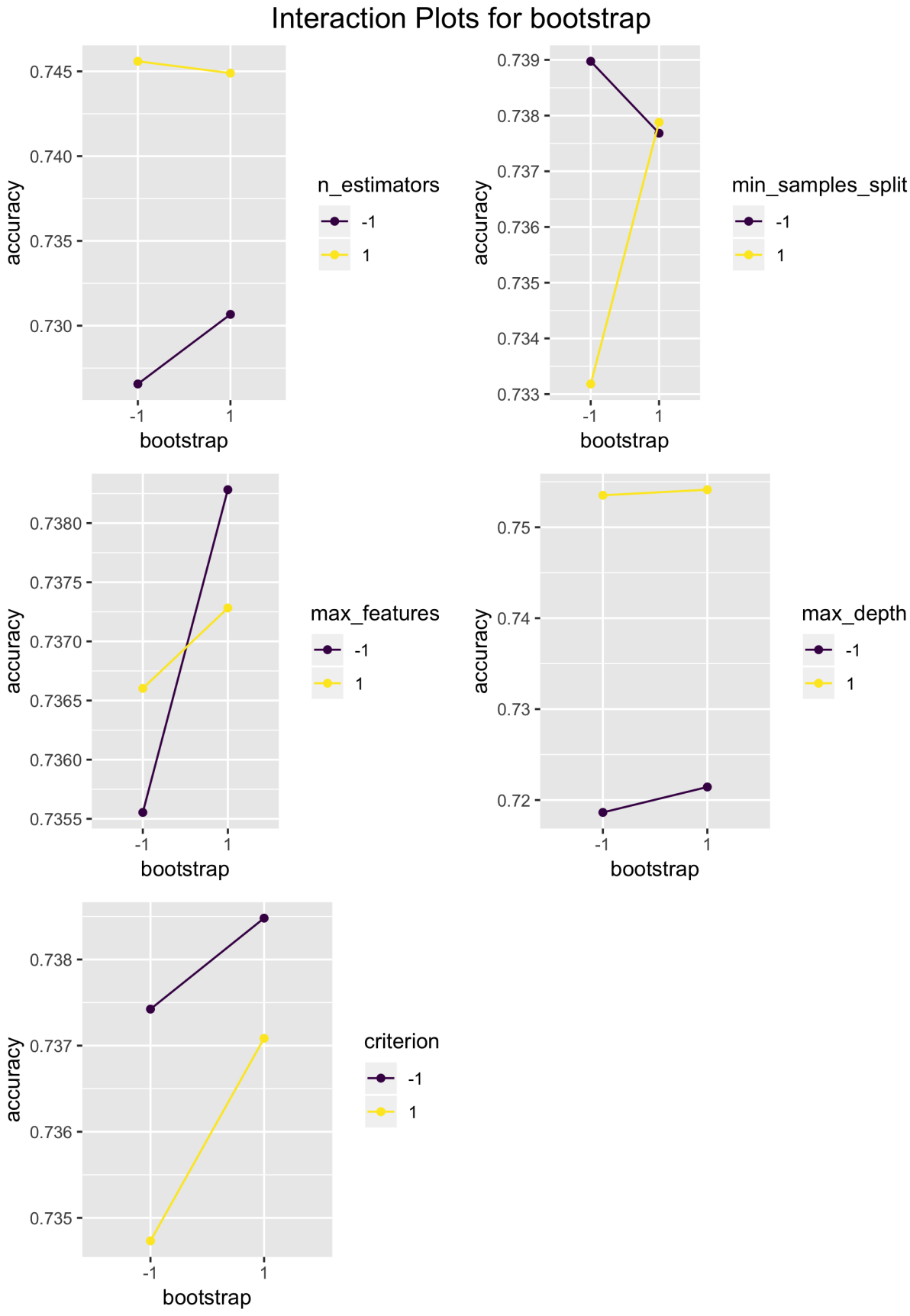
Analysis of Interactions

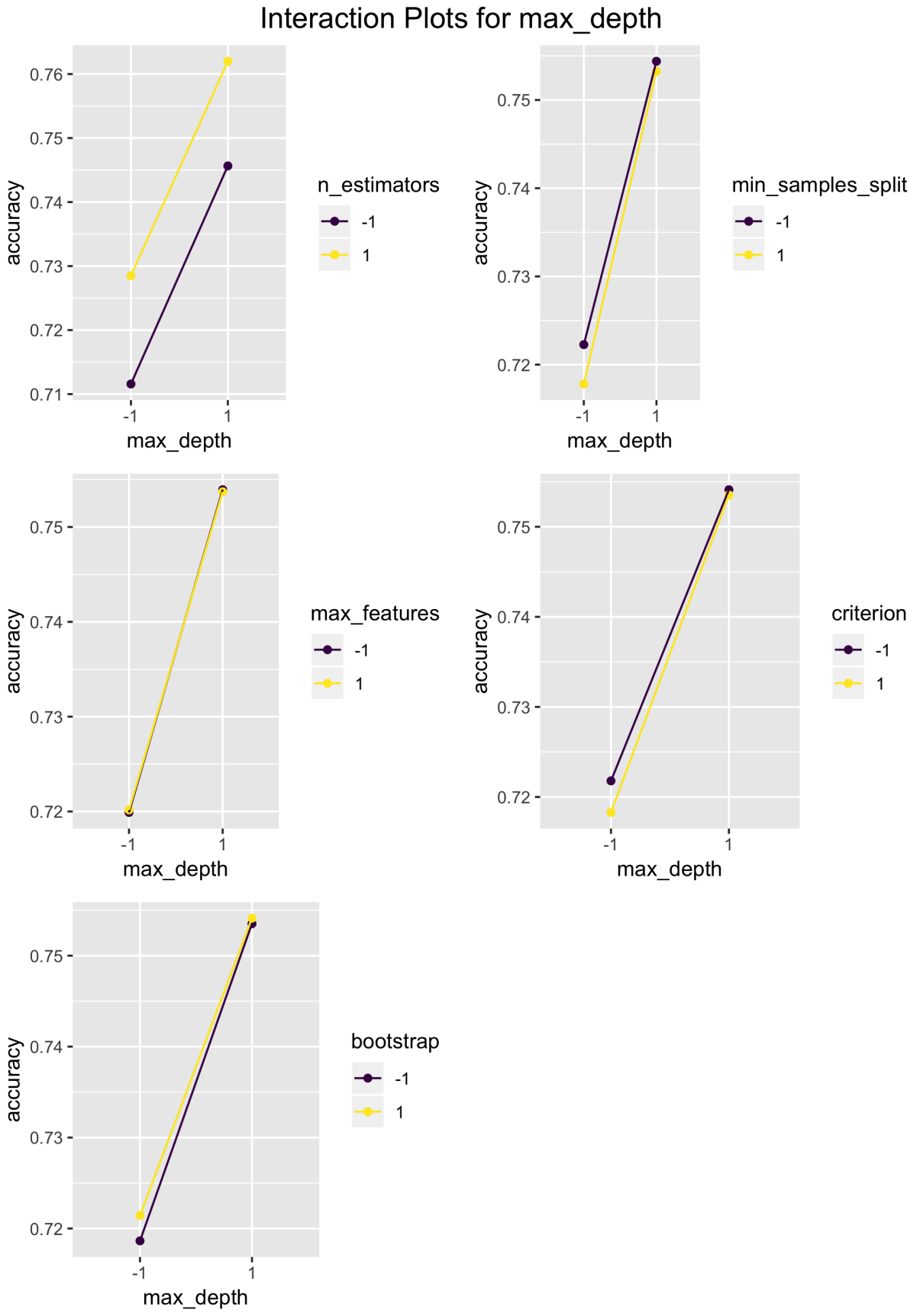
Visual Analysis of Interactions

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

Table 5: Empirical Assessment of Pairwise Interactions

Variables	`bootstrap`	`max_depth`	`max_features`	`min_samples_split`	`criterion`	`n_estimators`
`bootstrap`	*NA*	-	**X**	**X**	-	**X**
`max_depth`	-	*NA*	-	-	-	-
`max_features`	**X**	-	*NA*	**X**	**X**	-
`min_samples_split`	**X**	-	**X**	*NA*	**X**	-
`criterion`	-	-	**X**	**X**	*NA*	-
`n_estimators`	**X**	-	-	-	-	*NA*


Interaction Plots for bootstrap
accuracy
bootstrap
n_estimators
min_samples_split
max_features
max_depth
criterion
-1
1
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


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

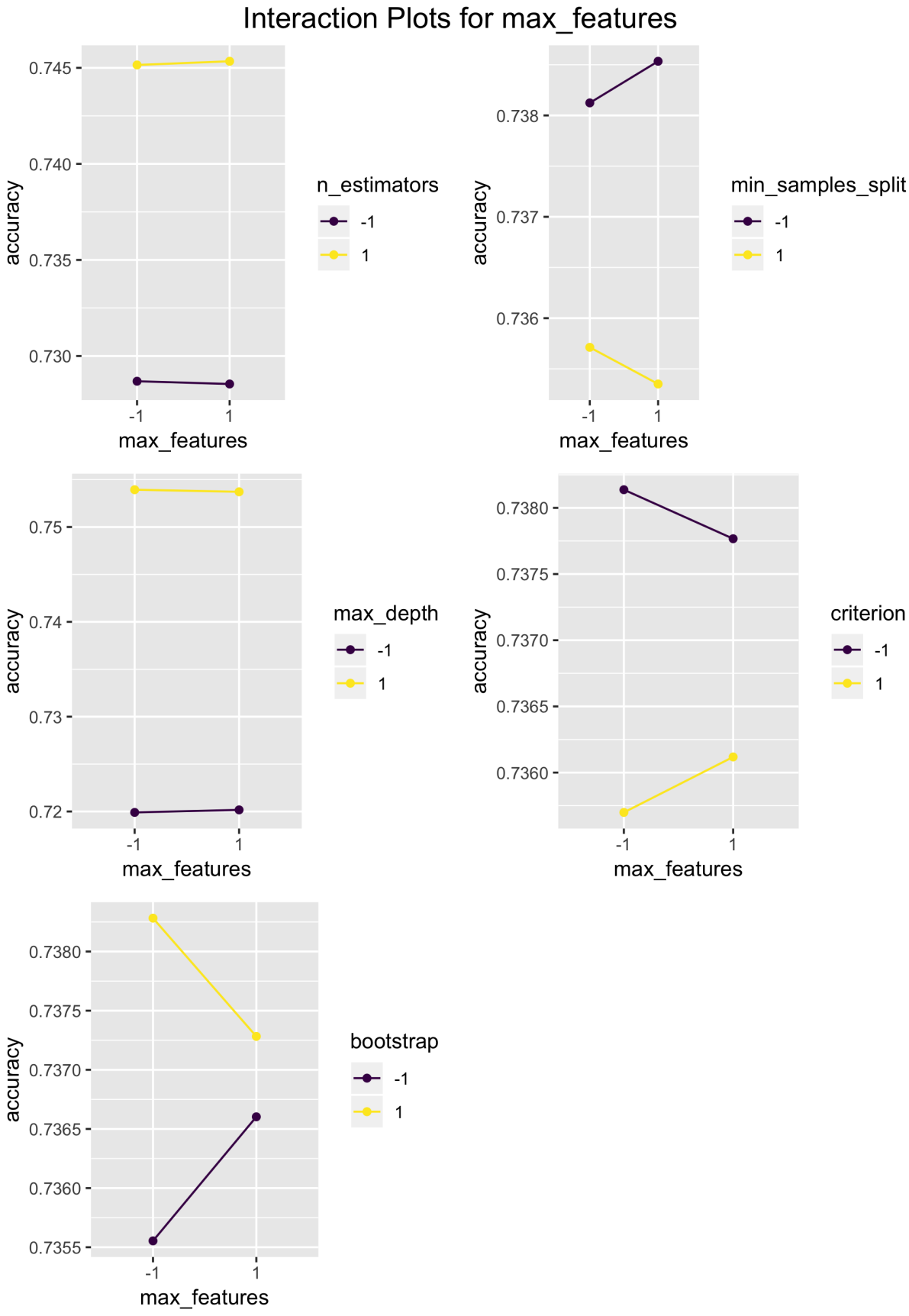

Interaction Plots for max_features
accuracy
max_features
n_estimators
min_samples_split
max_depth
criterion
bootstrap
-1
1
0.745
0.740
0.735
0.730
0.738
0.737
0.736
0.75
0.74
0.73
0.72
0.7380
0.7375
0.7370
0.7365
0.7360
0.7355

Figure 3:

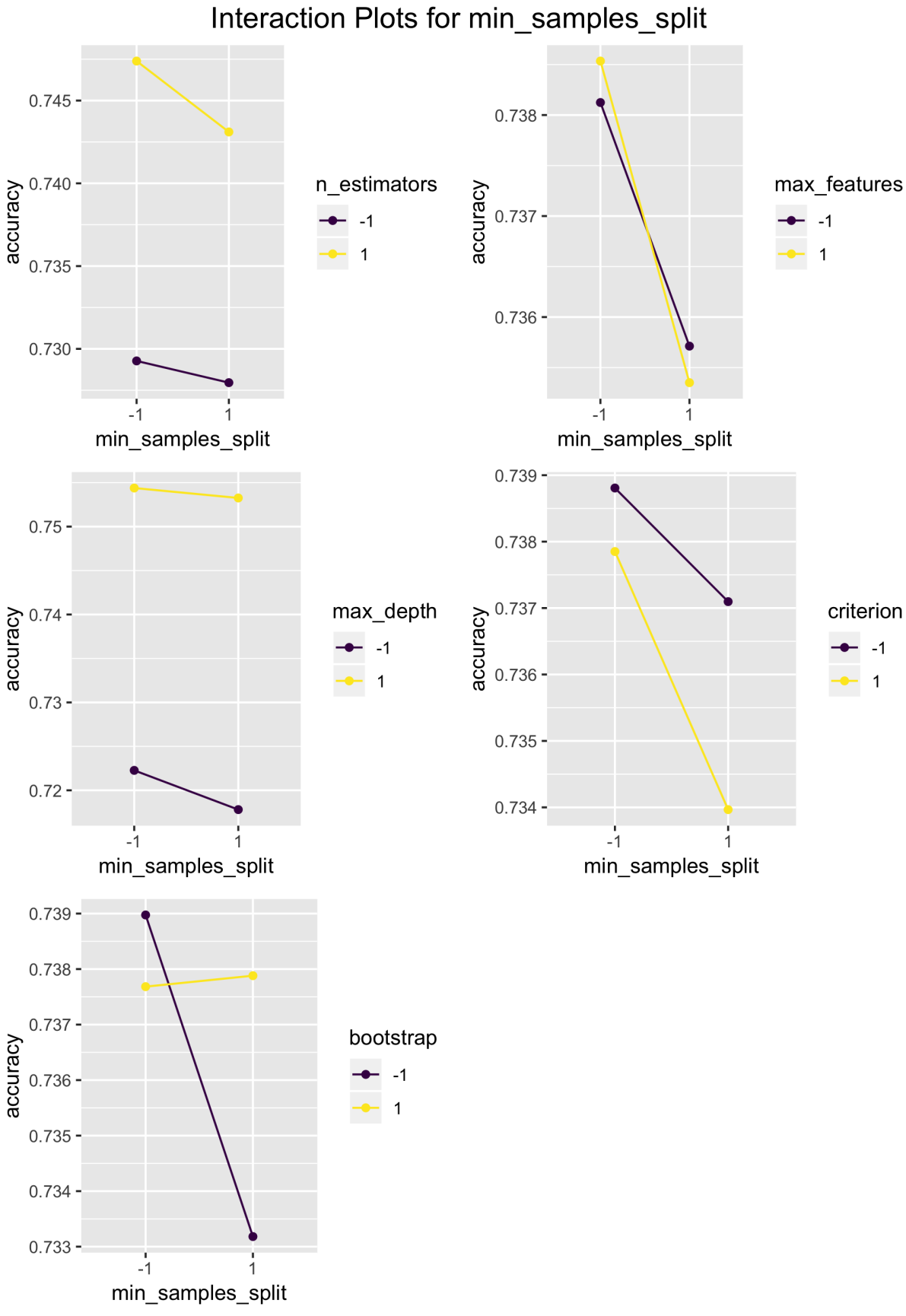

Interaction Plots for min_samples_split
accuracy
min_samples_split
n_estimators
max_features
max_depth
criterion
bootstrap
-1
1
0.745
0.740
0.735
0.730
0.738
0.737
0.736
0.75
0.74
0.73
0.72
0.739
0.738
0.737
0.736
0.735
0.734
0.733

Figure 4:

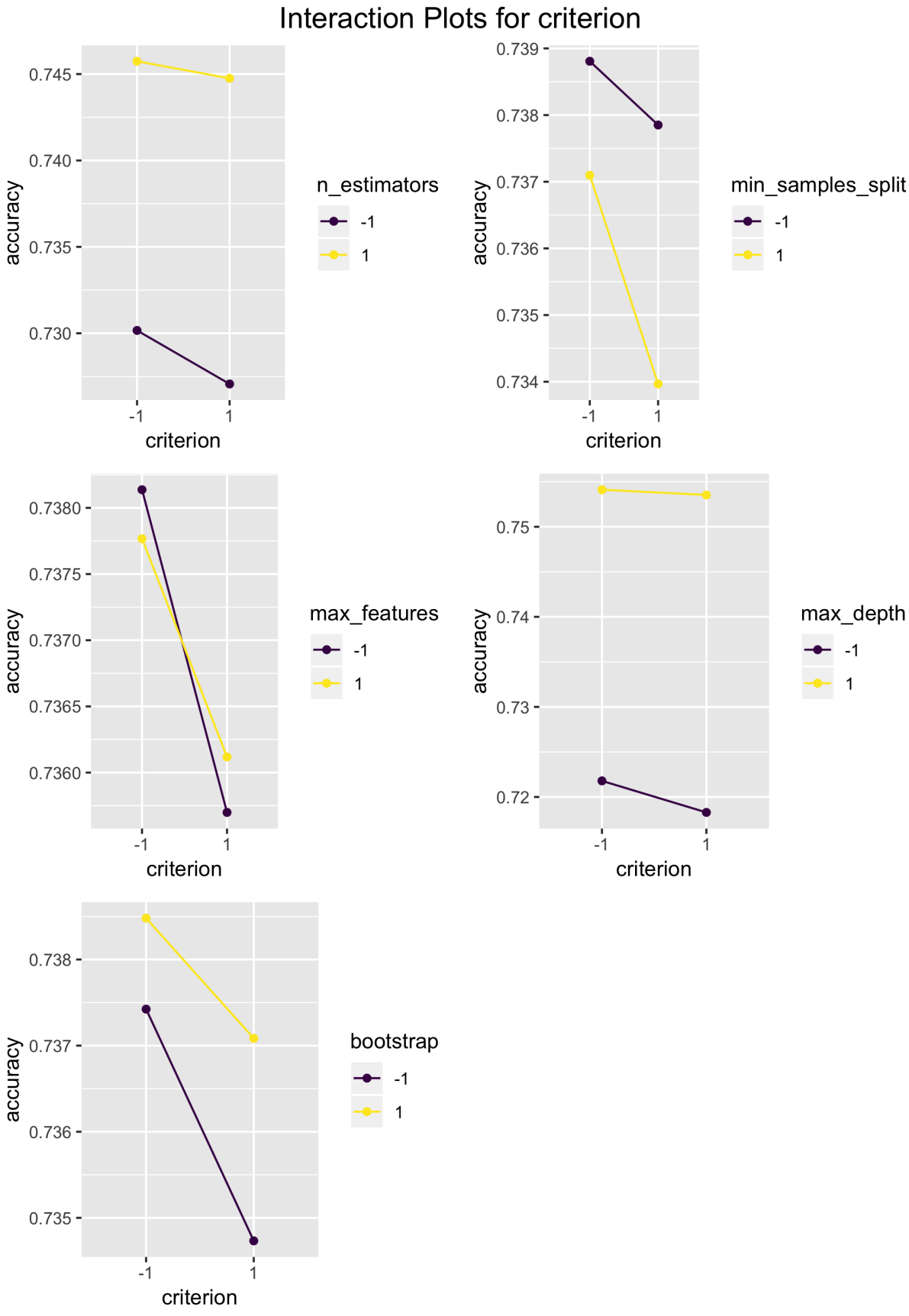

Interaction Plots for criterion
accuracy
criterion
n_estimators
min_samples_split
max_features
max_depth
bootstrap
-1
1
0.745
0.740
0.735
0.730
0.739
0.738
0.737
0.736
0.734
0.7380
0.7375
0.7370
0.7365
0.7360
0.75
0.74
0.73
0.72

Figure 5:

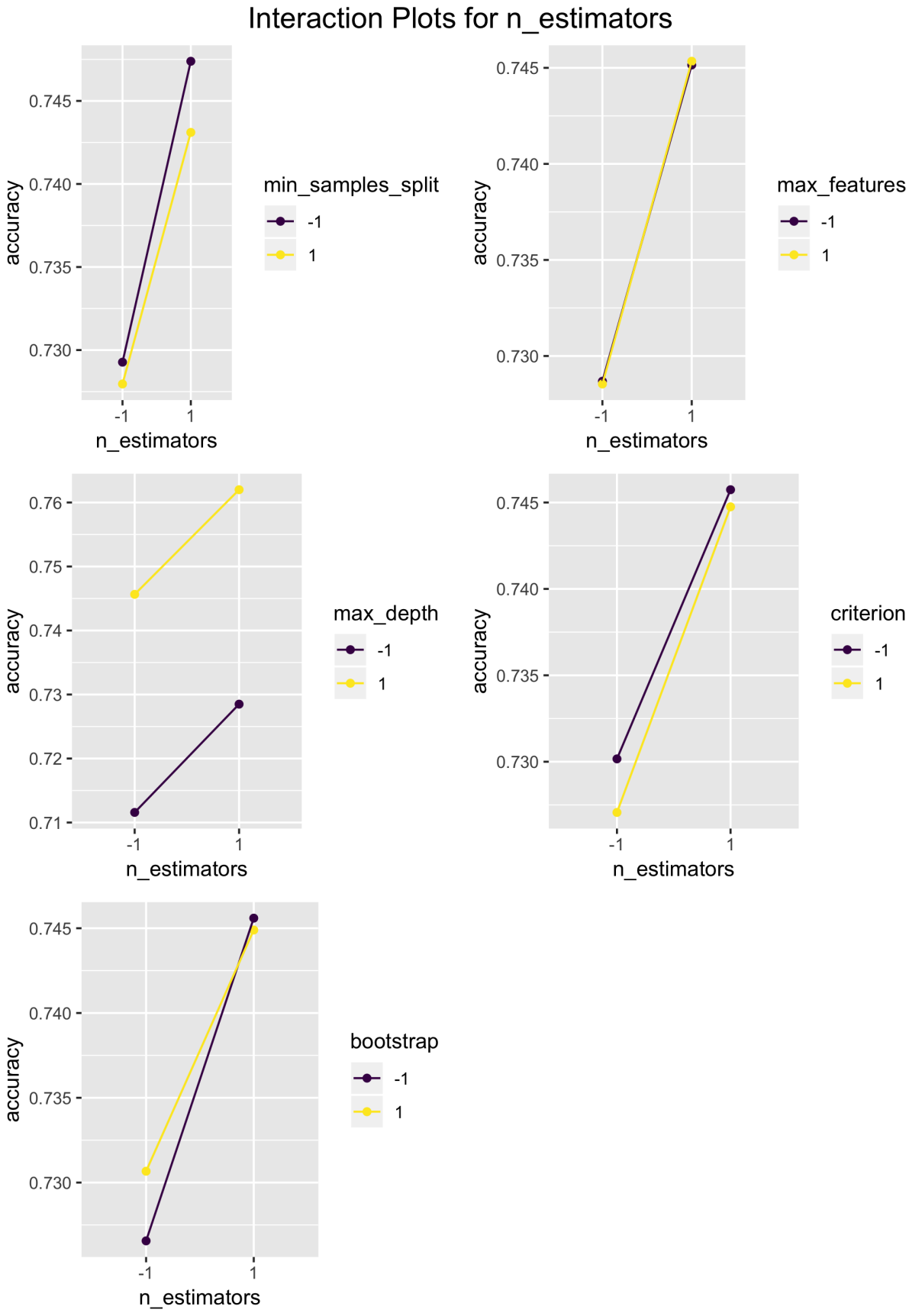

Interaction Plots for n_estimators
accuracy
n_estimators
min_samples_split
max_features
max_depth
criterion
bootstrap
-1
1
0.745
0.740
0.735
0.730
0.76
0.75
0.74
0.73
0.72
0.71

Figure 6: