Data Matrix Modeling

Version: BINARY CLASSIFIER

Updated 12/18/2017 by E. Gates. See readme on github

WIP: semi-supervised feature and genetic variable selection

WIP: custom model selection

WIP: Compare to null model for imbalanced datasets

BUG: preprocess crashes when trying too many image features

BUG: statement "No univariate significant values" not printing.

Compiled: 2018-Jan-23 11:28:09

Target Variable: MutationalStatus

Input File: DF_mutation_Dec182017/pyradiomicsout.csv

Target and inputs are column headings in csv file, everything else is ignored

Pre-processing data:

By default removes columns with zero variance and discards variables correlated >0.8

Pre-Processing results:

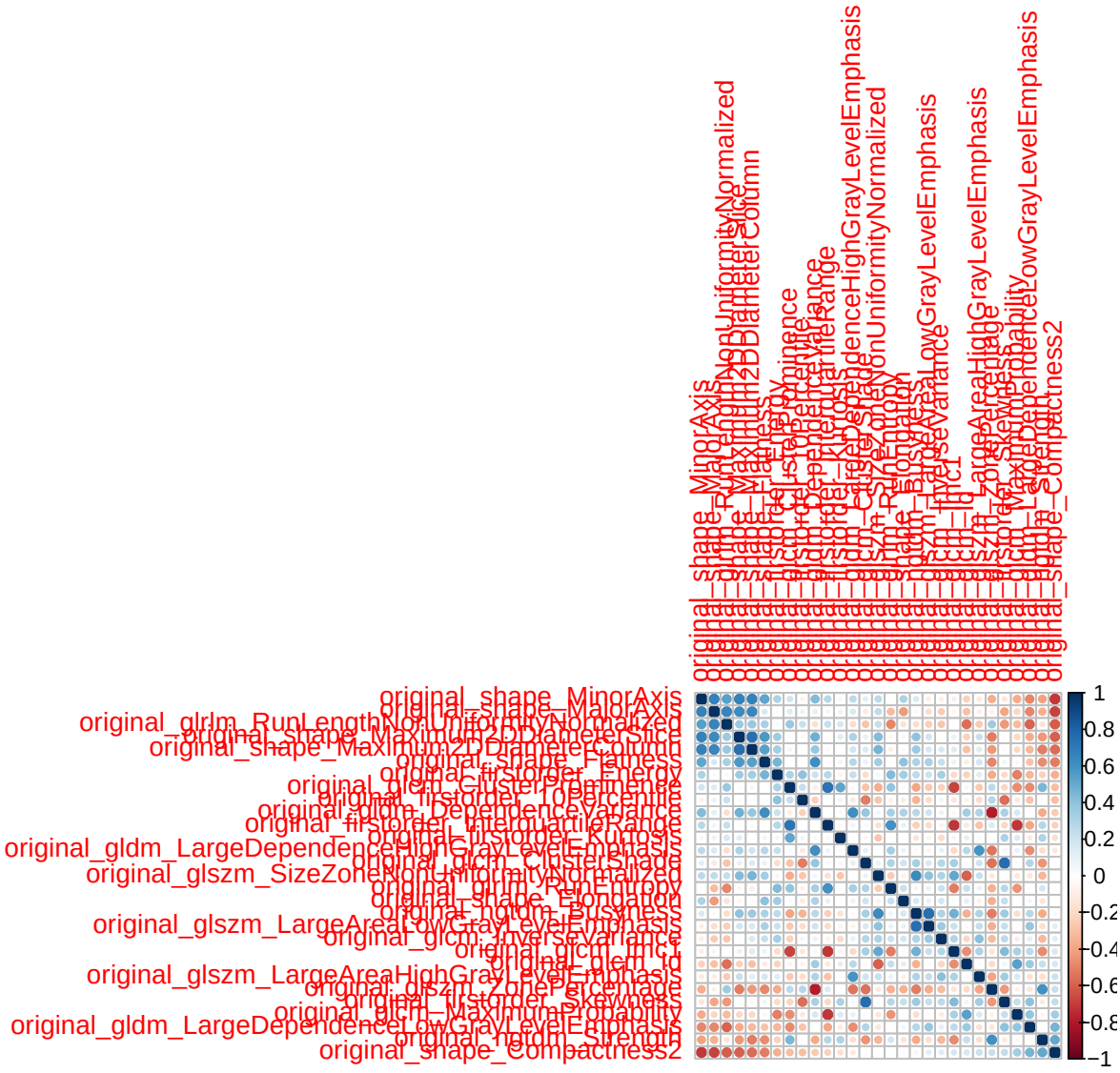
Started with 113 non-NA variables.

```
## Created from 29 samples and 113 variables
##
## Pre-processing:
##   - centered (29)
##   - ignored (0)
##   - removed (84)
##   - scaled (29)
```

29 remained after pre-processing

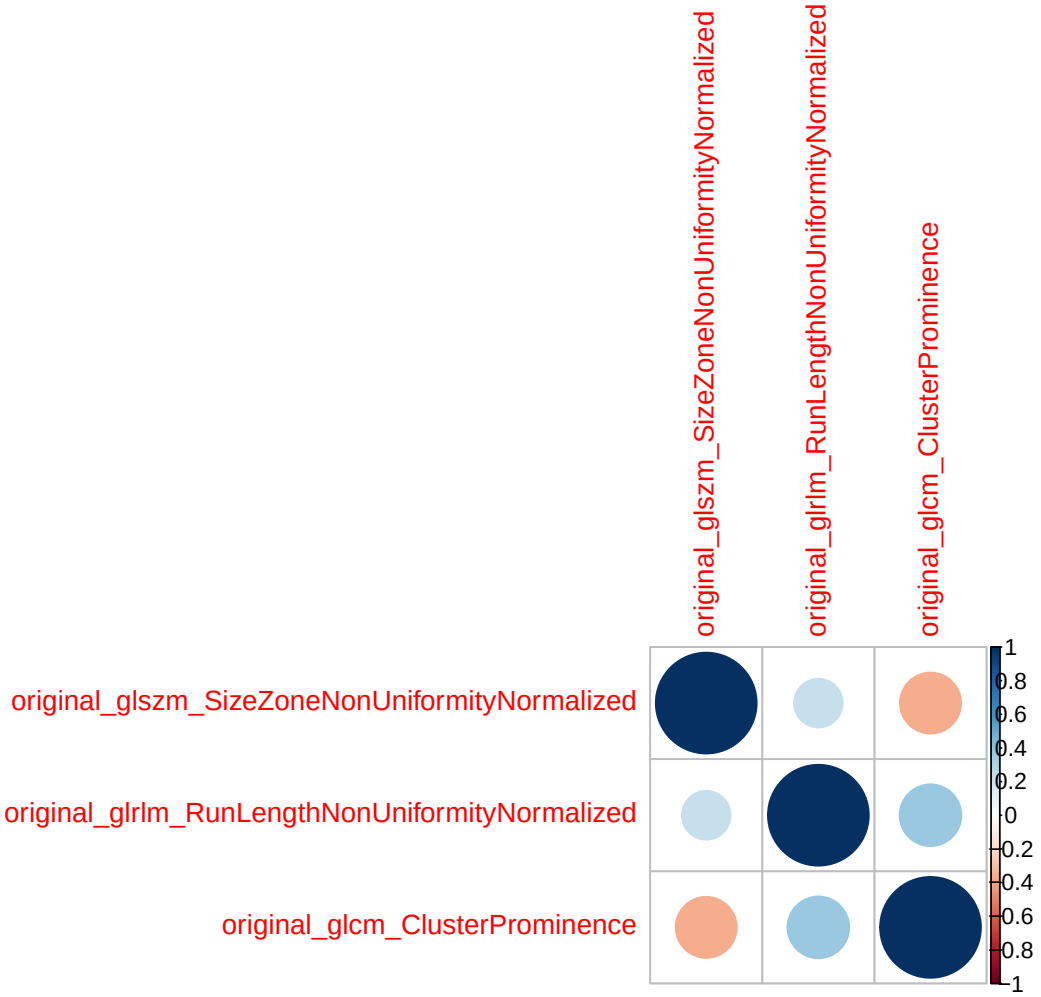
Variable Selection:

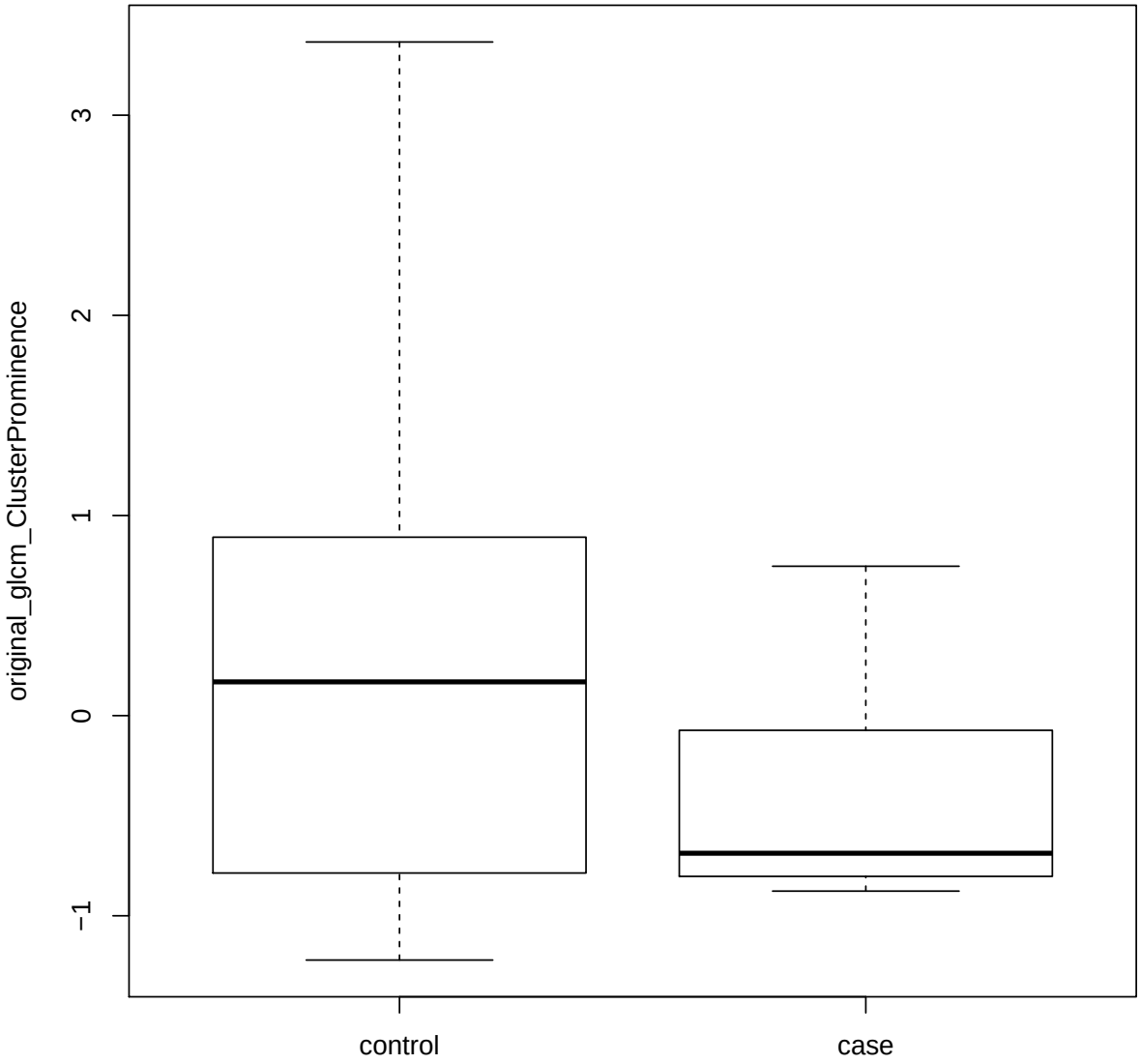
Default is Boruta and Wilcoxon test (P value cutoff 0.20/ 29). Wilcoxon currently only tests numeric input variables

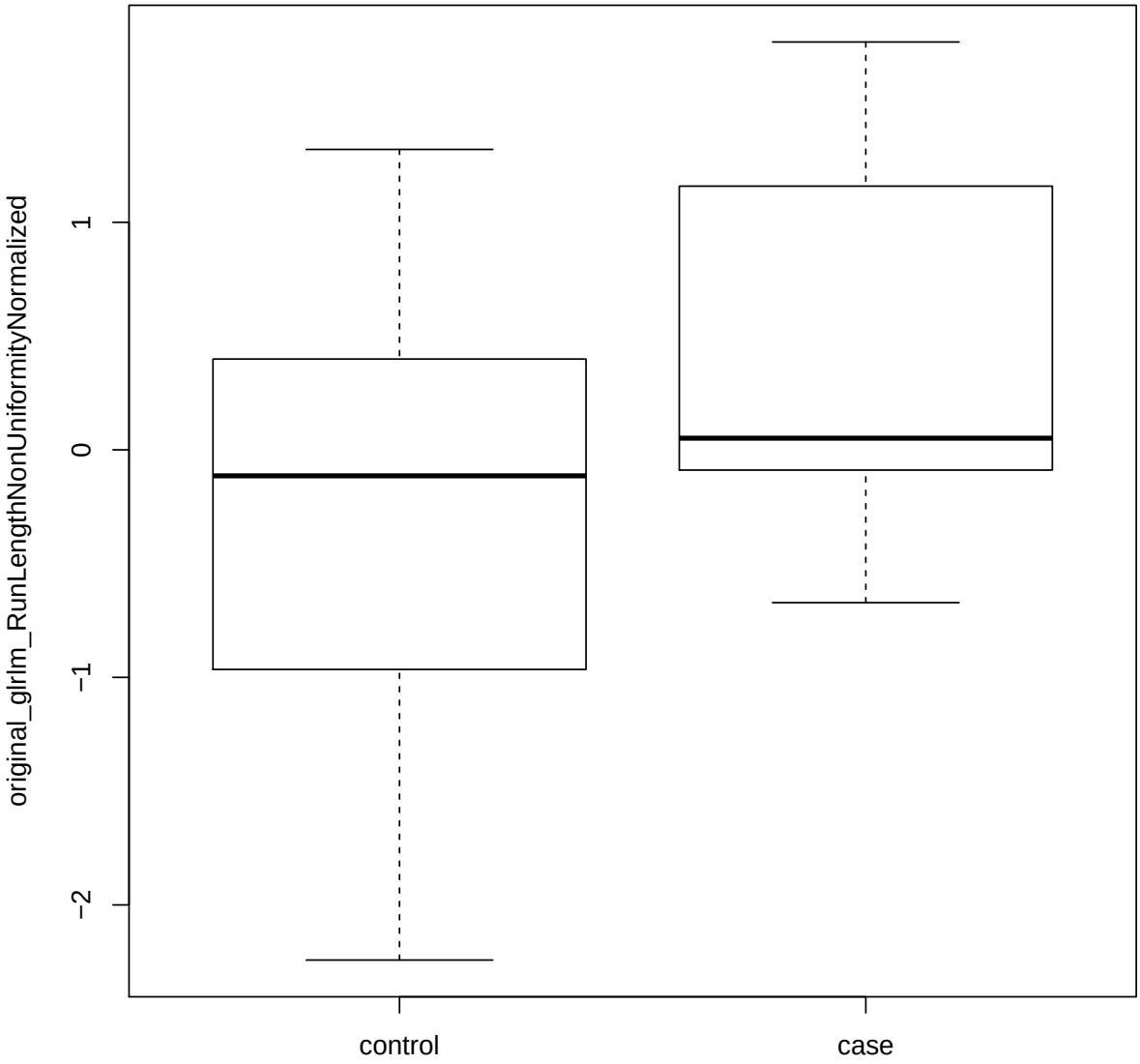


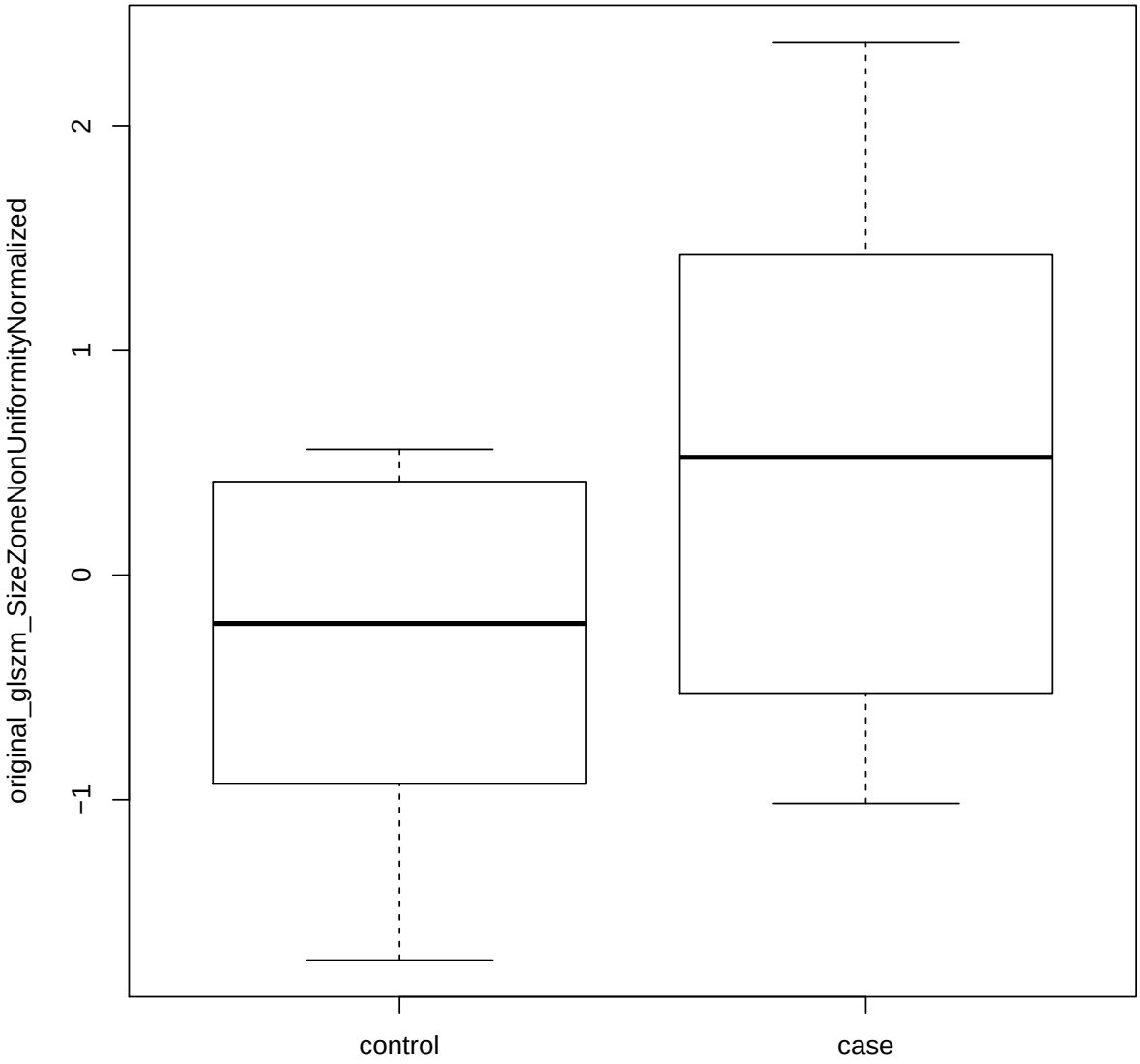

```
## [1] "Finished Boruta variable selection"
## Boruta performed 382 iterations in 3.991 secs.
##  3 attributes confirmed important:
## original_glcm_ClusterProminence,
## original_glrlm_RunLengthNonUniformityNormalized,
## original_glszm_SizeZoneNonUniformityNormalized;
##  26 attributes confirmed unimportant:
## original_firstorder_10Percentile, original_firstorder_Energy,
## original_firstorder_InterquartileRange,
## original_firstorder_Kurtosis, original_firstorder_Skewness and 21
## more;
```

Correlations for Boruta method




original_glcm_ClusterProminence
3
2
1
0
−1
control
case


original_glrlm_RunLengthNonUniformityNormalized
1
0
−1
−2
control
case

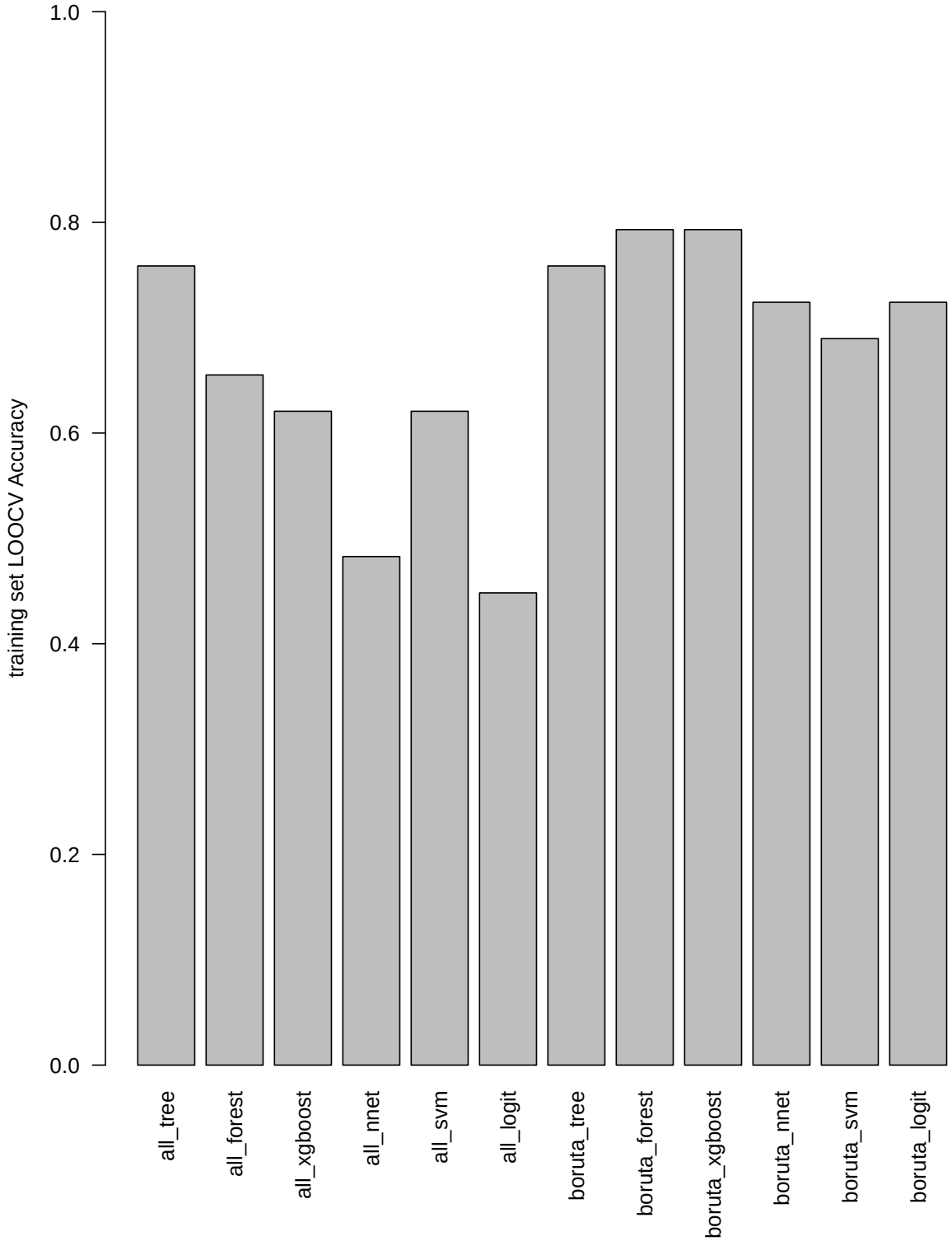



```
## No P values < 0.00690
```

Modeling using tree, forest, xgboost, nnet, svm, logit.

Use Leave-one-out cross validation: TRUE

```
## note: only 2 unique complexity parameters in default grid. Truncating the grid to 2 .
```



Best model(s): boruta_forest, boruta_xgboost

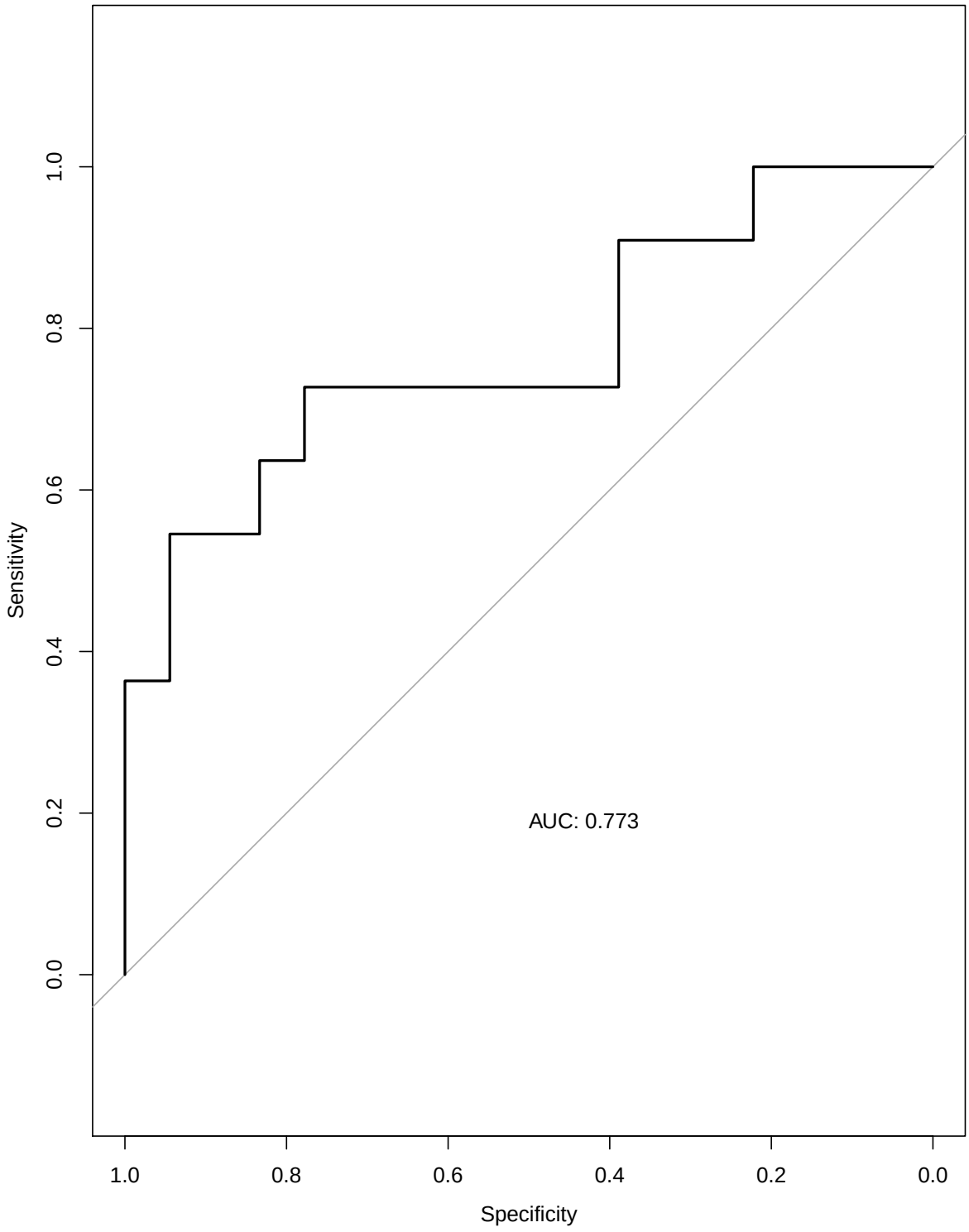
Accuracy: 0.7931

```
## [[1]]
```

```
## Random Forest
##
## 29 samples
##  3 predictor
##  2 classes: 'control', 'case'
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 28, 28, 28, 28, 28, 28, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.7931034  0.5271739
##   3     0.7586207  0.4586667
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 2.
##
## [[2]]
## eXtreme Gradient Boosting
##
## 29 samples
##  3 predictor
##  2 classes: 'control', 'case'
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 28, 28, 28, 28, 28, 28, ...
## Resampling results across tuning parameters:
##
##   nrounds  lambda  alpha  Accuracy   Kappa
##    50      0e+00   0e+00  0.7931034  0.5445026
##    50      0e+00   1e-04  0.7931034  0.5445026
##    50      0e+00   1e-01  0.7586207  0.4586667
##    50      1e-04   0e+00  0.7931034  0.5445026
##    50      1e-04   1e-04  0.7931034  0.5445026
##    50      1e-04   1e-01  0.7586207  0.4586667
##    50      1e-01   0e+00  0.7586207  0.4586667
##    50      1e-01   1e-04  0.7586207  0.4586667
##    50      1e-01   1e-01  0.7586207  0.4586667
##   100      0e+00   0e+00  0.7931034  0.5445026
##   100      0e+00   1e-04  0.7931034  0.5445026
##   100      0e+00   1e-01  0.7586207  0.4586667
##   100      1e-04   0e+00  0.7931034  0.5445026
##   100      1e-04   1e-04  0.7931034  0.5445026
##   100      1e-04   1e-01  0.7586207  0.4586667
##   100      1e-01   0e+00  0.7586207  0.4586667
##   100      1e-01   1e-04  0.7586207  0.4586667
##   100      1e-01   1e-01  0.7586207  0.4586667
##   150      0e+00   0e+00  0.7586207  0.4781491
##   150      0e+00   1e-04  0.7931034  0.5445026
##   150      0e+00   1e-01  0.7586207  0.4586667
##   150      1e-04   0e+00  0.7586207  0.4781491
##   150      1e-04   1e-04  0.7931034  0.5445026
```

```
##   150      1e-04   1e-01  0.7586207  0.4586667
##   150      1e-01   0e+00  0.7241379  0.3926702
##   150      1e-01   1e-04  0.7241379  0.3926702
##   150      1e-01   1e-01  0.7586207  0.4586667
##
## Tuning parameter 'eta' was held constant at a value of 0.3
## Accuracy was used to select the optimal model using  the largest value.
## The final values used for the model were nrounds = 50, lambda = 0, alpha
##  = 0 and eta = 0.3.
```

```
## [1] "Building ROC Curve for model boruta_forest"
## Confusion Matrix and Statistics
##
##           Reference
## Prediction control case
##    control      17    5
##    case          1    6
##
##                Accuracy : 0.7931
##                  95% CI : (0.6028, 0.9201)
##     No Information Rate : 0.6207
##     P-Value [Acc > NIR] : 0.03859
##
##                   Kappa : 0.5272
##  Mcnemar's Test P-Value : 0.22067
##
##             Sensitivity : 0.5455
##             Specificity : 0.9444
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.7727
##              Prevalence : 0.3793
##          Detection Rate : 0.2069
##    Detection Prevalence : 0.2414
##       Balanced Accuracy : 0.7449
##
##        'Positive' Class : case
##
##
## Call:
## roc.default(response = results$obs, predictor = results$case)
##
## Data: results$case in 18 controls (results$obs control) < 11 cases (results$obs case).
## Area under the curve: 0.7727
```



threshold	0.4000000
sensitivity	0.7272727
specificity	0.7777778