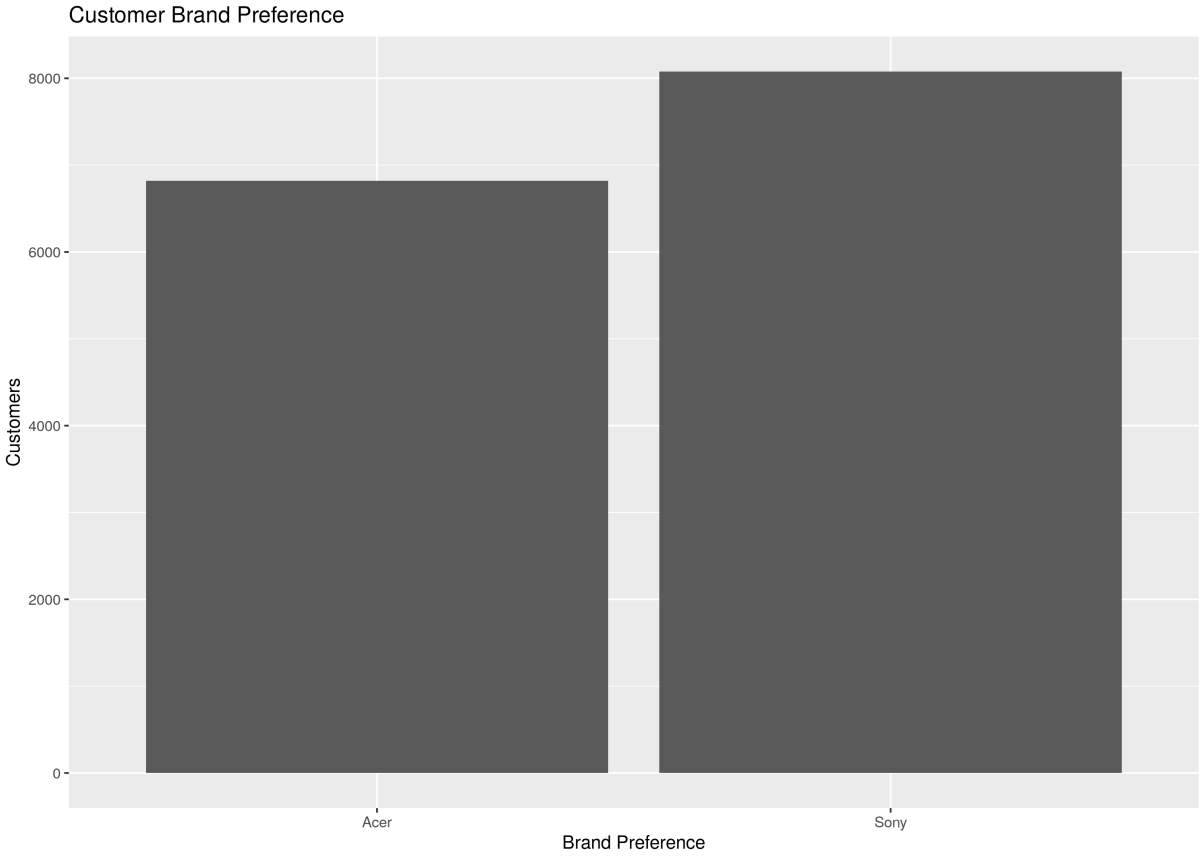
Data Analytics 2 - Predicting Brand Preference

Tuomo Kareoja

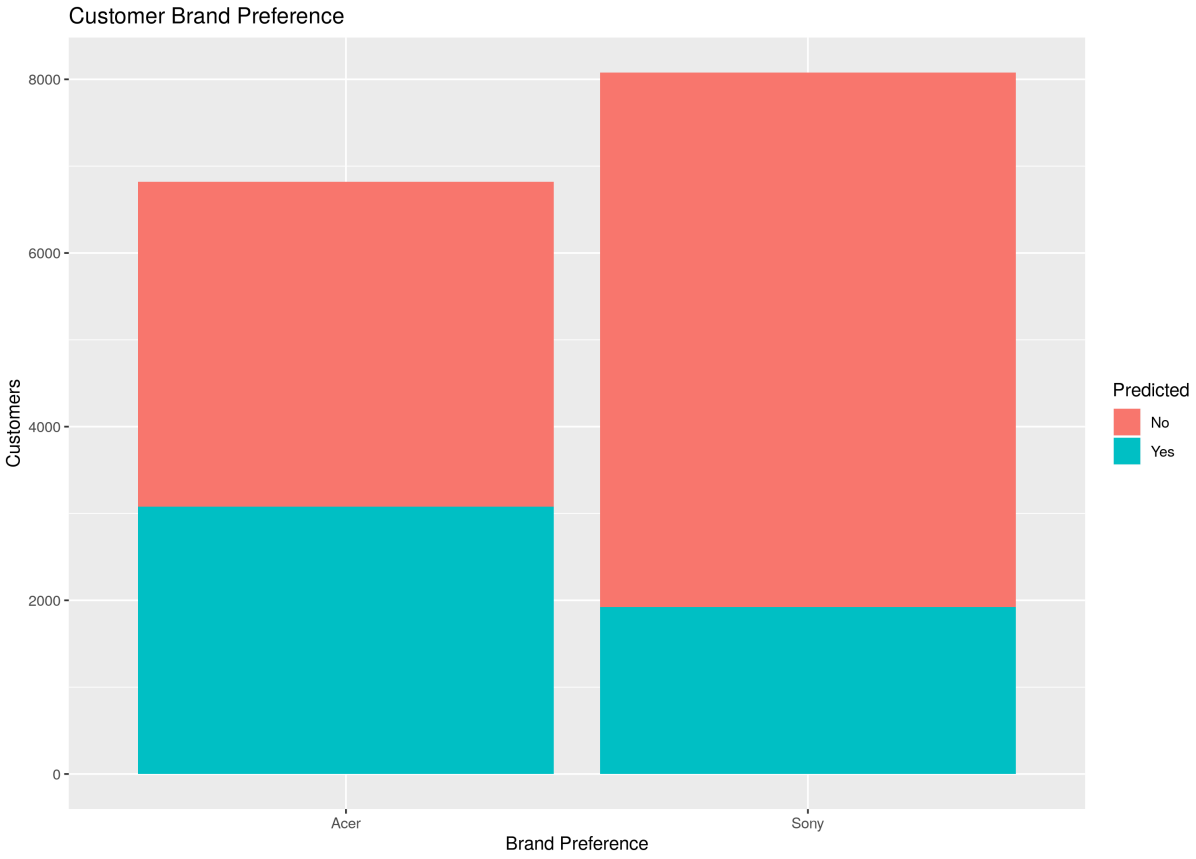
Table 1: Version history

Version Number	Changes	Date
0.1	Created basic outline without content	01.08.2019
0.2	Added visualizations, printouts and text outlines	02.08.2019

Fixed predictions

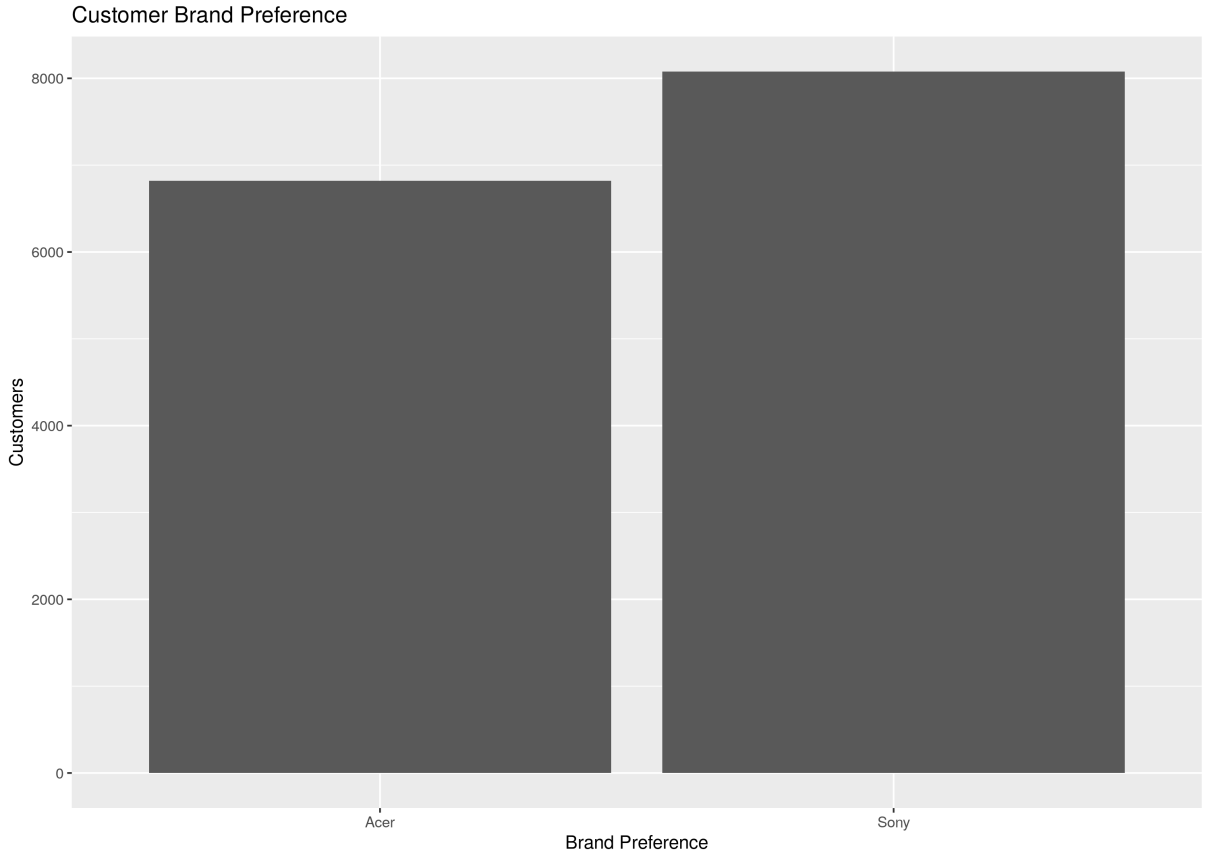


These are results after we added predictions



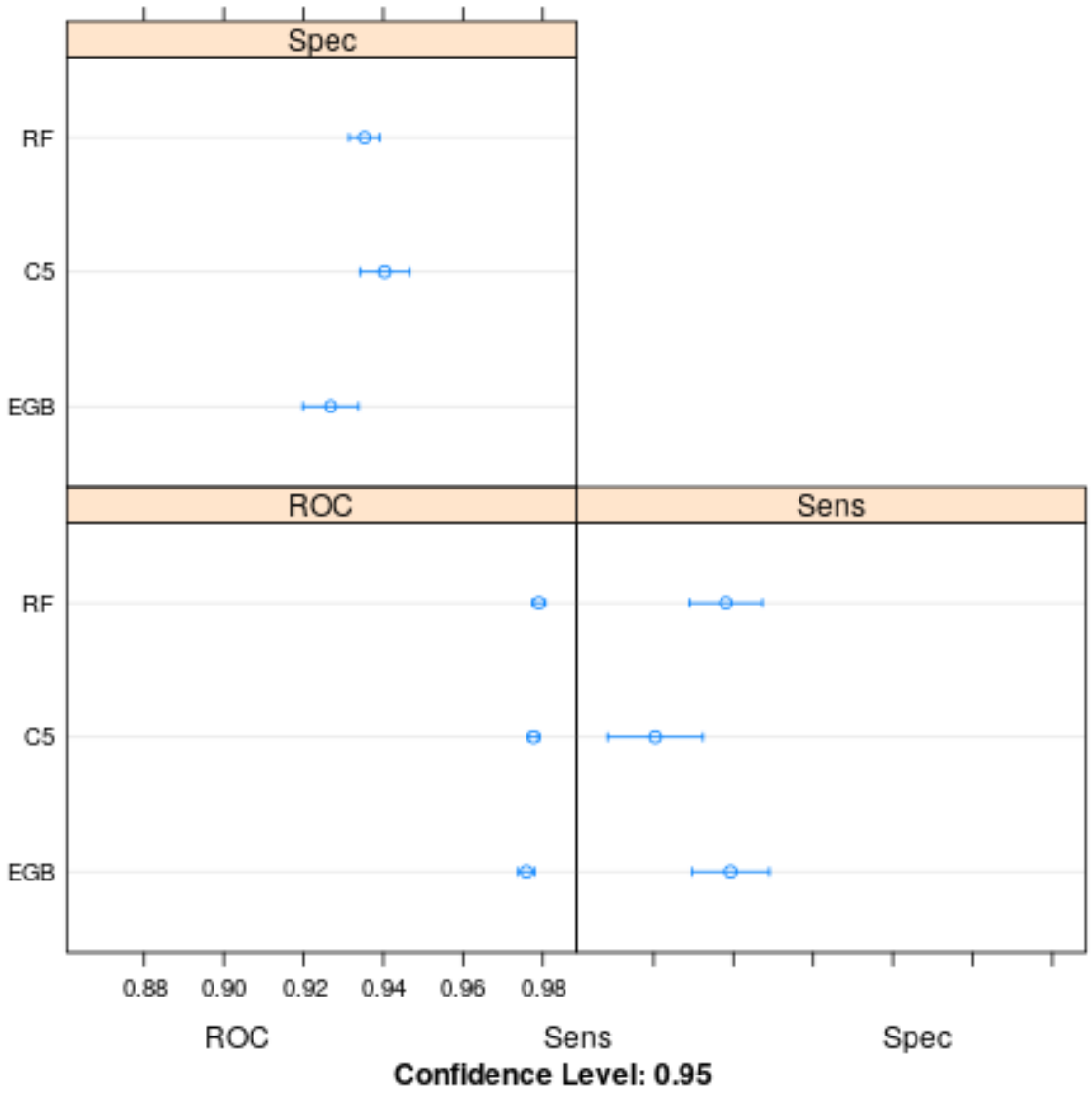
We can see that the predicted values include a higher proportion of Acer preferences than there are in the data with values

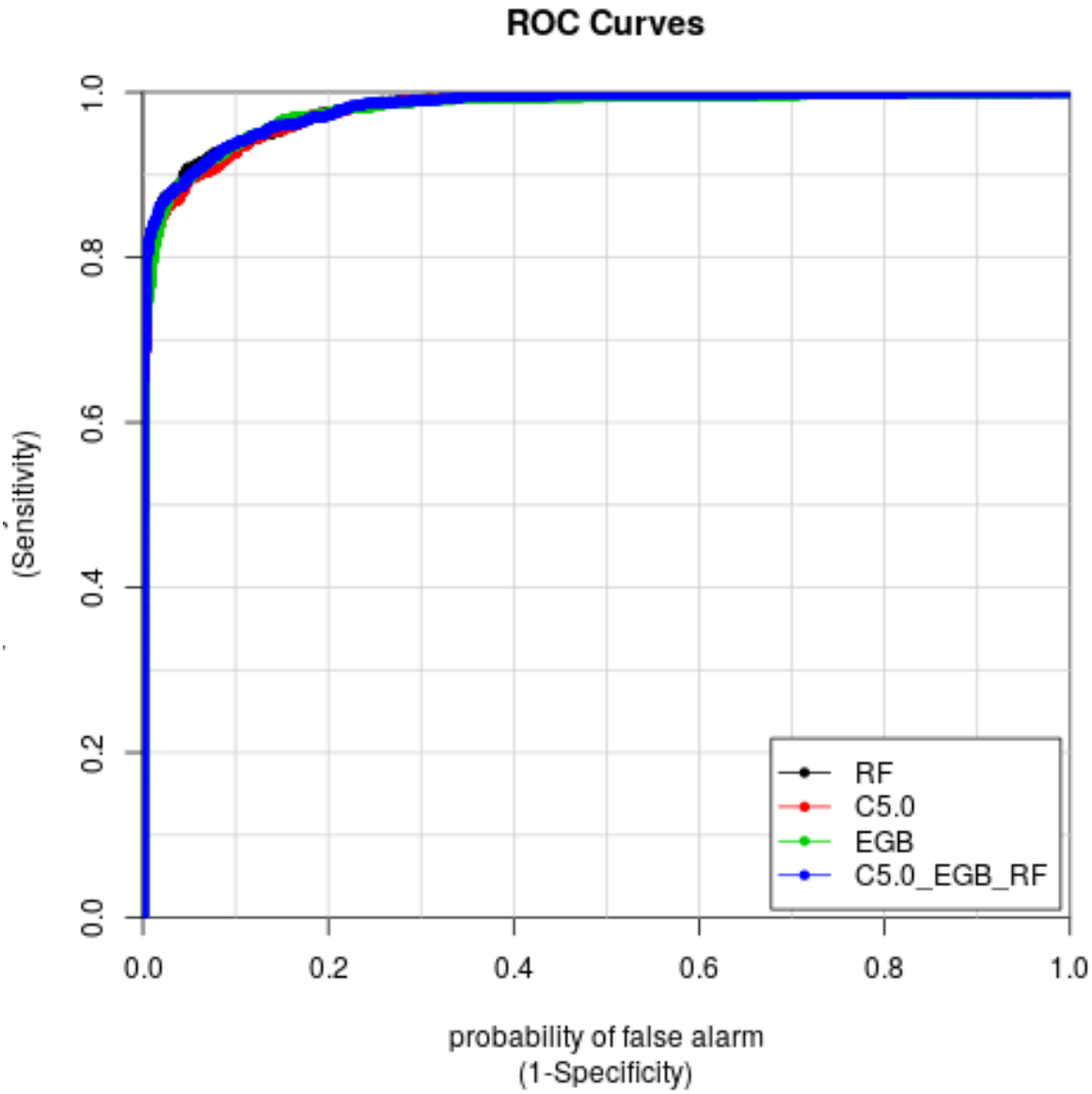
Chosen Model and Its' Performance



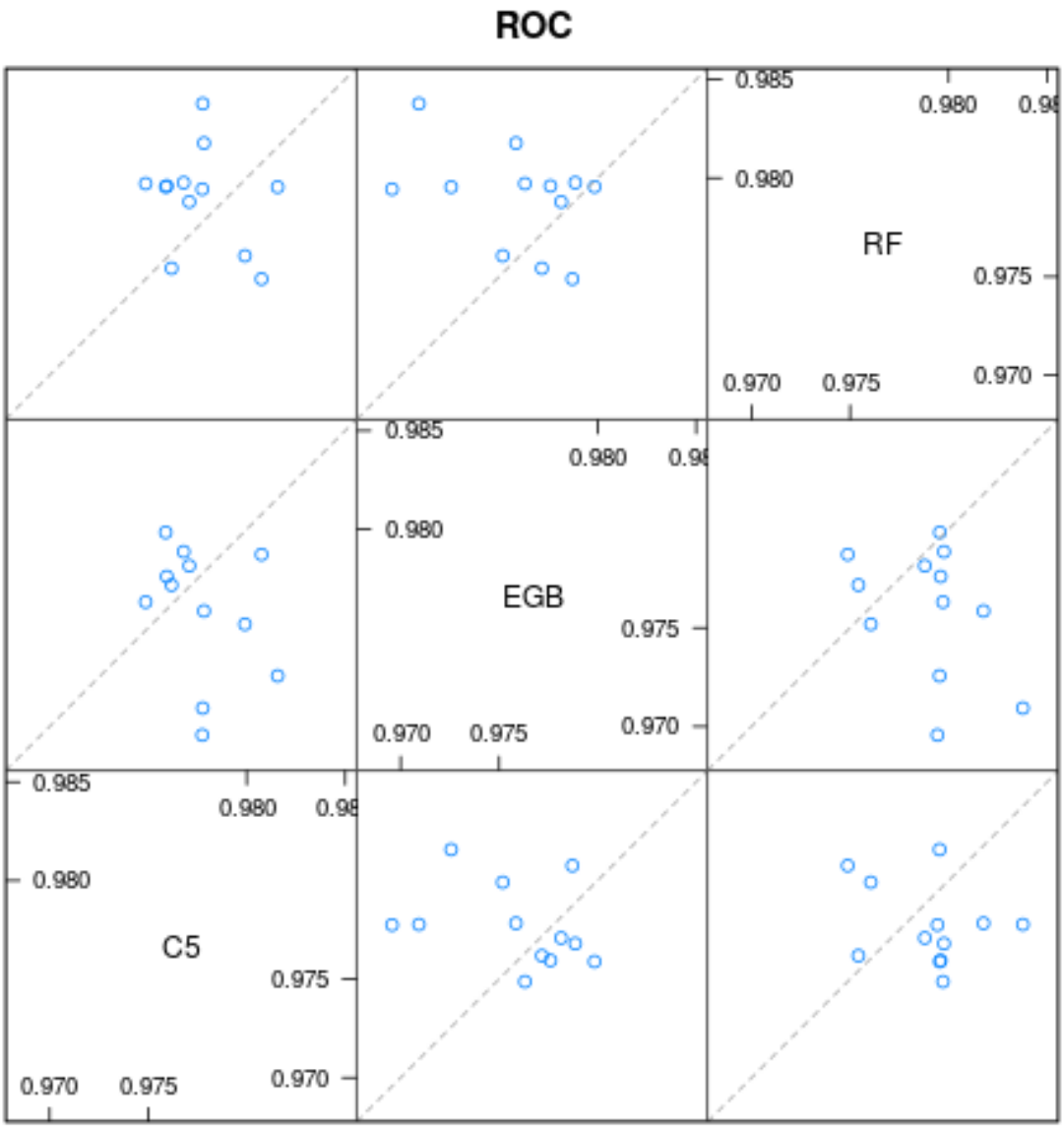
Here we explain what is our model and how it performs

Model Comparison and Performance





Differences between the models are very small


ROC
RF
EGB
C5
0.985
0.980
0.975
0.970
Scatter Plot Matrix

Listing 1: C5.0

```
> model_c5
C5.0

6929 samples
  37 predictor
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Bootstrapped (12 reps)
Summary of sample sizes: 5774, 5775, 5774, 5774, 5774, 5774, ...
Resampling results across tuning parameters:

  model  winnow  trials  ROC        Sens       Spec
  rules  FALSE     1      0.9090473  0.9095760  0.8905525
  rules  FALSE    10      0.9755175  0.8693338  0.9398793
  rules  FALSE    20      0.9753839  0.8802030  0.9370938
  rules   TRUE     1      0.9095039  0.9097671  0.8890436
  rules   TRUE    10      0.9769554  0.8710330  0.9437094
  rules   TRUE    20      0.9776982  0.8803823  0.9403435
  tree   FALSE     1      0.9415051  0.8821165  0.9073816
  tree   FALSE    10      0.9764442  0.8935450  0.9305942
  tree   FALSE    20      0.9760400  0.9036636  0.9261838
  tree    TRUE     1      0.9396531  0.8842154  0.9041318
  tree    TRUE    10      0.9760833  0.8849660  0.9357010
  tree    TRUE    20      0.9757764  0.8935516  0.9316388

ROC was used to select the optimal model using the largest value.
The final values used for the model were trials = 20, model = rules and win

> sink()
```

Listing 2: Extreme Gradient Boosting

```
> model_egb
eXtreme Gradient Boosting

6929 samples
  37 predictor
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Bootstrapped (12 reps)
Summary of sample sizes: 5774, 5775, 5774, 5774, 5774, 5774, ...
Resampling results across tuning parameters:

  eta  max_depth  colsample_bytree  subsample  nrounds
ROC        Sens       Spec
  0.3  1          0.6               0.50        50
0.7847144  0.6495611  0.7689183
  0.3  1          0.6               0.50       100
0.7805679  0.6396436  0.7785515
  0.3  1          0.6               0.50       150
0.7809150  0.6350631  0.7787837
  0.3  1          0.6               0.75        50
0.7868286  0.6522334  0.7728644
  0.3  1          0.6               0.75       100
0.7839463  0.6438394  0.7787837
  0.3  1          0.6               0.75       150
0.7831192  0.6383053  0.7801764
  0.3  1          0.6               1.00        50
0.7885789  0.6491854  0.7758821
  0.3  1          0.6               1.00       100
0.7868308  0.6386902  0.7785515
  0.3  1          0.6               1.00       150
0.7866826  0.6358280  0.7806407
  0.3  1          0.8               0.50        50
0.7865338  0.6571958  0.7685701
  0.3  1          0.8               0.50       100
0.7830051  0.6436508  0.7749536
  0.3  1          0.8               0.50       150
0.7823076  0.6421275  0.7772748
  0.3  1          0.8               0.75        50
0.7865843  0.6564374  0.7703110
  0.3  1          0.8               0.75       100
0.7828460  0.6451803  0.7769266
  0.3  1          0.8               0.75       150
0.7831444  0.6400303  0.7807567
  0.3  1          0.8               1.00        50
0.7900142  0.6470886  0.7759981
```

Listing 3: Random Forest

```
> model_rf
Random Forest

6929 samples
  37 predictor
   2 classes: 'Acer', 'Sony'

Pre-processing: centered (37), scaled (37)
Resampling: Bootstrapped (12 reps)
Summary of sample sizes: 5774, 5774, 5775, 5774, 5774, 5774, ...
Resampling results across tuning parameters:

  mtry  splitrule   ROC        Sens         Spec
   2    gini        0.8642404  0.0242286510 0.9988394
   2    extratrees  0.7967037  0.0003813883 1.0000000
  19    gini        0.9780926  0.9015681530 0.9369777
  19    extratrees  0.9749603  0.8895478625 0.9340761
  37    gini        0.9787809  0.8962282801 0.9336119
  37    extratrees  0.9790243  0.8981321598 0.9352368

Tuning parameter 'min.node.size' was held constant at a value of 1
ROC was used to select the optimal model using the largest value.
The final values used for the model were mtry = 37, splitrule = extratrees

> sink()
```

Listing 4: GLM Ensemble (C5.0 + EGB + RF)

```
> model_c5_egb_rf
A glm ensemble of 2 base models: xgbTree, ranger

Ensemble results:
Generalized Linear Model

13858 samples
    2 predictor
    2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Bootstrapped (12 reps)
Summary of sample sizes: 5774, 5774, 5775, 5774, 5774, 5774, ...
Resampling results:

  ROC        Sens       Spec
  0.9792814  0.9505129  0.884273


> sink()
```