Music Preference Prediction Using Random Forest Model

This dataset was made by our external consultant Alexander Makhratchev. He found about 500 songs he liked and 500 songs he disliked, and downloaded information about them through the Spotify API. There are 16 attribute columns, such as song name, danceability, and speechiness. You can see a sample of the dataset below.

In [1]:

```
%%bash
head spotify.csv
```

```
,artist,album,track_name,track_id,danceability,energy,key,loudness,mode,speechiness,instr
umentalness,liveness,valence,tempo,duration_ms,time_signature,like
0,Tiësto,BLUE (Remixes),BLUE - Mike Williams Remix,10WrTQMhZu2gocF8UB6obr,0.644,0.9209999
999999999,11,-3.201,1,0.045,0.000371,0.355,0.5770000000000001,128.015,191733,4,True
1,RICCI,Whistle,Whistle,0s7TF4xqdNcXn8U8cWXrhC,0.7120000000000001,0.934,2,-4.769,1,0.0595
,6.22e-06,0.0542,0.53,120.024,196030,4,True
2,Harris & Ford,"Freitag, Samstag","Freitag, Samstag",35WEFAhw47XLju1gu40cjT,0.638,0.961,
10,-3.8280000000000003,0,0.121,0.0055899999999999995,0.0663,0.353,138.034,156356,4,True
3,ILLENIUM,Awake,Feel Good,0e0UxWGgjXoYAYUFhJgwji,0.625,0.7070000000000001,2,-4.761,1,0.0
337,0.0,0.213,0.479,138.064,248156,4,True
4,TJR,Bounce Generation,Bounce Generation - Radio Edit,3l3wjXneWieRL0yKd4Tihf,0.687,0.998
,2,-1.304,1,0.29100000000000004,0.0431,0.32899999999999996,0.129,128.007,152813,4,True
5,Rush,Moving Pictures (2011 Remaster),Tom Sawyer,3QZ7uX97s82HFYSmQUAN1D,0.536,0.90099999
99999999,9,-7.211,1,0.0374,0.0186,0.06,0.6659999999999999,87.559,276880,4,False
6,Tina Turner,Tina!,River Deep - Mountain High,19jo0UT2vqD4pNVfIqTy4R,0.621,0.972,8,-3.79
10000000000004,1,0.0724,0.000585,0.195,0.866,155.113,244160,4,False
7,Drake,Dark Lane Demo Tapes,Toosie Slide,466cKvZn1j45IpxDdYZqdA,0.83,0.49,1,-8.82,0,0.20
9,3.04e-06,0.113,0.845,81.604,247059,4,True
8,Axel Rudi Pell,Wings of the Storm,Wings of the Storm,0oPS4seBoSMmz0M1OnL0l1,0.492000000
00000005,0.836,8,-6.922999999999999,0,0.0356,0.00043,0.128,0.384,82.07600000000001,347720
,4,False
```

In this notebook our goal is to simply demonstrate the Random Forest model. Using the -f RF command we can specify the exact predictor we would like to use.

In [2]:

```
! btc spotify.csv -f RF --yes
```

```
WARNING: Could not detect a GPU. Neural Network generation will be slow.

Brainome Table Compiler 0.991
Copyright (c) 2019-2021 Brainome, Inc. All Rights Reserved.
Licensed to:                 Alexander Makhratchev  (Evaluation)
Expiration Date:             2021-04-30   44 days left
Maximum File Size:           30 GB
Maximum Instances:           unlimited
Maximum Attributes:          unlimited
Maximum Classes:             unlimited
Connected to:                daimensions.brainome.ai  (local execution)

Command:
    btc spotify.csv -f RF --yes

Start Time:                 03/17/2021, 04:49 UTC




Pre-training Measurements
Data:
    Input:                      spotify.csv
    Target Column:              like
```

```
    Number of instances:       1224
    Number of attributes:        17 out of 17
    Number of classes:            2

Class Balance:
                           True: 43.38%
                          False: 56.62%

Learnability:
    Best guess accuracy:          56.62%
    Data Sufficiency:             Not enough data to generalize. [red]

Capacity Progression:             at [ 5%, 10%, 20%, 40%, 80%, 100% ]
    Ideal Machine Learner:              6,   7,   8,   9,   9,  10



Expected Generalization:
    Decision Tree:                 1.99 bits/bit
    Neural Network:               24.57 bits/bit
    Random Forest:                12.24 bits/bit

Expected Accuracy:              Training           Validation
    Decision Tree:               100.00%               50.41%
    Neural Network:               56.30%               54.16%
    Random Forest:               100.00%               83.69%

Recommendations:
    Warning: Data has high information density. Using effort 5 and larger ( -e 5 ) can im
prove results.
    Note: Model type RF given by user.




Predictor:                        a.py
    Classifier Type:              Random Forest
    System Type:                  Binary classifier
    Training / Validation Split:  50% : 50%
    Accuracy:
      Best-guess accuracy:        56.61%
      Training accuracy:         100.00% (612/612 correct)
      Validation Accuracy:        85.78% (525/612 correct)
      Combined Model Accuracy:    92.89% (1137/1224 correct)

    Model Capacity (MEC):         12    bits

    Generalization Ratio:         50.54 bits/bit
    Generalization Index:         25.15
    Percent of Data Memorized:     3.98%
    Resilience to Noise:          -1.71 dB

    Training Confusion Matrix:
              Actual | Predicted
                True |   272     0
               False |     0   340

    Validation Confusion Matrix:
              Actual | Predicted
                True |   214    45
               False |    42   311

    Combined Confusion Matrix:
              Actual | Predicted
                True |   486    45
               False |    42   651

    Training Accuracy by Class:
               class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                True |   272     0   340     0  100.00%  100.00%  100.00%  100.00%  100.
```

```
00%  100.00%
               False |   340     0   272     0  100.00%  100.00%  100.00%  100.00%  100.
00%  100.00%

    Validation Accuracy by Class:
               class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                True |   214    42   311    45   82.63%   87.36%   83.59%   87.36%   83.
11%   71.10%
               False |   311    45   214    42   88.10%   83.59%   87.36%   83.59%   87.
73%   78.14%

    Combined Accuracy by Class:
               class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                True |   486    42   651    45   91.53%   93.53%   92.05%   93.53%   91.
78%   84.82%
               False |   651    45   486    42   93.94%   92.05%   93.53%   92.05%   93.
74%   88.21%

    Attribute Ranking:
                     danceability :   22.70%
                      speechiness :   10.05%
                            tempo :    8.51%
                           energy :    6.96%
                         loudness :    6.53%
                             mode :    6.36%
                      duration_ms :    5.57%
                           artist :    5.18%
                          valence :    5.13%
                   time_signature :    4.27%
                              key :    4.10%
                 instrumentalness :    2.75%
                                  :    2.73%
                         liveness :    2.48%
                         track_id :    2.44%
                       track_name :    2.31%
                            album :    1.93%



End Time:           03/17/2021, 04:49 UTC
Runtime Duration:   21s
```

From the measurements, we can see that the random forest predictor has better generalization and memory equvalent capacity than the decision tree. However, it is still fairly far off from the neural network. The accuracy of the random forest predictor on the validation set is 84.96%, which is about a 30% improvement on best guess.

Using Attribute Ranking

Now we will use the -rank command in order to select the attributes that are most correlated to the target class. This will allow us to find the needle in the haystack.

In [3]:

```
! btc spotify.csv -f RF -rank --yes
```

```
WARNING: Could not detect a GPU. Neural Network generation will be slow.

Brainome Table Compiler 0.991
Copyright (c) 2019-2021 Brainome, Inc. All Rights Reserved.
Licensed to:                Alexander Makhratchev  (Evaluation)
Expiration Date:            2021-04-30   44 days left
Maximum File Size:          30 GB
Maximum Instances:          unlimited
Maximum Attributes:         unlimited
Maximum Classes:            unlimited
Connected to:               daimensions.brainome.ai  (local execution)
```

Command:
```
    btc spotify.csv -f RF -rank --yes

Start Time:                 03/17/2021, 04:49 UTC
```

Attribute Ranking:
```
    Columns selected:           artist, , loudness,
    Risk of coincidental column correlation:    0.0%

    Test Accuracy Progression:
                            artist :   72.06%
                                   :   72.63% change   +0.57%
                          loudness :   72.79% change   +0.16%
```

Pre-training Measurements

Data:
```
    Input:                      spotify.csv
    Target Column:              like
    Number of instances:       1224
    Number of attributes:         3 out of 17
    Number of classes:            2

Class Balance:
                            True: 43.38%
                           False: 56.62%

Learnability:
    Best guess accuracy:          56.62%
    Data Sufficiency:             Maybe enough data to generalize. [yellow]

Capacity Progression:             at [ 5%, 10%, 20%, 40%, 80%, 100% ]
    Ideal Machine Learner:              6,   7,   8,   8,   9,   9


Expected Generalization:
    Decision Tree:                 3.63 bits/bit
    Neural Network:              346.00 bits/bit
    Random Forest:                 5.80 bits/bit

Expected Accuracy:              Training           Validation
    Decision Tree:               100.00%               72.79%
    Neural Network:               56.63%               56.61%
    Random Forest:               100.00%               65.58%

Recommendations:
    Note: Model type RF given by user.
```

Predictor:
```
    Classifier Type:              Random Forest
    System Type:                  Binary classifier
    Training / Validation Split:  60% : 40%
    Accuracy:
      Best-guess accuracy:        56.61%
      Training accuracy:          99.86% (733/734 correct)
      Validation Accuracy:        67.34% (330/490 correct)
      Combined Model Accuracy:    86.84% (1063/1224 correct)

    Model Capacity (MEC):         15    bits

    Generalization Ratio:         48.35 bits/bit
```

```
    Generalization Index:          24.06
    Percent of Data Memorized:      4.16%
    Resilience to Noise:           -1.69 dB

    Training Confusion Matrix:
              Actual | Predicted
                True |   322     1
               False |     0   411

    Validation Confusion Matrix:
              Actual | Predicted
                True |   121    87
               False |    73   209

    Combined Confusion Matrix:
              Actual | Predicted
                True |   443    88
               False |    73   620

    Training Accuracy by Class:
                class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                 True |   322     0   411     1   99.69%   99.76%  100.00%   99.76%   99.
84%   99.69%
                False |   411     1   322     0  100.00%  100.00%   99.76%  100.00%   99.
88%   99.76%

    Validation Accuracy by Class:
                class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                 True |   121    73   209    87   58.17%   70.61%   62.37%   70.61%   60.
20%   43.06%
                False |   209    87   121    73   74.11%   62.37%   70.61%   62.37%   72.
32%   56.64%

    Combined Accuracy by Class:
                class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                 True |   443    73   620    88   83.43%   87.57%   85.85%   87.57%   84.
62%   73.34%
                False |   620    88   443    73   89.47%   85.85%   87.57%   85.85%   88.
51%   79.39%

    Attribute Ranking:
                              album :   38.79%
                             artist :   34.53%
                                    :   26.68%



End Time:           03/17/2021, 04:50 UTC
Runtime Duration:   26s
```

The two most important columns that were selected by attribute ranking were artist and loudness. Surprisingly, the -rank command lowered out validation accuracy to 66.73%. This might suggest that the target class is dependent on many more columns, and a neural network might be most effective here as indicated by the measurements.

Using -ignoreclasses

One of the attributes that the BTC identified and was important to the target class was the artist. However, we do not want our predictor to use that column, so we can utilize the -ignoreclasses command in order.

In [6]:

```
! btc spotify.csv -f RF --yes -ignoreclasses artist
```

```
WARNING: Could not detect a GPU. Neural Network generation will be slow.
```

Brainome Table Compiler 0.991

Licensed to: Alexander Makhratchev (Evaluation)
Expiration Date: 2021-04-30 44 days left
Maximum File Size: 30 GB
Maximum Instances: unlimited
Maximum Attributes: unlimited
Maximum Classes: unlimited
Connected to: daimensions.brainome.ai (local execution)

Command:
```
btc spotify.csv -f RF --yes -ignoreclasses artist
```

Start Time: 03/17/2021, 04:59 UTC

Pre-training Measurements
Data:
- Input: spotify.csv
- Target Column: like
- Number of instances: 1224
- Number of attributes: 17 out of 17
- Number of classes: 2

Class Balance:
- True: 43.38%
- False: 56.62%

Learnability:
- Best guess accuracy: 56.62%
- Data Sufficiency: Not enough data to generalize. [red]

Capacity Progression: at [5%, 10%, 20%, 40%, 80%, 100%]
- Ideal Machine Learner: 6, 7, 8, 9, 9, 10

Expected Generalization:
- Decision Tree: 1.99 bits/bit
- Neural Network: 26.54 bits/bit
- Random Forest: 12.12 bits/bit

Expected Accuracy:	Training	Validation
Decision Tree:	100.00%	50.41%
Neural Network:	56.46%	54.49%
Random Forest:	100.00%	83.52%

Recommendations:
- Warning: Data has high information density. Using effort 5 and larger (-e 5) can improve results.
- Note: Model type RF given by user.

Predictor: a.py
- Classifier Type: Random Forest
- System Type: Binary classifier
- Training / Validation Split: 50% : 50%
- Accuracy:
  - Best-guess accuracy: 56.61%
  - Training accuracy: 100.00% (612/612 correct)
  - Validation Accuracy: 84.47% (517/612 correct)
  - Combined Model Accuracy: 92.23% (1129/1224 correct)

- Model Capacity (MEC): 12 bits

```
    Generalization Ratio:          50.54 bits/bit
    Generalization Index:          25.15
    Percent of Data Memorized:      3.98%
    Resilience to Noise:           -1.71 dB

    Training Confusion Matrix:
              Actual | Predicted
                True |   272     0
               False |     0   340

    Validation Confusion Matrix:
              Actual | Predicted
                True |   220    39
               False |    56   297

    Combined Confusion Matrix:
              Actual | Predicted
                True |   492    39
               False |    56   637

    Training Accuracy by Class:
                class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                 True |   272     0   340     0  100.00%  100.00%  100.00%  100.00%  100.
00%  100.00%
                False |   340     0   272     0  100.00%  100.00%  100.00%  100.00%  100.
00%  100.00%

    Validation Accuracy by Class:
                class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                 True |   220    56   297    39   84.94%   88.39%   79.71%   88.39%   82.
24%   69.84%
                False |   297    39   220    56   84.14%   79.71%   88.39%   79.71%   86.
21%   75.77%

    Combined Accuracy by Class:
                class |    TP    FP    TN    FN     TPR      TNR      PPV      NPV
F1       TS
                 True |   492    56   637    39   92.66%   94.23%   89.78%   94.23%   91.
20%   83.82%
                False |   637    39   492    56   91.92%   89.78%   94.23%   89.78%   93.
06%   87.02%

    Attribute Ranking:
                      danceability :   24.53%
                       speechiness :    7.77%
                             tempo :    7.58%
                              mode :    6.58%
                          loudness :    6.48%
                            energy :    6.05%
                           valence :    5.44%
                          track_id :    4.69%
                    time_signature :    4.29%
                       duration_ms :    4.23%
                               key :    4.05%
                            artist :    3.68%
                        track_name :    3.66%
                             album :    3.18%
                          liveness :    2.66%
                  instrumentalness :    2.65%
                                   :    2.47%




End Time:           03/17/2021, 05:00 UTC
Runtime Duration:   21s
```

We get 83.52% accuracy on the validation set, which is almost identical to the first time we ran the compiler on

the dataset. Also, our accuracy is similar across classes, beacuse the dataset is balanced.