Popularity_of_Music_Records.R

Suchitra

```
songs = read.csv("songs.csv")
str(songs)
```

```
## 'data.frame':    7574 obs. of  39 variables:
##  $ year                    : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ songtitle               : Factor w/ 7141 levels "'03 Bonnie & Clyde",..: 6204 5522 241 309
8 47 607 254 4419 2887 6756 ...
##  $ artistname              : Factor w/ 1032 levels "50 Cent","98 Degrees",..: 3 3 3 3 3 3 3 3
 3 12 ...
##  $ songID                  : Factor w/ 7549 levels "SOAACNI1315CD4AC42",..: 595 5439 5252 171
6 3431 1020 1831 3964 6904 2473 ...
##  $ artistID                : Factor w/ 1047 levels "AR00B1I1187FB433EB",..: 671 671 671 671 6
71 671 671 671 671 671 507 ...
##  $ timesignature           : int  3 4 4 4 4 4 4 4 4 4 ...
##  $ timesignature_confidence: num  0.853 1 1 1 0.788 1 0.968 0.861 0.622 0.938 ...
##  $ loudness                : num  -4.26 -4.05 -3.57 -3.81 -4.71 ...
##  $ tempo                   : num  91.5 140 160.5 97.5 140.1 ...
##  $ tempo_confidence        : num  0.953 0.921 0.489 0.794 0.286 0.347 0.273 0.83 0.018 0.929
 ...
##  $ key                     : int  11 10 2 1 6 4 10 5 9 11 ...
##  $ key_confidence          : num  0.453 0.469 0.209 0.632 0.483 0.627 0.715 0.423 0.751 0.602
 ...
##  $ energy                  : num  0.967 0.985 0.99 0.939 0.988 ...
##  $ pitch                   : num  0.024 0.025 0.026 0.013 0.063 0.038 0.026 0.033 0.027 0.004
 ...
##  $ timbre_0_min            : num  0.002 0 0.003 0 0 ...
##  $ timbre_0_max            : num  57.3 57.4 57.4 57.8 56.9 ...
##  $ timbre_1_min            : num  -6.5 -37.4 -17.2 -32.1 -223.9 ...
##  $ timbre_1_max            : num  171 171 171 221 171 ...
##  $ timbre_2_min            : num  -81.7 -149.6 -72.9 -138.6 -147.2 ...
##  $ timbre_2_max            : num  95.1 180.3 157.9 173.4 166 ...
##  $ timbre_3_min            : num  -285 -380.1 -204 -73.5 -128.1 ...
##  $ timbre_3_max            : num  259 384 251 373 389 ...
##  $ timbre_4_min            : num  -40.4 -48.7 -66 -55.6 -43.9 ...
##  $ timbre_4_max            : num  73.6 100.4 152.1 119.2 99.3 ...
##  $ timbre_5_min            : num  -104.7 -87.3 -98.7 -77.5 -96.1 ...
##  $ timbre_5_max            : num  183.1 42.8 141.4 141.2 38.3 ...
##  $ timbre_6_min            : num  -88.8 -86.9 -88.9 -70.8 -110.8 ...
##  $ timbre_6_max            : num  73.5 75.5 66.5 64.5 72.4 ...
##  $ timbre_7_min            : num  -71.1 -65.8 -67.4 -63.7 -55.9 ...
##  $ timbre_7_max            : num  82.5 106.9 80.6 96.7 110.3 ...
##  $ timbre_8_min            : num  -52 -61.3 -59.8 -78.7 -56.5 ...
##  $ timbre_8_max            : num  39.1 35.4 46 41.1 37.6 ...
##  $ timbre_9_min            : num  -35.4 -81.9 -46.3 -49.2 -48.6 ...
##  $ timbre_9_max            : num  71.6 74.6 59.9 95.4 67.6 ...
##  $ timbre_10_min           : num  -126.4 -103.8 -108.3 -102.7 -52.8 ...
##  $ timbre_10_max           : num  18.7 121.9 33.3 46.4 22.9 ...
##  $ timbre_11_min           : num  -44.8 -38.9 -43.7 -59.4 -50.4 ...
##  $ timbre_11_max           : num  26 22.5 25.7 37.1 32.8 ...
##  $ Top10                   : int  0 0 0 0 0 0 0 0 0 1 ...
```

```
View(head(songs))

#number of observations (songs)  from the year 2010
songs_2010 = subset(songs, year == 2010)
nrow(songs_2010)
```

```
## [1] 373
```

```
#songs of Michael Jackson
songs_mj = subset (songs,  artistname =="Michael Jackson" )
nrow(songs_mj)
```

```
## [1] 18
```

```
View(songs_mj)

table(songs$timesignature)
```

```
##
##    0    1    3    4    5    7
##   10  143  503 6787  112   19
```

```
#song with the highest tempo
which.max(songs$tempo)
```

```
## [1] 6206
```

```
songs[6206,]
```

```
##      year                   songtitle      artistname             songID
## 6206 1995 Wanna Be Startin' Somethin' Michael Jackson SONHIQM13738B7BE80
##                artistID timesignature timesignature_confidence loudness
## 6206 ARXPPEY1187FB51DF4             3                         1  -14.528
##        tempo tempo_confidence key key_confidence    energy pitch
## 6206 244.307            0.566   6           0.44 0.6379941 0.009
##      timbre_0_min timbre_0_max timbre_1_min timbre_1_max timbre_2_min
## 6206        11.84       44.378        -80.4      187.038      -106.47
##      timbre_2_max timbre_3_min timbre_3_max timbre_4_min timbre_4_max
## 6206      220.751      -79.722      199.699      -77.158      147.564
##      timbre_5_min timbre_5_max timbre_6_min timbre_6_max timbre_7_min
## 6206      -68.229      186.391     -110.029       63.148      -58.978
##      timbre_7_max timbre_8_min timbre_8_max timbre_9_min timbre_9_max
## 6206         93.6      -52.012       95.827      -63.554       84.129
##      timbre_10_min timbre_10_max timbre_11_min timbre_11_max Top10
## 6206       -53.492        67.001       -73.421        67.308     0
```

```
#subsetting the data
SongsTrain = subset(songs, year <= 2009)
SongsTest = subset(songs, year == 2010)
nrow(SongsTrain)
```

```
## [1] 7201
```

```
#removing non-numeric colnames
nonvars = c("year", "songtitle", "artistname", "songID", "artistID")

SongsTrain = SongsTrain[ , !(names(SongsTrain) %in% nonvars) ]

SongsTest = SongsTest[ , !(names(SongsTest) %in% nonvars) ]



SongsLog1 = glm(Top10 ~ ., data=SongsTrain, family=binomial)
summary(SongsLog1)
```

```
##
## Call:
## glm(formula = Top10 ~ ., family = binomial, data = SongsTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9220  -0.5399  -0.3459  -0.1845   3.0770
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.470e+01  1.806e+00   8.138 4.03e-16 ***
## timesignature             1.264e-01  8.674e-02   1.457 0.145050
## timesignature_confidence  7.450e-01  1.953e-01   3.815 0.000136 ***
## loudness                  2.999e-01  2.917e-02  10.282  < 2e-16 ***
## tempo                     3.634e-04  1.691e-03   0.215 0.829889
## tempo_confidence          4.732e-01  1.422e-01   3.329 0.000873 ***
## key                       1.588e-02  1.039e-02   1.529 0.126349
## key_confidence            3.087e-01  1.412e-01   2.187 0.028760 *
## energy                   -1.502e+00  3.099e-01  -4.847 1.25e-06 ***
## pitch                    -4.491e+01  6.835e+00  -6.570 5.02e-11 ***
## timbre_0_min              2.316e-02  4.256e-03   5.441 5.29e-08 ***
## timbre_0_max             -3.310e-01  2.569e-02 -12.882  < 2e-16 ***
## timbre_1_min              5.881e-03  7.798e-04   7.542 4.64e-14 ***
## timbre_1_max             -2.449e-04  7.152e-04  -0.342 0.732087
## timbre_2_min             -2.127e-03  1.126e-03  -1.889 0.058843 .
## timbre_2_max              6.586e-04  9.066e-04   0.726 0.467571
## timbre_3_min              6.920e-04  5.985e-04   1.156 0.247583
## timbre_3_max             -2.967e-03  5.815e-04  -5.103 3.34e-07 ***
## timbre_4_min              1.040e-02  1.985e-03   5.237 1.63e-07 ***
## timbre_4_max              6.110e-03  1.550e-03   3.942 8.10e-05 ***
## timbre_5_min             -5.598e-03  1.277e-03  -4.385 1.16e-05 ***
## timbre_5_max              7.736e-05  7.935e-04   0.097 0.922337
## timbre_6_min             -1.686e-02  2.264e-03  -7.445 9.66e-14 ***
## timbre_6_max              3.668e-03  2.190e-03   1.675 0.093875 .
## timbre_7_min             -4.549e-03  1.781e-03  -2.554 0.010661 *
## timbre_7_max             -3.774e-03  1.832e-03  -2.060 0.039408 *
## timbre_8_min              3.911e-03  2.851e-03   1.372 0.170123
## timbre_8_max              4.011e-03  3.003e-03   1.336 0.181620
## timbre_9_min              1.367e-03  2.998e-03   0.456 0.648356
## timbre_9_max              1.603e-03  2.434e-03   0.659 0.510188
## timbre_10_min             4.126e-03  1.839e-03   2.244 0.024852 *
## timbre_10_max             5.825e-03  1.769e-03   3.292 0.000995 ***
## timbre_11_min            -2.625e-02  3.693e-03  -7.108 1.18e-12 ***
## timbre_11_max             1.967e-02  3.385e-03   5.811 6.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6017.5  on 7200  degrees of freedom
## Residual deviance: 4759.2  on 7167  degrees of freedom
## AIC: 4827.2
```

```
##
## Number of Fisher Scoring iterations: 6
```

```
cor(SongsTrain$loudness, SongsTrain$energy)
```

```
## [1] 0.7399067
```

```
#avoiding correlation
SongsLog2 = glm(Top10 ~ . - loudness, data=SongsTrain, family=binomial)
summary(SongsLog2)
```

```
##
## Call:
## glm(formula = Top10 ~ . - loudness, family = binomial, data = SongsTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0983  -0.5607  -0.3602  -0.1902   3.3107
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.241e+00  7.465e-01  -3.002 0.002686 **
## timesignature              1.625e-01  8.734e-02   1.860 0.062873 .
## timesignature_confidence   6.885e-01  1.924e-01   3.578 0.000346 ***
## tempo                      5.521e-04  1.665e-03   0.332 0.740226
## tempo_confidence           5.497e-01  1.407e-01   3.906 9.40e-05 ***
## key                        1.740e-02  1.026e-02   1.697 0.089740 .
## key_confidence             2.954e-01  1.394e-01   2.118 0.034163 *
## energy                     1.813e-01  2.608e-01   0.695 0.486991
## pitch                     -5.150e+01  6.857e+00  -7.511 5.87e-14 ***
## timbre_0_min               2.479e-02  4.240e-03   5.847 5.01e-09 ***
## timbre_0_max              -1.007e-01  1.178e-02  -8.551  < 2e-16 ***
## timbre_1_min               7.143e-03  7.710e-04   9.265  < 2e-16 ***
## timbre_1_max              -7.830e-04  7.064e-04  -1.108 0.267650
## timbre_2_min              -1.579e-03  1.109e-03  -1.424 0.154531
## timbre_2_max               3.889e-04  8.964e-04   0.434 0.664427
## timbre_3_min               6.500e-04  5.949e-04   1.093 0.274524
## timbre_3_max              -2.462e-03  5.674e-04  -4.339 1.43e-05 ***
## timbre_4_min               9.115e-03  1.952e-03   4.670 3.02e-06 ***
## timbre_4_max               6.306e-03  1.532e-03   4.115 3.87e-05 ***
## timbre_5_min              -5.641e-03  1.255e-03  -4.495 6.95e-06 ***
## timbre_5_max               6.937e-04  7.807e-04   0.889 0.374256
## timbre_6_min              -1.612e-02  2.235e-03  -7.214 5.45e-13 ***
## timbre_6_max               3.814e-03  2.157e-03   1.768 0.076982 .
## timbre_7_min              -5.102e-03  1.755e-03  -2.907 0.003644 **
## timbre_7_max              -3.158e-03  1.811e-03  -1.744 0.081090 .
## timbre_8_min               4.488e-03  2.810e-03   1.597 0.110254
## timbre_8_max               6.423e-03  2.950e-03   2.177 0.029497 *
## timbre_9_min              -4.282e-04  2.955e-03  -0.145 0.884792
## timbre_9_max               3.525e-03  2.377e-03   1.483 0.138017
## timbre_10_min              2.993e-03  1.804e-03   1.660 0.097004 .
## timbre_10_max              7.367e-03  1.731e-03   4.255 2.09e-05 ***
## timbre_11_min             -2.837e-02  3.630e-03  -7.815 5.48e-15 ***
## timbre_11_max              1.829e-02  3.341e-03   5.476 4.34e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6017.5  on 7200  degrees of freedom
## Residual deviance: 4871.8  on 7168  degrees of freedom
## AIC: 4937.8
##
## Number of Fisher Scoring iterations: 6
```

```
SongsLog3 = glm(Top10 ~ . - energy, data=SongsTrain, family=binomial)
summary(SongsLog3)
```

```
##
## Call:
## glm(formula = Top10 ~ . - energy, family = binomial, data = SongsTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9182  -0.5417  -0.3481  -0.1874   3.4171
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.196e+01  1.714e+00   6.977 3.01e-12 ***
## timesignature              1.151e-01  8.726e-02   1.319 0.187183
## timesignature_confidence   7.143e-01  1.946e-01   3.670 0.000242 ***
## loudness                   2.306e-01  2.528e-02   9.120  < 2e-16 ***
## tempo                     -6.460e-04  1.665e-03  -0.388 0.698107
## tempo_confidence           3.841e-01  1.398e-01   2.747 0.006019 **
## key                        1.649e-02  1.035e-02   1.593 0.111056
## key_confidence             3.394e-01  1.409e-01   2.409 0.015984 *
## pitch                     -5.328e+01  6.733e+00  -7.914 2.49e-15 ***
## timbre_0_min               2.205e-02  4.239e-03   5.200 1.99e-07 ***
## timbre_0_max              -3.105e-01  2.537e-02 -12.240  < 2e-16 ***
## timbre_1_min               5.416e-03  7.643e-04   7.086 1.38e-12 ***
## timbre_1_max              -5.115e-04  7.110e-04  -0.719 0.471928
## timbre_2_min              -2.254e-03  1.120e-03  -2.012 0.044190 *
## timbre_2_max               4.119e-04  9.020e-04   0.457 0.647915
## timbre_3_min               3.179e-04  5.869e-04   0.542 0.588083
## timbre_3_max              -2.964e-03  5.758e-04  -5.147 2.64e-07 ***
## timbre_4_min               1.105e-02  1.978e-03   5.585 2.34e-08 ***
## timbre_4_max               6.467e-03  1.541e-03   4.196 2.72e-05 ***
## timbre_5_min              -5.135e-03  1.269e-03  -4.046 5.21e-05 ***
## timbre_5_max               2.979e-04  7.855e-04   0.379 0.704526
## timbre_6_min              -1.784e-02  2.246e-03  -7.945 1.94e-15 ***
## timbre_6_max               3.447e-03  2.182e-03   1.580 0.114203
## timbre_7_min              -5.128e-03  1.768e-03  -2.900 0.003733 **
## timbre_7_max              -3.394e-03  1.820e-03  -1.865 0.062208 .
## timbre_8_min               3.686e-03  2.833e-03   1.301 0.193229
## timbre_8_max               4.658e-03  2.988e-03   1.559 0.119022
## timbre_9_min              -9.318e-05  2.957e-03  -0.032 0.974859
## timbre_9_max               1.342e-03  2.424e-03   0.554 0.579900
## timbre_10_min              4.050e-03  1.827e-03   2.217 0.026637 *
## timbre_10_max              5.793e-03  1.759e-03   3.294 0.000988 ***
## timbre_11_min             -2.638e-02  3.683e-03  -7.162 7.96e-13 ***
## timbre_11_max              1.984e-02  3.365e-03   5.896 3.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6017.5  on 7200  degrees of freedom
## Residual deviance: 4782.7  on 7168  degrees of freedom
## AIC: 4848.7
##
## Number of Fisher Scoring iterations: 6
```

```
#validating the model

testPredict = predict(SongsLog3, newdata=SongsTest, type="response")
table(SongsTest$Top10, testPredict >= 0.45)
```

```
##
##     FALSE TRUE
##   0   309    5
##   1    40   19
```

```
#baseline model

table(SongsTest$Top10)
```

```
##
##   0   1
## 314  59
```

```
#accuracy

314/(314+59)
```

```
## [1] 0.8418231
```

```
#0.8418231


table(SongsTest$Top10, testPredict >= 0.45)
```

```
##
##     FALSE TRUE
##   0   309    5
##   1    40   19
```

```
#sensitivity = 19/(19+40) = 0.3220339
#specificity = 309/(309+5) = 0.9840764


#Model 3 favors specificity over sensitivity.
#Model 3 provides conservative predictions, and predicts that a song will make it to the Top 10
 very rarely.
#So while it detects less than half of the Top 10 songs,
#we can be very confident in the songs that it does predict to be Top 10 hits.



#Model 3 has a very high specificity, meaning that it favors specificity over sensitivity.
#While Model 3 only captures less than half of the Top 10 songs,
#it still can offer a competitive edge, since it is very conservative in its predictions.
```