Predicting Song Popularity

Raunak Pednekar

June 22, 2017

Introduction

The music industry has a well-developed market with a global annual revenue around $15 billion. The recording industry is highly competitive and is dominated by three big production companies which make up nearly 82% of the total annual album sales.

Artists are at the core of the music industry and record labels provide them with the necessary resources to sell their music on a large scale. A record label incurs numerous costs (studio recording, marketing, distribution, and touring) in exchange for a percentage of the profits from album sales, singles and concert tickets.

Unfortunately, the success of an artist's release is highly uncertain: a single may be extremely popular, resulting in widespread radio play and digital downloads, while another single may turn out quite unpopular, and therefore unprofitable. Knowing the competitive nature of the recording industry, record labels face the fundamental decision problem of which musical releases to support to maximize their financial success. In this report I will use song characteristics to attempt to predict whether a particular song will be in the Top 10 or not

Building a Model

Since we're predicting a binary variable (whether or not a song entered the top 10), we will use logistical regression for this problem.

```
Call:
glm(formula = Top10 ~ ., family = binomial, data = songs_train)

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
```

Interestingly, our model reveals that the most statistically significant and large effect size variables are not the signature, key and tempo of the song, but the *confidence* in the signature, key, and the tempo. Higher confidence suggests less complex music. This rather plausibly suggests that listeners like less complex songs.

Somewhat more surprising is the fact that the 'loudness' and 'energy' have opposite signs. It is possible that we're dealing with multicollinearity here

The correlation between loudness and energy is 0.74. We will expel energy for our new model and test it on new data.

Testing the Model

Our model has an accuracy of 0.88. The baseline model (of always predicting the more common reponse) has an accuracy of 0.85. This seemingly incremental gain is misleadingly small. This is because the baseline model *never* predicts a song being in the Top 10. Our model predicts 19 such songs correcty.

Confusion Matrix

	FALSE	TRUE
0	309	5
1	40	19

Our model (at threshold = 0.45) has a sensitivity of 0.32 and a specifity of 0.98.It 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. The threshold value can be played around with to make it less conservative.

Data

The dataset consists of all songs which made it to the Top 10 of the Billboard Hot 100 Chart from 1990-2010 plus a sample of additional songs that didn't make the Top 10. This data comes from three sources: Wikipedia, Billboard.com, and EchoNest.

The variables included in the dataset either describe the artist or the song, or they are associated with the following song attributes: time signature, loudness, key, pitch, tempo, and timbre.

Here's a detailed description of the variables:

- year = the year the song was released
- songtitle = the title of the song
- artistname = the name of the artist of the song
- songID and artistID = identifying variables for the song and artist
- timesignature and timesignature_confidence = a variable estimating the time signature of the song, and the confidence in the estimate
- loudness = a continuous variable indicating the average amplitude of the audio in decibels
- tempo and tempo_confidence = a variable indicating the estimated beats per minute of the song, and the confidence in the estimate
- key and key_confidence = a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B), and the confidence in the estimate
- energy = a variable that represents the overall acoustic energy of the song, using a mix of features such as loudness
- pitch = a continuous variable that indicates the pitch of the song
- timbre_0_min, timbre_0_max, timbre_1_min, timbre_1_max, . . . , timbre_11_min, and timbre_11_max = variables that indicate the minimum/maximum values over all segments for each of the twelve values in the timbre vector (resulting in 24 continuous variables)
- Top10 = a binary variable indicating whether or not the song made it to the Top 10 of the Billboard Hot 100 Chart (1 if it was in the top 10, and 0 if it was not)