POPULARITY OF MUSIC RECORDS

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.

How can we use analytics to predict the popularity of a song? In this assignment, we challenge ourselves to predict whether a song will reach a spot in the Top 10 of the Billboard Hot 100 Chart.

Taking an analytics approach, we aim to use information about a song's properties to predict its popularity. The dataset songs.csv 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)

PROBLEM 1.2 - UNDERSTANDING THE DATA (1/1 point)

How many songs does the dataset include for which the artist name is "Michael Jackson"?

18

You have used 1 of 3 submissions

PROBLEM 1.3 - UNDERSTANDING THE DATA (1/1 point)

Which of these songs by Michael Jackson made it to the Top 10? Select all that apply.

- □ Beat It
- ☑ You Rock My World
- □ Billie Jean
- ☑ You Are Not Alone

You have used 1 of 2 submissions

PROBLEM 1.4 - UNDERSTANDING THE DATA (2/2 points)

The variable corresponding to the estimated time signature (timesignature) is discrete, meaning that it only takes integer values (0, 1, 2, 3, . . .). What are the values of this variable that occur in our dataset? Select all that apply.

- ☑ 0
- ☑ 1
- □ 2
- ☑ 3
- ☑ 4
- ☑ 5
- □ 6
- ☑ 7
- □ 8

Which timesignature value is the most frequent among songs in our dataset?

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ● 4 ✓
- ○ 5
- ○ 6
- ○ 7
- ○ 8

You have used 1 of 2 submissions

PROBLEM 1.5 - UNDERSTANDING THE DATA (1/1 point)

Out of all of the songs in our dataset, the song with the highest tempo is one of the following songs. Which one is it?

- ○ Until The Day I Die
- ● Wanna Be Startin' Somethin' ✓
- ○ My Happy Ending
- ○ You Make Me Wanna...

You have used 1 of 2 submissions

PROBLEM 2.1 - CREATING OUR PREDICTION MODEL (1/1 point)

We wish to predict whether or not a song will make it to the Top 10. To do this, first use the subset function to split the data into a training set "SongsTrain" consisting of all the observations up to and including 2009 song releases, and a testing set "SongsTest", consisting of the 2010 song releases.

How many observations (songs) are in the training set?

7201

You have used 1 of 3 submissions

PROBLEM 2.2 - CREATING OUR PREDICTION MODEL (2/2 points)

In this problem, our outcome variable is "Top10" - we are trying to predict whether or not a song will make it to the Top 10 of the Billboard Hot 100 Chart. Since the outcome variable is binary, we will build a logistic regression model. We'll start by using all song attributes as our independent variables, which we'll call Model 1.

We will only use the variables in our dataset that describe the numerical attributes of the song in our logistic regression model. So we won't use the variables "year", "songtitle", "artistname", "songID" or "artistID".

We have seen in the lecture that, to build the logistic regression model, we would normally explicitly input the formula including all the independent variables in R. However, in this case, this is a tedious amount of work since we have a large number of independent variables.

There is a nice trick to avoid doing so. Let's suppose that, except for the outcome variable Top10, all other variables in the training set are inputs to Model 1. Then, we can use the formula

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

to build our model. Notice that the "." is used in place of enumerating all the independent variables. (Also, keep in mind that you can choose to put quotes around binomial, or leave out the quotes. R can understand this argument either way.)

However, in our case, we want to exclude some of the variables in our dataset from being used as independent variables ("year", "songtitle", "artistname", "songID", and "artistID"). To do this, we can use the following trick. First define a vector of variable names called nonvars - these are the variables that we won't use in our model.

```
nonvars = c("year", "songtitle", "artistname", "songID", "artistID")
```

To remove these variables from your training and testing sets, type the following commands in your R console:

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

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

Now, use the glm function to build a logistic regression model to predict Top10 using all of the other variables as the independent variables. You should use SongsTrain to build the model.

Looking at the summary of your model, what is the value of the Akaike Information Criterion (AIC)?

4827.2

You have used 1 of 5 submissions

PROBLEM 2.3 - CREATING OUR PREDICTION MODEL (1/1 point)

Let's now think about the variables in our dataset related to the confidence of the time signature, key and tempo (timesignature_confidence, key_confidence, and tempo_confidence). Our model seems to indicate that these confidence variables are significant (rather than the variables timesignature, key and tempo themselves). What does the model suggest?

○ The lower our confidence about time signature, key and tempo, the more likely the song is to be in the Top 10

● The higher our confidence about time signature, key and tempo, the more likely the song is to be in the Top 10 ✓

You have used 1 of 1 submissions

PROBLEM 2.4 - CREATING OUR PREDICTION MODEL (1/1 point)

In general, if the confidence is low for the time signature, tempo, and key, then the song is more likely to be complex. What does Model 1 suggest in terms of complexity?

○ Mainstream listeners tend to prefer more complex songs

● Mainstream listeners tend to prefer less complex songs ✓

You have used 1 of 1 submissions

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

New Post

© All Rights Reserved

About Blog News FAQs Contact Jobs Donate Sitemap

Terms of Service & Honor Code Privacy Policy Accessibility Policy

© edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.

POWERED BY OPENedX

Download on the App Store

ANDROID APP ON Google play