edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

zhushun0008 (/dashboard) ▼

Courseware (/courses/MITx/15.071x/1T2014/courseware)   Course Info (/courses/MITx/15.071x/1T2014/info)

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum)   Progress (/courses/MITx/15.071x/1T2014/progress)

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

Help

POPULARITY OF MUSIC RECORDS

The music industry has a well-developed market with a global revenue valued at $16.5 billion in 2012. 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 (/c4x/MITx/15.071x/asset/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 (http://en.wikipedia.org/wiki/Billboard_Hot_100), Billboard.com (http://www.billboard.com/), and EchoNest (http://echonest.com/).

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 1.1 - UNDERSTANDING THE DATA (1 point possible)

Use the read.csv function to load the dataset "songs.csv" into R.

How many observations (songs) are from the year 2010?

[   ]

[   ]

Show Answer   *You have used 0 of 3 submissions*

PROBLEM 1.2 - UNDERSTANDING THE DATA (1 point possible)

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

[   ]

[   ]

Show Answer   *You have used 0 of 3 submissions*

PROBLEM 1.3 - UNDERSTANDING THE DATA (1 point possible)

Which of these songs by Michael Jackson made it to the Top 10?

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

Show Answer   *You have used 0 of 2 submissions*

PROBLEM 1.4 - UNDERSTANDING THE DATA (2 points possible)

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

<u>Show Answer</u>   *You have used 0 of 2 submissions*

PROBLEM 1.5 - UNDERSTANDING THE DATA (1 point possible)

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...

<u>Show Answer</u>   *You have used 0 of 2 submissions*

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

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?

[   ]

[   ]

<u>Show Answer</u>   *You have used 0 of 3 submissions*

PROBLEM 2.2 - CREATING OUR PREDICTION MODEL (1 point possible)

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 variables 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)?

Show Answer   *You have used 0 of 3 submissions*

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

Let's now revisit the variables related to the confidence we have about 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

Show Answer   *You have used 0 of 1 submissions*

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

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

Show Answer	*You have used 0 of 1 submissions*

PROBLEM 2.5 - CREATING OUR PREDICTION MODEL (2 points possible)

Songs with heavier instrumentation tend to be louder (have higher values in the variable "loudness") and more energetic (have higher values in the variable "energy").

By inspecting the coefficient of the variable "loudness", what does Model 1 suggest?

   ○ Mainstream listeners prefer songs with heavy instrumentation
   ○ Mainstream listeners prefer songs with light instrumentation

By inspecting the coefficient of the variable "energy", do we draw the same conclusions as above?

[▾]

Show Answer	*You have used 0 of 1 submissions*

PROBLEM 3.1 - BEWARE OF MULTICOLLINEARITY ISSUES! (1 point possible)

What is the correlation between the variables "loudness" and "energy" in the training set?

[   ]

[   ]

Given that these two variables are highly correlated, Model 1 suffers from multicollinearity. To avoid this issue, we will omit one of these two variables and rerun the logistic regression. In the rest of this problem, we'll build two variations of our original model: Model 2, in which we keep "energy" and omit "loudness", and Model 3, in which we keep "loudness" and omit "energy".

Show Answer	*You have used 0 of 3 submissions*

PROBLEM 3.2 - BEWARE OF MULTICOLLINEARITY ISSUES! (1 point possible)

Create Model 2, which is Model 1 without the independent variable "loudness". This can be done with the following command:

SongsLog2 = glm(Top10 ~ . - loudness, data=SongsTrain, family=binomial)

We just subtracted the variable loudness. We couldn't do this with the variables "songtitle" and "artistname", because they are not numeric variables, and we might get different values in the test set that the training set has never seen. But this approach will work when you want to remove variables that you could feasibly use in your model.

Look at the summary of SongsLog2, and inspect the coefficient of the variable "energy". What do you observe?

   ○ Model 2 suggests that songs with high energy levels tend to be more popular. This contradicts our observation in Model 1.
   ○ Model 2 suggests that, similarly to Model 1, songs with low energy levels tend to be more popular.

PROBLEM 3.3 - BEWARE OF MULTICOLLINEARITY ISSUES! (1 point possible)

Now, create Model 3, which should be exactly like Model 1, but without the variable "energy".

Look at the summary of Model 3 and inspect the coefficient of the variable "loudness". Remembering that higher loudness and energy both occur in songs with heavier instrumentation, do we make the same observation about the popularity of heavy instrumentation as we did with Model 2?

[⌄]

In the remainder of this problem, we'll just use Model 3.

PROBLEM 4.1 - VALIDATING OUR MODEL (1 point possible)

Make predictions on the test set using Model 3. What is the accuracy of Model 3 on the test set, using a threshold of 0.45? (Compute the accuracy as a number between 0 and 1.)

[   ]

[   ]

PROBLEM 4.2 - VALIDATING OUR MODEL (1 point possible)

Let's check if there's any incremental benefit in using Model 3 instead of a baseline model. Given the difficulty of guessing which song is going to be a hit, an easier model would be to pick the most frequent outcome (a song is not a Top 10 hit) for all songs. What would be the fraction of accuracy using this baseline model for the test set?

[   ]

[   ]

PROBLEM 4.3 - VALIDATING OUR MODEL (2 points possible)

It seems that Model 3 gives us a small improvement over the baseline model. Still, does it create an edge?

Let's view the two models from an investment perspective. A production company is interested in investing in songs that are highly likely to make it to the Top 10. The company's objective is to minimize its risk of financial losses attributed to investing in songs that end up unpopular.

A competitive edge can therefore be achieved if we can provide the production company a list of songs that are highly likely to end up in the Top 10. We note that the baseline model does not prove useful, as it simply does not label any song as a hit. Let us see what our model has to offer.

How many songs does Model 3 correctly predict as Top 10 hits in 2010, using a threshold of 0.45?

How many non-hit songs does Model 3 predict will be Top 10 hits, using a threshold of 0.45?

Show Answer   *You have used 0 of 3 submissions*

PROBLEM 4.4 - VALIDATING OUR MODEL (2 points possible)

What is the sensitivity of Model 3 on the test set, using a threshold of 0.45?

What is the specificity of Model 3 on the test set, using a threshold of 0.45?

Show Answer   *You have used 0 of 3 submissions*

PROBLEM 4.5 - VALIDATING OUR MODEL (1 point possible)

What conclusions can you make about our model (select all that apply)?

☐ Model 3 favors specificity over sensitivity.

☐ Model 3 favors sensitivity over specificity.

☐ Model 3 captures less than half of Top 10 songs in 2010. Model 3 therefore does not provide a useful list of candidate songs to investors, and hence offers no competitive edge.

☐ 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.

Show Answer   *You have used 0 of 2 submissions*

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

Show Discussion

New Post

(http://www.meetup.com/edX-Global-Community/)

(http://www.facebook.com/EdxOnline)

(https://twitter.com/edXOnline)

(https://plus.google.com/108235383044095082

(http://youtube.com/user/edxonline)