Quiz Instructions

Logistic Regression

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) (Links to an external site.) 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.

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 = a variable estimating the time signature of the song
- timesignature_confidence = the confidence in the estimate
- loudness = a continuous variable indicating the average amplitude of the audio in decibels
- tempo = a variable indicating the estimated beats per minute of the song
- tempo_confidence = the confidence in the estimate
- key = a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B)

- key_confidence = 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)
- NOTEs
  - For all questions, you will use SciKit API and Python and will show ALL INPUT and OUTPUT required to understand the logistic regression model.
  - For each question you need to show how you came up with output/solution.

1.1 - Understanding the Data

Read input (songs.csv) and load it into a variable called dataset.

Show all of the work involved to get the answers

1.1 How many observations (songs) are there in total?

1.2 How many Top-10?

1.3 How many non-Top-10?

1.2 - Understanding the Data

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

2.1 - Creating Prediction Model

We wish to predict whether or not a song will make it to the Top 10.

To do this, first use the filter 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?

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

2.2 - Creating Prediction Model

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 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. However, in this case, this is a tedious amount of work since we have a large number of independent variables.

Step 1: we want to exclude some of the variables in our dataset from being used as independent variables ("year", "songtitle", "artistname", "songID", and "artistID").

Step 2: build a logistic regression model to predict Top10 using the training data. You may enumerate all the remaining independent variables

Problem 2.3 - Creating Prediction Model

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?

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

Group of answer choices

- A, B

2.4 - Creating Prediction Model

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 our model suggest in terms of complexity?

A. Mainstream listeners tend to prefer more complex songs
B. Mainstream listeners tend to prefer less complex songs

2.5 - Creating Prediction Model

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

By inspecting the coefficient of the variable "loudness", what does our model suggest?

A. Mainstream listeners prefer songs with heavy instrumentation
B. Mainstream listeners prefer songs with light instrumentation

3.1 - Validating Model

Make predictions on the test set using our model. What is the accuracy of our model on the test set (Compute the accuracy as a number between 0 and 1.)

3.2 - Validating Model

What is the True Positive Rate of our model on the test set?

3.3 - Validating Model

What is the False Positive Rate of our model on the test set?

Jupyter Notebook Solution

- All questions must be answered in the provided sections.
- But you have a chance to upload your entire solution (as a Jupyter notebook) here