Individual Assignment:

Predicting the Popularity of Music Records

The music industry has a well-developed market with global annual revenue of 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 companies provide them with the necessary resources to sell their music on a large scale.

A record company 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 companies 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 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.

The variables are described as follows

- **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

Individual Assignment:

Predicting the Popularity of Music Records

- **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)

Questions:

1. **[15 pts]:** Fit a logistic regression model using all variables
2. **[15 pts]:** Predict the popularity of records in the testing set.
3. **[15 pts]:** Generate the ROC curve
4. **[30 pts]:** Improve the prediction performance of your model. For example, you may try transforming some predictors, and/or perform variable selection, or other approaches. Explain all steps!
5. **[25 pts]:** Choose 5 coefficients from the finally chosen model and interpret them.