Capstone Project Proposal

Top 10 hit song predictive model using lyrical and song data

Objective
For my project proposal, I want to create a Hot 100 Predictive Model to predict if an album can make it onto the year-end Billboard hot 100 charts during the year it is released. I will treat this as a supervised machine learning problem and analyze the 5 commonalities among the top 100 songs (by sales) from each year over the last ten years. This will provide artists and music labels insight into what drive songs to generate revenue. My goal is to obtain a statistically significant accuracy in the prediction.

The Billboard Year-End charts are a cumulative measure of a single or album's performance in the United States, based upon the Billboard magazine charts during any given chart year. Billboard's "chart year" runs from the first Billboard "week" of December to the final week in November, but because the Billboard week is dated in advance of publication, the last calendar week for which sales are counted is usually the third week in November.The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine.[1] The year-end charts are now calculated by a very straightforward cumulative total of yearlong sales (or sales and airplay) points. This gives a more accurate picture of any given year's most popular titles, as an entry that hypothetically spent nine weeks at number one in the spring could possibly have earned fewer cumulative points than one spending six weeks at number three in January.

Due to this methodology, albums at the peak of their popularity at the time of the November/December chart-year cutoff many times end up ranked lower than one would expect on a year-end tally, yet are ranked on the following year's chart as well, as their cumulative points are split between the two chart years.

Goals
Music producers are consistently under pressure to understand what creates a hit song. Considering this, a predictive model that can quantify characteristics pre-release could save artists and record companies millions of dollars. I want to address the following:

1. Is the percent chance for striking ahit single affected by previously released hit singles and the time of the release?
2. Is it possible to analyze a certain number of factors to predict if a given song will be a hit single?
3. What are the top three factors that can be correlated with a song making it into the Billboard Hot 100?

Approach - Predicting Successful Songs
1. Understand the most important features to structure the model around
2. Use scraping to collect data from the internet, self selected features (record label, lyric features, artist features, etc.) and a support vector machine model.
3. Clean data
4. Explore and construct visualizations of findings

I will use a dataset of_____, with x amount of those songs having made it onto the Hot 100 charts, and the remaining songs not making it onto the charts.

I'll train the support vector machine on the deliverable characteristics, and then run the classifier on a test dataset of a songs (b successful and c unsuccessful)

Data Sources
I'll use data from the following sources among others not yet listed:
Lyrics.com
Billboard.com

Deliverables

A visualization of the month and season each album was released
A visualization of how many hit singles were released before the album was fully released
A visualization showing the 2 overarching themes from each album
Brand names referenced in the albums
Profanity visualization mentioned in the albums
Visualization of data run through SVM

Notes

Github = cathjunping88

Hypothesis testing -
Machine learning -
Data Source

Citations

[1]https://en.wikipedia.org/wiki/Billboard_Year-End