HitPredict: Predicting Billboard Hits Using Spotify Data

Abstract

- The Billboard Hot 100 Chart[1] remains one of the definitive ways to measure the success of a popular song. We investigated using machine learning techniques to **predict which songs will become Billboard Hot 100 Hits**.
- We were able to predict the Billboard success of a song with **~75% accuracy** using machine-learning algorithms including **Logistic Regression, GDA, SVM, Decision Trees** and **Neural Networks**.

Features and Data

- Ten audio features were extracted from the Spotify API[4] (Table 1).
- We created the Artist Score metric, assigning a score of 1 to a song if the artist previously had a Billboard hit, and 0 otherwise.

Audio Features	
Danceability	Liveness
Instrumentalness	Speechiness
Acousticness	Loudness
Valence	Tempo
Energy	Artist Score

Table 1. Audio features extracted from Spotify's API. Spotify assigns each song a value between 0 and 1 for these features, except loudness which is measured in decibels.

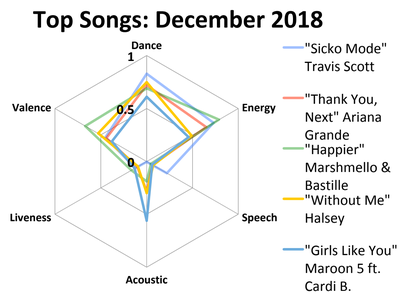


Figure 1. Illustration of audio features for the 5 top tracks of December 2018. Our algorithm predicted their Billboard success with 100% accuracy.

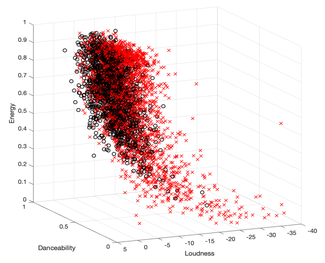
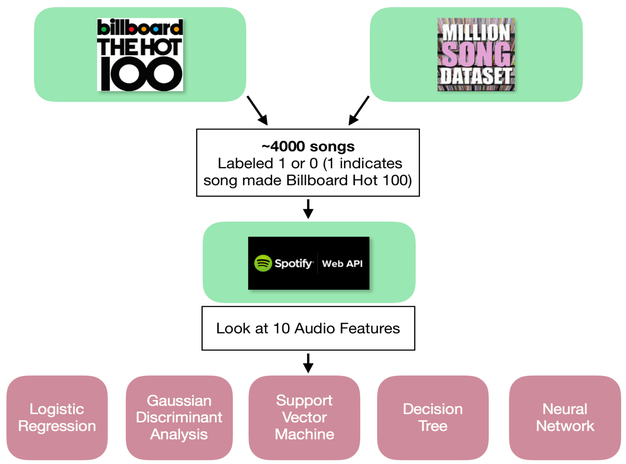


Figure 2. A plot of songs' danceability vs. energy vs. loudness (dB). Black circles represent Billboard hits and red marks represent non-hits.

Methods



- Data for ~4000 songs was collected from Billboard.com[3] and the Million Song Dataset[5]. Songs were from 1990-2018.
- Songs were labeled 1 or 0 based on Billboard success.
- Audio features for each song were extracted from the Spotify Web API[4].
- Five machine-learning algorithms were used to predict a song's Billboard success.

Algorithms

- **Supervised Learning**: data split 75/25 into training/ validation. **Logistic Regression** and **GDA** yielded the strongest results.
- Bagging using random forests corrected **SVM** from over-fitting.
- **Decision Tree** performs poorly as it suffers from severe over-fitting.
- **Neural Network** with regularization, using one hidden layer of six units with the sigmoid activation function. The L_2 regularization function was applied to the cost function to avoid over-fitting.

Results

Increasing order of importance

Logistic Regression		Neural Network	
Feature	Accuracy	Feature	Accuracy
Artist Score	72.9%	Danceability	65.3%
Instrumental	73.2%	Acousticness	69.6%
Danceability	73.2%	Speechiness	73.0%
Accousticness	75.3%	Valence	73.4%
Speechiness	75.8%	Energy	74.9%
Loudness	75.8%	Artist Score	74.6%
Tempo	**75.9%**	Instrumental	75.1%
Valence	75.7%	**Tempo**	**76.5%**
Energy	74.0%	Liveness	76.4%
Liveness	74.3%	Loudness	72.7%

Table 2. Error analysis for the two strongest-performing algorithms. The features at the end of the list decreased the accuracy of predictions.

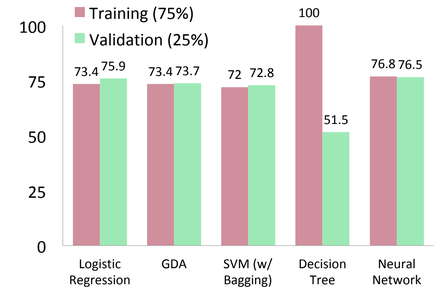


Figure 3. Billboard hit prediction accuracy results for five machine-learning algorithms. LR and NN give the highest prediction accuracy on the validation set.

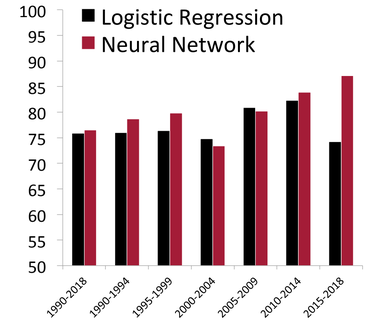


Figure 4. Algorithms yield higher accuracy for more recent songs. Features of pop songs are unique to their time period.

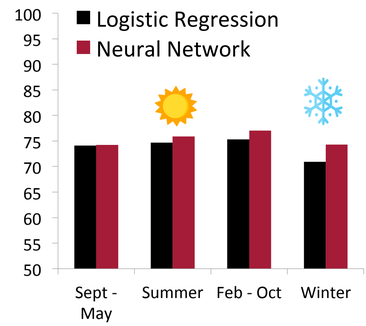


Figure 5. Features of songs released in winter vary from features of other songs. We did not observe the same trends for song of summer.