Spotify x Billboard Predictor Problem Statement

How can we build a model with at least 75% accuracy to identify songs that are likely to become Billboard successes based on their Spotify APIS through a systemic analysis of song records from 1990-2019?

1 Context

With the emergence of social media platforms in the late 1990s, the popularity of independent artists (musicians without record label contracts/representation) and single-song production has risen. As a result, revenues in the music industry have declined causing many artists to reconsider the benefits of partnership with record labels. This sentiment has pressured executives at record label companies to consider two propositions to remain competitive with independent artists, maintain their current client base, and attract new clients: assist in developing strong social media presence, or develop a data-driven model capable of identifying songs worthy of Billboard status (i.e. resulting in increased streams and revenue).

2 Criteria for success

Based on the datasets I will be using, the focus will be on:

- Successful identification of attributes that distinguish Billboard success (i.e. 'Hit' or 'Flop')
- Building an equation/model that can accurately (accuracy of at least 75%) predict a 'Hit Song'

These two outcomes will put record labels in a better position to consult artists on the viability of their success (maintaining/growing the client base) and allow the record labels to focus on releasing songs that meet the attribute thresholds of successful songs.

3 Scope of solution space

Timeframe of historical data: 1990-2019

Billboard Success: Distinguished as Hit (target = 1) or Flop (target = 0)

Identification of Attributes:
- Descriptive statistics: histograms, boxplots
- Predictive: feature importance, correlation matrix

Predictive Model:
- Logistic Regression or Random Forest
- Confusion Matrix
- Accuracy report

4 Constraints within solution space

- Trends in attributes may change vastly over the decades (i.e. something that was vastly popular in the 1990s may be an outlier in 2010s data)
- Trends in attributes may vary based on genre

5 Key data sources

The Spotify Hit Predictor Dataset (1960-2019)

- dataset-of-90s
- dataset-of-00s
- dataset-of-10s