Billboard Top 100: Predicting Song Popularity

James Parks

1 Background

For over 60 years, the Billboard Hot 100 chart has been used as the standard measurement of popular music success. On the week ending November 12, 1955, Billboard published the Top 100 chart for the first time. It combined all aspects of a single's performance (sales, airplay and jukebox activity), based on a point system that typically gave sales (purchases) more weight than radio airplay. The Billboard Hot 100 is still the standard by which a song's popularity is measured today. The Hot 100 is ranked by radio airplay audience impressions as measured by Nielsen BDS, sales data compiled by Nielsen Soundscan (both at retail and digitally) and streaming activity provided by online music sources.

With the advent of music streaming services, we now have a lot of data about not just the chart performance of songs, but their audio metadata features as well. In particular, Spotify has a number of measurements of different audio features of songs. Certainly artist name recognition is a big indicator of whether or not a song will find success on the Hot 100 chart.

The goal of this project is to develop a model to predict the popularity of songs based on their audio metadata properties as well as lyrical metadata.

1.1 The Client

There are at least two different groups of people who would be interested in the outcome of this project. The first group are music producers who will find it very beneficial to see how musical trends are changing over time when deciding which artists to sign and promote and current artists looking to have their songs chart could benefit from this knowledge during the writing process.

The second group is music streaming service companies like Spotify since their business model is based around the selection of their library and the power of their music recommendation system. Along with the audio metadata features taken from Spotify's API, we also use lyrical metadata in our model, which could be useful to implement in a music recommendation system.

2 The Data

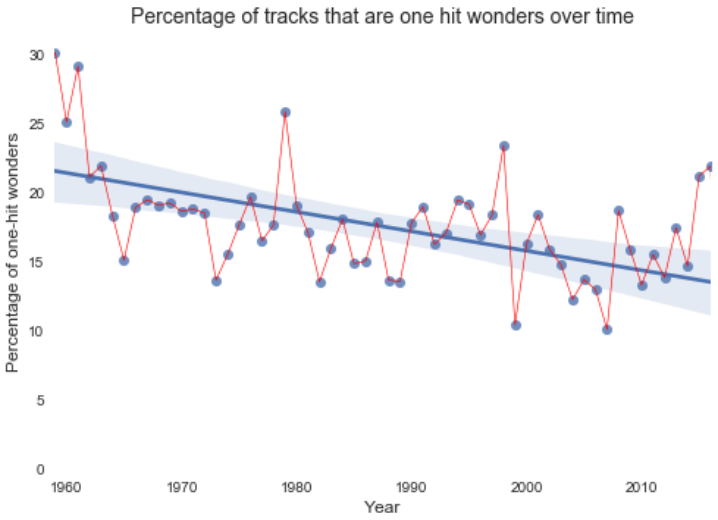
The Billboard Hot 100 dataset is located at the Ultimate Music Database. However we use the dataset by Brady Fowler at Decibels & Decimals that contains the Top 100 tracks ranked by popularity for every week between September 8, 1958 and January 8, 2017 as well as Spotify audio metadata for 78.3% of the tracks. We also scraped the lyrics for 82.6% of the tracks from several lyrics websites.

3 Exploratory Analysis

From the Billboard dataset we have the following features.

- **chartDate**: Always on a Saturday. Represents the ranking for the preceding week.
- **title**: Title of the song (Note that this title might be inconsistent with the title from the Spotify dataset.)
- **artist**: Artist of the song (Note that if the track features another artist that will be counted separately.)
- **peakPos**: The highest position the track ever reached on the charts.
- **lastPos**: The previous position on the track. A value of zero could mean that this is a new song that never charted before or that the song is reentering the charts.
- **weeks**: There is some inconsistency with this data. For some entries it is the number of weeks the song has been on the chart up to that point, but for some older entries it is the total number of weeks the song was on the charts.
- **rank**: The current rank of the song for that week.
- **change**: The change in the rank since the previous week. Songs that were not on the charts the previous week are either "New", "Re-Entry" or "Hot Shot Debut".
- **spotifyID**: The SpotifyID for the track for the Billboard dataset.

We find that a very slight majority (53.37%) of artists that appear in the Billboard Hot 100 only have one song, so-called one-hit wonders. We find that the number of one-hit wonders has experienced a slight decrease over the years.



The number of tracks released in a year that are one-hit wonders is usually between 10 and 25 percent. This is a sizeable amount of songs whose introduction into the Billboard Hot 100 chart is less likely to be caused by artist name recognition.

From the Spotify audio metadata we have the following features.

- **acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **duration_ms**: The duration of the track in milliseconds.
- **energy**: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- **id**: The Spotify ID for the track for the Spotify audio features dataset.
- **instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values

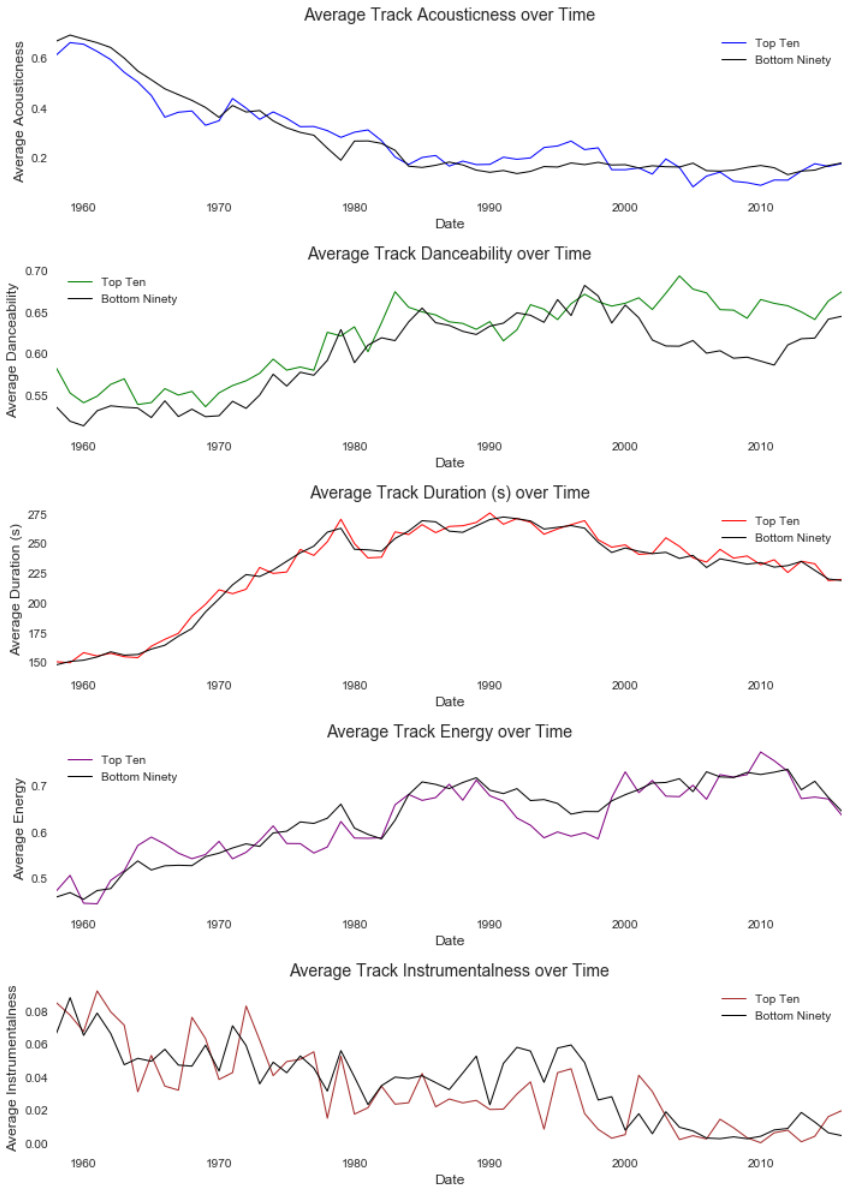
above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

- **key**: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C# or Db, 2 = D, and so on.
- **liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **loudness**: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
- **mode**: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- **speechiness**: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **time signature**: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- **valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Furthermore, lyrical diversity is defined to be the number of unique words in the lyrics divided by the total number of the words.

3.1 Audio Trends Over Time

We now consider the average rate of the Spotify audio metadata features over time. Here we are subsetting the data so that songs are each song is only counted once in the average for the first year it was released.



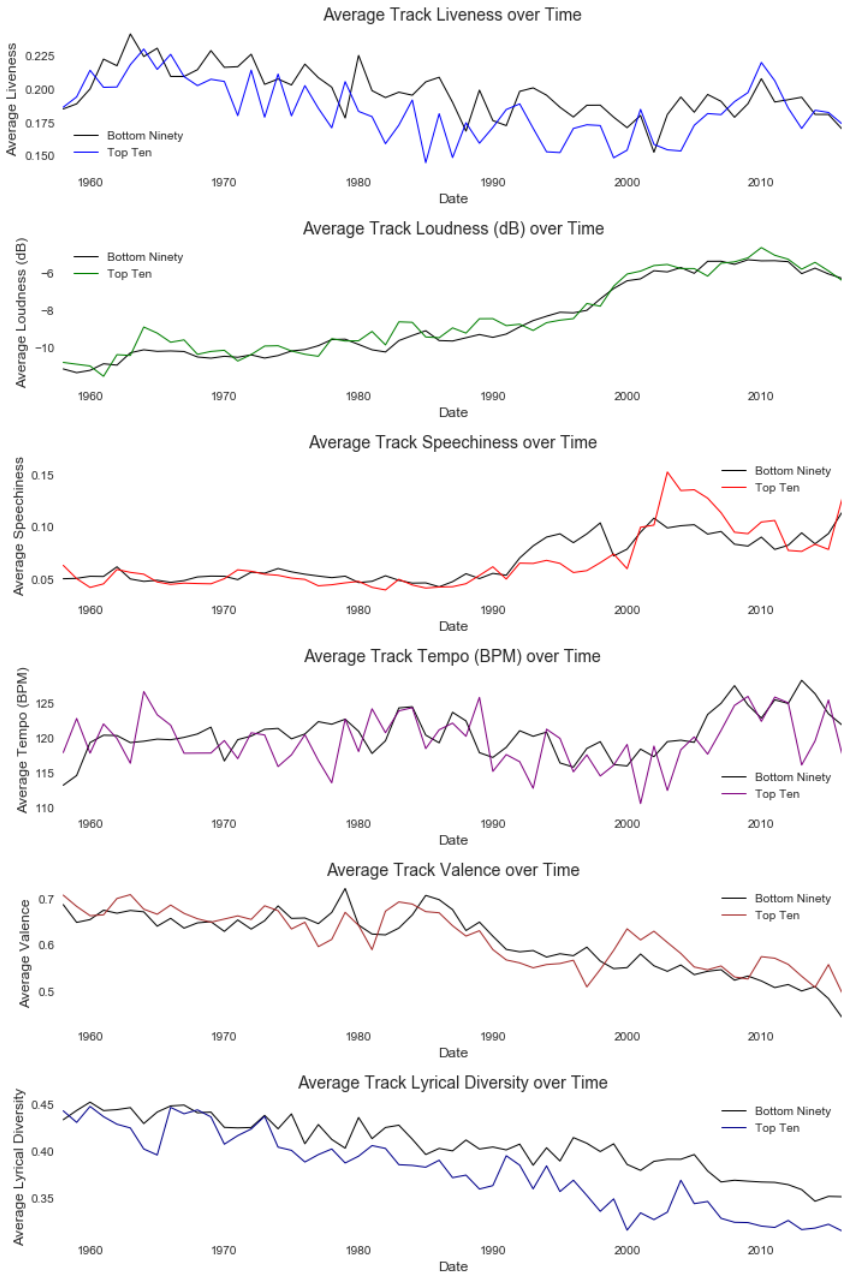
- *Acousticness*: Most songs had a higher acousticness score in the 60's possibly due to the popular of rockabilly and rock and roll which steadily

declined through the 70's and 80's as more electronic instruments were added.

- *Danceability*: Steadily increased through the 70's possibly due to the popularity of disco and continued increasing through the 80's and 90's. After 2000 the danceability of tracks in the top ten have been much higher than the bottom ninety.

- *Duration*: The average length of a song was around 2 and half minutes in the 1960's possibly tied to 7" EP format that singles were released on and increased up to a lenghth around 4 and a half minutes through the 70's until the 80's possibly as a result of the popularity of more complicated rock songs.

- *Energy* and *Instrumentalness*: Songs have become more energetic over time, possibly related to becoming more danceable and less instrumental possibly related to the increase in speechiness.



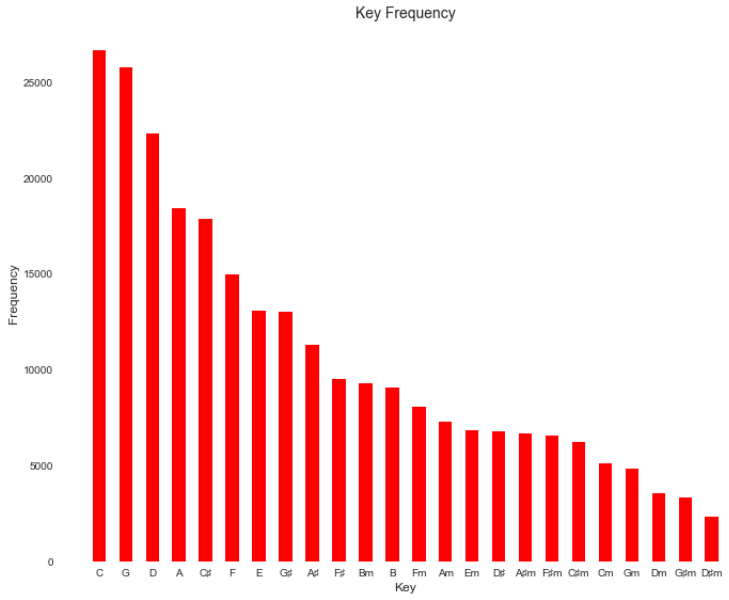
- *Liveness*: The liveness of tracks in the top ten have almost always been less on average than the bottom ninety.
- *Loudness*: We see a sharp increase in the 90's of Loudness which can be attributed to the popularity of the CD format and producers making songs

louder so that tracks would stand out more on the radio.

- *Speechiness* and *Valence*: The speechiness of top ten tracks was much less than the bottom ninety in the 90's, but this switched in 2001 and for the duration of the 2000's it was much higher. A similar trend is found with valence. Top ten songs had a lower valence from the mid 80's until the late 90's (which could be somewhat related to the popularity of grunge?). Then in the 2000's the top ten tracks began to have higher valence than the bottom ninety.
- *Lyrical Diversity*: The lyrical diversity of top ten hits have always been more repetitive than tracks in the bottom ninety and since the late 90's the gap between the lyrical diversity of top ten tracks and bottom ninety tracks has grown even wider.

3.2 Song Key

The key a song is written in has a big influence on how it sounds. For example songs in minor keys are generally sadder vs songs in major keys that sound more upbeat.

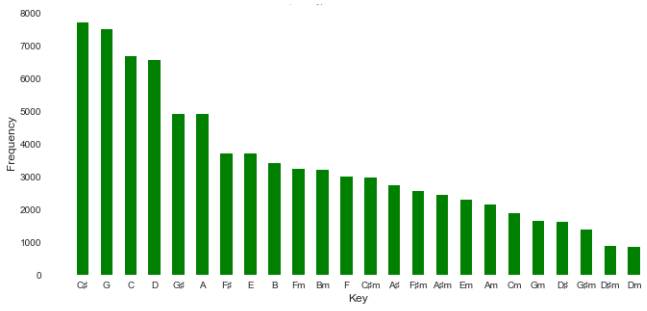


It's clear that major keys are much more popular than minor keys and we see that the most popular songs are in the key of C major followed closely by G major. This is likely a result of the fact that these keys are easy to play on both piano and guitar. In fact many songs are not only in the same key, but also the same chord progression as well. See this song by the Axis of Awesome

for reference. **Note**: that this differs slightly from the analysis of all songs on Spotify which found that the most popular key is G major.

Using K-Means clustering of key frequency by year we detect four distinct clusters. One cluster represents the earlier years on the charts from 1959 until the early 80's, a second cluster that represents every year since 2000 a third cluster that contains some of the 80's and every year in the 90's and a fourth cluster containing the incomplete years 1958 and 2017.

Figure 1: Key frequency for the years 2000-2016.



One interesting discovery is that the key C# has become much more popular over time and is now the most popular key whereas D# has become the least popular major key and is now less popular than almost all minor keys. Now that rock music is no longer the dominant genre it once was, it's possible that there are less songs being written with piano and guitar in mind. This may explain the decrease in the popularity of songs written in C and G major.

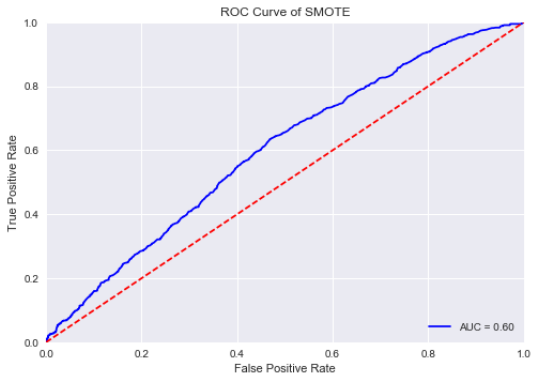
4 Predictive Modeling

We consider song popularity as a binary classification problem. To separate the data into more popular and less popular songs, we say that a song is popular if it becomes a top ten hit. This is given by songs with 'peakPos' <= 10. For the songs for which the Spotify audio metadata is present, we find that 20.484% of the songs are top ten hits. Thus, in the modeling phase we can obtain a 79.516% accuracy score by always predicting that the song is not a top ten hit. To account for this imbalance in positive and negative samples we use oversampling to ensure that relative class frequencies are approximately preserved in each train and validation fold. In particular, we use SMOTE - Synthetic Minority Over-sampling Technique to account for our imbalanced dataset.

Note: We only count each song once, since if we allow for songs to be counted with multiplicity then it might be possible that every song in the test data was actually already present in the training data, which could give a misleading high accuracy score.

We try multiple machine learning models like logistic regression, K-nearest neighbors and neural networks. The most accurate models we obtain come from a Random Forest Classifier, Support Vector Machine Classifier and a Gradient Boosted Machine Tree Classifier.

A random forest is an ensemble method that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control for over-fitting. The Random Forest Classification model results in 75.6% accuracy and an area under the curve of 0.6. The most important feature is mode, followed by year and key. Above we saw that key popularity has changed over time and different keys are much more popular than others. In particular major keys are more popular than minor keys so this model is consistent with that.



Support vector machines are a set of supervised learning methods used for classification, regression and outliers detection. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. The Support Vector Machine Classifier gives a 79.2% accuracy, which is even better than the Random Forest Classifier. However, the area under the curve is only 0.51 and one of the downfalls of SVC is that it does not tell us much about which features are most important in the model.

Gradient Boosting is a sequential technique which works on the principle of ensemble. It combines a set of weak learners and delivers improved prediction accuracy. GBM Tree gives a 70.7% accuracy with an area under the curve of

0.6. Similar to the Random Forest Classifier key, mode and year are the most important features.

4.1 Conclusion

From our investigation into key popularity we have found that the most popular keys have changed over time from being focused on the key of C and G to C now being the most popular key. Throughout the history of the Billboard Hot 100, major keys have been more popular than minor keys. From our machine learning models we find that these features turn out to be the most important features when predicting song popularity.

On the other hand lyrical diversity has steadily decreased on average over the years, and recently the most popular songs feature much more repetition than the other songs on the Billboard Hot 100. However, lyrical diversity did not show up as a strong indicator of popularity in any of the models we considered.

The best performing models were Support Vector Machine Classifier and Random Forest Classifer although neither were able to obtain a model with a better than 80% accuracy, so there is room for improvement.

In the Million Song Dataset there is a measurement of artist "hotness" and genre, for future work this could be nice to incorporate into our model to help improve it. For this project we did not get a chance to fully make use of the lyrics data so it could be interesting to explore how lyric sentiment influences song popularity.

5 Recommendations

- For songwriters looking to write popular songs one suggestion is to write lyrics with a lot of repetition in them and to write songs in major keys.

These are not exactly new revelations, but it is interesting to see how the lyrical diversity and key popularity has changed over the years and this could be taken into account in the song writing process.

- Producers can influence things like danceability and loudness and these features have a small influence on song popularity, but are worthwhile to take into account.
- For music streaming services, it would be interesting to take lyrical diversity into account in their recommendation model since it seems to be related to popularity, but this feature will hopefully be explored further in the future.