Finding the Formula to Create the Hits of Tomorrow

An exploration by Sebastian Engels

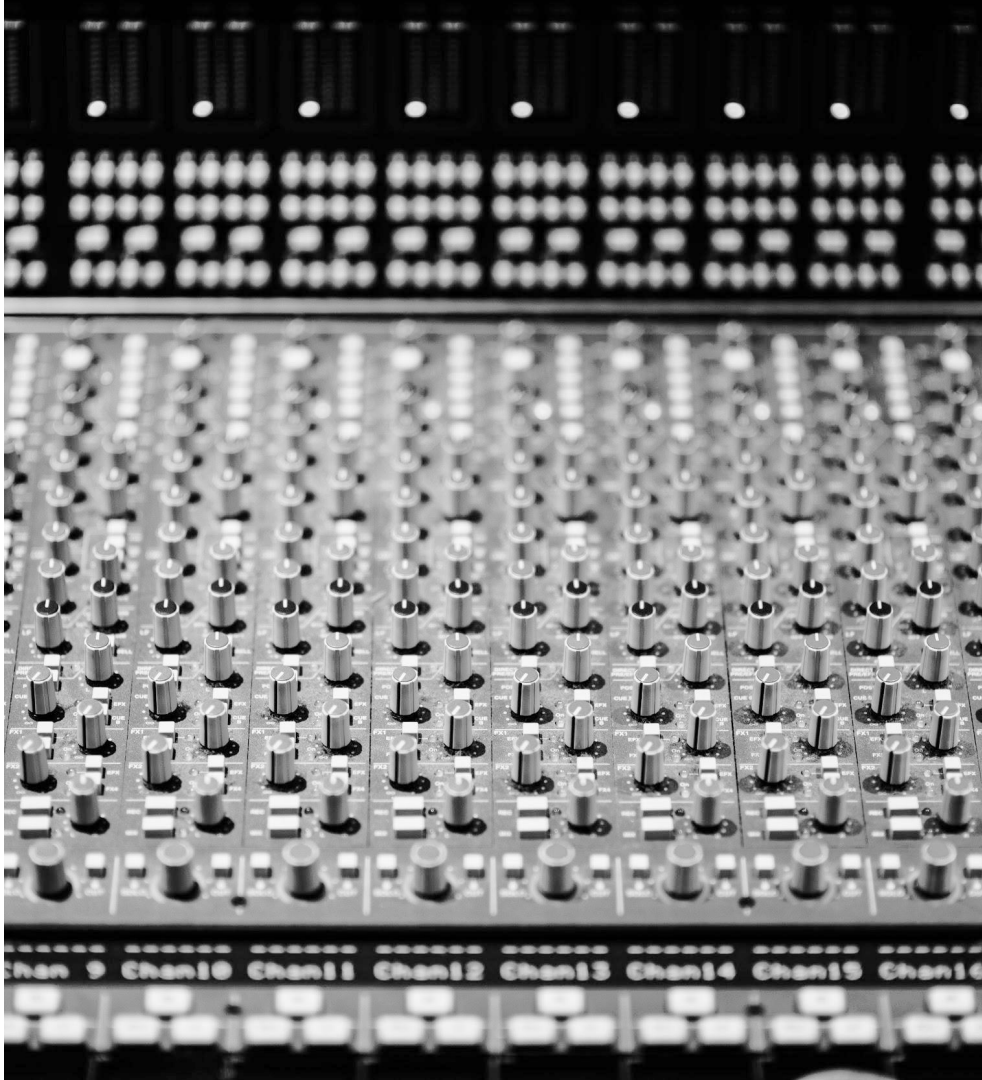
Every year >$4.5 Billion are invested into signing artists.

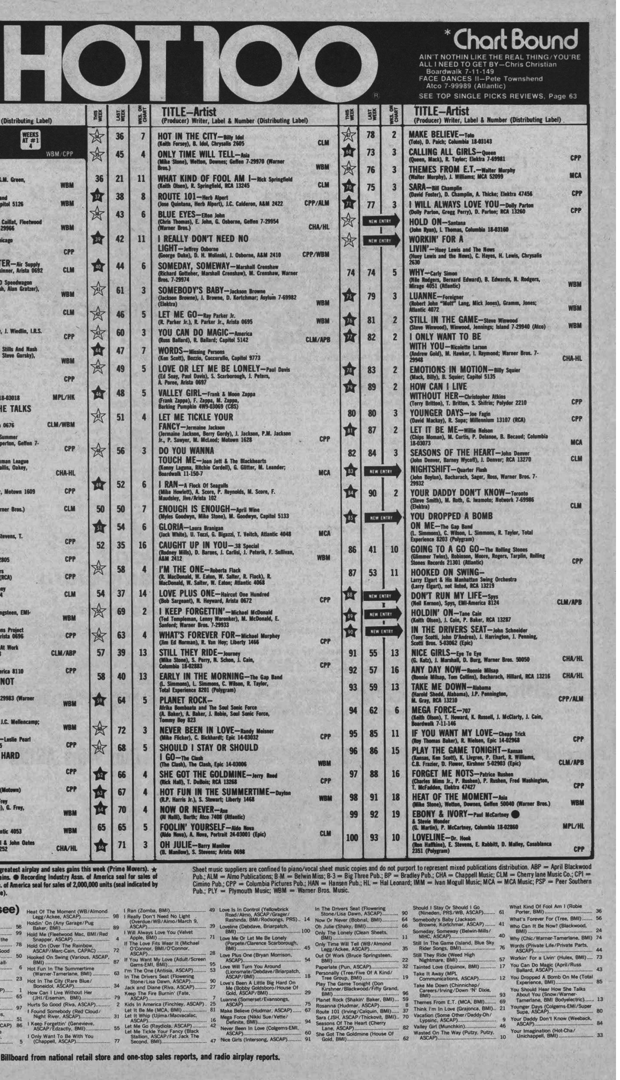
What would you do to lighten that burden?

Know the Future! It's as simple as that.

(Well, sort of. But we got off to a start.)

Intrinsic features can be used to **IDENTIFY SONGS THAT BECOME POPULAR** to some extent. From Justin Bieber to Nicky Minaj.





What is Popularity?

Popularity is hard to define but we often have an intuitive understanding of what's popular after the fact - Justin Bieber is. The IRS is not.

For the music industry popularity is often business values (i.e. brand awareness or records sold)

The Hot 100 measures popularity using commercial factors of popularity.

The Data

The main source of popular songs was the Hot 100. Non-Hits and Audio Features were sampled from the Spotify API.

More than 60 Years of Historic Popularity Data

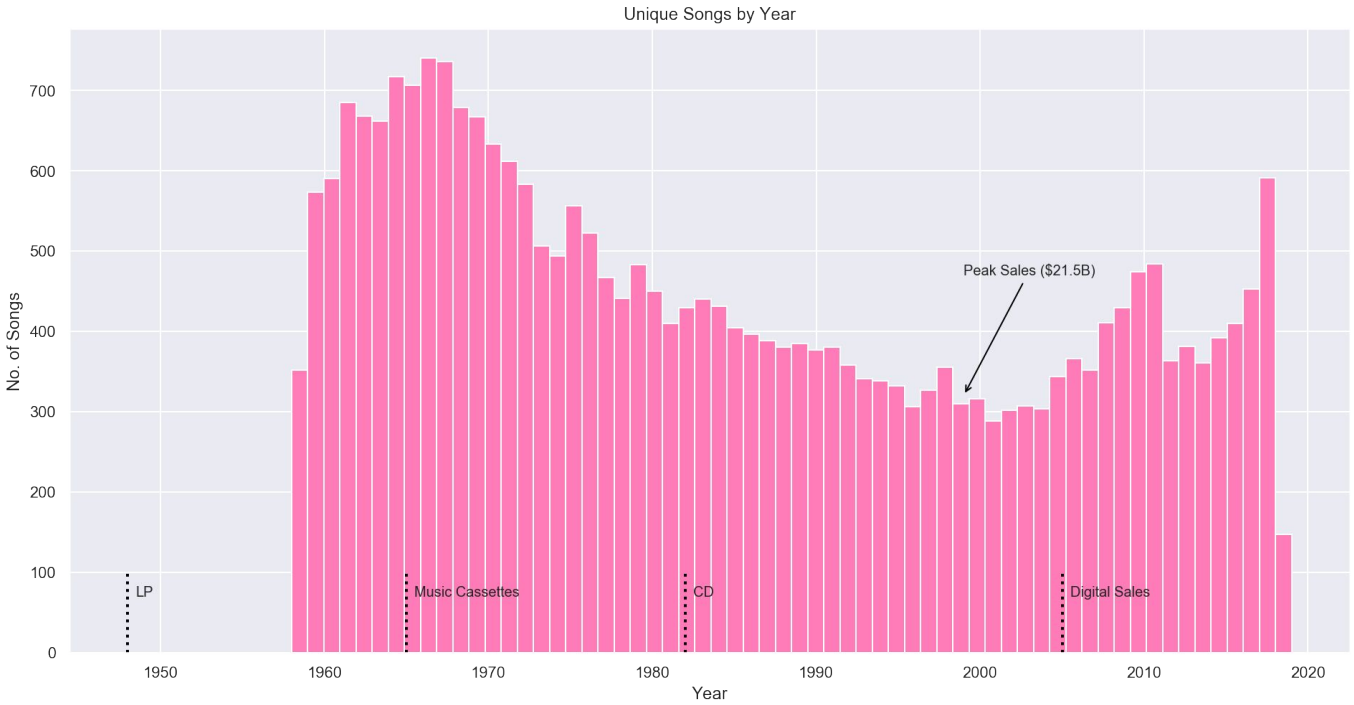
Weekly Snapshots of the 100 most popular songs

30 Million Songs on Spotify to sample Non-Hits

13 Audio Features to describe a song in numerical terms

Understanding the Hot 100 and its Hits.

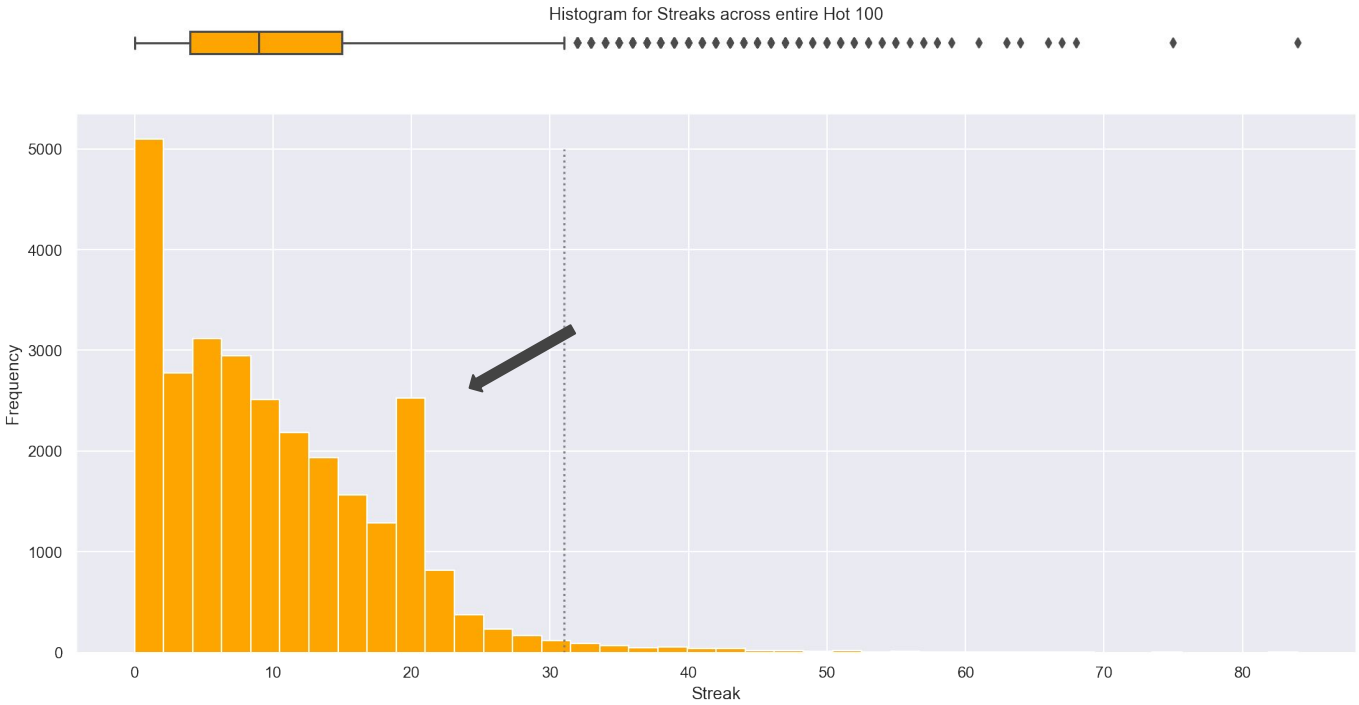
A look underneath the surface.

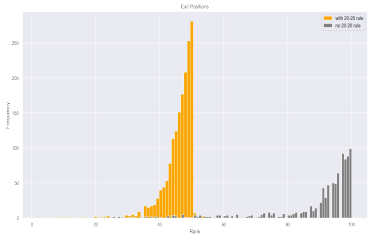


How many songs have been on the Hot 100?

Since 1958 there were *28083* songs on the Billboard Hot 100.

Declining Diversity of Unique Songs since the 1970s. Recent uptick though.

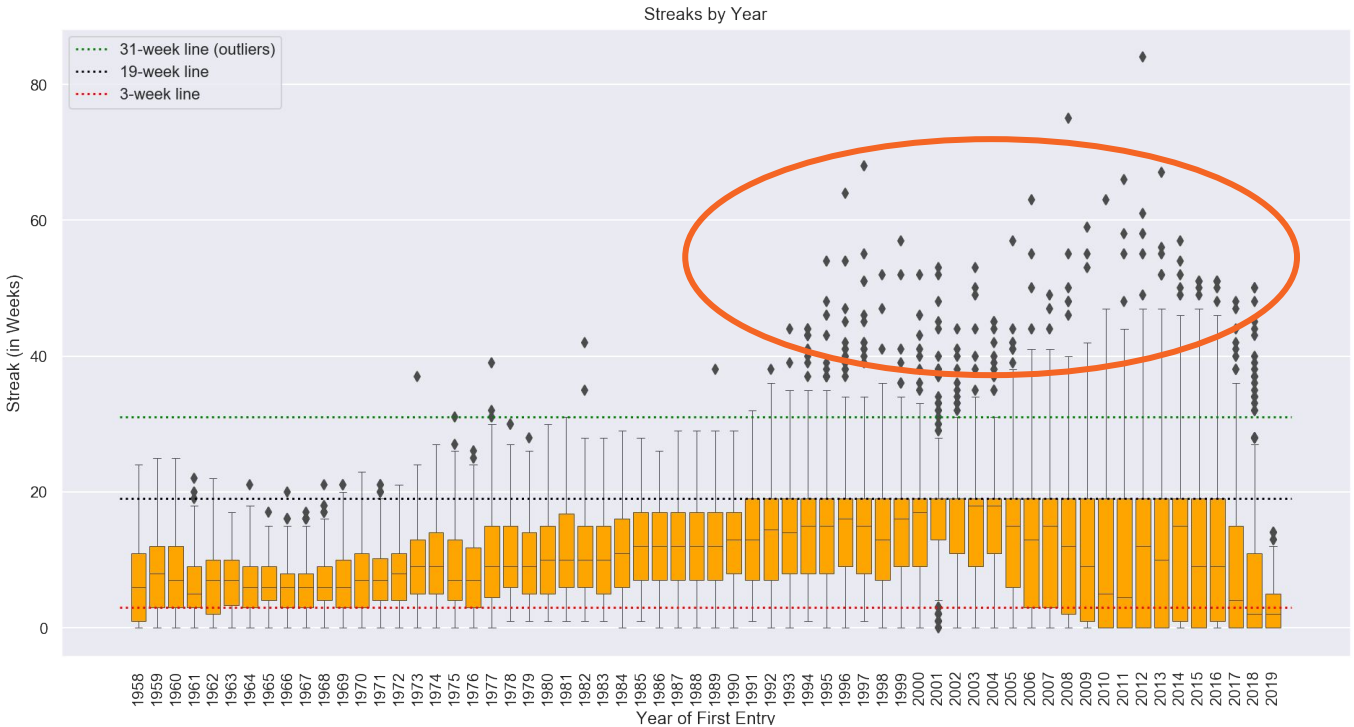




What is the 20-20 Rule?

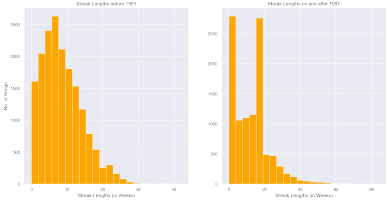
Introduced with the intent to accelerate fluctuation, the rule states that songs that fall out of the top 20 will be removed after 20 weeks. The rule was relaxed in 1992 and 1993 to falling out of the top 40 and top 50 respectively.

The 20-20 Rule is a formulaic quirk of the Hot 100 and clearly visible



Nielsen Soundscan

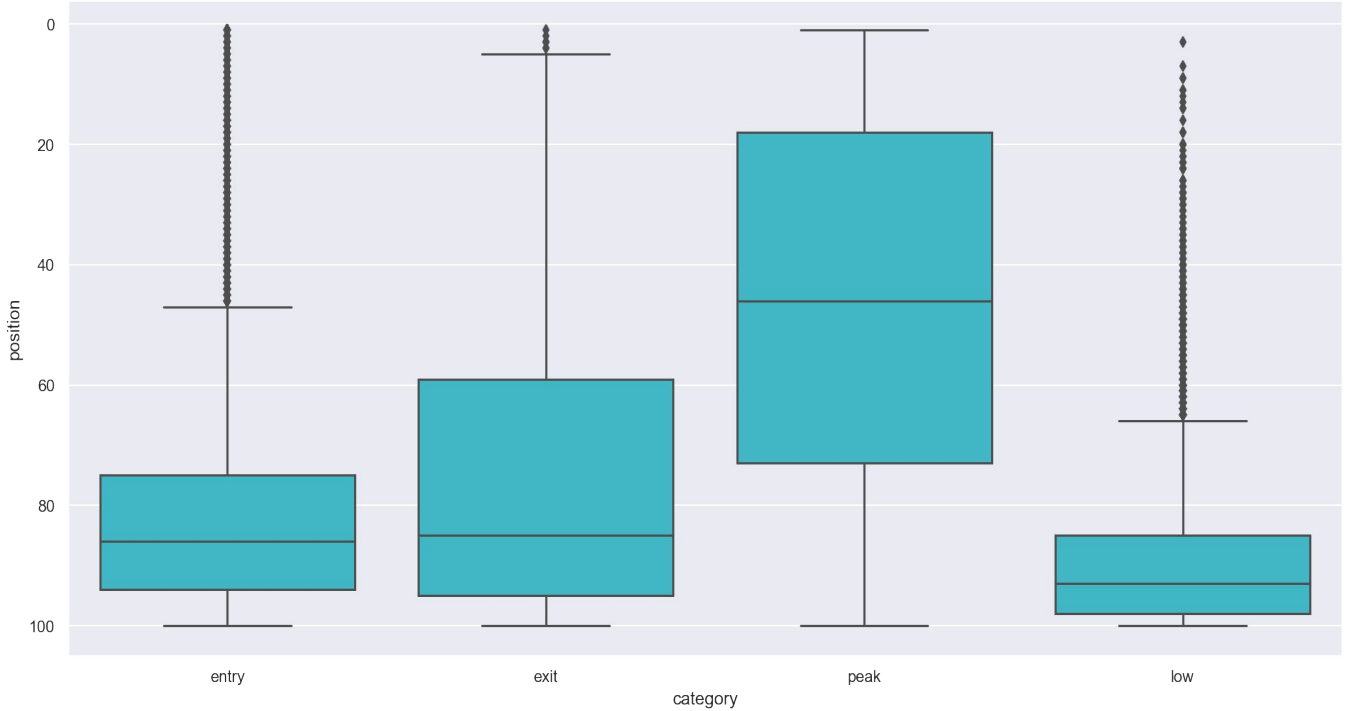
In 1991, the Hot 100 are starting to use Nielsen Soundscan to track their sales data. This change heavily impacted the structure and distribution of Hot 100 songs.



What songs stayed the longest?

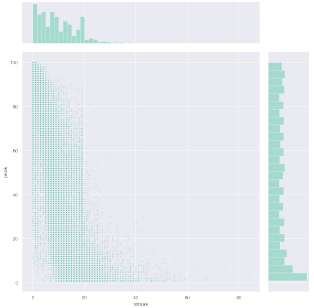
Radioactive by Imagine Dragons was continuously on the Hot 100 for 85 weeks. Jason Mraz's I'm Yours comes in second with 76 weeks. Kudos!

"Super Songs" emerge after 1991 i.e. songs that stay on the Hot 100 for >6 months.

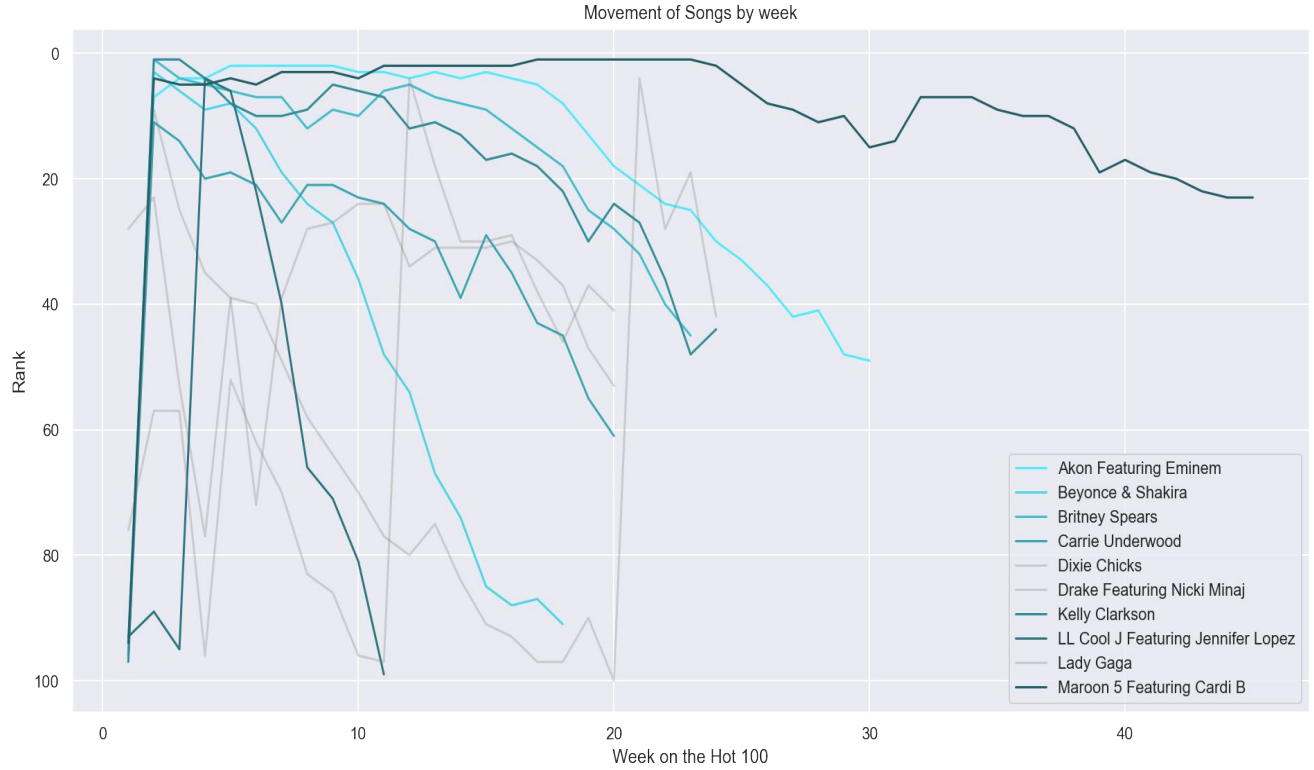


The Curious Nr.1 Position

Below, we can see that the higher a song peaks the longer its streak tends to last. Another interesting observation is the the histogram to the right shows that most positions are almost uniformly distributed but that the pole position has the highest fluctuation rate. Despacito (Justin Bieber) and One Sweet Day (Mariah Carey) have both lasted the longest on Nr. 1, a very respectable 16 weeks.



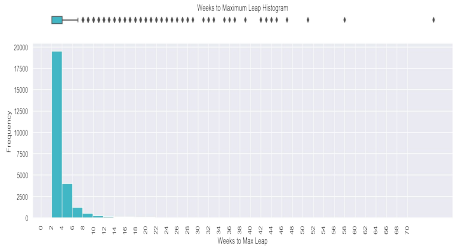
Less than 50% make it into the Top 40 and most Enter and Exit at the Bottom.



The Biggest Leaps (>90)

There are 6 songs that leaped more than 90 positions from one week to another:

Dixie Chicks - Not Ready To Make Nice
Kelly Clarkson - My Life Would Suck Without You
Britney Spears - Womanizer
Lady Gaga - Million Reasons
LL Cool J Feat. . . . - Control Myself Beyonce & Shakira - Beautiful Liar



Most Max Leaps happen within the first 4 weeks of breaking onto the Hot 100.

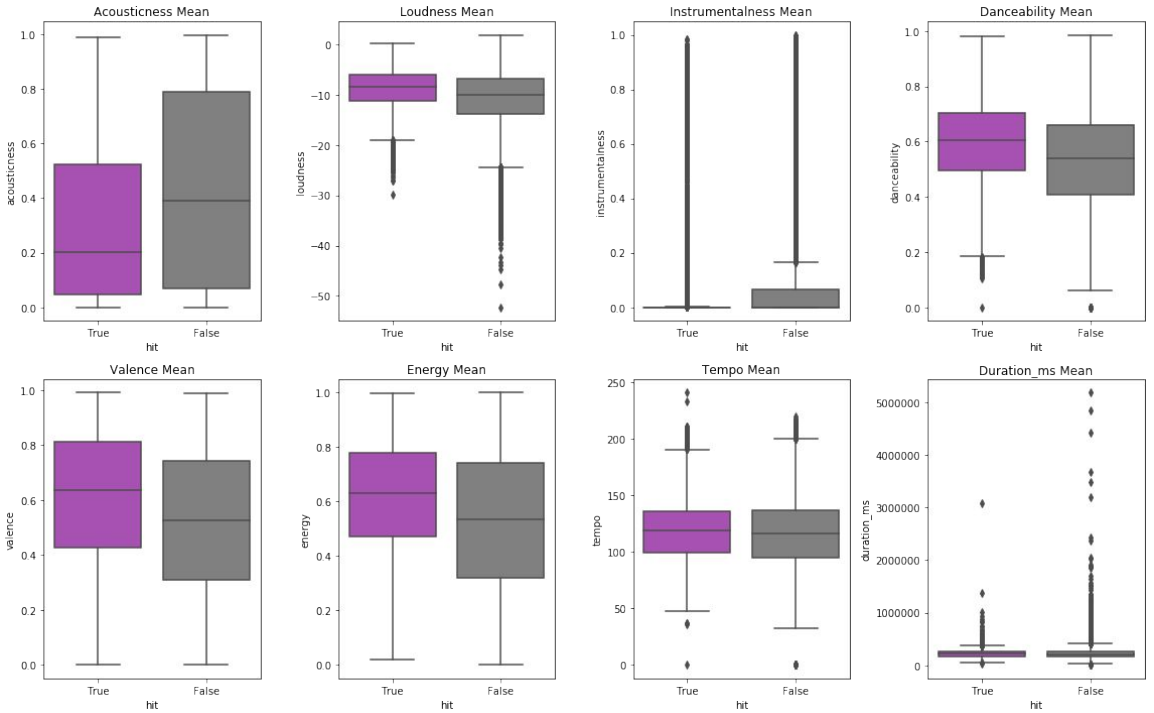
The Inner Workings of Hits.

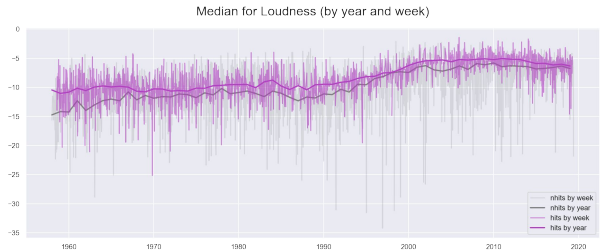
Using Numerical Audio Features.

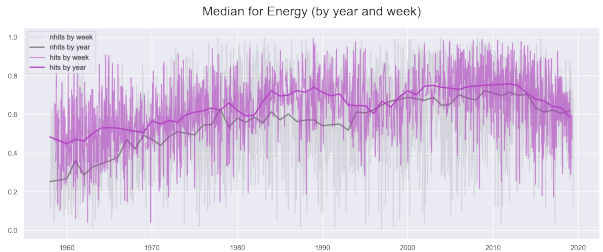
The Feature Overview

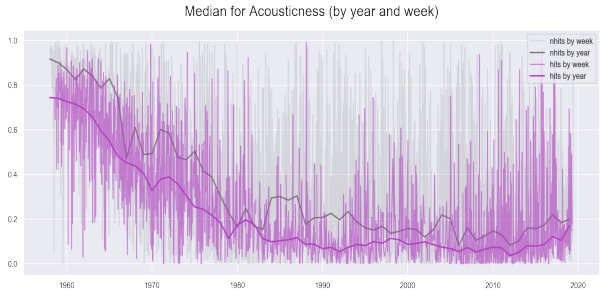
The features that weren't removed in the first step and are being looked at more closely are

- Acousticness
- Loudness
- Instrumentalness
- Danceability
- Valence (Positivity)
- Energy
- Tempo
- Duration (in ms)









Loudness, Energy and Acousticness

Loudness and Energy are strongly related features and are showing the rise of louder and more energetic music.

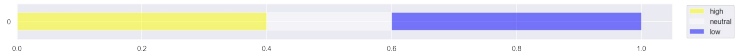
Acousticness shows this as well but from the opposite end the drop in acoustic songs.

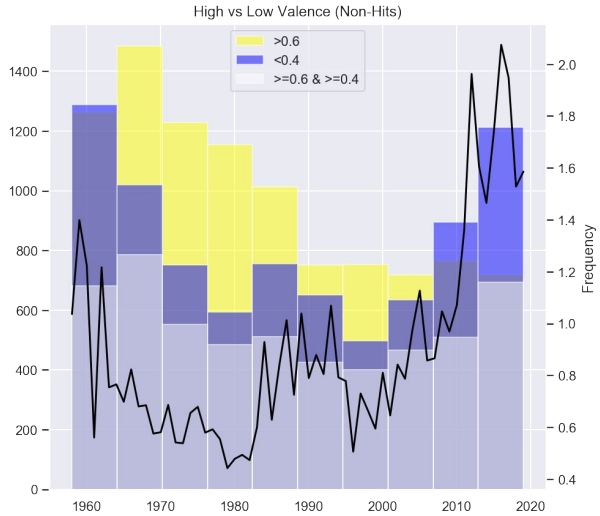
All features show that Hits (purple) tend to reinforce the trend by being persistently above or below the Non-Hits.

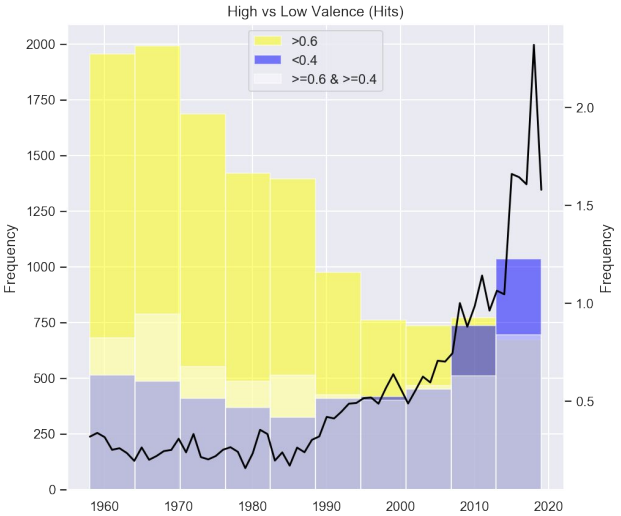
Our Music Has Become Less Positive

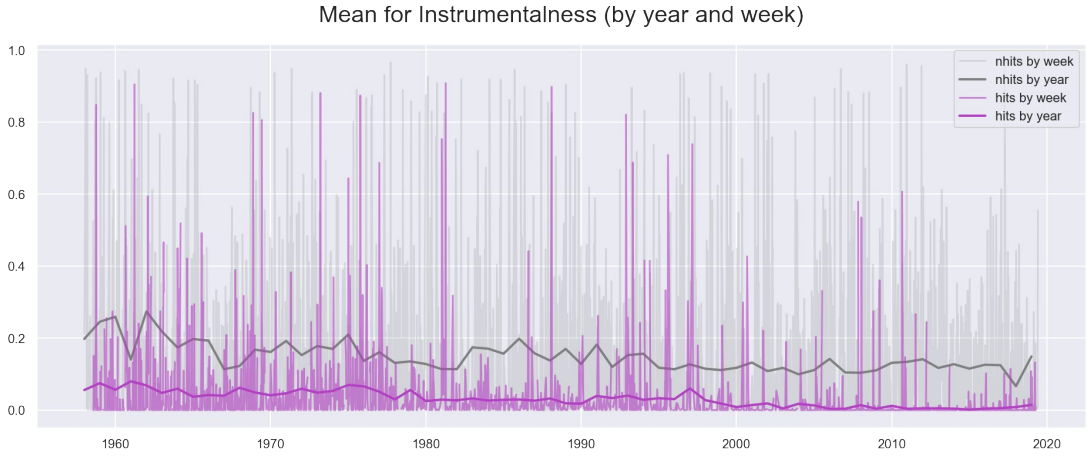
Valence measures the positiveness of a song.

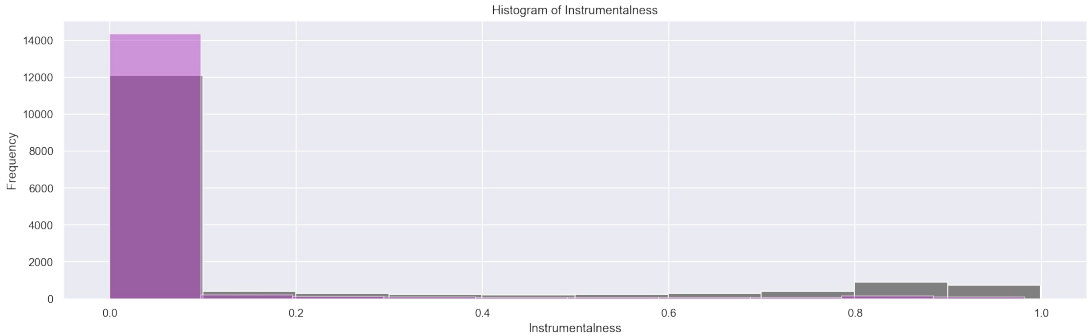
Hits and Non-Hits have increasingly become less positive since 1958.











Instrumentalness is not Popular

The Instrumentalness feature shows the clear bias of popular music toward Non-Instrumentalness.

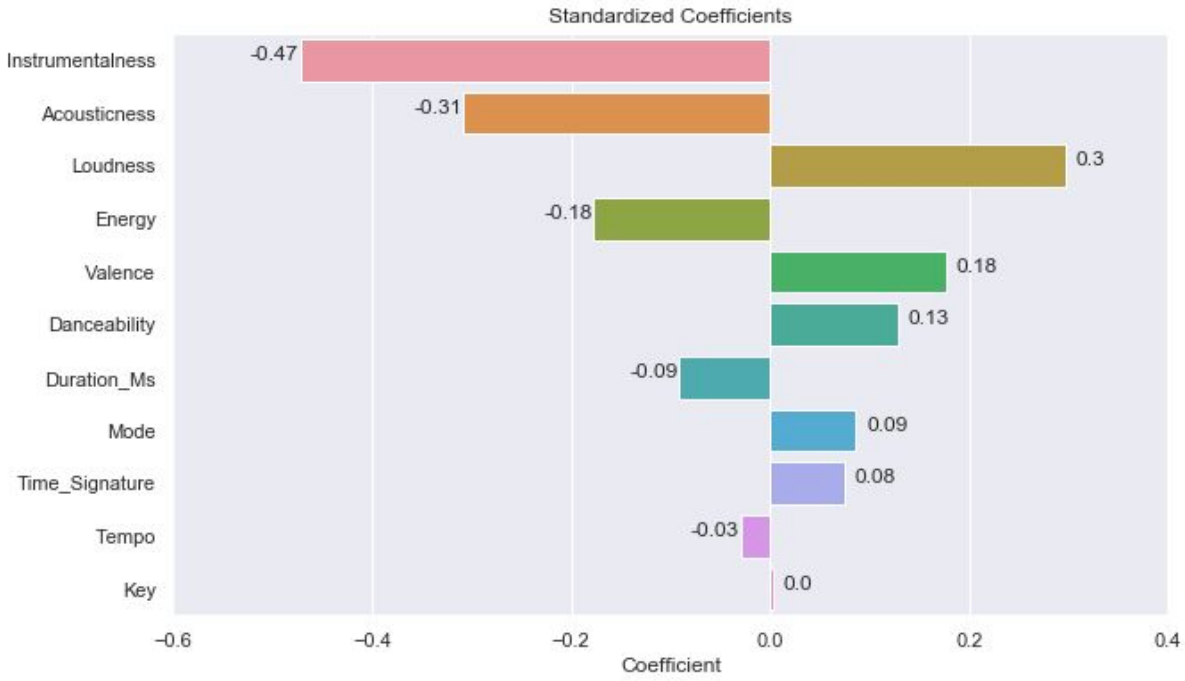
The Feature Importance

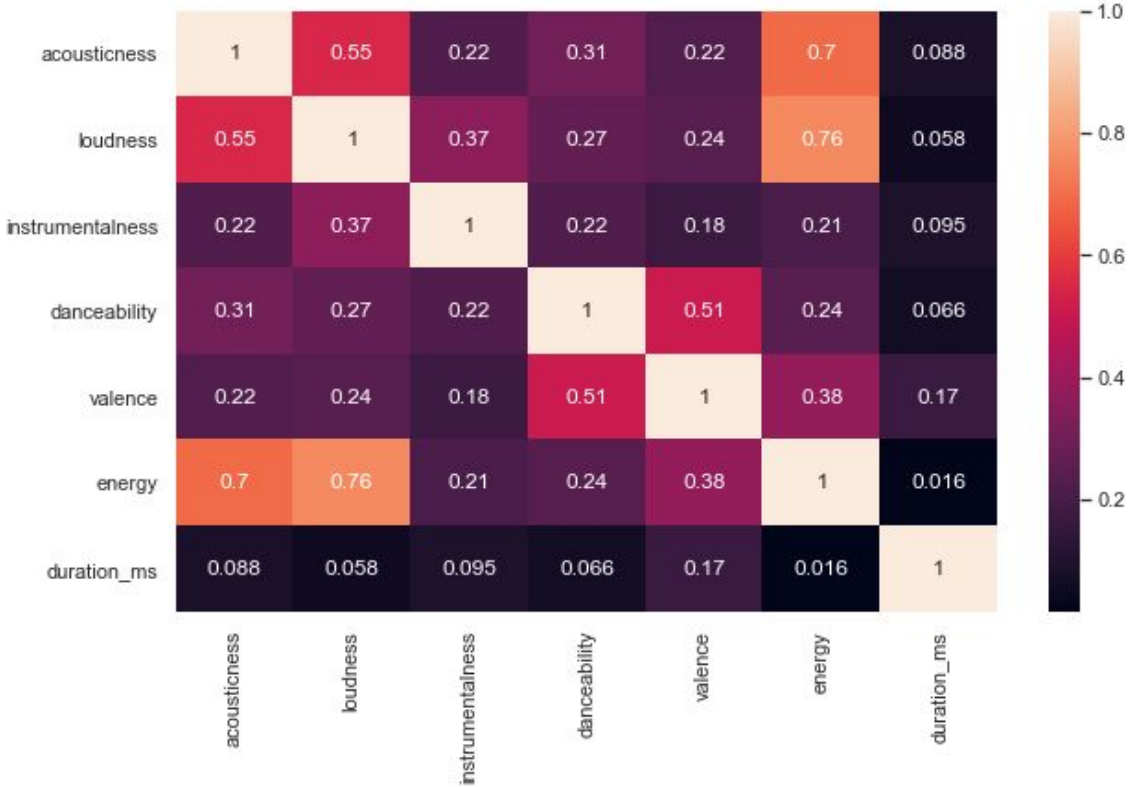
Using Logistic Regression we're looking at the Beta coefficient of standardized features to understand what features provide most explanation of variance.

Winners are Instrumentalness, Acousticness and Loudness.

Losers are Tempo and the discrete variables Key, Mode and Time Signature.

The Losers will be removed.





Collinearity

Correlation between predictor features should be avoided as it comes with commonly known down-sides.

We're seeing strong correlation between Energy, Loudness and Acousticness.

Furthermore, there is a moderate correlation between Acousticness and Loudness as well as Valence and Danceability.

From features to formula

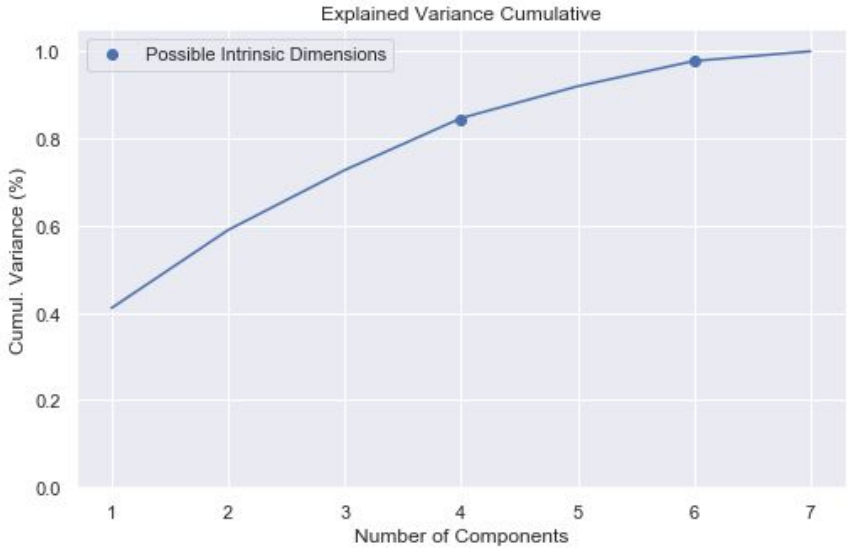
Machine Learning allows us to test the validity of our model and create incremental improvements by re-evaluating the model through time with new predictions.

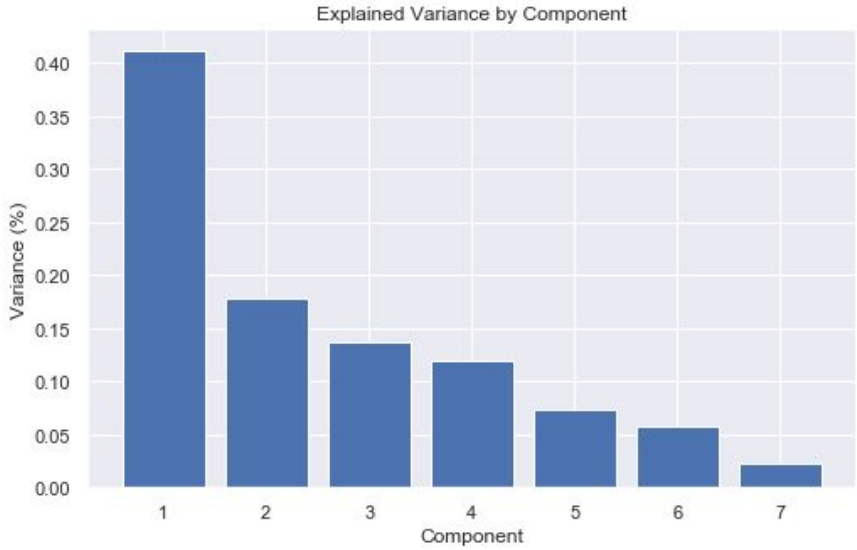
Dimensionality Reduction

Principal Component Analysis (PCA) allows us to remove collinearity of features by attempting to reduce the feature set to their intrinsic dimension.

Using PCA we lose the ability to explain via features how an individual decisions came about.

Our model is going to reduce dimensions to 4 components.





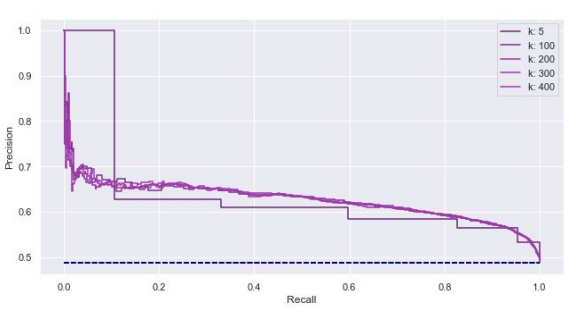
Hyperparameter Optimization

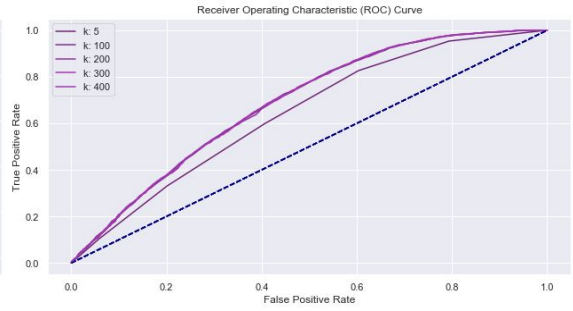
We've used k-Nearest Neighbor (kNN) and Random Forest Classifiers (RF). Both are popular choices.

kNN yielded significantly better Recall results (+15 points in comparison). Precision was slightly worse than Random Forest (-1.5 points in comparison).

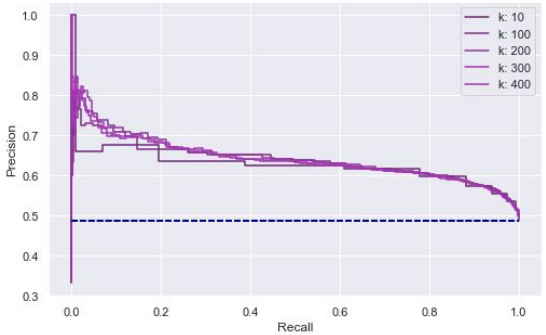
Optimal k-Neighbors Hyperparameter: 200

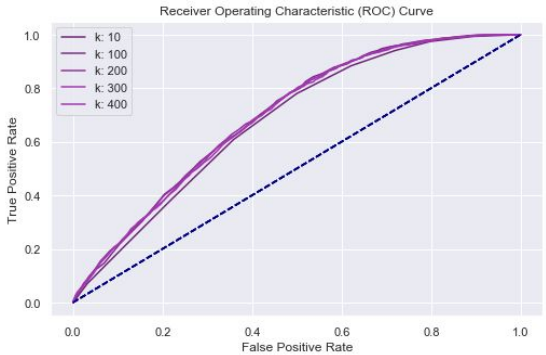
Random Forest Classifier





k-Nearest Neighbor Classifier





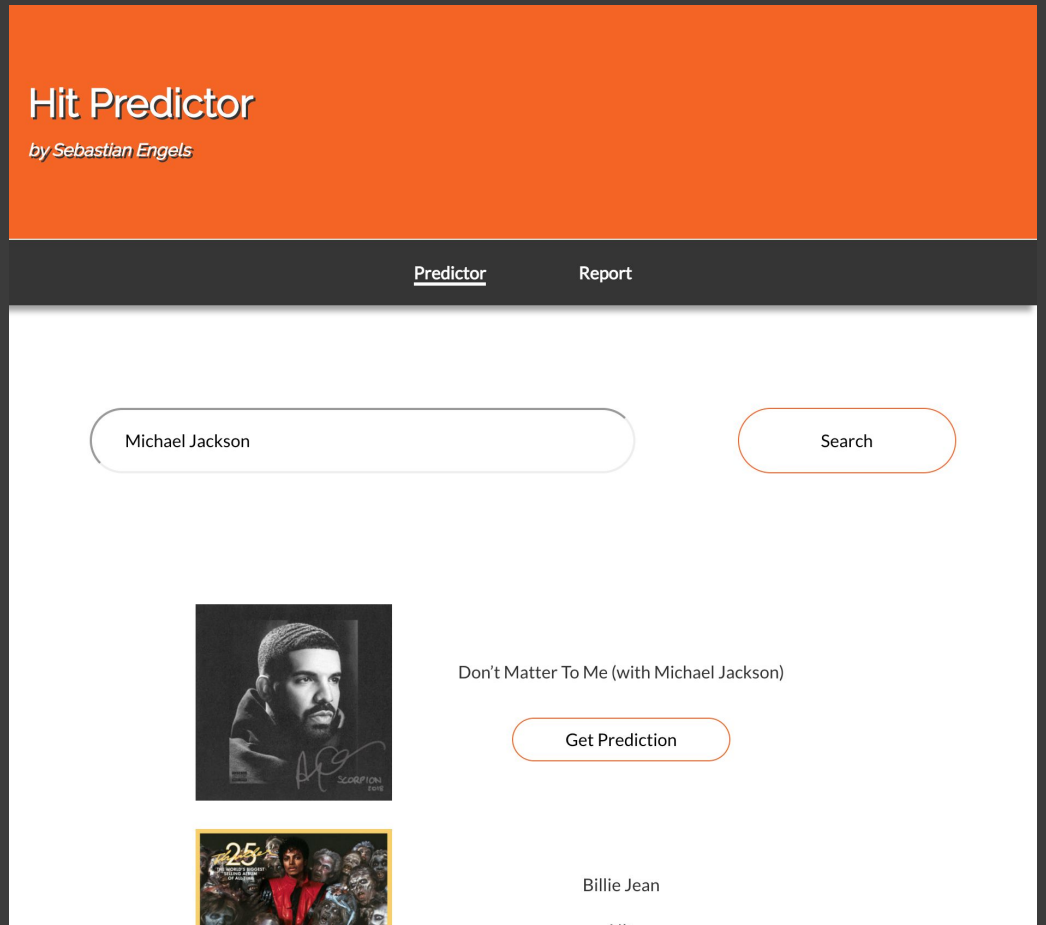
A model that isn't perfect but shows that popularity to some extent is ingrained in the music itself.

The model is 59% of the time correct and finds 83% of all Hits.

This model made the attempt to explain a part of hit popularity.

But that's only one piece of the cake. Popularity is also the Artist's Brand, Network Buzz, Fan-base, Marketing Budget etc. …

Iteration is necessary to create a more powerful model.



Try out the Prediction App

In the spirit of continuous learning and improvement the model was deployed and is available at *https://goofy-shirley-91ba1b.netlify.com/*