Music Popularity Predictor

Greg Welliver

Problem Statement & Objective

Problem Statement & Objective

- Question: How possible is it to predict whether or not a song will make the Top Hits in a year, given all of the attributes that we know about the song?

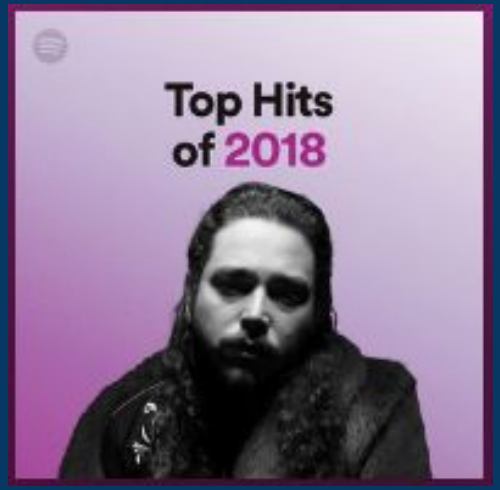
- The objective of this project was to investigate the characteristics of songs that make the Top 100 list, identifying patterns or commonalities for what gets a song in the Top 100.

- The end result is a predictive model where a song can be input into the model, and the model will return whether or not this song will make the Top 100 list.


Top Hits
of 2001


Top Hits
of 2008


Top Hits
of 2016


Top Hits
of 2018

Data Overview

Data

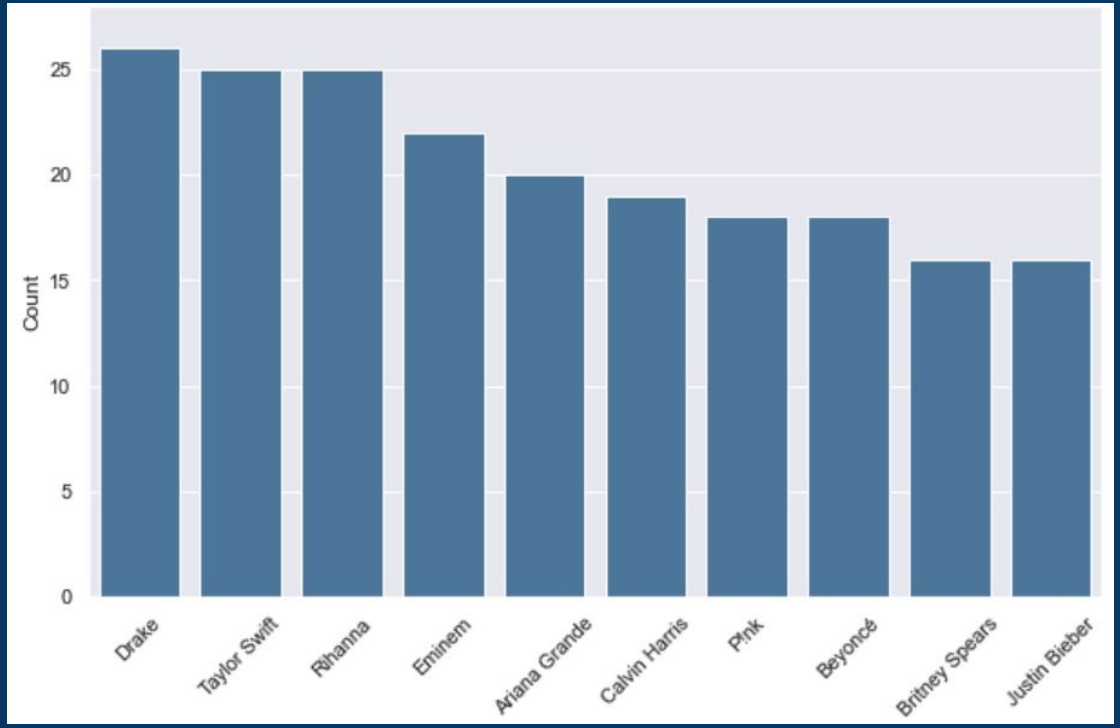
- Data sources: Spotify API, Publicly available datasets on Kaggle

- Data highlights:
  - Collected information on 255.096 tracks; included 1,863 tracks that made the Top 100 playlist in their respective years (89% coverage)
  - Time period was 2000-2021

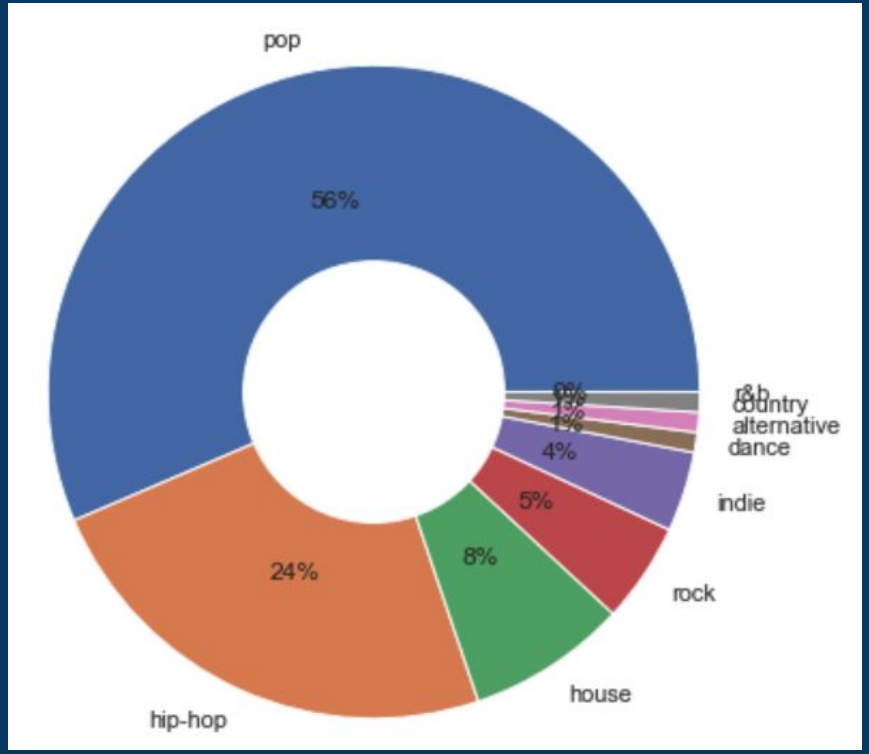
- Data for each track included:
  - Informational metadata such as track name, year, genre, artist
  - Advanced music data such as Key, Acousticness, Danceability

Exploratory Data Analysis

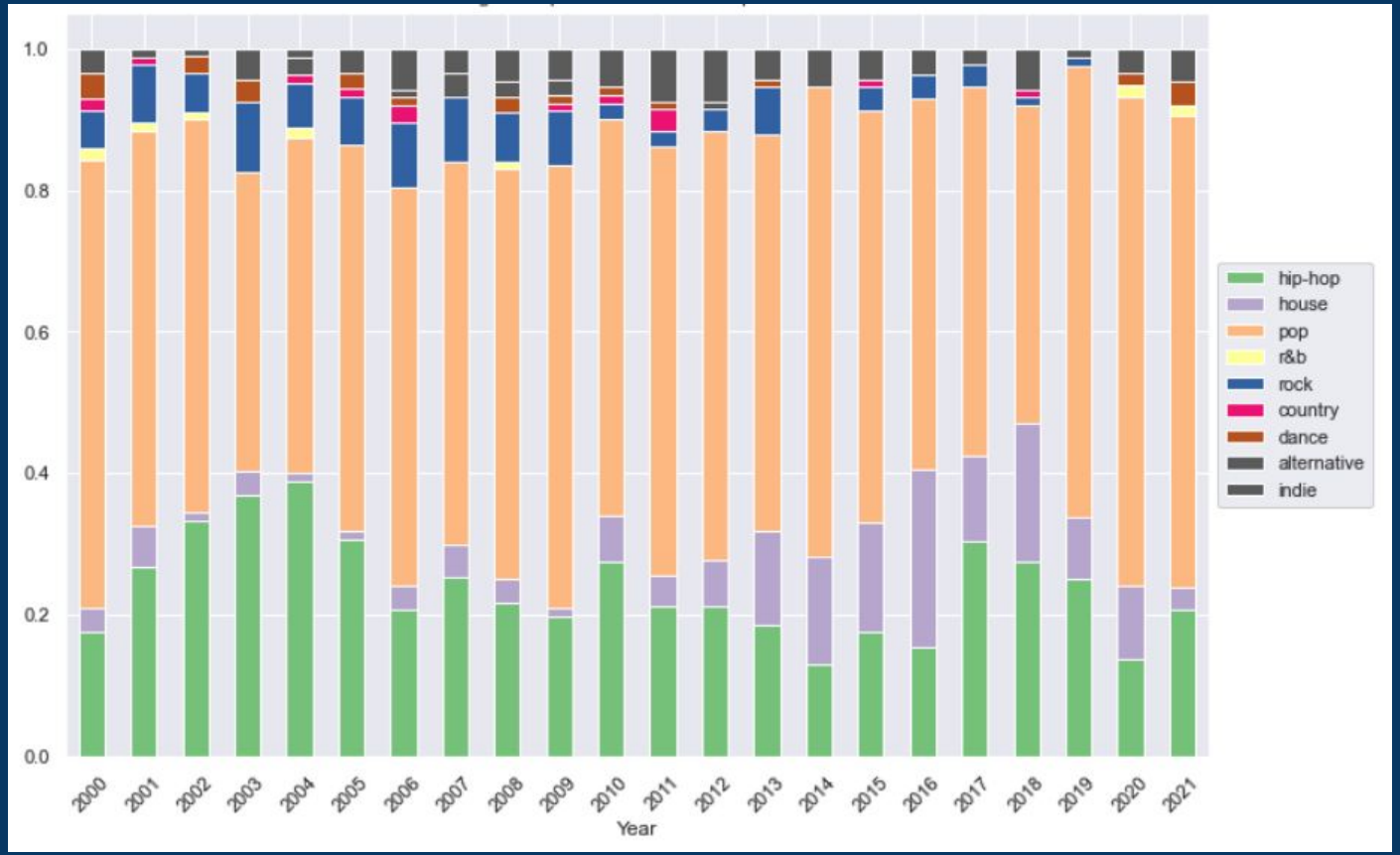
Artists with the most songs in the Top 100



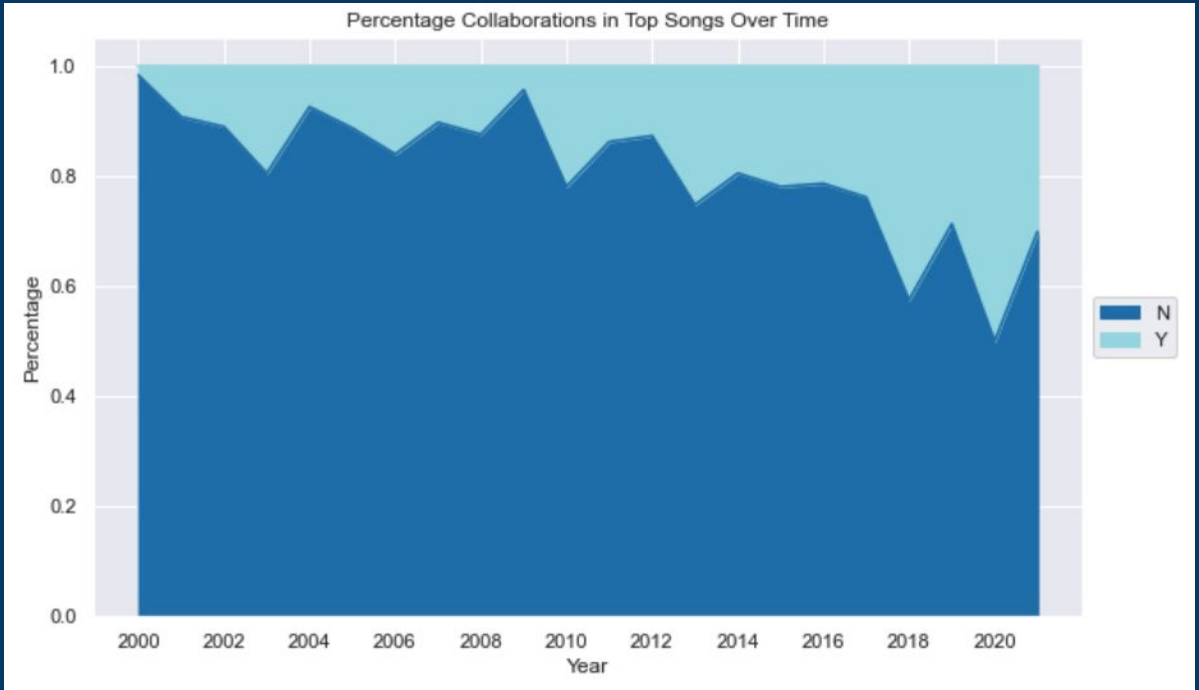
Distribution of genres in the Top 100 playlists



Trend in genre popularity over time



Trend in collaborations over time



Correlation heatmap for audio features

Algorithms and Modeling

Model Selection

- I chose to implement the following algorithms in my pursuit of the most effective model for this classification problem. All are similar in terms of complexity, with the exception of Random Forest, which is more complex than the rest.
    - Logistic Regression
    - Random Forest
    - Decision Tree
    - XGBoost
    - Naive Bayes
    - K-Nearest Neighbor

Metric Selection

- The most important outcome that we are looking for is the ability to correctly predict a hit song. Accordingly, I rank the outcomes in order of importance here:
    - True Positives: the ability to predict a hit song would be incredibly valuable to artists, management, and labels
    - False Positives: it could be costly to invest heavily in artists, only for it to turn out that they did not produce a hit
    - False Negatives: This could be detrimental as it presents a missed opportunity to promote an artist
    - True Negatives: since the vast majority of songs will not make the Top 100 hits, predicting this outcome is not especially useful
- **Based on the above ranking, it would follow that Precision would be our most important metric.**

Model Metrics - Sorted By Precision Test Score

Algorithm	Precision		Recall		F1		Accuracy		
	Training	Test	Training	Test	Training	Test	Training	Test	Comments
Logistic Regression	0.92	0.87	0.92	0.87	0.92	0.87	0.92	0.87	Suitable
Random Forest	0.86	0.85	0.85	0.83	0.85	0.83	0.85	0.83	Suitable
XGBoost	1.00	0.86	1.00	0.86	1.00	0.86	1.00	0.86	Overfitting
KNN	0.90	0.83	0.89	0.83	0.89	0.83	0.89	0.83	Suitable
Decision Tree									
Gini model - no max_depth	1.00	0.83	1.00	0.83	1.00	0.83	1.00	0.83	Overfitting
Entropy model - max_depth 3	0.85	0.83	0.85	0.82	0.85	0.82	0.85	0.82	Suitable
Gini model - max_depth 3	0.85	0.83	0.85	0.82	0.85	0.82	0.85	0.82	Suitable
Entropy model - no max_depth	1.00	0.81	1.00	0.81	1.00	0.81	1.00	0.81	Overfitting
Naive Bayes	0.94	0.78	0.93	0.73	0.93	0.71	0.93	0.72	Overfitting

Future Improvements, Opportunities for Pursuit

Future Improvements

- There are a number of additional data points that I believe could have predictive value, but were excluded in this initial analysis due to difficulties in data collection. These include:
    - Whether an artist has previously appeared in a Top 100 playlist
    - Whether a collaboration artist has previously appeared in the Top 100
    - Country of origin for artist
    - Time of year of track release
    - Did the artist go on tour in the given year?
    - Did the track appear in a film or television show?

- I intentionally reduced the number of tracks that were not Top 100 hits in order to avoid class imbalance issues. It may be useful to include some additional data to see if this improves model performance, while not creating any issues with class imbalance.