Music Popularity Predictor - Summary Report

Music Popularity Predictor - Summary Report

Github Repository

Tables of Contents

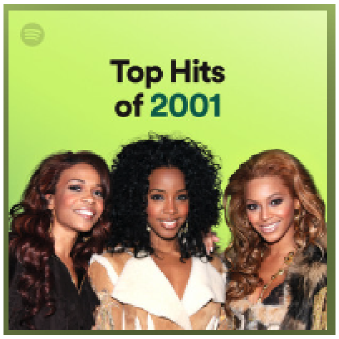
Introduction

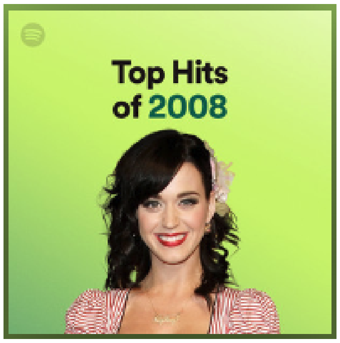
Music is a powerful artistic form that permeates cultures globally. Music streaming has been a driver of the industry for the last two decades; currently, there are over 600 million users on streaming services worldwide, and over $17 billion in revenue.

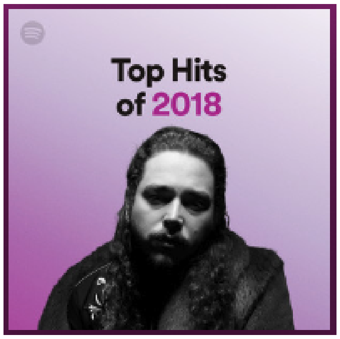
Spotify is one of the premier music streaming services in the world, and boasts some impressive statistics which underscores its stature in the global streaming industry:

- Approximately 200 million users.
- Global catalog of 100 million songs, with an estimated 60,000 new songs added every day.
- 4 billion user-generated playlists, and 3,000 Spotify editorial playlists

Each year, Spotify publishes a playlist of the Top 100 songs for that year - this playlist is called the “Top Hits of <year>”







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.

Potential users of the model

- Streaming services - use cases such as content curation, selecting artists for promotion
- Artist management - help understand what makes certain music popular, which could be used in collaboration with artists
- Artists - become informed about characteristics that make for popular music, if their desire is to become more popular
- Marketing teams interested in collaborating with artists - identify artists that could be emerging in the space, which would provide great value in establishing a partnership before they become famous

Data

| Data Collection and Refinement Notebook **|**

As mentioned above, Spotify boasts an enormous catalog of songs, with metadata like Artist, Album Name, and Genre. Additionally, Spotify also tracks advanced audio features for songs, such as acousticness, danceability, and tempo. One of the biggest challenges of the project was using Spotify's API to pull useful information. I enlisted the help of a library called Spotipy, and referenced some projects done by others in the developer community for help.

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 that I collected for each song:

- Informational metadata:
  - Track id, Track name, Track Year, Track Genre, Artist Name, Artist ID, Artist Country, Record Label

- Advanced music data
  - Key, Acousticness, Danceability, Duration, Energy, Instrumentalness, Liveness, Loudness, Modality, Speechiness, Tempo, Time Signature, Valance
  - This data is mostly numeric in nature, with the exception of Key, Modality and Time Signature, which are categorical

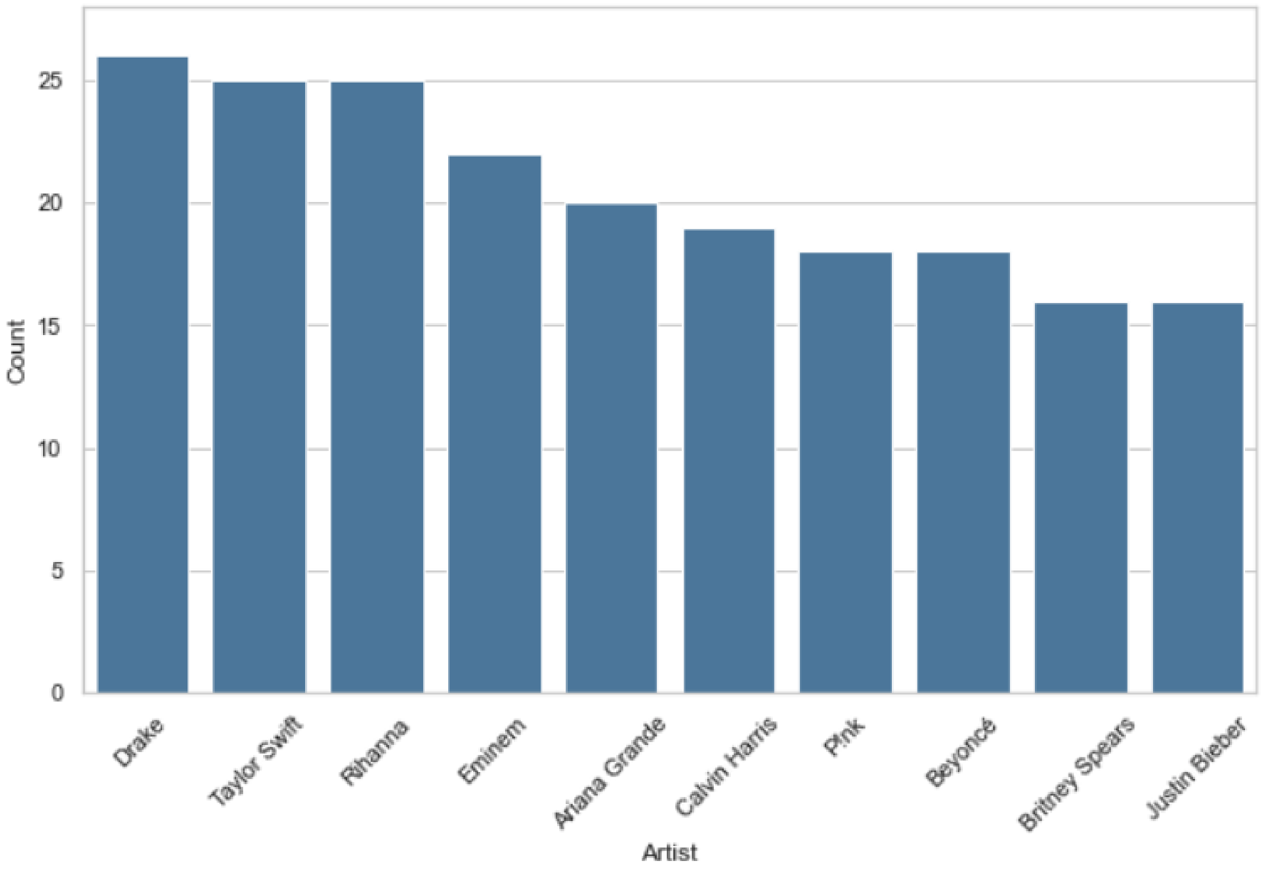
- Features that I created:
  - Whether the song was a collaboration with another artist

Exploratory Data Analysis

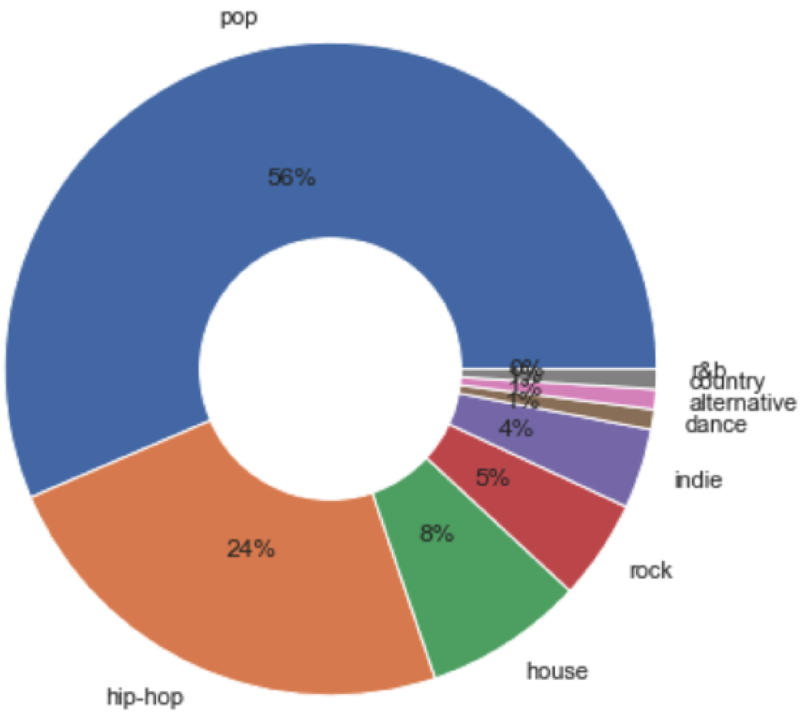
| EDA Notebook **|**

Below I will highlight a few of the many interesting trends that emerged from the data. Generally, a few influential items were apparent in the dataset: the dominance of a limited set of artists, the prevalence of pop and hip-hop music, and the growing presence of collaborations.

Artists with the most songs in the Top 100 playlist from 2000-2021

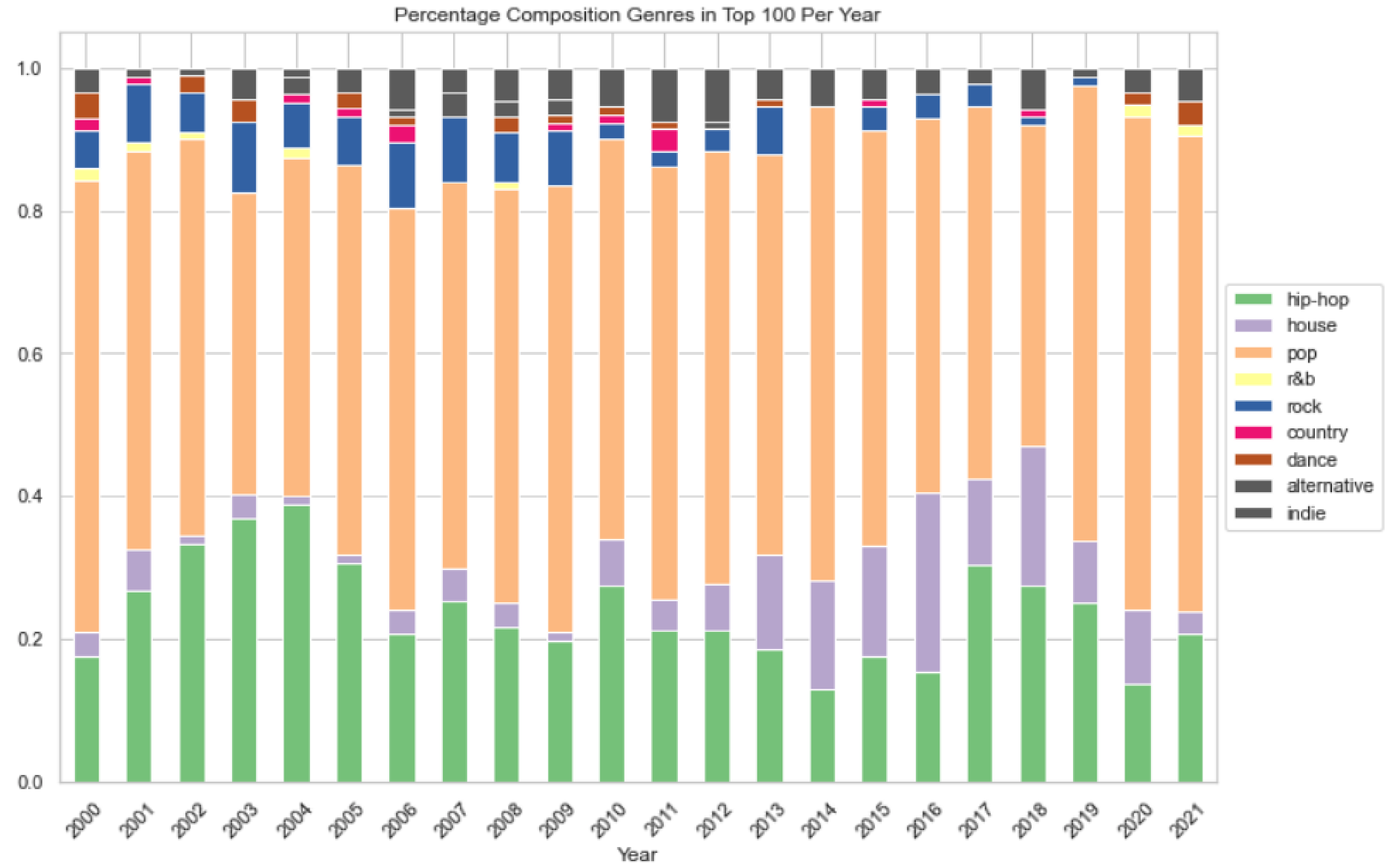


Distribution of genres in the Top 100 playlists



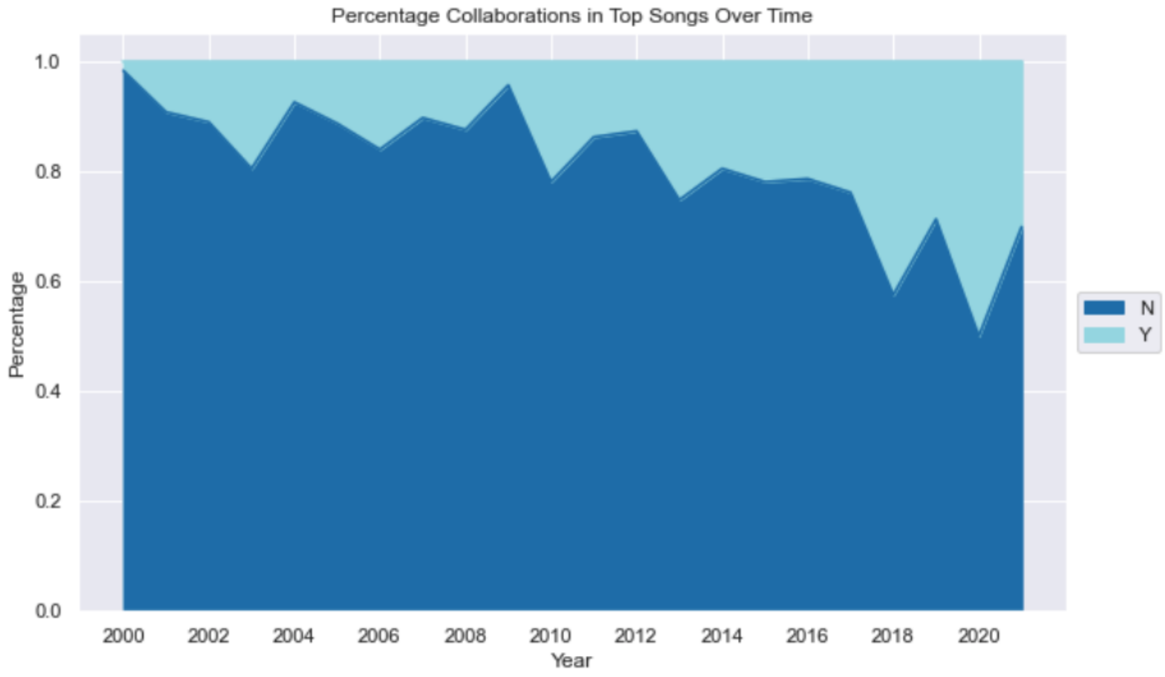
We can see that pop and hip-hop take the lion's share of Top 100 spots

Trend in genre popularity over time

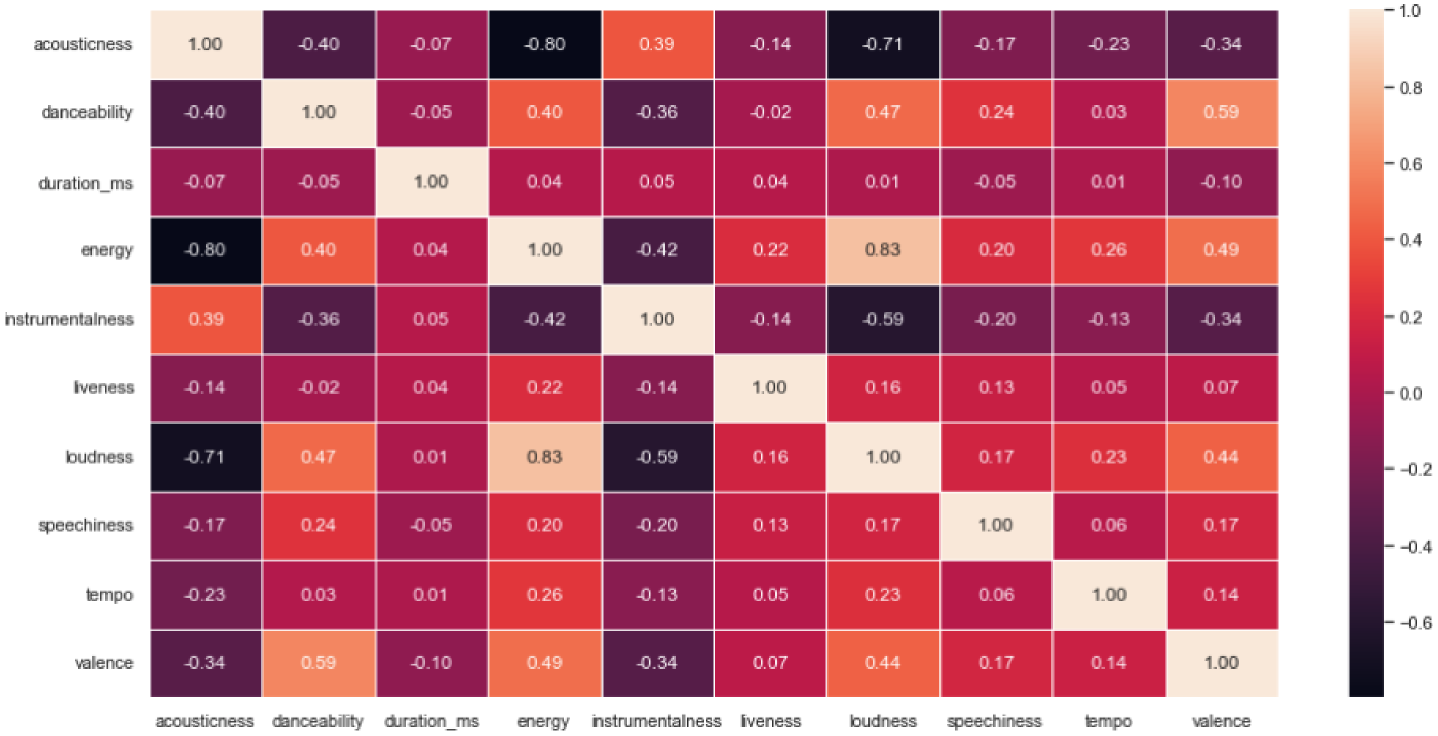


- Pop and hip-hop have always maintained a sizeable portion of the Top Songs
- House music really began to hold a meaningful position starting in 2010
- Rock and Country music has faded in popularity in recent years

Trend in collaborations over time



Correlation heatmap for audio features



Algorithms and Modeling

| Algorithms and Modeling Notebook |

| Pre-processing and Training Data Development |

| Metrics Table |

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

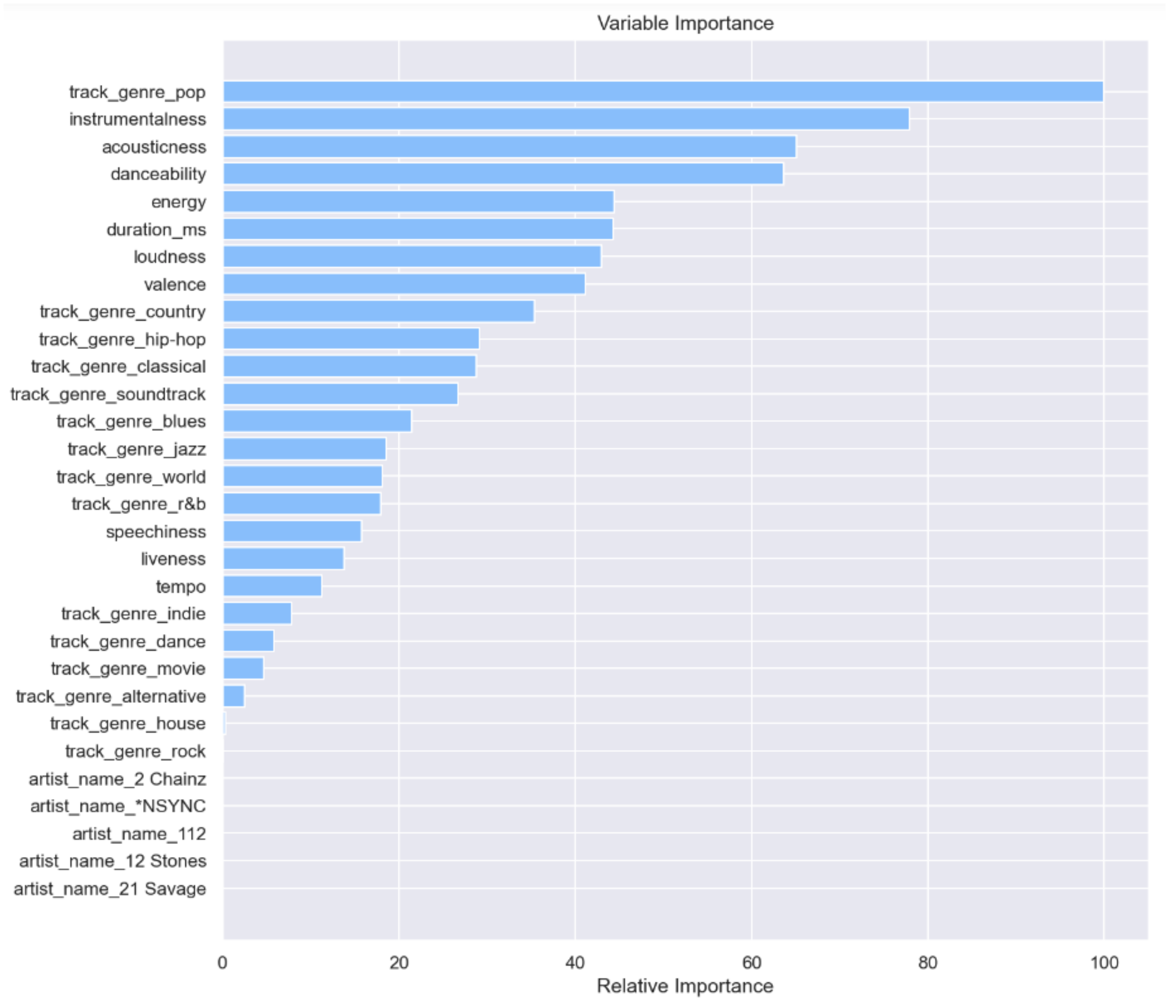
Based on the above ranking, it would follow that Precision would be our most important metric.

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
Naive Bayes	0.94	0.78	0.93	0.73	0.93	0.71	0.93	0.72	Overfitting

The table above shows a comparison of metrics across the different models, as well as comments on suitability and overfitting. **Based on the results, Logistic Regression, Random Forest, and KNN are all models with strong metrics, and not showing evidence of overfitting.**

Feature Importance from Random Forest Model



Future Improvements, Opportunities for Pursuit

- 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.