Uncover Music Recommendation System for Spotify

A Machine Learning Approach to Song Recommendation, Clustering, and Popularity Prediction in the Streaming Era

Presenters: Team 6
Andrew Rafael James
Doris Liang
Hunter Guo
Monica Ko
Pang Leesuravanich

ADSP 31017 IP09 Machine Learning I

Agenda

Problem Statement	02
Data Source	03
Song Clustering	08
Song Popularity Prediction	09
Song Recommendation System	10
Conclusion & Business Values	12

Problem Statement

Spotify®

Background

With the rise of music streaming, Spotify, with over 600M users, have reshaped how people discover music, making it crucial to understand what makes a song popular

Challenge

For over 100M songs on Spotify, it's quite difficult to analyze which audio characteristics such as tempo, energy, and danceability influence popularity, and whether trends have shifted across eras

Goal

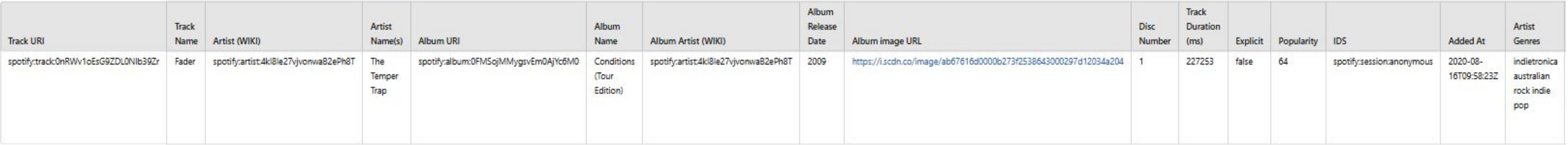
Develop a personalized song recommendation system based on the user's favorite song features, by completing two milestones: (1) categorize songs with common characteristics (2) predict if a song will be popular

Data Source

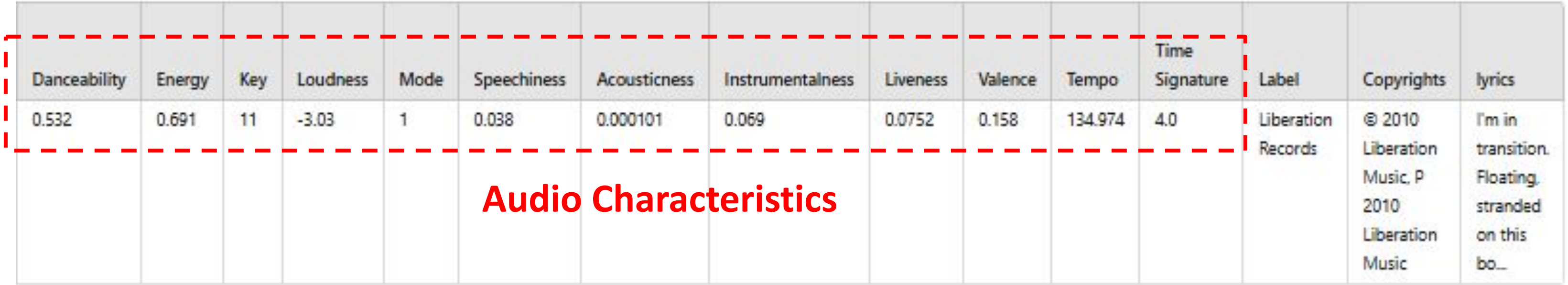
Top 10000 Songs on Spotify 1950-Now

The best and biggest songs from ARIA & Billboard charts spanning 7 decades.

https://www.kaggle.com/datasets/joebeachcapital/top-10000-spotify-songs-1960-now/



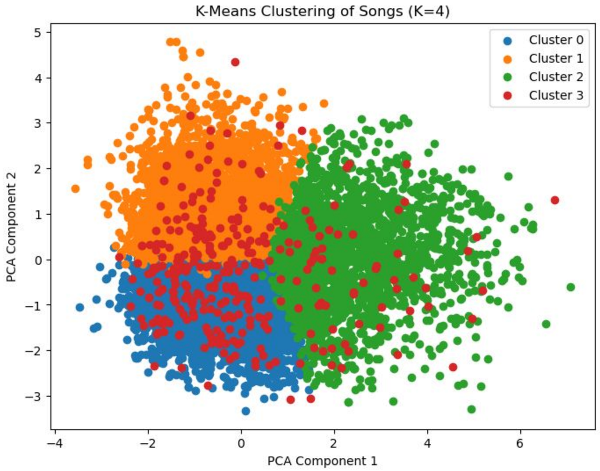
Track URI	Track Name	Artist (WIKI)	Artist Name(s)	Album URI	Album Name	Album Artist (WIKI)	Album Release Date	Album image URL	Disc Number	Track Duration (ms)	Explicit	Popularity	IDS	Added At	Artist Genres
spotify:track:0nRWv1oEsG9ZDL0NIb39Zr	Fader	spotify:artist:4kI8le27vjvonwaB2ePh8T	The Temper Trap	spotify:album:0FMSojMMygsvEm0AjYc6M0	Conditions (Tour Edition)	spotify:artist:4kI8le27vjvonwaB2ePh8T	2009	https://i.scdn.co/image/ab67616d0000b273f2538643000297d12034a204	1	227253	false	64	spotify:session:anonymous	2020-08-16T09:58:23Z	indietronica australian rock indie pop

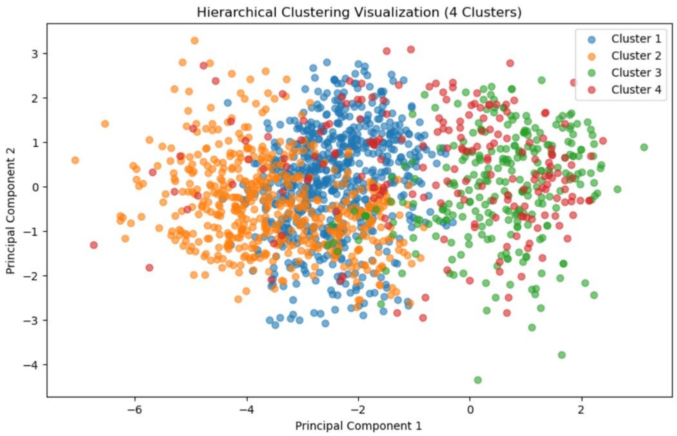


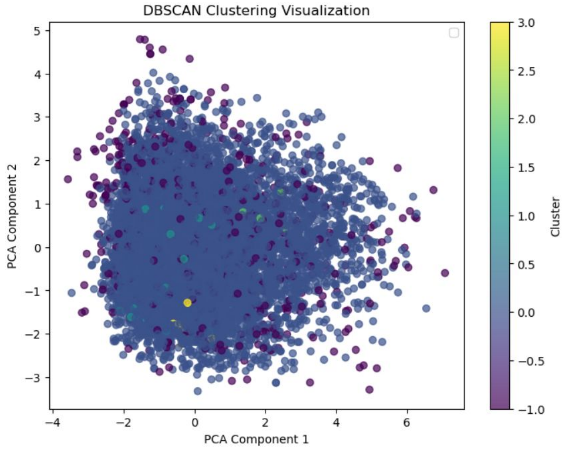
Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Time Signature	Label	Copyrights	lyrics
0.532	0.691	11	-3.03	1	0.038	0.000101	0.069	0.0752	0.158	134.974	4.0	Liberation Records	© 2010 Liberation Music, P 2010 Liberation Music	I'm in transition. Floating, stranded on this bo...

Audio Characteristics

Song Clustering







Danceable & Vocal: 4251
Energetic & Electric: 3256
Acoustic & Quiet: 2169
Instrumental & Pure Music: 317

Danceable & Vocal: 7598
Energetic & Happy: 759
Acoustic & Quiet:1172
Instrumental & Electric: 464

Danceable & Vocal: 9150
Energetic & Instrumental: 85
Electric & Quiet: 9
Danceable & Happy: 18
Noise & Outliers: 731

K-Means Outperforms Than Other Two Based On Balanced Sizes, Well-Separated, Dense Clusters

Song Popularity Prediction

Spotify offers plenty of songs for users to enjoy. From a business perspective, it is crucial for Spotify to predict whether a song will become popular. We classify songs as popular or not based on audio features and external artist information with supervised models.

Popularity Threshold: We classified songs as popular (popularity score ≥ 50) or not popular (< 50)

Data Pre-processing

- Drop Duplication

- Extract Release Year

- Binary Popularity Scores

- Encoded Categorical Variables

e.g. track characteristics

Feature Engineering

- StandardScaler

- SMOTE

- Created additional binary indicators

e.g. Famous Artist (top 20), Genre Encoding(12), Track Clustering

Supervised Models

- Logistic Regression

- Random Forest

- Gradient Boosting

- XGBoost

- LightGBM

- SVM

Model Evaluation

- Accuracy

- Precision

- Recall

- F1 Score

- ROC-AUC

Song Popularity Prediction - Model Evaluation

Spotify®

Model	Accuracy	F1 Score	Precision	Recall	ROC-AUC
LightGBM	0.58	**0.56**	0.43	0.82	**0.67**
Gradient Boosting	0.53	0.55	0.41	**0.88**	0.66
XGBoost	0.56	0.55	0.42	0.82	0.66
SVM	0.57	0.54	0.42	0.76	0.64
Logistic Regression	0.58	0.42	0.39	0.45	0.59
Random Forest	**0.67**	0.29	**0.50**	0.21	0.66

- **LightGBM, Gradient Boosting**, and **XGBoost** achieve high recall (~0.82–0.88), identifying popular songs but misclassifying some non-popular ones.

- **Random Forest** has the highest precision (0.50) but suffers from low recall (0.21), meaning it fails to capture a significant number of actual popular songs.

- **LightGBM, XGBoost, and Gradient Boosting** provide a better balance between precision and recall, making them more suitable overall.

LightGBM is the suitable choice, but low precision suggests the need for additional features like user behavior, demographic data, and real-time streaming trends.

Song Recommendation System - Methodology

Content-based filtering algorithm using cosine similarity based on the following three components:

Audio features, popularity score, and sentiment	lyrics	Artist information

Final Similarity Score = 0.1 * Artist Info Score + 0.25 * Lyrics Score + 0.65 Audio Score

Audio Similarity Score

Features:

1. Audio features including Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Key, Mode, and Time Signature.
2. Popularity score
3. Clustering Label
4. Sentiment of lyrics using **TextBlob** library

Data preprocessing:

1. Standardized numeric features using StandardScaler (mean = 0, std = 1)
2. Computed cosine similarity to measure song relationships
3. Normalized similarity scores to scale values between 0 and 1

Song Recommendation System - Methodology

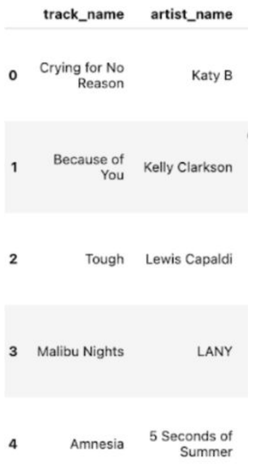
Lyrics Similarity Score

1. **Data Preprocessing**
   - Keywords extraction using **RAKE** library
   - Removed unnecessary characters, such as ? or /

2. Transform Lyrics data using 3 **word-embedding** approaches
   - TF-IDF
   - Word2Vec
   - **BERT**

3. Compute **cosine similarity** from the song vectors and chose **BERT** as it produces a **more balanced distribution of similarity score** and captures **semantic meaning of sentences** of lyrics

Songs with lyrics similarity of “**Sad**” by “**Maroon 5**” are about heartbreak, regret, and emotional pain



	track_name	artist_name
0	Crying for No Reason	Katy B
1	Because of You	Kelly Clarkson
2	Tough	Lewis Capaldi
3	Malibu Nights	LANY
4	Amnesia	5 Seconds of Summer

Artist Info Similarity Score

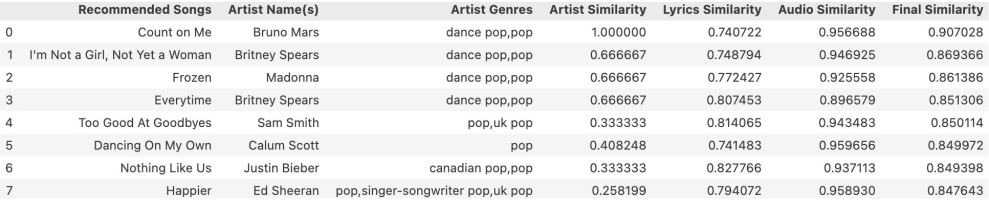
1. **Data Preprocessing**
   - Combine artist name and artist genres
   - Remove stopwords

2. Transform artist info text data using **CountVectorizer**

3. Compute cosine similarity

Song Recommendation System - Output

Input: “**When I Was Your Man** ” by “**Bruno Mars** ”

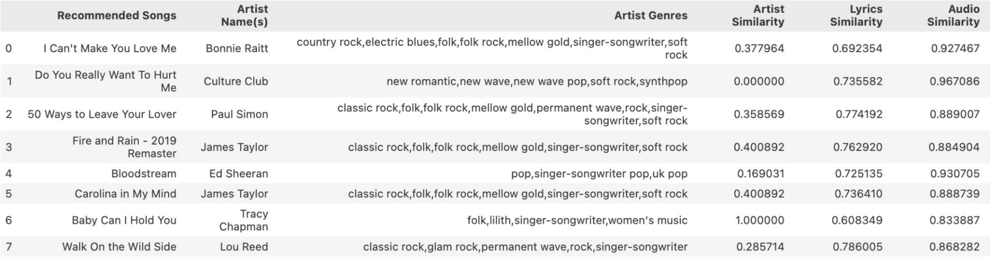


	Recommended Songs	Artist Name(s)	Artist Genres	Artist Similarity	Lyrics Similarity	Audio Similarity	Final Similarity
0	Count on Me	Bruno Mars	dance pop,pop	1.000000	0.740722	0.956688	0.907028
1	I'm Not a Girl, Not Yet a Woman	Britney Spears	dance pop,pop	0.666667	0.748794	0.946925	0.869366
2	Frozen	Madonna	dance pop,pop	0.666667	0.772427	0.925558	0.861386
3	Everytime	Britney Spears	dance pop,pop	0.666667	0.807453	0.896579	0.851306
4	Too Good At Goodbyes	Sam Smith	pop,uk pop	0.333333	0.814065	0.943483	0.850114
5	Dancing On My Own	Calum Scott	pop	0.408248	0.741483	0.959656	0.849972
6	Nothing Like Us	Justin Bieber	canadian pop,pop	0.333333	0.827766	0.937113	0.849398
7	Happier	Ed Sheeran	pop,singer-songwriter pop,uk pop	0.258199	0.794072	0.958930	0.847643

1. The first recommended song is **Count on Me** by Bruno Mars, driven by similarity of genre, **pop/soul style** with that of When I Was Your Man.

2. Additionally, both songs explore themes of romantic relationships

Input: “**Fast Car** ” by “**Tracy Chapman** ”



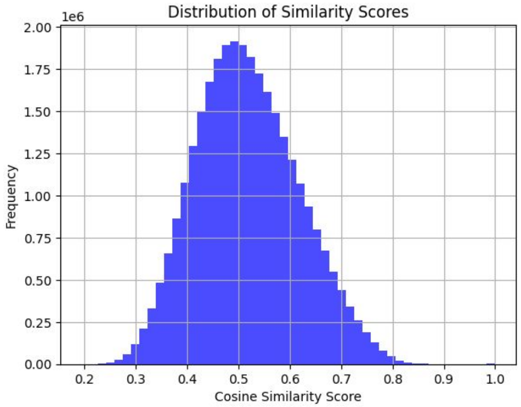
	Recommended Songs	Artist Name(s)	Artist Genres	Artist Similarity	Lyrics Similarity	Audio Similarity
0	I Can't Make You Love Me	Bonnie Raitt	country rock,electric blues,folk,folk rock,mellow gold,singer-songwriter,soft rock	0.377964	0.692354	0.927467
1	Do You Really Want To Hurt Me	Culture Club	new romantic,new wave,new wave pop,soft rock,synthpop	0.000000	0.735582	0.967086
2	50 Ways to Leave Your Lover	Paul Simon	classic rock,folk,folk rock,mellow gold,permanent wave,rock,singer-songwriter,soft rock	0.358569	0.774192	0.889007
3	Fire and Rain - 2019 Remaster	James Taylor	classic rock,folk,folk rock,mellow gold,singer-songwriter,soft rock	0.400892	0.762920	0.884904
4	Bloodstream	Ed Sheeran	pop,singer-songwriter pop,uk pop	0.169031	0.725135	0.930705
5	Carolina in My Mind	James Taylor	classic rock,folk,folk rock,mellow gold,singer-songwriter,soft rock	0.400892	0.736410	0.888739
6	Baby Can I Hold You	Tracy Chapman	folk,lilith,singer-songwriter,women's music	1.000000	0.608349	0.833887
7	Walk On the Wild Side	Lou Reed	classic rock,glam rock,permanent wave,rock,singer-songwriter	0.285714	0.786005	0.868282

1. The majority of other songs share similar artist genres, namely **rock and country**

2. Song recommendations were released in the **1980s and 1990s**, during which **Fast Car** was also released in **1982**

Song Recommendation System - Model Evaluation

Model-based evaluation



- Balanced similarity distribution centered around 0.5.
- No concentration near 0, reducing irrelevant recommendations.
- No over-concentration near 1, ensuring song differences are captured.

User-based evaluation

- **A/B testing** with real users can be conducted after deployment.
- Metrics like **Skip Rate** will assess recommendation effectiveness.

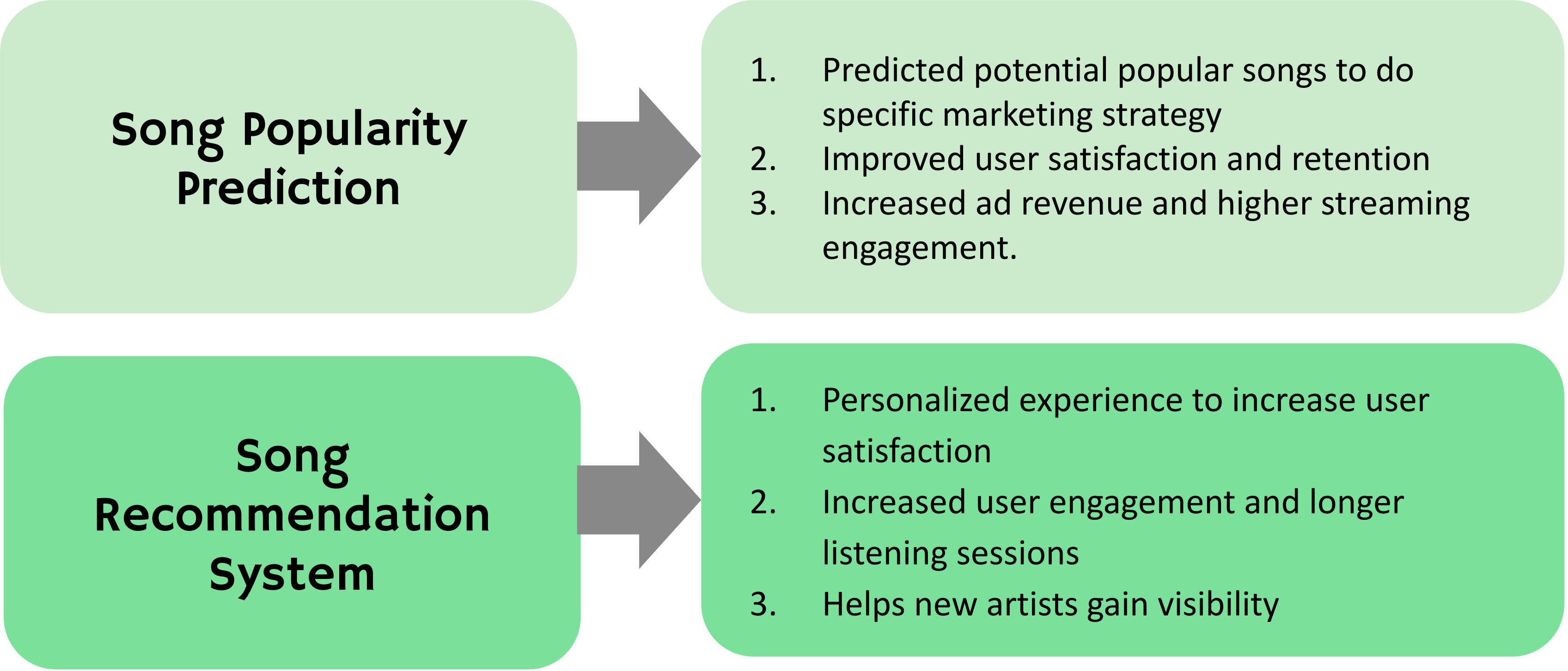
Conclusion

Our Clustering Puts 10,000 Songs into 4 Categories

Our Model Correctly Predicts 58% Popular Songs

Our System Offers Recommendation to 100M Users

Business Value



Song Popularity Prediction

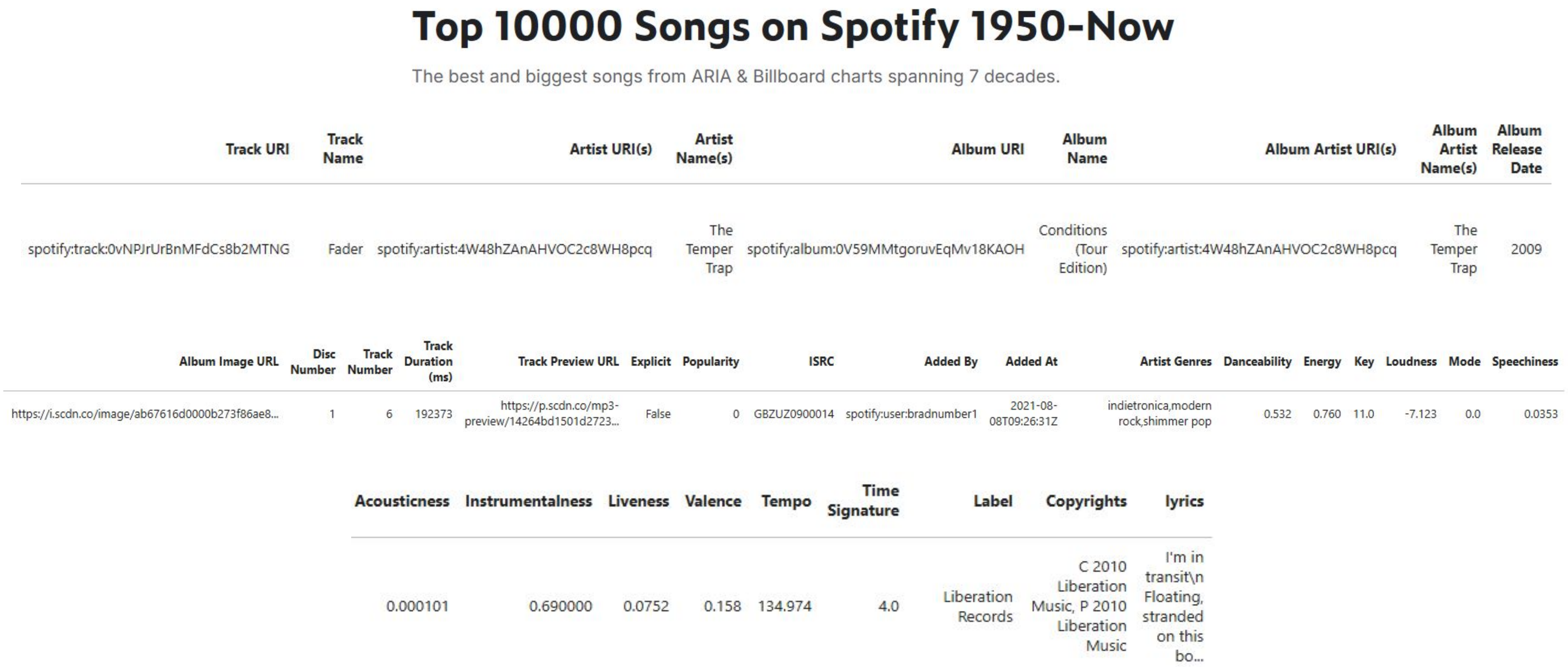
1. Predicted potential popular songs to do specific marketing strategy
2. Improved user satisfaction and retention
3. Increased ad revenue and higher streaming engagement.

Song Recommendation System

1. Personalized experience to increase user satisfaction
2. Increased user engagement and longer listening sessions
3. Helps new artists gain visibility

Thank you! Q&A

Data Source



Top 10000 Songs on Spotify 1950-Now

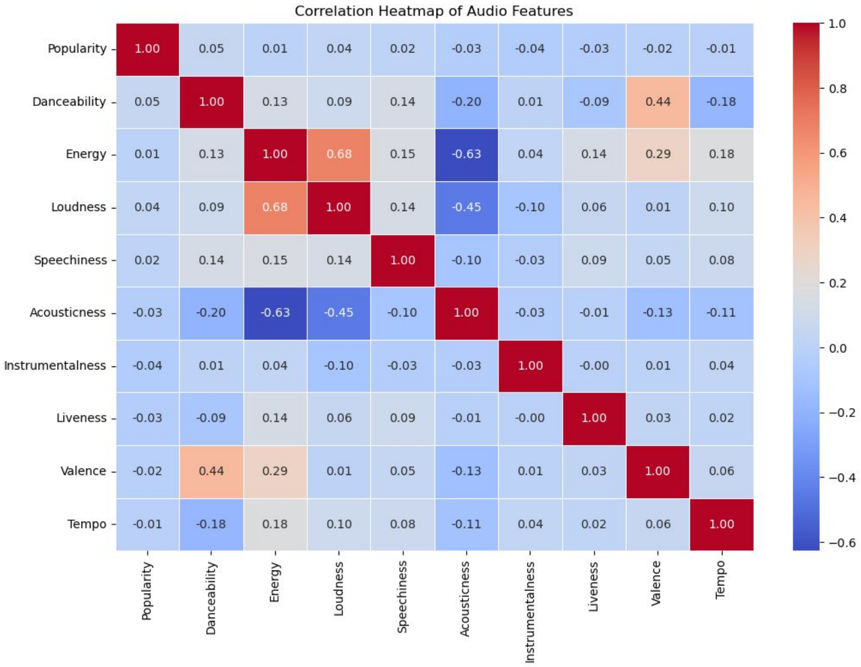
The best and biggest songs from ARIA & Billboard charts spanning 7 decades.

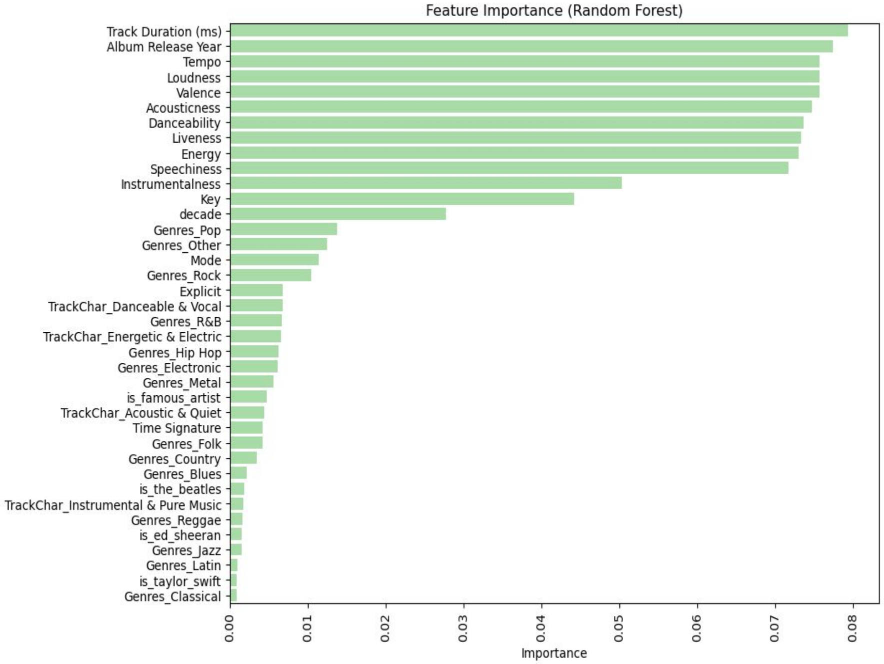
Track URI	Track Name	Artist URI(s)	Artist Name(s)	Album URI	Album Name	Album Artist URI(s)	Album Artist Name(s)	Album Release Date
spotify:track:0vNPJrUrBnMFdCs8b2MTNG	Fader	spotify:artist:4W48hZAnAHVOC2c8WH8pcq	The Temper Trap	spotify:album:0V59MMtgoruvEqMv18KAOH	Conditions (Tour Edition)	spotify:artist:4W48hZAnAHVOC2c8WH8pcq	The Temper Trap	2009

Album Image URL	Disc Number	Track Number	Track Duration (ms)	Track Preview URL	Explicit	Popularity	ISRC	Added By	Added At	Artist Genres	Danceability	Energy	Key	Loudness	Mode	Speechiness
https://i.scdn.co/image/ab67616d0000b273f86ae8...	1	6	192373	https://p.scdn.co/mp3-preview/14264bd1501d2723...	False	0	GBZUZ0900014	spotify:user:bradnumber1	2021-08-08T09:26:31Z	indietronica,modern rock,shimmer pop	0.532	0.760	11.0	-7.123	0.0	0.0353

Acousticness	Instrumentalness	Liveness	Valence	Tempo	Time Signature	Label	Copyrights	lyrics
0.000101	0.690000	0.0752	0.158	134.974	4.0	Liberation Records	C 2010 Liberation Music, P 2010 Liberation Music	I'm in transit\n Floating, stranded on this bo...

Song Popularity Prediction - Feature Importance


Correlation Heatmap of Audio Features
Popularity 1.00 0.05 0.01 0.04 0.02 -0.03 -0.04 -0.03 -0.02 -0.01
Danceability 0.05 1.00 0.13 0.09 0.14 -0.20 0.01 -0.09 0.44 -0.18
Energy 0.01 0.13 1.00 0.68 0.15 -0.63 0.04 0.14 0.29 0.18
Loudness 0.04 0.09 0.68 1.00 0.14 -0.45 -0.10 0.06 0.01 0.10
Speechiness 0.02 0.14 0.15 0.14 1.00 -0.10 -0.03 0.09 0.05 0.08
Acousticness -0.03 -0.20 -0.63 -0.45 -0.10 1.00 -0.03 -0.01 -0.13 -0.11
Instrumentalness -0.04 0.01 0.04 -0.10 -0.03 -0.03 1.00 -0.00 0.01 0.04
Liveness -0.03 -0.09 0.14 0.06 0.09 -0.01 -0.00 1.00 0.03 0.02
Valence -0.02 0.44 0.29 0.01 0.05 -0.13 0.01 0.03 1.00 0.06
Tempo -0.01 -0.18 0.18 0.10 0.08 -0.11 0.04 0.02 0.06 1.00
Popularity
Danceability
Energy
Loudness
Speechiness
Acousticness
Instrumentalness
Liveness
Valence
Tempo


Feature Importance (Random Forest)
Features
Track Duration (ms)
Album Release Year
Tempo
Loudness
Valence
Acousticness
Danceability
Liveness
Energy
Speechiness
Instrumentalness
Key
decade
Genres_Pop
Genres_Other
Mode
Genres_Rock
Explicit
TrackChar_Danceable & Vocal
Genres_R&B
TrackChar_Energetic & Electric
Genres_Hip Hop
Genres_Electronic
Genres_Metal
is_famous_artist
TrackChar_Acoustic & Quiet
Time Signature
Genres_Folk
Genres_Country
Genres_Blues
is_the_beatles
TrackChar_Instrumental & Pure Music
Genres_Reggae
is_ed_sheeran
Genres_Jazz
Genres_Latin
is_taylor_swift
Genres_Classical
Importance