Phase I – Problem Identification
Spotify Track Popularity & Feature Analysis

Shreevishnu 22BCE2213
Karen Joseph Williams 22BCE2797

Domain:Entertainment Industry
Dataset: Spotify Tracks Attributes and Popularity (Kaggle)
LINK: https://www.kaggle.com/datasets/melissamonfared/spotify-tracks-attributes-and-popularity
Submission Date: 11-Aug-2025

1. Introduction

In today's streaming-first era, platforms like Spotify rely on data-driven systems to recommend music, surface new artists, and keep users engaged. Yet predicting a song's popularity and modeling listener engagement remain challenging because musical tastes are highly diverse, trends shift quickly, and contextual factors (e.g., mood, activity, seasonality) play a significant role.

This project studies Spotify track attributes—such as danceability, energy, tempo, loudness, acousticness, valence, and release year—to understand what characteristics influence popularity and how these insights can support better recommendations. The findings can help artists and labels craft production and marketing strategies, while streaming platforms can improve personalization and cold-start handling for new tracks. Listeners ultimately benefit from higher quality, more relevant suggestions.

Relevance to society & industry:
• Artists & Producers: Identify audio attributes linked to audience appeal and optimize song design.
• Streaming Platforms: Enhance recommendation accuracy and user retention via improved popularity modeling.
• Music Industry: Inform A&R, playlist curation, and promotional strategies with evidence-based insights.

2. Study Section – Comparative Analysis

We review typical approaches used for song popularity prediction and recommendation. The table below summarizes strengths and limitations observed in prior work and common practice.

Aspect	Strengths of Existing Models	Limitations of Existing Models
Feature Usage	Audio descriptors (danceability, tempo, energy, loudness, valence) are predictive and easy to compute.	Often omit contextual/cultural signals (region, seasonality, virality) that affect popularity.
Algorithms	Tree-based models (Random Forest/XGBoost) and linear models provide interpretability and solid baselines.	Limited in capturing non-linear temporal/genre interactions; generalization to niche/novel trends can drop.
Deep Learning	CNN/RNN/Transformers on spectrograms capture complex timbral and temporal patterns.	Data- and compute-intensive; risk of overfitting and reduced interpretability on smaller curated datasets.
Data Availability	Public sources (Spotify API, Kaggle datasets) provide scalable access to track-level features.	Bias toward mainstream genres/artists; metadata gaps; lack of user demographics and listening context.
Recommenders	Collaborative filtering personalizes suggestions using user-item interactions.	Cold-start for new songs/artists; struggles when history is sparse or rapidly evolving.

Novel observations and data trends we aim to validate on the Kaggle dataset:

• Danceability and energy often show positive association with popularity; tempo may have a non-linear effect.
• Loudness and valence can interact with genre—e.g., higher loudness in pop/EDM vs. lower in acoustic/indie.
• Recency bias: newer releases tend to dominate popularity scores; re-releases/viral revivals are exceptions.
• Genre overlap and collaborations can lift visibility and playlist inclusion rates.

Deep learning techniques applied in this domain (for future phases):

• CNNs on Mel-spectrograms for timbre/harmonic textures.
• Bi-LSTM/Transformers for sequence modeling of frames/beats.
• Hybrid models combining content features with metadata and simple interaction histories.
• Contrastive learning to align audio embeddings with metadata/playlist contexts.

3. Objectives & Justification

Each student defines two objectives. Justifications reference the above pros/cons to motivate improvements.

Student: Shreevishnu

Objective 1: Build a supervised ML model to predict track popularity (0–100) using audio attributes and metadata (e.g., release year, artist).

Justification: Traditional baselines underuse combined content + metadata. A tuned model (e.g., Gradient Boosting/XGBoost) can better capture non-linear interactions while remaining interpretable for stakeholders.

Objective 2: Quantify the marginal influence of key features (danceability, energy, tempo, valence, loudness) via feature importance and partial-dependence/SHAP analyses.

Justification: Interpretable diagnostics address industry needs—artists and curators gain clear levers to optimize production and playlisting decisions.

Student: Karen

Objective 1: Prototype a content-aware recommendation approach that blends audio features with simple nearest-neighbor/cosine similarity to surface similar tracks for cold-start cases.

Justification: Complements collaborative filtering by avoiding reliance on user history, mitigating cold-start for new songs and emerging artists.

Objective 2: Perform a trend analysis across release years and genres to identify evolving patterns in popular tracks (e.g., shifts in energy/tempo/danceability).

Justification: Addresses data bias and recency effects, informing A&R and marketing on where audience taste is moving.

4. Standard Dataset

Dataset: Spotify Tracks Attributes and Popularity (Kaggle) by Melissa Monfared.
Contains: audio features (danceability, energy, loudness, tempo, acousticness, valence), popularity score, and metadata (track/artist, release year, genre). Suitable for supervised prediction and exploratory trend analysis.

Prepared for Phase I submission (Problem Identification).

APA References

Datasets:

1. Monfared, M. (2024). *Spotify Tracks Attributes and Popularity*. Kaggle. Retrieved from https://www.kaggle.com/datasets/melissamonfared/spotify-tracks-attributes-and-popularity

2. Bhuyan, R. P. (2025). *Spotify Dataset for ML Practice*. Kaggle. Retrieved from https://www.kaggle.com/datasets/rudraprasadbhuyan/spotify-dataset-for-ml-practice

Academic Papers & Research:

3. Middlebrook, K., & Sheik, K. (2019). Song hit prediction: Predicting Billboard hits using Spotify data. *arXiv*. Retrieved from arXiv repository arXiv

4. Adeagbo, A. (2020). Predicting Afrobeats hit songs using Spotify data. *arXiv*. Retrieved from arXiv repository arXiv

5. Clifton, A., Pappu, A., Reddy, S., Yu, Y., Karlgren, J., Carterette, B., & Jones, R. (2020). The Spotify podcast dataset. *arXiv*. Retrieved from arXiv repository arXiv

Surveys & Reviews:

6. Ferraro, A., Tkalčič, M., & Schedl, M. (2020). Recommender systems for music: A survey. *ACM Computing Surveys (CSUR)*, 53(5), 1–36. https://doi.org/10.1145/3407198

Deep Learning & Music Analysis:

7. Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems* (NeurIPS), 26, 2643–2651.

8. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2392–2396. https://doi.org/10.1109/ICASSP.2017.7952585

Exploratory Data Analysis & Projects:

9. Sharma, B. (2024, November 1). Building a machine learning model using the Spotify Song Attributes dataset. *Medium*. Retrieved from https://medium.com Medium

10. Luo, K. (2018). *Machine learning approach for genre prediction on Spotify top ranking songs* (Master's thesis). University of North Carolina at Chapel Hill. Retrieved from UNC repositor