Predicting Song Popularity & Discovering Musical Structure

CIS 5450 - Final Project

Amogh Channashetti, Binoy Patel, Yohan Vergis Vinu

Penn Engineering

Objective & Value Proposition

The Problem

- Music streaming platforms like Spotify host millions of tracks
- Record labels, artists, and playlist curators need to identify potential "hits" early
- Traditional genre labels don't capture the full spectrum of musical similarity

Our Objectives

- Build a predictive model to classify songs as "hits" (popularity ≥ 70) vs. non-hits
- Discover latent musical structures using unsupervised learning

Value Proposition

Stakeholder	Benefit
Record Labels	Early hit identification for marketing investment
Playlist Curators	Data-driven song selection beyond genre labels
Artists/Producers	Understand audio characteristics of successful music
Streaming Platforms	Better recommendation algorithms

Dataset Overview

Source: Spotify Tracks Dataset from Kaggle (via Spotify Web API)

Dataset Specifications:
- Original Size: 114,000 tracks × 21 columns
- After Cleaning: 89,238 tracks × 25 columns

Features Used:
- Audio Features: danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo
- Metadata: artists, album_name, track_name, duration_ms, explicit
- Target Variable: `is_hit` (1 if popularity ≥ 70, else 0)

Data Cleaning Highlights:
- Removed ~25,000 duplicate tracks
- Filtered duration outliers (>15 min = likely podcasts)
- Fixed invalid tempo values (<20 BPM → median imputation)
- Consolidated 100+ genres into top 20 + "Other"

Some Engineered Features:
- num_artists: Count of collaborators per track
- primary_artist: First listed artist
- is_hit: Binary classification target

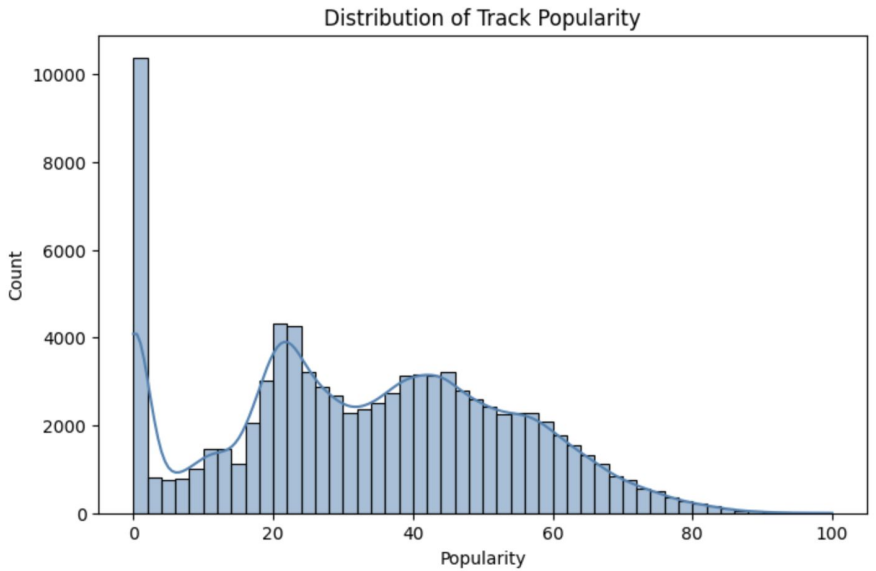
EDA Chart 1 - Popularity Distribution

Key Insights:

- Right-skewed distribution - most tracks have low popularity
- Only ~3.5% of tracks are "hits" (popularity ≥ 70)
- This creates a severe class imbalance (28:1 ratio)

Why This Matters:

- Standard classifiers would just predict "non-hit" for everything
- Need specialized techniques: class weighting, SMOTE, or undersampling
- Evaluation metrics must go beyond accuracy (use Precision, Recall, PR-AUC)



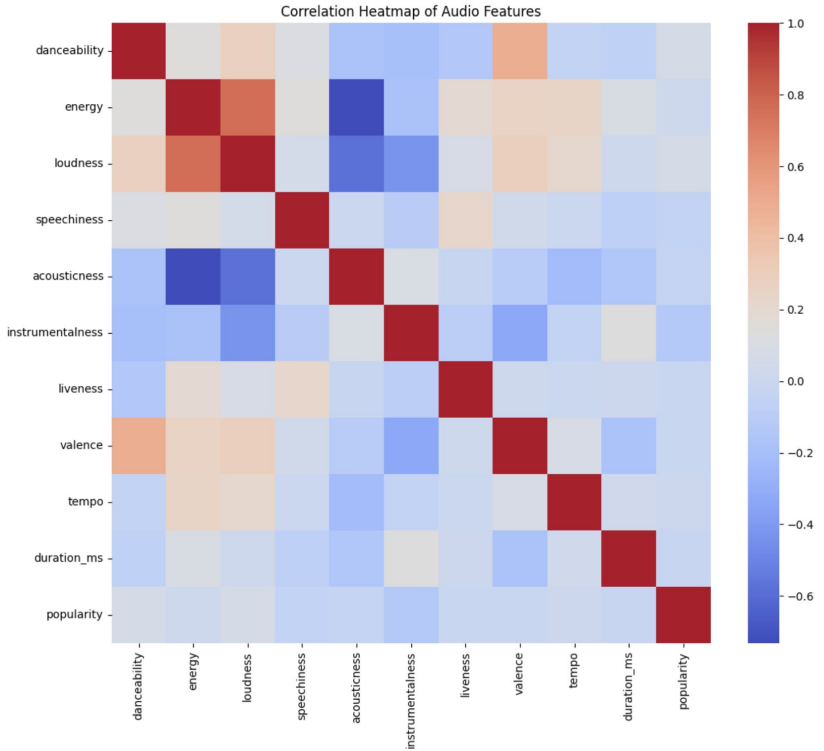
EDA Chart 2 - Correlation Heatmap

Key Insights:

- Strong positive correlation:
  Energy ↔ Loudness (r ≈ 0.75)
- Strong negative correlation:
  Acousticness ↔ Energy/Loudness
- Weak correlation with popularity:
  No single feature strongly predicts hits
- Danceability shows modest positive association

Implications for Modeling:

- No "silver bullet" feature for hit prediction
- Multi-feature models are necessary
- External factors (marketing, artist fame) likely drive popularity more than audio characteristics



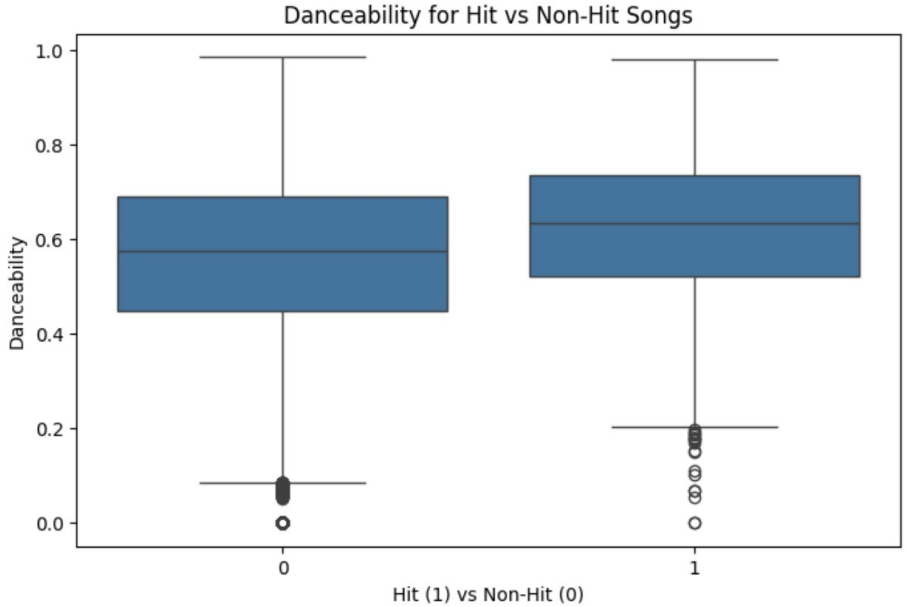
EDA Chart 3 - Danceability: Hits vs Non-Hits

Key Insights:

- Hits have slightly higher median danceability than non-hits
- However, distributions overlap significantly - danceability alone can't separate hits
- Wide variance in both groups - some hits have low danceability, some non-hits have high danceability

Implications:

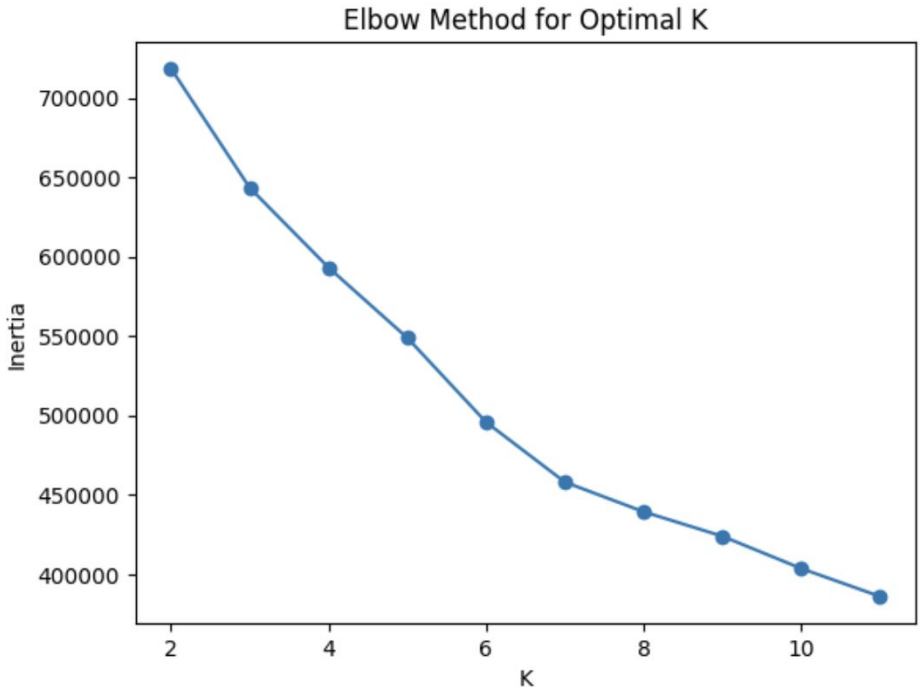
- Danceability is a weak signal for hit prediction
- Confirms what correlation heatmap showed - no single feature strongly predicts success
- Multi-feature approach is necessary



Elbow Method for Clustering

Key Insights:

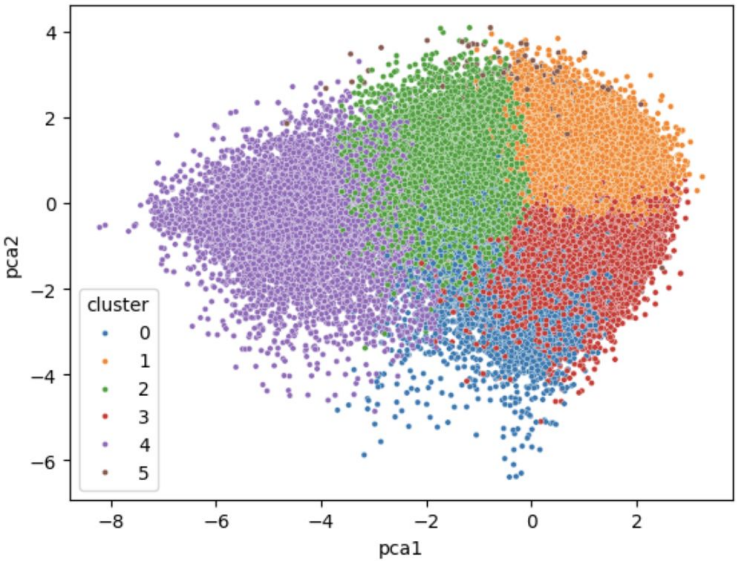
- Clear "elbow" at k = 6 clusters
- Diminishing returns after 6 clusters
- 6 is interpretable and meaningful for music industry application



6 Discovered Music Archetypes

Key Insight:
These clusters go above traditional genre labels - a 'pop' song and a 'rock' song might cluster together if they share similar audio DNA.

Cluster	Name	Key Characteristics
0	Instrumental Electronica	High instrumentalness, moderate energy, long duration
1	Dance/Pop Mainstream	High danceability, high valence (happiness), short songs
2	Acoustic Ballads	High accousticness, low energy, lowe loudness
3	Rock/Metal Intensity	High energy, low danceability, higher liveliness
4	Ambient/Classical	Very high instrumentalness, very low energy
5	Spoken Word/Live	Very high speechiness, very high liveness



Modeling Approach & Results

Key Findings

- Best Model: Gradient Boosting (ROC-AUC = 0.75)
- Precision-Recall Trade-off: High recall (75%) but low precision (6-7%)
- Class balancing techniques** (SMOTE, undersampling) didn't significantly improve performance
- Hyperparameter tuning provided marginal gains

Interpretation:

- Models can identify ~75% of actual hits, but generate many false positives
- For every true hit predicted, ~14 non-hits are incorrectly flagged
- Audio features alone have limited predictive power

Model	ROC-AUC	PR-AUC	Precision (Hit)	Recall (Hit)
Logistic Regression (baseline)	0.724	0.082	6.0%	75.4%
Random Forest	0.722	0.084	9.8%	2.6%
Gradient Boosting (baseline)	0.750	0.091	6.6%	75.2%
Gradient Boosting (tuned)	0.751	0.091	6.6%	74.0%
LR + SMOTE	0.664	0.066	5.3%	61.5%
LR + Undersampling	0.724	0.082	5.9%	76.4%

Implications & Actionable Insights

For Supervised Learning (Hit Prediction):

- Audio features alone are insufficient - Popularity driven by external factors (marketing, artist fame, playlist placement, viral trends)
- High recall achievable - Models useful for screening, narrow down candidate pool, then apply human judgment
- Low precision is acceptable for initial filter - Better to flag too many than miss potential hits
- Use the model as a first-pass filter to identify potential hits from thousands of new releases, then have A&R teams do deeper evaluation on the shortlist

For Unsupervised Learning (Clustering):

- 6 distinct musical archetypes discovered - playlist curation beyond genre labels
- Clusters transcend traditional genres - new way to organize music libraries
- Audio features capture meaningful similarity - Foundation for recommendation systems
- Create playlists based on audio profiles rather than genre tags - 'High Energy Workout' could pull from rock, EDM, and hip-hop clusters.

Challenges, Limitations & Future Work

Challenges Faced

- Severe Class Imbalance (28:1) - Required specialized handling
- Weak Feature-Target Correlations - Audio features don't strongly predict popularity
- No External Data - Missing crucial factors like marketing spend, playlist placement, social media buzz

Current Limitations

- Audio features only - Can't capture artist fame, marketing, virality
- Cross-sectional data - Can't track trends over time
- No release date - Can't account for recency bias in popularity scores
- Genre labels incomplete - "Other" category too large

Future Work

- Incorporate external data: Social media mentions, playlist appearances, artist follower counts
- Time-series analysis: Track popularity trajectories over time
- Deep learning: CNN/RNN on raw audio spectrograms
- A/B testing: Deploy model in production and measure actual hit rate
- Feature interactions: Explore non-linear relationships between audio features

Summary & Key Takeaways

What We Learned

1. Hit prediction from audio alone is limited - external factors dominate
2. Gradient Boosting outperforms simpler models - (ROC-AUC = 0.75)
3. K-Means successfully discovered 6 meaningful musical archetypes
4. Class imbalance is a critical challenge - requires careful handling
5. Clustering reveals structure invisible to genre labels

While we can't perfectly predict hits, we've built tools that help stakeholders make more informed decisions - from early hit screening to audio-based playlist curation.