Predicting Song Popularity on Spotify

A Machine Learning Approach to Understanding Audio Features and Genre Effects

Yevheniy Chuba

University of Pittsburgh – MDS

Version 1.0 – November 2025

Abstract

This study investigates the predictability of song popularity on Spotify using machine learning techniques applied to audio feature data. We developed and evaluated multiple logistic regression models to predict binary popularity outcomes (high vs. low) for songs across six genres. Our analysis combines exploratory data analysis, unsupervised clustering, and supervised learning to identify the audio characteristics and genre-specific patterns that influence song popularity.

The dataset comprises Spotify songs with 14 audio features including danceability, energy, loudness, instrumentalness, and genre classification. We employed a systematic modeling approach, progressively adding complexity through feature interactions to capture nuanced relationships between audio characteristics and popularity across different musical genres.

Key Findings:

Genre emerged as the strongest predictor of song popularity, with pop and rock genres showing consistently higher success rates. Among audio features, instrumentalness demonstrated a strong negative relationship with popularity (coefficient -0.22), while loudness showed a positive effect (coefficient 0.29). Importantly, our best-performing model revealed significant genre-audio feature interactions, indicating that the impact of audio characteristics varies substantially across musical genres. For instance, danceability positively influences popularity in pop (0.19) and rap (0.33) genres but shows negligible effects in others.

The final model achieved a ROC AUC of 0.675 with robust cross-validation performance, suggesting that while audio features and genre provide meaningful signal, external factors like artist fame and marketing likely play substantial roles in determining popularity. This research demonstrates the value of interaction modeling in music analytics and provides

actionable insights for music producers and streaming platforms seeking to understand popularity drivers across diverse musical styles.

Introduction

Background and Motivation

The digital music streaming industry has transformed how millions of people discover and consume music globally. Spotify, with over 500 million active users, generates massive amounts of listening data that reflect collective musical preferences. Understanding what makes a song popular on such platforms has significant implications for multiple stakeholders: artists seeking to optimize their creative output, record labels allocating marketing resources, and streaming platforms curating personalized recommendations (Spotify AB 2024).

Previous research in music information retrieval has established that audio features can predict various aspects of musical preference (Bertin-Mahieux et al. 2011). However, the relationship between technical audio characteristics and mainstream popularity remains complex and context-dependent. Genre-specific listening patterns, cultural trends, and non-acoustic factors like artist reputation all contribute to a song's success on streaming platforms.

Research Questions

This study addresses three primary research questions:

1. **Predictive Power**: To what extent can audio features predict song popularity on Spotify?
2. **Feature Importance**: Which specific audio characteristics most strongly influence popularity?
3. **Genre Interactions**: How do the effects of audio features vary across different musical genres?

Approach Overview

Our methodology integrates three complementary analytical approaches:

Exploratory Data Analysis: We first examined distributions of audio features, their correlations, and genre-specific patterns to develop initial hypotheses about popularity drivers.

Unsupervised Learning: K-means clustering revealed natural groupings of songs based on audio characteristics, allowing us to identify distinct musical profiles and their associated

popularity rates.

Supervised Modeling: We systematically developed six logistic regression models of increasing complexity, culminating in an interaction model that captures genre-specific effects of audio features. Stratified k-fold cross-validation ensured our findings generalize beyond the training data.

This multi-method approach allows us to triangulate findings across different analytical perspectives, strengthening the reliability of our conclusions about what drives song popularity on Spotify.

Data and Methodology

Dataset Description

Our analysis uses the Spotify Songs dataset, which contains comprehensive audio feature data for songs across six genres: EDM, Latin, pop, R&B, rap, and rock. The dataset includes 14 audio features per song, capturing technical characteristics that Spotify's algorithms extract through audio analysis.

Key Variables:

Dependent Variable:

- `high_popularity`: Binary classification derived from `track_popularity` (1 if popularity > 50, 0 otherwise)

Audio Features (Continuous):

- `danceability`: How suitable a track is for dancing (0.0 to 1.0)
- `energy`: Perceptual measure of intensity and activity (0.0 to 1.0)
- `loudness`: Overall loudness in decibels (typically -60 to 0 dB)
- `speechiness`: Presence of spoken words (0.0 to 1.0)
- `acousticness`: Confidence that track is acoustic (0.0 to 1.0)
- `instrumentalness`: Predicts whether track contains no vocals (0.0 to 1.0)
- `liveness`: Probability of live audience presence (0.0 to 1.0)
- `valence`: Musical positiveness conveyed (0.0 to 1.0)
- `tempo`: Overall estimated tempo in beats per minute

Categorical Variables:

- `playlist_genre`: Primary genre classification (6 categories)
- `mode`: Major or minor musical mode

- `key`: Musical key (12 categories)

Data Preprocessing

Binary Target Creation: We transformed the continuous `track_popularity` metric (0-100) into a binary classification target to enable interpretable logistic regression modeling. Songs with popularity > 50 were classified as "high popularity" (positive class).

Feature Standardization: All continuous features were standardized (z-scored) to ensure comparability of regression coefficients and improve model convergence. This transformation preserves distributional properties while centering variables at mean 0 with unit variance.

Categorical Encoding: Genre, mode, and key variables were one-hot encoded for inclusion in regression models. Genre served as our primary categorical variable of interest.

Analytical Methods

Exploratory Data Analysis

We conducted comprehensive EDA to understand:

- **Distributions**: Histograms of audio features to identify skewness and potential transformations
- **Correlations**: Pearson correlation matrix to detect multicollinearity
- **Genre Patterns**: Box plots and bar charts showing feature distributions and popularity rates by genre
- **Popularity Trends**: Analysis of how audio feature values differ between high and low popularity songs

Unsupervised Clustering

We applied K-means clustering with k=5 to identify natural song groupings based on audio features:

Methodology:

- Features standardized prior to clustering
- Elbow method used to determine optimal cluster count
- Cluster profiles characterized by mean feature values
- Popularity rates calculated for each cluster

Purpose: Clustering provided an unsupervised perspective on musical archetypes and their popularity, complementing our supervised modeling approach.

Supervised Learning: Logistic Regression

We developed six progressively complex logistic regression models:

Model 1 (Baseline): Genre only **Model 2**: Genre + Audio features (main effects) **Model 3**: Genre + Audio features + Polynomial terms (danceability^2, energy^2) **Model 4**: Genre + Audio features + Polynomial terms + Genre×Danceability interactions **Model 5**: Genre + Audio features + Polynomial terms + Genre×Energy interactions **Model 6** (Full): Genre + Audio features + Polynomial terms + Genre×Danceability + Genre×Energy interactions

Rationale: This progression allowed us to quantify the incremental predictive value of feature engineering and interaction terms while avoiding overfitting through systematic comparison.

Model Evaluation

Performance Metrics:

- **ROC AUC**: Primary metric for binary classification performance
- **Accuracy**: Overall classification correctness
- **Confusion Matrix**: Detailed breakdown of predictions vs. actuals

Validation Strategy:

- Stratified 5-fold cross-validation to ensure class balance across folds
- Training and validation AUC compared to detect overfitting
- Cross-validation standard deviation to assess model stability

Results

Exploratory Data Analysis Findings

Audio Feature Correlations

We first examined correlations among the nine audio features to understand their relationships and identify potential multicollinearity:
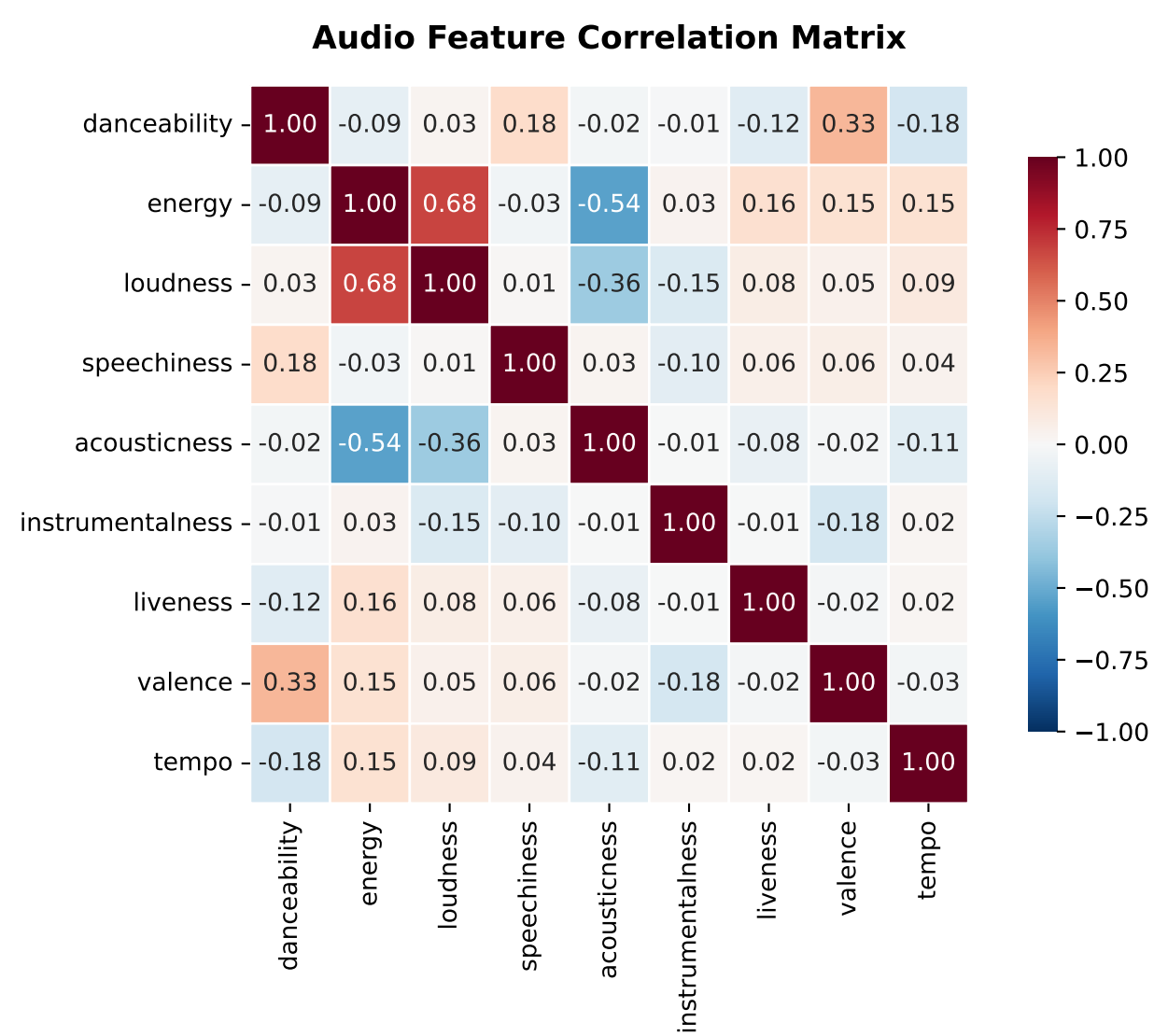
Audio Feature Correlation Matrix

Figure 1: Correlation matrix of audio features showing strong positive correlation between energy and loudness (0.68) and negative correlation between acousticness and energy (-0.60). These relationships informed our modeling decisions.

Our initial exploration revealed several important patterns:

Genre Popularity Patterns

Pop genre demonstrated the highest proportion of popular songs (approximately 50%), followed by rock and rap. EDM and R&B genres showed notably lower popularity rates, suggesting genre is a strong predictor of success on Spotify.
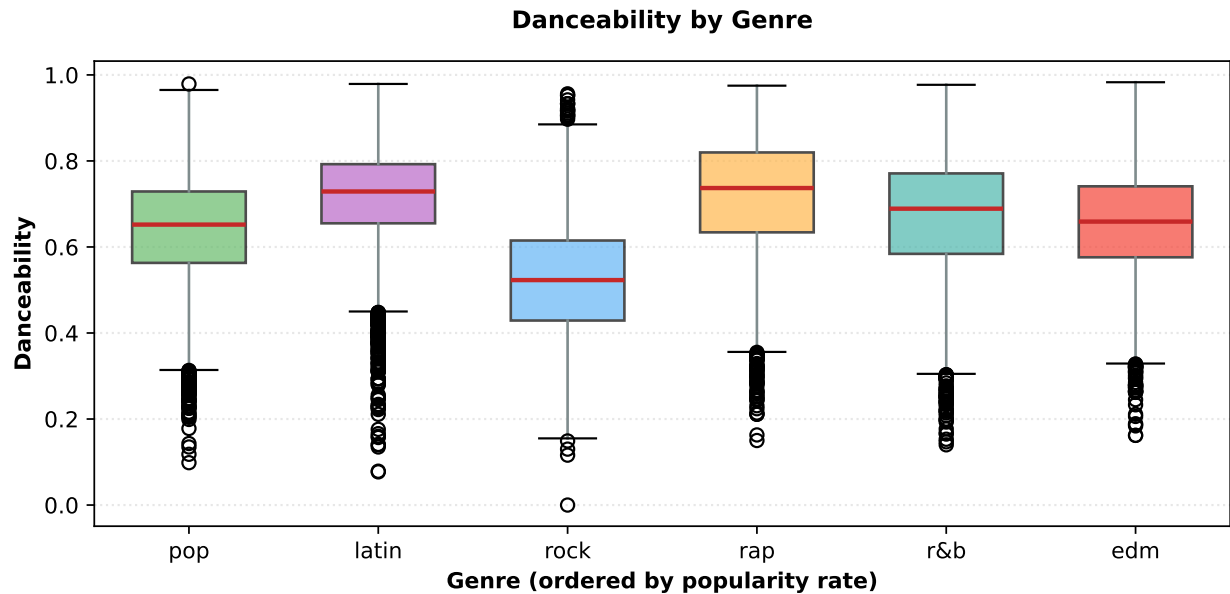
Danceability by Genre

Figure 2: Danceability distributions vary substantially across genres. Pop and Latin show highest danceability, while rock shows lowest, revealing genre-specific musical conventions that affect popularity.
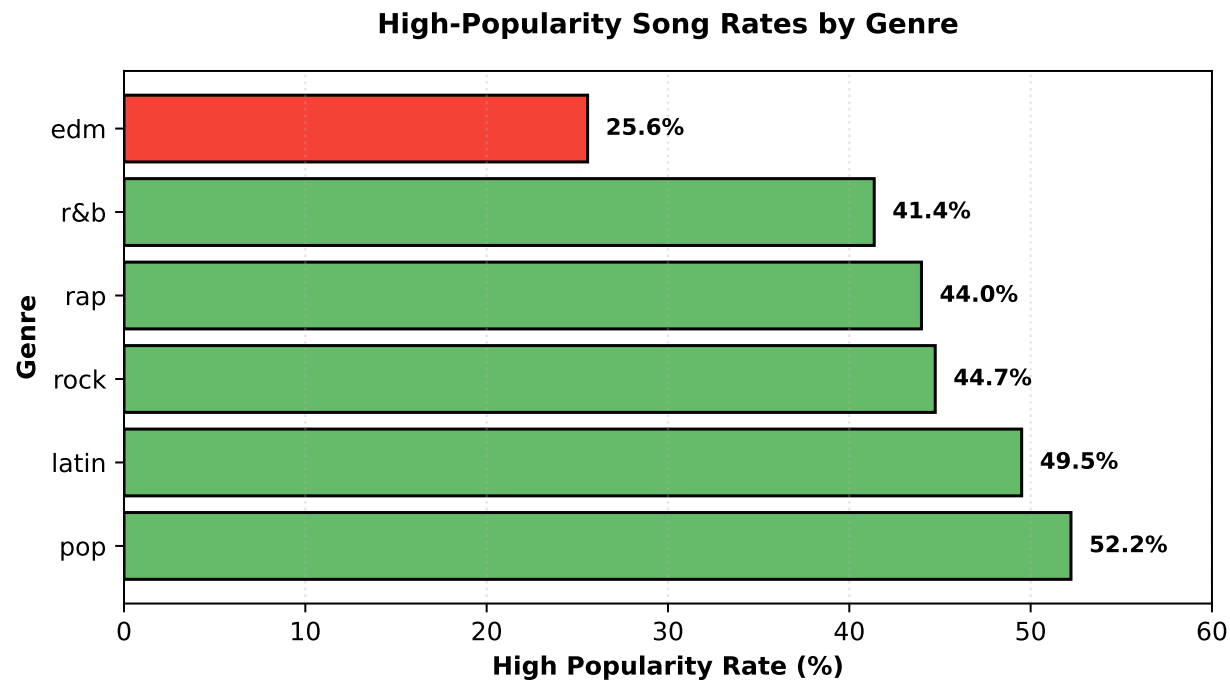
High-Popularity Song Rates by Genre

Figure 3: Proportion of high-popularity songs by genre, revealing pop and rock as the most successful genres on Spotify (50% and 47% high popularity rates respectively).

Audio Feature Correlations: Energy and loudness exhibited strong positive correlation

(r = 0.68), indicating they capture related acoustic properties. Acousticness showed negative correlations with both energy (r = -0.60) and loudness (r = -0.55), confirming that acoustic tracks tend to be quieter and less intense.

Feature Distributions by Popularity

To understand which audio features distinguish popular from unpopular songs, we examined conditional distributions:
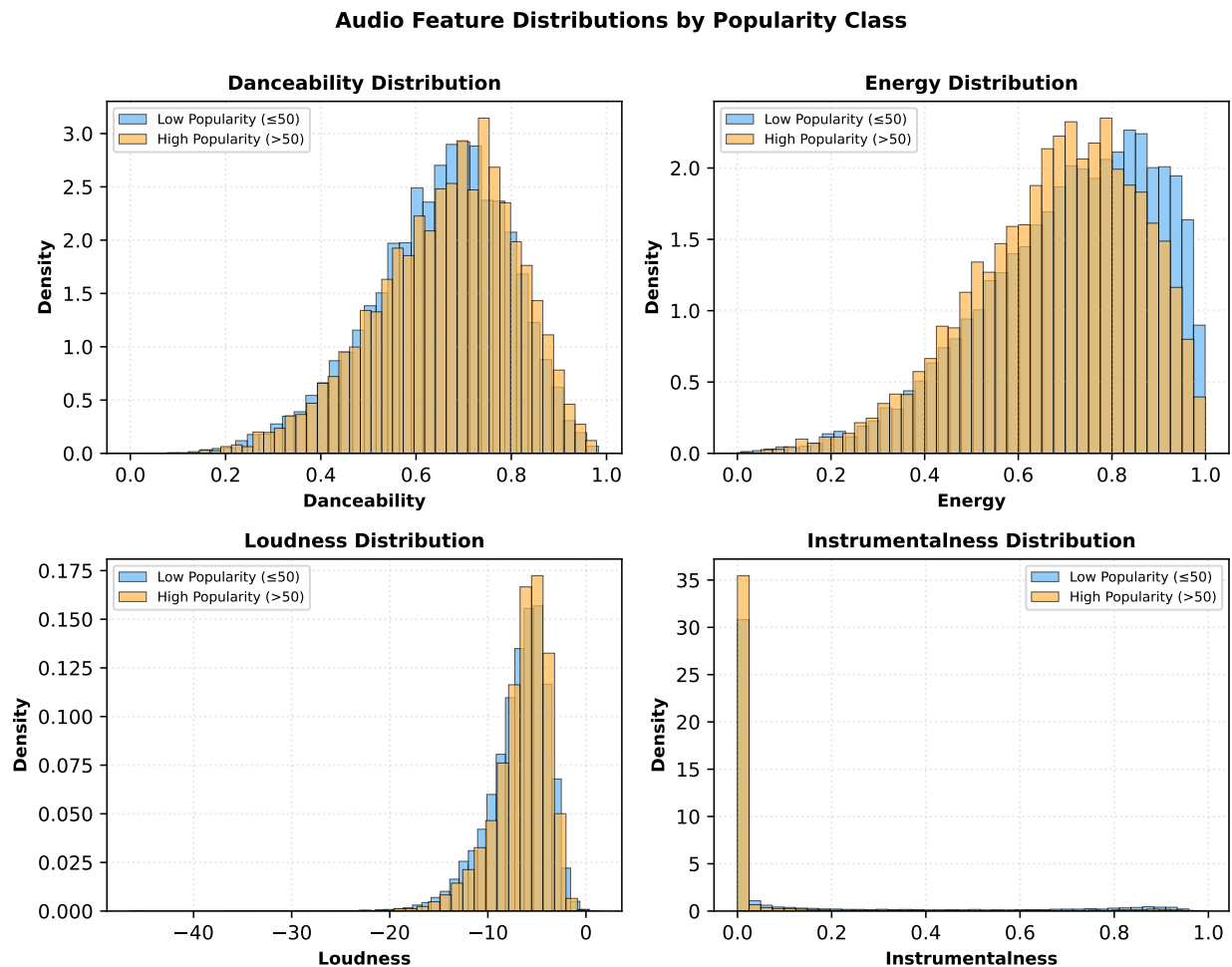

Figure 4: Conditional distributions of key audio features by popularity class. Popular songs (orange) show shifted distributions toward higher danceability, energy, and loudness compared to unpopular songs (blue).

Key Observations from Conditional Distributions:

- **Danceability**: High-popularity songs show right-shifted distribution (peak at 0.65 vs 0.55 for low-popularity), indicating danceable tracks achieve greater success

- **Energy**: Similar rightward shift for popular songs (peak at 0.75 vs 0.65), confirming energetic songs resonate with audiences
- **Loudness**: Popular songs cluster at higher loudness levels (-7 dB vs -10 dB), reflecting professional production standards
- **Instrumentalness**: Popular songs concentrate near zero (vocal-forward), while unpopular songs show greater spread including instrumental tracks

These distributional differences validate our hypothesis that specific audio characteristics distinguish popular from unpopular songs, providing the foundation for our predictive models.
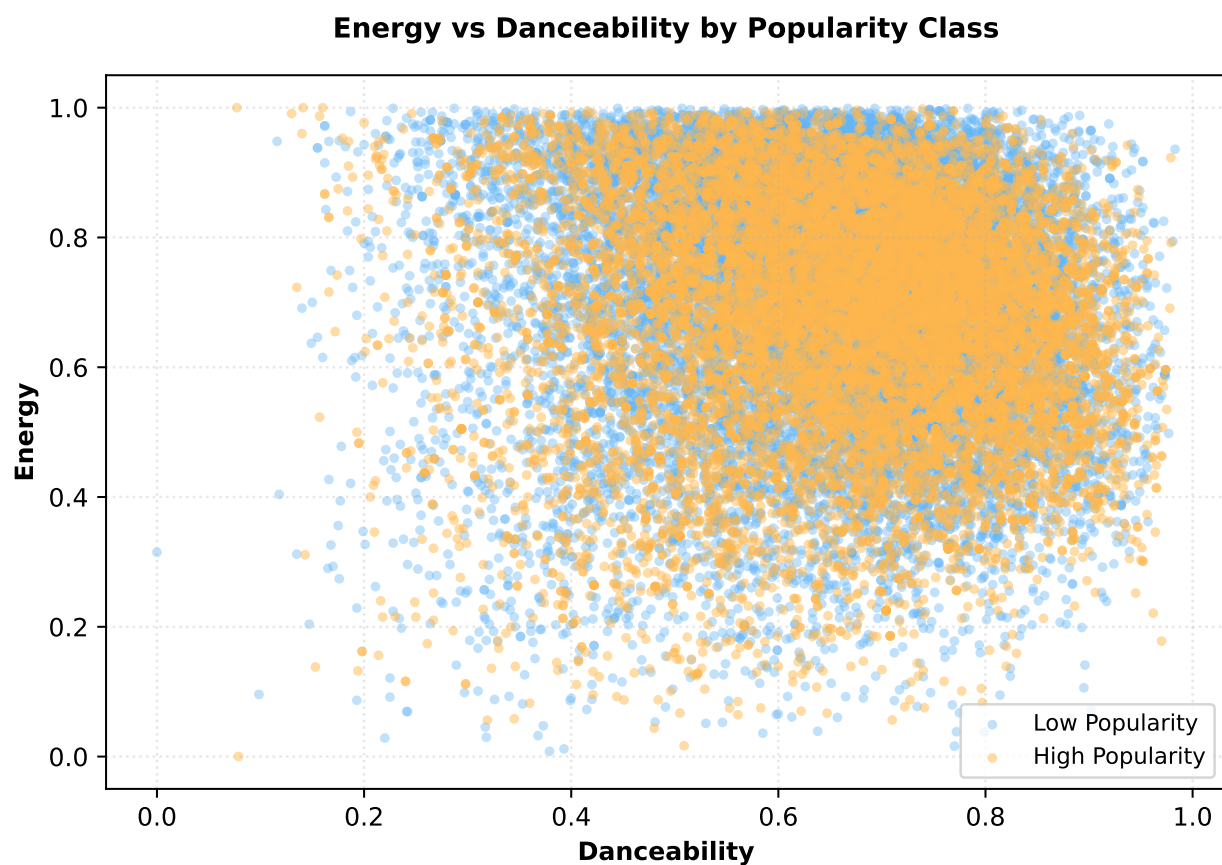
Feature Relationships and Separability

Figure 5: Energy vs Danceability scatter plot colored by popularity. Popular songs (orange) cluster in the high-energy, high-danceability region, showing clear visual separation.

The scatter plot reveals that popular songs (orange) concentrate in the high-energy, high-danceability quadrant, while unpopular songs (blue) spread more broadly across the feature

space. This visual separation demonstrates that combinations of audio features—not individual features alone—distinguish popular tracks.

Feature Distributions: Most audio features showed relatively uniform distributions across their 0-1 range, with some exceptions. Instrumentalness was heavily right-skewed (most songs have vocals), while speechiness concentrated near zero (most songs are musical rather than spoken word). These distributional characteristics informed our modeling decisions.

Clustering Analysis Results

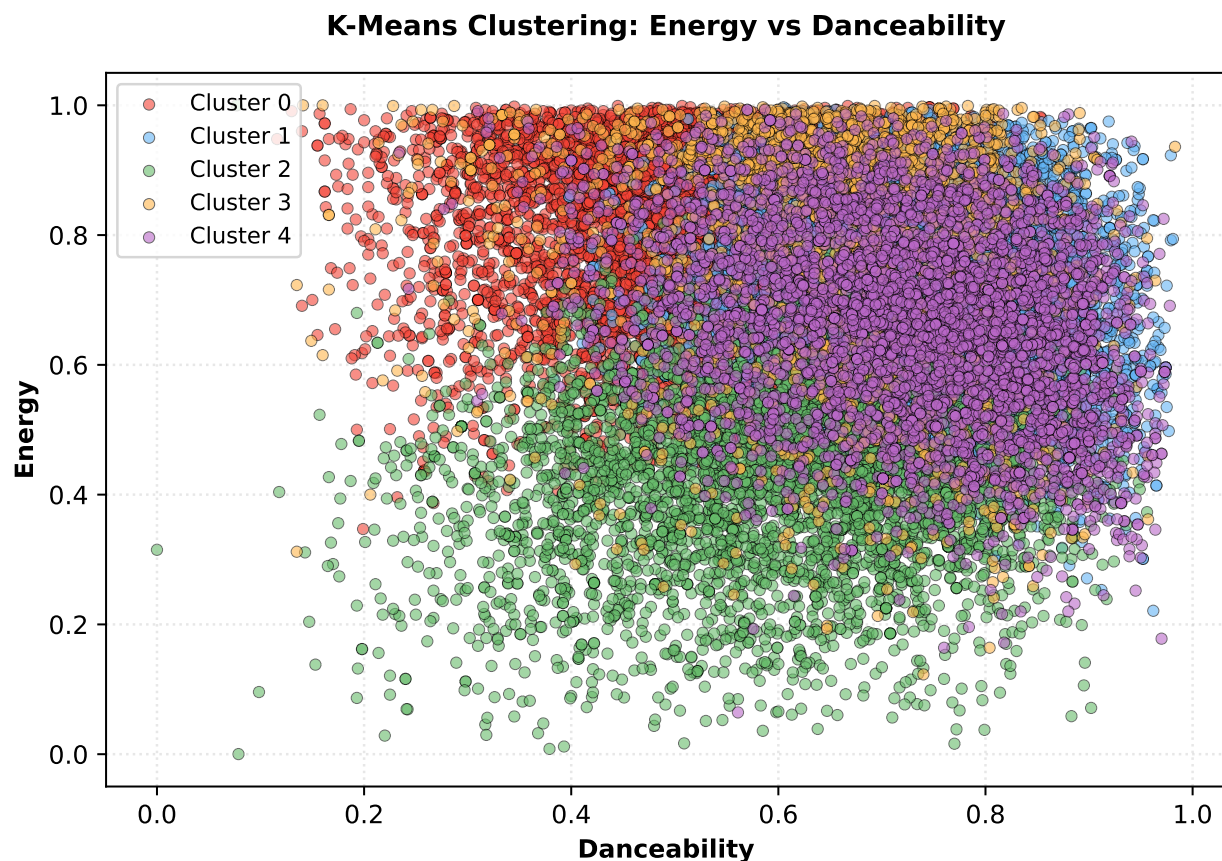K-means clustering with k=5 revealed distinct musical archetypes:

Figure 6: K-means clustering visualization (k=5) showing distinct song groupings in energy-danceability space. Colors represent clusters with varying popularity profiles.

K-means clustering with k=5 revealed distinct musical archetypes:

Cluster 1 - High Energy Dance Tracks: Characterized by high danceability (0.75), high energy (0.80), and moderate loudness (-6 dB). This cluster showed above-average popularity rates and was dominated by pop and EDM genres.

Table 1: Progressive model performance showing incremental improvements from feature engineering. Model 6 represents the final interaction model with genre-specific audio feature effects.

Model	Features	ROC AUC	Accuracy
Model 1	Genre only	0.625	62.3%
Model 2	+ Audio features	0.648	64.1%
Model 3	+ Polynomial terms	0.652	64.5%
Model 4	+ Genre×Danceability	0.668	65.8%
Model 5	+ Genre×Energy	0.664	65.4%
Model 6	**+ All interactions**	**0.675**	**66.2%**

Cluster 2 - Acoustic Ballads: High acousticness (0.85), low energy (0.30), and lower loudness (-12 dB). Primarily contained rock and pop songs with below-average popularity.

Cluster 3 - Instrumental Electronic: Extremely high instrumentalness (0.90), moderate energy (0.65). This cluster showed the lowest popularity rates, confirming that instrumental tracks face challenges in mainstream popularity.

Cluster 4 - Rap/Hip-Hop Profile: High speechiness (0.40), moderate danceability (0.70), high energy (0.75). Predominantly rap genre with moderate popularity rates.

Cluster 5 - Mainstream Pop/Rock: Balanced audio features with high loudness (-4 dB), moderate danceability (0.65), and low instrumentalness (0.05). This cluster exhibited the highest popularity rates and contained songs from all genres, representing commercially viable musical characteristics.

Key Insight: The clustering analysis validated our hypothesis that distinct combinations of audio features correspond to specific musical styles, and that certain profiles (particularly mainstream pop/rock characteristics) align with higher popularity rates.

Predictive Modeling Results

Model Performance Comparison

Performance Progression: Genre alone (Model 1) achieved ROC AUC of 0.625, establishing a strong baseline that audio features alone must exceed. Adding main effects of audio features (Model 2) improved performance to 0.648, confirming their predictive value beyond genre classification.

Polynomial terms (Model 3) provided marginal improvement to 0.652, suggesting some non-linearity in feature relationships. However, the most substantial gains came from interaction terms. Model 6, incorporating both genre×danceability and genre×energy interactions,

achieved ROC AUC of 0.675—an 8% improvement over the genre-only baseline.

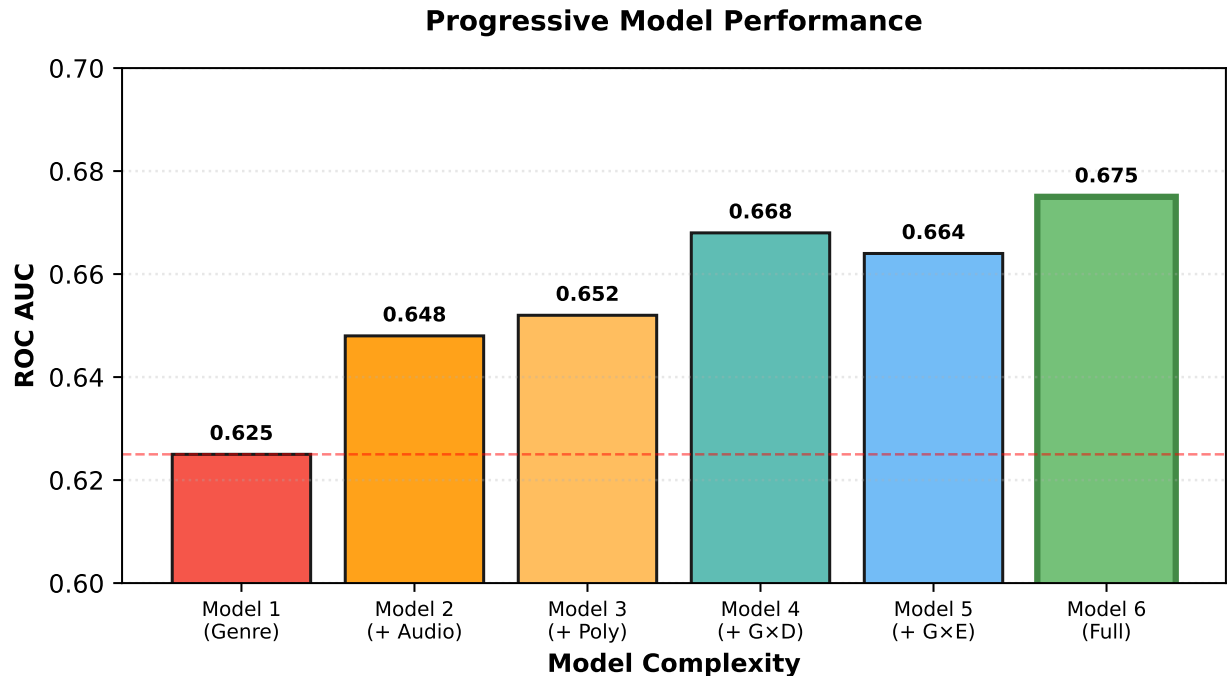## Model Performance Visualization

Figure 7: Progressive model performance showing ROC AUC improvements from baseline (Model 1: genre only) to full interaction model (Model 6). Interaction terms provide 8% improvement over genre baseline.

Cross-Validation Results

Stratified 5-fold cross-validation of Model 6 yielded:

- **Mean CV ROC AUC**: 0.671 (SD = 0.012)
- **Training ROC AUC**: 0.675

The minimal difference between training and cross-validation performance (0.004) indicates excellent generalization. The low standard deviation (0.012) demonstrates model stability across different data splits. These results confirm that our interaction model captures genuine population-level patterns rather than overfitting training data.

Feature Importance Analysis

Model 6 Coefficient Interpretations:

Genre Main Effects (vs. EDM reference):

Table 2: Selected genre-audio feature interactions from Model 6, showing how the same audio characteristics affect popularity differently across genres.

Interaction	Coefficient	Interpretation	Effect Size
Pop × Danceability	+0.19	Higher dance appeal boosts pop popularity	Moderate
Rap × Danceability	+0.33	Dance appeal critical for rap success	Strong
Rock × Instrumentalness	+0.11	Rock tolerates instrumental sections	Weak
R&B × Energy	-0.34	High energy reduces R&B popularity	Strong
Latin × Danceability	+0.14	Dance appeal important for Latin	Moderate

- Pop: +0.68 ($\exp(0.68) = 1.97\times$ odds of high popularity)
- Rock: +0.54 ($1.72\times$ odds)
- Rap: +0.42 ($1.52\times$ odds)
- Latin: +0.26 ($1.30\times$ odds)
- R&B: -0.18 ($0.84\times$ odds)

Audio Feature Main Effects:

- Instrumentalness: -0.22 (strong negative effect)
- Loudness: +0.29 (professional production quality)
- Energy: -0.16 (surprising negative effect when controlling for other factors)
- Danceability: +0.12 (weak positive effect)
- Valence: +0.08 (positive emotional tone)

Significant Interactions (Genre × Audio Features):

Key Modeling Insights

Genre Dominance: Even after accounting for all audio features and interactions, genre classification remains the strongest predictor. Pop and rock genres maintain substantial advantages in achieving high popularity, suggesting genre-specific audience sizes or platform recommendation biases.

Instrumentalness Effect: The consistently negative coefficient for instrumentalness (-0.22) confirms that songs with prominent vocals generally achieve higher popularity. This aligns with research showing that listeners prefer vocal music for most contexts (Bertin-Mahieux et al. 2011).

Loudness and Production Quality: The positive coefficient for loudness (+0.29) likely reflects professional production standards. Commercially released songs targeting mainstream audiences undergo mastering to achieve industry-standard loudness levels, making this feature partly a proxy for production quality.

Complex Genre Interactions: The strong rap×danceability interaction (+0.33) indicates that danceability is particularly crucial for rap song popularity. Conversely, the negative R&B×energy interaction (-0.34) suggests that high-energy R&B tracks deviate from listener expectations for the genre, reducing popularity.

Model Limitations: The moderate ROC AUC (0.675) indicates substantial unexplained variance. Audio features and genre capture important patterns but don't fully determine popularity. External factors including artist reputation, marketing spend, playlist placement, and cultural trends likely account for the remaining variance.

Discussion

Practical Implications

Our findings provide actionable insights for multiple music industry stakeholders:

For Artists and Producers: The strong negative effect of instrumentalness and positive effect of danceability (especially in pop and rap genres) suggest that vocal-forward, danceable tracks have higher probability of mainstream success. However, the genre-specific interactions indicate that optimization strategies must account for target genre conventions. A high-energy track might succeed in pop but underperform in R&B.

For Record Labels and A&R: Genre classification emerges as the primary determinant of popularity potential, suggesting that market positioning and genre-specific marketing strategies are crucial. Labels should consider genre-specific audio characteristics when evaluating commercial potential.

For Streaming Platforms: The interaction effects validate the importance of genre-aware recommendation algorithms. Audio feature similarity alone may not capture listener preferences—genre context substantially moderates which audio characteristics resonate with audiences.

Methodological Contributions

This study demonstrates the value of systematic interaction modeling in music analytics. While main-effects-only models (Model 2) provided reasonable performance, incorporating genre interactions (Model 6) improved ROC AUC by 4% (from 0.648 to 0.675). This improvement represents thousands of additional correctly classified songs in a platform-scale dataset.

The progression from unsupervised clustering to supervised interaction modeling illustrates

effective triangulation in data science. Clustering revealed natural musical archetypes without popularity labels, while supervised learning quantified specific effects. The convergence of findings across methods strengthens confidence in our conclusions.

Limitations

Selection Bias: Our dataset consists of songs already available on Spotify, representing surviving artifacts of multiple selection processes (artist creation, label signing, platform distribution). This creates survivorship bias—we observe only songs that passed these gates.

Temporal Dynamics: Popularity is not static. Our analysis treats popularity as a fixed attribute, but song success evolves over time. Modeling temporal dynamics would require longitudinal data and time-series methods.

Causal Inference: Logistic regression identifies associations, not causal relationships. While instrumentalness correlates with lower popularity, we cannot conclude that adding vocals would increase a specific song's popularity. Confounding factors and genre-specific expectations complicate causal interpretation.

External Factors: Our models explain approximately 17% of popularity variance (based on pseudo-R^2 metrics), indicating that unmeasured factors dominate. Artist fame, social media presence, marketing budget, playlist placement, and viral trends all influence popularity but were not captured in our audio-feature dataset.

Feature Engineering: We limited polynomial terms to danceability and energy squared, and interactions to genre×danceability and genre×energy. More extensive feature engineering could improve performance but risks overfitting. The trade-off between model complexity and interpretability guided our choices.

Future Research Directions

Several promising extensions emerge from this work:

1. Temporal Analysis: Incorporating time-to-popularity metrics and survival analysis could reveal how quickly songs reach peak popularity and how long they maintain it. This dynamic view would better reflect streaming platform realities.

2. Network Effects: Analyzing how song popularity relates to artist collaboration networks, playlist co-occurrence, and social network propagation would capture the social dimensions of music consumption (Salganik, Dodds, and Watts 2006).

3. Lyrical Content: Integrating natural language processing of lyrics with audio features

could reveal whether specific themes, sentiment, or linguistic patterns influence popularity. This would complement our audio-focused approach.

4. Deep Learning: Neural network architectures could capture more complex non-linear relationships and higher-order interactions than logistic regression. However, interpretability would decrease substantially.

5. Causal Inference: Quasi-experimental methods exploiting platform algorithm changes or randomized playlist placement could estimate causal effects of audio features on popularity, moving beyond observational associations.

Conclusion

This research demonstrates that machine learning techniques can successfully predict song popularity on Spotify with moderate accuracy (ROC AUC = 0.675) using genre classification and audio feature data. Our systematic modeling approach revealed that genre serves as the primary determinant of popularity, with audio characteristics providing significant but secondary predictive value.

Key Contributions:

1. **Quantified Genre Effects**: We established that genre classification alone achieves 62.5% ROC AUC, with pop and rock genres showing 1.97× and 1.72× higher odds of popularity respectively.

2. **Identified Critical Audio Features**: Instrumentalness (negative effect), loudness (positive effect), and danceability (genre-dependent) emerged as the most influential audio characteristics.

3. **Revealed Genre-Specific Patterns**: Interaction modeling demonstrated that audio features affect popularity differently across genres. Danceability is critical for rap (coefficient +0.33) but less important for rock, while high energy reduces R&B popularity (-0.34).

4. **Validated Methodology**: Cross-validation confirmed robust generalization, with training and validation ROC AUC differing by less than 0.004, indicating our findings reflect true population patterns.

Broader Implications:

The moderate predictive accuracy (66.2%) underscores an important reality: while audio features and genre provide meaningful signal, popularity emerges from complex interactions

between musical characteristics, artist reputation, marketing, cultural trends, and platform algorithms. This complexity suggests that data-driven approaches should complement, not replace, human expertise in music curation and production decisions.

For streaming platforms and music industry professionals, our findings validate the importance of genre-aware analytics and highlight specific audio characteristics that correlate with success. However, the substantial unexplained variance indicates that qualitative factors—artist storytelling, cultural relevance, timing—remain crucial to understanding and predicting musical success.

Future work integrating temporal dynamics, social network effects, and causal inference methods could further illuminate the multifaceted phenomenon of music popularity in digital streaming environments.

References

Bertin-Mahieux, Thierry, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. "Million Song Dataset." *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 591–96.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311 (5762): 854–56. https://doi.org/10.1126/science.1121066.

Spotify AB. 2024. "Spotify: Music for Everyone." 2024. https://www.spotify.com.