What Makes a Song a Hit?

How Simpson's Paradox Reveals That Genre Changes Everything About Music Popularity

The Question

Spotify publishes audio measurements for every song in its catalog: how danceable it is, how energetic, how acoustic, how loud, and more. With over 113,000 tracks across 114 genres, we asked a straightforward question: **can these audio characteristics predict whether a song will be popular?**

The intuitive answer is yes. Surely there's a recipe — some combination of tempo, energy, and danceability that makes a hit. We set out to find it.

The Surprise: It Depends on the Genre

When we looked at all songs together, audio features appeared to have almost no relationship with popularity. Correlations hovered near zero. A regression model using all audio features explained just 2.3% of popularity differences. The data seemed to say: **how a song sounds doesn't matter.**

But that conclusion was wrong — and the reason why is a well-known statistical phenomenon called **Simpson's Paradox.**

When we split the data by genre and looked again, audio features suddenly became strong predictors. The catch: the *same feature* often predicted popularity in *opposite directions* depending on the genre. High danceability predicts hits in pop, but not in classical. High acousticness helps in folk, but hurts in hip-hop. When all genres are mixed together, these opposite effects cancel each other out, making it appear as though nothing matters.

> **"Genre doesn't just change how popular a song is — it changes what makes a song popular."**

The Evidence: Three Phases, One Finding

We tested this finding across three analytical approaches, and it held up every time:

Phase	Approach	Without Genre	With Genre	Improvement
Exploration	Correlation analysis	Near zero	**Strong (varied)**	Patterns emerge
Regression	Statistical modeling	$R^2 = 0.023$	**$R^2 = 0.313$**	13× better
Classification	Machine learning	AUC = 0.75	**AUC = 0.87**	+16% accuracy

The pattern was consistent: audio features alone told us very little, but adding genre context transformed prediction quality. In the regression phase, a model that allowed feature effects to vary by genre explained 13 times more variance than one that treated all genres the same. In the machine learning phase, a gradient-boosted classifier improved from AUC 0.75 to 0.87 when genre was included.

A Deeper Layer

We also found that the improvement from genre depended on the type of model. Simple linear models (like logistic regression) gained almost nothing from genre, because they can only use it as a baseline shift — each genre gets a different starting point, but the audio features affect all genres the same way. Tree-based models like XGBoost gained dramatically, because they can learn that danceability matters *differently* in pop than in classical. Genre must be modeled as an **interaction**, not just a category.

When we trained separate models for individual genres, we found that hit predictability varied widely. Some genres had clear audio signatures for success — the "right" sound exists and the model can learn it. Other genres were nearly unpredictable from audio alone, suggesting that popularity there is driven by factors outside our data: who the artist is, whether the song landed on a major playlist, or whether it went viral on social media.

Why This Matters

For the music industry: Any tool claiming to predict hits from audio analysis alone is fundamentally limited — and its accuracy depends heavily on genre. Genres with strong audio conventions are predictable; genres where non-audio factors dominate are not.

For recommendation systems: Algorithms that ignore genre context or treat it as a simple filter are missing the most important signal in the data. Genre changes *which* features matter, not just *how much* they matter.

For data analysis broadly: Simpson's Paradox is not just a textbook curiosity. In this dataset, ignoring it would have led to the conclusion that audio features are useless. Accounting for it revealed meaningful predictive power. This is a reminder that aggregate statistics can be deeply misleading when subgroup effects vary — a lesson that applies far beyond music.

Dataset: Spotify Tracks (113,393 tracks, 114 genres) | **Tools:** Python, scikit-learn, XGBoost, statsmodels |
Author: [Your Name], MS Business Analytics, UT Arlington | **Date:** February 2026