HOME

PLAYLIST

GROUP 4

Universal Music Spotify

The Anatomy of a Hit: Unlocking Trends in Music Data!

DSO 528: Blended Data Business Analytics for Efficient Decisions

Dec 4, 2024

0:00

15:00

Problem Statement

- Universal Music struggles to efficiently predict and promote songs that will become popular, leading to missed revenue opportunities.
- Popular songs generate significantly more revenue compared to non-popular songs, making accurate predictions critical.

Why It Matters?

- Popular songs average $150K/year, while unpopular songs only generate $20K/year.
- Optimizing promotional investment can boost ROI and improve artist support.

Business Objectives

- Predict song popularity using historical and recent Spotify data.
- Optimize promotional investments for better ROI.
- Provide actionable insights for Universal Music's promotional strategies.

Impact

- Improved revenue generation.
- Data-driven decision-making for artist support and marketing.

Data Overview

Feature	Dataset 1	Dataset 2
Total Rows	26,266	953
Columns	19	17
Key Features	danceability, tempo, energy	streams, artist_count, key
Missing Values	3 entries in metadata	10% missing in key

Data Merging

Step 1

Data Alignment

- Renamed columns in df2 (e.g., released_year → release_year, artist(s)_name → track_artist).
- Added missing columns: playlist_genre and duration_ms.

Step 2

Metadata Creation

Added source column:
- pre-COVID for df1
- post-COVID for df2

Step 3

Dataset Concatenation

Merged both datasets using concat function

Step 4

Validation

Rows:
- Dataset 1: 26,266
- Dataset 2: 953
- Combined: 27,219

Step 5

Missing Values Check

Missing columns: duration_ms, streams, track_id, playlist_genre.

Specific gaps:
- key (95 values missing)
- track_name and track_artist (3 missing)

Data Cleaning

Missing Values: Imputed key based on artist trends; replaced invalid numeric values with medians.

Data Types: Converted release_year, release_month, and streams to appropriate formats.

Duplicates & Outliers: Removed duplicates and normalized streams with log transformation.

Validation: Ensured critical columns like key and streams had no missing values.

Feature Engineering

Derived Features: Created interaction metrics like dance_energy and positive_vibe to capture feature relationships.

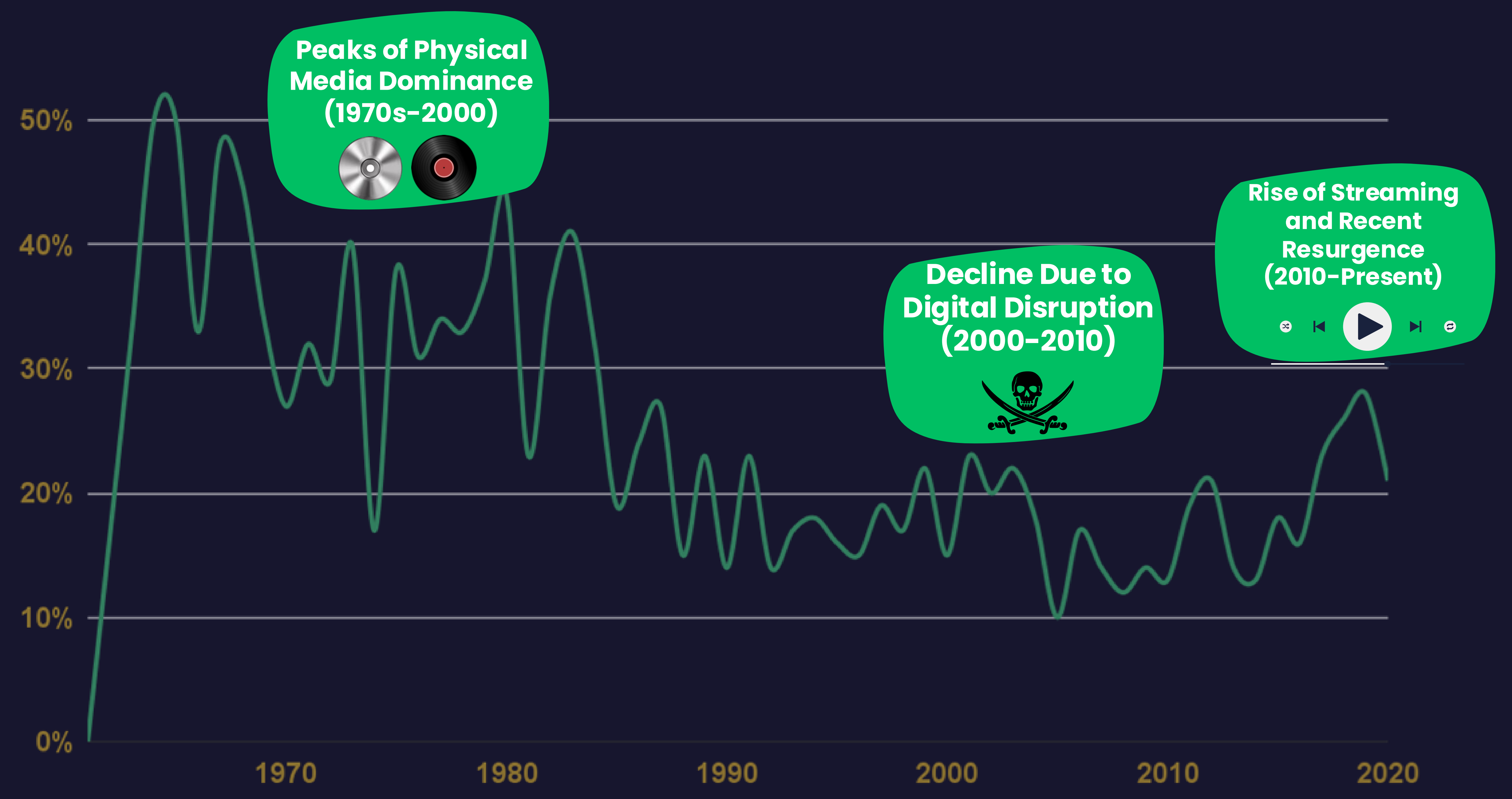
Temporal Features: Added days_since_release, release_season, and song_age for time-based insights.

Binning: Categorized tempo and streams into meaningful bins for interpretability.

Aggregations: Calculated artist-level metrics like artist_song_count and artist_total_streams.

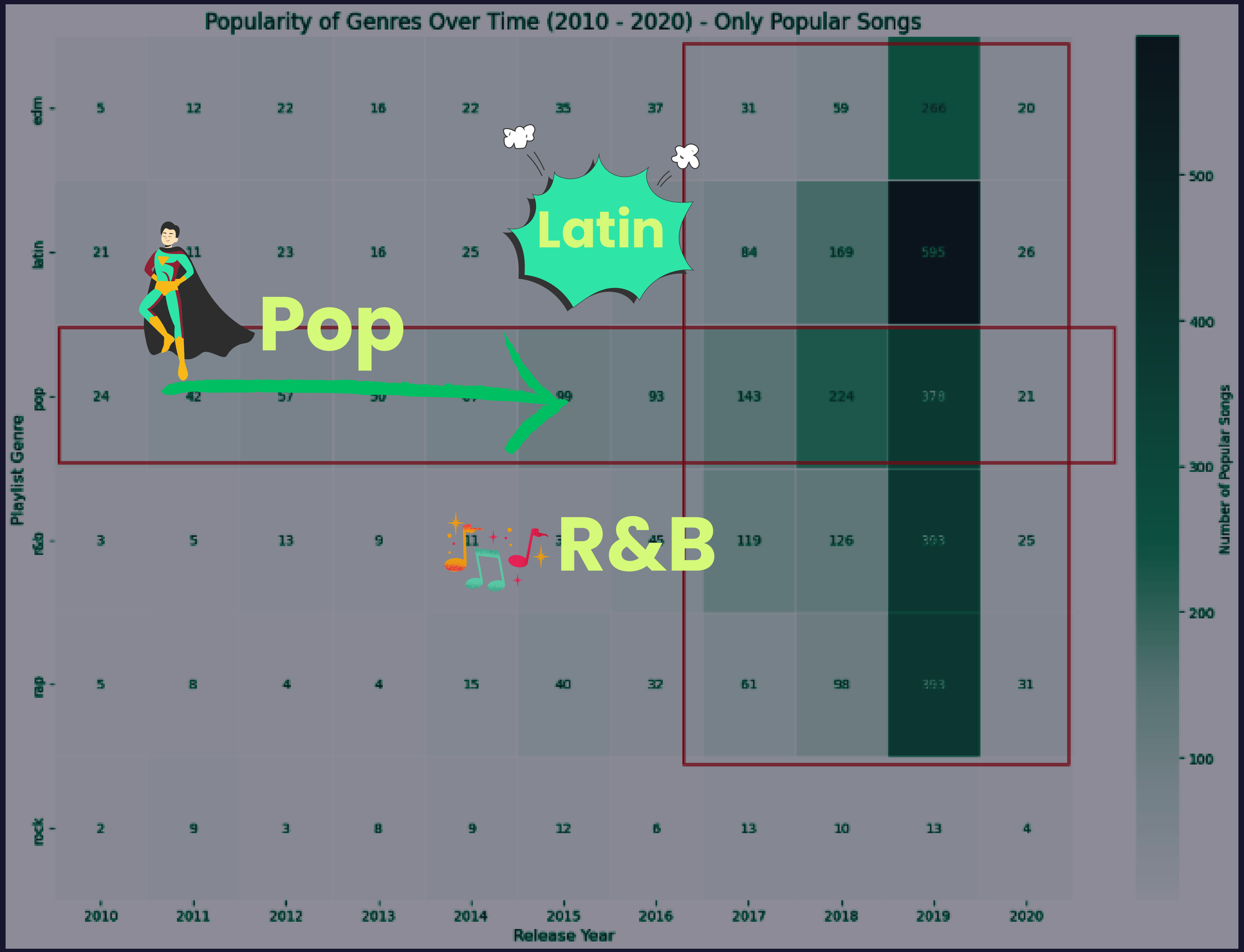
Evolution of Music Popularity: A Year-by-Year Analysis

Trends in Song Popularity Over Time (1970-2020)



Genre Fusion and Popularity: How Genres evolved

Trends in Song Popularity Over Time (1970-2020)



Pop remains a dominant genre

Latin influences in pop culture

Increase in Hybrid and Mixed Genres

Technological Advancements

Breaking Down the Success Formula

Data Path Followed to derive key insights

EDA →

Feature Selection

Pop and Latin songs score higher in danceability

Popularity has steadily increased over the years, with more popular songs released after 2015

Most frequent: Rap

Release Season: Songs released in Summer and Fall are slightly more popular on average.

Built-in Feature Importance

Random Forest

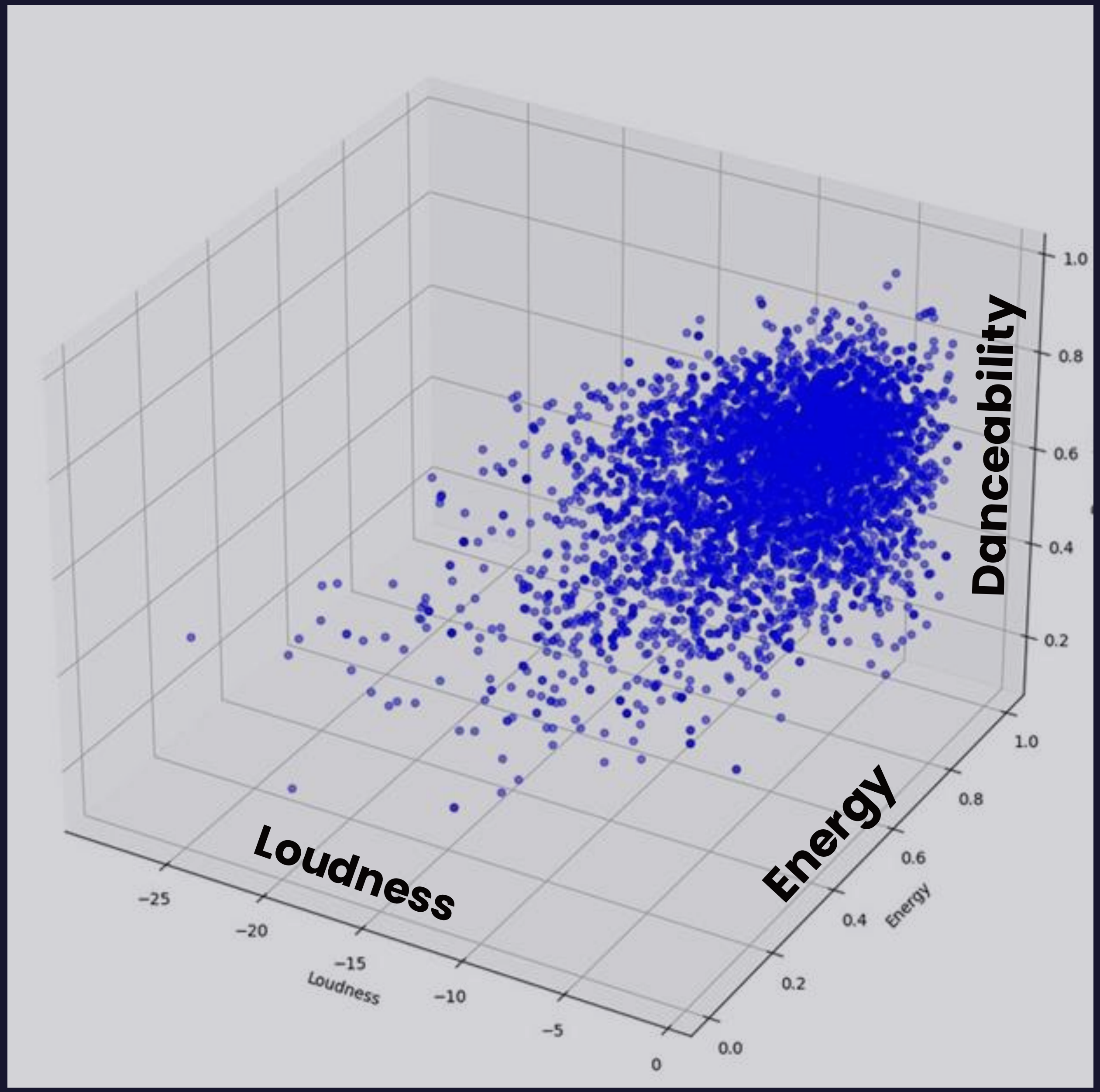
Energy

Danceability **Tempo** **Loudness**

Valence

Breaking Down the Success Formula

Spotify's winner Triplet



(PIS)
???

Energy — Danceability — Loudness

High Energy

High Danceability

Moderate Loudness

KPI Popularity Impact Score = (w1 x Energy)+(w2 x Danceability)+(w3x Loudness)

Models Explored and Hyperparameter Tuning

Initial Model Testing

Logistic Regression

Random Forest

XG Boost

LGBM

Hyperparameter Tuning

Number of Trees

Maximum Dept

Minimum Samples Split

Maximum Features

Cross-Validation Accuracy

Model Evaluation and Selection

Model Selection

Chosen Model: Random Forest

Reason for selection?
High accuracy, balanced precision & recall, robust to overfitting.

Performance Metrics

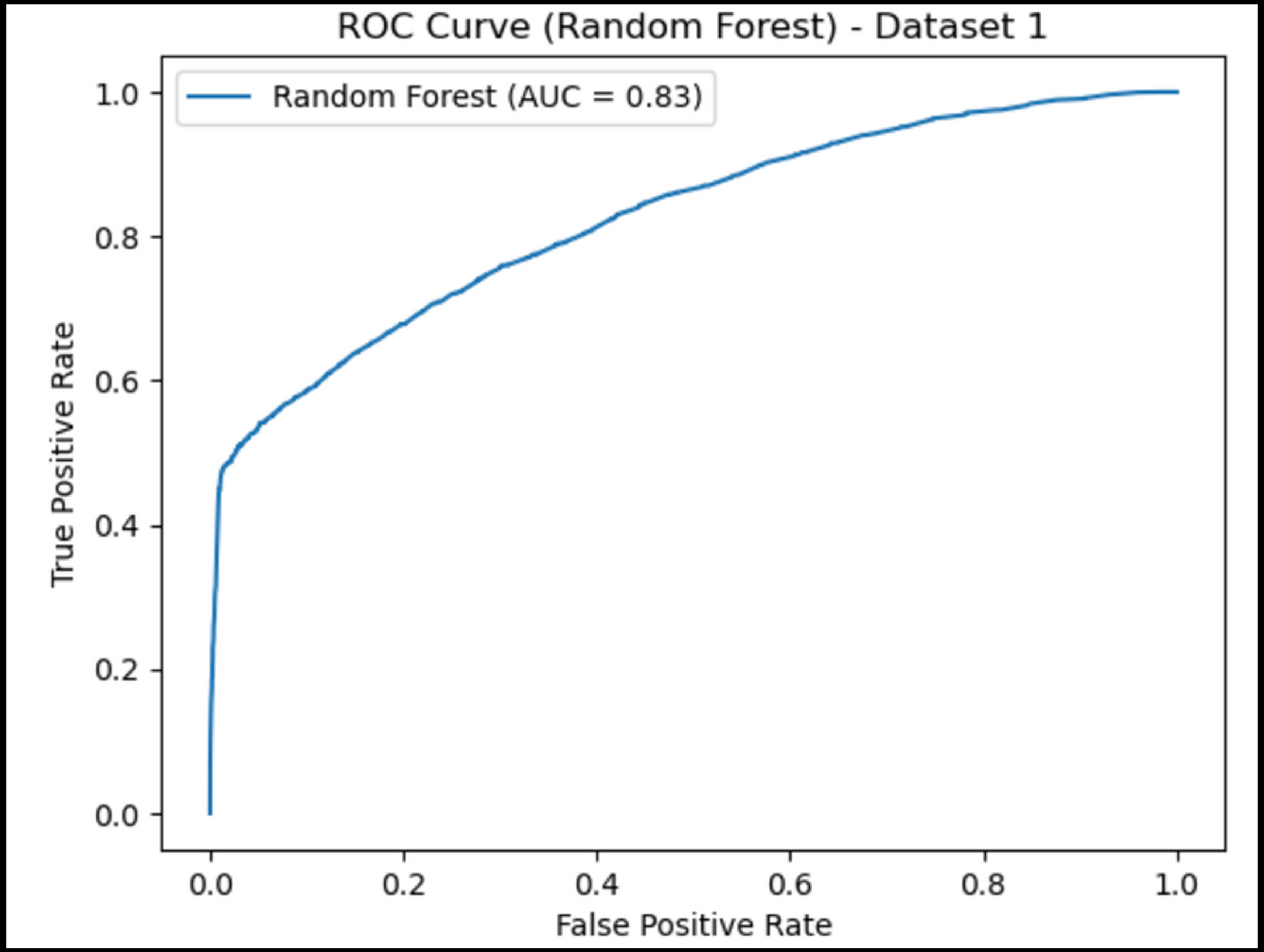
Accuracy: 87.1%

Precision: 87.5%

Recall: 87.3%

F1-score: 86.4%

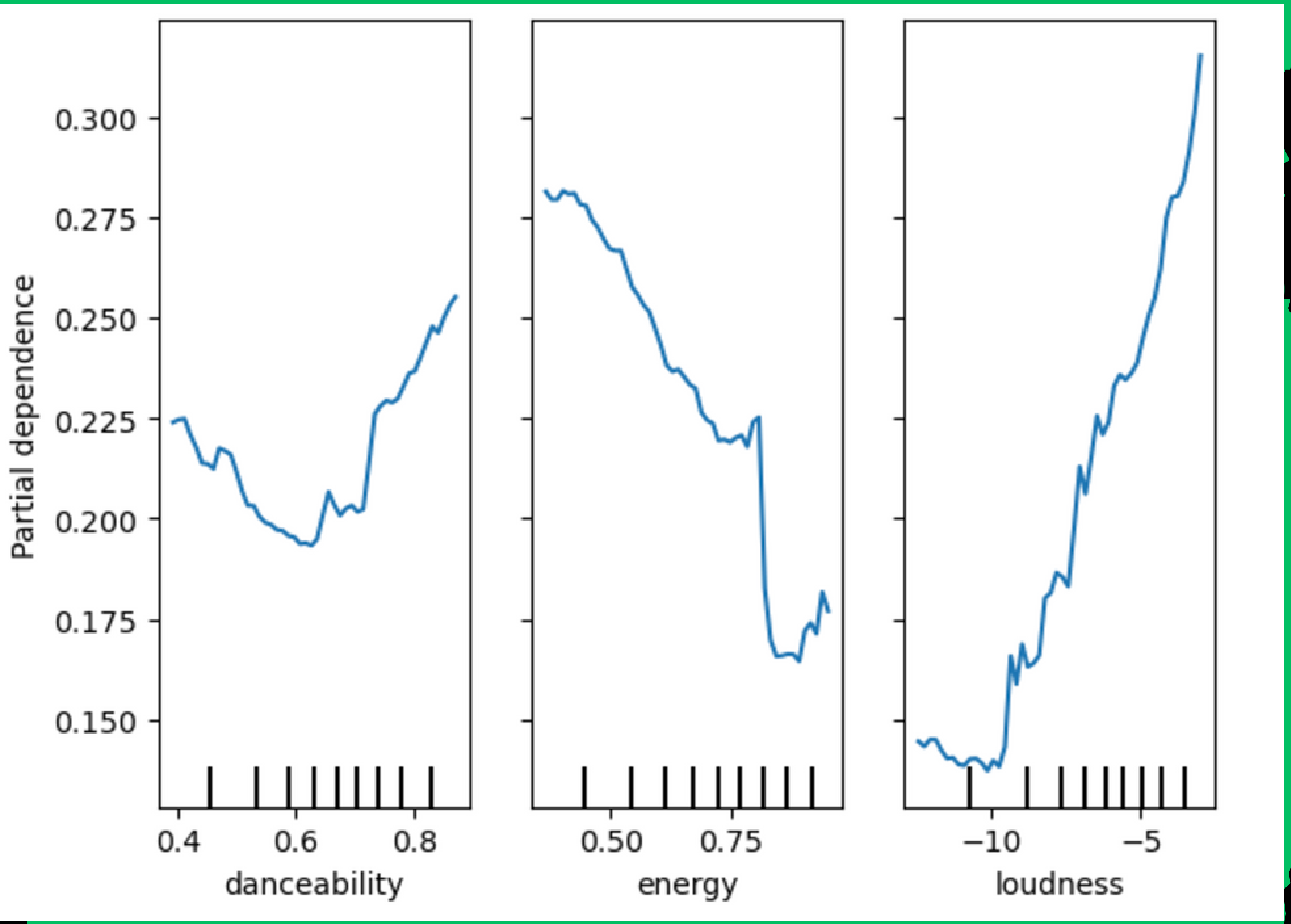
ROC Curve Results



Model Interpretation: Advanced Insights

Briefly mention examples of songs that were misclassified and why. For instance:
- False Positives: Songs predicted as popular but were

Partial Dependence Plots



Edge Cases

False Positive

Songs predicted as popular but turned out to be unpopular.

Impact: Leads to unnecessary investment in low-performing tracks, reducing ROI.

False Negative

Songs predicted as unpopular but were actually popular

Impact: Missed opportunities to capitalize on high-performing tracks.

Model Deployment Plan

Does the model make business sense?

Business goals Alignment: Helps select profitable tracks.

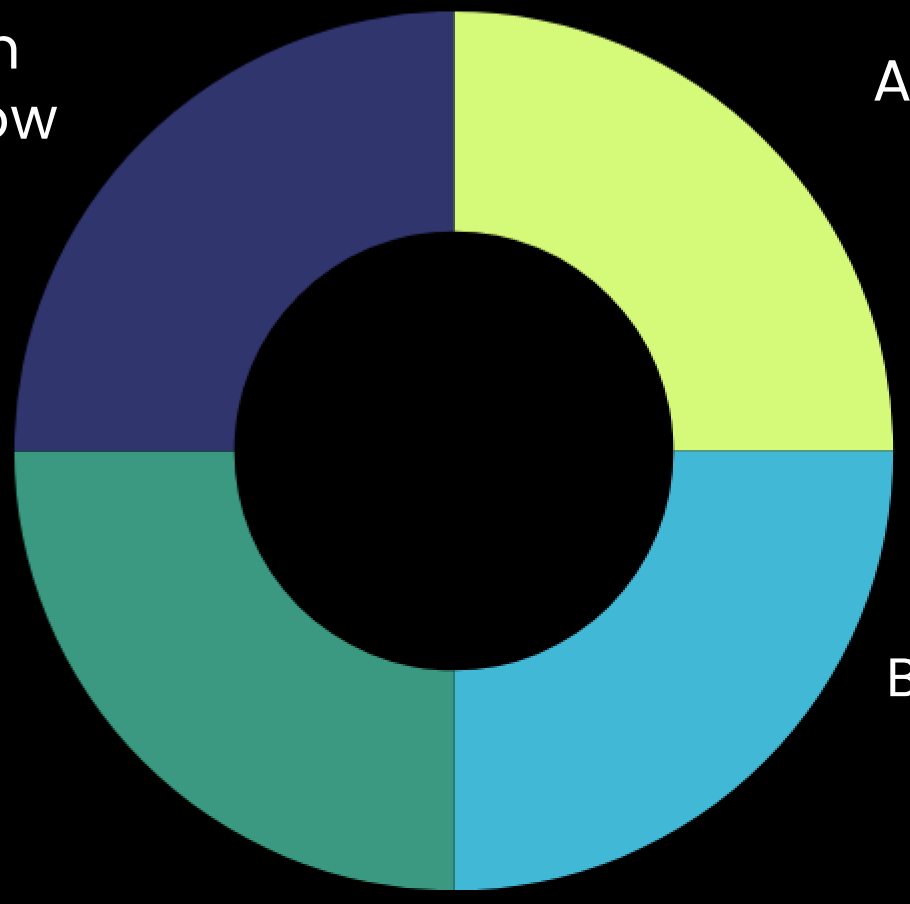
High predictive power: Reliable 88.8% accuracy

Actionable insights: Data-driven song selection

Deployment Plan

Step 1
Integration with Business Workflow

Step 2
Access and Usability



Step 3
Monitoring and Maintainence

Step 4
Business Impact and Scaling

Expected Outcomes

Improved ROI

Reduced $ Losses

Practical Considerations

Strengths

Weakness

Expected Payout Analysis

With the Modification of 20% Probability of an Unpopular Song becoming Popular

RFM(REVISED)	PREDICTED			
		0	1	Total
ACTUAL	0	6032	105	6137
	1	860	1169	2029
	TOTAL	6892	1274	8166

Revenue of a Popular Song (R_pop) = 150k – 30k = 120k

Revenue of an Unpopular Song (R_unpop) = 20k – 30k = -10k

*Expected Payout = TP * R_pop + FP * R_unpop*

*Expected Payout = 1169 * 150k + 105 * (-10k)*

$ 174,300,000

Expected Payout Analysis

RFM	PREDICTED			
		0	1	Total
ACTUAL	0	6032	131	6163
	1	860	1143	2003
	TOTAL	6892	1274	8166

Revenue of a Popular Song (R_pop) = 150k – 30k = 120k

Revenue of an Unpopular Song (R_unpop) = 20k – 30k = -10k

*Expected Payout = TP * R_pop + FP * R_unpop*

*Expected Payout = 1143 * 150k + 131 * (-10k)*

135,850,000

Recommendation Evaluation Matrix

Recommendation	Impact on Revenue	Implementation Cost	Time to Implement	Risk Level	Feasibility
Leverage danceability to create targeted playlists for high-engagement events	High	High	Medium	Low	High
Invest in mood-based playlist strategies (based on valence)	High	Low	Short	Medium	High
Use short-form video platforms to amplify tracks with optimal loudness and valence	High	High	Medium	Low	High
Refine energy-based segmentation to recommend songs for different activities (workouts, relaxation)	Medium	High	Medium	Low	Moderate
Develop predictive analytics to optimize artist collaborations based on interaction effects	Medium	Low	Medium	Low	High
Create fitness-focused playlists leveraging energy and loudness synergy	Low	High	Short	Low	Moderate

Recommendation Evaluation Matrix

	High Maturity	Low Maturity
High Business Value	Leverage danceability to create targeted playlists for high-engagement events	Invest in mood-based playlist strategies (based on valence)
Low Business Value	Refine energy-based segmentation to recommend songs for different activities	Create fitness-focused playlists leveraging energy and loudness synergy

Thank You!

0:00

15:00