IDENTIFYING SINGLES VS ALBUMS: FEATURE INSIGHTS & MODEL STRATEGY

A MACHINE LEARNING ANALYSIS OF THE SPOTIFY-YOUTUBE DATASET

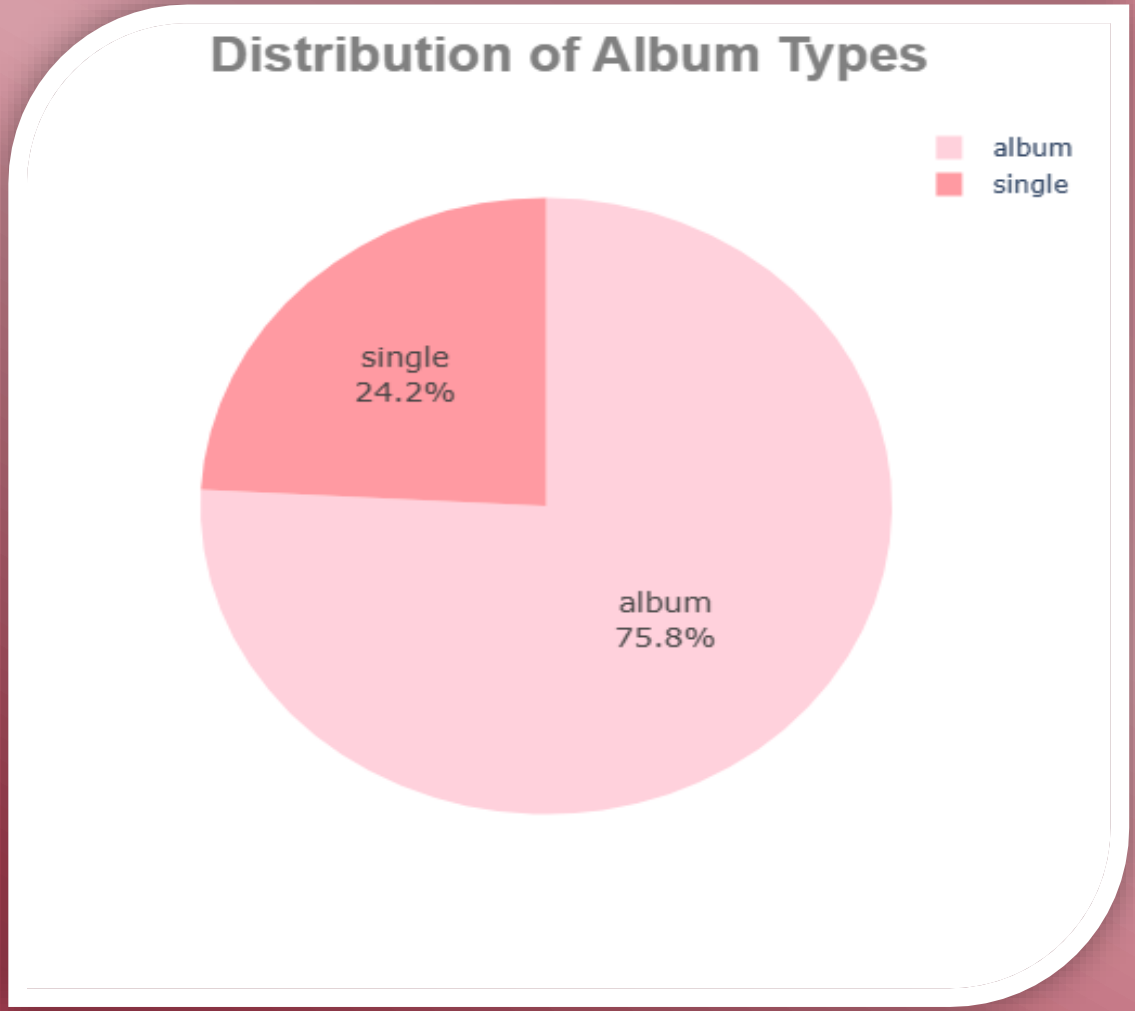
PRESENTED BY: NORA AND LAYLA

DATA CLEANING & FEATURE ENGINEERING

Section	Content
Cleaning	- Dropped ~700 rows with missing critical values (e.g. Views, Duration) - Filled missing Likes/Comments with 0 to reflect no engagement
Transformation	- Applied log(x+1) to Views, Likes, Streams to reduce skew - Ensured numerical stability for ratio features
Engineered Features	- Likes_per_View: engagement metric - Stream_per_Minute: pacing efficiency - Danceability × Valence: emotional rhythm combo
Why it matters	Cleaning first avoids invalid calculations. Feature engineering captures deeper behavioral and musical patterns. **Improved model performance:** • **SVM & GBoost** benefitted from smoother, scaled features • **Random Forest** leveraged clearer splits from engineered features

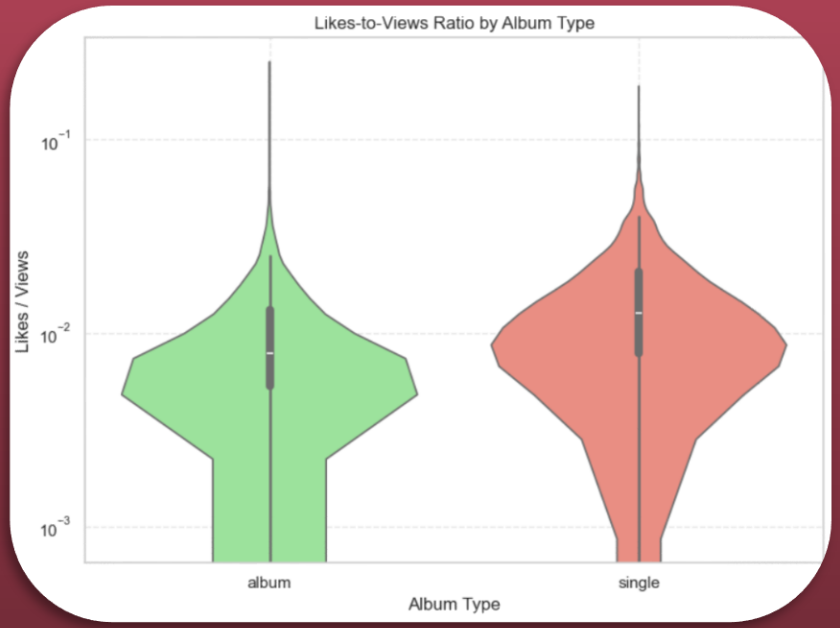
CLASS IMBALANCE IN ALBUM TYPE

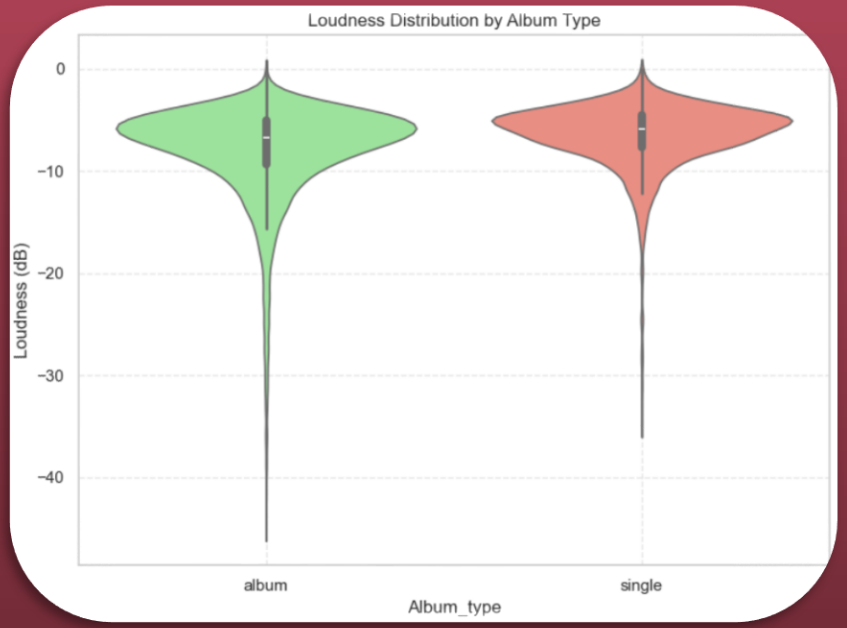
- ~75% of songs are labelled as **album**
- **Singles** are significantly underrepresented in the data
- Models tend to favor the majority class by default
- Can lead to **misleading accuracy** and **poor F1 for singles**
- Must address via feature engineering, class weighting, or evaluation metrics

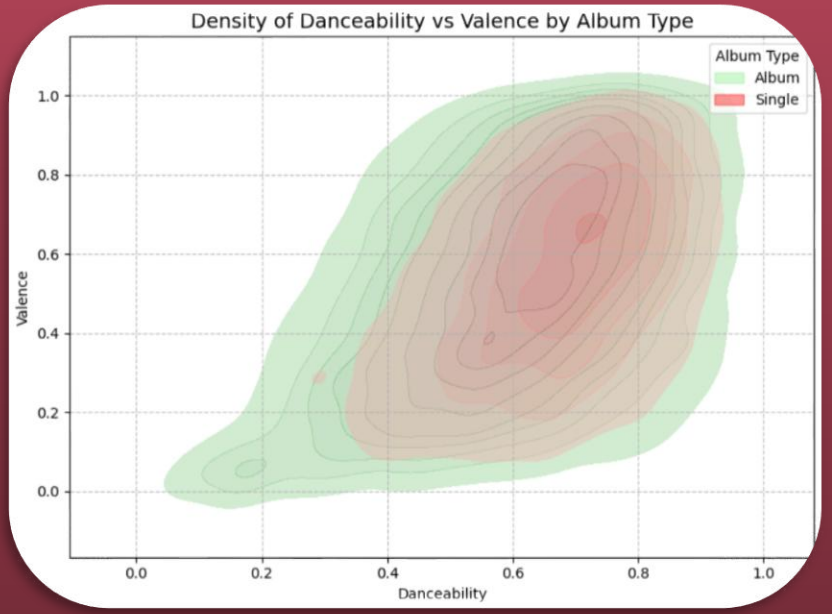


KEY FEATURE INSIGHTS: WHAT MAKES A SINGLE?

- Singles tend to be louder and more danceable
- Likes-to-Views ratio is often higher for singles, suggesting stronger engagement.
- Danceability + Valence: Singles cluster around higher emotional and rhythmic energy







MODEL SELECTION AND EVALUATION

Models Tested:
Random Forest, SVM, and Gradient Boosting — chosen for their ability to handle non-linear patterns and feature interactions.

Metric Used:
Macro F1 Score – gives equal weight to both classes (single vs. album), unlike accuracy which is skewed by class imbalance.

Best Model:
Random Forest performed best, especially in detecting singles — our main challenge. It effectively handled imbalance and used engineered features well.

Tuning:
Used **GridSearchCV** on validation set; test set remained untouched for unbiased evaluation.

Key Insight:
Tree-based models (RF, GB) outperformed SVM by better capturing complex relationships and mixed feature types.

Model	Accuracy	Macro F1
Random Forest	87.3%	0.87
GBoost	86.5%	0.86
SVM	86.0%	0.86

CLUSTERING SONGS USING KMEANS + PCA PROJECTION

What We Did	Why We Did It
Chose key musical features (e.g. Danceability, Valence, Energy)	To capture traits that differentiate singles from albums
Applied StandardScaler to normalize feature ranges	So that all features contribute equally to clustering
Clustered songs into 3–5 groups using KMeans	To uncover natural groupings in the audio space
Used PCA (2 components) **only for visualization**	To plot high-dimensional clusters in 2D and inspect separation patterns visually
Examined how singles/albums are distributed across clusters	To validate feature-based patterns and model assumptions