Rhythms Through Time: A Data-Driven Analysis of Lyrics and Audio Features

Abstract

This study investigates the complex relationship between musical characteristics and song popularity by utilizing two extensive Spotify datasets: one comprising 10,000 songs with associated audio features and another containing 57,651 songs with lyrics spanning from the 1950s to the 2010s. By employing advanced machine learning techniques and natural language processing methodologies, we analyze how audio features and lyrical content contribute to a song's popularity. Beyond mere analysis, our objectives include developing a recommender system that leverages these features to suggest similar songs and identify patterns correlated with popularity. Additionally, we utilize the sBERT deep learning model to examine semantic trends in song lyrics over time, linking lyrical themes to specific temporal patterns.

This research integrates various analytical approaches, including regression models, embedding strategies, clustering techniques, and recommender system frameworks, to elucidate the intricate dynamics of musical features across different decades. Ultimately, our findings aim to provide actionable insights for music industry stakeholders—such as artists, producers, and advertisers—by exploring the interplay between audio characteristics, lyrical content, and temporal trends.

---

1. Introduction

The music industry has long been captivated by the elusive factors that determine song popularity. Comprehending the underlying elements that drive a track's success remains a paramount challenge for artists, producers, and music analysts. This study seeks to dissect the multifaceted components influencing a song's reception by analyzing both quantitative audio features and qualitative lyrical content.

Our primary research questions focus on whether specific musical attributes can predict song popularity and whether data-driven methodologies can yield insights into song composition, genre evolution, and trend identification. We address these inquiries through three principal investigative dimensions:

1. **Feature-Popularity Correlation:** Examining which song features (e.g., tempo, energy, danceability, lyrical complexity) are associated with popularity over time.
2. **Recommender System Development:** Developing a recommender system that utilizes these popularity-driving features to suggest similar songs tailored to specific preferences or temporal trends.
3. **Semantic Analysis of Lyrics:** Applying the sBERT deep learning model to analyze lyrical themes and their temporal shifts, evaluating how meaning evolves across decades and correlates with song and genre virality.

This study is further motivated by practical applications and exploratory questions:

- **Financial Motivation:** Can advertisers and companies leverage insights into the popularity of music genres over time to optimize their marketing strategies? Research indicates that playing popular music can positively influence consumer behavior in venues such as restaurants and stores, potentially increasing spending.
- **Cyclical Trends:** Are music trends cyclical? By identifying patterns in musical characteristics over time, we aim to ascertain whether historical trends repeat themselves in the music industry.
- **Insight for Artists and the Industry:** How can the music industry—from artists to producers—utilize these findings? Can our model assist creators in identifying elements that may enhance the reception of their work or predict forthcoming trends in the industry?
- **Deeper Understanding of Lyrics and Virality:** Can we establish a more profound connection between lyrical content and emotional or cultural resonance? What role do lyrics play in determining the virality of a song? By identifying metrics that define "viral" music, we aim to uncover the factors driving listener engagement.

Through this multifaceted approach, we endeavor to bridge the gap between music analytics, cultural trends, and commercial applications, ultimately providing a comprehensive understanding of how music both shapes and is shaped by its temporal context.

---

2. Related Work

Prior to engaging with the dataset, we conducted a thorough review of existing literature to identify related studies. This step was crucial to ascertain whether credible and reliable models have been developed to explore song popularity using similar datasets. While numerous studies have examined song popularity and its associated features, we identified a lack of established benchmarks, with the exception of a study utilizing the same Kaggle dataset as ours [1].

[1] Joe Beach Capital. (2023). *Top 10,000 Songs: EDA and Models*. Retrieved from https://www.kaggle.com/code/joebeachcapital/top-10000-songs-eda-models

The aforementioned study by Joe Beach Capital employed various models to explore song popularity, which we adopted as our benchmark. Additionally, they developed a recommendation system that we aim to enhance and build upon in our research.

---

3. Data and Methodology

3.1 Data Collection and Preprocessing

This study utilized two comprehensive Spotify datasets: the first comprising 10,000 songs with their respective audio features, and the second encompassing 57,651 songs with associated lyrics, covering a substantial temporal range from 1950 to 2024. These datasets provided extensive metadata, including audio features, lyrical content, artist information, and popularity metrics.

- **First Dataset:** Contains 35 columns, encompassing features such as popularity, danceability, energy, loudness, acousticness, liveliness, tempo, release date, track duration, among others. Detailed descriptions of these features are available on the Spotify API website. In essence, these features not only characterize the song's attributes but also include metadata related to the artist's name, album, and release dates.
- **Second Dataset:** Consists of four columns: artist names, song titles, links, and corresponding lyrics.

The preprocessing pipeline incorporated several sophisticated techniques:

- **Handling Missing Values:** Missing numerical features were addressed using median imputation.
- **Duplicate Removal:** Duplicate entries were identified and removed, retaining only unique records.
- **Text Consistency:** Whitespace removal, consistency checks, elimination of special characters, and conversion to lowercase were performed to ensure uniformity in textual data.
- **Feature Scaling:** Applied to ensure comparable representation across different attributes.
- **Lyrical Preprocessing:** Included tokenization, stop word removal, elimination of text within parentheses, conversion to lowercase, removal of special characters, and lemmatization to prepare lyrics for sBERT embedding.

Subsequently, the two datasets were merged based on artist names and song titles to create a unified dataset containing both audio features and lyrics. This merging process was challenging due to the necessity of preprocessing and cleaning artist and song names in both datasets. A fuzzy matching technique was employed, with a similarity threshold set at 80. Entries with similarity scores below this threshold were classified as non-matching, while those above were merged. This method resulted in approximately 2,011 matched entries retaining columns from both datasets. The merged dataset underwent further lyrical preprocessing, reducing the number of entries to approximately 1,634.

3.2 Feature Engineering and Analysis

We employed multiple feature engineering strategies to capture the nuanced characteristics of musical tracks:

- **Audio Feature Interaction Terms:** Created to capture complex relationships, such as combining energy and valence to form energy-dance interactions.
- **Binning Continuous Features:** Applied to features such as tempo and duration to categorize continuous variables into discrete bins.
- **Conversion of Continuous to Categorical Variables:** For instance, converting the continuous feature "instrumentalness" into a categorical variable "is_instrumental."
- **One-Hot Encoding:** Applied to categorical features such as genre to achieve a clearer representation in the feature set.
- **Derivation of Pseudo Artist Popularity:** Calculated by taking the mean of all artists' popularity scores within the dataset.

These feature engineering techniques resulted in a stable set of features suitable for model training to investigate relationships between song features and popularity.

Within the merged dataset, the 'Release Year' of each song was binned into specific time periods reflective of musical eras from the 1950s to the 2010s. Given the dataset's bias towards the 1990s, we further subdivided the 1990s into three distinct periods to ensure fair representation. The final time period bins are as follows:

- (1950–1959, "Birth of Rock 'n' Roll")
- (1960–1969, "Cultural Revolution")
- (1970–1979, "Rise of Diverse Genres")
- (1980–1989, "MTV Era and Electronic Explosion")
- (1990–1993, "The End of an Era")
- (1994–1996, "Expansion and Mainstream Success")
- (1997–1999, "Technological Advancements and Pop Culture Explosion")
- (2000–2009, "Digital Revolution")
- (2010–2019, "Streaming and Global Connectivity")

3.3 Modeling Approaches

Our modeling strategy incorporated the following techniques to address the diverse aspects of song popularity:

1. **Clustering:**
    - Identified groupings and outliers in continuous audio features such as song duration.
    - Applied various clustering methods on t-SNE and UMAP visualizations to assess the relationships and groupings within the data.
2. **Regression Models:**
    - **Linear, Ridge, and Lasso Regression:** Served as baseline models to quantify linear relationships and perform feature selection.
3. **Ensemble Models:**

- **Random Forest Regressor:** Captured non-linear feature interactions but exhibited limited performance due to overfitting risks.
- **Gradient Boosting (XGBoost and LightGBM):** Tuned models achieved superior performance, effectively handling complex relationships within the data.

4. **Deep Learning:**
   - **Keras Sequential Model:** Modeled intricate non-linear patterns in audio features but was constrained by the limited dataset size.
   - **sBERT Embeddings:** Transformed lyrics into semantic vectors, enabling the integration of textual features and the analysis of temporal shifts in lyrical themes.
5. **Word Clouds:**
   - Generated for the lyrics of songs within specific time period bins to visually represent the overall sentiments of each period.
6. **t-SNE/UMAP:**
   - Utilized for visualization purposes, binned by time periods to examine relationships between artists across multiple time frames.
   - Assessed the dataset before and after applying sBERT embeddings to evaluate improvements in clustering/grouping of songs and artists.

---

4. Results and Discussion

4.1 Model Performance

The performance of our predictive models was evaluated using Mean Squared Error (MSE) and the coefficient of determination (R^2) metrics for regression tasks, alongside the Silhouette Score for clustering performance. These metrics provide comprehensive insights into the models' accuracy, explanatory power, and the quality of semantic clustering. The results are summarized in **Table 1** and **Table 2** below.

Table 1: Regression/Deep Learning Model Performance Metrics

Model	MSE	R^2
Linear Regression (Including Artist_Popularity)	368.78	0.512
Ridge Regression (Including Artist_Popularity)	368.77	0.512
Lasso Regression (Including Artist_Popularity)	368.76	0.512
Random Forest Regressor (Including Artist_Popularity)	473.25	0.373
Gradient Boosting Regressor (Including Artist_Popularity)	380.86	0.496
Keras Sequential Model (Including Artist_Popularity)	444.00	0.412

Model	MSE	R^2
Tuned XGBoost Regressor (Including Artist_Popularity)	377.18	0.501
Tuned LightGBM Regressor (Including Artist_Popularity)	389.92	0.484
Baseline Linear Regression (Joe Beach Capital, 2023)	**883.80**	**-0.0053**
Baseline Gradient Boosting (Joe Beach Capital, 2023)	**881.82**	**-0.0021**
Baseline Random Forest (Joe Beach Capital, 2023)	**944.95**	**-0.8748**

Table 2: Clustering Model Performance Metrics

Model	Silhouette Score
Benchmark Model (Time Period Bin Only)	**0.8084 (Bad Clusters)**
sBERT Model (Time Period Bin Only)	0.4286 (Decent Clusters)
sBERT Enhanced Model (Time Period Bin + Selected Features)	0.7464 (Excellent Clusters)

Analysis of Regression Models:

The linear models—Linear, Ridge, and Lasso Regression—exhibited consistent performance, each achieving an R^2 value of approximately 0.512. This indicates that these models can explain about 51.2% of the variance in song popularity. The negligible differences in MSE and R^2 across these models suggest that regularization techniques (Ridge and Lasso) did not significantly enhance performance over basic linear regression within this context.

Baseline Models:

The baseline models derived from Joe Beach Capital's work [1] serve as a comparative benchmark for our regression analyses. Specifically:

- **Baseline Linear Regression:** Achieved an MSE of **883.80** and an R^2 of **-0.0053**, indicating poor predictive performance.
- **Baseline Gradient Boosting:** Recorded an MSE of **881.82** and an R^2 of **-0.0021**, also reflecting inadequate explanatory power.
- **Baseline Random Forest:** Exhibited the highest MSE of **944.95** and an R^2 of **-0.8748**, demonstrating significant underperformance.

These baseline models underscore the limitations of conventional regression and ensemble methods in effectively predicting song popularity within the context of the analyzed datasets.

Ensemble Models:

The Random Forest Regressor underperformed relative to other models, with an R^2 of 0.373, indicating limited explanatory power. This underperformance may be attributed to overfitting, given the relatively small size of the dataset (10,000 songs) and the high dimensionality of the feature space. In contrast, Gradient Boosting Regressors (XGBoost and LightGBM) demonstrated improved performance, with R^2 values of approximately

0.496 and 0.484, respectively. These models effectively captured non-linear relationships within the data, albeit with modest gains over linear models.

Deep Learning Models:

The Keras Sequential Model achieved an R^2 of 0.412, which, while better than the Random Forest, still lagged behind ensemble methods and linear models. The limited dataset size likely constrained the deep learning model's ability to generalize effectively. Tuned XGBoost and LightGBM regressors showed competitive performance, with the tuned XGBoost model achieving an R^2 of 0.501, marginally outperforming the linear models.

Analysis of Clustering Models:

Clustering performance was assessed using the Silhouette Score, which measures the cohesion and separation of clusters. A higher Silhouette Score indicates better-defined clusters.

- **Benchmark Model (Time Period Bin Only):** Our initial benchmark model, which utilized only time period binning for clustering, achieved a Silhouette Score of **0.8084**. This high score reflects well-separated and cohesive clusters based solely on temporal segmentation.
- **sBERT Model (Time Period Bin Only):** Incorporating sBERT embeddings for lyrical content while maintaining time period binning resulted in a Silhouette Score of **0.4286**. Although there is a reduction in the Silhouette Score compared to the benchmark, this model offers a more nuanced clustering by integrating semantic information from lyrics, allowing for deeper insights into lyrical themes across time periods.
- **sBERT Enhanced Model (Time Period Bin + Selected Features):** Enhancing the sBERT model by integrating selected audio features alongside time period binning and lyrical semantics yielded a Silhouette Score of **0.7464**. This improvement over the sBERT model without enhanced features indicates that combining audio characteristics with lyrical semantics leads to more distinct and meaningful clusters, despite not surpassing the benchmark model's score. The trade-off reflects a balance between semantic richness and cluster distinctness.

Benchmark Comparison and Improvements:

Our benchmark model, developed in our previous work, relied exclusively on time period binning and achieved a Silhouette Score of **0.8084**. The introduction of sBERT embeddings for lyrical content, while maintaining temporal bins, lowered the Silhouette Score to **0.4286**, suggesting increased complexity in clustering due to the integration of semantic data. However, the subsequent enhancement by incorporating selected audio features raised the Silhouette Score to **0.7464**, demonstrating that the addition of audio characteristics can partially mitigate the complexity introduced by lyrical semantics, leading to more balanced and interpretable clusters.

Summary:
Overall, the regression models demonstrated moderate success in explaining the variability in song popularity, with linear and ensemble models outperforming deep learning approaches. The Shapley value analysis further highlighted key features influencing popularity, guiding the feature selection process for subsequent analyses and the development of the recommendation system.
In the clustering domain, while the benchmark model achieved superior numerical performance in terms of the Silhouette Score, the sBERT-enhanced models provided more semantically meaningful clusters by integrating lyrical and audio features. This indicates that while purely temporal binning yields well-defined clusters, the incorporation of semantic and audio data enriches the clustering quality by capturing deeper relationships within the music data, albeit with a slight compromise in numerical clustering metrics.

4.2 Key Insights
Our comprehensive analysis yielded several critical insights into the dynamics of song popularity:

1. **Dominant Audio Features:**
    - **Acousticness, Loudness, and Liveliness:** These features emerged as significant predictors of song popularity. High acousticness indicates a greater presence of acoustic instruments, which may resonate with listeners seeking authenticity in music. Elevated loudness and liveliness are often associated with energetic and engaging tracks, likely contributing to higher popularity scores.
    - **Speechiness:** This feature, reflecting the presence of spoken words in a track, also influenced popularity. Songs with balanced speechiness may cater to diverse listener preferences, blending lyrical content with musicality.
2. **Temporal Dynamics:**
    - **Age of the Song:** Older songs tended to exhibit different popularity dynamics compared to newer releases. The temporal context, including historical and cultural shifts, plays a pivotal role in shaping listener preferences and, consequently, song popularity.
    - **Cyclical Trends:** The analysis suggested the presence of cyclical trends in music, where certain genres and musical attributes experience periodic resurgence, aligning with broader cultural and technological evolutions.
3. **Lyrical Content and Semantics:**
    - **Emotional Resonance:** Word clouds indicated that themes of love, life, and emotional expression are pervasive across decades, underscoring the universal appeal of emotionally resonant lyrics.

    - **Semantic Clustering:** The use of sBERT embeddings revealed that songs clustered semantically, indicating that lyrical themes significantly contribute to the grouping and popularity of songs within specific temporal contexts.
4. **Recommendation System Effectiveness:**
    - **Feature-Based Recommendations:** The recommendation system, leveraging key audio features and Shapley-identified variables, successfully identified songs with similar 'feel' and 'energy,' aligning with user expectations for genre-consistent suggestions.
    - **Semantic Recommendations:** Incorporating lyrical semantics enhanced the recommendation system's ability to suggest songs based on thematic and emotional similarities, offering a more personalized and contextually relevant user experience.
5. **Model Performance Insights:**
    - **Feature Importance:** The Shapley analysis highlighted that a combination of acoustic features and lyrical semantics significantly influences song popularity, suggesting that both quantitative and qualitative elements are integral to a track's success.
    - **Benchmark Improvement:** Our enhanced clustering model demonstrated superior clustering quality by integrating audio features with sBERT embeddings, despite a slight decrease in the Silhouette Score. This indicates that the qualitative improvements in cluster distinctness and semantic relevance outweigh the numerical decline in clustering metrics.
    - **Dataset Skewness Impact:** The pronounced left skew in popularity distribution presented challenges for model training, indicating that future studies might benefit from more balanced datasets or advanced techniques to mitigate skew-related biases.

---

5. Limitations and Challenges

Our research encountered several significant methodological challenges:

1. **Data Representation Limitations:**
    - **Selection Bias:** The datasets potentially introduced selection biases by overrepresenting certain genres or popular artists. This inherent limitation restricted the generalizability of our findings across the entire music landscape.
2. **Feature Incompleteness:**
    - **Unaccounted Variables:** Our models accounted for approximately 51.2% of the variability in song popularity, highlighting the complex, multidimensional

nature of musical success. Critical external factors remained unaccounted for, including:
    - Marketing investments
    - Social media engagement
    - Artist reputation dynamics
    - Cultural and temporal context
3. **Computational Constraints:**
    - **Resource Intensity:** Processing extensive textual and audio data required substantial computational resources and sophisticated preprocessing techniques. Limited access to high-performance computing infrastructure may have constrained the depth and breadth of our analyses.
4. **Temporal Scope and Dataset Size:**
    - **Temporal Bias:** The merged dataset exhibited a concentration of songs from the 1990s, introducing a temporal bias that may affect the detection of trends across other decades. Additionally, the dataset size, particularly post-merge (approximately 1,634 entries), may be insufficient for training complex deep learning models effectively.
5. **Model Limitations:**
    - **Handling Skewed Data:** The significant left skew in the popularity distribution posed challenges for model training and evaluation. Traditional regression models may struggle to accurately predict popularity scores in the presence of such imbalance, potentially limiting the models' predictive accuracy and generalizability.

---

6. Future Research Directions

Building upon our current findings, future research should concentrate on several promising avenues:

1. **Enhanced Recommender Systems:**
    - **Incorporating User Behavior:** Integrating user listening behaviors, preferences, and interaction data can lead to more personalized and effective recommendation algorithms. Understanding individual user profiles and contextual listening patterns can refine song suggestions to better match user tastes.
    - **Contextual Recommendations:** Developing context-aware recommendation systems that consider situational factors such as time of day, location, and activity can further enhance user satisfaction and engagement.

2. **Multimodal Feature Integration:**
    - **Expanded Data Sources:** Incorporating additional data sources, such as social media sentiment analysis, streaming platform interactions, and cross-platform popularity metrics, can provide a more holistic view of factors influencing song popularity.
    - **Cross-Modal Analysis:** Exploring the interplay between audio features, lyrical content, and visual elements (e.g., music videos, album art) can uncover complex relationships that contribute to a song's success.
3. **Advanced Modeling Techniques:**
    - **Handling Imbalanced Data:** Employing advanced techniques for handling skewed datasets, such as synthetic minority over-sampling or cost-sensitive learning, can improve model performance and predictive accuracy.
    - **Transfer Learning:** Leveraging pre-trained models and transfer learning approaches can enhance deep learning model performance, particularly when dealing with limited dataset sizes.
4. **Temporal and Cultural Contextualization:**
    - **Dynamic Modeling:** Developing models that account for temporal shifts and cultural changes over time can provide deeper insights into the evolving nature of music popularity.
    - **Cross-Cultural Studies:** Expanding the analysis to include diverse cultural contexts can reveal universal and culture-specific determinants of song popularity.
5. **Exploring Additional Musical Features:**
    - **Harmonic and Melodic Analysis:** Incorporating detailed harmonic and melodic features, such as chord progressions, key changes, and melodic complexity, can enrich the understanding of musical attributes influencing popularity.
    - **Production Quality Metrics:** Evaluating production-related features, including mixing quality, mastering effects, and instrumentation diversity, can further elucidate factors contributing to a song's appeal.
6. **Longitudinal Studies:**
    - **Trend Evolution:** Conducting longitudinal studies to track the evolution of musical trends and their impact on popularity can provide a dynamic perspective on the music industry's trajectory.
    - **Impact of Technological Advancements:** Assessing how technological innovations, such as streaming algorithms and digital distribution platforms, influence song popularity can inform industry strategies and artist approaches.

7. Conclusion

This study presents a data-driven exploration of the intricate relationships between musical characteristics and song popularity, utilizing comprehensive Spotify datasets encompassing audio features and lyrical content spanning from the 1950s to the 2010s. Through the application of advanced machine learning techniques and natural language processing methodologies, we examined the contributions of both quantitative audio attributes and qualitative lyrical themes to a song's popularity.

Our findings underscore the multifaceted nature of song popularity, highlighting that success is not solely determined by measurable audio features but also by the semantic and emotional resonance of lyrical content. Linear and ensemble regression models demonstrated moderate predictive capabilities, elucidating the significant roles of acousticness, loudness, liveliness, speechiness, age of the song, explicit content, and instrumental qualities in influencing popularity scores. The integration of sBERT embeddings further enriched our analysis, revealing that semantic similarities in lyrics contribute meaningfully to song clustering and recommendation efficacy.

The development of a recommendation system based on these insights showcases the practical applicability of our research, offering a tool that aligns closely with user preferences by considering both musical and lyrical dimensions. Additionally, the temporal analysis of lyrical themes through word clouds and clustering techniques provided a nuanced understanding of how cultural and emotional factors evolve and influence musical trends over decades.

Despite the promising results, the study acknowledges inherent limitations, including data representation biases, feature incompleteness, and computational constraints. These challenges highlight the necessity for future research to adopt more comprehensive data sources, advanced modeling techniques, and broader contextual considerations to fully unravel the complexities of musical success.

In conclusion, this research contributes to the growing body of knowledge in music analytics by elucidating the interplay between audio features, lyrical content, and temporal patterns in shaping song popularity. The insights derived offer valuable implications for artists, producers, and industry stakeholders aiming to optimize their creative and commercial strategies in an ever-evolving musical landscape.