EXECUTIVE SUMMARY

Music Streams Prediction Model

As competition for listener attention intensifies in the expanding streaming market, the music industry seeks precise insights into what drives a song's popularity. Traditional promotional methods often rely on subjective assessments, which may not fully capture audience preferences. However, a data-driven predictive model offers a strategic approach to understanding and prioritizing the key factors influencing song success. By aligning production, marketing, and engagement strategies with these insights, industry stakeholders can make informed decisions that more effectively capture audience interest.

This report investigates the factors contributing to a song's popularity on Spotify, aiming to develop a robust predictive model for streaming success. By analyzing song attributes and engagement metrics across Spotify and YouTube, we derive actionable insights to support informed decisions in music production, marketing, and platform management.

Data Overview

Dataset Composition
The dataset includes over 20,000 records, capturing 27 key variables such as:

- **Musical Attributes**: Variables like danceability, acousticness, liveness, and speechiness reflect the intrinsic qualities of the music.
- **Engagement Metrics**: Views, likes, and comments on YouTube represent the social interaction surrounding each song.
- **Technical and Social Attributes**: Categorical variables such as "Licensed" and "Album_single" reflect production and licensing status, influencing a song's visibility and accessibility on streaming platforms.

Appendix: Variable Description Table

Data Preprocessing
Data preparation involved handling missing values, applying log transformations to highly skewed variables (e.g., views, likes, and comments), and standardizing continuous features to ensure comparability. Feature selection was executed through methods like Boruta SHAP and Variance Inflation Factor (VIF) analysis, refining the model by reducing multicollinear and low-impact variables.

Methodology

Model Formulation
We built a regression model to predict log_Stream, the log-transformed count of song streams, as our dependent variable. Logarithmic transformation allowed us to normalize the data, accommodating skewed distributions in variables like views and comments, and provided a more stable model by reducing variance.

Incorporating Non-Linear and Interaction Terms
To capture the intricate relationships within the data, we introduced non-linear elements, including quadratic and interaction terms, allowing for more nuanced modeling. For instance:

- **Quadratic Terms**: Features such as log_Duration_ms^2 and log_Comments^2 were included to account for non-linear effects, capturing the diminishing or increasing returns on streams for extreme values of song duration and comment engagement.
- **Interaction Terms**: Terms like log_Duration_ms * Liveness and log_Comments * Licensed highlight how combinations of specific features amplify or mitigate their effects on stream counts. For example, songs with high liveness and optimal duration exhibit a stronger impact on popularity, suggesting a synergistic effect.

Model Equation:

$$\begin{aligned} log\_Stream = {} & \mathbf{0.0634} - \mathbf{0.0191} \times Acousticness - \mathbf{0.0278} \times Liveness - \mathbf{0.0674} \times Speechiness \\ & - \mathbf{0.0131} \times Instrumentalness\_logit - \mathbf{0.1284} \times Licensed \\ & - \mathbf{0.0527} \times log\_Duration\_ms - \mathbf{0.0202} \times Valence + \mathbf{0.5866} \times log\_Comments \\ & - \mathbf{0.3604} \times Album\_single + \mathbf{0.0035} \times (log\_Duration\_ms \times Liveness) \\ & + \mathbf{0.1130} \times (log\_Comments \times Licensed) - \mathbf{0.0088} \times (Album\_single \times Speechiness) \\ & - \mathbf{0.0230} \times (log\_Duration\_ms2) + \mathbf{0.1599} \times (log\_Comments2) \end{aligned}$$

Model Evaluation and Optimization

Several diagnostic and evaluation steps were performed to validate the model and ensure its robustness:

- **Residual Analysis**: We generated residual plots to assess the model's fit and verify that residuals were randomly dispersed around zero, confirming that the model adequately captured the patterns in the data.

- **Breusch-Pagan Test for Heteroskedasticity**: This test indicated the presence of heteroskedasticity, suggesting variance in the errors was not constant across levels of the independent variables. This finding highlighted areas where further model refinement could improve reliability.

- **Ramsey RESET Test**: To check for model specification errors, we used the Ramsey RESET test, which suggested the presence of omitted non-linear relationships, indicating that additional non-linear transformations or interaction terms could be beneficial.

- **Cook’s Distance Analysis**: We used Cook's Distance to identify influential points that may unduly impact the regression coefficients. Points with high Cook’s Distance values were reviewed to ensure they did not disproportionately affect model stability. This step helped optimize the model by minimizing the influence of potential outliers.

These evaluation and optimization techniques allowed us to refine the model to ensure that it provided a robust framework for predicting song popularity, while also indicating areas for future enhancements based on the observed non-linear effects and interactions.

Key Findings and Insights

1. **Social Engagement is a Strong Predictor**
   The analysis reveals that the number of comments (log_Comments) significantly correlates with stream counts, suggesting that user engagement drives popularity. Encouraging comments and social interactions around a song could be an effective strategy to boost its reach.

2. **Optimal Duration and Acoustic Qualities**
   Moderate song durations and favorable values of acoustic features like danceability and valence were shown to contribute positively to popularity. For example, songs with a balance of danceability and positivity appeal to listeners, aligning with established trends favoring high-energy, upbeat music.

3. **Impact of Complex Interactions**
   The inclusion of interaction terms, such as log_Duration_ms * Liveness, allowed us to capture how certain features work in tandem to influence stream counts. For instance, high liveness combined with optimal duration enhances a song's appeal, suggesting a compounded effect on listener engagement.

4. **Non-Linear Relationships in Popularity**
   Quadratic terms in the model reveal a non-linear impact of attributes like duration and comment frequency. For example, moderate values of log_Duration_ms correlate with higher stream counts, while excessive values lead to diminishing returns. This insight underscores the importance of balancing attributes to maximize listener retention without causing fatigue.

Business Implications

Marketing and Promotion
Focus on promoting songs that show high engagement potential by creating campaigns that encourage listeners to comment, share, and engage with the track on social media. Run targeted ads that highlight these songs and leverage influencers or fan communities to boost interaction. Increased social engagement can significantly enhance visibility, driving more listeners to stream the track and increasing its chance of success.

Production Strategy
Artists and producers can align their music with listener preferences by optimizing track length and ensuring a balanced mix of danceability and energy. For example, creating upbeat, danceable songs of moderate length can improve listener retention and appeal. Additionally, using these insights early in the production process allows music labels to tailor tracks for maximum engagement, ultimately leading to stronger streaming performance and audience reach.

Platform Curation and Personalization
Streaming platforms can improve playlist effectiveness by curating songs that match user tastes, particularly those with high energy, engaging qualities, and social appeal. For instance, playlists could prioritize tracks with high engagement potential or songs that fit trending musical attributes. This personalized approach can enhance user satisfaction, boost playlist interaction rates, and increase overall listener retention, as audiences find content more aligned with their interests.

Conclusion

The Music Streams Prediction Model provides critical insights into the dynamics driving song popularity, offering a data-backed foundation for strategic decisions across the music industry. By uncovering key predictors—such as listener engagement metrics, song attributes, and unique interactions among features—this model enables stakeholders to navigate the complexities of music streaming with greater precision.

Beyond mere predictions, this model shines a light on nuanced audience behaviors, revealing how factors like song duration, social engagement, and acoustic elements impact a song's success. For instance, the inclusion of non-linear and interaction terms allows the model to capture how optimal combinations of attributes, such as duration and liveness, significantly boost stream counts, helping artists, producers, and marketers fine-tune content to resonate with listeners.

This approach not only enhances playlist curation and marketing strategies but also equips music platforms with actionable data to drive personalized recommendations, increase listener retention, and foster loyalty. In a competitive streaming landscape, where attention spans are short and choices are vast, understanding these relationships can help stakeholders craft content and campaigns that capture and sustain listener interest.

Looking ahead, the model paves the way for continuous innovation by incorporating more complex techniques, such as machine learning and real-time analysis, that could further refine predictions and adapt to evolving listener trends. Ultimately, this model empowers industry players to align creative and business goals with data insights, ensuring that they remain agile and responsive in an ever-changing digital music ecosystem.

APPENDIX

Variable	Description
log_Stream	Log-transformed count of song streams, used as the dependent variable to predict song popularity.
Acousticness	A measure of how acoustic the track is, with higher values indicating a more acoustic sound.
Liveness	Reflects the presence of an audience in the recording. Higher values indicate a live performance feel.
Speechiness	Detects the presence of spoken words in a track. Higher values suggest that the track is more speech-like.
Instrumentalness_logit	Logit-transformed value of instrumentalness, with higher values indicating less vocal content in the track.
Licensed	Binary variable indicating if the track is licensed for use on the platform (1 = licensed, 0 = not licensed).
log_Duration_ms	Log-transformed duration of the track in milliseconds, capturing the length of the song.
Valence	Describes the musical positiveness of the track, with higher values indicating a more positive mood.
log_Comments	Log-transformed count of comments on the track, representing listener engagement and interaction.
Album_single	Binary variable indicating if the track is a single (1 = single, 0 = not a single).
log_Duration_ms * Liveness	Interaction term capturing the combined effect of track duration and liveness on stream counts.
log_Comments * Licensed	Interaction term representing the impact of comment count when a track is licensed.
Album_single * Speechiness	Interaction term showing the effect of a track being a single combined with its speech-like characteristics.
log_Duration_ms^2	Quadratic term for log_Duration_ms, capturing non-linear effects of song duration on stream counts.
log_Comments^2	Quadratic term for log_Comments, capturing non-linear effects of comment engagement on stream counts.