Final Paper

**St. John’s University**

**LightGBM Classifier with Polynomial Features for Hit Song Prediction**

*Analysis and Discussion of Spotify Audio Characteristics*

**Sebastian Torres, Payal Moorti, Jomo Smith**

**Data Mining 2**

**Professor Parisi**

**4/23/2025**

## **Abstract**

This study investigates whether audio characteristics alone can predict commercial success in popular music, combining polynomial feature expansion with a LightGBM classifier. By merging Billboard Hot 100 chart data (2010-2024) with Spotify audio features and applying a customized "hit\_score" threshold of 120 to classify songs as hits or non-hits, we evaluate the predictive power of musical attributes independent of external factors such as artist reputation or marketing. Our methodology employs second-degree polynomial feature interactions to capture complex relationships between musical elements, balanced with SMOTE-Tomek sampling to address class imbalance. The model achieves 67.8% accuracy with stronger performance in identifying hits (F1=0.76) than non-hits (F1=0.50). Analysis reveals that successful songs exhibit specific feature interactions: particularly between danceability and liveness, key and acousticness, and valence and duration. These findings suggest that while audio features alone demonstrate moderate predictive capability, the interaction between multiple musical elements, rather than any single characteristic, best explains commercial success. The results offer practical insights for music creators and industry professionals while acknowledging the inherent limitations of excluding non-musical factors from hit prediction.

## **Introduction/Background**

This study investigates whether audio characteristics alone can predict commercial success in popular music by analyzing patterns in hit songs independent of market-driven influences. Our research eliminates external variables such as artist reputation, genre classification, chart position, and listener behavior by creating a custom "hit\_score" threshold of 120 to classify songs as hits or non-hits. Our approach combines second-degree polynomial feature expansion with a LightGBM classifier to capture complex relationships between musical elements like danceability, energy, key, and valence. While previous research has shown variable success in predicting hits based solely on audio features, our study aims to determine whether feature interactions, rather than individual characteristics, better explain commercial success in music.

## **Literature Review**

The research titled "Musical trends and predictability of success in contemporary songs in and out of the top charts" analyzes over 500,00 songs released in the UK between 1985 and 2015. The researchers define a song's commercial success as making it into the top charts and correlate that with the musical attributes the song possesses. Over time, the study uncovered a downward trend in songs that exhibited "happiness" and "brightness" as well as an upward trend in "sadness"(Interiano et al., 2018). Additionally, it was noticed that songs were becoming less "male" over the decades. When explicitly examining songs that topped the charts, Interniano et al. (2018) discovered that they tend to be "happier", "party-like", less "relaxed", and more "female" than songs that were not particularly chart toppers. Specific musical attributes that successful songs contained seemed to predict future trends in all songs, while other songs reflected patterns and trends of the past (Interiano et al., 2018).

To examine the predictive power musical attributes have on a song's success, researchers implemented random forest models in two settings: the first one only using musical attributes, and the other including a "superstar" variable that indicated if the song's artist has had chart success in the past. This approach allowed researchers to determine how significant of a role purely musical attributes play in a song's success. While using only musical attributes did yield some prediction success, the results showed that the "superstar" variable did improve prediction accuracy compared to the model not using the "superstar" variable. The researchers achieved a prediction accuracy of about 74% when using solely musical attributes to predict. Adding the "superstar" variable increased prediction accuracy to about 86%, which suggests that an artist's reputation does play a role in a song's commercial success (Interiano et al., 2018).

Another study by Nicholas Borg and George Hokkanen titled " What Makes for a Hit Pop Song? What Makes for a Pop Song?" investigates the possibility of developing a hit song prediction algorithm. The researchers used the Million Song Dataset subset, which contains pre-extracted features for 10,000 songs, and accompanied this with YouTube view counts to measure popularity (Borg & Hokkanen, n.d.). Their methodology included several models, primarily Support Machine Vectors with varying feature vector lengths (Borg & Hokkanen, n.d.). Trying to predict song popularity based solely on musical attributes did not result in much success. Prediction accuracy was never able to exceed 55% using musical attributes alone, which deviates significantly from the findings of the study performed by Interiano et al. (2018), which resulted in more promising results in musical attributes' capability to predict a song's success. Borg and Hokkanen (n.d..) emphasized the "inherent unpredictability of cultural markets" as the main limitation. When they shifted genre classification, they tested multiple models such as SVMs, Random Forests, and K Nearest Neighbors. Random Forests achieved the best ten-genre classification with 56.8% accuracy and over 83% accuracy with both Random Forest and K-Means with SVM for four-genre classification (Borg & Hokkanen, n.d.). The contrast in performance between genre classification and hit prediction led to the conclusion that audio features can differentiate music styles but lack predictive power for commercial success.

The study by Reisz, Servedio, and Thurner (2024) titled "Quantifying the impact of homophily and influencer networks on song popularity prediction" explores the role of homophily and influencer networks in predicting song popularity. This study examines how social influence can determine a song's success. The researchers analyzed a dataset from last.fm, which included 300 million listening events and a network of 2.7 million users. They identify a strong presence of musical homophily, where users and their friends share similar music preferences, and some users significantly impact their friends' listening behaviors. This study introduces an influence score based on these interactions and evaluates its predictive power in machine learning models (Reisz et al., 2024). To assess the predictive strength of social influence, the researchers developed three machine learning models: one using traditional factors such as artist popularity and song metadata, another using social network-based parameters, and a third combining both approaches. The results show that incorporating homophily-based influence scores improves hit song predictions. While conventional factors like an artist's past popularity remain influential, the findings suggest that social dynamics are crucial in determining a song's commercial success (Reisz et al., 2024). The research suggests that integrating social influence data could optimize music marketing and recommendation algorithms.

In the "Hit Song Science is not yet a Science" research, Pachet and Roy (2008) investigate the possibility of predicting a song's popularity using machine-learning techniques applied to musical and human-generated features (Pachet & Roy, 2008). The researchers constructed a dataset of 32,000 songs with detailed metadata and utilized three different feature sets: generic acoustic features, specific acoustic features, and human-annotated metadata. Their methodology involved training classifiers, primarily Support Vector Machines (SVMs), on these feature sets and testing them to determine whether song popularity could be accurately predicted. The results showed that while some subjective musical attributes, such as mood, could be reasonably learned by classifiers, song popularity remained largely unpredictable (Pachet & Roy, 2008). Even when refining feature selection and increasing training dataset sizes, the models failed to capture meaningful statistical patterns linking musical attributes to popularity. The researchers also trained classifiers using human-generated metadata, excluding popularity labels (Pachet & Roy, 2008), to validate their findings further. While certain labels showed high correlations and were accurately modeled, popularity remained an outlier, showing no apparent relationship with musical characteristics. These findings suggest that while audio features can effectively categorize musical attributes, they do not possess predictive power for commercial success.

## **Methodology**

This project aims to determine whether a machine learning model could accurately predict whether a song would become a hit based solely on its audio features, excluding metadata such as artist identity, genre, chart position, or listener behavior. By focusing exclusively on musical attributes, the study aims to uncover intrinsic patterns in hit songs that go beyond market-driven influences.

The data collection process for this project begins with web scraping the Billboard Hot 100 charts from January 1st, 2010, to December 31st, 2024, using Python's requests and BeautifulSoup libraries. This involves iterating through weekly charts to extract rankings, artist names, and other data, which is then structured and stored using pandas. The extracted data is stored in CSV format to create a record of song popularity. Another feature added to this dataset is the custom "hit\_score" derived from an algorithm that considers the number of weeks the song remained on the chart and its peak position. Songs with scores greater than or equal to 120 are classified as hits, establishing a benchmark for song success in the analysis.

To examine the factors contributing to a song's popularity, the Billboard dataset will be merged with Spotify audio features, including tempo, danceability, valence, and instrumentality. The merge will be performed using pandas by cross-referencing songs based on artist names and track titles, with preprocessing techniques such as case normalization, whitespace trimming, and fuzzy matching applied to ensure accurate joins between datasets in case of variations in formatting. The merged dataset will then go under machine learning classification to determine which audio features and artist characteristics most strongly correlate with chart success. The approach will utilize correlation analysis, principal component analysis, and feature importance ranking derived from machine learning models like Random Forest and Gradient Boosting.

This methodology builds upon prior research as it aims to identify whether specific audio characteristics consistently appear in hit songs based on the "hit\_score" feature and if there are notable trends in these characteristics over time, potentially offering insights into evolving musical preferences.

The dataset used was created by merging Billboard song chart performance, specifically the "Hot 100" chart over the span of 15 years (2010-2024), with Spotify audio feature data. The target variable, "is\_hit", was defined as a binary label indicating whether a song had a "hit\_score" value of over 120 (1) or under 120 (0). Twelve audio-based variables were chosen as input features: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration (ms) (Eishkaran, 2021, adapted).

To capture complex relationships between musical elements, second-degree polynomial feature expansion was applied using PolynomialFeatures(degree=2). This allowed the model to learn interactions such as danceability × liveness, key × acousticness, and squared terms like valence². Feature expansion enriched the feature space, allowing the model to identify complex, non-additive patterns that might distinguish hits from non-hits.

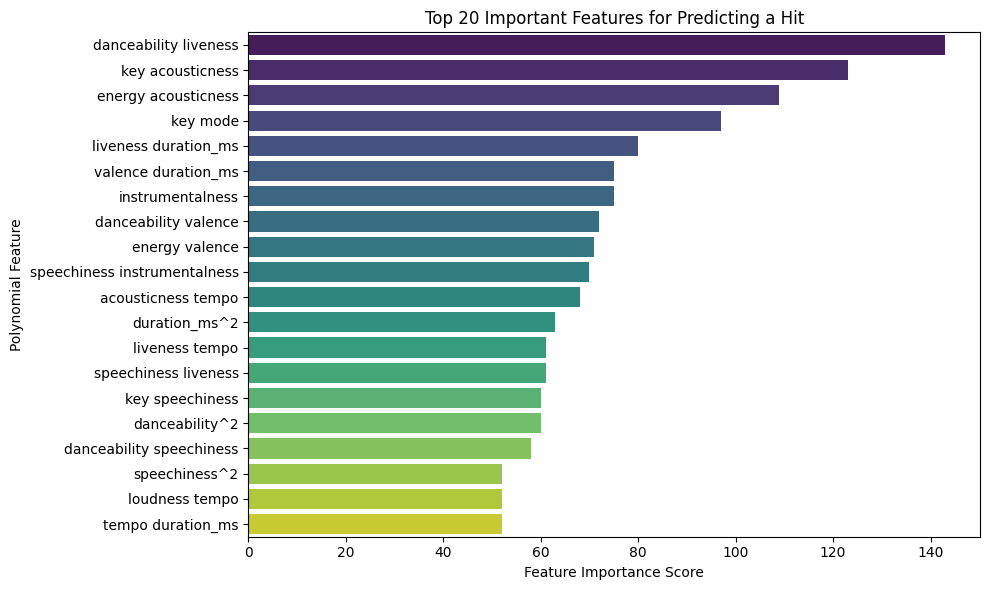
Given the natural imbalance between hit and non-hit songs, we employed SMOTE-Tomek, an integrated oversampling and undersampling technique. SMOTE generated synthetic examples for the minority class (non-hits), while Tomek links were used to remove ambiguous overlapping data points, improving class separation (Brownlee, 2020).

A LightGBM classifier with structured data was chosen for its efficiency. To reduce bias, the model was configured with 300 estimators, a maximum depth of 5, a learning rate of 0.03, and balanced class weights. The decision threshold was tuned to enhance classification performance by evaluating macro-averaged F1 scores across a range of thresholds (0.1 to 0.9). The threshold maximizing the macro F1 score was selected, ensuring balanced performance for both hit and non-hit predictions.

Model performance was evaluated using the following metrics: Accuracy, F1 score, ROC AUC, and PR AUC. A detailed classification report provided each class's precision, recall, and F1 scores. Visualizations such as a confusion matrix, ROC curve, and precision-recall curve were generated to interpret the model's behavior. Additionally, feature importance analysis highlighted the most influential polynomial features, revealing which musical attributes and their interactions were most predictive of a song's hit status.

## **Analysis and Results**

### **Analysis: Top Polynomial Feature Combinations Predicting a Hit**



The following are the most significant polynomial feature interactions identified in the model. These combinations offer insights into musical elements that drive listener engagement and commercial success:

Danceability × Liveness: Songs that are both highly danceable and possess a "live" feel are more likely to resonate with listeners and become hits.

Key × Acousticness: Specific musical keys may evoke emotional responses that increase a track's appeal when paired with acoustic textures.

Energy × Acousticness: The contrast between energetic and acoustic elements, such as in "indie pop with drive," can make songs more compelling.

Key × Mode: Tonal structure, particularly the interplay between major and minor modes in specific keys, plays a significant role in listener preference.

Valence × Duration (ms): Shorter songs with a positive emotional tone (valence) tend to perform better, possibly due to repeatability and listener mood alignment.

Danceability × Valence: Joyful, danceable music has more potential for virality and mass appeal.

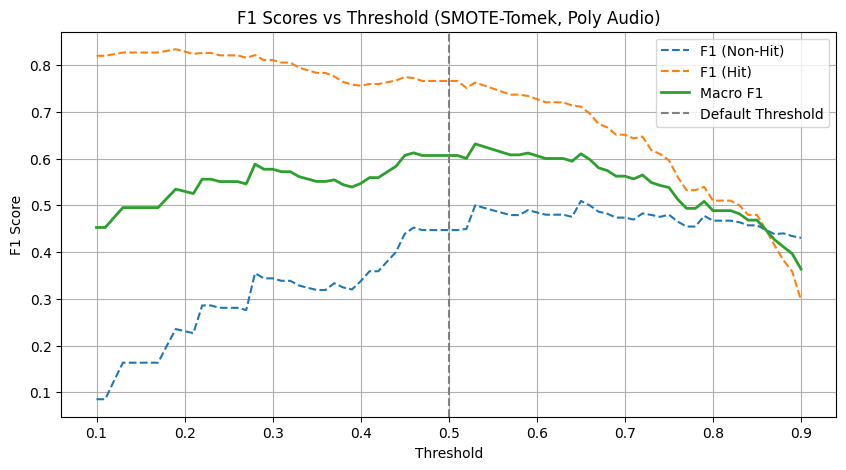
Instrumentalness, Speechiness: While lyrics and vocals matter, their interaction with instrumental texture is critical in defining a hit.

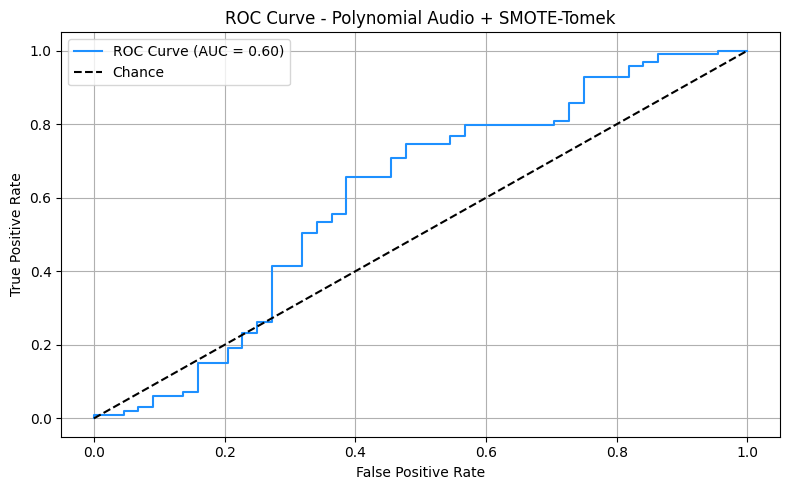
Duration², Tempo × Duration: Song pacing and length exhibit nonlinear effects — tracks that are too slow or too long may underperform compared to those with optimal pacing.

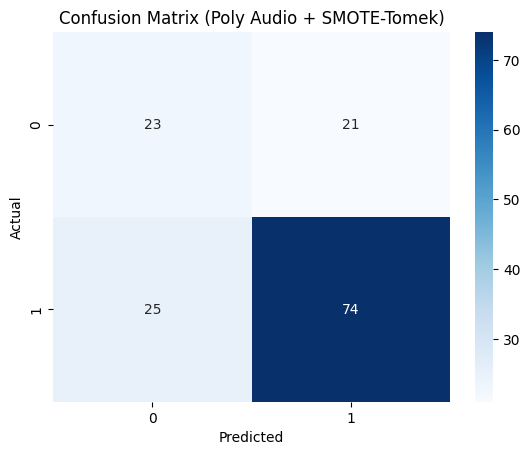
### **Results**

| Metric | Value |
| --- | --- |
| Accuracy | 0.678 |
| Hit F1 Score | 0.76 |
| Non-Hit F1 Score | 0.50 |
| Macro F1 | 0.63 |
| ROC AUC | 0.602 |
| PR AUC | 0.716 |
| Best Threshold | 0.53 |

### The model achieved moderate predictive accuracy (67.8%), correctly classifying both hit and non-hit songs. However, given the dataset's class imbalance, accuracy alone proved insufficient for comprehensive evaluation. The F1 scores revealed an imbalanced performance for hits and non-hits: while hits were identified effectively (F1 = 0.76), non-hit classification proved more challenging (F1 = 0.50), reflecting both the diversity of unsuccessful songs and potential limitations in feature representation. The macro F1 score (0.63) demonstrated reasonable balance between classes. Discrimination capability, measured by ROC AUC (0.602), exceeded chance-level performance (AUC > 0.5) but left room for improvement. Providing more evidence, the PR AUC (0.716) confirmed strong precision-recall characteristics for hit prediction—a critical consideration given the imbalanced nature of musical success. Optimal model performance was achieved through threshold tuning (0.53), which maximized the macro F1 score by strategically balancing sensitivity and precision across both classes.







## 

## 

## **Conclusion/Discussion**

### **Model Strengths**

This version of the model focuses exclusively on musical and audio-based data, avoiding potential biases introduced by external factors such as artist popularity, record label affiliation, or social media metrics. To address the inherent class imbalance in hit versus non-hit data, the model employs SMOTE-Tomek. This hybrid technique balances the dataset by oversampling the minority class and cleaning overlapping examples (Brownlee, 2020). A calibrated decision threshold is also selected to ensure balanced performance across both classes, rather than optimizing solely for accuracy or favoring one outcome. The model demonstrates strong generalization capabilities, avoiding overfitting and learning meaningful distinctions between hit songs and hits and non-hits. This comprehensive learning approach enables more reliable predictions across various musical inputs.

### **Limitations**

While the model provides valuable insights into hit songs' structural and audio characteristics, several limitations should be acknowledged. First, music is inherently subjective; factors such as listener preferences, cultural trends, timing, and an artist's reputation significantly influence a song's success, yet these variables are not captured in the dataset. Additionally, the Spotify audio features used in this study are aggregated as averages across the track and do not account for dynamic changes, lyrical content, or the evolution of sonic elements throughout a song. The model also does not incorporate raw audio data, thereby missing critical aspects such as melodic motifs, vocal performance nuances, and production quality, often essential elements in differentiating a hit. Moreover, the definition of a "hit" is treated as a binary classification, which may oversimplify the spectrum of commercial success and overlook songs that perform well without reaching top-tier chart status. Finally, the model is constrained by the available Spotify feature set, which lacks information such as genre, release timing, playlist placement, and marketing efforts, which can significantly affect a song's popularity.

### **Discussion**

For musicians looking to create commercially successful tracks, the findings suggest that no feature, such as energy or danceability, drives a song's popularity, but rather the interaction between multiple musical attributes. Songs that are shorter in duration, emotionally upbeat, highly danceable, and incorporate "live" performance elements show the strongest correlation with hit status. Emotional depth and tonal complexity, particularly the interplay between key and mode, also emerge as influential factors. These results imply that hit songs are likelier to evoke strong musical emotions and exhibit a nuanced tonal structure than non-hits.

This model is a practical tool for musicians, producers, and industry professionals to assess a song's "hit potential" based on its structural and audio features. Identifying key feature interactions—such as danceability, valence, and tonal complexity—provides actionable insights into the musical elements most associated with commercial success. The model could evolve into a comprehensive predictive system with further refinements, such as integrating Mel-frequency cepstral coefficients (MFCCs), lyrical sentiment analysis, or genre-specific variables. Such a tool would be highly valuable for A&R teams, independent artists, and producers, offering data-driven guidance during the creative and selection processes in music production and distribution.

**References**

Brownlee, J. (2020, January 17). Imbalanced classification in Python: SMOTE-Tomek links method. Towards Data Science.<https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc>

Borg, N., & Hokkanen, G. (n.d.). What makes for a hit pop song? What makes for a pop song? Retrieved from<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d40feafdf2fe60f942873d1b295db8f4d28930fa>

Eishkaran. (2021). Spotify music recommendation system dataset [Data set; adapted]. Kaggle.<https://www.kaggle.com/code/eishkaran/spotify-music-recommendation-system/input> Note: Modified by Jomo Smith, Sebastian Torres, Payal Moorti by merging with Billboard chart data.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. Royal Society Open Science, 5(5).<https://doi.org/10.1098/rsos.171274>

Pacher, F., & Roy, P. (2008). Hit song science is not yet a science. ResearchGate. Retrieved from<https://www.researchgate.net/publication/220723429_Hit_Song_Science_Is_Not_Yet_a_Science>

Reisz, N. V., & Thurner, S. (2024). Quantifying the impact of homophily and influencer networks on song popularity prediction. Scientific Reports, 14(1).<https://doi.org/10.1038/s41598-024-58969-w>