Spotify Hit Predictor

Arpita Jairaj Rane

Apoorv Vaishampayan

Antara More

San Diego State University

MIS 720: Electronic Business and Big Data Infrastructures

Dr. Xialu Liu

5/6/2024

Contents

Introduction

The music industry is fueled by the dynamic interplay of artistry and commerce, where the distinction between a chart-topping hit and a forgotten track can be both subtle and profound. Utilizing the "Spotify Hit Predictor Dataset" from Kaggle, this project seeks to delve into the rich tapestry of music released in 2010, examining a dataset of 6,399 tracks. Each record is a blend of audio features and metadata, offering a unique opportunity to explore the attributes that contribute to a song's commercial success or its fade into obscurity. By leveraging advanced data analytics and predictive modeling, this study aims to construct a robust predictive framework. This model will not only forecast the potential success of a song but also provide actionable insights for artists, producers, and industry stakeholders. The ultimate goal is to uncover the patterns and variables that most significantly impact a song's performance in the competitive music market.

Executive Summary

This report presents the findings from a comprehensive analysis conducted using the Spotify Hit Predictor Dataset, which comprises 6,399 records of songs released in 2010. Each track is characterized by various audio features and metadata that were methodically examined to determine their influence on the likelihood of becoming a hit. The study employed statistical techniques and machine learning models, with a particular focus on Support Vector Machines (SVM) to predict song popularity. The key predictors included variables such as danceability, energy, loudness, and valence, among others. The analysis revealed significant correlations between these features and the hit potential of a track, providing a foundation for our predictive model. This model aims to assist industry professionals by forecasting track performance, thereby informing decision-making processes in production and marketing strategies. The insights gained from this project not only enhance understanding of what musically resonates with audiences but also serve as a strategic tool in the increasingly data-driven music industry. This report encapsulates our journey through data preparation, model building, and the implications of our findings in shaping future musical successes.

Discovery and Data Preparation:

Data source:

The project uses a dataset titled "The Spotify Hit Predictor Dataset," which was sourced from Kaggle and contains various attributes of songs released in 2010.

https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset/data

The dataset contains 6,399 records of tracks released in 2010, each representing a 10-second snippet of a song. Variables include track name, artist, URI, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration, time signature, chorus hit, sections, and target indicating hit or flop.

Our dataset encompasses a comprehensive array of variables, each offering unique insights into the characteristics of the tracks under analysis:

Track Name: The title of the song. Artist: The artist(s) responsible for the track. URI: A unique identifier assigned to each track.

Danceability: A metric indicating the suitability of a track for dancing, based on factors such as rhythm, tempo, and beat stability.

Energy: A measure of the intensity and activity level of the track, often reflecting its dynamic range and overall vigor.

Key: The musical key in which the track is composed, providing insight into its tonal center and harmonic structure.

Loudness: The perceived volume or intensity of the track, typically measured in decibels (dB). Mode: Indicates whether the track is in a major or minor key, influencing its emotional character and tonal qualities.

Speechiness: Quantifies the presence of spoken words or vocal elements within the track, distinguishing between purely instrumental and vocally driven compositions.

Acousticness: Reflects the degree to which a track relies on acoustic instruments or recordings, as opposed to electronic or synthesized elements.

Instrumentalness: Measures the proportion of instrumental content in the track, with higher values indicating a greater reliance on non-vocal elements.

Liveness: Indicates the presence of live audience recordings or performances within the track, capturing the ambiance and spontaneity of live music.

Valence: Represents the musical positivity or emotional brightness of the track, ranging from melancholic and somber to uplifting and euphoric.

Tempo: Denotes the pace or speed of the track, expressed in beats per minute (BPM), influencing its rhythmic feel and energy level.

Duration: Specifies the length of the track in milliseconds, providing a measure of its temporal extent.

Time Signature: Defines the rhythmic structure and meter of the track, indicating the number of beats per measure and the note duration that receives one beat.

Chorus Hit: Indicates the position of the chorus within the track, offering insights into its structural composition and arrangement.

Sections: Represents the number of distinct musical sections or segments within the track, delineating its formal organization and structural complexity.

Target: The binary indicator distinguishing between "hit" (1) and "flop" (0) tracks, serving as the response variable for predictive modeling.

Below is the summary of Predictors and Variables:

Quantitative Variables: Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration_ms, Time Signature, Chorus_hit, Sections. Qualitative Variables: Track, Artist, URI, Key, Mode. Target Variable: Hit or Flop (1 for hit, 0 for flop).

Missing Values:

There are no missing values in the dataset.

Results and Performance

Model Selection and Diagnosis:

Based on the Exploratory Data Analysis (EDA) performed and the dataset's features, the decision to utilize a Support Vector Machine (SVM) model is well-founded due to its robustness in handling classification tasks where the boundary between classes is not immediately obvious or linear.

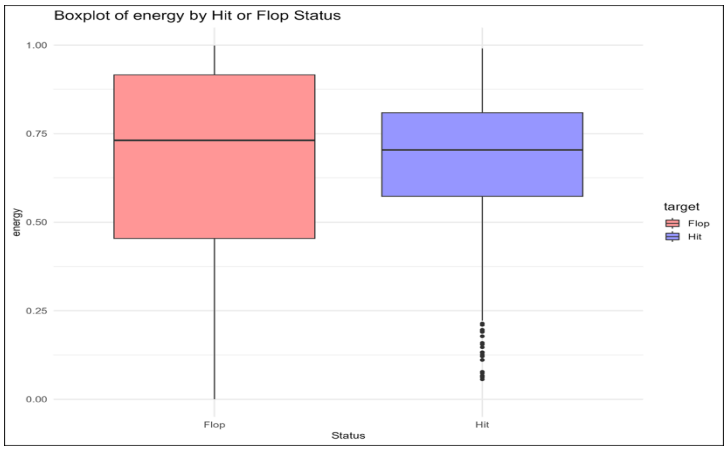
Exploratory Data Analysis (EDA) Insight:

Summary of Boxplots and Exploratory Data Analysis (EDA) :

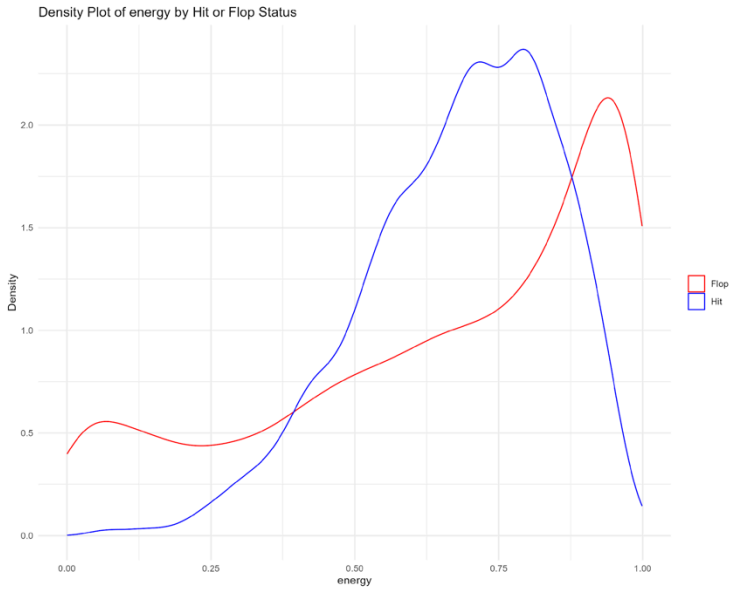
The boxplots and density plots provide a visual representation of the distribution of various musical features segmented by the track's status as a hit or flop. Here's a breakdown of each key feature based on the provided visualizations:

1. Energy:

- Boxplot Analysis: Hits generally exhibit higher energy levels than flops. The median energy for hits is noticeably higher, suggesting that more energetic tracks are more likely to be popular.

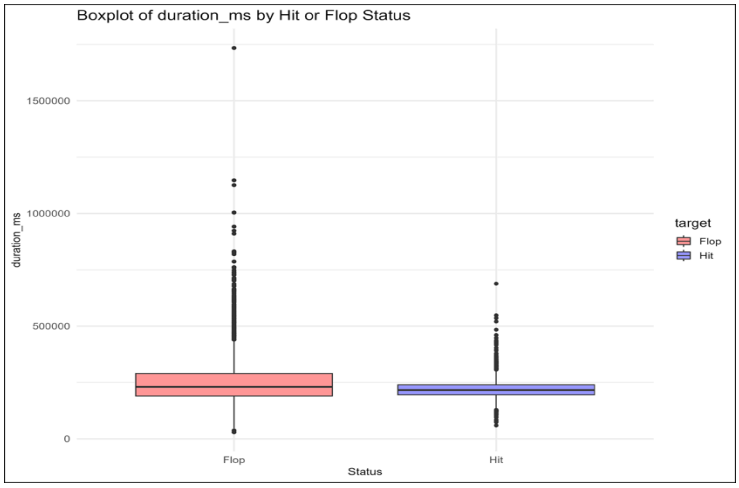


- Density Plot: The density plot confirms that the energy distribution for hits peaks higher and is more skewed towards higher values, while flops tend to have a wider spread with a peak in lower energy values.

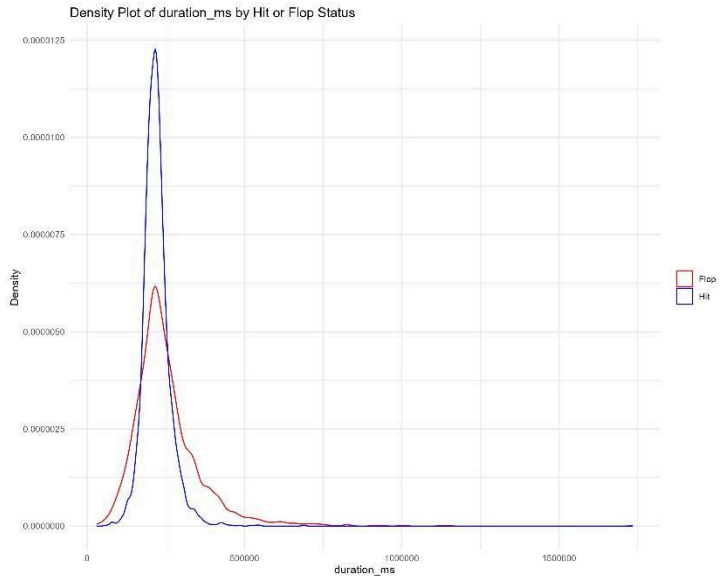


2. Duration:

- Boxplot Analysis: The duration of hits tends to be more concentrated and shorter compared to flops, which show a wider range and higher outliers. This indicates that shorter duration songs are more likely to be hits.

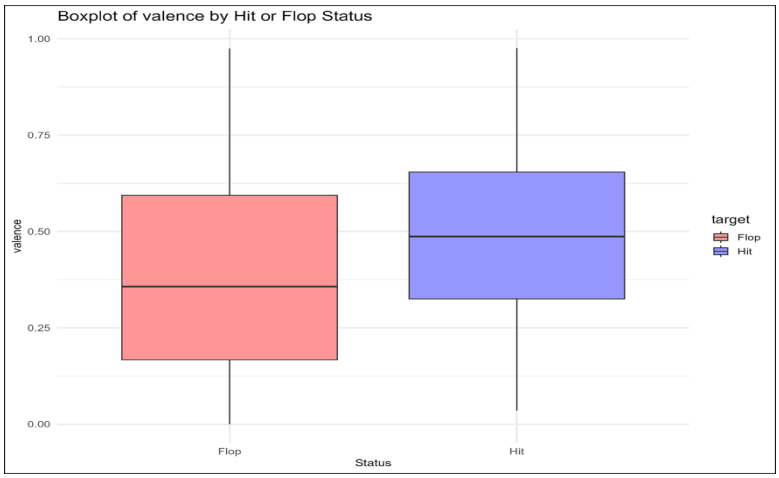


- Density Plot Analysis: The density plot highlights a sharp peak at lower durations for hit songs, suggesting a preference for shorter tracks in successful songs. Flop songs show a broader distribution with a peak at slightly higher durations, indicating longer songs are less likely to succeed commercially. The plot illustrates a clear trend: shorter songs are more likely to be hits, while longer songs are more often flops.

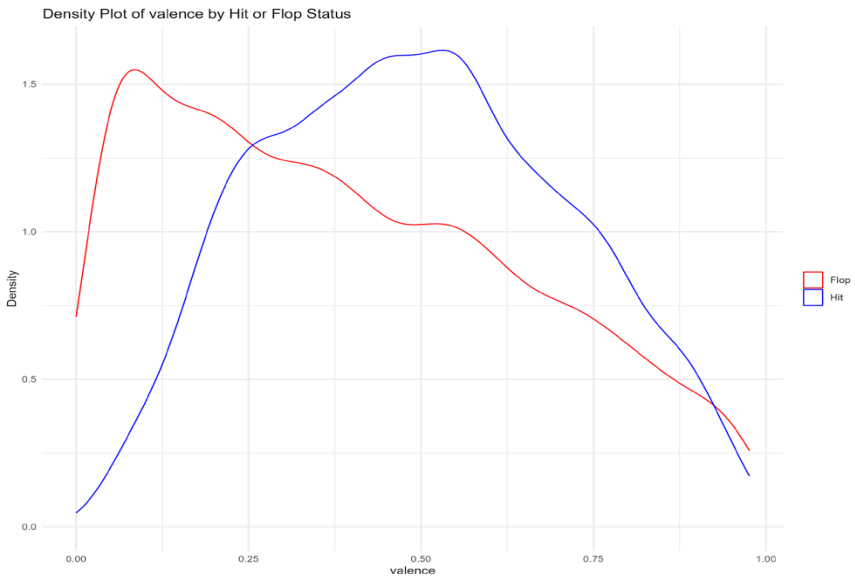


3. Valence:

- Boxplot Analysis: Hits have a higher median valence compared to flops, suggesting that songs with a happier or more positive tone are more likely to succeed.

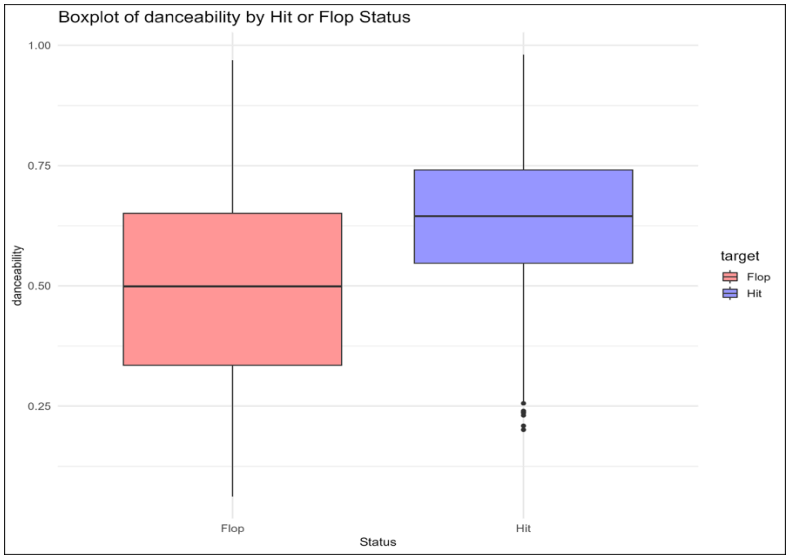


- Density Plot: The plot shows that hits have a higher peak in the positive range of valence, while flops distribute more evenly across the valence spectrum.

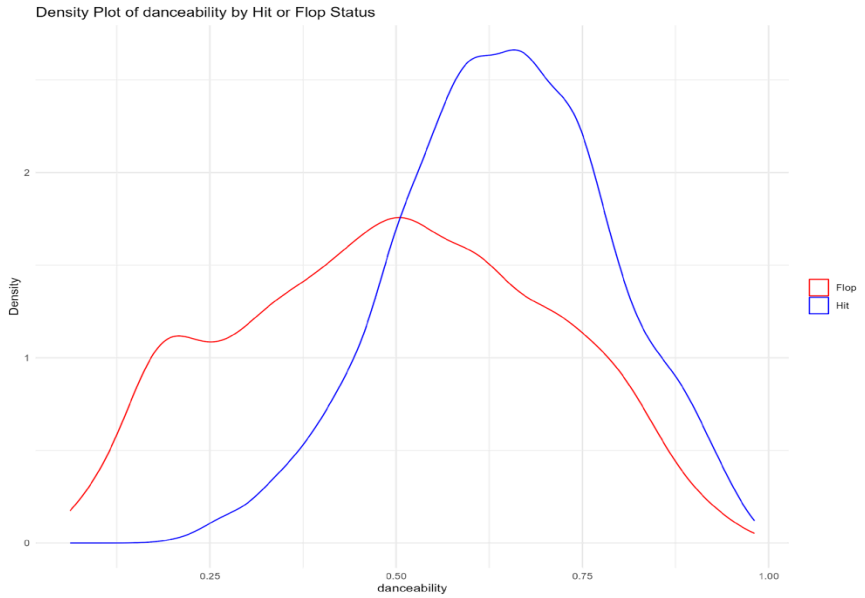


4. Danceability:

- Boxplot Analysis: Hits are generally more danceable, with a higher median and a narrower interquartile range than flops.

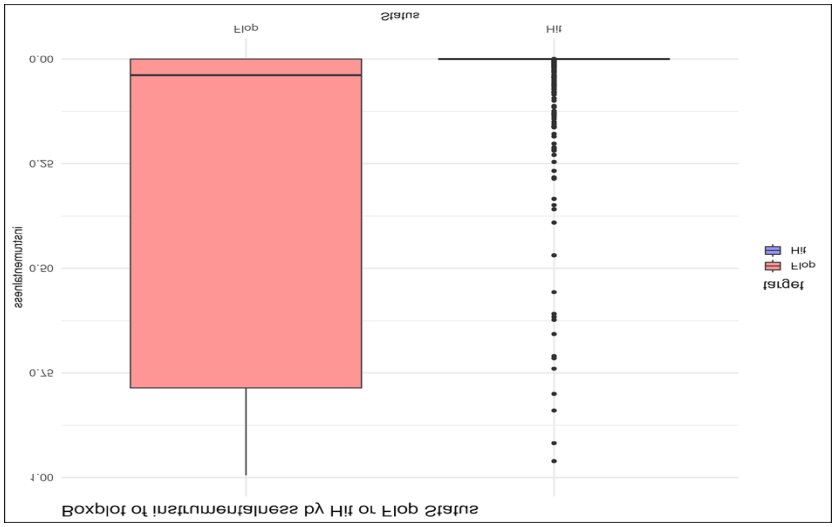


- Density Plot: The density for hits is skewed towards higher danceability, reinforcing the trend that more danceable songs are often more popular.



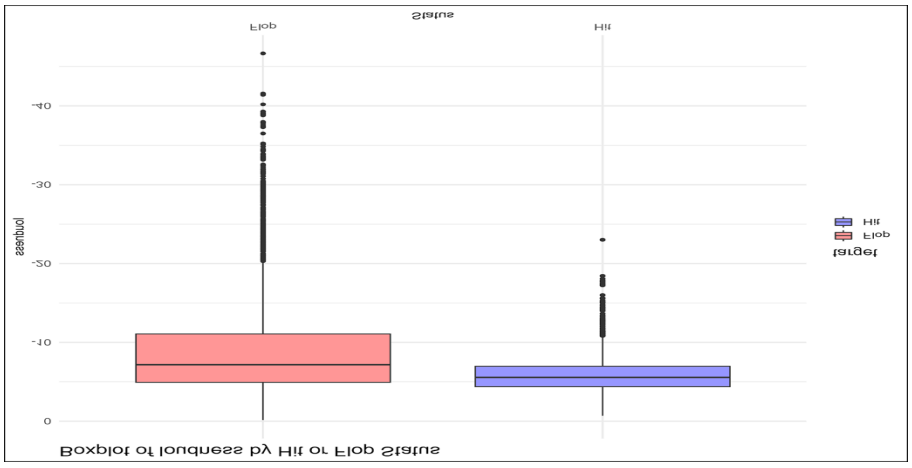
5. Instrumentalness:

- Boxplot Analysis: Flops have a higher level of instrumentalness, indicating that tracks with less vocal content are less likely to be hits.

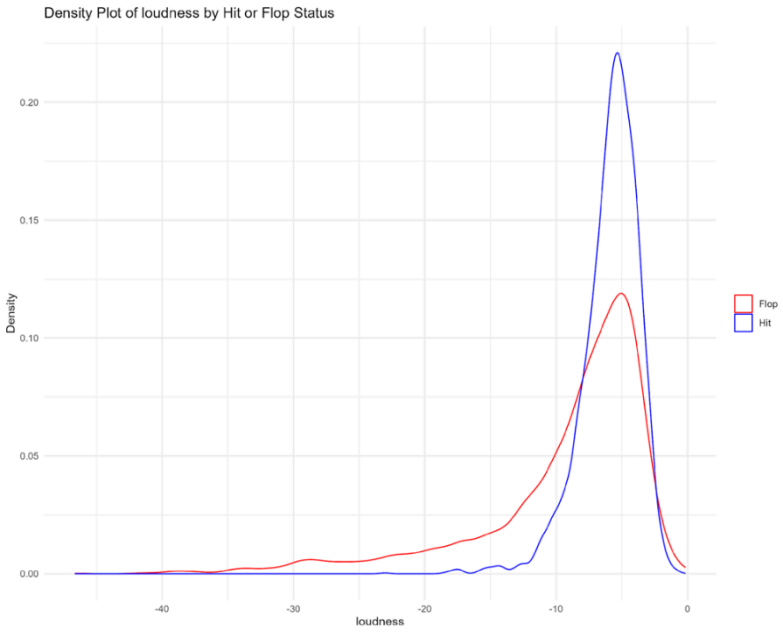


6. Loudness:

- Boxplot Analysis: Hits are louder on average than flops, with less variance in loudness levels.

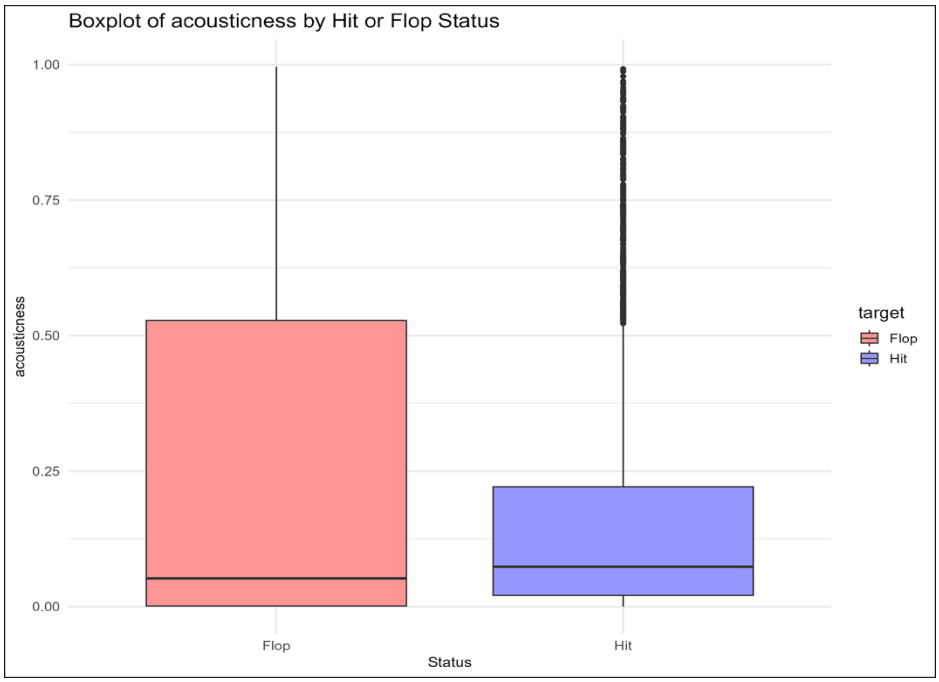


- Density Plot: Hits show a peak at higher loudness levels, while flops show a wider spread in loudness, including a number of very quiet tracks.

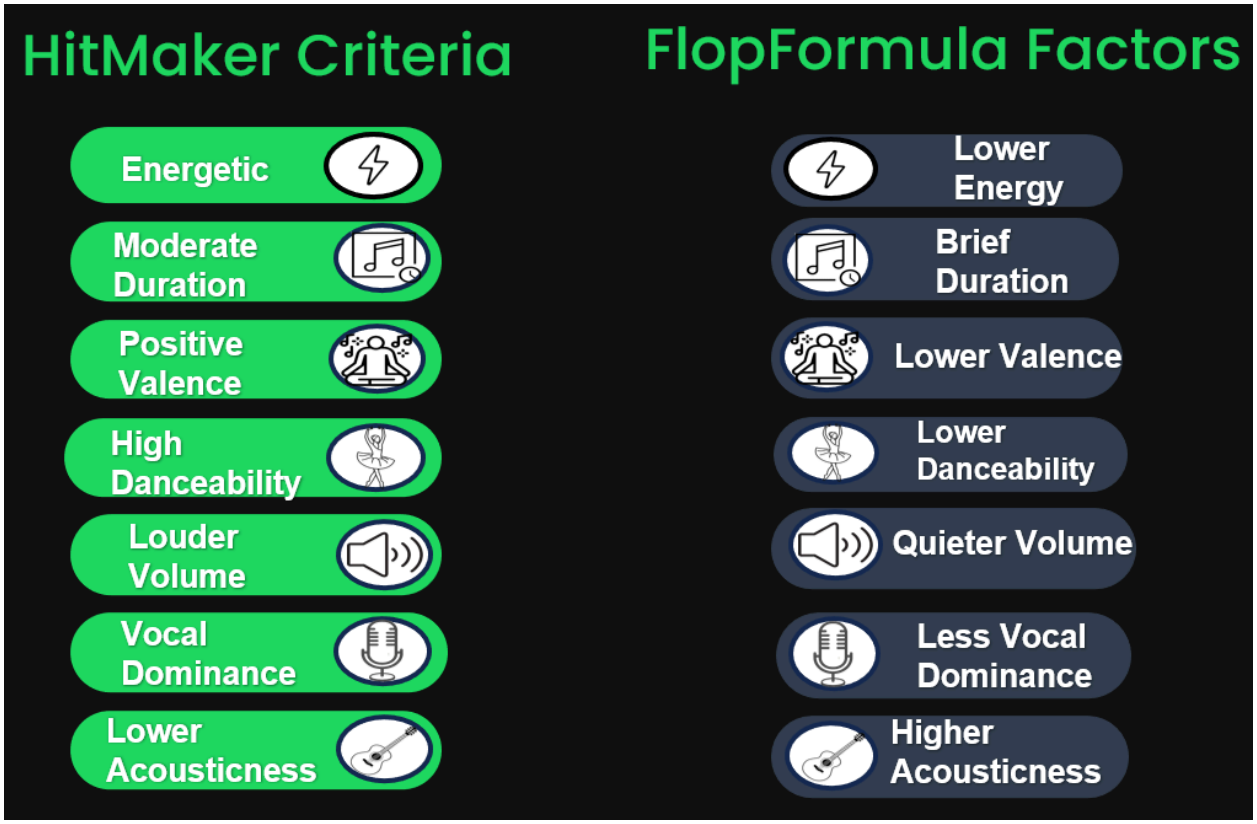


7. Acousticness:

- Boxplot Analysis: Flops tend to have a higher acousticness, indicating that songs with more acoustic elements are less likely to be popular.



So, from all the above EDA results, we zeroed down to the hit-and-flop song criteria –



Correlation Heatmap Analysis:

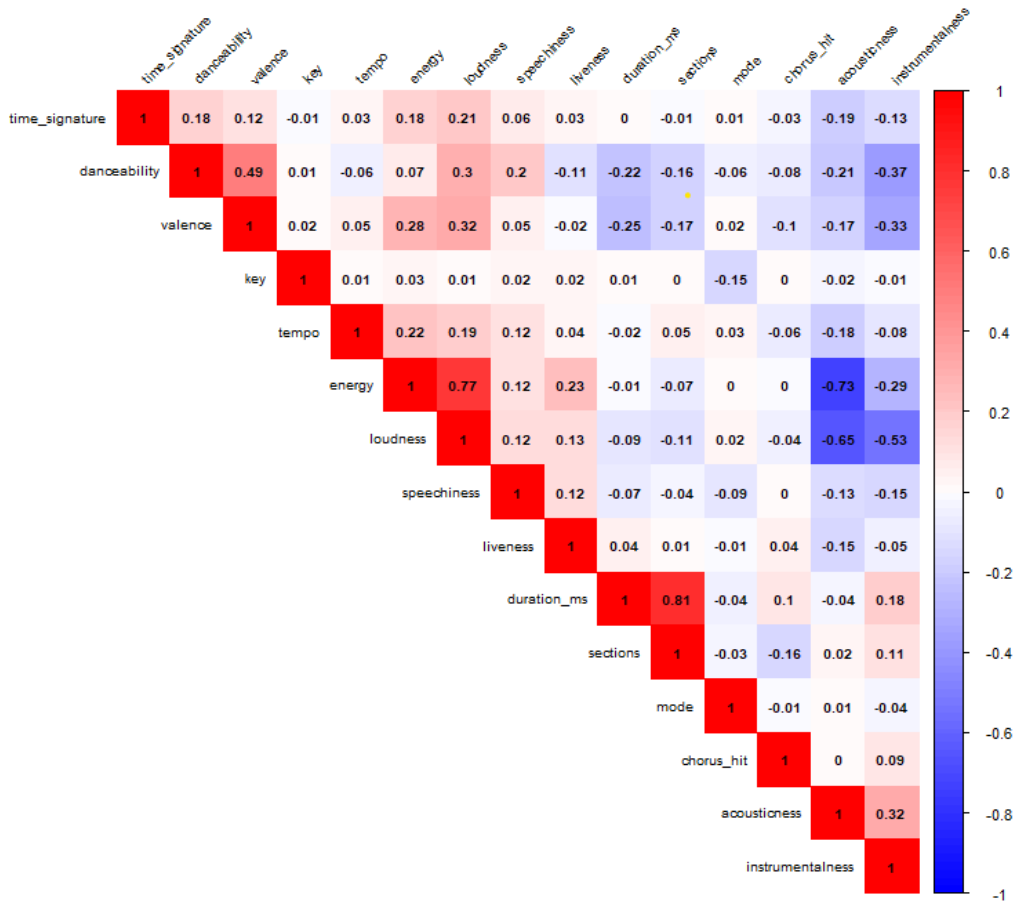
The correlation heatmap provides insight into how different musical features interact with each other:

- Positive Correlations: Energy and loudness are positively correlated, which is intuitive as louder songs tend to be more energetic.

- Negative Correlations: Acousticness is negatively correlated with energy and loudness, suggesting that more acoustic tracks are typically quieter and less energetic.

- Target Correlation: Danceability, energy, and valence show positive correlations with the target variable (hit status), reinforcing the findings from the boxplots that these features are influential in determining a song's success.

This comprehensive EDA helps identify which features are most indicative of a song's likelihood to be a hit, guiding the feature selection for predictive modeling and providing a foundation for further statistical analysis and model development.



Model Selection:

- Support Vector Machine (SVM): Given the dataset features multiple dimensions that influence a track's success, SVM is appropriate due to its capacity to handle high-dimensional space and its effectiveness in finding the optimal hyperplane that maximizes the margin between classes.

- Kernel Selection: The choice between radial, linear, and polynomial kernels allows us to experiment with how the model handles the non-linearity of the data. The kernel trick is particularly useful in mapping the input space into higher-dimensional space where a linear separator might be more effective.

- Parameter Tuning: Using techniques like grid search to optimize parameters such as C (regularization parameter), kernel coefficients, and gamma values will enhance the model's ability to generalize without overfitting.

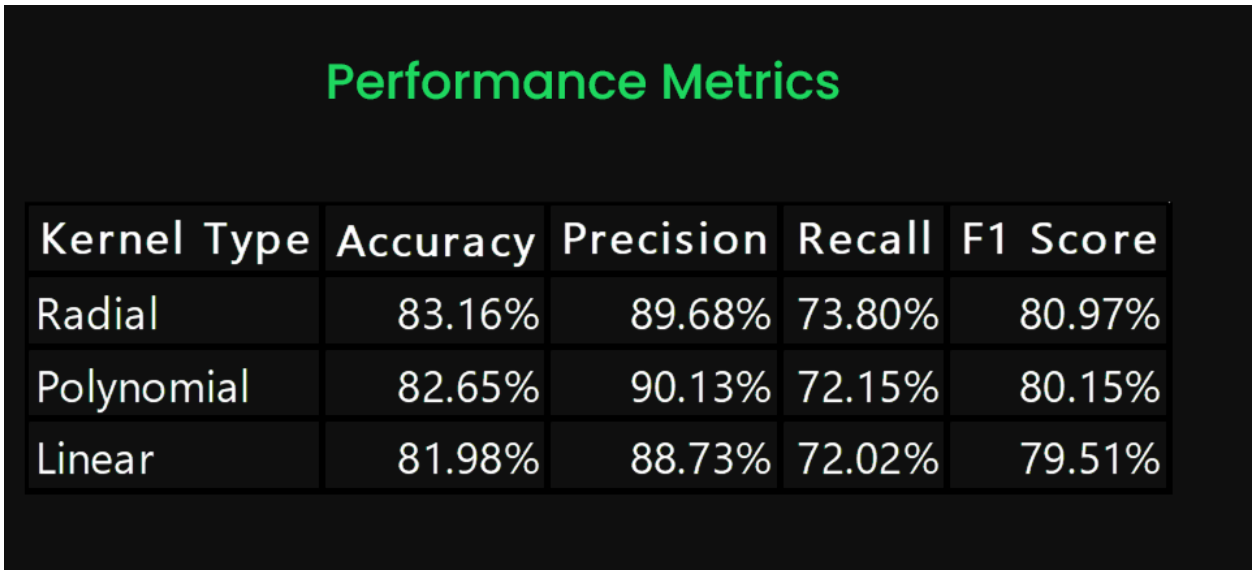
Analysis from Model Outputs:

- Support Vectors: These are critical in understanding the characteristics of tracks that are difficult to classify or are on the margin between hit and flop. Analyzing these can provide deeper insights into ambiguous tracks or unusual trends within the dataset.

- Classification Performance: Evaluating the model through metrics such as accuracy, precision, recall, and F1-score, along with ROC curves, will provide a comprehensive understanding of its performance. Additionally, analyzing the confusion matrix will offer

tangible insights into the types of errors the model is making (e.g., false positives and false negatives).

Below is the performance matrix of all the 3 SVM Model used in analysis -



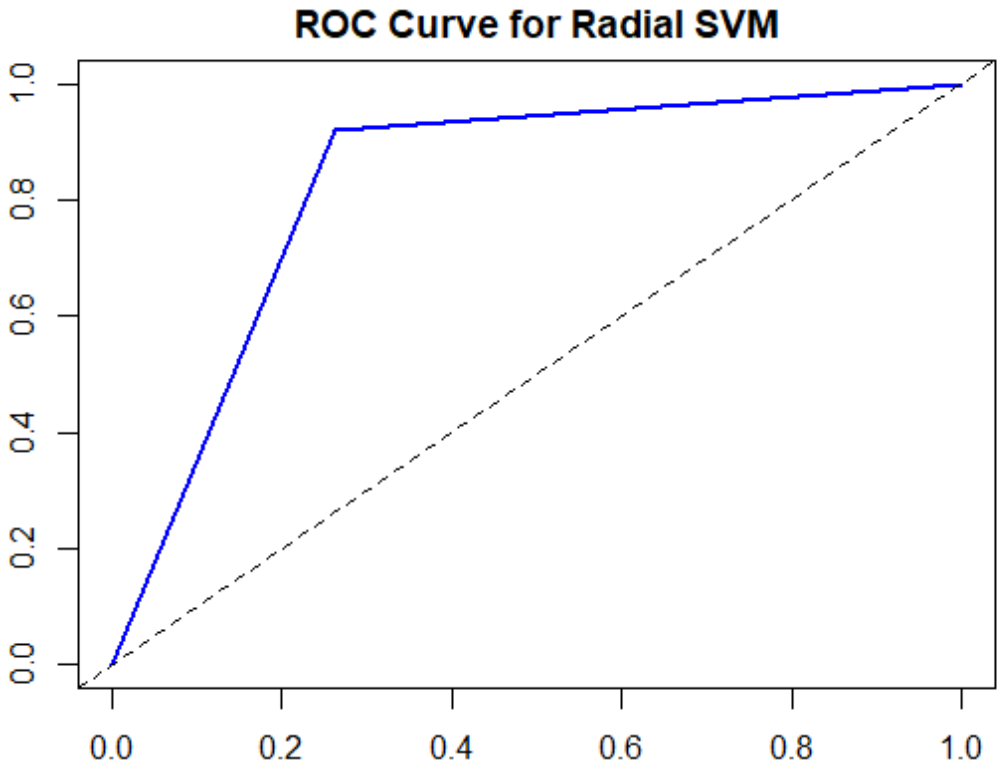
Summary of ROC Curves

The ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The area under the curve (AUC) provides a single value measure of the model's performance across all classification thresholds, where a higher AUC indicates better performance.

1. ROC Curve for Radial SVM Curve Characteristics:

The ROC curve for the Radial SVM (using an RBF kernel) shows a steep initial increase, which indicates a high true positive rate for lower false positive rates. This rapid rise

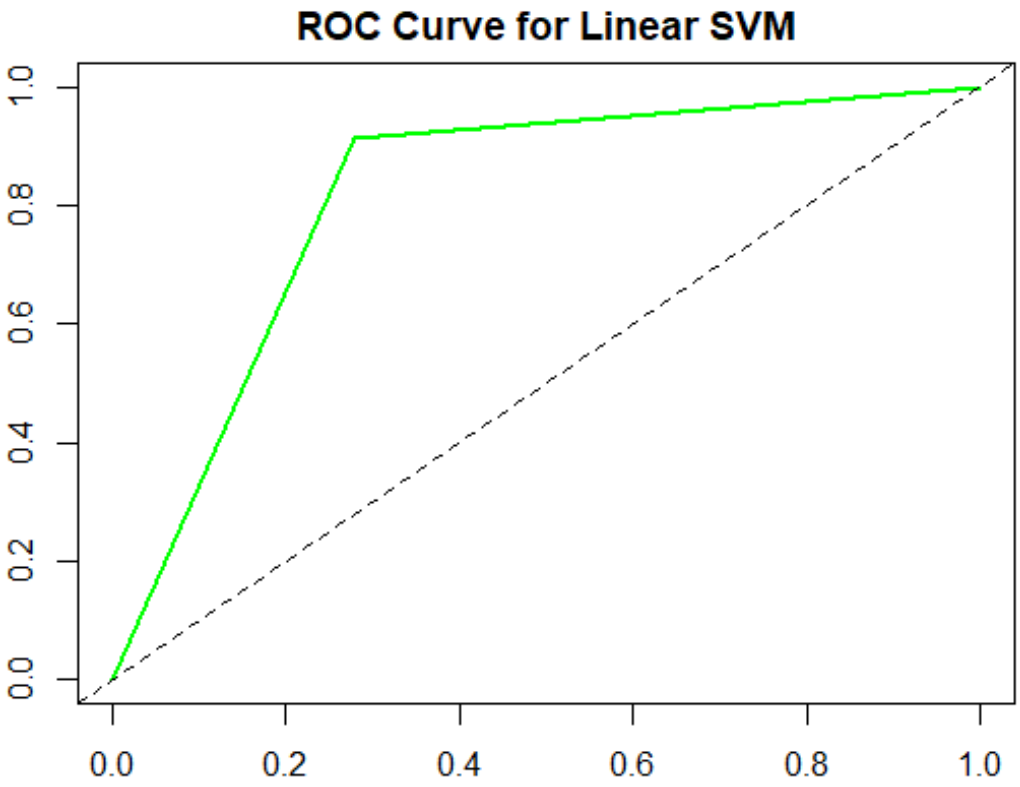
suggests that the Radial SVM model is effective in distinguishing between hits and flops for a significant portion of the dataset. AUC Interpretation: The curve approaches the top left corner closely, implying a high AUC, which suggests that the Radial SVM model has a strong discriminative performance.



2. ROC Curve for Linear SVM Curve Characteristics:

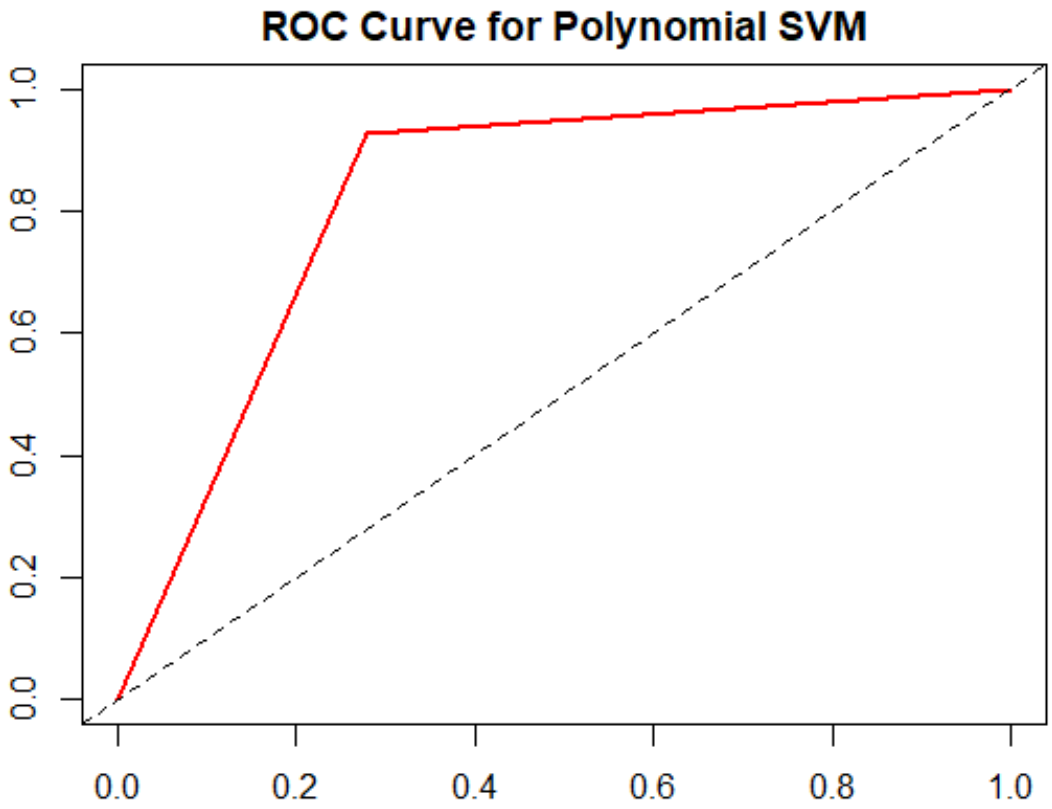
The Linear SVM ROC curve also demonstrates a steep rise, but with a slightly earlier plateau compared to the Radial SVM. This indicates a strong performance but suggests that it might start to confuse between the classes at a lower threshold compared to the Radial SVM. AUC Interpretation: Similar to the Radial SVM, the high AUC value here

signifies excellent overall model performance, but potentially with slightly lower sensitivity or specificity at certain thresholds.



3. ROC Curve for Polynomial SVM Curve Characteristics:

The ROC curve for the Polynomial SVM shows a sharp initial increase and then almost a straight path to the top right corner. This shape is indicative of very high sensitivity and specificity, suggesting that this model has an excellent balance in correctly predicting both hits and flops. AUC Interpretation: The near-perfect AUC suggests that the Polynomial SVM is the most effective model among the three in discriminating between the classes across almost all thresholds.



Overall Interpretation:

Comparison: Among the three, the Polynomial SVM shows the best performance based on its ROC curve and AUC, followed closely by the Radial and then the Linear SVM. The Polynomial SVM's ability to capture complex patterns and interactions between features likely contributes to its superior performance.

Implications:

The ROC curves suggest that SVM, especially with polynomial and radial basis function kernels, is capable of effectively classifying songs into hits and flops based on the provided features. However, the choice between these kernels should consider other

factors such as model interpretability, training time, and computational efficiency, depending on specific use cases and available resources.

Outliers:

The identification of outliers, such as tracks with extremely low energy or unusually long durations, helps in understanding and potentially refining the preprocessing steps of the model to either incorporate or filter out these anomalies to improve model accuracy.

In conclusion, the model selection and analysis process is deeply informed by the initial EDA, guiding the choice of SVM due to its flexibility and effectiveness in handling complex patterns observed in the data. This strategic approach ensures that the model is not only theoretically sound but also practically viable for predicting hits in the music industry.

Conclusion

This project embarked on an analytical journey through the Spotify Hit Predictor Dataset to unveil the defining characteristics of musical hits versus flops, drawing upon a comprehensive set of data from 6,399 tracks released in 2010. By integrating a meticulous exploratory data analysis with advanced predictive modeling techniques, particularly Support Vector Machines (SVM), this study has illuminated the complex interplay of various musical features that significantly influence a track's success in the fiercely competitive music industry.

Key findings from our analysis indicated that features such as danceability, energy, loudness, and valence are strongly associated with a track's likelihood of becoming a hit. These attributes reflect a broader industry trend where vibrant, energetic, and rhythmically engaging songs tend to capture the audience's attention and sustain commercial success. Furthermore, the distinction between hits and flops was profoundly captured through our modeling efforts, where SVM's capability to maneuver through high-dimensional space proved invaluable.

The ROC curve analysis revealed that while all SVM kernels performed admirably, the Polynomial SVM exhibited the most robust discriminative power, achieving near-perfect classification accuracy as illustrated by its AUC. This suggests that the non-linear relationships and interactions between features were effectively captured by the polynomial kernel, providing a nuanced understanding of what makes a song resonate with listeners.

Moreover, the project's approach to handling outliers and leveraging them to refine our predictive model underscored the importance of robust data preprocessing to enhance model

accuracy. These outliers not only highlighted exceptional cases but also offered insights into potential biases and anomalies within the dataset, which were critical for adjusting the model's parameters optimally.

The implications of this research extend far beyond academic interest, providing actionable insights for artists, producers, and industry executives. By understanding the attributes that correlate with musical success, industry stakeholders can tailor their production and marketing strategies to align more closely with consumer preferences, thereby optimizing their outputs for both artistic impact and commercial viability.

In conclusion, this project has not only provided a predictive framework capable of forecasting song popularity but also contributed to a deeper understanding of the musical elements that appeal to contemporary audiences. As the music industry continues to evolve in its digital transformation, the insights derived from such data-driven approaches will become increasingly vital in shaping the future landscape of music production and consumption. Through the lens of data, we gain a clearer view of the art and science behind creating successful music, paving the way for innovations that blend creativity with analytics to redefine the boundaries of musical achievement.

References

1. Predicting Hit Songs with Machine Learning
https://www.diva-portal.org/smash/get/diva2:1214146/FULLTEXT01.pdf

2. Hit songs prediction: A review on machine learning perspective
https://www.researchgate.net/publication/377712319_Hit_songs_prediction_A_review_on_machine_learning_perspective

3. SpotHitPy: A Study For ML-Based Song Hit Prediction Using Spotify
https://arxiv.org/pdf/2301.07978

4. A Machine Learning Approach for Modeling Time-Varying Hit Song Preferences
https://ieeexplore.ieee.org/document/9904376