Bop or Flop

15072: Advanced Analytics Edge

Mackenzie Lees · Joe Kajon · Christian Ingersoll · Katherine Mendyk

8 December 2023

1 Introduction

1.1 Context and Purpose

This project consisted of a comprehensive analysis to predict song popularity, focusing on contrasting attributes that define successful tracks on Spotify and TikTok. We used advanced analytical methods like Sentiment Analysis, Topic Modeling, and predictive models including Random Forest, CART, Logistic Regression, and XGBoost to dissect the elements that contribute to a song's success across various genres. Our findings reveal that while certain attributes hold across platforms, Spotify's diverse user base also values genre-specific features. The project's insights provide a nuanced understanding of song popularity dynamics, offering valuable guidance for targeted music production and marketing strategies.

1.2 Scope and Approach

Digital music consumption is continually evolving, with platforms like Spotify and TikTok shaping how songs gain popularity. Recognizing the need to understand these dynamics, this project aimed to analyze the key features contributing to a song's success across different platforms and genres. We analyzed Pop, Rock, Latin, Rap, EDM, and R&B, with a focus on identifying the distinguishing elements that resonate with listeners on Spotify and TikTok.

Sentiment analysis was applied to song lyrics to extract emotional and psychological tones, while Latent Dirichlet Allocation (LDA) helped in uncovering thematic patterns. Our predictive analysis involved Random Forest, CART, Logistic Regression, and XGBoost models, examining features influencing song popularity within each genre. These features led to a better understanding of the musical elements that captivate listeners' interests.

1.3 Data

In this project, datasets from Spotify and TikTok were combined to focus on a comprehensive range of musical attributes. The columns in the Spotify data included track_id, track_name, artist_id, artist_name, album_id, duration, release_date, track_popularity, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. Additionally, there were playlist-related data (playlist_id, playlist_name), temporal aspects (month, year, day_of_week), duration_mins, genre, and a binary "Popular" feature that was

created from track_popularity. Tiktok data included all columns except for artist data. Popularity was defined specifically as track popularity above the third quartile within a given genre. To ensure data quality, the data was cleaned by removing rows with N/A in the release date and, for repeated songs, the row with the maximum popularity score was chosen. This data selection and cleaning process provided a solid foundation for analysis, focusing on the intersection of music popularity and its various characteristics across these two major platforms.

In initial exploratory Spotify data analysis, we discovered that artist features were highly correlated. Also, acousticness was highly correlated with energy and loudness (Figure 5). As for seasonality, we found that most popular songs are released in January and November and on Fridays (Figure 7, 8).

2 Analysis/Methods

2.1 Sentiment Analysis and Topic Modeling

In terms of feature engineering, we leveraged various text mining and natural language processing techniques to derive meaningful insights from song lyrics. First, we employed Latent Dirichlet Allocation (LDA) for topic modeling. The model, set to identify five distinct topics, allowed us to understand the underlying themes within the song lyrics. We extracted the top 10 terms from each topic and identified the dominant topic for each song, merging this information back into the original dataset (Figure 6).

Next, we did sentiment analysis on song lyrics using two methodologies: Bing and NRC. The Bing sentiment analysis classified each song into Positive, Negative, or Neutral categories based on the aggregation of positive and negative word frequencies. In contrast, the NRC sentiment analysis focused on emotional content, categorizing songs into emotions such as joy, sadness, anger, and disgust. This classification was based on the frequency of words associated with each emotion within the lyrics (Figure 2, 4).

Finally, the results from the sentiment analysis and topic modeling were integrated back into the original dataset as additional features. This integration provided a more comprehensive view of each song, detailing its dominant topic and emotional classification.

2.2 Predictive Models

Our initial models predicted popularity for all genres, but we quickly discovered creating separate models for each genre would perform better as it better identifies nuances that influence popularity within different genres (i.e. what makes a rap song popular is different than what makes a rock song popular). After doing this, our AUC values increased. However, it is important to note the challenges that still exist in our modeling. First, what is popular in a genre one year may be quite different a few years down the line, especially when considering all the subgenres within each genre, so finding the perfect universal formula is quite difficult. Also, our data was smaller for certain genres, such as EDM and Latin, which resulted in lower predictability power as there was less to train on.

We experimented with Logistic Regression, classification trees (CART), Random Forest, and XGBoost predictive models on each genre's dataset. Our reasoning for picking these models was to aim for a balance of performance and interpretability. While XGBoost and Random Forests commonly perform better than Logistic Regression and CART models, the latter are far more interpretable and it is much easier to grasp the relationship between features and popularity. We assessed the success of our models through the AUC value. We chose AUC rather than accuracy because our dataset was inherently imbalanced (75% had a Popularity value of 0). We utilized a baseline score of 0.5 (random assignment of 0 and 1 for the target variable, popularity) to compare our models against. We also examined the impact of features on the target value.

When building Logistic regression models and choosing which features to use, we made sure to examine feature correlations as highly correlated features can lead to insignificant or counterintuitive coefficients (Figure 5). Even when accounting for this, it was difficult to find reasonably predictive and significant coefficients through this method. Our CART models had similar, unimpressive performance to the Logistic regressions. Our random forests and XGBoosts performed much better, with XGBoost being the best model for all genres (Figure 16). XGBoost's wide margin of performance improvement over CART and Logistic Regression was worth the decrease in interpretability. Further, we were able to examine feature importance for each XGBoost model to gain intuition towards which features are most impactful on a song's popularity. In comparison to the Spotify data, Tiktok data had an AUC of 0.6676 using XGBoost with its top features being loudness, tempo, speechiness, acousticness, and valence.

Genre	Model	AUC	Top Variable Importance
Rap	XGBoost	0.8357	Tracks/X100.Million/energy/danceability/Feats
R&B	XGBoost	0.7199	Tracks/speechiness/Lead.Streams/tempo/danceability
Pop	XGBoost	0.7127	One.Billion/Lead.Streams/energy/Tracks/Feats
Rock	XGBoost	0.6487	Lead.Streams/Tracks/energy/danceability/acousticness
EDM	XGBoost	0.6485	Loudness/speechiness/acousticness/Tracks/valence
Latin	XGBoost	0.6018	Tracks/loudness/danceability/Feats/valence

Figure 0: AUC and Variable Importance Results by Model

3 Conclusion

We gathered several insights from this project. Firstly, a significant driver of what makes a song popular is the current popularity of an artist. The most important features for our best models were consistently characteristics of the artists; thus, if a musician is already a “superstar,” it is more likely that their song will be popular, regardless of actual characteristics of their songs. While this can be discouraging for smaller artists, there are still valuable insights to be learned from modeling for non-superstars. For instance, what factors they focus on for the aim of increasing popularity will depend on what genre of music they’re producing. For example, a rock band will have more of a return on their investment for including “acousticness” in their songs, whereas a rap artist should expect more return from increasing their amount of features. Overall, identifying what makes a song popular is difficult with current data. While we can gather some results through modeling, much of what causes a song to be popular is still unexplained.

Spotify and Tiktok data shared several but not all important variables in predicting popularity. The overlap suggests a universal appeal for these attributes across platforms. However, genre-dependent predictors on Spotify indicate that its user base has more diverse musical preferences compared to TikTok. Our results highlight the importance of understanding platform-specific listener preferences, as they influence artists and producers in tailoring their music to meet the tastes of diverse audiences. In future work, we suggest examining subgenres as well as expanding on sentiment analysis and topic modeling from song lyrics.

Appendix

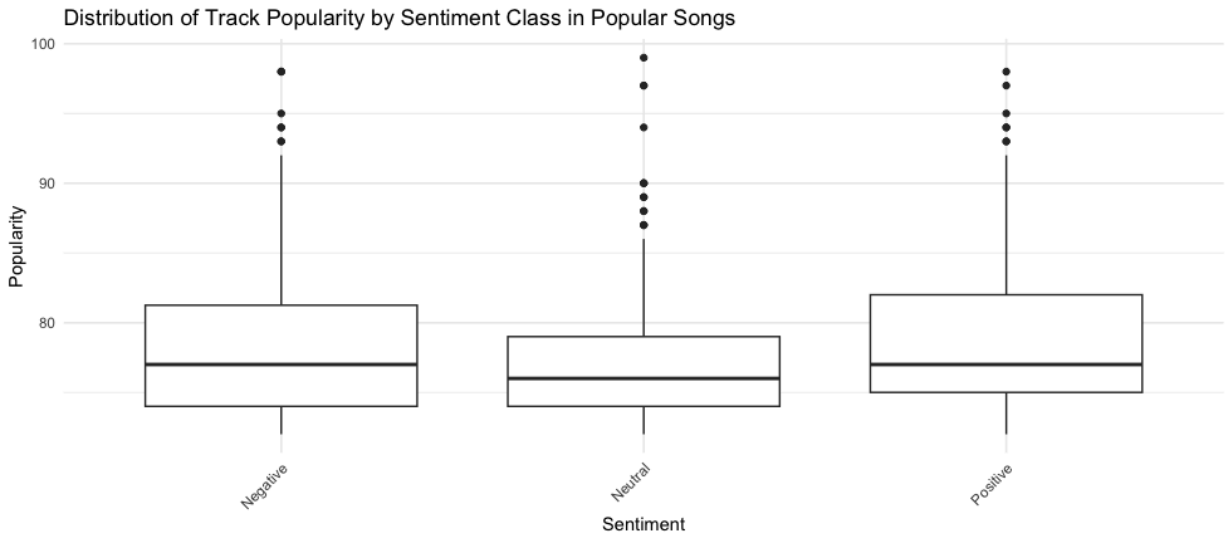

Figure 1: Distribution of Track Popularity by Sentiment Class in Popular Songs

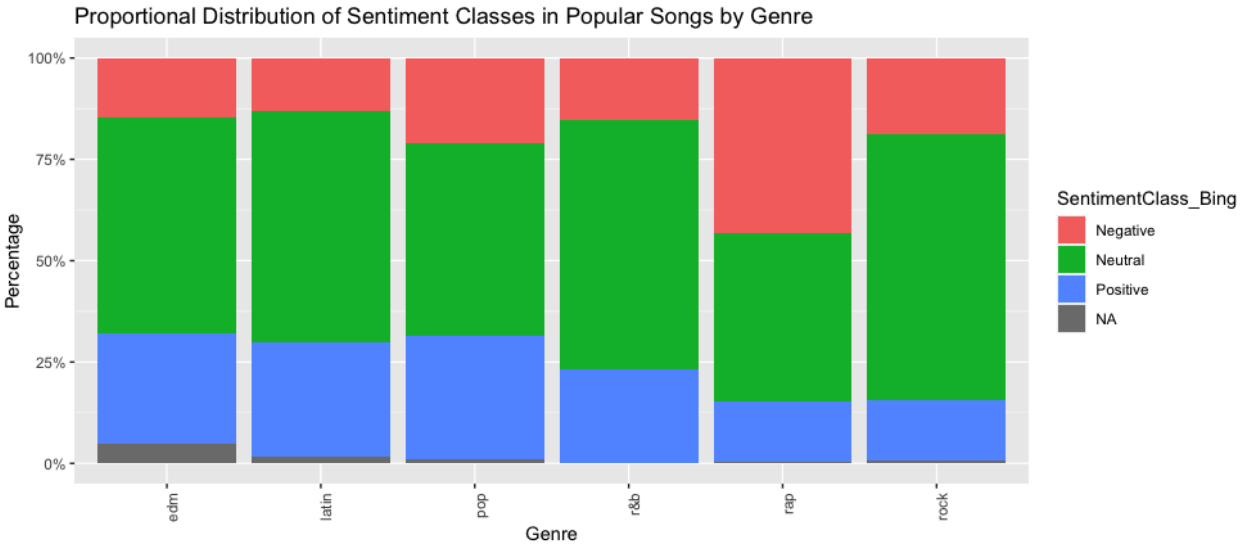

Figure 2: Proportional Distribution of Sentiment Classes in Popular Songs by Genre

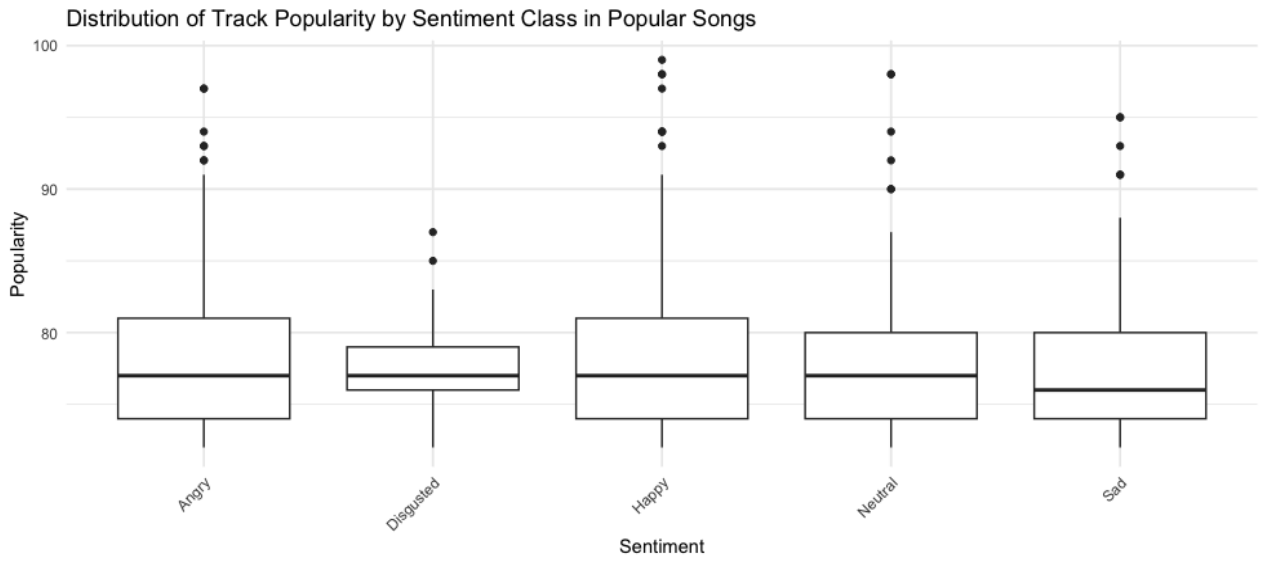

Figure 3: Distribution of Popularity by Sentiment Class in Popular Songs

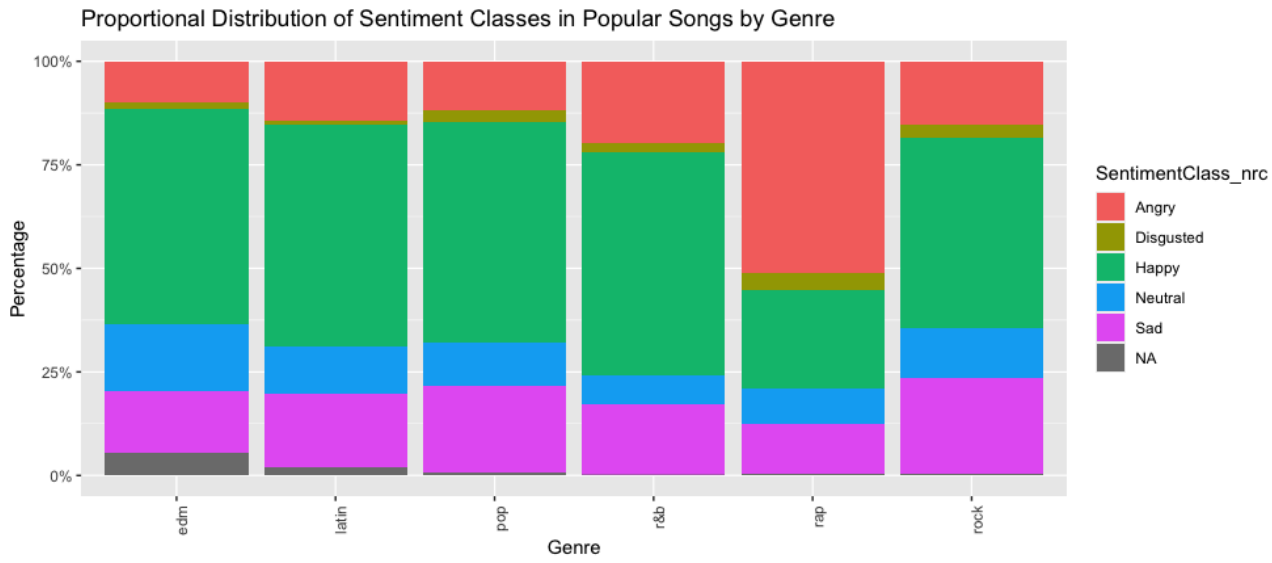

Figure 4: Proportional Distribution of Sentiment Classes in Popular Songs by Genre

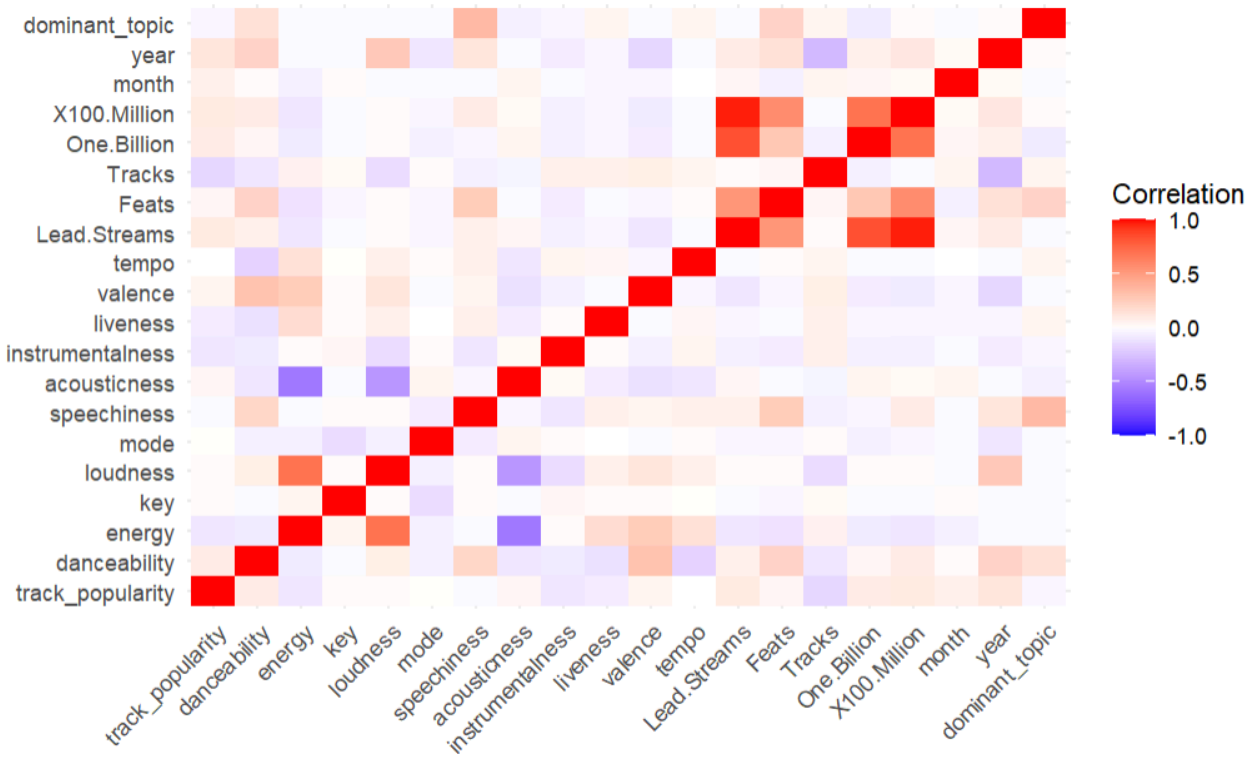

Figure 5: Correlation Matrix

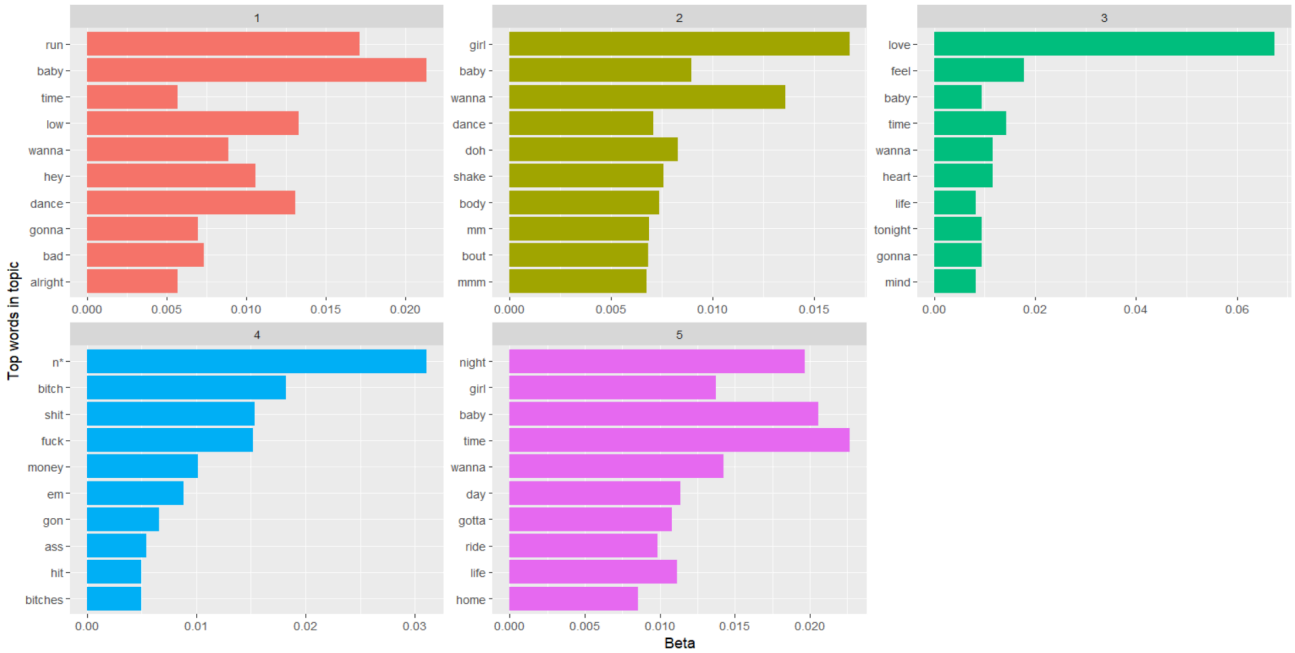

Figure 6: Top words in each topic from LDA

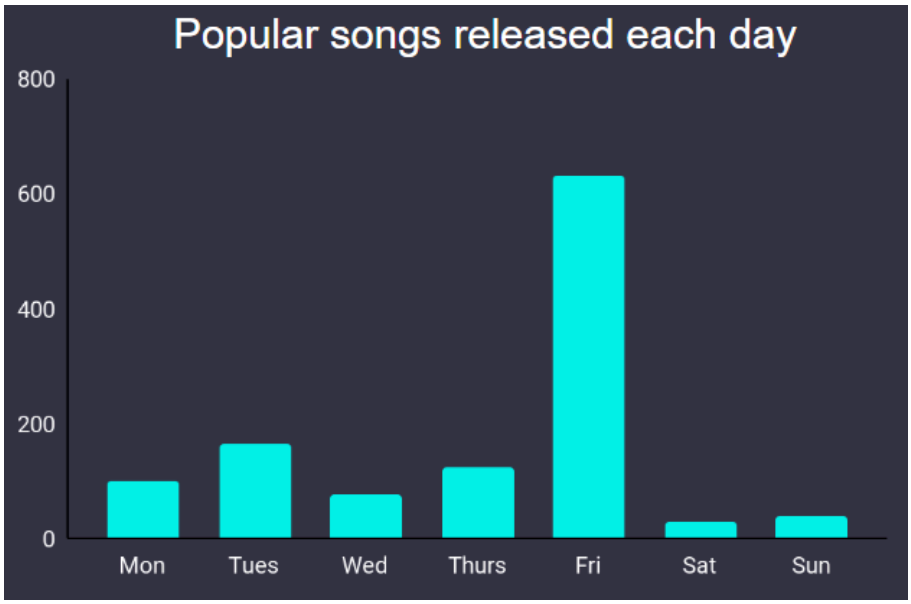

Figure 7: Number of Popular Songs Released by Day of Week

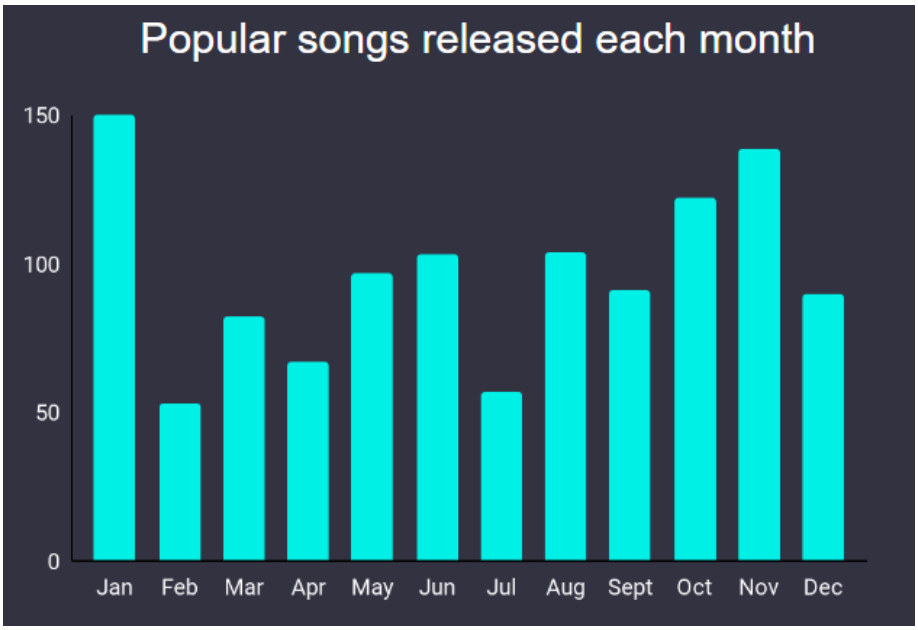

Figure 8: Number of Popular Songs Released by Month

	Overall <dbl>
One.Billion	0.239918244
Lead.Streams	0.123776189
energy	0.094127144
Tracks	0.080793040
Feats	0.077365006
SentimentClass_BingNeutral	0.052379758
acousticness	0.052277001
valence	0.045885388
X100.Million	0.032713966
speechiness	0.030092149

Figure 9: XGBoost Pop Model Variable Importance Results

	Overall <dbl>
Tracks	0.219656253
loudness	0.167547925
danceability	0.154788621
Feats	0.106891827
valence	0.060533344
X100.Million	0.052344514
Lead.Streams	0.048516957
energy	0.047570226
One.Billion	0.036592483
liveness	0.031269974

Figure 10: XGBoost Latin Model Variable Importance Results

	Overall <dbl>
Tracks	0.169681765
speechiness	0.091765218
Lead.Streams	0.086738777
tempo	0.071000700
danceability	0.066755643
acousticness	0.066743160
energy	0.065649690
loudness	0.063179212
valence	0.055668376
instrumentalness	0.051663899

Figure 11: XGBoost R&B Variable Importance Results

	Overall <dbl>
loudness	0.098012691
speechiness	0.095473048
acousticness	0.088940302
Tracks	0.085139599
valence	0.083799653
danceability	0.075337681
Lead.Streams	0.074453655
tempo	0.064789788
instrumentalness	0.057670996
X100.Million	0.057085428

Figure 12: XGBoost EDM Model Variable Importance Results

	Overall <dbl>
Tracks	0.242826549
X100.Million	0.106827383
energy	0.094388336
danceability	0.088753491
Lead.Streams	0.079876082
Feats	0.073403760
acousticness	0.057397513
loudness	0.051227867
liveness	0.048714081
valence	0.047474701

Figure 13: XGBoost Rap Model Variable Importance

	Overall <dbl>
Lead.Streams	0.168522209
Tracks	0.133260396
energy	0.101975414
danceability	0.092460878
acousticness	0.086360480
valence	0.081667622
tempo	0.063533151
instrumentalness	0.051053602
loudness	0.047538397
liveness	0.044878811

Figure 14: XGBoost Rock Model Variable Importance Results

```
xgbTree variable importance

                 Overall
loudness         0.16203
tempo            0.13839
speechiness      0.11569
acousticness     0.10986
valence          0.10140
liveness         0.10009
energy           0.09341
danceability     0.08158
instrumentalness 0.06439
mode             0.01681
key              0.01634
```

Figure 15: XGBoost TikTok Variable Importance Results

Genre	Model	AUC	Significant/important variables
Pop	Logistic Regression	0.6487	Energy, Sentiment_classBing
	Random Forest	0.6789	One.Billion, energy
	CART	0.6532	Lead.Streams,Tracks
Latin	Logistic Regression	0.5562	Dominant topic, speechiness, danceability
	Random Forest	0.5713	Tracks, loudness
	CART	0.5204	Loudness, Feats
R&B	Logistic Regression	0.6078	Day_of_weekThursday, OneHundred.Million
	Random Forest	0.6631	Month, loudness, acousticness
	CART	0.5559	OneHundred.Million, day_of_week, liveness
EDM	Logistic Regression	0.5703	Feats, Tracks
	Random Forest	0.5788	Acoustiness, feats, Lead.Streams
	CART	0.5206	Acoustiness, month, tempo
Rap	Logistic Regression	0.6847	Speechiness, tempo, instrumentalness
	Random Forest	0.8049	Loudness, instrumentalness, speechiness
	CART	0.7132	Energy, loudness, instrumentalness
Rock	Logistic Regression	0.5876	Tempo, instrumentalness, energy,
	Random Forest	0.5846	Loudness, instrumentalness, energy
	CART	0.5714	Energy, loudness, instrumentalness,

Figure 16: Results of Logistic Regression, CART, and Random Forests