DATA SCIENCE FOR BUSINESS FINAL REPORT

# TrackTracer: Analyzing Audio Features for Social Media Track Popularity Estimation

Section: C | Team: 59
Haider Ali (ha134), Dhruv Arora (da257), Roshni Balasubramanian (rb470), Camilla Kang (dk264)

## 1. BUSINESS UNDERSTANDING

Short video advertising on social media platforms, such as TikTok and Reels, has emerged as a crucial marketing strategy for building brand awareness. Studies (2021) find that relative to other platforms, ads on TikTok are significantly more likely to be seen and processed: 99% of the time, TikTok users are fixated on their screens while an ad is playing, compared to 76% of the time on other platforms. In response, platforms like TikTok have sought to make creating in-feed ads easier by offering templates and intuitive video creation tools that include background music (BGM), theme, and transitions. For example, TikTok has a long partnership with Canva to utilize existing TikTok templates on Canva, explore best practices for creating TikTok ads, and hear from a TikTok expert as they analyze best-performing designs. By utilizing templates, advertisers can recreate visuals quickly and easily.

When creating in-feed video ads, the choice of BGM is vital in engaging our audience. smooth and trendy BGM captivates user's attention in the first few seconds, which is crucial for short videos as that is the time users scroll on. Research shows that there is a 16% increase in impressions with background music compared to none. Well-engaged BGM have the potential to reach a massive audience within a short span of time and helps establish a sonic identity for the brand.

Advertising is all about trends. Ads are more likely to be watched in their entirety when they align with current trends. By identifying and utilizing upcoming viral songs, you can tap into the existing momentum and leverage the potential for your content to go viral as well. For advertising companies, this gives a head start in amplifiying your reach and increases the likelihood of your brand being discovered and shared by a wider audience. For platforms like TikTok, they could create ad templates that have potential and buy copyrights ahead of time. Understanding and predicting trendy soundtrack in the future have huge potential when it comes to advertising.

In the present report, we aim to investigate what songs are best used for commercialization through data mining on Spotify. Successful BGM contributes much to an advertisement, and alternatively, some commercials have even lead the trend of hit music so much that they become part of brand identity. Therefore, in order to capitalize short video advertising and be ahead of the trend, predicting hit songs allows leverage to platforms like TikTok to make a template made for them ahead of time. The costs associated with making a

right prediction could be the cost taken currently to make a template within a shorter time frame with the amount of turnaround of a template in a short amount of time. Further, identifying hits early offers to social media platforms a chance to secure favorable licensing deals, as these commercial music for templates often are offered royalty-free (TikTok for Business, 2023)**.** The cost of an incorrect prediction will be the amount of resources wasted in order to produce the template and ads as well as licensing deals. Thus, analyzing and predicting potential hit soundtracks that are best used for reels and tiktoks commercialization is crucial in short video marketing.

## 2. DATA UNDERSTANDING & DATA PREPARATION

In the world of advertising, music plays a huge role in grabbing the audience's attention. Knowing which song might be the next big hit can give advertisers a big advantage. So, our team decided to use the API of Spotify, one of the most popular music platforms, to help us predict the next trending songs. Spotify has information on millions of songs, which gives us a lot of data to work with.

The spotify library on Python, enables us to get the audio features of any playlist. Upon entering the URL for this playlist, the API fetches audio features of each song on the playlist. We looked at various playlists on Spotify (some official while some created by other users) to curate a list of hit and flop songs. For hit songs, we used playlists such as “Big on the Internet”, “Viral Hits Playlist”, “TikTok hit songs”, etc. As a proxy for songs that are not performing well, we searched for playlists that are ‘Underrated Songs on TikTok’.

This includes attributes such as the following:

Song Attributes:

| | | |
|---|---|---|
| Track Name | Artist Name | Track Popularity |
| Artist Popularity | Release Date | Album Name |
| URI of Song | Song Duration | URI of Artist |

Audio Features:

| | | | |
|---|---|---|---|
| Danceability | Acousticness | Mode | Key |
| Tempo | Loudness | Speechiness | Instrumentalness |
| Energy | Valence | Explicit | Liveness |

For our analysis, we focused on the 'TrackPopularity' metric – a direct representation of a song's global appeal. Further, in order to overcome any bias in a users’ listening tastes of what a

‘underrated song’ is and whether or not it is objectively a hit song, we focussed on the ‘TrackPopularity’ attribute rather than assigning every song on the “Hits” Playlist as a hit and every song on the “Underrated” Playlist as a flop. In this way, we are removing the bias of a particular user and making a data driven decision on what classifies a song as a hit. In our model, a song with a popularity score above 50 is categorized as a hit, while those below this threshold are deemed flops.

| Track Popularity Value | Count of Songs |
|---|---|
| >= 50 | 595 |
| < 50 | 496 |

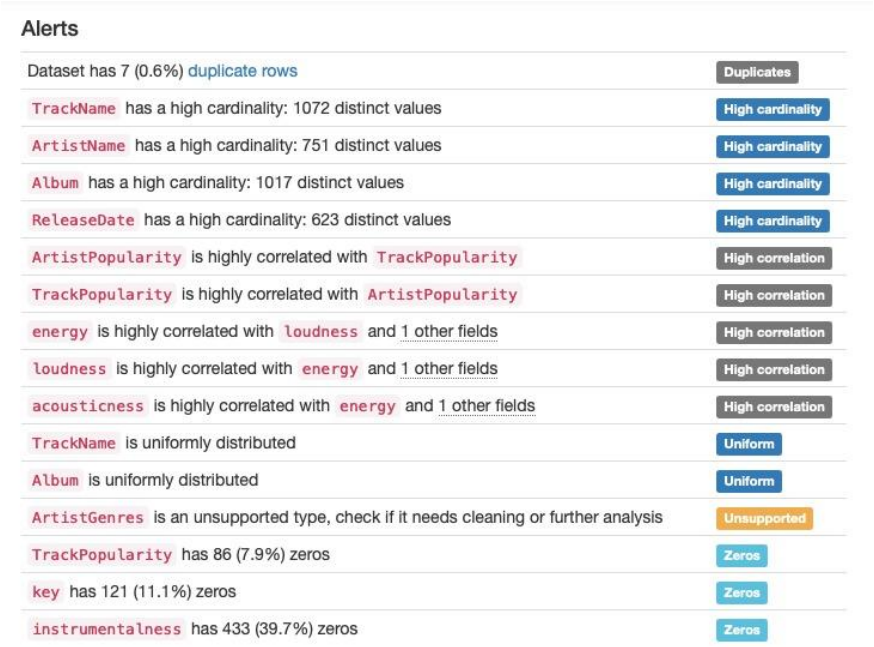
To understand the factors influencing this popularity, we delved deep into the dataset. Features like 'Danceability', 'Tempo', 'Valence', and 'Energy', among others, serve as our independent variables. These provide insights into the song's characteristics and its potential resonance with audiences. Attributes such as 'ArtistPopularity', 'ReleaseDate', and 'time_signature' offer valuable background context.

While the data from Spotify is incredibly rich and insightful, refining it is essential for focused analysis. Redundant columns such as 'URI', 'ArtistURI', and 'analysis_url' were removed. These, while integral for Spotify's operations, don't add value to our predictive model.

The data was collected from 17 different playlists. We were able to collect data for 1091 unique songs.

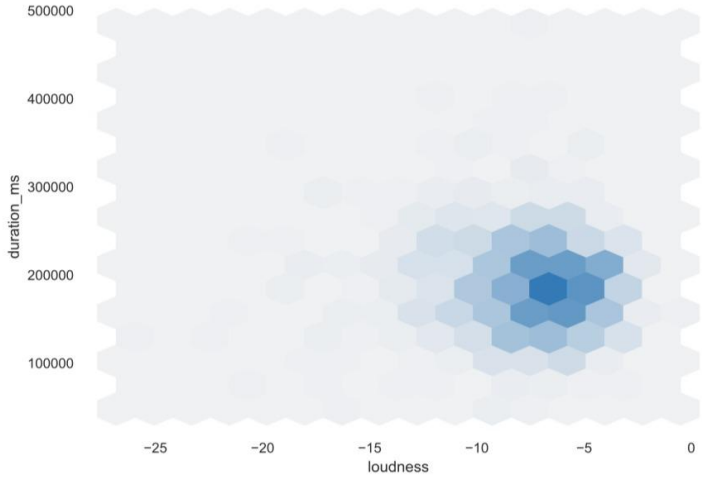
**Exploratory Data Analysis**

To start off with studying our dataset, we created a summary report using the *pandas-profiling* python library. It highlighted certain alerts to help us get started with which variables do we investigate some more.



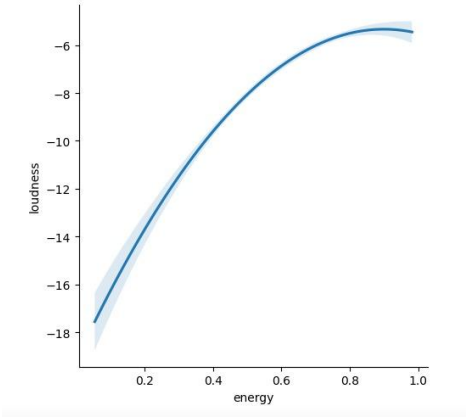

To study our dataset some more, we analyzed a few questions we wanted to understand to help us with modelling. The below are the questions along with their visualizations for the same.

**Question 1: Are louder songs shorter or longer in duration?**



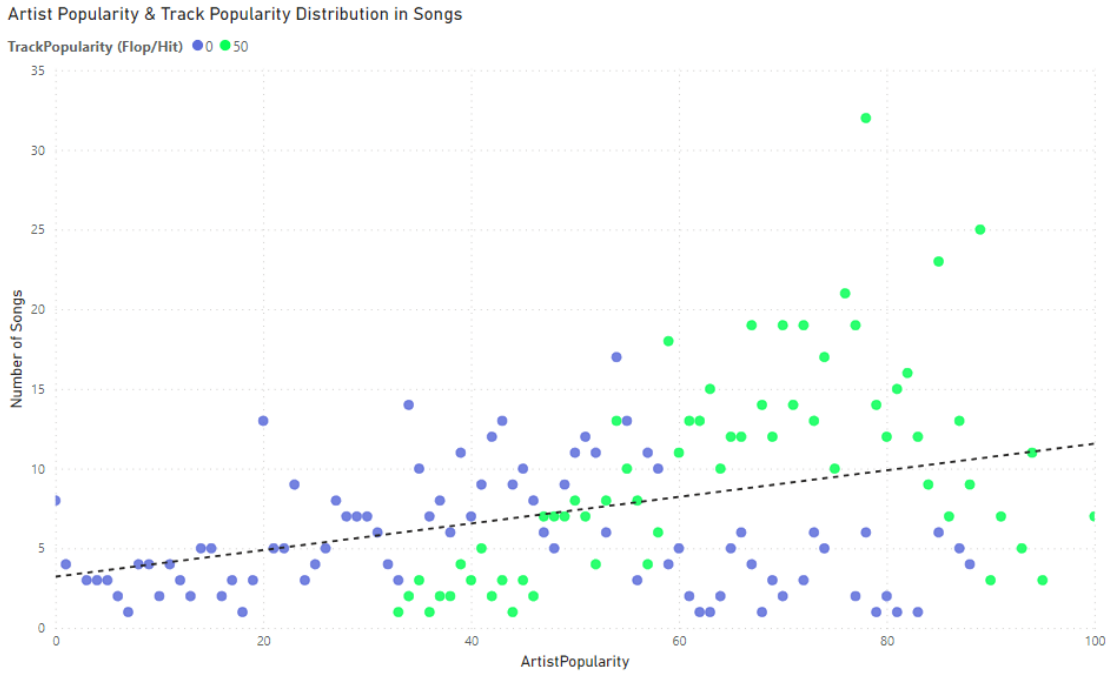

This visualization highly surprised us. This shows that there is an ideal (loudness, duration_ms) combination where most song tracks are concentrated. The further we move away from this highly concentrated region (by increasing/decreasing loudness or duration_ms), the lesser tracks we find are similar to it. Therefore, the hexagonal blocks keeps getting lighter the further they move away from the sweet spot (the darkest block).

**Question 2: Are more energetic songs louder too?**




This regression plot tell us that yes, more energetic tracks are louder too. However, after a certain extent, even if we keep increasing the energy of the song, the loudness starts coming to a standstill. Therefore, this tells us a song cannot increase its loudness beyond a certain extent.

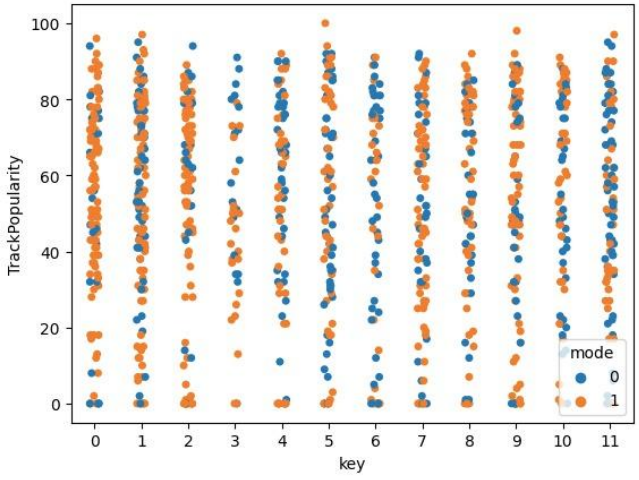
**Question 3: Does an artist's popularity affect the track popularity?**




In the above visualization, the green portion of the column indicates the number of hit songs created by the artist and the X axis is the distribution of artist's popularity in the dataset of songs collected. The conclusion that we come up with is that there could be a correlation between the artist popularity and the potential of popularity for a track. This is evident from the larger presence of hit songs (green data points) towards the higher artist popularity portion of the graph.

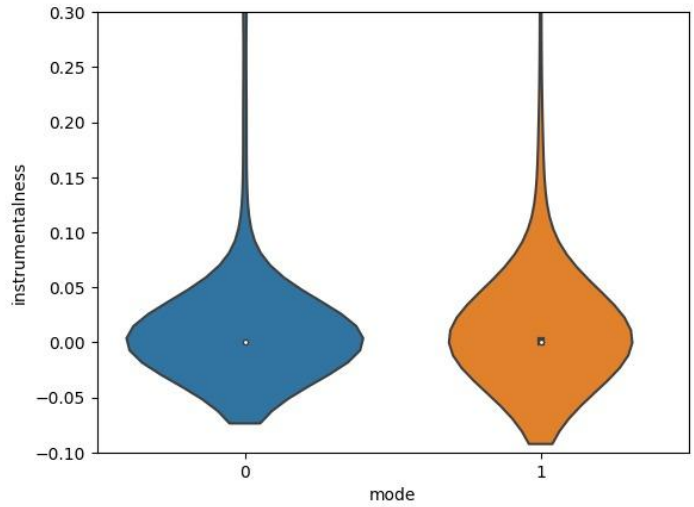
**Question 4: Do happier songs (having mode = 1) have a higher track popularity in general than a more dark/serious song (mode = 0)?**

While analyzing the relationship between Track Popularity and Mode generated no fruitful insight, what made it interesting was when we put another variable into the mix, which was Key (Like C, D, E, C#, etc represented as integers in the data).



Therefore, a happier song with a lower key is more likely to be popular than a happier song with a higher key. Similarly, a darker/more serious song with a higher key is more likely to be popular than a darker song with a lower key. This can be seen as the presence of orange data points is more on the lower keys end of the spectrum whereas the blue data points (mode = 1) become more frequent on the higher keys part of the visualization.

**Question 5: Does the happiness or sadness of a song affect how instrumental it is?**




For sadder songs (mode = 0), the intrumentalness points are more concentrated at zero when compared to happier (mode = 1) songs. This can be ascertained by looking at the belly of the violin plot. In fact, the happier songs' instrumentalness data points in general seem to be more widely distributed when compared to the sad track ones. Therefore, for a sad song, it is more likely that it will be less instrumental. The same cannot be said for a happy song.

**Interaction Terms**

Another step done in data preparation, beside dropping rows from our data that are not useful for our core tasks, was to introduce interactions that were suspected to prove useful from our data understanding and visualizations.
Upon analysis, we observe that Artist's popularity could influence whether a song becomes popular or not. Furthermore, attributes like energy, loudness, and acousticness may exhibit interaction effects, as these characteristics can be genre-dependent and influenced by the choice of instruments. These attributes can, in turn, influence the likelihood of a specific audience engaging with songs within a particular category.

## 3. MODELING & EVALUATION

The following tasks were identified and explored:

### 1. Predicting the Track Popularity for a new song:

This problem falls under the task of Regression. By setting the track's popularity as the dependent variable and using all other numeric/boolean values as the dependent variable, we are able to predict for a new song's popularity.

In order to carry out this task, various regression models can be employed. First, linear regression was applied. Second, random forest regressor was used.

In Python, the library 'sklearn' allows us to use the Linear Regression linear model. Upon importing this model, we also use a train/test split of 80%/20%. This allows us to train the model on 80% of the data and then test it on 20% of the data, to understand the performance. Upon fitting the model on the 80% train data, a prediction can be made for the 20% of the data set out as test-data.

From this, we can observe that the best model within this task is Random Forest Regressor, which is able to predict values with a mean squared error of 328.162.

### 2. Classification of Hit/Flop Song:

Since, it might be of more value as far as our business to predict whether a song is going to be a hit or a not hit, we can convert this to a classification problem by setting a threshold for the track popularity. A threshold we have set is 50 as a half way mark of popularity. Based on this, this problem has now been converted into a classification task.

| Model | Interaction | MSE | $R^2$ Score |
|---|---|---|---|
| Linear Regression | No | 347.162 | 0.404 |
| Random Forest Regressor | No | 332.036 | 0.522 |
| Linear Regression | Yes | 359.364 | 0.483 |
| **Random Forest Regressor** | **Yes** | **328.162** | **0.528** |

For this task, three models were built. A logistic regression model, a random forest classifier and a boosting model called AdaBoost classifier.

The logistic regression model was built using the sklearn library's function 'LogisticRegression'. A train/test size of 80/20% was used after which the model was trained and tested on the test set.

AdaBoost Classifier was used using the sklearn.ensemble library. Random Forest Classifier was used from sklearn.ensemble as well. These models produced the results consolidated in the table below:

| Model | Interaction | Validation Method | Accuracy |
|---|---|---|---|
| Logistic Regression | No | Train/Test Split | 0.7716 |
| AdaBoost Classifier | No | Train/Test Split | 0.8082 |
| **Random Forest Classifier** | **No** | **Train/Test Split** | **0.8173** |
| Logistic Regression | Yes | Train/Test Split | 0.8264 |
| **AdaBoost Classifier** | **Yes** | **Train/Test Split** | **0.8447** |
| Random Forest Classifier | Yes | Train/Test Split | 0.8219 |
| Logistic Regression | Yes | 10-Fold Cross Validation | 0.8103 |
| AdaBoost Classifier | Yes | 10-Fold Cross Validation | 0.7735 |
| **Random Forest Classifier** | **Yes** | **10-Fold Cross Validation** | **0.8038** |

### 3. Scoring Based on Similarity of Songs' AudioFeatures

For this, we used Cosine Similarity to get the most similar songs to every song in the test set and then used those to estimate the TrackPopularity of the song in the test set. For every song in the test set, we found the 10 most similar songs in the training set to that particular song. Then, we took the mean of the TrackPopularity of those 10 similar songs and estimated that average as the Track popularity of the song in the test set.
Mean Squared Error for all out of sample records: 0.1693
Even though the evaluation metric (MSE) gave a good result, this method is still not good for our current use-case as our data is limited. We are very likely to encounter many different types of songs (as in much different than the songs in our training set) which might make this method falter.

### 4. Deep Learning

Even though our dataset was small and Deep Learning models are data hungry, we still tried to use a Multi Layer Perceptron Neural Network to train our dataset. As expected, it gave a sub-par score in our evaluation metrics. The mean square error on the out of sample test set was 388.91 and the R^2 value was 0.4412.

**Evaluation of Best Model:**

We ultimately got 3 best models:-

- For train/test split without interactions: Random Forest Classifier
- For train/test split with interactions: AdaBoost Classifier
- For 10 fold CV with interactions: Random Forest Classifier

We have decided to go with the 3rd option: **Random Forest Classifier with interactions and 10 Fold Cross Validation** for the following reasons:-

- Our dataset was limited. We are more likely to encounter outliers and other noise in the real world. This model will be more robust in such a scenario.
- This model will not tend to overfit with the training data.

**NOTE:** One hypothesis that came out of our evaluation and one we would've liked to test was that if we had more data, would AdaBoost have performed better and would it then be a better choice as compared to Random Forest. The reason we thought about this was because AdaBoost is better at recognizing patterns in the data as compared to Random Forest and gives more consistent feature importance measurements when there are correlated features in the dataset (which is true for our Spotify dataset). Lastly, AdaBoost is better suited to Binary Classification problems (as in our case) as compared to Random Forest (which is more suited to Multiple Classification problems).

## 4. DEPLOYMENT

After extensively mining data, our predictive model is ready to identify potential hit songs. The next phase involves deploying these insights in a way that advertisers can act on them. We envision a digital dashboard available to advertisers, updated in real-time with songs that are trending. This tool will allow them to adjust their advertising content to match current music trends.

However, deploying this model comes with its challenges. We need to ensure our model integrates smoothly with existing advertising platforms, avoiding technical issues. There is also the risk of Data Drift, as the constantly changing nature of music means our model requires frequent updates as well as rigorous monitoring, demanding both time and resources. As more

advertisers use our service, the system must handle the increased demand without slowing down. Ethically, a major challenge is data privacy. Although our model primarily harnesses public data from Spotify, safeguarding any user-specific data against misuse is crucial.

If our team were a strategy consulting agency offering insights to TikTok, our strategy would involve curating a selection of songs that we anticipate will become popular among the recently added tracks on Spotify. This curated list of songs holds substantial potential for both TikTok and its content creators. TikTok could elevate its platform by either introducing templates that align with these potential hits or actively promoting these songs, thus simplifying the process for creators. Content creators, armed with advance knowledge of these predicted hit songs, could seamlessly incorporate them into their content, potentially expanding their audience reach. This symbiotic relationship presents a significant opportunity for TikTok to thrive.

Promoting content in line with musical trends can increase user engagement. Even a 5% increase in engagement, assuming each percentage point translates to an additional $1 million in ad revenue, could boost TikTok's coffers by an estimated $5 million. Additionally, identifying hits early offers TikTok a chance to secure favorable licensing deals. If negotiations reduce licensing costs by 10%, with the average hit song's license costing around $100,000, TikTok could save approximately $10,000 per song.

For Spotify, this venture offers two potential revenue streams. One is a subscription model where platforms like TikTok pay, perhaps $500,000 annually, for these insights. Another approach is commission-based, where Spotify earns a cut from TikTok's licensing cost savings, potentially netting Spotify around $1,000 per song.

To put our model into action, we tested our model on new songs added onto Spotify. We produce the following insights to TikTok:

The Top 3 Songs from the Provided List of New Songs:

a. SAY MY GRACE by Offset ft. Travis Scott
b. Hope You Know by Kodak Black
c. My Simple Jeep by Eyedress(feat. Mac DeMarco)

By allocating resources towards the development of templates and proactively promoting these songs to advertisers, TikTok can position itself as a frontrunner in staying ahead of music trends.

## APPENDIX

**ChatGPT Sessions**

- https://chat.openai.com/share/e842ff29-154e-4a24-bdd0-f1830170c018
- https://chat.openai.com/share/b11048cf-7e24-495a-806a-83af4ff846fe

**GitHub Repository: Link**

**References:**

Neurons. (2021). TikTok Marketing Science Global Ad Attention and Brand Building Study 2021 conducted by Neurons. https://ads.tiktok.com/business/creativecenter/quicktok/online/5_creative_tips/pc/en

Canva. (2022) Youtube video: Brand awareness ads on TikTok | Make effective ads with Canva's video editor. https://www.youtube.com/watch?v=QnQcoOzBYZA

TikTok for Business. https://www.tiktok.com/business/en-US/getstarted

Spotipy Documentation: https://spotipy.readthedocs.io/en/2.19.0/

**Team Contributions**

Roshni Balasubramanian: Data Collection, Data Understanding and Prep, Modelling, Evaluation
Dhruv Arora: Data Understanding and Prep, Modelling, Evaluation
Danming (Camilla) Kang: Business Understanding, Deployment Ideas, Presentation
Haider Ali: Deployment