EDA and Analysis of Spotify top 100 in 2023

Logan Keet

November 2023

Introduction

The data set that was analysed was collected from Kaggle. It is a collection of the most streamed songs on Spotify in 2023. The data set contains multiple features including a variety of audio features. Using these audio features a variety of machine learning algorithms were used to gain insight into what features have the largest impact in a songs chart success in Spotify. An EDA was also performed to gain insights into the data.

Data Cleaning

Data Features

- track name
- artist(s) name
- artist count
- released year
- relased month
- released dat
- in Spotify playlists
- in Spotify charts
- streams
- in apple charts
- in apple playlists
- in deezer charts
- in deezer playlists
- in shazam charts
- bpm
- key
- mode
- danceablitiy percentage
- valence percentage
- energy percentage
- acousticness percentage
- instrumentalness percentage
- liveness percentage
- speechiness percentage

Cleaning

There was found to be 95 values elements in the key feature. The songs with the missing values were removed from the data set. The keys and the modes were also mapped to numeric values so they can be used in machine learning algorithms. The data set was then filtered to be only the top 100 in the Spotify charts. Apple, Deezer and Shazam columns were removed and the date columns were combined into one. Two new columns were created one for ranking the songs by most streamed and one for songs in the most playlists.

EDA

Table 1: Most and Least Streamed Song

Features	Most Streamed	Least Streamed
Rank_By_Streams	1	813
Ranked_by_in_spotify_playlists	18	636
track_name	Shape of You	Que Vuelvas
artist(s)_name	Ed Sheeran	Carin Leon, Grupo Frontera
artist_count	1	2
in_spotify_playlists	32181	763
in_spotify_charts	10	26
streams(billions)	3.56254389	0.2762
release_date	2017-01-06 00:00:00	2022-12-09 00:00:00

Table 2: In most and Least of Playlists

Features	In Most	In Least
Rank_By_Streams	116	796
Ranked_by_in_spotify_playlists	1	814
track_name	Get Lucky - Radio Edit	Still With You
artist(s)_name	Pharrell Williams, Nile Rodgers, Daft Punk	Jung Kook
artist_count	3	1
in_spotify_playlists	52898	31
in_spotify_charts	0	39
streams(billions	0.933815613	0.038411956

Table 3: Top 10 Most streamed Artists

artist(s)_name	streams(billion)
Taylor Swift	11.85115108
Ed Sheeran	11.05125201
Bad Bunny	8.582384095
Eminem	6.183805596
The Weeknd	6.038640754
Harry Styles	6.033490512
Imagine Dragons	5.27248465
Adele	4.50874659
SZA	4.197341485
Bruno Mars	4.18573328
Coldplay	3.825176058
Olivia Rodrigo	3.55696115
Avicii	3.426754746
Dua Lipa	3.100230046
Arctic Monkeys	3.055659795
Kendrick Lamar	3.033135947
Linkin Park	2.985590613
Post Malone, Swae Lee	2.80809655
Justin Bieber	2.752482785
Drake, WizKid, Kyla	2.71392235

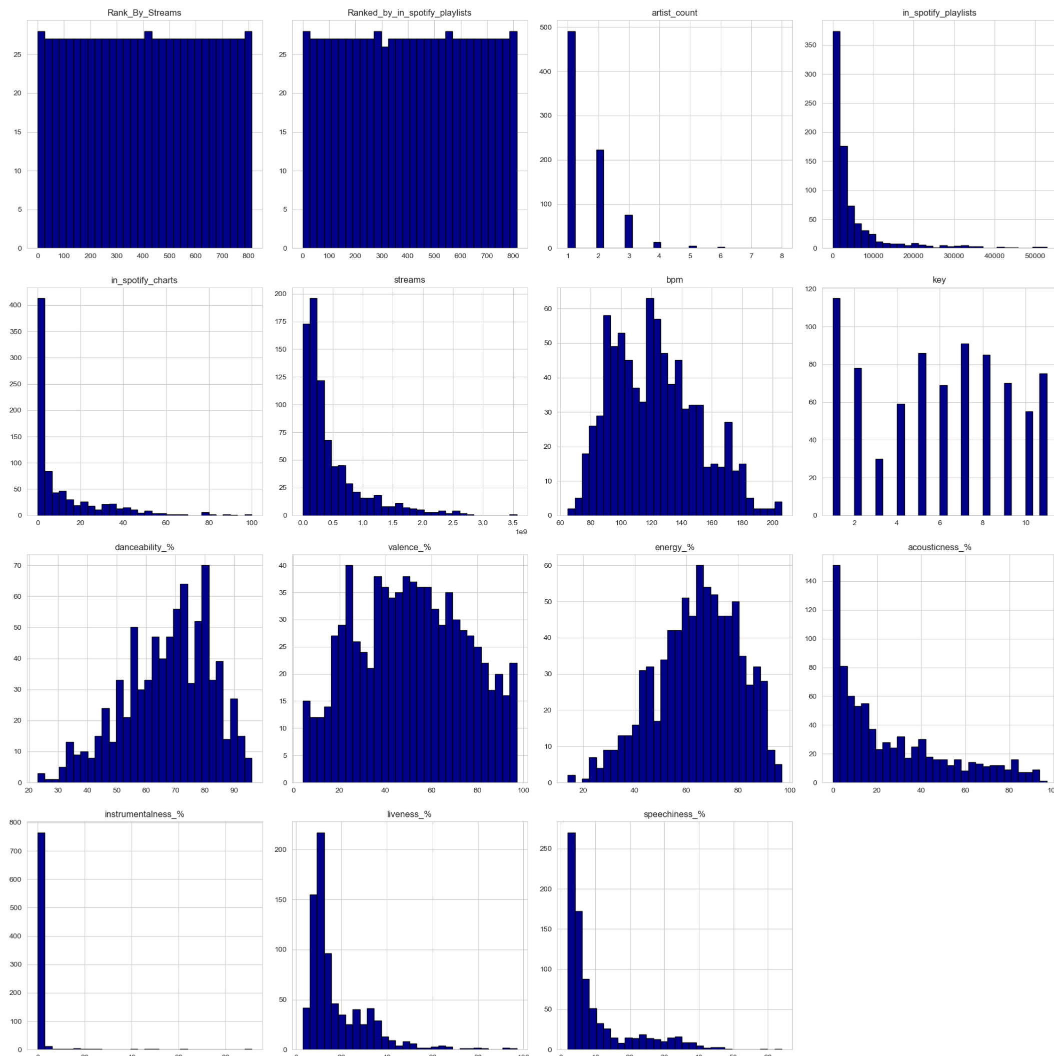

Figure 1: Distributions of the various features in the Data set

It is seen that most songs have a single artist. For the audio features it is seen that most songs have have little acousticness, instrumentalness, liveness or speechiness. The other features seem to have a fairly normal distribution of percentages.

Machine Learning Analysis

Key, BPM and Mode

The first three features that were explored using machine learning algorithms were Key, BPM and Mode. First a linear regression was performed on each of them separately.

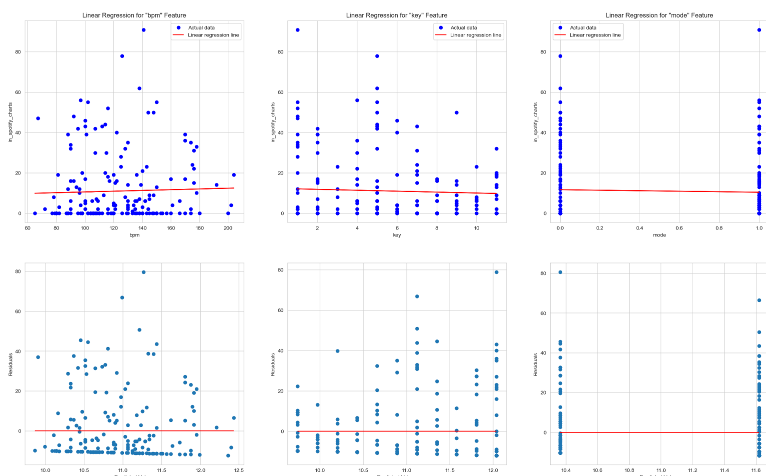

Figure 2: Plots of Key, BPM and Modes. The coefficient for Key is -0.23, BPM is 0.018 and mode is -1.26

The evaluation results of this model were

Table 4: Evaluation Results for Linear Regression of Key, BPM and Mode separately with means

Feature	MSE	R2 Score
bpm	348.8263171	-0.01611716
key	344.9559527	-0.004842943
mode	349.2622117	-0.017386904

The mean BPM was found to be $122.57 bpm$. All the models R^2 scores were negative which indicates that Key, BPM and Mode are not good predictors of chart success. Looking at the residual plots of these features it is clear that Key and Mode are not good indicators of chart success. Looking at the residual plot of BPM there might be a correlation between BPM and chart success.

Table 5: Linear Regression Model Using different BPM ranges

BPM Range	Predicted Success
65-85	10.31074084
85-105	10.43378662
105-125	10.55683241
125-145	10.6798782
145-165	10.80292399
165-185	10.92596977
185-206	11.05209171

This table was created by running the different BPM Range's through the linear regression model for BPM. The predicted success is what the model predicts it would be on the charts. There is very little difference between each test this indicates that BPM is not a good indicator of chart success.

Next I applied various Machine Learning Algorithms to the data set to see if there was a better predictive model.

Table 6: Different Machine Learning Models for Key, BPM, and Mode

Feature	Importance (Decision Tree)	Importance (Random Forest)	Importance (Gradient Boosting)	Importance (Linear Regression)
bpm	0.839325834	0.675831256	0.602238427	0.002971023
key	0.160674166	0.225068169	0.319117139	0.083250493
mode	0	0.099100575	0.078644434	0.913778484
MSE	362.408622	359.8348348	398.4740873	344.0749057
R2 Score	-0.055681873	-0.048184534	-0.160739136	-0.002276488

These models do not make any better predictions all the $R2$ scores are still negative but the most important factor still seems to be BPM.

Table 7: Decision Tree and Random Forest Importance using different BPM Ranges

BPM Range	Predicted Success Decision Tree	Predicted Success Random Forest
65-85	0	3.414393304
85-105	11.67647059	11.01320941
105-125	11.67647059	10.54416105
125-145	8.389830508	9.350957702
145-165	8.389830508	8.549521973
165-185	8.389830508	8.132539543
185-206	26.33333333	28.51730315

This predicts that songs have the most success in the range $65 - 85bpm$. Given the MSE and $R2$ scores of these models it is not the greatest predictor of a songs success in the charts.

Non-Engineered Audio Features

The Audio Features tested were

1. danceablitiy percentage
2. valence percentage
3. energy percentage
4. acousticness percentage
5. instrumentalness percentage
6. liveness percentage
7. speechiness percentage

First a Linear Regression was performed on each of the features separately.

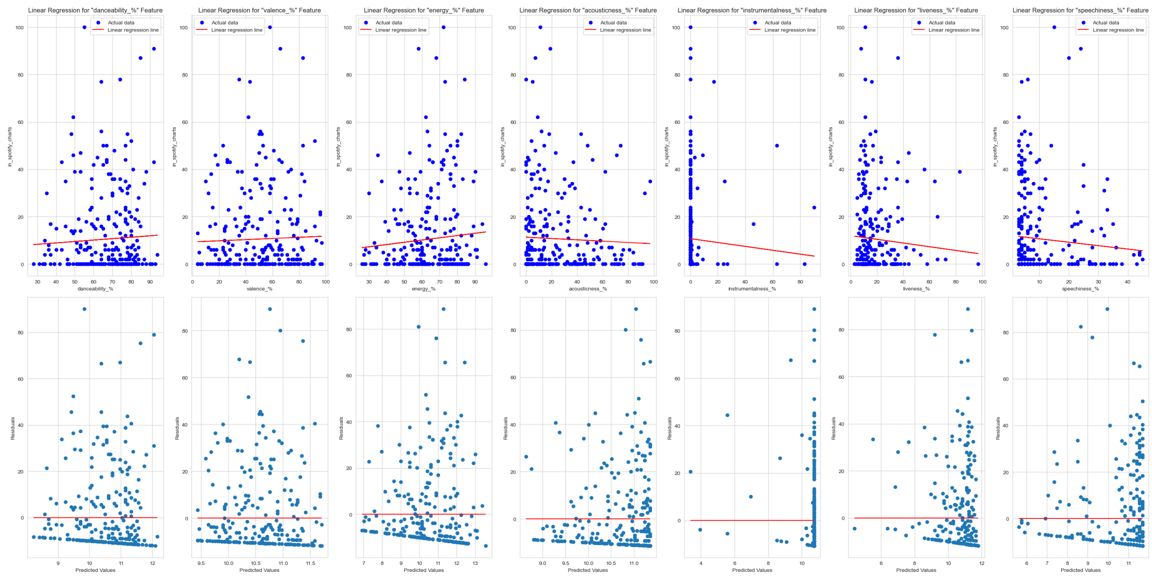

Figure 3: Linear Regression plot of the Audio features with residual plots

Table 8: Means and Coefficient results of Linear Regression

Feature	Means	Coefficients
danceability_%	67.41451415	0.059697246
valence_%	51.11685117	0.02433339
energy_%	64.32226322	0.095009237
acousticness_%	26.36900369	-0.028012126
instrumentalness_%	1.685116851	-0.081733765
liveness_%	18.14883149	-0.078481017
speechiness_%	10.55596556	-0.144386861

Table 9: Evaluations of the models

Feature	MSE	R2
danceability_%	348.2479391	-0.014432368
valence_%	348.6569715	-0.015623863
energy_%	345.3723099	-0.006055775
acousticness_%	346.1499651	-0.008321054
instrumentalness_%	351.831779	-0.024871951
liveness_%	351.2486857	-0.023173424
speechiness_%	346.527514	-0.009420839

All the $R2$ scores are negative and the MSE scores are fairly high this indicates that these features alone are not the greatest indicator of chart success of a song. The residual plots of the data for, danceablity, valence, energy and accousticness suggest there might be weak correlation for chart success. The other features it is clear there is no linear relationship in the data. Using the mentioned four features I ran various percentage ranges through the model to predict its songs success.

Table 10: Top 10 Results of Predicted Success

Feature	Percentage Range	Predicted Success
energy_%	0-20	5.398096615
danceability_%	0-20	7.147584791
energy_%	20-40	7.29828136
danceability_%	20-40	8.34152971
acousticness_%	80-100	8.833504013
energy_%	40-60	9.198466104
acousticness_%	60-80	9.393746524
danceability_%	40-60	9.535474629
valence_%	0-20	9.588768008
acousticness_%	40-60	9.953989035

The predicted success would be its estimated spot of the charts.

Engineered Audio Features

The features were created by adding two features together and then dividing them by two to get the mean average percentage between them.

$$feature_{eng} = \frac{feature_i + feature_j}{2} \tag{1}$$

All the features are

Features
danceability_%_+_valence_%
danceability_%_+_energy_%
danceability_%_+_acousticness_%
danceability_%_+_instrumentalness_%
danceability_%_+_liveness_%
danceability_%_+_speechiness_%
valence_%_+_energy_%
valence_%_+_acousticness_%
valence_%_+_instrumentalness_%
valence_%_+_liveness_%
valence_%_+_speechiness_%
energy_%_+_acousticness_%
energy_%_+_instrumentalness_%
energy_%_+_liveness_%
energy_%_+_speechiness_%
acousticness_%_+_instrumentalness_%
acousticness_%_+_liveness_%
acousticness_%_+_speechiness_%
instrumentalness_%_+_liveness_%
instrumentalness_%_+_speechiness_%
liveness_%_+_speechiness_%

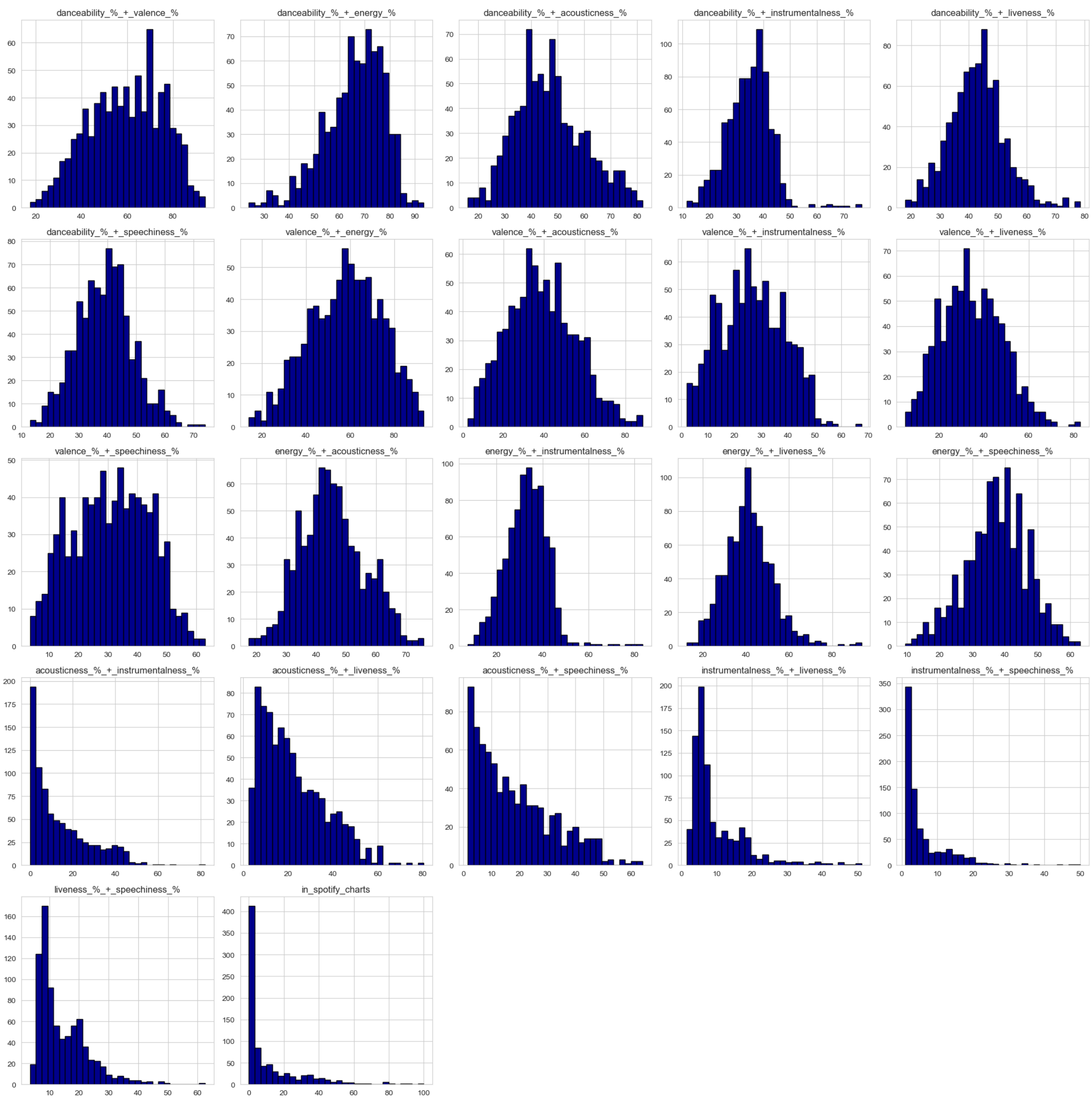

Figure 4: Distribution of the Engineered Features

Most of the features have a close to a normal distribution. Next performed Linear Regression on all of the features separately.

Table 11: Results for Linear Regression Models Separately

Feature	Means	Coefficients	R2	MSE
danceability_%_+_valence_%	59.26568266	0.051063571	-0.014369792	348.2264572
danceability_%_+_energy_%	65.86838868	0.140232773	-0.006338781	345.4694641
danceability_%_+_acousticness_%	46.89175892	-0.015928356	-0.015133446	348.4886144
danceability_%_+_instrumentalness_%	34.5498155	0.060908383	-0.013460414	347.9142737
danceability_%_+_liveness_%	42.78167282	-0.005699272	-0.017628532	349.3451609
danceability_%_+_speechiness_%	38.98523985	-0.011007028	-0.01724886	349.2148221
valence_%_+_energy_%	57.7195572	0.071048548	-0.010712253	346.970848
valence_%_+_acousticness_%	38.74292743	-0.008862186	-0.016191421	348.8518102
valence_%_+_instrumentalness_%	26.40098401	0.029584826	-0.015083187	348.4713608
valence_%_+_liveness_%	34.63284133	-0.001202627	-0.017375559	349.2583169
valence_%_+_speechiness_%	30.83640836	-0.004858112	-0.017327708	349.2418902
energy_%_+_acousticness_%	45.34563346	0.02995744	-0.020372212	350.2870484
energy_%_+_instrumentalness_%	33.00369004	0.126414116	-0.005184166	345.0730923
energy_%_+_liveness_%	41.23554736	0.045006609	-0.013120633	347.7976291
energy_%_+_speechiness_%	37.43911439	0.058410427	-0.015669438	348.6726168
acousticness_%_+_instrumentalness_%	14.02706027	-0.06336439	-0.009830582	346.6681763
acousticness_%_+_liveness_%	22.25891759	-0.080961083	-0.007319151	345.8060185
acousticness_%_+_speechiness_%	18.46248462	-0.089103845	-0.00053155	343.47588
instrumentalness_%_+_liveness_%	9.91697417	-0.165935936	-0.031237902	354.0171679
instrumentalness_%_+_speechiness_%	6.120541205	-0.267516424	-0.023320191	351.2990699
liveness_%_+_speechiness_%	14.35239852	-0.215442196	-0.020393669	350.2944144

None of the results had positive $R2$ scores. Looking at some of the residual plots(can view in notebook) there does seem to be slight correlation. A variety of ranges of percentages was ran through each model here are the ten most successful. Models where it was clear from the residual plots there was no correlation were not included.

Feature	Percentage Range	Predicted Success
danceability_%_+_energy_%	0-20	2.70187349
danceability_%_+_energy_%	20-40	5.506528951
valence_%_+_energy_%	0-20	7.174884954
danceability_%_+_valence_%	0-20	8.06023054
danceability_%_+_energy_%	40-60	8.311184412
valence_%_+_energy_%	20-40	8.595855909
danceability_%_+_valence_%	20-40	9.081501952
energy_%_+_acousticness_%	0-20	9.532472142
danceability_%_+_acousticness_%	80-100	9.925094852
valence_%_+_energy_%	40-60	10.01682686

The predicted success would be its estimated spot of the charts.

The data was then ran through various machine learning algorithms with all features together.

Table 12: Success Metrics of the Models

Model	MSE	R2 Score
Decision Tree	330.2744319	-0.052529096
Random Forest	312.391909	0.004459498
Gradient Boosting	353.2789165	-0.125840522
Linear Regression	310.8894583	0.009247556

The most successful model was the random forest model. When Linear Regression was tested on a random set of data the predicted results were outside of what would be suspected for chart success. From the random Forest Model the most important features are Most successful predictions from test model available on GitHub

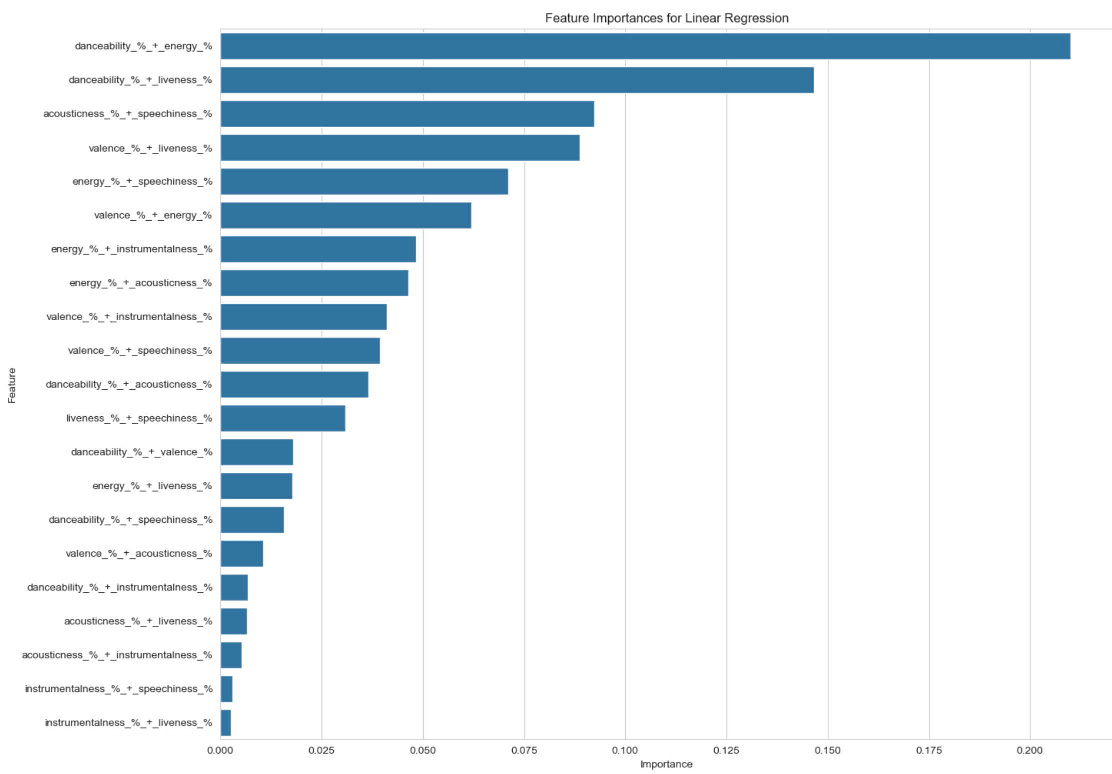

Figure 5: Feature Importance visualization