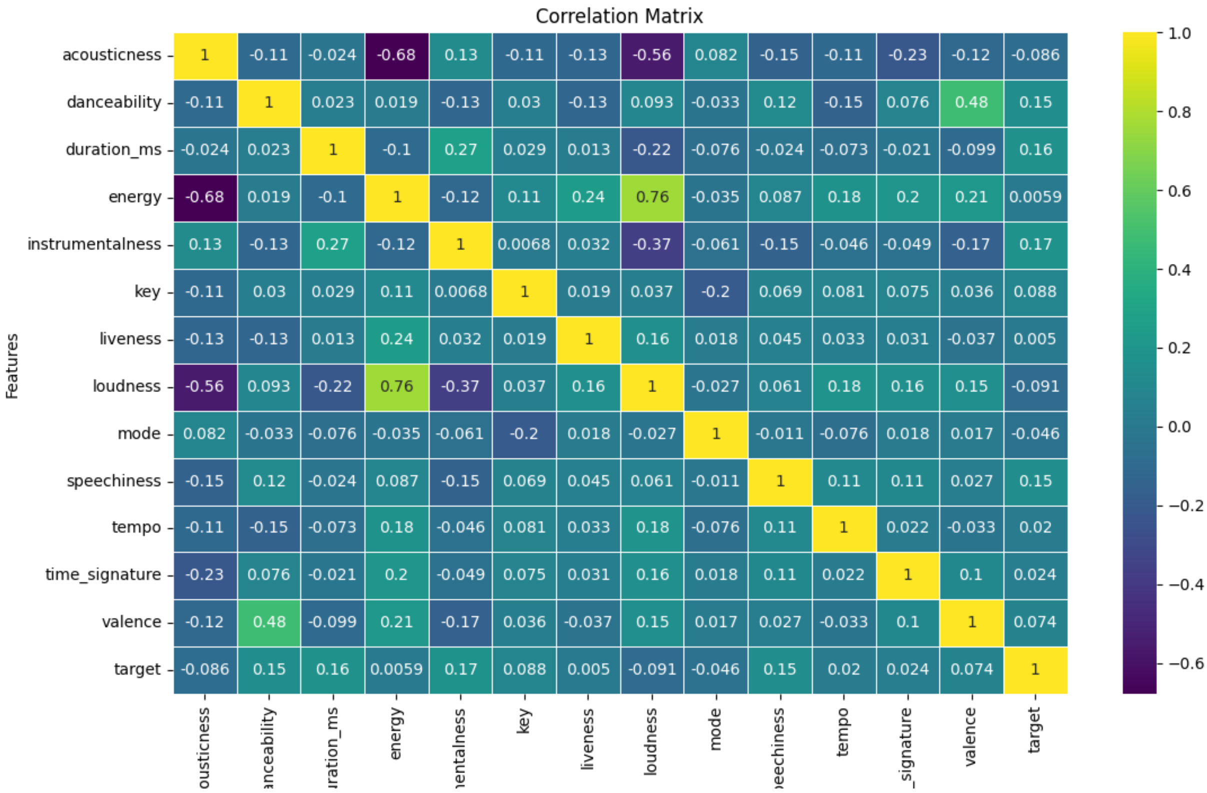
# Machine Learning Project on Music Analysis

## *Figuring out features that contribute to a song being successful*




Correlation Matrix

## Abstract

This report explores the relationship between various audio features of songs and their potential to become chart-topping hits. The dataset used in this analysis includes 13 key features extracted from Spotify, such as energy, danceability, valence, and tempo, for both hit and non-hit songs. The primary objective is to identify which of these features contribute most significantly to a song's success on the charts.

To conduct this analysis, the data was preprocessed and scaled, and various machine learning techniques, including logistic regression and random forest, were used to assess the importance of each feature. visualisations such as line graphs and feature importance cluster analysis were employed to interpret the results.

The analysis reveals that certain features, like instrumentalness, duration, and loudness, have a strong positive correlation with a song's hit potential, while liveliness and tempo were less significant. Interestingly, feature reduction did not improve model accuracy, indicating that most of the features contribute meaningfully to the classification task. These findings offer practical insights for music producers and marketers aiming to craft commercially successful tracks.


## Introduction

In the context of this report *"HIT"* are the songs that made it to the top 50 songs on Spotify.

The objective of this analysis is to examine a set of 13 audio features, including energy, danceability, tempo, and others, to determine which of these features most strongly influence a song's potential to become a hit.

By using statistical and machine learning methods, the analysis aims to quantify the contribution of each feature and understand how various audio characteristics correlate with a song's success on the charts. Ultimately, the goal is to provide data-driven insights into what makes a song stand out and achieve commercial success.

*Source:* https://www.kaggle.com/code/aeryan/spotify-music-analysis

## Data and Features Description

The data is a .csv file of over 2100 songs each having Author, song_title and 13 other features describing about the song.

The feature are: *'energy', 'danceability', 'acousticness', 'instrumentalness', 'liveness', 'loudness', 'speechiness', 'valence', 'tempo', 'key', 'time_signature', 'mode', 'duration_ms'*.

Also a final column target, which has the data regarding if the song was hit or not.

The data was sourced from Spotify via Kaggle, which may introduce certain biases based on the types of music that are popular on the platform. Future studies could consider collecting a more diverse dataset to account for genre-based variations in feature importance.

## Methodology and Observations

The data is first divided into 3 parts 70% the training part 20% the validation part and finally 10% the testing part.

The data obtained and stored in 3 different .csvfiles was then loaded for further analysis.

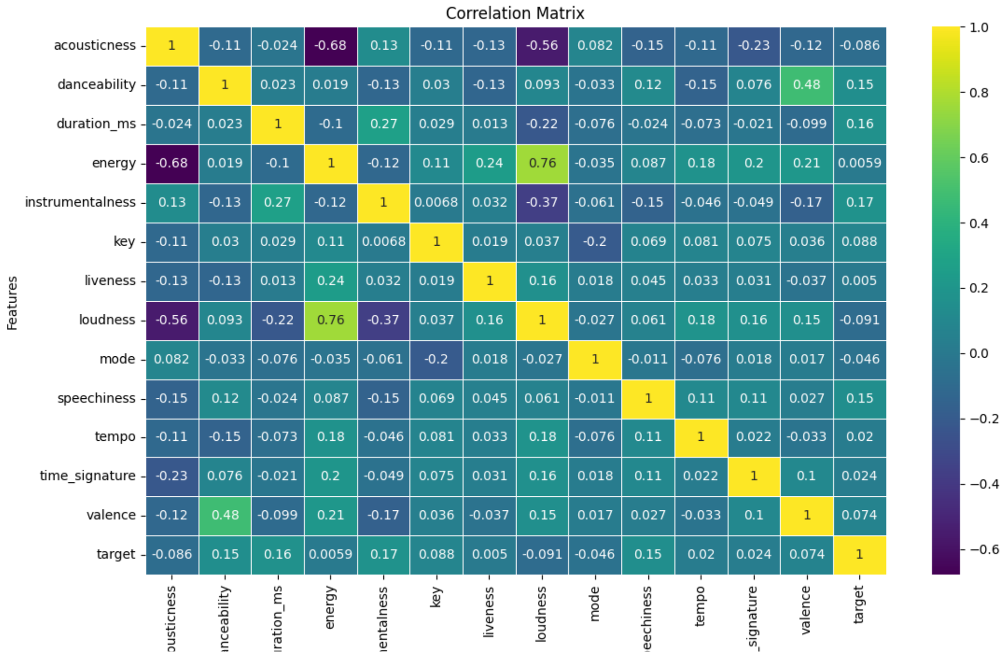
After this one of the 2 major things was to plot the individual features vs target to see it there was any viable difference in the data set. The second being Correlation Matrix one being with the target and the second without the target.

The link to the file with all the plots.

When you observe the plots related to the individual features vs the target it is quite clear that there is no significant observable difference relating to any other 13 features.

When we look at the Co-relation Matrix we see that instrumentalness, duration , danceability and speechiness show positive correlation. While loudness, acousticness and mode show negative correlation and the other 6 show small positive correlation with the target.

Various other patterns can be observed such as energy and loudness having a positive correlation. Valence and danceability having a positive




correlation. Acousticness having a negative correlation with energy and loudness.

Before starting with feature selection which is the Main Goal for this task. Let us first examine how well our classification works prior to feature selection. The classification algorithm that are being used are GradientBoostingClassifier and RandomForestClassifier.

*GradientBoostingClassifier*

Gradient Boosting Accuracy (Validation): 0.7770700636942676

Gradient Boosting Accuracy (Test): 0.765676567656765

The default case for 7 features

Gradient Boosting Accuracy (Validation): 0.7707006369426752

Gradient Boosting Accuracy (Test): 0.735973597359736

The best case scenario calculated with all the features.

Best parameters found: {'max_depth': 11, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 500}

Gradient Boosting Best Model Accuracy (Validation): 0.7770700636942676

Gradient Boosting Best Model Accuracy (Test): 0.7656765676567657

The best case scenario calculated with 7 features.

Best parameters found: {'max_depth': 3, 'max_features': None, 'min_samples_leaf': 6, 'min_samples_split': 2, 'n_estimators': 100}

Gradient Boosting Best Model Accuracy (Validation): 0.7579617834394905

Gradient Boosting Best Model Accuracy (Test): 0.7194719471947195

As you can see that the accuracy on the validation and test set reduced. Indication that the other 6 features with was playing a role in the classification task and was not external noise.

The same was observed in the other 2 methods as well.

*RandomForestClassifier*

The default case scenario.

With all 13

Random Forest Best Model Accuracy (Validation): 0.7802547770700637

Random Forest Best Model Accuracy (Test): 0.7838283828382838

with the best 7

Random Forest Best Model Accuracy (Validation): 0.7515923566878981

Random Forest Best Model Accuracy (Test): 0.7458745874587459

The best case scenario.

Best parameters found: {'max_depth': 20, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}

Random Forest Best Model Accuracy (Validation): 0.7707006369426752

Random Forest Best Model Accuracy (Test): 0.7739273927392739

for the best 7 features

Best parameters found: {'max_depth': 30, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}

Random Forest Best Model Accuracy (Validation): 0.7547770700636943

Random Forest Best Model Accuracy (Test): 0.740924092409241

It is clear from the above examples that choosing the K best features doesn't help improve the performance of the model indicating that all the features play a role. To calculate the feature that contribute the most towards a song being successful can be calculated using various methods.

Some of them are LinearRegression, Ridge, Lasso, RandomForestRegression. Of All these RandomForestRegression had the best R^2 and MSE value

*RandomForestRegressor*

Random Forest $R^2$ validation : 0.3648993119370132

Random Forest MSE validation : 0.15867210889950462

Random Forest $R^2$: 0.31245394251395653

Random Forest MSE: 0.171819114503117

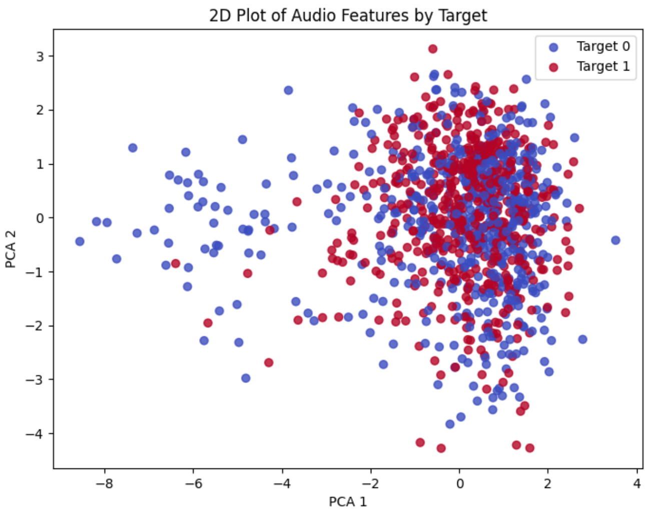
The feature importance of RandomForestRegression model is:

| | |
|---|---|
| speechiness | 0.126530 |
| instrumentalness | 0.122276 |
| loudness | 0.109353 |
| duration_ms | 0.104678 |
| energy | 0.103670 |
| danceability | 0.095080 |
| acousticness | 0.092643 |
| valence | 0.088084 |
| tempo | 0.060844 |
| liveness | 0.058636 |
| key | 0.028300 |
| mode | 0.006978 |
| time_signature | 0.002927 |

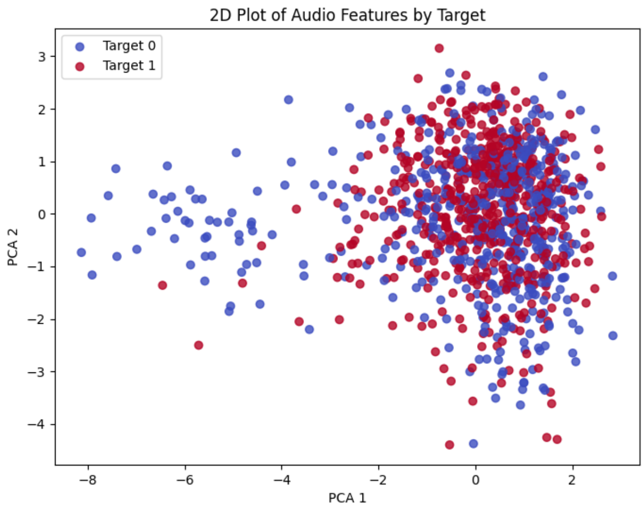
From this all the features with values above 0.09 were taken hence the best 7. #All the other 3 algorithms with their values are in the code file.

It is observed that the maximum accuracy of any models no matter the parameters tuning doesn't cross 80% accuracy. In order to figure out why a PCA is used to reduce the n - dimensional features to a 2d plane.

To understand better this has been done and plotted for 13 features , 10 features, 7 features, and 4 features all chosen based on the K best based on RandomForestRegression model.

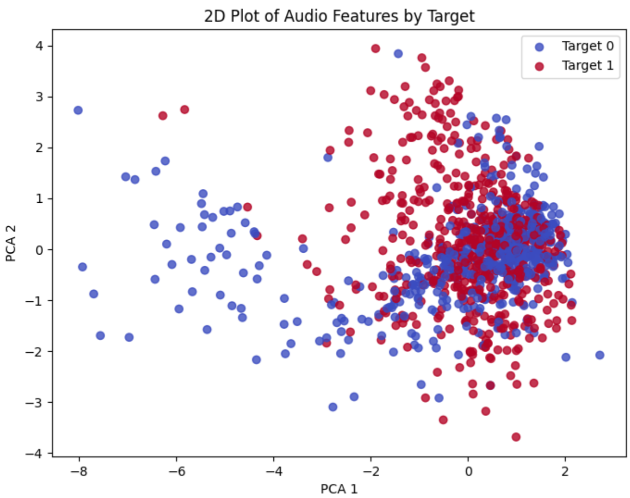



13 features

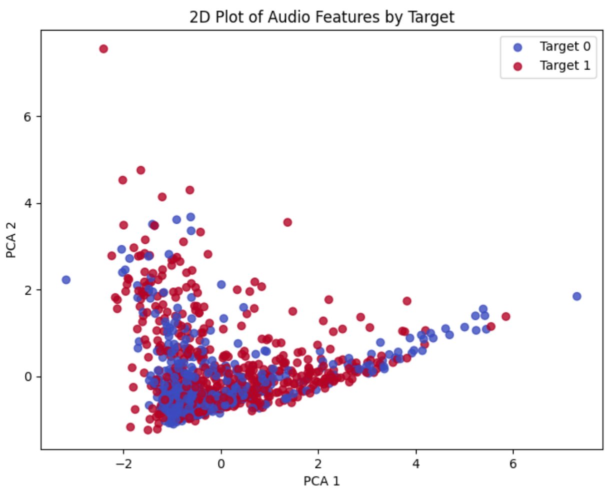



10 features

If you observe the 2 images they are almost identical with slight changes over the edges. However if you observed the cluster the red and blue are not easily separable resulting in the lower accuracy.




7 features




4 features

If you observe these 2 images they are quite different fro the above 2 images but still there isn't a clear separation between target 0 and target 1.

## Conclusion

This analysis of the audio features of songs and their correlation to chart success reveals critical insights into what makes a song a "hit." By examining 13 key attributes extracted from a Spotify dataset—ranging from energy, danceability, and instrumentalness, to loudness and valence—the study aimed to understand the factors contributing to a song's commercial performance. Using advanced machine learning algorithms like Random Forest and Gradient Boosting Classifiers, we were able to assess the predictive power of these features.

The findings suggest that certain attributes—such as instrumentalness, loudness, and speechiness—have a noticeable positive correlation with a song's likelihood of becoming a hit. Interestingly, liveliness and tempo appeared less significant in this context. While the classification accuracy of the models was around 77-78%, it is clear that no single feature alone can determine success. Instead, the interplay of multiple features contributes to the final outcome. This is further evidenced by the decrease in model performance when the number of features was reduced, emphasizing the importance of a holistic approach to feature analysis.

One notable limitation was the inability to improve model accuracy significantly through feature reduction. This suggests that while some features are more important, others still contribute enough to justify their inclusion. Additionally, the results from PCA (Principal Component Analysis) demonstrate that, even with feature reduction, clear distinctions between hit and non-hit songs were challenging to achieve, pointing to the complexity of musical success and its subjective nature.

Future research could extend this analysis by including more nuanced variables, such as listener demographics or social media trends, to further refine predictions. Additionally, experimenting with deep learning techniques, which may better capture the latent patterns within musical data, could offer improvements in predictive accuracy.

In conclusion, while this study provides valuable insights into the characteristics of chart-topping music, it also highlights the complexity of predicting musical success. Producers, artists, and marketers can leverage these findings to optimise song attributes, but they must also remain aware of the inherent unpredictability in what resonates with audiences. Ultimately, a combination of artistic creativity and data-driven strategy will likely be the most effective path to creating a hit song.

Github link : https://github.com/GHAUTHAM2509/Spotify_top50_analyser.git