CPSC 66 Final Report:
Determining what makes a song popular on Spotify

Brian Xiang BXIANG1@SWARTHMORE.EDU
Orhun Kolgeli OKOLGEL1@SWARTHMORE.EDU

Abstract

The music industry nowadays is fully digitalized and has an immense amount of content. There is more content on any one streaming service than humanly consumable, therefore it is important to be able to identify popular songs, artists, users, and streaming services.

We explore the possibility of determining song popularity among several different sub-genres. Popularity is predicted from danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveliness, valence, tempo, and duration using several machine learning models. In our findings, we determine that certain genres are easier to predict popularity than others. We suspect that the qualities of specific genres are more quantifiable than others.

1. Introduction

1.1. Motivations:

The music industry has undergone a paradigm shift with the major distributor of music shifting from large record label companies to online streaming services such as Spotify. With the digitization of the industry, artists nowadays have the potential to spread their work to millions of listeners. However the expansion of the music industry will inevitably cause a data deluge problem where there is more content on any one streaming service than humanly consumable. Data deluge clogs up recommendation systems, causes a decrease in user retention, and increases storage requirements. Therefore it is important to be able to identify songs that are ”good” for both users and streaming services.

Theoretically the increase in the number of songs should decrease the variance in the qualities of songs that are ”good” within a sub-genre. We therefore believe in the possibility of identifying recurring qualities that shape a ”good” song. In our experiments, we use popularity as a measure to determine how ”good” a song is.

Our goal is to identify the key factors that make a song popular or unpopular on Spotify, utilizing the 30000 Spotify Songs dataset from Kaggle. We explore the possibility of determining song popularity among several different sub-genres. We believe bieing able to predict song popularity will have impact on music artists, users, and streaming platforms. An understanding of listener preferences is one of the top considerations for musicians that want to reach a wider audience. Such an understanding would also immensely help streaming platforms and listeners alike in offering song recommendations. This would greatly improve the music streaming experience of users.

1.2. Dataset:

The 30000 Spotify Songs data set (Arviddson, 2023) offers a wide range of information about tens of thousands of different songs. Popularity is predicted from danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveliness, valence, tempo, and duration using several machine learning models such as KNN, Decision Tree, Boosting, Logistic Regression, Random Forest, Linear SVM. Acousticness accounts for how little the sounds have gone through post processing. Danceability informs how suitable a track is for dancing. Energy represents a perceptual measure of intensity and activity. Valence describes the musical positiveness conveyed by a track. Tempo estimates the beats per minute (BPM) or pace of the track. Loudness measures the average level of decibels (dB) for a song. There is also non-musical features included in the dataset such as artist, album, and release date.

2. Methods

We processed the data by categorizing songs, labeling them as either popular or unpopular and then balanced the dataset using these labels. We analyzed feature correlations, measured accuracies of numerous machine learning models, and examined feature weights.

CPSC 66 Machine Learning Proceedings, Swarthmore College, Spring 2021.

2.1. Data Processing

2.1.1. Labeling

We separated songs based on sub-genres because we reasoned that different genres cater to a different set of preferences. We then labeled all songs with a popularity score of greater than 75 as popular. Songs with a popularity score of 75 were within the top 10% of songs on Spotify, so we decided that 75 would be an appropriate cutoff for determining song popularity.

2.1.2. Dimensionality Reduction

We first filtered out unimportant features such as key. The key, the main pitch, of a song is very unlikely to have any impact on how much listeners will like that song. Any song can be reproduced in any possible key while preserving the relative distance of the notes from each other, and these distances are likely what makes a song popular, not the main pitch.

2.1.3. Normalization

We also normalized numeric features and balanced out the dataset to have an even number of labels that are popular and unpopular to help us measure how good our models are performing.

2.1.4. Data Balancing

As a result of the way we are labelling the data, we have unbalanced data. We solve this with two approaches: we can under sample the label with more data or we can over sample the label with less data. Under sampling is done without replacement as we do not want to over represent underlying variables and over sampling is done with replacement because we want to artificially create more data. We understand that over sampling has the risk of over representing specific examples and so all analysis done on the over sampled data is taken with a grain of salt though we hope over sampling mimics a future where ”good” songs are less variable.

2.2. Exploration

We examined the relationship some features have with popularity and plotted a correlation heat map to identify features that are codependent.

We also used unsupervised learning such as k-means clustering to search for any trends and examined feature weight of specific models to better understand how popularity is determined.

2.3. Models

We used scikit-learn’s implementation of KNN, Decision Tree, Gradient Boost, Logistic Regression, Random Forest, and Linear SVM and hyper parameterized them in 5-folds.

We run each model 5 times with the best hyper parameters and find the average to get the accuracy of each model

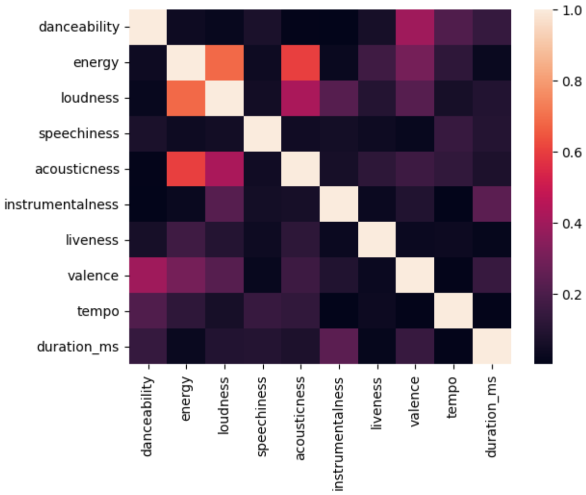
3. Results and Discussion

After applying k-means clustering with 8 clusters we noticed that there were not many distinctions between the 8 groups. We realized that popularity classification is probably non trivial and that features may be codependent.

3.1. Correlation Matrices

We decided to examine the correlations between each of the features across each genre.

Figure 1. Correlation Heatmap for Pop Songs



From Figures 1 and 4 we examine that these seem to have the most codependent features. Figures 3 and 5 seem to be the group with the next most codependent features. We suspect that the more codependent features, the harder the task of predicting popularity will be and so we hypothesize that Rap and EDM music will be the easiest to predict popularity.

Figure 2. Correlation Heatmap for Rap Songs

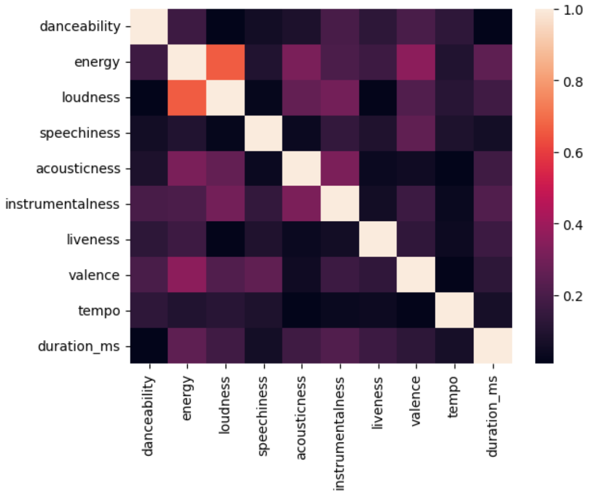


Figure 3. Correlation Heatmap for R&b Songs

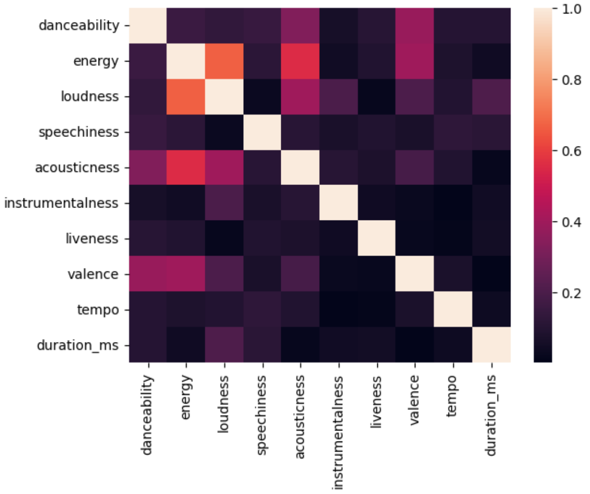


Figure 4. Correlation Heatmap for Rock Songs

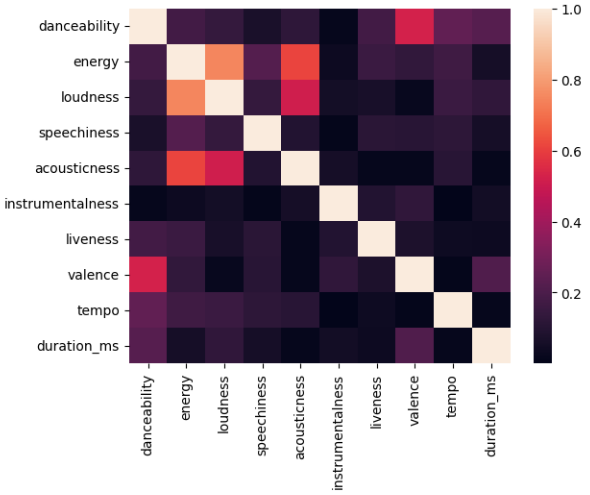


Figure 5. Correlation Heatmap for Latin Songs

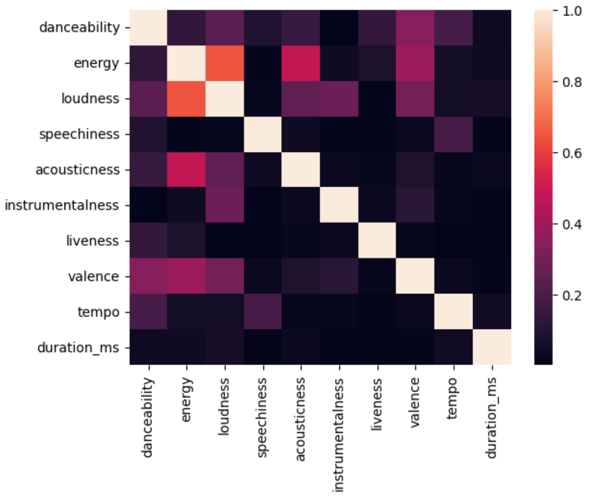
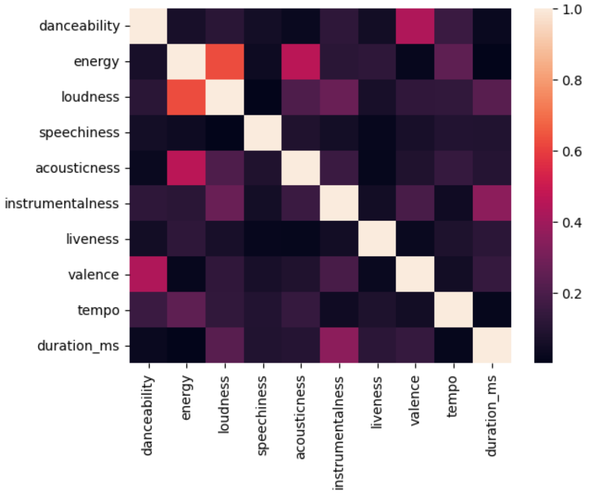


Figure 6. Correlation Heatmap for EDM Songs



3.2. Accuracies

We examine in Table 1 that EDM performs the best across all classifiers. This confirms half of our hypothesis. Rap is seemingly not much easier to predict than Rock or Latin, implying that correlation between features is not the only factor affecting predictor performance. Alternatively, as we expected, Pop and Rock were the hardest genres to predict. We suspect that rather than our hypothesis being completely wrong, Rap may have external factors in determining popularity that are not encompassed within the data set, making it an exception.

Next we examine Table 2. As expected all classifiers perform substantially better across all genres. Since over sampling causes many training and testing examples to be re-

Table 1. Accuracies of Each Classifier for Each Genre with Under Sample Balancing

Genre	KNN	Linear SVM	Logistic Regression	Decision Tree	Random Forest	Gradient Boost
Pop	0.60	0.64	0.65	0.63	0.67	0.66
Rap	0.63	0.68	0.67	0.64	0.68	0.67
Rock	0.58	0.60	0.61	0.62	0.63	0.62
Latin	0.58	0.67	0.65	0.65	0.68	0.68
R&b	0.64	0.71	0.70	0.69	0.73	0.73
EDM	0.70	0.77	0.77	0.74	0.78	0.77

Table 2. Accuracies of Each Classifier for Each Genre with Over Sample Balancing

Genre	KNN	Linear SVM	Logistic Regression	Decision Tree	Random Forest	Gradient Boost
Pop	0.77	0.65	0.65	0.77	0.85	0.88
Rap	0.87	0.68	0.68	0.82	0.90	0.95
Rock	0.90	0.62	0.62	0.77	0.88	0.96
Latin	0.80	0.68	0.68	0.81	0.87	0.91
R&b	0.83	0.72	0.71	0.85	0.90	0.92
EDM	0.92	0.77	0.77	0.92	0.95	0.97

peated, this implies that in order for classifier performance to improve, current songs need to share more similarities to decrease variance. With hyper parameter tuning, we suspect the under sampled models have low bias and high variance, causing lower accuracies.

3.3. Feature Importance

From our models, we identified several features that are seemingly more important than others. We explored the feature importances for EDM, since the genre was the easiest to predict popularity. In the EDM genre, liveliness, energy, and danceability were the most consistent predictors for popularity. The classifiers for which we can examine the feature importances (Linear SVM, Logistic Regression, Decision Tree, Random Forest, Gradient Boost) generally considered these features to be the most important. We therefore believe these to be the most important factors for the EDM genre. For other genres, the trends become less clear and so we are less confident about making conclusions.

4. Social Implications

4.1. Ethical Stakeholder Analysis

4.1.1. Key Stakeholders

The key stakeholders are artists, streaming platforms and users of these platforms, although we would like to acknowledge that this certainly is not an exhaustive list.

4.1.2. Power Dynamics

We believe that being able to predict song popularity and therefore a deeper understanding of listener preferences may help musicians reach wider audiences.

This ability may also benefit streaming platforms since they would be able to make better recommendations to their users, which, in turn, may increase their ad revenue and potentially the number of monthly users.

Finally, listeners can benefit from the aforementioned better song recommendations as well, which would enhance their streaming experience.

Overall, this work appears to empower stakeholders more than it disempowers them, but we acknowledge that it is possible that a stakeholder analysis from a different point of view might come to a different conclusion.

4.2. Concerns

We have concerns that in the real world, the power to predict song popularity might make artists more inclined to use a set of features that would make the song more popular, which, we believe, might decrease the amount of creativity that is involved in the process of making a song.

5. Conclusions

Overall we conclude that predicting song popularity is difficult and does not necessarily have a clear answer. It is likely that song popularity cannot be fully determined. Although some features such as energy, danceability, and liveliness may be strong indicators towards popularity, ultimately the qualities of a song cannot be easily quantified.

With that said, we found some success in certain genres such as EDM where the important qualities are easier to quantify. We hope this work confounds general trends in each of the explored genres and informs artists on the qualities of popular music. We believe that majority of our error comes from high variance. Therefore in order to improve classifier performance, we suggest either taking a different approach towards classification to limit the scope of the problem so that songs are less variable, or to refine the features to be more independent. We do not suggest using more features as this would increase variance.

References

Arviddson, Joakim. 30000 spotify songs, 2023. https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs.