Ipsita Bisht

Professor Marcela Alfaro-Córdoba

STAT 155

9 June 2025

Spotify Dataset Exploration Report

Introduction

Music today has been deeply ingrained into our daily lives, from listening on our phones, on the radio, or going to shows. As a frequent listener of music on platforms like Spotify, I have been curious about musical features and the unique qualities of each genre which is what motivated me to explore the topic for this project. How do pop and jazz songs differ from each other and what makes one song popular in its genre compared to another. These are the types of questions that will be explored in this project with the use of a Spotify tracklist dataset containing a multitude of audio features like ‘danceability’, ‘loudness’, ‘valence’, etc. The three guiding questions for this research are as follows:

1. What audio features best predict a track's popularity within specific genres?
2. How do these success factors differ across genres?
3. Could track lists that deviate from the typical features of their genre still be successful?

Data Collection, Processing, and Exploration

To source the data, I first explored using the Spotify API to extract information from my own listening activity and then extract the audio features using the HTTP endpoints. However, after learning that the endpoints for extracting audio features was deprecated, I resorted to using the Spotify Track dataset from Kaggle. This dataset contained track information such as artists, album name, track name etc. It also contained information such as ‘danceability’, ‘energy’,

‘loudness’, ‘mode’, ‘speechiness’, ‘acousticness’, ‘instrumentalness’, ‘liveness’ , ‘valence’, and tempo’ which contained values from 0 to 1. In order to use this dataset for classifying songs as popular or not based on a hit score of 75 or above, I added in a new column called ‘is_hit’ that contains 0 or 1 depending on if it is a hit (1) or not (0). After creating this column, I removed duplicates from the dataset.

To understand the relationships between the data variables and understand what I need to explore to answer my questions, I used three different kinds of graphs. To understand the popularity density, I created a classic histogram visualization to see where most songs lie on the popularity scale which helped with narrowing down on a popularity threshold value for the ‘is_hit’ column. Next, since my exploration was focused on popularity based on features within genres, I wanted to explore what the data looked like across a couple of the features to see if there were any positive or negative relationships. Based on the findings discussed in the eda.qmd file, ‘danceability’, ‘energy’. ‘valence’, and ‘accousticness’ weren’t the best indicators of popularity of a song which further reinforces the idea that there are a lot of cross-feature interactions that need to be taken into account, not just singular feature relationships. Then I created the correlation matrix to map out which features were highly correlated to each other. Across most of the features, there weren’t features that were extremely strongly correlated which helps to ensure that the data is not redundant.

Modeling

My questions centered around popularity across genres and due to the number of genres (125) in the dataset, I decided to focus mostly on the top genres with most number of songs as well as genres that I found to be interesting and potentially have different genre characteristics: 'jazz', 'country', 'rock', 'dubstep', 'pop', 'heavy-metal', 'bluegrass', 'soul', 'reggaeton', 'house', 'techno', and

'k-pop'. The goal with modeling was to train the model on multiple music features ('danceability', 'energy', 'valence', 'acousticness', 'speechiness', 'instrumentalness', 'tempo', 'loudness') across the genres mentioned earlier. In order to do this, I first used one-hot encoding to categorize the genres as discrete numeric values. The model performance was not ideal due to the sampling size for the minority class (‘is_hit’). As a result, an oversampling method from the Imbalanced Learn library was added to the preprocessing steps after encoding the data. This step increases the volume of hit songs for the model to train on. This increased the precision, recall, and f1-score where it was initially around 0.09 - 0.12 and now 0.23-0.27. These findings are still not the best, which is why I believe the ‘is_hit’ threshold can be adjusted to be a little more flexible and be around 60 instead. That way there can be more hit songs for the model to train with that aren’t artificially created with the oversampling method.

The predictive model alone cannot answer all the questions listed so that is where the Permutation Importance method was introduced. This method reshuffles the features and retrains the model (the artificial neural network model mentioned earlier) to see if there is an impact on the model’s performance. As a result of running this method on the genres mentioned above, the unique feature fingerprint for each genre can be interpreted and used to define what features are important for the model’s success in predicting a song’s hit status. Some genres require more features than others, which can be interpreted as some features are useful for the model’s predictive performance while other features create noise.

In order to answer my third question about songs deviating from genre fingerprint, the K Mean clustering model presented as an ideal candidate. The clusters in this model represent the genre norm so from there I can test my question on if songs that stick to the status quo can have a better or worse chance of being successful. Then I created a deviation score which measures the

distance between each song from the genre cluster center. This score helps reflect how different the song is from the genre norm. Using density graphs, I analyzed the distribution of deviation scores for both hit and non-hit songs within each genre to assess whether uniqueness correlated with popularity or not. My findings varied across the genres where there were outliers for either hit or non-hit songs based on the tail of the distribution, however they did follow a similar pattern of having overlapping distribution curves. This reflects that even if the songs deviated from the center (more unique), that didn't make a difference in popularity score since both the hit and non hit curves were pretty much the same.

Monte Carlo Simulation

During the modeling phase, one challenge emerged where the dataset did not contain a lot of popular or hit songs. This leads to class imbalances which skews the models performance. This challenge is what inspired the design of this experiment so that there can be a clear understanding of how class imbalances impact model performance. To evaluate the varying proportions of hit songs, I decided to compare the results of the experiment across two different models, ANN and Random Forest, and look into metrics like f1-score, precision, and recall. I generated a dataset for each hit proportion level of 1%, 20%, and 55% and each dataset contained 5000 samples. At each level, I ran 20 simulations per model and then collected the average values for precision, recall, f1-score, and accuracy.

Accuracy remained pretty high for 1% and 20% levels due to the prevalence of non-hit songs. Precision tended to be low meaning many of the predicted results ended up being false positives. Recall increased with a higher level of hit songs which shows how sensitive it is to true positives. This experiment reveals the limitations involved with working with an imbalanced dataset and how it can contribute to misleading model performance metrics. With the support of

oversampling methods, resampling, and training splits, the performance can be improved on which can be something to note for future experimentation.

Results and Ethical Dilemma

This project explored three core questions at the intersection of audio features, genre identity, and song popularity. Reflecting on our first question which aims to explore the best features for predicting success of a song within a genre, our results show that predictive features vary across genres. For example, the model's predictive capabilities depend on loudness, instrumentalness, and acousticness for jazz, while danceability and energy were more influential for pop. To answer how those success factors differ across the genres, we can see from the permutation importance graphs how every genre relies on a different mix of features at different levels of influence as well. This further emphasizes the importance of genre-specific feature analysis when analyzing musical data. And for our last question, I wanted to explore how song deviation can impact a song's performance within the genre. The data suggested that song deviation or 'weirdness' did not specifically influence if a song would be a hit or not based on the deviation distribution plots.

My project's goal was to simply explore the characteristics of genres within the Spotify dataset however it can raise a couple ethical concerns. For starters, in a field where creativity is valued, using an algorithm to artificially predict song success and using that as a producer can be viewed as going against what music is known for - self-expression. No algorithm can replace those values and experiences an artist uses to craft their music. With the use of a predictive model measuring popularity, artists can feel pressured to conform to certain sounds or audio features associated with their respective genre. There can also be a risk of bias where songs that are considered popular will overrepresent certain genres and impact chances for other sounds or

artists to emerge in those spaces. This tool can be used as an exploratory tool to understand music but not for music creation. By understanding these ethical concerns, we can responsibly use these tools to understand the music we listen to rather than limit creative expression.

Conclusion

As a result of the work conducted throughout the quarter, I have learned how to take a simple question and turn it into a meaningful collection of data to support my curiosity. From sourcing the data, visualizing relationships, creating predictive models, and simulations, I have explored and answered my questions regarding genre specific characteristics and popularity. My findings helped guide me to explore tools to further improve my model's performance and understand how to work with imbalance data. Aside from the analytical side of the project, this topic and the use of machine learning has broadened my understanding and reflection on the impact of using tools like this in creative fields.