Song Genre and Popularity Prediction

Group members: Alec Patterson 1920-1293, Shalini Patel 1960-9004

Synopsis

The goal of our project is to predict the genre and popularity of a song based on its audio features. A potential use of song genre and popularity prediction is to provide musical artists with information about their work. Popularity prediction could help an artist select songs for an album, event, or a marketing campaign. By knowing which songs are likely to be popular, an artist can focus their attention and resources on the likely popular songs. Genre prediction can also assist musical artists by simply providing the genre classification information. A new artist may not be sure which genre they themselves belong to and could therefore use genre prediction to learn about themselves. On the other hand, an artist may be confident in their genre, but produce a song that is likely to be popular and in another genre. With that information, an artist could market a single song to a new audience.

To accomplish our goal, we will apply data science techniques. First, through data exploration we will review statistical descriptions and data visualizations, paying close attention to attribute correlation to genre and popularity. After reviewing the information gained from statistical methods, we will perform data wrangling. During the data wrangling process, we will seek to reduce the dimensionality, convert attributes to numeric values, and smooth values. With the data exploration and wrangling complete, the next step is to use baseline techniques whose results will be compared to the more complex models' results. The comparison between the baselines and more complex models will allow for evaluation of the complex models' improvement over the baselines. We plan to use both Random Forest and Neural Network models. The models will be trained, parameter tuned, and then evaluated against the baselines.

Details

Dataset

The Spotify dataset from Kaggle provides a comprehensive collection of information about Spotify tracks across various genres. Each track is represented by several attributes that offer insights into the characteristics and categorization of the songs.

The dataset includes fundamental attributes such as track_id, artists, album_name, and track_name, which provide identification and descriptive details about each track. The popularity attribute indicates the track's popularity level on a scale of 0 to 100, calculated based on the total number of plays and recency of plays. This popularity value helps assess the track's overall appeal and current trending status.

The audio features of the tracks play a significant role in understanding their musical properties. Danceability reflects how suitable a track is for dancing, considering factors like tempo, rhythm stability, and beat strength. Energy measures the intensity and activity level of a track, with higher values associated with more energetic and lively music. The key attribute represents the pitch or tonality of the track, allowing for the identification of tracks with similar musical characteristics or user filtering based on preferred keys. Loudness measures the overall volume of a track in decibels (dB), while the mode attribute indicates whether the track is in a major or minor key.

Additional audio features include speechiness, which detects the presence of spoken words in a track and helps distinguish between instrumental and vocal-focused tracks. Acousticness quantifies the likelihood of a track being acoustic, with higher values indicating a greater chance of the track being recorded with acoustic instruments. Instrumentals assess the probability of a track containing no vocals, and liveness detects the presence of an audience in the recording, indicating the likelihood of a live performance. Valence represents the musical positiveness conveyed by a track, with higher values indicating more positive emotions. Tempo signifies the speed or pace of a track in beats per minute (BPM), and time_signature estimates the number of beats in each bar or measure, allowing for an understanding of the track's rhythmic structure.

The track_genre attribute captures the genre to which a track belongs, categorizing songs based on their musical style, characteristics, and influences. This attribute helps organize the dataset by genre, allowing for genre-specific analysis and recommendation systems.

Techniques

Data exploration through statistical analysis may provide valuable information before model training begins. This exploration can include calculating descriptive statistics such as mean, median, and standard deviation for numerical attributes like popularity, duration, danceability, energy, etc. It also involves examining correlations between these attributes using techniques such as correlation matrices or scatter plots to identify potential relationships and dependencies between the variables. This data exploration helps in gaining insights into the dataset's distribution, identifying outliers, and understanding the initial patterns and associations among the features.

Baseline techniques can be used as a reference point to evaluate the performance of the more complex models. This will allow us to identify if the more complex models are providing significantly improved results. For genre prediction, the most frequent genre in the dataset gives us a baseline accuracy for comparison. Another baseline is randomly guessing the genre which can provide an idea of the lower limit accuracy. For popularity prediction we can use the average and median popularity. Both values can provide insight into the common values and ranges that are expected. The final baseline technique we'll use is a decision tree classifier. The basic decision tree model provides a more complex prediction yet is still less complex than the final models. When making predictions, it is prudent to know the most likely results.

After completing data exploration and conducting the baseline techniques, more advanced models will be trained, tuned, and evaluated. A Random Forest model will be employed.

Random Forest combines multiple trees to mitigate the limitations of singular decision trees. There are several key benefits in relation to the chosen dataset. Random Forest can handle both numerical and categorical features and can handle high-dimensional spaces. The chosen dataset has both high dimensionality and both numerical and categorical features.

Another advanced model we will employ is neural networks. Neural networks are effective for a wide range of tasks and provide benefits that work well with the chosen dataset. Neural networks can handle high-dimensional spaces and have the ability to find non-linear relationships.

Evaluation

To evaluate the results, we will use metrics such as accuracy, precision, confusion matrix. Accuracy is simply the proportion of correction predictions out of all predictions and may be skewed if our classes are imbalanced. Precision is the proportion of true positives predictions out of all positive predictions, which makes it more appropriate than accuracy if the classes are imbalanced. Lastly, through confusion matrixes, we'll produce breakdowns of the predictions for each class. This may show our models predict some genres more accurately then others.

Novelty

Most data science applications to song databases are typically focused on creating recommendations of songs that the user would enjoy. In contrast, this project is focusing on empowering musical artists to learn more about themselves, the art they create, and to assist their commercial success.

Dataset: 🎹 Spotify Tracks Dataset | Kaggle