Spotify Trend Analysis Final Report

Jahneulie Weste, Grecia-Melany Morales, Gabriel Carson, & Sean Ward

Introduction

This project focuses on understanding what makes a song popular on Spotify within individual countries and the world. First, we planned to look at musical features and see which were the most strongly correlated to a song's popularity in different regions. After examining the dataset, we saw that the audio features did not clearly explain what makes global success as closely as we expected. Therefore, we changed direction and built a machine learning model to predict whether a song would become a global hit or not.

Objective

The original objective of this project was to determine which musical features are strongly correlated with worldwide song popularity and how these features differ across geographic regions. The objective pivoted to determining which features determined whether a country would be a global hit through a machine learning model and displaying data trends in a Flask dashboard.

Data

The data used was the "Top Spotify Songs in 73 Countries (Daily Updated)" dataset, obtained from Kaggle. The dataset contains approximately 2.1 million records, representing the daily top 50 songs across 73 countries during the years of 2023, 2024, and 2025.

We processed the data by removing records with missing values. The only column where missing values were expected was the Country column, because a null value meant the song was a Global Hit. To determine a song's global hit status, we condensed songs with multiple information into one row summarizing that song's data. For example, if a song was a global hit and a country hit we labeled it as a global hit.

We also used the features provided by the dataset to create derived features such as number of positive and negative weeks on the charts, worst and best rank, average movement across the charts, etc. These derived features were very important for our results and overall model performance. We validated a subset of the data to establish trust and credibility when using it by using Spotify's API to verify the song's features.

Models

We used a random forest classifier for our machine learning model. Our team was familiar with this model and believed it would provide a strong model for our use case of determining whether a song would be a global hit. We also knew that the important features attribute within the RandomForestClassifier would allow us to see what features actually contributed to a global hit's success. Our model had about a 98% validation accuracy overall.

Other analyses provided on the dashboard are: missing values, top features, top features, feature importances, top correlated features, streams by country, popularity by region, top genres by popularity, and many others.

Results

From our trend analysis, we discovered that the specific audio features of the songs (e.g. danceability, tempo, mode, etc) were not as important as we expected. So, we pivoted to creating a machine learning model with the goal to predict whether a song would be a global hit. We used a Random Forest Model with two outputs: Global Hit or Not Global Hit. We created multiple derived features and found that they better predicted a song's hit status than the original features of the dataset. We used the Random Forest Classifier's important features attribute to determine the most important features for prediction.

We found that the top five important features were the following: number of positive weeks on Top 50 charts (18.8%), popularity (18%), average movement (9.3%), worst rank (8.8%), and negative weeks (8.6%). We found it interesting that some of the top five important features are negative indicators such as worst rank and negative weeks. Upon further investigation through a confusion matrix and accuracy scores, we found that our model performed better at determining that a song would not be a global hit instead of if it would be. The chart below compares the accuracy scores for precision, recall, and f1-score.

Global Hit?	Precision	Recall	F1-Score
False	0.99	1.00	0.99
True	0.86	0.72	0.79

Primary Issues Encountered During the Project

We did not encounter any difficult issues when conducting the project. Since the Kaggle dataset was pretty thorough, we were able to prepare it for our analysis without seeking any outside information. When we discovered that the song attributes were pretty insignificant, we had to pivot to make something more meaningful like the machine learning model. We did run into one hiccup when running the dashboard application, when the data failed to load on some member's computers. We were able to troubleshoot this problem.

Reflecting on the process, ways we could have mitigated these issues before they started were to improve code documentation to give all team members ample resources to know how to run the software without running into issues. With more documentation along the development process we could provide more insight for both contributors and users upon gaining access to the project files.

Future Work

Based on our findings that the song's attributes are not that important for a global hit, it would be interesting to determine if they play a role in becoming a country's or region's local hit.

This could be determined by comparing the attributes of the songs found on the Top 50 list and comparing it to similar songs that did not make the charts. A local evaluation may find that the song's attributes are important in popularity or at least identify trends in the country's interest.

Organizational Chart

- September - Mid-October:
  - Sean Ward conducted EDA on the dataset.
- Mid-October - Mid-November:
  - Jahneulie Weste created the machine learning model.
  - Gabriel Carson developed the frontend dashboard.
- Mid-November - Early-December:
  - Grecia-Melany Morales created the final presentation and worked on the final report.