DSME5110 (IT)

Statistical Analysis

Project Report

Bombarding the Popularity:

Unveiling the Secrets Behind Hit Songs

Group 7
LI Yunran 1155201157
LIN Rui 1155208160
LIANG Huixiu 1155201557
Lau Cheuk Ki Jake 1155201163
YU Yangyang 1155201568
LIN Dengdeng 1155208104

1. Introduction

Music permeates every corner of society, with diverse preferences among individuals. This prevalence leads to an extensive array of music genres and an abundance of songs for listeners to explore. Despite this vast selection, only a few songs achieve near-instant recognition among the majority. As enthusiasts of music, we are intrigued by the mechanisms behind a song's rise to popularity and what sets certain songs apart, making them significantly more popular than others. Additionally, we are curious to explore whether there are common characteristics among these widely acclaimed songs.

In the dynamic realm of music streaming, the popularity of songs on platforms like Spotify is influenced by a myriad of factors. Understanding these factors is crucial not only for artists and record labels but also for music enthusiasts, researchers, and industry professionals. This report delves into the multifaceted landscape of Spotify song popularity, exploring the diverse elements that contribute to the success and resonance of top songs on this global platform. Through a comprehensive analysis, we aim to uncover patterns, trends, and insights that illuminate the intricate web of factors shaping the contemporary music landscape on Spotify.

We collected our data from Kaggle, the world's biggest data science community, using the data set "*Best and Worst Spotify songs from 2000-2023" (Eden, 2023)*[1]. This data set contains full-range popular songs in this time period. Popularity is our core dependent variable, while dB, energy, valence and other 9 features (see Appendices 1) in total are our independent variables.

2. Research Questions

- What characteristics do popular songs possess each year?
- How does the popularity of songs in different genres change annually?
- Which characteristics are associated with highly popular songs, and do these traits impact the popularity of songs in different genres differently?
- Is it possible to develop models to predict the popularity of songs?
- Can we further construct a model to recommend similar songs to users.

3. Data Description

3.1. Descriptive statistics

Based on the descriptive statistics shown in Appendix 1.2, the data of popularity, energy, danceability, liveness, valence, and acousticness are all within the range of 0 to 100, with standard deviations mostly greater than 15, indicating that the data distributions are relatively discrete, which ensures data diversity. However, the means lack representativeness. The std of speechiness is 11.4, denoting that its data is concentrated compared to the previous variables. Bpm ranges from 41 to 248,

[1] Eden, C. V. (2023). *Best and Worst Spotify songs from 2000-2023*. Kaggle. https://www.kaggle.com/datasets/conorvaneden/best-and-worst-spotify-songs-from-2000-2023/data

duration ranges from 29 to 688, and dB ranges from -26 to 5. It is worth mentioning that the std of dB is only 3, which means distribution is relatively focused.

3.2. Yearly Change in Features of Top 50 Popular Songs

What characteristics do popular songs possess each year? To solve this problem, we select the top 50 songs of each year, calculate the average values for 9 features respectively, and draw a line graph that shows how public preferences change over time (Appendix 2.2.4).

From a general perspective, the clear downward trends of valence, energy, and dB are probably due to the increasing social pressure in recent years, under which people prefer songs with a depressed and quiet style. It can also be inferred that the economic downturn leads to a return to traditional beliefs, causing an increase in acousticness (non-electronic). Danceability exhibits a slight upward trend, which could be attributed to the popularity of K-pop and dance challenges on Tiktok. In the fast-paced era, people tend to enjoy short videos and certainly short songs, which may lead to a decreasing tendency in duration. 2020 witnesses an obvious turning point - the outbreak of COVID-19. We can find significant fluctuations in people's music preferences at this time, across several graphs.

3.3. Differences of popularity in different genres

For further exploration of the data, we raised another question: How does the popularity of songs in different genres change annually?

We grouped our original data of over 500 genres into 9 major genres based on keywords, calculated average popularity by genre, and observed the trends over years. For instance, genres containing "hip" or "rap" are reclassified as "Hip Hop/Rap". The WordCloud (Appendix 2.2.5) provides a clear visualization of the distribution of data grouped by new genres. Apparently, Electronic/Dance and Hip Hop/rap occupy the majority, both containing more than 1000 pieces of data, followed by pop and rock. Latin remains the most popular for many years. Next comes pop, R&B and country folk, which experience stable trends. The popularity of rock, Electronic/Dance and Hip Hop/Rap decline year by year, despite a dramatic growth from 2020. Noisy music becomes less popular in general. Jazz always lacks popularity, but also becomes more and more popular in recent years. All categories of songs have been trending upwards since 2020. So, we can infer that the pandemic reduced other forms of entertainment and increased people's time spent listening to music.

3.4. Relation between 9 features and popularity

According to "relationship between 9 features and popularity" (Appendix 2.2.3), energy, danceability and dB are positively related to popularity, while speechiness, acousticness and liveness are negatively related to popularity. It seems like that bpm, duration, and valence have no significant correlation with popularity.

The correlation Matrix (Appendix 2.3) shows a comprehensive relation between all variables. Color and value in every cell represent the correlation coefficient between 2 variables, which ranges from -1 to 1. We can see that energy has a strong positive correlation with dB (0.72) and a strong negative correlation with acousticness (-0.58). dB has a slight positive correlation with danceability (0.13), a strong negative correlation with acousticness (-0.52). Duration and popularity have low correlation with most other features.

It is apparent that dB has strong positive correlation with energy (0.72), and valence is positively related to energy (0.32), danceability (0.34), dB (0.21). Besides, dB is the variable with the strongest correlation (0.31) with popularity in the table, so we choose it as the main variable temporarily. The three features that have the highest correlation with dB are energy (0.82), danceability (0.13), valence (0.21), so it's worth exploring the relation through Lowess. It is our common sense that the larger the dB, the higher the song energy. Music under high dB also arouses a desire to dance. Results show that dB has strong positive correlation with energy, and slight positive correlations with danceability and energy, which means we should handle the multicollinear problems in the modelling parts.

3.5. Hypothesis

Based on the analysis above, we expect that bpm (0.00) and duration (0.01) are not related to popularity; dB (0.31), danceability (0.11), energy (0.07) and valence (0.05) are positively correlated to popularity, while liveness (-0.05), acousticness (-0.19), and speechiness (-0.07) are negatively correlated to popularity, shown in the causal map in Appendix 2.4.

4. Data Analysis and Results

4.1. Baseline Regression

This research used the original data without logarithmic or standardized processing. The decision was made because the dataset was already processed by the provider, with most variables ranging from 0 to 100 and having a relatively stable data span. The benchmark regression method used in this study was multiple linear regression.

Based on the research hypothesis, the regression equation was formulated as follows:

$$Popularity = \sum_{i=1}^{9} \beta_i feature_i + \varepsilon$$

The regression analysis employed multiple linear regression with the hypothesis that quantifiable music indicators can explain music popularity. The model had a significant P value and an R-squared value of 16.8%, indicating practical significance and moderate explanatory power.

Variables such as bpm, danceability, and duration were found to have P-values exceeding 0.1 and were eliminated due to multicollinearity. The remaining variables had P-values below 1% and were considered significant contributors to music popularity.

Positive coefficients were observed for "danceability", "dB", "valence", indicating that people prefer songs with rhythm, high volume, and positive emotions. Negative coefficients were found for "liveness", "acousticness", "speechiness", and "energy", suggesting that popular songs tend to have clear recordings, moderate lyrics, and lower energy. However, the negative coefficient for energy raised doubts, suggesting a potential quadratic relationship with popularity, where an optimal level of energy generates excitement but excessive energy might have a negative impact on listeners' emotions.

The findings suggest that quantifiable music indicators play a role in determining music popularity, with factors like rhythm, volume, and emotional tone influencing listener preferences.

4.2. Preliminary Study

4.2.1. Introduce the Squared Term of Energy

After including the energy square term in the analysis, it was observed that the coefficient was significantly negative, indicating that the relationship between energy and popularity is not linear. This suggests that the impact of energy on popularity follows an inverted U-shaped curve, with an inflection point where the effect starts to decline. The wide variation in energy levels in music and the subjective nature of the criteria used to judge it contribute to this finding. Additionally, there may be a diminishing marginal benefit in terms of the intake of "positive emotions" when it comes to music popularity.

4.2.2. Group Regression based on Songs Type

We classified songs into 8 main genres: Country/Folk, Electronic/Dance, Hip Hop/Rap, Jazz/Blues, Latin, Pop, R&B/Soul and Rock. Electronic/Dance songs, which are highly popular, showed similar coefficient patterns as the baseline model, but shorter durations were preferred. But R&B/Soul songs didn't require excessive volume or rhythmic elements. Rock songs followed a U-shaped pattern similar to the main model.

These findings highlight the distinct popularity characteristics among different music genres, with Hip Hop/Rap, Electronic/Dance, and Rock showing higher model fit.

4.2.3. LASSO Regression

In the feature selection process, we removed highly collinear variables and non-significant ones. We found that higher volume, high-quality recordings, and moderate lyrical content are associated with greater music popularity. LASSO regression confirmed these findings and identified the key variables as "dB", "acousticness", "speechiness", and "danceability". The combination of manual selection and LASSO improved the accuracy and explanatory power of the model.

4.3. Prediction

For prediction and comparison purposes, data from 2023 was utilized. Both Regression models (Multiple Linear Regression, LASSO Regression) and Machine Learning Models (Decision Tree, Random Forest, Neural Network) were employed. The data was divided into 80% for training and 20% for testing. Parametric tuning and nonparametric tuning methods were used in Machine Learning Models for model training, except the Neural Network. All numerical features were included, along with an additional categorical variable called Modified Genre. Parametric tuning methods yielded the lower RMSE and were used for Decision Tree and Random Forest.

4.3.1. Predictive Performances

4.3.1.1. Regression Models

Based on the Manual Filter in Modelling & Explanation, features dB, liveness, Acoustiness, Speechiness were utilized for Multiple Linear Regression Models (MLR). The model tended to overestimate popularity with actual value below 60 underestimate for values above 60. The MSE and RMSE were 665.09 and 25.79, respectively.

All the numerical variables were included in the LASSO Regression model as it searches for best parameters. Similar to the result from MLR, the overcasting and undercasting regions were comparable. The predicted values were more concentrated within the range 40-60. The performance was slightly worse than MLR, with 686.46 MSE and 26.2 RMSE.

4.3.1.2. Machine Learning Models

The graph of Decision Tree indicated a more diverse distribution compared to the Regression models. There were parallel points within 60 which were the maximum predicted value.The graph for Random Forest and Neural Network were similar and this was supported by their similar MSE and RMSE.

4.3.2. Overview

	MLR	Lasso Regression	Decsion Tree	Random Forest	Neural Network
Amount of X Variables	4	10	9	9	10
MSE	665.09	686.46	615.67	542.75	599.38
RMSE	25.79	26.20	24.81	23.30	24.48

To summarize, the models performed similarly in terms of RMSE. Random Forest exhibited slightly better performance with the lowest RMSE, but it utilized nine song features, making it relatively complex. On the other hand, the MLR only employed four features, demonstrating simplicity. In this case, the simpler the better, the best model for prediction was the MLR model.

5. Songs Recommendation

Music recommendation stands as a pivotal sector within the music industry. Tailoring song suggestions to align with a user's listening preferences is a key strategy for sustaining user engagement. Leveraging existing data, we embarked on a preliminary exploration in this domain.

Limited by the lack of comprehensive features in the dataset and insufficient computing power, we were unable to implement complex models and methods. Consequently, we adopted a simple PCA (Principal Component Analysis) approach to reduce the dimensions of the original data, projecting the information of nine numerical dimensions in the original data into three-dimensional feature space. In this space, we can plot a scatter diagram of the songs in the song dataset, which facilitates the observation of differences in song characteristics. We can measure the differences between songs by calculating the angle between the feature vectors of the song samples or the Euclidean distance between the sample points. For ease of interpretation, we chose Euclidean distance: the greater the distance, the more significant the difference between the songs; the smaller the distance, the more similar the songs.

Thus, we constructed a simple song recommendation algorithm: The feature information of the last song listened to by the user is reduced to three-dimensional information through PCA. The target song is then represented as a point in the feature space. We can find the nearest songs to it by calculating the Euclidean distance in the feature space, forming a recommended playlist.

Taking the song "un x100to" as an example, its vector representation in the feature space is $[1.40046855, 1.29014391, 1.35021049]^T$, marked with a red dot in the feature space. By calculating

the Euclidean distance, the three nearest songs were identified as "I Wonder," "Beige," and "Die for You," marked with purple dots in the feature space. The three farthest songs were "Booty Gang," "Glob Waterfall," and "Bible And A K," marked with blue dots.

After testing and listening, we subjectively evaluated the results not bad. However, due to the limited features considered, there may be significant inaccuracies in the recommendations between songs.

6. Summary and Outlook

In this research, we delved into the attributes that define popular songs across various years, uncovering a distinct downward trend in dB, energy, and valence, coupled with a slight increase in danceability. When examining song popularity across genres, we observed that Latin music consistently maintains high popularity. Conversely, Jazz, with its niche appeal, has not enjoyed widespread popularity, though it's noteworthy that it's gradually gaining broader acceptance in recent years. Genres like Rock, Electronic, and Hip-Hop/Rap are experiencing a decline in popularity, whereas Pop, R&B, and Country Folk exhibit more stability. In terms of characteristics that correlate with high song popularity, we identified a pattern where highly popular songs typically feature high loudness, positive tones, enhanced danceability, and a specific range of high energy. These songs are often recorded in studios, tend to eschew or minimally use electronic tunes, and generally contain fewer lyrics. Our analysis also revealed that factors such as BPM and duration do not significantly influence a song’s popularity. Further, through group regression analysis, we discovered that these features have varying impacts on the popularity of songs across different genres. Employing multiple regression and Lasso models, we extended our research to include machine learning models like decision trees, random forests, and neural networks for predicting song popularity. Additionally, we ventured to develop a rudimentary song recommendation system utilizing the PCA algorithm, aiming to enhance the user experience in music selection.

Nevertheless, it is essential to acknowledge several limitations in our research. Firstly, the quality of our data is suboptimal. The limited data size, the lack of more musical features, and the mysterious popularity index that Spotify directly offers may cause potential problems in our analysis. Adhering to the principle of 'garbage in, garbage out,' we recognize that the insights derived from this dataset may lack robustness. Secondly, our study does not incorporate causal or mechanistic analyses. We have not delved into how specific features influence a song's popularity, presenting an avenue for future exploration. Thirdly, the parameter tuning in our machine learning models was somewhat cursory. We did not engage in extensive fine-tuning, which resulted in the models performing below their potential. These aspects offer significant opportunities for enhancement in future iterations of our research. Finally, our model does not incorporate factors such as the artist or the album, which are crucial to a song's popularity. Users may exhibit a preference for songs by popular artists or from trending albums, rather than focusing solely on the inherent characteristics of the songs themselves. This oversight highlights a critical aspect that could significantly influence the effectiveness of our prediction and recommendations and should be considered in future enhancements of our model.