CIS 3715: Principles of Data Science

Predicting Music Popularity Using Logistic Regression

Tasnia Kader

Apr. 28, 2025

Introduction

Music is a universal language that uplifts, unites, heals, and celebrates. It transcends cultural and linguistic barriers, creating a shared experience for people around the world. With today's technology, music is more easily accessible than ever before, with people all over the world listening to their favorite tunes on YouTube, Spotify, and other streaming services. From upbeat melodies to soothing harmonies, music has found its place in our constant companionship, impacting the moods, social lives, and even identities of people.

While the mass appeal of music cannot be ignored, what drives its popularity is still a complex puzzle. There are hundreds of thousands of songs being put out daily, so it is hard to determine exactly what separates a hit from all the others. Popularity is ultimately a mix of individual preferences, social trends, and, sometimes, luck. It remains a mystery how one song takes off with millions while another does not.

Knowing why a song is popular is significant for a number of reasons. For producers and artists, it can inform the production of music that has a better chance of engaging listeners and becoming successful in a crowded marketplace. With an industry drowning in millions of songs at the click of a button, being in a position to forecast the drivers behind popularity can help artists make their art more sensibly and ensure it finds its way into the hands of more people. It also helps music industry staff, such as streaming organizations and record labels, to market the right people and bring rising stars who could otherwise fly under the radar.

Also, studying the difference between popular and unpopular music enables us to better understand why some music has such a lasting influence on a global scale. If we look at the most significant features of popular songs, we can identify patterns that show the relationship between the music and its audience. This project seeks to forecast the most contributing factors to a song's popularity, providing insights into the factors that make one song stand out while others remain overlooked.

In a similar effort, a study by Niklas Sebastian Jung and Florian Mayer from the University of Innsbruck utilized algorithms such as Ordinary Least Squares (OLS), Multivariate Adaptive Regression Splines (MARS), Random Forest, and XGBoost to analyze song characteristics. This project will complement their work by applying the Logistic Regression model, allowing for a comparison of results and furthering our understanding of the key factors that drive music popularity.

Approach

I decided to use the "Spotify Music Dataset" from Kaggle for my binary classification project. The link for the data is provided in the references. This dataset was collected using Spotify’s API and contains popular and unpopular songs and their associated features. The data includes a total of 14 descriptive and audio features. I considered 4831 total entries.

The table below describes the descriptive features I have decided to include in the project, along with their respective descriptions:

Feature	Description	Data Type
Track Artist	Artist(s) performing the track	object
Genre	Main genre of the track (e.g., pop, rock, etc.)	object
Subgenre	More specific subgenre related to the track (e.g., indie pop, punk rock)	object

The next table represents the audio features for the tracks:

Feature	Description	Data Type
Energy	Intensity and liveliness of the track; energetic tracks are often fast, loud, and dynamic.	float64
Tempo	Speed of the track, measured in beats per minute (BPM).	float64
Danceability	Score for how danceable the track is; based on tempo, rhythm, and regularity.	float64
Loudness	Volume of track; measured in decibels (dB)	float64
Valence	Emotional tone of track; higher values → more positive, upbeat moods; lower values → sadness or anger.	float64
Instrumentalness	Probability that the track is purely instrumental; values closer to 1.0 are tracks with no vocals.	float64
Speechiness	Represents the use of spoken words in the song	float64
Mode	Indicates whether track is in a major or minor key.	float64
Key	Musical key in standard Pitch class notation between 0 and 11	float64
Duration_ms	Length of the track in milliseconds.	float64
Acousticness	Score indicating whether the track is acoustic (1) or not (0).	float64

After selecting the dataset, I used Python to load and preprocess the data. First, I loaded both the high popularity and low popularity data and removed irrelevant features from both. Then, I included a new column called “popularity” where 1 represents the high popularity songs and 0 represents the low popularity songs. Next, I handled missing values by filling numerical features with the median and categorical features with a new "unknown" category. I then applied one-hot encoding to the categorical features, including the track artist, genre, and subgenre columns, to convert them into numerical values. Finally, I plotted the number of popular songs and unpopular songs in my data using a bar graph. The results are shown in Figure 1.

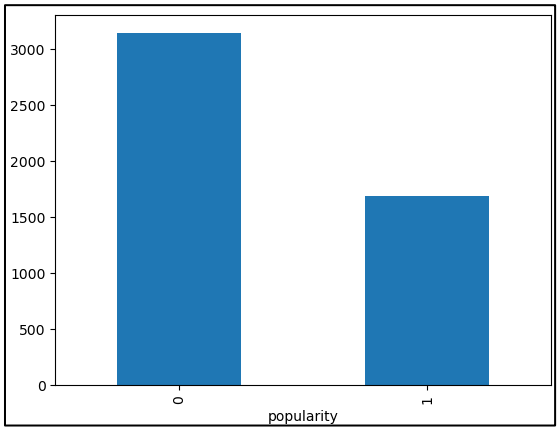

Figure 1 Distribution of High Popularity and Low Popularity Songs

There are significantly more unpopular songs than popular songs, which suggests the data is imbalanced.

I used logistic regression with L2 (Ridge) regularization to train my data. I split the data in the following way: 90% of the data was used for training (4347 entries) and the other 10% was used for testing (484 entries). Then, I normalized the features for both the training and testing data using StandardScaler. I applied nine fold cross validation. The regularization coefficient was chosen from the following set $[10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 20, 50, 100]$. The hyperparameter was chosen based on the best average F1-score. The average accuracy score was not used because the data was imbalanced. Also, since the data was imbalanced, I used class_weight="balanced" for the logistic regression, which adjusts the weights inversely proportional to class frequencies.

Results

Figure 2 shows the regularization coefficients and their corresponding average F1 scores:

```
reg_coeff: 1000.0, f1: 0.805
reg_coeff: 100.0, f1: 0.807
reg_coeff: 10.0, f1: 0.787
reg_coeff: 1.0, f1: 0.763
reg_coeff: 0.1, f1: 0.723
reg_coeff: 0.05, f1: 0.702
reg_coeff: 0.02, f1: 0.666
reg_coeff: 0.01, f1: 0.660
```

Figure 2 Regularization Coefficients and F1-scores

Thus, the best regularization coefficient is 100 with an F1 score of 0.807. I used this hyperparameter to retrain the model and evaluate it on the testing set. Then, I computed the following evaluation metrics: F1 score, recall, and precision. The results are shown in Figure 3.

```
recall: 0.835, precision: 0.813, f1: 0.824
```

Figure 3 Recall, Precision, F1-score for the Best Hyperparameter

All of these values are around 0.8 which indicates that the model fits the data moderately well. The recall score of 0.835 suggests that out of all the truly popular songs, the model correctly identified about 83.5% of them. The precision score of 0.813 means that among the songs predicted as popular, 81.3% were actually popular. Figure 4 shows the results of plotting the model coefficients using a bar graph.

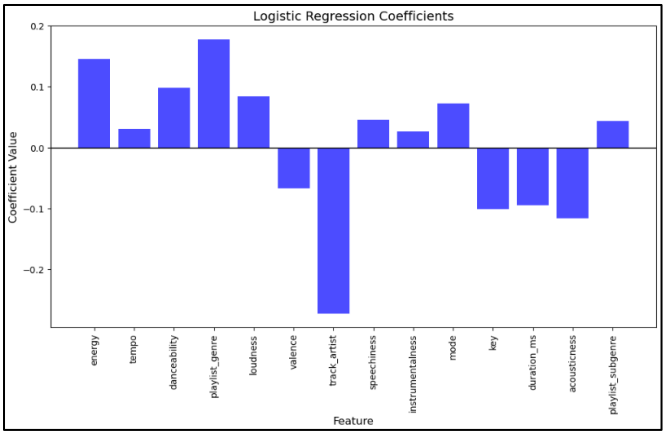

Figure 4 Logistic Regression Coefficients

According to this graph, artist, genre, and energy had the largest influence on predicting a song's popularity. This result aligns with intuition: if an artist is relatively unknown, it is unlikely that their songs will become popular. Similarly, songs belonging to popular genres tend to perform better. With the absence of the artist feature, Jung and Mayer used their XGBoost model to similarly conclude that genre of a song is a large predictor of song popularity. High-energy songs are often played at events like parties and dances, making them more likely to become popular.

On the other hand, instrumentalness, tempo, playlist subgenre, and speechiness were the features that had the least effect on a song's popularity. These features are more related to the internal composition of a song. Jung and Mayer revealed using their Random Forest and OLS models that instrumentalness was the least significant song characteristic. Interestingly, although genre was important, subgenre was not as influential. This suggests that broader genre categories matter more for popularity predictions than specific subgenres.

I also wanted to investigate how the model's behavior would change if the descriptive features (artist, genre, and subgenre) were removed. After dropping these features, the average F1 scores became much lower than before. The new regularization coefficients and their corresponding F1 scores are shown in Figure 5.

```
reg_coeff: 1000.0, f1: 0.580
reg_coeff: 100.0, f1: 0.591
reg_coeff: 10.0, f1: 0.597
reg_coeff: 1.0, f1: 0.596
reg_coeff: 0.1, f1: 0.596
reg_coeff: 0.05, f1: 0.596
reg_coeff: 0.02, f1: 0.596
reg_coeff: 0.01, f1: 0.596
```

Figure 5 Regularization Coefficients and F1-scores

Thus, the best hyperparameter is 10 with an F1 score of 0.597. After retraining the model using this hyperparameter and evaluating it on the testing set, the following metrics were obtained as shown in Figure 6.

```
recall: 0.852, precision: 0.517, f1: 0.643
```

Figure 6 Recall, Precision, F1-score without Descriptive Features

All of these metrics were significantly lower compared to when descriptive features were included, except for recall, which increased. This suggests that without artist and genre information, the model became better at identifying actual popular songs (higher recall) but worse at correctly labeling songs as popular when they truly were (lower precision). Consequently, the model became more biased toward predicting songs as 'popular,' catching more true popular songs but also mistakenly labeling more unpopular songs as popular. I again graphed the coefficients of the features using a bar plot which is shown in Figure 7.

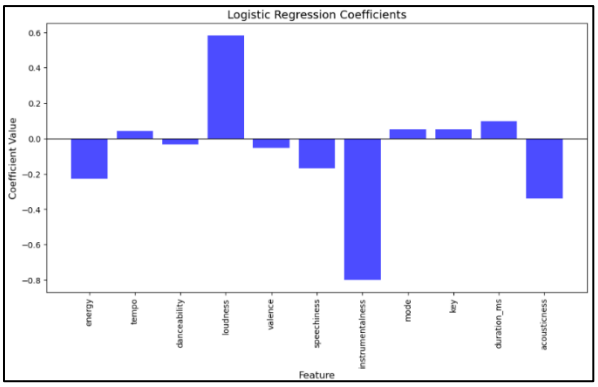

Figure 7 Logistic Regression Coefficients without Descriptive Features

When examining feature importance after dropping the descriptive features, instrumentalness became the most important feature, even though it was previously one of the least significant. Loudness and acousticness also became major factors, whereas they were only moderately important before. This shift makes sense because without descriptive identifiers like artist and genre, the model has to rely solely on the intrinsic audio characteristics of the songs. This also matches the results obtained by Jung and Mayer with their Random Forest model where loudness was ranked the highest in terms of importance. In their OLS model, instrumentalness was the third most important feature behind energy and loudness. Features like instrumentalness, loudness, and acousticness now help differentiate popular songs from unpopular ones, even if they are weaker signals compared to artist or genre.

The least important features after dropping the descriptive ones were danceability, followed by valence, tempo, mode, and key, all of which had similarly low importance. This suggests that these features, although they contribute to the feel of a song, are less predictive of popularity when considered alone. It makes sense because factors like danceability or mode may influence the style of a song but not necessarily its mainstream success without additional context like artist fame or genre trends.

Conclusion

This project aimed to explore the factors that contribute to a song's popularity, using data from Spotify's music dataset to predict whether a song would be categorized as popular or unpopular. By analyzing descriptive features such as artist, genre, and subgenre, as well as audio characteristics like energy, tempo, and loudness, the model identified key predictors of a song's success. Through the application of logistic regression with ridge regularization, the best model achieved a relatively strong F1 score of 0.807, indicating that factors like artist, genre, and energy played a significant role in predicting a song's popularity. Furthermore, the results highlighted that without descriptive features like artist and genre, the model's performance dropped, with increased recall but lower

precision. In this case, audio features like instrumentalness, loudness, and acousticness became the most impactful predictors.

These findings align closely with the results obtained by Jung and Mayer, who also identified genre as a major driver of song popularity and found that, in the absence of descriptive information, audio features such as loudness and instrumentalness gained importance. This consistency across different modeling approaches strengthens the validity of the conclusions drawn.

The outcome of this project provides valuable insights for artists, producers, and music industry professionals on which factors to consider when crafting songs that could resonate with a wider audience. However, the study's limitations, such as the imbalance in the dataset and the exclusion of other potential variables like social media trends or cultural shifts, suggest that there is room for improvement. Future studies could explore additional data sources, such as social media metrics or listener demographics, to create a more complete model. Also, experimenting with more advanced machine learning techniques, such as neural networks or ensemble methods, could further enhance prediction accuracy. While this project offers an initial framework for understanding what makes a song popular, it opens the door for deeper analysis and exploration in the rapidly evolving music industry.

References

Ameh, Solomon. *Spotify Music Dataset*. Kaggle, www.kaggle.com/datasets/solomonameh/spotify-music-dataset.

Jung, Niklas Sebastian, and Florian Mayer. *Beyond Beats: A Recipe to Song Popularity? A Machine Learning Approach.* University of Innsbruck, arxiv.org/pdf/2403.12079.