Feature Analysis in Music

Julie Walsh, Paddy Lavin, Sofia Lambro, Samantha Burkard, Kyle Bridger

December 11, 2025

1 Introduction

This project uses a Kaggle dataset sourced from Spotify's API containing two CSV files: a "high-popularity" subset (calculated popularity ≥ 68) and a "low-popularity" subset (< 68, totaling 4,831 songs with 30 audio, genre, and temporal features. Our analysis was guided by two central questions:

1. *What factors play into how popular a song becomes?*
2. *Which audio or genre-related features are most predictive of a hit?*

To answer these, we implemented a full machine-learning workflow including data cleaning, feature scaling, exploratory analyses, and both regression and classification modeling. Our goals were to identify audio features meaningfully linked to popularity and test whether long-term historical patterns within genres and subgenres provide stronger predictive power.

2 Analysis

2.1 Regression Models

After cleaning, scaling, and exploring the dataset, we compared three supervised models for predicting continuous popularity (Fig. 1): Linear Regression, k-Nearest Neighbors (k=5), and a Random Forest Regressor. Linear Regression assumes a mostly linear relationship; kNN makes local predictions but is sensitive to noise; Random Forest captures complex nonlinear interactions but may overfit. All models used the same preprocessing pipeline (median imputation, standardization, one-hot encoding) and were evaluated on a consistent train–test split.

Using RMSE, MAE, and R^2, Linear Regression performed the weakest, indicating that popularity is not well approximated linearly. kNN performed moderately but inconsistently, suggesting instability in high-dimensional space. Random Forest achieved the lowest RMSE (13.91) and was selected as the base model.

	model	mse_train	mse_test	rmse_train	rmse_test	mae_train	mae_test	r2_train	r2_test
2	RandomForestRegressor	29.290002	193.690540	5.412024	13.917275	3.993330	10.263686	0.926282	0.485101
0	LinearRegression	268.725073	256.318596	16.392836	16.009953	12.640921	12.420396	0.323668	0.318613
1	KNNRegressor	199.387236	275.177226	14.120455	16.588467	10.561439	12.223602	0.498179	0.268479

Figure 1

To justify hyperparameters, we used GridSearchCV with 3-fold cross-validation. We tuned the number of trees (50, 100, 200) and maximum depth (None, 10, 20). The best combination (Fig. 2) was 200 trees with unlimited depth, yielding the lowest cross-validated RMSE. On the unseen test set, the tuned model achieved RMSE = 13.86, meaning predictions are typically within 14 popularity points of the true score. Applying the model to the full dataset, we compared predicted labels with true high-popularity outcomes (Fig. 3).

To improve interpretability, we complemented regression with a logistic classification model (Fig. 4). Coefficient analysis (Fig. 5) revealed strong negative effects for instrumentalness and release year, and

```
Tuned RF - test RMSE: 13.863292235763735
Tuned RF - test MAE: 10.207189095928227
Tuned RF - test R²: 0.4890871893402142
```

Figure 2

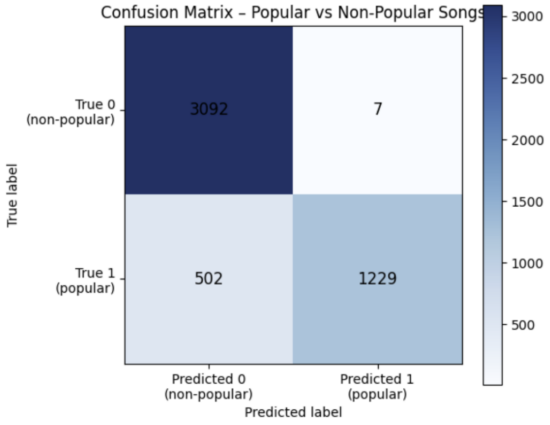

Figure 3

positive effects for loudness, energy, and danceability. This suggests highly popular songs tend to be recent, vocal, energetic, and professionally produced rather than acoustic or experimental. Limitations include missing non-audio predictors, a static popularity snapshot, and arbitrariness in the 68-point threshold. Nonetheless, results highlight meaningful structure linking audio features to popularity.

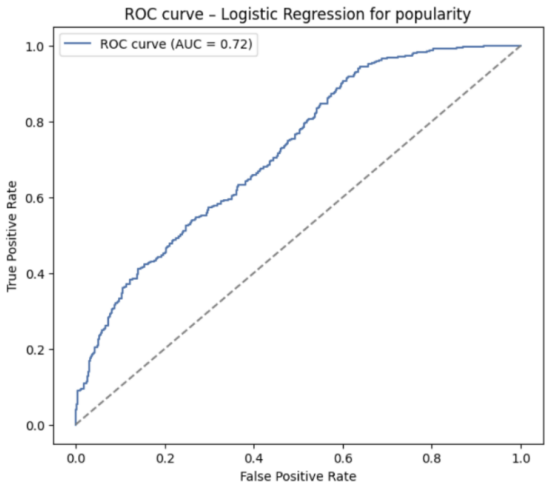

Figure 4

	feature	coefficient	abs_coefficient
3	instrumentalness	-1.023932	1.023932
8	release_year	-0.636443	0.636443
5	valence	-0.313916	0.313916
1	energy	0.299667	0.299667
0	danceability	0.213272	0.213272
6	tempo	0.112396	0.112396
7	duration_min	-0.112105	0.112105
4	liveness	-0.076099	0.076099
2	speechiness	-0.075506	0.075506

Figure 5

2.2 Historical Momentum Analysis

We examined whether long-term historical popularity rates at the genre and subgenre level are the most powerful predictors of "highly popular" songs, hypothesizing that momentum dominates individual song attributes.

We computed log-odds hit rates over 5-year windows for each genre and subgenre, producing two engineered variables: `logodds_sub_5yr` and `logodds_sub_5yr`. After merging these features and training several models, Random Forest again produced the best performance (Fig. 6), achieving AUC = 0.955. Feature importance confirmed that historical patterns dominate prediction: `logodds_sub_5yr` (0.2278) was strongest, followed by `logodds_sub_5yr` (0.1427).

These results illustrate a structural shift: while broad genres predicted popularity in earlier decades, the streaming era favors narrower subgenre identities, such as "mainstream," which consistently produce hits. Subgenre momentum captures the new structure of Spotify popularity.

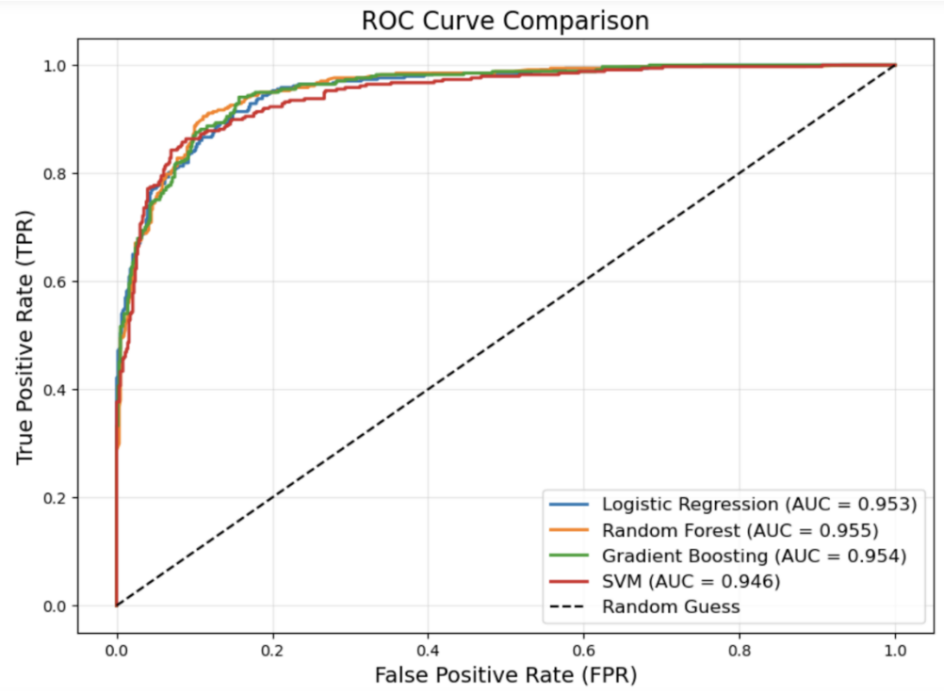

Figure 6

3 Conclusion

Spotify song popularity is shaped by audio characteristics and, more strongly, by long-term historical momentum within subgenres. Audio-feature analysis is limited by missing non-audio factors such as playlist placement or marketing exposure; future work could integrate user- and platform-level behavioral data. Historical-trend analysis is constrained by inconsistent genre and subgenre labels; future work could cluster songs by audio features and compute 5-year log-odds within fixed groups, tracking popularity trends more cleanly across decades.
Word Count: 598

Post-Conclusion Reflection

Member	Proposal	Coding	Presentation	Report
Julie Walsh	1	1	1	1
Kyle Bridger	1	1	1	1
Paddy Lavin	1	1	1	1
Samantha Burkard	1	1	1	1
Sofia Lambro	1	1	1	1

Note: In the chart above, **1** = full contribution, **0.1–0.9** = partial, **0** = no contribution.