Performance Assessment for
D214: Data Analytics Graduate Capstone
Task 3

Drew Mendez
MSDA Western Governors University
May 21st, 2025

D214 Executive Summary

May 21, 2025

0.1 Problem Statement and Hypothesis

In a streaming landscape with millions of songs competing for visibility, there is no objective, data-driven method to predict which songs will achieve high popularity. This study investigates the extent to which quantifiable audio features explain the variation in the popularity of songs on Spotify.

Using Lasso regression for simultaneous modeling and feature selection, the hypothesis seeks to test whether at least one of these features has a statistically significant influence on popularity scores (Brownlee, 2021). Formally:

Null Hypothesis: The explanatory variables do not have a statistically significant effect on the popularity of Spotify songs.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Alternative Hypothesis: At least one explanatory variable has a statistically significant effect on the popularity of Spotify songs.

$$H_a : \text{there exists an } i \text{ such that } \beta_i \neq 0$$

The aim is to identify which musical characteristics correlate with success, offering practical insight for artists, producers, and marketers navigating an increasingly data-driven industry.

0.2 Data-Analysis Process

This analysis used a publicly available Kaggle CSV dataset of 160,000+ songs (Ay, 2020), which was imported into a `pandas` data frame. After the data frame was checked for nulls, duplicates, and missing values, non-numeric identifiers (artist, song ID, song name) were dropped. Additionally, release year was converted to song age, and musical key was circularly encoded.

The data was split 80/20 into train/test sets, then features were standardized in `scikit-learn` pipelines that fit a multiple linear regression (MLR) model and a Lasso regression model (with 5-fold cross-validation).

0.3 Findings

The models were evaluated and compared using R^2, MAE, and RMSE. Both models demonstrated strong and nearly identical performance:

Metric	Lasso	MLR
R^2	0.7535	0.7535
MAE	8.0057	8.0065
RMSE	10.8039	10.8046
Best α	0.0188	N/A

These results indicate nearly identical predictive performance between the two models, with Lasso performing slightly better across all metrics. With an R^2 of 0.7535, both models explain a substantial portion of the variance in song popularity. The MAE and RMSE values suggest that the models predict popularity scores within about 8–11 points on average, on a scale of 0–100.

The Lasso model selected an optimal alpha of 0.0188, applying only mild regularization. As a result, all features were retained except `key_sin`. This low alpha value suggests that most features contribute meaningful signal to the target variable, and more aggressive regularization would have hurt performance.

Top predictors included song age (negative correlation), acousticness, instrumentalness, speechiness, and danceability.

0.4 Limitations

A limitation of this analysis is that Lasso regression may arbitrarily exclude correlated features and can misrepresent cyclical variables if partial encodings are dropped (`key_sin` was removed, compromising circular encoding). Additionally, only linear models were used, which may fail to detect non-linear relationships or interaction effects present in the variables.

This analysis is further limited by the proprietary nature of the Spotify popularity score, which may embed confounding factors (such as playlist placement and editorial curation) not captured in the dataset. Furthermore, the dataset lacks contextual metadata (artist fame, marketing efforts, etc) that likely influence popularity.

0.5 Proposed Actions

- Integrate the Lasso model into pre-release evaluation or recommendation systems to flag high-potential songs.
- Augment with contextual variables (social metrics, label support, etc) and test non-linear algorithms (random forests, neural nets, etc).
- Revisit key encoding methods or assess `key`'s predictive value separately.

0.6 Expected Benefits

- **Resource efficiency**: By using the model to pre-screen songs, Artists & Repertoire workload can be reduced by ~75% based on model variance explained, so only about 25% of incoming songs would need to be evaluated.
- **Quantitative guidance**: Target production efforts toward features correlated with a 5–10-point uplift in predicted popularity.
- **Scalability**: Enhanced models with additional features may push R^2 above 0.80, refining song-tier segmentation and boosting playlist recommendation precision.

0.7 Sources

Ay, Y. E. (2020). Spotify Dataset 1921-2020, 160k+ Tracks. Kaggle.com. https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-1921-2020-160k-tracks

Brownlee, J. (2020, October 11). How to Develop LASSO Regression Models in Python. Machine Learning Mastery. https://machinelearningmastery.com/lasso-regression-with-python/