Data Analytics Capstone Topic Approval Form

Student Name: Drew Mendez
Student ID: 010426487

Capstone Project Name: Modeling Song Popularity using Lasso Regression

Project Topic:
This project investigates the relationship between Spotify song popularity and audio features using a machine learning approach. Leveraging Lasso regression for feature selection, the analysis will identify which musical characteristics most significantly influence a song's popularity score. The resulting model will offer interpretable insights for artists, producers, and music marketers seeking to understand the ingredients of a hit song.

☑ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: To what extent do the explanatory variables (listed below) affect the popularity of Spotify songs?

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.

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

Context:
In today's music industry, streaming platforms like Spotify serve as the primary distribution and discovery channels for new music. With millions of tracks competing for attention, artists, producers, and record labels face uncertainty when trying to forecast which songs will perform well. While subjective opinions and marketing budgets have traditionally influenced decisions, data analytics offers a more objective, evidence-based approach.

Spotify collects detailed metadata for every track in its database, including attributes like danceability, energy, tempo, and acousticness. These features reflect the song's acoustic profile and can be used to analyze listener preferences and engagement patterns. Understanding how these variables correlate with song popularity could help industry stakeholders make more informed decisions—from designing production strategies to targeting promotional efforts.

By applying Lasso regression, this project will identify the most influential predictors of popularity and create a model that can estimate a song's popularity based on its characteristics. Lasso is particularly well-suited for this analysis because it not only builds a predictive model but also performs feature selection, reducing the noise from less relevant variables. The resulting insights could help reduce uncertainty in the music development process, offering a competitive advantage in a highly saturated market.

Data: This dataset includes variables relevant to this analysis such as:

- Popularity (target variable)
- Danceability
- Energy
- Loudness
- Speechiness
- Acousticness
- Instrumentalness
- Liveness
- Valence
- Tempo
- Duration
- Release year
- Key
- Mode
- Time Signature

The dataset originates from Spotify's public API and has been compiled into downloadable formats by third-party platforms like Kaggle. The data is publicly shared under terms that permit non-commercial use for educational and research purposes. It is distributed under the Community Data License Agreement – Sharing – Version 1.0, which allows use for academic projects like this capstone, provided credit is given and no commercial redistribution occurs. Thus the dataset is appropriate and legally acceptable for use in the MSDA capstone.

Data Gathering:
For this project, data will be obtained from an existing, publicly available dataset hosted on Kaggle titled *"Spotify Dataset 1921–2020, 160k+ Tracks"* (by user Yamac Eren Ay). The dataset includes over 160,000 Spotify tracks, each with a wide range of audio features and a popularity score assigned by Spotify's internal algorithm (on a scale of 0 to 100). The dataset will be downloaded as a CSV file and imported into a Python environment using the pandas library.

Data Analytics Tools and Techniques:
This project will use Lasso regression as the primary data-analysis technique. Data preparation and modeling will be conducted using Python and relevant libraries, including:

- pandas for data cleaning and manipulation
- scikit-learn for implementing Lasso regression, model evaluation, and train-test splitting
- matplotlib and seaborn for data visualization
- NumPy for numerical operations

Justification of Tools/Techniques:
Lasso regression is a technique for predictive modeling and feature selection in datasets with many explanatory variables. In the context of Spotify song data, where numerous audio features are available, Lasso is especially useful because it applies an L1 regularization penalty that shrinks less important feature coefficients to zero. This not only helps prevent overfitting, but also identifies which features are most influential in predicting song popularity.

Compared to traditional linear regression, Lasso improves model interpretability and handles multicollinearity between features more effectively. This is particularly important in music data, where features like energy, loudness, and danceability may be correlated. By selecting only the most predictive variables, Lasso ensures a more robust and generalizable model.

Python and its libraries are widely used in the data science industry, and scikit-learn provides efficient implementations of Lasso regression, cross-validation, and model evaluation metrics. These tools are appropriate for building a reproducible, scalable, and interpretable solution to the research question.

Project Outcomes:
The key anticipated outcomes and deliverables of this project include:

1. **Cleaned and Preprocessed Dataset**
   A Spotify song dataset with relevant features selected, cleaned for missing values, formatted appropriately, and prepared for regression modeling.
2. **Feature Selection Results via Lasso Regression**
   A trained Lasso regression model that identifies the subset of audio features most influential in predicting song popularity. Features with zero coefficients will be excluded from the final interpretation.
3. **Predictive Model**
   A regression model that estimates a song's popularity score based on its audio characteristics. The model will be evaluated using metrics such as R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) on a test set.
4. **Visualizations and Interpretations**
   Clear data visualizations to support the interpretation of which features matter most and how well the model performs.
5. **Business Insight Summary**
   A concise interpretation of the model's results, highlighting what they imply for music producers, labels, or marketers—such as which types of songs are more likely to succeed based on quantifiable traits.
6. **Final Report**
   A comprehensive report summarizing the methodology, analysis, results, and business implications, suitable for academic evaluation and for stakeholders interested in the music streaming industry.

Projected Project End Date: June 30th, 2025

Sources: https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-1921-2020-160k-tracks

Course Instructor Signature/Date:

☑ The research is exempt from an IRB Review.

□ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: May 3, 2025

Reviewed by: Daniel J. Smith, PhD, MBA

Comments: