**Predicting the Popularity of a Song**

BUAN 6312.S01: Applied Econometrics and Time Series Analysis

Fall 2023

A musical notes flying in the air

Description automatically generated

**GROUP SIX**

Indrasena Reddy Vukanti - ixv220001

Sai Teja Narahari - sxn220028

Shreya Dasu - sxd220064

Shreshta Reddy Aekula - sxa220116

Neha Reddy Chegi Reddy - nxc210051

Vssmk Sumanth Chilamakuri - svc200000

Poorna Srinivas Kanakamedala - pxk220047

Sai Pranay Reddy Aare - sxa220072

A logo of a university of texas

Description automatically generated with low confidence

December 14th, 2023

**Background:** ​

In the expansive universe of music streaming platforms like Spotify, the dynamics of song popularity remain an enigmatic puzzle. Artists, with their unique voice, styles, and genres, often produce albums that contain a mix of chart-toppers and hidden gems. What distinguishes these songs in terms of their popularity? Are there elusive factors that significantly influence a listener's inclination to embrace a particular track? This research aims to unravel these mysteries through extensive Exploratory Data Analysis (EDA) and hypothesis testing.​

​

**Objective: ​**

The research aims to decipher the underlying factors influencing song popularity within Spotify's music streaming platform. Through a comprehensive examination of song attributes and statistical modeling, the study endeavors to shed light on the elusive aspects contributing to a song's standing in terms of popularity.

**Literature Review- Our Contribution:**

Previous studies on predicting song popularity, such as the works by Salganik, Dodds, and Watts, highlight the partial influence of song quality on its popularity, emphasizing the significant role played by other factors. This project endeavors not only to construct a robust predictive model for song popularity but also to interpret the effects of various features on a song's standing.

**References:**

* Predicting Popularity on Spotify — When Data Needs Culture More than Culture Needs Data (Philip Peker, Towards Data Science)
* Prediction of product success: explaining song popularity by audio features from Spotify data (R. Nijkamp, UTwente)
* P4KxSpotify: A Dataset of Pitchfork Music Reviews and Spotify Musical Features (A. Pinter, J. Paul, J. Smith, J. Brubaker, AAAI Conference on Web and Social Media)

**Attributes:**

The dataset contains a comprehensive collection of attributes associated with each song, providing multifaceted insights into the characteristics influencing song popularity. These attributes include:

* artist\_name : Name of the artist who performed the track.
* track\_id : A unique identifier for each track.
* track\_name : The title or name of the track.
* acousticness : Measure of how acoustic or non-electronic the track is (ranging from 0 to 1, where 1 indicates high acousticness).
* danceability : Suitability of a track for dancing (ranging from 0 to 1, where 1 signifies high danceability).
* duration\_ms : Duration of the track in milliseconds.
* energy : Represents the intensity and activity of the track (ranging from 0 to 1, where 1 indicates highly energetic).
* instrumentalness : Likelihood that a track is instrumental (ranging from 0 to 1).
* key : The key in which the track is represented numerically.
* liveness : Indicates the presence of an audience in the recording (ranging from 0 to 1, where 1 indicates a live recording).
* loudness : The overall loudness of the track in decibels (LUFS).
* mode : Indicates the modality (major or minor) of the track (0 represents minor, and 1 represents major).
* speechiness : Detects the presence of spoken words in the track (ranging from 0 to 1).
* tempo : The estimated tempo of the track in beats per minute (BPM).
* time\_signature : An estimated overall time signature of the track.
* valence : Describes the musical positiveness conveyed by a track (ranging from 0 to 1, where 1 indicates high valence).
* popularity : A measure of how popular the track is, often based on factors like the number of listens or streams.

**Data:**

The dataset utilized for this research, denoted as SpotifySongPolularityAPIExtract.csv, represents a comprehensive compilation of song attributes sourced from the Spotify API. This dataset serves as the cornerstone for exploring the intricate dynamics underlying song popularity within the Spotify music streaming platform. The dataset, SpotifySongPolularityAPIExtract.csv, was obtained from the Spotify API, providing a rich repository of song attributes, allowing for comprehensive analysis and modeling to unravel the intricacies of song popularity on the platform.

A table of numbers with different colored numbers

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated

**Exploratory Data Analysis (EDA):**

The EDA phase involved meticulous exploration of the dataset to identify patterns, trends, and anomalies that could influence subsequent analysis. Key findings and actions taken during EDA:

* **Anomaly Identification:** Detection and removal of anomalies, such as tracks exceeding 10 minutes in duration, zero values in popularity, and inconsistencies in tempo and time signature, ensuring data integrity for analysis.
* **Correlation Analysis:** Examination of correlations between attributes revealed significant relationships among features, influencing subsequent feature engineering and model construction.
* **Feature Engineering:** Creation of new variables like 'noise\_level', 'total\_songs\_by\_artist', 'popular\_key', and 'is\_new\_artist' based on attribute correlations to enhance predictive modeling capabilities.
* **Data Cleaning:** Addressing inconsistencies, outliers, and missing values through data cleaning procedures to ensure the dataset's reliability and suitability for analysis.

A table of numbers and symbols

Description automatically generated with medium confidenceA graph of a danceability

Description automatically generated A graph of a normal distribution

Description automatically generated A graph of energy and energy

Description automatically generated A graph of a function

Description automatically generated with medium confidence A graph with blue lines

Description automatically generatedA graph of a number of cells

Description automatically generatedA graph of a frequency

Description automatically generatedA graph of a speech

Description automatically generated

A graph of a function

Description automatically generatedA graph of a function

Description automatically generatedA graph of a number of people

Description automatically generatedA graph of a number of numbers

Description automatically generated with medium confidenceA graph of a number of dots

Description automatically generated

A graph of different colored squares

Description automatically generated

**Heat Map Analysis:**

The construction of a correlation heat map allowed for a visual examination of the interrelationships between various predictor variables within the dataset. The heatmap provided insights into the strength and direction of correlations, aiding in identifying potential multicollinearity issues among attributes.

Observations:

* High Correlation with Loudness: The heat map revealed a notably high correlation between 'loudness' and other attributes, notably 'energy', 'instrumentalness', and 'acousticness'.
* Impact on Model Stability: Such high correlations among predictor variables can lead to multicollinearity, potentially resulting in unstable coefficient estimates. This situation complicates the interpretation of individual variable effects within the model.

Action Taken:

- Removal of 'Loudness': In response to the observed high correlation, the decision was made to remove the 'loudness' feature from the dataset. This strategic removal aimed to alleviate multicollinearity concerns, enhancing the stability and interpretability of subsequent model analyses.

The heat map analysis provided crucial insights into attribute relationships, prompting the deliberate exclusion of 'loudness' to ensure a more robust and reliable modeling process.

A screenshot of a graph

Description automatically generated

**Feature Engineering:**

Feature engineering plays a pivotal role in refining the dataset and enhancing predictive modeling capabilities. In our research, this process was meticulously conducted to address multicollinearity concerns and introduce new variables aimed at improving the predictive power of the model.

The feature engineering phase played a pivotal role in refining the dataset for subsequent analysis. By addressing multicollinearity issues through attribute removal and introducing purposeful new variables, the dataset was optimized to enhance the predictive power of our model. This process significantly contributed to a more reliable and robust analysis of song popularity factors within the Spotify platform.

**Handling High Correlation:**

Initially, the analysis revealed that the 'loudness' feature exhibited high correlation with several other attributes. To mitigate multicollinearity issues within the dataset, 'loudness' was deliberately removed, ensuring a more stable and reliable modeling process.

**Introduction of New Variables:**

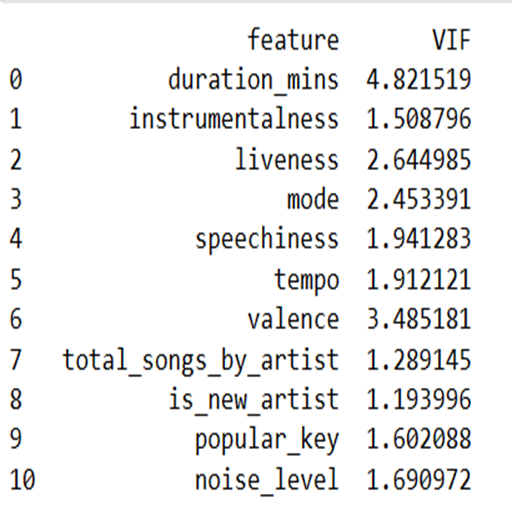
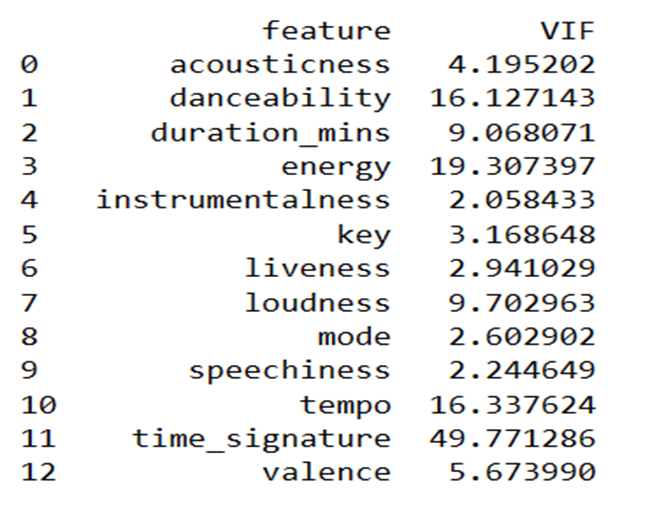
To augment the model's predictive ability and further mitigate multicollinearity, new variables were ingeniously engineered:

* 'noise\_level': This variable was derived from correlated features such as 'energy' and 'acousticness', offering a distinct representation of the sound profile.
* 'total\_songs\_by\_artist': Representing the total number of songs by an artist within the dataset, providing insights into an artist's presence and influence.
* 'popular\_key': A binary indicator highlighting specific keys [0.0, 1.0, 8.0, 11.0] known to be popular, facilitating easier interpretation within the model.
* 'is\_new\_artist': A binary variable distinguishing artists with only one song from those with multiple songs in the dataset, offering insights into artist novelty.

**Variance Inflation Factor (VIF) Analysis:**

VIF values were calculated before and after the feature engineering process to gauge multicollinearity among the variables. The results underscored the effectiveness of the feature engineering efforts:

* Before Feature Engineering: Several attributes exhibited high VIF values, signaling significant multicollinearity concerns and the necessity for refinement.
* After Feature Engineering: The introduction of new variables and the removal of 'loudness' substantially reduced multicollinearity, evidenced by lowered VIF values across the attributes, ensuring improved model stability and performance.



**Data Cleansing for Enhanced Analysis:**

In pursuit of data refinement and improved analytical robustness, specific steps were dedicated solely to data cleansing procedures:

1. **Addressing Multicollinearity**:

   - Removed attributes ('loudness', 'danceability', 'time\_signature', 'acousticness', and 'energy') exhibiting high multicollinearity to ensure model stability and reliability in subsequent analyses.

2. **Enhancing Attribute Representation:**

   - Transformed 'tempo' into discrete categories (-1, 0, 1, 2) for better model representation, facilitating a more accurate analysis of tempo-related influences on song popularity.

These focused data cleansing steps significantly improved the dataset's quality, eliminating multicollinearity concerns and refining attribute representation for more precise analytical insights.

**Predictive Models for Song Popularity**:

This section delves into the predictive models developed to discern the relationship between song attributes and their respective popularity within the Spotify platform.

These models elucidate the relationships between song attributes and their impact on song popularity within the Spotify ecosystem. Each model offers distinctive insights into the predictive nature of various song features, providing a comprehensive understanding of factors influencing song popularity metrics.

**1. Model with Features and Dependent Variable:**

   - The model equation is represented as follows:

     - Popularity = 36.034 - 1.07 log(duration\_min) - 8.45 instrumentalness - 6.634 liveness - 0.875 mode - 5.145 speechiness - 0.325 tempo - 1.596 valence - 0.007 total\_songs\_by\_artist - 6.503 is\_new\_artist + 0.6759 popular\_key + 4.439 noise\_level

A screenshot of a computer

Description automatically generated

**2. Model with Log of Dependent Variable:**

   - The equation for this model is depicted as:

     - Log(Popularity) = 3.312 + 0.006 log(duration\_min) – 0.323 instrumentalness – 0.258 liveness – 0.027 mode – 0.376 speechiness – 0.008 tempo – 0.086 valence – 0.0004 total\_songs\_by\_artist – 0.335 is\_new\_artist + 0.0218 popular\_key + 0.2709 noise\_level

A screenshot of a computer

Description automatically generated

**3. Model with Quadratic Term and Log of the Dependent Variable:**

   - This model equation is formulated as:

     - Log(Popularity) = 2.906 + 0.7655 log(duration\_min) – 0.273 instrumentalness – 0.233 liveness – 0.024 mode – 0.351 speechiness – 0.008 tempo – 0.0999 valence – 0.0004 total\_songs\_by\_artist – 0.34 is\_new\_artist + 0.02 popular\_key + 0.239 noise\_level – 0.332 log(duration\_min\_squared)

A screenshot of a computer

Description automatically generated

**Normality Assessment of Model Residuals:**

An evaluation of the model's residuals was conducted to ascertain the normal distribution and validity of statistical inferences derived from these residuals. The Shapiro-Wilk test yielded the following results:

- **Shapiro-Wilk Test Statistic:** 0.984377920627594

- **p-value:** 0.28617727756500244

The objective of this analysis was to validate if the residuals conform to a normal distribution. A normal distribution of residuals is essential to ensure the validity of statistical tests and confidence intervals derived from these residuals.

In this case, the Shapiro-Wilk test, indicated by a p-value of 0.286, suggests that there is no significant departure from normality in the distribution of residuals. Therefore, it can be inferred that the residuals reasonably approximate a normal distribution, validating the reliability of statistical inferences and confidence intervals derived from the model.

A graph of different sizes and shapes

Description automatically generated with medium confidence

**Hypothesis Testing: Impact of Song Duration and Noise Levels on Popularity:**

A hypothesis was formulated to investigate whether individuals display a preference for songs with shorter durations and higher noise levels. The null hypothesis (H0) posits that the coefficients of both 'duration\_mins' and 'noise\_level' are equal to zero, indicating an absence of significant impact on 'popularity'. Conversely, the alternative hypothesis (H1) suggests that at least one coefficient differs from zero, signifying a substantial effect on 'popularity'.

 Hypothesis Statements and Statistical Inference:

* H0: The coefficients for ‘duration\_mins’ and ‘noise\_level’ have no significant effect on ‘popularity’.
* H1: At least one coefficient (‘duration\_mins’ or ‘noise\_level’) significantly impacts ‘popularity’.
* Interpretation of p-value: A p-value less than α (significance level) rejects the null hypothesis in favor of the alternative hypothesis.
* F Test Result: The F statistic is 498.82 with a remarkably small p-value of approximately 2.83e-216, denoting strong evidence against the null hypothesis.
* Conclusion: Based on the coefficients, evidence supports the assertion that song duration and noise levels are associated with changes in popularity.

**Inference from Analysis:**

The statistical evidence derived from the F test rejects the null hypothesis, indicating a substantial relationship between song duration, noise levels, and their influence on song popularity. Notably, it suggests that for songs exceeding 4 minutes, higher noise levels and shorter durations tend to correlate with increased popularity. Conversely, songs less than 4 minutes witness an upsurge in popularity as the duration increases.

**Future Scope and Potential Extensions:**

The study has identified several variables that, despite their relevance, were not encompassed within the current model due to data limitations. These variables present potential avenues for future research, contributing to a more comprehensive understanding of song popularity dynamics within the Spotify ecosystem.

**Variables for Consideration in Future Research:**

* Social Media Following for Artists: Investigate the impact of an artist's social media presence on song popularity, providing insights into the correlation between an artist's online following and song performance.
* Artist's Record Label Affiliation: Explore how an artist's association with specific record labels influences song popularity, unraveling the potential impact of different label affiliations on audience reception.
* Artist's Network Value Metric: Develop a metric to quantify an artist's network value, potentially utilizing measures akin to a "Who do you know here?" score, providing deeper insights into an artist's influence and reach.
* Nostalgia Score:  Introduce a metric to gauge the nostalgic value of a song, exploring the influence of nostalgia on song popularity metrics.
* Historical Artist Data: Incorporate historical data for artists, enabling longitudinal analysis to understand how an artist's evolution impacts song popularity over time.
* Song Playback Frequency: Analyze the correlation between the number of times a song is played and its popularity, unraveling the impact of repeated listening on a song's reception.

**Subsetting Popularity for Deeper Analysis:**

Exploring subsets of popularity metrics based on various factors could enhance the depth of analysis:

* Regional Variation: Investigate how popularity metrics vary across different regions, unveiling regional preferences and trends.
* Demographic Segmentation:  Examine popularity variations based on different demographics, providing insights into preferences among diverse listener groups.
* Streaming Device Analysis: Assess popularity differences based on the streaming device used, uncovering potential disparities in audience reception across platforms.
* Spotify Account Sharing:  Investigate the influence of shared Spotify accounts on song popularity, shedding light on collaborative listening trends.

**Potential Impact:**

The integration of these variables and subsets into future research endeavors could significantly enrich our understanding of song popularity determinants within Spotify's music landscape. Such investigations hold the potential to unveil nuanced insights, aiding music industry professionals, artists, and streaming platforms in optimizing content and strategies tailored to diverse audience preferences.

**Conclusion**

This research provides insights into the complex interplay of song attributes influencing popularity on Spotify. Despite significant strides in understanding these dynamics, there remains a wealth of unexplored variables and subsets that could further enrich our comprehension of what makes a song resonate with audiences.

A black and white drawing of a guitar

Description automatically generated