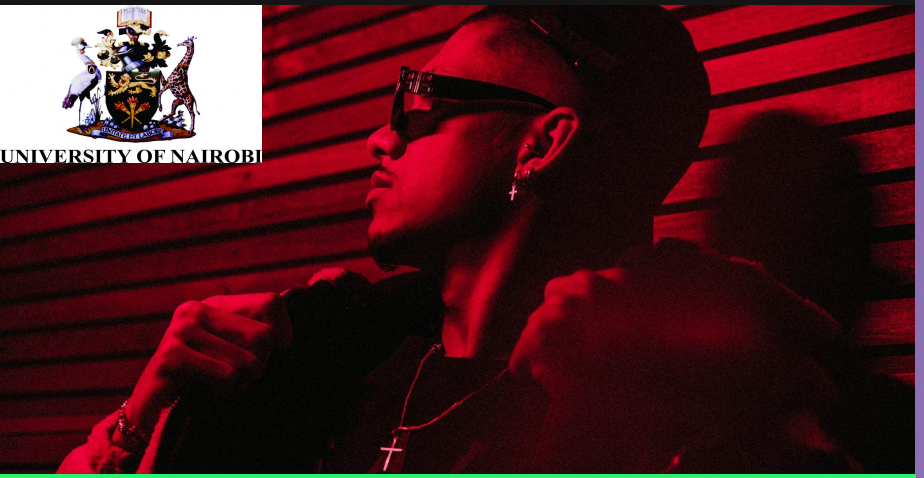


UNIVERSITY OF NAIROBI

A MACHINE LEARNING APPROACH TO PREDICTING SPOTIFY SONG POPULARITY USING INTRINSIC AUDIO FEATURES

Presented by:

Wendy Josephine Awuor - I63/4881/2021

Hannah Sururu - I63/4878/2021

Daniella Ngii Muli - I63/4877/2021

Kiluva James Kitenye - I63/5754/2021

Supervised by : Prof J . I . Mwaniki

NEXT

INTRODUCTION

Understanding Digital Music & The Science of Popularity

- **Music's Big Change:** The music industry has undergone a significant transformation, shifting from physical formats like CDs to digital music streaming platforms such as Spotify which has millions of songs.
- **The Big Question:** **With millions of tracks available, what makes a song popular?**

INTRODUCTION

Understanding Digital Music & The Science of Popularity

- **In Our Quest to Answer This:** We leveraged a large dataset from Spotify to identify measurable audio features that contribute to popularity. Our idea was to look at song features such as:
  - Danceability (how good it is for dancing)
  - Energy (how intense it sounds)
  - Loudness (how loud it is)
- Our research contributes to fields like 'Music Information Retrieval (MIR)' (analyzing music data) and 'Hit Song Science' (predicting song success) by finding these key popularity factors.

Objectives of the Study

“From Performance to Prediction – Mapping Our Goals”

- **Main Objective:**
  - To build a predictive model that estimates a song's likelihood of success using its audio features.
- **Specific Objectives:**
  - Analyze how different song features contribute to popularity.
  - Compare various machine learning approaches (like regression, classification) to find the most accurate prediction method.
  - Evaluate our model's performance using key metrics like R-squared and AUC.
  - Build an Application tool to demonstrate and visualize these predictions.
- **Significance of Study:**
  - For Artists & Producers: As understanding song popularity helps them tailor and optimize their music for today's market.
  - For the Music Industry: Provides valuable insights for streaming platforms, record labels, and music marketers to optimize their strategies and engage users better.

Literature Review Summary

What Research Says About Hit Songs & Machine Learning

Before we dive into our own research, it's essential to understand what's already been explored in the quest to predict song popularity

- **Initial Explorations (2022):**
  - Duman et al. (2022) specifically studied dance music features (energy, danceability, loudness) and their impact on listener preferences, noting their role in mood regulation.
- **Challenges Emerge (2023):** As studies continued, limitations in purely audio-based predictions surfaced.
  - Studies (e.g., Dong et al., Nijkamp) found out that audio features alone show weak links to popularity, suggesting that other external factors like marketing and social media may be crucial in predicting song popularity.
  - Due to these weak linear links, *Xing (2023)*, began exploring nonlinear approaches for better fit.

Literature Review Summary

What Research Says About Hit Songs & Machine Learning

- **Embracing Nonlinear Models (2023-2024):** Recognizing complex data, researchers moved to **nonlinear Machine Learning models**
  - *Beesa et al. (2024)* applied ML algorithms to song audio features to understand their role in song success and audience engagement.
  - *Sardana (2024)* further aimed for high-accuracy models by addressing dataset issues and ***re-framing prediction as a classification problem***, using diverse ML and neural network architectures.

This collective understanding from previous studies directly shaped our own approach, leading us to investigate these complexities further.


RESEARCH METHODOLOGY

Introduction: This chapter outlines the methodology used for the study, the evaluation metrics ar
the key assumptions for the study.

- To build our machine learning models, we would employ:

Regression Approach:

- **Objective:** To predict a **continuous numerical output** (e.g., song popularity score).
- **How it works:** Establishes and models relationships between independent variables (song features) and a dependent variable (popularity) to forecast its value.

Classification Approach:

- **Objective:** To assign data points to one of several predefined categories or classes.
- **How it works:** Converts the continuous popularity variable into categorical data (Very Low, Low, High, Very High) and predicts which category a song belongs to based on its attributes.

Machine Learning Models

Random Forest (RF):

- An ensemble, tree-based method that combines many decision tree classifiers.
- Decision trees are grown independently on random bootstrap samples.
- Known to be protected from overfitting, even with a large number of trees.

Concept	Equation
Single Tree Prediction	$\hat{f}_b(x) = \sum_{m=1}^{M} c_m \cdot \mathbb{I}(x \in R_m)$

RF builds B trees on **bootstrapped samples** (random subsets of data with replacement). The prediction is:

RF Regression	$\hat{f}_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$
RF Classification	$\hat{y} = \text{mode}\{\hat{f}_1(x), \ldots, \hat{f}_B(x)\}$

where:

- c_m = majority class in region R_m,
- $\mathbb{I}(\cdot)$ = indicator function.

For regression, c_m is the average of training outputs in R_m

XGBoost:

- A highly optimized gradient boosting framework.
- Builds decision trees sequentially, with each tree attempting to correct the mistakes of the previous one.
- The final prediction is the sum of predictions from all tree**s.**

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

- $\hat{y}_i$: Predicted output for sample x_i.
- f_k: k-th decision tree (weak learner).
- K: Total number of trees.

Gradient Boost Machine (GBM):

- Builds an ensemble of shallow, weak successive trees.
- Each new tree learns from and improves upon the errors of the previous ones.

Concept	Equation/Explanation
Additive Model	$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$

- $\hat{y}_i$: Predicted output for sample x_i.
- f_k: k-th decision tree (weak learner).
- K: Total number of trees.

Concept	Equation/Explanation
Final Prediction	$\hat{y}_i = \hat{y}_i^{(0)} + \eta \sum_{k=1}^{K} f_k(x_i)$

- $\hat{y}_i^{(0)}$: Initial prediction (e.g., mean for regression).
- η: Learning rate (shrinkage factor to prevent overfitting).

Evaluation Metrics - Regression

R-squared (R^2) - Coefficient of Determination:

Measures the proportion of variance in the dependent variable (popularity) explained by independent variables (song features).

Ranges from 0 to 1; higher values indicate a better fit.

Formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

- $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: **Sum of Squared Residuals** (errors between true and predicted values).
- $SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$: **Total Sum of Squares** (variance in the true data).
- y_i: True value of the i-th observation.
- $\hat{y}_i$: Predicted value of the i-th observation.
- $\bar{y}$: Mean of the true values ($\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$).
- n: Number of observations.

Root Mean Square Error (RMSE):

Measures the average difference between predicted and actual values.

Aims for a low value (close to 0) for accurate predictions.

Formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

- y_i: True value of the i-th observation.
- $\hat{y}_i$: Predicted value of the i-th observation.
- n: Number of observations.

Evaluation Metrics - Classification

1. **Receiver Operating Characteristic (ROC) Curve:**

- Visualizes the trade-off between a classification model's ability to correctly identify positive cases (True Positives) and incorrectly classify negative cases (False Positives).
- Plots True Positive Rate (TPR) against False Positive Rate (FPR) at various classification thresholds.

True Positive Rate (TPR) / Sensitivity:

- The proportion of actual positive cases correctly identified by the model.

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

2. **Area Under the Curve (AUC):**

- Provides a single-number summary of the ROC curve, representing the model's overall ability to discriminate between positive and negative classes.
- Ranges from 0 to 1:

AUC = 1: Perfect distinction.

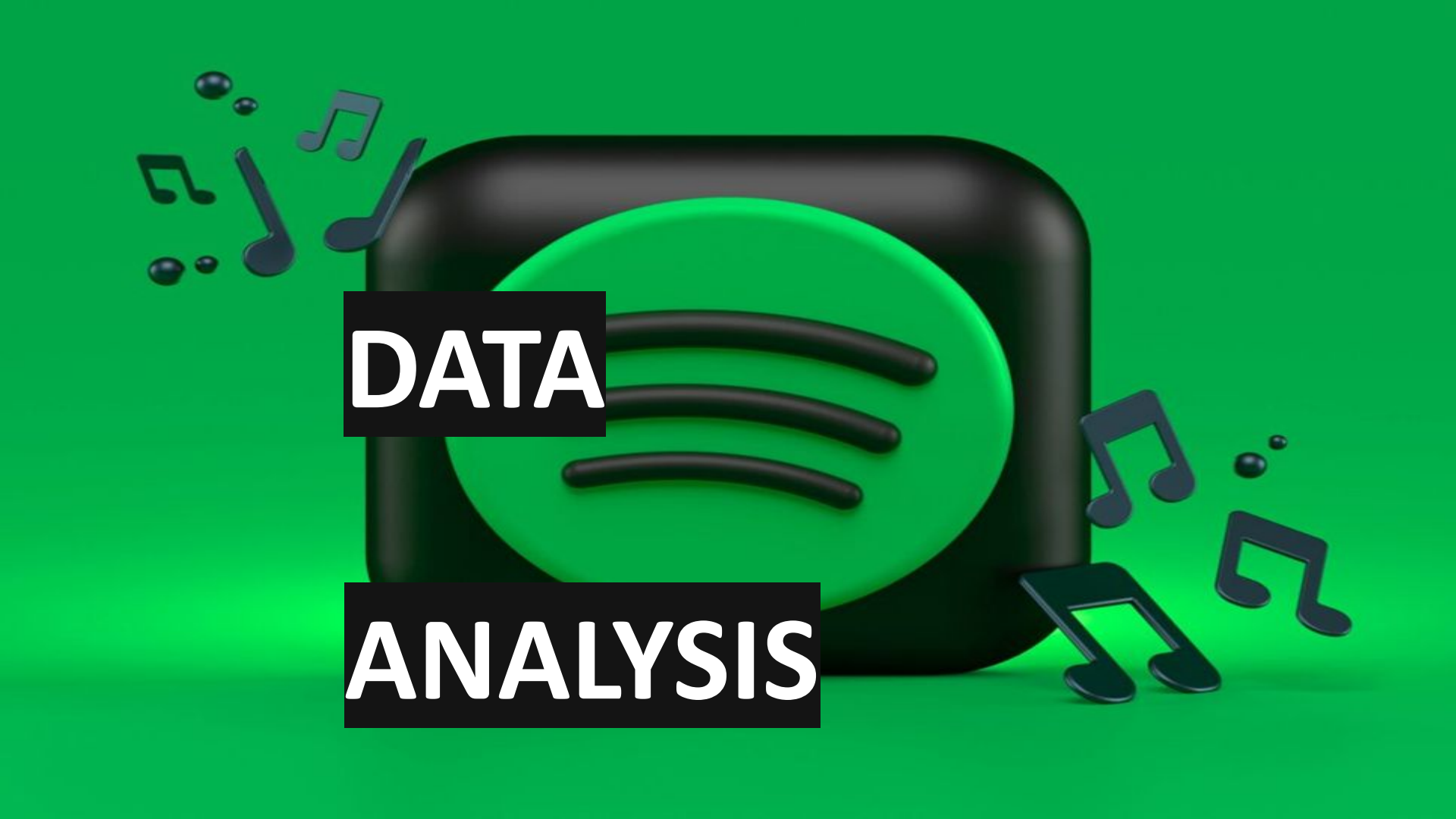
AUC = 0.5: No better than random guessing.

AUC > 0.5: Better than random guessing.

➔ In this study, a high AUC means the model effectively separates high-popularity songs from low-popularity ones.

Key Study Assumptions

1. **Data Representativeness:** Assumes the Spotify 2024 dataset represents all songs and Spotify listener preferences.
2. **Feature Relevance:** Assumes Spotify audio features are the primary factors influencing song popularity, acknowledging other external factors (marketing, social media) may also play a role.
3. **Model Generalization:** Predictions should apply to unseen songs, though differences in data distributions may affect accuracy.
4. **Independence of Observations:** Each song is assumed to be independent, though artist/genre dependencies may exist.


DATA
ANALYSIS

Understanding the Data

Introduction: This chapter outlines the methods used for the study, including the data source, its features, and the methods employed for research.

Data Source:

- Retrieved from Kaggle, originally extracted from the Spotify Application Programming Interface (API).
- Initial dataset: 1,750,032 records and 25 attributes.
- Processed dataset: 20 key attributes after removing irrelevant ones (e.g., song names, artist names).

Dependent Variable:

Popularity: - A value from 0-100, where 100 indicates the most popular.

- Calculated algorithmically based on total and recent plays.

Independent Variables:

The other 19 song features, such as danceability, loudness, energy, key, acousticness, mode, speechiness, instrumentalness, valence, time signature, liveness, and tempo.

Independent Variables - Key Feature Descriptions

x1 **Danceability:** Suitability for dancing (0.0 least, 1.0 most).
x2 **Energy:** Perceptual measure of intensity and activity (0.0 to 1.0).
x3 **Key:** Main note around which the song revolves (0 for C, 1 for C#, etc.).
- **Loudness:** Overall loudness in decibels (typically -60 to 0 dB).
- **Mode:** Indicates major (1) or minor (0) tonality.
- **Speechiness:** Presence of spoken words (closer to 1.0 for more speech).
- **Acousticness:** Confidence measure (0.0 to 1.0) of whether the track is acoustic.
- **Instrumentalness:** Lack of vocals (0.0 prominent vocals, 1.0 purely instrumental).
- **Liveness:** Likelihood of being a live recording (0.0 studio, 1.0 live).
- **Valence:** Emotional positivity (0.0 negative, 1.0 positive).
- **Tempo:** Speed of the musical piece in beats per minute.
- **Time signature:** Number of beats per musical bar.
- **Explicitness:** Indicates explicit lyrics.

xn **Days since release**: Days since song production.

Overall Statistical Analysis Methods:

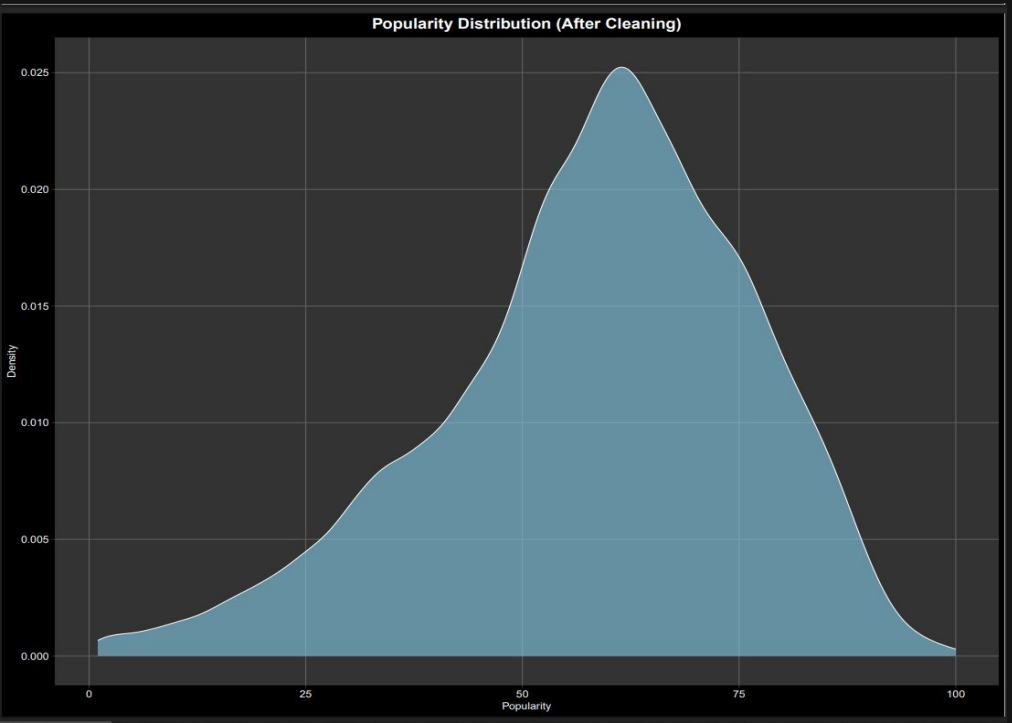
Initial Data Cleaning involved:

- Removing duplicate data entries.
- Missing values were imputed with the median due to left-skewed data and outliers.
- Songs with a popularity score of 0 were removed to focus on factors influencing relative success.

There after we did statistical analysis:

- **Exploratory Data Analysis (EDA)**: Summary statistics, distribution analysis, and correlation examination.
- **Machine Learning Models**: Regression and classification analysis to identify the best prediction approach.

Exploratory Data Analysis - Popularity Distribution



The histogram illustrates the distribution of song popularity.
Most frequent popularity scores fall between 60 and 65.
The distribution is left-skewed, meaning there are more songs with moderately lower popularity scores than extremely high ones.
There are very few songs with extreme popularity scores.

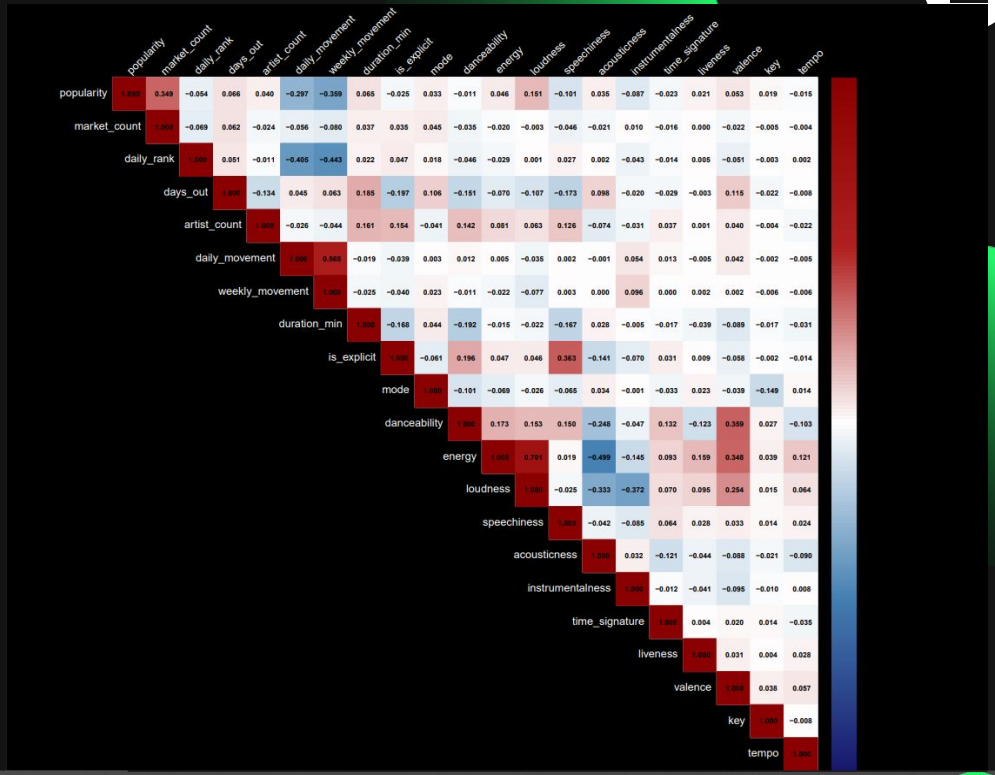
Correlation Analysis

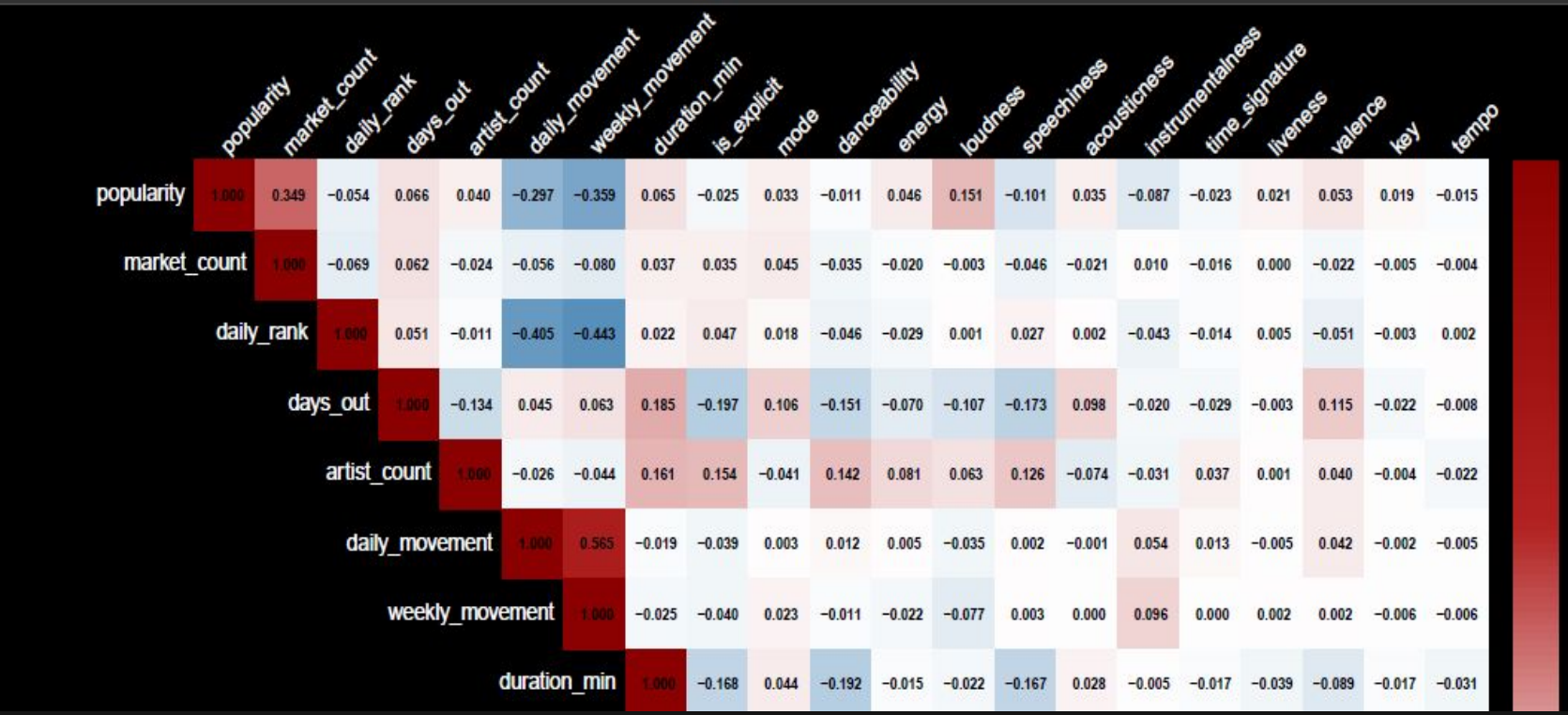
- Correlation heatmap was plotted first to study the relationships between popularity and the audio features.

1. **Correlation Heatmap** :

Shows that while some variables exhibit moderate correlations with "popularity," many pairs have absolute correlation values below 0.1, signaling weak or insignificant linear associations.

The Pearson correlation coefficient, measuring linear dependence, struggles to capture the nonlinear relationships.




popularity market_count daily_rank days_out artist_count daily_movement weekly_movement duration_min is_explicit mode danceability energy loudness speechiness acousticness instrumentalness time_signature liveness valence key tempo
popularity 1.000 0.349 −0.054 0.066 0.040 −0.297 −0.359 0.065 −0.025 0.033 −0.011 0.046 0.151 −0.101 0.035 −0.087 −0.023 0.021 0.053 0.019 −0.015
market_count 1.000 −0.069 0.062 −0.024 −0.056 −0.080 0.037 0.035 0.045 −0.035 −0.020 −0.003 −0.046 −0.021 0.010 −0.016 0.000 −0.022 −0.005 −0.004
daily_rank 1.000 0.051 −0.011 −0.405 −0.443 0.022 0.047 0.018 −0.046 −0.029 0.001 0.027 0.002 −0.043 −0.014 0.005 −0.051 −0.003 0.002
days_out 1.000 −0.134 0.045 0.063 0.185 −0.197 0.106 −0.151 −0.070 −0.107 −0.173 0.098 −0.020 −0.029 −0.003 0.115 −0.022 −0.008
artist_count 1.000 −0.026 −0.044 0.161 0.154 −0.041 0.142 0.081 0.063 0.126 −0.074 −0.031 0.037 0.001 0.040 −0.004 −0.022
daily_movement 1.000 0.565 −0.019 −0.039 0.003 0.012 0.005 −0.035 0.002 −0.001 0.054 0.013 −0.005 0.042 −0.002 −0.005
weekly_movement 1.000 −0.025 −0.040 0.023 −0.011 −0.022 −0.077 0.003 0.000 0.096 0.000 0.002 0.002 −0.006 −0.006
duration_min 1.000 −0.168 0.044 −0.192 −0.015 −0.022 −0.167 0.028 −0.005 −0.017 −0.039 −0.089 −0.017 −0.031

Predictive Modeling - The Regression Approach

Rationale: Due to observed nonlinear data tendencies, we employed machine learning for popularity prediction, starting with a regression approach.

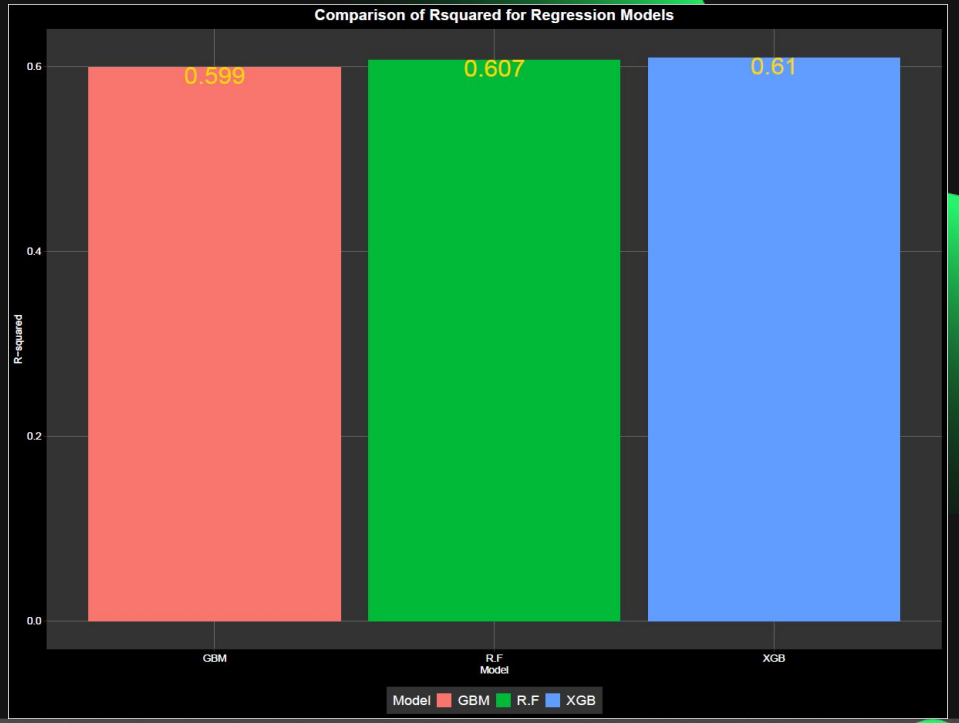
Methodology:

- **Data Split:** 80% training, 20% testing.
- **Robustness:** 5-fold cross-validation on training data to ensure reliable model evaluation.
- **Models:** Random Forest, XGBoost, and Gradient Boosted Machines (GBM).
- **Metrics:** R-squared and RMSE.

Regression Results - Performance Comparison

R-squared Comparison :

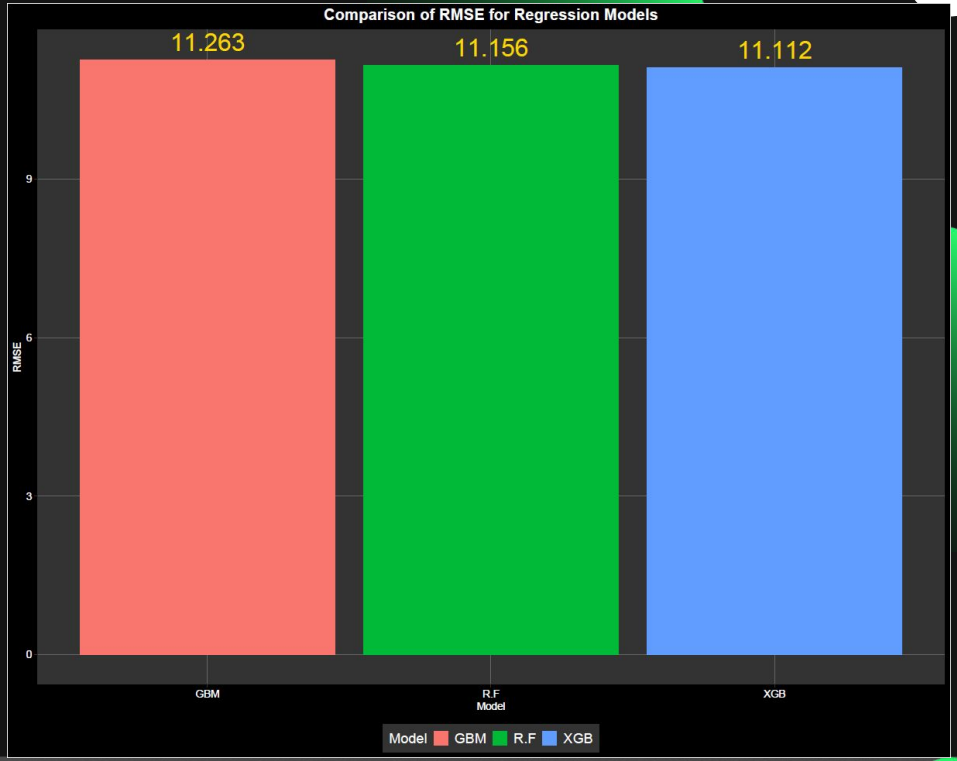
- Our regression analysis showed comparable R-squared values (about 0.59-0.61) across Random Forest (RF), XGBoost (XGB), and Gradient Boosted Machines (GBM) in predicting song popularity.
- This implies that these features account for more than half of the variation in popularity, but other unmodeled factors likely also have an impact.



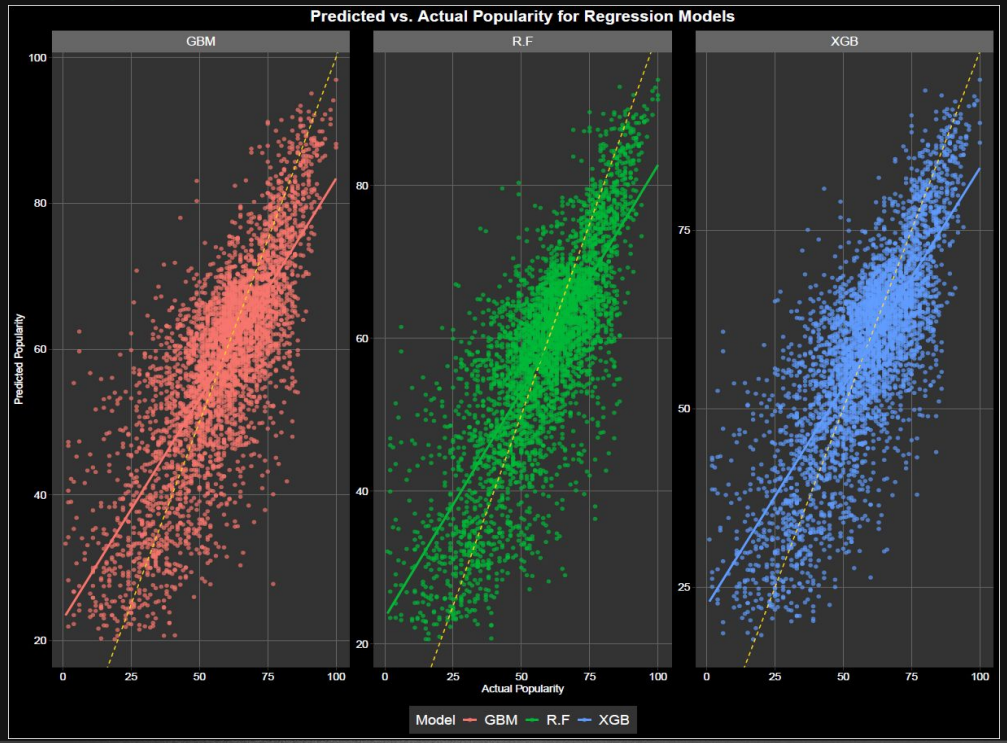
Regression Results - Performance Comparison

RMSE Comparison :

- Lower RMSE values indicate better model performance and smaller average prediction errors.
- The RMSE values were: Random Forest (RF) at 11.156, XGBoost (XGB) at 11.112, and Gradient Boosted Trees (GBM) at 11.263.



Regression Results - Actual vs. Predicted Popularity Plots



- Beyond RMSE and R-squared, examining plots provides insight into model behavior and potential biases.
- The dotted line represents y = x (perfect prediction).
- All models' plots are not dispersed too far from this line, indicating a reasonable alignment.
- XGBoost is relatively better in terms of how its predicted values align with actual values.

XGBoost was our best model, explaining 61% of popularity variations with audio features. However,this is not a perfect explanation, therefore we explored a classification approach.

Predictive Modeling - The Classification Approach

Objective: Enhance predictive accuracy by classifying songs into four popularity tiers: Very Low (1-24), Low (25-49), High (50-74), and Very High (75-100).

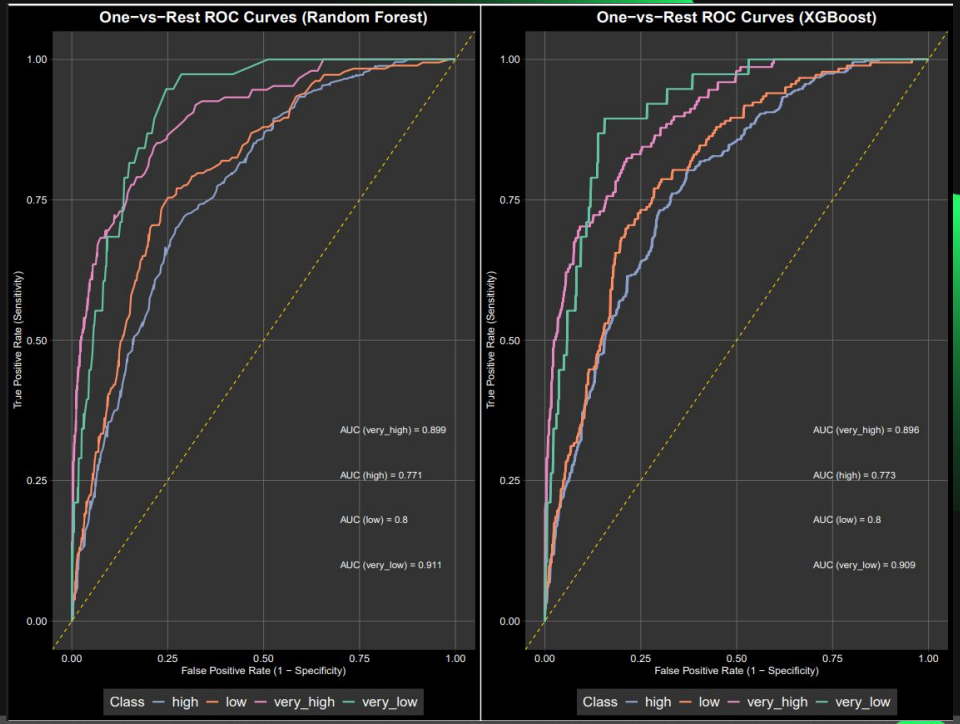
Methodology:

- **Categorization:** Continuous popularity converted to these four discrete classes.
- **Evaluation Strategy:** Used the same 80/20 train-test split and 5-fold cross-validation.
- **Metrics:** Model performance assessed using ROC curves and AUC for each category (one-vs-rest strategy).

Classification Results - ROC and AUC (Random Forest & XGBoost)

ROC and AUC for Random Forest :

- AUC values for Random Forest and XGBoost exhibit strong predictive power across all categories. They have AUC values ranging from 0.77 to 0.91
- The curve for "very_low" shows a steeper rise towards the top-left corner, suggesting a strong ability to achieve high sensitivity with a low false positive rate for this category.
- The models have almost similar results.

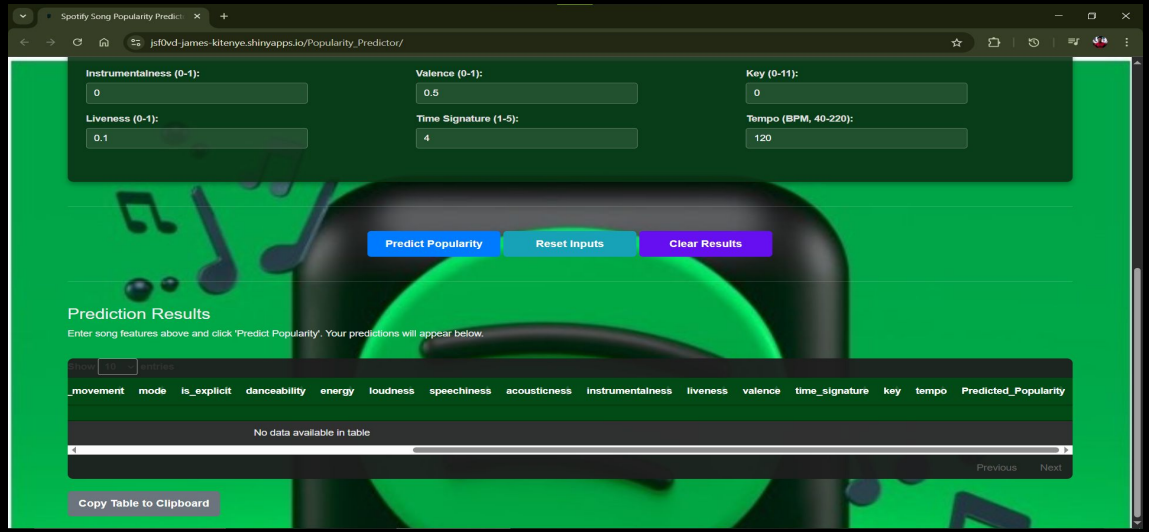



Practical Application Of Our Project Outcome
THE APP (Popularity_Predictor)

Practical Application: Prediction Tool

Song Popularity Prediction Tool

- Built using the **XGBoost model trained on Spotify Charts 2024 dataset**. Allows users to enter **audio features** → generates **predicted popularity score**.



APP LINK: https://jsf0vd-james-kitenye.shinyapps.io/Popularity_Predictor/


CONCLUSION

SUMMARY OF FINDINGS

Exploratory Data Analysis:

- Song popularity exhibited a non-normal distribution.
- We observed a complex relationships between audio features and popularity.

Machine Learning Approaches:

- Both regression and classification models showed predictive capabilities.
- **Classification** was more effective at distinguishing between popularity levels.

Model Performance:

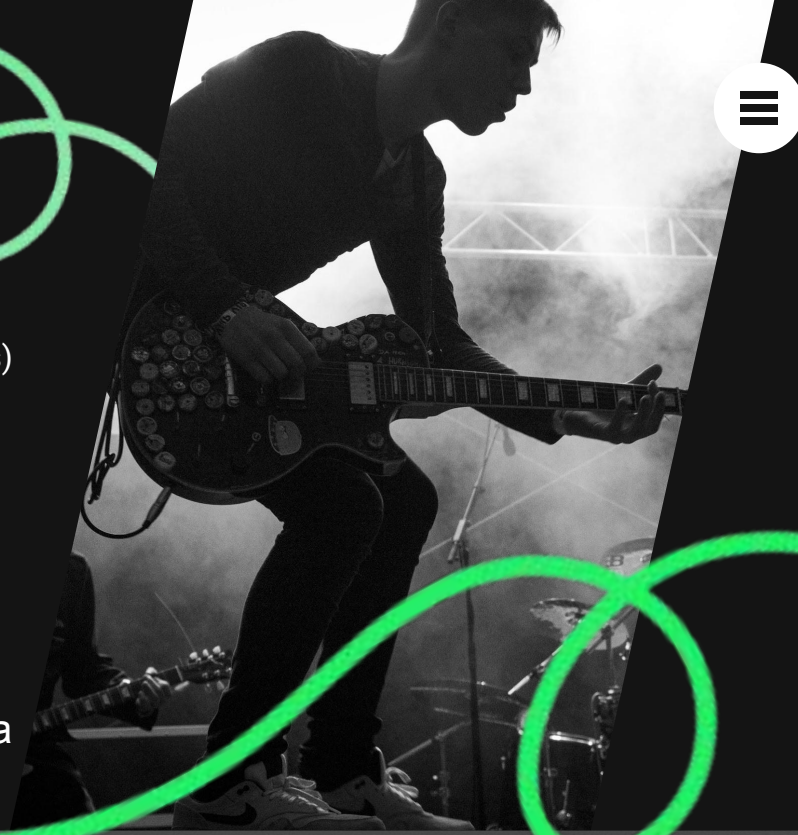
- **Regression (XGBoost):** Moderate predictive power (R^2 approx 0.61).
- **Classification (Random Forest):** Strong performance (high AUC values).

CONCLUSIONS

- Our analysis confirmed that audio features are crucial in understanding musical appeal that is providing valuable insights into their contribution to popularity.
- We successfully developed machine learning models that leverage audio features to estimate a song's potential popularity level.
- We compared regression and classification models, finding classification (especially Random Forest with high AUC) offered superior accuracy for discerning popularity levels.
- We built an App that predicts song popularity from Spotify audio features.This tool not only serves as a practical demonstration of the project's findings but also offers artists, producers, and record labels a data-driven glimpse into a song's potential reception in the competitive music market.

Study Limitations

- **Dataset Constraints:**
  - Some audio features (e.g., acousticness, instrumentalness) may have weak correlations with popularity, adding noise.
- **Low Correlations:**
  - Weak feature interactions **challenge model effectiveness**.
- **Missing External Factors:**
  - Important external influences on popularity (marketing, artist fame, playlisting, social media trends, release timing) were not include.



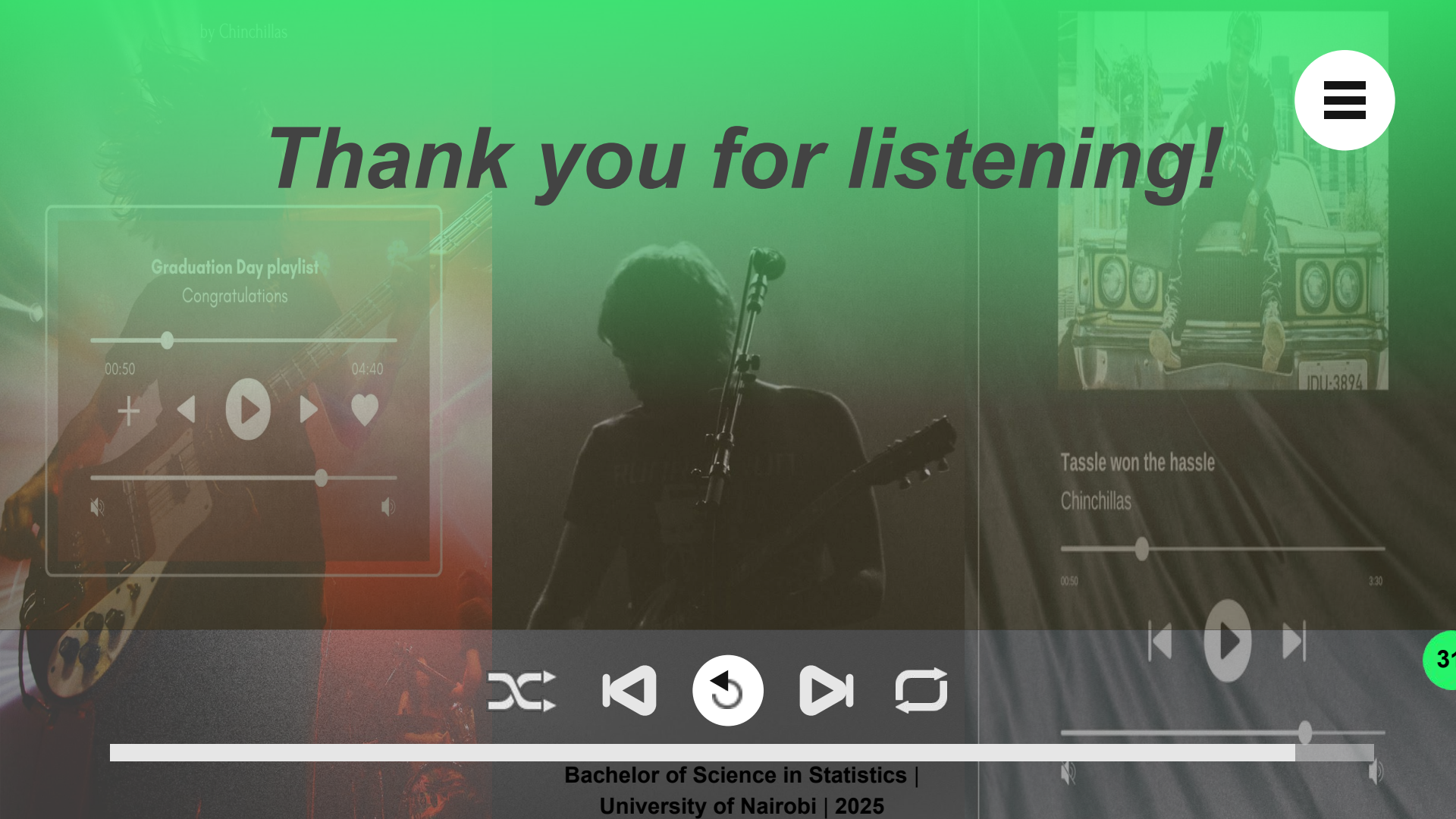
Future Works & Recommendations

- **Expanding Data Sources:**
  - Incorporate **social media trends, listener demographics, & contextual factors**.
- **Enhanced Feature Engineering:**
  - Investigate **interaction effects between independent variables**.
- Develop **more efficient data cleaning techniques** to optimize predictions
- **Time Series Analysis:** Model popularity changes over time.
- **Explore deep learning & hybrid models** for better accuracy

REFERENCES

1. *Beesa, P., Naregavi,* V., *Imandar, J., & Thatte, S. (2023)*. Songs Popularity Analysis Using Spotify Data: An exploratory study. Vidhyayana-An International Multidisciplinary Peer-Reviewed E-Journal-ISSN 2454-8596, 8(si7), 211-223.
2. *Dong, A., Qiu, R., & Ye, Z. (2023)*. Regression Analysis of Song Popularity based on Ridge,K-Nearest Neighbors and Multiple-Layers Neural Networks. Highlights in Science,Engineering And Technology, 39, 609–617.
3. *Kaur, Priya Pinder, and Sukhdev Singh*. “Random Forest Classifier Used for Modelling and Classification of Herbal Plants Considering Different Features Using Machine Learning.” Lecture Notes in Networks and Systems, 2022, pp. 83–94
4. *Kuhn, Max, and Kjell Johnson. Applied Predictive Modeling*. New York, Ny, Springer New York, 2013
5. https://github.com/kiteluva/popularity_predictor.

Thank you for listening!



Bachelor of Science in Statistics |
University of Nairobi | 2025