Carnegie Mellon University

Final Project:
Exploring the Dynamics of Music Popularity and Genre Classification Through Spotify's Data

Developed by Team 8:

Nicole Ma;

Chenwen Dong;

Xulin Wang;

Fei Xia;

Hanying Qiao

19603: Data Science for Technology, Innovation and Policy

April 26, 2024

Contents

1. Background and Motivation

In the dynamic music industry, understanding what drives song popularity and genre preferences is crucial for artists and producers aiming to optimize production and marketing strategies. Leveraging Spotify's comprehensive dataset, which includes detailed audio features like danceability, energy, and acousticness, our project aims to predict song popularity and genre classification. This analysis could provide actionable insights for the music industry, enhancing audience engagement and potentially increasing revenue by aligning music production with listener preferences.

2. Research Question

We are wondering - Can the popularity and genre of a song be accurately predicted based on its audio features using data from the Spotify API?

- **Objective 1:** Predict song popularity by audio features from Spotify data
- **Objective 2:** Predict song genres by audio features from Spotify data

3. Literature Review

The intersection of music analytics and predictive modeling has been an area of growing interest, with previous studies leveraging machine learning techniques to uncover patterns in music consumption and production.

4. Data Source

The Spotify Database API offers a robust framework for accessing detailed audio features and track popularity metrics. We leverage this API to gather data on thousands of tracks across various genres, providing a rich dataset to train our predictive models.

We utilized Python for data extraction, as detailed in our Spotify_Download_API.ipynb. The data includes features like Acousticness, Danceability, Energy, etc., scaled between 0 and 1.

5. Data Pre-Processing for Objective 1

Our report is structured to sequentially address each of our two main objectives. Each section first explores Objective 1—predicting song popularity—before delving into Objective 2—classifying song genres, thereby providing a comprehensive and parallel examination of both research questions.

5.1. Exploratory Data Analysis (EDA)

xxxx

Distribution of Popularity

We began our analysis by examining the distribution of song popularity using histogram (Figure 1 (a)). The distribution proved to be left-skewed, indicating a concentration of non-popular tracks. This skewness also sheds light on the limitation with regression model to accurately predict exact popularity scores.

The average popularity score in our dataset is identified at 16. To further probe this dynamic, we analyzed how many songs fall below or exceed this average (Figure 1 (b)). Given these insights, we are prompted to consider a classification approach. By categorizing songs into 'popular' or 'not popular' based on whether their popularity exceeds the average, we aim to develop a more robust model

Correlation Analysis

In our correlation analysis, we noted that the 'Energy' feature exhibited high correlations with multiple other audio features (Figure 2). To avoid potential multicollinearity, which could skew the predictive accuracy of our models, we opted to remove the 'Energy' variable from our dataset.

5.2. Feature Selection

To ensure the efficiency and accuracy of our models, a thorough feature selection process was implemented. We utilized the Random Forest algorithm to evaluate the importance of each feature within our dataset. The Random Forest model was trained using the caret package with a 5-fold cross-validation to ensure the stability and reliability of the feature importance scores (Figure 3 (a)).

To further refine our feature selection, we employed the Boruta algorithm (Figure 3 (b)). The Boruta process confirmed key variables as essential for predicting popularity: "Length, Acousticness, Danceability, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Tempo, Time_signature, valence"

6. Model Analysis for Objective 1

xxx

7. Result Interpretation for Objective 1

xxx

8. Data Pre-Processing for Objective 2

Having explored the predictive modeling of song popularity, we now shift our focus to Objective 2, which involves the classification of song genres based on audio features, to further our understanding of music dynamics on Spotify.

8.1. Exploratory Data Analysis (EDA)

In our initial exploration, we employed box plots (Figure XX) to analyze the distribution of various audio features (12 audio features in total) across different music genres. This approach helped us identify distinctive patterns and outliers within specific genres, guiding further analysis and feature selection.

Key Findings:

Acousticness: Jazz consistently exhibited the highest levels of acousticness, indicating its reliance on acoustic instrumentation. In contrast, electronic music showed the lowest levels, reflecting its synthetic production style.

Danceability: Hip-hop stood out with the highest danceability scores, aligning with the genre's rhythmic and dynamic nature. Conversely, rock displayed the lowest danceability, indicating less emphasis on rhythm-centric music production.

Valence: Pop music demonstrated high valence, suggesting a prevalence of positive, upbeat tracks within this genre. Electronic music, on the other hand, had lower valence, potentially indicating a broader emotional range, including more subdued or complex emotional expressions. These observations form the basis of our subsequent data engineering efforts and feature selection, ensuring our models capture the essential characteristics that distinguish between genres.

8.2. Data Normalization

To prepare for effective model training, particularly for algorithms like SVM, we normalized the numerical features to a standard range using the formula $\frac{x-\mathsf{min}(x)}{\mathsf{max}(x)-\mathsf{min}(x)}$.

8.3. Data Engineering

To enhance the predictive power of our models, we engineered interaction features. To understand the relationships better, we utilized scatter plots (Figures xxx) as a part of our exploratory data analysis. These plots help to uncover potential interaction effects that are vital for refining our data engineering strategies.

From Scatter Plots, we find:

- **Danceability vs. Energy:** High levels of danceability and energy are typically associated with upbeat and fast-paced music, evident in genres like Electronic and Hip-hop.

- **Acousticness vs. Instrumentalness:** These features effectively differentiate genres known for live performances from more electronically influenced music. Jazz, for instance, shows higher levels of both acousticness and Instrumentalness.

- **Speechiness vs. Valence:** This analysis is intriguing as it allows us to explore how genres that feature more lyrical content vary in emotional tone. Hip-hop, with its high speechiness levels, exhibits a wide range of emotional expressions, contrasting with genres like Jazz and Electronic, where instrumental elements predominate.

These analyses are integral to our data engineering process, as they inform the creation of features that can significantly enhance model accuracy by capturing complex patterns within the data.

8.4. Feature Selection

Using a decision tree as a base model for feature selection, we identified which features most significantly predict music genres.

Based on the horizontal bar chart for feature importance (Figure XX), Time_signature, Liveness, Tempo and Release_year are the least important features, which we exclude to simplify the model, reducing complexity without sacrificing accuracy.

9. Model Analysis for Objective 2

9.1. Model Choice for Genre Classification

In our endeavor to classify song genres from Spotify's audio features, we explored several machine learning models, each chosen for their distinct advantages:

- **Random Forest:** We opted for this model for its ensemble approach and robustness, which is crucial given the complexity and high dimensionality of our dataset. It also offers feature importance scores, allowing us to gauge the relevance of each audio feature for genre classification.

- **Support Vector Machine (SVM):** We utilized both radial and polynomial kernels to investigate which would best capture the nuances of our dataset. The SVM models are known for their capability to define complex hyperplanes in multi-dimensional space, which is beneficial for distinguishing nuanced genre differences.

- **Neural Network:** Due to its ability to model non-linear relationships, a neural network was a logical choice for capturing the intricate patterns that might be present between audio features and music genres.

9.2. Model Fitting

The fitting of our models was an integral process, which began by partitioning our dataset into training and testing sets to evaluate model performance. We keep 80% of our data for training, where the models learn the relationships between features and genres, and reserve 20% for testing, where the learned relationships are evaluated for accuracy.

We used functions like **train** from the **caret** package, which provides a streamlined way to apply complex machine learning algorithms to our data.

9.3. Model Validation

To ensure the robustness of our models, we utilized 5-fold cross-validation within our training control settings. This method enhances the validation process by dividing the training dataset into five parts, training the model on four parts, and validating it on the fifth. This cycle is repeated five times to minimize the variability in the validation process and ensure the model's performance is not dependent on a particular division of the data.

9.4. Model Tuning

For **Random Forest**, we tuned the 'mtry' parameter, which determines the number of features to consider when making splits, and 'ntree', which is the number of trees in the forest. After grid tuning, we found that mtry = 4 and ntree = 500 yielded the best balance between bias and variance.

In the case of **SVM**, we applied different tuning grids for the radial and polynomial kernels. For the radial kernel, our tuning grid searched through sigma values of 0.001 and 0.01 and C values of 10, 100, and 1000. For the polynomial kernel, we explored a degree of 2 and scale and C values in the ranges of 0.01, 0.1, 1 and 5, 10, respectively. Our SVM model comparison indicated that the Radial Basis Function (RBF) kernel had better accuracy than the Polynomial kernel. Therefore, in the later section of result interpretation, we only account for RBF kernel for comparison.

For **Neural Networks**, we experimented with 'size', which determines the number of units in the hidden layer, and 'decay', which is a parameter for regularization to prevent overfitting. We finalized on size = 10 and decay = 0.01 to balance model complexity and learning capacity.

10. Result Interpretation for Objective 2

In interpreting our results, we prioritized accuracy as the metric for comparison, which reflects the proportion of the total number of predictions that were correct. Accuracy is a natural choice for multi-class classification, providing a quick snapshot of model effectiveness. [Figures xxx]

The Random Forest model showed a slight edge over the others, with an accuracy of approximately 63.22%. This suggests that it might be more capable of managing the complexity inherent in multi-genre classification without overfitting.

Beyond overall accuracy, we evaluated the 'balanced accuracy' across genres [Figures xxx], which accounts for any imbalances in the class distribution. The accompanying bar chart reveals how each model fared in recognizing distinct features across genres, such as electronic and hip-hop, which were predicted with higher accuracy. Conversely, genres with overlapping characteristics presented more of a challenge, as depicted in the varied balanced accuracy scores. This nuanced understanding of model performance by genre is instrumental for future model improvements and applications.

11. Limitation

xxxxx

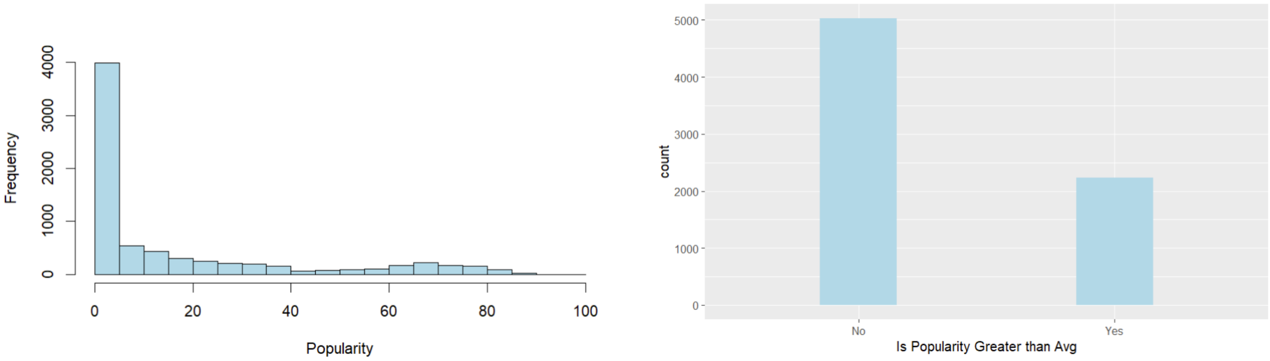
12. Conclusion

xxxxx

13. Reference

Guidelines and examples of this style are available at: purdue.edu/owl/chicago This section should start on a new page and be single spaced. All references should use hanging indentation

14. Appendix



(a) Distribution of Popularity Score

(b) Popularity Classification Relative to Average

Figure 1: Distribution of Popularity

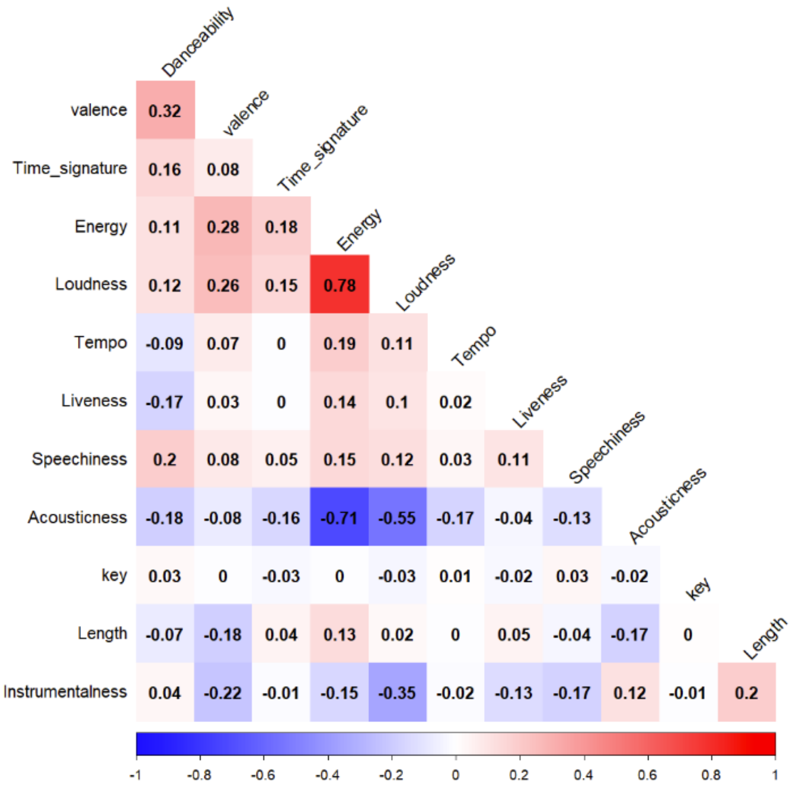
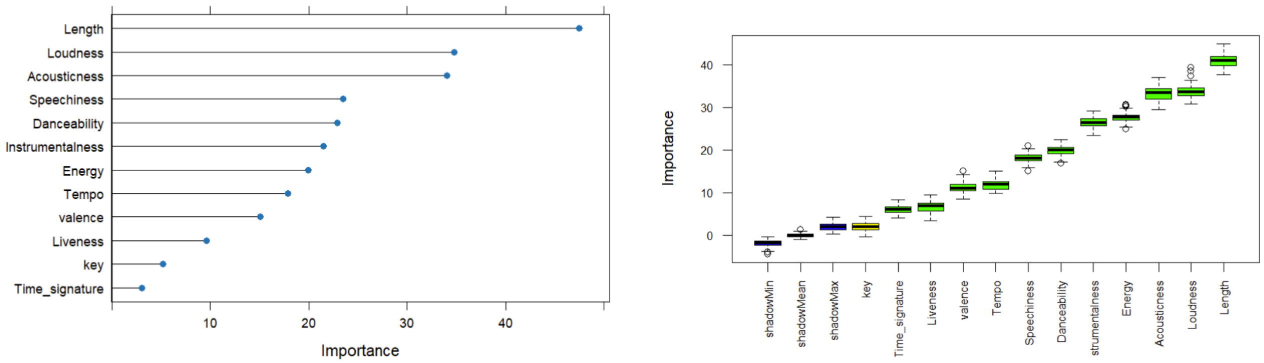


Figure 2: Correlation Matrix of Audio Features



(a) Feature importance by Random Forest

(b) Feature importance by Boruta

Figure 3: Feature importance for Feature Selection

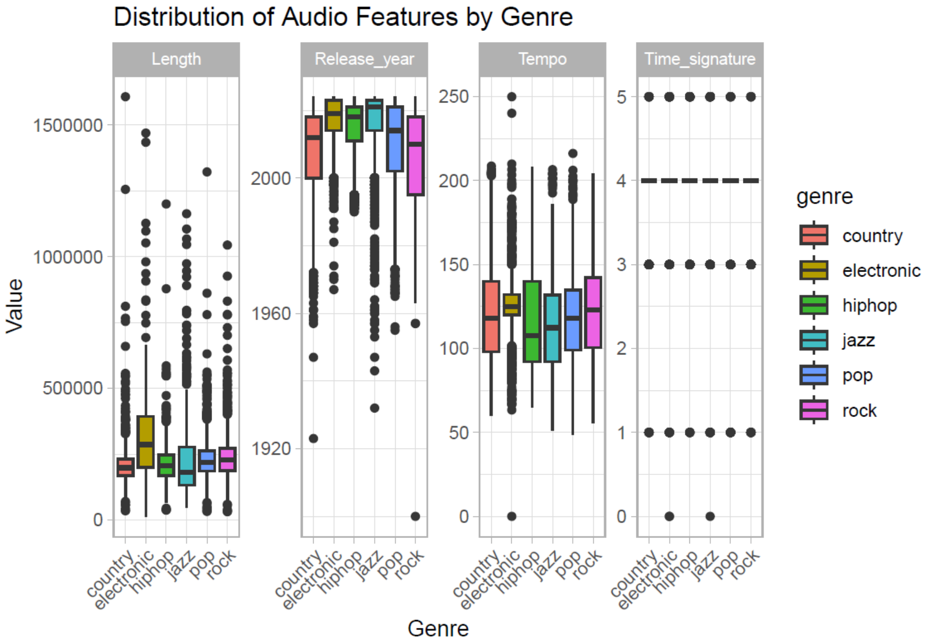


Figure 4: Boxplot 1: Audio Features by Genre

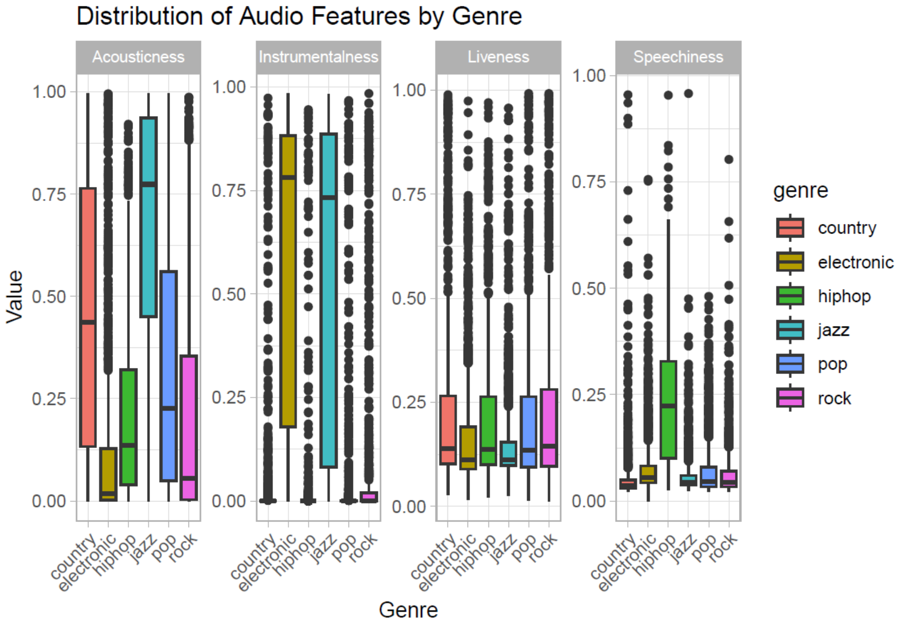


Figure 5: Boxplot 2: Audio Features by Genre

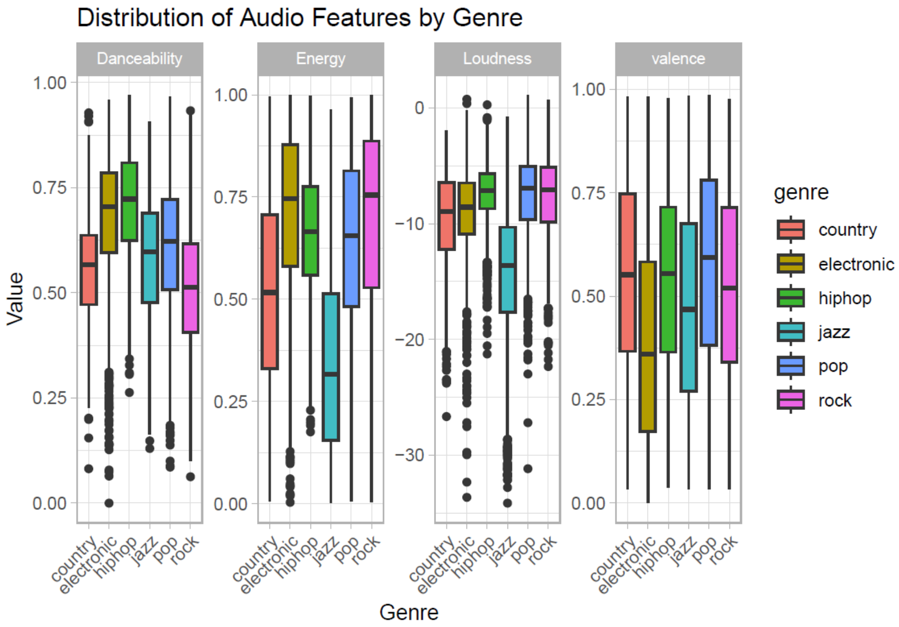


Figure 6: Boxplot 3: Audio Features by Genre

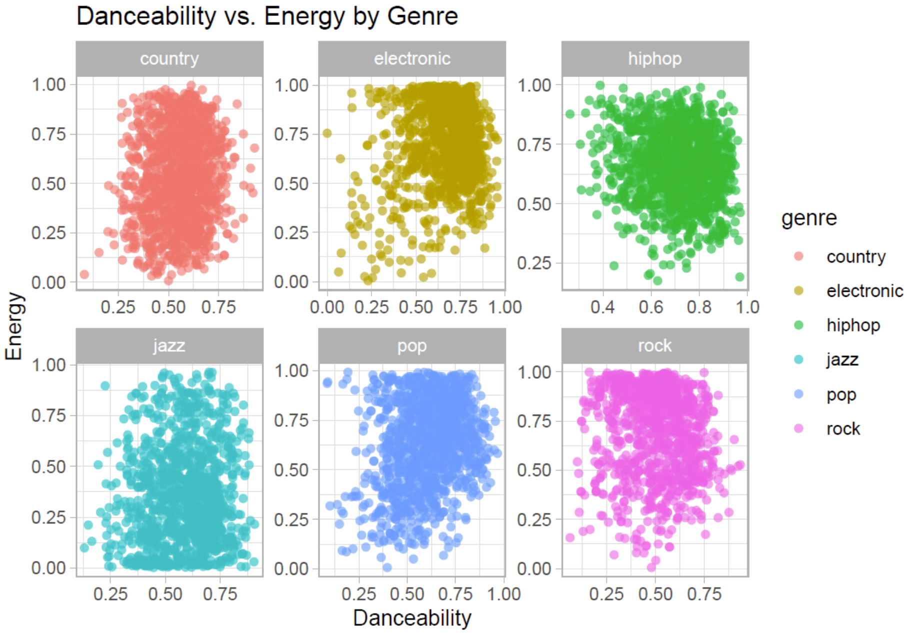


Figure 7: Scatter plot 1: Danceability vs. Energy by Genre

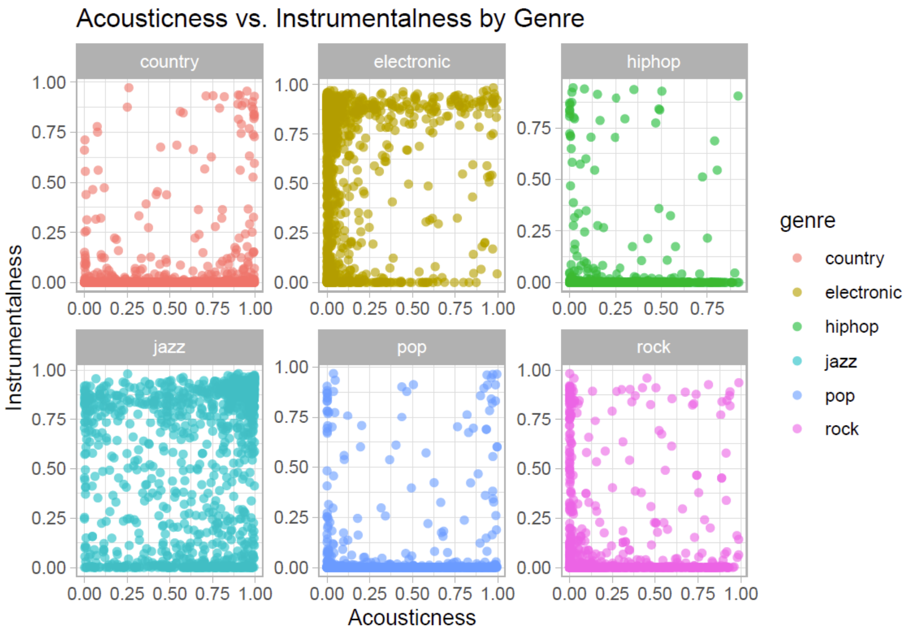


Figure 8: Scatter plot 2: Acousticness vs. Instrumentalness by Genre

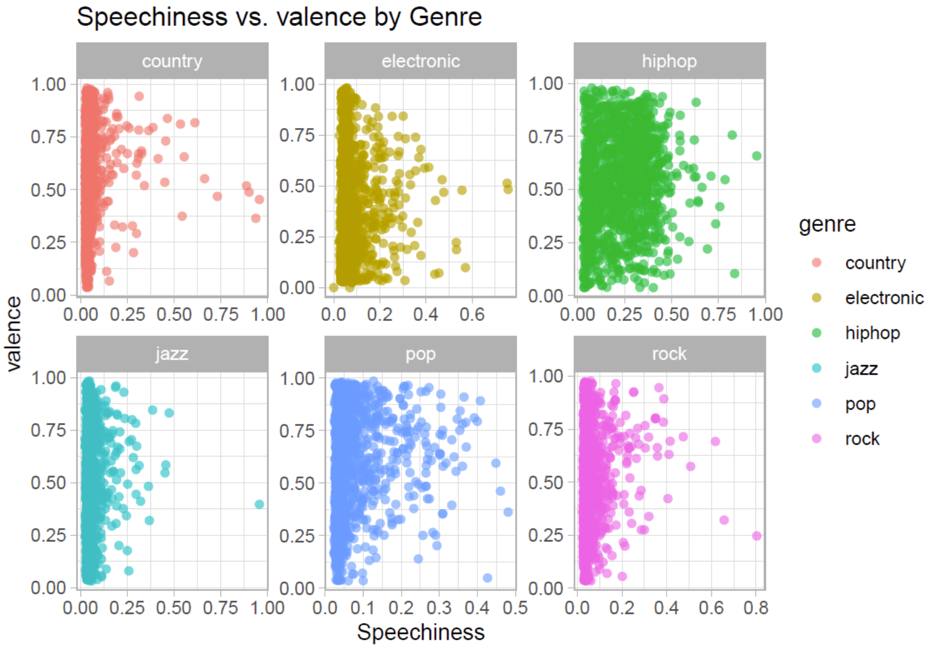


Figure 9: Scatter plot 3: Speechiness vs. valence by Genre

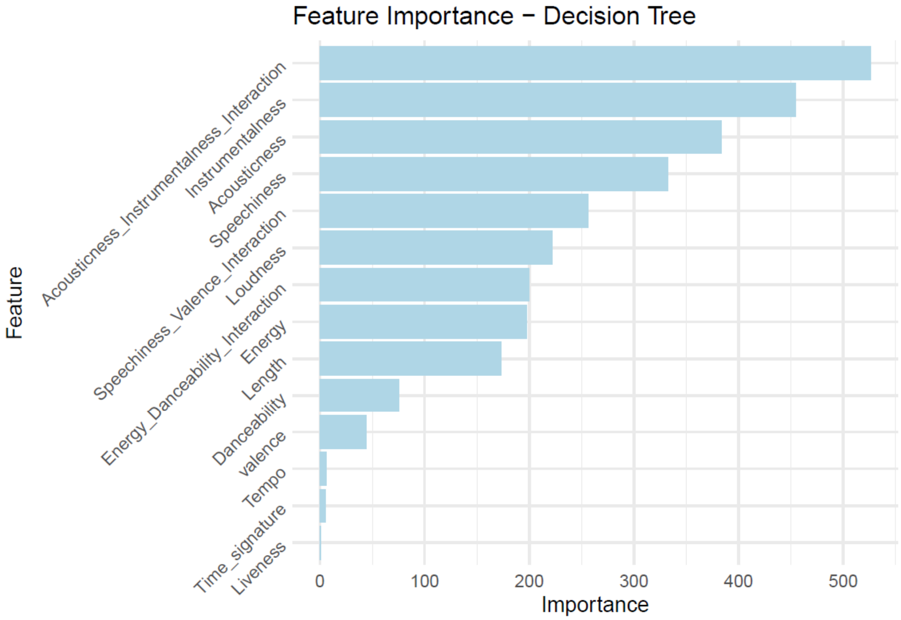


Figure 10: Feature Importance - Decision Tree

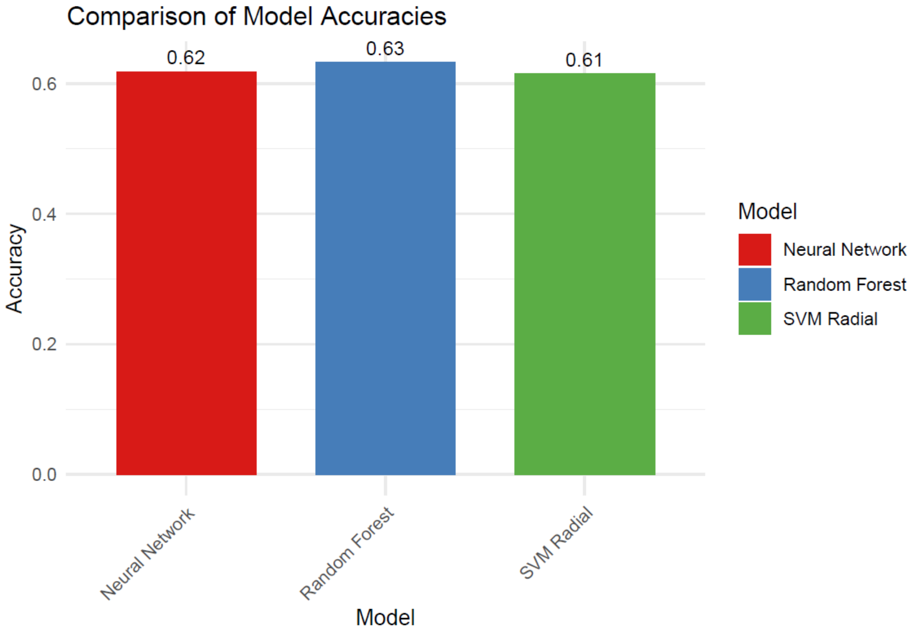


Figure 11: Comparison of Accuracy

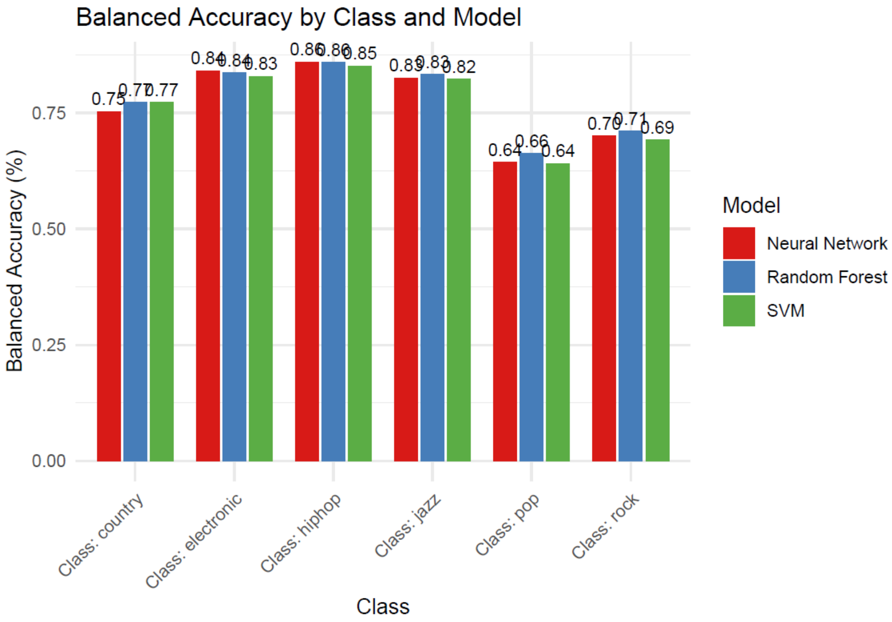


Figure 12: Comparison of Balanced Accuracy