Harmony Through Numbers: Classifying Spotify Data Across Genres and Decades

AUTHORS

Billy McGloin, Jorge Bris Moreno, Kangheng Liu, & Isfar Baset

AFFILIATIONS

Special thanks to our professors in the Data Science and Analytics Program at Georgetown University.

We employ a range of data science techniques—including logistic regression, support vector machines, neural networks, random forests, and extreme gradient boosting—to classify music data by genre and decade at both the artist and track levels.

01 Introduction

In this project, our group endeavors to bridge the gap between music and data science by employing Spotify's extensive datasets to perform assorted classification tasks. Using various metrics provided by Spotify, such as danceability, energy, and loudness, we aimed to categorize music tracks into genres and decades. Our goal was to implement sophisticated data analysis techniques learned in class to perform classification at both the artist and track levels. This project explores the feasibility of such classifications and tests the effectiveness of different machine-learning models in handling uniquely complex data like music.

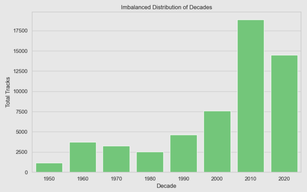
02 Our Data

To tackle this project, we utilized the spotifyr package to extract data for the top 50 artists across ten popular genres, resulting in an initial dataset that included over 400 artists. We then expanded our dataset by extracting every track for these artists, although we had to clean and preprocess the data due to duplicates and corrupted information. This process left us with a final dataset comprising approximately 50,000 tracks from roughly 400 artists. Additionally, we introduced a new variable, 'decade,' by binning song release dates, which was crucial for our subsequent temporal analysis.

03 Methodology

Our methodology involved several key steps: data aggregation, cleaning, and establishing classification labels. For artist-level data, we averaged the Spotify metrics to create a singular profile per artist, assuming a consistent style across their work. For genre classification, we started with multiple genres per artist, but narrowed it down to one primary genre due to significant overlap and the limitations of our dataset size. We applied various machine learning models—logistic regression, support vector machines, neural networks, random forests, and XGBoost, all with hyperparameter tuning—to predict genres and decades, adapting our approach based on the peculiarities of each dataset level (artist vs. track).

04 Exploratory Data Analysis

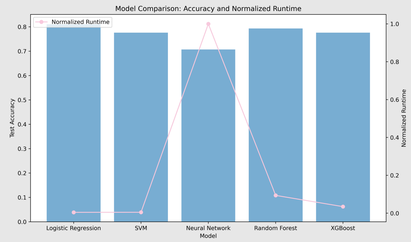


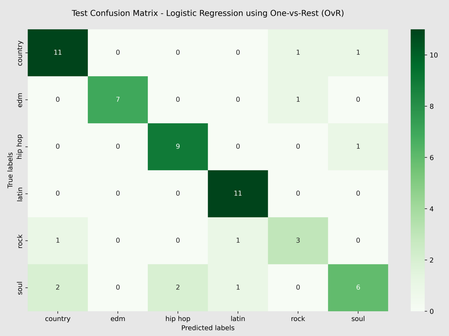
The above graph displays the number of tracks in each decade within our dataset. One can see that there is a severely imbalanced distribution, with many newer songs. To address this problem, we used various sampling methods to balance the training data for our models.

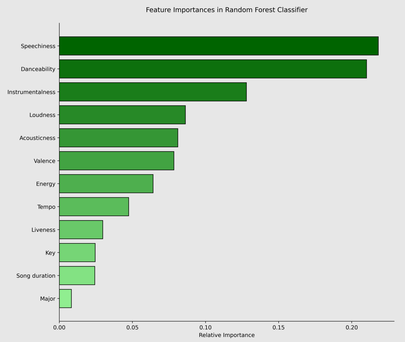
This heat map allows us to explore values of the assorted metrics, compare them between genres, and gain a better understanding of the various distributions. With many of the metrics having similar values across genres, classification becomes a harder task.

05 Classification: Genre

The genre classification at the artist level proved to be quite successful, routinely achieving test accuracies over 75%, with logistic regression peaking at 81%. This was significantly higher than a baseline model of random guessing, indicative of the effectiveness of our feature selection and machine learning techniques. The track-level genre classification, however, performed less optimally due to the inherent variability in songs by the same artist, peaking at 63% accuracy with XGBoost.

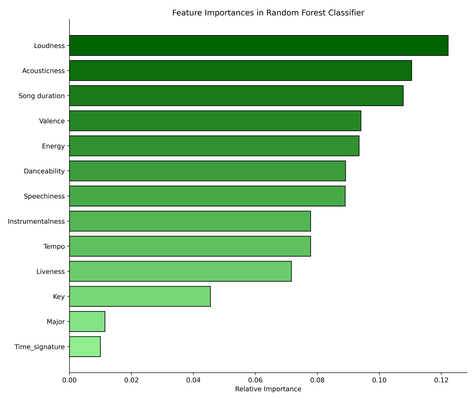


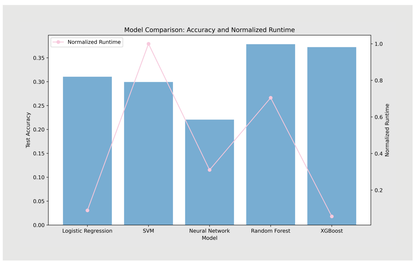




06 Classification: Decade

Decade classification posed a more significant challenge, primarily due to the imbalanced nature of our data—most tracks were from recent decades. After downsampling to even out the class distribution for the training data, we achieved modest accuracies of around 38%, with random forests and XGBoost being the top performers. Notably, when we filtered the dataset to include only rock genre tracks, the accuracy improved, suggesting genre-specific characteristics can significantly influence the success of decade classification.





The Random Forest Classifier provides insights into which variables most significantly impact the determination of a track's decade, highlighting the notable changes over time. Additionally, the plot above also allows for the comparison of various models' accuracies and runtimes. While these models aren't highly accurate, they notably outperform a random classifier, doubling accuracy in some instances.

07 Conclusion

Our findings demonstrate a promising but challenging path forward in classifying music data. While we achieved notable success in classifying artists into genres, decade classification could have performed better, highlighting the complexity of temporal influences in music. The variability within an artist's work across different songs suggests a potential for future research to refine track-level classifications. Overall, this project underscores the potential of machine learning in transforming our understanding of music through data, opening avenues for deeper exploration into the analytics of sound.

QR Code for More Information

Best Performance for Each Classification Task

TASK	ACC.
Artists into Genres	81.03%
Tracks into Genres	63.03%
All Tracks into Decades	41.23%
Rock Tracks into Decades	43.99%

08 Recommendations

Based on our project's findings and the challenges we encountered, we recommend the following strategies for future projects in music classification using machine learning:

1. **Expand Data Collection:** To enhance the models' generalizability, future projects should consider expanding the dataset to include a more diverse array of artists and tracks from additional genres and less represented decades. This expansion is crucial for mitigating issues related to data imbalances and providing a richer basis for classification.
2. **Refine Data Labeling Techniques:** In our project, track-level genre data was unavailable, necessitating the use of an artist's primary genre to label their songs. Further research could help investigate how an artist's style changes over time, possibly incorporating more granular and dynamic labeling methods that reflect these evolutions.
3. **Utilize Unsupervised Learning:** Unsupervised learning techniques such as clustering and PCA could be explored to better understand the underlying structure of the data and identify patterns without predefined labels. This approach might also uncover interesting insights about genre and decade classifications that are not immediately apparent.

By addressing these recommendations, future projects can build on our work's foundation and push the boundaries of what's possible at the intersection of data science and music analysis.