Exploring Feature Importance in Predicting Music Preference

Jackie Zhou
400186858

Department of Psychology, Neuroscience & Behaviour

Friday, December 6, 2024

Outline

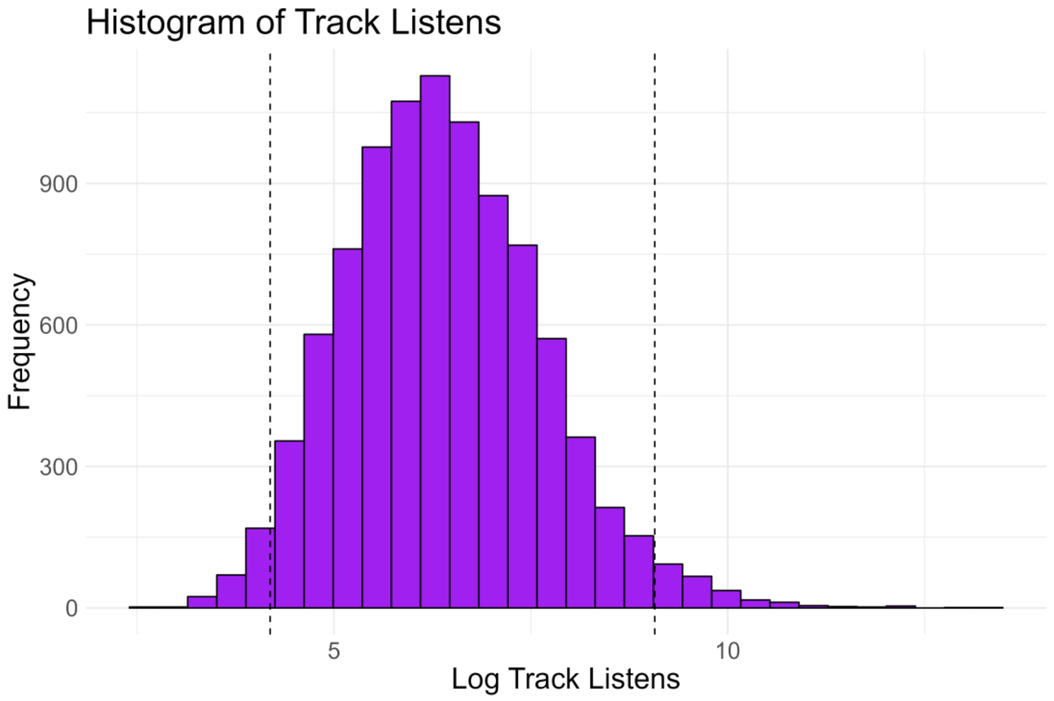
- Introduction + Motivation
- Data: Free Music Archive
- Methods
  - Principal Components Analysis
  - Regression Analyses: Random Forest and Boosting
- Results
  - Regression Models Comparison
  - PCA Feature Loadings
- Discussion

Introduction + Motivation

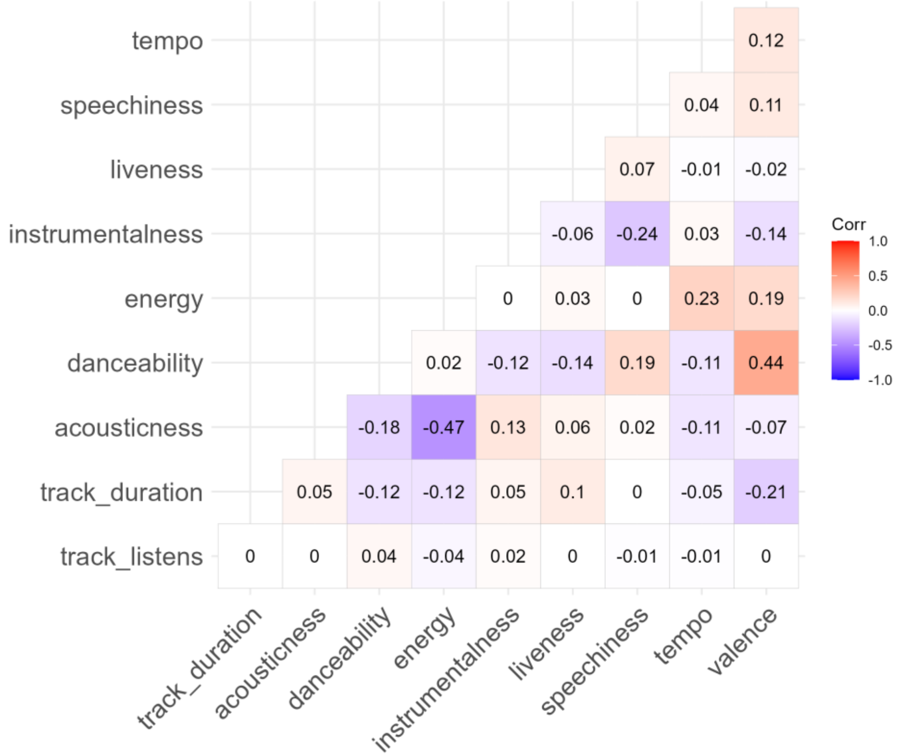
- Music popularity is influenced by a complex interplay of factors
    - Artist recognition, marketing, etc.
    - Qualities and acoustic features of music
- By analyzing acoustic features, we can determine specific musical features that predict the popularity and success of the song
- Further implications include identifying key features to prioritize for music recommendation, improving listener engagements

Data: Free Music Archive

- The Free Music Archive (FMA) dataset is a publicly available collection of over 106,000 tracks, each annotated with metadata
  - Popularity/success determined by the number of listens per track
  - Bottom and top 2.5% removed

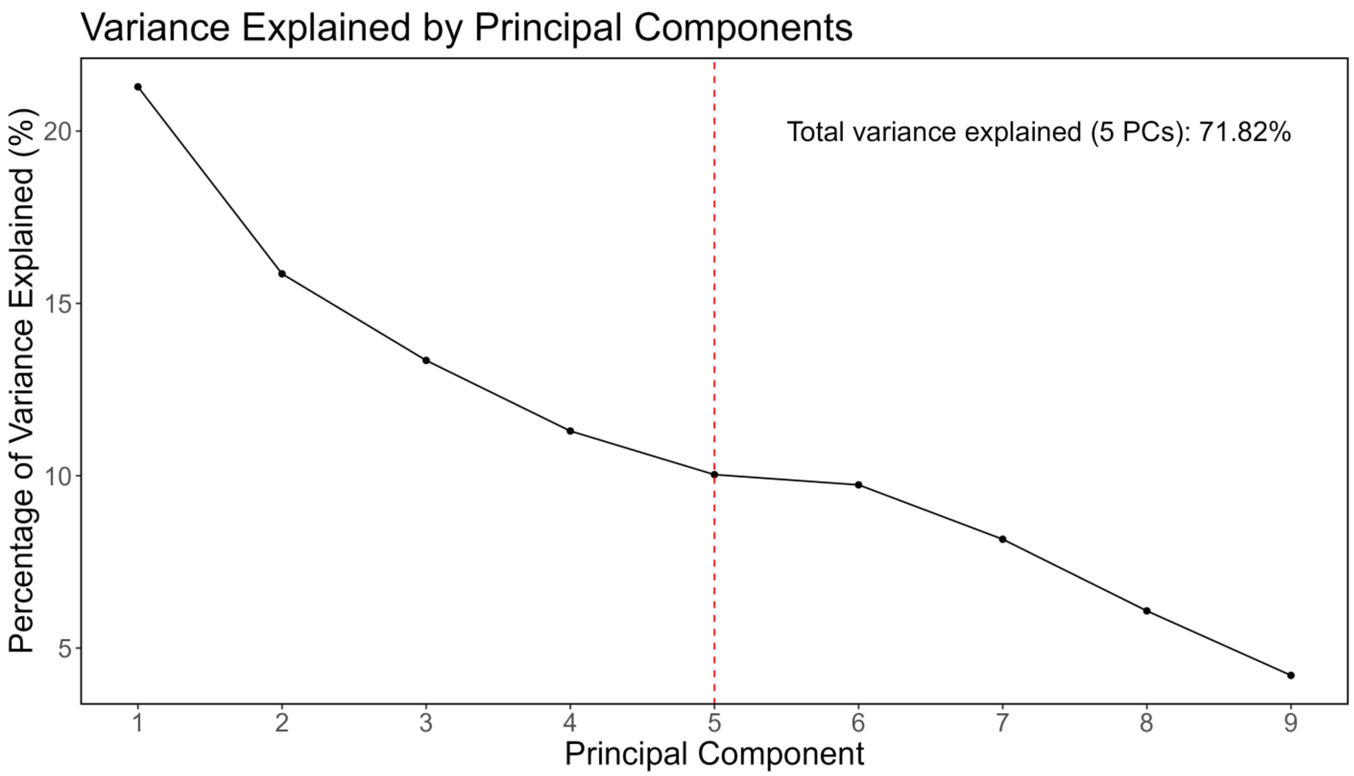


- ~9000 tracks included audio features extracted by Echo nest (Spotify's music AI platform)



Methods: PCA

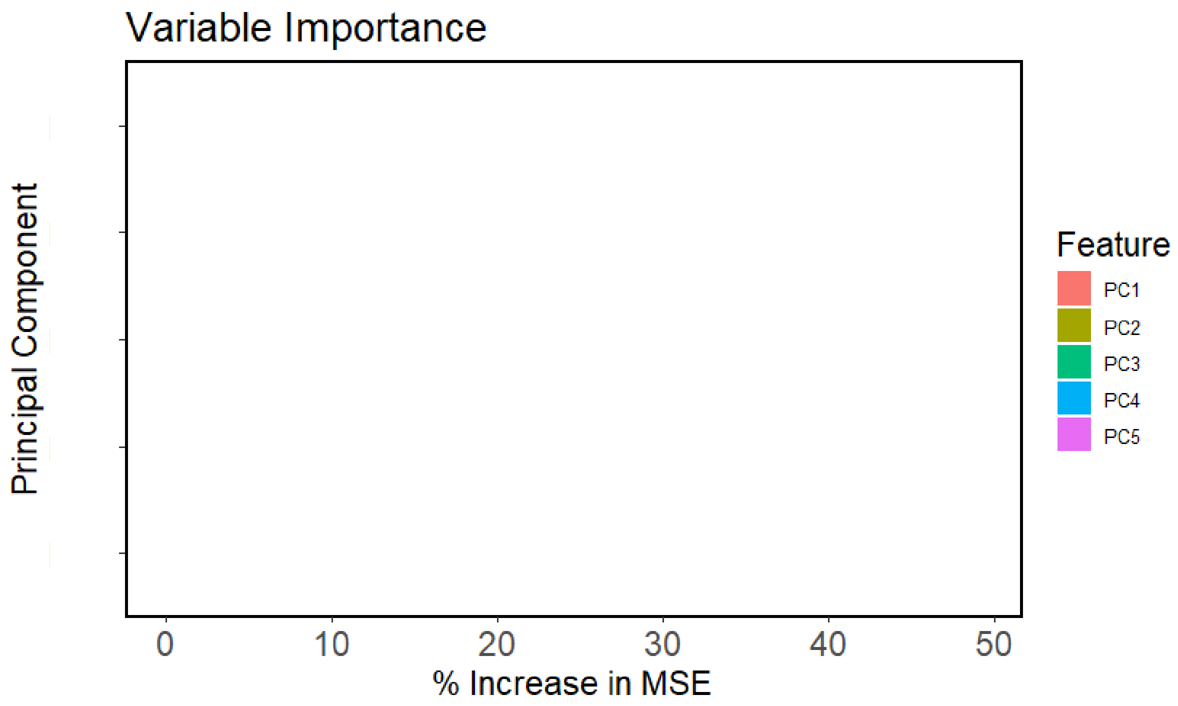
- Principal Components Analysis used for dimensionality reduction
- Used first 5 PCs as they explain ~72% of the total variance
  - log_track_listens ~ PC1 + PC2 + PC3 + PC4 + PC5



Methods: Models

Exploring the predictive power of PCs on music popularity using regression models

- 80% for Training and 20% for Testing
- Model Evaluation using
  - Feature (PC) Importance
  - RMSE and R-squared



Random Forest Regression

- Random Forest Parameter Tuning:
  - 500 Trees
  - mtry = 1

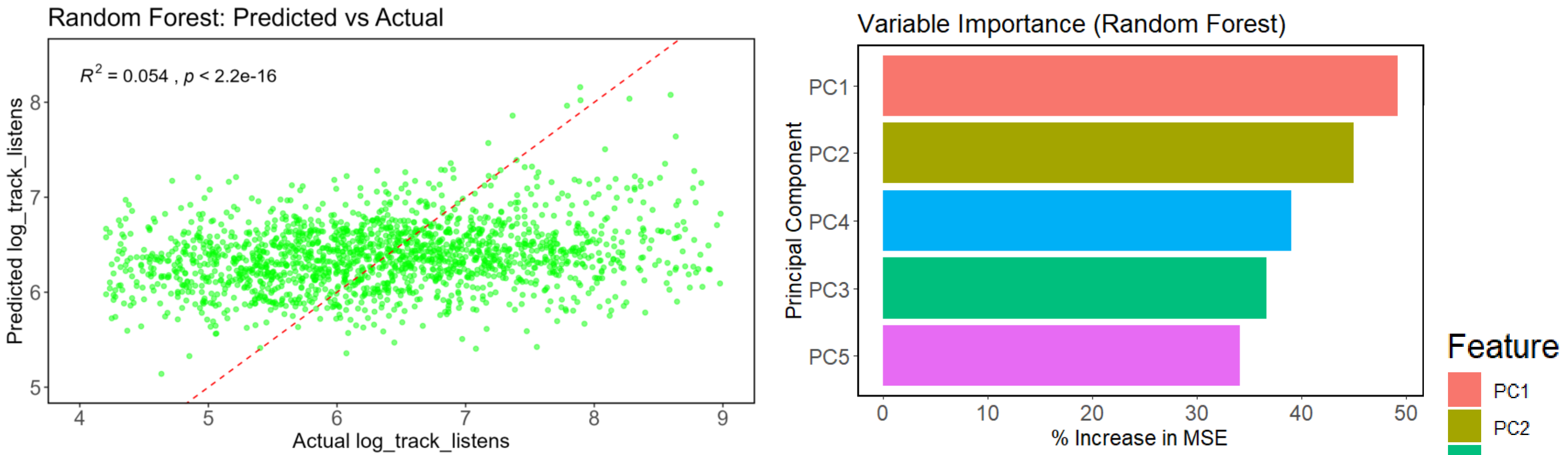
Regression with Boosting

- XGBoost Parameter Tuning:
  - 200 Trees (rounds of boosting)
  - Tested depth from 1-5
  - Minimizing test RMSE while using the lowest max_depth: depth of 3 at boost iteration 137

Results: Model Comparison & PCA Loadings

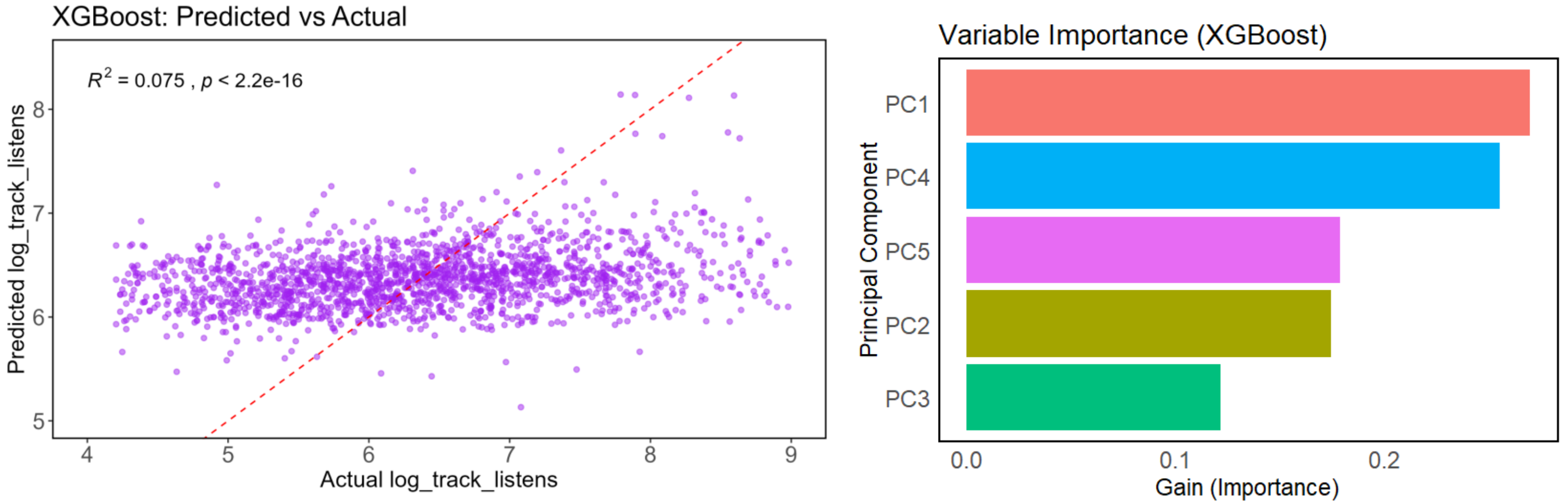
Random Forest Regression

- RMSE: 1.043
- R-squared = 0.054***
- Most Important: PC1, PC2



Regression with Boosting

- RMSE: 1.026
- R-squared = 0.075***
- Most Important: PC1, PC4



Discussion: PC and Feature Loadings

High RMSEs suggest that **acoustic features alone cannot predict** much variance in music preference, however, there was still a **small significant association** with the first 5 PCs on the number of listens the piece had in both models

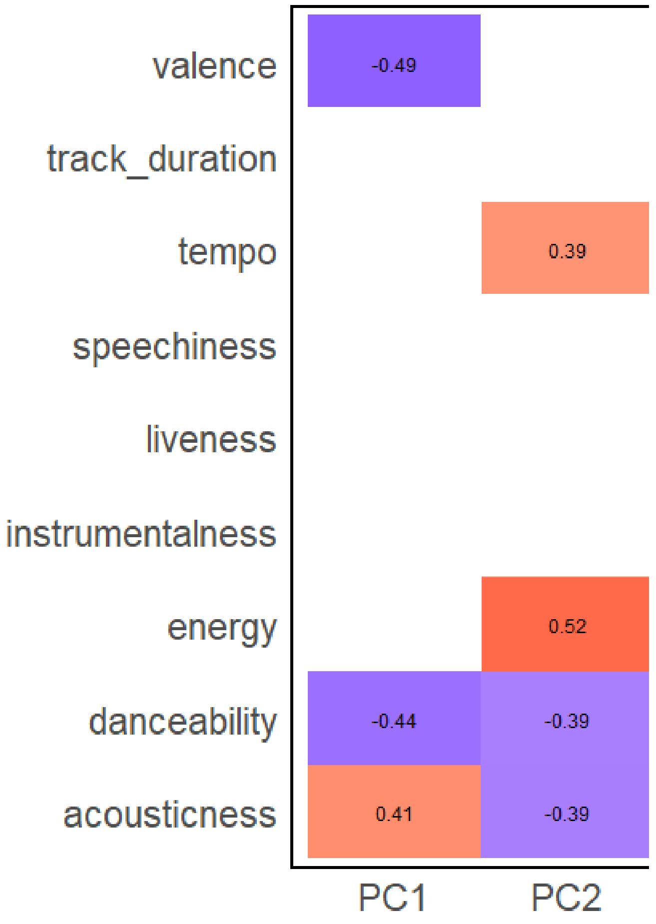
Random Forest Regression

- **PC1:** Low valence, low danceability, high acousticness
  - *Sad, slow, acoustic music*
- **PC2:** High energy & fast tempo, low danceability, low acousticness
  - *Fast, energetic, electronic music*

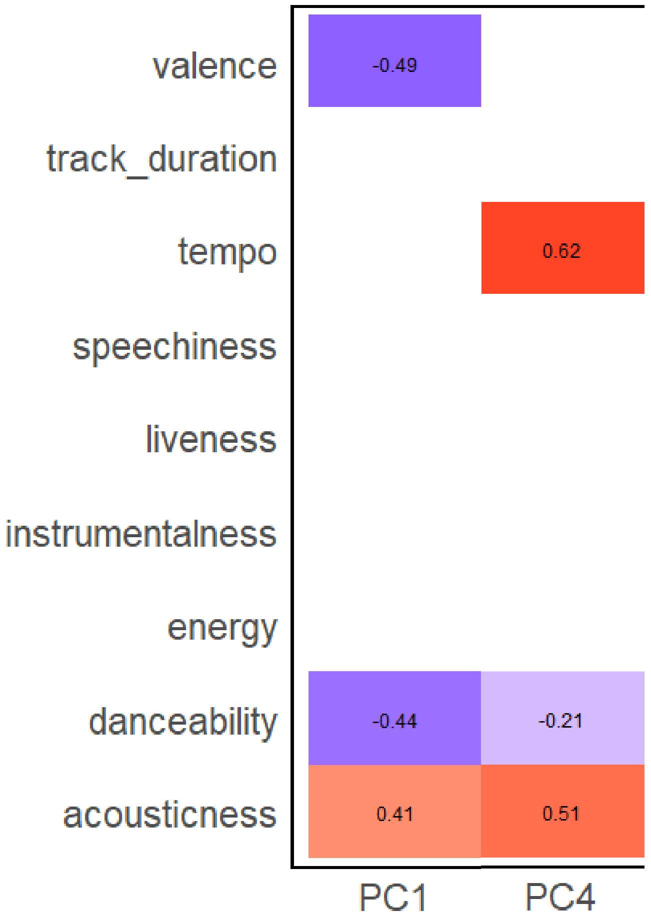
Regression with Boosting

- **PC1:** *Sad, slow, acoustic music*
- **PC4:** Fast tempo, high acousticness, short duration
  - *Short, upbeat, acoustic music*

Random Forest Regression



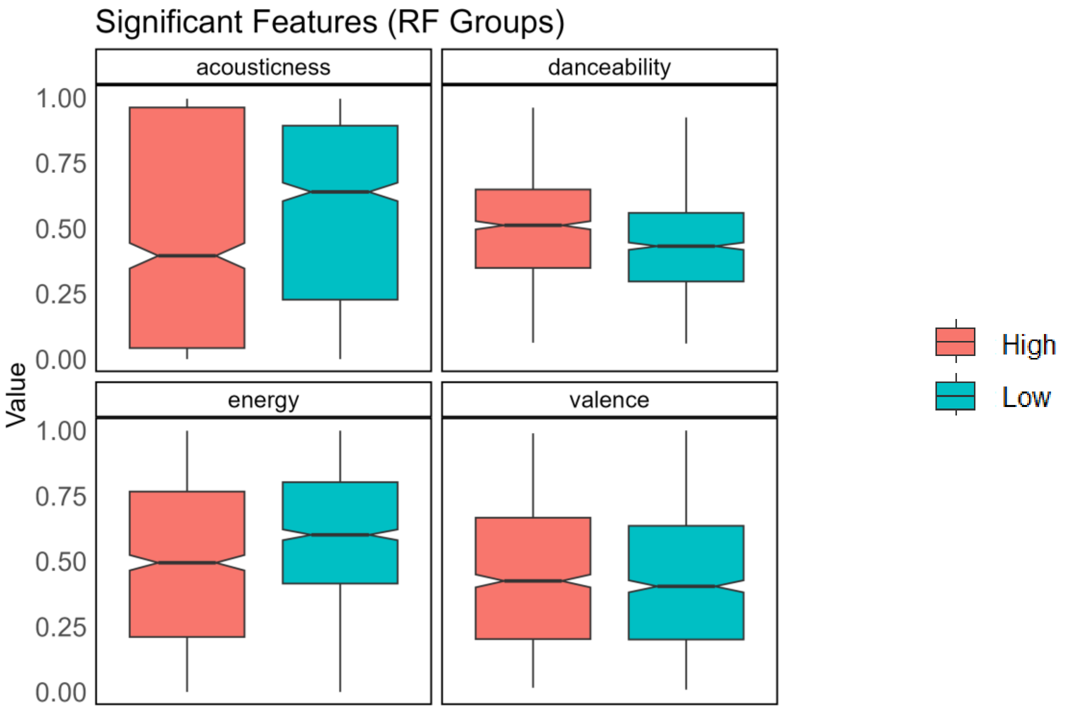
Regression with Boosting



Discussion: PC and Feature Loadings

Random Forest Regression

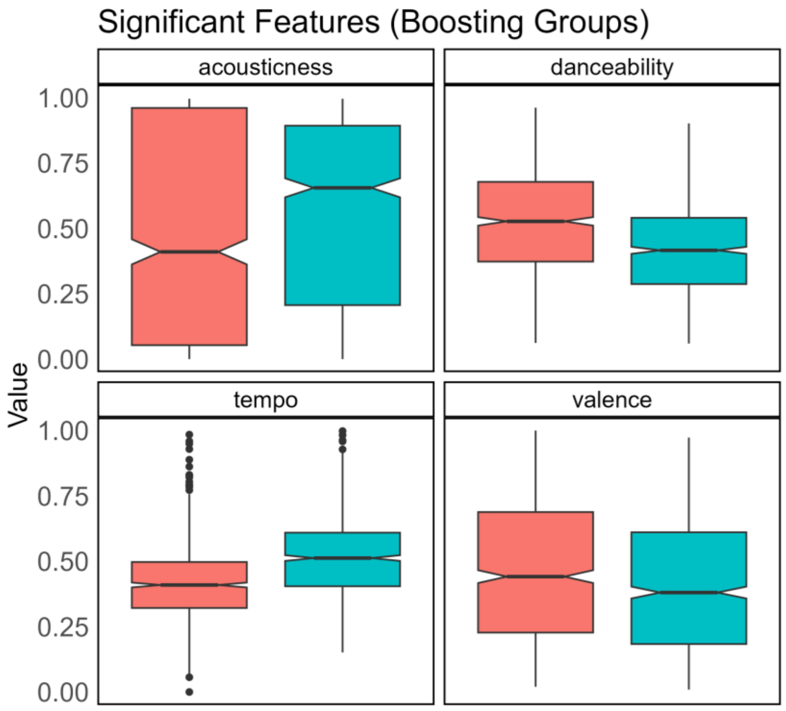
- Valence, Energy, Danceability, Acousticness



According to the RF regression model, listeners prefer music with less acoustics, more danceable, moderate to low energy, and slightly more positive valence.

Regression with Boosting

- Valence, Tempo, Danceability, Acousticness



According to the boosting regression model, listeners prefer music with less acoustics, more danceable, moderate to slower tempo, and positive valence.

Thank you!

Questions?

References

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Defferrard, Michaël, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2016. "FMA: A Dataset for Music Analysis." arXiv Preprint arXiv:1612.01840.

Ellis, Daniel PW, Brian Whitman, Tristan Jehan, and Paul Lamere. 2010. "The Echo Nest Musical fingerprint."

Liaw A, Wiener M (2002). "Classification and Regression by randomForest." R News, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.

Lin, Ning, Ping-Chia Tsai, Yu-An Chen, and Homer H Chen. 2014. "Music Recommendation Based on Artist Novelty and Similarity." In 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), 1–6. IEEE.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686.

Zhang, Jurui, Shan Yu, Raymond Liu, Guang-Xin Xie, and Leon Zurawicki. 2024. "Unveiling the Melodic Matrix: Exploring Genre-and-Audio Dynamics in the Digital Music Popularity Using Machine Learning Techniques." Marketing Intelligence & Planning