STATS/CSE 780 - Final Project Proposal

Jackie Zhou (400186858)

2024-10-31

Introduction

Music popularity is influenced by a complex interplay of factors, including not only artist recognition and marketing but also the inherent qualities of the music itself. Vast amounts of research has investigated correlates of popularity, including investigating acoustic features, to artist popularity (Lin et al. (2014); Zhang et al. (2024)). By analyzing acoustic features alongside relevant metadata, it is possible to gain insights into the characteristics that make certain songs resonate more widely with audiences.

The Free Music Archive (FMA) dataset (Defferrard et al. (2016)) is a publicly available collection designed to support music information retrieval research and other studies within the field of computational musicology. The FMA dataset includes over 106,000 tracks, each annotated with extensive metadata, such as track titles, artist names, genre classifications, and listening statistics. This makes the FMA dataset a valuable resource for examining trends and patterns in music, allowing researchers to analyze relationships between musical content, metadata, and listener behavior.

One unique feature of the FMA dataset is its inclusion of audio features extracted by Echonest (Ellis et al. (2010)), a music intelligence platform acquired by Spotify. Echonest uses advanced algorithms to analyze tracks, providing a set of eight audio features: acousticness, danceability, energy, instrumentalness, liveness, speechiness, tempo, and valence. These audio features have been widely used in MIR tasks, such as genre classification and music recommendation, as they offer insights into the sonic characteristics that differentiate musical styles and appeal to listeners.

In this study, we aim to investigate whether a song's *musical features* (Echonest features) and *metadata* (e.g., genre, artist, and duration) can accurately predict its popularity. Here, *popularity* is defined by the number of times a track has been listened to, with tracks categorized into different popularity levels based on their `track_listens` counts. Understanding the predictive power of these musical characteristics in relation to popularity could provide valuable insights for music producers, artists, and industry professionals who aim to create music that resonates with a broad audience. Moreover, insights from this research could have practical applications in music recommendation systems, where algorithms aim to deliver tracks that match listener preferences and have a higher likelihood of popularity. By identifying key audio features and metadata that contribute to popularity, recommendation algorithms could prioritize these characteristics when suggesting music, potentially improving listener engagement.

Methods

All data processing, analyses, and visualization was performed in R (R Core Team (2020)). The tidy dataset was saved in an R data format. During data preparation, missing values in the `track_genre_top` column were removed. Our final dataset has a total of 9355 tracks. The 8 Echonest features are continuous numerical values, and categorical variables include artist names, genres, and popularity. Track popularity was equally binned into 10 categories based on track_listens, which provided a balanced categorical target variable.

Dimensionality Reduction with PCA: The eight Echonest features may contain overlapping information. PCA reduces these features to a smaller set of principal components, which capture the most variance in the data while minimizing redundancy. These components, along with relevant metadata (e.g., track_duration), were then used as inputs for the random forest classifier.

Random Forest for Classification: Random forest provides feature importance scores, allowing us to assess the contribution of each principal component or raw features and original metadata to the prediction task. Random forest was chosen for its balance of accuracy, interpretability, and ability to handle complex interactions among features. By comparing these two models, we aimed to balance predictive performance with interpretability. We hypothesized that the random forest with PCA components would achieve slightly better performance by reducing noise and redundancy, while the random forest with features would provide greater insight into the contribution of each individual feature to track popularity.

To investigate the predictive power of musical features and metadata on popularity, we implemented two approaches: (1) Random Forest with PCA Components and (2) Random Forest with Original Raw Features.

Preliminary Analysis

Figure 1 (in Supplementary) shows the distribution of each Echonest feature across different genres. This highlights that different genres exhibit distinct patterns in their acoustic characteristics. Figure 2 (in Supplementary) shows the distribution of `track_listens`, or the number of times each track has been listened to. The distribution is highly positively skewed, with most tracks having relatively low listen counts, and only a few tracks reaching higher popularity levels. To address this and create

balanced target categories, we used10 equal-sized bins of tracks, ensuring that each "popularity" category has an equal number of tracks for a fair comparison across different levels of popularity. A correlation matrix (Figure 3 in Supplementary) shows the relationships among the Echonest features and their associations with popularity. No single feature shows a strong correlation with popularity, though there are correlations among some Echonest features. For example, there is a correlation between energy and acousticness (-0.47) and between valence and danceability (0.44). These relationships suggest that certain musical characteristics may interact in meaningful ways, which may influence a track's potential for popularity.

Given the moderate correlations between some features and the diversity in genre-specific attributes, PCA will be used to reduce dimensionality by identifying the most significant components among the metadata and audio features. This transformation will help to capture the main variability within the data, allowing us to reduce redundancy and focus on the most informative aspects of each track's acoustic profile. After dimensionality reduction, we will use random forests to classify tracks into popularity categories based on these principal components and relevant metadata. The random forest model provides an out-of-bag (OOB) error estimate, which we will use to measure accuracy. This OOB error, derived from the data samples left out during the training of each tree, gives a reliable measure of the model's performance. In addition to accuracy, feature importance scores from the random forest will allow us to interpret which components or features contribute most significantly to track popularity. By comparing the performance of random forest models trained with (1) the PCA components and (2) the raw features, we aim to balance interpretability and predictive power, gaining insight into the factors most strongly associated with track popularity.

Project Timeline

I have set the following milestones: I will be familiarizing myself with PCA and the random forest method in the next 2 weeks, aiming to finish my results and conclusions by November 14. That gives me a week to prepare the presentation slides for November 21. I will be writing and practicing a script for my oral presentation for the following week. As I finish my results, and prepare for the presentation, I will be continually working on writing notes and drafting the final report. I will be finalizing and cleaning up my code for the final project report by the beginning of December. I will give myself all of December to finalize my report (writing and plots) to hand in by December 10.

References

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."

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/.

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*.

Supplementary Materials

```
# FMA Dataset
# https://github.com/mdeff/fma
# Pull Data
temp <- paste(tempfile(), ".zip", sep = "")
options(timeout = 60 * 10)
download.file("https://os.unil.cloud.switch.ch/fma/fma_metadata.zip", temp)

# Feature Data
# Consolidate multiline header
echonest_colnames <- unz(temp, "fma_metadata/echonest.csv") %>%
  read_csv(n_max = 0, skip = 2) %>%
  rename(track_ID = "...1") %>%
  names()
# Read the data
echonest_raw <- unz(temp, "fma_metadata/echonest.csv") %>%
  read_csv(skip = 4, col_names = echonest_colnames) %>%
  # Transform track_ID to integer for tibble merging
  mutate(track_ID = as.integer(track_ID))

# Remove temporal features
echonest <- echonest_raw[, c(1:26)]

# Metadata
# Consolidate multiline header
metadata_colnames_a <- unz(temp, "fma_metadata/tracks.csv") %>%
  read_csv(n_max = 0, skip = 1) %>%
  rename(track_ID = `...1`) %>%
  names() %>%
  # Removing strange encoding
  sub("\\...*", "", .)
```

```r
metadata_colnames_b <- unz(temp, "fma_metadata/tracks.csv") %>%
  read_csv(n_max = 0) %>%
  names() %>%
  # Removing strange encoding
  sub("\\...*", "", .)

metadata_colnames <- paste(metadata_colnames_b, metadata_colnames_a, sep = "_")

# Read the data
metadata_raw <- unz(temp, "fma_metadata/tracks.csv") %>%
  read_csv(skip = 3, col_names = metadata_colnames) %>%
  rename(track_ID = `_track_ID`) %>%
  # Transform track_ID to integer for tibble merging
  mutate(track_ID = as.integer(track_ID))
# Combine the data and metadata
data <- inner_join(metadata_raw, echonest, by = "track_ID")

# Genre data
genres <- unz(temp, "fma_metadata/genres.csv") %>%
  read_csv()

# Clean up downloaded files
unlink(temp)

# Tidy Data
df_tidy <- data %>%
  select(c("track_ID", "artist_id", "artist_name.x",
           "track_duration", "track_genre_top", "track_genres",
           "track_listens", "acousticness", "danceability", "energy",
           "instrumentalness", "liveness", "speechiness", "tempo", "valence")
         )
```

```r
# Adding names to track_genres based on genres dataset
df_tidy <- df_tidy %>%
  # Separate `track_genres` into rows (one genre per row) by removing brackets and splitting
  mutate(track_genres = str_remove_all(track_genres, "\\[|\\]")) %>%
  separate_rows(track_genres, sep = ",") %>%
  mutate(track_genres = as.integer(track_genres)) %>%
  # Join with genres.csv to get genre names
  left_join(genres, by = c("track_genres" = "genre_id")) %>%
  # Group back by track_ID and collapse genre names into a single string
  group_by(track_ID) %>%
  mutate(track_genres_named = str_c(title, collapse = ", ")) %>%
  ungroup() %>%
  # Select relevant columns and drop duplicates
  select(-title, -track_genres, -parent, -top_level, -`#tracks`) %>%
  distinct() %>%
  # Dropping rows with missing genre
  drop_na(., track_genre_top)

# Write .RData
save(df_tidy, file = "df_tidy.RData")
```

```r
# Load data
load("df_tidy.RData")
```

```r
# Normalize Tempo
df_tidy <- df_tidy %>%
  mutate(tempo = (tempo - min(tempo)) / (max(tempo) - min(tempo)))



# Summary of features by genre
ggplot(df_tidy %>%
```

```
        gather(feature, val, 7:14),
      aes(x = feature, y = val, colour = feature)) +
  geom_boxplot() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  theme(legend.position = "none") +
  ylab("Value") +
  facet_wrap(~ track_genre_top, nrow = 2)
```

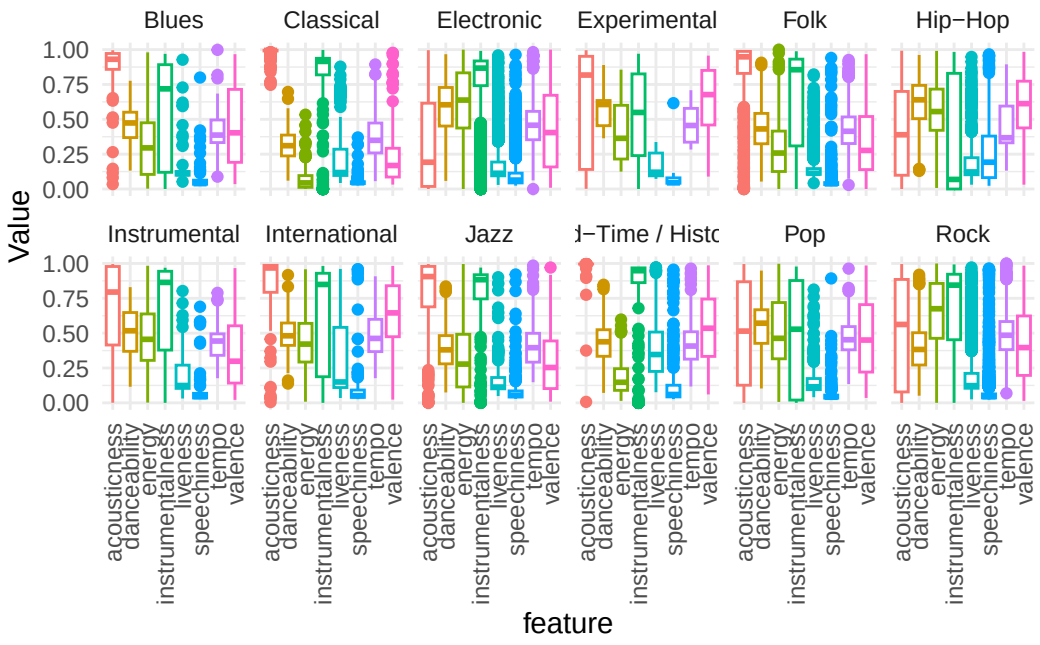

Figure 1: Summary of features by genre. Echonest features are normalized on the x-axis while mean values and boxplot summaries are displayed.

```
# Plot histogram of bottom 95% of track_listens
ggplot(df_tidy %>% filter(track_listens < quantile(track_listens, 0.95)), aes(x = track_list
  geom_histogram(binwidth = 200, fill = "purple", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Track Listens (Filtered)",
       x = "Track Listens",
       y = "Frequency")
```

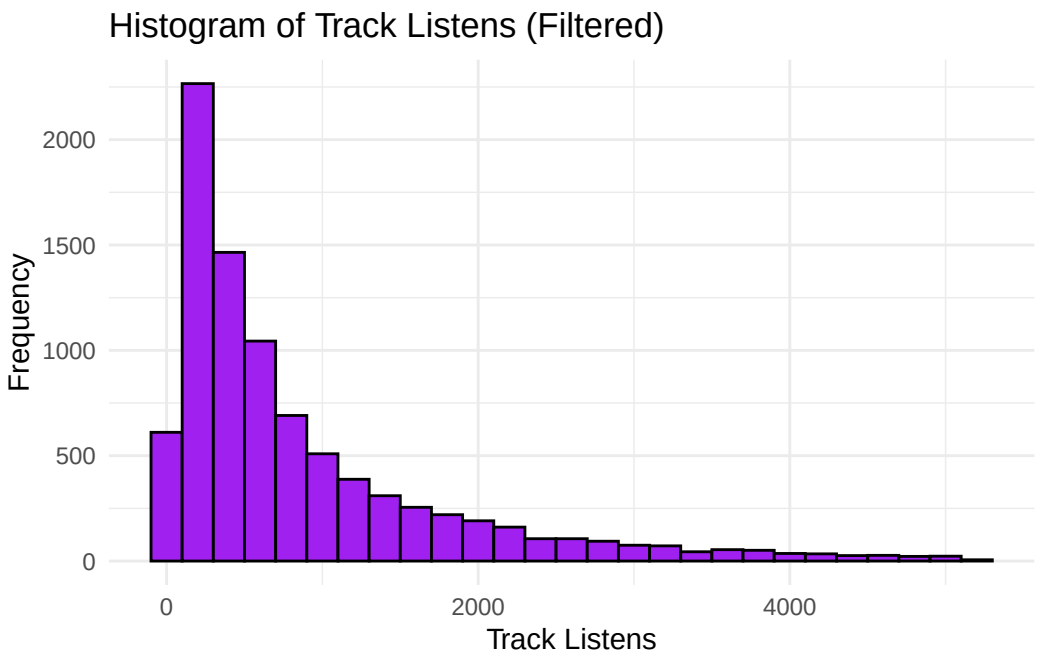

Figure 2: Summary distribution of tracks by the number of times they are listened to. Filtered by the bottom 95% of tracks to visually remove outliers.

```
# Convert popularity to a numeric value from 1 to 10
df_tidy <- df_tidy %>%
  mutate(popularity_numeric = as.numeric(popularity))


# Correlation matrix
numeric_features <- df_tidy %>%
  select_if(is.numeric) %>%
  select(-track_listens, -track_ID, -artist_id)
cor_matrix <- cor(numeric_features)
ggcorrplot(cor_matrix, lab = TRUE, lab_size = 2.5)+
  theme(axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10))
```

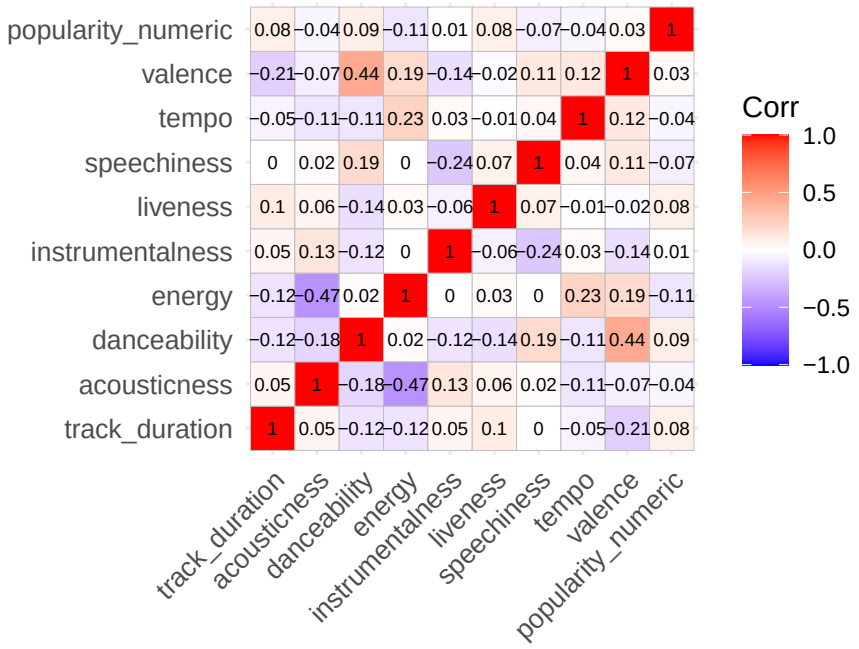

Figure 3: Correlation matrix between all 8 Echonest features, duration of track, and popularity metric.