Executive Summary: Harmonic Features for Song Recommendations
Team: Iliyana Dobreva, Matthew Dykes, Elizabeth Rizor, Joshua Ruiter, and Juan Salinas

Introduction

Music has served as an important medium for human expression for tens of thousands of years. Songs can be broken down into quantifiable harmonic patterns that can signal a cultural moment in time or belonging to a specific subculture. Mapping these harmonic features to markers of cultural relevance are useful for musicians, researchers, and businesses who wish to better recommend songs to individuals. Because of this need, we aimed to explore how harmony shapes the sound of popular music, and determine whether harmonic "fingerprints" can accurately classify songs by genre, decade, and whether a song had ever reached the Billboard Hot 100.

Stakeholders and KPIs

Our stakeholders consist of the following:

- **Researchers and theorists** who care about interpretable insights into harmonic change
- **Musicians** who want to see how harmonic conventions differ by style and era
- **Streaming platforms** who are interested in practical classification and prediction

Primary KPI:

- Classification accuracy (percentage of correctly classified songs)

Secondary KPI:

- F1 score (balances precision/recall)

Datasets

Our primary dataset was the *Chordonomicon* dataset, which includes over 600,000 chord-annotated songs, each labeled with release decade, genre, and Spotify ID. Using the Spotify API, we retrieved track and artist names associated with each Spotify ID. We then cross-referenced these songs with the *Billboard Hot 100* dataset (catalogs every track that has appeared on the U.S. charts since the list's inception) to determine whether each song had ever been featured on the chart.

Methods

EDA: EDA revealed decade and genre class imbalance. We excluded data from before the 1950s due to extreme class imbalance and focused on later decades. While some genre imbalance remained, all genre categories were retained since their representation was sufficient. Additionally, songs with fewer than six total chords (about 0.2% of the data) were removed to ensure compatibility with analyses. Class imbalance was also present for Billboard Hot 100 presence, as only 3% of the songs had ever been on the Hot 100. This was accounted for in model training by oversampling the minority class.

Feature Engineering and Selection: We engineered a large variety of harmonic features. All features were ultimately derived from the original "chords" feature in the *Chordonomicon* dataset (list of chords in a song are recorded as a string). These strings were converted to length 12 binary vectors (representing the 12 notes from the 12TET scale as defined by western music theory). Our final list of model features fits broadly into two categories: n-grams and "holistic" features. N-grams are chord sequences of length 3, 4, and 5; these binary features indicate whether a song contains this chord sequence (or its harmonic equivalents) or not. Holistic features describe chord patterns throughout a song. A holistic feature example is "Unique Chord Density", which describes the ratio of unique chords to total chords in a song. For each target, we picked our final list of model features based on avoiding multicollinearity and ranking feature importance rankings with Lasso L1 Regularization.

Results

Multiple models were trained and compared to baseline models in order to select the most appropriate for each target. Models were then tested on a held-out test set that did not overlap with the training set.

Hot 100: The best performing model (Random Forest) performed well with our primary KPI (~80% accuracy) when compared to the baseline model (50%). However, this model did not perform well with our secondary KPI (F1 score < 10%). These results show that the model has promise to predict whether a song was on the Hot 100, but may need further tuning to improve model precision.

Decade: The best-performing model (XGBoost) achieved 43.5% accuracy, which only slightly outperformed the baseline dummy classifier (42.7%). These results indicate that our current features alone are insufficient for accurately predicting release decade.

Genre: Similarly to decade, our features showed limited predictive power. N-gram only models (27% accuracy) and the best performing random forest model (28%) only barely outperformed the baseline model (25%). Overall, chord sequences alone did not provide enough information to reliably predict genre.

Future Directions

Because our models could be improved, we have a few suggestions for future directions. Early EDA showed small trends for decade vs. various features that indicate song complexity (such as "Unique Chord Density"), but we didn't explore this thoroughly. Future models could consider this trend. Future modeling could also consider using the modulation-invariant Levenshtein distance between chord sequences (possibly most common n-gram) to cluster harmonically similar songs (KNN similarity analysis).