Harmonic Features for Song Recommendations

Erdős Institute Data Science Boot Camp

Iliyana Dobreva, Matthew Dykes, Elizabeth Rizor, Joshua Ruiter, and Juan Salinas

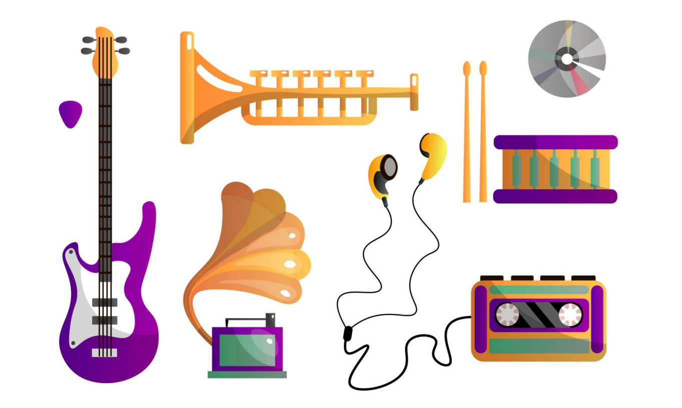
11/05/25

Introduction

- Songs can be broken down into quantifiable harmonic patterns

- Mapping these harmonic features to markers of cultural relevance are useful for those (such as streaming platforms) who wish to better recommend songs to individuals

- We aimed to determine whether harmonic "fingerprints" can accurately classify songs by **genre, decade, and popularity**



Dataset Curation and Cleaning

Chordonomicon Dataset: 680,000 chord-annotated songs sourced from users on UltimateGuitar.com with other identifying information including release date, genre and artist name.

Data Cleaning Process:

- Used artist variable to merge song popularity measures from Billboard
- Eliminated entries missing target values and restricted dataset to songs released after 1950 and having greater than 5 chords in the progression
- Reduced from 679,807→300,713 entries
- 85-15 test-train split

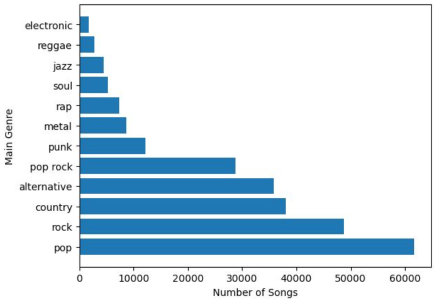
Summary of training data before featurization

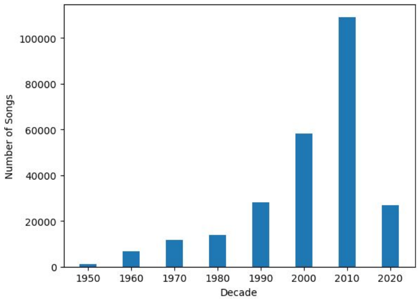
Raw features: Chords (string), 253,666 chord progressions

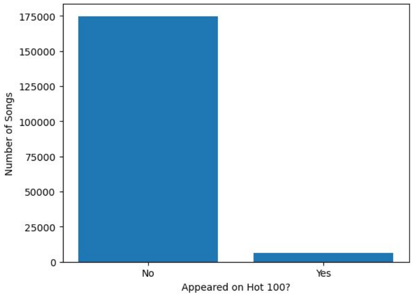
- Eg. ‘Gmin,Ab,Cmin,Bb,Emin,F,Amin,G’

Target variables:

- Decade (categorical), 8 decades in the range of 1950’s-2020’s
- Genre (categorical), 12 genres classified as “main” genres
- Hot 100 popularity (binary), indicates whether of not a song has ever appeared on the Billboard Hot 100

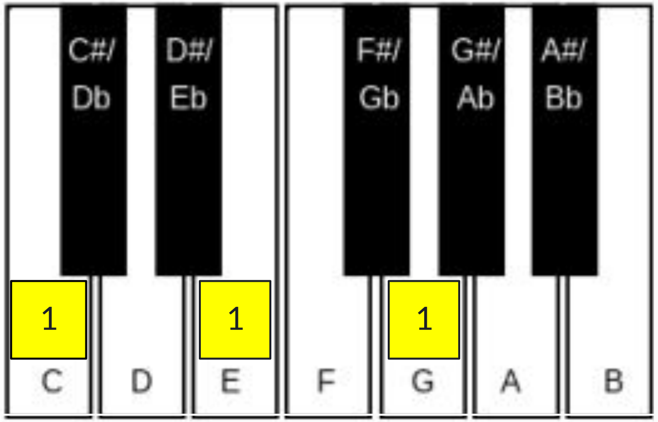






Turning chords into features

- In western music theory, a chord is a subset of the notes from the 12TET scale
- Chords are stored as length 12 binary vectors, indexed with the first entry corresponding to C
- The data set contained a dictionary for string->vector conversion
- From chord sequences, we extracted various features



C major: [1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0]

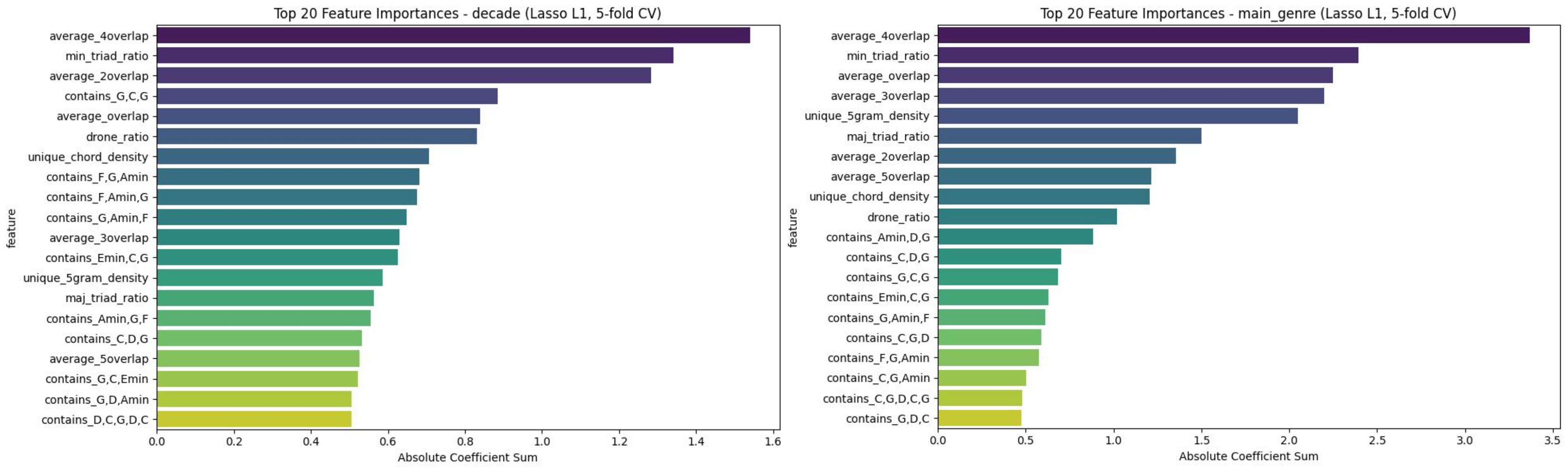
D minor: [0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0]

Turning chords into features

- N-gram features
  - **contains_x** - around 100 binary features of the form "**contains_x**", where x is a chord sequence of length 3, 4, or 5
- Holistic features
  - **missing_notes** - counting notes not used by any chord
  - **drone_ratio** - measure of how close a song is to having a single note played throughout
  - **average_overlap** - Note overlap between sequential chords (e.g. overlap for "C,Amin" is 2, as C and A minor share the notes E and C)
  - **average_overlap[*n*]** - similarities between a chord and the chord *n* after it, averaged over all possible such pairs in a song, for n∈{2,3,4,5}
  - **maj_triad_ratio** - fraction of chords in a song that are major triads
  - **min_triad_ratio** - fraction of chords in a song that are minor triads
  - **unique_chord_density** - distinct chords divided by total number of chords
  - **unique_5_gram_density** - distinct 5-grams divided by the total number of chords

Feature Selection

- **Lasso (L1) Regularization (5 Folds):** ranks features by shrinking less important coefficients to zero
- **Variance Threshold:** removes features with little variation
- **Multicollinearity Check:** identifies and drops highly correlated features

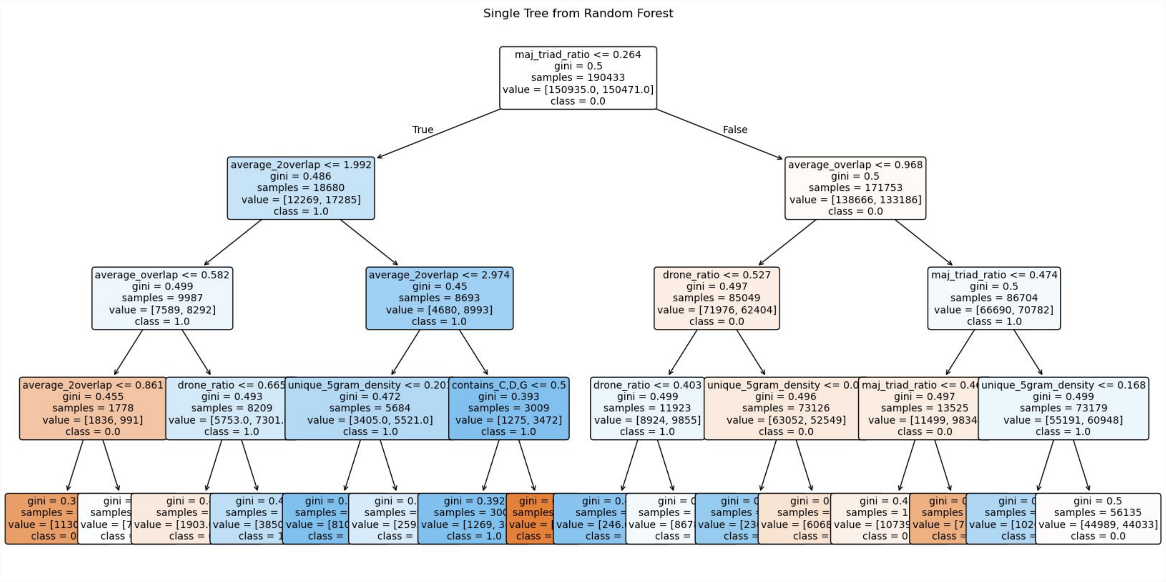


Top 20 features for predicting **Decade** and **Genre** using Lasso (L1) Regularization

Modeling

Model Options for categorical targets:

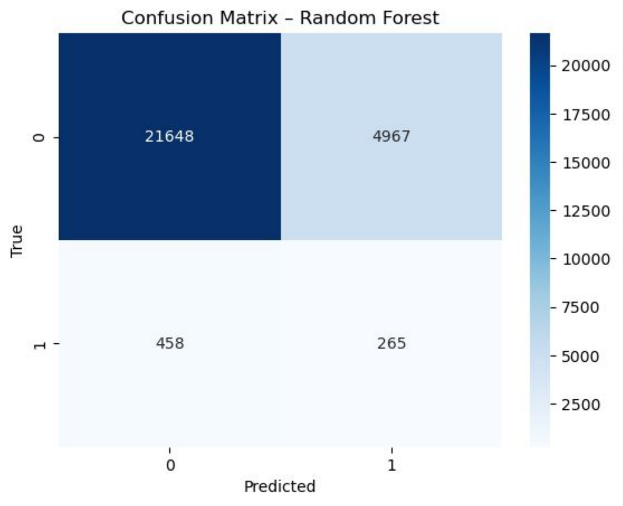
- **Base model:** sklearn Dummy Classifier (most frequent)
- **Logistic Regression**
- **Lasso** (Logistic Regression with L1 penalty)
- **Random Forest** Classifier
- **XGBoost** Classifier



Modeling - Hot 100

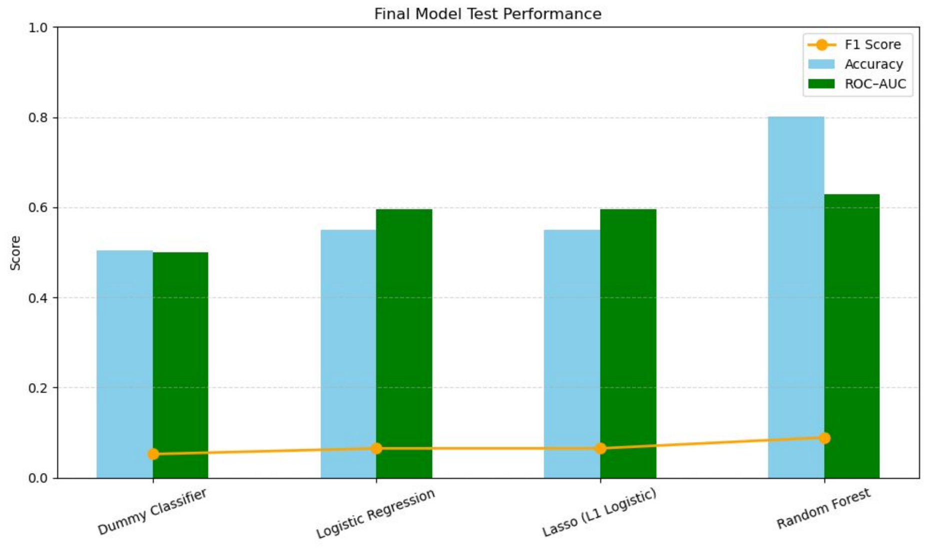
Model comparison (5-fold cross validation) with logistic, lasso, and random forest

- Training data oversampled to address class imbalance
- Random forest was best (**accuracy = 0.72, F1 = 0.69**)
- Hyperparameter grid search found optimal param combination, max_depth limited to 10 to avoid overfitting
- Features limited to top 7 for simplicity (5 holistic, 2 n-gram)
- Decades limited to 1990s-2020s



Final performance:

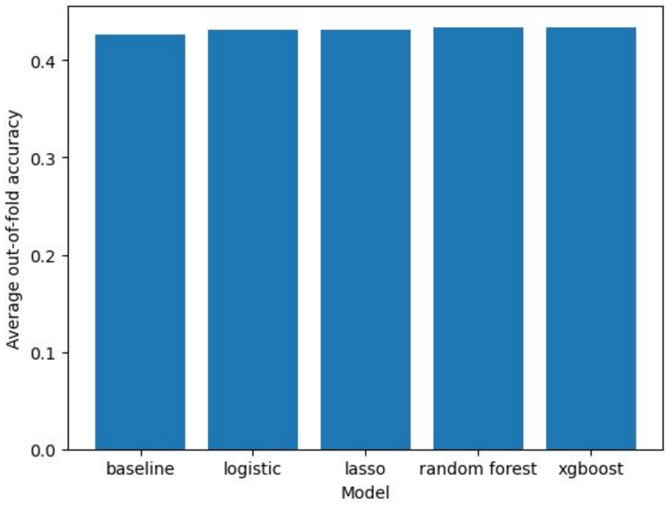
- Random forest still best (**Uniform dummy = 50%, RF = 80%**)
- High accuracy, poor F1 score due to low precision (many false positives)



Modeling - Decade

- Tested random forest, logistic regression and XGBoost with 5-fold stratified cross validation with hyperparameter tuning
- Tested on 20 most-important features determined by lasso
- Tested against baseline which predicts the most frequent decade
- All models performed negligibly better than the baseline

Best-performing model:
XGBoost with 50 estimators and max depth of 4
Average out-of-fold accuracy of baseline model: 0.4269
Average-out-of-fold accuracy of best-performing model: 0.4341
Test accuracy of best-performing model: 0.4354



Modeling - Genre

Our best model for genre: random forest utilizing 8 holistic features below
maj_triad_ratio, min_triad_ratio, unique_chord_density, drone_ratio, and average_Noverlap for N=2,3,4,5

This model still only performs marginally better than a “most frequent” dummy classifier.

Dummy	24.1%	accuracy
Random forest	28.46%	accuracy

Also considered: logistic + lasso regressions

Also considered: n-gram features
We were very hopeful that n-gram features would be useful to predicting genre, but they did not end up being usefully predictive.