Creating the Perfect Playlist: Content-Based Generation of Spotify Playlists

Sonu Mehta, Laura Ware, Omar Abboud

GOAL

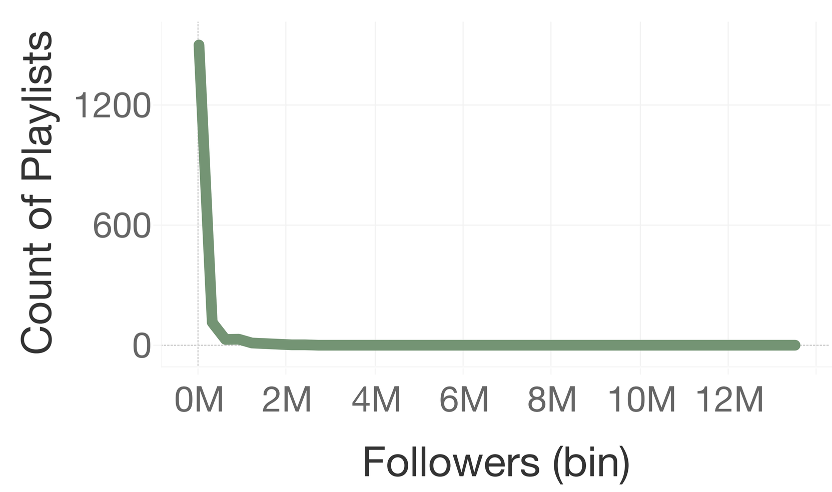
Explore methods for predicting the success of a Spotify-curated playlist, based only on data about the songs that comprise that playlist, and use these models to develop novel processes for curating successful Spotify playlists.

Features:

- Predict the success (# of followers) of a given playlist
- Identify songs that are acoustically similar to a seed song
- Generate a playlist that optimizes the grouping and sequence of tracks

DATASETS

Only a small number of playlists have over 100k followers



Spotify API

Includes data about each Spotify-curated playlist (e.g. total tracks, sequence, no. of followers) and individual tracks (e.g. audio features, popularity)

Raw Audio

30-second samples available from the Spotify API for ~25% of the 20k tracks in our dataset.

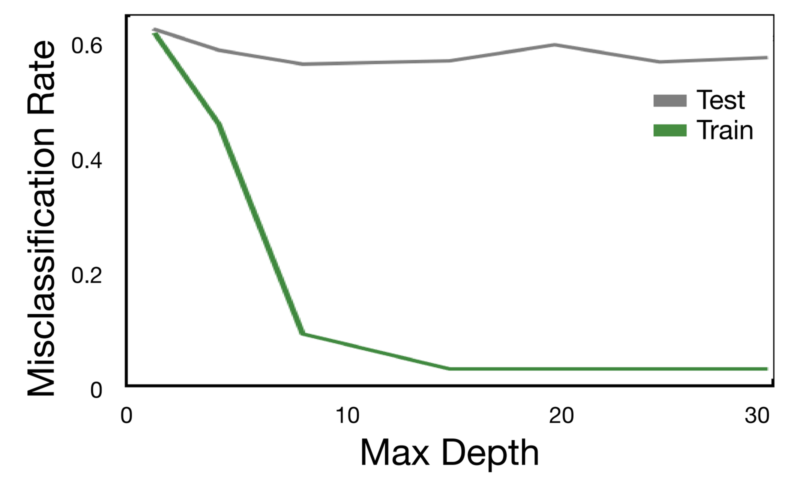
MODELS

Random Forest to Predict Number of Playlist Followers

Inputs: Acoustic features (danceability, loudness, energy, liveness, etc.), duration, popularity of songs, track order

Output: No. of playlist followers, divided into 5 bins

Tuning Random Forest Max Depth

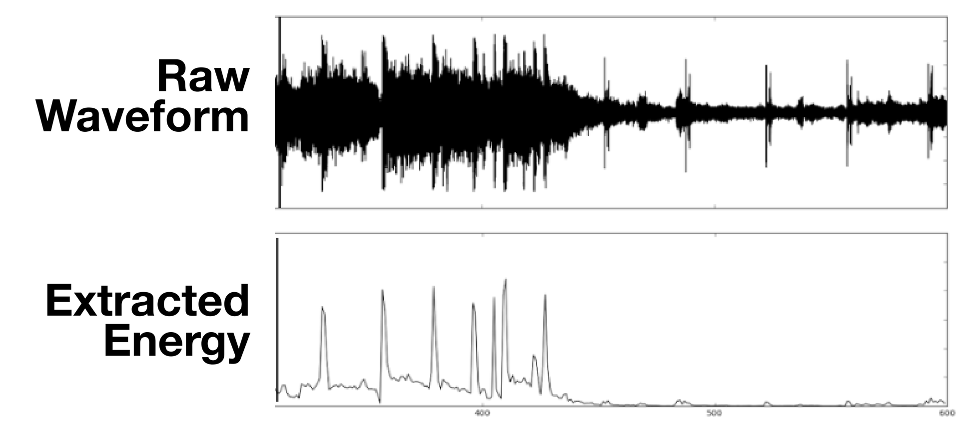


Using Raw Audio to Predict Track Popularity

Inputs: MFCCs, Chroma coefficients,Energy at each timestamp

Output: Popularity divided into 6 bins

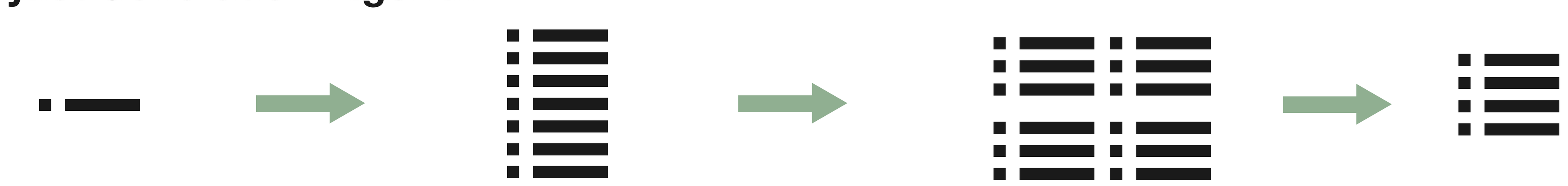
Performance: Random forest classifier gave 75% accuracy, F-1 score of 1 for the top class



Methods for Determining Acoustic Similarity

Manhattan distance with features from Spotify API **+** *Earth Mover's Distance* based on KL Divergence and raw audio features, with vantage points to optimize the search process.

Playlist Generation Algorithm



1. Input a seed track
2. Identify candidate tracks from either: A) top tracks from related artists, B) similar tracks from related artists or C) acoustically similar songs from a vantage point database
3. Perform simulated annealing, starting with a random sample of 30 tracks
4. Return the playlist that optimizes predicted number of followers.

RESULTS

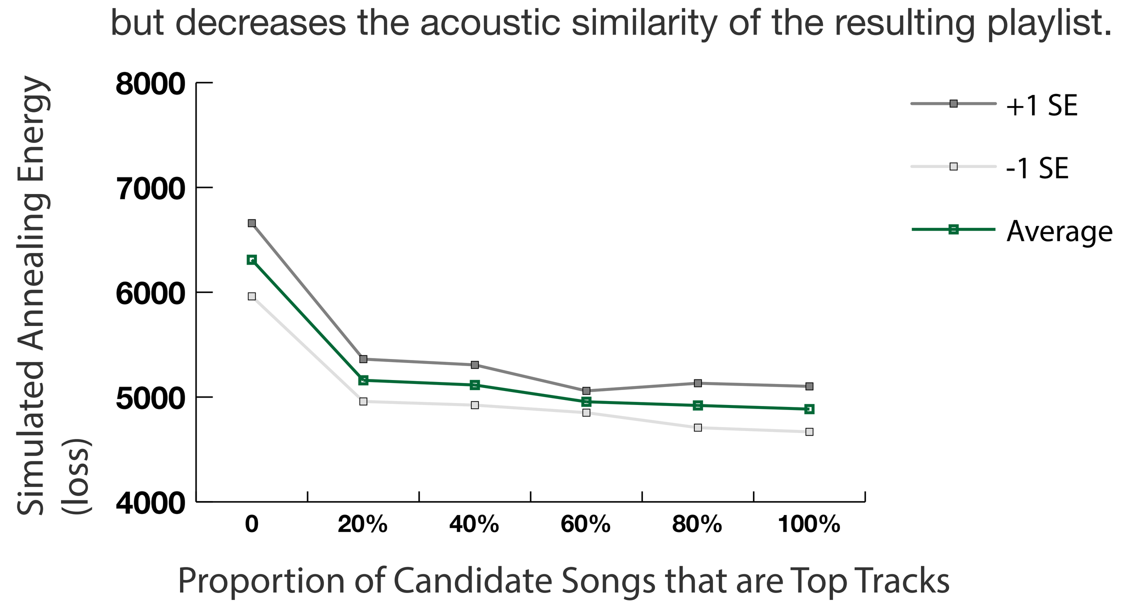
Classification Model Comparison

The mean popularity of a playlist's tracks alone is not sufficient to predict overall playlist followers; this metric is much stronger when combined with acoustic features.

Predictor Set	Classification Accuracy
Mean Popularity Only	0.35
Spotify Acoustic Features Only	0.66
Mean Popularity + Acoustic Features	0.78

Weighting of Top Tracks vs. Performance:

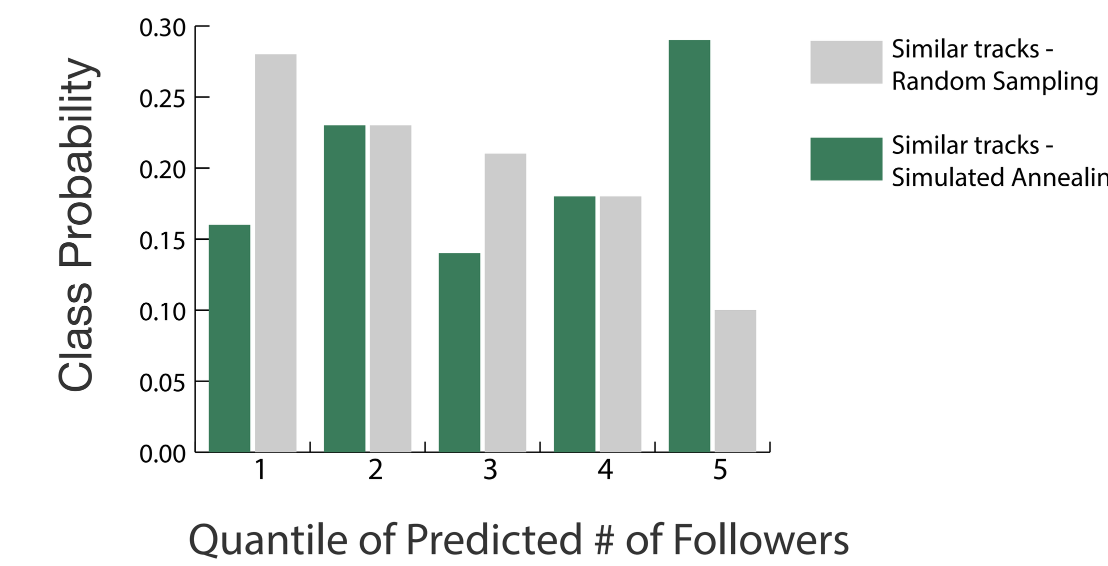
Using a higher percentage of top tracks improves performance but decreases the acoustic similarity of the resulting playlist.



CONCLUSIONS

- Combinations of songs can be reasonably optimized for popularity using simulated annealing supported by raw audio-based similarity metrics and a Random Forest predictive model of popularity.

Simulated Annealing increases optimization performance significantly



Future Work

- Take steps to mitigate the overfitting that affected our random forest classifier, which over-predicted unpopular playlists in the testing set.
- Include more features as predictors of popularity to ensure that the model is not capturing information such as release date or other non-intrinsic characteristics.

REFERENCES

Berenzweig, Adam, Beth Logan, Daniel P.W. Ellis, & Brian Whitman. A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures. Proceedings of the ISMIR International Conference on Music Information Retrieval (Baltimore, MD), 2003, pp. 99–105.

Logan, B., "A Content-Based Music Similarity Function," (Report CRL 2001/02) Compaq Computer Corporation Cambridge Research Laboratory, Technical Report Series (Jun. 2001).