Predicting Song Popularity using Spotify's Music Dataset

Denisha Saviela

OVERVIEW

The Spotify music dataset contains numeric metrics generated by Spotify which measure the songs' danceability, mood, liveness, etc. The goal of this project was to predict the hit songs in future years from multiple artists based on the aiven dataset.

PRE-PROCESSING

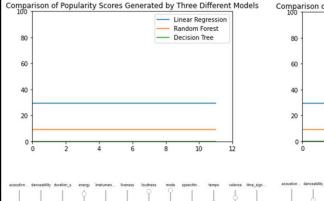
REDUCING NUMBER OF SONGS

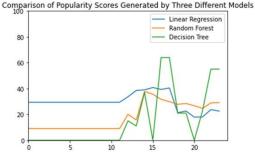
The database contains more than 232,000 songs. Following the data cleaning process, the number of songs was reduced to 176,000.

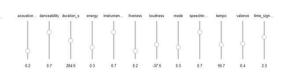
REMOVING FEATURES

Dropping duplicates of track_id and converted time_signature from type object to integer.

FEATURE MIXTURE



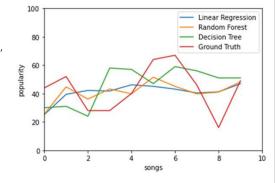




A feature mixer can be used to identify which features are most important in predicting the target variable. The features can be manipulated which will change the predication popularity score. The graph on the left is a representation of the current features below it. Based on the given input, the linear regression gave a ranking of 35/100 for the first song, the random forest gave a ranking of 15/100, and the decision tree gave a ranking of 0/100. Changing the feature extractor will give a different graph and result of the song. The graph on the right shows the new popularity score based on the changed features below it.

RESULTS AND **EVALUATION**

The red line represents the songs' popularity score, while the other three lines represent the models' prediction. Higher distance from the red line represents a lower accuracy. All three models didn't perfectly fit the red line, but the random forest provided the closest predictions to the ground



MODEL SELECTION

LINEAR REGRESSION testing score: 0.199

RANDOM FOREST

DECISION TRFF

training score: 0.202

training score: 0.906

testing score: 0.341

training score: 0.998

testing score: -0.350