Action Plan for the Song Popularity Prediction Project

Project Goal Definition

Build a regression model using Random Forest and Gradient Boosting to predict song popularity from audio features.

Dataset Acquisition

Use the "Song Popularity Dataset" from Kaggle.

Environment Setup

Python with Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. Alternatively, R with relevant packages.

Exploratory Data Analysis (EDA)

Examine feature distributions, check for missing values, and explore relationships between features and popularity scores.

Data Preprocessing

Handle missing values, encode categorical variables if any, and scale numerical features.

Feature Engineering

Create new features that could improve model performance, such as interaction terms between tempo and energy.

Model Selection

Train Random Forest and Gradient Boosting regressors, and compare their performance with baseline models.

Model Training and Evaluation

Split data into training and testing sets, use regression metrics like RMSE, R-squared, and MAE for evaluation, and perform cross-validation.

Hyperparameter Tuning

Use GridSearchCV or RandomizedSearchCV to find optimal model parameters.

Feature Importance Analysis

Identify the most influential audio features in predicting popularity.

Timeline

Week 1: Data acquisition and environment setup.
Week 2: EDA and preprocessing.
Week 3: Model building and tuning.
Week 4: Evaluation and reporting.