Predicting Track Popularity with ML

Derya Er
07.11.2025

Project Overview

The goal was to predict a song’s popularity score using regression models. The insights could help artists and platforms understand what drives success on Spotify.

Objective: Predict track popularity from Spotify data
ML Type: Supervised - Regression
Dataset: 230K+ tracks with 18+ features
Impact: Help artists, marketers, and music platforms

Data Selection & Preparation

DATASET

Source: Kaggle
Ultimate Spotify Tracks DB

EDA

Checked missing values, duplicates.

EDA

Converted duration, explored categorical columns (genre, key, mode, time signature)

EDA

Correlation & feature relationships

Feature Engineering & Selection

On this part, encoded categorical columns, I tested different scalers. RobustScaler worked best, mainly because it handles outliers well. Then I used feature importance to identify the strongest predictors.

Encoding → **Scalers** → **Train Test Split**

Encoding

Encoded categorical features:
'genre', 'key', 'time_signature'

Scalers

Tested 3 scalers:

StandardScaler → best
MinMaxScaler
RobustScaler

Train Test Split

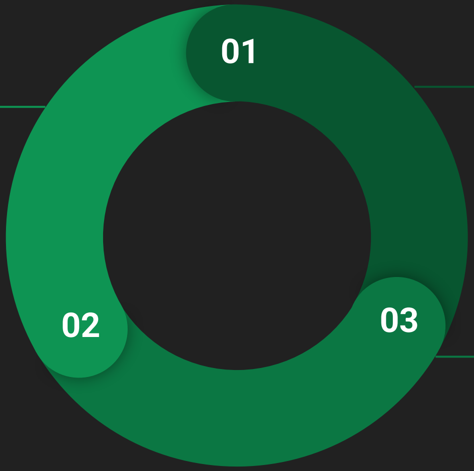
Created datasets for 3 scalers and save datasets to a pickle file (you can save any data with this library) for later use

Model Building

I tried multiple regression algorithms. The tree-based ensemble models generally outperformed the simpler ones, with XGBoost giving the best results even before optimization.

Models trained:

KNN, Linear Regression, Decision Trees, Random Forest, XGBoost, LightGBM,



Split:

80% train / 20% test

Used Metrics:

R^2, MAE, MSE, RMSE

Model Performances

Tested 6 regression algorithms under 3 scaling methods.

Models

- KNN
- Lineer Regression
- Random Forest
- XGBoost
- Decision Tree
- LightGBM

Scaler

- Standart scaler
- Robust scaler
- MinMax Scaler

Best model is Random Forest with the robust scaler. You can see how model performance improves as we move to more complex algorithms. The tuned Random Forest model reached an R^2 of 0.72, explaining about 72% of the variation in popularity.

Hyperparameter Tuning

I optimized the Random Forest model with Randomized Search.

model	dataset	R2 score	MAE	MSE	RMSE
KNN	with robust	0.656364	8.020695	113.722180	10.664060
Linear Regression	with robust	0.625915	8.354649	123.798607	11.126482
Random Forest	with robust	0.709140	7.355473	96.256539	9.811042
Decision Tree	with robust	0.395440	10.321094	200.071655	14.144669
XGBoost	with robust	0.707421	7.385321	96.825451	9.839992
LightGBM	with robust	0.705746	7.406970	97.379763	9.868118
KNN	with minmax	0.662783	7.884683	111.597768	10.563984
Linear Regression	with minmax	0.625915	8.354649	123.798607	11.126482
Random Forest	with minmax	0.709185	7.360767	96.241499	9.810275
Decision Tree	with minmax	0.401861	10.282988	197.946486	14.069346
XGBoost	with minmax	0.707421	7.385321	96.825451	9.839992
LightGBM	with minmax	0.705738	7.405376	97.382473	9.868256
KNN	with standart	0.666331	7.860270	110.423494	10.508258
Linear Regression	with standart	0.625915	8.354649	123.798607	11.126482
Random Forest	with standart	0.708129	7.361734	96.591007	9.828072
Decision Tree	with standart	0.397073	10.321746	199.531265	14.125554
XGBoost	with standart	0.707421	7.385321	96.825451	9.839992
LightGBM	with standart	0.706032	7.399171	97.285181	9.863325
XGBoost with best hyperparameters	with standart	0.714956	7.287075	94.331866	9.712459
RandomForest with best hyperparameters	with robust	0.718050	7.257531	93.307686	9.659590

After hyperparameter tuning, the Random Forest model with robust scaling achieved the best performance with an R^2 score of 0.718 and RMSE of 9.66, improving accuracy by approximately 1.3%. This indicates the model effectively captures complex, non-linear relationships in the Spotify dataset and provides the most reliable predictions overall.

Key Findings

The model assigns the highest importance to genre_Movie, meaning that whether a track belongs to the “Movie” genre has the largest impact on the model’s predictions. Similarly, Pop, Rap, and Hip-Hop also play strong roles

The most important features are respectively:

Group	Description	Share of Total Importance
Top 5 features	`Movie`, `Pop`, `Children’s Music`, `Rap`, `Hip-Hop`	**47.5%**
Top 10 features	Add `Rock`, `Indie`, `Folk`, `R&B`, `Opera`	**68.3%**

As a result, 68% of the model’s predictive power comes from just 10 features, most of which are genre-related.

Challenges & Learnings

I faced a few challenges like

- Model tuning time,
- The order of the workflow,
- Data cleaning issues

I learned how important preprocessing, data cleaning and even small tunings improvements matter.

THANK YOU!

Derya Er