Spotify billboard Hit Predictor

By:
Frederika Cook - Duru Demirbag - Hannah McAuley - Zhen Chen

Introduction

- We are interested in the common features and replicability of hit songs.
- If hit songs follow certain patterns, this could have major commercial value in the music industry.
- Spotify provides rich audio feature data for machine learning.
- We want to explore whether hit songs can be predicted and reproduced using data.

Research Questions

- Can machine learning predict a hit song using audio features?
- Which audio features are most influential in predicting?
- What common features do hit songs share?
- Which classification model gives the best performance?

Data Overview

Sources:
Billboard (GitHub), Spotify features (Kaggle).
Contain 40560 unique songs and 26 variables – only 16 selected for modelling.

Predictor Features:
danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature, chorus_hit, sections.

Dependent Variable:
on_billboard

Limitation:
Genre/time bias in non-hit sampling.
Non-hit label may reflect niche, not failure
Spans 60 years—blurs temporal music trends

Methodology

Data integration:
Combined two datasets using a custom Python script, and created a binary target label on billboard.

Data processing:
Remove duplicates, missing value handling, standardization, or log transformation for numerical variables.

EDA:
Analysed feature distributions, correlations, and class balance to gain initial insights

Comparing techniques:
Classical: Logistic Regression.
Modern: Random Forest, XGBoost, Support Vector Machines (SVMs).

Technique Comparison

Logistic Regression:

Linear classifier that models the relationship between input features and binary target using logistic function. Fast and simple to train but assumes linearity.

Random Forest:

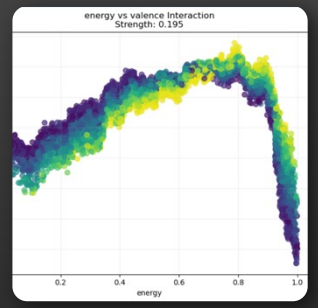
Ensemble method that builds on the weaknesses of individual decision trees by averaging the predictions of many trees trained on bootstrapped samples of the data and random subsets of features. Reduces overfitting and improves generalisation but can become computationally expensive.

XGBoost:

Gradient-boosted ensemble method that builds trees sequentially, each one correcting the errors of its predecessor. Fast and accurate in capturing subtle patterns and complex feature interactions, with its flexibility with regularisation helping to control overfitting, making it suitable for datasets with imbalance or noise. Though, the need for careful tuning increases difficulty to interpret and deploy.

Support Vector Machine (SVM):

Classifies data by finding the optimal hyperplane that maximally separates the classes. Can model complex, non-linear boundaries. However, can struggle with large datasets and can, consequently, become computationally expensive. Additionally, sensitive to feature scaling and less interpretable than other methods.



Best models for each technique

Results

	Logistic Regression	Random Forest	XGBoost	Support Vector Machines
AUC	0.812	**0.866**	**0.862**	0.840
Accuracy	0.739	**0.789**	**0.781**	0.760

Conclusion

01 Reflecting on our research questions, machine learning models can be used to predict a hit song using audio features, alongside other relevant features like duration.

02 In particular, the audio features instrumentalness, acousticness, and danceability were the most influential audio features in predicting a song's hit status.

03 Random Forest and XGBoost models were found to be best at successfully classifying between hit and non-hit songs. Though, both are computationally expensive, and data must be tuned accurately.

Thank You

Slide Chef

Powerpoint theme created by