ie UNIVERSITY SCHOOL OF SCIENCE & TECHNOLOGY

Predicting Song Popularity

A classification model based on spotify data to predict the popularity of a song

Proposed by:

Alejandro Gutierrez Werner, Apilash Balasingham, Fhadah Alromy, Lucia Pellicer Cascales, Luis Soto Guareschi, Ignacio Sahonero Vadillo

Professor:

Alvaro Jose Mendez Lopez

AGENDA

1. Executive Summary
2. Data Audit & Cleaning
   - 2.1 Genre Distribution
3. Model Selection Methodology
   - 3.1 XGBoost Configuration
4. Model Performance
   - 4.1 Evaluation: ROC Curve
   - 4.2 Confusion Matrix Insight
5. Business Applications
6. Limitations
7. Demo

1. Executive Summary

The Objective

The goal is to build a classification model to predict the future success of songs based on audio features and artist data. We analyzed a dataset of **2,200 songs** to create a discrete "popularity" target.

Key Findings

After testing multiple algorithms, **XGBoost** proved to be the most robust solution, achieving an accuracy of **78.6%**. It effectively distinguishes potential hits from the rest.

2. Data Audit & Cleaning

The Dataset

2,200 records containing audio features like danceability, energy, and loudness (19 columns). No duplicate entries were found.

Processing

Dropped minor missing values. Transformed 'Artist Genres' via one-hot encoding for ML compatibility. 22 columns in total after data processing

Target Variable

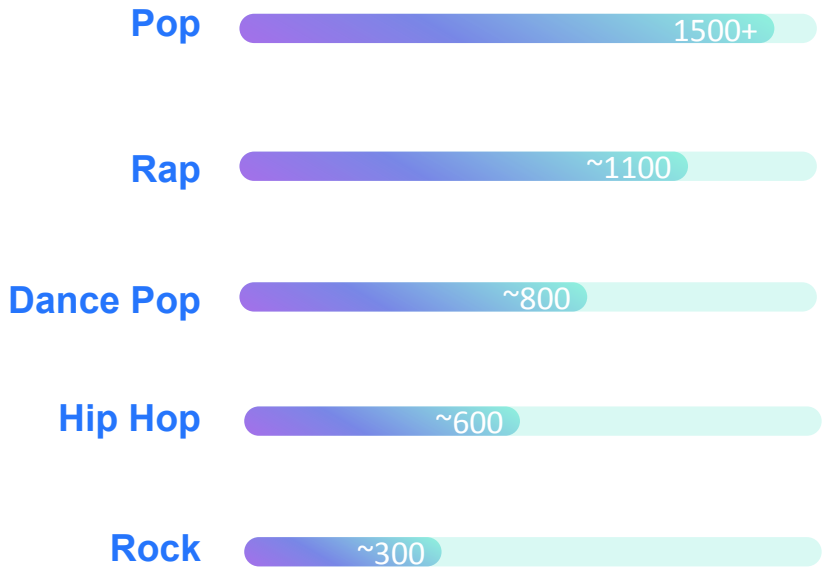
Created a binary 'Popular' variable. Songs in the top 25% quantile labeled **True**, others False.

2.1 Genre Distribution

Dominance of Pop

Pop dominates the dataset, appearing in over 1,500 tracks, followed significantly by Rap and Dance Pop.

This imbalance suggests the model may learn pop-characteristics more effectively than niche genres.



3. Model Selection Methodology

Tested Algorithms

- Decision Trees
- Random Forest
- Naive Bayes
- K-Nearest Neighbors (KNN)
- Boosting Methods

Winner: XGBoost

XGBoost was selected as the final model because:

- It handles complex, non-linear relationships in audio data well.
- Built-in regularization prevents overfitting.
- Consistently delivered the highest performance metrics during testing.

3.1 XGBoost Configuration

Initial Model: 99.9% train accuracy, 75.6% test accuracy → clear overfitting.

After Optimization using **GridSearchCV** to find the optimal hyperparameters: 85.5% train, 78.6% test → +3.9% test improvement.

Result: Better generalization, reduced overfitting, and robust real-world performance while maintaining strong predictive power.

4. Model Performance

78.6%

Test Accuracy

Improved from 75.6% base

0.74

ROC - AUC Score

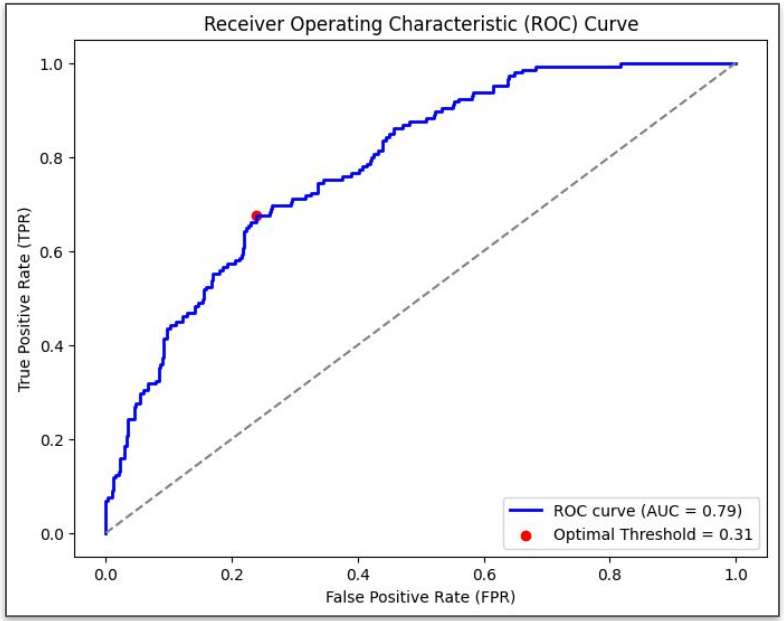
Strong predictive power

0.31

Optimal Threshold

Maximized TPR - FPR

4.1 Evaluation: ROC Curve



Interpretation

The curve rises steeply, indicating good discrimination between popular and non-popular songs.

An AUC of **0.74** confirms the model is significantly better than random guessing (0.5).

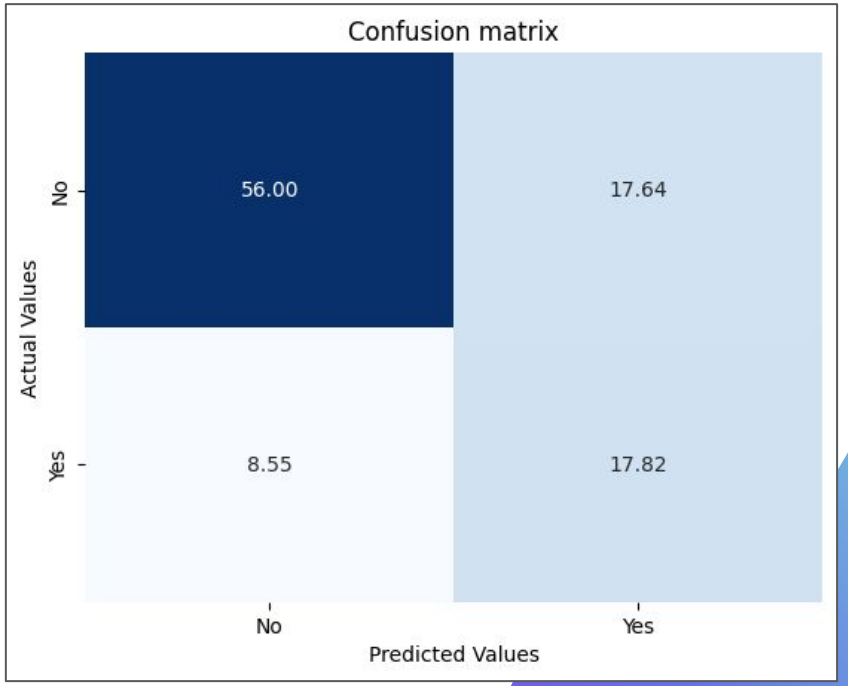
4.2 Confusion Matrix Insight

High Recall, Lower Precision

The model prioritizes sensitivity. It correctly identifies most songs that become popular (True Positives: 17.82%).

The Trade-off

It is tuned to minimize "missed hits" (False Negatives: 8.55%). This results in some non-popular tracks being flagged as potential hits (False Positives: 17.64%). This is acceptable for discovery-focused applications.



5. Business Applications

A&R Screening

Filter incoming demos to prioritize high-potential tracks for human review.

Playlist Opt.

Curate algorithmic playlists to align new releases with listener preferences.

Budgeting

Allocate marketing spend to tracks with the highest predicted probability of success.

6. Limitations

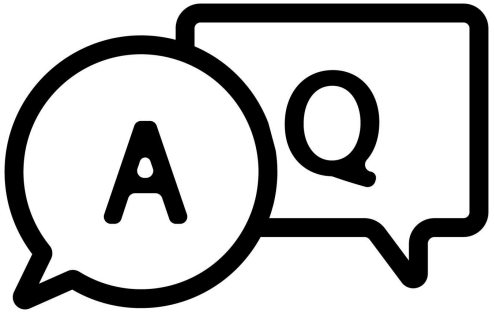
- **Data dependency:** Relies on Spotify-specific metrics (e.g., danceability, energy) — limits cross-platform use.
- **Temporal relevance:** Reflects a fixed time; trends change → needs periodic retraining.
- **Performance trade-off:** High recall, moderate precision → finds most hits but overpredicts popularity.
- **Bias risk:** Genre imbalance and Spotify-derived features may favor mainstream music, and disfavor less popular and new genres.
- **Model Interpretability:** Prediction-oriented model, it does not allows us to capture the importance of the features clearly.
- **Ethical note:** Should support, not replace, human creativity and artistic judgment.

"Predictive analytics can bridge artistic creativity and business intelligence, transforming subjective intuition into quantifiable evidence."

— Group 5

Demo


15:03:27 -- Smoke decrease: 2.85, (total = -0.0017721176147461), smokeeffect = -0.0017551302909851
15:03:27 -- Smoke decrease: 2.85, (total = -0.0006122231434595), smokeeffect = -0.0006027102470397 9
15:03:27 -- Smoke decrease: 2.85, (total = -0.0007087111473835), smokeeffect = -0.0006998777389526 4
15:03:27 -- Smoke decrease: 2.85, (total = -0.0013943195343018), smokeeffect = -0.001380729675293
15:03:27 -- Smoke decrease: 2.85, (total = -0.0009044051170349 1), smokeeffect = -0.0008874177932739 3
15:03:27 -- Smoke decrease: 2.85, (total = -0.0010253548622131), smokeeffect = -0.0010083675384521
15:03:27 -- Smoke decrease: 2.85, (total = -0.0017204761505127), smokeeffect = -0.001700770854 95
15:03:27 -- Smoke decrease: 2.85, (total = -0.0011272788047791), smokeeffect = -0.0011136889457703
15:03:27 -- Smoke decrease: 2.85, (total = -0.0021193385124207), smokeeffect = -0.0021105051040649
15:03:27 -- Smoke decrease: 2.85, (total = -0.0009023666381859), smokeeffect = -0.0008846998214721 7
15:03:27 -- Smoke decrease: 2.85, (total = -0.0008962512016296 4), smokeeffect = -0.0008819818496704 1
15:03:27 -- Smoke decrease: 2.85, (total = -0.0008045196533203 1), smokeeffect = -0.0007922887802124
15:03:27 -- Smoke decrease: 2.85, (total = -0.0005768895149231), smokeeffect = -0.0005666971206665
15:03:27 -- Smoke decrease: 2.85, (total = -0.0017340660095215), smokeeffect = -0.0017191171646118
15:03:27 -- Smoke decrease: 2.85, (total = -0.0005476713180542), smokeeffect = -0.0005401968955993 7
15:03:27 -- Smoke decrease: 2.85, (total = -0.0006033897399902 3), smokeeffect = -0.0005925178527832
15:03:27 -- Smoke decrease: 2.85, (total = -0.0011150479316711), smokeeffect = -0.0011014580726624
15:03:27 -- Smoke decrease: 2.85, (total = -0.0007392883300781 3), smokeeffect = -0.0007284164428711 9
15:03:27 -- Smoke decrease: 2.85, (total = -0.0009594440460205 1), smokeeffect = -0.0009533286094665 5
15:03:27 -- Smoke decrease: 2.85, (total = -0.0007827758789625), smokeeffect = -0.0007691860198974 6
15:03:27 -- Smoke decrease: 2.85, (total = -0.0010919451713562), smokeeffect = -0.0010769963264465
15:03:27 -- Smoke decrease: 2.85, (total = -0.0005700945854187), smokeeffect = -0.0005619406700134 3
15:03:27 -- Smoke decrease: 2.85, (total = -0.0006285309791564 9), smokeeffect = -0.0006190180778503 4
15:03:27 -- Smoke decrease: 2.85, (total = -0.0032309889793396), smokeeffect = -0.003203880929 1322
15:03:27 -- Smoke decrease: 2.85, (total = -0.0009478926658630 4), smokeeffect = -0.0009349822998046 9
15:03:27 -- Smoke decrease: 2.85, (total = -0.0009343028068542 5), smokeeffect = -0.0009159564971923 8
15:03:27 -- Smoke decrease: 2.85, (total = -0.0006976318359375), smokeeffect = -0.0006890058517456 1
15:03:27 -- Smoke decrease: 2.85, (total = -0.0006020307540893 6), smokeeffect = -0.0005938768386840 8
15:03:27 -- Smoke decrease: 2.85, (total = -0.0017320775985718), smokeeffect = -0.0017191171646118
15:03:27 -- Smoke decrease: 2.85, (total = -0.0007121086120605 5), smokeeffect = -0.0007012367248535 2
15:03:27 -- Smoke decrease: 2.85, (total = -0.0009084820747375 5), smokeeffect = -0.0008976101875305 2
15:03:27 -- Smoke decrease: 2.85, (total = -0.0006101846694946 3), smokeeffect = -0.0006027102470397 9
15:03:28 -- Smoke decrease: 2.85, (total = -0.0008575201034545 9), smokeeffect = -0.0008466482162475 6
15:03:28 -- Smoke decrease: 2.85, (total = -0.0013970375061035), smokeeffect = -0.0013895630836487
15:03:28 -- Smoke decrease: 2.85, (total = -0.0005979537963867 2), smokeeffect = -0.0005864024162292 5
15:03:28 -- Smoke decrease: 2.85, (total = -0.0005694150924682 6), smokeeffect = -0.0005612611770629 9
15:03:28 -- Smoke decrease: 2.85, (total = -0.0005782485008239 7), smokeeffect = -0.0005700945854187
15:03:28 -- Smoke decrease: 2.85, (total = -0.0026153635787963 4), smokeeffect = -0.0025956630706787
15:03:28 -- Smoke decrease: 2.85, (total = -0.0016498088836670), smokeeffect = -0.0016382575035095
15:03:28 -- Smoke decrease: 2.85, (total = -0.0015254616737366), smokeeffect = -0.0015016794204712
15:03:28 -- Smoke decrease: 2.85, (total = -0.0009816713333130 1), smokeeffect = -0.0009689569473266 6
15:03:28 -- Smoke decrease: 2.85, (total = -0.0006400823593139 6), smokeeffect = -0.0006312489509582 5
15:03:28 -- Smoke decrease: 2.85, (total = -0.0023544430732727), smokeeffect = -0.0023333787918091
15:03:28 -- Smoke decrease: 2.85, (total = -0.0015241026878357), smokeeffect = -0.0015125513076782
15:03:28 -- Smoke decrease: 2.85, (total = -0.0012509465217590), smokeeffect = -0.0012346386909485



Thank You