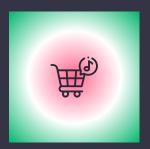


Problem

- Difficulty in forecasting the success of a song.
- The music industry is increasingly data-driven. It's critical to understand the impact of musical characteristics, artist popularity, exposure, and industry recognition on the success of songs.



Why it matters?

- Strategic decision-making for record labels and artists
- Revenue forecasting
- Reducing risk in a hit-driven industry









Data Structure

Spotify Top Songs and Audio Features (Kaggle)

Outcome Variable: Streams (Integer)

Key Predictors:

- **Audio Features**
 - Tempo, Danceability, Energy, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Loudness, Duration
- Categorical Song Features
 - Key, Mode, Time Signatures
- **Artist Popularity Features**
 - Artist Popularity, Number of Collaborators









OData Structure

Spotify Top Songs and Audio Features (Kaggle)

Outcome Variable: Streams (Integer)

Dataset:

- Rows: 6513 unique songs
- Columns: 25 variables (full dataset)
- Selected Features for Modeling: 20 variables (3 categorical, 17 numerical)
- Variable types: integer, string, decimal
- Training/Testing: 80/20









Feature Engineering

Added Popularity Features:



Number of collaborators

Artist Followers



Artist popularity

Total Grammys



Total Nominations

Tiktok Viral





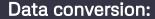
Data Preprocessing











Converted TikTok_Viral Viral \rightarrow 1, Not Viral \rightarrow 0



Removed identifiers:

streams, id, artist_names, track_name, source



Dummy variables:

key, mode, time_signature

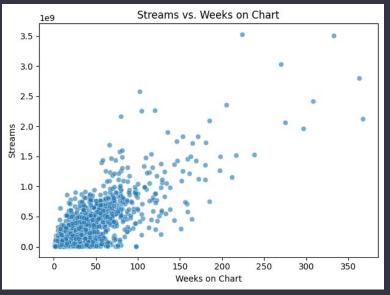


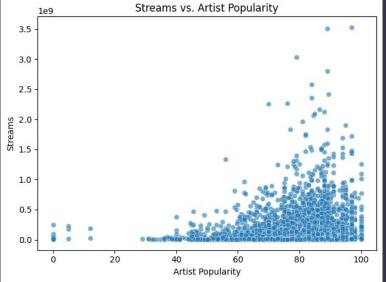






Data Exploration



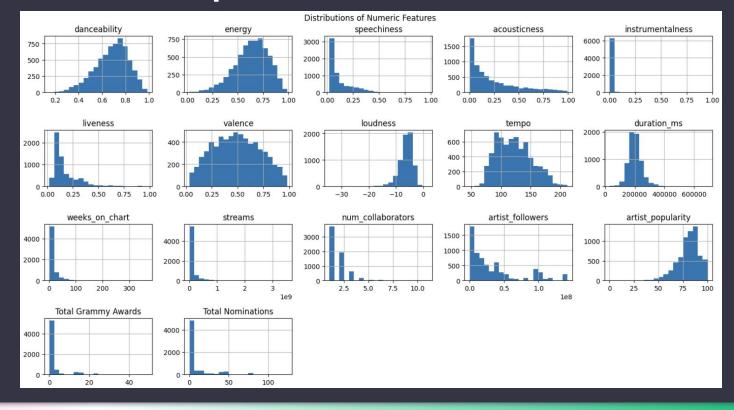








Data Exploration







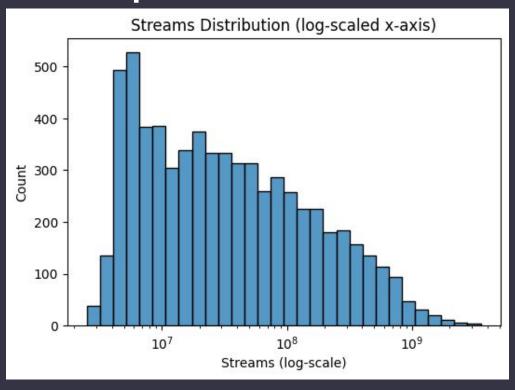


Data Exploration



















								Corre	lation N	1atrix									- 1.0
danceability -	1.00	0.13	0.18		-0.08	-0.08	0.35	0.19	-0.06	-0.16	0.02	0.01	0.15	-0.06	-0.08	-0.10	-0.07		1.0
energy -	0.13	1.00	-0.01	-0.52	-0.07	0.14	0.38	0.73	0.10	0.01	0.01	0.00	0.16	-0.11	-0.11	-0.09	-0.11		
speechiness -	0.18	-0.01	1.00	-0.08	-0.05	0.06	0.04	-0.06	0.12	-0.04	-0.09	-0.09	0.12	-0.00	0.02	0.05	0.08		- 0.8
acousticness -		-0.52	-0.08	1.00	0.06	-0.05	-0.08	-0.43	-0.07	-0.03	0.00	0.01	-0.12	0.04	-0.01	0.01	-0.01		
instrumentalness -	-0.08	-0.07	-0.05	0.06	1.00	-0.01	-0.11		-0.01	-0.03	-0.02	-0.02	-0.06	-0.00	0.00	-0.01	-0.02		- 0.6
liveness -	-0.08	0.14	0.06	-0.05	-0.01	1.00	0.04	0.06	-0.00	0.00	-0.05	-0.04	0.03	-0.01	0.00	0.06	0.06		
valence -	0.35	0.38	0.04	-0.08	-0.11	0.04	1.00	0.30	0.05	-0.15	0.04	0.04	0.12	-0.13		-0.11	-0.14		- 0.4
loudness -	0.19	0.73	-0.06	-0.43		0.06	0.30	1.00	0.07	0.02	0.05	0.04	0.17	-0.09	-0.05	-0.10	-0.10		
tempo -	-0.06	0.10	0.12	-0.07	-0.01	-0.00	0.05	0.07	1.00	-0.02	-0.02	-0.01	0.01	-0.03	0.01	-0.03	-0.04		
duration_ms -	-0.16	0.01	-0.04	-0.03	-0.03	0.00	-0.15	0.02	-0.02	1.00	0.03	0.01	0.17	0.20	0.17	0.21	0.20		- 0.2
weeks_on_chart -	0.02	0.01	-0.09	0.00	-0.02	-0.05	0.04	0.05	-0.02	0.03	1.00	0.86	-0.02	0.06	0.06	-0.04	-0.04		
streams -	0.01	0.00	-0.09	0.01	-0.02	-0.04	0.04	0.04	-0.01	0.01	0.86	1.00	-0.02	0.08	0.08	-0.02	-0.01		- 0.0
num_collaborators -	0.15	0.16	0.12	-0.12	-0.06	0.03	0.12	0.17	0.01	0.17	-0.02	-0.02	1.00	-0.01		-0.02	-0.00		
artist_followers -	-0.06	-0.11	-0.00	0.04	-0.00	-0.01	-0.13	-0.09	-0.03	0.20	0.06	0.08	-0.01	1.00	0.61	0.45	0.58		-0.2
artist_popularity -	-0.08	-0.11	0.02	-0.01	0.00	0.00		-0.05	0.01	0.17	0.06	0.08		0.61	1.00	0.32	0.40		
Total Grammy Awards -	-0.10	-0.09	0.05	0.01	-0.01	0.06	-0.11	-0.10	-0.03	0.21	-0.04	-0.02	-0.02	0.45	0.32	1.00	0.88		-0.4
Total Nominations -	-0.07	-0.11	0.08	-0.01	-0.02	0.06	-0.14	-0.10	-0.04	0.20	-0.04	-0.01	-0.00	0.58	0.40	0.88	1.00		
,	danceability -	energy -	speechiness -	acousticness -	instrumentalness -	liveness -	valence -	loudness -	- tempo	duration_ms -	weeks_on_chart -	streams -	num_collaborators -	artist_followers -	artist_popularity -	Total Grammy Awards -	Total Nominations -		









Overall Approach

1. Initial Model Exploration

 Ran 10 models to find promising models for predicting Spotify streams.

2. Structured Model Evaluation with Nested Cross-Validation

 Used nested cross-validation to tune hyperparameters and select the best-performing model based on generalization performance.

3. Hyperparameter Optimization and Final Model Assessment

Optimized learning rate, number of estimators, and tree depth;
 evaluated final model performance using MSE and R² scores.









Models

- Baseline Model
- Linear Regression with CV
- Ridge Regression
- Lasso Regression
- Elastic Net
- K-Nearest Neighbor
- Decision Tree with pruning
- Random Forest
- AdaBoosting
- **Gradient Boosting**

Baseline Model

MSE: 2.2136

R²: -9.7698e-10











Ridge

- alpha = 100
- Test MSE = 1.0758
- $R^2 = 0.5140$

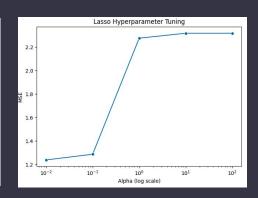
Alpha vs. MSE (Traning Set) 1.2428 1.2426 1.2424 1.2422 1.2420 10² 10^{-2} 10¹ Alpha (log-scaled)

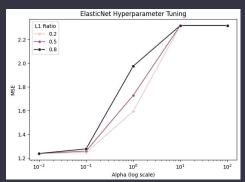
Lasso

- alpha = 0.01
- Test MSE = 1.0707
- $R^2 = 0.5163$

Elastic Net

- alpha = 0.01
- I1 ratio = 0.8
- Test MSE = 1.0714
- $R^2 = 0.5160$













ADA Boosting

Parameter Grid:

of Estimators

50, 100, 150

Learning Rate

0.01, 0.1, 0.5

Max Depth

2, 3, 5

Test MSE

0.3664

Test R²









Gradient Boosting

Parameter Grid:

of Estimators

100, 200

Learning Rate

0.01, 0.1

Max Depth

• 3,5

Test MSE

0.3654

Test R²

0.8349

Nested CV Test MSE

0.4292

Nested CV Test R²









Pruned Decision Tree

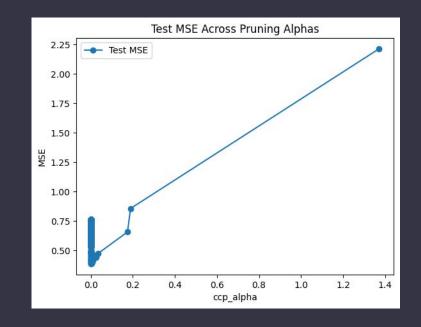
Test MSE

0.3860

Test R²

0.8256

Alpha











Random Forest

Parameter Grid:

of Estimators

100, 200, 300

Max Depth

None, 10, 20, 30

Test MSE

0.3760

Test R²

Comparing All Models







	Best Hyperparameter	Test MSE	R²
Baseline Model		2.2136	-9.7698e-10
Linear Regression		1.0719	0.5158
Ridge Regression	alpha = 100	1.0758	0.5140
Lasso Regression	alpha = 0.01	1.0707	0.5163
Elastic Net	alpha = 0.01, I1 ratio = 0.8	1.0714	0.5160
K-Nearest Neighbor	k = 16	1.5263	0.3105
Decision Tree with pruning	ccp_alpha = 0.002465	0.3860	0.8256
Random Forest	max_depth: 10, n_estimators: 300	0.3760	0.8301
AdaBoosting	Max_depth: 5, learning rate = 0.01, n_estimators: 50	0.3664	0.8345
Gradient Boosting	Max_depth: 3, learning rate = 0.1, n_estimators: 100	0.3654	0.8349









Best Model: Gradient Boosting

With the lowest MSE and highest R² Gradient Boosting outperformed every other model we ran!

Best Parameters

- # of Estimators: 100
 - Enough trees to reduce bias without overfitting
- Learning Rate: 0.1
 - To improve steadily and avoid overshooting
- Max Depth: 3
 - Deep enough to capture complex relationships but shallow enough to avoid overfitting

Test MSE:

0.3654

Test R²:









Challenges

- Lack of Popularity Features
 - Added 7 new columns to predict the influence of artist popularity.
- Raw stream number were large and skewed Log-transform stream to improve model stability and interpretation.
- Nested CV long runtime
 - Limited the hyperparameter grid, yet still experience long runtime.















Experimentation and Nested CV Matters



Ensemble Methods



Stream forecasting to drive decision making in the music industry











Thank you! Questions?