2023 Spotify Streaming Feature Analysis

Group 6: Katherine Liang, Yuwei Tang, Zhengning Li, Sharon Pang

Introduction

Our analysis on Spotify, a premier music streaming service, explores the determinants of song popularity using the 'Most Streamed Spotify Songs of 2023' dataset from Kaggle. Initial focus on musical features alone proved insufficient, leading to the incorporation of non-musical features. We applied four models: Random Forest, Gradient Boosting, SVM, and PCA. Findings indicate that 'In_Spotify_playlists' is the most critical factor influencing its popularity and the Random Forest model performed best.

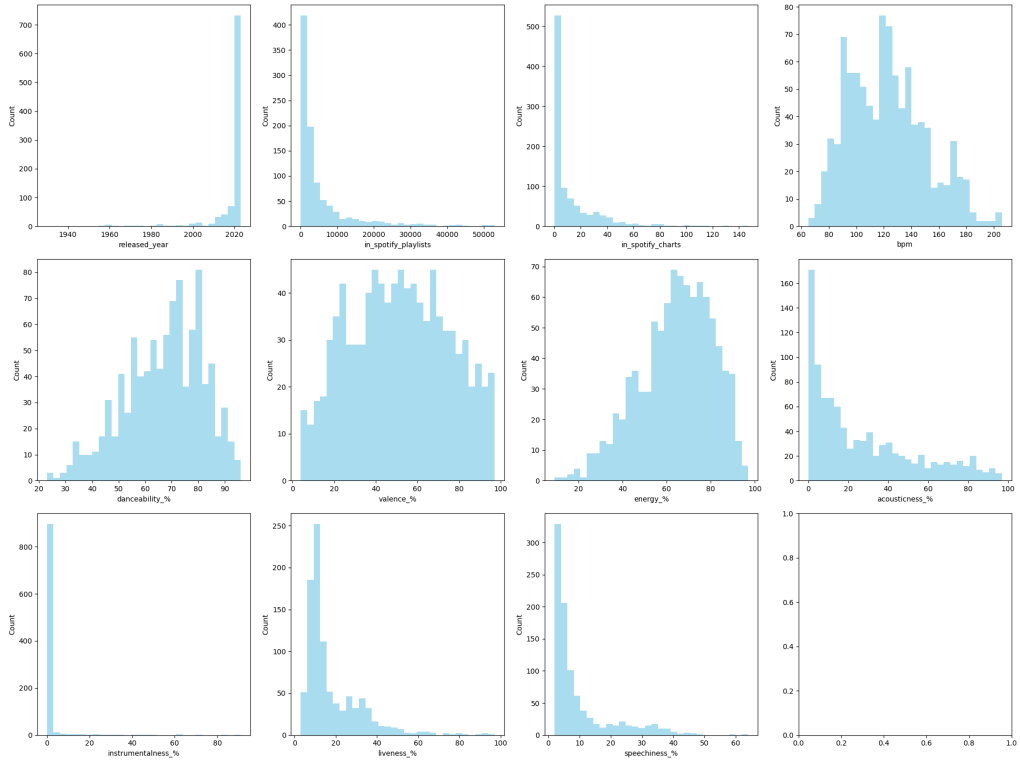
Data

In the dataset, it has 952 songs, each described by 11 variables after removing irrelevant variables such as artist_count, converting streams into millions.

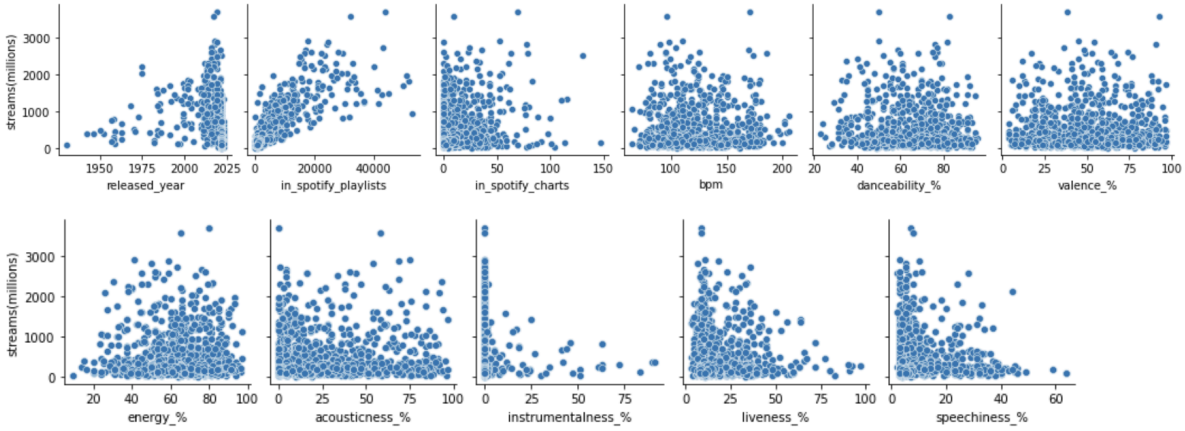
track_name	artist(s)_name	streams(millions)	released_year	in_spotify_playlists	in_spotify_charts	bpm
Seven (feat. Latto) (Explicit Ver.)	Latto, Jung Kook	141.382	2023	553	147	125
LALA	Myke Towers	133.716	2023	1474	48	92
vampire	Olivia Rodrigo	140.004	2023	1397	113	138
Cruel Summer	Taylor Swift	800.841	2019	7858	100	170
WHERE SHE GOES	Bad Bunny	303.236	2023	3133	50	144

danceability_%	valence_%	energy_%	acousticness_%	instrumentalness_%	liveness_%	speechiness_%
80	89	83	31	0	8	4
71	61	74	7	0	10	4
51	32	53	17	0	31	6
55	58	72	11	0	11	15
65	23	80	14	63	11	6

The histograms represent the frequency of variables in our dataset, providing a brief overview of data distribution. For example, the bpm graph shows several peaks at 90, 125, and 140, suggesting songs with these tempos are more likely to be popular.



Similarly, we also applied scatter plots to overlook the general patterns and associations. There isn't a linear relationship in any of them, explaining why we choose all nonlinear regression models to infer the features that contribute to a song's popularity.

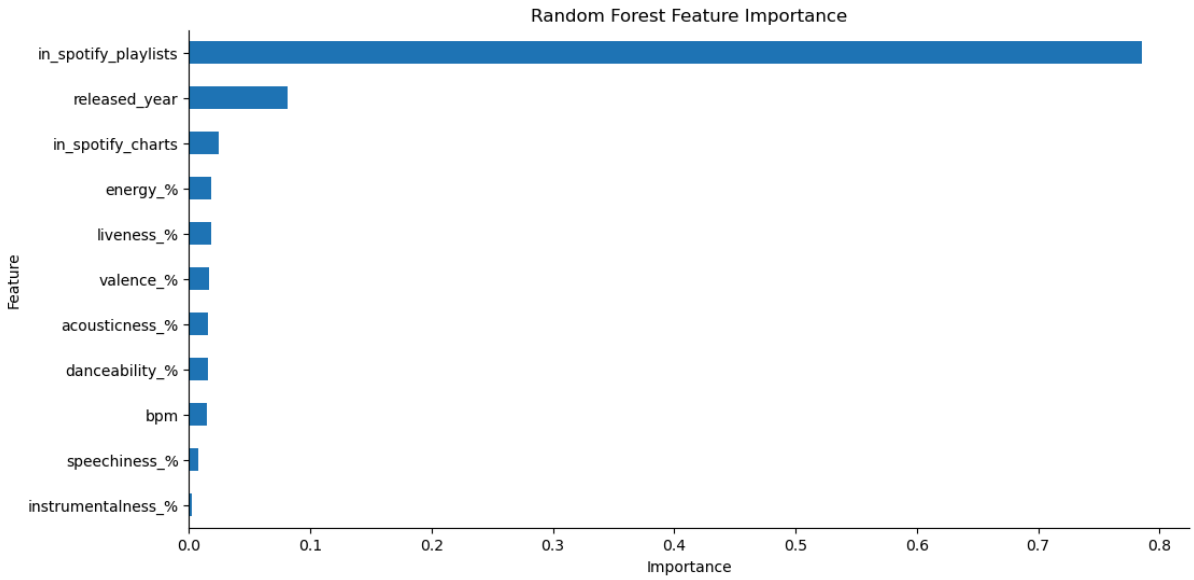


To sort the data, we split it into 80% training, 20% testing to prevent overfitting, then use below models to select the features that are more related to Spotify song's streams.

Method

Random Forest

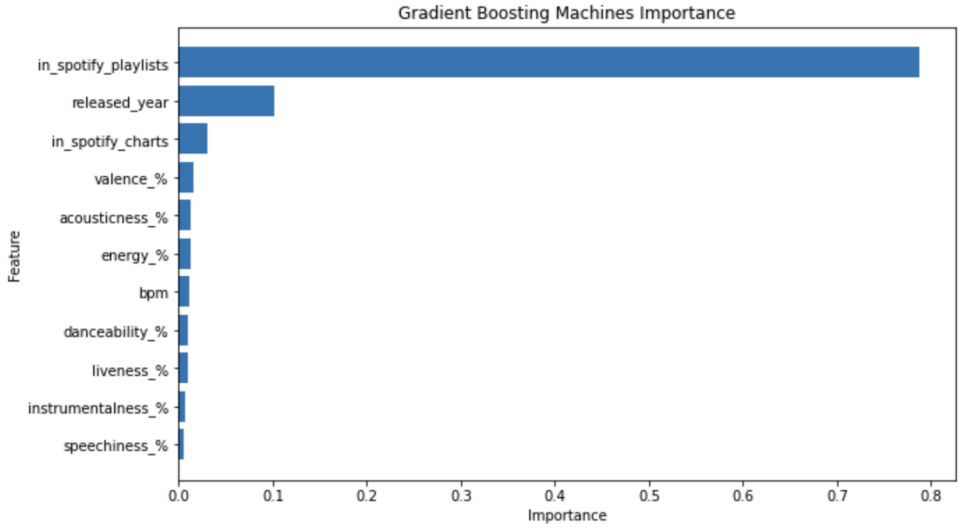
We used a random forest regression to reduce overfitting, as shown in a feature importance graph from the model. This graph ranks factors by their influence on the number of Spotify streams a song receives. 'In_spotify_playlists' is the top feature, followed by 'released_year' and 'in_spotify_charts'. Attributes like 'energy' and 'danceability' have moderate importance, while 'speechiness' and 'instrumentalness' are less significant. The model has a low mean square error and an accuracy of 0.8, indicating strong predictive performance.



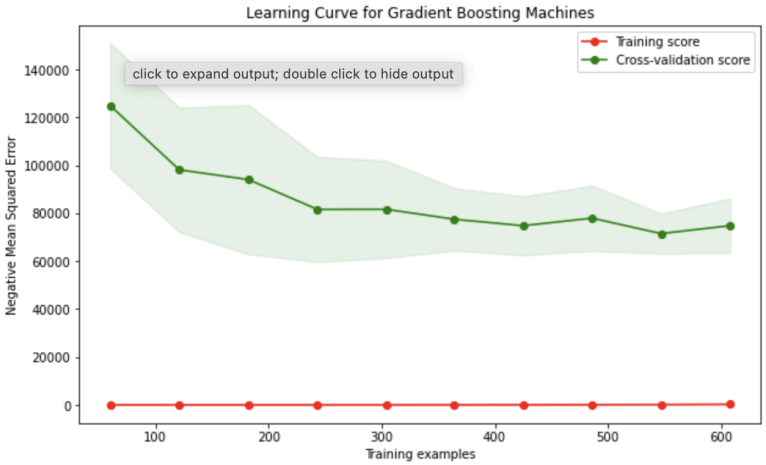
Gradient Boosting

Because of the non-linearity of the data, we then applied gradient boosting to build sequential models aimed at reducing underfitting. Upon examining the feature importance table, we discovered that the results closely resembled those obtained with random forests. The model

exhibited excellent performance, boasting a minimal mean squared error of 54,507 and an accuracy score of 0.78, demonstrating its precise predictive capabilities.



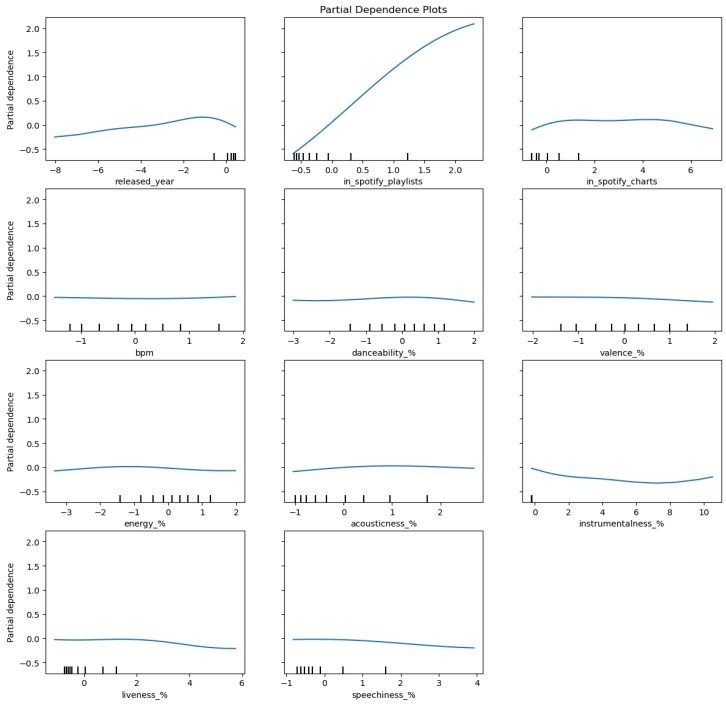
To further evaluate the performance, we plot the learning curve, which shows how the accuracy score changes as the iterations increase. The learning curve indicates the model might be underfitting and it is not generalizing well to unseen data.



SVM

The third model to rank the most relevant features is SVM regression, attempting to fit observations while limiting margin violations when: non- linear relationships, which is our case;

prediction with small to medium sized datasets; robust regression with outliers; high dimensional spaces; and cases requiring model flexibility. In the visualization, most features are slightly positive or negative associated with streams, except for the frequency a song appears in playlists, it's positively correlated with the stream. To conclude, it has a MSE of 100,082.32 and accuracy of 0.84, which are slightly off compared to previous models.

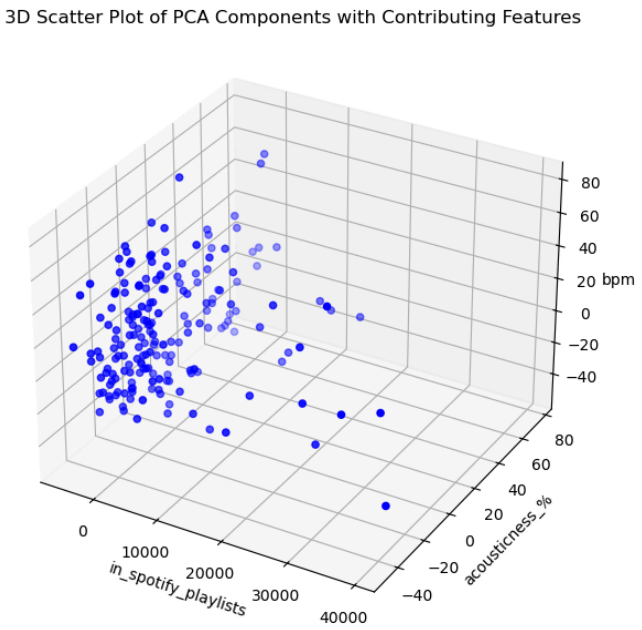


PCA

In our analysis, we used Principal Component Analysis (PCA) to reduce the dimensionality of our 11-variable dataset. Following the PCA transformation, we utilized a decision tree to determine the optimal number of principal components (n). We chose n=3 because the change in the explained variance ratio was not substantial, allowing us to rely more on improvements in MSE and R^2 values for decision-making. Although n=9 exhibited better performance metrics, we opted for n=3 due to its greater effectiveness in reducing the number of variables, thus better aligning with our goals of simplification and visualization.

n_components	MSE	R2	Explained Variance Ratio	top_features
1	190442.204	0.222	0.999948	[in_spotify_playlists]
2	163073.441	0.334	0.999962	[in_spotify_playlists, acousticness_%]
3	107537.600	0.561	0.999974	[in_spotify_playlists, bpm, acousticness_%]
4	122244.574	0.501	0.999983	[in_spotify_playlists, acousticness_%, valence_%, bpm]
5	128093.048	0.477	0.999989	[in_spotify_playlists, acousticness_%, valence_%, bpm, in_spotify_charts]
6	137177.859	0.440	0.999992	[in_spotify_playlists, acousticness_%, valence_%, bpm, in_spotify_charts, danceability_%]
7	171627.112	0.299	0.999994	[in_spotify_playlists, acousticness_%, valence_%, bpm, danceability_%, in_spotify_charts, liveness_%]
8	110176.436	0.550	0.999996	[in_spotify_playlists, acousticness_%, valence_%, bpm, danceability_%, energy_%, in_spotify_charts, liveness_%]
9	96812.289	0.605	0.999998	[in_spotify_playlists, acousticness_%, valence_%, danceability_%, bpm, energy_%, in_spotify_charts, released_year, liveness_%]
10	105689.100	0.568	0.999999	[in_spotify_playlists, acousticness_%, valence_%, danceability_%, bpm, energy_%, in_spotify_charts, released_year, liveness_%, speechiness_%]
11	110927.754	0.547	1.000000	[in_spotify_playlists, acousticness_%, valence_%, danceability_%, bpm, energy_%, in_spotify_charts, released_year, liveness_%, speechiness_%, instrumentalness_%]

After setting n=3, we identified the most influential features using the loadings calculated from the PCA components and explained variance, which highlighted 'in_spotify_playlists', 'acousticness', and 'bpm' as the top features. These were used to generate a 3D scatter plot of the PCA-transformed dataset, illustrating that 'in_spotify_playlists' significantly impacts dataset variance, while 'acousticness' and 'bpm' also contribute to variability, but to a lesser extent.



Conclusion

In conclusion, Random Forest and gradient boosting models exhibit superior performance, with low mean squared error and high model scores falling within the ideal range. "In_Spotify_playlists" emerges as the most influential feature across all models, alongside "released_year" and "in_spotify_charts."

Moving forward, we aim to expand the dataset beyond the top 1000 songs to include all Spotify songs up to 2023. Additionally, refining our analysis by considering internal song rankings will provide deeper insights into feature importance.

	Random Forest	Gradient Boosting	SVM Regression	PCA
MSE	49029.03	54407.78	100082.32	107537.60
Model Score	.80	.78	.67	.56

Citation:

1. Elgiriyewithana, Nidula. “Most Streamed Spotify Songs 2023.” *Kaggle*, 26 Aug. 2023, www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023.

Post- Conclusion

Member	Proposal	Coding	Presentation	Report
Katherine Liang	1	1	1	1
Yuwei Tang	1	1	1	1
Zhengning Li	1	1	1	1
Sharon Pang	1	1	1	1

Notes:

– Everyone in our group is very good and highly contribution!