Analytical Report on Music Streaming Data Analysis

Data Analysis Overview

The dataset used for this analysis contains comprehensive information about the most streamed Spotify songs in 2024, including metadata such as track names, artists, album details, streaming statistics across multiple platforms (Spotify, YouTube, TikTok, etc.), and playlist reach. Additionally, it includes explicit track indicators and popularity metrics. The analysis aimed to explore patterns in explicit vs. non-explicit tracks, assess popularity trends, and predict explicit content using machine learning models.

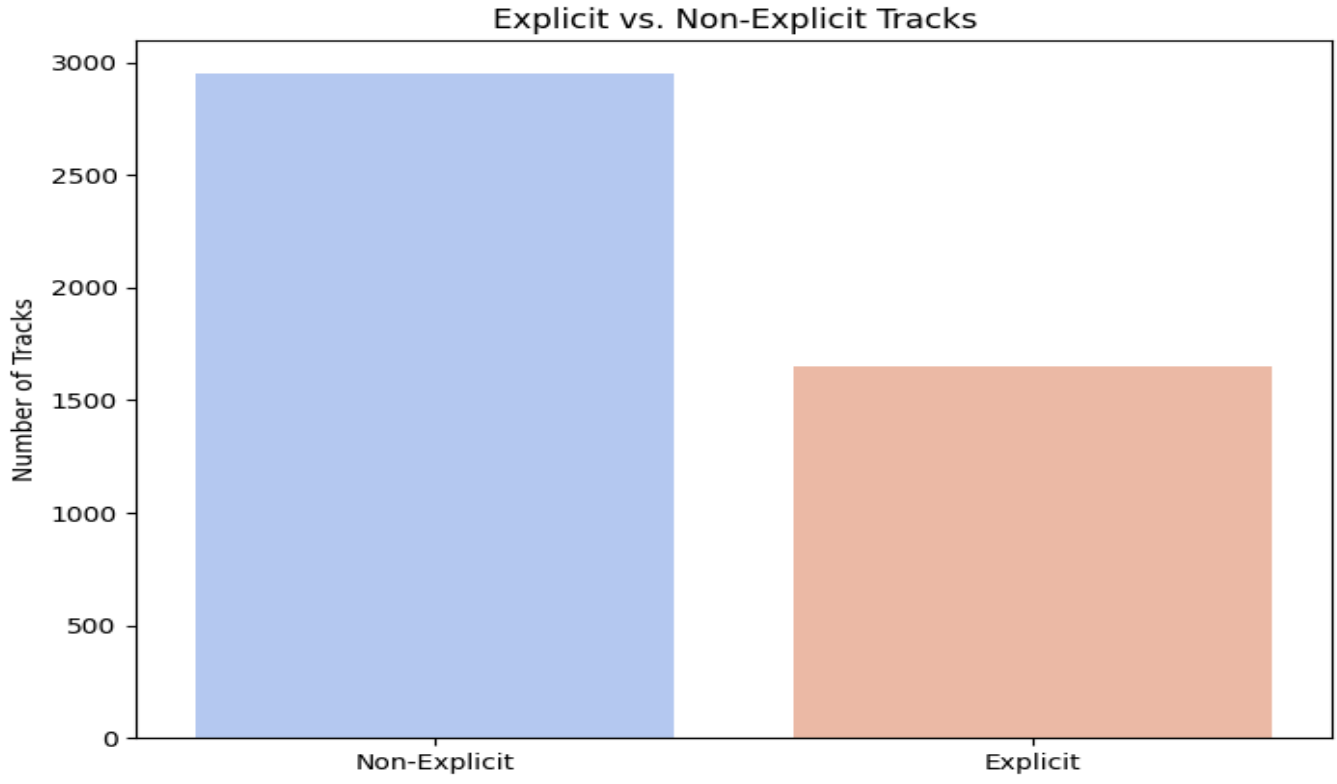
Exploratory Data Analysis (EDA)

---

1. Explicit vs. Non-Explicit Tracks

A bar chart comparison of explicit and non-explicit tracks revealed:

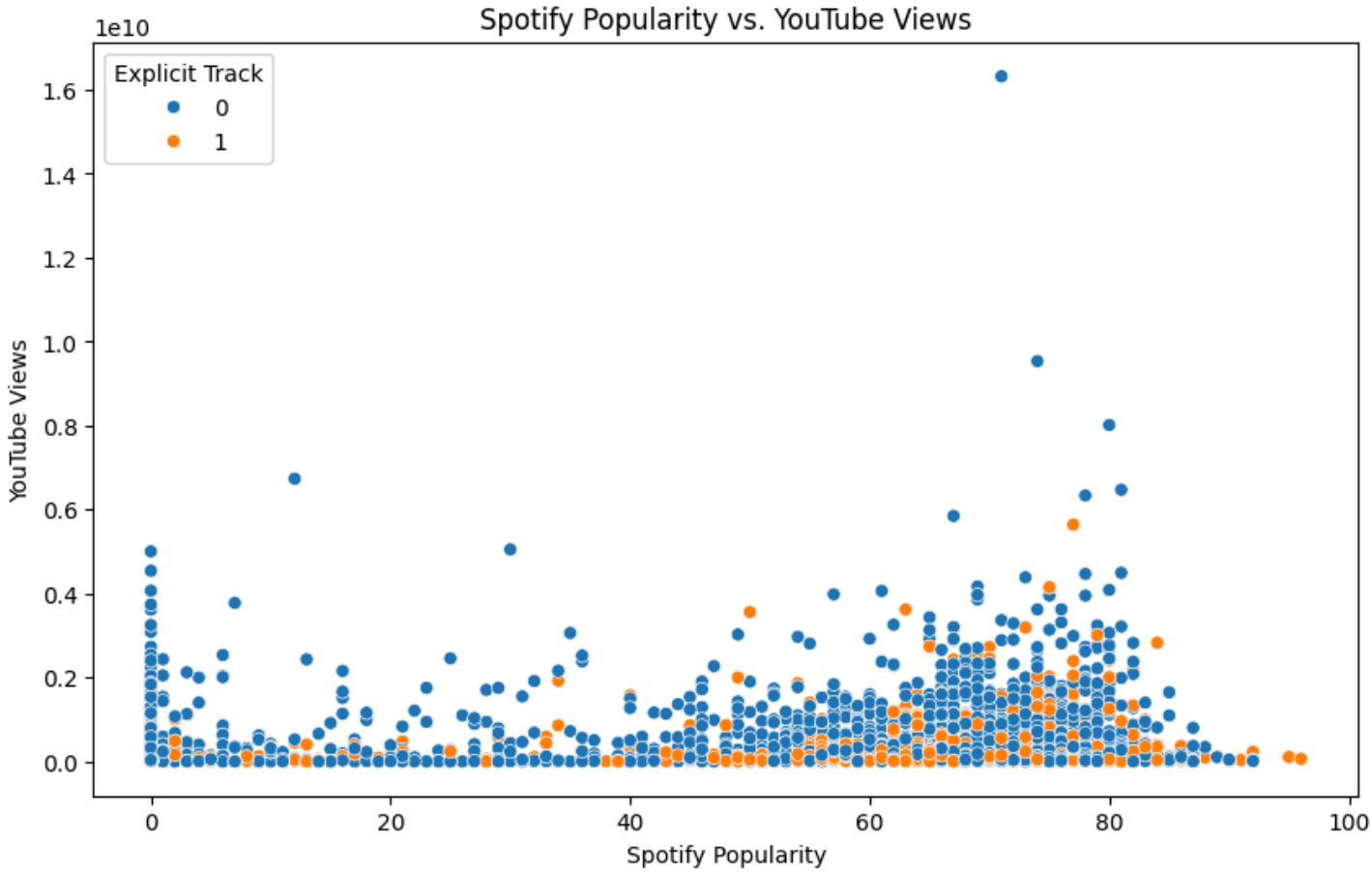
- Non-explicit tracks dominate the dataset with around 3000 entries, while explicit tracks are fewer, with approximately 1500 entries.
- This highlights a general preference for non-explicit content among listeners or a tendency for artists to produce more non-explicit music.



2. Spotify Popularity vs. YouTube Views

A scatter plot analyzed the relationship between Spotify popularity and YouTube views, categorized by explicitness:

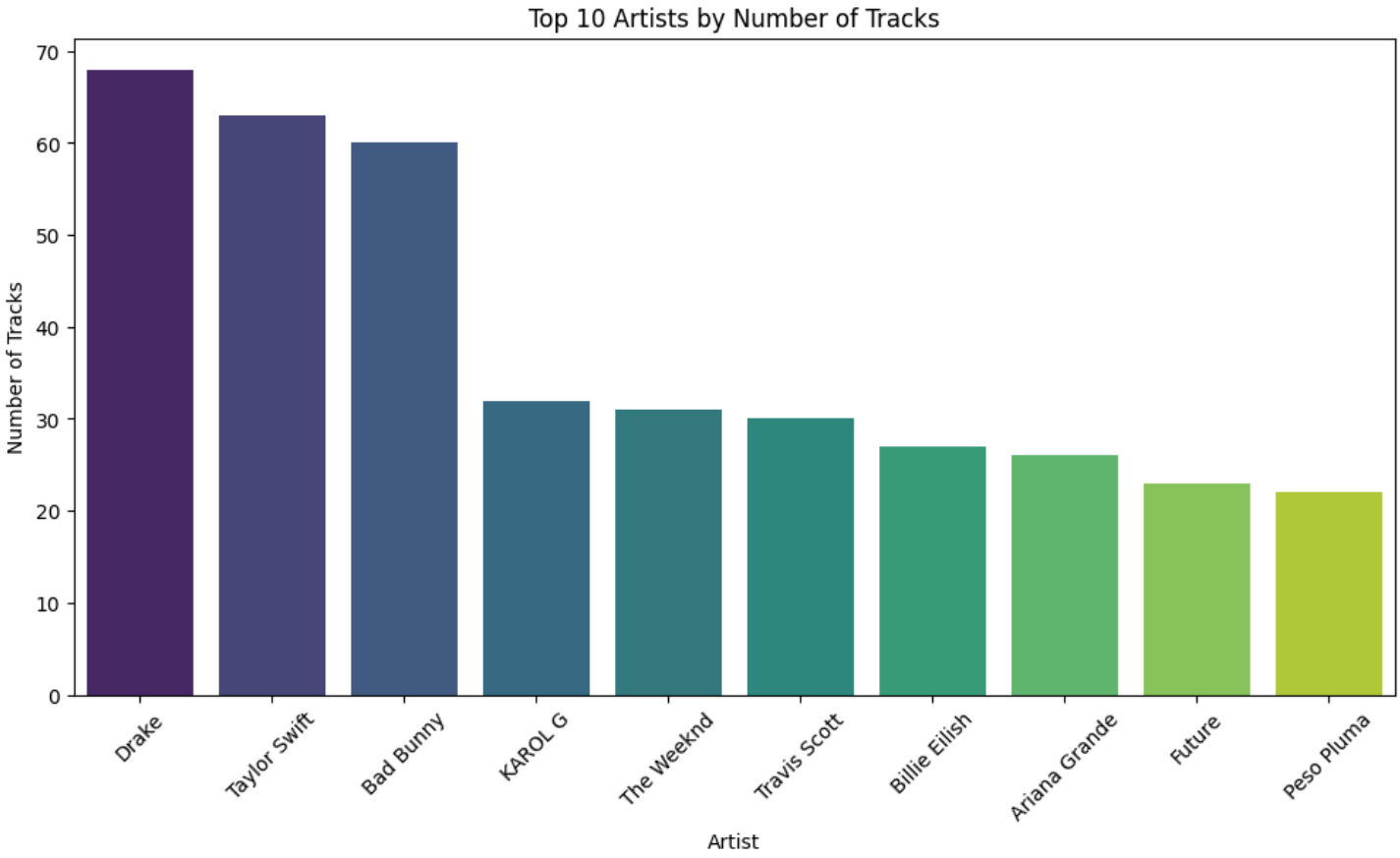
- Most tracks, whether explicit or not, cluster in the lower popularity and view ranges.
- Explicit tracks (orange points) show a slightly broader distribution across higher popularity scores, though non-explicit tracks dominate the extremes of YouTube views.
- The relationship between Spotify popularity and YouTube views appears non-linear, suggesting varying audience behaviors across platforms.



3. Top 10 Artists by Number of Tracks

A bar chart showcasing the top 10 artists revealed:

- **Drake**, **Taylor Swift**, and **Bad Bunny** lead the dataset with the highest number of tracks.
- Other notable artists include **KAROL G**, **The Weeknd**, and **Billie Eilish**, indicating their significant influence in streaming platforms.
- This highlights the dominance of certain artists in streaming metrics, likely due to their consistent releases and global appeal.



Machine Learning Models

To predict whether a track is explicit based on the provided features, several machine learning models were implemented and evaluated. Below are the results:

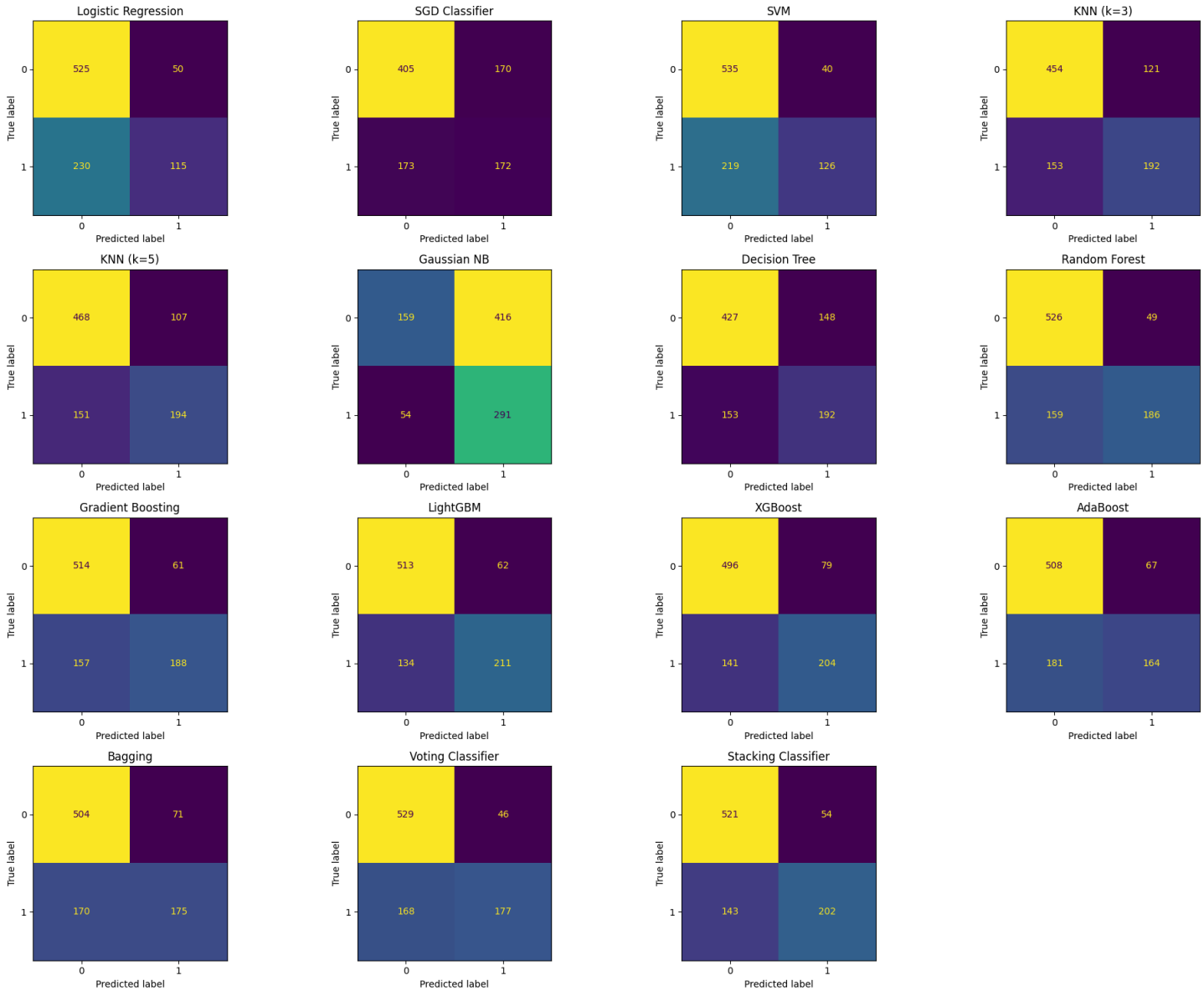
Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	MAE	MAPE	RMSE	R^2 Score
Logistic Regression	0.6957	0.6970	0.3333	0.4510	0.3043	2.45e+14	0.5517	-0.2986
SGD Classifier	0.6272	0.5029	0.4986	0.5007	0.3728	8.32e+14	0.6106	-0.5907

SVM	0.7185	0.7590	0.3652	0.4932	0.2815	1.96e+14	0.5306	-0.2012
KNN (k=3)	0.7022	0.6134	0.5565	0.5836	0.2978	5.92e+14	0.5457	-0.2707
KNN (k=5)	0.7196	0.6445	0.5623	0.6006	0.2804	5.24e+14	0.5296	-0.1965
Gaussian NB	0.4891	0.4116	0.8435	0.5532	0.5109	2.04e+15	0.7148	-1.1797
Decision Tree	0.6576	0.5431	0.5478	0.5455	0.3424	7.78e+14	0.5851	-0.4609
Random Forest	0.7739	0.7915	0.5391	0.6414	0.2261	2.40e+14	0.4755	0.0354
Gradient Boosting	0.7630	0.7550	0.5449	0.6330	0.2370	2.99e+14	0.4868	-0.0110
LightGBM	**0.7870**	0.7729	**0.6116**	**0.6828**	**0.2130**	3.04e+14	**0.4616**	**0.0910**
XGBoost	0.7609	0.7208	0.5913	0.6497	0.2391	3.87e+14	0.4890	-0.0203
AdaBoost	0.7304	0.7100	0.4754	0.5694	0.2696	3.28e+14	0.5192	-0.1501
Bagging	0.7380	0.7114	0.5072	0.5922	0.2620	3.48e+14	0.5118	-0.1177
Voting Classifier	0.7674	**0.7937**	0.5130	0.6232	0.2326	2.25e+14	0.4823	0.0075
Stacking Classifier	0.7859	0.7891	0.5855	0.6722	0.2141	2.64e+14	0.4627	0.0864

Insights from Model Performance

1. **LightGBM** emerged as the best-performing model, achieving the highest accuracy (78.7%), recall (61.16%), and F1 score (68.28%).
2. **Random Forest** and **Gradient Boosting** also performed well, demonstrating robust predictive capabilities.
3. Simpler models like Logistic Regression and Gaussian NB struggled with recall and R^2 scores, indicating limitations in capturing the complexity of the data.
4. Ensemble methods (e.g., Voting and Stacking Classifiers) provided balanced performance, leveraging the strengths of multiple models.



Recommendations

---

1. **Focus on Non-Explicit Tracks:** Given their dominance in the dataset, non-explicit tracks should be prioritized for playlist curation and promotional strategies.
2. **Leverage Multi-Platform Metrics:** The non-linear relationship between Spotify popularity and YouTube views suggests opportunities to target platform-specific audiences.
3. **Adopt LightGBM for Prediction:** Its superior performance makes it an ideal choice for predicting explicit content.
4. **Promote Top Artists:** Capitalize on the popularity of leading artists like Drake and Taylor Swift to maximize audience engagement.

Conclusion

This analysis provides actionable insights into the streaming landscape, highlighting trends in explicit content, platform-specific behaviors, and artist dominance. The predictive modeling results demonstrate the potential of machine learning to enhance decision-making in the music industry. Future work could explore deeper feature engineering and hyperparameter tuning to further improve model performance.