Simon Xia
sx2393@columbia.edu

Data Analysis Report

1. Actionable Insights:

- **Monthly downloads** and **favorites** are the most important key indicators of high plays counts.
- **Cumulative add-to-playlist** is the best key indicator for a **sudden increase** in plays.
- **Trend Analysis:** The majority of plays for the top 10 hit songs on Audiomack occurred in Sub-Saharan Africa. This trend could be associated with Audiomack's business model, which focuses on Afrobeats and Latin music (Table 1). Hence, I will search for hidden gems specifically in Sub-Saharan Africa.
- **Top 10 Hidden Gem Songs** that will be future hits in Sub-Saharan Africa :
    a. Goodsin by olivetheboy
    b. Butta My Bread by JZyNo
    c. Daddy Wey Dey Pamper by Moses Bliss
    d. Call of Duty by Zinoleesky
    e. Aseda by Nacee
    f. Too Faithful by Moses Bliss
    g. Personal by Zinoleesky
    h. Many Things by Zinoleesky
    i. Hope by XXXtentation
    j. F.N by LilTJay

2. Justification:

I captured early risers by calculating ratios and cumulative data from engagement metrics and by populating an additional column for the difference in plays for each region. The ratio reflects the "conversion" rate for newly uploaded songs, while cumulative data captures the lagging and compounding effects on play counts. For trend analysis, I mapped countries to regions for more interpretable classification.

Next, I built an XGBoost Regressor and used SHAP values to evaluate which features are most positively relevant. Given the time series nature of the data and the need to capture certain spikes, the XGBoost Regressor is sufficiently sophisticated and more explainable than an LSTM.

Interesting findings: Download over plays ratio for certain songs exceeds its plays. How could more people download a song without listening to it? My speculation is that when people downloaded the entire album, all songs count for 1 download, but only hit songs in that album would be played most frequently.

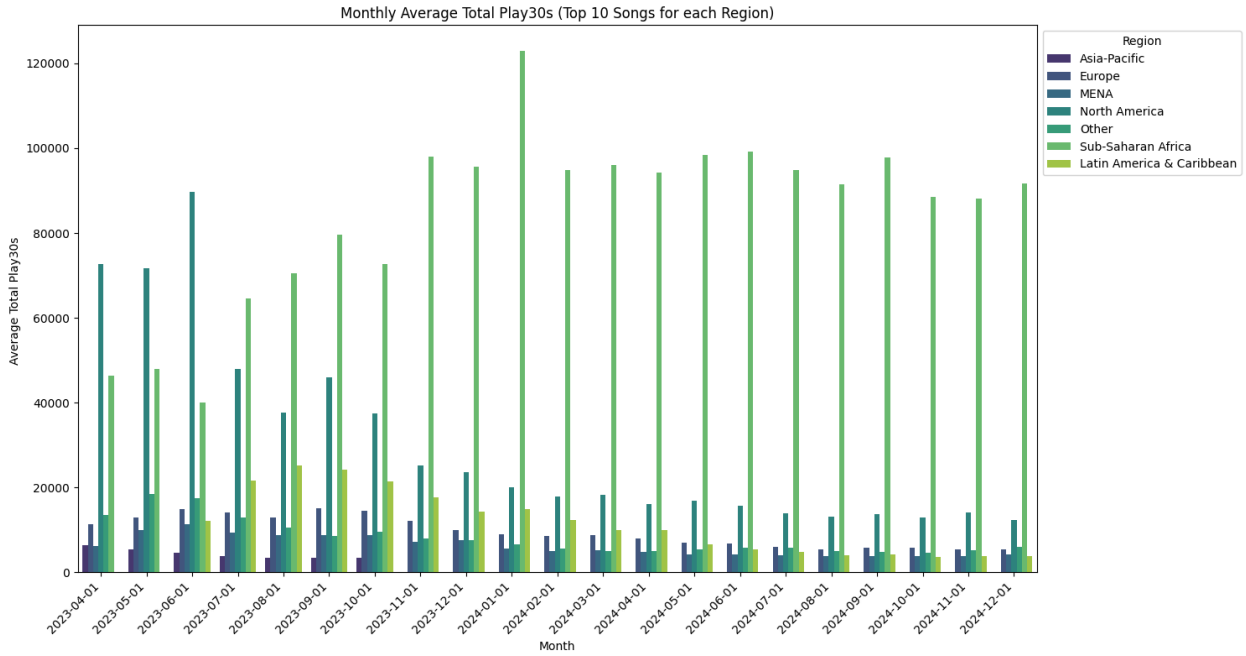
3. Data for better decision making

- Introduce first 30 seconds retention rate data (similar to what youtube is doing).
- Collaborate with Shazm to capture songs that suddenly got recognized significantly.
- Improve granularity of data to day instead of months to check out first 72 hours download rates for new songs uploaded.
- Introducing news data, for example, beef between kendrick and drake provided 'context' for the song 'Not Like Us' to the peak which later won awards in grammy. If we can incorporate news data into our analysis of potential hit songs, it might generate faster predictions than looking at engagement data.

Simon Xia
sx2393@columbia.edu

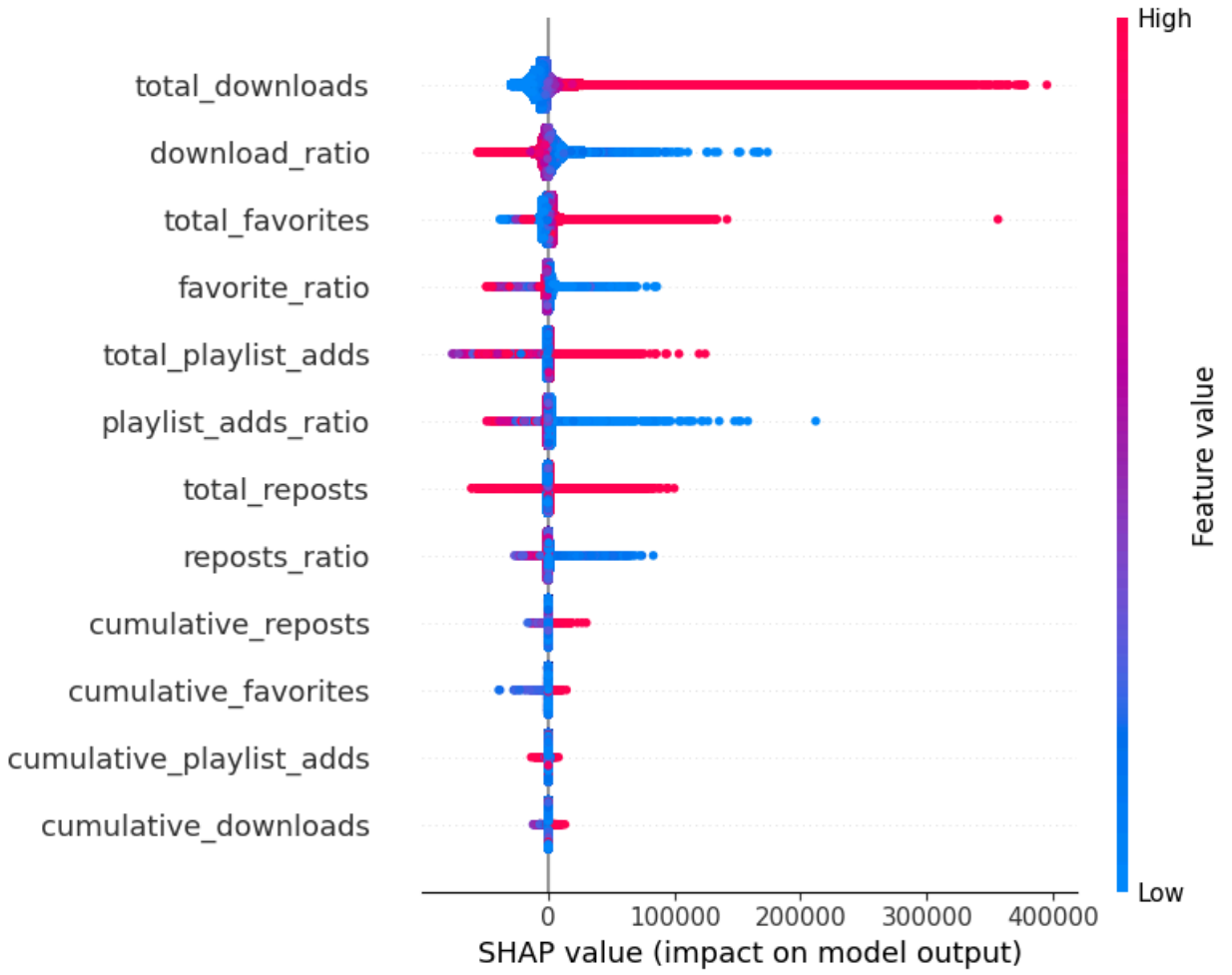
Appendix:

Table1:



Simon Xia
sx2393@columbia.edu

Table 2:



Simon Xia
sx2393@columbia.edu

Table3: