Project: Predicting Viral Songs Using Spotify & TikTok Data

1. Introduction

This project explores whether a song's audio features can help predict if it will go viral.

I combined Spotify audio features with TikTok viral song data and created a dataset with over 120,000 tracks labeled as viral or nonviral.

2. Data Preparation

I cleaned the datasets, matched songs by track ID, removed missing values, and unified duration units.

Key features included energy, valence, danceability, loudness, tempo, speechiness, instrumentalness, popularity, and others.

I also created a binary viral label and split the data into training and test sets.

3. Exploratory Analysis

Basic EDA showed clear differences between viral and nonviral tracks.

Viral songs tended to be louder, happier (higher valence), faster, and had more speech-like elements.

Nonviral songs were generally more danceable, acoustic, and longer.

4. Model

I trained a logistic regression model using the selected audio features.

Most features were statistically significant. Positive predictors of virality included loudness, speechiness, tempo, valence, and popularity.

Negative predictors included danceability, energy, acousticness, and duration.

5. Performance

Test accuracy was 94.9%, but because the dataset is imbalanced, AUC was a better measure.

The model achieved an AUC of 0.886, showing strong ability to separate viral from nonviral songs.

6. Conclusion

Audio features alone cannot fully capture social factors behind viral success, but they still provide strong predictive power.

This project demonstrates how audio data can be used to model virality and could be extended with TikTok engagement metrics or with more advanced models.