Forecasting Top 10 Music Genres: A Comparative Analysis of Time Series and Machine Learning Models

Mikayel Mikayelyan

American University of Armenia

BS Data Science

May 9, 2025

Abstract

We have seen a rapid evolution of the music industry in the last decade. This change was mainly driven by the accessibility to music provided by streaming platforms such as Spotify and Apple Music. Another big change has been social media companies like TikTok, which have had a massive influence on how streaming numbers behave. We have seen numerous songs from the 70s being resurrected, small independent artists reaching mainstream popularity overnight, and other unpredictable behaviors of music popularity. This creates the need for a reliable predictor of streaming data so music corporations can adjust and know which artists and genres to push. This project focuses on predicting the weekly popularity of music genres by forecasting the top 10 genres based on Spotify streams. An end-to-end data pipeline was developed, extracting data from multiple sources (Spotify charts, Deezer metadata, AcousticBrainz audio features, and Last.fm user engagement data), cleaning and preparing that data, feature engineering, and exploratory data analysis (EDA). A list of forecasting models was implemented and evaluated, including baseline methods (Naive, Moving Average), classical time series models (SARIMA, Prophet), and machine learning approaches (Random Forest, LightGBM). This report details the methodologies employed, presents a comparative analysis of model performance, and discusses the key findings. The results aim to identify the most effective models for genre popularity forecasting and provide insights into the dynamic nature of music trends. Keywords: Music Genre Forecasting, Time Series Analysis, Machine Learning, Spotify, Deezer, Last.fm, AcousticBrainz, Feature Engineering, Prophet, SARIMA, Random Forest, LightGBM, Exploratory Data Analysis.

Contents

1. Introduction

The music industry has undergone a huge shift with the introduction of digital streaming platforms. As a result of these platforms democratizing access to music, they generate an unprecedented volume of data reflecting listener preferences and consumption patterns. Understanding and forecasting these trends, particularly at the genre level, is crucial for artists, record labels, streaming services, and marketing agencies. Accurate genre popularity forecasts can inform content acquisition, playlist curation, promotional strategies, and resource allocation.

This capstone project addresses forecasting weekly music genre popularity, with a specific focus on the top 10 genres based on Spotify streams. To achieve this, data was collected and integrated from diverse sources including Spotify chart data, Deezer for rich track metadata, AcousticBrainz for audio features, and Last.fm. An ETL pipeline was developed to clean, process, and engineer a robust feature set suitable for time series and machine learning models.

A key component of this project was a thorough Exploratory Data Analysis (EDA) to uncover underlying patterns, seasonality, trends, and correlations within the data. Following the EDA, a range of forecasting models were implemented and evaluated. These models span from simple baselines like Naive and Moving Average forecasts to classical time series methods such as SARIMA and Prophet, and finally to more complex machine learning algorithms, including Random Forest and LightGBM.

This report details the entire project lifecycle: the data acquisition and preparation stages, the feature engineering strategies, the insights derived from the EDA, the methodologies behind each forecasting model, and a comparative evaluation of their performance. The primary evaluation metric used is the Mean Absolute Percentage Error (MAPE), supplemented by MAE and RMSE, to determine the most accurate and reliable models for this specific forecasting task. The findings aim to provide actionable insights and a robust framework for music genre trend prediction.

The remainder of this paper is structured as follows: Section 2 provides a review of relevant literature. Section 3 details the overall methodology, including data sources, the ETL pipeline, feature engineering, and the forecasting models employed. Section 4 summarizes the key findings from the Exploratory Data Analysis. Section 5 presents the performance results of the different models and their comparative analysis. Section 6 concludes the paper, and Section 7 suggests potential avenues for future research.

2. Literature Review

"Music Trend Prediction Based on Improved LSTM and Random Forest Algorithm" (Liu, 2022)

- **Overview:** Chen and colleagues combine an enhanced LSTM network with an attention layer and a Random Forest algorithm to predict pop music trends from large-scale user data. Their method preprocesses and normalizes the data, selects key features via Random Forest (aided by fuzzy clustering), and uses the improved LSTM to forecast future playback volumes, significantly reducing RMSE and MAE.
- **Relevance:** This study highlights the power of integrating deep learning with ensemble methods for accurate music trend forecasting, providing a strong methodological basis for our predictive modeling efforts in the music industry.

"Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify" (Pareek et al., 2022)

- **Overview:** Pareek, Shankar, Pathak, and Sakariya (2022) investigate the use of machine learning to forecast song popularity by analyzing Spotify's music metrics. Their study employs multiple classifiers—including Random Forest, K-Nearest Neighbor, and Linear Support Vector Classifier—on a dataset of 232,725 English songs from Kaggle. The results reveal that the Random Forest classifier outperforms the others, achieving an accuracy of 89%.
- **Relevance:** This research is significant for our work as it demonstrates how quantifiable audio features (e.g., energy, loudness, acoustics) can be leveraged to predict music success, providing a robust methodological framework for data-driven decision-making in the digital music industry.

3. Methodology

This section outlines the overall methodology employed in this project, from data acquisition to model development and evaluation.

3.1 Data Sources and Collection

The dataset for this project was constructed by integrating information from multiple sources:

- **Spotify Chart Data:** Weekly chart information, including stream counts and chart positions, served as the primary source for the target variable (genre-level Spotify streams). This was collected with a scraping script using Selenium.
- **Deezer API:** Used to enrich track information with metadata such as release date, album details (track count, duration), BPM, and primary genre classifications
- **AcousticBrainz API:** Provides detailed low-level and high-level audio features for tracks, offering insights into their acoustic properties (e.g., danceability, energy, mood probabilities).

- **Last.fm API:** Supplied user engagement data, including listener counts, play counts, album information, and user-generated tags for tracks.

3.2 ETL Pipeline and Feature Engineering

A multi-stage Extract, Transform, Load (ETL) pipeline was developed to process and integrate the raw data into a model-ready dataset (model_dataset_weekly.csv). Key steps included:

- **Spotify Data Scraping** (00_scrape_spotify.py): Collected weekly Spotify chart data, including track names, artists, stream counts, and chart positions. The output was a raw CSV file (e.g., spotify_charts_data_raw.csv).
- **Initial Cleaning and Aggregation** (10_cleaning.py): Cleaned and standardized Spotify data, aggregating it to a weekly level. Key outputs:
  - **unique_tracks.csv:** List of unique artist-song combinations.
  - **merged_weekly.csv:** Weekly aggregated Spotify metrics per track.
- **Deezer Enrichment** (20_enrich_deezer.py): Enriched the dataset with metadata from the Deezer API, including:
  - Release date, album details, track duration, BPM, rank, genres, and record type.
  - **Output:** unique_tracks_deezer_enriched.csv
- **Last.fm Enrichment** (30_enrich_lastfm.py): Added Last.fm metrics such as listeners, playcount, track duration, album, and tags.
  - **Output:** lastfm_enriched_tracks.csv.
- **MusicBrainz/AcousticBrainz Enrichment** (40_enrich_musicbrainz.py): Retrieved standardized track identifiers (MBIDs) and detailed audio features from AcousticBrainz.
  - **Outputs:** unique_tracks_with_mbid.csv and acousticbrainz_api_enriched_tracks.csv.
- **Feature Consolidation** (50_merge_features.py): Merged all enrichment data into a master feature table. Key steps:
  - Prefixed columns by source (e.g., deezer_*, acoustic_*).
  - One-hot encoded Deezer genres and top Last.fm tags.
  - Converted zero values in numeric Last.fm features to NaN
  - **Output:** master_feature_table.csv.
- Time Series Dataset Creation (60_merge_timeseries.py): Final dataset preparation combining static features with weekly Spotify data. Key engineering steps:

- Column normalization and date conversion.
- Derivation of release date-based features (e.g., song age, release year).
- Binarization of AcousticBrainz probabilities and cleanup of low-variance features. • Generation of lagged and rolling features for Spotify streams.
- One-hot encoding of album record types.
- Final cleaning and filtering of incomplete records.
- **Final Output:** model_dataset_weekly.csv.

3.3 Forecasting Models

A variety of models were implemented to forecast weekly genre-level Spotify streams.

- **Baseline Models**

  - **Naive Model** (14_genre_naive_model.py): Assumes that the stream count for the next week will be the same as the current week's streams for each genre.
  - **MovingAverageModel** (15_genre_moving_average_model.py): Forecasts future streams based on the average of the last N weeks' streams. A 4-week moving average (MA4) was selected as a reasonable configuration.

- **Time Series Models**

  These models were applied to aggregated genre-level time series.

  - **SARIMA** (12_genre_sarima_model.py): Seasonal Autoregressive Integrated Moving Average (SARIMA) models were fitted for each genre using pmdarima.auto_arima to automatically select the optimal (p,d,q) (P, D, Q)s parameters based on the Akaike Information Criterion (AIC). A seasonal period of m = 52 was used to reflect weekly seasonality.
  - **Prophet** (13_genre_prophet_model.py): Facebook's Prophet model was applied with yearly and weekly seasonality components. Hyperparameters such as changepoint_prior_scale and seasonality_prior_scale were tuned individually for each genre to improve forecasting accuracy.

- **Machine Learning Models**

  These models were trained on the feature set in model_dataset_weekly.csv at the song-week level. Predictions were then aggregated by date and genre to produce genre-level forecasts.

  - **Random Forest Regressor** (20_random_forest_regressor.py): An ensemble of decision trees trained to capture non-linear relationships between features and weekly streams. Hyperparameters were optimized for each song using RandomizedSearchCV with TimeSeriesSplit cross-validation.

- ○ **LightGBM Regressor** (21_lightgbm_regressor.py): A gradient boosting framework designed for efficiency and high performance on large datasets. The hyperparameters were tuned similarly per song using RandomizedSearchCV with TimeSeriesSplit cross-validation.

3.4 Model Evaluation

Models were evaluated on their ability to forecast genre-level Spotify streams for a hold-out test period (typically 8 weeks). The primary metric for comparison was the Mean Absolute Percentage Error (MAPE). Other metrics included:

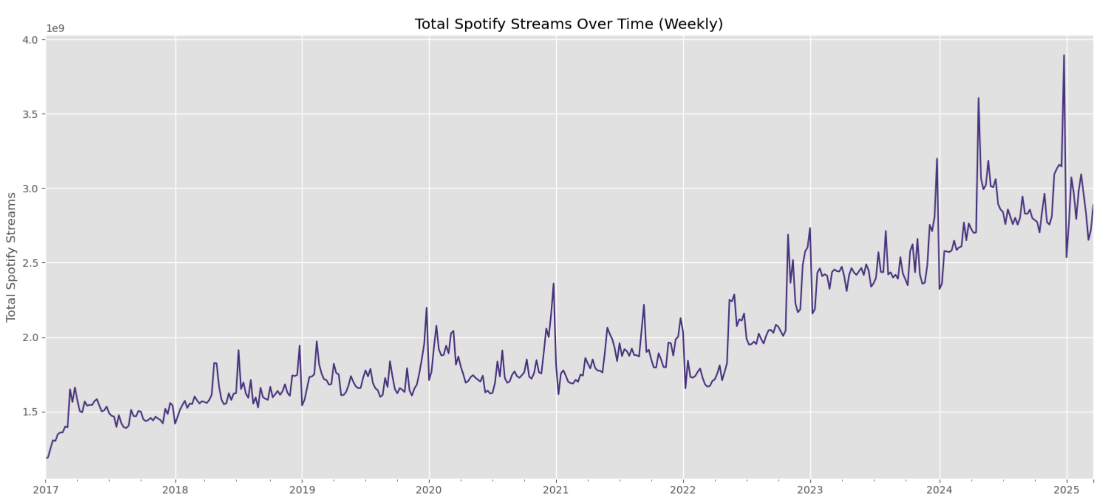
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Dynamic Time Warping Distance (DTW)

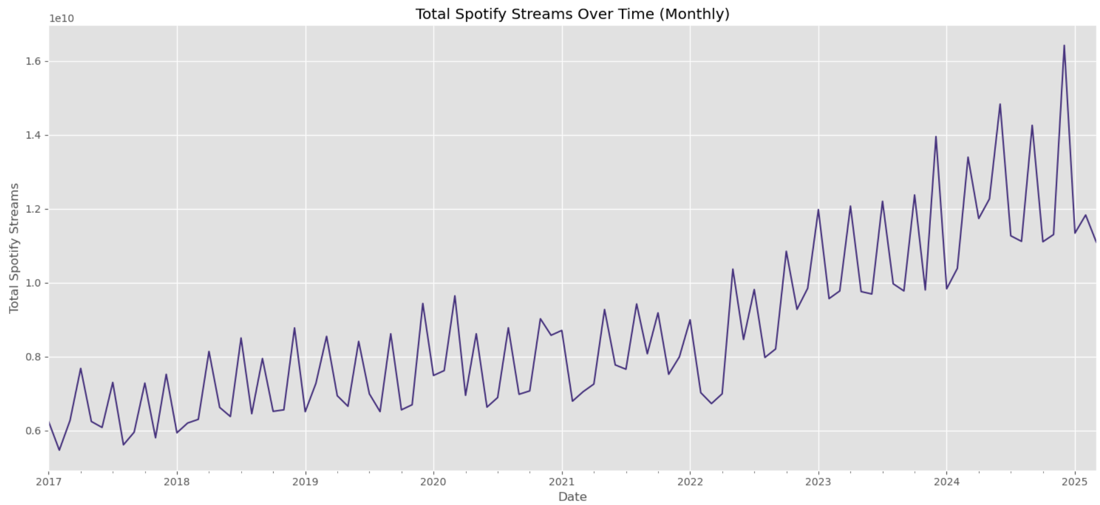
The combine_metrics.py, combine_predictions.py, and summarize_dtw.py scripts were developed to aggregate genre-level evaluation metrics from all forecasting models.

4. Exploratory Data Analysis (EDA)

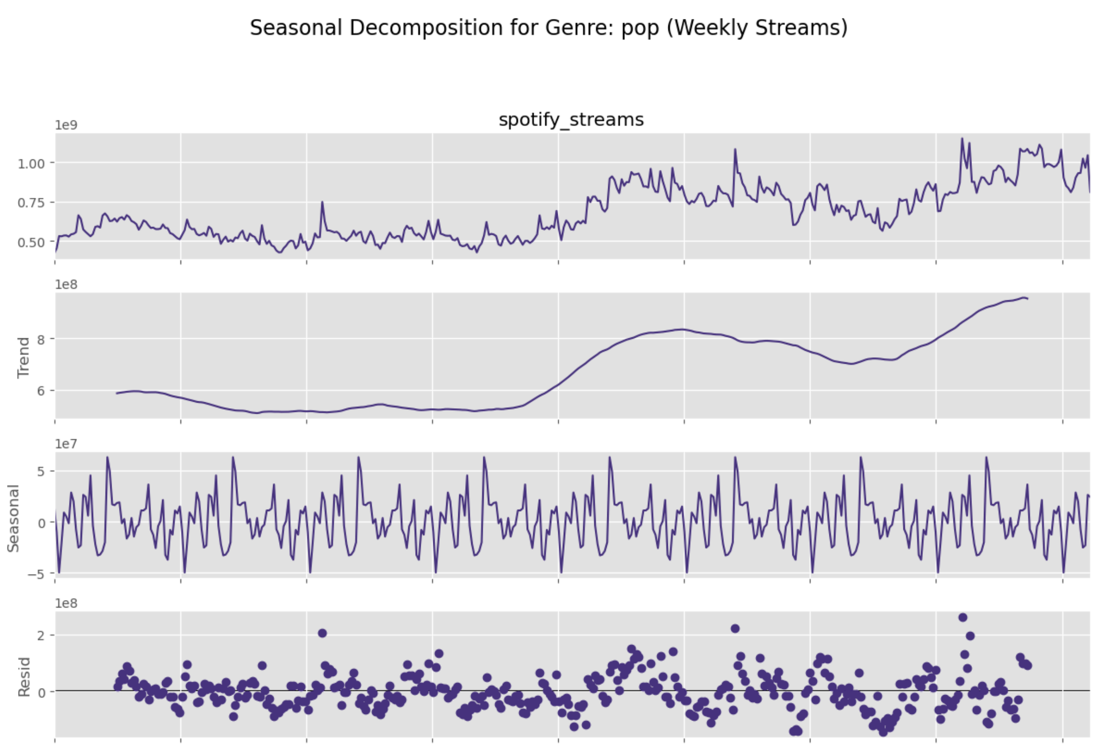
A comprehensive Exploratory Data Analysis (EDA) was conducted on the model_dataset_weekly.csv to understand the dataset, evaluate feature behavior, and identify meaningful patterns to inform modeling strategies. The analysis focused on understanding the behavior of streams and genre distributions.

- **Target Variable Overview:**
    - ○ The primary target variable is Spotify streams.
    - ○ Visual inspection shows a positive trend and some seasonality

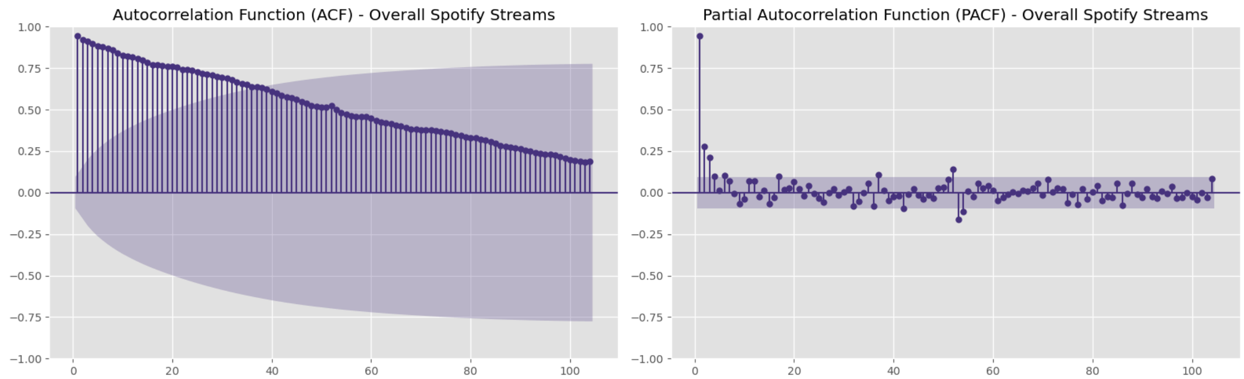




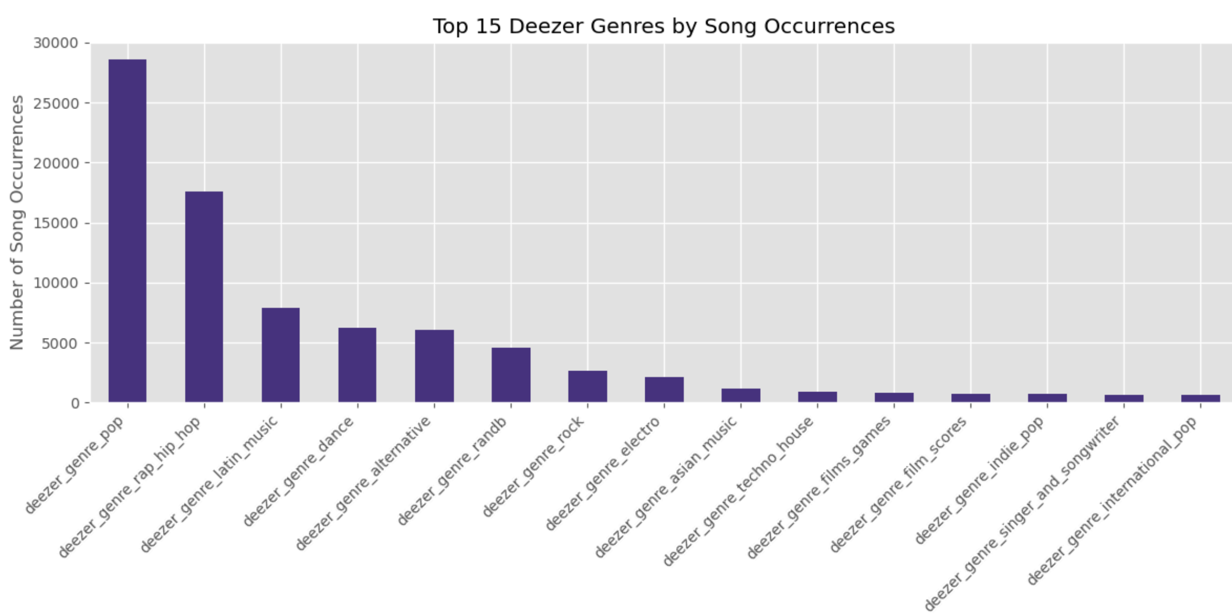
- **Seasonal Decomposition of Top Genre:**
  - Performed analysis of seasonal decomposition of the top genre (pop) to see if the previously observed trend and seasonality translate to the individual genres.
  - The overall weekly streams for Pop show variability with notable peaks and dips.
  - There is a steady upward trend starting around 2021, reflecting growing popularity and engagement with pop music.
  - Strong recurring seasonal patterns are present, indicating regular listening cycles or release patterns.



- **ACF and PACF of Top 5 Genres:**
  - Autocorrelation (ACF) and Partial Autocorrelation (PACF) function plots were examined for the top 5 genres to identify potential autoregressive (AR) and moving average (MA) orders for time series models. An example plot is shown below.

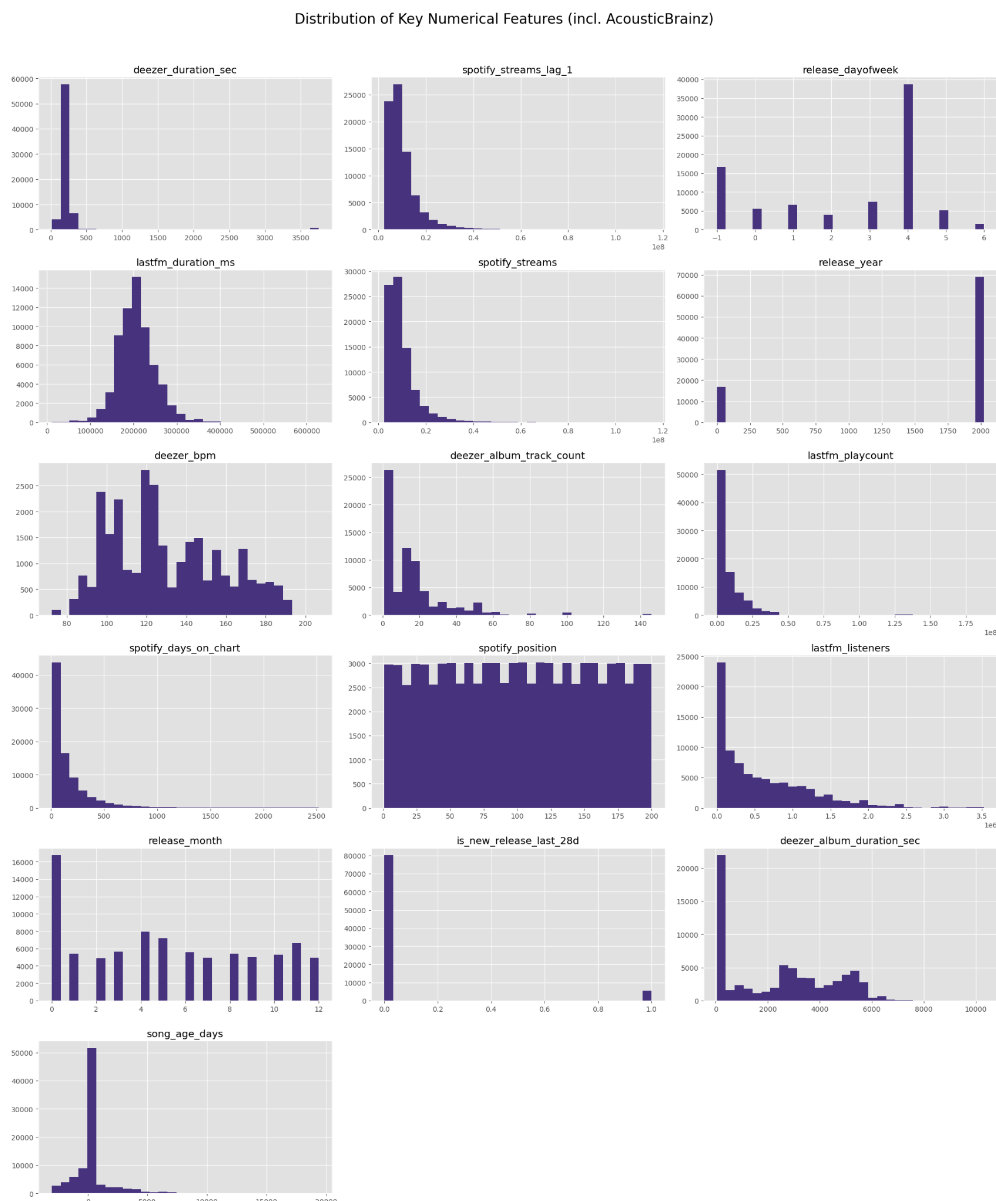


- **Genre Distribution:**
  - Looking at the number of streams per top genres to understand the distribution.
  - Pop, Rap/Hip-Hop, and Latin Music are the leading genres in terms of total Spotify streams, with Pop being nearly double Rap/Hip-Hop.
  - Also used boxplots to visualize the outliers.



- **Distribution of Key Numerical Features**

The plot of all of the Numeric Features highlighted the heavily right-skewed nature of streaming-related metrics such as Spotify Streams, lastfm_playcount, and spotify_days_on_chart, indicating that a small number of tracks dominate in popularity. Positional and categorical features like spotify_position and release_month show more uniform or cyclical distributions. Album-level metadata, such as deezer_album_track_count and deezer_album_duration_sec, display as you would expect according to most album structure patterns. Overall, the plot reveals clear scale imbalances across features, suggesting that normalization or log-scaling may benefit downstream modeling.



- **Correlation Analysis**
  - Spotify streams are highly correlated with their own lagged values (`lag_1`, `lag_4`) and rolling mean (`roll_mean_4`), confirming that recent performance is a strong predictor of current streams.
  - Moderate correlation with `lastfm_playcount` suggests that cross-platform engagement is somewhat aligned.

- **EDA Conclusions**

The exploratory data revealed that music streaming data has clear temporal patterns, including a long-term positive trend and seasonality, especially in dominant genres. Strong autocorrelation structures show that past performance is highly predictive of future streams, supporting using time series and lag-based features in forecasting models. On the other hand, acoustic and album metadata showed limited correlation with streams. These insights informed the feature selection and modeling strategies adopted in the forecasting analysis.

5. Forecasting Analysis

- **RMSE, MAE, and MAPE Comparison Summary Across Top 10 Genres**

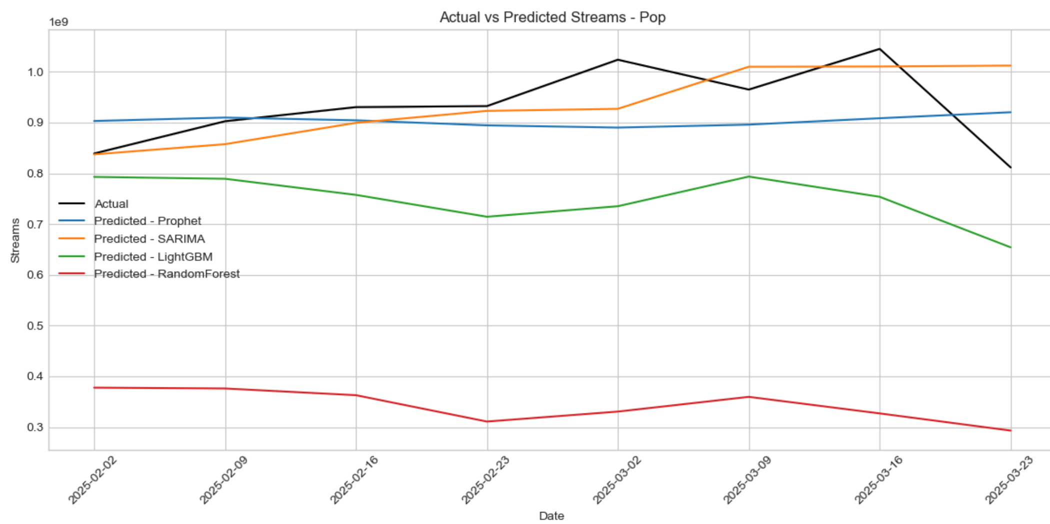
Random Forest consistently achieved the lowest values in almost all genres, as it is best at minimizing magnitude-based errors. LightGBM followed closely, ranking second in both metrics. SARIMA performed moderately well, improving upon the naive but falling short of the machine learning models. Prophet, Naive, and Moving Average models consistently were the weakest in absolute error performance. Some genres, like **Films & Games** and **Rap/Hip-Hop,** exhibited particularly high RMSE and MAE across all models.

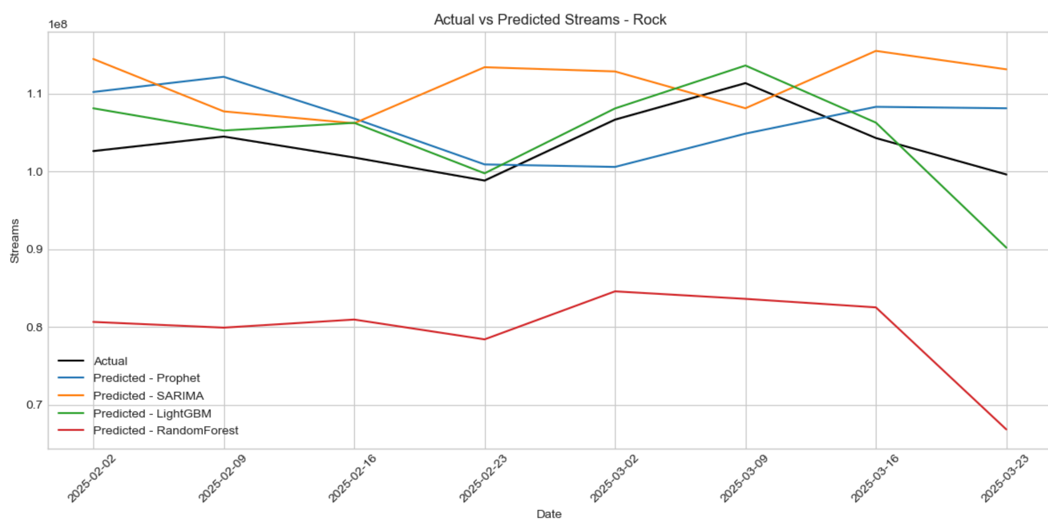
- **Temporal Shape Alignment (DTW)**

To evaluate how well each model captured the temporal shape of the actual stream patterns, the Dynamic Time Warping (DTW) distance was computed for all models and genres. Lower DTW distances indicate better temporal alignment. SARIMA and Prophet achieved the lowest DTW distances in most genres, suggesting that they better preserved the shape and fluctuations of actual streams. In contrast, LightGBM and Random Forest exhibited higher DTW distances, which means that while they reduced absolute error, they failed to capture the temporal dynamics effectively.

- **Actual vs Predicted Top 10 Genres**

The lower DTW is apparent when looking at predictions on top of the actual test data. SARIMA visually follows the shape of the actual data much better than the other models. Prophet generally produced flat forecasts with poor responsiveness to trends. Random Forest and LightGBM seem wildly off most of the time, but in the case of Rock music specifically, LightGBM did a great job at following the Actual Streams line.





- **Forecast Analysis Conclusions**

  - **Best Performing Models:**

    i. SARIMA: Most consistent in capturing temporal structure, achieving low DTW and competitive MAPE.
    ii. Random Forest: Best in reducing absolute errors (RMSE and MAE), but weaker in capturing temporal dynamics.

  - **Model Limitations:**

    i. LightGBM and Random Forest: Produced flattened forecasts, missing short-term variability.
    ii. Prophet, Naive, and Moving Average: Consistently underperformed across all metrics.

  - **Genre Difficulty:**

    i. Films & Games and Latin Music: Most difficult to forecast accurately, showing high errors across all models.

  - **Overall Recommendations:**

    SARIMA is the most balanced model for genre-level forecasting when both trend fidelity and error minimization are considered. Random Forest and LightGBM are best for applications where minimizing absolute error is the priority. Prophet, Naive, and Moving Average models are not recommended due to consistently poor performance.

6. Conclusions and Future Work

In this project, a range of forecasting models was successfully developed, implemented, and evaluated to predict weekly music genre popularity based on Spotify streaming data. By integrating multiple data sources (Spotify charts, Deezer metadata, AcousticBrainz audio features, and Last.fm), a robust dataset was created that is capable of supporting both statistical and machine learning approaches. The evaluation showed that SARIMA and Random Forest models consistently outperformed the others, with SARIMA excelling at capturing the temporal structure of genre trends and Random Forest minimizing absolute forecast errors. LightGBM showed promise in certain cases, but its inconsistent performance across genres suggests the need for further tuning. Simple baselines such as Naive, Moving Average, and Prophet consistently underperformed across all evaluation metrics and are obviously not recommended for production use.

Despite these advances, several limitations were identified. Models were unstable when forecasting for genres with fewer streams, such as Films & Games, and the tree-based models were unable to accurately capture temporal dynamics. These limitations present opportunities for future research. Expanding machine learning training to cover a larger sample of tracks or finding more metadata about each track with a higher correlation to streams could improve generalizability. Incorporating external features, such as social media trends, artist-level popularity, or event-driven data, may further enhance forecasting accuracy. Advanced deep learning models like LSTMs or Transformers could address the downsides of tree-based models. Finally, the development of ensemble methods combining the strengths of statistical and machine learning approaches holds the potential for delivering even more robust and accurate forecasts. Focusing on these directions, future work can provide valuable tools for industry stakeholders trying to capitalize on emerging music trends.