Spotify

The Sound of Streaming Behaviour

Uncovering Patterns in Spotify Listening Using Data Mining & Machine Learning

• • •

- Welcome
- About Us
- Our Team
- Agenda
- Our Services

- Achievements
- Our Goals
- Gallery
- Pricing
- Contact Us

‹ ›

Premium

bhavya ▼

Our Team

Aamir Khan

Artist

Ajay Karthick

Artist

Bhavya Saladhi

Artist

Surya Yoganathan

Artist

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

Premium

bhavya ▼

What Can Spotify Learn From How We Listen?

Business Question:
Can we uncover behavioural patterns in Spotify streaming data that help predict, segment, and personalize the user experience?

Goals:

- Understand when and how user listen
- Predict skip behaviour
- Segment listener by behaviour
- Forecast peak times
- Create smart playlists using listening history

01 Classification

02 Clustering

03 Time Series

04 Association rules

- Welcome
- About Us
- Our Teams
- Agenda
- Our Services

- Achievements
- Our Goals
- Gallery
- Pricing
- Contact Us

Premium

bhavya ▼

Cleaning up the noise

Data snapshots

140k	11	230	1.8k
listening records	features used	missing values	duplicates
Achieved	Achieved	Achieved	Achieved

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

Premium

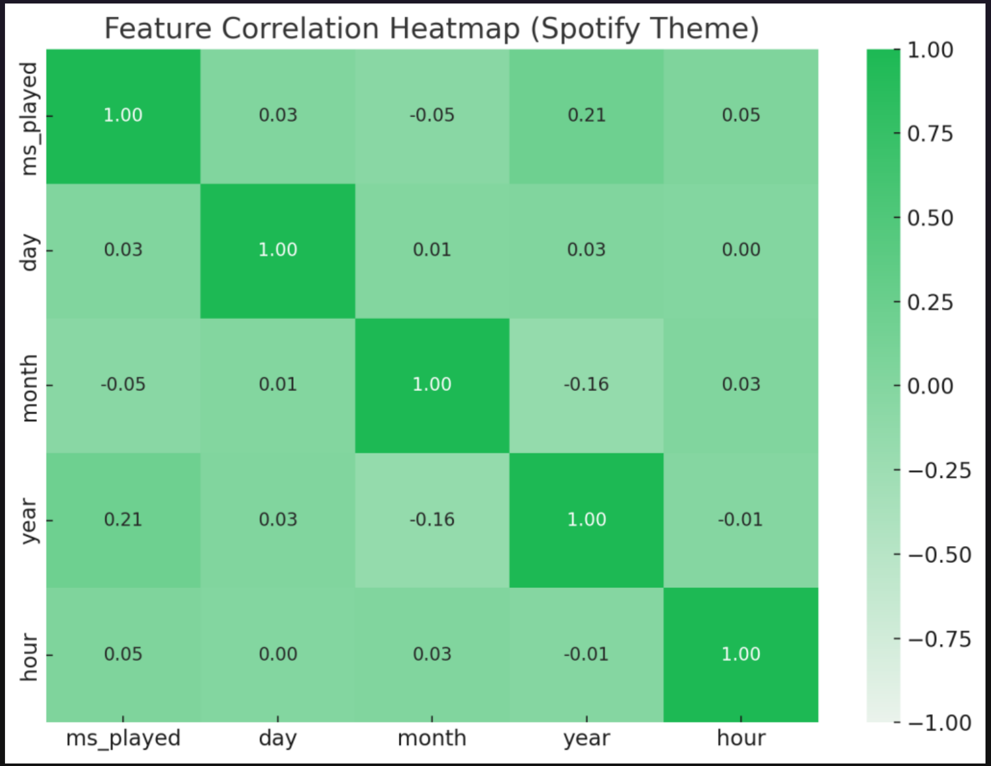
bhavya

CORRELATION

Play

Follow

- No strong correlations exist in the dataset.
- Listening time (ms_played) shows a mild positive trend over the years.
- Other numerical features (day, month, hour) do not strongly influence each other.



Feature Correlation Heatmap (Spotify Theme)

	ms_played	day	month	year	hour
ms_played	1.00	0.03	-0.05	0.21	0.05
day	0.03	1.00	0.01	0.03	0.00
month	-0.05	0.01	1.00	-0.16	0.03
year	0.21	0.03	-0.16	1.00	-0.01
hour	0.05	0.00	0.03	-0.01	1.00

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

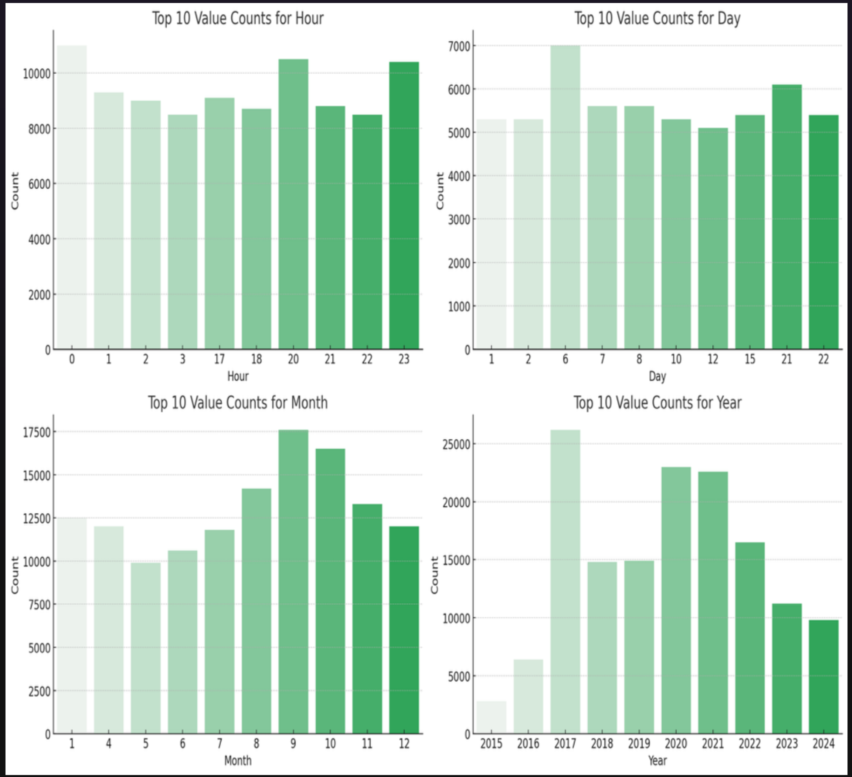
Our Goals

Gallery

Pricing

Contact Us

Listening Behaviour Trends Across Time Dimensions (Hour, Day, Month, Year)



Top 10 Value Counts for Hour

Top 10 Value Counts for Day

Top 10 Value Counts for Month

Top 10 Value Counts for Year

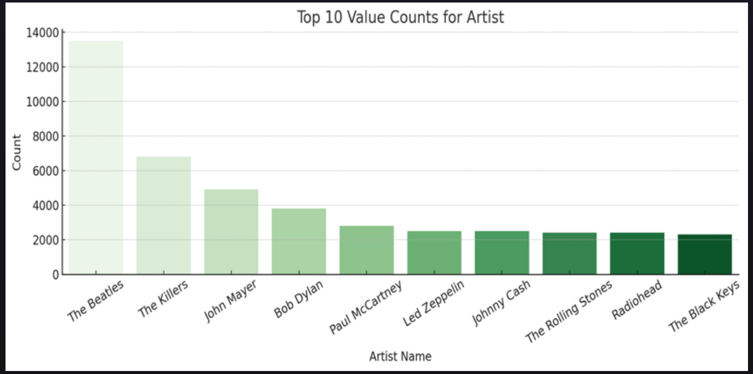
- **Hour of Day**: Highest listening activity occurs around midnight and late evenings (20:00–23:00), indicating user preference for nighttime streaming.
- **Day of Month**: Listening is fairly spread out across the month, with no strong peak pattern, suggesting consistent engagement.
- **Month of Year**: August and September show the highest activity, likely due to holidays, releases, or seasonal preferences.
- **Yearly Trend**: Overall streaming activity increased till 2020–2021, followed by a slight drop, possibly due to changes in behaviour post-pandemic.

Premium

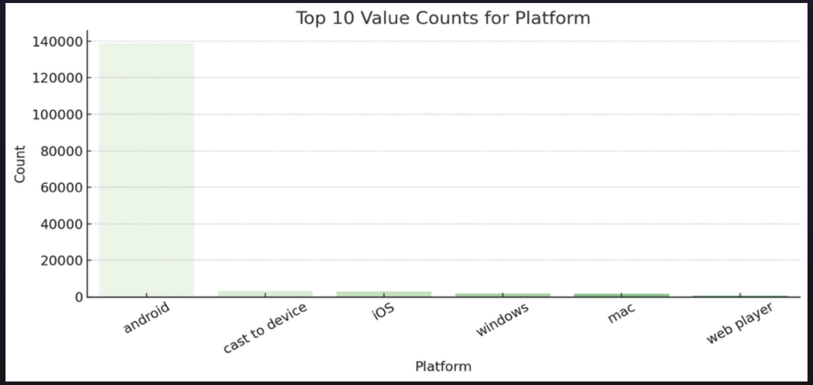
bhavya ▼

Top Categories in Listening Behaviour: Artists, Tracks & Platforms

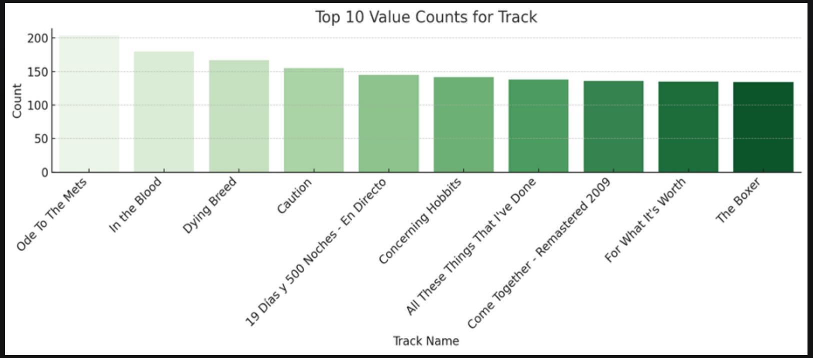
The Beatles are the most-streamed artist by a significant margin, showing high user engagement with timeless classics.



A few standout tracks like *"Ode To The Mets"* and *"In The Blood"* drive a large portion of engagement, highlighting potential hits for playlist curation.



Android dominates as the listening platform, accounting for the vast majority of streams, far ahead of iOS and other platforms.

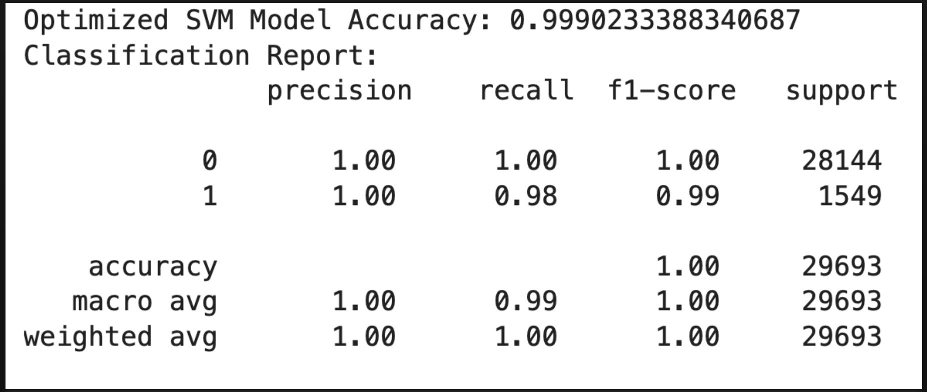


Welcome
About Us
Our Teams
Agenda
Our Services
Achievements
Our Goals
Gallery
Pricing
Contact Us

Classification

SVM

To identify what drives skip behaviour, we trained classification models using features like session time, device, and play duration. Our **Support Vector Machine (SVM)** model delivered the best performance, reaching **99.9% accuracy**.



```
Optimized SVM Model Accuracy: 0.9990233388340687
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28144
           1       1.00      0.98      0.99      1549

    accuracy                           1.00     29693
   macro avg       1.00      0.99      1.00     29693
weighted avg       1.00      1.00      1.00     29693
```

•Minority class (1) is well-handled with **Recall = 0.98**, indicating minimal false negatives.

•Balanced performance reflected in **Macro F1-score = 1.00**, confirming the model generalizes well across classes.

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

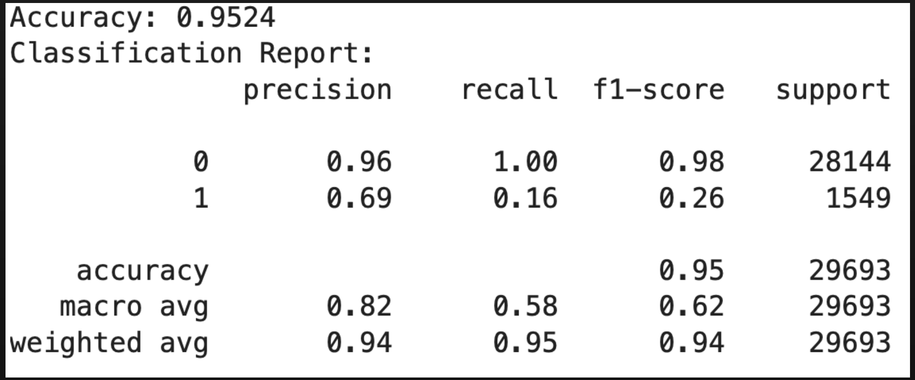
Contact Us

Premium

bhavya

Classification

Logistic Regression



```
Accuracy: 0.9524
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98     28144
           1       0.69      0.16      0.26      1549

    accuracy                           0.95     29693
   macro avg       0.82      0.58      0.62     29693
weighted avg       0.94      0.95      0.94     29693
```

Overall accuracy is decent at 95.24%, but performance drops sharply on class 1 (positive class).

Recall for class 1 is only 0.16, meaning the model fails to detect most positive cases.

Macro F1-score = 0.62, indicating poor balance and possible bias toward the majority class.

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

Premium

bhavya

Classification

SVM vs Logistic Regression

Model Comparison Table

Metric	Optimized SVM Model	Baseline Model	Better Model
Accuracy	99.90%	95.24%	Optimized SVM
Class 1 Recall	0.98	0.16	Optimized SVM
Class 1 F1-Score	0.99	0.26	Optimized SVM
Macro Avg F1-Score	1.00	0.62	Optimized SVM
Bias Toward Class 0?	No	Yes	Optimized SVM

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

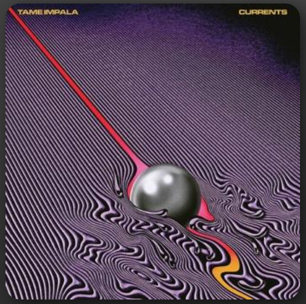
Gallery

Pricing

Contact Us

Premium

bhavya

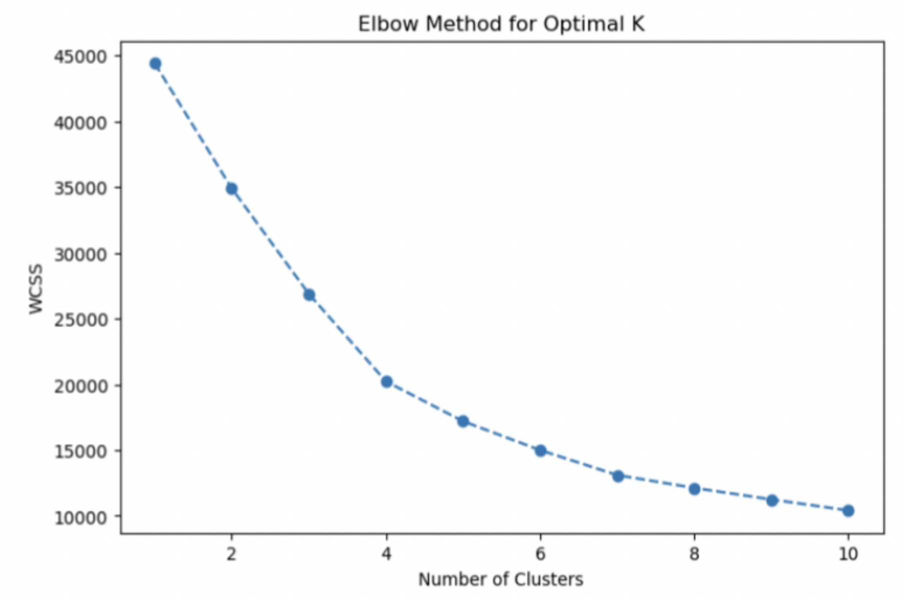


Clustering

Listening habits, unplugged

We grouped user activity into sessions and applied **K-Means clustering** to uncover hidden behaviour patterns. Session-level features revealed four natural listening styles based purely on engagement and interaction.

To determine the optimal number of clusters, e used the Elbow method, which revealed a natural bend at k=4. This balance allowed us to capture meaningful variation in listener behaviour while avoiding overfitting or overly granular segmentation.



Premium

bhavya

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

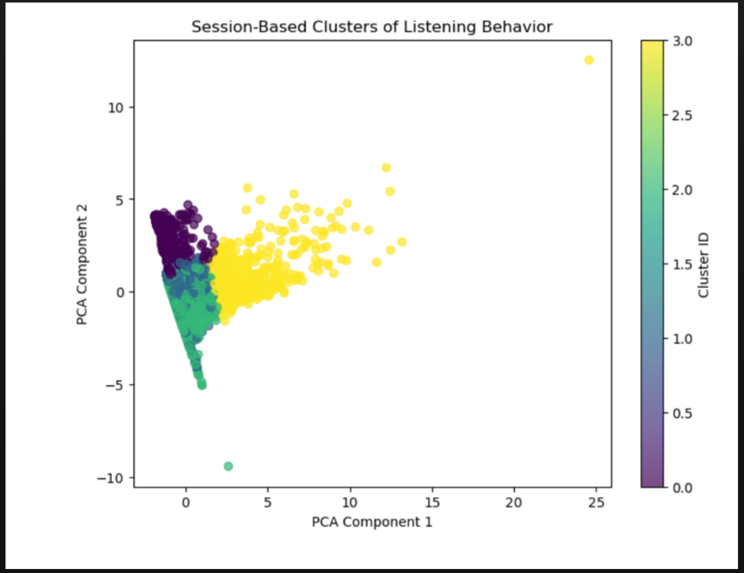
Contact Us

Clustering

What kind of listener are you?

After clustering sessions with K-Means, we used PCA to reduce dimensions and visualize user behaviour.

- ● Loyal Listeners: Long sessions, rarely skip tracks.
- ● Skimmers: Short sessions, frequently skip tracks.
- ● Repeaters: Small sessions with repeated tracks.
- ● Focused Listeners: Long, intentional sessions with minimal skipping.



Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

Premium

bhavya ▼

Time Series

Forecasting listening trends

We used time series forecasting to understand when users are most active on Spotify. Using ARIMA and Prophet models, we analysed hourly, weekly, and monthly trends to predict future streaming behaviour and identify peak engagement periods.

```
model = ARIMA(hourly_data, order=(2,1,2))
results = model.fit()
forecast = results.forecast(steps=24)
```

```
model = Prophet()
model.fit(df_prophet)
forecast = model.predict(future)
```

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

Premium

bhavya ▼

Time Series

Prophet vs ARIMA

What is Prophet?

Prophet is a time series forecasting tool by Facebook that models data using trend, seasonality, an events. It’s easy to use, requires minimal tuning, and handles outliers and missing values well.

Why is it good for Spotify data?

Spotify data has strong daily and weekly patterns, occasional spikes (e.g., album drops), and some gaps. Prophet captures these patterns and handles irregularities effectively.

Why is it better than ARIMA?

Prophet supports multiple seasonality's, is more robust to noise, and easier to configure. ARIMA needs clean, stationary data and complex tuning—making Prophet more practical for Spotify's dynamic usage data.

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

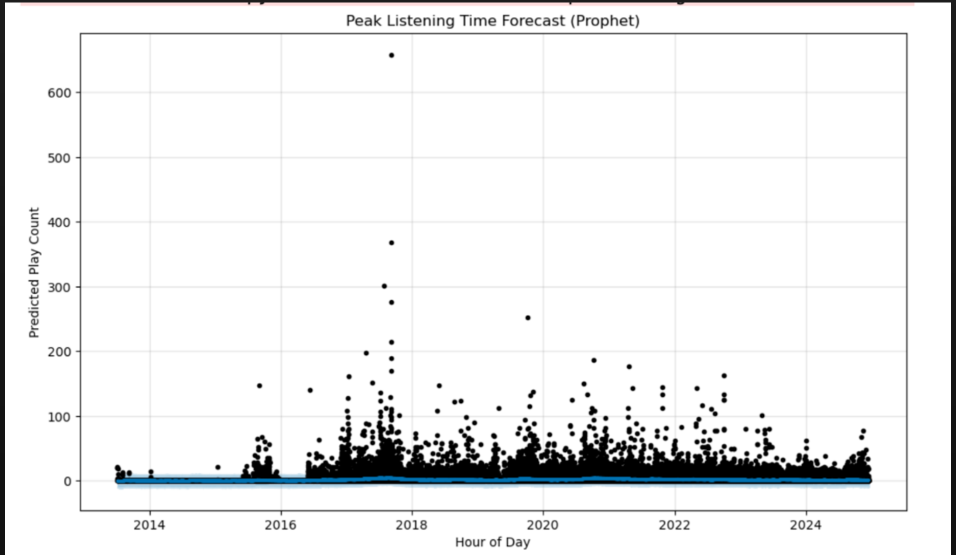
Contact Us

Premium

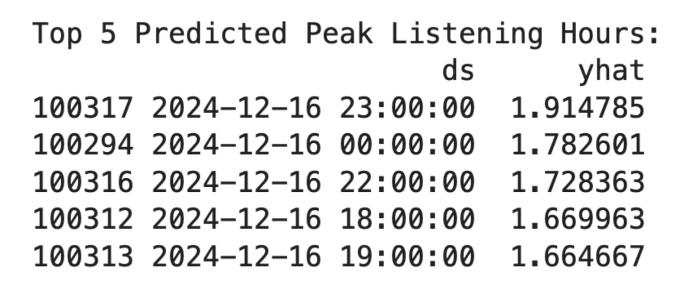
bhavya

Time Series

When do people listen?



- Peak Hours: Highest engagement observed between 18:00–23:00 on Dec 16, 2024.
- Behavioural Trends: Overall listening shows a gradual increase, with fluctuations and spikes tied to special events.
- Business Impact: Platforms can optimize ads and recommendations during high engagement periods.



```
Top 5 Predicted Peak Listening Hours:
                        ds       yhat
100317 2024-12-16 23:00:00  1.914785
100294 2024-12-16 00:00:00  1.782601
100316 2024-12-16 22:00:00  1.728363
100312 2024-12-16 18:00:00  1.669963
100313 2024-12-16 19:00:00  1.664667
```

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

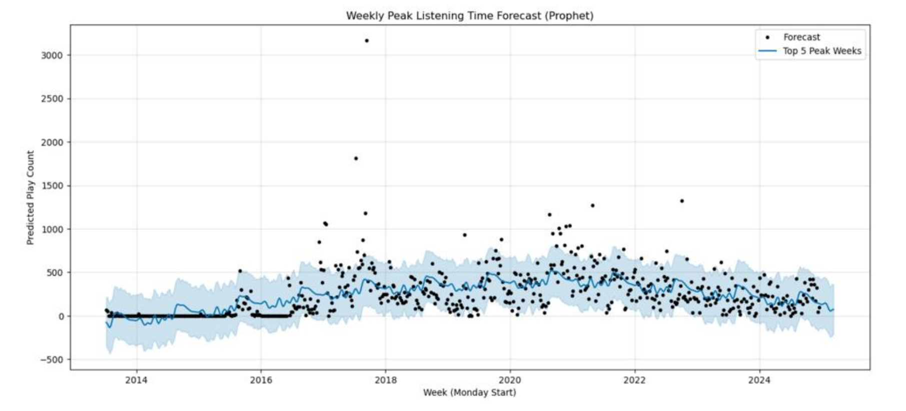
Premium

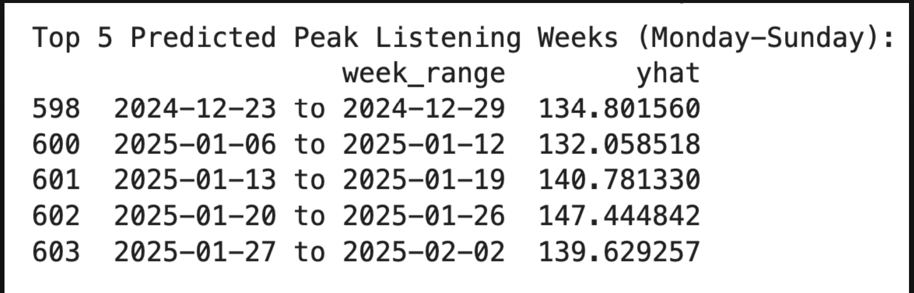
bhavya ▼

Time Series

When do people listen?

- Prophet model predicts weekly listening trends from 2013 to 2025.
- Listening peaked in early 2025, with the highest during Jan 20–26, 2025.
- Engagement gradually increased until 2021, then fluctuated with spikes.
- Top 5 peak weeks occur between Dec 2024 and Feb 2025.
- Ideal for launching promotions, ads, or exclusive content during these periods.





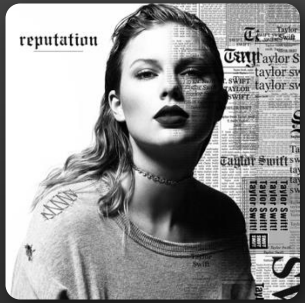
```
Top 5 Predicted Peak Listening Weeks (Monday–Sunday):
                      week_range          yhat
598  2024-12-23 to 2024-12-29  134.801560
600  2025-01-06 to 2025-01-12  132.058518
601  2025-01-13 to 2025-01-19  140.781330
602  2025-01-20 to 2025-01-26  147.444842
603  2025-01-27 to 2025-02-02  139.629257
```

- Welcome
- About Us
- Our Teams
- Agenda
- Our Services

- Achievements
- Our Goals
- Gallery
- Pricing
- Contact Us

Premium

bhavya

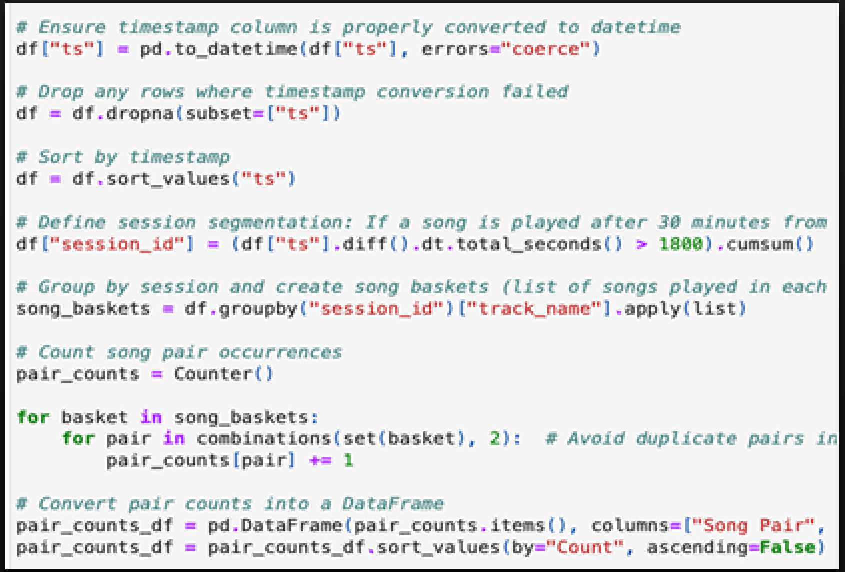


Association

What songs go together?

We used association rule mining to identify which songs are frequently played together in the same session. These frequent pairs were then used to build playlists that reflect real user behaviour — not just genre or artist preference.

- Created session-level song baskets
- Identified frequent song pairs using combinations
- Calculated support, lift, and confidence
- Applied a diversity rule to avoid same-artist repeats



```
# Ensure timestamp column is properly converted to datetime
df["ts"] = pd.to_datetime(df["ts"], errors="coerce")

# Drop any rows where timestamp conversion failed
df = df.dropna(subset=["ts"])

# Sort by timestamp
df = df.sort_values("ts")

# Define session segmentation: If a song is played after 30 minutes from
df["session_id"] = (df["ts"].diff().dt.total_seconds() > 1800).cumsum()

# Group by session and create song baskets (list of songs played in each
song_baskets = df.groupby("session_id")["track_name"].apply(list)

# Count song pair occurrences
pair_counts = Counter()

for basket in song_baskets:
    for pair in combinations(set(basket), 2):  # Avoid duplicate pairs in
        pair_counts[pair] += 1

# Convert pair counts into a DataFrame
pair_counts_df = pd.DataFrame(pair_counts.items(), columns=["Song Pair",
pair_counts_df = pair_counts_df.sort_values(by="Count", ascending=False)
```

Welcome

About Us

Our Teams

Agenda

Our Services

Achievements

Our Goals

Gallery

Pricing

Contact Us

Premium

bhavya ▼

Played together

Top 5

#	Song pair	Times played together	🕒
1	Nuovo Cinema Paradiso + Claudia’s Theme	70	06:46
2	Mia & Sebastian’s Theme + Engagement Party	68	07:33
3	The Road Goes Ever On... + The Return of the King	63	08:23
4	The Breaking of the Fellowship + In Dreams	61	07:55
5	VoluptatesTotò e Alfredo (v2) + Tema D’Amore Per Nata	59	06:58

Spotify

Generated Playlist

- Nuovo Cinema Paradiso

The Great Eye

Raglan Road

Birds

Bittersweet

Craving – Acoustic Version

No Such Thing

Romeo And Jullet

Shake Your Hips

Bell Bottom Blues

Desolation Row

① Old Brown Shoe- Remastered 2009

① One Of These Days - Remastered 2011

Crawl

I Don't Care (with Justin Bieber)

Culpable

Waitin'On The Day

The Gun

Give A Little Bit

Hot House of Omagarashid

Curated to match every mood – one skipless stream at a time.

- Welcome
- About Us
- Our Teams
- Agenda
- Our Services

- Achievements
- Our Goals
- Gallery
- Pricing
- Contact Us

Premium

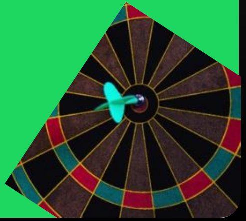
bhavya ▼

Recommendation

Smarter recommendations

Use real-time skip prediction to improve track suggestions and reduce user disengagement.

Listener segmentation



Cluster listeners into behavioural groups to personalize content, promotions, and features.

Optimised release timing

Use time series forecasts to identify peak listening hours and schedule new releases accordingly.

Playlist generation

Generate organic playlists from frequently co-played songs without relying on genres or likes.

That’s a wrap!

Curated by behaviour. Powered by data. Streamed by everyone.