Project Report: Predicting User Subscription Behavior in an EdTech Startup Using Classification Models

Date: 5th August, 2025
To: Stakeholders
From: Shijin Ramesh, Data Analyst
---

1. Introduction

Title: Predicting User Subscription Behavior in an EdTech Startup Using Classification Models
Business Question: Can we predict what type of subscription plan a user is most likely to choose based on their learning intent, dancer type, location, and other factors?

The goal of this project is to explore whether it is possible to predict the type of subscription plan a user is likely to choose on the IDALS platform. IDALS is an Indian dance-based ed-tech startup offering various subscription plans to learners based on their goals, experience level, and preferences.

Understanding the likelihood of a user selecting a particular plan helps the business tailor marketing strategies, optimize pricing, and personalize user experience.

2. Data Description

The dataset consists of 14,842 user profiles who signed up on the IDALS platform. It includes features like:

- subscription_plan – Target variable with 5 classes (Free, Weekly, Monthly, Half-Yearly and Yearly)
- type_of_dancer – Experience level (Beginner, Intermediate, Advance)
- genres_of_interest – Preferred genres (Single or Multi genre)
- purpose_of_learning – Motivation behind learning (Passion, Career, Hobby)
- future_aspirations – Career goals (Choreographer, Freelancer, etc.)
- repeated_subscriber – Indicates if the user has subscribed before (extracted)
- signup_month, signup_day, signup_hour – Temporal features extracted from timestamp

3. Data Cleaning (Performed in Excel & Snowflake)

- Removed unnecessary columns
- More than 2000 records of gibberish data were removed
- Country column was standardized as using title case and whitespace stripping. Many values were correctly mapped due to incorrect data entry by the user while registering on the website.

- Derived a new column of full_name which was created merging first_name and last_name.
- New columns like subscribed, subscriptions and repeated_subscriber were created for better analysis.
- **plan.name** was renamed as **subscription_plan** and values were changed for better understanding:
    - **Basic – Free**
    - **Full Access 7 – Weekly**
    - **Full Access 30 – Monthly**
    - **Full Access 180 – Half Yearly**
    - **Full Access 365 – Yearly**

SAMPLE SQL QUERY

// 1. Merging first_name and last_name

```
ALTER TABLE
  idalsuserdata
ADD
  COLUMN full_name STRING;
UPDATE
  idalsuserdata
SET
  full_name = CONCAT(first_name, ' ', last_name) // Dropping first_name and last_name
ALTER TABLE
  idalsuserdata DROP COLUMN first_name;
ALTER TABLE
  idalsuserdata DROP COLUMN last_name;
```

// 2. Identifying gibberish names

```
UPDATE
  idalsuserdatanew
SET
  is_gibberish = TRUE
WHERE
  full_name_cleaned IS NOT NULL
  AND (
    -- Very long single words (often fake)
    REGEXP_LIKE(full_name_cleaned, '\\b\\w{13,}\\b')
    OR -- High uppercase ratio (junk names often have too many caps)
    LENGTH(REGEXP_REPLACE(full_name_cleaned, '[^A-Z]', '')) > 6
    AND LENGTH(full_name_cleaned) <= 25
    OR -- Low vowel count (e.g., less than 2 vowels in long name)
```

```
    (
      LENGTH(
        REGEXP_REPLACE(full_name_cleaned, '[^aeiouAEIOU]', '')
      ) < 2
      AND LENGTH(full_name_cleaned) >= 10
    )
    OR -- Names with no space (looks like gibberish if both parts are fake)
    NOT full_name_cleaned LIKE '% %'
    AND LENGTH(full_name_cleaned) > 14
  );
```

// 3. Standardize Country Names

```
UPDATE
  idalsuserdatanew
SET
  country = 'India'
WHERE
  LOWER(TRIM(country)) IN (
    'india',
    'india.',
    'india ',
    'indai',
    'ind',
    'indai.',
    'in',
    'bharat'
  );
```

// 4. Updating the subscriptions column

```
UPDATE
  idalsuserdatanew u
SET
  subscriptions = s.sub_count
FROM
  (
    SELECT
      email,
      COUNT(*) AS sub_count
    FROM
      subscribedusers
```

```
    GROUP BY
      email
  ) s
WHERE
  TRIM(LOWER(u.email)) = TRIM(LOWER(s.email));
```

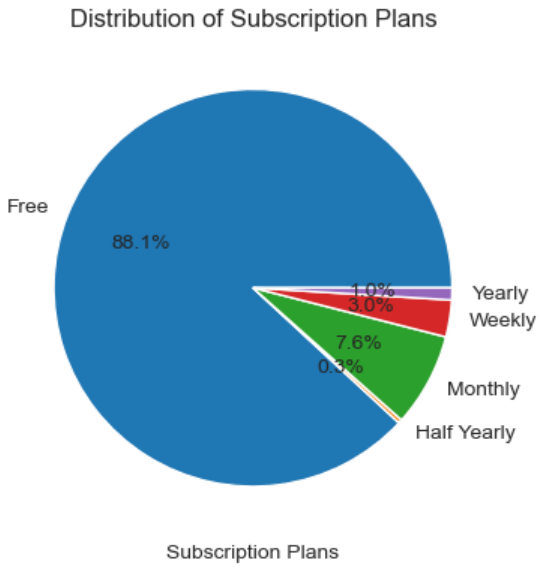
// 5. Updating the column of repeated_subscriber

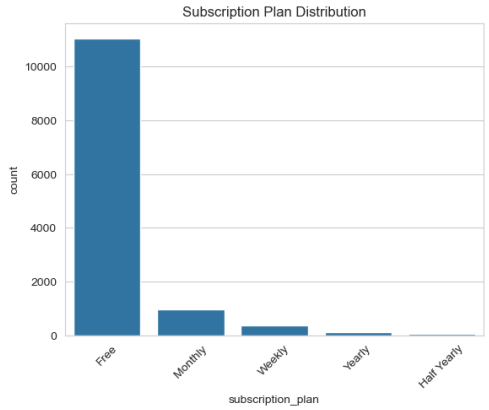
```
UPDATE
  idalsuserdatanew
SET
  repeated_subscriber = CASE
    WHEN subscriptions > 1 THEN 'Yes'
    WHEN subscriptions = 1 THEN 'No'
    ELSE 'No'
  END;
```

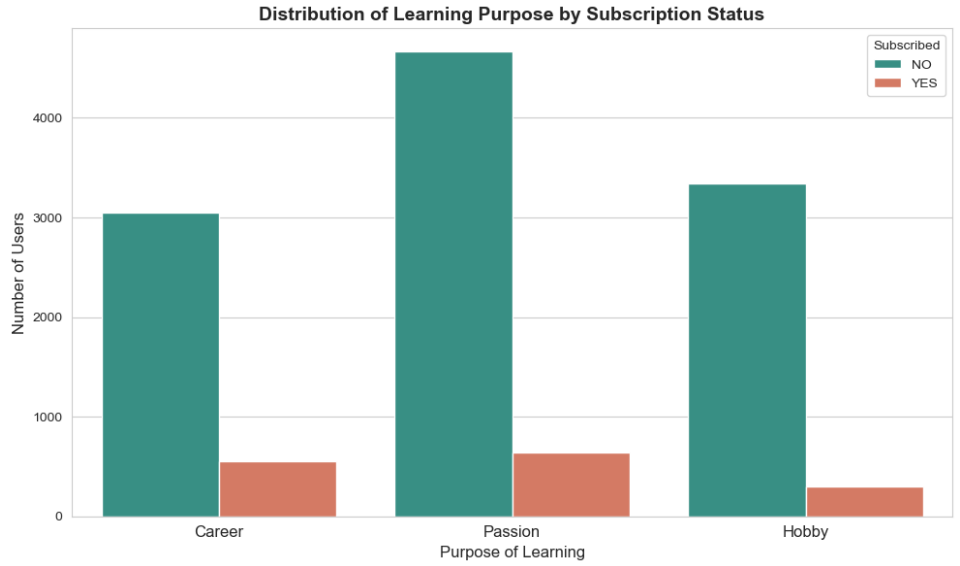
PLAN_EXPIRES_AT	LOCATION	COUNTRY	CREATED_AT	TYPE_OF_DANCER	GENRES_OF_INTEREST	PURPOSE_OF_LEARNING	FUTURE_ASPIRATIONS	FULL_NAME	SUBSCRIBED	SUBSCRIPTIONS	REPEATED_SUBSCRIBEI
null	Bharuch, gujarat	India	1/1/2021 10:11	Advance	Multi Genre	Career	Freelancer or Professional Dancer	Viraj Suthar	NO	null	No
null	Thane	India	1/1/2021 10:20	Beginner	Multi Genre	Career	Freelancer or Professional Dancer	Prarthana Pawar	NO	null	No
null	Surat	India	1/1/2021 10:39	Intermediate	Single Genre	Passion	Others	Sandip Pariyar	NO	null	No
null	Korba Chhattisgarh	India	1/1/2021 10:39	Intermediate	Multi Genre	Hobby	Choreographer	Harish Kanwar	NO	null	No
null	Lonavla	India	1/1/2021 10:53	Intermediate	Single Genre	Hobby	Choreographer	Tanuj Mehta	NO	null	No
null	Delhi	India	1/1/2021 10:55	Advance	Single Genre	Hobby	Art/Dance Director	Ankit Singh	NO	null	No
null	Delhi	India	1/1/2021 10:59	Advance	Single Genre	Passion	Choreographer	Ankit Singh	NO	null	No
null	Rajasthan	India	1/1/2021 11:16	Beginner	Single Genre	Passion	Instructor	Sachin Sharma	NO	null	No
null	Trivandrum	India	1/1/2021 11:28	Beginner	Multi Genre	Hobby	Choreographer	Ajay Sm	NO	null	No
null	Lucknow	India	1/1/2021 13:27	Intermediate	Multi Genre	Passion	Choreographer	Kumar Shivam	NO	null	No
null	Kota	India	1/1/2021 14:58	Advance	Multi Genre	Career	Instructor	Umesh Sharma	NO	null	No
null	Kuchamn	India	1/1/2021 15:27	Beginner	Single Genre	Career	Freelancer or Professional Dancer	Ricky Soni	NO	null	No
null	Kharagpur	India	1/1/2021 16:27	Intermediate	Single Genre	Career	Choreographer	Vinod Honey	NO	null	No
null	Keonjhar	India	1/1/2021 16:46	Intermediate	Multi Genre	Hobby	Freelancer or Professional Dancer	Gogo Op	NO	null	No
3/23/2023 10:32	Bulandshahr	India	1/1/2021 17:29	Intermediate	Multi Genre	Career	Choreographer	Sanju Teotia	YES	2	Yes
null	Hyderabad	India	1/1/2021 17:35	Intermediate	Single Genre	Passion	Choreographer	Utkarsh Nilegaonk	NO	null	No
null	churu	India	1/1/2021 19:31	Intermediate	Multi Genre	Hobby	Choreographer	Rohit Verma	NO	null	No
null	Surat	India	1/1/2021 19:52	Beginner	Multi Genre	Career	Freelancer or Professional Dancer	Venus Choudhary	NO	null	No
null	Sindhanoor	India	1/1/2021 21:07	Beginner	Single Genre	Passion	Freelancer or Professional Dancer	Nagaraj R	NO	null	No
null	Dhaka	BANGLADESH	1/1/2021 21:26	Intermediate	Multi Genre	Career	Choreographer	Rakib Ahmed Tom	NO	null	No
null	Pune	India	1/1/2021 21:27	Intermediate	Single Genre	Hobby	Freelancer or Professional Dancer	Yash Jain	NO	null	No
null	mlshrlkh	India	1/1/2021 22:14	Intermediate	Single Genre	Hobby	Freelancer or Professional Dancer	Akash Shukla	NO	null	No
null	MATHURA	India	1/1/2021 22:18	Beginner	Multi Genre	Passion	Choreographer	Poornima Gupta	NO	null	No
null	Bangalore	India	1/1/2021 22:21	Intermediate	Multi Genre	Hobby	Choreographer	Anand Pandey	NO	null	No
null	Buldhana	India	1/1/2021 22:21	Beginner	Multi Genre	Passion	Others	Vinayak Gawali	NO	null	No
null	Atlanta	UNITED STATES	1/1/2021 22:33	Beginner	Multi Genre	Hobby	Freelancer or Professional Dancer	Arjun Sharma	NO	null	No
null	Bareilly	India	1/1/2021 22:37	Intermediate	Single Genre	Career	Choreographer	Aman Txz	NO	null	No
null	Indore	India	1/1/2021 23:19	Intermediate	Multi Genre	Passion	Choreographer	Harshdeep Goyal	NO	null	No
2/2/2021 0:05	Pune	India	1/1/2021 23:54	Intermediate	Single Genre	Passion	Freelancer or Professional Dancer	Onkar More	YES	1	No

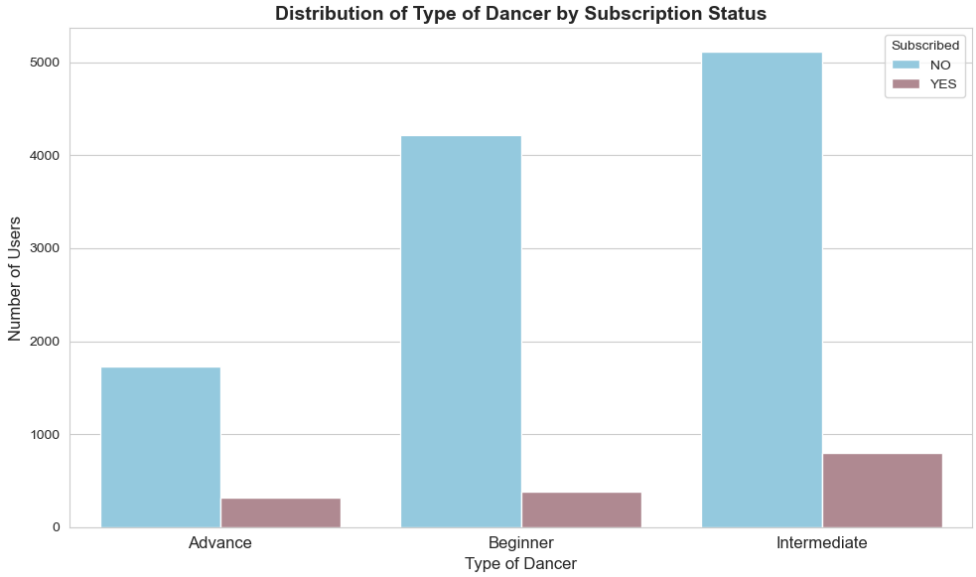
3. Exploratory Data Analysis and Key Patterns

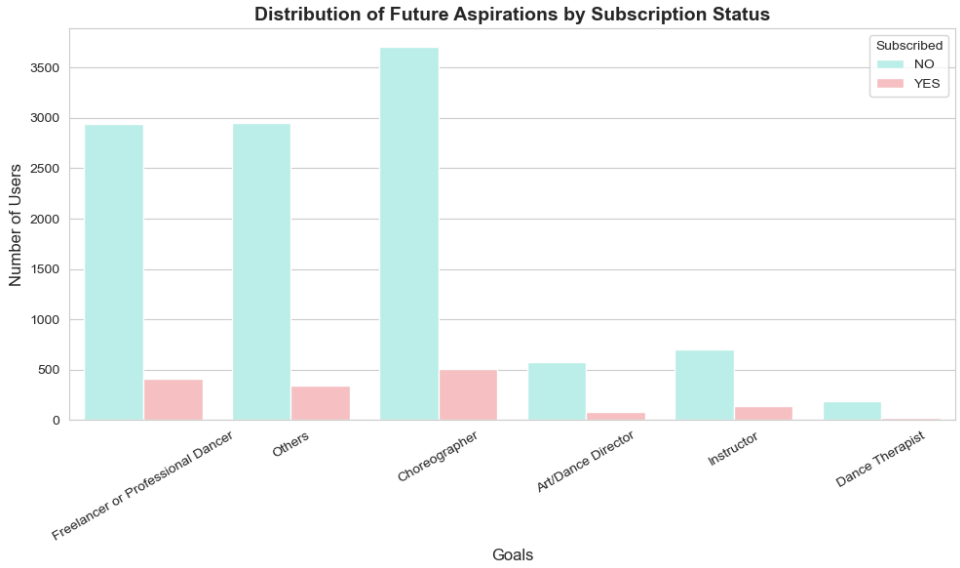
Majority of the users were under Free plan (~88%)

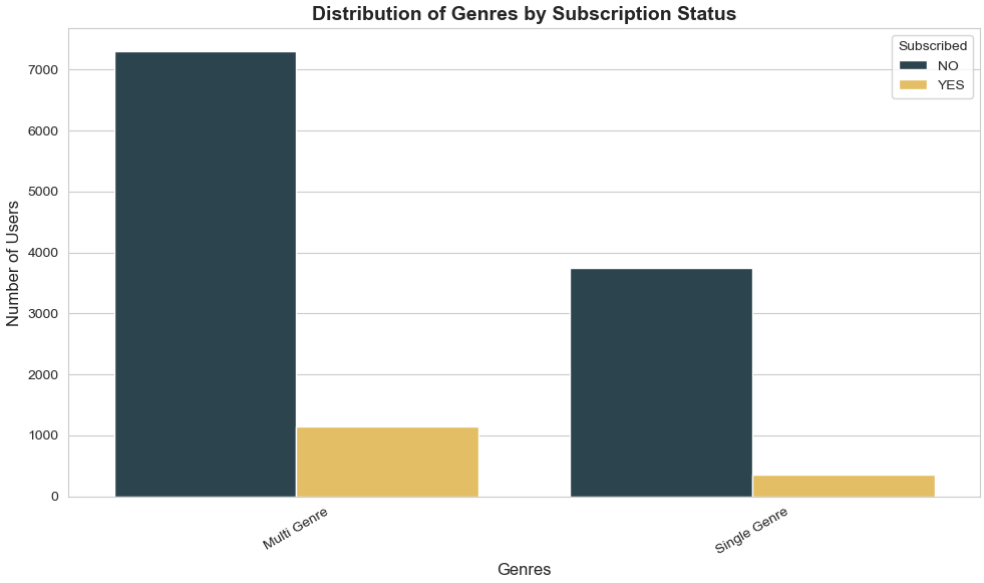

Distribution of Subscription Plans
Free
88.1%
1.0%
3.0%
7.6%
0.3%
Yearly
Weekly
Monthly
Half Yearly
Subscription Plans

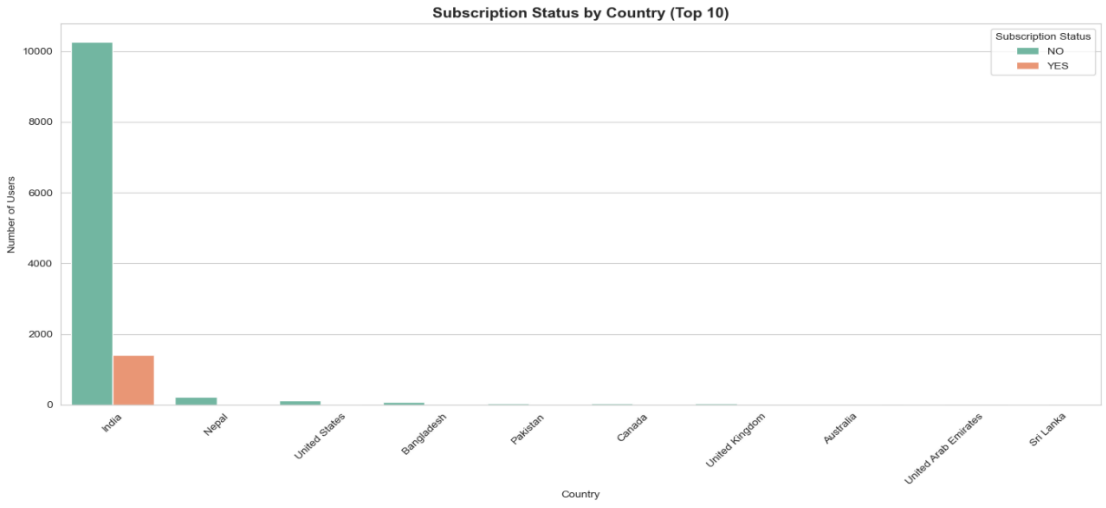

Subscription Plan Distribution
count
10000
8000
6000
4000
2000
0
Free
Monthly
Weekly
Yearly
Half Yearly
subscription_plan

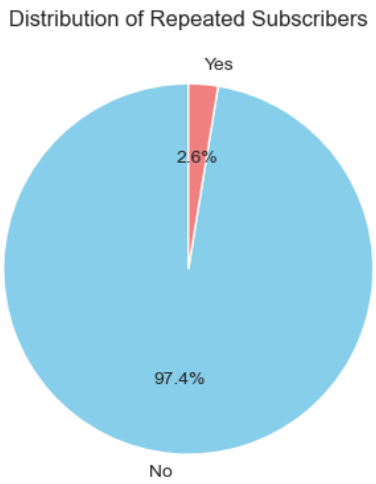

Distribution of Learning Purpose by Subscription Status
Subscribed
NO
YES
Number of Users
0
1000
2000
3000
4000
Career
Passion
Hobby
Purpose of Learning


Distribution of Type of Dancer by Subscription Status
Subscribed
NO
YES
Number of Users
0
1000
2000
3000
4000
5000
Advance
Beginner
Intermediate
Type of Dancer

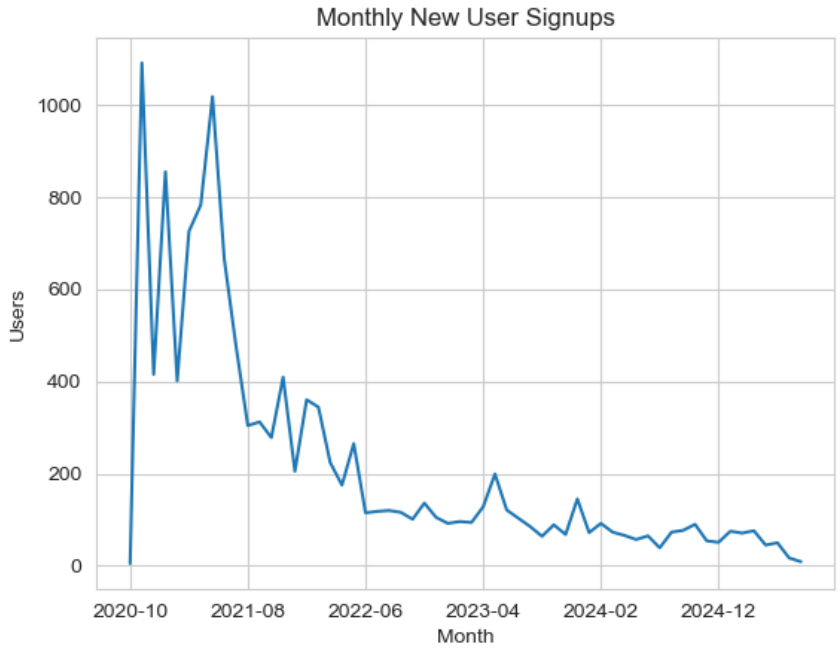

Distribution of Future Aspirations by Subscription Status
Subscribed
NO
YES
Number of Users
0
500
1000
1500
2000
2500
3000
3500
Freelancer or Professional Dancer
Others
Choreographer
Art/Dance Director
Instructor
Dance Therapist
Goals

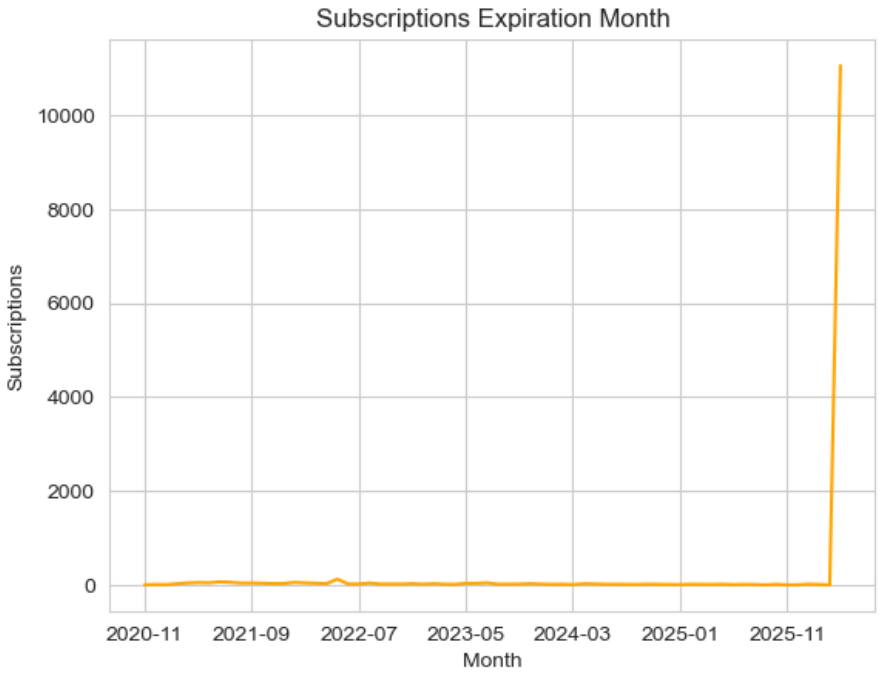

Distribution of Genres by Subscription Status
Subscribed
NO
YES
Number of Users
0
1000
2000
3000
4000
5000
6000
7000
Multi Genre
Single Genre
Genres

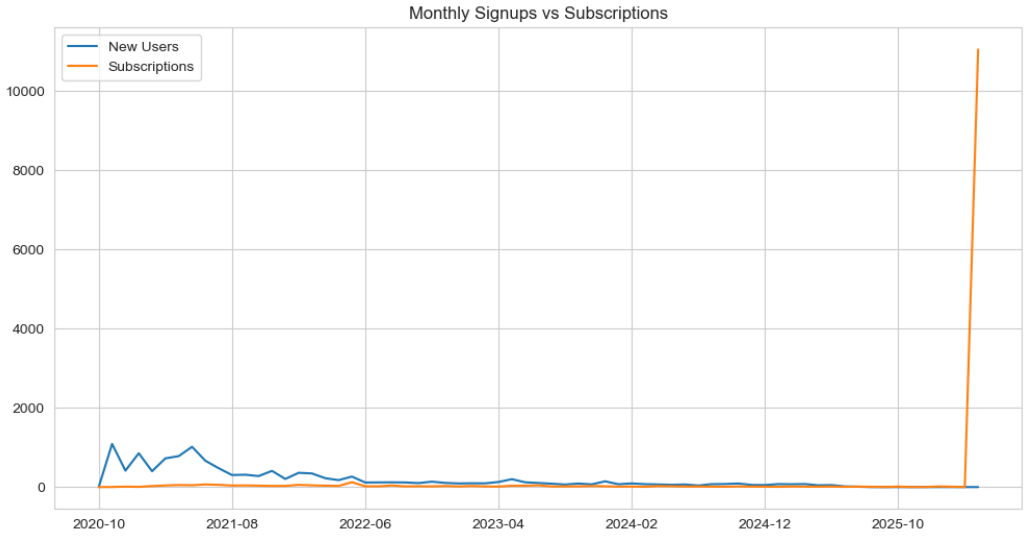

Subscription Status by Country (Top 10)
Subscription Status
NO
YES
Number of Users
0
2000
4000
6000
8000
10000
India
Nepal
United States
Bangladesh
Pakistan
Canada
United Kingdom
Australia
United Arab Emirates
Sri Lanka
Country

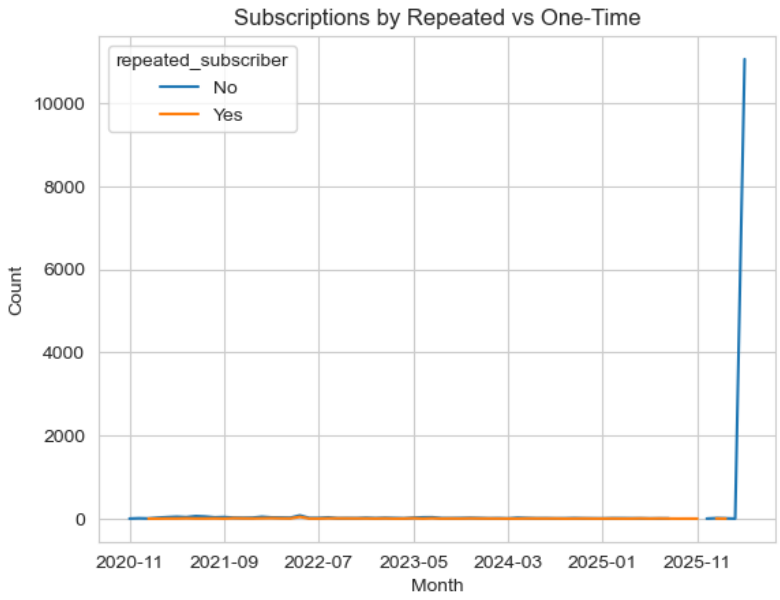


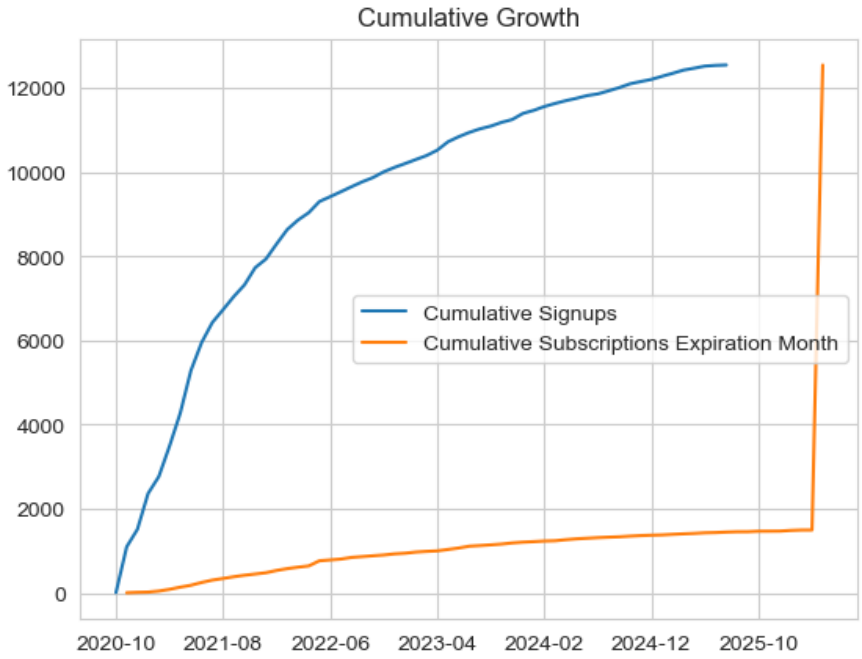
Time-Series Analysis




Subscriptions Expiration Month
Subscriptions
10000
8000
6000
4000
2000
0
2020-11
2021-09
2022-07
2023-05
2024-03
2025-01
2025-11
Month


Monthly Signups vs Subscriptions
New Users
Subscriptions
10000
8000
6000
4000
2000
0
2020-10
2021-08
2022-06
2023-04
2024-02
2024-12
2025-10





Key Observations:

- The data was cleaned and prepared by combining user information with subscription details, and new columns like subscriptions and repeated_subscriber were added to support deeper analysis.
- The insights showed that more than 90% of the users were from India.
- A significantly high number of intermediate-level dancers indicated that intermediate-level dancers are more inclined to invest in subscriptions to refine and elevate their skills further as artists.

- The data also showed that users with career or passion-driven motivations are more likely to invest in a subscription, aligning their intent with actionable commitment.
- More than 4000 users wished to become a Choreographer followed by the career option of Freelancer or Professional Dancer.
- 2.6% of the users were repeated subscribers and 1057 users had subscribed only once.
- Time-series analysis showed that most user sign-ups occurred during the platform's launch in late 2020, likely due to the nationwide lockdown that encouraged online learning.
- A steady decline in sign-ups began after mid-2021, which aligns with the reopening of offline classes. This suggests that many users may have returned to physical dance studios or the platform faced challenges in user acquisition.
- A sharp increase in subscription expirations around November 2025 was observed, possibly due to a promotional annual plan offered during the startup's anniversary. This spike represents the end of subscription periods, not the beginning.
- Overall, while the platform gained strong initial traction, user retention and growth appear to have slowed over time, highlighting the need for improved engagement or marketing strategies.

4. Modeling & Evaluation

Models Used:

- **Logistic Regression**
- **Random Forest**
- **XGBoost Classifier**

Problem:

The target variable was highly imbalanced with the **Free** plan dominating the data.

Baseline Model (Logistic Regression)

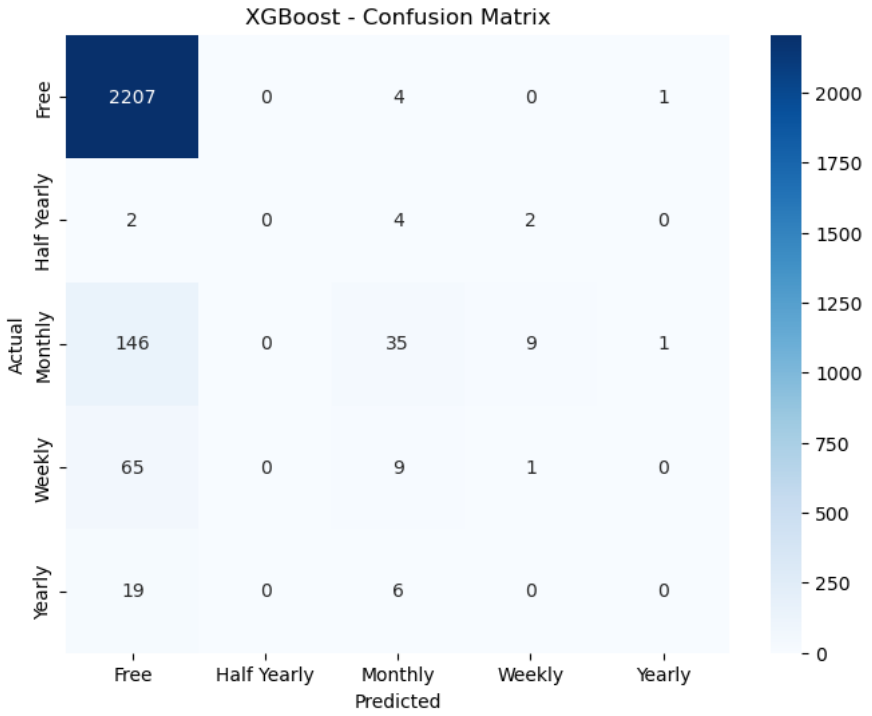
- The model achieved an overall accuracy of ~46%, heavily skewed toward predicting the dominant class "Free", which accounts for nearly 88% of the dataset.
- Confusion Matrix showed most predictions were biased toward **Free**.
- Other classes (Weekly, Monthly, Yearly) had poor precision and recall.

Improved Model (Random Forest)

- The model achieved a high overall accuracy of ~88.5%, driven primarily by strong performance in predicting the 'Free' subscription class.
- Better generalization, but still biased toward majority class.

Final Model (XGBoost Classifier)

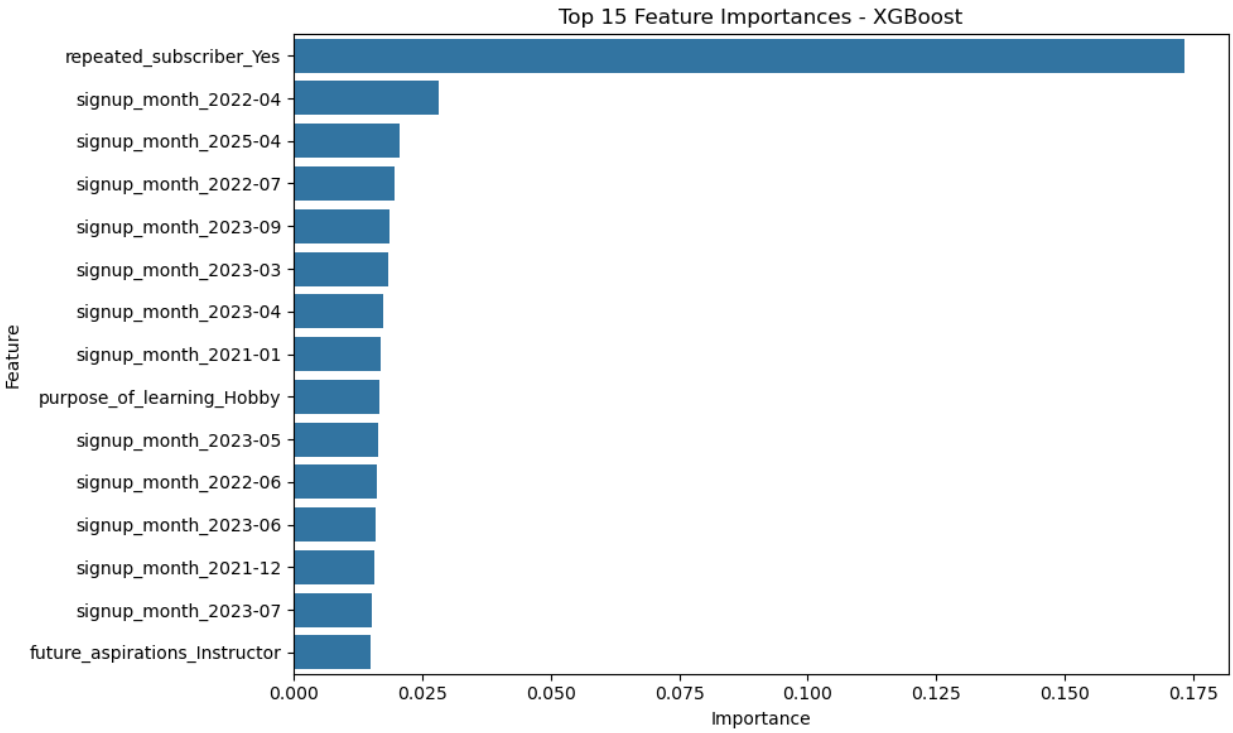
- The accuracy score is ~89.32%, which seems high, but this is misleading due to class imbalance.
- Class 0 (Free) has: Precision: 0.90, Recall: 1.00, F1-Score: 0.95 indicating model is excellent in predicting Free class.
- All other classes (Half Yearly, Monthly, Weekly, Yearly) have: Very low recall (model missed most of them) & F1-scores near zero, indicating poor predictive ability.
- XGBoost is heavily biased towards the Free plan (majority class). This is because dataset is imbalanced towards Free plans and the features don't strongly differentiate between subscription types.



5. Feature Importance

Top predictors from Random Forest:

- Repeated_subscriber



10. Conclusion & Recommendations

- Despite exploring advanced classifiers and improving accuracy, all models were heavily biased toward the majority class — the Free subscription.
- This suggests that the predictor variables do not hold strong discriminatory power to distinguish between all 5 subscription categories.
- There's a need to collect more actionable user data, such as:
  1. Interaction behavior (time spent, classes viewed)
  2. Payment method or pricing sensitivity
  3. User reviews or feedback

Recommendation for Stakeholders:

A. **Subscription Skew Toward Free Users**
   - The majority of users opt for the Free plan, with minimal engagement in paid plans like Yearly or Half-Yearly.
   - This indicates a low conversion rate from free to paid users.
   - Introduce targeted upselling strategies (e.g., limited-time offers, premium content previews, or bundled discounts) to encourage upgrades.

B. **Low Predictive Power of Existing Features**
   - The current user profile features (e.g., dancer type, aspirations, genre interest) are not strong predictors of subscription level.
   - Collect additional data points such as: Engagement metrics (e.g., time spent learning, class completion rate), Demographic data (e.g., age) and User Behavior History (e.g., login frequency, video views)

C. **Rethink User Segmentation Strategy**
   - While current user segmentation helps identify aspirants (e.g., Career-focused, Freelancers), it doesn't correlate strongly with paid conversions.
   - Run a detailed customer segmentation analysis combining learning goals, time of signup, and future aspirations to tailor marketing campaigns.

D. **Feature Insights (from XGBoost Importance)**
   - Features like signup_month, repeated_subscriber and purpose_of_learning had relatively higher importance**.**
   - Use these features to trigger personalized retention campaigns. For instance, returning users (repeated_subscriber = Yes) can be offered loyalty benefits.

E. **Need for Continuous Feedback Loop**
   - The model and insights are constrained by limited feature variety and data volume.
   - Establish a system for continuous data collection and model retraining every quarter to improve targeting accuracy.