Darvix
DATA ANALYTICS, RESEARCH AND VISUALISATION AT XLRI

Strategix 2.0
Decode. Decide. Deliver.

StreamMax OTT Platform

BACKGROUND

StreamMax is a fast-growing over-the-top (OTT) video streaming platform operating in a highly competitive market. The company has millions of active subscribers and continues to expand its content library. However, like all streaming services, StreamMax faces a critical business challenge: users are cancelling subscriptions, and by the time the platform identifies them, it's often too late to intervene.

Traditional churn prediction models flag users only after they've stopped using the platform entirely. At that point, re-engagement becomes extremely difficult and costly. StreamMax's business intelligence team has been studying this problem and believes there's a better approach.

The company wants to move from reactive to proactive retention strategies. Instead of waiting for users to churn completely, StreamMax aims to identify "engagement fatigue" early. These are users whose interaction patterns are declining and who show warning signs of future disengagement. If caught early, targeted interventions like personalized content recommendations, special offers, or new feature rollouts can prevent them from leaving.

YOUR CHALLENGE

You have been brought in as analytics consulting team to help StreamMax solve this problem. Your task is twofold: build a predictive model that accurately identifies at-risk users, and develop practical business strategies that the company can implement to retain them.

DATASET OVERVIEW

You have been provided with user behavior data from StreamMax's platform covering a recent snapshot period. The dataset includes behavioral metrics, content consumption patterns, and engagement indicators for 10,000 users.

Files Provided

1. Training Dataset (ott_train.csv): 8,000 users

- Complete user profiles with engagement metrics
- Includes the target variable: fatigue_label (0 = Engaged, 1 = At Risk of Disengagement)
- Use this data to build and validate your predictive model

2. Test Dataset (ott_test.csv): 2,000 users

- User profiles identical in structure to training data
- Does NOT include the target variable (fatigue_label)
- You must predict whether each user is at risk of disengagement

DATA DICTIONARY

User Identification & Account Information

- ***user_id*** **:** A unique text identifier for each user in the dataset. You'll use this to match your predictions with the test set.
- ***tenure_days*** **:** The number of days since the user first subscribed to StreamMax. Values range from 7 days (very new users) to 1,095 days (three-year subscribers). This tells you how long someone has been a customer.
- ***subscription_tier*** **:** The user's current subscription plan. There are three options: Basic (entry-level), Standard (mid-tier), and Premium (top-tier with features like 4K streaming and multiple screens).

- ***avg_daily_minutes_last_7d*** : The average number of minutes per day this user spent watching content over the past week. Values range from 0 (didn't watch anything) to 300 minutes (5 hours per day). This shows you their current level of platform usage.
- ***sessions_last_7d*** : How many separate times the user opened the app or website to watch content in the past week. A session starts when they begin watching and ends after 30 minutes of inactivity. Values range from 0 to 21 sessions.
- ***avg_daily_minutes_last_30d*** : The average number of minutes per day this user watched content over the past month. Values range from 0 to 300 minutes.
- ***sessions_last_30d*** : The total number of viewing sessions over the past 30 days. Values range from 0 to 140 sessions. This helps you understand their typical viewing patterns.
- ***avg_completion_rate*** : What percentage of content does this user typically finish watching? For example, 0.75 means they watch 75% of a movie or show on average before stopping. Values range from 0.0 to 1.0. This metric reflects content satisfaction.
- ***unique_genres_watched_30d*** : How many different content genres (Action, Comedy, Drama, Documentary, etc.) did this user watch in the past month? Values range from 1 to 15. StreamMax offers 15 total genres.
- ***days_since_last_session*** : How many days has it been since this user last logged into StreamMax? Values range from 0 (they watched today) to 30 (haven't logged in for a month). This is a critical recency metric.
- ***binge_sessions_last_30d*** : How many times in the past month did this user binge-watch content? We define a binge session as watching 3 or more episodes in a row or spending 2+ continuous hours watching. Values range from 0 to 30.
- ***peak_hour_viewing_pct*** : What percentage of this user's total watch time happens during peak evening hours (7:00 PM to 11:00 PM)? Values range from 0 to 100. For example, 80 means 80% of their viewing is in the evening.

- ***original_content_pct*** : What percentage of watch time does this user spend on StreamMax original productions versus content licensed from other studios? Values range from 0 to 100. Original content is exclusive to StreamMax.
- ***recommendation_click_rate*** : When StreamMax's algorithm recommends content to this user, what percentage do they actually click on and watch? Values range from 0.0 to 1.0. For example, 0.25 means they watch 25% of recommended content

Target Variable (Training Data Only)

- ***fatigue_label***

This is what you need to predict:

- ***0 = Engaged:*** The user is maintaining healthy platform usage and is not at risk
- ***1 = Fatigued:*** The user shows declining engagement patterns and is at high risk of reducing usage or cancelling within the next 30 days

This label appears only in the training dataset. Your task is to predict it for all users in the test dataset.

DELIVERABLES

Submit the following 3 components by the competition deadline:

1) Prediction File (MANDATORY)

Format: csv file

Filename: TeamName_Predictions.csv

Required Structure (exactly 2,000 rows):

user_id	predicted_fatigue_probability
U000001	0.561
U000002	0.234
U000003	0.765
...	...

Requirements:

- ***user_id*** **:** Must exactly match test dataset IDs
- ***predicted_fatigue_probability*** **:** should be a number between 0.0 and 1.0 representing the probability that each user is at risk.
- **0.0 =** Very low risk (highly engaged)
- **1.0 =** Very high risk (severely fatigued)
- No missing values, all 2,000 users included

2. Python Notebook / Script (MANDATORY)

Format: Jupyter Notebook (.ipynb)

Filename: TeamName_Analysis.ipynb

Must include:

- **Data understanding and preprocessing :** Show how you explored the dataset, handled any data quality issues, and prepared it for modelling. Include summary statistics, visualizations, and any transformations you applied.

- **Feature analysis and engineering :** Derived features, correlations, selection rationale
- **Model development :** Selection process, training approach, validation metrics
- **Final predictions :** Clear code generating test set predictions and export csv

3) Presentation (MANDATORY)

Format: PowerPoint (.pptx) or PDF (.pdf)

Length: 4–6 slides

Slides can include:

- **Problem Understanding :** Business challenge and analytical approach
- **Key Insights :** Compelling patterns differentiating engaged vs. fatigued users (2-3 visualizations)
- **Modelling Approach :** Model selection, key features, performance metrics (in business terms)
- **Business Interpretation :** User segments at risk, factors predicting disengagement
- **Strategic Recommendations :** Specific, actionable strategies (targeting, interventions, product changes)
- **Implementation Roadmap :** Deployment strategy, monitoring, limitations

EVALUATION CRITERIA

1) Technical Performance (50%)

Your prediction file will be evaluated using following metric:

- **AUC-ROC:** This measures how well your model ranks users by risk. This is the main metric used for leaderboard ranking of AI model

2) Business Strategy (50%)

SUBMISSION INSTRUCTIONS

Required Files:

- TeamName_Predictions.csv
- TeamName_Analysis.ipynb
- TeamName_Presentation.pptx or .pdf

Replace "TeamName" with your actual team’s name (no spaces, use underscores).

FREQUENTLY ASKED QUESTIONS

Q: Can we use external data?
A: No. Use only the provided datasets.

Q: Should we predict 0/1 labels or probabilities?
A: Submit probabilities (values b/w 0 and 1), not binary labels

Q: Will we present to judges?
A: Yes. Finalists will present to a panel

Darvix

DATA ANALYTICS, RESEARCH AND VISUALISATION AT XLRI

StrategiX 2.0

Decode. Decide. Deliver.

GOOD LUCK!

WE LOOK FORWARD TO SEEING YOUR INNOVATIVE APPROACHES TO SOLVING STREAMMAX'S ENGAGEMENT CHALLENGE.