Customer Behavior Analysis

for Subscription Retention and Value Estimation

Course: [01276343] ML Project for Case Analysis
Topic: Numerical Forecasting & Categorical Prediction

OFFICIAL PROJECT HANDBOOK
Technical Specification, Role Definitions & Grading Strategy

Prepared by:
Lead Data & AI Infrastructure Engineer

Date: February 3, 2026

Abstract

This handbook details the architectural and theoretical framework for using the **Alibaba User Behavior Dataset** to simulate a Subscription Retention system. It maps the complex e-commerce interaction data to ”Retention” and ”Value” metrics and defines the specific responsibilities of each team member to meet the ”From Scratch” implementation requirements of the course.

1. Project Definition & Justification

The Challenge

The project requires analyzing customer behavior to predict **Retention** (Classification) and **Value** (Regression). While traditional examples use monthly contracts (e.g., Netflix), this project tackles the more complex challenge of **”Non-Contractual Retention”** in the Attention Economy (e.g., TikTok, Alibaba).

Justifying the Title for the Professor

We are using the **Alibaba E-commerce User Behavior Dataset**. Here is how we justify this choice against the project title to ensure full marks for ”Problem Setting” and ”Storytelling”:

1. **Customer Behavior Analysis:** Unlike simple transaction logs, this dataset captures high-frequency behavior: scrolling depth, click-through rates (CTR), dwell time, and decision sequences. We analyze *how* users interact, not just what they buy.

2. **Subscription Retention (The ”Creative Interpretation”):** In modern apps, **Retention = Engagement**.

    - We treat the user’s active **Session** as the ”Subscription.”
    - **Retained:** As long as the user scrolls to the next page (Column 15 `terminal` = 0), they are ”renewing” their subscription to the app’s content.
    - **Churned:** When they stop scrolling or leave the app (Column 15 `terminal` = 1), they have churned.
    - **Goal:** Predict the probability of Churn at step t.

3. **Value Estimation:** We map specific columns to financial value.

    - **Feature:** Column 5 (User Purchase Power) and Column 13 (Purchase Amount).
    - **Goal:** Predict the **Total Session Revenue** (LTV) a user will generate before they churn.

2. The Data: Alibaba User Behavior (RL)

Source: Alibaba E-commerce User Behavior Dataset (Used for Reinforcement Learning Research).
Link: https://drive.google.com/file/d/14OtIC8eiDkzoWCTtaUZHcb7eB-bUmtTT/view

Detailed Data Format

The dataset is complex (concatenated lists inside cells), satisfying the requirement for ”Data Engineering” and ”Feature Insight.”

Col	Name	Description & Project Mapping
1	Page ID	The sequence ID (0-11). Represents the **Time Step** of the session.
2	Hour	24hr format. Used for behavioral patterns.
3–5	User Profile	Age, Gender, and **Purchase Power**. (Crucial User State feature).
6	Position List	List of positions for 50 items shown.

7–9	Predictions	Platform's predicted CTR/CVR/Price.
10	**Is Click**	List of 0/1. Summing this gives **Engagement Score**.
11	Is Cart	Strong intent signal.
12	Is Fav	Wishlist signal.
13	**Purchase Amount**	List of Float values (Yuan). Summing this gives **Session Value (Target for Regression)**.
14	State Feature	Optional RL feature.
15	**Terminal**	**Churn Label (Target for Classification)**. 0=Stay, 1=Leave.

3. Team Roles & Responsibilities

To tackle the 100-point score, responsibilities are divided to ensure deep focus on ”From Scratch” implementation.

1. Lead Data & AI Engineer (You)

Focus: Infrastructure, Architecture, Pipelines.

- **Job:** Build the ”Factory.” Set up Docker, Airflow, and MinIO. Write the ETL pipelines that ingest the raw text data and parse the complex columns into clean Parquet files.

- **Grading Requirement Met:** ”Case analysis and preparation of structured data” (30 pts). Handing complex Big Data proves engineering capability.

2. Lead Data Scientist (The Storyteller)

Focus: EDA, Feature Engineering, Strategy.

- **Job:** Analyze the clean data. Determine which 5+ features (e.g., Purchase Power, Hour, Clicks) drive Churn. Design the ”Problem Statement” presentation.

- **Grading Requirement Met:** ”Finalize feature selection and insight” (Storytelling) and ”Highly creative problem setting” (+5 Extra Points).

3. ML Engineer A: Regression Specialist

Focus: Numerical Forecasting (Value Estimation).

- **Job:** Build **Linear Regression** and **Polynomial Regression** classes *from scratch* (using NumPy, no Scikit-learn logic). Implement the ”Better Model” (e.g., Gradient Boosting Regressor from scratch).

- **Target:** Predict `page_value` (Revenue).

- **Grading Requirement Met:** ”Task: Regression (all models built from scratch)” and ”A better regression model” (Mandatory).

4. ML Engineer B: Classification Specialist

Focus: Categorical Prediction (Subscription Retention).

- **Job:** Build **Logistic Regression** and **Decision Tree** classes *from scratch*. Implement the ”Better Model” (e.g., Random Forest from scratch).

- **Target:** Predict `is_churn` (Terminal = 1).
- **Grading Requirement Met:** ”Task: Classification (all models built from scratch)” and ”Quantitative results (Confusion Matrix, ROC, F1)”.

4. How to Set Up the Project (Mac/Laptop)

Follow these exact steps to join the project infrastructure.

Prerequisites

1. **Install Docker Desktop:** `https://www.docker.com/products/docker-desktop/` 2. **Install Git.**

Step-by-Step Installation

1. Clone the Repository

```
git clone <REPO_URL>
cd customer-churn-mlops
```

2. Download & Place Data

- Download the data from the link in Section 2.
- Rename the file to: `alibaba_behavior.txt` (Crucial for the pipeline).
- Move it to: `customer-churn-mlops/data/raw/alibaba_behavior.txt`.

3. Start the Engine Run this in your terminal:

```
docker-compose up -d
```

Wait 2 minutes. This launches Airflow (Orchestrator), MinIO (Storage), and MLflow (Tracking).

4. Generate the Training Data

- Go to `http://localhost:8080` (User: `airflow`, Pass: `airflow`).
- Toggle **ON** the DAG: `alibaba_ingestion_pipeline`. Click **Play**.
- Toggle **ON** the DAG: `feature_engineering_pipeline`. Click **Play**.
- *Result:* The `processed-data` bucket in MinIO will fill with clean training data.

5. Run Your Models (For Engineers) Run your specific python scripts locally to train your Scratch models:

```
# Regression Engineer
python3 src/regression/value_predictor.py

# Classification Engineer
python3 src/classification/churn_predictor.py
```