Predicting User Churn for an E-commerce Platform

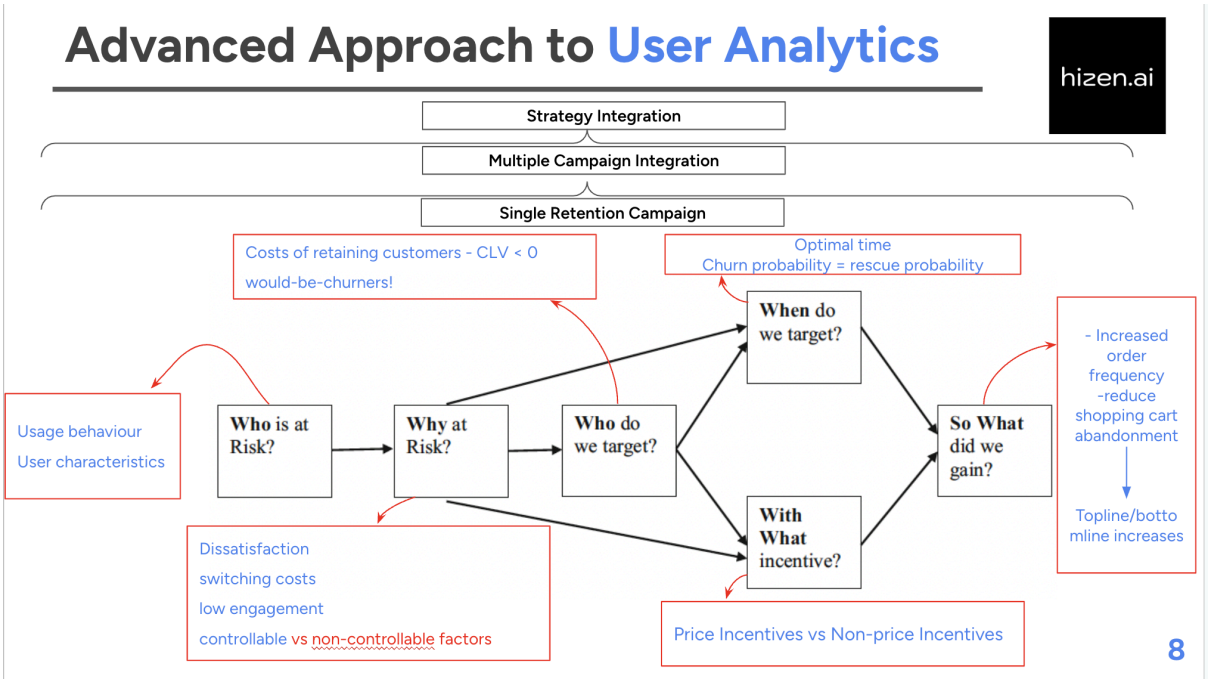
1. Objective

You are given an events dataset containing user activities (view, cart, purchase) on an e-commerce platform. Your goal is twofold:

1. **Predict** which users are most likely to churn (i.e., stop returning or purchasing).
2. Provide **insights** on the **why** behind their churn, focusing on actionable business takeaways.

The hunt is for **strong problem solvers**—candidates who can go beyond just building a model and demonstrate how they **think about the problem**, **define churn**, and **create features** that capture meaningful aspects of user behavior.

Reference Research Paper: Given is the research paper on **user retention** that outlines theoretical approaches and industry best practices. You are **encouraged** to draw inspiration or methodologies from it to inform your approach.



2. Dataset

File Name: events.csv

Columns:

1. **event_time**: Datetime of the event (e.g., 2020-12-01 09:00:00)

2. **event_type**: Type of the event, one of view, cart, or purchase
3. **product_id**: Identifier of the product
4. **category_id**: Identifier of the product's category
5. **category_code**: Human-readable category code (if available)
6. **brand**: Brand of the product
7. **price**: Price of the product (numerical)
8. **user_id**: Identifier of the user who performed the event
9. **user_session**: Identifier of the user's session

---

3. Expected Deliverables

1. **Exploratory Data Analysis (EDA) Report**
    - Overview of the dataset and user behavior.
    - Key patterns or anomalies in the data.
2. **Churn Definition & Reasoning**
    - Propose a clear **definition of churn** for this e-commerce context (e.g., no purchases or site visits after X days).
    - Explain why you chose this definition and how you handle edge cases or special user types.
3. **Feature Engineering**
    - **Create user-level features** that capture churn signals from the event stream. This is a **key part** of the assignment.
    - Possible feature areas include:
        - **RFM metrics** (Recency, Frequency, Monetary).
        - **Session-based metrics** (session count, average session duration, bounce rate, etc.).
        - **Product/brand preferences** (most viewed brand, top categories viewed vs. purchased).
        - **Behavioral patterns** (view-to-cart ratio, cart-to-purchase ratio, etc.).
    - Document your feature transformation steps comprehensively. You can also read the given research paper to know more about the possible features.
4. **Predictive Modeling**
    - Build a churn prediction model using any ML algorithm(s) you find appropriate
    - Describe your model selection process and how you tune hyperparameters.
    - Demonstrate **performance metrics** (AUC, precision-recall, F1, etc.) to evaluate your model.
5. **Interpretability & Insights**
    - Identify which features are most influential in predicting churn (e.g., with feature importance or SHAP values).
    - Provide a rationale for why these features matter.

    - Highlight **which** products or categories may drive churn and **why**.
6. **Business Recommendations**
    - Connect your modeling insights back to real-world strategies for **user retention**.
    - Suggest possible interventions (personalized offers, product improvements, marketing campaigns, etc.).
7. **Code & Documentation**
    - Well-organized, reproducible code in a Jupyter notebook or script.
    - A short write-up summarizing your main findings, methodology, and recommendations.
8. **Reference Integration**
    - Briefly mention how concepts or methods from the **provided research paper** influenced your approach to churn modeling or feature creation.

4. Example Framework

1. **Understanding the Problem**
    - Restate the goal in your own words.
    - Show how you interpret the research paper's concepts in this context.
2. **Data Inspection & Preprocessing**
    - Load and clean the dataset.
    - Handle missing, duplicate, or inconsistent entries.
    - Parse event_time and ensure correct data types.
3. **Exploratory Data Analysis (EDA)**
    - Event distributions (view, cart, purchase) over time.
    - Brand and category popularity.
    - User-level summaries (total spend, frequency of visits, time between visits, etc.).
4. **Defining Churn**
    - Propose a threshold-based or time-based definition (e.g., no purchase in last 30 days).
    - Justify the choice with logical or business-based reasoning.
5. **Feature Engineering** (**Important**)
    - Explain the **logic** behind your feature creation.
    - Show how each feature might help identify at-risk users.
    - Give examples (e.g., "Users who viewed over 5 products but never carted any item have a higher churn probability.").
6. **Modeling**
    - Train at least one model; consider comparing multiple.
    - Use an appropriate validation scheme (e.g., train/test split, cross-validation).
    - Optimize hyperparameters if time and resources allow.
7. **Interpretation & Explanation**
    - Use feature importance plots, partial dependence plots, or SHAP.
    - Link model results to user behavior themes from the research paper.
8. **Recommendations & Conclusions**

- Summarize how your model can be integrated into **user retention** strategies.
- Suggest targeted or personalized interventions to reduce churn.

5. Evaluation Criteria

1. **The best model metrics(40%)**
   - Model's metrics such as Precision, Recall, F1 score, accuracy, log loss, brier loss etc.
2. **Problem-Solving Approach (20%)**
   - How effectively do you define the problem, propose a solution strategy, and iterate through the data?
3. **Feature Engineering (20%)**
   - Are the features you create from the event stream meaningful, creative, and relevant to churn prediction?
4. **Code Quality & Documentation(10%)**
   - Is your code well-structured, clearly commented, and reproducible?
5. **Reference Research Utilization(10%)**
   - Did you leverage insights or methodologies from the provided research paper in a thoughtful way?

6. Tips and Best Practices for Candidates

1. **Think Business & Technology**: Combine **domain insight** (from the paper + your own analysis) with robust ML practices.
2. **Creativity in Feature Engineering**: This is where you can differentiate yourself. Show us your ability to turn raw event data into powerful predictive signals.
3. **Logical Reasoning**: Justify your assumptions, model choices, and recommendations.
4. **Keep it Reproducible**: Make sure your code can be run end-to-end. (using constant seed wherever required)
5. **Be Concise & Clear**: We value clarity over complexity.

7. Submission Instructions

1. **Format**: Submit one Jupyter notebook (preferred) or a GitHub repository link.
2. **Content**:
   - All relevant code and a step-by-step approach.
   - A short summary/report (PDF or markdown) explaining the approach and the final scores.
   - A readme.md explaining clearly to run all the code files.
3. **Deadline**: **9th January 11:59 P.M**
4. We will evaluate the submitted solutions, followed by an interview round.

Note: We are more interested in understanding your approach to the problem statement than in receiving complete solutions. Can you dig deep and conduct an ultra-fine-grained analysis? This is what will set you apart from the rest of the candidates.

All the very best!

SUBMIT YOUR SOLUTIONS HERE.