1. Exploratory Data Analysis (EDA) Report

- **Overview of the Dataset**:
    - Number of rows and columns.
    - Description of each column (e.g., `event_time`, `event_type`, `price`, etc.).
    - Missing values and how they were handled (e.g., replacing `brand` with `Unknown`).
- **Key Findings**:
    - Distribution of `event_type` (e.g., views, carts, purchases).
    - Most popular brands and product categories.
    - User behavior patterns (e.g., frequency of purchases, total spend, active days).

Sample Insights:

- 70% of events are product views, while purchases make up only 5%.
- Top brands include `Brand A` and `Brand B`, contributing 30% of the purchases.
- 60% of users are active for less than a week, indicating short engagement cycles.

2. Churn Definition & Reasoning

Definition:

- Churned users are defined as those who have not made a purchase in the last 30 days.
- Justification: In e-commerce, the lack of purchases over a period suggests disengagement, especially when the average time between purchases is less than 30 days.

Handling Edge Cases:

- **New Users**: Users with only recent events but no purchases are excluded from churn calculations.
- **Bulk Purchasers**: High-value customers with infrequent but large purchases are flagged separately to avoid misclassification.

3. Feature Engineering

Features Created:

1. **RFM Metrics**:
    - **Recency**: Days since the last purchase.
    - **Frequency**: Total number of purchases.
    - **Monetary Value**: Total spend on purchases.
2. **Session-Based Metrics**:
    - **Session Count**: Number of sessions.
    - **Average Session Duration**: Average time spent per session.
    - **Bounce Rate**: Sessions with no purchase or cart activity.
3. **Behavioral Patterns**:

- **View-to-Cart Ratio**: Proportion of views that resulted in a cart action.
- **Cart-to-Purchase Ratio**: Proportion of carts that resulted in a purchase.

4. **Preferences**:
   - **Top Brand Interactions**: Number of interactions with the most popular brands.
   - **Category Diversity**: Number of unique categories a user interacted with.

4. Predictive Modeling

Model Used: XGBoost

- **Hyperparameter Tuning**: GridSearchCV was used to identify the best parameters:
  - Best Parameters: `{'colsample_bytree': 0.8, 'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.8}`.

Results:

- **Metrics**:
  - ROC AUC Score: `1.0`
  - Accuracy: `85.7%`.
  - Precision: `84.5%`
  - Recall: `100%`
  - F1 Score: `91.6%`
  - Log Loss: `0.292`
- **Classification Report**:
  - Class `1` (Churn): High recall (100%) but moderate precision (84.5%) indicates the model prioritizes identifying churners over avoiding false positives.

5. Interpretability & Insights

Feature Importance:

- The top features identified by XGBoost were:
  1. **Recency**: Strong predictor of disengagement.
  2. **Frequency**: Frequent purchasers are less likely to churn.
  3. **Cart-to-Purchase Ratio**: High ratios indicate higher engagement.

Insights:

- Users who haven't purchased in the last 30 days are at the highest risk of churn.
- Frequent cart actions without purchases signal hesitation; targeted campaigns (e.g., discounts) may reduce churn.

6. Business Recommendations

Recommendations:

1. **Personalized Discounts**:
    - Offer discounts to users with high `view-to-cart` ratios but no purchases.
2. **Engagement Campaigns**:
    - Re-engage users with low purchase frequency using personalized email reminders.
3. **Loyalty Programs**:
    - Incentivize high-value customers with loyalty rewards to maintain engagement.
4. **Product Improvements**:
    - Analyze most abandoned cart products and address common issues (e.g., pricing, shipping).

8. Reference Integration

The research paper highlights the importance of managing churn as a key aspect of customer relationship management, defining it as a discontinuity in transactions. It emphasizes leveraging transactional and behavioral data, along with advanced machine learning techniques, to predict churn. Inspired by this, features like RFM metrics, session-based patterns, and behavioral indicators were engineered for the model. The study also stresses the balance between proactive (preventing churn) and reactive (winning back churned users) strategies, influencing the definition of churn as inactivity over a set period. Evaluation metrics like AUC, precision, and recall were prioritized to ensure effective churn management. Additionally, actionable retention strategies, such as personalized incentives and customer segmentation, were derived from the paper's insights to link model outputs to business outcomes.