Internal Project Documentation: Customer Churn Analysis

A Self-Reference Guide for Methodology, Insights, and Business Value

Prepared by: Deepak Malviya
Data Analyst Intern

1. Project Overview

Business Problem: The project addresses the critical issue of low customer retention on a major Brazilian e-commerce platform. The primary goal was to move beyond surface-level metrics and diagnose the root causes of customer churn.

Focus on Churn: Customer churn was selected as the key performance indicator because it directly measures the health of the customer relationship and has a significant impact on long-term profitability. In e-commerce, acquiring new customers is substantially more expensive than retaining existing ones, making churn a primary lever for sustainable growth.

Expected Business Impact: The analysis was designed to produce actionable recommendations to reduce churn, improve customer lifetime value (CLV), and optimize operational and marketing expenditures by focusing on retention-driving factors.

2. Dataset & Business Context

Dataset Source and Scope: The analysis utilized the "Brazilian E-Commerce Public Dataset by Olist," available on Kaggle. This is a real-world, anonymized dataset.

Time Period and Scale: The data covers approximately 100,000 orders from 2016 to 2018, providing a robust sample size for statistically significant analysis of customer behavior over a two-year period.

Suitability for Churn Analysis: The dataset's relational structure, containing detailed information on customers, orders, payments, reviews, and products, makes it ideal for a comprehensive churn analysis. It allows for the creation of a 360-degree customer view by joining disparate data sources.

3. Data Architecture & Modeling Logic

Relational Structure: The data is organized in a relational model, with distinct tables for entities like customers, orders, and items. This structure minimizes data redundancy and enhances data integrity. The core of the analysis involved joining these tables to synthesize a complete customer journey.

Role of Each Table:

- **customers:** Contains customer identifiers and location data.
- **orders:** The central table, linking customers to transactions and tracking order timelines.
- **order_items:** Details the products and financials (price, freight) within each order.
- **order_payments:** Records payment methods, installments, and values.
- **order_reviews:** Captures customer feedback scores and comments.

Key Identifier for Aggregation: The `customer_unique_id` was chosen as the primary key for aggregation. While `customer_id` is generated for each order, `customer_unique_id` remains constant for an individual across all their orders, making it the only reliable identifier for analyzing customer lifetime behavior and repeat purchases.

4. Data Ingestion & Preparation Approach

Tools Used: The data pipeline was built using Python (Pandas, SQLAlchemy) and a MySQL database. SQL was the primary language for data transformation and feature engineering.

Data Loading Strategy: Each raw CSV file was read into a Pandas DataFrame and then loaded into a dedicated table within a MySQL database (`customerchurn_db`). This approach created a centralized, queryable repository for the raw data.

Data Quality and Filtering: A critical preparation step was to filter for orders with a status of 'delivered'. This ensures that the analysis is based on completed transactions, as churn and satisfaction can only be meaningfully assessed for customers who have received their products.

5. Exploratory Analysis Rationale

The initial exploratory data analysis (EDA) was designed to answer foundational business questions and establish a baseline understanding of the platform's ecosystem.

- **Sales Trend Analysis:** To identify revenue patterns, seasonality (e.g., Black Friday spikes), and the overall growth trajectory of the business.
- **Customer Distribution Analysis:** To understand the geographic concentration of the customer base, revealing key markets and potential regional dependencies.
- **Product Category Performance:** To determine which product categories are the primary revenue drivers, informing inventory and marketing strategies.
- **Logistics Performance Evaluation:** To get a preliminary measure of operational efficiency by calculating the rate of delayed orders, a suspected driver of dissatisfaction.

6. Churn Definition Strategy

Challenge: The dataset lacks an explicit "churn" or "account closed" flag. Therefore, a proxy definition was required.

Custom Churn Definition: A customer was defined as **churned** if they had not made a new purchase within **180 days (6 months)** of their last purchase, relative to the final date in the dataset.

Justification: A 6-month window was chosen as a reasonable timeframe for a non-subscription e-commerce model. It is long enough to account for infrequent purchasers but short enough to identify a clear drop-off in engagement.

Limitations: This definition is an assumption. A customer might intend to return after 6 months. However, for a large-scale analysis, this time-based cohorting is a standard and effective industry practice for identifying at-risk customers.

7. Feature Engineering Decisions

To move from raw data to predictive insights, several features were engineered to quantify the customer experience. These were aggregated at the `customer_unique_id` level.

- **Delivery Delay:** (Actual Delivery Date - Estimated Delivery Date).
  - **Business Meaning:** Measures the gap between the company's promise and its actual performance.
  - **Why it Matters:** A positive delay (late delivery) is a direct breach of trust and a primary source of frustration.
- **Freight Ratio:** (Freight Cost / Total Order Value).
  - **Business Meaning:** Represents the proportion of the total bill that is shipping cost.
  - **Why it Matters:** High ratios can lead to "sticker shock," making the customer feel the transaction is unfair, especially for low-value items.
- **Average Review Score:** The mean of a customer's review scores across all orders.
  - **Business Meaning:** A direct, quantitative measure of a customer's expressed satisfaction.
  - **Why it Matters:** Low scores are a clear signal of dissatisfaction and a strong predictor of churn.
- **Order Frequency:** The total number of orders placed by a customer.
  - **Business Meaning:** A simple measure of customer loyalty and engagement.
  - **Why it Matters:** Repeat purchasers are the backbone of a healthy business; a lack of frequency indicates a retention problem.

8. Key Analytical Findings

> The analysis distilled several critical, data-backed insights into the drivers of customer churn. Charts and code were omitted here to focus purely on the "so what."

- **Impact of Delivery Delays:** Delivery performance is the single most significant factor driving churn. A clear, positive correlation exists between the length of a delivery delay and the probability of a customer churning. Customers experiencing "Very Late" deliveries (>5 days) had a churn rate of 85.9%, compared to 59.5% for those with "On Time" deliveries.
- **Price Sensitivity to Freight Costs:** Customers who churned had a demonstrably higher average ";freight ratio" than those who were retained. This confirms that the perceived cost of shipping, relative to the item's price, is a major friction point that kills loyalty.
- **Regional Churn Patterns:** Churn is not uniform across Brazil. Northern and Northeastern states show significantly higher churn rates, which correlate strongly with poorer average delivery performance in those regions. The business is systematically failing customers in these areas.
- **Critically Low Repeat Customer Rate:** Only 3.12% of customers ever make a second purchase. This reveals a "one-and-done" business model heavily reliant on costly customer acquisition, which is an unsustainable long-term strategy.

9. Business Implications

The findings translate into direct consequences for the business:

- **Operations:** Inefficient logistics are not just a cost center; they are actively destroying customer relationships and future revenue streams. The failure of regional logistics partners is a primary operational liability.
- **Marketing:** Marketing spend on customer acquisition in regions with poor logistics is largely wasted. The business is paying to acquire customers it is almost guaranteed to lose, resulting in a negative return on investment.
- **Customer Experience:** The customer journey is broken at two critical touchpoints: the delivery promise and the price perception. This erodes trust and makes it difficult to build a loyal customer base.
- **Financial Impact:** The business is losing money by failing to retain customers, especially high-value customers who have a bad delivery experience. The low repeat purchase rate indicates a significant untapped revenue potential in the existing customer base.

10. Strategic Recommendations

Based on the analysis, the following prioritized actions were recommended to address the root causes of churn:

1. **Optimize Logistics Partnerships (High Priority):**

    - **Recommendation:** Immediately audit and replace underperforming logistics carriers in high-churn states.
    - **Expected Outcome:** Reduced delivery delays, leading to a direct decrease in the churn rate for affected regions.
    - **Metric to Track:** Average delivery delay per state; churn rate per state.
2. **Implement Proactive Customer Recovery (High Priority):**
    - **Recommendation:** Create an automated "Apology Protocol" that sends a discount voucher to customers whose orders are predicted to be delayed.
    - **Expected Outcome:** Mitigate customer frustration from a negative experience, potentially salvaging the relationship and encouraging a future purchase.
    - **Metric to Track:** Churn rate of customers who received an apology voucher vs. those who did not.
3. **Restructure Shipping Costs (Medium Priority):**
    - **Recommendation:** Introduce free shipping above a certain cart value and consider absorbing shipping costs into the item price for low-value goods.
    - **Expected Outcome:** Reduced "sticker shock" from high freight ratios, leading to improved conversion and retention for price-sensitive customers.
    - **Metric to Track:** Average freight ratio; cart abandonment rate.

11. Assumptions & Limitations

- **Data Limitations:** The analysis is confined to the 2016-2018 period. Customer behavior may have evolved since. The dataset also lacks information on marketing campaigns or customer demographics beyond location.
- **Churn Definition Assumption:** The 6-month inactivity window is a reasoned assumption but may not perfectly capture every customer's churn intent.
- **External Factors:** The model does not account for external factors like competitor actions, macroeconomic conditions, or local holidays that could influence purchasing behavior.
- **Correlation vs. Causation:** While strong correlations were found (e.g., delay and churn), this analysis does not definitively prove causation. However, the logical link is strong enough to warrant business action.

12. Future Enhancements

- **Incorporate Additional Data:** Enriching the dataset with customer demographics, web browsing behavior, or marketing interaction data would yield more nuanced insights.
- **Advanced Modeling:** Implement more sophisticated machine learning models (e.g., Gradient Boosting, Neural Networks) to improve churn prediction accuracy. Survival analysis could also be used to model the *time* to churn.

- **Real-Time Churn Monitoring:** Develop a real-time dashboard that tracks leading indicators of churn (e.g., delivery delays, low review scores) to enable proactive intervention.
- **Production-Level Implementation:** Deploy the churn model into a production environment to score customers daily, feeding high-risk segments directly into automated marketing and retention campaigns.

13. Final Summary

Achievement: This project successfully moved beyond a surface-level understanding of churn to diagnose its primary drivers: logistical failures and negative pricing perceptions. It transformed raw, multi-table data into a cohesive, customer-centric analytical base table and derived actionable business intelligence.

Value and Capability: The project demonstrates end-to-end analytical capability, from data ingestion and complex SQL-based feature engineering to hypothesis-driven analysis and the formulation of data-backed strategic recommendations. It provides a clear, evidence-based roadmap for the business to address a critical vulnerability (low retention) and unlock sustainable, long-term growth.