Customer Churn Data Analysis and Feature Engineering Report

Date: 2025-06-01  
Author: [Kasandika Andariefli]

# 1. Introduction

The objective of this analysis is to investigate customer churn behaviors using a multi-source dataset. We consolidated data from several sources to create a unified dataset that captures customer demographics, transactional behavior, service interactions, and digital engagement. This report outlines the data processing steps, feature engineering, exploratory data analysis, correlation analysis, anomaly detection, and key insights for churn prediction.

# 2. Data Sources & Join Strategy

The dataset comprises five sheets:  
- Customer\_Demographics: Basic customer attributes  
- Transaction\_History: Purchase records  
- Customer\_Service: Customer support interactions  
- Online\_Activity: Digital engagement metrics  
- Churn\_Status: Churn label (1=churned, 0=retained)  
  
We aggregated and merged these tables on the CustomerID field to build a comprehensive dataset.

# 3. Data Preprocessing & Feature Engineering

We aggregated the Transaction\_History data by CustomerID to calculate total spent, average spent, and number of transactions. Similarly, Customer\_Service data was aggregated to count interactions and identify the most common resolution status.  
  
Categorical features like Gender, MaritalStatus, IncomeLevel, and ServiceUsage were encoded to numerical formats for analysis. Merging these tables provided a unified view of each customer's behavior.

# 4. Exploratory Data Analysis

Initial exploration of numerical features was performed using descriptive statistics, boxplots, and histograms. Key insights include identifying variability in average spending behavior and understanding the distribution of login frequency and interaction counts.

# 5. Correlation Analysis

Correlation analysis showed that features like LoginFrequency, NumInteractions, and ServiceUsage have the strongest negative correlation with churn. Features like AvgSpent and Age have a weak positive correlation with churn. TotalSpent and other features show near-zero correlation.

# 6. Anomaly Detection

We used Z-score based anomaly detection to identify outliers in numerical features. Only AvgSpent showed significant outliers, highlighting customers with unusually high or low average spending. These outliers are flagged for further analysis, as they may indicate either errors or key behavioral patterns.

# 7. High Spender Analysis

An in-depth examination of high spenders revealed several churned customers with above-average spending behavior. This suggests that high spending alone does not guarantee loyalty, and these customers may require targeted retention strategies.

# 8. Conclusion & Recommendations

This analysis demonstrates that churn behavior is influenced by multiple subtle factors, including digital engagement and service interactions. High spenders who churn may represent valuable segments for personalized retention efforts. Next steps involve building predictive churn models incorporating these insights and testing targeted interventions.