02_Preprocessing

January 29, 2026

Connected to Python 3.10.11

1 Preprocessing & Feature Strategy

Based on the exploratory data analysis, we will define our preprocessing and feature strategy as follows:

1. Which features to keep/drop
2. How to handle missing values
3. How to encode categorical variables
4. Which features need special care (leakage / interpretation)

1.1 Target & Identifier

Target:

Churn

- Binary classification (0 / 1)
- Imbalanced (~16.8% churn rate)

Identifier:

CustomerID

- CustomerID is a unique identifier and is excluded from modeling.

1.2 Feature Selection

A. Numeric features (candidates) - Tenure - OrderCount - DaySinceLastOrder - leakage risk* - HourSpendOnApp - SatisfactionScore - WarehouseToHome - CouponUsed - CashbackAmount

DaySinceLastOrder will be excluded from baseline models due to potential leakage and reconsidered only after a strict prediction timestamp is defined.

B. Categorical features (candidates) - PreferredLoginDevice - phone & mobile phone combined - PreferredPaymentMode - CC & Credit Card combined + COD & Cash on Delivery combined - Gender - PreferredOrderCategory - MaritalStatus - CityTier

1.3 Missing Value Strategy

There are missing values in numeric features.

Missing values will be imputed using median. For selected numeric features where missingness may be informative, binary missing-value indicators will be added.

1.4 Categorical Encoding Strategy

Categorical features will be encoded using one-hot encoding to allow non-ordinal representation. High-cardinality categorical features are limited, so one-hot encoding is unlikely to introduce excessive dimensionality.

CityTier will be treated as categorical for baseline models, even though it is ordinal. Ordinal encoding may be explored in later experiments.

1.5 Scaling

Required for: - Logistic Regression - Distance-based models

Not required for: - Tree-based models (e.g., Random Forest, Gradient Boosting)

Feature scaling will be applied where required depending on the model family.

1.6 Evaluation-aware decisions

Because churn is imbalanced (16.8%):

Accuracy is not sufficient to evaluate model performance. Focus will be on: - Recall - few false positives (not to waste retention actions on customers who wouldn't churn) - Precision - few false negatives (to catch most at-risk customers before they leave) - F1-score - balance between precision and recall - ROC-AUC - overall model quality

1.7 Special Feature Considerations

- **DaySinceLastOrder**: Potential leakage risk depending on churn definition. Will be handled cautiously.
- **SatisfactionScore**: Counterintuitive relationship with churn. Will be monitored for model interpretability.
- **Imbalanced Target**: Churn is imbalanced, so evaluation metrics will account for this (e.g., using AUC-ROC, F1-score).
- **Feature Interactions**: Some features may have predictive power only in combination. Feature interactions will be explored in later modeling stages.