Feature Engineering – Simplified Spec

Guardrails

- Only use data available **on or before** `asof_date` (to prevent leakage).
- The **label is** `churned`. Do **not** use it as a feature.

Data Cleaning & Fixes

- **Dates** → parse as proper dates: `asof_date`, `signup_date`, `last_login_date`.
- **Auto-renew flags** → collapse into one:
  - `auto_renew_enabled = 1` if `auto_renew_enabled_True == True`
  - `auto_renew_enabled = 0` if `auto_renew_enabled_False == True`
  - Add `auto_renew_enabled_missing` if neither present.
- **Plans** → one-hot encode `subscription_plan_le` → `plan_0`, `plan_1`, `plan_2`.
- **Drop**: `auto_renew_enabled_False`, `auto_renew_enabled_True`, `subscription_plan_le` (optional), `ingest_ts`.

Missing Data Rules

- `last_login_date` missing →
  - `last_login_missing = 1`

    - `days_since_last_login = 9999` (sentinel)
- For other numeric columns: impute later (median or zero) + create `<col>_missing` flag.

Row-Level Derived Features

- `tenure_days = asof_date - signup_date`
- `days_since_last_login = asof_date - last_login_date` (or `9999` if missing)
- `inactive_30d = 1` if `days_since_last_login > 30`
- `inactive_90d = 1` if `days_since_last_login > 90`
- `new_user_30d = 1` if `tenure_days ≤ 30`
- `tickets_per_30d = support_tickets_last_90d / 3`
- `session_hours = avg_session_length_minutes / 60`
- `email_open_rate_30d = email_opens_last_30d` (already in [0,1] if rate)

Seasonality Features

- `asof_month = month(asof_date)` ∈ {1..12}
- Cyclical encoding:
    - `asof_month_sin = sin(2π * asof_month / 12)`
    - `asof_month_cos = cos(2π * asof_month / 12)`

Auto-Renew Logic

- `auto_renew_off = 1` if `auto_renew_enabled == 0`
- `auto_renew_off_and_inactive_30d = auto_renew_off AND inactive_30d`

Interaction Features

- `tenure_x_auto_renew_off = tenure_days * auto_renew_off`
- `inactivity_x_email = days_since_last_login * (1 - email_open_rate_30d)`
- `tickets_x_recency = tickets_per_30d * inactive_30d`

Trend Features (if multiple snapshots per customer)

- Sort by `asof_date` within `customer_id`
- Changes vs. previous snapshot:
  - `delta_email_opens_30d`
  - `delta_session_hours`
  - `delta_tickets_30d`
- Rolling means (3 snapshots, min 2):
  - `rollmean_email_3`, `rollmean_session_3`, `rollmean_tickets_3`

Scaling / Normalization

- **For linear/KNN/SVM models:**
    - Z-score: `session_hours`, `email_open_rate_30d`, `tickets_per_30d`, `monthly_spend` (if not already in [0,1])
    - Robust/log1p + clip at P99: `tenure_days`, `days_since_last_login` (ignore sentinel 9999)
- **For tree/boosting models:**
    - Only clip extreme outliers at P99
    - Keep binary/one-hot as 0/1

Final Column Groups (besides keys/label)

- **Keys:** `customer_id`, `asof_date`
- **Dates:** `signup_date`, `last_login_date`
- **One-hot:** `plan_0`, `plan_1`, `plan_2`
- **Engagement:** `session_hours`, `email_open_rate_30d`, `inactive_30d`, `inactive_90d`, `new_user_30d`
- **Support:** `tickets_per_30d`
- **Tenure/Recency:** `tenure_days`, `days_since_last_login`, `last_login_missing`
- **Auto-renew:** `auto_renew_enabled`, `auto_renew_off`, `auto_renew_off_and_inactive_30d`
- **Interactions:** `tenure_x_auto_renew_off`, `inactivity_x_email`, `tickets_x_recency`
- **Trends (if available):** deltas + rollmeans

- **Missing flags:** `<col>_missing` as needed
- **Label (separate):** `churned`