Churn Prediction Model Based on Beta Distribution

Description

This model is designed to predict user churn by modeling churn time (T) as a discrete random variable, influenced by a hidden continuous rebilling probability (p), which follows a Beta distribution.

Base Model

Rebilling Probability

$p \in (0, 1)$ — a continuous hidden random variable representing the probability of a user rebilling.

Churn Time

$T \in \{0, 1, 2, 3, 4, 5, 6\}$ — the discrete time period at which churn occurs.

Probability of Churn at Time t for Fixed p

$$P(T = t \mid p) = p^t \cdot (1 - p)$$

Prior Distribution for p

$$p_i \sim \mathrm{Beta}(\alpha(x_i), \beta(x_i))$$

Marginalized Distribution

$$P(T = t_i \mid \alpha(x_i), \beta(x_i)) = \frac{B(\alpha(x_i) + t_i, \beta(x_i) + 1)}{B(\alpha(x_i), \beta(x_i))}$$

Censored Observations

$$P(T > t_i \mid \alpha(x_i), \beta(x_i)) = \frac{B(\alpha(x_i) + t_i + 1, \beta(x_i))}{B(\alpha(x_i), \beta(x_i))}$$

Logarithm of the Beta Function

$$\log B(\alpha, \beta) = \log \Gamma(\alpha) + \log \Gamma(\beta) - \log \Gamma(\alpha + \beta)$$

Loss Function

The final loss function for the model, which predicts $(\alpha(x_i), \beta(x_i))$, considers both observed and censored data:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [\log B(\alpha(x_i) + t_i + c_i, \beta(x_i) + 1 - c_i) - \log B(\alpha(x_i), \beta(x_i))]$$

where:

- t_i is the observed churn time
- c_i is the censoring indicator (1 if the observation is censored, 0 otherwise)

Beta-Discrete-Weibull Model

The Beta-Discrete-Weibull (BdW) model extends the basic Beta model by introducing a Weibull scaling factor $\gamma(x_i)$.

Conditional Distribution

$$P(T = t \mid p) = p^{t^\gamma} - p^{(t+1)^\gamma}$$

Prior Distribution

$$p_i \sim \mathrm{Beta}(\alpha(x_i), \beta(x_i))$$

Censored Probability

$$P(T > t \mid \alpha(x_i), \beta(x_i)) = \frac{B(\alpha(x_i) + (t+1)^{\gamma(x_i)}, \beta(x_i))}{B(\alpha(x_i), \beta(x_i))}$$

Marginalized Distribution

$$P(T = t \mid \alpha(x_i), \beta(x_i)) = \frac{B(\alpha(x_i) + t^{\gamma(x_i)}, \beta(x_i)) - B(\alpha(x_i) + (t+1)^{\gamma(x_i)}, \beta(x_i))}{B(\alpha(x_i), \beta(x_i))}$$

Log-Likelihood Function

$$\mathcal{L} = \sum_{i=1}^{n} \Big[\log \Big(B(\alpha(x_i) + t_i^{\gamma(x_i)}, \beta(x_i)) - B(\alpha(x_i) + (t_i+1)^{\gamma(x_i)}, \beta(x_i))\Big) \cdot \mathbb{I}_{c_i=0}$$

$$+ \log B(\alpha(x_i) + (t_i+1)^{\gamma(x_i)}, \beta(x_i)) \cdot \mathbb{I}_{c_i=1} - \log B(\alpha(x_i), \beta(x_i))$$