Нейросетевые подходы к решению задачи оттока абонентов

Е. В. Батарин

Московский физико-технический институт

27 марта 2025 г.

Цели исследования

Задача

Создать модель предсказания оттока абонентов на основе методов анализа выживаемости

Требуется

Предложить метод, который:

1) Учитывает неполноту информации о факте оттока
2) Упорядочивает абонентов в зависимости от их времени оттока
3) Является интерпретируемым

Решение

Использовать нейросетевую модель с дискретным временем

Обозначения

$\mathcal{T} = \{0, \dots, T_{\max}\}$ - дискретное время
$\mathcal{K} = \{\emptyset, 1, \cdots, K\}$ - множество событий
$\tau^i = \min(T^i, C^i) \in \mathcal{T}$ - право-цензурированные отсчеты времени
$\mathcal{X}^i(t) = \{\mathbf{x}^i(t_j^i) : 0 \le t_j^i \le t \text{ for } j = 1, \cdots, J^i\}$ - вектора признаков
$\mathcal{D} = \{(\mathcal{X}^i, \tau^i, k^i)\}_{i=1}^N$ - обучающая выборка

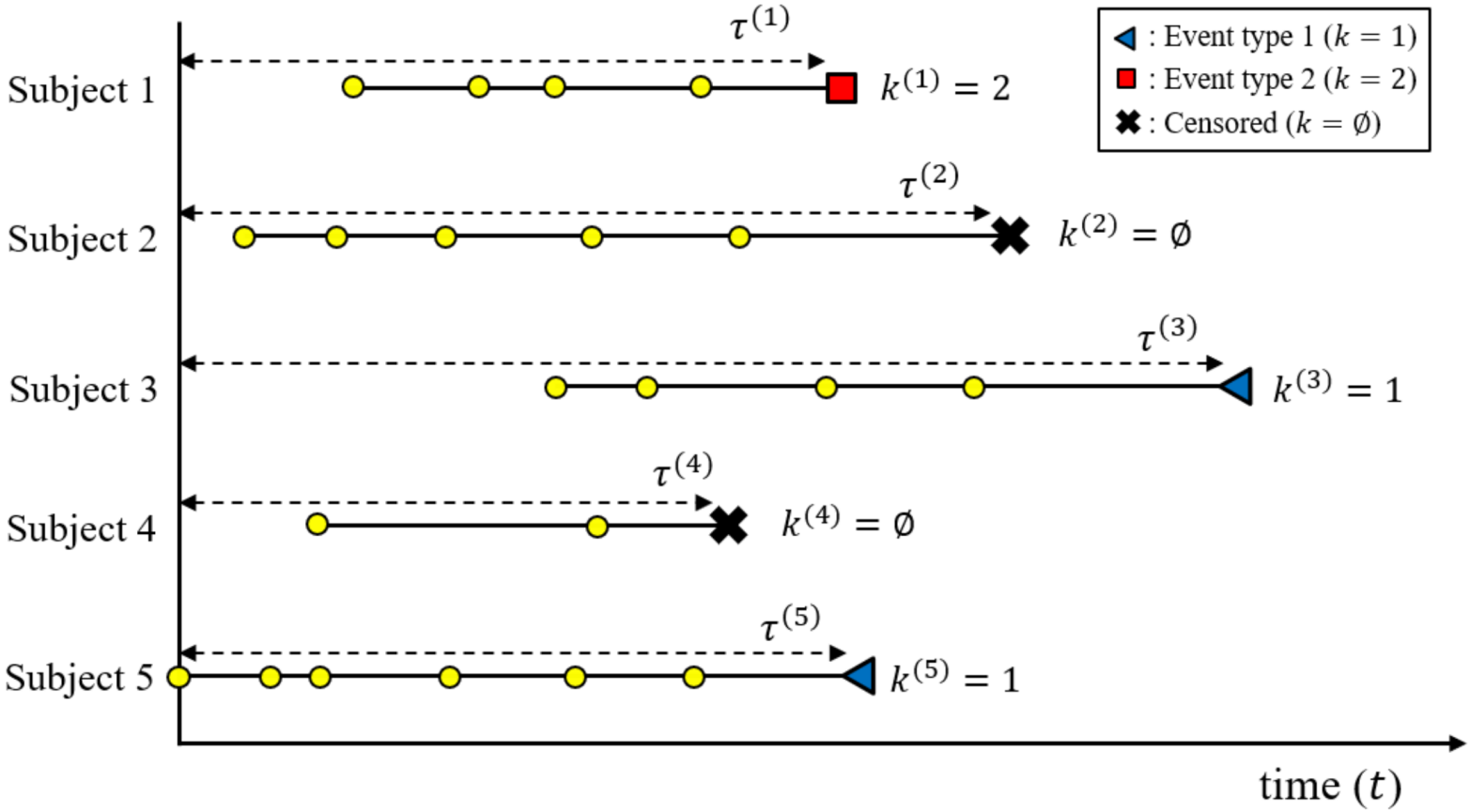
$$F_{k^*}(\tau^*|\mathcal{X}^*) = P(T \le \tau^*, k = k^*|\mathcal{X}^*, T > t^*_{J^*}) = \sum_{\tau \le \tau^*} P(T = \tau, k = k^*|\mathcal{X}^*, T > t^*_{J^*}).$$

Функция распределения для события k^*

$$S(\tau^*|\mathcal{X}^*) = P(T > \tau^*|\mathcal{X}^*, T > t^*_{J^*}) = 1 - \sum_{k \ne \emptyset} F_k(\tau^*|\mathcal{X}^*).$$

Функция выживания

Иллюстрация правого цензурирования



Цензурирование абонентов

Предложенный метод

Постановка задачи

На основе обучающей выборки $\mathcal{D}$ построить аппроксимации функции распределения и функции выживаемости: $\hat{F}_{k^*}(\tau^*|\mathcal{X}^*)$ и $\hat{S}(\tau^*|\mathcal{X}^*) = 1 - \sum_{k\neq\emptyset} \hat{F}_{k^*}(\tau^*|\mathcal{X}^*)$

Функция потерь

Задача сводится к минимизации функции $\mathcal{L}_{\text{Total}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$, которая состоит из слагаемых:

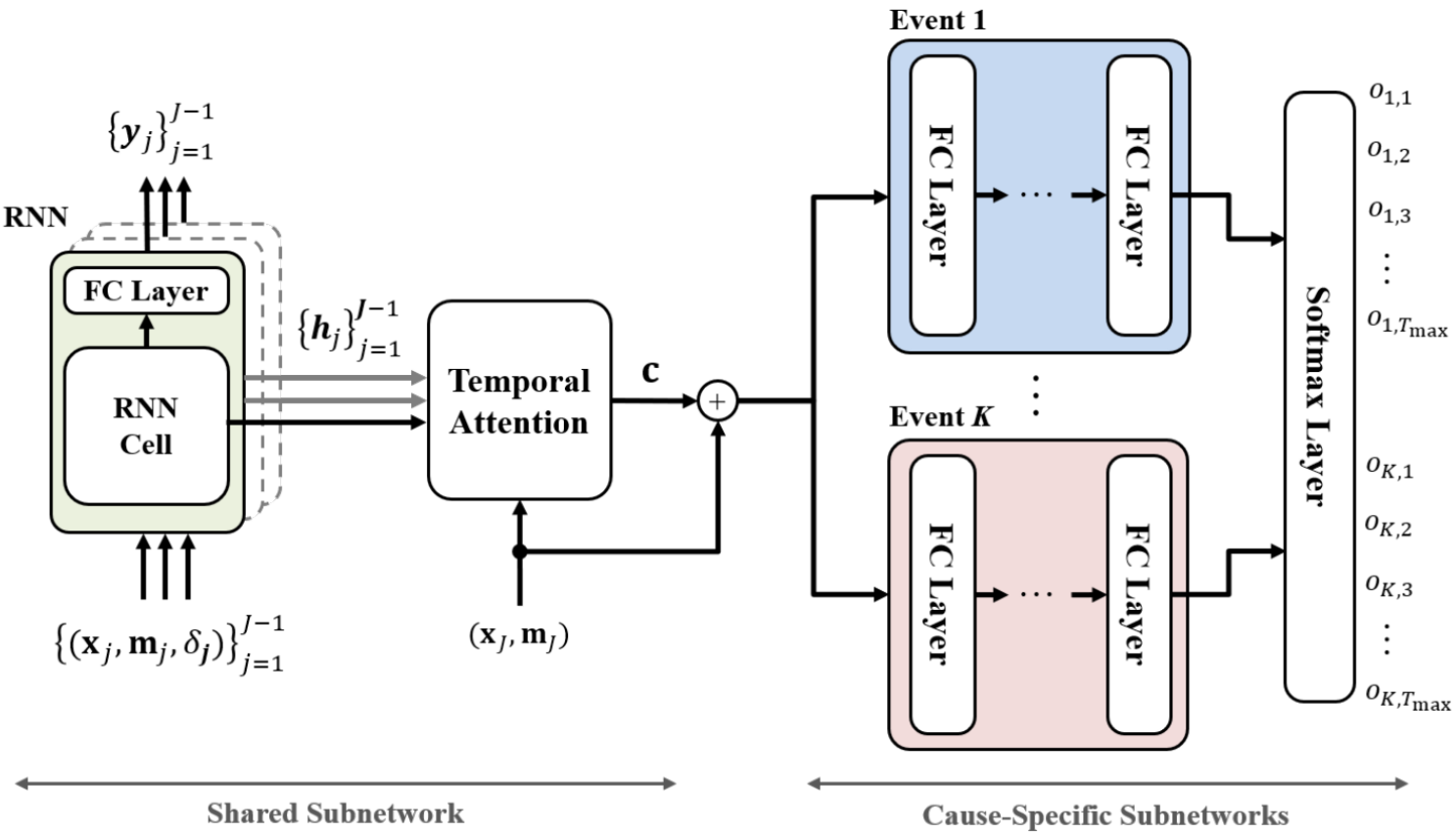
$$\mathcal{L}_1 = -\sum_{i=1}^{N}\left[\mathbb{1}(k^i \neq \emptyset)\cdot \log\left(\frac{o^i_{k^i,\tau^i}}{1-\sum_{k\neq\emptyset}\sum_{n\leq t^i_{Ji}} o^i_{k,n}}\right) + \mathbb{1}(k^i=\emptyset)\cdot\log\left(1-\sum_{k\neq\emptyset}\hat{F}_k(\tau^i|\mathcal{X}^i)\right)\right]$$

$$\mathcal{L}_2 = \sum_{k=1}^{K}\alpha_k\sum_{i\neq j}A_{kij}\cdot\eta\left(\hat{F}_k(s^i+t^i_{Ji}|\mathcal{X}^i), \hat{F}_k(s^i+t^j_{Jj}|\mathcal{X}^j)\right)$$

$$\mathcal{L}_3 = \beta\cdot\sum_{i=1}^{N}\sum_{j=0}^{J^i-1}\sum_{d\in\mathcal{I}}(1-m^i_{j+1,d})\cdot\zeta(x^i_{j+1,d}, y^i_{j,d})$$

где $s^i = \tau^i - t^i_{Ji}$, $A_{kij} = \mathbb{1}(k^i = k, s^i < s^j)$, $\eta(a,b) = \exp\left(-\frac{a-b}{\sigma}\right)$, $o^i_{k,n}$ - выходы модели

Архитектура модели



Критерии качества модели анализа выживаемости

Определение C-индекса

Пусть для пары абонентов (i, j) определены $\tau_i < \tau_j$ - моменты времени (возможно, цензурированные), δ_i - индекс цензурирования, равный 0, если τ_i право-цензурировано и 1 - в противном случае. Также обозначим через $\hat{F}_{k,i}$ и $\hat{F}_{k,j}$ вероятности оттока до момента по событию k времени τ_i для абонентов i и j соответственно (оцененные моделью функции распределения). Тогда C-индекс определяется следующим образом:

$$C\text{-}index = \frac{\sum_{i,j} \mathbf{1}_{\tau_i < \tau_j} \cdot \mathbf{1}_{\hat{F}_{k,i} > \hat{F}_{k,j}} \cdot \delta_i}{\sum_{i,j} \mathbf{1}_{\tau_i < \tau_j} \cdot \delta_i}$$

Понятие верно упорядоченной пары

Пара абонентов (i, j) считается верно упорядоченной (отранжированной), если для оцененных моделью функций распределений $\hat{F}_{k,i}$ и $\hat{F}_{k,j}$ выполнено неравенство:

$$\hat{F}_{k,i}(\tau_i) > \hat{F}_{k,j}(\tau_i)$$

Свойство добавки $\mathcal{L}_2$

Определение отступов

Пусть (i, j) - пара абонентов, для которых произошло событие оттока k в моменты времени $\tau^{(i)}$ и $\tau^{(j)}$ соответственно, причем $\tau^{(i)} < \tau^{(j)}$. Тогда $M_{k,i,j}$ определяется как:

$$M_{k,i,j} = \hat{F}_k(s^i + t^i_{j^i}|\mathcal{X}^i) - \hat{F}_k(s^i + t^j_{j^j}|\mathcal{X}^j)$$

Основное свойство

Пусть

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \eta\left(\hat{F}_k(s^i + t^i_{j^i}|\mathcal{X}^i), \hat{F}_k(s^i + t^j_{j^j}|\mathcal{X}^j))\right)$$

- добавка к функции потерь. Тогда $\mathcal{L}_2$ является убывающей от отступов функцией.

Выводы

1. Предложена нейросетевая модель анализа выживаемости с дискретным временем для решения задачи предсказания оттока абонентов
2. Построена специфическая для задачи функция потерь и обосновано свойство ее добавки $\mathcal{L}_2$
3. Проведен вычислительный эксперимент, в котором показано, что применение модели анализа выживаемости позволяет увеличить по сравнению с моделями градиентного бустинга:
   - Точность предсказания месяца оттока,
   - C-индекс.