Нейросетевые подходы к решению задачи оттока абонентов

A Preprint

Батарин Егор Владиславович
Кафедра алгоритмов и технологий программирования
Московский физико-технический институт
Москва
batarin.ev@phystech.edu

Джумакаев Тимур Казбекович
Мегафон
Москва

Abstract

В работе решается задача прогнозирования оттока абонентов компании Мегафон. Задача рассматривается как многоклассовая классификация, где в качестве меток класса выбраны факты оттока в будущие месяцы и факт отсутствия оттока в эти месяцы. Предлагаются различные подходы к решению задачи, как классические подходы: градиентный бустинг и модель Кокса, так и более современные подходы, связанные с применением методов глубокого обучения в моделях выживаемости. Проводится сравнение различных подходов с точки зрения принятых в работе критериев качества. В роли критериев качества модели выступают метрики Precision, Recall, F1, вычисленные при различных вероятностных порогах - числах, позволяющих перевести вероятности классов в метки классов. Цель работы заключается в поиске новых подходов к решению задачи оттока, которые покажут более высокие результаты по выбранным критериям качества, чем у текущих бейзлайнов. Эксперименты проведены на внутренних абонентских данных Мегафона.

Keywords CatBoost · Модель Кокса · Анализ выживаемости · DeepHit · Dynamic DeepHit

1 Введение

Данная работа посвящена решению задачи прогнозирования оттока в контексте телекоммуникационной отрасли в рамках проекта компании Мегафон. Данная задача позникает в большом количестве различных индустрий и для ее решения используются разные подходы. Ahn et al. [2020]. Среди всех таких подходов, включающих в себя разные современные методы машинного обучения (SVM, Random Forest, Gradient Boosting) основным направлением исследований для решения данной задачи был выбран анализ выживаемости.

Классической работой по анализу выживаемости является модель пропорциональных рисков Cox [1972]. В настоящее время появилось много новых подходов, так или иначе задействующих глубокое обучение Wiegrebe et al. [2024]. Эти подходы расширяют классические модели анализа выживаемости, позволяя обойти некоторые предположения о данных, которые на практике редко выполняются, например, предположение о пропорциональности рисков. Все модели анализа выживаемости можно разделить на две категории с зависимости от того, считается ли время непрерывным или дискретным. Поскольку специфика проекта Мегафона требует рассмотрения дискретного времени, то именно такой случай фигурирует в постановке задачи.

Одной из известных дискретных моделей выживаемости является модель DeepHit Lee et al. [2018], а также ее развитие - модель Dynamic DeepHit Lee et al. [2020]. В архитектурах обеих моделей используются нейросети, причем во второй модели помимо многослойного перцептрона применяется рекуррентная нейросеть RNN, а также механизм внимания. Данные работы взяты за основу для исследования в данной работе, их оригинальные версии и модицикации сравниваются для выявления лучшей модели.

Одним из самых распространенных подходов к задаче является градиентый бустинг Ahmad et al. [2019]. В данной работе в качестве базового подхода рассматривается категориальный бустинг от компании Яндекс Dorogush et al. [2018]. Он получил большое распространение на российском рынке, поэтому исследуемые в работе подходы, основанные на анализе выживаемости, сравниваются с CatBoost.

В работе используются внутренние датасеты компании Мегафон, собранные на основе данных в КХД. В роли критериев качества модели выступают метрики Precision, Recall, F1, вычисленные при различных вероятностных порогах - числах, позволяющих перевести вероятности классов в метки классов.

2 Постановка задачи

2.1 Общая постановка задачи анализа выживаемости

В дискретном случае время имеет вид $\mathcal{T} = \{0, \dots, T_{\max}\}$, где $T_{\max}$ - это максимальный горизонт предсказания (например, максимально возможное время жизни абонента). В любой их этих моментов может произойти событие $\mathcal{K} = \{\emptyset, 1, \cdots, K\}$, где все события $\{1, \cdots, K\}$ соответствуют факту оттока по одной из K возможных причин в некоторый момент времени τ, а событие $\emptyset$ означает факт правого цензурирования - информация о том, что отток произойдет не раньше того времени τ, когда произошло цензурирование, но точно неизвестно когда именно. Для каждого момента времени, таким образом, мы можем написать $\tau^i = \min(T^i, C^i)$, где $T^i \in \mathcal{T}$ - это времена наступлений одного из событий $\{1, \cdots, K\}$, а $C^i \in \mathcal{T}$ соответствует право-цензурированным событиям $\emptyset$. Имея в распоряжении информацию о произошедших событиях (включая цензурирования) $\{\tau^i, k^i\}_{i=1}^N$ и некоторую дополнительную информацию о признаках абонентов, мы хотим научиться предсказывать вероятности наступления события из $\mathcal{K}$ в будущем. В зависимости от того, рассматриваем ли мы признаки абонентов только в один момент времени (как в случае модели DeepHit) или рассматриваем для каждого абонента временной ряд соотвествующих ему признаков (как в случае модели Dynamic DeepHit), будет немного различаться математическая постановка задачи.

Специфичным для анализа выживаемости критерием качества является C-индекс (concordante index) Alabdallah et al. [2022], являющийся аналогом широкого известного критерия качества AUC. Он численно равен доле пар абонентов, которые модель верно упорядочила с точки зрения функции выживаемости. Пара (i, j), где $\tau_i < \tau_j$ - моменты наступлений события оттока k, считается верно упорядоченной, если для оцененных моделью условных функций выживания $\hat{F}_{k,i}$ и $\hat{F}_{k,j}$ при условии наступления признаков i и j соответственно, выполнено неравенство:

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

Этот индекс отражает то разумное требование к моделям выживаемости, которое заключается в том, что для абонентов, которые оттекают рано, оценочная условная функция распределения должна возрастать быстрее, так как для них целевое событие наступило при меньших значениях времени.

2.2 Модель DeepHit

В данной модели предполагается, что абонент полностью описывается вектором $\mathbf{x} \in X$, соответственно обучающая выборка имеет вид $\mathcal{D} = \{(\mathbf{x}^{(i)}, \tau^{(i)}, k^{(i)})\}_{i=1}^N$, вероятности нецензурированных событий $k^* \neq \emptyset$ имеют вид $P(\tau = \tau^*, k = k^* | \mathbf{x} = \mathbf{x}^*)$, а функция распределения имеет вид:

$$F_{k^*}(t^*|\mathbf{x}^*) = P(\tau \leq t^*, k = k^*|\mathbf{x} = \mathbf{x}^*) = \sum_{\tau^*=0}^{t^*} P(\tau = \tau^*, k = k^*|\mathbf{x} = \mathbf{x}^*). \tag{1}$$

Поскольку теоретическая функция распределения неизвестна, то рассматривается ее оценка

$$\hat{F}_{k^*}(\tau^*|\mathbf{x}^*) = \sum_{m=0}^{\tau^*} o^*_{k,m} \tag{2}$$

на основе ответов модели DeepHit: $\mathbf{o} = [o_{1,1}, \cdots, o_{1,T_{\max}}, \cdots, o_{K,1}, \cdots, o_{K,T_{\max}}]$, где $o_{k,\tau}$ - оценка моделью DeepHit вероятности того, что событие оттока k произойдет в момент времени τ.

В качестве функции потерь используется сумма двух слагаемых $\mathcal{L}_{\text{Total}} = \mathcal{L}_1 + \mathcal{L}_2$, в которой первое слагаемое имеет вид:

$$\mathcal{L}_1 = -\sum_{i=1}^{N}\left[\mathbb{1}(k^{(i)} \neq \emptyset)\cdot\log\left(y^{(i)}_{k^{(i)},\tau^{(i)}}\right) + \mathbb{1}(k^{(i)} = \emptyset)\cdot\log\left(1-\sum_{k=1}^{K}\hat{F}_k(\tau^{(i)}|\mathbf{x}^{(i)})\right)\right] \tag{3}$$

и оно отвечает за логарифмическое правдоподобие, а второе слагаемое имеет вид:

$$\mathcal{L}_2 = \sum_{k=1}^{K}\alpha_k\cdot\sum_{i\neq j}A_{k,i,j}\cdot\eta\left(\hat{F}_k(\tau^{(i)}|\mathbf{x}^{(i)}),\hat{F}_k(\tau^{(i)}|\mathbf{x}^{(j)})\right) \tag{4}$$

где $A_{k,i,j} = \mathbb{I}(k^{(i)} = k, \tau^{(i)} < \tau^{(j)})$ - индикатор того, что событие k наступает для j-ого абонента позже, чем для i-ого и функция $\eta(x,y) = \exp\left(\frac{-(x-y)}{\sigma}\right)$. Эта добавка тем меньше, чем лучше модель упорядочивает абонентов с точки зрения C-индекса.

2.3 Модель Dynamic DeepHit

Эта модель является расширением предыдущей и для каждого абонента i в ней рассматривается уже не единственный вектор-признак $\mathbf{x}_i$, а временной ряд векторов-признаков:

$$\mathcal{X}^i(t) = \{\mathbf{x}^i(t^i_j) : 0 \leq t^i_j \leq t \text{ for } j = 1,\cdots,J^i\}$$

, где $\mathbf{x}^i(t_j)$ - это вектор признак i-ого абонента, замеренный в момент времени t_j и равный $\mathbf{x}^i_j = [x^i_{j,1},\cdots,x^i_{j,d_r}]$. Кроме того, в этой модели для каждого абонента i вводится набор векторов-флагов $\mathrm{M}^i = \{\mathrm{m}^i_1,\cdots,\mathrm{m}^i_{J^i}\}$, $\mathbf{m}^i_j = [m^i_{j,1},\cdots,m^i_{j,d_x}]$, которые сигнализируют о пропущенных значениях: $m^i_{j,d} = 1$ тогда и только тогда, когда $x^i_{j,d}$ пропущено, иначе $m^i_{j,d} = 0$. Напоследок, вводится последовательность векторов временных интервалов между замерами времени $\Delta_i = \{\delta^i_1,\delta^i_2\cdots,\delta^i_{J^i}\}$, где $\delta^i_j = t^i_{j+1} - t^i_j$ для всех $1 \leq j < J^i$ и $\delta^i_{J^i} = 0$. В итоге получается обучающая выборка: $\mathcal{D} = \{(\mathbf{X}^i,\mathbf{M}^i,\Delta^i,\tau^i,k^i)\}_{i=1}^N$.

Теоретическая функция распределения в модели Dynamic DeepHit принимает вид:

$$\begin{aligned} F_{k^*}(\tau^*|\mathcal{X}^*) &= P(T\leq\tau^*, k=k^*|\mathcal{X}^*, T>t^*_{J^*}) = \\ &= \sum_{\tau\leq\tau^*}P(T=\tau,k=k^*|\mathcal{X}^*,T>t^*_{J^*}). \end{aligned} \tag{5}$$

Теоретическая функция выживания вычисляется следующим образом:

$$\begin{aligned} S(\tau^*|\mathcal{X}^*) &= P(T>\tau^*|\mathcal{X}^*,T>t^*_{J^*}) = \\ &= 1-\sum_{k\neq\emptyset}F_k(\tau^*|\mathcal{X}^*) \end{aligned} \tag{6}$$

Поскольку теоретические функции неизвестны, мы можем пользоваться только оценочными. Оценочная функция распределения выражается через ответы модели Dynamic DeepHit:

$$\hat{F}_{k^*}(\tau^*|\mathcal{X}^*) = \frac{\sum_{t^*_{J^*}<\tau\leq\tau^*}o^*_{k^*,\tau}}{1-\sum_{k\neq\emptyset}\sum_{n\leq t^*_{J^*}}o^*_{k,n}} \tag{7}$$

Функция потерь состоит из трех частей: $\mathcal{L}_{\text{Total}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$. Слагаемые $\mathcal{L}_1$ и $\mathcal{L}_2$ аналогичны соответствующим слагаемым из модели DeepHit и имеют вид:

$$\begin{aligned} \mathcal{L}_1 = -\sum_{i=1}^{N}\Bigg[&\mathbb{1}(k^i\neq\emptyset)\cdot\log\left(\frac{o^i_{k^i,\tau^i}}{1-\sum_{k\neq\emptyset}\sum_{n\leq t^i_{J_i}}o^i_{k,n}}\right)\Bigg] \\ &+\mathbb{1}(k^i=\emptyset)\cdot\log\left(1-\sum_{k\neq\emptyset}\hat{F}_k(\tau^i|\mathcal{X}^i)\right)\Bigg] \end{aligned} \tag{8}$$

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} A_{kij} \cdot \eta \left(\hat{F}_k(s^i + t^i_{J^i} | \mathcal{X}^i), \hat{F}_k(s^i + t^j_{J^j} | \mathcal{X}^j) \right) \tag{9}$$

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

Третье слагаемое в общей функции потерь является новым и отвечает за регуляризацию временных рядов:

$$\mathcal{L}_3 = \beta \cdot \sum_{i=1}^{N} \sum_{j=0}^{J^i-1} \sum_{d \in \mathcal{I}} (1 - m^i_{j+1,d}) \cdot \zeta(x^i_{j+1,d}, y^i_{j,d}) \tag{10}$$

Здесь $\mathcal{I}$ определяет подмножество зависящих от времени признаков абонентов, по которым мы хотим провести регуляризацию.

Список литературы

Jaehyun Ahn, Junsik Hwang, Doyoung Kim, Hyukgeun Choi, and Shinjin Kang. A survey on churn analysis in various business domains. IEEE Access, 8:220816–220839, 2020. ISSN 2169-3536. doi:10.1109/access.2020.3042657.

D. R. Cox. Regression models and life-tables. Journal of the Royal Statistical Society Series B: Statistical Methodology, 34(2):187–202, January 1972. ISSN 1467-9868. doi:10.1111/j.2517-6161.1972.tb00899.x.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. Artificial Intelligence Review, 57(3), February 2024. ISSN 1573-7462. doi:10.1007/s10462-023-10681-3.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), April 2018. ISSN 2159-5399. doi:10.1609/aaai.v32i1.11842.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Transactions on Biomedical Engineering, 67(1):122–133, January 2020. ISSN 1558-2531. doi:10.1109/tbme.2019.2909027.

Abdelrahim Kasem Ahmad, Assef Jafar, and Kadan Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), March 2019. ISSN 2196-1115. doi:10.1186/s40537-019-0191-6.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support, 2018.

Abdallah Alabdallah, Mattias Ohlsson, Sepideh Pashami, and Thorsteinn Rögnvaldsson. The concordance index decomposition: A measure for a deeper understanding of survival prediction models. 2022. doi:10.48550/ARXIV.2203.00144.