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.

2.2 C-индекс и его связь с ранжированием абонентов

Специфичным для анализа выживаемости критерием качества является C-индекс (concordante index) Alabdallah et al. [2022], являющийся аналогом широкого известного критерия качества AUC. Дадим его строгое определение:

Определение 1. 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:

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

Нам также понадобится определение верно упорядоченной пары абонентов:

Определение 2. 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:

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

Он численно равен доле пар абонентов, которые модель верно упорядочила.

Этот индекс отражает то разумное требование к моделям выживаемости, которое заключается в том, что если абонент i оттекает раньше абонента j, то оцениваемая моделью вероятность его оттока до момента его фактического оттока будет больше, чем аналогичная вероятность для j абонента, рассматриваемая для того же момента времени - фактического оттока i-ого абонента.

2.3 Постановка с одним вектором-признаком

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:

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

2.4 Поставнока в временным рядом векторов-признаков

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:

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

, где $\mathbf{x}^i(t_j)$ - это вектор признак i-ого абонента, замеренный в момент времени t_j и равный $\mathbf{x}_j^i = [x_{j,1}^i, \cdots, x_{j,d_r}^i]$. Кроме того, в этой модели для каждого абонента i вводится набор векторов-флагов $\mathrm{M}^i = \{\mathrm{m}_1^i, \cdots, \mathrm{m}_{J^i}^i\}$, $\mathbf{m}_j^i = [m_{j,1}^i, \cdots, m_{j,d_x}^i]$, которые сигнализируют о пропущенных значениях: $m_{j,d}^i = 1$ тогда и только тогда, когда $x_{j,d}^i$ пропущено, иначе $m_{j,d}^i = 0$. Напоследок, вводится последовательность векторов временных интервалов между замерами времени $\Delta_i = \{\delta_1^i, \delta_2^i \cdots, \delta_{J^i}^i\}$, где $\delta_j^i = t_{j+1}^i - t_j^i$ для всех $1 \leq j < J^i$ и $\delta_{J^i}^i = 0$. В итоге получается обучающая выборка: $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{M}^i, \Delta^i, \tau^i, k^i)\}_{i=1}^{N}$.

3 Описание подходов к решению задач

3.1 Решение задачи с одним вектором-признаком

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:

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

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. Ниже мы обсудим это свойство более подробно и формально его докажем.

3.2 Решение задачи для временного ряда векторов-признаков

Теоретическая функция распределения в модели Dynamic DeepHit принимает вид:

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

Теоретическая функция выживания вычисляется следующим образом:

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

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:

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

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

, 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.

3.3 Свойства добавок $\mathcal{L}_2$

Выше мы упомянули, что $\mathcal{L}_2$ обладает ранжирующим свойством с точки зрения C-индекса. Сформулируем строго это утверждение и докажем его для двух вышеописанных постановок задач. Прежде введем понятие отступа, которое аналогично по смыслу понятию отступа в задаче классификации:

Определение 3. 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 в случае постановки с одним вектором-признаком отступ $M_{k,i,j}$ определяется как:

$$M_{k,i,j} = \hat{F}_k(\tau^{(i)}|\mathbf{x}^{(i)}) - \hat{F}_k(\tau^{(i)}|\mathbf{x}^{(j)})$$

С случае постановки для временного ряда векторов-признаков отступ $M_{k,i,j}$ определяется как:

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

В задаче классификации отступ является мерой того, насколько уверенно модель правильно классифицирует объекты. В нашей постановке анализа выживаемости отступ является мерой того, насколько модель уверенно правильно упорядочивает абонентов. Теоремы ниже являются точным выражением того, что добавка $\mathcal{L}_2$ тем ниже, чем выше отступы $M_{k,i,j}$.

Теорема 1. Пусть $\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \eta \left(\hat{F}_k(\tau^{(i)}|\mathbf{x}^{(i)}), \hat{F}_k(\tau^{(i)}|\mathbf{x}^{(j)}) \right)$ - добавка к функции потерь в смысле постановки с одним вектором-признаком. Тогда $\mathcal{L}_2$ является убывающей от отступов функцией.

Теорема 2. 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 в смысле постановки для временного ряда векторов-признаков. Тогда $\mathcal{L}_2$ является убывающей от отступов функцией.

Доказательство. В обеих постановках добавка принимает вид:

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \exp\left(-\frac{M_{k,i,j}}{\sigma} \right)$$

поэтому выражения для производных принимают вид:

$$\frac{\partial \mathcal{L}_2}{\partial M_{k,i,j}} = -\frac{1}{\sigma} \sum_{k=1}^{K} \alpha_k \cdot A_{k,i,j} \cdot \exp\left(-\frac{M_{k,i,j}}{\sigma} \right) < 0$$

откуда следуют утверждения теорем. □

Таким образом, смысл добавки $\mathcal{L}_2$ сводится к штрафованию модели появлению маленьких отступов, которые связаны с нарушением верной упорядоченности абонентов и понижением C-индекса.

4 Численные эксперименты

4.1 Описание поставноки задачи

В численных экспериментах рассматривался частный случай общей поставноки задачи при котором $K = 1$ - события оттока не различаются между собой. При этом производительность моделей сравнивалась на задачах бинарной и многоклассовой классификации:

1. В задаче бинарной классификации предсказываются вероятности двух классов - сохранение абонента в следующий месяц (таргет $q = 0$) и отток абонента в следующий месяц (таргет $q = 1$)

2. В задаче многоклассовой классификации предсказываются вероятности следующих четырех классов: отток в конец текущего месяца (таргет $q = 0$), отток в конец следующего месяца (таргет $q = 1$), отток в конец послеследующего месяца (таргет $q = 3$) и сохранение абонента на конец послеследующего месяца (таргет $q = 4$)

4.2 Описание данных

Эксперименты проводились на данных компании Мегафон, собранных с начала апреля 2024 до конца октября 2024.

Обучающая выборка состоит из 2.7 миллионов примеров, валидационная и тестовая выборки содержат по 600 и 800 тыс. обучающих примеров соответственно.

Для модели DeepHit обучающая выборка предварительно была нормирована. Для CatBoost нормировка не проводилась. Для подбора гиперпараметров CatBoost была использована библиотека optuna. Для архитектуры DeepHit был использован многослойный перцептрон.

4.3 Критерии качества

Для анализа качества была использована функция reports из модуля scoring, которая строит отчеты-таблицы по моделям. Отчет-таблица демонстрирует, насколько хорошо модель отделяет каждый из

классов от всех остальных (One-vs-All подход). Опишем структуру этого отчета. Для начала введем понятие топ перцентиля:

Определение 4. Топ-p перцентиль для некоторого класса q определяется как квантиль уровня $1-p$ для предсказанных моделью вероятностей класса q.

В рамках One-vs-All подхода, по определению будем считать, что если предсказанная моделью вероятность класса q больше порога в топ-p перцентиль для некоторого заранее фиксированного p, то ответом модели будет класс q, иначе ответом будут все классы, кроме q.

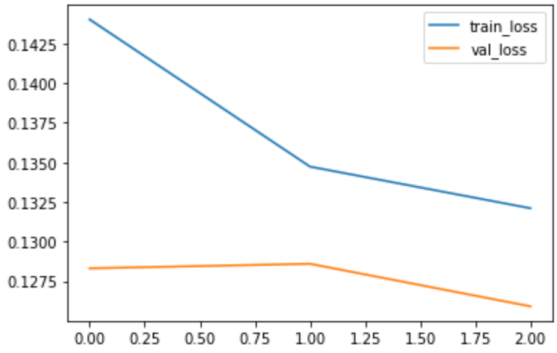
В случае задачи бинарной классификации рассматриваются только топ перцентили для класса $q = 1$, соответствующему оттоку абонента, а в случае многоклассовой классификации - топ-перцентили для всех классов. При этом в многоклассовой постановке каждому классу соответствует свой отчет, показывающий, как хорошо данный класс отделяется от всех остальных.

Стоблцы таблицы соответствуют широко распространенным метрикам классификации - Precision, Recall, F1 и AUC. Строки таблицы соответствуют различным топ-p перцентилям для данного класса. Таким образом можно сравнивать метрики в зависимости от различных порогов вероятностей.

4.4 Сравнение CatBoost и DeepHit в задаче бинарной классификации

4.4.1 Число эпох обучения = 3

Для начала было использовано три эпохи обучения. График функции потерь имеет вид:



Метрики на ТОП 10% оказались у DeepHit оказались хуже, чем у бейзлайновой CatBoost.

	metrics	precision	recall	f1	roc-auc
score date	percentile				
1	0.500	0.986	0.010	0.019	0.505
	1.000	0.982	0.019	0.038	0.510
	2.000	0.977	0.039	0.074	0.519
	3.000	0.975	0.058	0.109	0.528
	4.000	0.972	0.077	0.142	0.537
	5.000	0.969	0.096	0.174	0.546
	10.000	0.959	0.190	0.317	0.591

	metrics	precision	recall	f1	roc-auc
score date	percentile				
1	0.500	0.997	0.010	0.020	0.505
	1.000	0.994	0.020	0.039	0.510
	2.000	0.992	0.039	0.076	0.519
	3.000	0.990	0.059	0.111	0.529
	4.000	0.988	0.078	0.145	0.539
	5.000	0.987	0.098	0.178	0.548
	10.000	0.979	0.194	0.323	0.595

4.4.2 Число эпох обучения = 50

На этом этапе мы заметно увеличили число эпох обучения DeepHit.

4.5 Сравнение CatBoost и DeepHit в задаче многоклассовой классификации

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.