Нейросетевые модели анализа выживаемости для решения задачи предсказания оттока абонентов

Е. В. Батарин

Московский физико-технический институт

24 июня 2025 г.

Нейросетевые модели анализа выживаемости для решения задачи предсказания оттока абонентов

Задача

Исследовать качество различных моделей 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 со специально подобранной функцией потерь

Обозначения

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

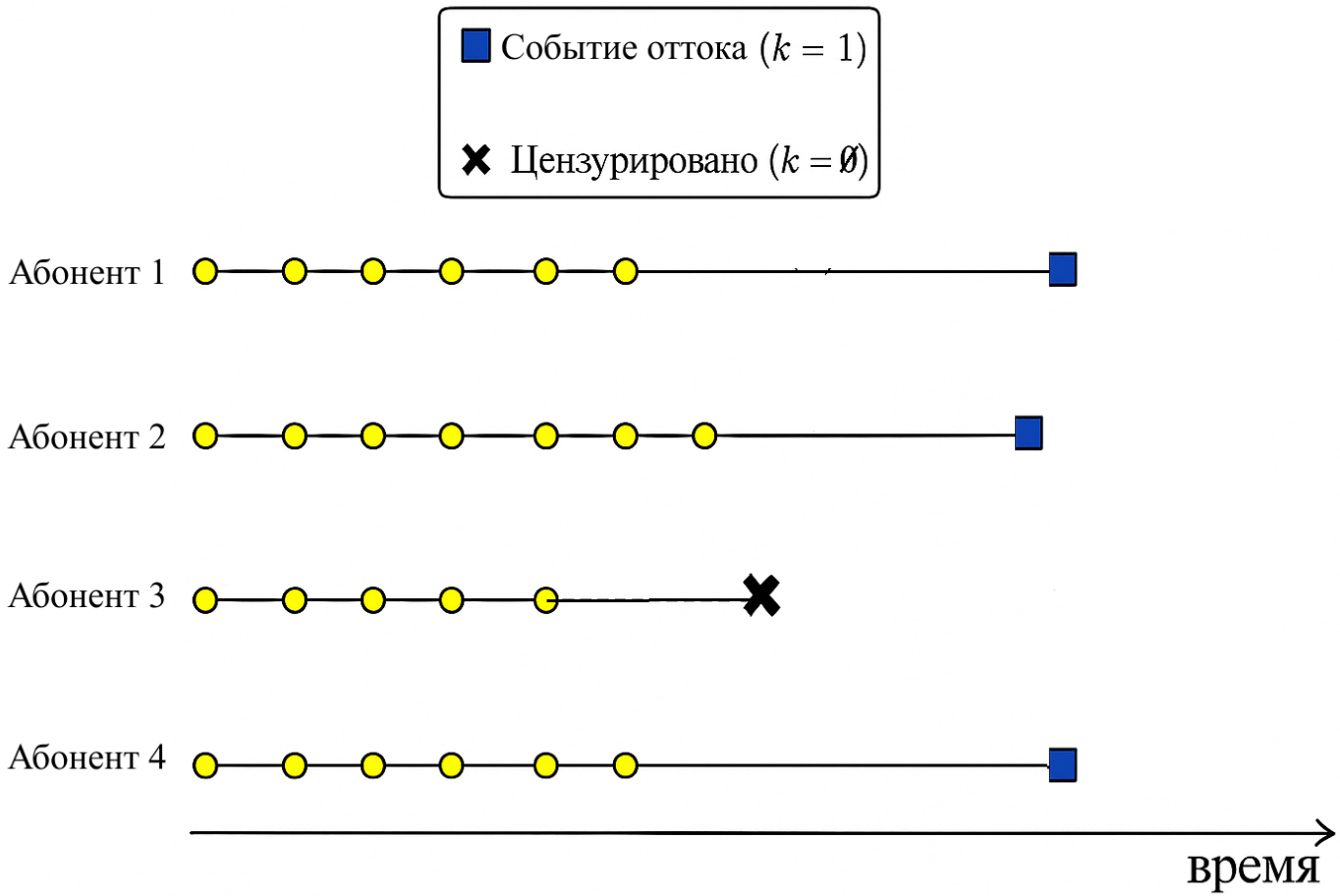
$$F_{k^*}(\tau^*|\mathcal{X}^*) = P(T \leq \tau^*, k = k^*|\mathcal{X}^*, T > t^*_{J^*}) = \sum_{\tau \leq \tau^*} P(T = \tau, k = k^*|\mathcal{X}^*, T > t^*_{J^*}).$$

Функция распределения для события k^*

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

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:

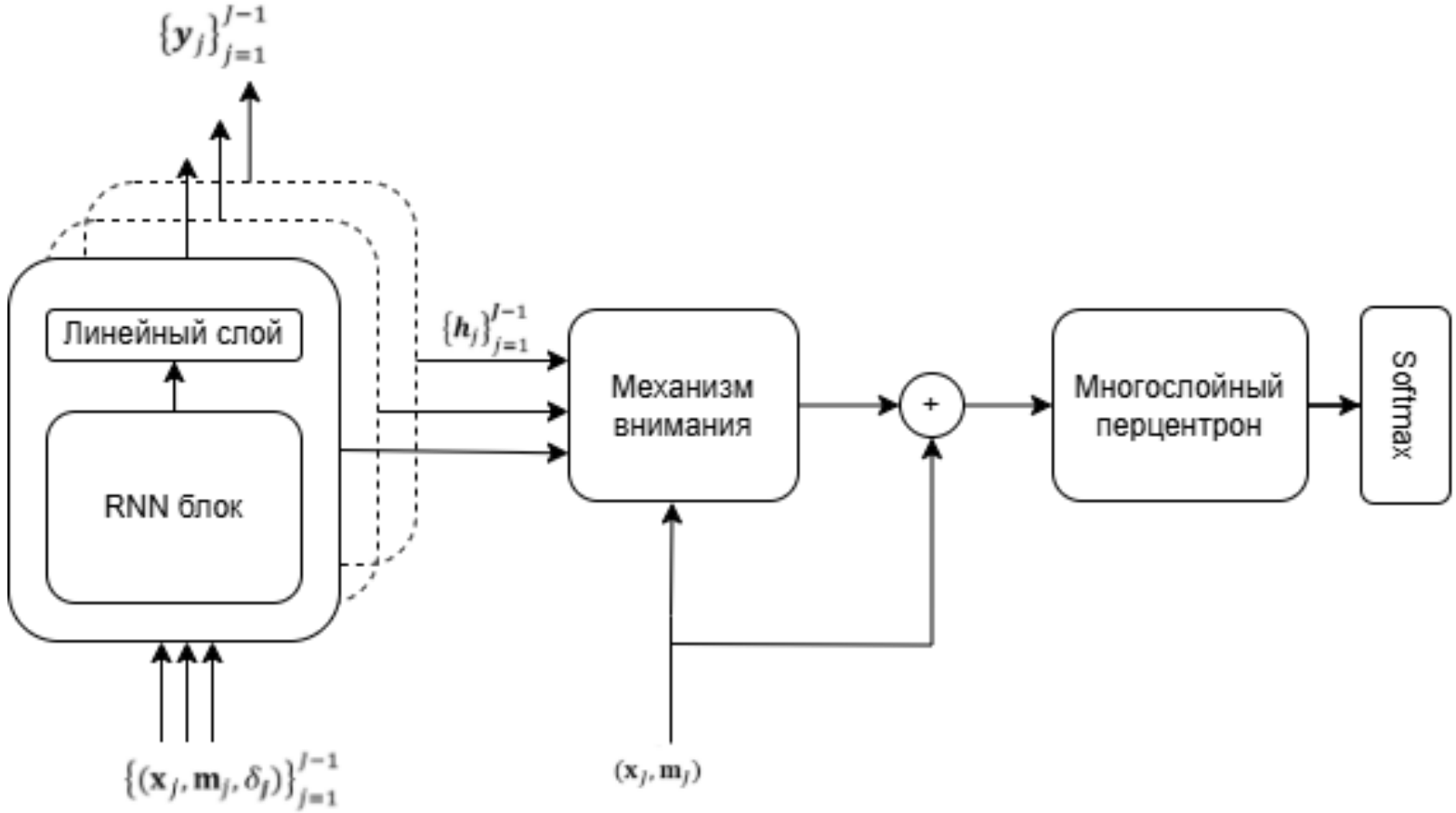
$$\mathcal{L}_1 = -\sum_{i=1}^{N} \left[\mathbb{1}(k^i \neq \emptyset) \cdot \log \left(\frac{o^i_{k^i, \tau^i}}{1 - \sum_{k \neq \emptyset} \sum_{n \leq t^i_{Ji}} o^i_{k,n}} \right) + \mathbb{1}(k^i = \emptyset) \cdot \log \left(1 - \sum_{k \neq \emptyset} \hat{F}_k(\tau^i|\mathcal{X}^i) \right) \right]$$

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

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

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



Для обучения используются процедуры Grid Search и Learning Rate Finder

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:

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

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.

Описание вычислительного эксперимента

Цель эксперимента

1) Проследить влияние новых признаков - выходов предложенной модели - на точность решения задачи классификации абонентов
2) Оценить значимость новых признаков

Описание датасета

1) Базовый сегмент: обучение - B2C абоненты Мегафона в июне и июле 2024, тест - в августе 2024. Обучающая выборка содержит 12 миллионов примеров и сбалансирована. Тестовая выборка содержит 3 миллиона примеров и не сбалансирована
2) Целевое событие: 4 класса - отток в конце 1, 2 и 3 месяц и отсутствие оттока
3) Исходные признаки: 50 признаков абоненской активности

Выходной вектор модели

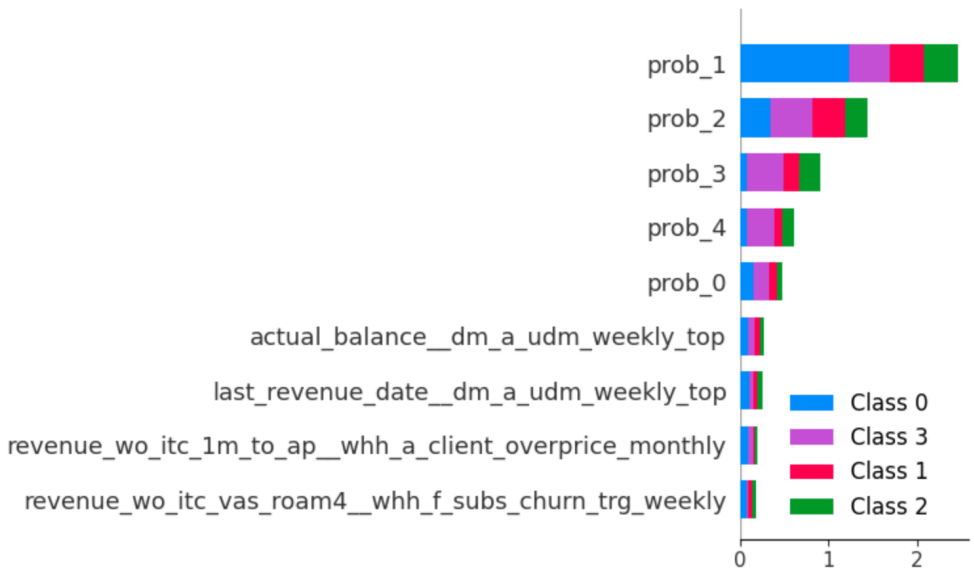
Новые признаки, генерируемые моделью, представляют собой вектор $(x_0, x_1, x_2, x_3, x_4) \in \mathbb{R}^5$, где x_i - вероятность выживания в i-ый месяц, где за 0 месяц берется июнь 2024

Результаты вычислительных экспериментов

Точность предсказания

Название модели	Исходная точность	Точность на 5 признаках	Потеря точности	C-индекс
DeepHit	80.6%	73.1%	7.5%	0.74
PMF		75.2%	5.4%	0.75
Nnet-survival		74.8%	5.8%	0.77
Наша модель		78.5%	2.1%	0.81

Значимость признаков



Интерпретация плотностей вероятностей - июнь и июль

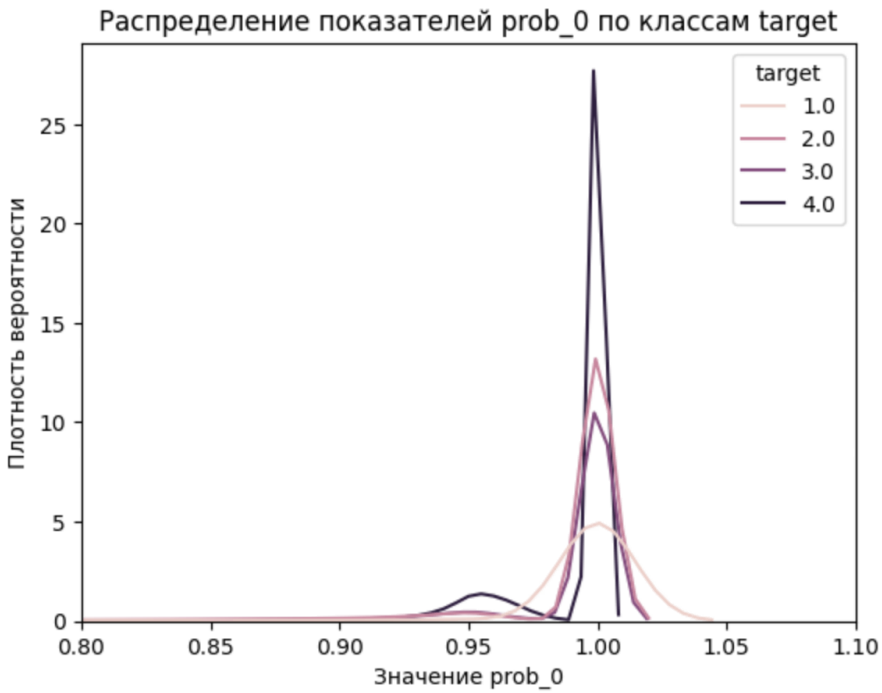


Рис.: Июнь 2024

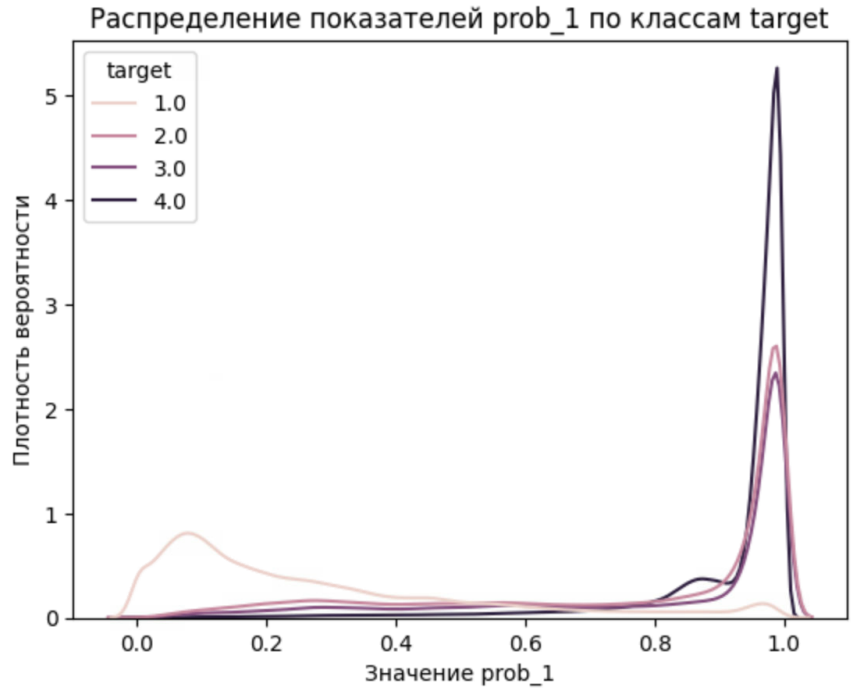


Рис.: Июль 2024

Интерпретация плотностей вероятностей - август и сентябрь

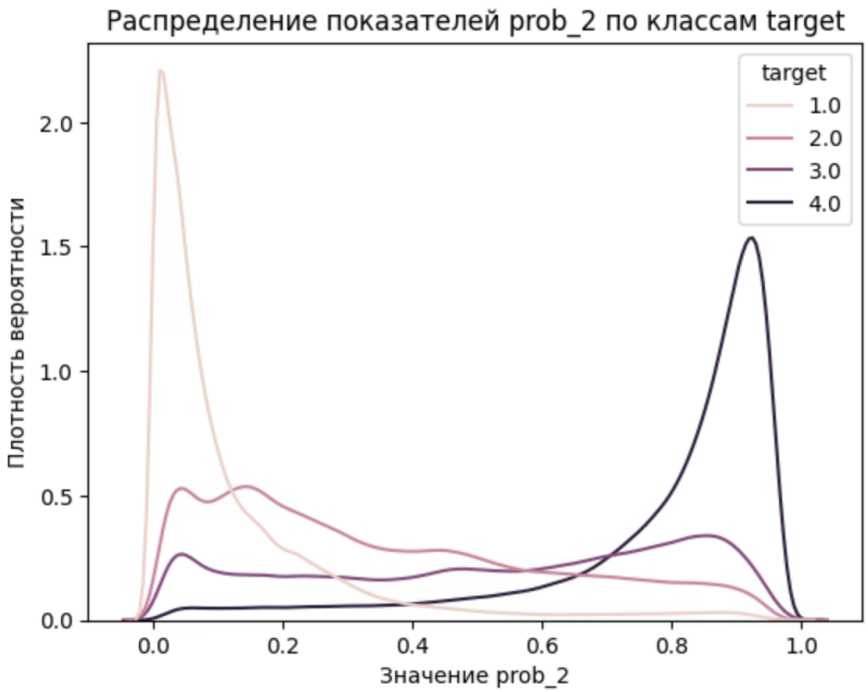


Рис.: Август 2024

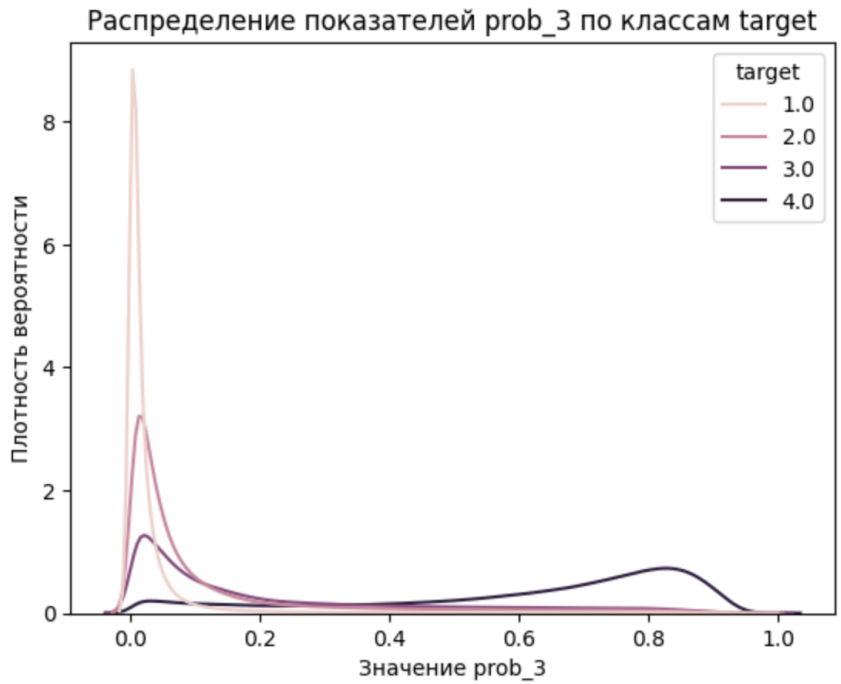


Рис.: Сентябрь 2024

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 получить новые признаки, которые:
   - Имеют наибольшую SHAP-значимость
   - Вносят вклад в точность предсказаний, сопоставимый с исходными признаками, причем потеря в точности минимальна по сравнению с остальными моделями
4. Исследованы свойства модели по кластеризации абонентов:
   - Модель хорошо отделяет абонентов отточников от абонентов сохраняющихся
   - Модель плохо отделяет отточников между собой, что связано с медленным изменением абоненстких данных на месячном периоде