Оглавление

[Оглавление 1](#_Toc39176081)

[Введение 3](#_Toc39176082)

[Глава 1. Описание и анализ предметной области 4](#_Toc39176083)

[1.1. Онлайн-реклама 4](#_Toc39176084)

[1.2. Онлайн аукцион 4](#_Toc39176085)

[1.3. Машинное обучение 6](#_Toc39176086)

[1.4. Survival analysis 7](#_Toc39176087)

[1.5. Постановка задачи 7](#_Toc39176088)

[Выводы по главе 1 8](#_Toc39176089)

[Глава 2. Теоретическое исследование 9](#_Toc39176090)

[2.1. Анализ существующих решений 9](#_Toc39176091)

[2.1.1. Аналитическое решение 9](#_Toc39176092)

[2.1.2. Решение с использованием решающих деревьев 10](#_Toc39176093)

[2.2. Решение с использованием глубокого обучения. 11](#_Toc39176094)

[2.2.1. Сведение Survival Analysis к задаче предсказания вероятностей 11](#_Toc39176095)

[2.2.2. Описание работы нейронной сети 13](#_Toc39176096)

[2.3. Описание решения 14](#_Toc39176097)

[2.3.1. Устройство нейронной сети 14](#_Toc39176098)

[2.3.2. Обучение нейронной сети 15](#_Toc39176099)

[2.4. Подходы к улучшению показателей нейронное сети 17](#_Toc39176100)

[2.4.1. Механизм внимания 17](#_Toc39176101)

[2.4.2. Pruning нейронной сети 18](#_Toc39176102)

[2.4.3. Улучшение loss-функции 18](#_Toc39176103)

[2.5. Использование альтернативного решения из Survival Analysis 18](#_Toc39176104)

[Выводы по главе 2 18](#_Toc39176105)

[Глава 3. Практическое исследование 19](#_Toc39176106)

[3.1. Описание датасета 19](#_Toc39176107)

[3.2. Метрика качества модели 19](#_Toc39176108)

[3.3. Сравнение с актуальным решением 19](#_Toc39176109)

[3.4. Результаты экспериментов 21](#_Toc39176110)

[Выводы по главе 3 21](#_Toc39176111)

[Заключение 22](#_Toc39176112)

[Библиографический список 23](#_Toc39176113)

[Приложения 24](#_Toc39176114)

Введение

Аукцион в реальном времени (англ. RTB) появился в 2009 году и стал одним из самых важных механизмов в онлайн-рекламе. Рекламы начали платить за каждый аукцион отдельно, основываясь на своих стратегий торгов. От качества выбранной стратегии напрямую зависит эффективность рекламы.

Ежедневно в сети проходит огромное количество аукционов и проследить стратегию каждого отдельного игрока не представляется возможным. Поэтому можно подстроить алгоритм, который будет предсказывать количество выигранных аукционов для каждой ставки. Основываясь на этой статистике, рекламодатели или рекламные агентства смогут построить свою стратегию ведения торгов для повышения показателей эффективности каждого рекламного объявления.

# Описание и анализ предметной области

Введем основные понятия и определения из предметной области, которые необходимы для описания постановленной задачи.

## Онлайн-реклама

Определение 1. Онлайн-реклама – это форма маркетинга, которая использует интернет, чтобы доставить маркетинговое сообщения до потенциального покупателя. [1]

Онлайн-рекламу можно увидеть на большинстве интернет-страницах в виде баннера (что приведено на рисунке ниже). Контекст и тематика рекламы чаще всего совпадает со сферой интересов посетителей сайта, чтобы увеличить показатели эффективности объявления.

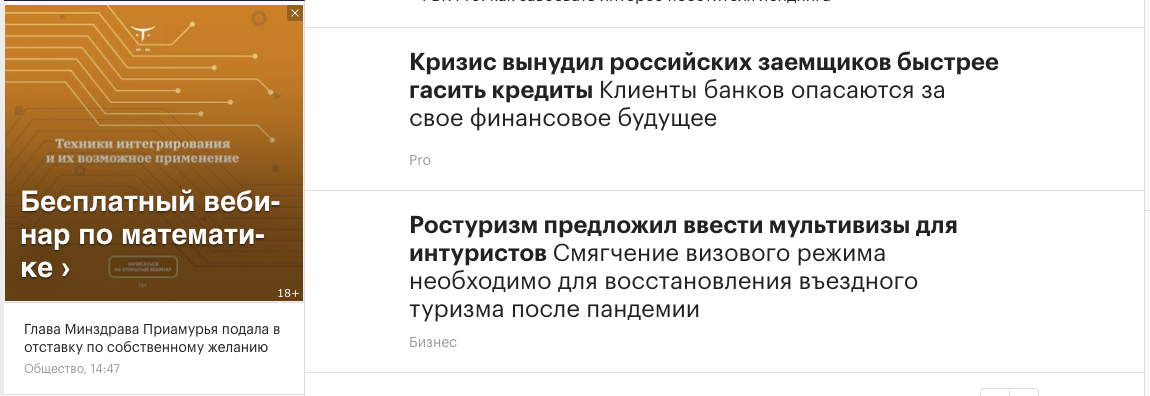


Рисунок 1 – Пример рекламного баннера на интернет-странице rbc.ru

## Онлайн аукцион

Определение 2. Онлайн-аукцион – процесс выбора рекламного объявления для показа пользователю. Кандидаты на показ предлагают свою ставку и выигрывает участник с наибольшей ценой.

Аукционы могут быть открытыми или закрытыми. В закрытом аукционе участники не видят ставку своих оппонентов и не могут изменять свои ставки, в отличии от открытого аукциона, где все участники видят ставки друг друга. [2]

Определение 3. Аукцион первой цены – закрытый аукцион, при котором победителем является участник с самой высокой ценой и именно эта цена подлежит уплате. [2]

Определение 4. Аукцион второй цены – закрытый аукцион, при котором победителем является участник с самой высокой ценой, но уплатить он должен «вторую цену», то есть цену своего ближайшего конкурента. [2]

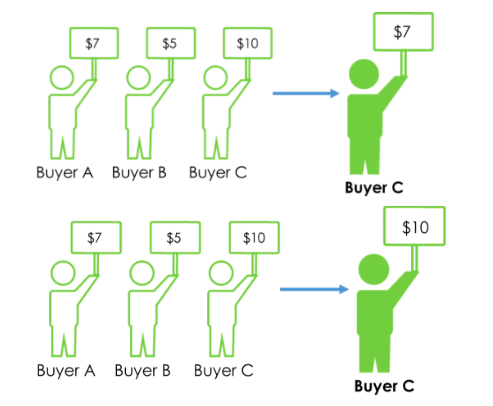


Рисунок 2 – Пример аукциона второй и первой цены [3]

Определение 5. Рыночная цена – цена, которую платит победитель аукциона.

Определение 6. RTB (Real Time Bidding) – технология закупки медийной рекламы посредством программируемых онлайн-аукционов.

RTB фокусируется непосредственно на показах целевым посетителям, а не планированию резервов рекламных площадей на определенных сайтах. Каждый показ выкупается за доли секунды – во время загрузки страницы – система. RTB мгновенно проводит аукцион. В результате лучшее предложение от рекламодателей появляется на глазах пользователя, которому оно наиболее интересно. [4]



Рисунок 3 – Механика работы RTB [4]

Ставка в аукционе основывается на информации о пользователе, который заходит на веб-страницу, и потенциальной выгоде, которую он может принести.

О пользователе рекламодатели знают информацию о дате и времени захода на сайт, регион, URL сайта, где будет выставлена реклама, размер рекламного баннера и интересы пользователя (таргеринг). Далее эту информацию будем называть как запрос для ставки.

Выгода от рекламного объявления является позитивным ответом пользователя, показатели CTR и CVR.

Определение. CTR (англ. Click-through rate) – метрика в интернет-маркетинге, которая определяется как отношение числа кликов на баннер или рекламное объявление к числу показов. [5]

Определение. CVR (англ. Conversion rate) – конверсия.

## Машинное обучение

Определение 7. Машинное обучение – подраздел искусственного интеллекта, изучающий обучающиеся методы построения алгоритмов. [6]

Определение 8. Обучение с учителем (англ. Supervised learning) – один из способов машинного обучения, в ходе которого испытуемая система обучается с помощью пары «объект, ответ». Цель обучения в восстановлении зависимости между множеством «объектов» и «ответов».

(Будет написано дополнительно про Нейронные сети, RNN, внимание, кластеризацию, KL-divergence, метрики (точность и т.д.))

## Survival analysis

Survival analysis – это класс статических моделей, позволяющий оценить вероятность наступления событий. [7][8]

Цель анализа в оценке времени, когда произойдет интересующее событие.

Функция выживания (англ. Survival Function) определяет вероятность того, что интересующее событие не произойдет в момент времени .

Функция – плотность распределения вероятности наступления события в момент времени . В Survival analysis называется, как rate of death или failure events per unit time.

Функция условной вероятности определяет, что событие произойдет в рассматриваемый интервал времени, при условии, что оно не произошло до этого интервала.

Определение. Оценка Каплана-Мейера – непараметрическая функция, используется для оценки Функции выживания (**вставить номер)**. [9]

## Постановка задачи

Цель данной работы улучшить предсказание плотности распределения рыночной цены в аукционе второй цены.

Более формально. По запросу (информации о пользователе) получить распределение рыночной цены .

Для решения будет использоваться истории ставок в аукционе, каждый пример можно представить как тройку значений , где – информация о пользователе, – рыночная цена аукциона, b – ставка рекламодателя.

Аукцион закрытый, поэтому рыночная цена известна, только если рекламодатель выиграет аукцион. То есть в случае .

С помощью плотности распределения можно будет рассчитывать вероятность выигрыша и проигрыша аукциона для любой интересующей ставки.

Будем решать сведением задачи к Survival analysis и использованием глубоких нейронных сетей.

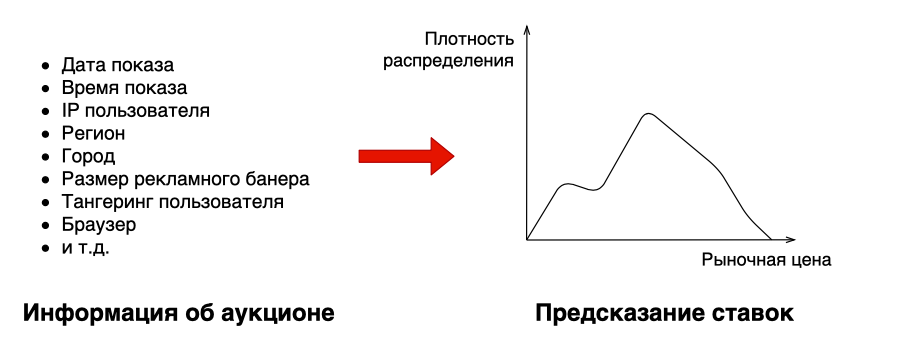


Рисунок 4 – Задача предсказания ставок в аукционе

Выводы по главе 1

В данной главе были определены основные понятия и определения об онлайн-рекламе и онлайн-аукционе, рассмотрена задача Survival analysis и поставлена задача.

# Теоретическое исследование

В первой главы были введены основные определения и сформулирована задача данной работы. В этой главе будут рассмотрены существующие решения этой задачи и новые подходы для улучшения существующих результатов.

## Анализ существующих решений

В этом разделе будут описаны решения задачи предсказания ставок в аукционе, с помощью сведения к Survival Analysis и без использования нейронных сетей.

### Аналитическое решение

Рассмотрим статью 2016 года, которая была опубликована авторами Weinan Zhang, Tianxiong Zhou, Jun Wang, Jian Xu. В этой статье не использовалась информация о пользователе, а учитывалась только ставка рекламодателя и рыночная цена аукциона. [11]

В этой статье решили задачу нахождения вероятности проигрыша/выигрыша аукциона с заданной ставкой с помощью сведения к Survival analysis.

Переведем тройку значений , где – ставка, – рыночная цена, – выигрыш/проигрыш аукциона, в формат . Где – число выигранных аукционов с рыночной ценой равной (по аналогии, интересующее нам событие происходит в день ). – число аукционов, которые не могут быть выиграны со ставкой , то есть число выигранных аукционов с рыночной ценой не менее и число проигранных, в которых ставка была не менее . А саму ставку можно рассматривать как день наблюдения.

Тогда вероятность проигрыша аукциона со ставкой , используя формулу (**номер формы КМ**), будет равна:

Что в Survival Analysis аналогично вероятности того, что интересующее событие не произойдет в интервале от 1 до .

Вероятность выигрыша будет равна:

Плотность распределения рыночной цены:

### Решение с использованием решающих деревьев

Большой недостаток решения, описанного в прошлом разделе, в том, что не используется информация о пользователе. В статье, опубликованной в 2016 году, для это была использована задача кластеризации. [12]

Для получения распределение рыночной цены используется бинарное дерево решений . Главная идея в кластеризации данных, используя информацию о пользователях. Для этого используется метод k-средних и расстояние Кульбака-Лейблера.

При построении дерева в каждом узле будет выбираться разбиение с максимальным расстоянием Кульбака-Лейблера . Математически алгоритм разбиения будет выглядеть следующим образом.

Где и – это вероятность распределения разбиения, – наибольшая рыночная цена, – наибольшее расстояние Кульбака-Лейблера среди всех признаков в -ом узле, - количество признаков, – количество разбиений.

Тогда для предсказание будет выбираться кластер, проходом по дереву, и в листе считаться .Сама плотность распределения считается аналогично статье **(номер статьи)**, также задача сводиться к Survival analysis, только статистические данные считаются для каждого кластера отдельно.

## Решение с использованием глубокого обучения.

Главная цель этой работы состоит в улучшении показателей решения, которого будет описано в этом разделе. В 2019 году студенты Китайского университета Шанхая опубликовали статью про использование глубоких рекуррентных нейронных сетей в Survival Analysis, и также сделали отдельную публикацию использования этого решения в предсказании рыночной цены в аукционах. [13][14]

В первой статье были опубликованы показатели применения нового решения в разных областях: в медицине, музыкальной индустрии и предсказании ставок в аукционе. По результатам эксперимента это решение выдала лучшие показатели ROC-кривой (как бинарного классификатора) и ANLP (average negative log probability) среди всех известных на тот момент решений.

В следующей главе, это решение будет взято за baseline и именно его мы будем пытаться улучшить.

### Сведение Survival Analysis к задаче предсказания вероятностей

Во второй статье было описано как задачу Survival Analysis свели к нахождению распределения вероятности рыночной цены и поставленную задачу решили с помощью глубоких нейронных сетей.

Для введения вероятностей выигрыша и проигрыша в аукционе с заданной ставкой была использована формула **(номер формулы)** Survival function.

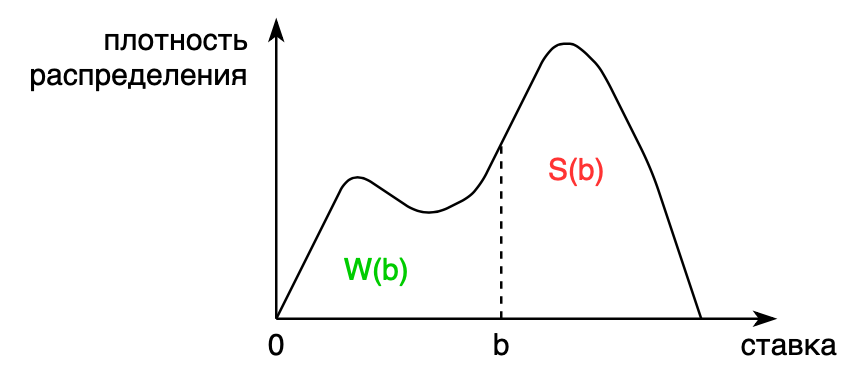


Рисунок 5 – Плотность распределения в непрерывном пространстве

Чтобы нейронная сеть смогла обучаться предсказывать плотность распределения, необходимо перейти из непрерывного пространства в дискретное.

Для этого представим ставки как последовательность от наименьшей до наибольшей и выберем блоков таких, что для блок .

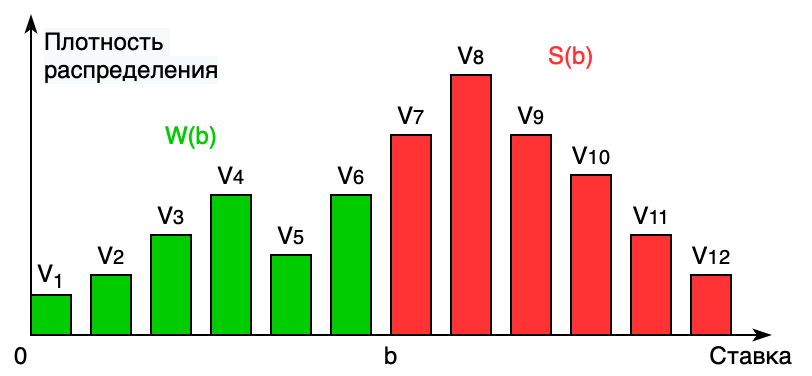


Рисунок 6 – Плотность распределения в дискретном пространстве

Все ставки принадлежал множеству натуральных чисел и для лучшего предсказания .

Тогда формулы вероятностей будут иметь следующий вид.

Введем формулу условной вероятности выигрыша в аукционе для ставки . Это значение будет предсказывать нейронная сеть.

### Описание работы нейронной сети

Для предсказания условной вероятности используется – функция рекуррентной нейронной сети, которая приминается пару значений , где – информация о пользователе, а – ставка.

В выходе нейронной сети блоков – условная вероятность выигрыша аукциона для пользователя с информацией и ставкой . Используя значения об условной вероятности, можно найти вероятности выигрыша, проигрыша и плотность распределения рыночной цены.

Формула условной вероятности для и блока . – скрытый вектор из предыдущего рекуррентного блока.

Вероятность проигрыша и выигрыша в аукционе для примера , со ставкой будет равны:

И плотность распределения рыночной цены будет равна:

## Описание решения

В данном разделе будет описания решения, которое реализовано на данный момент, устройство нейронной сети, обучение и метрики для проверки качества.

### Устройство нейронной сети

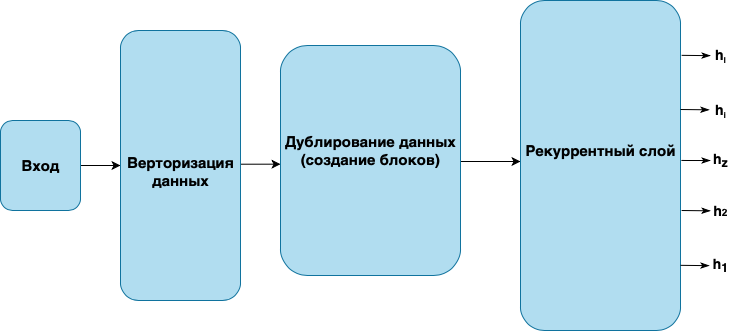


Рисунок 7 – Иллюстрация устройства нейронной сети.

Для наглядности разобьем нашу нейронную сеть на три части.

В первой части модели происходит перевод переданных примеров в формат, необходимый для вычисления. Этот процесс выполняют первые два слоя: Embedding переводит слова в векторное представление, Dense слой уменьшает размерность примеров, для увеличения производительности.

Во второй части переводим данных в формат, который будет удобен для предсказания ставки. Переданные данные дублируются ровно столько раз, сколько мы хотим предсказать условных вероятностей. То есть в этом слое создаются блоки ставок, которые были описаны выше в разделе 2.2.2. После дублирования к каждому блоку добавляется еще один параметр – номер блока, который является значением ставки в этом блоке.

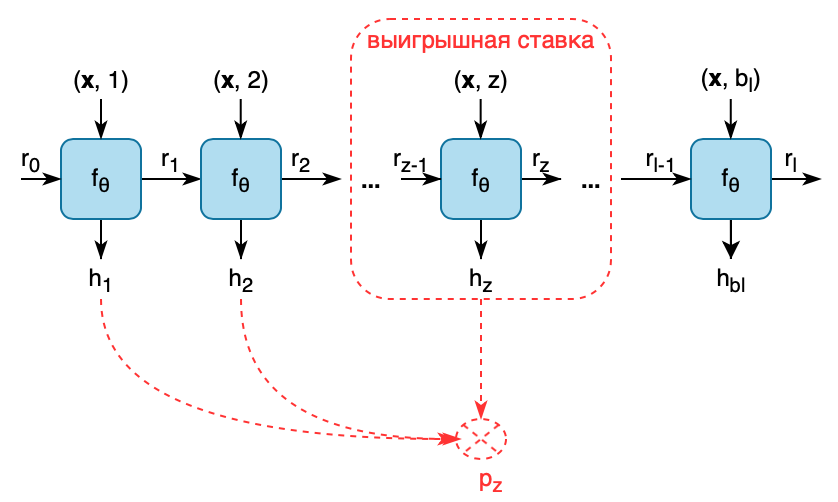


Рисунок 8 – Иллюстрация работы рекуррентного слоя

В третей части подсказываем условную вероятность для каждого блока. Для этого используется рекуррентная LSTM сеть. Так как все вероятности должны быть в интервале , то на выходе из этого слоя, используется монотонно возрастающая функция сигмоида.

На выходе мы получается условные вероятности проигрыша для каждой ставки. С помощью формул **(номер формулы)** можно посчитать вероятности выигрыша, проигрыша и плотности распределения рыночной цены.

### Обучение нейронной сети

Для обучения используется две функции ошибки. Первая основана на плотности распределения и ее основная цель минимизировать ANLP для всех выигрышных случаев (обозначим их как ).

Где – номер интервала, в который входит рыночная цена для -ого примера.

Вторая ошибка основывается на функции распределения и разбита на две части. В случае выигрыша аукциона необходимо «занижать» вероятность выигрыша в интервале , и «завышаем» в интервале .

В случае проигрыша рыночная цена не известна, поэтому можно только «занизить» вероятность выигрыша в интервале .

Чтобы использовать обе ошибки вместе введем простую формулу , которая зависит только от ставки и рыночной цены в аукционе.

Тогда вторую функцию ошибки можно расписать как сумму и .

Для обучения модели будет использоваться комбинация двух формул.

Где и гиперпараметры, которые контролируют величину градиента для стабилизации обучения модели.

## Подходы к улучшению показателей нейронное сети

### Механизм внимания

Один из первых подходов для улучшения модели было применение механизма внимания.

Будем применять механизм для скрытого состояния рекуррентного слоя для создания вектора контекста. После применим этот вектор для каждого блока ставки. С помощью этого, наша сеть получит доступ к необходимой информации о том каким параметрам уделять больше внимания при предсказании распределения, в каждом скрытом состоянии.

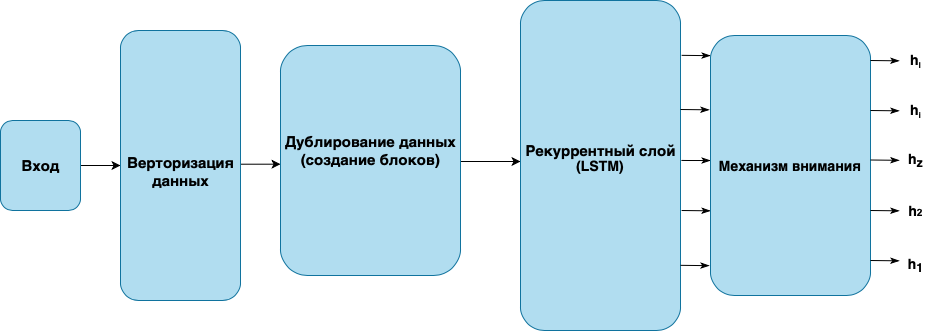


Рисунок 7 – Иллюстрация устройства нейронной сети с механизмом внимания.

### Pruning нейронной сети

Основная идея в облегчении модели нейронной сети, путем удаления некоторого количества нейронов и связей, при этом с сохранением качества модели. Элементы сети, которые оказывают небольшое влияние на ошибку аппроксимации, будут исключаться без значительного ухудшения качества модели.

### Улучшение loss-функции

Хотим добавить дополнительную функцию ошибки для регуляризации всего процесса обучения. По аналогии с решениями из Survival analysis, где используется третья функция ошибся для регуляции при различных замерах показателей при наблюдении за пациентом.

## Использование альтернативного решения из Survival Analysis

Основные тезисы из альтернативного решения задачи:

1. Подход оценивает условную функцию из Survival analysis как смесь отдельных параметрических распределений.
2. Мы не делаем сильных предположений о пропорциональных опасностях и не допускаем обучения с изменяющимися во времени рисками.
3. Подход позволяет изучать, распределенные представления входных ковариаций, помогая передаче знаний по множеству конкурирующих рисков.

Выводы по главе 2

В данном разделе были описаны существующие решения поставленной задачи, решение с помощью нейронных сетей, которое взято за baseline, также было описано реализованное решение и подходы к него улучшению.

# Практическое исследование

## Описание датасета

Для обучения и тестирования использовался открытый датасет компании iPinYou. Этот датасет публичный, что дает возможность сравнивать результаты алгоритмов из других статей. [15]

В датасете содержатся информация об аукционах для 9 различных рекламных объявлений, каждые из которых находятся в разных сферах. Данные разделены на две части: датасет для обучения (train) и для тестирования (test).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер объявления | Выигрышные аукционы | | Проигрышные аукционы | |
| Для train | Для test | Для train | Для test |
| 1458 | 997247 | 119397 | 2085809 | 495241 |
| 2259 | 296657 | 99626 | 538899 | 317571 |
| 2261 | 221454 | 100477 | 466163 | 243385 |
| 2821 | 142697 | 86136 | 1179864 | 575828 |
| 2997 | 43803 | 26944 | 268634 | 129119 |
| 3358 | 278637 | 36373 | 1463467 | 264555 |
| 3386 | 683319 | 136128 | 2164483 | 409293 |
| 3427 | 501868 | 153121 | 2091897 | 383674 |
| 3476 | 644951 | 78896 | 1325409 | 444952 |

Таблица 1 – iPinYou датасет

## Метрика качества модели

Для метрики качества используются показатели ANLP и ROC-кривая. Первая – это функция ошибки, которая используется при обучении, вторая – метрика для классификаторов, которая позволяет оценить точность и качество модели.

## Сравнение с актуальным решением

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № объявления | 2997 | | 3476 | | 1458 | 2259 | 2261 | 2821 | 3358 |
| Модель | baseline | Текущее решение | baseline | Текущее решение | Будет добавлено |  |  |  |  |
| Общая ошибка | 2.2743 | **0.5439** | 1.4686 | **1.4222** |  |  |  |  |  |
| Точность (AUC) | **0.9258** | 0.9236 | 0.9233 | **0.924** |  |  |  |  |  |

|  |  |
| --- | --- |
| 3386 | 3427 |
|  |  |
|  |  |

Таблица 2 – Сравнение baseline с текущим решением

(Будет добавлено для всех объявлений, и таблица будет приведена в более красивый вид)



График 1 - Результаты на датасете iPinYou 3476

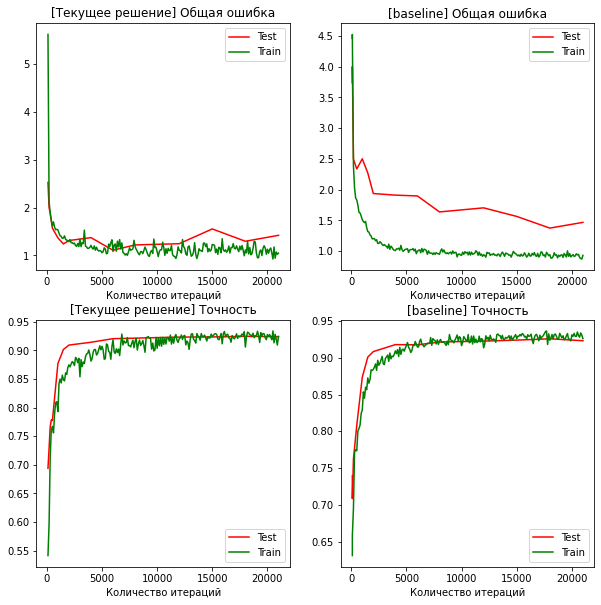


График 2 - Результаты на датасете iPinYou 3476

## Результаты экспериментов

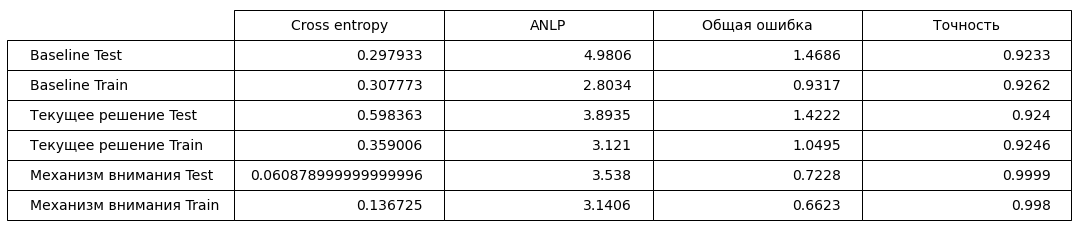


Таблица 3 – Результаты применения механизма внимания

на датасете iPinYou 3476

Выводы по главе 3

В этой главе был описан датасет, на котором происходит обучение и тестирования моделей. Введены метрики для проверки качества модели. Опубликованы первые результаты работы алгоритма.

Заключение

Разделы «Введение» и «Заключение» не нумеруются.

Библиографический список

1. <https://en.wikipedia.org/wiki/Online_advertising>
2. <https://ru.wikipedia.org/wiki/Аукцион>
3. <http://www.adnews.com.au/opinion/how-to-better-understand-auction-dynamics-for-video-ad-campaigns>
4. <http://rtb-media.ru/wiki/>
5. <https://ru.wikipedia.org/wiki/CTR_(Интернет)>
6. <https://ru.wikipedia.org/wiki/Обучение_с_учителем>
7. <https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e>
8. <https://ru.wikipedia.org/wiki/Анализ_выживаемости>
9. <https://en.wikipedia.org/wiki/Kaplan–Meier_estimator>
10. Cox, D. R. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological)
11. <http://wnzhang.net/papers/unbias-kdd.pdf>
12. <http://apex.sjtu.edu.cn/public/files/members/20160929/functional-bid-lands.pdf>
13. <https://www.aaai.org/ojs/index.php/AAAI/article/view/4407>
14. <https://arxiv.org/pdf/1905.03028.pdf>
15. <http://contest.ipinyou.com/ipinyou-dataset.pdf>

Приложения

Если приложение одно, то оно включается в раздел «Приложение» без подразделов.