Масштабное тестирование для предсказания эффекта влияния объекта и UpLift Моделирование

Eustache Diemert

Artem Betlei

Christophe Renaudin

Massih-Reza AminiThéophane Gregoir

Thibaud Rahier

November 22, 2021

Тезисы

Прогнозирование индивидуального эффекта влияния является важной областью исследований в области машинного обучения, которая направлена на объяснение и оценку причинно-следственных связей действия на детальном уровне. Это представляет собой проблему, вызывающую растущий интерес во многих областях применения, таких как здравоохранение, онлайн-реклама или социально-экономическая наука. Чтобы стимулировать исследования по этой теме, мы публикуем общедоступную коллекцию из 13,9 миллионов образцов, собранных в ходе нескольких рандомизированных контрольных исследований, увеличивая ранее доступные наборы данных в 210 раз. Мы предоставляем подробную информацию о сборе данных и выполняем проверки на работоспособность, чтобы подтвердить использование этих данных для задач причинно-следственного вывода. Сначала мы формализуем задачу Моделирования Прироста, которая может быть выполнена с этими данными, наряду с соответствующими оценочными показателями. Затем мы предлагаем поверхности синтетического отклика и назначение гетерогенной обработки, обеспечивающие общую настройку для прогнозирования индивидуального эффекта влияния. Наконец, мы докладываем об экспериментах по проверке ключевых характеристик набора данных, используя его размер для оценки и сравнения − с высокой статистической значимостью − выбора базовых методов прогнозирования Моделирования Прироста и индивидуального эффекта влияния.

# Введение

Область причинно-следственных связей лежит в основе многих научных исследований, поскольку исследователи выявляют и количественно оценивают причинно-следственное влияние различных факторов в сложных системах. Совсем недавно ученые изучали причинно-следственный эффект методов влияния на отдельных объектах в таких приложениях, как персонализированная медицина, онлайн-реклама и целевое профессиональное обучение. Воспроизводимые и более убедительные исследования в этой области стали возможны благодаря наличию эталонных наборов данных из реальных экспериментов, в основном с использованием данных, собранных в 1980-х годах для эконометрики и медицины, где размеры выборки относительно невелики. С появлением Интернета дешевые контролируемые эксперименты стали мейнстримом [36] и делает возможными гораздо более масштабные исследования (обычно используются миллионы образцов). К сожалению, большинство таких наборов данных остаются недоступными для открытой науки, поскольку компании, проводящие эти эксперименты, держат их в секрете. Такой размер набора данных создает уникальные задачи и является более репрезентативным для современных приложений. В этой работе мы представляем набор данных из 14 миллионов экземпляров, содержащий данные, полученные от рекламной онлайн-компании Criteo, и обладающий следующими отличительными характеристиками: дисбаланс воздействия, бинарные и непрерывные, анонимизированные признаки и низкие показатели исходов. Набор данных общедоступен на веб-сайте Criteo AI Lab. С выпуском этого набора данных мы надеемся привнести современный масштаб, надлежащую гарантию качества и современную документацию [27] в области причинно-следственных связей и, особенно, прогнозирования индивидуального эффекта воздействия и моделирования прироста.

**Мотивация для создания нового набора данных**

Вообще говоря, преимущество наличия большего количества наборов данных под рукой заключается в том, что это позволяет делать более обоснованные выводы при экспериментировании с новыми методами, поскольку алгоритмы запускаются в различных условиях с различными характеристиками. В частности, когда доступно очень мало тестов в очень активной области, такой как причинно-следственный вывод, всегда существует риск "концептуального переоснащения", поскольку исследовательское сообщество пытается превзойти уровень техники на нескольких наборах данных, представляющих лишь очень узкий спектр реальных приложений. Более того, отсутствие общедоступных, реалистичных наборов данных иногда побуждает исследователей публиковать результаты на основе частных данных, делая таким образом невоспроизводимые заявления. Наличие крупномасштабных наборов данных также является преимуществом для исследователей, поскольку они могут проводить эксперименты с новыми методами с большей вероятностью зафиксировать значительные различия в производительности, поскольку дисперсия показателей уменьшается с увеличением размера.

**Сопутствующие работы**

Новаторские работы в области прогнозирования индивидуального эффекта влияния в 2000-х и 2010-х годах, такие как [32] или [ 54], используют два набора ориентировочных данных: Международная программа по человеческим измерениям, рандомизированное исследование развития младенцев (<1000 выборок) и работ [40] (около 3000 выборок), рандомизированное исследование профессиональной подготовки, оба собраны в 1980-е годы. Совсем недавно опубликованная работа в области прогнозирования индивидуального эффекта влияния по сей день использует почти исключительно эти наборы данных [54; 43; 56]. Кроме того, для тестирования нескольких методов с 2016 по 2019 год проводился ACIC испытания [24] с использованием набора данных Международной программы по человеческим измерениям, в сочетании с различными процессами генерации. В области моделирования прироста заметным исключением из частных наборов данных является исследование Хиллстрома [33] (64 000 выборок), в котором собраны результаты продаж маркетинговой кампании по электронной почте 2000-х годов. Несмотря на то, что это хорошая положительная динамика, оно недостаточно документировано и все еще отстает с точки зрения масштаба и размерности ковариаций (см. таблицу 1) по сравнению с современными наборами данных в других областях, таких как компьютерное зрение [20] или рекомендательные системы [7], где распространены богатые возможности и ориентир в миллион выборок.

**Материал 1: крупномасштабный набор данных реального мира для моделирования прироста**

Этот набор данных обеспечивает увеличение значимости на 2 порядка для моделирования прироста по сравнению с установленными эталонными показателями. С точки зрения ковариационной размерности, это значительно усложняет настройку, поскольку некоторые функции имеют тысячи возможных значений, что более характерно для современных проблем в веб-приложениях. Вдобавок к этому, набор данных указывает на проблемы, связанные как с целевыми показателями, так и с дисбалансом в воздействии, поскольку лишь небольшая часть пользователей отнесена к контрольной группе и низкий процент положительных результатов. Стратегия анонимизации, используемая в этом наборе данных с показателями, представленными в виде хэшированных токенов, возможно, также является характеристикой будущих приложений, где широко распространены технологии повышения конфиденциальности. Наконец, набор данных и его документация поддерживаются и совершенствуются с течением времени, поскольку, например, новая версия была выпущена через год после первоначальной, чтобы исправить проблему, потенциально препятствующую объективной оценке моделей.

**Материал 2: реалистичное тестирование для прогноза индивидуального эффекта воздействия**

В частности, для прогнозирования индивидуального эффекта воздействия размер нашего набора данных увеличивается на 4 порядка по сравнению с другими тестами. Обычно используется для разработки полусинтетических экспериментов с использованием реальных объектов и смоделированных результатов, определяемых простыми поверхностями отклика (постоянными или экспоненциальными) [32]. Здесь мы предлагаем дополнительные реалистичные поверхности, которые соответствуют наблюдаемым закономерностям в реальных данных и обогащают разнообразие существующих критериев.

**План**

В разделе 2 мы обсуждаем показатели и протоколы оценки для индивидуального прогнозирования эффекта воздействия и моделирования прироста. В духе [27] мы подробно описываем ключевые элементы из таблицы набора данных в разделе 3. Затем, в разделе 4, мы представляем новые синтетические поверхности отклика, вдохновленные реальными наблюдениями, которые позволяют использовать наш крупномасштабный набор данных для прогнозирования индивидуального эффекта воздействия. Наконец, в разделе 5 мы сообщаем об экспериментах по проверке ключевых характеристик набора данных

# Предпосылки

Как правило, как прогнозирование индивидуального эффекта лечения, так и моделирование прироста имеют общую цель - определить, как изменение воздействия влияет на результат. Однако в процессе получения данных можно проследить тонкую разницу между ними: методы, разработанные для прогнозирования эффекта индивидуального воздействия, обычно предполагают данные наблюдений и, следовательно, дополнительно должны решать проблему предвзятости отбора. С другой стороны, при моделировании прироста учитываются экспериментальные данные рандомизированного контрольного исследования.

## 2.1 Обозначения/Структура

Чтобы формализовать понятия индивидуального эффекта воздействия и прироста, мы вводим следующие обозначения: мы предполагаем, что у нас есть доступ к набору данных , содержащему n независимо одинаково распределенных реализаций случайных величин X (ковариат), T (воздействие) и Y (исход). Точнее, для данного индивида принадлежит некоторому d-мерному пространству X и содержит характеристики индивида, содержит статус воздействия, т.е. , если я получил воздействие, и в противном случае, и это значение результата, соответствующего i. Следуя схеме потенциальных исходов [51], у каждого индивидуума i есть два потенциальных исхода: (если я не получу воздействие) и (если я получу воздействие), из которых наблюдается только − фактический результат. Мы обозначаем и лежащие в основе случайные величины. Для заинтересованного читателя подробные обзоры областей прогнозирования индивидуального эффекта лечения и моделирования повышения эффективности приведены в [57; 30].

## Индивидуальный эффект воздействия

**Определение**

Индивидуальный эффект воздействия индивида i затем определяется разницей его потенциальных результатов , которые никогда не наблюдаются одновременно в наборе данных D. Поскольку индивиды описываются векторами признаков x, нас скорее интересует условный средний эффект лечения (CATE), определенный для любого как:

Для любого мы определяем поверхности отклика − или условные потенциальные результаты − как:

Так, что:

**Метрики**

Для оценки оценок индивидуального эффекта воздействия/причинно-следственного эффекта, можно использовать адаптированную версию среднеквадратичной ошибки, а именно точность оценки гетерогенных эффектов (PEHE) [32], которая определяется для модели как:

Где ожидаемое значение берется относительно распределений как X, так и D.

В реальном наборе данных, таком как D, можно наблюдать только фактический результат , но не его контрфактический результат − это называется фундаментальной проблемой причинно−следственного вывода (FPCI), которая препятствует доступу к истинному значению индивидуального эффекта воздействия и, следовательно, делает вычисление PEHE невозможным. Большинство индивидуальных эффектов воздействия срабатывают, поэтому экспериментируйте только с (полу)синтетическими данными. Разумной альтернативой для измерения эффективности прогнозирования индивидуального эффекта воздействия без использования контрфактов является использование показателя политического риска [54], который оценивает потерю целевого воздействия в соответствии с прогнозами модели оцениваемого индивидуального эффекта воздействия при различных соотношениях объема выборки.

**Методы**

Стандартные подходы к прогнозированию индивидуального эффекта включают методы, не зависящие от модели (мета-обучающие [38; 44], модифицированные методы оценки результатов [5] и их комбинации [35]), которые неявно предсказывают CATE, а также подходы на основе дерева, которые особенно подходят для оценки индивидуального эффекта воздействия для прямой оценки эффекта воздействия [4; 55]. Была предложена многочисленная серия все более эффективных алгоритмов, нацеленных на прогнозирование индивидуального эффекта воздействия, с использованием различных методов для корректировки ковариационного сдвига, таких как генеративные состязательные сети [56], автокодеры [43], двойное машинное обучение [15], обучение представлению [54 ; 58] или сбивающие с толку алгоритмы балансировки [37]. Другой недавней тенденцией является изучение теоретических пределов в прогнозировании индивидуального эффекта воздействия и особенно границ обобщения [2].

**Протокол оценки: от необработанных данных до вычисления показателей**

Модели индивидуального эффекта воздейсвтяи обычно оцениваются с использованием полусинтетических наборов данных, которые позволяют получить доступ к истинному значению индивидуального эффекта воздействия и к показателю PEHE, например, предложенным в [32 ; 25]. Двумя основными вариантами проектирования в таких полусинтетических наборах данных являются поверхности отклика для потенциальных исходов и , определенные в (3), и механизм назначения лечения или оценка склонности [19] . Этот выбор оказывает прямое влияние на сложность соответствующей индивидуальной проблемы эффекта воздействия и на тип модели, которая будет наилучшим образом выполнять такие аспекты, как баланс, перекрытие, выравнивание или гетерогенность эффекта воздействия, как описано в [25]. Например, в популярном полусинтетическом фреймворке IHDP [32] выбранные поверхности отклика являются соответственно экспоненциальными и линейными для контрольной и тестовой выборки, а назначение воздействия смещено в сторону выбранной доли пользователей. В случае [25] тестируется несколько поверхностей отклика с различными уровнями нелинейности, а также несколько схем смещения выбора. В менее распространенных сценариях, когда доступны только реальные данные и выполняется FPCI, можно оценить и выбрать индивидуальные модели эффекта воздействия либо с помощью показателя риска политики, либо применяя альтернативные методы оценки модели [3] и выбора модели [53].

## Моделирование прироста

**Определение**

Практико−ориентированное направление работы по оценке причинно-следственного эффекта, называемое моделированием прироста [47], фокусируется на условиях рандомизированного контрольного исследования, в которых люди случайным образом делятся на группу воздействия и контрольную группу. Такие условия позволяют обойти проблему идентификации причинно-следственной связи (избегая какой-либо предвзятости при назначении воздействия) и гарантирует, что оценка (или прирост) по праву определяется разницей между следующими условными ожиданиями:

**Метрики**

При изучении моделей прироста в реальных условиях (PEHE) не может быть вычислена, поскольку достоверный прирост не наблюдается (из-за FPCI). Однако предположение рандомизированного контрольного исследования допускает использование показателей, основанных на ранжировании, таких как кривая прироста [52], которая рассматривает сортировку индивидов в соответствии с прогнозируемым показателем прироста в порядке убывания и измеряет групповой прирост как функцию процента индивидов, занявших первое место. Интуитивно, хорошие модели прироста будут ранжировать наблюдения с большим повышением достоверности, и наоборот (аналогично интуиции политического риска, где соотношение выборок также основано на ранжировании величины индивидуального эффекта воздействия, а пользователи с более высоким прогнозируемым индивидуальным эффектом воздействия вызывают меньшие потери). В литературе существует несколько версий кривых прироста, отличающихся тем, как рассчитывается групповой прирост [21]. В этом ключе площадь под кривой прироста (AUC) [34] представляет собой наиболее популярный показатель в сообществе.

**Методы**

Методы моделирования прироста часто пересекаются с методами прогнозирования эффекта индивидуального воздействия или изобретаются независимо, поскольку первое является подзадачей второго. Например, метод с двумя моделями описан как в литературе индивидуального эффекта воздействия [38], так и в моделировании прироста [31], преобразование переменной класса [34] является частным случаем [5], а метод на основе дерева из сообщества моделирования прироста [52; 48] лишь незначительно отличаются от индивидуальных эффектов воздействия. Однако существуют методы, специфичные для моделирования прироста, которые либо расширяют метод двух моделей, используя определенные представления [8], либо направлены непосредственно на максимизацию AUUC [39; 21; 10]

**Протокол оценки**

В большинстве случаев в литературе по моделированию прироста в качестве показателей эффективности приводятся различные версии AUC. Как для настройки гиперпараметров, так и для построения доверительных интервалов обычно используются популярные методы контролируемого обучения, такие как стратифицированная перекрестная проверка [34; 21] или случайная стратифицированная выборка [8; 21], также применяется набор тестов, привязанный к AUUC [10], с целью обеспечения доверительных интервалов на основе разделения на один тест.

# Набор данных CRITEO-UPLIFTv2

Набор данных находится в открытом доступе на веб-сайте Criteo; код, использованный в экспериментах и валидации, доступен на Github Criteo Research.

**Мотивация и поддерживающая организация**

Criteo - частная компания, которая долгое время занималась развитием науки о воспроизводимой рекламе и за последние 7 лет выпустила 7 крупномасштабных наборов данных, некоторые из которых стали отраслевыми и академическими стандартами. В целом, эти наборы данных интересны тем, что они демонстрируют проблемы на стыке современной теории и практики машинного обучения, например, многомерные задачи тераскейлового прогнозирования [18], контрфактическое обучение [41] и обучение присвоению зачетных единиц [23]. Чтобы обеспечить реалистичное тестирование для моделирования прироста, лаборатория искусственного интеллекта Criteo создала набор данных с помощью набора онлайн-контролируемых экспериментов (A/B-тестов) в Интернете, чтобы лучше изучить индивидуальное влияние рекламы на клики по объявлениям и продажи. Точнее, набор данных создается путем объединения данных, полученных в результате нескольких тестов на инкрементальность, особой процедуры рандомизированного контрольного исследования, при которой часть онлайн-аудитории не подвергается рекламной атаке, в то время как другая часть подвергается ей.

**Описание системы**

Система может быть формально описана путем введения следующих переменных: для данного пользователя, X содержит его показатели, T - двоичная переменная воздействия, такая, что T = 1 для пользователей в группе воздействия и T = 0 для пользователя в контрольной группе, а E, V и C являются двоичными переменными соответственно, указывающие, было ли у пользователя хотя бы одно ознакомление с рекламой, посещение веб-сайта или конверсия в течение периода A/B тестирования (см. рисунок 1 для примера временных рамок таких пользователей). На рисунке 2 мы представляем лежащий в основе причинно-следственный график [45], связанный с этой системой. Она содержит как условную независимость, так и причинно-следственные связи. Например, из причинно-следственного графика мы можем сделать вывод, что воздействие (T) не зависит от пользовательских показателей (X), гарантируя правильную идентификацию причинно-следственных связей, даже если существуют ненаблюдаемые переменные (U), которые могут влиять на E, V и C.

**Структурные аспекты рекламного приложения**

Существует пара характеристик набора данных, которые структурно определяются приложением для онлайн-рекламы. Во-первых, дисбаланс в воздействии объясняется нежеланием терять доход при отказе от показа рекламы в контрольной группе. Мы полагаем, что другие приложения могут столкнуться с такой же проблемой, связанной с отсутствием достаточного бюджета на эксперименты, чтобы сбалансировать методы воздействия для всей выборки. Кроме того, дисбаланс меток объясняется здесь относительно низким коэффициентом конверсии, который типичен для приложений онлайн-рекламы. Наконец, переменные учитывают следующие ограничения − исключительно из-за их определения в контексте онлайн-рекламы:

*отсутствие воздействия рекламы в контрольной группе*

*никакое обращение не может произойти без посещения*

Контекст онлайн−рекламы предполагает некоторые дополнительные предположения, которые позволяют более эффективно прогнозировать индивидуальный эффект воздействия − например, что влияние T на C или V зависит только от E [49] - которые мы не будем подробно описывать далее в этой работе.

**Сбор данных**

Как показано на рисунке 1, пользователи, идентифицируемые файлом cookie браузера, оставляют онлайн−следы посредством событий рекламодателя, таких как посещение веб−сайта или просмотр продукта [36]. Для данного рекламодателя в заранее определенный момент времени случайным образом каждый пользователь попадает либо в группу воздействия, либо в контрольную группу. Период до этого назначения используется для сбора характеристик пользователя (в основном связанных с предыдущей активностью пользователя). Набор показателей был выбран таким образом, чтобы он предсказывал последующие пользовательские события, и мы можем легко проверить с технической точки зрения, что все они были зафиксированы в законный период. После назначения воздействия пользователи либо подвергаются персонализированной рекламе (если было воздействие), либо нет (если находятся в группе контроля) до окончания периода сбора данных. В течение первых двух недель после назначения воздействия регистрируются посещения рекламы и онлайн-конверсии на веб-сайте рекламодателя. Затем характеристики, наблюдаемые во время назначения воздействия, объединяются со статусом назначения воздействия, эффективным показом рекламы, наблюдаемыми посещениями и метками конверсий. Наконец, данные для нескольких таких тестов рекламодателей объединяются для получения необработанного набора данных. Современные веб-платформы обычно проводят множество рандомизированных экспериментов одновременно, однако пользователи, пользующиеся этими сервисами, обычно не полностью осведомлены об этом. В нашем случае мы соблюдали Политику конфиденциальности Criteo, позволяющую пользователям отказаться от эксперимента в любой момент. Единственный недостаток, который мы можем придумать для пользователей, участвовавших в эксперименте, заключался в том, что они избегали просмотра рекламы, что, вероятно, благоприятно для большинства из нас.

**Анонимизация**

Для защиты промышленных активов Criteo и конфиденциальности пользователей не разглашается ни происхождение тестируемого рекламодателя, ни названия и значение характеристик. Более того, значения показателей были хэшированы в случайном векторном пространстве, что сделало их практически невозможными для восстановления при сохранении их предсказательной способности. Неоднородная отрицательный сдвиг на результатах был выполнен таким образом, что исходный уровень приращения не может быть определен при сохранении реалистичного, сложного эталона.

**Соображения, позволяющие избежать временной путаницы**

Особой характеристикой современных рекламных систем является то, что они динамически ориентируются на пользователей на основе наблюдаемых взаимодействий с течением времени [6]. Это означает, что даже в рандомизированном контрольном исследовании (A/B тест) взаимодействие с системой влияет на последующий показ рекламы посредством корректировки ставок на основе реакций пользователей. Примечательно, что на взаимодействие после первого сеанса влияет как группа воздействия, так и предыдущие взаимодействия. Возможные решения, позволяющие избежать этой временной путаницы, включают рассмотрение только первого взаимодействия пользователя во время A/B-теста или регистрацию пользовательских переменных в начале теста и наблюдение за вознаграждением во время теста. Мы выбрали последнее, поскольку оно обеспечивает одновременное ведение журнала характеристик для всех пользователей, сводя к минимуму вероятность наблюдения временных сдвигов в их распространении по таким причинам, как сезонность продаж или эволюция производственной платформы.

**Рекомендации по объединению данных из разных тестов**

Тесты на инкрементальность у разных рекламодателей имели разные коэффициенты воздействия, а это означает, что характеристики, а также прирост коррелировали с фактом участия в данном тесте. Другими словами, (неизданный) идентификатор теста был скрытой помехой в распределении (характеристики, метки) x воздействие. Чтобы обеспечить правильное использование для прироста и прогнозирования индивидуального эффекта воздействия, нам нужно было, чтобы все экземпляры в конечном наборе данных были выведены в независимом дискретном распределении. из того же распределения . Если нет, то модель прогнозирования могла бы получить хороший результат, научившись предсказывать, из какого теста был получен тот или иной экземпляр, и используя тот факт, что некоторые тесты были сильно несбалансированы с точки зрения воздействия, чтобы угадать, более вероятен положительный результат воздействия или нет. Это противоречило бы цели набора данных такой как реалистичное тестирование для моделирования прироста или индивидуального воздействия. Чтобы исправить эту ситуацию, мы провели выборку всех тестов на инкрементальность с одинаковым глобальным коэффициентом воздействия. Таким образом, масштаб набора данных сохраняется, а задача остается приближенной к реальности. Это изменение баланса является ключевым отличием между предыдущей версией (v1) и v2 набора данных [22] и было подтверждено (см. раздел 5.1).

**Описание и анализ набора данных**

Окончательный набор данных (версия 2), далее именуемый CRITEO-UPLIFTv2, состоит из 14 миллионов строк, каждая из которых представляет пользователя с 12 характеристиками, индикатором воздействия, индикатором эффективного показа рекламы и 2 бинарными метками (посещения и конверсии). Глобальный коэффициент охвата составляет 85%, что означает, что в контрольной группе наблюдается лишь небольшая часть пользователей, для которых Criteo не размещает рекламу. Типично, что рекламодатели удерживают только небольшую контрольную группу, поскольку это обходится им в потенциальный доход. Положительные метки означают, что пользователь посетил/купил веб-сайт рекламодателя в течение тестового периода (2 недели). Положительный показ рекламы означает, что пользователь действительно увидел рекламу рекламодателя в течение периода сбора меток. Среди 12 переменных 4 являются непрерывными, а 8 - категориальными с большим количеством модальностей. Чтобы оценить влияние каждого признака на результат посещения V, модель случайного леса (сформированная из 100 оценщиков) обучается на каждой из групп воздействия и контрольной выборки для прогнозирования V. Затем для каждого признака вычисляется среднее уменьшение примеси (MDI) [11] для обеих моделей, и соответствующий средний MDI представлен в таблице 2. Согласно этому эксперименту, f0, f2, f8, f9, по-видимому, значительно влияют на V, в то время как f1, f5, f11 оказывают меньшее влияние.

**Дополнительные подробности и этические соображения**

Более подробная информация доступна в разделе 2 Дополнительного документа; он включает в себя полное описание набора данных, предложенное Гебру и др. [27], которое также охватывает подробные этические вопросы и вопросы конфиденциальности.

# CRITEO-ITE: создание синтетических поверхностей для индивидуального эффекта обработки

В духе [32] мы предлагаем класс синтетических поверхностей отклика, а также метод для разработки смешанного назначения воздействия, предоставляя полусинтетическую версию нашего набора данных, названную CRITEO-ITE, которую можно использовать в качестве эталона для оценки моделей индивидуального эффекта воздействия.

## Поверхности отклика

Мы добавляем два класса синтетических поверхностей отклика для критерия. Сначала мы воспроизводим популярные полусинтетические настройки из [32]. В случае "A" постоянный эффект обработки генерируется двумя линейными поверхностями отклика, а именно , где - вектор коэффициентов, причем каждый компонент выбирается из одного и того же многочленного распределения. В случае "B’ используется экспоненциальная поверхность отклика управления и линейная поверхность отклика воздействия , здесь - матрица фиксированного смещения, а - действительное число, скорректированное таким образом, чтобы средний эффект воздействия на обработанных (ATT) соответствует реальным показателям. Во-вторых, мы предлагаем новый класс поверхностей отклика с несколькими вершинами (немонотонных) в духе интерполяции радиальной базисной функции, вдохновленный наблюдениями, сделанными на проекциях реальной поверхности прироста (см. рис. 3), которые определяют как новую, так и сложную задачу моделирования индивидуального эффекта воздействия. Формально мы предполагаем, что X снабжен нормой ||.|| и определяем для и :

где C - набор опорных точек, - веса, а соответствуют ширине влияния, связанной с каждой из этих точек. Таким образом, для любого связанный CATE задается формулой

Если расстояние между различными опорными точками велико по сравнению со значениями , значение CATE для каждого равно и для любого , является взвешенной суммой с весами , экспоненциально уменьшающимися с соотношениями .

## Механизм назначения воздействия

Чтобы смоделировать условия наблюдения, мы разрабатываем гетерогенную функцию назначения лечения таким образом, чтобы смешивалось с результатом (обратите внимание, что случай, когда является постоянным, соответствует настройке RCT). Мы предлагаем простой способ ввести смещение при назначении воздействия, поставив в зависимость от компонента , который обладает наибольшей прогностической силой в отношении результата . В частности, для заданного малого мы определяем

где − разреженный d-мерный вектор, для которого единственной ненулевой компонентой является та, которая соответствует компоненте наивысшей важности для предсказания ]. Такой выбор механизма назначения воздействия гарантирует, что выполняется строгое предположение о незнании [50], поскольку для всех , и что все помехи между и содержатся в , гарантируя, что .

# Эксперименты

В этой части мы представим эксперименты, проведенные с наборами данных CRITEO-UPLIFTv2 и CRITEO-ITE. Сначала выполняется проверка работоспособности, чтобы подтвердить правильность сбора и создания CRITEO-UPLIFT v2. Затем мы проводим эксперименты на CRITEO-UPLIFTv2, демонстрирующие влияние размера набора данных на разделимость для задачи моделирования прироста. Наконец, мы предоставляем тест методов прогнозирования индивидуального эффекта воздействия на классических поверхностях отклика и наших.

## Механизм назначения воздействия

Мы выполняем несколько проверок на работоспособность, чтобы проверить свойства нашего набора данных. Во-первых, мы гарантируем, что обработка действительно независима от признаков, определяя нулевую гипотезу . Удобный способ проверить это предположение - выполнить тест классификатора с двумя выборками (C2ST) [42]: при классификатор, обученный предсказывать по признакам , должен работает не лучше, чем случайный классификатор. В таблице 3 приведены результаты теста с использованием логарифмических потерь для оценки производительности классификаторов. Эмпирическая потеря классификатора изученного воздействия существенно не отличается от потери случайного классификатора, что отражается довольно высоким значением для одностороннего теста: это не позволяет отклонить и подтверждает, что . Во-вторых, мы удостоверяемся, что зарегистрированные признаки (X) являются информативными и релевантными для прогнозирования результатов (посещение V и конверсия C). Это не обязательно тривиально, поскольку мы отобрали признаки, которые было технически легко зарегистрировать, и обезличили их. В таблице 4 представлено увеличение производительности классификаторов, изученных с использованием посещений и конверсий в качестве меток, по сравнению с базовыми классификаторами, выводящими постоянное значение (среднее значение). Мы наблюдаем значительное увеличение производительности, указывающее на то, что признаки действительно информативны для задачи.

## Моделирование прироста

**Признаки**

Чтобы достичь разумного времени работы при сохранении высокой сложности объектов CRITEO-UPLIFTv2, используемые здесь объекты формируются из 4 исходных непрерывных объектов и 100 проекций на случайные векторы категориальных объектов, которые затем кодируются однократно.

**Целевые переменные**

Для обучения моделей прироста в качестве меток доступны как посещения, так и конверсии. Однако, как представлено здесь, мы предлагаем практикующим специалистам моделировать прирост в первую очередь на посещениях, поскольку сигнал прироста конверсии кажется слишком слабым из-за высокого дисбаланса в метке.

**Метрики**

Мы выбрали в качестве показателя эффективности "отдельные, относительные" оценки AUUC [21], которые показали устойчивость этой версии к дисбалансу воздействия и ее способность отражать предполагаемое использование моделей прироста для таргетирования будущих методов воздействия. Доверительные интервалы вычисляются с использованием привязки набора тестов UUC [9].

**Протокол**

Цель этого эксперимента заключается не в предоставлении наилучших исходных данных, а скорее в том, чтобы подчеркнуть тот факт, что CRITEO-UPLIFTv2 является значимой альтернативой существующим наборам данных моделирования прироста, расширяя масштаб задачи и позволяя получить статистическую значимость результатов. По этой причине мы используем 80%/20% разбиений на обучение/тест, и показатели AUUC сравнивались на тестовых подвыборках пропорциональных размеров с существующими наборами данных, а именно 1000 (IHDP), 5000 (Jobs), 50000 (Hillstrom), 1M и целыми тестовыми данными. Кроме того, обучающий набор используется для настройки базовых моделей с помощью поиска по сетке в сочетании со стратифицированной 5-кратной перекрестной проверкой (для сохранения дисбаланса как в лечении, так и в результатах).

**Модели**

В качестве базовых были использованы четыре модели прироста: двухмодельная [31], преобразование переменных класса [ 34], модифицированный метод конечных результатов [ 5] и совместное представление данных [8]. Для всех моделей были настроены условия регуляризации (см. Дополнение).

**Результаты**

На рисунке 4 представлены результаты эксперимента. Для размеров теста до 1 м все представленные методы неотличимы по показателю AUC, поскольку их доверительные интервалы почти полностью перекрываются. Однако, начиная с 1 м баллов и далее, можно разделить подходы и выполнить выбор модели. Следовательно, это оправдывает необходимость в большом наборе данных для такой сложной задачи.