

УДК 338.2

5.2.2. Математические, статистические и инструментальные методы экономики (экономические науки)

ОСНОВНЫЕ ТИПЫ НЕОПРЕДЕЛЕННОСТИ В А/В-ТЕСТИРОВАНИИ И ИХ ВЛИЯНИЕ НА ПОСЛЕДУЮЩЕЕ UPLIFT-МОДЕЛИРОВАНИЕ В АНАЛИТИКЕ МАРКЕТИНГОВЫХ КОММУНИКАЦИЙ

Шишкин Алексей Евгеньевич
Аспирант кафедры прикладной математики
aleksej.shishkin.02@gmail.com
Университет «Синергия», Россия, Москва 129090, Мещанская, д. 9/14, стр. 1

В статье рассматриваются А/В-тесты как источник данных для оценки среднего эффекта маркетингового воздействия и дальнейшего моделирования uplift-моделей для оценки персонализированного эффекта и персонализации маркетинговых коммуникаций. Цель статьи – формирование типологии источников неопределенности в А/В-тестах и демонстрация возможного влияния различных причин неопределенности на последующее обучение, оценку и применение uplift-моделей. На основании анализа литературы выделены восемь типов источников неопределенности: постановка вопроса эксперимента и ключевой оцениваемой величины; случайная статистическая неопределенность; причинная интерпретация; качество данных и исполнения эксперимента; соответствие измерений эксперимента долгосрочным целям его проведения; сезонность и контекст эксперимента; ошибки анализа эксперимента; переход от модели к управленческому решению. Научная новизна статьи заключается в формировании подробной типологии, соединяющей неопределенность в А/В-тестировании с последующим построением uplift-моделей на экспериментальных данных. Для каждого типа неопределенности приведены примеры влияния в контексте аналитики маркетинговых воздействий и описаны механизмы влияния неопределенности на последующее построение uplift-моделей

Ключевые слова: А/В-ТЕСТИРОВАНИЕ, UPLIFT-МОДЕЛИРОВАНИЕ, НЕОПРЕДЕЛЕННОСТЬ, УГРОЗЫ ВАЛИДНОСТИ, МАРКЕТИНГОВАЯ АНАЛИТИКА

<http://dx.doi.org/10.21515/1990-4665-221-033>

UDC 338.2

5.2.2. Mathematical, statistical and instrumental methods of economics (economic sciences)

THE MAIN TYPES OF UNCERTAINTY IN A/B-TESTING AND THEIR IMPACT ON SUBSEQUENT UPLIFT MODELING IN MARKETING COMMUNICATIONS ANALYTICS

Shishkin Alexey Evgenyevich
Graduate student of the Department of Applied Mathematics
aleksej.shishkin.02@gmail.com
Университет «Синергия», Россия, Москва 129090, Мещанская, д. 9/14, стр. 1

The article considers A/B-tests as a data source for assessing the average effect of marketing impact and further modeling uplift models to assess personalized effect and personalization of marketing communications. The purpose of the article is to form a typology of sources of uncertainty in A/B tests and demonstrate the possible influence of different causes of uncertainty on subsequent training, evaluation and application of uplift models. Based on the literature analysis, eight types of sources of uncertainty have been identified: asking the question of the experiment and the key evaluated quantity; random statistical uncertainty; causal interpretation; data quality and execution of the experiment; Consistency of experimental measurements with long-term objectives; seasonality and context of the experiment; errors in experimental analysis; transition from model to management decision. The scientific novelty of the article consists in forming a detailed typology that combines uncertainty in A/B testing with subsequent building uplift models on experimental data. For each type of uncertainty, examples of influence in the context of marketing impact analysis are given and mechanisms of influence of uncertainty on subsequent construction of uplift-models are described

Keywords: A/B-TESTING, UPLIFT-MODELING, UNCERTAINTY, VALIDITY THREATS, MARKETING ANALYSIS

Введение

А/В-тестирование широко применяется различными компаниями для оценки эффекта нововведений, наиболее частыми сценариями применения является аналитика изменений в цифровых продуктах и анализ изменений в политике коммуникаций в клиентами. В классическом формате эксперимент сравнивает заранее определенные показатели в двух группах – контрольной (отсутствие изменений/воздействия) и целевой (под влиянием изменений/воздействия) и оценивает итоговый средний эффект воздействия как различие в показателях по группам. Однако, данная постановка позволяет ответить только на вопрос «эффективно ли нововведение/воздействие», в то время как зачастую не менее важным вопросом остается «какие именно пользователи демонстрируют наиболее сильную положительную реакцию на изменение/воздействие?». Данный вопрос ставит задачу моделирования неоднородности эффекта воздействия в разрезе отдельных пользователей.

Uplift-моделирование направлено на оценку того, насколько измеряемые метрики по конкретному пользователю изменяются в результате воздействия по сравнению с ситуацией отсутствия воздействия. В литературе задачу uplift-моделирования часто связывают с оценкой индивидуального эффекта воздействия – ITE (individual treatment effect) и оценкой условного среднего эффекта воздействия – CATE (conditional average treatment effect) [41, 51]. В контексте маркетинга итоговый вопрос, решаемый методами uplift-моделирования, можно сформулировать следующим образом: «у каких клиентов вероятность целевого действия увеличится в результате маркетингового воздействия?» [20, 23].

Основанием для построения uplift-моделей, зачастую, являются результаты проводимых компанией ранее А/В-тестов. В связи с этим, источники неопределенности, влияющие на качество проводимых экспериментов,

оказывают также значительное воздействие на последующее моделирование инкрементального эффекта.

Цель статьи – создание типологии источников неопределенности в A/B-тестах с демонстрацией потенциального влияния данной неопределённости на последующее uplift-моделирование. Работа выполнена в формате качественного анализа релевантных источников, систематизирующего представленные в работах описания неопределенности в A/B-тестировании в единую аналитическую схему [3, 6, 8].

Материалы исследования и методология анализа.

В рамках статьи проведен анализ трех направлений в статистике и анализе данных.

Первое направление – методология проведения контролируемых онлайн A/B-экспериментов. В данном блоке проведен анализ работ о статистических особенностях различных методов сравнения групп, проверке качества разделения групп, инструментах снижения дисперсии, возможных ошибках в метриках, кросс-влиянии тестируемых объектов и итоговом формировании вывода на основании A/B-эксперимента.

Второе направление – исследования о выборе целевой оцениваемой величины в эксперименте, как о факторе влияния на исход эксперимента.

Третьим направлением является литература, посвященная инструментам uplift-моделирования и самой методологии как способу таргетирования воздействий в маркетинге и иных областях.

Основной фокус в выборе источников сделан на публикациях 2021 года и более поздних. Основанием данного выбора является динамическое состояние научной области, характеризующееся активным появлением новых исследований, дополняющих, переосмысливающих и/или

критикующих ранее описанные подходы. В то же время исследование затрагивает и более ранние работы в случае, если они описывают теоретический базис исследуемой области и/или конкретных исследуемых объектов. Так, например, анализ строится на статьях, описывающих CUPED, SRM, Ratio-метрики и других [14, 15, 18, 22, 34].

Взаимосвязь А/В-экспериментов и uplift-моделирования.

Связующей точкой между методологиями А/В-тестирования и uplift-моделирования является оценка результата воздействия. Каждый пользователь теоретически может демонстрировать два состояния метрик – при получении воздействия и при его отсутствии. Однако, в реальности каждый пользователь одновременно может находиться только в одном состоянии: пользователь либо получает воздействие, либо же находится в контрольной группе. В связи с этим, оценка индивидуального эффекта рассчитывается опосредовано – через сравнение групп со схожими характеристиками [34, 41, 48, 51].

В базовом варианте А/В-тестирования производится оценка средних изменений в метриках между группами – конверсия, retention, LTV или иные показатели, важные для исследователей. В то же время Uplift-моделирование фокусируется на более детальной задаче – модели в данном подходе строят функцию, связывающую признаки пользователя до получения воздействия с ожидаемым приростом целевых показателей в результате назначения воздействия. В практических задачах за счет данных моделей часто производится ранжирование клиентов по результатам скоринга – с выявлением наиболее позитивно реагирующих на коммуникацию потребителей [8, 41].

В контексте маркетинговой аналитики важным является не только оценить вероятность клиентом покупки в результате маркетингового воздействия, но также и оценить изменение данной вероятности под

воздействием маркетинговых коммуникаций. Клиент с высокой вероятностью совершения целевого действия, возможно, совершил бы его независимо от маркетингового воздействия. В то же время клиент, с низкой исходной вероятностью, но высокой позитивной реакцией на маркетинговые коммуникации является первоочередной аудиторией для целевых воздействий [20].

A/B-тест формирует данные для uplift-моделирования за счет документирования реакции похожих потребителей в ситуациях наличия/отсутствия воздействия. До назначения воздействия фиксируются признаки пользователя, важные для исследователей: частота покупок, средний чек трат, канал привлечения и другие признаки, не являющиеся результатом назначаемого в рамках эксперимента воздействия. На следующем шаге пользователь случайно назначается в контрольную или тестовую группу. Наконец, производится измерение метрик пользователей во время эксперимента.

Uplift-модель, в свою очередь, использует эти данные для определения признаков, связанных с наибольшими различиями между метриками контрольной и тестовой групп. Качество подобной модели сильно зависит от исходных экспериментальных данных, поскольку если в сборе данных или проведении эксперимента будет совершена ошибка, то это неизбежно создаст дополнительные ложные сигналы, которые зафиксирует uplift-модель [8, 41].

Типология источников неопределенности в A/B-тестировании и их влияние на uplift-моделирование.

Неопределенность постановки экспериментального вопроса и оцениваемой величины.

Первый источник неопределенности проявляется на этапе формирования дизайна эксперимента и связан с ответом на вопрос «какой

именно эффект оценивается?». В современной литературе существует понятие оцениваемой величины (estimand), связывающее воедино назначаемое воздействие, оцениваемую совокупность, измеряемые изменения, контекст оценки – время и иные условия, а также способ итогового обобщения результата [30, 31].

В контексте А/В-тестирования внимание может быть сфокусировано как на оценке эффекта на пользователях, получивших воздействие (TOT – treatment on treated), так и в целом на оценке общего состояния системы после проведения воздействия на конкретных пользователей (ATE – average treatment effect).

В контексте маркетинговой аналитики данный источник неопределенности может проявиться в различии теоретического назначения воздействия потребителю в целевой группе и фактическим получением воздействия. Например, потребитель может получить push-уведомление от приложения об акции, но не увидеть его из-за специфических настроек устройства или же получить предложение о скидке, но на самом деле, так и не зайти в раздел приложения/сайта, где отображается данное предложение. В работах, фокусирующихся на анализе взаимодействия потребителей и назначенного воздействия/нововведения, подчеркивается, что данные ситуации требуют внимательного контроля и могут влиять как на мощность эксперимента, так и в принципе на содержательный смысл оцениваемого воздействия [14, 19].

При последующем построении uplift-моделей данная неопределенность может приводить к расхождениям в требованиях компании к uplift-модели и фактически имеющимися данными. Так, компания может захотеть выделять потребителей, наиболее позитивно реагирующих на маркетинговое воздействие, однако, данные А/В-

экспериментов, в то же время, будут подходить только для оценки вероятности взаимодействия с назначенным воздействием.

Случайная статистическая неопределенность полученных результатов

Второй источник неопределенности – стохастический характер процессов, наблюдаемых в ходе А/В-экспериментов – метрики потребителей характеризуются определенной дисперсией, происходящей из многообразия различных сценариев взаимодействия потребителя с продуктом – например, одни потребители могут часто совершать недорогие покупки в то время как другие покупают редко но сразу тратят большие суммы. Следствием этого является невозможность абсолютно достоверного описания генеральной совокупности потребителей на основании анализа ограниченных выборок.

В данном контексте, даже при корректном разделении потребителей в контрольную и целевые группы и при корректном проведении эксперимента итоговая оценка прироста метрик в целевой группе будет иметь как собственную оценку вероятности (p -value – вероятность наблюдать текущую или более экстремальную разницу между тестовыми группами в эксперименте при фактическом отсутствии различий) так и погрешность оценки – доверительный интервал. Низкая мощность эксперимента, может проявиться как в высоком p -value, не позволяющем сказать есть ли статистически значимая разница, так и в широких доверительных интервалах, снижающих нашу уверенность в размере итогового эффекта. Причинами низкой мощности могут стать малое число наблюдений в группах, высокая дисперсия метрики, редкость наблюдаемого события и эффект размытия влияния эксперимента посторонними факторами.

В маркетинговых А/В-экспериментах данная проблема ярко проявляется при анализе нововведений в методы коммуникации с исходно низкой конверсией и иных случаях, когда наблюдаемые исходы могут быть крайне редки. Вторым сценарием риска является анализ метрик с неравномерным распределением и «тяжелыми» хвостами. Примером может служить средний чек по клиенту в экспериментальной группе - множество клиентов может либо вовсе не совершить покупку, либо потратить небольшие суммы, одновременно с этим малый процент потребителей может потратить на покупки значительно больше всех остальных. Актуальность проблемы статистической неопределенности приводит к появлению в статистической литературе отдельных направлений, посвященных работе с повышением мощности эксперимента за счет снижения дисперсии метрик [13, 18, 24]. Также внимание уделяется анализу способов обработки метрик с распределениями, сильно отличающимися от стандартного [54].

Построение uplift-модели на основании А/В-тестов с недостаточной статистической мощностью может привести к низкой предсказательной силе полученного инструмента. Происходит это по причине того, что в основе uplift-моделирования лежит анализ разницы в целевых исходах (например, в конверсии в покупку) в разрезе различных факторов влияния, заданных исследователями. Недостаточная выборка А/В-теста, высокая дисперсия исследуемых метрик или высокая разрежённость данных может приводить к тому что в отдельных сегментах выборки эксперимента (например, пользователи приложения из Москвы с системой iOS в возрасте 50+ лет) не будет достаточно данных для достоверной оценки влияния назначенного воздействия. В связи с этим, при планировании А/В-теста как последующего источника данных для uplift-моделирования, важно обеспечивать достаточное число наблюдений и иные условия, позволяющие оценивать результат в разрезе сегментов [8, 41].

Неопределенность каузальной интерпретации назначенного эффекта воздействия.

Третий тип неопределенности связан с частными случаями невозможности интерпретации разницы между экспериментальными группами как каузального эффекта назначения воздействия, когда по итогу эксперимента исследователи не могут однозначно утверждать, что разница в показателях групп обеспечена исключительно экспериментальным воздействием на целевую группу. Основными причинами обычно являются нарушение рандомизации и наличие взаимосвязей между исследуемыми потребителями из контрольной и целевой группы.

Например, при проведении рандомизации на основании cookie-файлов возможна ситуация, когда один и тот же пользователь, заходящий с разных устройств, получит назначение как в контрольную, так и в целевую группу. В продуктах, предусматривающих контакты между пользователями, возможна ситуация что потребитель из контрольной группы расскажет об изменениях пользователю из тестовой группы, что также повлияет на его поведение. Ещё одним возможным сценарием является нахождение всех пользователей продукта во взаимосвязанной системе – так, например, в случае маркетплейсов, изменение условий для продавцов может повлиять на цену и ассортимент их товаров, что, в свою очередь, может повлиять на действия покупателей, не участвующих в эксперименте [27, 36]. Подобные ситуации требуют отдельных подходов, таких как switchback-тестирование, в ходе которого вся система многократно переключается из одного состояния в другое (например, условия меняются сразу для всех продавцов) и последующее сравнение идет между временными периодами активности старой и новой версии системы.

В контексте uplift-моделирования нарушение каузальности в исходном A/B-тесте означает наличие в обучающей базе дополнительных

сигналов, не связанных с исходной постановкой задачи. Так, при построении uplift-модели предполагается наличие взаимосвязи между характеристиками потребителя и его итоговой реакцией на воздействие. Однако, нарушение каузальной предпосылки в A/B-тесте приводит к ситуации, когда итоговая реакция потребителя зависит также от дополнительных факторов или вовсе определяется исключительно ими. Построенная на подобных данных модель может корректно определять структуру эффектов в обучающей выборке, но при изменении контекста применения она может неожиданно для исследователей перестать работать. Например, если мы нашли способ атрибутировать разные устройства к одному конкретному пользователю, то uplift-модель, обученная по A/B-тестам с cookie-рандомизацией перестанет быть эффективной. Произойдет этого не только из-за изменения объекта анализа с устройства на конкретного человека, но в первую очередь из-за того, что модель вероятно строит предсказание также на «особом характере взаимодействия» между разными потребителями, которые, как мы видим теперь, являются одним человеком [41, 48, 51].

Неопределенность качества собранных данных и фактического исполнения эксперимента.

Ошибки при проведении A/B-эксперимента и/или сборе данных также могут приводят к появлению неопределённости и усложнению дальнейшего построения uplift-моделей. В то время как второй выделенный в данной работе тип неопределенности порожден случайным характером исследуемых величин, а третий происходит из неучтенных при планировании эксперимента зависимостей в данных, то четвертый возникает отдельно именно как следствие технических ошибок в ходе эксперимента.

К ошибкам в данном случае в первую очередь относятся ошибки рандомизации, приводящие к несоответствию долей выборки (sample ratio mismatch – SRM) – ситуации, когда группы или сегменты групп отличаются от заранее заданных значений – например, группа А может быть больше группы В или в одной из групп будет значительно больше потребителей с высоким средним чеком. Данную ошибку возможно диагностировать с помощью статистического теста Хи-квадрат, однако, её обнаружение делает невозможным дальнейший анализ.

Также, к данной группе относятся потери данных о ходе эксперимента, искажение собираемых данных (например, неточная фиксация времени пользователя на сайте) и попадание в выборку эксперимента нерелевантных потребителей (например ботов или аудитории, на которую эксперимент исходно не нацелен). Данные ошибки могут быть менее заметными чем SRM, однако, они также вносят в данные искажения, препятствующие оценке эффективности назначенного воздействия [9, 22].

В контексте маркетинга данные ошибки могут возникать, как и в случае с другими областями из-за ошибок в программном обеспечении и/или личных ошибок сотрудников, участвующих в проведении эксперимента. Ошибка в CRM-системе может привести к неверному распределению потребителей в группы эксперимента, а нововведения, анализируемые в тесте, могут быть нестабильны для потребителей из разных каналов привлечения (например, новый рекламный баннер или сообщение могут некорректно открываться в определенных браузерах и/или на определенных устройствах).

С точки зрения uplift-моделирования существует два возможных риска. Во-первых, нарушения в процессе проведения эксперимента могут привести к невозможности проведения анализа назначенного воздействия, что не позволит использовать полученные результаты в базе обучения

модели. Во-вторых, потенциально более опасной является ситуация, в которой аналитики не выявили существующие проблемы при проведении эксперимента, провели анализ и зафиксировали результат в базе знаний компании. В данном случае в базу обучения uplift-модели могут попасть наблюдения, в которых итоговые изменения целевых метрик потребителя не зависят от его характеристик, а определяются неучтенными ошибками сбора данных. Как и в случае с неучтенными при планировании эксперимента зависимостями, обученная на некорректно собранных данных модель может использовать в предсказаниях сигналы, которые не будут существовать при практической эксплуатации модели или примут другие формы. В результате, предсказательная способность модели на реальных данных может быть значительно более низкой по сравнению с тестами перед запуском.

Неопределенность выбора экспериментальных метрик

Одной из ключевых задач во время дизайна А/В-эксперимента является формирование списка проверяемых показателей или метрик. Основной проблемой для исследователей является баланс между измеримостью изменения и вкладом в общее развитие бизнеса. Итоговой стратегической целью компании является увеличение прибыли, соответственно, изменения внедряемые благодаря А/В-эксперименту зачастую предполагаются как положительно влияющие на прибыль. Однако, измерить то, как новая маркетинговая коммуникация в конечном итоге повлияет на бизнес, зачастую является невозможной задачей – к моменту подведения итогов периода появляется слишком много факторов влияния на итоговую прибыль. В связи с этим исследователи в А/В-тестах часто применяют промежуточные или прокси-метрики: кликабельность объявления/акции, открытие письма/сообщения с маркетинговой коммуникацией, средний чек потребителя во время проведения акции и

другие. Данные метрики удобные тем, что могут быть быстро и однозначно оценены, однако, рост данных метрик в краткосрочном периоде не гарантирует долгосрочного роста выручки, удовлетворенности потребителей, удержания их в продукте и иных стратегических показателей [5, 17, 21].

При проведении А/В-теста маркетинговых коммуникаций исследователи могут столкнуться с тем, что направленные потребителям сообщения повысят посещаемость продукта, но в конечном итоге приведут к разочарованию и оттоку потребителей. Агрессивная скидочная политика может дать рост количества покупок/подписок в короткий срок, но имеет риск сформировать у потребителей привычку ожидания акций. В связи с этим, исследователи стремятся минимизировать возможные негативные эффекты за счет измерения их через барьерные (guardrail) метрики и предварительного формирования нововведения с учетом потенциальных негативных результатов.

Подбирая базу для обучения uplift-модели важно понимать какую задачу с её помощью планируется решить. Ответ на этот вопрос позволит определить, какие из проведенных экспериментов возможно использовать как источники наблюдений. Ключевым моментом является то, что даже если какой-то А/В-тест содержит данные об изменении метрики потребителя, которую планируется предсказывать uplift-моделью, данный эксперимент не обязательно является релевантным по причине возможно расхождения цели исследований. Возможна ситуация, когда А/В-тест новых коммуникаций предполагал применение радикально новых текстов и образов с целью определения наиболее эффективных коммуникаций для кратковременного привлечения внимания. В таком случае, если uplift-модель должна отвечать на вопрос «какому потребителю направить коммуникацию чтобы вырастить итоговый LTV» данный тест будет нерелевантным источником данных, поскольку обучение модели на таких

наблюдениях приведет к тому что она будет оценивать потребителей по краткосрочной реакции но не по долгосрочному росту целевой метрики. Иными словами, при переходе от A/B-тестов к uplift-модели важно понимать содержательный смысл роста целевой метрики потребителя, который рассчитывает модель [8, 20, 23].

Неопределенность устойчивости долгосрочного эффекта нововведения

Одним из ограничений A/B-тестирования является неспособность учета всех возможных контекстов существования тестируемой системы. Наиболее частым примером смены контекста является сезонность. Формируя дизайн A/B-теста исследователи определяют измеримые метрики, по которым ожидается увидеть рост и барьерные метрики, снижения которых нельзя допускать. Каждая метрика имеет свой горизонт изменчивости – например, посещаемость сайта может быть выше по вечерам и по выходным. Исследователи держат во внимании эту изменчивость и определяют срок проведения эксперимента, достаточный для объективной оценки метрики, а не её частного сезонного колебания. Однако, даже корректно проведённый эксперимент не способен учитывать, что между экспериментом и внедрением нововведения могут пройти изменения среды: изменятся состав ключевых конкурентов, портрет потребителя на рынке, насыщенность информационных каналов тематическими коммуникациями или, например, ожидания потребителей [1, 12, 43, 45].

В A/B тестах маркетинговых коммуникаций могут возникать ограничения в долгосрочной применимости разработанных нововведений. Так, предложение об акции может потерять актуальность при нововведениях со стороны конкурентов, а разработанная программа онбординга в продукте окажется неэффективной при увеличении числа

регистраций пользователей с новых каналов/категорий устройств. В таком случае успешные тесты, показавшие значимый положительный эффект от нововведения, могут не привести к полномасштабному внедрению изменения из-за изменений контекста существования продукта. Альтернативным сценарием является принятие изменений несмотря на смену контекста, что может привести к снижению показателей.

В контексте uplift-моделирование данное ограничение означает, что исследователи должны учитывать контекст построения модели, контекст её применения и, что особенно важно, понимать контекст проведенных A/B-экспериментов, данные из которых планируется использовать при обучении uplift-моделей. Возможными решениями становятся: временная валидация, мониторинг сдвигов в данных, использование долгосрочных метрик и переобучение модели в случае её несоответствия изменившемуся контексту [5, 12, 21, 29].

Неопределенность аналитики результатов эксперимента

Потенциальным источником неопределенности является также сама процедура статистического анализа полученных в ходе эксперимента результатов. A/B-тест с корректным дизайном и процедурой проведения может показать ложные результаты в случае если анализ результатов не соответствует заранее определенным требованиям. Частым случаем является «подглядывание» - промежуточная проверка результатов теста, приводящая к некорректной досрочной остановке тестирования. Также исследователи могут не учитывать необходимость поправки множественных проверок, изменять целевую метрику исследования при отсутствии эффекта на исходно заданной в дизайне метрике и дополнительно искать эффект нововведения в разрезах выборки, не определенных исходной методологией [4, 16, 32, 33].

В контексте маркетингового анализа, как и в других сферах, досрочная остановка эксперимента по итогам промежуточной оценки действительно возможна, если она соответствует методологии group sequential testing – например, когда для досрочного вывода рассчитывается требуемый уровень p -value, учитывающий объем собранных данных и возможную статистическую погрешность оценки на текущем объеме наблюдений. Однако, нарушение данной методологии может привести к тому, что в исторической базе экспериментов будет зафиксирован «успешный» A/B-тест с ложноположительным или ложноотрицательным результатом за счет некорректной ранней остановки. Также ложноположительный итог может быть зафиксирован в случае отсутствия контроля за множественными сравнениями в рамках одного эксперимента и сменой измеряемой целевой метрики с целью поиска статистически значимых изменений.

С точки зрения uplift-моделирования данный тип неопределенности пересекается с предыдущим с точки зрения требований к пониманию контекста проведенных экспериментов, используемых как источники данных для модели. Важно понимать, что данные о конкретном маркетинговом воздействии могут быть ограниченными по причине досрочной остановки теста, предполагавшей однозначно положительное/негативное влияние нововведения на не попавших в итоговый тест потребителей. Однако, данное предположение при повторной проверке может оказаться ошибочным и, соответственно, достоверной информации о влиянии маркетингового воздействия на всю базу, в данных может не оказаться.

Отдельно важно понимать, что найденный значимый в рамках эксперимента эффект воздействия на конкретного потребителя мог быть обнаружен в ходе перебора вариантов без поправки на множественные сравнения и на самом деле не является значимым. Это может потребовать

дополнительной переоценки фактических эффектов воздействия на конкретные сегменты потребителей за счет дополнительных А/В-тестов. В противном случае, существует риск, что случайно найденный шум будет передан в uplift-модель как статистически значимое событие, что снизит предсказательную способность итогового инструмента [8, 41].

Неопределенность перехода от модели к управленческому решению

Как А/В-тесты так и uplift-модель применяются с целью формирования аналитической базы для принятия управленческого решения. В случае с А/В-тестированием может быть принято решение о массовом внедрении нововведения. Uplift-модель, в свою очередь позволяет решить каким именно потребителям назначать воздействие для максимизации требуемого эффекта по всей аудитории. Итоговый вывод принимается с учетом стоимости реализации решения, возможных рисков и наличия потенциальных эффективных альтернатив [10, 23, 44].

В маркетинговом анализе возможна ситуация, что А/В-тест показал значимый рост метрики для клиентов, получивших определенную коммуникацию. Например, для клиентов, получивших звонок оператора колл-центра, значимо выросла конверсия в покупку. Однако, данный эксперимент может не учитывать стоимость работы оператора и, соответственно, итоговую выгоду от использования данного канала коммуникации. Которая, более того, может изменяться со временем как с точки зрения реакции потребителей, так и с точки зрения стоимости использования канала.

В связи с этим, при построении uplift-модели важно учитывать, какое итоговое управленческое решение планируется принимать с её помощью и понимать, что высокая оценка эффекта воздействия на конкретного потребителя может не являться достаточным основанием для принятия

решении о назначении воздействия. При построении модели необходимо учитывать стоимость итогового воздействия на потребителя и учитывать различные варианты возможного взаимодействия [23, 52, 53].

Итоговый вывод: как неопределенность A/B-теста находит выражение в ошибках uplift-модели

Представленная в статье типология демонстрирует неоднородность влияния неопределенности A/B-тестов на uplift-модели, построенные на данных экспериментов. Одни источники неопределенности увеличивают статистическую ошибку оценки и снижают надежность ранжирования клиентов моделью. Другие выражаются в смещении взаимосвязей эффектов в A/B-тесте что делает uplift-модель каузально некорректной. Неопределенность может выражаться в несоответствии метрик эксперимента целям uplift-моделирования, что приведет к построению качественной статистически, но фактически не применимой модели. Наконец неопределенность может проявляться на этапе принятия итогового решения по результатам uplift-моделирования.

С точки зрения практики это означает что использование результатов A/B-тестов в uplift-модели должно включать в себя многофакторную проверку качества данных. Ещё до обучения uplift-модели необходимо однозначно определить оцениваемую величину и убедиться, что она соответствует требованиям компании. Важно установить контекст моделирования и барьерные метрики. В самих A/B-экспериментах важно проверить качество проведенной рандомизации и корректность сбора данных и расчета метрик. Наконец, важно оценить достаточность объема данных для построения модели и проверить устойчивость эффекта воздействия на уровне отдельных сегментов выборки.

Научная значимость представленной в работе типологии состоит в объединении описания неопределенности A/B-тестирования и

пригодности полученных экспериментальных данных для uplift-моделирования. Типология может применяться как основа описания рисков для формирования дизайна структуры экспериментов, направленных на сбор данных для последующего построения uplift-модели.

В таблице 1 представлена общая типология неопределенности в A/B-тестировании, описаны механизмы проявления каждого типа и описаны примеры в маркетинговой аналитике.

В таблице 2 демонстрируется как каждый вид неопределенности может влиять на последующее uplift-моделирование и какие шаги предпринимать для обеспечения эффективности использования экспериментальных данных.

Таблица 1 – Типы неопределенности A/B-тестов

Тип неопределенности	Механизм проявления	Пример в маркетинговой аналитике
1. Постановка экспериментального вопроса и оцениваемая величина	Низкая конверсия в целевой группе во взаимодействие с нововведением/коммуникацией; Различие в расчете метрик, рассчитанных в тесте и требуемых в uplift-модели.	Потребитель из тестовой группы получил рассылку о новой акции, но по итогу так и не открыл письмо на почте; Тест считал конверсию в покупку на горизонте 7 дней с момента коммуникации, а uplift-модель требуется для оценки клиентов с наибольшей вероятностью покупки в следующие 30 дней после воздействия.
2. Случайная статистическая неопределенность	Аналитика редких событий, малозаметных эффектов, метрик с высокой дисперсией и/или тяжелыми хвостами.	Тестирование коммуникаций на сегмент с низкой конверсией; анализ влияния на средний чек потребителя; Отсутствие статистической мощности для обнаружения эффекта или, наоборот, обнаружение эффекта при фактическом его отсутствии.
3.	Нарушение	Рандомизация теста на уровне

Неопределенность каузальной интерпретации	предпосылок каузальной интерпретации: некорректная единица рандомизации, наличие взаимного влияния между участниками из разных групп эксперимента.	отдельного устройства, в то время как по факту потребитель может использовать приложение с разных устройств, попадая в обе группы теста; Потребители обмениваются информацией об акциях, и участники тестовой группы пересылают информацию участникам контрольной группы
4. Качество собранных данных и проведения эксперимента	Фактическая реализация эксперимента отличается от описанной в дизайн-документе – ошибки назначения участников в группы, ошибка сбора данных и/или расчета метрик.	Ошибка кода сайта/приложения, отвечающем за назначение группы; Ошибки в механизмах фиксации/хранения результатов
5. Выбор метрик и их валидность	Выбранные метрики не отвечают долгосрочным требованиям внедрения нововведения.	Новый формат коммуникаций повышает конверсию в открытие сообщение, но негативно влияет на итоговый LTV клиента. Новый вариант страницы сайта/приложения замедляет загрузку у некоторых потребителей
6. Устойчивость долгосрочного эффекта воздействия	Эффект измеренного в эксперименте нововведения имеет сезонный характер и может не сохраниться при внедрении нововведения после эксперимента.	Новый формат коммуникаций демонстрирует краткосрочный положительный эффект на конверсию за счет новизны – потребителям интересно посмотреть что-то новое, однако долгосрочного эффекта не наблюдается; В А/В-тесте не учтено влияние конкретного временного периода (например, сезонное повышение спроса на консалтинг в сфере аудита) в

		результате чего делается ложный вывод об эффективности или неэффективности нововведения/коммуникации.
7. Аналитика результатов эксперимента	Нарушение методологии анализа результатов эксперимента.	Досрочная остановка эксперимента в результате досрочной проверки метрик эксперимента без учета GST-корректировки; Проведение множественных парных сравнений без соответствующей поправки уровня значимости; Поиск значимого эффекта в дополнительных разрезах аудитории, не установленных в изначальном дизайне эксперимента.
8. Переход к управленческому решению.	Принятие решения на основании результатов эксперимента без учета общего контекста внедрения нововведения.	Новый формат коммуникации с потребителем внедряется несмотря на его низкую относительно других вариантов маржинальности при сравнительно равном влиянии на метрики потребителя; Принятие нововведения, направленного на краткосрочный рост метрик в конечном счете оказало негативное влияние на стратегические цели компании.

Таблица 2 – Влияние типов неопределенности А/В тестов на uplift-модель и возможные решения

Тип неопределенности	Как отражается на uplift-модели	Способ преодоления
1. Постановка экспериментального вопроса и оцениваемая величина	Модель обучается на одном каузальном эффекте (например, вероятность открытия письма), а	Формирование единого плана А/В-экспериментов и дальнейшего построения uplift-модели с использованием полученных данных, унифицирующего

	применяется для предсказания другого (вероятность покупки после получения письма).	рассчитываемые метрики; При использовании в обучении модели результатов экспериментов, не согласованных заранее – проверка соответствия фактически проанализированных/измеренных метрики постановке задачи для uplift-модели.
2. Случайная статистическая неопределенность	Модель не имеет достаточной предсказательной силы для оценки uplift-эффекта в целом или в отдельных сегментах	На этапе проведения экспериментов определять требования к объему выборки с учетом последующего uplift-моделирования; При апостериорном применении результатов А/В-теста – пересчет прироста метрик с использованием инструментов понижения дисперсии (CUPED; CUPAC)
3. Неопределенность каузальной интерпретации	Модель обучается на скрытых сигналах в данных, вероятно, коррелирующих с отдельными показателями. Это может приводить как к низкой предсказательной силе модели, так и к неустойчивым предсказаниям, в случае исчезновения истинного фактора влияния.	Согласование дизайна А/В-теста и последующего uplift-моделирования; Проверка предшествующих моделированию тестов на адекватность каузальной постановки задачи и соответствия данной постановке итогового дизайна эксперимента – сбора и анализа данных.
4. Качество собранных данных и проведения эксперимента	Модель обучается на ложных сигналах, вызванных ошибками в эксперименте. В	Проверка проведенных А/В-экспериментов на соблюдение SRM и качества сбора данных, в том числе, анализ метрик сбоев системы проведения

	результате возможно как низкое качество модели, так и переобучение на эти ошибки вместо реальных сигналов, что снизит фактическое качество модели.	экспериментов за период эксперимента
5. Выбор метрик и их валидность	Модель обучается на невалидных метриках, что может привести как к низкому качеству полученной uplift-модели, так и к высокому качеству одновременно с несоответствием поставленной перед моделью цели.	Проверка метрик, используемых в А/В-экспериментах. При обнаружении несоответствия метрик требованиям задачи перерасчет результатов в контексте требуемых метрик (в случае если исходный дизайн эксперимента допускает такую возможность за счет обеспечения каузальности).
6. Устойчивость долгосрочного эффекта воздействия	Модель демонстрирует высокое качество на тестовых данных и, вероятно, первое время при запуске в работу, однако со временем качество деградирует с изменением поступающих данных.	Использовать временную валидацию построенной модели, проводить регулярный мониторинг качества модели на предмет сдвигов, переобучать модель в случае выявления снижения качества.
7. Аналитика результатов эксперимента	Наличие ложноположительных и ложноотрицательных выводов о результатах экспериментов может исказить обучающую	Проверять качество анализа данных в экспериментах, используемых в базе обучения uplift-модели. При необходимости производить переоценку результатов экспериментов с целью подбора достаточной базы как по значимым, так и по

	выборку uplift-модели за счет фактического превалирования данных по значимым/не значимым экспериментам. Это может негативно повлиять на предсказательную силу будущей модели.	незначимым изменениям.
8. Переход к управленческому решению.	Высокое предсказательное качество модели и высокий предсказанный рост метрики по конкретным юзерам в результате назначенного воздействия не гарантирует позитивного влияния реализуемой коммуникации на цели компании.	Принимать итоговое управленческое решение учитывая наиболее полный контекст – возможные негативные последствия проведения коммуникации, стоимости коммуникации и наличия возможных альтернативных решений.

Литература

1. Adam H., Heidemanns M., Hains D., McQueen J. Scalable Heterogeneity Detection in Online Experiments. Amazon Science. 2024.
2. Ankargren S., Frånberg M., Schultzberg M. It is About Time: What A/B Test Metrics Estimate. arXiv:2411.06150. 2024.
3. Auer F., Ros R., Kaltenbrunner L., Runeson P., Felderer M. Controlled Experimentation in Continuous Experimentation: Knowledge and Challenges // Information and Software Technology. 2021. Vol. 134. Article 106551. DOI: 10.1016/j.infsof.2021.106551.
4. Berman R., Van den Bulte C. False Discovery in A/B Testing // Management Science. 2022. Vol. 68, No. 9. P. 6762-6782. DOI: 10.1287/mnsc.2021.4207.

5. Bibaut A., Kallus N., Ejdemyr S., Zhao M. Long-Term Causal Inference with Imperfect Surrogates Using Many Weak Experiments, Proxies, and Cross-Fold Moments. arXiv:2311.04657. 2023.
6. Bojinov I., Gupta S. Online Experimentation: Benefits, Operational and Methodological Challenges, and Scaling Guide // Harvard Data Science Review. 2022. Vol. 4, No. 3. DOI: 10.1162/99608f92.a579756e.
7. Bojinov I., Simchi-Levi D., Zhao J. Design and Analysis of Switchback Experiments // Management Science. 2023. Vol. 69, No. 7. P. 3759-3777. DOI: 10.1287/mnsc.2022.4583.
8. Bokelmann B., Lessmann S. Improving Uplift Model Evaluation on Randomized Controlled Trial Data // European Journal of Operational Research. 2024. Vol. 313, No. 2. P. 691-707. DOI: 10.1016/j.ejor.2023.09.018.
9. Chen N., Liu M., Xu Y. How A/B Tests Could Go Wrong: Automatic Diagnosis of Invalid Online Experiments // Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM 2019). 2019. P. 501-509. DOI: 10.1145/3289600.3291000.
10. Chou W., Gray C., Kallus N., Bibaut A., Ejdemyr S. Evaluating Decision Rules Across Many Weak Experiments // Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025). 2025. DOI: 10.1145/3711896.3737217.
11. Coey D., Bailey M. People and Cookies: Imperfect Treatment Assignment in Online Experiments // Proceedings of the 25th International Conference on World Wide Web (WWW 2016). 2016. P. 1103-1111. DOI: 10.1145/2872427.2882981.
12. Degtiar I., Rose S. A Review of Generalizability and Transportability // Annual Review of Statistics and Its Application. 2023. Vol. 10. P. 501-524. DOI: 10.1146/annurev-statistics-042522-103837.
13. Deng A., Du M., Matlin A., Zhang Q. Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes // Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023). 2023. P. 3937-3946. DOI: 10.1145/3580305.3599928.
14. Deng A., Hu V. Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015). 2015. P. 349-358. DOI: 10.1145/2684822.2685307.
15. Deng A., Knoblich U., Lu J. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2018). 2018. P. 233-242. DOI: 10.1145/3219819.3219919.
16. Deng A., Li Y., Lu J., Ramamurthy V. On Post-Selection Inference in A/B Testing // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2021). 2021. DOI: 10.1145/3447548.3467129.
17. Deng A., Shi X. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016). 2016. P. 77-86. DOI: 10.1145/2939672.2939709.
18. Deng A., Xu Y., Kohavi R., Walker T. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data // Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM 2013). 2013. P. 123-132. DOI: 10.1145/2433396.2433413.
19. Deng A., Yuan L.-H., Kanai N., Salama-Manteau A. Zero to Hero: Exploiting Null Effects to Achieve Variance Reduction in Experiments with One-Sided Triggering //

Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM 2023). 2023. P. 823-831. DOI: 10.1145/3539597.3570413.

20. Devriendt F., Berrevoets J., Verbeke W. Why You Should Stop Predicting Customer Churn and Start Using Uplift Models // Information Sciences. 2021. Vol. 548. P. 497-515. DOI: 10.1016/j.ins.2019.12.075.

21. Duan W., Ba S., Zhang C. Online Experimentation with Surrogate Metrics: Guidelines and a Case Study // Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM 2021). 2021. P. 193-201. DOI: 10.1145/3437963.3441737.

22. Fabijan A., Gupchup J., Gupta S., Omhover J., Qin W., Vermeer L., Dmitriev P. Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2019). 2019. P. 2156-2164. DOI: 10.1145/3292500.3330722.

23. Gubela R. M., Lessmann S., Stöcker B. Multiple Treatment Modeling for Target Marketing Campaigns: A Large-Scale Benchmark Study // Information Systems Frontiers. 2024. Vol. 26. P. 875-898. DOI: 10.1007/s10796-022-10283-4.

24. Guo Y., Coey D., Konutgan M., Li W., Schoener C., Goldman M. Machine Learning for Variance Reduction in Online Experiments // Advances in Neural Information Processing Systems. 2021. Vol. 34. P. 8637-8648.

25. Gupchup J., Hosseinkashi Y., Dmitriev P., Schneider D., Cutler R., Jefremov A., Ellis M. Trustworthy Experimentation Under Telemetry Loss // Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018). 2018. DOI: 10.1145/3269206.3271747.

26. Gupta S., Kohavi R., Tang D., Xu Y. Top Challenges from the First Practical Online Controlled Experiments Summit // ACM SIGKDD Explorations Newsletter. 2019. Vol. 21, No. 1. P. 20-35. DOI: 10.1145/3331651.3331655.

27. Han K., Li S., Mao J., Wu H. Detecting Interference in Online Controlled Experiments with Increasing Allocation // Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023). 2023. DOI: 10.1145/3580305.3599308.

28. Jeunen O. A Common Misassumption in Online Experiments with Machine Learning Models // SIGIR Forum. 2023. Vol. 57, No. 1. Article 13. DOI: 10.1145/3636341.3636358.

29. Jeunen O., Baweja S., Pokharna N., Ustimenko A. Powerful A/B-Testing Metrics and Where to Find Them // Proceedings of the 18th ACM Conference on Recommender Systems (RecSys 2024). 2024. DOI: 10.1145/3640457.3688036.

30. Kahan B. C., Hindley J., Edwards M., Cro S., Morris T. P. The Estimands Framework: A Primer on the ICH E9(R1) Addendum // BMJ. 2024. Vol. 384. e076316. DOI: 10.1136/bmj-2023-076316.

31. Keene O. N., Lynggaard H., Englert S., Lanius V., Wright D. Why Estimands Are Needed to Define Treatment Effects in Clinical Trials // BMC Medicine. 2023. Vol. 21. Article 276. DOI: 10.1186/s12916-023-02969-6.

32. Kohavi R., Chen N. False Positives in A/B Tests // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024). 2024. DOI: 10.1145/3637528.3671631.

33. Kohavi R., Deng A., Vermeer L. A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments // Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022). 2022. P. 3168-3177. DOI: 10.1145/3534678.3539160.

34. Künzel S. R., Sekhon J. S., Bickel P. J., Yu B. Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning // *Proceedings of the National Academy of Sciences*. 2019. Vol. 116, No. 10. P. 4156-4165. DOI: 10.1073/pnas.1804597116.
35. Larsen N., Stallrich J. W., Sengupta S., Deng A., Kohavi R., Stevens N. T. Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology // *The American Statistician*. 2024. Vol. 78, No. 2. P. 135-149. DOI: 10.1080/00031305.2023.2257237.
36. Li H., Zhao G., Johari R., Weintraub G. Y. Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms // *Proceedings of the ACM Web Conference 2022*. 2022. P. 182-192. DOI: 10.1145/3485447.3512063.
37. Maharaj A. V., Sinha R., Arbour D., Waudby-Smith I., Liu S. Z., Sinha M., Addanki R., Ramdas A., Garg M., Swaminathan V. Anytime-Valid Confidence Sequences in an Enterprise A/B Testing Platform // *Companion Proceedings of the ACM Web Conference 2023*. 2023. DOI: 10.1145/3543873.3584635.
38. Nie K., Zhang Z., Xu B., Yuan T. Ensure A/B Test Quality at Scale with Automated Randomization Validation and Sample Ratio Mismatch Detection. *arXiv:2208.07766*. 2022.
39. Quin F., Weyns D., Galster M., Silva C. C. A/B Testing: A Systematic Literature Review // *Journal of Systems and Software*. 2024. Vol. 211. Article 112011. DOI: 10.1016/j.jss.2024.112011.
40. Robertson D. S., Wason J. M. S., Ramdas A. Online Multiple Hypothesis Testing // *Statistical Science*. 2023. Vol. 38, No. 4. DOI: 10.1214/23-STS901.
41. Rößler J., Schoder D. Bridging the Gap: A Systematic Benchmarking of Uplift Modeling and Heterogeneous Treatment Effects Methods // *Journal of Interactive Marketing*. 2022. Vol. 57, No. 4. P. 629-650. DOI: 10.1177/10949968221111083.
42. Rzepakowski P., Jaroszewicz S. Decision Trees for Uplift Modeling with Single and Multiple Treatments // *Knowledge and Information Systems*. 2012. Vol. 32, No. 2. P. 303-327. DOI: 10.1007/s10115-011-0434-0.
43. Sadeghi S., Gupta S., Gramatovici S., Lu J., Ai H., Zhang R. Novelty and Primacy: A Long-Term Estimator for Online Experiments // *Technometrics*. 2022. Vol. 64, No. 4. P. 524-534. DOI: 10.1080/00401706.2022.2124309.
44. Schultzberg M., Ankargren S., Frånberg M. Risk-Aware Product Decisions in A/B Tests with Multiple Metrics // *Journal of Statistical Planning and Inference*. 2026. Vol. 245. Article 106393. DOI: 10.1016/j.jspi.2026.106393.
45. Tran A., Bibaut A., Kallus N. Inferring the Long-Term Causal Effects of Long-Term Treatments from Short-Term Experiments // *Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research*. 2024. Vol. 235. P. 48565-48577.
46. Tripuraneni N., Richardson L., D'Amour A., Soriano J., Yadlowsky S. Choosing a Proxy Metric from Past Experiments // *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)*. 2024. P. 5803-5812. DOI: 10.1145/3637528.3671543.
47. Vermeer L., Anderson K., Acebal M. Automated Sample Ratio Mismatch (SRM) Detection and Analysis // *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering (EASE 2022)*. 2022. DOI: 10.1145/3530019.3534982.
48. Wager S., Athey S. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests // *Journal of the American Statistical Association*. 2018. Vol. 113, No. 523. P. 1228-1242. DOI: 10.1080/01621459.2017.1319839.

49. Wan R., Liu Y., McQueen J., Hains D., Song R. Experimentation Platforms Meet Reinforcement Learning: Bayesian Sequential Decision-Making for Continuous Monitoring // Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023). 2023. DOI: 10.1145/3580305.3599818.
50. Wu Y., Zheng Z., Zhang G., Zhang Z., Wang C. Non-Stationary A/B Tests // Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022). 2022. DOI: 10.1145/3534678.3539325.
51. Zhang W., Li J., Liu L. A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling // ACM Computing Surveys. 2021. Vol. 54, No. 8. Article 162. DOI: 10.1145/3466818.
52. Zhao Y., Fang X., Simchi-Levi D. Uplift Modeling with Multiple Treatments and General Response Types // Proceedings of the 2017 SIAM International Conference on Data Mining. 2017. P. 588-596. DOI: 10.1137/1.9781611974973.66.
53. Zhao Z., Harinen T. Uplift Modeling for Multiple Treatments with Cost Optimization // Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA 2019). 2019. P. 422-431. DOI: 10.1109/DSAA.2019.00057.
54. Zhou H., Sun K., Li S., Fan Y., Jiang G., Zheng J., Li T. STATE: A Robust ATE Estimator of Heavy-Tailed Metrics for Variance Reduction in Online Controlled Experiments // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024). 2024. DOI: 10.1145/3637528.3672352.

References

1. Adam H., Heidemanns M., Hains D., McQueen J. Scalable Heterogeneity Detection in Online Experiments. Amazon Science. 2024.
2. Ankargren S., Frånberg M., Schultzberg M. It is About Time: What A/B Test Metrics Estimate. arXiv:2411.06150. 2024.
3. Auer F., Ros R., Kaltenbrunner L., Runeson P., Felderer M. Controlled Experimentation in Continuous Experimentation: Knowledge and Challenges // Information and Software Technology. 2021. Vol. 134. Article 106551. DOI: 10.1016/j.infsof.2021.106551.
4. Berman R., Van den Bulte C. False Discovery in A/B Testing // Management Science. 2022. Vol. 68, No. 9. P. 6762-6782. DOI: 10.1287/mnsc.2021.4207.
5. Bibaut A., Kallus N., Ejdeymyr S., Zhao M. Long-Term Causal Inference with Imperfect Surrogates Using Many Weak Experiments, Proxies, and Cross-Fold Moments. arXiv:2311.04657. 2023.
6. Bojinov I., Gupta S. Online Experimentation: Benefits, Operational and Methodological Challenges, and Scaling Guide // Harvard Data Science Review. 2022. Vol. 4, No. 3. DOI: 10.1162/99608f92.a579756e.
7. Bojinov I., Simchi-Levi D., Zhao J. Design and Analysis of Switchback Experiments // Management Science. 2023. Vol. 69, No. 7. P. 3759-3777. DOI: 10.1287/mnsc.2022.4583.
8. Bokelmann B., Lessmann S. Improving Uplift Model Evaluation on Randomized Controlled Trial Data // European Journal of Operational Research. 2024. Vol. 313, No. 2. P. 691-707. DOI: 10.1016/j.ejor.2023.09.018.
9. Chen N., Liu M., Xu Y. How A/B Tests Could Go Wrong: Automatic Diagnosis of Invalid Online Experiments // Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM 2019). 2019. P. 501-509. DOI: 10.1145/3289600.3291000.
10. Chou W., Gray C., Kallus N., Bibaut A., Ejdeymyr S. Evaluating Decision Rules Across Many Weak Experiments // Proceedings of the 31st ACM SIGKDD Conference

on Knowledge Discovery and Data Mining (KDD 2025). 2025. DOI: 10.1145/3711896.3737217.

11. Coey D., Bailey M. People and Cookies: Imperfect Treatment Assignment in Online Experiments // Proceedings of the 25th International Conference on World Wide Web (WWW 2016). 2016. P. 1103-1111. DOI: 10.1145/2872427.2882981.

12. Degtiar I., Rose S. A Review of Generalizability and Transportability // Annual Review of Statistics and Its Application. 2023. Vol. 10. P. 501-524. DOI: 10.1146/annurev-statistics-042522-103837.

13. Deng A., Du M., Matlin A., Zhang Q. Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes // Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023). 2023. P. 3937-3946. DOI: 10.1145/3580305.3599928.

14. Deng A., Hu V. Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015). 2015. P. 349-358. DOI: 10.1145/2684822.2685307.

15. Deng A., Knoblich U., Lu J. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2018). 2018. P. 233-242. DOI: 10.1145/3219819.3219919.

16. Deng A., Li Y., Lu J., Ramamurthy V. On Post-Selection Inference in A/B Testing // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2021). 2021. DOI: 10.1145/3447548.3467129.

17. Deng A., Shi X. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016). 2016. P. 77-86. DOI: 10.1145/2939672.2939709.

18. Deng A., Xu Y., Kohavi R., Walker T. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data // Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM 2013). 2013. P. 123-132. DOI: 10.1145/2433396.2433413.

19. Deng A., Yuan L.-H., Kanai N., Salama-Manteau A. Zero to Hero: Exploiting Null Effects to Achieve Variance Reduction in Experiments with One-Sided Triggering // Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM 2023). 2023. P. 823-831. DOI: 10.1145/3539597.3570413.

20. Devriendt F., Berrevoets J., Verbeke W. Why You Should Stop Predicting Customer Churn and Start Using Uplift Models // Information Sciences. 2021. Vol. 548. P. 497-515. DOI: 10.1016/j.ins.2019.12.075.

21. Duan W., Ba S., Zhang C. Online Experimentation with Surrogate Metrics: Guidelines and a Case Study // Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM 2021). 2021. P. 193-201. DOI: 10.1145/3437963.3441737.

22. Fabijan A., Gupchup J., Gupta S., Omhover J., Qin W., Vermeer L., Dmitriev P. Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2019). 2019. P. 2156-2164. DOI: 10.1145/3292500.3330722.

23. Gubela R. M., Lessmann S., Stöcker B. Multiple Treatment Modeling for Target Marketing Campaigns: A Large-Scale Benchmark Study // Information Systems Frontiers. 2024. Vol. 26. P. 875-898. DOI: 10.1007/s10796-022-10283-4.

24. Guo Y., Coey D., Konutgan M., Li W., Schoener C., Goldman M. Machine Learning for Variance Reduction in Online Experiments // *Advances in Neural Information Processing Systems*. 2021. Vol. 34. P. 8637-8648.
25. Gupchup J., Hosseinkashi Y., Dmitriev P., Schneider D., Cutler R., Jefremov A., Ellis M. Trustworthy Experimentation Under Telemetry Loss // *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*. 2018. DOI: 10.1145/3269206.3271747.
26. Gupta S., Kohavi R., Tang D., Xu Y. Top Challenges from the First Practical Online Controlled Experiments Summit // *ACM SIGKDD Explorations Newsletter*. 2019. Vol. 21, No. 1. P. 20-35. DOI: 10.1145/3331651.3331655.
27. Han K., Li S., Mao J., Wu H. Detecting Interference in Online Controlled Experiments with Increasing Allocation // *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023)*. 2023. DOI: 10.1145/3580305.3599308.
28. Jeunen O. A Common Misassumption in Online Experiments with Machine Learning Models // *SIGIR Forum*. 2023. Vol. 57, No. 1. Article 13. DOI: 10.1145/3636341.3636358.
29. Jeunen O., Baweja S., Pokharna N., Ustimenko A. Powerful A/B-Testing Metrics and Where to Find Them // *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys 2024)*. 2024. DOI: 10.1145/3640457.3688036.
30. Kahan B. C., Hindley J., Edwards M., Cro S., Morris T. P. The Estimands Framework: A Primer on the ICH E9(R1) Addendum // *BMJ*. 2024. Vol. 384. e076316. DOI: 10.1136/bmj-2023-076316.
31. Keene O. N., Lynggaard H., Englert S., Lanius V., Wright D. Why Estimands Are Needed to Define Treatment Effects in Clinical Trials // *BMC Medicine*. 2023. Vol. 21. Article 276. DOI: 10.1186/s12916-023-02969-6.
32. Kohavi R., Chen N. False Positives in A/B Tests // *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)*. 2024. DOI: 10.1145/3637528.3671631.
33. Kohavi R., Deng A., Vermeer L. A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments // *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*. 2022. P. 3168-3177. DOI: 10.1145/3534678.3539160.
34. Künzel S. R., Sekhon J. S., Bickel P. J., Yu B. Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning // *Proceedings of the National Academy of Sciences*. 2019. Vol. 116, No. 10. P. 4156-4165. DOI: 10.1073/pnas.1804597116.
35. Larsen N., Stallrich J. W., Sengupta S., Deng A., Kohavi R., Stevens N. T. Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology // *The American Statistician*. 2024. Vol. 78, No. 2. P. 135-149. DOI: 10.1080/00031305.2023.2257237.
36. Li H., Zhao G., Johari R., Weintraub G. Y. Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms // *Proceedings of the ACM Web Conference 2022*. 2022. P. 182-192. DOI: 10.1145/3485447.3512063.
37. Maharaj A. V., Sinha R., Arbour D., Waudby-Smith I., Liu S. Z., Sinha M., Addanki R., Ramdas A., Garg M., Swaminathan V. Anytime-Valid Confidence Sequences in an Enterprise A/B Testing Platform // *Companion Proceedings of the ACM Web Conference 2023*. 2023. DOI: 10.1145/3543873.3584635.

38. Nie K., Zhang Z., Xu B., Yuan T. Ensure A/B Test Quality at Scale with Automated Randomization Validation and Sample Ratio Mismatch Detection. arXiv:2208.07766. 2022.
39. Quin F., Weyns D., Galster M., Silva C. C. A/B Testing: A Systematic Literature Review // *Journal of Systems and Software*. 2024. Vol. 211. Article 112011. DOI: 10.1016/j.jss.2024.112011.
40. Robertson D. S., Wason J. M. S., Ramdas A. Online Multiple Hypothesis Testing // *Statistical Science*. 2023. Vol. 38, No. 4. DOI: 10.1214/23-STS901.
41. Rößler J., Schoder D. Bridging the Gap: A Systematic Benchmarking of Uplift Modeling and Heterogeneous Treatment Effects Methods // *Journal of Interactive Marketing*. 2022. Vol. 57, No. 4. P. 629-650. DOI: 10.1177/10949968221111083.
42. Rzepakowski P., Jaroszewicz S. Decision Trees for Uplift Modeling with Single and Multiple Treatments // *Knowledge and Information Systems*. 2012. Vol. 32, No. 2. P. 303-327. DOI: 10.1007/s10115-011-0434-0.
43. Sadeghi S., Gupta S., Gramatovici S., Lu J., Ai H., Zhang R. Novelty and Primacy: A Long-Term Estimator for Online Experiments // *Technometrics*. 2022. Vol. 64, No. 4. P. 524-534. DOI: 10.1080/00401706.2022.2124309.
44. Schultzberg M., Ankargren S., Frånberg M. Risk-Aware Product Decisions in A/B Tests with Multiple Metrics // *Journal of Statistical Planning and Inference*. 2026. Vol. 245. Article 106393. DOI: 10.1016/j.jspi.2026.106393.
45. Tran A., Bibaut A., Kallus N. Inferring the Long-Term Causal Effects of Long-Term Treatments from Short-Term Experiments // *Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research*. 2024. Vol. 235. P. 48565-48577.
46. Tripuraneni N., Richardson L., D'Amour A., Soriano J., Yadlowsky S. Choosing a Proxy Metric from Past Experiments // *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)*. 2024. P. 5803-5812. DOI: 10.1145/3637528.3671543.
47. Vermeer L., Anderson K., Acebal M. Automated Sample Ratio Mismatch (SRM) Detection and Analysis // *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering (EASE 2022)*. 2022. DOI: 10.1145/3530019.3534982.
48. Wager S., Athey S. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests // *Journal of the American Statistical Association*. 2018. Vol. 113, No. 523. P. 1228-1242. DOI: 10.1080/01621459.2017.1319839.
49. Wan R., Liu Y., McQueen J., Hains D., Song R. Experimentation Platforms Meet Reinforcement Learning: Bayesian Sequential Decision-Making for Continuous Monitoring // *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023)*. 2023. DOI: 10.1145/3580305.3599818.
50. Wu Y., Zheng Z., Zhang G., Zhang Z., Wang C. Non-Stationary A/B Tests // *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*. 2022. DOI: 10.1145/3534678.3539325.
51. Zhang W., Li J., Liu L. A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling // *ACM Computing Surveys*. 2021. Vol. 54, No. 8. Article 162. DOI: 10.1145/3466818.
52. Zhao Y., Fang X., Simchi-Levi D. Uplift Modeling with Multiple Treatments and General Response Types // *Proceedings of the 2017 SIAM International Conference on Data Mining*. 2017. P. 588-596. DOI: 10.1137/1.9781611974973.66.

53. Zhao Z., Harinen T. Uplift Modeling for Multiple Treatments with Cost Optimization // Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA 2019). 2019. P. 422-431. DOI: 10.1109/DSAA.2019.00057.

54. Zhou H., Sun K., Li S., Fan Y., Jiang G., Zheng J., Li T. STATE: A Robust ATE Estimator of Heavy-Tailed Metrics for Variance Reduction in Online Controlled Experiments // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024). 2024. DOI: 10.1145/3637528.3672352.