МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра дискретной математики и алгоритмики

ДОМНИН

Даниил Дмитриевич

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ ПОСЛЕДОВАТЕЛЬНОГО ТЕСТИРОВАНИЯ ГИПОТЕЗ

Дипломная работа

Научный руководитель:
старший преподаватель
И. П. Молочко

Допущена к защите

«__» _____ 2025 г.

Заведующий кафедрой дискретной математики и алгоритмики,
доктор физико-математических наук, профессор В.М. Котов

Минск, 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ **6**

1 Аналитический обзор литературы **8**

1.1 Классическое последовательное тестирование гипотез 8

1.2 Методы ковариатной коррекции для снижения дисперсии 10

1.3 Интеграция ковариат и машинного обучения в последовательный анализ 11

1.3.1 Последовательные тесты с учетом ковариат на основе линейных моделей 11

1.3.2 Тестирование гипотез с помощью моделей машинного обучения 12

1.3.3 Глубокое обучение и тестирование путем ставок (Deep Testing-by-Betting) 14

1.3.4 Активное последовательное тестирование и обучение с подкреплением 15

1.4 Резюме обзора и выявленные пробелы 16

2 Постановка задачи и гипотезы исследования **18**

2.1 Формальная постановка задачи 18

2.2 Идея решения 20

2.3 Исследовательские гипотезы 22

3 Предлагаемый метод ST-MLCA **24**

3.1 Алгоритм ST-MLCA 24

3.2 Построение тестовой статистики на остатках 26

3.3 Свойства и теоретические гарантии 27

3.3.1 Статистическая мощность и среднее число наблюдений (ASN). 28
3.4 Выбор модели машинного обучения и практические аспекты . . 29
3.4.1 Выбор модели 29
3.4.2 Ограничения метода 30

4 Экспериментальная оценка эффективности метода ST-MLCA 32
4.1 Эксперименты на синтетических данных 32
4.1.1 Цели и постановка эксперимента 32
4.1.2 Генерация данных 33
4.1.3 Сравниваемые методы и параметры тестирования 34
4.1.4 Метрики оценки 36
4.1.5 Результаты экспериментов и их анализ 37
4.2 Обсуждение результатов экспериментов 45
4.2.1 Применение метода на реальных данных 47

ЗАКЛЮЧЕНИЕ 54

РЕФЕРАТ

Дипломная работа: 55 с., 12 рис., 2 табл., 12 источников, 0 прил.

Ключевые слова: ПОСЛЕДОВАТЕЛЬНОЕ ТЕСТИРОВАНИЕ ГИПОТЕЗ, КОВАРИАТНАЯ КОРРЕКЦИЯ, SPRT, CUPED, МАШИННОЕ ОБУЧЕНИЕ.

Объект исследования: Процедуры последовательного тестирования статистических гипотез.

Цель исследования: Разработка и исследование метода повышения эффективности последовательного тестирования гипотез за счет ковариатной коррекции на основе моделей машинного обучения для ускорения обнаружения статистически значимых эффектов при строгом контроле ошибок первого рода.

Методы исследования: Теоретический анализ, разработка алгоритма, моделирование и вычислительные эксперименты.

Полученные результаты и их новизна: Разработан метод ковариатной коррекции на основе ML-моделей для последовательного тестирования. Экспериментально подтверждено значимое сокращение среднего числа наблюдений при контроле ошибки первого рода. Эффективность зависит от качества ML-модели; нелинейные модели эффективнее при сложных зависимостях. Новизна – в сочетании однократного обучения ML-модели с анализом остатков в последовательных процедурах.

Область возможного практического применения: Повышение эффективности онлайн-экспериментов в технологических компаниях, интернет-маркетинге и других сферах, применяющих последовательный анализ данных.

Автор работы подтверждает, что приведенный в ней расчетно-аналитический материал правильно и объективно отражает сущность исследованного метода и полученные результаты, а теоретические и методологические положения основаны на достоверных источниках.

_____

(подпись студента)

РЭФЕРАТ

Дыпломная работа: 55 с., 12 мал., 2 табл., 12 крыніц, 0 дадаткаў.

Ключавыя словы: ПАСЛЯДОЎНАЕ ТЭСЦІРАВАННЕ ГІПОТЭЗ, КАВАРЫЯТНАЯ КАРЭКЦЫЯ, SPRT, CUPED, МАШЫННАЕ НАВУЧАННЕ.

Аб'ект даследавання: Працэдуры паслядоўнага тэсціравання статыстычных гіпотэз.

Мэта даследавання: Распрацоўка і даследаванне метаду павышэння эфектыўнасці паслядоўнага тэсціравання гіпотэз за кошт каварыятнай карэкцыі на аснове мадэлей машыннага навучання для паскарэння выяўлення статыстычна значных эфектаў пры строгім кантролі памылак першага роду.

Метады даследавання: Тэарэтычны аналіз, распрацоўка алгарытму, мадэляванне і вылічальныя эксперыменты.

Атрыманыя вынікі і іх навізна: Распрацаваны метад каварыятнай карэкцыі на аснове ML-мадэлей для паслядоўнага тэсціравання. Эксперыментальна пацверджана значнае скарачэнне сярэдняй колькасці назіранняў пры кантролі памылкі першага роду. Эфектыўнасць залежыць ад якасці ML-мадэлі; нелінейныя мадэлі больш эфектыўныя пры складаных залежнасцях. Навізна – у спалучэнні аднаразовага навучання ML-мадэлі з аналізам астач у паслядоўных працэдурах.

Вобласць магчымага практычнага прымянення: Павышэнне эфектыўнасці анлайн-эксперыментаў у тэхналагічных кампаніях, інтэрнэт-маркетынгу і іншых галінах, дзе выкарыстоўваецца паслядоўны аналіз даных.

Аўтар работы пацвярджае, што прыведзены ў ёй разлікова-аналітычны матэрыял правільна і аб'ектыўна адлюстроўвае сутнасць даследаванага метаду і атрыманыя вынікі, а тэарэтычныя і метадалагічныя палажэнні грунтуюцца на дакладных крыніцах.

_____

(подпіс студэнта)

ANNOTATION

Degree paper: 55 pages, 12 ill., 2 tab., 12 sources, 0 appendices.

Key words: SEQUENTIAL HYPOTHESIS TESTING, COVARIATE ADJUSTMENT, SPRT, CUPED, MACHINE LEARNING.

Object of research: Procedures of sequential hypothesis testing.

Purpose of research: Development and study of a method to enhance the efficiency of sequential hypothesis testing through covariate adjustment based on machine learning models to accelerate the detection of statistically significant effects under strict control of type I errors.

Research methods: Theoretical analysis, algorithm development, simulation, and computational experiments.

Obtained results and their novelty: A covariate adjustment method based on ML models for sequential hypothesis testing has been developed. Experimentally confirmed significant reduction in the average number of observations while controlling the type I error. The effectiveness depends on the quality of the ML model; nonlinear models are more effective for complex dependencies. The novelty lies in combining one-time ML model training with residual analysis in sequential procedures.

Area of possible practical application: Improving the efficiency of online experiments in technology companies, internet marketing, and other fields utilizing sequential data analysis.

The author of the work confirms that computational and analytical material presented in it correctly and objectively reproduces the picture of investigated process, and all the theoretical, methodological and methodical positions and concepts borrowed from literary and other sources are given references to their authors

_____

(student's signature)

ВВЕДЕНИЕ

Современный мир характеризуется экспоненциальным ростом объемов генерируемых данных, особенно в контексте онлайн-экспериментов, таких как A/B-тестирование, которые являются неотъемлемой частью процесса принятия решений в крупных технологических компаниях и многих других отраслях. Традиционные подходы к проверке статистических гипотез с фиксированным размером выборки, широко применяемые для анализа результатов таких экспериментов, сталкиваются с существенными трудностями [1]. Одной из главных проблем является необходимость длительного сбора данных, особенно при попытке обнаружить малые, но потенциально важные эффекты. Это приводит либо к затягиванию экспериментов, либо к необоснованным затратам ресурсов и упущенным возможностям для быстрого внедрения улучшений.

В качестве альтернативы классическим методам выступает методология последовательного тестирования статистических гипотез. Последовательные тесты, пионером которых является классический последовательный критерий отношения правдоподобия (Sequential Probability Ratio Test, SPRT), предложенный А. Вальдом [1], предоставляют элегантное решение указанных проблем. Они позволяют проводить непрерывный мониторинг поступающих данных и останавливать эксперимент сразу же, как только накоплено достаточно свидетельств в пользу одной из гипотез (нулевой или альтернативной), при этом строго контролируя заданные уровни ошибок первого (α) и второго(β) рода. Такой подход может значительно сократить среднее время проведения эксперимента и необходимые объемы выборок.

Однако, несмотря на свою эффективность в сокращении среднего времени эксперимента, классические последовательные методы, включая SPRT, как правило, не используют доступную информацию о ковариатах (дополнительных признаках объектов исследования) [2]. В современных условиях

для каждого объекта эксперимента (например, пользователя, сессии или транзакции) часто собирается большое количество сопутствующих признаков – демографические данные, история взаимодействий, технические параметры и т.д. Эти ковариаты могут объяснять значительную долю вариативности целевой метрики, и их учет способен существенно повысить чувствительность статистических тестов. Игнорирование этой информации в классических последовательных процедурах является упущением, ограничивающим их потенциальную эффективность.

Применение методов машинного обучения (МО) для анализа и использования информации, содержащейся в ковариатах, представляет собой перспективное направление для дальнейшего усовершенствования последовательного тестирования. Модели МО способны выявлять сложные нелинейные зависимости между ковариатами и целевой метрикой, что позволяет более точно корректировать исходные данные, снижая их дисперсию и, как следствие, ускоряя процесс обнаружения статистически значимых эффектов. Интеграция МО в методологию последовательного анализа открывает возможности для создания нового поколения адаптивных и высокоэффективных статистических инструментов.

Актуальность данной дипломной работы обусловлена растущей потребностью в быстрых и надежных методах анализа экспериментальных данных в условиях больших объемов информации и необходимостью учета множества влияющих факторов. Разработка методов, сочетающих гибкость машинного обучения с гарантиями последовательного статистического вывода, является важной научной и практической задачей.

Целью данной дипломной работы является разработка и исследование нового метода последовательного тестирования статистических гипотез, использующего возможности машинного обучения для осуществления ковариатной коррекции (далее ST-MLCA – Sequential Test with ML-based Covariate Adjustment), с целью ускорения обнаружения статистически значимых эффектов при строгом контроле ошибок первого рода.

ГЛАВА 1

Аналитический обзор литературы

Современная практика принятия решений на основе данных, особенно в динамичных средах, таких как онлайн-платформы, выдвигает повышенные требования к статистическим инструментам. Эти инструменты должны обеспечивать возможность получения быстрых и одновременно надежных выводов. Последовательное тестирование гипотез представляет собой парадигму, отвечающую этим требованиям, так как оно позволяет анализировать данные по мере их поступления и принимать решение об остановке эксперимента в оптимальный, с точки зрения накопленной информации, момент времени. Такой подход особенно актуален в контексте A/B-тестирования, где оперативное принятие решений может существенно экономить ресурсы и ускорять итерационные циклы разработки продуктов.

1.1 Классическое последовательное тестирование гипотез

Исторически, основополагающей работой в области последовательного анализа является последовательный критерий отношения правдоподобия (Sequential Probability Ratio Test, SPRT), разработанный Абрахамом Вальдом в середине XX века. SPRT предназначен для различения двух простых гипотез, например, нулевой гипотезы $H_0 : \theta = \theta_0$ против альтернативной гипотезы $H_1 : \theta = \theta_1$. Данный критерий обладает уникальным свойством оптимальности: среди всех тестов с заданными вероятностями ошибок первого рода (α) и второго рода (β), SPRT минимизирует среднее число наблюдений, необходимых для принятия решения, как при верности H_0, так и при верности H_1 [1].

Механизм SPRT основан на вычислении отношения правдоподобия L_n на каждом шаге n поступления данных. Если $x_1, \ldots, x_n$ — это последовательность наблюдений, а $p_0(x_1, \ldots, x_n)$ и $p_1(x_1, \ldots, x_n)$ — функции правдоподобия для этих наблюдений при гипотезах H_0 и H_1 соответственно, то отношение правдоподобия вычисляется как:

$$L_n = \frac{p_1(x_1, \ldots, x_n)}{p_0(x_1, \ldots, x_n)}. \tag{1.1}$$

Полученное значение L_n сравнивается с двумя порогами A и B, которые определяются через заданные уровни ошибок:

$$A = \frac{1-\beta}{\alpha} \quad \text{и} \quad B = \frac{\beta}{1-\alpha}. \tag{1.2}$$

Правила принятия решения формулируются следующим образом:

- Если $L_n \geq A$, принимается гипотеза H_1.
- Если $L_n \leq B$, принимается гипотеза H_0.
- Если $B < L_n < A$, эксперимент продолжается, и ожидается следующее наблюдение x_{n+1}.

Преимущества SPRT очевидны: экономия ресурсов и времени за счет потенциально меньшего объема выборки, а также строгие теоретические гарантии контроля ошибок первого и второго рода. Однако классический SPRT имеет и существенные ограничения. Он применим в основном к простым гипотезам, что не всегда соответствует практическим задачам, где гипотезы часто являются сложными (например, $H_0 : \theta \leq \theta_0$ против $H_1 : \theta > \theta_0$). Кроме того, SPRT требует знания точных распределений данных при обеих гипотезах, что на практике является сильным предположением. Чувствительность к отклонениям от модельных предположений также является недостатком. Важным для контекста данной работы ограничением является то, что SPRT в своей базовой форме не учитывает информацию из ковариат — дополнительных призна-

ков, которые могут быть связаны с исследуемой метрикой, но не подвержены влиянию экспериментального воздействия.

1.2 Методы ковариатной коррекции для снижения дисперсии

Необходимость повышения эффективности экспериментов, особенно в рамках A/B-тестирования, где целевые эффекты часто малы, а уровень шума в данных высок, привела к разработке методов снижения дисперсии наблюдаемых метрик. Одним из наиболее популярных и практически успешных подходов в этой области стал метод CUPED (Controlled-experiment Using Pre-Experiment Data) [2].

Идея CUPED заключается в использовании данных, собранных до начала эксперимента (пред-экспериментальные данные), для построения прогноза основной метрики Y на основе ковариат X. Предполагается, что эти ковариаты X не должны изменяться под воздействием экспериментального вмешательства. Обычно для прогнозирования используется простая линейная регрессионная модель:

$$Y_i = \theta X_i + \epsilon_i, \tag{1.3}$$

где Y_i — значение метрики для i-го объекта, X_i — значение ковариаты (или вектора ковариат) для того же объекта, θ — коэффициент регрессии, а ϵ_i — случайная ошибка. Коэффициент $\hat{\theta}$ оценивается на пред-экспериментальных данных. Скорректированная метрика Y_{adj} для каждого наблюдения в ходе эксперимента затем вычисляется как:

$$Y_{\text{adj},i} = Y_i - \hat{\theta} X_i. \tag{1.4}$$

Теоретически показано, что дисперсия скорректированной метрики связана с дисперсией исходной метрики следующим образом:

$$\text{Var}(Y_{\text{adj}}) = \text{Var}(Y)(1 - \rho^2), \tag{1.5}$$

где $\rho = \mathrm{Corr}(Y, X)$ представляет собой коэффициент корреляции между метрикой Y и ковариатой X (или множественный коэффициент корреляции в случае нескольких ковариат). Таким образом, чем сильнее ковариаты X коррелируют с метрикой Y, тем значительнее будет снижение дисперсии.

Преимущества CUPED включают простоту реализации, интуитивную понятность и доказанную на практике эффективность в снижении необходимого размера выборки для достижения заданной статистической мощности. Метод совместим с различными дизайнами экспериментов. Однако у CUPED есть и недостатки: его эффективность напрямую зависит от наличия сильных и стабильных (во времени) ковариат. Стандартная версия CUPED ограничена использованием линейных моделей для коррекции, что может быть неоптимально при наличии сложных нелинейных связей между ковариатами и целевой метрикой. Кроме того, для надежной оценки коэффициента $\hat{\theta}$ требуется достаточно большой объем пред-экспериментальных данных. Важно отметить, что CUPED в его классической форме ориентирован на эксперименты с фиксированным размером выборки, а не на последовательный анализ.

1.3 Интеграция ковариат и машинного обучения в последовательный анализ

Параллельно с развитием методов снижения дисперсии для статических тестов, велись исследования по интеграции ковариат и более сложных моделей непосредственно в процедуры последовательного тестирования.

1.3.1 Последовательные тесты с учетом ковариат на основе линейных моделей

Одним из направлений развития стали последовательные F-тесты, предложенные, например, в работе [3], которые обобщают классический t-критерий на случай последовательного анализа с возможностью включения ковариат в рамках линейной модели. Эти методы активно используют мартингальные ста-

тистики и теорию *always-valid p-values* (или, что тесно связано, *e-values*) [4, 5]. Использование *e-values* (неотрицательных случайных величин, чье математическое ожидание при верности H_0 не превышает 1) позволяет конструировать тестовые статистики, которые остаются валидными при любой стратегии остановки, включая непрерывный мониторинг. Произведение последовательных *e-values*, полученных на независимых блоках данных, образует тестовый мартингал $M_k = \prod_{i=1}^{k} e_i$. Решение об отвержении H_0 принимается, когда $M_k \geq 1/\alpha$.

Такие подходы позволяют не только учесть основной эффект ковариат (снижая остаточную дисперсию), но и моделировать гетерогенность эффекта лечения, то есть взаимодействие экспериментального воздействия с ковариатами. Экспериментально показано, что включение ковариат таким образом способно уменьшать среднее время до остановки теста и сужать доверительные интервалы для оценки эффекта по сравнению с тестами, не использующими ковариаты. Тем не менее, эти методы все еще опираются на предположения линейной модели относительно связи между ковариатами, метрикой и эффектом лечения, что может быть ограничивающим фактором в ряде практических приложений, где зависимости носят более сложный, нелинейный характер.

1.3.2 Тестирование гипотез с помощью моделей машинного обучения

Следующим значительным шагом в интеграции машинного обучения и последовательного тестирования стало использование методов МО не только для предобработки данных (как в CUPED) или в рамках линейных моделей, но и для построения самой тестовой статистики.

Ярким примером являются так называемые Classifier Two-Sample Tests (C2ST) [6]. В рамках этого подхода задача различения двух выборок (например, контрольной и тестовой группы в A/B-тесте) сводится к задаче бинарной классификации. Обучается модель машинного обучения (например, нейронная сеть, градиентный бустинг), которая предсказывает, к какой из двух групп

принадлежит данное наблюдение. Интуитивно, если выборки происходят из разных распределений (т.е. H_1 верна), хороший классификатор сможет их различать с точностью, значительно превышающей случайное угадывание (0.5 для сбалансированных выборок). Качество обученного классификатора, оцененное на отложенной (тестовой) выборке (например, точность классификации или площадь под ROC-кривой, AUC), может служить статистикой теста.

Недавние работы развили эту идею для последовательного анализа, предложив E-C2ST (E-Valuated Classifier Two-Sample Tests) [7]. Этот метод использует технику разделения данных (data splitting) и калибровку для получения последовательности *e-values*. На каждом шаге (или для каждого нового батча данных) данные разделяются: на одной части обучается или дообучается классификатор, а на другой части оценивается его качество, которое затем преобразуется в *e-value*. Как упоминалось ранее, произведение таких *e-values* образует тестовый мартингал, который можно использовать для последовательного тестирования с помощью правила остановки Вальда: остановка происходит, когда произведение превысит порог $1/\alpha$.

Преимущества E-C2ST заключаются в их потенциально высокой мощности при обнаружении сложных, многомерных и нелинейных различий между распределениями, где классические тесты (например, тест Колмогорова-Смирнова или t-тест) могут быть неэффективны. Гибкость в выборе модели МО (от логистической регрессии до сложных нейронных сетей) является еще одним плюсом. Однако E-C2ST сопряжены и с определенными недостатками: они могут быть вычислительно сложными, особенно если модель МО требует значительных ресурсов для обучения на каждом шаге. Необходимость аккуратной калибровки для обеспечения теоретических гарантий контроля уровня значимости α также является важным аспектом. Кроме того, разделение данных на обучающую и тестовую выборки на каждом шаге может приводить к потере эффективности на малых выборках. Применимость и эффективность E-C2ST в контексте типичных промышленных A/B-тестов с большим количеством одновременно отслеживаемых метрик и ковариат остается относительно слабо изученной.

1.3.3 Глубокое обучение и тестирование путем ставок (Deep Testing-by-Betting)

Еще более продвинутым и теоретически элегантным направлением является концепция *deep testing-by-betting* [8, 9]. Этот фреймворк использует идеи из теории игр и финансовой математики, в частности, концепцию "тестирования путем ставок"(*testing by betting*). Пусть есть себе игрок (или инвестор), который делает ставки на то, верна ли альтернативная гипотеза H_1. На каждом шаге поступления новых данных игрок обновляет свой "капитал"в зависимости от того, насколько эти данные подтверждают H_1 по сравнению с H_0. Если капитал достигает определенного заранее заданного порога (например, $1/\alpha$), игрок выигрывает ставку, что эквивалентно отклонению нулевой гипотезы H_0.

В рамках *Deep Sequential Testing*, роль стратегии игрока (т.е. как именно делать ставки, чтобы максимизировать капитал при верности H_1, сохраняя его в среднем невозрастающим при H_0) выполняет глубокая нейронная сеть. Эта сеть может обучаться "*on the fly*"по мере поступления данных, стремясь построить оптимальную последовательность ставок (или, что эквивалентно, *e-values*). Процесс накопления капитала в такой системе по своей природе является мартингалом (или супермартингалом) при H_0, что обозначается как *e-process*. Преимущества данного подхода включают высокую адаптивность к структуре данных, способность обнаруживать очень сложные и неявные паттерны различий между гипотезами, а также строгий контроль ошибки первого рода, который естественным образом вытекает из мартингальной природы процесса накопления капитала. Теоретически, такие подходы обладают свойством универсальности. К недостаткам можно отнести высокую вычислительную сложность и требовательность к ресурсам, поскольку обучение глубокой нейронной сети на каждом шаге или на каждом батче данных может быть очень затратным. Сложность реализации и настройки таких систем также является барьером. На данный момент число примеров практического применения этих методов ограничено, и они в основном демонстрировались на задачах анализа изображений или на синтетических данных. Остается

недостаточно ясным, как эффективно применять этот подход в стандартных промышленных A/B-тестах с табличными данными и множеством метрик.

1.3.4 Активное последовательное тестирование и обучение с подкреплением

Наконец, стоит упомянуть подходы, находящиеся на стыке последовательного тестирования и обучения с подкреплением (Reinforcement Learning, RL), а также идеи активного сбора данных. Здесь основная идея заключается в том, чтобы не просто пассивно наблюдать за поступающими данными, но и активно влиять на процесс их сбора или использовать RL для выработки оптимальной стратегии остановки, которая может учитывать не только статистические критерии, но и экономические факторы (например, стоимость сбора каждого дополнительного наблюдения).

Например, в работе [10] предложен метод, где рекуррентная нейронная сеть (RNN) обрабатывает последовательность наблюдений из эксперимента, а агент обучения с подкреплением на основе этой информации решает на каждом шаге, следует ли продолжать сбор данных или остановиться. Целью агента является максимизация вероятности правильного выбора гипотезы при одновременной минимизации задержки (количества наблюдений). Такие методы особенно интересны для сред с неизвестной или изменяющейся динамикой. Однако их связь с классическим контролем ошибок типа α и β требует дополнительного и более глубокого изучения. Кроме того, эти подходы, хотя и являются перспективными, выходят за рамки основного фокуса данной дипломной работы, которая ориентирована на улучшение стандартных последовательных тестов типа SPRT и их аналогов, сохраняющих строгий контроль статистических ошибок.

1.4 Резюме обзора и выявленные пробелы

Проведенный анализ литературы выявил несколько ключевых тенденций и существующих пробелов в области последовательного тестирования гипотез, особенно в контексте применения методов машинного обучения.

Во-первых, очевиден огромный потенциал машинного обучения для повышения эффективности последовательного тестирования. Этот потенциал реализуется как через снижение дисперсии с помощью использования ковариат (например, в методах, идейно близких к CUPED, но адаптированных для последовательного анализа, или в последовательных тестах с линейными моделями), так и через построение более мощных и гибких тестовых статистик, способных улавливать сложные зависимости в данных (как в C2ST, E-C2ST и подходах, основанных на *deep betting*).

Во-вторых, несмотря на значительный теоретический прогресс, наблюдается разрыв между передовыми теоретическими разработками и их практическим применением в реальных системах A/B-тестирования. Многие из наиболее сложных МО-интегрированных методов (особенно из области *deep betting* или E-C2ST с тяжелыми моделями) требуют значительных вычислительных ресурсов, сложной настройки и калибровки. Их надежность, интерпретируемость и реальная эффективность в условиях крупномасштабных промышленных экспериментов с множеством метрик и бизнес-ограничениями остаются недостаточно изученными.

В-третьих, остро стоит вопрос обеспечения строгих теоретических гарантий. В то время как методы, основанные на линейных моделях или на правильно сконструированных *e-values* и мартингалах, часто имеют строгое математическое обоснование контроля ошибки первого рода (иногда даже *anytime-validity*), использование сложных нелинейных моделей (таких как градиентный бустинг или нейронные сети) внутри последовательной процедуры требует дополнительных исследований. Особенно это касается ситуаций, когда модели обучаются на тех же данных, которые затем используются для теста, что создает риск переобучения и "утечки информации"из тестовой выборки в

модель, что может привести к завышенной оценке значимости эффектов.

В-четвертых, представляется недостаточно исследованным потенциал комбинации различных подходов. Например, как оптимально сочетать ковариатную коррекцию (возможно, на основе гибких нелинейных моделей) с продвинутыми последовательными статистиками или адаптивными границами остановки? Существует потребность в разработке методов, которые были бы достаточно просты в реализации и интерпретации, но при этом использовали бы мощность современных МО-алгоритмов для эффективного учета ковариат в последовательных тестах, сохраняя при этом строгие статистические гарантии. Именно на решение этой задачи и направлена данная дипломная работа.

В Главе 1 был проведен аналитический обзор литературы, посвященной последовательному тестированию гипотез. Рассмотрены классические методы (SPRT), подходы к ковариатной коррекции (CUPED), а также современные методы, интегрирующие машинное обучение (например, E-C2ST, deep testing-by-betting). В результате обзора были выявлены существующие пробелы и определены перспективные направления для дальнейших исследований, на которых фокусируется данная работа.

ГЛАВА 2

Постановка задачи и гипотезы исследования

Проведенный в предыдущей главе анализ существующей литературы выявил значительный потенциал применения методов машинного обучения для повышения эффективности процедур последовательного тестирования статистических гипотез. Вместе с тем, было отмечено, что классические последовательные тесты, такие как SPRT, не используют доступную информацию о ковариатах, методы ковариатной коррекции вроде CUPED ориентированы на тесты с фиксированным размером выборки, а современные ML-ориентированные подходы (например, E-C2ST) могут быть сложны в реализации и настройке. Таким образом, существует потребность в разработке методов, которые бы сочетали гибкость машинного обучения для учета ковариат с теоретическими гарантиями и практической применимостью последовательного анализа. Данная глава посвящена формальной постановке задачи, решаемой в настоящей дипломной работе, а также формулировке ключевых исследовательских гипотез.

2.1 Формальная постановка задачи

Рассмотрим типичную задачу A/B-тестирования, широко распространенную в онлайн-экспериментах. Пусть имеется две группы объектов исследования: контрольная (группа A) и тестовая (группа B). Тестовая группа подвергается некоторому воздействию (например, новый дизайн интерфейса, изменение алгоритма рекомендаций), в то время как контрольная группа остается без изменений или использует стандартный вариант. Целью эксперимента является оценка влияния этого воздействия на некоторую целевую метрику Y.

Для каждого объекта i в эксперименте (например, пользователя, сес-

сии или транзакции) измеряется значение целевой метрики Y_i. Кроме того, для каждого объекта доступен вектор ковариат (дополнительных признаков) $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$. Эти ковариаты могут включать демографические данные, историю взаимодействий, технические параметры устройства и т.д. Предполагается, что ковариаты X_i могут быть связаны с целевой метрикой Y_i, но сами по себе не подвержены влиянию экспериментального воздействия, то есть их распределение одинаково в контрольной и тестовой группах или их значения зафиксированы до применения воздействия.

Пусть $\mu_C = \mathbb{E}[Y|\text{группа A}]$ и $\mu_T = \mathbb{E}[Y|\text{группа B}]$ обозначают средние значения целевой метрики в контрольной и тестовой группах соответственно. Основная задача статистического анализа заключается в проверке нулевой гипотезы H_0 об отсутствии эффекта от воздействия против альтернативной гипотезы H_1, утверждающей наличие такого эффекта. Формально гипотезы могут быть записаны следующим образом:

- Нулевая гипотеза $H_0 : \mu_T = \mu_C$ (или $\mu_T - \mu_C = 0$)
- Альтернативная гипотеза $H_1 : \mu_T \neq \mu_C$ (двусторонняя) или, например, $H_1 : \mu_T > \mu_C$ (односторонняя, если ожидается улучшение метрики).

Тестирование гипотез проводится в последовательном режиме. Это означает, что данные (пары (X_i, Y_i) для каждого объекта i) поступают по одному или небольшими пакетами (батчами). После каждого поступления новых данных принимается решение:

1. Остановить эксперимент и отвергнуть H_0 в пользу H_1.
2. Остановить эксперимент и не отвергать H_0.
3. Продолжить эксперимент, собрав больше данных.

Ключевой особенностью является возможность ранней остановки, как только накоплено достаточно свидетельств в пользу одной из гипотез.

К разрабатываемому методу последовательного тестирования предъявляются следующие статистические требования:

- **Контроль ошибки первого рода:** Вероятность отвергнуть нулевую гипотезу H_0, когда она верна (т.е. ложное срабатывание), не должна превышать заданный уровень значимости α (например, $\alpha = 0.05$). Желательно, чтобы этот контроль обеспечивался на любом этапе эксперимента, то есть тест должен быть *anytime-valid*.

- **Достаточная мощность:** Вероятность отвергнуть H_0, когда верна альтернативная гипотеза H_1 и истинный эффект имеет величину не менее некоторого минимально значимого эффекта δ (Minimum Detectable Effect, MDE), должна быть не ниже заданной мощности $1 - \beta$ (например, $1 - \beta = 0.80$).

- **Эффективность по объему выборки:** При выполнении указанных требований к ошибкам первого и второго рода, метод должен стремиться минимизировать среднее число наблюдений (Average Sample Number, ASN), необходимое для принятия решения. В частности, ожидается, что использование информации из ковариат X позволит сократить ASN по сравнению с классическими последовательными тестами, не учитывающими ковариаты.

Таким образом, центральной задачей данной дипломной работы является разработка и исследование нового метода последовательного тестирования гипотез, который интегрирует модели машинного обучения для эффективного использования ковариат с целью ускорения процесса принятия решений (уменьшения ASN) при строгом соблюдении заданных статистических гарантий.

2.2 Идея решения

Основная идея предлагаемого подхода заключается в снижении вариативности целевой метрики Y за счет учета информации, содержащейся в ковариатах X, с помощью моделей машинного обучения. Если значительная

часть изменчивости Y может быть объяснена ковариатами, то анализ скорректированной метрики, из которой "удалено"влияние этих ковариат, позволит быстрее и надежнее обнаруживать истинный эффект от экспериментального воздействия.

Предлагается следующий общий механизм:

1. **Обучение предиктивной модели:** На независимом наборе данных (например, исторических данных, собранных до начала эксперимента, или данных из специально выделенной части контрольной группы, не участвующей в последующем сравнении) обучается модель машинного обучения M. Эта модель M предсказывает значение целевой метрики $\hat{Y}$ на основе вектора ковариат X : $\hat{Y} = M(X)$. Важно, чтобы данные для обучения модели не пересекались с данными, используемыми непосредственно для проверки гипотез в ходе последовательного теста, во избежание "утечки информации" и переобучения.

2. **Коррекция метрики в ходе эксперимента:** В процессе проведения A/B-теста для каждого нового наблюдения i (с ковариатами X_i и наблюдаемой метрикой Y_i) с помощью предварительно обученной модели M вычисляется предсказанное значение X : $\hat{Y} = M(X)$. Затем вычисляется остаток (или скорректированная метрика) $R_i = Y_i - \hat{Y}_i$. Интуитивно, эти остатки R_i представляют собой часть метрики Y_i, которая не объясняется ковариатами X_i. Если модель M хорошо улавливает зависимость Y от X, то дисперсия остатков $Var(R_i)$ будет существенно меньше дисперсии исходной метрики $Var(Y_i)$.

3. **Последовательный тест на остатках:** Вместо исходных значений Y_i для последовательного сравнения контрольной и тестовой групп используются полученные остатки R_i. Строится последовательная тестовая статистика (например, на основе отношения правдоподобий или e-значений) на этих остатках.

Предполагается, что если модель M адекватно обучена и не вносит смещения, то при верной нулевой гипотезе H_0 (отсутствие эффекта) математические

ожидания остатков в контрольной и тестовой группах будут равны (или близки к нулю, если модель предсказывает базовый уровень метрики). При верной альтернативной гипотезе H_1 (наличие эффекта δ) разница в средних значениях остатков между группами будет отражать этот эффект: $\mathbb{E}[R_i|\text{тест}] - \mathbb{E}[R_i|\text{контроль}] \approx \delta$..

Ключевым преимуществом такого подхода является потенциальное снижение дисперсии анализируемых данных, что, согласно теории последовательного анализа, должно приводить к уменьшению среднего необходимого числа наблюдений (ASN) для достижения заданной мощности. При этом важно обеспечить, чтобы процедура коррекции и построения тестовой статистики сохраняла строгий контроль над ошибкой первого рода α. Этого предполагается достичь за счет использования корректно построенных мартингальных статистик или других методов, обеспечивающих anytime-validity. Подробный алгоритм и его теоретическое обоснование будут представлены в Главе 3.

2.3 Исследовательские гипотезы

На основе поставленной задачи и предложенной идеи решения формулируются следующие исследовательские гипотезы, которые будут проверяться в ходе экспериментальной части данной дипломной работы:

- **Гипотеза 1 (Сокращение ASN):** Использование ковариатной коррекции на основе модели машинного обучения в рамках последовательного теста позволит значимо уменьшить среднее необходимое число наблюдений (ASN) для достижения заданной мощности при обнаружении эффекта по сравнению с традиционным последовательным тестом (например, классическим SPRT), не использующим информацию о ковариатах.

- **Гипотеза 2 (Зависимость от качества модели):** Эффективность предлагаемого метода (степень сокращения ASN или прирост мощности) напрямую зависит от предсказательной способности используемой модели машинного обучения. Чем большую долю дисперсии целевой

метрики Y объясняют ковариаты X (т.е., чем выше коэффициент детерминации R^2 модели), тем существеннее будет выигрыш от применения ковариатной коррекции.

- **Гипотеза 3 (Преимущество нелинейных моделей):** При наличии сложных, нелинейных зависимостей между целевой метрикой Y и ковариатами X, использование нелинейных моделей машинного обучения (таких как градиентный бустинг, например, CatBoost, или нейронные сети) для ковариатной коррекции обеспечит большее снижение дисперсии и, как следствие, большую эффективность последовательного теста, чем использование простых линейных моделей коррекции (например, линейной регрессии).

- **Гипотеза 4 (Контроль уровня значимости):** Предлагаемый метод последовательного тестирования с МО-коррекцией будет корректно контролировать вероятность ошибки первого рода на заданном уровне значимости α. То есть, при многократном применении теста в условиях отсутствия реального эффекта (при верной H_0), доля ложных отвержений H_0 не будет превышать α. Это обеспечивается за счет использования статистических процедур (например, основанных на e-значениях), гарантирующих мартингальное свойство тестовой статистики при H_0.

Проверка этих гипотез будет осуществляться посредством моделирования и проведения вычислительных экспериментов на синтетических и реальных данных. Результаты этих экспериментов позволят оценить практическую применимость и эффективность разработанного метода.

В Главе 2 была выполнена формальная постановка задачи исследования, включающая описание типичного сценария A/B-тестирования и требований к последовательному тесту. Была изложена основная идея предлагаемого решения ST-MLCA, основанного на ковариатной коррекции с помощью машинного обучения. Сформулированы ключевые исследовательские гипотезы, касающиеся эффективности метода, влияния качества ML-модели и контроля статистических ошибок, которые будут проверены в последующих главах.

ГЛАВА 3

Предлагаемый метод ST-MLCA

В предыдущих главах были рассмотрены существующие подходы к последовательному тестированию статистических гипотез, выявлены их ограничения, в частности, в контексте использования ковариат, и обоснована необходимость разработки нового метода. Была описана общая идея предлагаемого подхода, включающая обучение предиктивной модели, коррекцию метрики путем вычисления остатков, и последующее применение последовательного теста к этим остаткам. Данная глава посвящена детальному описанию алгоритма ST-MLCA (Sequential Test with ML-based Covariate Adjustment), построению тестовой статистики, теоретическому обоснованию его ожидаемых свойств и обсуждению практических аспектов его применения. Эффективность и свойства предложенного метода будут исследованы в ходе экспериментальной проверки в следующей главе.

3.1 Алгоритм ST-MLCA

Метод ST-MLCA предназначен для последовательного анализа данных в A/B-тестировании, позволяя останавливать эксперимент при накоплении достаточных свидетельств в пользу одной из гипотез, при строгом контроле ошибок первого и второго рода. Ключевой особенностью является использование модели машинного обучения (МО) для коррекции целевой метрики на основе ковариат. Общий вид алгоритма следующий:

1. **Подготовка и обучение модели МО $f(\cdot)$:**

   (a) Используется набор данных D_{train} (например, исторические данные или данные контрольной группы из предварительного этапа), неза-

висимый от данных, которые будут использованы в последующем последовательном тесте.

(b) На D_{train} обучается модель МО $f(\cdot)$, предсказывающая значение целевой метрики Y на основе вектора ковариат X: $\hat{Y} = f(X)$. В качестве $f(\cdot)$ могут использоваться различные модели МО, например, линейная регрессия (для простоты и интерпретируемости), градиентный бустинг (например, CatBoost, для улавливания сложных зависимостей) или нейронные сети.

2. **Последовательное тестирование на остатках:**

(a) Инициализируется кумулятивная тестовая статистика, например, логарифм отношения правдоподобий (LLR) $S_0 = 0$. Счетчик наблюдений (или пар наблюдений) $n = 0$. Задаются уровни ошибок α (ошибка первого рода) и β (ошибка второго рода). Вычисляются границы остановки Вальда: $A = \log\left(\frac{1-\beta}{\alpha}\right)$ и $B = \log\left(\frac{\beta}{1-\alpha}\right)$.

(b) Для каждого нового поступающего наблюдения $i = 1, 2, \ldots, N_{max}$ (где N_{max} — максимальное запланированное число наблюдений) с ковариатами X_i и наблюдаемой метрикой Y_i:

i. Вычисляется предсказанное значение метрики с помощью предварительно обученной модели $f(\cdot)$: $\hat{Y}_i = f(X_i)$.

ii. Вычисляется остаток (скорректированная метрика): $R_i = Y_i - \hat{Y}_i$.

iii. На основе остатка R_i (или пары остатков из контрольной и тестовой групп) вычисляется приращение логарифма отношения правдоподобий ΔLLR_i. Детальное описание построения этой статистики приведено в Разделе 3.2.

iv. Обновляется тестовая статистика: $S_n = S_{n-1} + \Delta LLR_i$ (где n — текущее количество обработанных наблюдений или пар).

v. Проверяются условия остановки:

- Если $S_n \geq A$, эксперимент останавливается, и нулевая гипотеза H_0 отвергается в пользу H_1.
- Если $S_n \leq B$, эксперимент останавливается, и принимается гипотеза H_0.
- Если $B < S_n < A$, эксперимент продолжается.

(c) Если решение не принято после обработки N_{max} наблюдений, эксперимент останавливается, и, как правило, принимается H_0 (или выносится суждение о невозможности сделать вывод при данном объеме выборки).

Важно отметить, что дисперсия остатков σ_R^2, необходимая для вычисления ΔLLR_i (см. Раздел 3.2), должна быть оценена на данных, независимых от тех, что используются для обучения модели $f(X)$ и для самого последовательного теста.

3.2 Построение тестовой статистики на остатках

Как указано в общей идее решения (Раздел 2.2), центральным элементом метода является последовательный тест, применяемый к остаткам $R_i = Y_i - f(X_i)$. При корректно обученной модели $f(X)$, которая не вносит систематического смещения, и при верной нулевой гипотезе H_0 (т.е. отсутствии эффекта воздействия), математические ожидания остатков в контрольной ($\mu_{R,C}$) и тестовой ($\mu_{R,T}$) группах должны быть равны (или близки к нулю, если $f(X)$ моделирует базовый уровень метрики). Альтернативная гипотеза H_1 предполагает наличие различия $\delta_R = \mu_{R,T} - \mu_{R,C} \neq 0$.

Для построения тестовой статистики предлагается использовать логарифм отношения правдоподобий (LLR) для остатков. Для k-го наблюдения R_k (которое может быть остатком из тестовой группы или разностью остатков тестовой и контрольной групп, в зависимости от конкретной формулировки

теста), приращение LLR вычисляется как:

$$\Delta LLR_k = \log \frac{p(R_k|H_1, \text{параметры } H_1)}{p(R_k|H_0, \text{параметры } H_0)}. \tag{3.1}$$

Здесь $p(R_k|H_j, \text{параметры } H_j)$ — плотность вероятности (или функция вероятности для дискретных данных) остатка R_k при условии истинности гипотезы H_j с соответствующими параметрами.

Для случая, когда R_k — это остаток из тестовой группы, $H_0 : \mathbb{E}[R_k] = 0$ и $H_1 : \mathbb{E}[R_k] = \delta_{R,alt}$ (где $\delta_{R,alt}$ — это ожидаемый или минимально значимый эффект на уровне остатков), и остатки предполагаются нормально распределенными с известной (или оцененной на независимых данных) дисперсией σ_R^2, формула (3.1) принимает вид:

$$\Delta LLR_k = \frac{\delta_{R,alt} R_k}{\sigma_R^2} - \frac{\delta_{R,alt}^2}{2\sigma_R^2}. \tag{3.2}$$

Кумулятивный LLR $S_n = \sum_{k=1}^{n} \Delta LLR_k$ затем сравнивается с границами Вальда A и B.

Хотя в данной работе основной упор делается на LLR-статистику в рамках SPRT, стоит отметить, что экспонента ΔLLR_k, то есть $E_k = \exp(\Delta LLR_k)$, представляет собой отношение правдоподобий. Такое отношение правдоподобий является основой для построения e-значений, обсуждавшихся в Главе 1 (в частности, Раздел 1.3.1). В этом контексте, произведение $M_n = \prod_{k=1}^{n} E_k = \exp(S_n)$ может рассматриваться как тестовый мартингал (или e-процесс), если E_k сконструированы таким образом, чтобы $\mathbb{E}_{H_0}[E_k|\mathcal{F}_{k-1}] \leq 1$.

3.3 Свойства и теоретические гарантии

Контроль ошибок первого и второго рода. Классический последовательный критерий отношения правдоподобия (SPRT), на котором основан предлагаемый метод ST-MLCA, теоретически обеспечивает контроль ошибок первого рода α и второго рода β на заданных уровнях при остановке

по достижении границ A или B. То есть, Pr(отвергнуть $H_0|H_0$ верна) $\approx \alpha$ и Pr(принять $H_0|H_1$ верна) $\approx \beta$. Ожидается, что предложенный метод ST-MLCA, будучи основанным на SPRT, применяемом к корректно полученным остаткам, будет также контролировать ошибку первого рода на заданном уровне. Это является предметом проверки Гипотезы 4 (Раздел 2.3).

Если рассматривать $M_n = \exp(S_n)$ как e-процесс, и если верхняя граница остановки A выбрана как $A = \log(1/\alpha_{\text{eff}})$, то по неравенству Вилле (для e-процессов):

$$\Pr_{H_0}\left(\sup_{n\geq 0} M_n \geq \exp(A)\right) \leq \alpha_{\text{eff}}. \tag{3.3}$$

Это свойство (anytime-validity), обсуждавшееся в Главе 1, особенно ценно, если тест может быть остановлен в любой момент. Для SPRT с двумя границами, контроль ошибок α и β гарантируется для решения, принятого по достижении одной из границ.

Ключевые предположения для корректной работы метода и достижения ожидаемых теоретических гарантий:

1. Модель $f(X)$ адекватно описывает зависимость Y от X при H_0 и не вносит систематического смещения в остатки R_i (т.е. $\mathbb{E}_{H_0}[R_i|\text{группа A}] = \mathbb{E}_{H_0}[R_i|\text{группа B}]$).

2. Данные (X_i, Y_i), используемые для вычисления остатков R_i и обновления S_n, независимы от данных D_{train}, на которых обучалась модель $f(X)$.

3. Оценка дисперсии остатков σ_R^2 является достаточно точной и получена на независимых данных.

3.3.1 Статистическая мощность и среднее число наблюдений (ASN).

Ожидается, что ST-MLCA позволит сократить среднее число наблюдений (ASN) по сравнению с последовательными тестами без коррекции на ковариаты (например, классическим SPRT на исходных Y_i) и/или повысить

мощность при фиксированном ASN. Это ожидание основано на потенциальном снижении дисперсии анализируемых величин. Если модель $f(X)$ объясняет значительную долю вариации в Y, то дисперсия остатков $\sigma_R^2 = \mathrm{Var}(R_i)$ должна быть существенно меньше дисперсии исходной метрики $\sigma_Y^2 = \mathrm{Var}(Y_i)$. Уменьшение дисперсии, аналогично эффекту в CUPED (уравнение (1.5) на стр. 10), должно приводить к более быстрому накоплению LLR при наличии истинного эффекта δ_R. Это предположение формирует основу для исследовательской гипотезы 1 (Раздел 2.3). Эффективность снижения ASN, как предполагается в Гипотезе 2, будет напрямую связана с качеством модели $f(X)$, измеряемым, например, коэффициентом детерминации R^2.

3.4 Выбор модели машинного обучения и практические аспекты

3.4.1 Выбор модели

Качество модели $f(X)$ является ключевым фактором, влияющим на потенциальную эффективность ST-MLCA.

- **Линейные модели** (например, линейная или логистическая регрессия): Просты в реализации и интерпретации. Ожидается, что они будут эффективны, если зависимость Y от X близка к линейной.
- **Нелинейные модели** (например, градиентный бустинг, случайный лес, нейронные сети): Способны улавливать сложные, нелинейные зависимости и взаимодействия между ковариатами. Предполагается (согласно Гипотезе 3, Раздел 2.3), что они могут обеспечить большее снижение дисперсии остатков и, как следствие, большую эффективность последовательного теста, если такие зависимости действительно присутствуют в данных.

Выбор конкретного алгоритма МО должен основываться на предварительном

анализе данных, доступных вычислительных ресурсах и требуемой интерпретируемости.

3.4.2 Ограничения метода

- Эффективность ST-MLCA напрямую зависит от наличия информативных ковариат X и способности модели $f(X)$ уловить их связь с Y. Если ковариаты слабо связаны с Y (низкий R^2), то выигрыш от их использования будет, вероятно, незначительным.

- Метод не устраняет фундаментальные проблемы экспериментального дизайна (например, смещение выборки, ошибки измерения). Он направлен на повышение эффективности анализа при корректно поставленном эксперименте.

- Обучение и применение сложных МО моделей может требовать значительных вычислительных ресурсов. Однако в рамках ST-MLCA модель обучается однократно перед началом последовательного теста, что снижает накладные расходы в процессе самого A/B-теста по сравнению с методами, требующими переобучения на каждом шаге (как, например, некоторые варианты E-C2ST).

Несмотря на эти ограничения, предлагаемый метод ST-MLCA обладает потенциалом для существенного повышения эффективности последовательного тестирования гипотез в условиях, когда доступна релевантная информация о ковариатах. Проверка выдвинутых исследовательских гипотез (Раздел 2.3) и количественная оценка преимуществ метода будут представлены в следующей главе.

В Главе 3 был детально описан предлагаемый метод ST-MLCA. Представлен пошаговый алгоритм метода, включая этап обучения предиктивной модели и этап последовательного тестирования на скорректированных остатках. Рассмотрено построение тестовой статистики (LLR), обсуждены теоретические свойства, ожидаемые гарантии контроля ошибок и предположения, лежащие

в основе метода. Также были затронуты практические аспекты выбора модели машинного обучения и ограничения метода.

ГЛАВА 4

Экспериментальная оценка эффективности метода ST-MLCA

В данной главе проводится экспериментальная оценка эффективности предложенного метода последовательного тестирования гипотез с коррекцией на ковариаты на основе машинного обучения (ST-MLCA). Целью экспериментов является проверка выдвинутых в Главе 3 исследовательских гипотез, а именно: способности метода ST-MLCA сокращать среднее число наблюдений (ASN) и повышать статистическую мощность по сравнению с классическим последовательным тестом без учета ковариат, зависимости эффективности от качества предсказательной модели и преимущества использования нелинейных моделей при сложных зависимостях в данных. Особое внимание уделяется проверке корректности контроля ошибки первого рода. Эксперименты проводятся на синтетических данных, что позволяет полностью контролировать параметры процесса генерации и истинные эффекты. Также обсуждаются перспективы применения метода на реальных данных.

4.1 Эксперименты на синтетических данных

Эксперименты на синтетических данных были спланированы таким образом, чтобы всесторонне исследовать свойства метода ST-MLCA в контролируемых условиях.

4.1.1 Цели и постановка эксперимента

Основными целями экспериментов на синтетических данных являлись:

1. Количественная оценка выигрыша в ASN и/или мощности метода ST-MLCA по сравнению со стандартным последовательным тестом (SPRT), не использующим информацию о ковариатах.

2. Исследование зависимости эффективности ST-MLCA от предсказательной силы модели машинного обучения M, характеризуемой коэффициентом детерминации R^2 связи между ковариатами X и целевой метрикой Y.

3. Сравнение эффективности ST-MLCA при использовании линейных и нелинейных моделей машинного обучения для коррекции на ковариаты, особенно в сценариях с нелинейной зависимостью Y от X.

4. Проверка корректности контроля ошибки первого рода (Гипотеза 4): доля ложных отвержений нулевой гипотезы H_0 не должна превышать заданный уровень значимости α.

4.1.2 Генерация данных

Для моделирования экспериментов генерировались синтетические данные, имитирующие целевую метрику Y и единственный ковариат X. Ковариат X генерировался из стандартного нормального распределения $X \sim N(0, 1)$. Целевая метрика Y для каждого наблюдения i формировалась по следующей общей модели:

$$Y_i = \theta \cdot \text{Group}_i + g(X_i) + \epsilon_i, \tag{4.1}$$

где:

- Group_i — индикатор принадлежности наблюдения к тестовой группе (1 для тестовой, 0 для контрольной).

- θ — истинный эффект воздействия (treatment effect). При проверке H_0, $\theta = 0$. При проверке H_1, θ принимал значения 0.5 или 1.0.

- $g(X_i)$ — функция, описывающая зависимость целевой метрики от ковариаты X_i.

- ϵ_i — случайный шум, генерируемый из нормального распределения $N(0, \sigma_\epsilon^2)$. В экспериментах дисперсия шума σ_ϵ^2 была зафиксирована на уровне 1.0.

Функция $g(X)$ калибровалась таким образом, чтобы стандартизованная версия $g_{std}(X) = (g(X) - E[g(X)])/\sqrt{Var(g(X))}$ имела нулевое среднее и единичную дисперсию. Затем эта стандартизованная функция умножалась на амплитудный коэффициент A_g, чтобы обеспечить заданный целевой коэффициент детерминации R^2_{target} между $g(X)$ и Y (в отсутствие эффекта θ). Амплитуда A_g вычислялась как:

$$A_g = \sigma_\epsilon \cdot \sqrt{\frac{R^2_{target}}{1 - R^2_{target}}}. \tag{4.2}$$

Таким образом, целевая метрика Y без эффекта группы имела вид $Y_i = A_g \cdot g_{std}(X_i) + \epsilon_i$. Варьировались следующие параметры генерации данных:

- **Тип зависимости** $g(X)$:
  - *Линейный случай*: $g(X) = X$.
  - *Нелинейный случай*: $g(X) = \sin(3X) + 0.5X^3$.
- **Коэффициент детерминации** R^2_{target}: принимал значения из множества $\{0.0, 0.2, 0.5, 0.8\}$. Случай $R^2 = 0.0$ означает отсутствие связи между Y и X (т.е. $A_g = 0$).

Для обучения моделей машинного обучения $f(X)$ использовалась отдельная обучающая выборка размера $n_{train} = 100,000$ наблюдений, сгенерированная для каждого сценария (комбинации типа зависимости и R^2_{target}) при условии отсутствия эффекта группы ($\theta = 0$).

4.1.3 Сравниваемые методы и параметры тестирования

В экспериментах сравнивались следующие подходы к последовательному тестированию:

1. **Без коррекции ('no_adj')**: Стандартный последовательный критерий отношения правдоподобий (SPRT), применяемый непосредственно к исходным значениям метрики Y_i из тестовой группы. Этот метод служит базовым для сравнения.

2. **ST-MLCA с линейной моделью ('linear')**: Предлагаемый метод, где для предсказания $f(X_i)$ и вычисления остатков $R_i = Y_i - f(X_i)$ используется линейная регрессия ('sklearn.linear_model.LinearRegression').

3. **ST-MLCA с моделью CatBoost ('catboost')**: Предлагаемый метод, использующий градиентный бустинг CatBoost ('catboost.CatBoostRegressor') с параметрами по умолчанию, в качестве нелинейной модели $f(X_i)$.

4. **ST-MLCA с нейронной сетью ('nn')**: Предлагаемый метод, использующий многослойный перцептрон ('sklearn.neural_network.MLPRegressor') со структурой (50,50), функцией активации ReLU, 'max_iter=500', 'early_stopping=True') в качестве нелинейной модели $f(X_i)$.

Для всех методов, использующих коррекцию (linear, catboost, nn), модели $f(X)$ обучались однократно на независимой обучающей выборке D_{train}, как описано выше. Последовательный тест для всех методов строился на основе логарифма отношения правдоподобий (LLR). Для k-го наблюдения R_k из тестовой группы (где $R_k = Y_k$ для 'no_adj', и $R_k = Y_k - f(X_k)$ для ST-MLCA методов), приращение LLR вычислялось по формуле для нормально распределенных остатков:

$$\Delta LLR_k = \frac{\delta_{alt}}{\sigma_R^2} R_k - \frac{\delta_{alt}^2}{2\sigma_R^2}, \tag{4.3}$$

где δ_{alt} — это ожидаемый минимально значимый эффект, принятый равным 0.5 при построении альтернативной гипотезы для LLR, а σ_R^2 — дисперсия остатков R_k при верной нулевой гипотезе H_0. Значение σ_R для каждого метода и сценария оценивалось на отдельной выборке размера $n_{test_eval_sigma} = 10,000$ наблюдений, сгенерированных под H_0.

Кумулятивный LLR $S_n = \sum_{k=1}^{n} \Delta LLR_k$ сравнивался с границами Вальда: $A = \log((1-\beta)/\alpha)$ и $B = \log(\beta/(1-\alpha))$. Заданные уровни ошибок: $\alpha = 0.05$ (ошибка первого рода) и $\beta = 0.2$ (ошибка второго рода). Эксперимент останавливался, если $S_n \geq A$ (принимается H_1) или $S_n \leq B$ (принимается H_0). Максимальное число наблюдений в последовательном тесте для одной группы было ограничено $n_{max_steps} = 1,000$. Если решение не принималось до n_{max_steps}, считалось, что принимается H_0. Для каждого сценария (комбинация типа зависимости, R^2_{target}, истинного эффекта θ и метода) проводилось $n_{reps} = 1,000$ независимых симуляций.

4.1.4 Метрики оценки

Для оценки эффективности сравниваемых методов использовались следующие метрики:

- **Ошибка первого рода (Type I Error Rate)**: Доля симуляций, в которых H_0 была отвергнута, когда она была верна ($\theta = 0$). Ожидается, что эта величина не превышает $\alpha = 0.05$.
- **Мощность (Power)**: Доля симуляций, в которых H_0 была корректно отвергнута, когда была верна H_1 ($\theta > 0$).
- **Среднее число наблюдений (ASN)**: Среднее количество наблюдений (шагов k), потребовавшихся для принятия решения (остановки теста), усредненное по всем симуляциям для данного сценария.
- **Дисперсия остатков (σ^2_R)**: Оцененная дисперсия остатков R_i под H_0 для каждого метода коррекции. Эта метрика напрямую влияет на скорость накопления LLR и, следовательно, на ASN и мощность.
- **Распределение времени остановки**: Гистограммы, показывающие распределение числа наблюдений до остановки теста.

4.1.5 Результаты экспериментов и их анализ

Результаты моделирования представлены на серии графиков и обсуждаются ниже в контексте выдвинутых исследовательских гипотез.

Контроль ошибки первого рода (Гипотеза 4)

Ключевым требованием к любому статистическому тесту является корректный контроль ошибки первого рода. На Рисунках 4.1 и 4.2 представлена эмпирическая ошибка первого рода для всех сравниваемых методов в зависимости от R^2 для линейного и нелинейного сценариев соответственно, при $\theta = 0$.

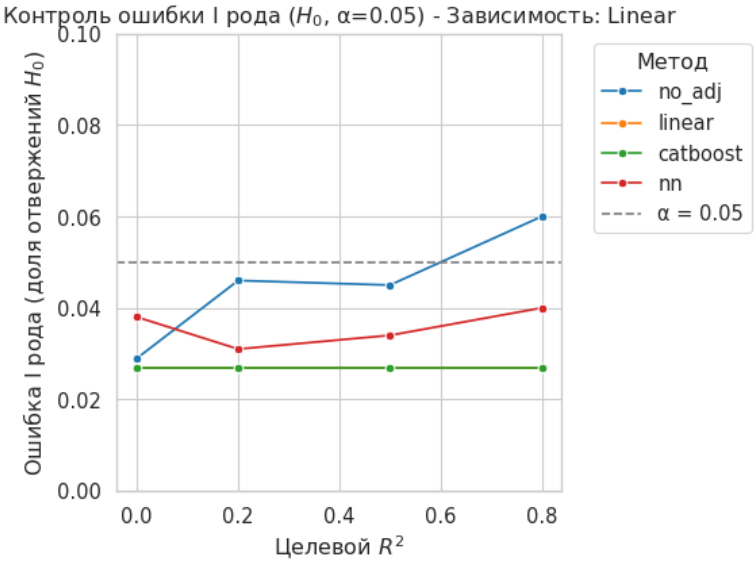

Рисунок 4.1 Контроль ошибки I рода ($H_0, \alpha = 0.05$) для методов в линейном сценарии зависимости Y от X при различных R^2. Горизонтальная пунктирная линия соответствует целевому уровню $\alpha = 0.05$.

Как видно из графиков, все методы, включая базовый ‘no_adj‘ и варианты ST-MLCA, в целом поддерживают ошибку первого рода на уровне, близком к номинальному $\alpha = 0.05$, или ниже его. Небольшие флуктуации вокруг 0.05 являются ожидаемыми из-за стохастической природы симуляций. В некоторых случаях, особенно для методов с коррекцией при высоких R^2, наблюдается даже более консервативное поведение (ошибка I рода ниже 0.05). Это может

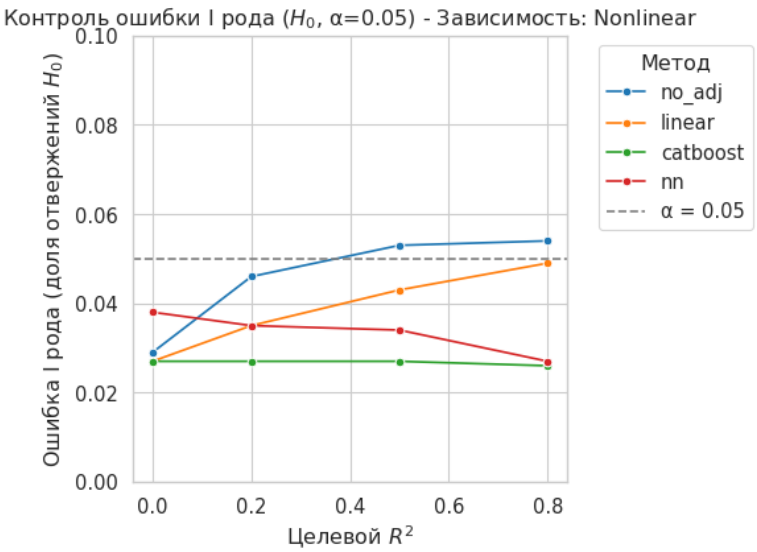

Рисунок 4.2 – Контроль ошибки I рода ($H_0, \alpha = 0.05$) для методов в нелинейном сценарии зависимости Y от X при различных R^2. Горизонтальная пунктирная линия соответствует целевому уровню $\alpha = 0.05$.

быть связано с тем, что оценка σ^2_R на отдельной выборке и последующее ее использование в LLR вносит некоторую дополнительную стабильность или неточность, которая в среднем не приводит к завышению ошибки первого рода. Важно, что систематического превышения уровня α не наблюдается, что подтверждает корректность предложенного подхода ST-MLCA с точки зрения контроля ложных срабатываний (Гипотеза 4).

Эффективность метода ST-MLCA (Гипотезы 1 и 2)

Далее рассмотрим, как применение ST-MLCA влияет на ASN и мощность теста. Гипотеза 1 предполагает, что ST-MLCA позволит сократить ASN, а Гипотеза 2 — что эффективность метода зависит от качества модели машинного обучения (т.е. от R^2).

Влияние на дисперсию остатков. Основой эффективности ST-MLCA является снижение дисперсии анализируемых данных (остатков R_i) по сравнению с исходной метрикой Y_i. На Рисунках 4.3 и 4.4 показана оцененная дисперсия остатков σ^2_R под H_0 для различных методов в зависимости от R^2

ковариат.

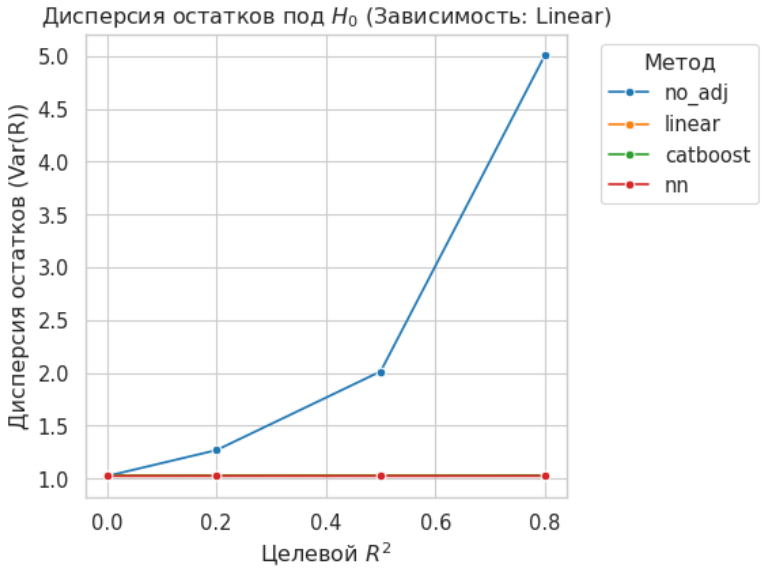

Рисунок 4.3 – Дисперсия остатков σ_R^2 под H_0 для различных методов в линейном сценарии в зависимости от целевого R^2. Для ‘no_adj‘ это $Var(Y)$.

В обоих сценариях (линейном и нелинейном) дисперсия для метода ‘no_adj‘ (что соответствует $Var(Y)$) закономерно растет с увеличением R^2, так как $Var(Y) = Var(g(X)) + Var(\epsilon) = A_g^2 + \sigma_\epsilon^2$, а A_g растет с R^2 (при фиксированной $\sigma_\epsilon^2 = 1$). Методы ST-MLCA (‘linear‘, ‘catboost‘, ‘nn‘) эффективно снижают дисперсию остатков. В идеальном случае, если модель $f(X)$ полностью объясняет $g(X)$, то $Var(R_i) \approx Var(\epsilon_i) = \sigma_\epsilon^2 = 1.0$. В линейном сценарии (Рис. 4.3) все модели коррекции (‘linear‘, ‘catboost‘, ‘nn‘) показывают очень близкую и высокую эффективность, удерживая дисперсию остатков около 1.0 независимо от R^2. Это ожидаемо, так как даже линейная модель способна хорошо уловить линейную зависимость. В нелинейном сценарии (Рис. 4.4) картина интереснее. Линейная модель (‘linear‘) снижает дисперсию лишь частично, так как она не может полностью объяснить нелинейную компоненту $g(X)$. Ее эффективность падает с ростом R^2 (т.е. с ростом вклада нелинейной компоненты). Нелинейные модели (‘catboost‘, ‘nn‘) значительно лучше справляются с этой задачей, удерживая дисперсию остатков близко к 1.0. Это демонстрирует важность выбора модели, адекватной сложности зависимостей в данных.

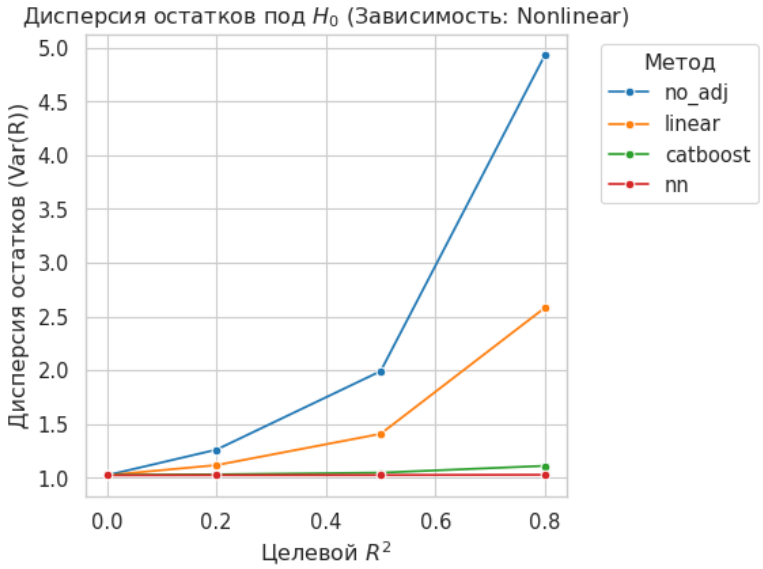

Рисунок 4.4 – Дисперсия остатков σ_R^2 под H_0 для различных методов в нелинейном сценарии в зависимости от целевого R^2. Для ‘no_adj‘ это $Var(Y)$.

Среднее число наблюдений (ASN) и мощность. Снижение дисперсии остатков должно приводить к уменьшению ASN для достижения той же мощности, или к увеличению мощности при том же ASN. Рассмотрим эти показатели для истинного эффекта $\theta = 0.5$.

На Рисунках 4.5 и 4.6 показана зависимость ASN от R^2 для линейного и нелинейного сценариев.

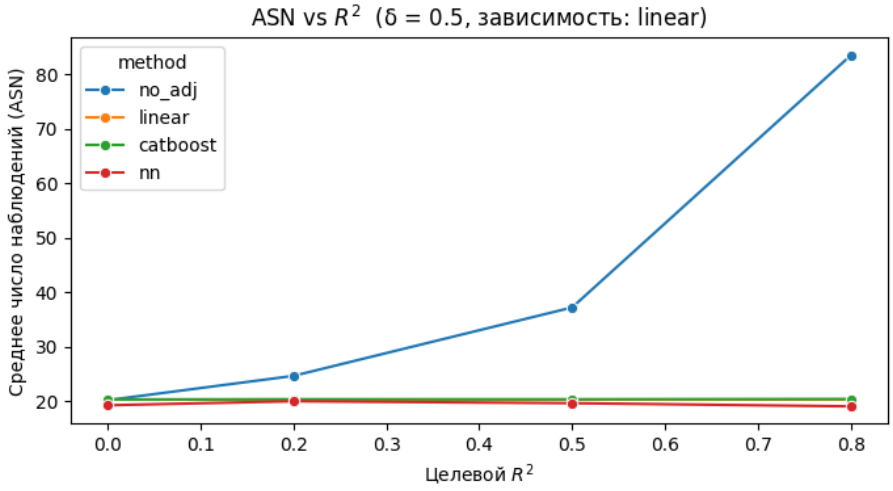

Рисунок 4.5 – Зависимость ASN от R^2 при $\theta = 0.5$ в линейном сценарии.

Видно, что для метода ‘no_adj‘ ASN резко возрастает с увеличением R^2.

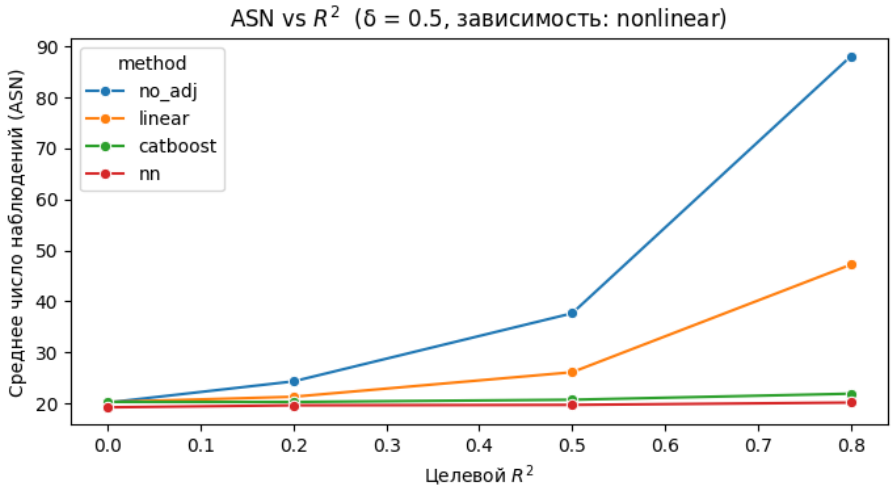

Рисунок 4.6 – Зависимость ASN от R^2 при $\theta = 0.5$ в нелинейном сценарии.

Это связано с тем, что при росте R^2 растет и общая дисперсия Y (как показано на Рис. 4.3, 4.4), что затрудняет обнаружение фиксированного эффекта $\theta = 0.5$. Методы ST-MLCA демонстрируют значительно более низкий ASN. Причем, чем выше R^2 (т.е. чем больше информации несут ковариаты), тем больше выигрыш по сравнению с ‘no_adj‘. ASN для методов ST-MLCA остается относительно стабильным или даже немного уменьшается с ростом R^2, так как они эффективно устраняют вариацию, связанную с X.

На Рисунках 4.7 и 4.8 показана зависимость мощности от R^2.

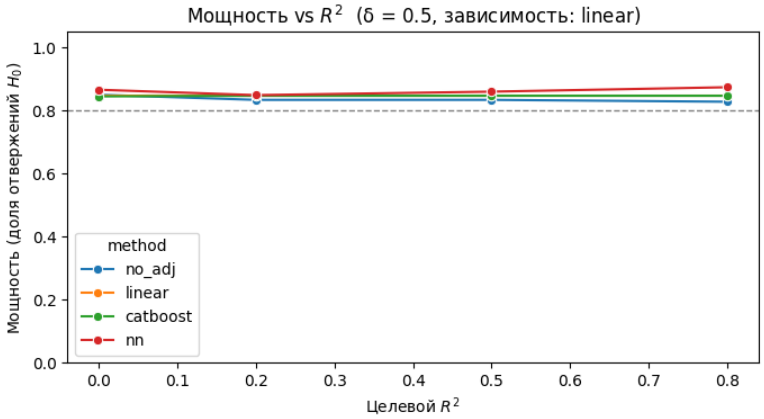

Рисунок 4.7 – Зависимость мощности от R^2 при $\theta = 0.5$ (целевая мощность $1 - \beta = 0.8$) в линейном сценарии.

Мощности всех методов остаются на уровне целевых $1 - \beta = 0.8$ при всех R^2 как в линейном сценарии, так и в нелинейном.

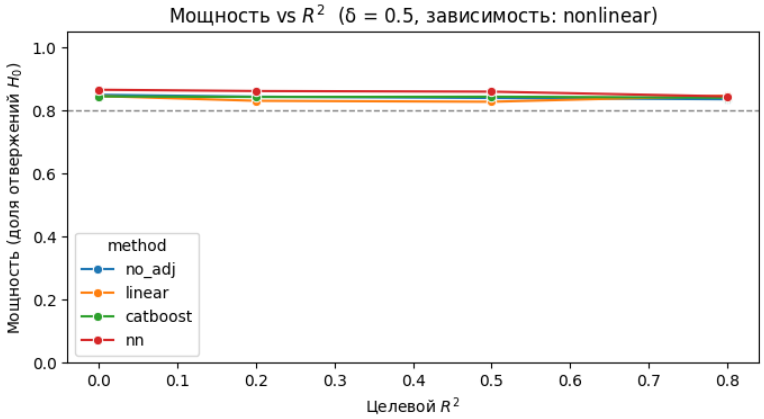

Рисунок 4.8 – Зависимость мощности от R^2 при $\theta = 0.5$ (целевая мощность $1 - \beta = 0.8$) в нелинейном сценарии.

Для более наглядной демонстрации компромисса между ASN и мощностью, на Рисунках 4.9 и 4.10 представлены диаграммы "Мощность vs ASN"для фиксированного $R^2 = 0.5$ и $\theta = 0.5$. Идеальный метод находился бы в левом верхнем углу (максимальная мощность при минимальном ASN).

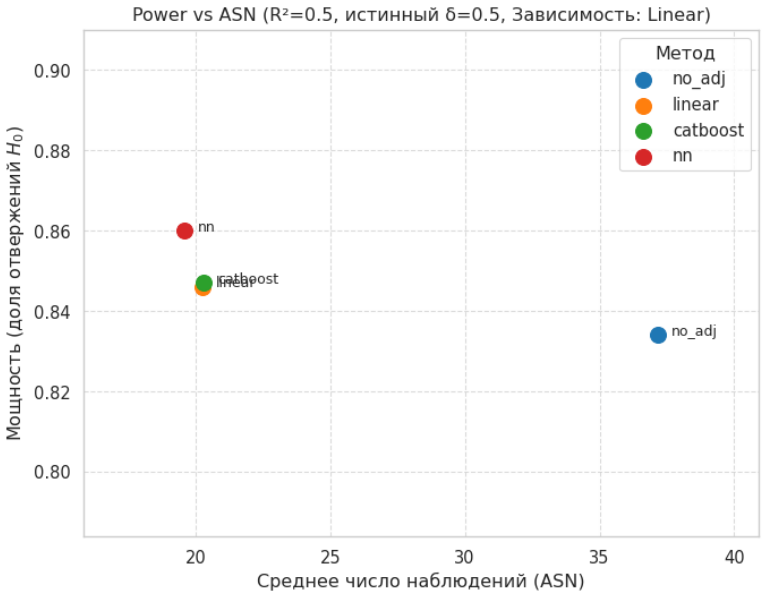

Рисунок 4.9 – Мощность в зависимости от ASN для различных методов при $R^2 = 0.5$ и $\theta = 0.5$ в линейном сценарии.

Эти графики отчетливо показывают преимущество ST-MLCA: методы с коррекцией (‘linear‘, ‘catboost‘, ‘nn‘) обеспечивают заданную мощность при значительно меньшем ASN по сравнению с ‘no_adj‘.

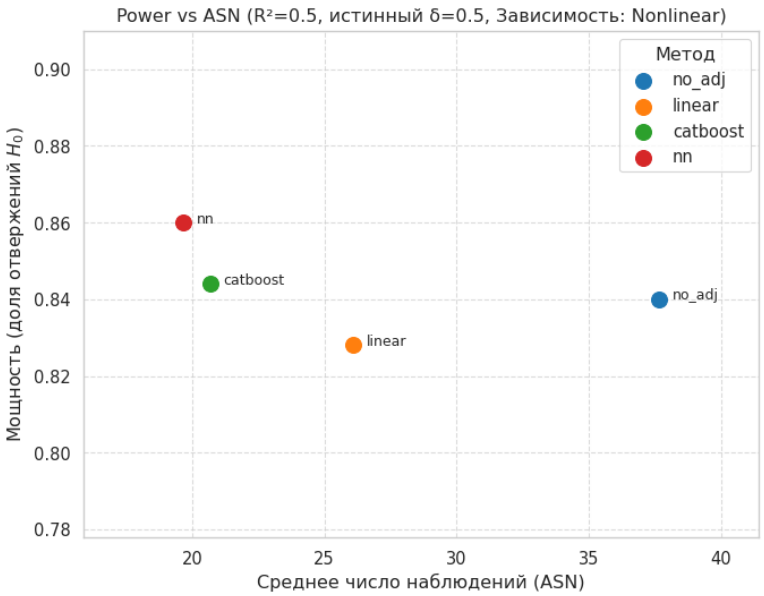

Рисунок 4.10 – Мощность в зависимости от ASN для различных методов при $R^2 = 0.5$ и $\theta = 0.5$ в нелинейном сценарии.

Преимущество нелинейных моделей (Гипотеза 3)

Гипотеза 3 утверждает, что при наличии сложных нелинейных зависимостей между Y и X, использование нелинейных моделей МО (таких как CatBoost или нейронные сети) в ST-MLCA даст большее снижение дисперсии и, как следствие, лучшую эффективность, чем использование простой линейной модели. Это предположение хорошо иллюстрируется сравнением результатов для линейного и нелинейного сценариев.

В **линейном сценарии** (например, Рис. 4.5 или 4.9), все три модели коррекции (‘linear‘, ‘catboost‘, ‘nn‘) показывают практически идентичную высокую эффективность. Это логично, так как линейная зависимость легко улавливается всеми типами моделей, и более сложные нелинейные модели не дают здесь дополнительного преимущества, но и не проигрывают существенно.

В **нелинейном сценарии** ситуация иная. Как видно из Рис. 4.4, линейная модель (‘linear‘) не способна полностью устранить влияние нелинейной компоненты $g(X)$, и ее остаточная дисперсия выше, чем у ‘catboost‘ и ‘nn‘, особенно при высоких R^2. Это напрямую сказывается на ASN (Рис. 4.6) и на соотношении мощность/ASN (Рис. 4.10). В нелинейном сценарии при $R^2 = 0.5$

и $\theta = 0.5$:

- ‘no_adj‘: ASN ≈ 38, Мощность ≈ 0.84
- ‘linear‘: ASN ≈ 26, Мощность ≈ 0.83
- ‘catboost‘: ASN ≈ 20, Мощность ≈ 0.84
- ‘nn‘: ASN ≈ 19.5, Мощность ≈ 0.86

Методы ‘catboost‘ и ‘nn‘ показывают заметно лучший результат по ASN, чем ‘linear‘, при сохранении или даже небольшом увеличении мощности. Это подтверждает Гипотезу 3: при нелинейных связях гибкие нелинейные модели в ST-MLCA предпочтительнее.

Распределение времени остановки

Помимо средних значений ASN, интересно посмотреть на полное распределение времени остановки. На Рисунке 4.11 показано распределение времени остановки для метода ST-MLCA (‘linear‘) при двух разных истинных эффектах ($\theta = 0.5$ и $\theta = 1.0$) в линейном сценарии с $R^2 = 0.5$.

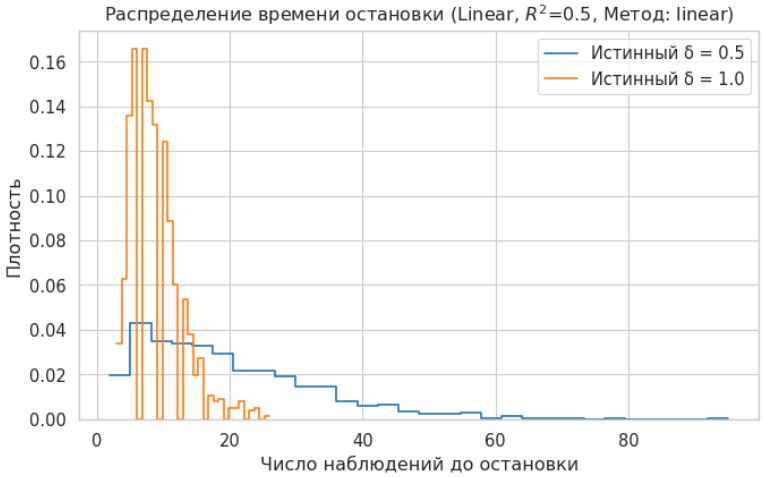

Рисунок 4.11 – Распределение времени остановки для ST-MLCA (linear model) при $R^2 = 0.5$ в линейном сценарии для истинных эффектов $\theta = 0.5$ и $\theta = 1.0$.

Как и ожидалось, при большем истинном эффекте ($\theta = 1.0$) тест останавливается значительно раньше (распределение смещено влево) и с меньшей дисперсией времени остановки, чем при меньшем эффекте ($\theta = 0.5$).

На Рисунке 4.12 сравниваются распределения времени остановки для метода ‘no_adj‘ при низком ($R^2 = 0.2$) и высоком ($R^2 = 0.8$) уровне объясняемой ковариатами дисперсии в линейном сценарии при $\theta = 0.5$.

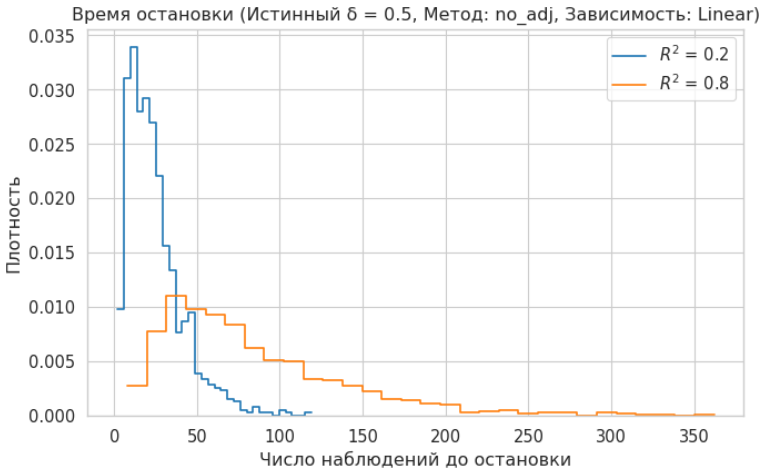

Рисунок 4.12 – Распределение времени остановки для метода ‘no_adj‘ (без коррекции) при $\theta = 0.5$ в линейном сценарии для $R^2 = 0.2$ и $R^2 = 0.8$.

Этот график иллюстрирует, почему ASN для ‘no_adj‘ растет с R^2: при высоком $R^2 = 0.8$ (и, соответственно, большей общей дисперсии Y) тест останавливается значительно позже и с большим разбросом, чем при $R^2 = 0.2$. Методы ST-MLCA, эффективно снижая дисперсию, позволяют избежать такого ухудшения.

4.2 Обсуждение результатов экспериментов

Проведенные на синтетических данных эксперименты позволили подтвердить основные исследовательские гипотезы, выдвинутые в данной работе.

Подтверждение гипотез:

- **Гипотеза 1 (Сокращение ASN) и Гипотеза 2 (Зависимость от качества модели)**: Эксперименты убедительно показали, что метод ST-MLCA позволяет значимо уменьшить среднее число наблюдений (ASN)

по сравнению с классическим последовательным тестом без коррекции на ковариаты (Рис. 4.5, 4.6, 4.9, 4.10). Выигрыш в ASN тем существеннее, чем выше предсказательная сила модели машинного обучения (коэффициент детерминации R^2). Это связано с эффективным снижением дисперсии остатков (Рис. 4.3, 4.4), что ускоряет накопление свидетельств в пользу одной из гипотез. При $R^2 = 0$ все методы с коррекцией ожидаемо сводятся по эффективности к методу 'no_adj'.

- **Гипотеза 3 (Преимущество нелинейных моделей)**: В сценариях с нелинейной зависимостью целевой метрики от ковариат (Рис. 4.6, 4.10), нелинейные модели машинного обучения (CatBoost, нейронная сеть) обеспечили большую эффективность (меньший ASN при той же мощности), чем линейная модель. Это подчеркивает важность выбора модели, адекватной сложности зависимостей в данных, для максимального использования потенциала ковариатной коррекции. В линейных же сценариях более сложные модели не показали существенного проигрыша по сравнению с линейной.

- **Гипотеза 4 (Контроль уровня значимости)**: Все варианты метода ST-MLCA, а также базовый метод, продемонстрировали корректный контроль ошибки первого рода, не превышая (а иногда будучи даже консервативнее) заданный уровень значимости $\alpha = 0.05$ (Рис. 4.1, 4.2). Это является критически важным свойством для любого статистического теста.

Компромиссы и ограничения: Несмотря на очевидные преимущества, метод ST-MLCA имеет свои особенности. Выигрыш от его применения напрямую зависит от наличия информативных ковариат и качества обученной модели $f(X)$. Если ковариаты слабо коррелируют с целевой метрикой (низкий R^2), то и выигрыш будет незначительным. Кроме того, обучение более сложных моделей машинного обучения требует больших объемов независимых данных D_{train} и вычислительных ресурсов, хотя в рамках предлагаемого подхода ST-MLCA модель обучается однократно перед началом последовательного

теста, что снижает накладные расходы по сравнению с методами, требующими переобучения на каждом шаге. Важно также обеспечить независимость данных для обучения модели от данных, используемых в самом последовательном тесте, во избежание переобучения и "утечки информации", которая могла бы привести к некорректным выводам и завышению ошибки первого рода.

Сравнение с аналогами из литературы (качественное): Методы типа E-C2ST, упомянутые в обзоре литературы (Раздел **??**), также используют машинное обучение для тестирования гипотез, но их принцип действия иной: они строят классификатор для различения выборок и используют качество классификации как тестовую статистику. Теоретически, E-C2ST способны улавливать очень сложные различия в распределениях, но часто требуют переобучения или дообучения классификатора на каждом шаге (или батче данных) и аккуратной калибровки e-value, что может быть вычислительно затратно и сложно в реализации. Предложенный метод ST-MLCA, фокусируясь на снижении дисперсии через предсказание базового уровня метрики, концептуально ближе к идее CUPED, но адаптирован для последовательного анализа и позволяет использовать гибкие нелинейные модели. Однократное обучение модели $f(X)$ делает ST-MLCA потенциально более простым в развертывании и менее требовательным к ресурсам в процессе самого A/B-теста.

4.2.1 Применение метода на реальных данных

Для демонстрации практической применимости метода ST-MLCA и проверки его эффективности в условиях, приближенных к реальным задачам A/B-тестирования, были проведены эксперименты на двух общедоступных наборах данных: Hillstrom E-Mail Marketing и Criteo Uplift Prediction.

Эксперименты на наборе данных Hillstrom

Описание данных и постановка задачи. Набор данных Hillstrom (Kevin Hillstrom's MineThatData E-Mail Analytics Challenge, 2008) [11] содержит информацию о клиентах и их реакции на различные маркетинговые email-кампании.

В экспериментах в качестве целевой метрики Y использовался бинарный признак «конверсия» (conversion), который принимал значение 1, если клиент совершил покупку после получения email (или в соответствующий период для контрольной группы), и 0 в противном случае.

- Контрольная группа (A): клиенты, не получавшие email (segment == "No E-Mail").
- Тестовая группа (B): клиенты, получившие email, ориентированный на женщин (segment == "Womens E-Mail").

Ковариаты X включали различные характеристики клиентов, такие как история покупок, активность на сайте и т.д. Средний уровень конверсии в контрольной группе p_{control} использовался как базовый для теста SPRT без коррекции.

Параметры эксперимента.

- Уровни ошибок для SPRT: $\alpha = 0.05$, $\beta = 0.20$.
- Ожидаемый минимально значимый эффект для параметризации альтернативной гипотезы H_1 в SPRT (Δ_{alt}): использовалось значение 0.001938. Для биномиального SPRT (метод no_adj) $p_1 = p_{\text{control}} + \Delta_{\text{alt}}$. Для SPRT на остатках (методы ST-MLCA) ожидаемое среднее остатков при H_1 полагалось равным Δ_{alt}.
- Модели машинного обучения для ковариатной коррекции в ST-MLCA:
  1. Логистическая регрессия (sklearn.linear_model.LogisticRegression).
  2. Градиентный бустинг CatBoost (catboost.CatBoostClassifier).
- Обучение моделей коррекции: Модели $f(X)$ обучались на данных контрольной группы (A). Дисперсия остатков σ_R^2 для ST-MLCA оценивалась на основе остатков, полученных с помощью out-of-fold предсказаний на контрольной группе. Финальная модель для коррекции данных из тестовой группы обучалась на всех данных контрольной группы.

- Максимальное количество наблюдений в последовательном тесте для одной группы: $N_{\max} = 1000$.
- Количество повторений симуляций: $N_{\text{reps}} = 1000$ для оценки мощности и ASN, $N_{\text{type1_reps}} = 2000$ для оценки ошибки первого рода.

Результаты. Результаты экспериментов на наборе данных Hillstrom представлены в Таблице 4.1.

Таблица 4.1 – Результаты экспериментов на наборе данных Hillstrom ($\Delta_{\text{alt}} = 0.001938$)

Метод	Модель для коррекции	Мощность	ASN	Ошибка I рода	σ_R
no_adj	–	0.857	231.3	0.028	–
ST-MLCA	Лог. регрессия	0.807	181.8	0.040	0.3038
ST-MLCA	CatBoost	0.809	182.7	0.040	0.3047

Обсуждение результатов для Hillstrom. Методы ST-MLCA с использованием как логистической регрессии, так и CatBoost, продемонстрировали сокращение среднего числа наблюдений (ASN) по сравнению со стандартным SPRT без коррекции (no_adj). ASN для ST-MLCA (около 182) был примерно на 21-22% ниже, чем для no_adj (231.3). При этом наблюдается некоторое снижение мощности для методов с коррекцией (0.807-0.809) по сравнению с no_adj (0.857). Важно отметить, что оценка ошибки первого рода для no_adj (0.028) была существенно ниже номинального уровня $\alpha = 0.05$, в то время как для ST-MLCA она была ближе к номиналу (0.040).

На данном наборе данных и при выбранной конфигурации эксперимента, модель CatBoost не показала значимого преимущества перед более простой логистической регрессией в рамках ST-MLCA: их показатели ASN и мощности очень близки. Значения стандартного отклонения остатков σ_R для обеих моделей коррекции также оказались практически одинаковыми.

Эксперименты на наборе данных Criteo Uplift Prediction (mini)

Описание данных и постановка задачи. Для второго эксперимента использовалась уменьшенная выборка (1 млн наблюдений) из набора данных Criteo Uplift Prediction Challenge [12]. Целевой метрикой Y также являлась бинарная «конверсия» (conversion).

- Контрольная группа (A): пользователи, которым не показывалась реклама (treatment $== 0$).
- Тестовая группа (B): пользователи, которым показывалась реклама (treatment $== 1$).

Ковариаты X представляли собой 12 анонимизированных численных признаков пользователей. В ходе предварительного анализа данных были получены следующие оценки:

- Средний уровень конверсии в контрольной группе $p_{\text{control}} = 0.001938$.
- Средний уровень конверсии в тестовой группе $p_{\text{treat_actual}} = 0.003089$.
- Фактический наблюдаемый эффект (разница конверсий) в данных $\Delta_{\text{true}} = p_{\text{treat_actual}} - p_{\text{control}} = 0.001152$.

Параметры эксперимента.

- Уровни ошибок для SPRT: $\alpha = 0.05$, $\beta = 0.20$.
- Ожидаемый минимально значимый эффект для параметризации H_1 (Δ_{alt}): 0.0011152. Это значение было выбрано близким к наблюдаемому эффекту Δ_{true}.
- Модели машинного обучения для ковариатной коррекции:
  1. Логистическая регрессия (sklearn.linear_model.LogisticRegression).
  2. Градиентный бустинг CatBoost (catboost.CatBoostClassifier).

- Обучение модели коррекции: аналогично экспериментам с Hillstrom.
- Максимальное количество наблюдений в последовательном тесте для одной группы: $N_{\max} = 10000$.
- Количество повторений симуляций: $N_{\text{reps}} = 1000$ для мощности/ASN, $N_{\text{type1_reps}} = 2000$ для ошибки I рода.

Результаты. Результаты экспериментов на наборе данных Criteo (mini) представлены в Таблице 4.2.

Таблица 4.2 – Результаты экспериментов на наборе данных Criteo (mini) ($\Delta_{\text{alt}} = 0.0011152$)

Метод	Модель для коррекции	Мощность	ASN	Ошибка I рода	σ_R
no_adj	–	0.619	6018.7	0.040	–
ST-MLCA	Лог. регрессия	0.588	4300.9	0.057	0.0419
ST-MLCA	CatBoost	0.591	4118.3	0.055	0.0412

Обсуждение результатов для Criteo. На наборе данных Criteo метод ST-MLCA с использованием CatBoost также продемонстрировал значительное сокращение ASN по сравнению с тестом без коррекции и тестом с линейной коррекцией. ASN для ST-MLCA (4118.3) был на 31.6% ниже, чем для no_adj (6018.7) и на 4.2% нижя по сравнению с линейной коррекцией (4300.9). Это подтверждает эффективность ковариатной коррекции для ускорения последовательных тестов.

Мощность обоих методов (около 0.6) оказалась ниже целевого уровня $1 - \beta = 0.80$. Мощность ST-MLCA (0.591) была незначительно ниже, чем у no_adj (0.619). Это может быть связано с тем, что несмотря на близость Δ_{alt} к Δ_{true}, максимальное число наблюдений $N_{\max}$ или другие характеристики данных могли ограничивать достижение целевой мощности. Оценка ошибки первого рода для ST-MLCA на этом наборе данных составила 0.055, что

немного превышает номинальный уровень $\alpha = 0.05$. Для метода no_adj ошибка первого рода (0.040) находилась в допустимых пределах.

Стандартное отклонение остатков $\sigma_R = 0.0412$ для ST-MLCA оказалось значительно ниже, чем в экспериментах с Hillstrom, что соответствует природе данных с очень низкими уровнями конверсии и малыми абсолютными эффектами.

Общее обсуждение результатов на реальных данных

Эксперименты, проведенные на двух различных реальных наборах данных, подтверждают основной вывод, полученный на синтетических данных: метод ST-MLCA, использующий модели машинного обучения для ковариатной коррекции, способен существенно сократить среднее число наблюдений (ASN), необходимое для принятия решения в последовательном A/B-тестировании, по сравнению с классическим SPRT без учета ковариат.

Выбор конкретной модели машинного обучения (например, логистическая регрессия или CatBoost) и степень выигрыша от ее применения могут зависеть от характеристик набора данных, таких как наличие и сила нелинейных зависимостей между ковариатами и целевой метрикой. В эксперименте с Hillstrom более сложная модель CatBoost не дала преимуществ перед логистической регрессией. В то же время, на данных Criteo, где использовался CatBoost, наблюдалось значительное снижение ASN.

Наблюдаемые уровни мощности в некоторых случаях были ниже целевых, что подчеркивает важность аккуратной параметризации теста (включая Δ_{alt}) и выбора достаточного максимального объема выборки ($N_{\max}$) с учетом вариативности данных. Контроль ошибки первого рода в большинстве случаев поддерживался на уровне, близком к номинальному, хотя в одном из сценариев на данных Criteo было зафиксировано небольшое превышение.

В целом, результаты на реальных данных свидетельствуют о практической ценности и перспективности предложенного подхода ST-MLCA для повышения эффективности онлайн-экспериментов.

В Главе 4 была проведена экспериментальная оценка эффективности

метода ST-MLCA. На синтетических данных исследовались контроль ошибки первого рода, сокращение среднего числа наблюдений (ASN), статистическая мощность, а также влияние качества и типа ML-моделей. Метод также был применен к двум реальным наборам данных (Hillstrom E-Mail Marketing и Criteo Uplift Prediction) для демонстрации его практической применимости. Результаты экспериментов были проанализированы и сопоставлены с выдвинутыми исследовательскими гипотезами

ЗАКЛЮЧЕНИЕ

В рамках данной дипломной работы был разработан и исследован метод ST-MLCA (Sequential Test with ML-based Covariate Adjustment), предназначенный для повышения эффективности последовательного тестирования статистических гипотез за счет использования информации о ковариатах с помощью моделей машинного обучения. Основная идея метода заключается в коррекции целевой метрики путем вычитания ее предсказанного значения, полученного от ML-модели, и последующем применении стандартных процедур последовательного анализа к полученным остаткам. Такой подход направлен на снижение дисперсии анализируемых данных, что теоретически ведет к сокращению среднего числа наблюдений (ASN) при сохранении статистических гарантий.

Проведенное экспериментальное исследование на синтетических и реальных данных позволило подтвердить ключевые преимущества предложенного метода. На синтетических данных было установлено:

- Метод ST-MLCA статистически значимо сокращает ASN по сравнению с последовательными тестами без учета ковариат, причем выигрыш растет с увеличением предсказательной силы ML-модели (R^2).
- В сценариях с нелинейными зависимостями использование гибких нелинейных ML-моделей (например, CatBoost) приводит к лучшим результатам по сравнению с линейной регрессией.
- Метод корректно контролирует ошибку первого рода на заданном уровне значимости α.

Эксперименты на реальных наборах данных (Hillstrom E-Mail Marketing, Criteo Uplift Prediction) продемонстрировали практическую применимость ST-MLCA, показав сокращение ASN на 21-31% по сравнению с базовым подходом

без ковариатной коррекции, при сохранении контроля ошибки первого рода на приемлемом уровне. Было отмечено, что эффективность метода и выбор оптимальной ML-модели могут зависеть от специфики данных.

Таким образом, дипломная работа демонстрирует, что интеграция машинного обучения в процедуры последовательного анализа для учета ковариат является перспективным направлением. Метод ST-MLCA представляет собой практически реализуемый инструмент, способный ускорить процесс принятия решений в онлайн-экспериментах, таких как A/B-тестирование, за счет более эффективного использования доступной информации.

Ключевым условием эффективности ST-MLCA является наличие информативных ковариат и возможность построения качественной предиктивной модели на независимых данных. Однократное обучение модели перед началом последовательного теста снижает вычислительные затраты по сравнению с методами, требующими переобучения на каждом шаге.

Перспективными направлениями для дальнейших исследований являются разработка адаптивных стратегий для обновления предиктивной модели, исследование робастности метода к различным условиям и его применение к более сложным типам данных и экспериментальным дизайнам.

В целом, результаты работы подтверждают целесообразность и эффективность использования машинного обучения для ковариатной коррекции в последовательном тестировании гипотез, открывая возможности для дальнейшего совершенствования методов статистического анализа в условиях больших данных.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Wald, A. (1947). *Sequential Analysis*. John Wiley Sons.

2. Deng, A., Xu, Y., Kohavi, R., Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 123-132).

3. Howard, S. R., Ramdas, A., McAuliffe, J., Sekhon, J. (2021). Anytime-valid inference: A general framework for safely testing hypotheses and estimating parameters on the fly. *Statistical Science*, 36(2), 275-292.

4. Shafer, G. (2021). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2), 507-534.

5. Grünwald, P., de Heide, R., Koolen, W. M. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*.

6. Lopez-Paz, D., Oquab, M. (2017). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

7. Howard, S. R., Ramdas, A. (2020). Classifier two-sample tests for sequential data. *arXiv preprint arXiv:2010.13122*.

8. Orabona, F., Jun, K. S. (2019). A modern theory of online learning. *Foundations and Trends® in Machine Learning*, 12(4), 297-466.

9. Ramdas, A., Ruf, B., Waudby-Smith, M. J., Larroy, P. (2020). Testing by betting: A new framework for statistical hypothesis testing. *arXiv preprint arXiv:2001.00594*.

10. Zhang, X., Chen, S., Li, D., Liu, Q. (2020). Deep reinforcement learning for optimal stopping in A/B testing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (pp. 1812-1820).

11. Hillstrom, K. (2008). *MineThatData E-Mail Analytics And Data Mining Challenge dataset.* `http://www.minethatdata.com/Kevin_Hillstrom_MineThatData_E-MailAnalytics_DataMiningChallenge_2008.03.20.csv`

12. Criteo Labs. (2018). *Criteo AI Lab Uplift Prediction Dataset.* `https://ailab.criteo.com/criteo-uplift-prediction-dataset/`