УДК 004.8

Николаев Иван Кириллович, аспирант, институт передовых информационных технологий, Тульский государственный педагогический университет им. Л. Н. Толстого, РФ, г. Тула

Привалов Александр Николаевич, профессор, директор института, институт информатики и информационных технологий, доктор технических наук, Тульский государственный педагогический университет им. Л. Н. Толстого, РФ, г. Тула

ОБЗОР МЕТОДОВ ОБЪЯСНЕНИЯ РЕШЕНИЙ ML-МОДЕЛЕЙ В ЗАДАЧАХ АУДИТА: ИНТЕРПРЕТИРУЕМЫЕ МОДЕЛИ, LIME/SHAP И КОНТРФАКТЫ, НА 2026 ГОД

Аннотация

В статье представлен обзор методов объяснимого машинного обучения (XAI) и контрфактуальных объяснений в контексте задач IT-аудита. Рассматривается проблема обеспечения прозрачности алгоритмических решений при анализе информационных систем, выявлении аномалий и оценке рисков. Проанализированы основные подходы: интерпретируемые модели, post-hoc методы (LIME, SHAP), контрфактуальные и гибридные методы. Показаны их преимущества и ограничения, включая проблемы стабильности, вычислительной сложности и реалистичности объяснений. Особое внимание уделено метрикам качества объяснений и ограничениям, связанным с приватностью данных и регуляторными требованиями. Сделан вывод о значимости XAI для повышения доверия к автоматизированным системам и эффективности IT-аудита.

Annotation

This article provides an overview of explainable machine learning (XAI)

methods and counterfactual explanations in the context of IT audit tasks. It examines the challenge of ensuring the transparency of algorithmic decisions when analysing information systems, detecting anomalies and assessing risks. The main approaches are analysed: interpretable models, post-hoc methods (LIME, SHAP), counterfactual and hybrid methods. Their advantages and limitations are highlighted, including issues of stability, computational complexity and the realism of explanations. Particular attention is paid to metrics for the quality of explanations and constraints related to data privacy and regulatory requirements. The conclusion is drawn regarding the significance of XAI for increasing trust in automated systems and the effectiveness of IT audits.

Ключевые слова: объяснимое машинное обучение, контрфактуальные объяснения, IT-аудит, интерпретируемые модели, LIME, SHAP, прозрачность алгоритмов, метрики качества объяснений

Keywords: explainable machine learning, counterfactual explanations, IT audit, interpretable models, LIME, SHAP, algorithmic transparency, explanation quality metrics

Введение. В последние годы методы объяснимого машинного обучения (XAI) стали неотъемлемой частью разработки и применения искусственного интеллекта в различных сферах, включая информационные технологии. Одной из важнейших задач в этой области является предоставление прозрачности алгоритмическим решениям, что особенно важно для областей, где результаты автоматических систем могут иметь серьёзные последствия, такие как IT-аудит. Аудит информационных систем и процессов требует не только детекции аномалий и рисков, но и способности объяснить, почему система приняла то или иное решение. В связи с этим важным направлением стало использование контрфактуальных объяснений (CE), которые позволяют не только прояснить процесс принятия решений, но и показать, какие минимальные изменения данных могут привести к иному результату.

В последние годы активно развиваются различные методы XAI, включая интерпретируемые модели (такие как деревья решений и линейные модели), пост-hoc методы (например, LIME и SHAP) и контрфактуальные объяснения, каждый из которых имеет свои преимущества и ограничения. Однако не все из этих методов подходят для применения в сфере IT-аудита, где важно учитывать как технические ограничения, так и специфические требования безопасности и конфиденциальности.

Целью данной статьи является обзор существующих методов объяснимого машинного обучения и контрфактуальных объяснений, с акцентом на их применимость в контексте IT-аудита. Мы рассматриваем не только теоретические подходы, но и практическую применимость методов в задачах анализа и проверки информационных систем, включая выявление аномалий, оценку рисков и оценку соответствия регуляторным требованиям.

Статья структурирована следующим образом: вначале мы представим методологию обзора и классификацию методов XAI и контрфактуальных объяснений. Далее мы детально рассмотрим основные методы, такие как интерпретируемые модели, LIME, SHAP и контрфактуальные объяснения, а также особенности их применения в IT-аудите. Мы также обсудим метрики и критерии качества объяснений, которые необходимы для оценки их применимости в сфере аудита. В заключение, мы сформулируем практические рекомендации для исследователей и практиков, а также обозначим открытые вопросы и направления для дальнейших исследований.

Классификация методов объяснения. В рамках анализа решений, принятых машинными моделями, существует несколько подходов, которые обеспечивают интерпретируемость их решений. Эти методы можно разделить на несколько категорий в зависимости от их принципа работы, области применения и возможностей для адаптации к различным типам данных и задач. Важнейшими из этих категорий являются интерпретируемые модели, пост-hoc методы, контрфактуальные объяснения и гибридные подходы. Каждый из этих методов имеет свои особенности, сильные и слабые стороны,

что делает их применимыми для разных сценариев, в том числе в области IT-аудита.

Интерпретируемые модели, такие как деревья решений, линейные модели и правила классификации, представляют собой методы, чья внутренняя структура понятна человеку. Эти модели обладают высокой степенью интерпретируемости, так как решения, принимаемые моделью, могут быть объяснены с помощью простых и прозрачных правил. Например, дерево решений отображает последовательность вопросов, каждый из которых приводит к финальному результату, и это дерево легко визуализируется. Линейные модели, такие как линейная и логистическая регрессия, также являются интерпретируемыми, так как их коэффициенты можно интерпретировать как влияние каждого признака на результат. Модели на основе правил классификации, такие как логические правила, могут быть полезны в задачах аудита, где важно понимать, по каким именно критериям принято решение. Однако, несмотря на свою простоту и прозрачность, такие модели ограничены в своей выразительности. Например, деревья решений могут быть слишком простыми и не способны захватывать сложные зависимости, а линейные модели ограничены только линейными взаимосвязями между признаками, что делает их менее эффективными при работе с более сложными данными.

Пост-hoc методы объяснения, такие как LIME, SHAP и методы оценки важности признаков, применяются к моделям, которые сами по себе не являются интерпретируемыми, например, к нейросетям и ансамблевым методам. Эти методы направлены на интерпретацию решений сложных моделей после их обучения, путем аппроксимации модели или анализа важности признаков. LIME, например, создаёт локальные линейные модели, которые аппроксимируют поведение сложной модели на конкретном примере. Это позволяет дать локальные объяснения решениям модели, что важно для задач, связанных с аудиторским анализом. SHAP использует теорию Шепли для вычисления вклада каждого признака в решение модели, обеспечивая как

локальные, так и глобальные объяснения. Методы оценки важности признаков, такие как оценка вклада каждого признака в предсказания модели, помогают выявить ключевые факторы, влияющие на принятие решения. Однако эти методы также имеют свои ограничения: LIME может быть нестабильным, давая разные объяснения для схожих данных, а методы, требующие дополнительного вычислительного времени, могут быть неэффективными при работе с большими наборами данных.

Контрфактуальные объяснения (CE) предоставляют более глубокое понимание того, какие минимальные изменения в данных могут привести к иному результату. Этот подход особенно полезен для объяснения решений, когда необходимо понять, что могло бы быть сделано иначе для получения другого результата. Алгоритмические подходы для генерации контрфактов включают поиск в пространстве изменений с минимизацией расстояния между текущими данными и изменёнными примерами, а также поиск с ограничениями, например, минимальные изменения только в допустимых признаках. Один из подходов, min-change, стремится найти минимальные изменения в данных, например, минимизируя L2-норму между исходным и контрфактуальным примерами. Контрфакты должны быть не только plausibles (реалистичными), но и actionable (осуществимыми), что особенно важно для IT-аудита, где изменения данных должны быть практически применимыми. Генерация контрфактов, однако, может быть вычислительно дорогой, особенно для сложных моделей, и также существует проблема с реалистичностью изменений, так как не все данные могут быть изменены в реальном мире.

Гибридные подходы комбинируют символические методы, такие как онтологии, и статистические методы машинного обучения. Например, использование онтологий для улучшения качества контрфактов позволяет повысить их plausibility, так как онтология может указать, какие изменения допустимы в рамках бизнес-правил или технических ограничений. Символические методы, такие как логика и правила, могут быть использованы

в сочетании с машинным обучением для создания более объяснимых решений. Такой подход может быть полезен в контексте аудита, где важно не только понять, как приняты решения, но и как эти решения соотносятся с установленными правилами и требованиями. Несмотря на свою эффективность, гибридные подходы требуют разработки обогащённых онтологий для специфических задач, что может быть трудоёмким процессом.

Метрики и критерии качества объяснений. Для оценки качества объяснений в XAI и контрфактуальных объяснений используются несколько метрик, которые помогают определить их эффективность и применимость. Одной из важных метрик является fidelity, которая оценивает, насколько хорошо объяснение соответствует поведению модели. Stability отражает, насколько устойчиво объяснение при изменении входных данных или модели. Compactness оценивает, насколько лаконичным является объяснение — чем проще и понятнее объяснение, тем лучше. Plausibility оценивает, насколько реалистичными являются изменения, предложенные в контрфактах. Наконец, actionability важна для определения, насколько предложенные изменения можно реально применить в рабочем процессе. Эти метрики являются критически важными для IT-аудита, где необходимо не только получить объяснение, но и убедиться, что оно полезно, понятно и применимо в контексте реальных данных и задач.

Особенности данных и задач IT-аудита. Для эффективного использования методов объяснимого машинного обучения (XAI) в контексте IT-аудита необходимо учитывать специфические особенности данных, с которыми работают аудиторы, а также учитывать ограничения, которые накладываются как на сами модели, так и на требования регуляторов. В IT-аудите основными типами данных являются логи, реестры и обращения. Логи содержат информацию о действиях пользователей, событиях и ошибках, реестры — данные о системных параметрах и конфигурации, а обращения — запросы пользователей к системе. Все эти данные являются высокоразмерными и часто имеют пропуски или шум, что создаёт

дополнительные сложности при применении традиционных машинных моделей. Важно отметить, что в задачах IT-аудита существует множество ограничений, таких как невозможность изменения части признаков (например, идентификационных данных пользователей), что усложняет задачу генерации объяснений.

Одним из ключевых аспектов работы с такими данными является соблюдение регуляторных требований, включая законы о защите данных и требования к безопасности, такие как GDPR. Эти нормы обязывают организации обеспечить максимальную прозрачность в использовании алгоритмов и моделей, а также минимизировать использование данных, которые могут идентифицировать личность пользователя. Приватность данных, таким образом, становится важным аспектом при выборе методов объяснения, особенно если речь идёт о контрфактуальных объяснениях, которые могут требовать изменения данных в их исходном виде. В этих условиях методы XAI, такие как интерпретируемые модели и пост-hoc методы, являются более предпочтительными, так как они позволяют объяснить результат без необходимости глубокого вмешательства в данные.

Применение XAI в IT-аудите подтверждается рядом литературных примеров. В одной из работ исследовалась прозрачность решений в системах мониторинга безопасности с использованием методов LIME для локального объяснения решений модели, основанной на нейросетях. Полученные результаты показали, что использование LIME позволило улучшить понимание решений модели и повысить доверие аудиторов к автоматизированным системам мониторинга. В другом примере, использующем SHAP, были проанализированы модели для обнаружения аномалий в сетевом трафике. Исследование показало, что SHAP способен детально объяснять, какие именно признаки наиболее важны для принятия решения о наличии аномалии, что способствует повышению точности в анализе сетевых инцидентов. Однако в обоих случаях возникла проблема нестабильности объяснений при работе с большими наборами данных, что

требует дополнительной настройки и оптимизации методов.

Контрфактуальные объяснения (CE) играют важную роль в IT-аудите, так как они помогают понять, какие минимальные изменения в данных могут изменить решение модели. Однако, несмотря на свою полезность, контрфакты сталкиваются с рядом проблем. Одна из них — реалистичность контрфактов. Не все изменения данных могут быть выполнимыми в реальном мире, особенно когда речь идет о данных, связанных с идентификацией пользователей или системных конфигураций. Генерация таких изменений требует от моделей способности учитывать ограничения данных и регуляторные требования, чтобы предложенные изменения были не только теоретически возможными, но и практически применимыми. Кроме того, важно учитывать валидацию объяснений экспертами, поскольку контрфакты могут быть интерпретированы по-разному, и в случае IT-аудита необходимо обеспечить, чтобы эти объяснения были действительно полезными для операторов и аудиторов. Проблема с валидацией объяснений также заключается в сложности оценки их реальной полезности в процессе аудита, так как многие из этих методов требуют проверки в реальных условиях работы.

Таблица 1.

Сравнительный анализ методов объясняемого машинного обучения (XAI) для применения в IT-аудите

Метод объяснения	Интерпретируемость	Применимость в IT-аудите	Прозрачность решений	Проблемы с реалистичностью контрфактов	Совместимость с регуляторными требованиями	Вычислительная сложность	Адаптивность к изменениям данных

Интерпретируемые модели (деревья, линейные модели)	Высокая	Высокая	Высокая	Не применимо	Легко соблюдается, данные не изменяются	Низкая	Ограниченная, эффективны только для линейных зависимостей
Post-hoc методы (LIME, SHAP)	Средняя (поясняется через аппроксимацию)	Средняя (для сложных моделей, например, нейросетей)	Средняя (локальная интерпретация)	Проблемы с устойчивостью и стабильностью	Требует минимизации изменений для защиты приватности	Средняя	Зависит от модели, может быть нестабильно при больших данных
Контрфактуальные объяснения (CE)	Средняя (требуют изменений в данных)	Высокая (позволяют оценить, какие изменения влияют на решение)	Высокая (объяснение "что если")	Высокие требования к реалистичности изменений	Может нарушать приватность данных, при изменении данных	Высокая (необходимы дополнительные вычисления)	Влияют на исходные данные, что ограничивает возможные изменения
Гибридные подходы (символика + ML, онтологии + CE)	Средняя (комбинированная интерпретируемость)	Высокая (повышают plausibility контрфактов)	Высокая (пояснение с учётом бизнес-логики)	Проблемы с созданием обогащённых онтологий	Требуют соблюдения дополнительных ограничений	Высокая	Высокая (учёт специфических ограничений через онтологии)

Тенденции на 2026 год

В рамках данного обзора были рассмотрены ключевые методы объяснимого машинного обучения (XAI) и контрфактуальных объяснений, которые могут быть применимы в сфере IT-аудита. Подробно проанализированы такие методы, как интерпретируемые модели, включая деревья решений и линейные модели, пост-hoc методы (LIME, SHAP, методы оценки важности признаков), а также контрфактуальные объяснения и гибридные подходы. Было подчеркнуто, что каждый из этих методов имеет свои особенности, преимущества и ограничения в контексте задач аудита, где прозрачность и интерпретируемость решений особенно важны для понимания

работы автоматизированных систем и обеспечения их доверия со стороны аудиторов.

Особое внимание было уделено контрфактуальным объяснениям, как важному инструменту для понимания, какие минимальные изменения в данных могут изменить результат принятого решения. Важно отметить, что, несмотря на значительный потенциал этих методов, они сталкиваются с проблемами, такими как трудности в генерации реалистичных контрфактов и сложность валидации объяснений экспертами. Кроме того, проблема приватности и соблюдения регуляторных требований, таких как защита персональных данных, остаётся важным ограничением при применении методов XAI в реальных условиях IT-аудита.

Применение методов XAI в аудите информационных систем позволяет не только повысить доверие к результатам автоматических решений, но и сделать процесс аудита более эффективным, обеспечив более точное выявление аномалий и рисков.

Заключение. В статье представлен обзор современных методов объяснимого машинного обучения и контрфактуальных объяснений, а также их применения в сфере IT-аудита. Рассмотрены как теоретические основы методов XAI, так и их практическая применимость в задачах анализа и проверки информационных систем. Методы, такие как интерпретируемые модели, LIME, SHAP и контрфактуальные объяснения, показали свою высокую эффективность в задачах, связанных с повышением прозрачности автоматизированных решений. Однако, несмотря на положительные результаты, остаются нерешённые вопросы, такие как реалистичность контрфактов, стабильность объяснений и их валидация экспертами, что требует дальнейших исследований в этой области.

Для успешного внедрения XAI в IT-аудит необходимо учитывать особенности данных, с которыми работают аудиторы, а также соблюдать требования конфиденциальности и безопасности. В дальнейшем потребуется разработка новых методов и подходов, которые смогут обеспечить большую

стабильность объяснений и улучшить практическую применимость контрфактуальных объяснений в реальных рабочих процессах.

Таким образом, данное направление представляет собой перспективную область для дальнейших исследований и разработки новых решений в области IT-аудита с использованием объяснимого машинного обучения и контрфактуальных объяснений.

Список литературы

1. Удай Камат, Джон Лю. Объяснимый искусственный интеллект: введение в интерпретируемое машинное обучение // Springer International Publishing. – 15.12.2021. С. 310.
2. Марко Тулио Рибейро, Самер Сингх, Карлос Густерен. «Почему я должен вам доверять?» Объяснение прогнозов любого классификатора – 09.08.2016. С. 10. – URL: https://arxiv.org/pdf/1602.04938 (дата обращения: 03.09.2025).
3. Скотт Лундберг и Су-Ин Ли. Единый подход к интерпретации прогнозов моделей – 09.12.2017. С. 10. – URL: https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions (дата обращения: 03.09.2025).
4. Сандра Вахтер, Брент Миттельштадт, Крис Рассел. Контрфактуальные объяснения без вскрытия «черного ящика»: автоматизированные решения и GDPR // Harvard Journal of Law & Technology. 2018. том 31, выпуск 2. С. 841–887.
5. Наута М., Триенес Дж., Патак С. и др. От эмпирических данных к количественным методам оценки: систематический. Обзор по оценке объяснимого ИИ // ACM Computing Surveys. 2023. № 51. С. 1–42.
6. Контрфактуальные объяснения и алгоритмические средства для машинного обучения: обзор // ACM Computing Surveys. – 15.11.2022. С. 23.

References

1. Uday Kamat, John Liu. Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning // Springer International Publishing. – 15 December 2021. P. 310.
2. Marco Tulio Ribeiro, Samer Singh, Carlos Gusteren. ‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier – 9 August 2016. p. 10. – URL: https://arxiv.org/pdf/1602.04938 (accessed: 3 September 2025).
3. Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions – 9 December 2017. p. 10. – URL: https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions (accessed: 3 September 2025).
4. Sandra Wachter, Brent Mittelstadt, Chris Russell. Counterfactual explanations without ‘black-box’ disclosure: automated decisions and the GDPR // Harvard Journal of Law & Technology. 2018. Vol. 31, No. 2. pp. 841–887.
5. Nauta M., Trienes J., Patak S. et al. From empirical data to quantitative evaluation methods: a systematic review on the evaluation of explainable AI // ACM Computing Surveys. 2023. No. 51. pp. 1–42.
6. Counterfactual explanations and algorithmic tools for machine learning: a review // ACM Computing Surveys. – 15 November 2022. p. 23.