

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

СРАВНЕНИЕ МЕТОДОВ ОБЪЯСНЕНИЯ РЕШЕНИЙ ML-МОДЕЛЕЙ В ЗАДАЧАХ АУДИТА: ИНТЕРПРЕТИРУЕМЫЕ МОДЕЛИ, LIME/SHAP И КОНТРАФАКТЫ

Аннотация

В статье рассматриваются методы объяснимого машинного обучения (ХАИ) в контексте IT-аудита. Обосновывается необходимость прозрачности и интерпретируемости решений моделей при анализе информационных систем, выявлении аномалий и оценке рисков. Проводится сравнение основных классов методов ХАИ: интерпретируемых моделей, post-hoc подходов (LIME, SHAP) и контрафактуальных объяснений. Показано, что каждый из подходов имеет свои преимущества и ограничения, что делает актуальным применение гибридных решений. Особое внимание уделено вопросам конфиденциальности данных, регуляторным требованиям и практической применимости объяснений в аудиторской практике.

Annotation

This article examines methods of explainable artificial intelligence (XAI) in the context of IT auditing. It argues for the need for transparency and interpretability

in model decisions when analysing information systems, detecting anomalies and assessing risks. A comparison is made of the main classes of XAI methods: interpretable models, post-hoc approaches (LIME, SHAP) and counterfactual explanations. It is shown that each approach has its own advantages and limitations, making the use of hybrid solutions particularly relevant. Particular attention is paid to data privacy issues, regulatory requirements and the practical applicability of explanations in auditing practice.

Ключевые слова: объяснимое машинное обучение, XAI, IT-аудит, интерпретируемость моделей, LIME, SHAP, контрфактуальные объяснения, гибридные методы, прозрачность алгоритмов, анализ рисков

Keywords: explainable machine learning, XAI, IT audit, model interpretability, LIME, SHAP, counterfactual explanations, hybrid methods, algorithm transparency, risk analysis

Введение. Современные методы машинного обучения (ML) находят широкое применение в различных сферах, включая информационные технологии, где они используются для решения задач, связанных с автоматизацией, прогнозированием, классификацией и обнаружением аномалий. Однако несмотря на высокую эффективность моделей, применяемых в этих областях, остаётся важная проблема: как предоставить объяснение решений, принимаемых этими моделями, особенно в контексте критически важных сфер, таких как IT-аудит. Прозрачность и интерпретируемость моделей имеют ключевое значение для доверия пользователей к автоматическим системам, особенно когда результаты их работы могут повлиять на безопасность данных или операционные процессы организации.

IT-аудит представляет собой комплексный процесс анализа и оценки эффективности информационных систем с целью выявления рисков и обеспечения соответствия нормативным требованиям. В условиях, когда

системы становятся всё более сложными, а данные обрабатываются с использованием «чёрных ящиков» машинного обучения, важно обеспечить возможность для аудиторов понять, каким образом были приняты те или иные решения, на основании каких факторов и в какой степени модель может быть доверена. Это открывает потребность в применении методов объяснимого машинного обучения (ХАИ), которые могут обеспечить прозрачность процессов принятия решений.

Целью данной статьи является систематизация существующих методов ХАИ с акцентом на их применимость для задач IT-аудита. В статье рассматриваются ключевые подходы к объяснению решений ML-моделей, такие как интерпретируемые модели (встроенные модели), post-hoc методы (например, LIME и SHAP) и контрфактуальные объяснения. Статья направлена на то, чтобы показать принципиальные различия между этими методами, выявить их сильные и слабые стороны, а также обосновать, почему ни один из них не является универсальным решением для всех задач аудита. Дополнительно, рассматривается необходимость комбинированных и прикладных подходов для решения задач, возникающих в ходе IT-аудита.

Задачи исследования включают:

- Обзор существующих методов ХАИ и их классификация по типам и возможностям.
- Сравнение методов интерпретируемых моделей, LIME/SHAP и контрфактуальных объяснений с учётом их применимости к задачам IT-аудита.
- Анализ ограничений и преимуществ каждого метода в контексте специфических требований аудиторских задач, таких как безопасность, приватность данных и соблюдение регуляторных норм.
- Формулировка выводов о необходимости использования гибридных методов и предложений для дальнейших исследований в этой области.

Для оценки применения различных методов в IT-аудите в рамках статьи рассматриваются как теоретические подходы, так и практическая применимость в решении реальных задач, таких как анализ и проверка

безопасности данных, а также соответствие систем нормативным требованиям.

Классы методов ХАИ и их принципиальные различия. В данной части систематически рассматриваются ключевые классы методов объяснимого машинного обучения (ХАИ) и выделяются их принципиальные отличия с акцентом на применимость в задачах IT-аудита. Анализ проводится через набор прикладных критериев, релевантных аудиторской практике: прозрачность (*interpretability*), устойчивость/стабильность объяснений (*stability*), реализуемость и вычислительные требования (*implementability*), соответствие регуляторным требованиям и приватности (*compliance & privacy*), а также практическая полезность/дееспособность (*actionability*). На основе этого анализа обосновывается выбор рассматриваемых методов (интерпретируемые встроенные модели, *post-hoc* методы LIME/SHAP, контрфактуальные объяснения) и даётся оценка их сильных и слабых сторон.

Критерии сравнения

● *Прозрачность (interpretability)*. Насколько прямо и однозначно объяснение передаёт логику принятия решения.

● *Локальность/глобальность объяснения*. Дает ли метод глобальную картину модели или объясняет отдельный прогноз.

● *Стабильность (stability)*. Насколько воспроизводимы и устойчивы объяснения при небольших изменениях входных данных или параметров метода.

● *Реализуемость и вычислительная сложность*. Ресурсоёмкость метода, потребность в дополнительной генерации примеров или сложной оптимизации.

● *Соответствие требованиям регуляторов и приватности*. Возможность объяснять решения без раскрытия персональных данных и соблюдение правовых ограничений.

● *Actionability / практическая применимость*. Насколько объяснение даёт аудитору конкретные и выполнимые указания (что изменить, где искать

проблему).

Общая таксономия классов методов

Методы ХАИ удобно группировать по двум независимым осям:

а) интерпретируемые встроенные модели (intrinsic/white-box) vs post-hoc объяснения применяемые к «чёрным ящикам»;

б) атрибутивные (feature-attribution, вклад признаков) vs примерные (example-based, в том числе контрфакты и прототипы). На пересечении этих осей оказываются рассматриваемые в статье группы: встроенные модели (деревья, линейные модели, правила) — интерпретируемые и атрибутивные/структурные; LIME/SHAP — post-hoc атрибутивные (локально/глобально); контрфактуальные объяснения — post-hoc примерные (example-based, «что-если»).

Интерпретируемые (встроенные) модели

Описание и примеры. Классы моделей, чья внутренняя структура читаема для человека: деревья решений, линейные и логистические модели, модели на правилах.

Характерные свойства. Объяснение вытекает из самой модели: правило, путь в дереве, коэффициенты линейной комбинации. Объяснения, как правило, глобальны (модельная логика) и/или локальны (путь до листа).

Преимущества для аудита:

- Высокая прозрачность и простота верификации; легко показать логику аудитору и регулятору;
- Низкие вычислительные требования; воспроизводимость объяснений;
- Хорошая совместимость с нормативными требованиями: мало скрытых процедур, можно формализовать соответствие правилам.

Ограничения:

- Ограниченная выразительная способность при сложных нелинейных зависимостях; снижение качества предсказаний на сложных задачах;
- Склонность к переобучению в глубоких деревьях или низкой обобщающей способности у простых моделей.

По критерию прозрачности (interpretability) эти модели оцениваются как высокопрозрачные: логика принятия решения представляется в виде явных правил, коэффициентов или путей в дереве, что обеспечивает однозначную интерпретацию.

По критерию локальности/глобальности интерпретируемые модели способны предоставлять как глобальную картину (структура модели, влияние признаков в общем), так и локальные объяснения (путь до листа дерева описывает конкретный прогноз).

По критерию стабильности (stability) такие модели демонстрируют, как правило, высокую воспроизводимость объяснений при незначительных вариациях входных данных, хотя в отдельных архитектурах (глубокие деревья, сложные наборы правил) возможна чувствительность к шуму и к изменениям выборки.

По критерию реализуемости и вычислительной сложности интерпретируемые модели характеризуются низкой ресурсоёмкостью при обучении и объяснении, что делает их удобными для практического развёртывания.

По критерию соответствия регуляторным требованиям и приватности встроенные модели имеют преимущество: отсутствие скрытых слоёв и прозрачная логика упрощают аудит и проверку на соответствие нормативам.

По критерию actionability / практическая применимость интерпретируемые модели обеспечивают средне-высокую практическую ценность: правила и коэффициенты легко переводятся в конкретные рекомендации, но в случае упрощённых моделей такие рекомендации могут оказаться недостаточно точными для сложных сценариев.

В сумме интерпретируемые модели представляют собой предпочтительную отправную точку в задачах, где требуется формальная отчётность и высокая степень доверия к логике модели, при условии принятия компромисса по точности в задачах высокой сложности

Post-hoc методы атрибуции: LIME и SHAP (локальные и глобальные

объяснения)

Описание и примеры. Методы, применяемые к обученным «чёрным ящикам»: LIME строит локальную аппроксимацию (обычно линейную) вокруг интересующего примера; SHAP вычисляет вклад признаков на основе величин Шепли, обеспечивая аддитивную разложимость.

Характерные свойства. Независимость от внутренней структуры модели (model-agnostic в большинстве реализаций); дают как локальные, так и агрегируемые глобальные объяснения (через суммирование/средние SHAP-значения).

Преимущества для аудита:

- Позволяют сохранять качество «чёрного ящика» при предоставлении объяснений; применимы к сложным моделям, востребованным для обнаружения аномалий;

- SHAP имеет сильную теоретическую основу (Shapley values), дающую интерпретационно привлекательную аппроксимацию вклада признаков.

Ограничения:

- Локальная аппроксимация (LIME) и численные оценки (SHAP) могут быть нестабильны, небольшие изменения в данных или в параметрах метода дают разный набор значимых признаков;

- Высокая вычислительная стоимость при больших моделях и больших объёмах данных (особенно для точных SHAP-оценок);

- Вопрос доверия. Объяснение аппроксимацией не всегда отражает истинную внутреннюю логику модели.

По прозрачности такие методы оцениваются как средние: объяснения получаются в виде оценок вклада признаков или локальных аппроксимаций, однако интерпретация этих величин требует понимания механики аппроксимации и агрегирования.

По локальности/глобальности, LIME обеспечивает локальные объяснения одиночного прогноза, а SHAP предоставляет как локальные, так и агрегируемые глобальные представления (через суммирование и средние

SHAP-значений).

По стабильности post-hoc методы показывают относительно низкую или среднюю устойчивость: LIME особенно чувствителен к выбору параметров и к шуму в данных, SHAP менее чувствителен, но точные оценки могут изменяться при приближениях и подвыборках.

По реализуемости и вычислительной сложности обе методики требуют дополнительных вычислений (генерация близких примеров, многократные прогонки модели), что повышает вычислительные затраты пропорционально размеру модели и объёму данных; для точного SHAP-вычисления затраты особенно высоки.

По соответствию регуляторным требованиям и приватности post-hoc методы имеют средний статус: они не раскрывают структуру модели, но могут оперировать персональными признаками в объяснениях, что требует осторожности и фильтрации.

По actionability такие методы дают ценную информацию о важности признаков, однако не всегда предлагают реалистичные или осуществимые изменения, следовательно практическая применимость оценивается как средняя.

Следует подчеркнуть, что post-hoc методы целесообразны при необходимости сохранить высокую предиктивную мощность модели, при этом требуя дополнительных процедур валидации и ограничения использования чувствительных признаков.

Контрфактуальные (example-based) объяснения

Описание и примеры. Методы, формирующие альтернативные входные примеры (контрфакты), приводящие модель к иному выводу: «что нужно изменить, чтобы результат был другим». Подходы включают оптимизационные процедуры, перебор в дискретных пространствах, генеративные модели.

Характерные свойства. Сдвиг фокуса: не «почему» в терминах вкладов признаков, а «какие изменения» приведут к желаемому результату.

Контрфакты, по определению, ориентированы на actionability.

Преимущества для аудита:

- Предоставляют аудитору конкретные сценарии корректировки или проверки: что нужно изменить в данных/конфигурации, чтобы модель дала иной вердикт;

- Высокая практическая ценность при расследовании инцидентов и формировании рекомендаций.

Ограничения:

- Plausibility/реалистичность. Сгенерированные контрфакты могут быть неосуществимы в реальном мире (например, изменение идентификаторов, временных меток);

- Ограничения приватности и регуляторики. Генерация и демонстрация контрфактов могут предполагать манипуляции с персональными данными;

- Вычислительная стоимость и необходимость учёта дискретных и связанных ограничений (например, зависимостей между признаками);

- Неоднозначность/множественность решений: существует множество контрфактов; выбор «наиболее релевантного» требует дополнительных критериев.

По прозрачности контрфакты оцениваются как средние: они демонстрируют, какие изменения в признаках приводят к иному выводу, но не обязательно раскрывают внутреннюю логику модели в терминах вклада каждого признака.

По локальности/глобальности контрфактуальные объяснения преимущественно локальны, поскольку формулируются в отношении конкретного прогноза или экземпляра.

По стабильности оценки зависят от алгоритма генерации: многие алгоритмы контрфактов дают нестабильные результаты при вариациях данных и параметров; следовательно стабильность считается варьирующейся от низкой до средней.

По реализуемости и вычислительной сложности контрфакты обычно

предъявляют высокие требования: оптимизационные процедуры, учёт дискретных и семантических ограничений и поиск в большом пространстве приводят к значительным вычислительным затратам.

По соответствию регуляторным требованиям и приватности контрфакты потенциально проблематичны: генерация и демонстрация альтернативных примеров может задействовать персональные данные и потребовать специальных мер анонимизации и фильтрации.

По actionability контрфакты обладают высоким потенциалом — при условии соблюдения plausibility (реалистичности) они предоставляют аудитору конкретные и выполнимые указания — однако практическая ценность напрямую зависит от корректности учёта ограничений предметной области.

В итоге контрфактуальные подходы представляют собой мощный инструмент операционной интерпретации, нуждающийся в строгой фильтрации и дополнительных механизмах валидации для применения в аудите.

Гибридные подходы

Описание и примеры. Сочетание статистических моделей с символической информацией: онтологии домена, бизнес-правила, логические ограничения, интеграция с контрфактами для повышения plausibility.

Характерные свойства. Позволяют формализовать допустимые трансформации и бизнес-знания, повышая правдоподобие объяснений и делая их совместимыми с регламентами.

Преимущества для аудита:

- Повышенная plausibility контрфактов; объяснения соотносятся с бизнес-правилами и юридическими ограничениями;

- Лучшая интерпретация сложных связей (например, связей между конфигурацией и журналами событий).

Ограничения:

- Создание и поддержка онтологий/правил требует экспертного труда и

может быть дорогостоящим;

- Увеличение сложности интеграции и требования к семантической совместимости данных.

.По прозрачности такие решения могут обеспечивать средне-высокий уровень объяснимости: встроенные правила и семантика облегчают восприятие, хотя статистическая часть остаётся частично непрозрачной.

По локальности/глобальности гибриды способны давать как глобальные, так и локальные объяснения в зависимости от архитектуры интеграции.

По стабильности интеграция формальных правил повышает воспроизводимость объяснений относительно чисто статистических post-hoc подходов.

По реализуемости и вычислительной сложности гибридные решения обычно более ресурсоёмки и требуют значительных усилий по созданию/поддержке онтологий и правил, а также по обеспечению семантической совместимости данных.

По соответствию регуляторным требованиям и приватности гибриды имеют преимущество за счёт явной формализации бизнес- и правовых ограничений.

По actionability такие подходы обеспечивают высокую практическую применимость: валидация и фильтрация контрафактов через онтологии повышает вероятность предложений, которые можно реализовать в операционной практике.

Следовательно гибридные методы представляются наиболее перспективными для перехода от теории к практической интеграции ХАИ в процессы ИТ-аудита.

Проблемы применения методов машинного обучения в ИТ-аудите. Методы машинного обучения (ML), успешно применяемые в различных сферах, не всегда могут быть напрямую перенесены на задачи ИТ-аудита. В частности, несмотря на высокую эффективность многих алгоритмов, их применение в условиях аудита, где требуется высокая степень объяснимости и

доверия к результатам, сталкивается с рядом специфических трудностей. Эти трудности обусловлены различиями в требованиях к обработке данных и интерпретации решений в контексте аудита по сравнению с типичными задачами машинного обучения. В данном разделе рассматриваются причины, по которым методы, успешные в ML в целом, могут не соответствовать требованиям IT-аудита, а также особенности, которые должны быть учтены при адаптации этих методов.

Основная причина заключается в том, что данные и задачи IT-аудита существенно отличаются от стандартных наборов данных, используемых в машинном обучении. Аудит включает работу с данными, которые имеют высокую степень чувствительности и требуют соблюдения строгих регуляторных стандартов. Например, данные, такие как логи, реестры и запросы пользователей, содержат информацию, которая может включать персональные данные и конфиденциальные сведения. Это накладывает особые ограничения на использование ML-моделей, так как стандартные методы объяснения, применяемые в ML, не всегда могут гарантировать безопасность данных или соблюдение стандартов конфиденциальности, что является важным аспектом в IT-аудите.

Еще одной ключевой проблемой является необходимость интерпретируемости решений. В отличие от многих задач машинного обучения, где модель может быть использована как чёрный ящик, в IT-аудите требуется чёткое и прозрачное объяснение того, как именно было принято то или иное решение. Это становится особенно важным в условиях, когда решения автоматизированных систем могут повлиять на безопасность или законность действий, проводимых на основе этих решений. Например, методы, такие как глубокие нейросети, которые эффективно решают сложные задачи, могут оказаться непригодными для аудита из-за своей низкой интерпретируемости и трудности в понимании того, какие именно признаки влияли на итоговое решение.

Регуляторные требования и приватность данных также создают

дополнительные ограничения. В области IT-аудита существует строгая необходимость соблюдения законодательных и корпоративных стандартов, таких как GDPR, который регулирует обработку персональных данных. Методы, которые требуют раскрытия внутренних параметров модели или манипуляций с данными, могут быть неприемлемыми с точки зрения конфиденциальности. Например, методы пост-хок атрибуции (LIME, SHAP), хотя и обеспечивают локальные объяснения, могут работать с персональными признаками, что требует дополнительной обработки и защиты данных.

Кроме того, методы, успешные в ML, не всегда учитывают специфику оценки рисков и аномалий в контексте IT-аудита. В аудите особое внимание уделяется выявлению рисков и аномальных ситуаций, где изменения в данных могут быть вызваны различными факторами, включая ошибки конфигурации или кибератаки. Стандартные методы ML ориентированы на оптимизацию предсказаний на основе исторических данных, но не всегда способны оперативно адаптироваться к новым, необычным ситуациям, что делает их менее эффективными для таких задач, как обнаружение и предотвращение угроз.

Наконец, методы ML, предназначенные для высокоточных предсказаний, часто не учитывают аспект практической применимости их выводов в реальных рабочих процессах. В IT-аудите важны не только предсказания, но и конкретные рекомендации относительно того, что нужно изменить в системе или данных для минимизации рисков. Контрфактуальные объяснения, как пример более прикладного подхода, могут быть значительно полезнее, так как они предлагают не просто объяснение, а возможные пути для изменения ситуации.

Таким образом, методы ML, которые демонстрируют высокую эффективность в стандартных приложениях, могут не всегда удовлетворять строгим требованиям, предъявляемым IT-аудитом. Для задач аудита необходимы методы, которые не только обеспечивают высокую точность и стабильность, но и предоставляют понятные, интерпретируемые объяснения,

соблюдают регуляторные требования, гарантируют практическую применимость и легко адаптируются к специфике проверки рисков и аномалий. В связи с этим необходимо развивать комбинированные и гибридные подходы, которые смогут интегрировать различные методы ХАИ для удовлетворения всех требований аудита.

Заключение. В ходе исследования методов объяснимого машинного обучения (ХАИ) в контексте IT-аудита были рассмотрены ключевые подходы, такие как интерпретируемые модели, post-hoc методы (LIME, SHAP) и контрфактуальные объяснения. Каждый из этих методов имеет свои сильные и слабые стороны, что подтверждает необходимость их применения в зависимости от конкретных задач аудита.

Методы, успешные в области машинного обучения, не всегда могут быть напрямую перенесены на задачи IT-аудита. Это связано с рядом специфических требований, таких как высокая степень чувствительности данных, соблюдение регуляторных стандартов, необходимость прозрачности решений и их интерпретируемости. В частности, стандартные методы, такие как глубокие нейросети, демонстрируют высокую предсказательную способность, но из-за своей сложности и низкой интерпретируемости становятся непригодными для задач, связанных с аудиторской проверкой.

Особое внимание следует уделить контрфактуальным объяснениям, которые помогают устранить часть проблем, связанных с пониманием решений моделей. Однако они также имеют ограничения, включая проблемы с реализуемостью изменений в реальных данных и вопросы приватности.

Для успешного применения методов ХАИ в IT-аудите необходимо учитывать не только точность и стабильность моделей, но и их соответствие специфическим требованиям безопасности, конфиденциальности и регуляторным нормам. Необходимость в более гибких и комбинированных подходах подтверждается текущими ограничениями существующих методов, что делает интеграцию различных методов ХАИ перспективным направлением

для дальнейших исследований и практического применения в области IT-аудита.

Таким образом, для эффективного решения задач IT-аудита требуется развитие комбинированных и гибридных методов, которые смогут объединить сильные стороны различных подходов ХАИ и соответствовать требованиям аудиторов и регуляторов.

Список литературы

1. Арриета, А. Б., Диас-Родригес, Н., Дель Сер, Х., и др. Объяснимый искусственный интеллект (ХАИ): концепции, классификации, возможности и вызовы на пути к ответственному ИИ // Information Fusion. 2020. № 58. С. 82–115.
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, Brent 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.