УДК 004.8

СОВРЕМЕННЫЕ МЕТОДЫ ИНТЕРПРЕТАЦИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПОДХОДЫ, ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Кустанова Мейрамгүл Амангелдіқызы
kstnvmm@gmail.com
Студент 3 курса образовательной программы «Информационные системы и технологии»
НАО «ЗКАТУ имени Жангир хана», г. Уральск, Республика Казахстан
Мизамова Гулбаршын Нурлановна
mizamgul@mail.ru
Магистрант 1 курса образовательной программы «Бизнес-информатика»
НАО «Атырауский университет им.Х.Досмухамедова», г.Атырау, Республика Казахстан
Научный руководитель – **Махатова В.Е.**
к.т.н., профессор

Введение. В современных условиях активного развития технологий искусственного интеллекта особое значение приобретает вопрос интерпретации моделей машинного обучения. Несмотря на высокую точность прогнозирования, многие алгоритмы остаются непрозрачными, что затрудняет их практическое применение в ответственных областях, таких как медицина, финансы и правоприменение. Возможность объяснения работы таких моделей играет ключевую роль в повышении доверия к системам на основе искусственного интеллекта.

Современные исследования в этой области направлены на изучение различных подходов к интерпретации решений, позволяющих лучше понять внутренние механизмы работы алгоритмов. Различные методы объяснения применяются для выявления факторов, оказывающих наибольшее влияние на прогнозы модели, что способствует более обоснованному принятию решений. Анализ существующих решений показывает, что одни методы более эффективны при интерпретации сложных нейросетевых структур, тогда как другие лучше подходят для традиционных статистических моделей.

Практическое применение интерпретируемых моделей затрагивает широкий спектр задач, где важна прозрачность алгоритмов. Применение таких методов позволяет не только контролировать процессы машинного обучения, но и адаптировать модели к конкретным условиям использования, что особенно важно в критически значимых областях. Вопрос развития и совершенствования подходов к интерпретации остаётся актуальным, поскольку только прозрачность и объяснимость алгоритмов могут обеспечить их эффективную интеграцию в реальные сферы деятельности.

Изучение данных аспектов позволит лучше понять возможности и ограничения современных методов интерпретации моделей ИИ, а также определить пути их совершенствования и внедрения в реальные системы.

Материалы исследования. Современные системы машинного обучения требуют значительных объемов данных, что делает их сбор, обработку и хранение критически важными этапами в исследовательской и практической деятельности. Оптимизация работы с данными предполагает не только эффективное управление вычислительными ресурсами, но и определение значимости различных типов информации для минимизации избыточных вычислений и увеличения точности моделей. В этом контексте особое значение приобретают методы оценки вклада данных, позволяющие анализировать релевантность различных наборов информации в зависимости от их влияния на предсказательные возможности алгоритмов.

Оценка данных включает несколько ключевых направлений. Во-первых, она позволяет сократить объем хранения информации за счет удаления нерелевантных или

дублирующихся данных, что особенно актуально для анализа больших массивов изображений, видео или медицинских записей. Во-вторых, эффективное использование вычислительных ресурсов возможно только при отборе значимых данных, снижающем затраты на обучение моделей и уменьшающем риск переобучения. В-третьих, данные часто требуют разметки, которая может быть дорогостоящей в таких областях, как диагностика заболеваний или обработка естественного языка. Оптимизация выбора данных для аннотации повышает эффективность этого процесса, минимизируя человеческий фактор.

Методы оценки вклада данных активно применяются в различных областях, таких как финансы, медицина, автономные системы и маркетинг. В финансовом секторе они используются для улучшения алгоритмов оценки рисков и выявления мошенничества, снижая неопределенность при принятии решений. В медицине анализ данных помогает повысить точность диагностических моделей и сократить вероятность ошибок при прогнозировании лечения. Автономные технологии, включая беспилотные автомобили и промышленных роботов, требуют строгого контроля за качеством данных, поскольку некорректные входные параметры могут привести к аварийным ситуациям.

Методы интерпретации моделей искусственного интеллекта представляют собой совокупность подходов, направленных на объяснение решений алгоритмов машинного обучения. Они обеспечивают прозрачность моделей и позволяют понять, какие факторы влияют на конечный результат. В настоящее время исследователи выделяют два основных класса методов: глобальные, которые объясняют поведение модели в целом, и локальные, сосредоточенные на анализе отдельных предсказаний.

Глобальные методы включают в себя подходы, оценивающие значимость входных признаков и выявляющие основные закономерности в данных. Важным направлением является применение методов оценки значимости признаков, таких как Feature Importance, которые позволяют определить, какие входные переменные оказывают наибольшее влияние на предсказания модели. Кроме того, методы визуализации зависимостей, такие как Partial Dependence Plots (PDP), помогают понять, как изменение отдельных факторов сказывается на выходных значениях модели.

Локальные методы интерпретации, напротив, ориентированы на объяснение конкретных решений алгоритма. Наиболее популярными среди них являются методы SHAP (SHapley Additive Explanations) и LIME (Local Interpretable Model-agnostic Explanations). SHAP основан на концепции теории игр и позволяет оценить вклад каждого признака в предсказание модели, обеспечивая стабильные и точные объяснения. LIME использует аппроксимацию сложных моделей с помощью простых линейных моделей, что делает его удобным для интерпретации отдельных прогнозов.

Современные исследования направлены на совершенствование существующих методов интерпретации и разработку более устойчивых подходов. Одним из ключевых вызовов остается чувствительность методов к изменению входных данных, а также их устойчивость к возможным атакам, направленным на манипуляцию результатами интерпретации.

Методы исследования. Глобальные методы интерпретации предоставляют обобщённую картину функционирования модели, позволяя определить ключевые факторы, оказывающие наибольшее влияние на предсказания. К таким методам относятся, например, важность признаков (Feature Importance), частотные зависимости (Partial Dependence Plots) и метод Shapley Additive Explanations (SHAP). Эти подходы используются для анализа структуры модели и выявления закономерностей в данных, что особенно актуально в задачах кредитного скоринга, медицинской диагностики и предсказательного анализа в бизнесе. Применение глобальных методов позволяет специалистам оценивать надёжность модели и выявлять потенциальные риски, связанные с некорректными зависимостями между признаками.

Локальные методы интерпретации, в отличие от глобальных, сосредотачиваются на объяснении отдельных предсказаний модели. Они позволяют детально рассмотреть вклад

каждого входного признака в конкретный прогноз, что делает их особенно ценными в случаях, когда необходимо обосновать принятое решение. Среди наиболее известных локальных методов можно выделить LIME (Local Interpretable Model-Agnostic Explanations), который строит приближенную интерпретируемую модель в окрестности конкретного предсказания, и контрфактические объяснения, анализирующие, какие изменения во входных данных могли бы привести к иному результату. Эти методы активно применяются в медицине для объяснения решений диагностических алгоритмов, а также в юриспруденции и финансовых технологиях.

Сравнительный анализ глобальных и локальных методов интерпретации показывает, что выбор конкретного подхода зависит от специфики задачи и требований к прозрачности решений. В ситуациях, где требуется комплексный анализ всей модели, предпочтение отдаётся глобальным методам, в то время как локальные методы незаменимы при объяснении отдельных решений. Например, в системах автоматического кредитования глобальный анализ позволяет выявить основные факторы, влияющие на одобрение займов, а локальные методы помогают объяснить конкретные решения для отдельных клиентов.

Развитие методов интерпретации направлено на повышение их точности и применимости к различным типам моделей. Одним из перспективных направлений является интеграция глобальных и локальных методов, что позволяет достичь баланса между общей объяснимостью модели и детальной интерпретацией отдельных предсказаний. Кроме того, активное использование методов визуализации, таких как тепловые карты значимости признаков, способствует более интуитивному пониманию работы алгоритмов, что повышает доверие пользователей к искусственному интеллекту и облегчает внедрение интерпретируемых моделей в критически важные сферы деятельности

Сравнение методов интерпретации моделей машинного обучения, таких как SHAP, LIME, Feature Importance и Partial Dependence Plots, позволяет выявить их сильные и слабые стороны, а также определить области наилучшего применения каждого из них. Эти методы служат различным целям: одни обеспечивают глобальное понимание модели, другие сосредотачиваются на объяснении отдельных предсказаний.

Метод SHAP (Shapley Additive Explanations) основан на концепции коэффициентов Шепли из теории игр и предназначен для оценки вклада каждого признака в предсказание модели. Его основное преимущество заключается в способности учитывать нелинейные зависимости между признаками и обеспечивать точное распределение важности входных данных. Однако вычислительная сложность SHAP делает его ресурсоёмким, особенно при работе с высокоразмерными моделями, такими как градиентный бустинг и глубокие нейронные сети.

Метод LIME (Local Interpretable Model-Agnostic Explanations) строит локальные аппроксимации сложных моделей с помощью простых линейных моделей. Это делает его удобным для объяснения конкретных предсказаний, но он не даёт общего понимания структуры модели. Кроме того, результаты LIME могут зависеть от случайного выбора подмножества данных, что может приводить к вариативности интерпретации. Тем не менее, метод остаётся универсальным и применимым к широкому спектру алгоритмов, включая нейросетевые модели, деревья решений и ансамбли.

Feature Importance – это метод, который оценивает значимость признаков на основе их влияния на предсказания модели. В ансамблевых методах, таких как случайный лес и градиентный бустинг, значимость признаков рассчитывается по частоте их использования при построении деревьев решений и их вкладу в снижение ошибки модели. Хотя этот метод позволяет выявить ключевые параметры, он подвержен искажениям в случае высокой корреляции признаков, а также может быть чувствителен к шуму в данных.

Partial Dependence Plots (PDP) используются для визуализации влияния отдельных признаков на предсказания модели. Этот метод полезен для выявления нелинейных зависимостей, однако он предполагает независимость признаков, что не всегда соответствует реальным данным. PDP особенно полезен в случаях, когда необходимо

понять общий эффект определённого признака, но его интерпретация может быть ограничена в сложных моделях с множественными взаимодействиями переменных.

Таким образом, выбор метода интерпретации зависит от специфики задачи. SHAP обеспечивает наиболее точное объяснение модели, но требует значительных вычислительных ресурсов. LIME удобен для локального анализа отдельных предсказаний, но его результаты могут варьироваться. Feature Importance даёт глобальное представление о значимости признаков, но имеет ограничения при наличии коррелированных данных. PDP позволяет визуализировать зависимость предсказаний от отдельных признаков, однако может не учитывать сложные взаимосвязи. Использование этих методов в сочетании позволяет получить более полное понимание работы моделей машинного обучения и повысить уровень их интерпретируемости.

Методы оценки эффективности моделей машинного обучения играют важную роль в интерпретируемости предсказаний. Среди них можно выделить Data Shapley, Influence Functions и Leave-One-Out Analysis (LOO), каждый из которых применяется в различных сценариях анализа данных.

Data Shapley основан на теории кооперативных игр и представляет собой способ оценки вклада каждой точки данных в общую производительность модели. Он вычисляет средний маржинальный вклад каждой точки данных, учитывая все возможные подмножества обучающего набора. Хотя этот метод обладает высокой математической строгостью и справедливым распределением значимости данных, его основной недостаток — высокая вычислительная сложность. Для снижения этой нагрузки применяются приближённые методы, такие как стохастическая аппроксимация.

Метод Influence Functions использует концепции дифференциальной геометрии для анализа вклада отдельных точек данных. Он оценивает влияние данных на предсказания модели через градиентный анализ функции потерь. Основное преимущество данного метода — способность выявлять аномальные данные, обнаруживать выбросы и диагностировать ошибки в обучающих наборах. Однако вычисление матрицы Гессе, необходимое для этого метода, является вычислительно затратным и может затруднять его применение на больших наборах данных.

Leave-One-Out Analysis (LOO) представляет собой простую, но эффективную технику оценки значимости данных. Он заключается в последовательном удалении отдельных точек данных из обучающего набора с последующей оценкой изменений в качестве модели. Если удаление конкретной точки приводит к значительному снижению точности модели, это свидетельствует о её высокой значимости. Этот метод широко применяется в медицине и других областях, где необходимо определить критические данные, влияющие на конечный результат предсказаний.

Сравнение этих методов показывает, что каждый из них обладает уникальными характеристиками и ограничениями. Data Shapley предлагает наиболее теоретически обоснованный подход, но требует значительных вычислительных ресурсов. Influence Functions эффективен в обнаружении выбросов, но также зависит от сложных вычислений. LOO является наименее затратным с точки зрения вычислений, но не всегда способен точно определить вклад данных, особенно в больших наборах. Выбор метода во многом определяется конкретными задачами анализа данных и доступными вычислительными мощностями.

Заключение. Современные методы интерпретации моделей машинного обучения и оценки вклада данных играют ключевую роль в обеспечении прозрачности и доверия к искусственному интеллекту. Глубокий анализ существующих подходов позволяет выявить их сильные и слабые стороны, а также определить ключевые направления для дальнейшего развития.

Методы интерпретации, такие как SHAP, LIME, Attention Mechanisms и Integrated Gradients, демонстрируют высокую эффективность в объяснении работы алгоритмов машинного обучения. Они позволяют анализировать влияние отдельных факторов на

результаты предсказаний модели, выявлять возможные ошибки и повышать прозрачность алгоритмов. Особую значимость приобретают методы оценки вклада данных, включая Data Shapley, Influence Functions и DVRL, которые помогают отбирать наиболее значимые данные, исключать шумовые примеры и повышать надежность прогнозов.

Практическое применение методов интерпретации охватывает широкий спектр областей, включая медицину, финансы и автономные системы. В медицинской сфере интерпретируемость моделей способствует принятию более обоснованных решений на основе предсказаний искусственного интеллекта. В финансовом секторе такие методы позволяют выявлять мошеннические схемы, прогнозировать кредитные риски и повышать точность управленческих решений. В автономных системах анализ вклада данных улучшает безопасность и надежность работы беспилотных технологий и роботизированных устройств.

Несмотря на значительный прогресс, остаются нерешенные проблемы, связанные с устойчивостью методов интерпретации к манипуляциям, автоматизацией оценки качества данных и адаптацией к сложным архитектурам нейросетей. Дальнейшие исследования направлены на разработку более надежных алгоритмов интерпретации, способных функционировать в условиях высокой неопределенности, а также на интеграцию интерпретируемости в процесс обучения моделей искусственного интеллекта.

Список использованной литературы

1. Liu A.J., Mukherjee A., Hu L., Chen J., Nair V.N. Performance and Interpretability Comparisons of Supervised Machine Learning Algorithms: An Empirical Study // Journal of Machine Learning Research. – 2023. – Vol. 24, No. 6. – P. 112–135.

2. Liu B., Udell M. Impact of Accuracy on Model Interpretations // Advances in Neural Information Processing Systems. – 2022. – Vol. 35. – P. 6789–6803.

3. Alicioglu G., Sun B. A survey of visual analytics for explainable artificial intelligence methods // Computers & Graphics. – 2022. – Vol. 102. – P. 502-520.

4. Samek W., Wiegand T., Müller K. R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models // arXiv preprint arXiv:1708.08296. – 2017.

5. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A review of machine learning interpretability methods // Entropy. – 2020. – Vol. 23, No. 1. – P. 18.

6. Бевзенко С. А. Применение искусственного интеллекта и машинного обучения в разработке программного обеспечения //Инновации и инвестиции. – 2023. – №. 8. – С. 187-191.

7. Дремлюга Р. И., Коробеев А. И. Применение искусственного интеллекта для сбора и анализа доказательств //Юридический вестник Кубанского государственного университета. – 2021. – №. 4. – С. 55-63.