УДК 004.89

ПОДХОДЫ К ОБЕСПЕЧЕНИЮ ПРОЗРАЧНОСТИ ФРЕЙМВОРКОВ МАШИННОГО ОБУЧЕНИЯ

Канарейко С.С., студент гр. 24.М02-мкн, I курс
Санкт-Петербургский государственный университет, г. Санкт-Петербург

Современное машинное обучение играет всё более значимую роль в разнообразных сферах человеческой деятельности, от медицины и финансов до государственного управления и автономных систем. В условиях расширяющегося применения моделей искусственного интеллекта особенно важной становится проблема прозрачности фреймворков машинного обучения, поскольку именно от уровня прозрачности зависит доверие пользователей, возможность аудита и качество принятия решений на основе вычислительных результатов. Прозрачность в данном контексте означает обеспечение открытости, интерпретируемости и воспроизводимости как самого фреймворка, так и моделей, создаваемых с его помощью. Настоящая статья посвящена рассмотрению ключевых подходов к достижению и поддержанию прозрачности в современных ML-фреймворках [1].

Первым направлением является открытость исходного кода и документации. Многие популярные ML-фреймворки, включая TensorFlow, PyTorch и scikit-learn, активно развиваются как проекты с открытым исходным кодом, что позволяет широкому сообществу исследователей и разработчиков контролировать и совершенствовать архитектуру, алгоритмы и реализации. Однако открытость кода сама по себе не гарантирует прозрачность, если отсутствуют понятные и полные описания архитектурных решений, алгоритмических подходов и параметров по умолчанию. Поэтому важным элементом является поддержка качественной и доступной документации, включающей пояснения по внутренним процессам и методам обработки данных.

Вторым ключевым аспектом является интеграция механизмов интерпретируемости и объяснимости. Современные фреймворки предоставляют встроенные или подключаемые модули, которые позволяют визуализировать важность признаков, анализировать влияние параметров модели и получать локальные объяснения конкретных предсказаний. Эти инструменты способствуют не только глубокому пониманию поведения модели, но и позволяют выявлять потенциальные ошибки и предвзятости.

Особенно важна стандартизация форматов таких объяснений, что обеспечивает их совместимость и возможность автоматической проверки в рамках аудита [2].

Третьим подходом выступает обеспечение воспроизводимости экспериментов и вычислений. Фреймворки должны поддерживать контроль версий данных, кода и параметров моделей, а также автоматическое логгирование всего процесса обучения и валидации. Воспроизводимость позволяет не только повторить результаты, но и детально проанализировать причины отклонений, что существенно повышает доверие к системам и облегчает выявление ошибок или нештатных ситуаций [3].

Четвёртым направлением является внедрение механизмов прозрачного мониторинга и аудита. Современные ML-системы зачастую развёрнуты в виде распределённых сервисов с непрерывной интеграцией и доставкой (CI/CD). Интеграция прозрачных инструментов мониторинга, фиксирующих метрики производительности, устойчивости к дистрибутивным сдвигам и соответствия этическим нормам, обеспечивает своевременное обнаружение проблем и своевременное реагирование на них. Важным элементом является создание аудиторских журналов с возможностью детального анализа операций и их последующего воспроизведения [4].

Наконец, следует отметить значимость взаимодействия с нормативными и этическими стандартами. Фреймворки, ориентированные на прозрачность, предусматривают встроенные механизмы проверки соответствия создаваемых моделей действующим законодательным требованиям и этическим принципам, включая защиту персональных данных, отсутствие дискриминации и обеспечение безопасности. Это достигается посредством внедрения автоматизированных проверок, генерации отчётов и инструментов поддержки принятия ответственных решений.

В итоге, подходы к обеспечению прозрачности фреймворков машинного обучения представляют собой комплекс взаимосвязанных решений, охватывающих открытость кода, интерпретируемость, воспроизводимость, мониторинг и нормативное соответствие. Такой многоуровневый подход создаёт основу для формирования доверия к системам ИИ, способствует их широкому внедрению в ответственные сферы и стимулирует дальнейшее развитие науки и практики в области надёжного и этичного машинного обучения [5].

Список литературы:

1. Mitchell, M. Model Cards for Model Reporting / M. Mitchell, S. Wu, A. Zaldivar [et al.] // Proceedings of the Conference on Fairness, Accountability, and Transparency. – 2019. – P. 220–229. – DOI 10.1145/3287560.3287596.

2. Gebru, T. Datasheets for Datasets / T. Gebru, J. Morgenstern, B. Vecchione [et al.] // Communications of the ACM. – 2021. – Vol. 64, № 12. – P. 86–92. – DOI 10.1145/3458723.

3. Hind, M. Increasing Trust in AI Services through Supplier's Declarations of Conformity / M. Hind, S. Mehta, A. Mojsilovic [et al.] // IBM Journal of Research and Development. – 2020. – Vol. 64, № 4/5. – P. 1–13. – DOI 10.1147/JRD.2020.2982288.

4. Arnold, M. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity / M. Arnold, R. K. E. Bellamy, M. Hind [et al.] // IBM Journal of Research and Development. – 2019. – Vol. 63, № 4/5. – P. 6:1–6:13. – DOI 10.1147/JRD.2019.2942288.

5. Протодьяконов, А. В. Алгоритмы Data Science и их практическая реализация на Python : учебное пособие / А. В. Протодьяконов, П. А. Пылов, В. Е. Садовников. – Москва : Инфра-Инженерия, 2022. – 392 с. – ISBN 978-5-9729-1006-9. – EDN BITUQD.