УДК 004.89

АЛГОРИТМИЧЕСКИЙ АУДИТ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ПРЕДМЕТ СОЦИАЛЬНОЙ СПРАВЕДЛИВОСТИ

Ахмед Х.М., студент гр. BIT15, II курс
Университет Кади Айяд, г. Марракеш

Алгоритмический аудит моделей машинного обучения на предмет социальной справедливости представляет собой важный инструмент, направленный на выявление и устранение предвзятости, дискриминации и неравенства в алгоритмических системах. В последние годы применение машинного обучения (ML) в различных областях, таких как кредитование, здравоохранение, правосудие и рекрутинг, значительно возросло, что привело к возникновению новой волны обеспокоенности по поводу возможных социальных последствий использования этих технологий [1].

В частности, если алгоритмы не учитывают важнейшие принципы социальной справедливости, они могут усилить существующие социальные различия, в том числе дискриминацию по признакам расы, пола, возраста, социального положения или этнической принадлежности. Алгоритмический аудит становится важным инструментом для гарантии того, что такие технологии используются этично и без ущерба для общественного благосостояния.

Основной целью алгоритмического аудита является не только проверка точности и производительности модели, но и тщательное исследование её воздействия на различные социальные группы. В контексте социальной справедливости это означает, что модель должна обеспечивать равные возможности для всех категорий пользователей, независимо от их социального фона, демографических характеристик или других чувствительных признаков.

Модели, игнорирующие эти аспекты, могут приводить к систематической дискриминации, создавая «цифровое неравенство» и усиливая существующие социальные барьеры. Например, система, которая не справедливо распределяет кредиты или медицинские ресурсы среди различных этнических групп, или система, которая не учитывает особенности работников в возрасте, может не только повредить отдельным индивидам, но и усилить социальную несправедливость [2].

Алгоритмический аудит социальной справедливости начинается с анализирования данных, на которых обучалась модель. Важно отметить, что данные часто содержат предвзятость, которая может быть передана модели.

Эти данные могут быть исторически искажены или неполны, что особенно актуально в случае с данными, касающимися таких чувствительных признаков, как раса, пол, возраст и другие факторы. Например, если модель обучается на данных, в которых зафиксированы исторически существующие предвзятости в области трудоустройства (например, когда определённые профессии преимущественно занимают мужчины), она может воспроизвести эти предвзятые шаблоны в своих предсказаниях. В рамках алгоритмического аудита важно проверить, насколько хорошо данные представляют все категории пользователей и нет ли в них скрытых предвзятостей, которые могут привести к дискриминации.

Затем проводится оценка справедливости самой модели. Для этого применяются различные метрики, которые измеряют, насколько равномерно распределяются результаты модели для различных групп. Например, используется демографический паритет, который проверяет, одинаково ли модель распределяет положительные результаты для различных групп. Важным аспектом является также равенство ошибок (equalized odds), которое оценивает, насколько одинаковы ошибки модели для разных демографических групп. Если, например, система одобряет кредит в большей степени для одной группы, а для другой она систематически отклоняет заявки, это может быть признаком дискриминации, требующим вмешательства [3].

Особое внимание в алгоритмическом аудите уделяется анализу причинно-следственных связей в модели, поскольку предвзятость может проявляться не только на уровне предсказаний, но и на уровне того, какие признаки модель использует для принятия решения. Для этого применяются методы интерпретируемости и объясняемости моделей, такие как LIME (Local Interpretable Model-agnostic Explanations) и SHAP (Shapley Additive Explanations). Эти подходы позволяют не только понять, какие признаки играют ключевую роль в принятии решений модели, но и проверить, используются ли чувствительные признаки (например, пол или этническая принадлежность) в принятии решений. Если такие признаки оказывают значительное влияние на решение модели, это может привести к нежелательной предвзятости и нарушению принципа социальной справедливости.

Важным этапом является также контрфактическое тестирование модели, когда изменяются данные, например, меняется пол, возраст или этническая принадлежность, чтобы проверить, как это влияет на предсказания. Если изменение одного признака существенно меняет результат модели, это может указывать на то, что модель делает слишком сильный акцент на этом признаке и может быть предвзятой. Это тестирование позволяет выявить скрытую зависимость, которую модель может не осознавать, но которая может иметь серьёзные этические последствия [4].

Одним из ключевых инструментов в алгоритмическом аудите является мониторинг модели в процессе её эксплуатации. Даже если модель была проверена на этапе разработки и показала соответствие стандартам социальной справедливости, это не гарантирует, что она останется справедливой в реальных условиях эксплуатации. Со временем модель может столкнуться с новыми данными, которые не были учтены в процессе её обучения, или её поведение может измениться из-за внешних факторов, таких как изменение социальной или экономической ситуации. Поэтому необходимо регулярно проводить мониторинг работы модели в реальном времени, а также анализировать её предсказания в условиях изменяющихся данных. Это позволит своевременно корректировать модель и предотвращать её возможное отклонение от принципов социальной справедливости.

В заключение, алгоритмический аудит моделей машинного обучения на предмет социальной справедливости играет важнейшую роль в создании этичных и справедливых систем искусственного интеллекта. В процессе такого аудита анализируются данные, алгоритмы, метрики справедливости, а также процессы, с помощью которых моделям обеспечивается независимость от предвзятых или дискриминирующих факторов.

Эффективный аудит позволяет не только повышать доверие к ИИ-системам, но и обеспечивать их соответствие моральным, этическим и юридическим стандартам. В конечном итоге, алгоритмический аудит способствует созданию технологий, которые могут быть справедливо использованы для улучшения жизни людей, без усиления существующих социально-экономических различий [5].

Список литературы:

1. Sandvig, C. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms / C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort // Data and Discrimination: Converting Critical Concerns into Productive Inquiry. – Seattle : University of Washington, 2014. – P. 1–23.

2. Bandy, J. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits / J. Bandy // Proceedings of the ACM on Human-Computer Interaction. – 2021. – Vol. 5, CSCW1. – P. 1–34. – DOI 10.1145/3449148.

3. Metaxa, D. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In / D. Metaxa, J. S. Park, J. A. Landay, J. Hancock // Foundations and Trends in Human-Computer Interaction. – 2021. – Vol. 14, № 4. – P. 272–344. – DOI 10.1561/1100000083.

4. Koshiyama, A. Towards Algorithm Auditing: Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms / A. Koshiyama, E.

Kazim, P. Treleaven [et al.] // SSRN Electronic Journal. – 2021. – DOI 10.2139/ssrn.3778998.

5. Пылов, П. А. Teamlead как разработчик и юридический лидер команды в одном лице / П. А. Пылов, А. В. Протодьяконов, А. И. Бобровских // Россия молодая : сб. материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года. – Кемерово : Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2022. – С. 94406.1–94406.7. – EDN TOAHDH.