УДК 004.89

РАЗРАБОТКА ИНСТРУМЕНТОВ АВТОМАТИЗИРОВАННОГО АНАЛИЗА СПРАВЕДЛИВОСТИ МОДЕЛЕЙ

Щербакова Э.Д., студент гр. 09-361, III курс
Казанский федеральный университет, г. Казань

Разработка инструментов автоматизированного анализа справедливости моделей машинного обучения становится важнейшей задачей для обеспечения этичности и прозрачности алгоритмических систем, особенно в условиях их широкого применения в различных сферах, таких как финансовые услуги, здравоохранение, правосудие и другие критически важные области [1].

Справедливость моделей машинного обучения означает, что алгоритмы должны принимать решения без дискриминации и предвзятости, обеспечивая равные возможности для всех пользователей независимо от их социального положения, расы, пола, возраста, этнической принадлежности или других чувствительных признаков. Однако даже с использованием самых передовых технологий и методов обучения, модели могут случайным образом или систематически проявлять предвзятость в своих решениях, что может приводить к неравенству в распределении ресурсов, несправедливым решениям или усилению социальных барьеров.

В связи с этим, автоматизированный анализ справедливости становится критически важным инструментом для обнаружения и устранения подобных предвзятостей.

Основной проблемой в области машинного обучения является то, что модели часто строятся на основе исторических данных, которые могут содержать предвзятость. Эти данные могут быть искажены социальными, экономическими или культурными факторами, которые затем передаются в обучающие алгоритмы. Например, если данные о трудоустройстве отражают существующую дискриминацию в отношении женщин или этнических меньшинств, модель, обученная на таких данных, может неосознанно воспроизводить эти предвзятые шаблоны в своей работе [2].

Автоматизированные инструменты анализа справедливости позволяют выявлять такие искажения на разных этапах разработки и эксплуатации модели, обеспечивая её соответствие принципам социальной справедливости и этическим нормам.

Разработка таких инструментов включает в себя создание метрик справедливости, которые измеряют, насколько равномерно модель распределяет

свои предсказания между различными группами пользователей. Эти метрики могут быть различными в зависимости от специфики задачи, но основными являются демографический паритет, равенство ошибок и равенство возможностей. Демографический паритет проверяет, одинаково ли модель распределяет положительные или отрицательные исходы между различными группами пользователей. Если модель, например, одобряет кредиты с разной вероятностью для мужчин и женщин, это может свидетельствовать о наличии предвзятости. Равенство ошибок, в свою очередь, оценивает, насколько одинаково модель делает ошибки (ложноположительные и ложные отрицательные) для различных групп. Если для одной группы наблюдается значительное увеличение ошибок, это также является признаком дискриминации. Равенство возможностей проверяет, насколько одинаково модель предсказывает шанс для успеха для разных групп, особенно когда речь идёт о возможностях доступа к ресурсам, таким как кредиты, рабочие места или медицинские услуги.

Для автоматизации этого процесса необходимо использование алгоритмов и программных средств, которые позволяют проводить регулярный мониторинг работы модели с точки зрения справедливости. Эти инструменты могут включать как стандартные библиотеки для вычисления справедливости, такие как AIF360 (AI Fairness 360), Fairness Indicators или Fairness-aware classifiers, так и более сложные платформы для интеграции справедливости в процесс разработки и эксплуатации моделей. Они предоставляют пользователю набор функций для анализа предсказаний модели, оценки метрик справедливости, а также диагностики возможных отклонений от эталонных значений. При этом алгоритмы анализа справедливости должны быть гибкими и адаптируемыми, чтобы учитывать специфику конкретной области применения и особенности данных [3].

Кроме того, важной составляющей таких инструментов является возможность их интеграции с уже существующими фреймворками машинного обучения. Применение таких инструментов в реальной практике помогает внедрять принципы справедливости непосредственно на этапе разработки, а не только в процессе тестирования или на стадии эксплуатации модели. Внедрение алгоритмов, автоматически проверяющих справедливость моделей, позволяет предотвратить дискриминацию на ранних этапах разработки, гарантируя, что принятые алгоритмические решения будут соответствовать современным этическим требованиям.

Одним из значимых аспектов разработки инструментов автоматизированного анализа справедливости является возможность постоянного мониторинга моделей в реальных условиях эксплуатации. Даже если модель прошла все этапы тестирования на справедливость в процессе разработки, она может столкнуться с новыми данными и изменениями, которые могут повлиять на её

производительность и вызвать появление новых форм предвзятости. Автоматизированные инструменты анализа справедливости позволяют отслеживать поведение модели в реальном времени, выявлять перекосы и предвзятость по мере того, как модель продолжает работать в динамично меняющихся условиях. Это особенно важно для таких задач, как предоставление кредитов или медицинская диагностика, где ошибка может повлечь серьёзные последствия для пользователей. Регулярный аудит с помощью автоматизированных инструментов позволяет не только оперативно выявлять такие отклонения, но и вносить корректировки в модель, улучшая её справедливость и точность.

Автоматизация анализа справедливости также позволяет ускорить процесс тестирования и улучшения моделей, делая его более масштабируемым и эффективным. В условиях быстро меняющихся технологий и данных, ручной анализ справедливости может быть трудоёмким и неэффективным, особенно если модели разрабатываются и обучаются на большом количестве данных. В то же время, автоматизированные системы способны в реальном времени проверять модель на множестве критериев справедливости и выдавать отчёты, которые могут быть использованы для дальнейшей работы по улучшению модели. Это особенно актуально для крупных организаций и предприятий, которые работают с огромными объёмами данных и должны обеспечивать не только точность своих моделей, но и их соответствие этическим нормам [4].

В завершение, разработка инструментов автоматизированного анализа справедливости моделей является необходимым шагом для того, чтобы искусственный интеллект и машинное обучение могли эффективно использоваться в различных областях, при этом обеспечивая справедливость, прозрачность и этичность решений.

Это способствует созданию более устойчивых и ответственных систем, которые минимизируют риск дискриминации и социальных перекосов, обеспечивая равные возможности для всех пользователей.

Внедрение таких инструментов в процесс разработки и эксплуатации моделей не только улучшает их производительность, но и укрепляет доверие общества к технологиям искусственного интеллекта [5].

Список литературы:

1. Bird, S. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI / S. Bird, M. Dudik, R. Edgar [et al.] // Microsoft Journal of Applied Research. – 2020. – Vol. 12. – P. 1–11.

2. Tramèr, F. FairTest: Discovering Unwarranted Associations in Data-Driven Applications / F. Tramèr, V. Atlidakis, R. Geambasu [et al.] // IEEE European Symposium on Security and Privacy. – 2017. – P. 401–416. – DOI 10.1109/EuroSP.2017.29.

3. Wexler, J. The What-If Tool: Interactive Probing of Machine Learning Models / J. Wexler, M. Pushkarna, T. Bolukbasi [et al.] // IEEE Transactions on Visualization and Computer Graphics. – 2020. – Vol. 26, № 1. – P. 56–65. – DOI 10.1109/TVCG.2019.2934619.

4. Babic, B. Algorithms on Regulatory Lockdown in Medicine / B. Babic, S. Gerke, T. Evgeniou, I. G. Cohen // Science. – 2019. – Vol. 366, № 6470. – P. 1202–1204. – DOI 10.1126/science.aay9547.

5. Venkatasubramanian, S. Algorithmic Fairness: Measures, Methods and Representations / S. Venkatasubramanian, A. Alfano // Proceedings of the ACM Conference on Economics and Computation. – 2020. – P. 1–2. – DOI 10.1145/3391403.3399504.