

УДК 004.89

ВЫЯВЛЕНИЕ ПРЕДВЗЯТОСТИ В МОДЕЛЯХ НА ОСНОВЕ АНАЛИЗА ОСТАТОЧНОЙ ОШИБКИ

Билан В.В., научный сотрудник

Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Выявление предвзятости в моделях на основе анализа остаточной ошибки является важным этапом в обеспечении справедливости и прозрачности алгоритмов машинного обучения, особенно когда речь идет о системах, которые оказывают влияние на ключевые решения в различных областях, таких как медицина, финансовые услуги, правосудие и трудоустройство [1].

Предвзятость в моделях может привести к несправедливым или дискриминационным результатам, которые затрудняют принятие справедливых и объективных решений. Одним из способов выявления скрытых предвзятостей является анализ остаточной ошибки, который позволяет исследовать, как модель справляется с различными подмножествами данных и какие группы оказываются недооценёнными или неправильно классифицированными.

Остаточная ошибка – это разница между предсказанными значениями модели и фактическими значениями, которые она пытается предсказать. В контексте обнаружения предвзятости анализ остаточной ошибки позволяет определить, какие конкретные группы или категории данных модель обрабатывает хуже, чем другие. Этот метод заключается в том, чтобы исследовать остаточные ошибки модели, то есть расхождения между предсказаниями и реальными результатами, для различных подгрупп данных.

Остаточные ошибки могут указывать на систематические отклонения в поведении модели, что может быть результатом предвзятости, проявляющейся по отношению к чувствительным признакам, таким как пол, раса, возраст или социальный статус [2].

Когда модель демонстрирует высокие остаточные ошибки для одной группы, это может быть сигналом о том, что она плохо обрабатывает данные этой группы. Например, если модель предсказывает кредиты или медицинские диагнозы для определённых групп с более высокими ошибками, чем для других, это может означать, что модель не учитывает специфические характеристики этих групп или делает на них необоснованные предположения. Анализ остаточных ошибок помогает выявить эти ситуации, поскольку остаточная

ошибка будет больше для тех данных, которые модель не смогла корректно интерпретировать или классифицировать.

Для начала анализа остаточных ошибок важно разделить данные на различные группы, которые могут быть различимы по чувствительным признакам. Например, можно разделить данные по таким признакам, как возраст, пол, этническая принадлежность или географическое местоположение. После этого проводится анализ предсказаний модели для каждой из этих групп, чтобы увидеть, имеются ли значительные различия в размере остаточных ошибок для разных категорий. Если модель демонстрирует значительные отклонения в предсказаниях для одной группы по сравнению с другой, это может служить индикатором предвзятости, и потребует дальнейшего вмешательства для устранения этих отклонений.

Особенно важно, что остаточные ошибки могут помочь выявить как явные, так и скрытые формы предвзятости. Явная предвзятость может проявляться в том, что модель систематически игнорирует или дискриминирует определённые группы, например, ошибается чаще в отношении определённой расовой или этнической группы. Скрытая предвзятость может проявляться более тонким образом, например, когда модель учитывает те или иные признаки, которые не должны играть роли в принятии решения, но при этом остаются неприметными. Эти скрытые формы предвзятости могут быть выявлены именно через анализ остаточных ошибок, когда показатели ошибок для разных групп, имеющих схожие характеристики, всё же различаются [3].

Методы, используемые для анализа остаточной ошибки, могут включать различные статистические тесты и визуализации. Один из самых популярных методов – это анализ ошибок по подгруппам, который позволяет разделить данные на несколько категорий и вычислить остаточную ошибку для каждой из них. Например, для задачи прогнозирования кредитного рейтинга можно разделить данные по полу и возрасту, чтобы проверить, не происходит ли существенного увеличения остаточной ошибки для женщин или пожилых людей. Этот подход позволяет не только выявить предвзятость в предсказаниях, но и определить, какие группы данных наиболее уязвимы к ошибкам модели.

Дополнительным инструментом является использование графиков остатков для визуализации различий в ошибках между группами. Такие графики могут наглядно продемонстрировать, как остаточная ошибка меняется в зависимости от характеристик данных. Например, можно построить график распределения ошибок модели по различным категориям и сравнить, есть ли значимые различия в предсказаниях для разных подгрупп. Это позволяет не только обнаружить предвзятость, но и понять, какие факторы оказывают наибольшее влияние на принятие решения моделью.

Кроме того, методы регуляризации и коррекция ошибок могут быть использованы для минимизации предвзятости, выявленной через остаточные ошибки. Например, если модель показывает высокие остаточные ошибки для определённой группы, можно применить методы обучения с учётом этих ошибок, чтобы сбалансировать её производительность для всех групп. Это может включать в себя такие подходы, как переподборка данных или переобучение модели с использованием весов для подгрупп, которые наиболее подвержены ошибкам.

Особое внимание стоит уделить тому, что анализ остаточных ошибок может быть полезен не только на этапе разработки модели, но и на этапе её эксплуатации. Модели, которые продолжают функционировать в реальных условиях, могут столкнуться с новыми данными, которые значительно отличаются от тех, на которых они были обучены. Анализ остаточных ошибок в реальном времени помогает выявлять такие отклонения и своевременно адаптировать модель, чтобы минимизировать предвзятость и улучшить её производительность для всех групп [4].

Важным аспектом является то, что анализ остаточных ошибок помогает обеспечить прозрачность и интерпретируемость модели. Когда пользователи или заинтересованные стороны могут видеть, как модель работает для различных групп и какие ошибки она совершает, это значительно повышает доверие к системе и её результатам. Прозрачность, в свою очередь, критически важна для создания этических и ответственных ИИ-систем, которые соответствуют общественным и правовым стандартам.

В заключение, выявление предвзятости в моделях на основе анализа остаточной ошибки представляет собой мощный и эффективный инструмент для повышения справедливости машинного обучения. Этот подход позволяет не только обнаружить предвзятость, которая может быть скрыта в результатах модели, но и предоставляет методы для её минимизации и корректировки.

В условиях всё более широкого применения ИИ в социальных, юридических и экономических сферах такой анализ становится неотъемлемой частью процесса обеспечения этичности, справедливости и прозрачности решений, принимаемых алгоритмами [5].

Список литературы:

1. Angwin, J. Machine Bias / J. Angwin, J. Larson, S. Mattu, L. Kirchner // ProPublica. – 2016. – 23 May. – URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (дата обращения: 09.06.2025).
2. Dressel, J. The Accuracy, Fairness, and Limits of Predicting Recidivism / J. Dressel, H. Farid // Science Advances. – 2018. – Vol. 4, № 1. – eaao5580. – DOI 10.1126/sciadv.aao5580.

3. Kleinberg, J. Inherent Trade-Offs in the Fair Determination of Risk Scores / J. Kleinberg, S. Mullainathan, M. Raghavan // Proceedings of Innovations in Theoretical Computer Science. – 2017. – P. 43:1–43:23. – DOI 10.4230/LIPIcs.ITCS.2017.43.

4. Pleiss, G. On Fairness and Calibration / G. Pleiss, M. Raghavan, F. Wu [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5680–5689.

5. Chen, I. Y. Why Is My Classifier Discriminatory? / I. Y. Chen, F. D. Johansson, D. Sontag // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 3543–3554.