УДК 004.89

ОЦЕНКА ПРЕДВЗЯТОСТИ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ FAIRNESS-МЕТРИК И ИНДИКАТОРОВ

Окафор Г., студент направления MSc Transport, Mobility and Innovation (EIT Urban Mobility), II курс
Королевский технологический институт, Стокгольм, Швеция

Оценка предвзятости в моделях машинного обучения является важной и сложной задачей, особенно когда речь идет о системах, которые принимают решения, влияющие на людей, такие как финансовые технологии, медицинские системы, правоохранительные органы и трудоустройство [1].

Скрытая предвзятость в таких системах может привести к несправедливым результатам, дискриминации или неэтичным решениям, что снижает доверие пользователей и может привести к юридическим последствиям.

Для эффективного решения этой проблемы активно разрабатываются и внедряются fairness-метрики и индикаторы, которые позволяют количественно оценить степень справедливости работы модели, выявлять предвзятости и минимизировать их влияние на конечные решения. Оценка предвзятости с использованием этих метрик стала важной частью процесса разработки и внедрения ИИ-систем, обеспечивая их прозрачность, этичность и соответствие правовым и социальным нормам.

Fairness-метрики представляют собой количественные показатели, которые позволяют измерить, насколько справедливо модель относится к различным группам или категориям данных, которые могут быть различимы по чувствительным признакам, таким как пол, раса, возраст, этническая принадлежность или социальный статус. Эти метрики помогают определить, есть ли у модели систематические ошибки для разных групп, а также могут выявить скрытые предвзятости, которые не очевидны при обычной оценке точности модели. Среди наиболее известных fairness-метрик можно выделить следующие:

1. Демографический паритет (Demographic Parity): Эта метрика проверяет, равна ли вероятность положительного исхода (например, одобрения кредита, принятия на работу) для разных групп. Идея заключается в том, чтобы модель не делала предсказания, которые зависят от чувствительных признаков, таких как пол или раса, если эти признаки не должны иметь значения для задачи. Если вероятность положительного исхода для разных групп существенно различается, то это может свидетельствовать о наличии предвзятости [2].

2. Эгалитарный паритет (Equal Opportunity): Это метрика, которая оценивает, насколько одинаковыми являются ошибки модели для разных групп в контексте благоприятных исходов. В отличие от демографического паритета, здесь анализируется, обеспечиваются ли одинаковые шансы на положительное предсказание для всех групп, что важно, например, в задачах с высокой социальной значимостью, таких как медицинская диагностика или уголовное правосудие.
3. Равенство шансов (Equalized Odds): Это более строгая метрика, которая проверяет, насколько одинаковы как ложноположительные, так и ложноотрицательные ошибки для различных групп. Этот критерий используется для оценки, насколько модель генерирует справедливые предсказания в отношении всех групп. Он требует, чтобы модель не только предоставляла равные шансы на правильные прогнозы, но и устраняла вероятность систематических ошибок для любой из групп.
4. Предвзятость прогнозируемой ошибки (Predictive Parity): Эта метрика оценивает, имеют ли различные группы одинаковую точность предсказаний модели. То есть проверяется, насколько точно модель предсказывает исход для различных групп, что особенно важно для задач с высоким риском ошибки, таких как кредитование или медицинская диагностика.
5. Статистическая паритетность (Statistical Parity): Это метрика, которая оценивает, насколько хорошо модель предсказывает позитивные исходы для разных демографических групп, не завися от скрытых или непредставленных признаков. Эта метрика позволяет понять, насколько зависимость модели от определённых признаков может влиять на справедливость предсказаний [3].

Для корректной оценки предвзятости необходимо также учитывать индивидуальные особенности задачи, поскольку понятие "справедливости" может различаться в зависимости от контекста. Например, в задачах кредитного скоринга требуется учитывать исторические данные, чтобы компенсировать возможные исторические дискриминации, в то время как в других областях, таких как рекомендательные системы, важно балансировать точность и разнообразие предлагаемых результатов.

Оценка fairness требует не только количественных метрик, но и интеграции с процессами интерпретируемости моделей. Часто метрики fairness сами по себе не позволяют глубоко понять, почему модель делает те или иные предсказания.

Именно поэтому комбинированные подходы, такие как использование методов интерпретируемости и анализа важности признаков, становятся важным инструментом для понимания, какие факторы могут приводить к

предвзятости. Методы, такие как LIME, SHAP или анализ коэффициентов в линейных моделях, позволяют исследовать, как влияние определённых признаков на решения модели может привести к несправедливости.

Кроме того, для обеспечения более справедливых решений можно использовать подходы предобучения и коррекции предвзятости. Например, при обучении моделей на несбалансированных данных можно использовать методы обучения с весами или генерацию синтетических данных, чтобы компенсировать недостаток представления некоторых групп в данных. Также активно исследуются методы, которые включают регуляризацию справедливости, где модель обучается таким образом, чтобы минимизировать предвзятость в дополнение к стандартным целям, таким как точность предсказаний [4].

Многие современные фреймворки и библиотеки, такие как Fairness Flow, AIF360 (AI Fairness 360 от IBM), что ориентированы на решения задач справедливости и этичности в ИИ, предоставляют встроенные методы и алгоритмы для оценки fairness-метрик. Эти инструменты активно интегрируются в процесс разработки ИИ-систем, обеспечивая систематическую проверку моделей на предвзятость и снижение рисков дискриминации.

Важно отметить, что метрики fairness – это не универсальные решения, и их применение требует внимательного подхода. В разных ситуациях эти метрики могут требовать разных интерпретаций и адаптаций. Например, для одних задач важнее баланс между ошибками ложных отрицаний и ложных положительных предсказаний, а для других – минимизация различий в вероятности положительного исхода между группами. Таким образом, выбор правильной метрики зависит от контекста задачи, целей модели и специфики данных.

В заключение, использование fairness-метрик и индикаторов в машинном обучении позволяет эффективно выявлять и минимизировать дискриминацию в обученных моделях.

Внедрение этих методов способствует созданию более справедливых и этичных искусственных интеллектов, что крайне важно для повышения доверия пользователей, соблюдения этических норм и минимизации социальных рисков [5].

Список литературы:

1. Verma, S. Fairness Definitions Explained / S. Verma, J. Rubin // IEEE/ACM International Workshop on Software Fairness. – New York : ACM, 2018. – P. 1–7. – DOI 10.1145/3194770.3194776.

2. Berk, R. Fairness in Criminal Justice Risk Assessments: The State of the Art / R. Berk, H. Heidari, S. Jabbari [et al.] // Sociological Methods & Research. – 2021. – Vol. 50, № 1. – P. 3–44. – DOI 10.1177/0049124118782533.

3. Pessach, D. A Review on Fairness in Machine Learning / D. Pessach, E. Shmueli // ACM Computing Surveys. – 2023. – Vol. 55, № 3. – Article 51. – DOI 10.1145/3494672.

4. Makhlouf, K. Survey on Causal-Based Machine Learning Fairness Notions / K. Makhlouf, S. Zhioua, C. Palamidessi // arXiv preprint. – 2020. – arXiv:2010.09553.

5. Jacobs, A. Z. Measurement and Fairness / A. Z. Jacobs, H. Wallach // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2021. – P. 375–385. – DOI 10.1145/3442188.3445901.