УДК 004.89

АНАЛИЗ КОРРЕЛЯЦИЙ МЕЖДУ ВЫХОДАМИ МОДЕЛЕЙ И ЧУВСТВИТЕЛЬНЫМИ АТРИБУТАМИ

Нуреддин Б.А., студент гр. BIT15, II курс
Университет Кади Айяд, Марокко, г. Марракеш

Анализ корреляций между выходами моделей и чувствительными атрибутами является важным инструментом для выявления предвзятости и обеспечения справедливости в алгоритмических решениях. Чувствительные атрибуты, такие как пол, раса, возраст, этническая принадлежность, сексуальная ориентация и другие, являются признаками, которые могут быть использованы для принятия решений, но которые, в идеале, не должны влиять на результаты моделей, если речь идет о честных и справедливых алгоритмических системах. Однако, несмотря на этические стандарты, такие атрибуты часто оказывают влияние на предсказания моделей, что может привести к нежелательным последствиям, таким как дискриминация или усиление существующих социальных неравенств. Важной частью работы с такими моделями является анализ корреляций между их выходами (предсказаниями) и чувствительными атрибутами, чтобы выявить скрытые зависимости, которые могут нарушать принципы справедливости [1].

Одной из ключевых целей такого анализа является определение, насколько сильно выходы модели (например, вероятность одобрения кредита, диагноз заболевания или рекомендация для трудоустройства) зависят от чувствительных признаков, которые, по сути, не должны играть роль в принятии решения. Например, в контексте кредитного скоринга или трудоустройства, если модель находит сильную корреляцию между полом или расой человека и вероятностью одобрения кредита или найма, это может указывать на то, что алгоритм несправедливо принимает во внимание эти признаки, что, в свою очередь, приводит к дискриминационным решениям. Таким образом, важно не только понимать, какие атрибуты могут повлиять на решение модели, но и оценить, в какой степени эти атрибуты используются в алгоритмическом процессе.

Методы анализа корреляций между выходами моделей и чувствительными атрибутами включают как статистические, так и машинно-обученческие подходы. Статистический анализ корреляции может быть проведён с использованием стандартных методов, таких как Коэффициент Пирсона, Спирмена или Кендала, которые измеряют степень связи между двумя переменными. В контексте машинного обучения это может быть полезно для выявления случаев,

когда модель использует чувствительные признаки для принятия решения, и где эта связь статистически значима [2].

Например, если модель имеет высокую корреляцию между расой и её предсказаниями, это сигнализирует о том, что данные о расе сильно влияют на итоговые решения.

Однако важно помнить, что даже если корреляция между чувствительными признаками и выходом модели невелика, это не обязательно означает отсутствие предвзятости. В некоторых случаях непрямые зависимостимежду признаками могут привести к проявлению скрытых форм дискриминации. Например, если модель использует такие признаки, как уровень образования, доход или место проживания, которые могут быть косвенно связаны с чувствительными признаками, это также может быть источником предвзятости. Для более глубокого анализа таких ситуаций часто используют методы многократного анализа или регрессионных моделей, чтобы выявить скрытые зависимости, которые могут не быть очевидными при простом расчете корреляции [3].

Одним из эффективных инструментов для оценки таких зависимостей является анализ значимости признаков. Этот подход заключается в вычислении важности каждого признака для предсказаний модели. В случае, если модель избыточно использует чувствительные признаки, такие как пол или раса, для принятия решения, это может означать, что алгоритм несправедливо ориентирован на эти атрибуты. Для этого можно применить методы интерпретируемости моделей, такие как SHAP (Shapley Additive Explanations) или LIME (Local Interpretable Model-Agnostic Explanations), которые помогают понять, как конкретные признаки влияют на решение модели. Если модели предсказывают результаты, сильно зависящие от чувствительных атрибутов, это может быть индикатором того, что модель работает с скрытыми предвзятостями.

Ключевым этапом анализа является выявление скрытых предвзятостей, которые могут быть неочевидны на уровне простых корреляций, но проявляться через более сложные взаимодействия между признаками. Это требует применения методов взаимодействия признаков и многоуровневых моделей, которые могут учитывать взаимодействия между признаками на более высоких уровнях, а не только на уровне их прямой связи с предсказаниями. Например, если модель классификации в медицинской сфере использует возраст, пол и состояние здоровья для постановки диагноза, но при этом возраст и пол не должны быть значимыми для определённых заболеваний, это может быть признаком того, что модель использует эти атрибуты ненадлежащим образом.

Анализ контрфактической справедливости – ещё один полезный инструмент для изучения, как изменения чувствительных признаков влияют на результат модели. Контрфактический анализ позволяет проверить, как изменение одного или нескольких признаков (например, изменение пола или расы) при

сохранении остальных факторов неизменными влияет на предсказания модели. Если при таких изменениях модель продолжает показывать значительное изменение результата, это указывает на то, что модель слишком сильно зависит от этих чувствительных признаков, что нарушает принципы справедливости [4].

Методы перераспределения весов также могут быть использованы для устранения предвзятости в модели. Если выявлена сильная зависимость между выходами модели и чувствительными признаками, можно применить корректирующие методы, такие как выравнивание вероятностей или корректировка на уровне обучающих данных. Эти методы позволяют уменьшить влияние чувствительных признаков на результат и сделать модель более справедливой, сводя к минимуму возможность дискриминации на основе атрибутов, которые не должны быть основанием для принятия решения.

Внедрение этих методов и стратегий в процессы разработки и тестирования моделей помогает не только выявить и устранить предвзятость, но и улучшить прозрачность и доверие к алгоритмическим решениям. Важно помнить, что анализ корреляций между выходами моделей и чувствительными атрибутами не только улучшает этичность моделей, но и способствует более ответственному и справедливому использованию технологий машинного обучения в обществе. Это помогает избежать неравенства и дискриминации в тех областях, где автоматизированные системы могут оказывать значительное влияние на жизни людей, и обеспечивает более справедливые и равные условия для всех групп пользователей [5].

Список литературы:

1. Kilbertus, N. Avoiding Discrimination through Causal Reasoning / N. Kilbertus, M. Rojas-Carulla, G. Parascandolo [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 656–666.

2. Madras, D. Learning Adversarially Fair and Transferable Representations / D. Madras, E. Creager, T. Pitassi, R. Zemel // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 3384–3393.

3. Edwards, H. Censoring Representations with an Adversary / H. Edwards, A. Storkey // International Conference on Learning Representations. – 2016. – P. 1–14.

4. Johndrow, J. E. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction / J. E. Johndrow, K. Lum // Annals of Applied Statistics. – 2019. – Vol. 13, № 1. – P. 189–220. – DOI 10.1214/18-AOAS1201.

5. Louizos, C. The Variational Fair Autoencoder / C. Louizos, K. Swersky, Y. Li [et al.] // International Conference on Learning Representations. – 2016. – P. 1–14.