лізувати та використовувати дані все більше визначатиме успіх кінокомпаній у глобальному масштабі.

Список використаних джерел

1. Kulkarni S. S., Vaishnavi M., Kusuma H. Predicting the Conceptual Appeal of Movies using Data Analytics. *International journal of engineering research & technology (IJERT).* Vol. 09, iss. 02 (February 2020).
2. Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success / P. C. Murschetz, C. Bruneel, J.-L. Guy, D. Haughton, N. Lemercier, M.-D. McLaughlin, K. Mentzer, et al. Nordic journal of media management. 2020. Vol. 1, № 3. P. 339–359.
3. The Role of Data Analytics in Cinema. URL: https://www.cloudthat.com/resources/blog/the-role-of-data-analytics-in-cinema

УДК 004.032.26:621.397.331

Лаптєва М. А., здобувачка вищої освіти,
Хмелівський Ю. С., асистент кафедри
інформаційних технологій

ВИЯВЛЕННЯ УПЕРЕДЖЕНЬ У ШІ ЗА ДОМОПОГОЮ СТАТИСТИЧНОГО НАВЧАННЯ

Донецький національний університет імені Василя Стуса, м. Вінниця

Штучний інтелект (ШІ) дедалі більше стає невід'ємною частиною процесів прийняття рішень у різних галузях – від повсякденних потреб, робочих питань до надважливих для держави сфер. Однак зі зростанням популярності систем штучного інтелекту пропорційно зростає занепокоєння щодо помилок і упереджень, які можуть виникнути під час роботи. Вони можуть виникати на різних етапах конвеєра машинного навчання, включно зі збором даних, попередньою обробкою, вибором функцій, навчанням моделі та розгортанням. Виявлення та пом'якшення цих упереджень має вирішальне значення для забезпечення справедливості, підзвітності та прозорості.

Знаходження упереджень передбачає поєднання статистичного аналізу, знань предметної області та критичного мислення. Використовуючи методи перевірки гіпотез, регресійного аналізу і метрики справедливості, статистичне навчання може допомогти оцінити, чи непропорційно впливають прогнози або класифікації ШІ-моделі на певні групи. У цій статті досліджується роль статистичного навчання у виявленні упередженості в ШІ, його методології та застосування.

Зміщення в моделях ШІ часто перетинаються з різними джерелами, що робить складним завдання їх детекції. Щоб розв'язати проблему, спочатку потрібно виявити її корінь.

По-перше, часто проблеми виникають з навчальними даними. Важливо ретельно переглянути їх на наявність будь-яких властивих упереджень, тому що якщо ці дані відображають наявні суспільні упередження, модель збереже їх і сприйме як щось регулярне. Наприклад, набори даних, що використовуються

в системах розпізнавання облич, можуть недопредставляти групи меншин, що призводить до зниження точності для цих груп населення. Тому вибір даних для навчання ШІ має бути зроблений ретельно.

По-друге, упередженість може бути вкорінена в процесі навчання моделі через алгоритмічний вибір, гіперпараметри або цілі оптимізації. Алгоритми можуть посилювати упередження під час оптимізації точності або інших показників продуктивності, нехтуючи міркуваннями справедливості. Це явище, відоме як «алгоритмічне упередження», може призвести до ненавмисної дискримінації.

І, по-третє, творці механізмів штучного інтелекту можуть вносити власні упередження у вибір дизайну та процеси оцінки [1].

Але у машинному навчанні існує деяка «Fairness Metrics», що в перекладі означає «Метрика неупередженості або чесності». Ось декілька пунктів звідти, які допоможуть уникнути упередженості:

1. Статистичний або демографічний паритет. Вимірює, чи рівномірно розподілені прогнози між групами.

How to Calculate Statistical Parity:

$$P(Outcome=1|Group=A)=P(Outcome=1|Group=B)$$

Рисунок 1 – Формула обрахунку статистичного паритету

2. Рівність шансів. Гарантує, що рівень помилок моделі є однаковим для різних груп.

How to Calculate Equality of Odds:

$$P(Outcome=1|Actual=0,Group=A)=P(Outcome=1|Actual=0,Group=B)$$
$$P(Outcome=1|Actual=0,Group=A)=P(Outcome=1|Actual=0,Group=B)$$

Рисунок 2 – Формула обрахунку рівності шансів

3. Калібрування. Перевіряє, чи збігаються прогнозовані ймовірності з фактичними результатами для всіх підгруп [2].

Після оцінки справедливості моделі наступним кроком є аналіз даних, щоб зрозуміти джерела упереджень. **Дослідницький аналіз даних (EDA)** є ключовим етапом, який допомагає виявити проблеми ще до навчання моделі. За допомогою візуальних інструментів, як-от гістограми або діаграми розсіювання, можна побачити, чи є в наборі даних недостатньо представлені групи. Описова статистика, як-от середнє значення або розподіл даних, дає уявлення про дисбаланс, а статистичні тести, наприклад, хі-квадрат або t-тести, допомагають визначити, чи є ці відмінності статистично значущими.

Коли упередження виявлені, потрібно збалансувати дані, щоб усі групи були належно представлені. Для цього використовують методи **надмірної ви-**

бірки або **недостатньої вибірки.** Надмірна вибірка, наприклад, за допомогою техніки SMOTE створює синтетичні дані для недостатньо представлених груп, збільшуючи їх кількість. Недостатня вибірка, навпаки, зменшує кількість точок у перепредставлених групах. Обидва підходи допомагають моделі навчатись на більш збалансованому наборі даних, зменшуючи ризик упередженості в результатах [3].

Ще однією важливою частиною є робота з ознаками, які можуть викликати упередженість у моделі. Наприклад, деякі ознаки можуть бути проксі для чутливих атрибутів, як-от раса чи стать. Поштовий індекс, наприклад, може приховувати соціально-економічний статус і впливати на прогнози моделі. Аналіз кореляції дає змогу виявити такі небажані зв'язки. Якщо виявлено, що певна ознака несе упередженість або мало впливає на якість прогнозу, її можна видалити чи трансформувати. Методи зменшення розмірності, як-от PCA (аналіз головних компонент), допомагають залишити лише ті ознаки, які мають значення для моделі.

Для зменшення упередженості під час навчання моделі використовуються статистичні методи **регуляризації.** Вони додають обмеження, які карають модель за надмірну залежність від чутливих ознак. Наприклад, якщо модель сильно спирається на проксі-ознаки, регуляризація зменшує їх вплив, допомагаючи зробити прогнози більш справедливими. Інтеграція обмежень справедливості в процес навчання гарантує, що боротьба з упередженнями є частиною самої розробки моделі, а не відбувається після її створення.

Після завершення навчання моделі статистичні інструменти дають змогу оцінити результати та перевірити, чи залишились у прогнозах залишкові упередження. Наприклад, **контрфактичний аналіз** дає змогу перевірити, як модель змінює свої прогнози для двох схожих випадків, які відрізняються лише в одному чутливому атрибуті (наприклад, стать чи раса). Якщо модель дає різні результати, це вказує на дискримінацію. До того ж перевірка калібрування допомагає впевнитися, що передбачені ймовірності відповідають реальним результатам у всіх групах [4].

У підсумку, статистичне навчання дає змогу не лише виявляти упередження, а й активно боротися з ними на кожному етапі – від аналізу даних і підготовки до навчання моделі та її оцінки. Це забезпечує створення більш справедливих і прозорих систем.

Список використаних джерел

1. Responsible AI Development: Bias Detection and Mitigation Strategies. Agiliway. URL: http://surl.li/lsvpjm (дата звернення 29.11.2024).

2. Fairness Metrics In Machine Learning. Aporia. URL: http://surl.li/ygmtib (дата звернення 30.11.2024).

3. Machine Learning Yearning. Andrew Ng. 2019. 118 p. (дата звернення 30.11.2024).

4. Interpretable Machine Learning: A Guide For Making Black Box Models Explainable. Christoph Molnar. 2022. 328 p. (дата звернення 30.11.2024).