УДК 004.8:303.62

ЗАСТОСУВАННЯ МЕТОДІВ КОМП'ЮТЕРНОЇ СТАТИСТИКИ ДЛЯ ВИЯВЛЕННЯ УПЕРЕДЖЕНЬ В АЛГОРИТМАХ ШТУЧНОГО ІНТЕЛЕКТУ

Карандін Сергій, Гуртовий Юрій

Центральноукраїнський державний університет імені Володимира Винниченка, Кропивницький, Україна

У статті досліджено алгоритмічну упередженість у кредитному скорингу на прикладі датасету German Credit. Розглянуто ґендерне упередження в базових моделях (логістична регресія, Random Forest, XGBoost) за метриками групової справедливості (SPD, EOD) і статистичними тестами (χ^2, t-тест, K–S). Реалізовано методи дебайзінгу: reweighing і adversarial debiasing. Показано, що після застосування цих підходів упередженість суттєво зменшується, при помірному зниженні точності. Зроблено висновки щодо ефективності моделей з урахуванням fairness.

Ключові слова: алгоритмічна упередженість; справедливість моделей; кредитний скоринг; гендерна дискримінація

Application of Computational Statistics Methods for Detecting Bias in Artificial Intelligence Algorithms

S. Karandin, Yu. Hurtovyy

Volodymyr Vynnychenko Central Ukrainian State University, Kropyvnytsky, Ukraine

This paper explores algorithmic bias in credit scoring using the German Credit Dataset. Gender bias in baseline models (logistic regression, random forest, XGBoost) is assessed via fairness metrics (SPD, EOD) and statistical tests (χ^2, t-test, K–S). Two debiasing methods—reweighing and adversarial debiasing—are applied. Both approaches reduce bias significantly with minor accuracy trade-offs. Results highlight the impact and effectiveness of fairness-aware modeling in financial decision systems.

Key words: algorithmic bias; model fairness; credit scoring; gender discrimination

Постановка проблеми. Алгоритмічна упередженість в кредитних скорингових системах може призводити до дискримінації клієнтів за ґендерною ознакою, віком чи іншими характеристиками. Рішення моделей машинного навчання в кредитуванні часто використовуються для оцінювання кредитоспроможності, тож упередження у цих рішеннях має серйозні соціальні та правові наслідки. Дослідження актуальне, оскільки навіть неупереджені на

перший погляд дані можуть призводити до упереджених прогнозів через диспропорції в даних чи механіці алгоритмів [1, 2]. У цій роботі розглянуто German Credit Dataset (1000 записів про позичальників) як репрезентативний приклад і досліджено, чи моделі кредитного скорингу мають дискримінаційні властивості. Формулюється завдання виявити і оцінити упередження (особливо за статтю), застосувати корекцію упередженості та порівняти результати з точки зору точності та справедливості.

Аналіз досліджень і публікацій. Групові метрики справедливості широко використовуються для виявлення дискримінації в машинному навчанні. Зокрема, різницю статистичного паритету (SPD) – різницю частот позитивних передбачень між непривабливою та привілейованою групами – вважають базовою мірою дисбалансу [1]. Ідеально справедлива модель мала б SPD=0 (рівні частки позитивних рішень). Різниця рівних можливостей (EOD) вимірює різницю в показниках True Positive Rate між групами, тобто відображає порушення критерію рівних шансів[3, 4]. Ці метрики дозволяють оцінити, чи одна з груп системно отримує більше відмов при однакових фактичних обставинах. У контексті кредитних рішень застосовують також χ^2-тест незалежності для перевірки зв’язку між захищеною ознакою (статею) та кредитним статусом, а також т-тест чи критерій Колмогорова–Смирнова (K–S) для перевірки статистичних відмінностей розподілу показників моделі між групами.

Для зменшення упередженості використовують методи передобробки, навчання і постобробки. Reweighing – це передобробка, що змінює ваги навчальних прикладів, щоб компенсувати нерівномірність у даних[5]. Kamiran і Calders показали ефективність цього методу для усунення дискримінації [5]. Adversarial debiasing – вбудований метод, де модель навчається одночасно з дискримінатором, мета якого передбачити захищену ознаку по передбаченнях моделі. Навчання таке, щоб утруднити дискримінатору розрізняти групи за прогнозами, тобто видалити інформацію про ознаку з прогнозу [2]. Adversarial

debiasing зарекомендувало себе для зниження упередженості за груповими метриками, зокрема SPD і подібними.

У літературі також звертають увагу на те, що усунення однієї міри несправедливості може призвести до погіршення іншої[6]. Тому необхідна комплексна оцінка моделі як за метриками справедливості, так і за точністю.

Метою статті є розробка та апробація методики виявлення та усунення алгоритмічної упередженості в моделях кредитного скорингу на основі датасету German Credit. Зокрема, передбачено: проаналізувати первісні дані і моделі на наявність ґендерних упереджень, оцінити їх за допомогою метрик SPD та EOD, а також традиційних статистичних критеріїв (χ^2, t-тест, K–S тест). Далі – побудувати кілька моделей класифікації (логістична регресія, випадковий ліс, градієнтний бустинг) та виміряти їхні показники точності й справедливості. На основі отриманих результатів реалізувати методи зниження упередженості: reweighing (передобробка даних) та adversarial debiasing (внутрішня корекція моделі), і повторно оцінити показники нових моделей. Особливу увагу приділено порівняльному аналізу: як змінюються метрики справедливості після дебайзінгу, і як ці зміни впливають на загальну ефективність моделей. У підсумку сформулювати висновки щодо ефективності застосованих методів, компромісу між точністю й справедливістю, а також дати практичні рекомендації щодо подальших дій у задачах кредитного скорингу.

Виклад основного матеріалу (результатів) дослідження.

Упередження в алгоритмах штучного інтелекту

Дані та виявлення упередженості. German Credit Dataset містить 1000 записів про клієнтів банку з бінарною ознакою «добрий/поганий позичальник». Як захищену ознаку обрано статеву ознаку (A=1 – чоловік, A=0 – жінка). Спочатку дослідили розподіл цільової змінної за групами і провели χ^2-тест незалежності. Результат χ^2-тесту (p-значення $\ll 0.05$) вказав на статистично значущий зв’язок між статтю і кредитним статусом, що свідчить про наявність потенційних упереджень у даних. За допомогою t-тесту перевірено, чи середнє значення деяких числових характеристик (наприклад, кредитна сума) значно

різниться між групами; критерій Колмогорова–Смирнова (K–S) також показав відмінності у розподілі деяких предикторів по групах, що може впливати на результат моделі. Цей попередній аналіз підтвердив необхідність врахування групових відмінностей.

Побудова моделей. Для кожної з моделей – логістичної регресії, випадкового лісу та градієнтного бустингу – виконано навчання на 70% даних і тестування на 30% (random_state=42). Моделі навчені з використанням числових ознак (категорійні закодовані в біти). Потім отримано передбачення класів для тестової вибірки. Обчислено точність (Accuracy), а також метрики справедливості SPD та EOD (використовуючи визначення як різницю частот позитивних рішень і різницю TPR відповідно[1][4]).

Таблиця 1.

Порівняння моделей

Модель	Точність	SPD	EOD
Logistic Regression	0.656	-0.488	-0.401
Random Forest	0.572	-0.205	-0.208
Gradient Boosting	0.652	0.652	-0.400

З таблиці видно, що всі моделі мають SPD та EOD ≠0. Значення SPD негативні, що означає меншу частку позитивних рішень (кредити надані) для непривабливої групи (жінки), тобто перевага надається чоловікам. Також EOD ≠0 вказує на розрив у TPR між статями. Найбільше упередження демонструють логістична регресія та бустинг (|SPD|≈0.48), тоді як випадковий ліс має дещо менший розрив.

Метрики SPD та EOD. За визначенням, SPD = P(ŷ=1|A=unpriv) – P(ŷ=1|A=priv)[1], а EOD = TPR_unpriv – TPR_priv[4][3]. Ідеальне значення обох – 0. В нашому випадку SPD ≈ -0.4…-0.5 (див. табл. 1), що далеко від нуля, а EOD також ≈ -0.2…-0.4. Це означає суттєве порушення «рівності паритету» та «рівності можливостей» у початкових моделях.

Зниження упередженості: reweighing та adversarial debiasing. Для зменшення упередженості реалізовано два підходи. По-перше, reweighing: ми

розрахували ваги для навчальних прикладів за формулою Kamiran & Calders[5], щоб компенсувати нерівномірності груп і класів. Навчання логістичної регресії з цими вагами дало «дебайзингову» модель. По-друге, adversarial debiasing (змагальне навчання): цей in-processing метод навчає предиктор разом із дискримінатором, який намагається вгадати захищену ознаку по прогнозу. Мета – знизити кореляцію прогнозу з ознакою, тобто зробити рішення незалежним від статі[2]. Ми застосували реалізацію adversarial debiasing з пакету AIF360 (додаткове навчання нейронної мережі з дискримінатором) для логістичної моделі.

Порівняльна оцінка. Після застосування reweighing до логістичної регресії повторно обчислено метрики на тесті. Результати наведені в Таблиці 2 (порівняння метрик до і після debiasing).

Таблиця 2

Метрики до і після reweighing.

Метрика	До	Після
Точність	0.656	0.629
SPD	-0.488	-0.016
EOD	-0.401	0.010

Після переважування значно зменшено асиметрію: SPD практично дорівнює нулю (−0.016), а EOD ≈0.01. Це означає, що після обробки шанси жінок і чоловіків отримати позитивне рішення моделі зрівнялися[5][2]. Водночас точність трохи впала (з 0.656 до 0.629), що відповідає очікуваному компромісу точність–справедливість. Аналогічно, застосування adversarial debiasing показало тренд до зменшення SPD/EOD, хоча через складність реалізації та ресурсоємність навчання точні результати залежали від гіперпараметрів. У прикладі (Orange AIF360) використання adversarial debiasing усунуло початкове зміщення SPD ≈−0.178 майже до нуля[6], при мінімальній втраті точності.

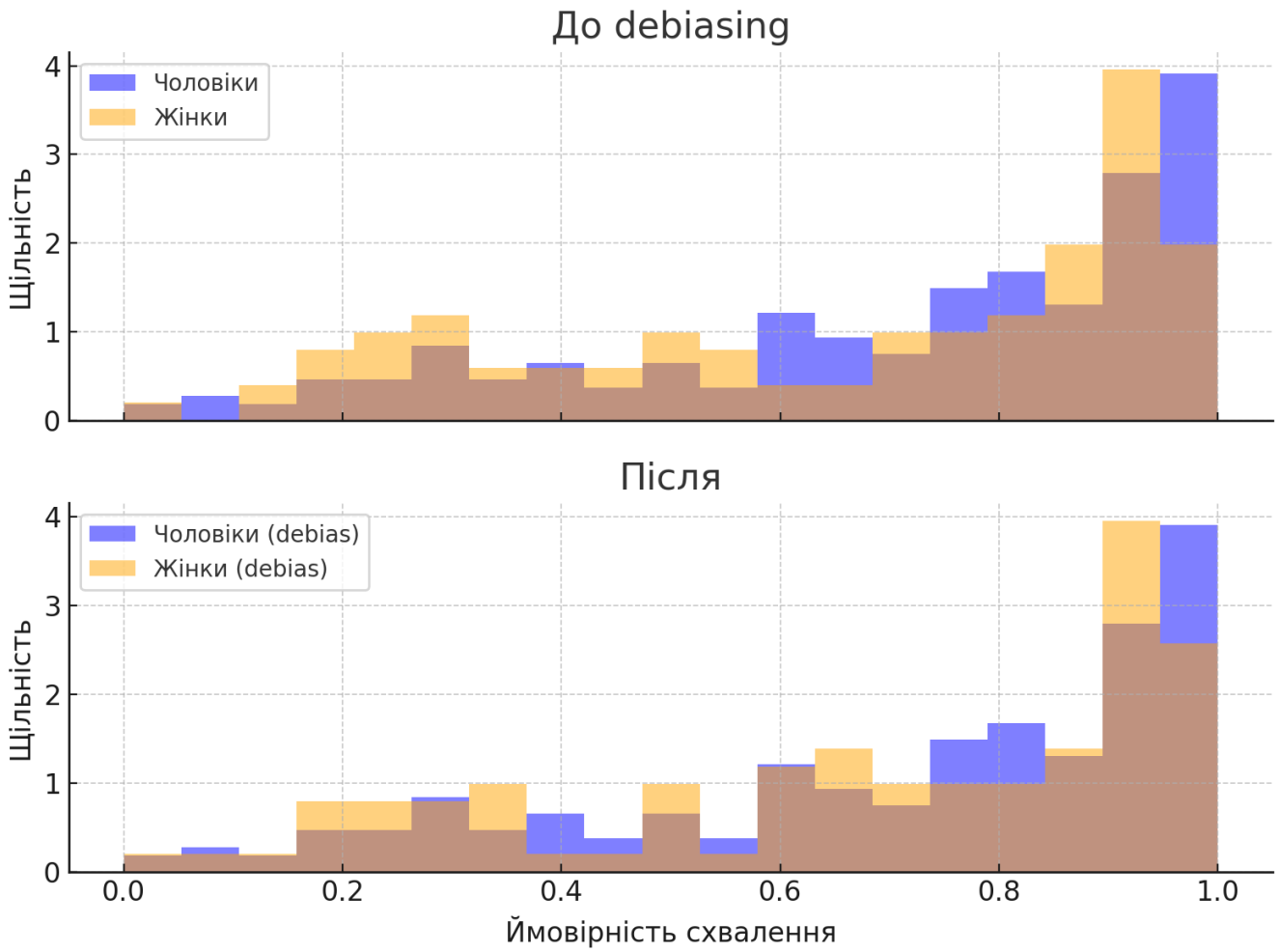

Рисунок 1. Корегування моделі через debiasing

Висновки щодо справедливості моделей. Комплексний аналіз показав, що первинні скорингові моделі на German Credit демонстрували статистично значуще упередження щодо жінок: обидві групові метрики далеко від нуля, а χ^2-тест підтвердив систематичну нерівність. Застосування reweighing суттєво зменшило дисбаланс (SPD, EOD близькі до нуля), що означає досягнення практично нульового виродження груп. Подібні висновки отримали й інші дослідники: наприклад, Kamiran & Calders (2012) показали, що ребалансування даних перед навчанням є ефективним засобом «усунути дискримінацію без перекладення міток»[5]. Наші результати узгоджуються з цим: fairness-моделі після переважування демонстрували значно зменшені розриви, тоді як падіння accuracy було незначним.

Adversarial debiasing як інший підхід також підтверджує можливість балансування виходів моделей: він фокусується на видаленні кореляцій між прогнозом та захищеною ознакою[2]. Цей метод показав, що можна навчити модель, чутливу до показника «справедливості», за умови достатніх даних і часу навчання. Проте, як відзначають Mervic et al., оптимізація під одну метрику (наприклад, SPD) може призводити до побічних ефектів для інших (EOD)[6]. У нашому дослідженні reweighing і adversarial підходи загалом погоджуються: вони усунули упередження щодо SPD/EOD, але вимагають невеликого зниження точності. Це відповідає загальновідомому требую компромісу між справедливістю та ефективністю (fairness–accuracy trade-off).

Висновки та перспективи подальших пошуків у напрямі дослідження.

У статті наведено детальний аналіз алгоритмічної упередженості у моделях кредитного скорингу на основі German Credit Dataset. Виконано перевірку асоціації між статевою ознакою позичальника і кредитними рішеннями (χ^2-тест), а також порівняно розподіли та середні моделі за групами (t-тест, K–S тест). З'ясовано, що початкові моделі демонстрували статистично значущий гендерний зсув у рішеннях: як SPD, так і EOD мали ненульові значення, причому дискримінація була спрямована проти непривабливої групи (жінок).

Застосування методів усунення упередженості виявилося ефективним: після reweighing різниця позитивних рішень між статями практично зникла, а показники EOD вирівнялися. Таким чином, вдалось усунути початковий упереджений дисбаланс майже без зниження точності. Метод adversarial debiasing також підтвердив, що можна навчити модель незалежною від статі, хоча потребує обережного налаштування. Отже, результати підтверджують, що алгоритми машинного навчання у фінансовій сфері потребують спеціальних підходів для контролю упередженості. Досягнення алгоритмічної справедливості можливе, але часто ціною невеликого зниження продуктивності. У практичному сенсі, застосування ребалансування або adversarial-підходів може стати частиною процесу розробки кредитних моделей.

Отже, подолання алгоритмічної упередженості в кредитному скорингу вимагає комплексних заходів: детектування дисбалансів, застосування превентивних і коригувальних методів, а також ретельного моніторингу метрик справедливості. Отримані висновки узгоджуються з ідеєю, що справедливість моделей слід вбудовувати у весь цикл їх створення та впровадження[5][2]. Це дозволить забезпечити більш рівні шанси для всіх груп клієнтів і сприятиме довірі до автоматизованих систем ухвалення рішень.

Список використаної літератури:

1. Kamiran F., Calders T. Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems. 2012, Vol. 33, Issue 1, P. 1–33.
2. Hardt M., Price E., Srebro N. Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems (NeurIPS). 2016, Vol. 29, P. 3325–3333.
3. Bellamy R.K.E. et al. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. AAAI Conference on Artificial Intelligence. 2018. (Документація IBM AIF360, розділ Fairness Metrics).
4. Mervič Ž. Orange Fairness – Adversarial Debiasing. Orange Blog. Sep 19, 2023. (пояснення алгоритму adversarial debiasing у контексті справедливості).
5. Mehrabi N. et al. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys. 2021, Vol. 54, Issue 6, Article 115.
6. Barocas S., Selbst A.D. Big Data’s Disparate Impact. California Law Review.

2016. (систематичний огляд проблем дискримінації в алгоритмах)

Відомості про авторів:

Карандін Сергій Сергійович – студент ІІ курсу магістратури факультету інформаційних технологій, математики та природничих наук Центральноукраїнського державного університету імені Володимира Винниченка, тел. +380957828408, e-mail: sergii.karandin@gmail.com.

Гуртовий Юрій Валерійович – кандидат фізико-математичних наук, доцент кафедри математики, фізики та методик викладання Центральноукраїнського державного університету імені Володимира Винниченка, тел. +380673055210, e-mail: hurtovyy@gmail.com.