Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

УДК 004.85:004.94
DOI: https://doi.org/10.30837/2522-9818.2025.1.005

Д. Герасимук, А. Поляков, В. Федорченко

ВИЯВЛЕННЯ КОМПРОМІСІВ МІЖ СПРАВЕДЛИВІСТЮ, СТАБІЛЬНІСТЮ І ТОЧНІСТЮ ДЛЯ ВІДПОВІДАЛЬНОГО ВИБОРУ МОДЕЛІ МАШИННОГО НАВЧАННЯ

Предметом дослідження в статті є процес вибору моделі машинного навчання, що виконується дата-саєнтистами для побудови моделей у критично важливих сферах. **Мета роботи**: 1) створити програмну бібліотеку для вимірювання точності, стабільності та справедливості моделей; 2) провести експерименти для виявлення компромісів між справедливістю, стабільністю і точністю; 3) запропонувати відповідальний процес вибору моделей для підвищення безпеки використання моделей машинного навчання. У статті передбачено розв'язати такі **завдання**: виміряти справедливість і стабільність моделей машинного навчання та дослідити їх взаємозв'язок з точністю моделей. Упроваджено такі **методи:** емпіричне оцінювання, теорія розкладання помилки моделі на упередження та дисперсію, теорія алгоритмічної справедливості та методи кількісного оцінювання невизначеності. **Досягнуті результати:** 1) запропоновано низьку передбачувальну мінливість як бажану властивість для забезпечення безпеки та рівність варіативності між різними соціальними групами як нову метрику справедливості моделей машинного навчання; 2) продемонстровано, як аналіз стабільності допомагає фахівцям долати виклики множинності моделей і обирати надійні, стійкі та справедливі моделі; 3) створено програмний фреймворк з відкритим кодом для використання спільнотою, який інтегрує вимірювання стабільності у процеси розроблення моделей. **Висновки**. У цій роботі запропоновано нову парадигму групової справедливості, що об'єднує питання правильності / якості та випадковості / стабільності з порядку денного досліджень відповідального штучного інтелекту. Застосування запропонованих підходів допомагає у відповідальному виборі моделей машинного навчання за умов множинності моделей, демонструє, що, хоча може існувати чимало моделей із зіставною точністю, є лише одна (або декілька) "найкраща" модель, що є надійною, справедливою і стабільною, як це й потрібно.

Ключові слова: відповідальний штучний інтелект; алгоритмічна справедливість; стабільність моделей; експериментальний аналіз.

Вступ

Упродовж останнього десятиліття моделі машинного навчання (МН) ширше застосовуються в реальних умовах. Ці передові моделі тепер використовуються для розв'язання різноманітних складних завдань у багатьох секторах, таких як перевірка заявок на кредит, ухвалення рішень щодо застави й умовно-дострокового звільнення, а також генерування тексту, подібного до людського, на основі зображень і текстової інформації. Зі збільшенням застосування МН у соціальних сферах стало очевидним, що, попри те, що ці моделі навчаються робити точні прогнози на основі вхідних характеристик, вони також мимоволі засвоюють дискримінаційні шаблони, що призводить до "несправедливості" в їх результатах. Чимало досліджень показали, що підходи, основані на даних, можуть ненавмисно підсилювати наявні людські упередження [1–4]. Зрештою справедливість стала ключовим питанням у розвитку відповідального штучного інтелекту в останні роки.

Занепокоєння щодо групової справедливості в дослідженнях МН зосереджуються на питанні: "Чи є прогнози моделі однаково якісними в середньому для різних соціальних груп?" Один із способів визначити "однаково якісні" полягає у вимозі, щоб модель допускала порівняні помилки на вибірках з різних груп. Це природно сприяло розробленню різноманітних метрик справедливості в науковій літературі, зазвичай визначених як різниця або співвідношення міри помилки моделі (таких як рівень правильних позитивних результатів або рівень правильних негативних результатів), агрегованих за привілейованими та непривілейованими соціальними групами відповідно [5–11].

Натомість занепокоєння щодо безпеки в дослідженнях МН зосереджуються на технічних вимогах, таких як стабільність і надійність. Стабільність тут означає мінливість і випадковість у прогнозах моделі, і проводяться цікаві дослідження на перетині цих напрямів, що вивчають взаємозв'язок між справедливістю та випадковістю в ухваленні

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

рішень у критичних соціальних сферах крізь призму множинності моделей [12–14].

Нормативне обґрунтування

Для прикладу розглянемо ухвалення рішень щодо прийому до юридичних шкіл. Наявні визначення справедливості, такі як демографічний паритет і баланс показників помилок, засуджували б рішення комітету, яке або систематично обирає кандидатів з однієї соціальної групи (диспропорція позитивної ставки), або систематично відхиляє перспективних кандидатів із соціально неблагополучної групи (диспропорція негативної ставки).

Однак розглянемо, як ці визначення ігнорують несправедливість через випадковість. Припустимо, що в нас є правило прийняття рішень, яке задовольняє баланс помилок, але є дуже нестабільним. Це означає, що це правило іноді приймає значно різні рішення щодо прийому для схожих кандидатів. Навряд чи правило, чиї рішення нестабільні, буде дуже точним, тому звітність про точність може відтворювати загальну стабільність. Далі розглянемо правило з високою точністю, яке є також досить стабільним у середньому. Однак це правило дуже випадкове для деяких вибірок, можливо, відповідних певній меншості соціальної групи. У такій ситуації правило задовольняє критерії стабільності, але не критерії стабільності-паритету. З погляду процедурної справедливості це стає моральною проблемою, оскільки правило, очевидно, застосовує дві різні процедури – більш систематичну для представників більшості та більш випадкову для представників меншості.

Диспропорція в стабільності може також призвести до несправедливості результатів. В умовах прийому до юридичних шкіл розглянемо здатність абітурієнтів звертатися за переглядом рішення. Якщо абітурієнт звертається до комітету з проханням пояснити причину певного рішення, то (1) надійність пояснення залежить від стабільності правила – як надійно пояснити правило, чиї прогнози непослідовні? (2) ефективність перегляду залежить від диспропорції стабільності правила: перегляд буде менш ефективним для представників меншості, оскільки навіть якщо вони покращують свої характеристики (значення ознак), диспропорція випадковості правила робить невизначеним, чи отримають вони бажаний результат.

У цій роботі ґрунтуємося на позиції Кріла та Геллмана [15]. На їхню думку, "проблема з випадковими алгоритмами полягає не в їх випадковості як такій, а в систематичності цієї випадковості". Ми стверджуємо, що дисперсія або випадковість сама по собі не є моральною проблемою, але диспропорція у випадковості є моральним питанням.

Стислий опис результатів

1. Запропоновано нову парадигму групової справедливості в машинному навчанні, таку, що досліджує не лише паритет у рівнях помилок, а також паритет в упередженні, дисперсії та шуму, на які ці помилки розкладаються. Першим кроком у цьому напрямі обрано паритет-дисперсії для різних соціальних груп як новий критерій справедливості.

2. За допомогою цього нового підходу визначено компроміси між справедливістю, точністю та стабільністю, спираючись на широкомасштабне емпіричне дослідження на реальних і синтетичних даних (https://github.com/denysgerasymuk799/fairness-variance).

3. Продемонстровано, як цей аналіз допомагає практикам долати труднощі множинності моделей, інтегруючи додаткові критерії стабільності під час вибору моделі.

4. У межах цієї роботи створено відкритий програмний фреймворк, що передбачає вимірювання справедливості та стабільності у процесі розроблення моделей для використання спільнотою (https://github.com/denysgerasymuk799/model-profiler).

Аналіз проблеми й наявних методів

Множинність моделей – це явище, в умовах якого існує кілька моделей з функціонально еквівалентною точністю для певного завдання, але ці моделі розрізняються внутрішньою структурою (що називається процедурною множинністю) або прогнозами на конкретних вибірках (що називається предикативною множинністю). У роботі демонструємо, як розуміння компромісів між справедливістю, стабільністю і точністю допомагає впоратися з множинністю моделей: включення нових критеріїв стабільності та паритету стабільності в процедури вибору моделі допомагає розрізнити моделі, які мають однакову точність (див. рис. 1).

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Це дослідження конкурує з нещодавніми студіями Купера, Лонга та ін. [16, 17].

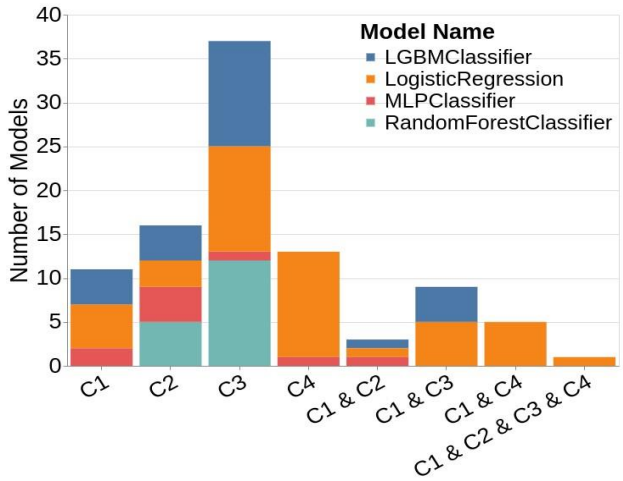


Рис. 1. Набір даних *Law school*: вибір моделей з найвищими показниками на основі декількох критеріїв. C1 – правильність (точність); C2 – паритет помилок (вирівняні шанси FPR); C3 – стабільність (*Label Stability*); C4 – паритет стабільності (співвідношення стабільності міток). З-поміж 84 натренованих моделей 11 мають бажану точність (C1), але лише одна модель відповідає всім чотирьом критеріям (C1&C2&C3&C4)

Так, Купер та ін. [16] досліджують справедливість за умов множинності моделей, розглядаючи розподіли можливих моделей замість окремих моделей, і пропонують нову міру випадковості, яка називається *самостійністю* (*self-consistency*). Їх експериментальна установка подібна до нашої: вони використовують простий підхід бутстрепу (*bootstrapping approach*) для вимірювання предикативної дисперсії. Важливо, що автори зосереджуються на методах зменшення випадковості (наприклад, за допомогою бегінгу (*bagging*)) та звітують про вплив на точність і справедливість. Це дослідження зосереджене на протилежній поведінці: як справедливість компрометується з точністю та стабільністю.

Лонг та ін. [17] розглядають групову справедливість у контексті стабільності. Крім того, як і наша робота, їхнє дослідження зосереджується на компромісах між справедливістю та стабільністю за умов множинності, пояснюючи наслідки як "вартість випадковості втручань для досягнення справедливості". Автори розглядають різноманітні метрики справедливості, тоді як ми обмежуємо наше обрамлення метриками паритету помилок, щоб дослідити загальні та специфічні для груп компроміси крізь призму розкладання помилки моделі на упередження та дисперсію. Отже, як фокус, так і висновки наших досліджень різняться.

Найважливіше, що обидві згадані роботи розглядають випадковість лише загалом, тоді як ми пропонуємо паритет стабільності як новий критерій справедливості. Крім того, і Купер, і Лонг та ін. [16, 17] пропонують процедури для зменшення випадковості під час розроблення моделей. Натомість ми зосереджуємося на включенні цілей стабільності в процесі вибору моделі для контекстів ухвалення рішень, де випадковість і диспропорція у випадковості мають моральне значення.

Технічне рішення

Для узгодження звітності про показники продуктивності, пов’язані зі справедливістю та стабільністю, ми створили бібліотеку *Python* спеціально для аудиту справедливості та стабільності моделей. Дизайн цієї бібліотеки керується трьома основними принципами:

1) спрощення розширення можливостей аналізу моделей;

2) забезпечення сумісності із широким спектром наборів даних і типів моделей;

3) надання можливості легкого створення метрик паритету відповідно до контексту використання.

Програмний каркас організовує аудит моделей у чіткі фази, зокрема *обчислення метрик підгруп*, *складання метрик для груп* та *візуалізацію метрик*. Такий структурований підхід забезпечує фахівцям з оброблення даних та практикам МН більший контроль і гнучкість, роблячи бібліотеку корисною як на етапі розроблення моделей, так і для моніторингу після їх впровадження.

Огляд архітектури

Рис. 2 ілюструє підхід бібліотеки до побудови конвеєра аналізу моделі. Вхідні змінні для інтерфейсу користувача позначені зеленим кольором, етапи конвеєра – синім, а вихід кожного етапу відтворюється фіолетовим. Далі надамо детальний опис кожного етапу.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

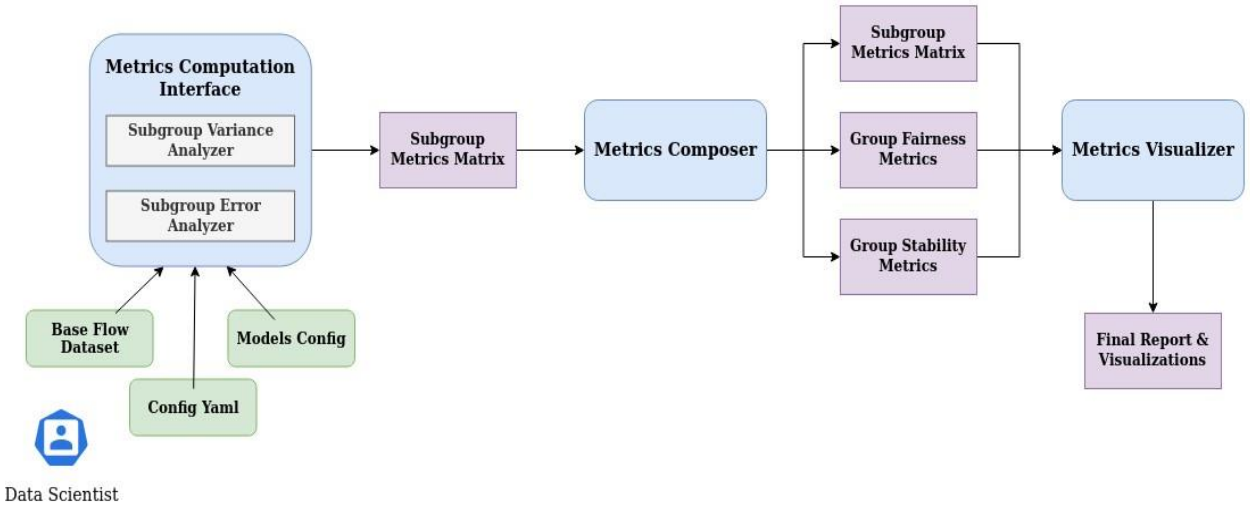


Рис. 2. Архітектура програмної бібліотеки. Вхідні змінні інтерфейсу користувача зображені зеленим кольором, етапи конвеєра – синім, а вихід кожного етапу – фіолетовим. Конвеєр аналізу передбачає три етапи оброблення: обчислення метрик підгруп, складання метрик груп і візуалізація та звітування метрик

Вхідні змінні. Для застосування бібліотеки користувачам необхідно надати три основні входи.

- **Основний набір даних (*Base flow dataset*).** Це спеціальний об'єкт, що є набором даних користувача, зокрема такі описові характеристики, як цільовий стовпець, числові та категоріальні стовпці, а також навчальні й тестові набори. Цей об'єкт походить від класу *BaseFlowDataset*, розробленого для зручності користувача. Застосування спільного базового класу дає змогу здійснити початкову валідацію введених даних користувача, спрощуючи логіку подальших обчислень метрик.
- **Файл конфігурації YAML (*Config YAML*).** Цей файл використовується з метою визначення параметрів конфігурації для інтерфейсів користувача фреймворку, які обробляють обчислення метрик. Даючи змогу користувачам визначати конфігурацію, цей підхід забезпечує більшу гнучкість, спрощуючи проведення експериментів. За допомогою одного файлу конфігурації YAML на експеримент користувачі можуть легко перемикатися між експериментами без додаткових змін в інтерфейсі користувача бібліотеки. Крім того, файл конфігурації є ефективним методом для вказівки складних налаштувань підгруп і об'єднання всіх параметрів, необхідних для аудиту моделей, в одному місці.
    - Конфігураційний файл YAML містить різні деталі, такі як кількість оцінювачів в ансамблі для аналізу дисперсії, частка вибірки для випадкового відбору з навчального набору, режим обчислення метрик та інші параметри. Важливим є те, що користувачі визначають підгрупи інтересу, просто передаючи словник із парами ключ-значення, де ключі є назвами стовпців, а значення відповідають невигідним значенням (або списку значень) для чутливого атрибута, уможливлюючи поділ підгруп. Такий підхід також дає змогу користувачам за потреби вказувати інтерсекційні групи.
- **Конфігурація моделей (*Models config*).** Цей словник *Python* пов'язує назви моделей з їх ініціалізованими екземплярами для аналізу. Такий підхід дає змогу проводити аудит кількох моделей і підтримує оцінювання різних типів моделей, надаючи гнучкість у їх комплексному оцінюванні.

Інтерфейси користувача

Обчислення метрик підгруп. Бібліотека надає кілька інтерфейсів користувача для обчислення метрик, зокрема інтерфейс для роботи з кількома моделями, інтерфейс для роботи з різними тестовими наборами та інтерфейс для збереження результатів у базі даних, визначеній користувачем. Після введення змінних в інтерфейс користувача основний набір даних обробляють аналізатори підгруп для отримання різноманітного набору метрик. Бібліотека використовує об'єктно орієнтовані шаблони проєктування, такі як *Factory Method, Abstract Factory, Facade* і *Template Method* [18]. Її склад містить *Subgroup Variance Analyzer* і *Subgroup Error Analyzer*, що узгоджуються з підходом розкладання помилки моделі, запропонованим

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

у праці [19], і розроблені таким чином, щоб їх було легко розширювати додатковими аналізаторами. Після завершення обчислення метрик виходи об'єднуються в *pandas dataframe*. Користувачі також мають змогу встановити параметр збереження, що сприяє збереженню *dataframe* метрик на диску або в базі даних за допомогою функції *df_writer*.

Subgroup Variance Analyzer відповідає за обчислення метрик стабільності як для всього тестового набору, так і для специфічних підгруп, визначених користувачем. Для визначення дисперсії оцінювача застосовується метод бутстрепінгу [20]. Окрім стандартного відхилення предикативних розподілів, також обчислюються додаткові метрики стабільності, такі як стабільність міток (*Label Stability*) [21], "тремтіння" (*Jitter*) [22] та IQR (міжквартильний розмах предикативної дисперсії). Аналогічно *Subgroup Error Analyzer* оцінює метрики помилок, зокрема точність, TPR (рівень правильних позитивних результатів), FPR (рівень хибних позитивних результатів), TNR (рівень правильних негативних результатів) і FNR (рівень хибних негативних результатів) – як для всього тестового набору, так і для зазначених підгруп інтересу.

Складання метрик груп. *Metrics Composer* керує другим етапом аудиту моделі, обчислюючи групові метрики справедливості та стабільності. Користувачі також можуть налаштовувати додаткові метрики за потреби. Наприклад, диспропорційний вплив, показник справедливості, обчислюється способом визначення співвідношення позитивних результатів для привілейованих (*priv*) і непривілейованих (*dis*) підгруп.

Візуалізація метрик. *Metrics Visualizer* об'єднує кілька етапів оброблення для складених метрик, створюючи формати даних, оптимізовані для візуалізації. Упроваджуючи методи цього класу, користувачі можуть створювати індивідуальні графіки, адаптовані до їхнього аналізу. Ці візуалізації також можна зібрати в інтерактивний звіт, що сприяє відповідальному вибору моделей.

Емпіричне оцінювання

Деталі експериментів

Набори даних

Досліджуємо компроміси між точністю, стабільністю та справедливістю (визначеними різними способами) на чотирьох наборах даних реального світу й завданнях, підсумованих у табл. 1. Ми свідомо обрали набори даних та завдання з контекстів політики, де паритет стабільності є моральною необхідністю.

Таблиця 1. *Опис наборів даних і завдань*

Набір даних	Галузь	Опис завдання	Групи
ACS Income	фінанси	передбачити, чи дохід людини перевищує національний медіанний дохід	стать, раса
ACS Public Coverage	охорона здоров'я	передбачити, чи має особа державне медичне страхування	стать, раса
Law school	освіта	передбачити, чи складе кандидат іспит на адвокатську практику	раса
Ricci v. De Stefano	юриспруденція	передбачити, чи отримає особа підвищення на основі результатів іспиту	раса

Для дослідження використали чотири популярні набори даних з метою аналізу справедливості, детально описані в огляді [26]. Нижче стисло подано кожний з них.

- ***Folktables*** **(**https://github.com/zykls/folktables). Цей еталонний набір є збіркою сучасних даних, отриманих з опитувань Бюро перепису населення США з 50 штатів за 2014–2018 рр. Він охоплює кілька завдань прогнозування, пов'язаних із доходом, охороною здоров'я, транспортом, зайнятістю та житлом. Ми обрали два з них (перелічені нижче). Для аналізу групової справедливості формуємо групи на основі двох чутливих атрибутів: статі (бінаризованої на "жінок" (незахищена група) і "чоловіків" (привілейована група)) та раси (бінаризованої на "небілих" (незахищена група) і "білих" (привілейована група)). Для нашого дослідження ми звітуємо про метрики диспропорцій щодо інтерсекційних груп, а саме "небілих жінок" (інтерсекційно незахищена група) порівняно з усіма іншими зразками.
    - ***ACS Income***. Це завдання бінарної класифікації для прогнозування: чи є дохід особи меншим за $50,000 (мітка 0), або перевищує, або дорівнює $50,000 (мітка 1). Цей набір даних містить вісім категоріальних і дві числові ознаки, зокрема освіту, професію та сімейний стан. Для нашого дослідження використано дані штату Джорджія

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

за 2018 р. і зменшено вибірку з 50,915 до 15,000 рядків для обчислювальної зручності.

- ***ACS Public Coverage***. Це завдання бінарної класифікації для прогнозування: чи має особа державне медичне страхування (мітка 1), чи ні (мітка 0). Набір даних містить 17 категоріальних і дві числові ознаки, зокрема освіту, інвалідність та сімейний стан. Для нашого дослідження застосовано показники штату Каліфорнія за 2018 р. і зменшено вибірку з 138,554 до 15,000 рядків для скорочення обчислювальних вимог.

- ***Law School*** (https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/law_school_clean.csv). Цей набір даних був зібраний за допомогою опитування, проведеного Радою з прийому до юридичних шкіл (LSAC) у 163 юридичних школах США 1991 р. Він містить записи про прийом до юридичних шкіл. Набір даних містить інформацію про 20,798 студентів, схарактеризованих 12 ознаками (три категоріальні, три бінарні та шість числових ознак). Завдання прогнозування полягає у визначенні: чи складе кандидат іспит на право (мітка 1), чи ні (мітка 0). Чутливі атрибути в цьому наборі даних – стать студента та його раса. Для нашого дослідження використовуємо повний набір даних і звітуємо про метрики диспропорції для груп, сформованих за расовою ознакою; "білі" (привілейовані) проти "небілих" (незахищені).

- ***Ricci*** (https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/ricci_race.csv). Цей набір даних походить з історичної справи *Ricci v. DeStefano*, у якій досліджували результати іспиту на підвищення в пожежній службі в листопаді 2003 р. Зазначений набір містить 118 результатів іспитів, де кожен зразок визначається шістьма ознаками (три числові та три бінарні ознаки). Хоча набір даних є відносно невеликим, він широко використовується у дослідженнях, спрямованих на забезпечення справедливості в класифікації. Завдання бінарної класифікації полягає в прогнозуванні: чи отримає особа підвищення на основі результатів іспиту. Єдиний чутливий атрибут – раса, що має три унікальні значення: чорний, білий і латиноамериканець. Відповідно до попередніх досліджень ми об'єднуємо "чорних" і "латиноамериканців" у "небілу" захищену групу та звітуємо про результати диспропорції для бінарних груп, сформованих за расовою ознакою. Оскільки набір даних дуже незначний, змінюємо наше співвідношення поділу на навчальну і тестову вибірки з 80:20 на 67:33.

Навчання та оцінювання моделі

Приймаючи орієнтований на дані підхід до стабільності, прагнемо виявити емпіричні компроміси між справедливістю, стабільністю та точністю, тому поєднуємо нашу систему оцінювання стабільності з інтервенцією передоброблення для справедливості, орієнтованої на дані. Зокрема застосовуємо *DisparateImpactRemover* – передову інтервенцію передоброблення, спрямовану на усунення дискримінаційних шаблонів у даних, реалізовану в інструментарії IBM AIF 360 [23].

Один експериментальний запуск відбувається таким чином: спочатку у випадковий спосіб ділимо набір даних на навчальний і тестовий набори (співвідношення 80:20). Потім ініціалізуємо процесор справедливості, тренуємо його на навчальному наборі й трансформуємо як навчальний, так і тестовий набори. Далі ініціалізуємо базову модель H і налаштовуємо гіперпараметри один раз для кожної пари рівень виправлення / тип моделі. Після цього запускаємо процедуру бутстрепу для створення наближеного ансамблю з $m = 200$ учасниками, до того ж кожна модель навчається з оптимізованими налаштуваннями гіперпараметрів базової моделі H. Повторюємо цю процедуру шість разів, щоразу з іншим випадковим поділом навчальний: тестовий набір, для кожної пари рівень виправлення / тип моделі.

Опис експериментів

Проведено два експерименти для кожного набору даних і завдання.

Експеримент 1. Одна модель, багато рівнів виправлення. Фіксуємо архітектуру моделі та змінюємо рівень виправлення α методу передоброблення для справедливості. Використано *Random Forest* як тип моделі для цього експерименту та тестували для $\alpha \in [0,1]$ з кроком 0,1. У цьому експерименті виявляємо компроміси між точністю та справедливістю (рис. 4), компроміси між стабільністю та справедливістю (рис. 5) і паритет у компромісах між стабільністю та справедливістю (рис. 3).

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Експеримент 2. Один рівень виправлення, багато моделей. Цей експеримент доповнює орієнтований на дані підхід експерименту 1 моделецентричним поглядом на стабільність. З огляду на експеримент 1 обираємо одне або два відповідні значення рівня виправлення справедливості α та навчаємо різні моделі для того самого рівня виправлення. Ми порівнюємо продуктивність лінійних моделей, моделей на основі дерев, ансамблевих моделей і нейронних мереж з бібліотеки *scikit-learn*.

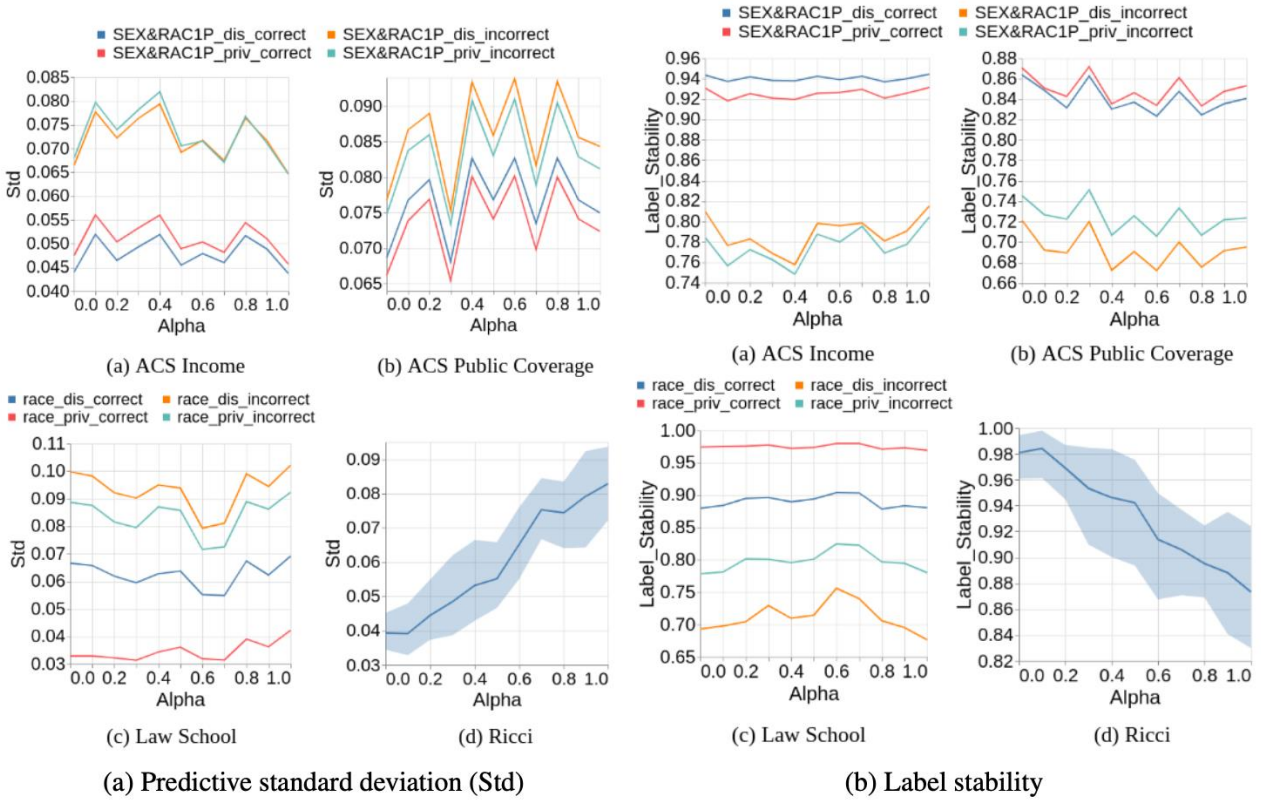


Рис. 3. Компроміси між справедливістю та стабільністю: метрика стабільності на осі y проти рівня справедливості (α) на осі x. Вищі значення *Std* і нижчі значення стабільності міток вказують на більшу нестабільність. Вище значення α означає сильнішу інтервенцію для забезпечення справедливості

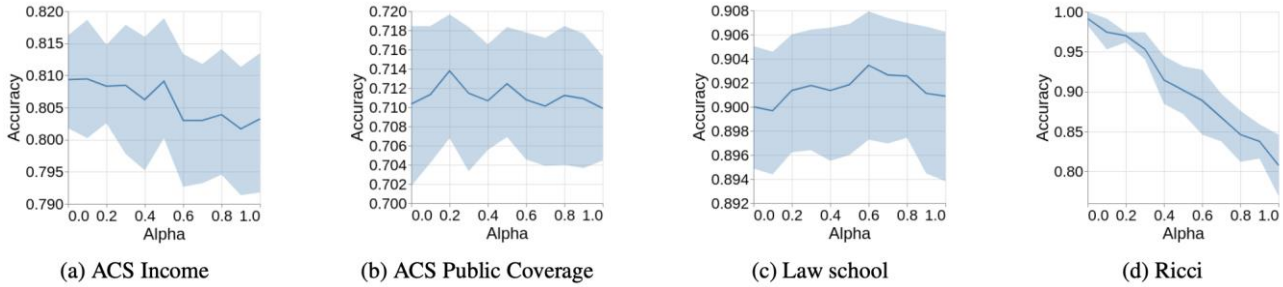


Рис. 4. Компроміси між точністю та справедливістю, де рівень справедливості α зображено на осі x, а точність – на осі y. У трьох із чотирьох наборів даних спостерігається незначний вплив на точність за умови зростання справедливості, що спричиняє множинність моделей

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

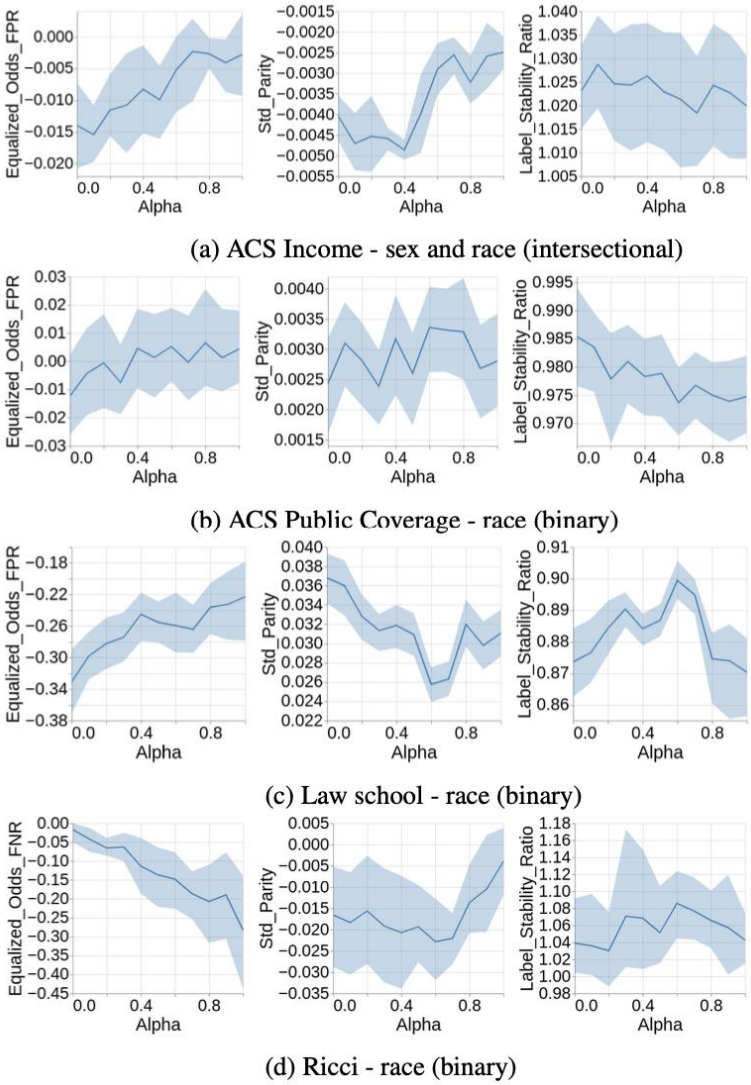


Рис. 5. Тренди для метрик паритету. Вісь y : (ліворуч) паритет у рівнях помилок, значення, близьке до 0, є бажаним; (центр) паритет у стандартному відхиленні прогнозів, значення, близьке до 0, є бажаним; (праворуч) коефіцієнт стабільності міток, значення, близьке до 1, є бажаним. Вісь x : рівень справедливості (α), де більше значення вказує на сильнішу інтервенцію для забезпечення справедливості

Метрики продуктивності моделі

Кількісно оцінюємо чотири аспекти продуктивності моделі, використовуючи широкий набір метрик.

- *Правильність*. Звітуємо про точність моделі як міру правильності її прогнозів.
- *Стабільність*. Звітуємо про стандартне відхилення прогнозів (квадратний корінь з дисперсії) та стабільність міток, усереднених по всіх вибірках, як міру нестабільності прогнозів моделі. Середнє значення *Std* 0,1, як зображено на рис. 3, означає, що основний прогноз 0,4, наприклад, може варіюватися між 0,3 і 0,5 через невеликі відмінності в навчальних даних. Аналогічно середня стабільність міток 0,8 означає, що очікується зміна прогнозованої мітки моделі у 20 % випадків за незначних змін у навчальних даних.
- *Паритет помилок* (класичний аспект справедливості). Звітуємо про диспропорцію в рівнях хибнопозитивних і хибнонегативних результатів для непривілейованих і привілейованих груп відповідно, як міру несправедливості моделі (класично визначену).
- *Паритет стабільності* (новий аспект справедливості). Також звітуємо про дві міри диспропорції, створені на основі наших показників стабільності, а саме паритет *Std* – різницю між

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

середнім значенням стандартного відхилення прогнозів для вибірок з непривілейованої та привілейованої груп відповідно, і коефіцієнт стабільності міток – співвідношення середньої стабільності міток для вибірок з непривілейованої та привілейованої груп відповідно. Ці метрики кількісно оцінюють диспропорцію у випадковості, яку пропонуємо як новий критерій справедливості. Диспропорція *Std*, що дорівнює 0, і коефіцієнт стабільності міток, що дорівнює 1, означають, що модель є рівномірно випадковою для різних соціальних груп у даних.

Експеримент 1. Справедливість – стабільність – точність

Точність. На рис. 4 продемонстровано графік залежності точності від рівня справедливості. Для трьох із чотирьох оцінених нами наборів даних вплив попереднього оброблення для справедливості на точність є незначним. Це означає, що ми отримуємо моделі, які порівняні за точністю, але розрізняються іншими характеристиками, такими як справедливість – саме це і є умовою множинності моделей. Винятком є набір даних *Ricci v. De Stefano*, де спостерігаємо класичний компроміс між справедливістю та точністю: підвищення рівня виправлення справедливості α погіршує точність. На цьому наборі даних ми не спостерігаємо множинності моделей, оскільки моделі, які є порівняно справедливими, мають нижчу точність.

Стабільність. Далі розглядаємо, як загальна стабільність залежить від попереднього оброблення для справедливості. Звітуємо про стандартне відхилення прогнозів і стабільність міток на рис. 3 для різних рівнів справедливості (α), а також розбиваємо за демографічними групами (привілейовані або непривілейовані) і типами помилок (правильні або неправильні). Набір даних *Ricci v. De Stefano* є занадто малим для такого поділу, тому для цього набору звітуємо лише про загальні тенденції.

Експериментальні результати демонструють, що стабільність моделі залежить від попереднього оброблення для справедливості, але тренд (погіршується або покращується) є непостійним у різних наборах даних. Цікаво, що модель завжди більш стабільна (нижча, ніж *Std*, і вища, ніж *Label Stability*) на правильно класифікованих вибірках. Далі, і в межах типів помилок, модель є більш стабільною для привілейованих, ніж для непривілейованих груп у наборах даних *ACS Public Coverage* та *Law school*. Ця тенденція для демографічних груп змінюється для *ACS Income*, де модель, як не дивно, є більш стабільною на непривілейованій групі, ніж на привілейованій.

Набір даних *Law school* показує найбільшу диспропорцію в стабільності для різних пар груп і типів помилок, тоді як *ACS Public Coverage* показує найбільшу варіацію стабільності для різних значень α.

Метрики паритету. У цій роботі приймаємо погляд на справедливість для груп і звітуємо про диспропорцію здебільшого для вибірок з непривілейованої групи порівняно з привілейованою групою. Класично диспропорція в рівнях помилок вважалася метрикою справедливості. Звітуємо про це на рис. 5 разом з нашими новими метриками паритету стабільності, такими як паритет *Std* і коефіцієнт стабільності міток (*Label Stability*). Як очікувалося, для всіх наборів даних підвищення рівня справедливості зменшує диспропорцію в рівнях помилок (класична справедливість). Для *Ricci v. De Stefano* неконтрольована модель досягає ідеальної точності, тому модель покращує справедливість унаслідок погіршення показників для привілейованої групи, а не завдяки покращенню показників для непривілейованої групи. Отже, застосування попереднього оброблення для забезпечення справедливості до цього набору даних призводить до негативної дискримінації, де диспропорція в помилках насправді зростає для вищих рівнів справедливості на користь непривілейованої групи.

Для *ACS Income* та *Law school* – наборів даних, на яких спостерігали множинність моделей, – підвищення рівнів справедливості покращує паритет стабільності (паритет *Std* наближається до 0, коефіцієнт стабільності міток наближається до 1). Для *ACS Public Coverage* паритет *Std* коливається навколо значення за умови $\alpha = 0$ (без втручання для забезпечення справедливості), тоді як коефіцієнт стабільності міток незначно погіршується (віддаляється від 1 зі збільшенням α).

У підсумку в нашому емпіричному аналізі зауважуємо певні тенденції.

1. Точність і стабільність зазвичай змінюються разом. Ми спостерігаємо незначний вплив на точність і незначний вплив на стабільність у *ACS Income* та *Law school*. Коли втручання для забезпечення справедливості впливають на точність, вони також позначаються на стабільності, як видно з *Ricci v. De Stefano*.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

2. За умов множинності, якщо вплив втручання для забезпечення справедливості на точність і стабільність не значний, то покращення справедливості (паритет у помилках) також покращує паритет стабільності, як видно з *ACS Income* та *Law school*.

3. Якщо точність залишається незмінною (також за умов множинності), але стабільність змінюється, то підвищення справедливості (паритет у помилках) не покращує паритет стабільності, як видно з *ACS Public Coverage*, і в цьому разі це погіршує (збільшує) диспропорцію стабільності.

4. Нарешті, якщо вплив на точність і стабільність є значним (тобто підвищення рівня справедливості погіршує і точність, і стабільність), то досягнення справедливості (покращення паритету в помилках) не покращує паритет стабільності, як видно з *Ricci*.

Експеримент 2. Модель – центричний підхід до стабільності

Далі досліджуємо, як втручання для забезпечення справедливості впливають на стабільність різних типів моделей. На рис. 6 наведені результати цього експерименту, де ми обираємо одне або два відповідні значення рівня справедливості (α), навчаємо різні моделі для одного й того самого рівня та порівнюємо їх продуктивність. Нейронні мережі демонструють найбільший вплив втручань для забезпечення справедливості на стабільність. Компроміс не є однозначним: для двох наборів даних стабільність покращується за умови вищих значень α, для інших двох – погіршується, але обидва ефекти є більшими, ніж для інших типів моделей.

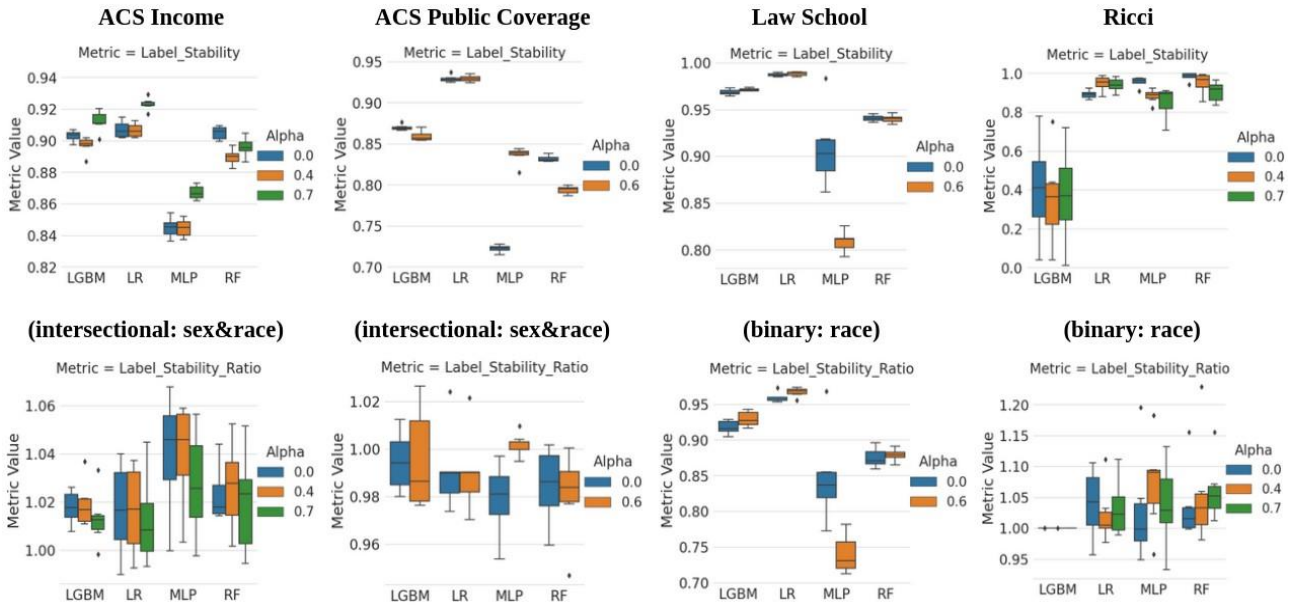


Рис. 6. Компроміси стабільності, орієнтовані на моделі. Вісь *x*: тип моделі, де LGBM – *LightGradientBoostingMachine*; LR – *LogisticRegression*; MLP – *MLPClassifier*; RF – *RandomForest*. Вісь *y*: стабільність міток (верхній ряд) і коефіцієнт стабільності міток (нижній ряд). Різні кольори відповідають рівням справедливості, де $\alpha = 0$ означає відсутність втручання для забезпечення справедливості

Нейронні мережі також мають найнижчу стабільність міток серед усіх моделей на всіх наборах даних, за винятком LGBM на наборі *Ricci*, ймовірно, через те, що ми тренуємо ансамблевий бустинг-класифікатор на дуже малому наборі даних (118 зразків). Нейронні мережі також демонструють найгірший паритет стабільності на більшості завдань порівняно з усіма оціненими типами моделей. Лінійні моделі (LR) є найбільш стабільними на всіх завданнях.

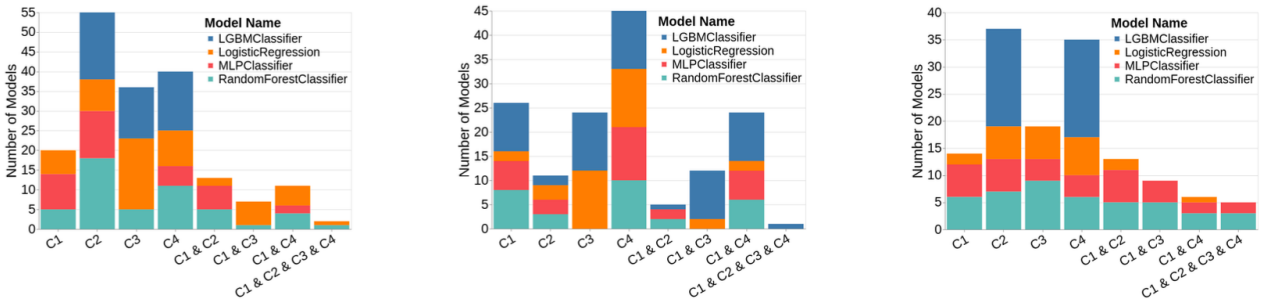
Цікаво, що на наборі даних, який не призводить до множинності – *Ricci*, більшість моделей є досить стабільними. Однак на наборах даних, де втручання для забезпечення справедливості призводять до множинності – *ACS Income*, *ACS Public Coverage* та *Law School*, спостерігаємо компроміси в стабільності, специфічні для моделей. З огляду на наш попередній результат про те, що втрати стабільності співвідносяться із втратами точності, це означає, що коли стабільність і точність погіршуються разом (як результат втручання для забезпечення справедливості), ефект не залежить від типу моделі. Однак, коли стабільність погіршується,

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

а точність залишається незмінною (як результат втручання для забезпечення справедливості), то ефект є високоспецифічним для кожної моделі.

Відповідальний вибір моделі

Нарешті, розглянемо, як включення метрик стабільності (як загальних, так і з огляду на диспропорції між групами) допомагає покращити процедури вибору моделей. Для цього моделюємо процес вибору моделі, порівнюючи всі натреновані моделі (для різних рівнів α в експерименті 1 та різних типів моделей в експерименті 2, із шістьма випадковими запусканнями для кожного налаштування). Ми порівнюємо приблизно 90 моделей для кожного набору даних (на основі різних виборів для експерименту 2). На рис. 7 продемонстровано кількість "найкращих" моделей для кожного критерію – точності, стабільності, паритету в помилках (класична метрика справедливості) та паритету в стабільності (наша нова метрика справедливості). Спочатку застосовуємо ці обмеження окремо, а потім – у поєднанні. Ми демонструємо, що для трьох завдань, де спостерігалася множинність після застосування інтервенції з метою забезпечення справедливості, існує кілька моделей з порівняною точністю; однак лише одна (або дві) серед них є одночасно точною, стабільною та справедливою щодо наявності хорошого паритету як у помилках, так і в стабільності між групами. Цікаво, що вибір моделі для набору даних *Ricci* не сприяє вибору єдиної найкращої моделі. Нагадаємо з огляду на рис. 5, що *Ricci* був єдиним набором даних у нашому дослідженні, на якому попереднє оброблення для забезпечення справедливості не викликала множинності, а, навпаки, погіршувала точність. Це підтверджує ідею, що множинність є можливістю, а не викликом [12].



(a) ACS Income (b) ACS Public Coverage (c) Ricci

Рис. 7. Відповідальний вибір моделі: включення нормативно бажаних цілей щодо справедливості та стабільності в процедури вибору моделі розв'язує проблему множинності моделей. Визначення критеріїв: C1 – правильність (точність); C2 – паритет помилок (рівень хибнопозитивних результатів); C3 – стабільність (стабільність міток); C4 – паритет стабільності (коефіцієнт стабільності міток)

На рис. 7 також звітуємо про кількість відібраних моделей для кожного типу моделі. Моделі на основі дерев і лінійні моделі вважаються найсучаснішими за точністю для табличних даних, особливо у сферах, де справедливість є важливим фактором. Наше дослідження підтверджує сказане, додатково показуючи, що ці типи моделей також найкраще відповідають критеріям стабільності.

Висновки

Створено програмний фреймворк, що інтегрує вимірювання справедливості та стабільності в конвеєр розроблення моделей. Запропоновано нову парадигму групової справедливості, яка об'єднує питання правильності / якості та випадковості / стабільності з порядку денного досліджень відповідального штучного інтелекту. Звернено увагу на відомі теореми про розкладання похибки класифікатора на упередження та дисперсії, щоб обґрунтувати новий критерій групової справедливості, що розглядає паритет у стабільності, і досліджено, як це співвідноситься з точністю, стабільністю та паритетом помилок під час втручань для забезпечення справедливості. Застосовано ці висновки для розроблення методу відповідального вибору моделей за умов множинності моделей, продемонстровано, що, хоча може існувати чимало моделей із зіставною точністю, є лише одна (або декілька) "найкраща" модель, яка є надійною, справедливою і стабільною, як це й потрібно.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

Перспективи подальших досліджень

У роботі впроваджено простий підхід бутстрепу для вимірювання стабільності моделі. Це передбачає тренування 200 моделей на бутстреп-вибірках навчального набору та використання їх для наближення стабільності однієї моделі. Хоча обрано цю процедуру завдяки її простоті та надійності, вона є дуже ресурсомісткою. Альтернативні підходи для надійного та ефективного вимірювання стабільності могли б посилити практичну значущість запропонованої процедури вибору моделей. Крім того, у дослідженні застосовано лише один метод попереднього оброблення для забезпечення справедливості. Було б цікаво розширити аналіз на інші втручання для забезпечення справедливості та дослідити, як інші властивості, зокрема невизначеність, співвідносяться зі стабільністю та справедливістю за умов множинності моделей.

Список літератури

1. Man is to computer programmer as woman is to homemaker? debiasing word embeddings / T. Bolukbasi et al. *Advances in neural information processing systems.* 2016. No. 29. P. 4356–4364. DOI: https://doi.org/10.48550/arXiv.1607.06520
2. Buolamwini J., Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency. 2018. P. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html
3. Caliskan A., Bryson J. J., Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. 2017. Vol. 356. P. 183–186. DOI: https://doi.org/10.1126/science.aal4230
4. Sweeney L. Discrimination in online ad delivery. *Communications of the ACM*. 2013. Vol. 56. No. 5. P. 44–54. DOI: https://doi.org/10.1145/2447976.2447990
5. Calders T., Verwer S. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery.* 2010. Vol. 21. No. 2. P. 277–292. DOI: https://doi.org/10.1007/s10618-010-0190-x
6. Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*. 2017. Vol. 5. No. 2. P. 153–163. DOI: https://doi.org/10.1089/big.2016.0047
7. Preserving statistical validity in adaptive data analysis / C. Dwork et al. *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015. P. 117–126. DOI: https://doi.org/10.1145/2746539.2746580
8. Fairness through awareness / C. Dwork et al. *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012. P. 214–226. DOI: https://doi.org/10.1145/2090236.2090255
9. Certifying and Removing Disparate Impact / M. Feldman et al. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015. P. 259–268. DOI: https://doi.org/10.1145/2783258.2783311
10. Kamishima T., Akaho S., Sakuma J. Fairness-aware Learning through Regularization Approach. *IEEE 11th International Conference on Data Mining Workshops*. 2011. P. 643–650. DOI: https://doi.org/10.1109/icdmw.2011.83
11. Kleinberg J. M., Mullainathan S., Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Innovations in Theoretical Computer Science Conference.* 2017. Vol. 67. P. 43:1–43:23. DOI: https://doi.org/10.4230/LIPIcs.ITCS.2017.43
12. Black E., Raghavan M., Barocas S. Model Multiplicity: Opportunities, Concerns, and Solutions. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 2022. P. 850–863. DOI: https://doi.org/10.1145/3531146.3533149
13. Underspecification presents challenges for credibility in modern machine learning / A. D'Amour et al. *The Journal of Machine Learning Research*. 2022. Vol. 23. No. 1. P. 10237–10297. DOI: https://doi.org/10.48550/arXiv.2011.03395
14. Marx C., Calmon F., Ustun B. Predictive multiplicity in classification. *International Conference on Machine Learning*. 2020. P. 6765–6774. URL: https://proceedings.mlr.press/v119/marx20a.html
15. Creel K., Hellman D. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic decision making systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021. 816 p. DOI: https://doi.org/10.1145/3442188.3445942
16. Arbitrariness and social prediction: The confounding role of variance in fair classification / A. F. Cooper et al. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024. Vol. 38. No. 20. P. 22004-22012. DOI: https://doi.org/10.1609/aaai.v38i20.30203
17. Arbitrariness Lies Beyond the Fairness-Accuracy Frontier / C. Long et al. *arXiv preprint arXiv:2306.09425*. 2023. DOI: https://doi.org/10.48550/arXiv.2306.09425
18. Shvets A. Dive into design patterns. Refactoring Guru. 2018. P. 22–29.

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

19. Domingos P. A unified bias-variance decomposition. *Proceedings of 17th international conference on machine learning.* 2000. P. 231–238. URL: https://www.scirp.org/reference/referencespapers?referenceid=2848771
20. Efron B., Tibshirani R. J. An introduction to the bootstrap. *CRC press*. 1994.
21. Darling M. C., Stracuzzi D. J. Toward Uncertainty Quantification for Supervised Classification. 2018. *OSTI.GOV*. DOI: https://doi.org/10.2172/1527311
22. Model Stability with Continuous Data Updates / H. Liu et al. *arXiv preprint arXiv:2201.05692*. 2022. DOI: https://doi.org/10.48550/arXiv.2201.05692
23. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias / R. K. E. Bellamy et al. *IBM Journal of Research and Development*. 2019. Vol. 63. No. 4/5. P. 4:1–4:15. DOI: https://doi.org/10.1147/jrd.2019.2942287
24. Certifying and removing disparate impact / M. Feldman et al. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015. P. 259–268. DOI: https://doi.org/10.1145/2783258.2783311
25. Preserving statistical validity in adaptive data analysis / C. Dwork et al. *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015. P. 117–126. DOI: https://doi.org/10.1145/2746539.2746580
26. A survey on datasets for fairness-aware machine learning / T. Le Quy et al. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2022. Vol. 12. No. 3. DOI: https://doi.org/10.1002/widm.1452

References

1. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., (2016), "Man is to computer programmer as woman is to homemaker? debiasing word embeddings", *Advances in neural information processing systems*, 29, P. 4356-4364. DOI: https://doi.org/10.48550/arXiv.1607.06520
2. Buolamwini, J. and Gebru, T., (2018), "Gender shades: Intersectional accuracy disparities in commercial gender classification", *In Conference on fairness, accountability and transparency* P. 77-91, PMLR. available at: https://proceedings.mlr.press/v81/buolamwini18a.html
3. Caliskan, A., Bryson, J.J. and Narayanan, A., (2017), "Semantics derived automatically from language corpora contain human-like biases", *Science*, 356(6334), P.183-186. DOI: https://doi.org/10.1126/science.aal4230
4. Sweeney, L., (2013), "Discrimination in online ad delivery", *Communications of the ACM*, 56(5), P.44-54. DOI: https://doi.org/10.1145/2447976.2447990
5. Calders, T., and Verwer, S. (2010), "Three Naive Bayes Approaches for Discrimination-Free Classification", *Data Min. Knowl. Discov.*, 21(2). P. 277–292. DOI: https://doi.org/10.1007/s10618-010-0190-x
6. Chouldechova, A. (2017), "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments", *Big Data*, 5(2). P. 153–163. DOI: https://doi.org/10.1089/big.2016.0047
7. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. (2015), "Preserving statistical validity in adaptive data analysis", *In Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, P. 117–126. DOI: https://doi.org/10.1145/2746539.2746580
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012), "Fairness through awareness", In Goldwasser, S., ed., *Innovations in Theoretical Computer Science 2012*, Cambridge, MA, USA, January 810, 2012, P. 214–226, ACM. DOI: https://doi.org/10.1145/2090236.2090255
9. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015), "Certifying and Removing Disparate Impact", In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, P. 259–268. New York, NY, USA: Association for Computing Machinery, ISBN 9781450336642. DOI: https://doi.org/10.1145/2783258.2783311
10. Kamishima, T., Akaho, S., and Sakuma, J. (2011), "Fairness-aware Learning through Regularization Approach", In *2011 IEEE 11th International Conference on Data Mining Workshops*, P. 643–650. DOI: https://doi.org/10.1109/icdmw.2011.83
11. Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017), "Inherent Trade-Offs in the Fair Determination of Risk Scores", In *Papadimitriou, C. H., ed., 8th Innovations in Theoretical Computer Science Conference, ITCS 2017*, January 9-11, 2017, Berkeley, CA, USA, Volume 67 of LIPIcs, P. 43:1–43:23, Schloss Dagstuhl Leibniz-Zentrum fur Informatik. DOI: https://doi.org/10.4230/LIPIcs.ITCS.2017.43

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)*

12. Black, E., Raghavan, M., and Barocas, S. (2022), "Model Multiplicity: Opportunities, Concerns, and Solutions", In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, P. 850–863. New York, NY, USA: Association for Computing Machinery, ISBN 9781450393522. DOI: https://doi.org/10.1145/3531146.3533149
13. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. 2022, "Underspecification presents challenges for credibility in modern machine learning", *The Journal of Machine Learning Research*, 23(1). P. 10237–10297. DOI: https://doi.org/10.48550/arXiv.2011.03395
14. Marx, C., Calmon, F., and Ustun, B. (2020), "Predictive multiplicity in classification", In *International Conference on Machine Learning*, P. 6765–6774, PMLR. available at: https://proceedings.mlr.press/v119/marx20a.html
15. Creel, K. and Hellman, D. (2021), "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic decision making Systems", In *Elish, M. C.; Isaac, W.; and Zemel, R. S., eds., FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event / Toronto, Canada, March 3-10, 816 p. ACM. DOI: https://doi.org/10.1145/3442188.3445942
16. Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., De Sa, C., Grimmelmann, J., Kleinberg, J., Sen, S., and Zhang, B. (2024), "Arbitrariness and social prediction: The confounding role of variance in fair classification", In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 20, P. 22004-22012. DOI: https://doi.org/10.1609/aaai.v38i20.30203
17. Long, C. X., Hsu, H., Alghamdi, W., and Calmon, F. P. (2023), "Arbitrariness Lies Beyond the Fairness-Accuracy Frontier", *arXiv preprint arXiv:2306.09425*. DOI: https://doi.org/10.48550/arXiv.2306.09425
18. Shvets, A. (2018), *Dive into design patterns*, Refactoring, Guru, P.22-29.
19. Domingos, P. (2000), "A unified bias-variance decomposition", In *Proceedings of 17th international conference on machine learning*. P. 231–238, Morgan Kaufmann Stanford. available at: https://www.scirp.org/reference/referencespapers?referenceid=2848771
20. Efron, B. and Tibshirani, R. J. (1994), "An introduction to the bootstrap", *CRC press*.
21. Darling, M. C. and Stracuzzi, D. J. (2018), "Toward Uncertainty Quantification for Supervised Classification". *OSTI.GOV*. DOI: https://doi.org/10.2172/1527311
22. Liu, H., Patwardhan, S., Grasch, P., Agarwal, S., et al. (2022), "Model Stability with Continuous Data Updates", *arXiv preprint arXiv:2201.05692*. DOI: https://doi.org/10.48550/arXiv.2201.05692
23. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2019), "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias", *IBM J. Res. Dev.*, 63(4/5). P. 4:1–4:15. DOI: https://doi.org/10.1147/jrd.2019.2942287
24. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015), "Certifying and removing disparate impact", In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, P. 259–268. DOI: https://doi.org/10.1145/2783258.2783311
25. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. (2015), "Preserving statistical validity in adaptive data analysis", In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, P. 117–126. DOI: https://doi.org/10.1145/2746539.2746580
26. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. (2022), "A survey on datasets for fairness-aware machine learning", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3): e1452. DOI: https://doi.org/10.1002/widm.1452

Надійшла (Received) 03.01.2025

Відомості про авторів / About the Authors

Герасимук Денис Вікторович – Харківський національний економічний університет імені Семена Кузнеця, магістр кафедри інформаційних систем, Харків, Україна; e-mail: Herasymuk.Denys@m.hneu.edu.ua; ORCID ID: https://orcid.org/0009-0009-3997-5458

Поляков Андрій Олександрович – кандидат технічних наук, доцент, Харківський національний економічний університет імені Семена Кузнеця, доцент кафедри інформаційних систем; Харківський національний університет радіоелектроніки, доцент кафедри прикладної математики, Харків, Україна; e-mail: Andrii.Poliakov@m.hneu.edu.ua; Andrii.Poliakov@nure.ua; ORCID ID: https://orcid.org/0000-0003-1805-9011

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Федорченко Володимир Миколайович – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри електронних обчислювальних машин, Харків, Україна; e-mail: volodymyr.fedorchenko@nure.ua; ORCID ID: http://orcid.org/0000-0001-7359-1460

Herasymuk Denys – Simon Kuznets Kharkiv National University of Economics, Master's Student at the Department of Information Systems, Kharkiv, Ukraine.

Poliakov Andrii – PhD (Engineering Sciences), Associate Professor, Simon Kuznets Kharkiv National University of Economics, Associated Professor at the Department of Information Systems; Kharkiv National University of Radio Electronics, Associated Professor at the Department of Applied Mathematics, Kharkiv, Ukraine.

Fedorchenko Volodymyr – PhD (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Electronic Computers, Kharkiv, Ukraine.

DETECTING TRADE-OFFS BETWEEN FAIRNESS, STABILITY, AND ACCURACY FOR RESPONSIBLE MACHINE LEARNING MODEL SELECTION

The subject of the study in the article is the process of machine learning model selection performed by data scientists to build models in critical areas. The purpose of the work: 1) create a software library for measuring the accuracy, stability, and fairness of models; 2) conduct experiments to identify trade-offs between fairness, stability, and accuracy; 3) propose a responsible model selection process to improve the safety of using machine learning models. The article provides for the following tasks: to measure the fairness and stability of machine learning models and to investigate their relationship with the accuracy of models. The following methods are introduced: empirical evaluation, the theory of decomposition of model error into bias and variance, the theory of algorithmic fairness, and methods for quantitative assessment of uncertainty. Results achieved: 1) low predictive variability is proposed as a desirable property to ensure safety and equality of variability between different social groups as a new metric of fairness of machine learning models; 2) it is demonstrated how stability analysis helps specialists overcome the challenges of model multiplicity and choose reliable, stable and fair models; 3) an open-source software framework for community use is created that integrates stability measurements into model development processes. Conclusions. This work proposes a new paradigm of group fairness that combines the issues of correctness/quality and randomness/stability from the research agenda of responsible artificial intelligence. The application of the proposed approaches helps in the responsible selection of machine learning models under conditions of model multiplicity, demonstrating that, although there may be many models with comparable accuracy, there is only one (or a few) "best" model that is reliable, fair and stable, as it should be.

Keywords: responsible artificial intelligence; algorithmic fairness; model stability; experimental analysis.

Бібліографічні описи / Bibliographic descriptions

Герасимук Д. В., Поляков А. О., Федорченко В. М. Виявлення компромісів між справедливістю, стабільністю і точністю для відповідального вибору моделі машинного навчання. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 1 (31). С. 5–19. DOI: https://doi.org/10.30837/2522-9818.2025.1.005

Herasymuk, D., Poliakov, A., Fedorchenko, V. (2025), "Detecting trade-offs between fairness, stability, and accuracy for responsible machine learning model selection", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (31), P. 5–19. DOI: https://doi.org/10.30837/2522-9818.2025.1.005