УДК 004.85.004.62

ОЦІНЮВАННЯ СТАНУ МОДЕЛІ МАШИННОГО НАВЧАННЯ ЗА УМОВ ВІДКЛАДЕНОЇ ВЕРИФІКАЦІЇ ПРОГНОЗІВ

Лобода Ю. Г., к.пед.н, доцент [1]
[1] Національний університет «Одеська юридична академія»
[1] Україна, Одеса
[1] loboda@onua.edu.ua

Анотація. *Розглянуто моніторинг моделі машинного навчання в період між формуванням прогнозів і надходженням достатньої кількості фактичних цільових значень для їх перевірки. Запропоновано процедуру, у якій можливість оцінити прогнозну якість визначає тип висновку, до верифікації формується попередній рівень ризику за змінами вхідних даних та розподілу прогнозів, а після верифікації визначається стан моделі з урахуванням контрольованих метрик якості. Процедуру реалізовано в інформаційній системі моніторингу та перевірено в контрольованих сценаріях. Показано, що відсутність виявленого дрейфу до верифікації не дає підстав вважати прогнозну якість стабільною.*

Ключові слова: *моніторинг моделей машинного навчання; відкладена верифікація; дрейф даних; деградація моделі; індекс стабільності популяції; оцінювання ризику.*

Вступна частина. Після розгортання моделі характеристики даних і зв'язки між ознаками та цільовою змінною можуть змінюватися, створюючи ризик погіршення прогнозної якості [1]. Безпосередньо оцінити якість можна лише за фактичними цільовими значеннями, які в низці задач стають відомими із затримкою, так результат медичного прогнозу встановлюють після періоду спостереження, а факт звільнення працівника після настання відповідної події [2].

У працях [3, 4] затримку надходження фактичних цільових значень розглянуто переважно як проблему навчання моделей на потоках даних. Інший напрям досліджень присвячено моніторингу змін вхідних даних і виходів моделі, які можна виявити ще до отримання фактичних значень [5, 6]. Проте такі зміни є лише непрямими сигналами можливого погіршення і не характеризують безпосередньо точність прогнозів. Отже, до верифікації можливе лише оцінювання ризику деградації, тоді як висновок про зміну прогнозної якості потребує зіставлення прогнозів із фактичними значеннями.

Мета дослідження полягає у розробленні процедури оцінювання стану моделі машинного навчання, яка враховує можливість верифікації прогнозів і забезпечує перехід від попередньої оцінки ризику до висновку за результатами верифікації.

Основна частина. У процедурі використано три сигнали: рівень дрейфу вхідних даних L_D, рівень зміни розподілу прогнозів L_P і рівень погіршення прогнозної якості L_M. Перші два визначають за індексом стабільності популяції (Population Stability Index, PSI), L_D – за найбільшим PSI серед контрольованих вхідних ознак, L_P – за PSI розподілу виходів моделі. Рівень L_M визначають за найбільшим відносним погіршенням контрольованих метрик порівняно з їх еталонними значеннями. Кожен сигнал набуває рівня 0, 1 або 2 відповідно до заданих порогів. Для PSI використано пороги 0,10 і 0,25. Вони мають евристичний характер, а їх інтерпретація залежить, зокрема, від обсягу вибірки [7], тому в системі ці пороги задають як параметри моделі.

Можливість оцінити прогнозну якість у періоді t описує індикатор A_t. Він дорівнює 1, якщо кількість спостережень із відомими фактичними цільовими значеннями не менша за встановлений мінімум і контрольовані метрики можна коректно обчислити. В інших випадках $A_t = 0$. За $A_t = 0$ рівень L_M є невизначеним: $L_M = \emptyset$. Це значення не можна замінювати на $L_M = 0$, оскільки неможливість перевірити якість не свідчить про відсутність її погіршення.

За $A_t = 0$ процедура формує попередній рівень ризику $R \in \{0, 1, 2\}$ на основі L_D і L_P. Одночасна зміна вхідних даних і розподілу прогнозів підвищує ризик до найвищого рівня. Водночас $R = 0$ означає лише те, що за використаними показниками змін не виявлено; прогнозна якість у цьому режимі залишається невідомою. Навіть високий попередній ризик не підтверджує деградацію і сам по собі не є підставою для автоматичного повторного навчання моделі. Він засвідчує потребу в посиленому моніторингу та, за можливості, пріоритетному отриманні фактичних значень.

За $A_t = 1$ стан моделі визначають за парою (L_D, L_M). Це дає змогу розрізнити чотири випадки: відсутність виявленого дрейфу й погіршення контрольованих метрик; дрейф даних без підтвердженої деградації; погіршення метрик без виявленого дрейфу; одночасну появу дрейфу й погіршення метрик. Попередні значення L_D, L_P і R для того самого періоду зберігають, щоб після верифікації зіставити оцінку ризику з визначеним станом.

Процедуру реалізовано в інформаційній системі моніторингу моделей. Для кожної перевірки зберігають моніторингові сигнали, значення A_t та сформований результат. Якщо прогнозну якість оцінити неможливо, L_M зберігають як відсутнє значення. На графіку зміни метрик такий період позначають розривом лінії: з'єднання точок через період без верифікації створювало б хибне враження, що якість у ньому була відома.

Процедуру перевірено в контрольованих сценаріях на моделі прогнозування наявності серцевого захворювання з використанням набору Heart Failure Prediction [8]. У стабільному сценарії до надходження фактичних значень отримано $L_D = 0$ і $L_P = 0$, проте прогнозну якість не оцінювали. Після верифікації 400 спостережень отримано $L_M = 0$, і стан моделі уточнено до стабільного.

У сценарії зі збільшеною часткою старших пацієнтів максимальний PSI вхідних ознак становив 0,5367. Разом із $L_P = 1$ це зумовило високий попередній ризик. Після верифікації 400 спостережень отримано $L_M = 0$: погіршення контрольованих метрик за встановленими порогами не виявлено. Зокрема, ROC AUC становила 0,9278 за еталонного значення 0,9076, а F-міра – 0,9234 за еталонного значення 0,8571. Отже, у цьому сценарії встановлено дрейф вхідних даних без підтвердженої деградації прогнозної якості.

Протилежний результат отримано в сценарії зі збільшеною часткою складних для моделі випадків. Максимальний PSI вхідних ознак становив лише 0,0715, тобто не досяг нижнього порога 0,10. Водночас ROC AUC знизилася з 0,9076 до 0,5949, а F-міра – з 0,8571 до 0,6316. PSI розподілу прогнозів у цьому сценарії становив 0,2373 і слугував додатковим непрямим сигналом зміни, проте встановити погіршення якості дала змогу саме верифікація прогнозів.

Обмеженість непрямих сигналів виявлено й для моделей прогнозування відтоку персоналу [9] та вартості медичного страхування [10]. У сценаріях їх деградації без виявленого дрейфу вхідних ознак PSI розподілу прогнозів становив відповідно 0,0497 і 0,0475, тобто був нижчим за поріг 0,10. Отже, до верифікації в таких сценаріях можливий рівень $R = 0$, хоча подальша перевірка прогнозів виявляє погіршення якості.

Висновки. Можливість верифікації прогнозів визначає тип результату моніторингу. Якщо фактичні цільові значення відсутні або їх недостатньо, процедура формує попередній

рівень ризику за змінами вхідних даних і розподілу прогнозів. Після надходження достатньої кількості фактичних значень стан моделі визначають з урахуванням контрольованих метрик якості. Розглянуті сценарії показали, що виражений дрейф може не супроводжуватися підтвердженою деградацією, а низький попередній ризик не гарантує збереження прогнозної якості.

Використання інструментів ШІ. Під час підготовки тексту тез використано ChatGPT (GPT-6) для граматичного редагування матеріалу, формулювання анотацій.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Bayram F., Ahmed B. S., Kassler A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, Vol. 245. 108632. https://doi.org/10.1016/j.knosys.2022.108632.

[2] Лобода, Ю. Г., Трофименко, О. Г., & Логінова Н. І. (2025). Інтегрована система аналізу відтоку персоналу на основі поєднання класифікаційних моделей та аналізу виживання. *Таврійський науковий вісник. Серія: Технічні науки*, (3), 228-238. https://doi.org/10.32782/tnv-tech.2025.3.25

[3] Gomes H. M., Grzenda M., Mello R., Read J., Le Nguyen M. H., Bifet A. (2022). A survey on semi-supervised learning for delayed partially labelled data streams. *ACM Computing Surveys*, Vol. 55, 4. Art. 75. https://doi.org/10.1145/3523055.

[4] Castellani A., Schmitt S., Hammer B. (2022). Stream-based active learning with verification latency in non-stationary environments. *Artificial Neural Networks and Machine Learning – ICANN 2022*, 260–272. https://doi.org/10.1007/978-3-031-15937-4_22.

[5] Miller T., Durlik I. (2026). When models fail: Trustworthy anomaly detection under distributional drift via dual-layer monitoring of data and AI behaviour. *Applied Sciences*, Vol. 16, No. 11. 5293.https://doi.org/10.3390/app16115293.

[6] Guan H., Zhou L. (2026). Detecting performance degradation under data shift in pathology vision-language model. *arXiv*. https://doi.org/10.48550/arXiv.2601.00716.

[7] Potgieter C. J., Van Zyl C., Schutte W. D., Lombard F. (2026). The population resemblance statistic: a chi-square measure of fit for banking. *Annals of Operations Research*, Vol. 361, P. 413–435. https://doi.org/10.1007/s10479-025-07024-6.

[8] Fedesoriano. (2021). Heart Failure Prediction Dataset. *Kaggle*. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[9] Pavansubhash. IBM HR Analytics Employee Attrition & Performance. *Kaggle*. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[10] Choi M. Medical Cost Personal Datasets. *Kaggle*. https://www.kaggle.com/datasets/mirichoi0218/insurance.

Loboda Yu.G.

Assessing the state of a machine learning model under delayed verification of predictions

Abstract. *The paper addresses the monitoring of a machine learning model between the time predictions are made and the time sufficient ground truth becomes available for their verification. A procedure is proposed in which the possibility of evaluating predictive performance determines the type of monitoring conclusion. Before verification, a preliminary risk level is assigned based on changes in input data and prediction distributions; after verification, the model state is determined using the monitored performance metrics. The procedure was implemented in a model monitoring information system and examined under controlled scenarios. The results show that the absence of detected drift before verification provides no grounds for considering predictive performance stable.*

Keywords: *machine learning model monitoring; delayed verification; data drift; model degradation; Population Stability Index; risk assessment*