УДК 004.62 DOI: 10.31673/2786-8362.2025.024380

Шаш М.С.; Звенигородський О.С., к.т.н.

МЕТОД ІТЕРАЦІЙНОГО СТВОРЕННЯ ОЗНАК НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ПРОГНОЗУВАННЯ ВІДТОКУ

Shash M.S., Zvenyhorodskyi O.S. Iterative feature generation method based on artificial intelligence for churn prediction. The article proposes an iterative feature generation method using large language models (LLMs) to improve the efficiency of customer churn prediction in SaaS platforms. The proposed approach combines an LLM generator and an LLM critic in a closed feedback loop, enabling multi-step refinement and selection of relevant features based on model performance metrics. Experimental results on streaming platform data demonstrated an increase in the F1-score compared to the baseline approach. The obtained experiment results confirm the effectiveness of the iterative use of LLMs for automating feature creation and enhancing the accuracy of churn prediction models.

Keywords: churn prediction, large language models, feature generation, machine learning, artificial intelligence

Шаш М.С., Звенигородський О.С. Метод ітераційного створення ознак на основі штучного інтелекту для прогнозування відтоку. У статті запропоновано метод ітераційного створення ознак із використанням великих мовних моделей (LLM) для підвищення ефективності прогнозування відтоку клієнтів у SaaS-платформах. Запропонований підхід поєднує LLM-генератор і LLM-критик у замкненому контурі зворотного зв'язку, що забезпечує багатокрокове вдосконалення та відбір релевантних ознак на основі показників якості моделі. Експериментальні результати на даних стрімінгової платформи показали підвищення F1-міри у порівнянні з базовим підходом. Експериментальні результати підтверджують ефективність ітераційного використання LLM для автоматизації створення ознак і підвищення точності моделей прогнозування відтоку користувачів.

Ключові слова: прогнозування відтоку, великі мовні моделі, створення ознак, машинне навчання, штучний інтелект

Вступ

Прогнозування відтоку користувачів є важливим завданням для компаній із підписковою моделлю, оскільки своєчасне виявлення ризику відтоку підвищує утримання клієнтів і прибутковість бізнесу. Основні труднощі впровадження моделей прогнозування відтоку користувачів пов'язані з сильним класовим дисбалансом [1], обмеженою кількістю або якістю релевантних ознак та потребою в прозорих [2], зрозумілих моделях. Додатково, серійні зміни поведінки клієнтів і трансформація бізнес-моделі в умовах цифрових сервісів вимагають, щоб моделі не лише передбачали, але й адаптувалися до нових паттернів та надавали інтерпретацію за своїми рішеннями [3].

Сучасні дослідження демонструють потенціал великих мовних моделей для автоматизованої генерації ознак із табличних даних [4]. Зокрема, у роботі показано, що такі моделі можуть створювати ефективні ознаки без додаткового ручного втручання. Проте більшість підходів не враховують ітераційний зворотний зв'язок між якістю моделі та створеними ознаками, що знижує їхню ефективність.

Постановка завдання. Завдання полягає у розробці метода прогнозування відтоку користувачів для SaaS-платформ з автоматизованою генерацією ознак за допомогою великих мовних моделей (LLM).

Аналіз останніх досліджень. У сучасних дослідженнях відтоку за допомогою машинного навчання значна увага приділяється автоматизації процесу створення ознак [5] (feature engineering), який традиційно вимагає глибокої предметної експертизи та значних людських зусиль. Розвиток великих мовних моделей (LLM) відкрив нові можливості для побудови систем, здатних генерувати, перевіряти та вдосконалювати ознаки на основі описів даних та попередніх результатів. Зокрема, у [4] запропоновано концепцію FeatLLM, де LLM використовується для синтезу нових ознак із табличних даних. Подібні підходи розвивають ідеї TabLLM [6] та подібні, що демонструють ефективність використання мовних моделей у задачах табличного навчання. Проте більшість досліджень зосереджені лише на одноразовій генерації ознак, без інтеграції ітераційного зворотного зв'язку на основі якості моделі.

Попри значний прогрес, залишається низка невирішених питань, пов'язаних із контролем якості згенерованих ознак, ризиком генерації хибних або некоректних правил, а також відсутністю адаптивних механізмів вдосконалення на основі показників ефективності (наприклад, F1-міри). Так, дослідження [7] підкреслюють необхідність залучення «LLM-критика» – другої моделі, що виконує роль фільтра, оцінюючи релевантність і безпечність створених ознак. [8, 9] вказують на доцільність комбінування LLM з класичними методами інтерпретованості, такими як SHAP або LIME, для підвищення точності та прозорості у процесі автоматизованого створення ознак. Однак інтеграція таких підходів у повноцінний замкнений цикл генерації, оцінки та відбору ознак залишається відкритою проблемою.

Актуальність даного дослідження зумовлена необхідністю розробки ітераційної системи створення ознак, здатної не лише генерувати, але й самостійно перевіряти, відбирати та вдосконалювати ознаки протягом кількох циклів, використовуючи як метрики ефективності моделі, так і критику від додаткової LLM. Представлена робота спрямована на вирішення цієї прогалини шляхом створення багатокрокового контуру зворотного зв'язку, у якому поєднано генерацію кандидатних наборів ознак, автоматизовану оцінку результатів і адаптивний вибір найкращих рішень для підвищення узагальнюючої здатності моделей прогнозування відтоку користувачів.

Метою роботи є підвищення ефективності прогнозування відтоку користувачів шляхом створення методу ітераційного створення ознак із використанням великих мовних моделей. Задача полягає в тому щоб створити автоматизований ітераційний метод, який забезпечує багатокрокове генерування, перевірку та відбір релевантних ознак на основі показників якості моделі та зворотного зв'язку від додаткової мовної моделі-критика.

Виклад основного матеріалу дослідження

У задачах прогнозування відтоку користувачів (churn prediction) мета полягає у створенні моделі, що здатна передбачити ймовірність припинення користування послугою на основі історичних характеристик клієнтів.

Нехай маємо набір даних із n спостережень, кожне з яких представлено вектором ознак $x_i \in R^m$ x_i та цільовою змінною $y_i \in \{0, 1\}$, що позначає факт відтоку (1 – клієнт відтік, 0 – залишився):

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}, \quad (1)$$

Мета моделі – знайти таку функцію $f(x) = (x,\ w)$, яка апроксимує залежність між вхідними ознаками та ймовірністю відтоку:

$$\hat{y} = f(x, w), \quad (2)$$

де w – параметри моделі, які навчаються в процесі оптимізації.

Функцію втрат $L(w)$ можна визначити як бінарну крос-ентропію:

$$L(w) = \frac{-1}{n} * \sum_{i=1}^{n} [y_i \log(\hat{y_i}) + (1 - y_i) \log(1 - \hat{y_i})], \quad (3)$$

Оптимізація параметрів здійснюється шляхом мінімізації функції втрат:

$$w^* = argmin\, L(w), \quad (4)$$

Запропонований метод побудований як багатоетапна ітераційна процедура, що поєднує класичні підходи машинного навчання та великі мовні моделі (LLM). Основна ідея полягає в автоматизованому створенні нових ознак (features) із вихідного табличного набору даних за допомогою семантичного аналізу та зворотного зв'язку моделі-критика. Процес роботи включає етапи, представлені на рис. 1.

Генерація нових ознак. Мовна модель M_g аналізує опис колонок та статистику даних і генерує множину нових ознак $F_g = \{f_1, f_2, \dots, f_k\}$ кожна з яких подана у вигляді функції:

$$f_j = \varphi_j(X), \quad (5)$$

де φ_j – правило трансформації, сформульоване LLM (наприклад, "якщо кількість звернень до підтримки > 2, то клієнт проблемний").

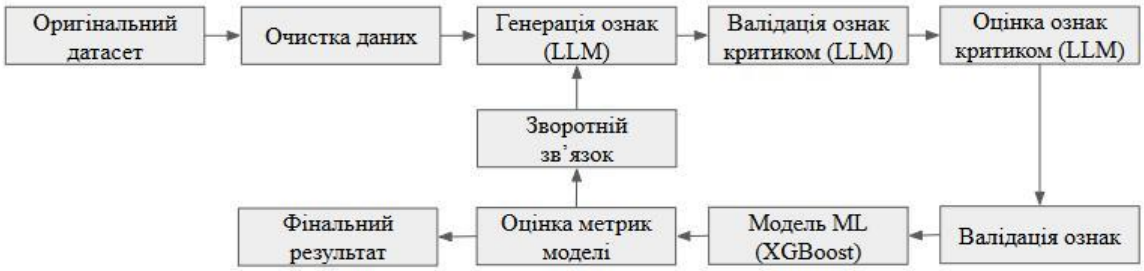

Рис 1. Метод ітеративної генерації ознак з LLM критиком та контуром зворотного зв'язку

Оцінка моделей на нових ознаках. Для кожної ітерації створюється модель класифікації $f_i$fi, яка навчається на наборі ознак $X_i = X_{base} \cup F_{g_i}$. Якість оцінюється за метрикою F1-score:

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}, \quad (6)$$

Зворотний зв'язок моделі-критика. Додаткова мовна модель M_c аналізує продуктивність кожної ітерації та пропонує уточнення чи відсів нерелевантних ознак на основі логів попередніх кроків.

Ітераційне оновлення. Система повторює кроки 1-3 до досягнення критерію зупинки:

$$\Delta F1 < \varepsilon, \quad (7)$$

де ε – мінімальний поріг покращення якості.

Архітектура експерименту. У ході експериментів використано набір даних із 245 000 користувачів цифрової платформи потокового відео [10]. Дані містили 21 змінну (вік акаунту, середні години перегляду, кількість звернень до підтримки, рейтинг користувача тощо). Модель базувалася на XGBoost, параметри якого підбиралися випадковим чином у межах кількох ітерацій. Зокрема, у кожному циклі випадково обиралися значення таких гіперпараметрів: $n\_estimators \in \{200, 300, 400, 500\}$, $learning\_rate \in \{0.01, 0.05, 0.1, 0.2\}$, max $\_depth \in \{3, 4, 5, 6, 7\}$, $colsample\_bytree \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$.

Також використовувався $random\_state = 42$ для відтворюваності результатів та $eval\_metric = 'logloss'$ як метрика оцінки моделі.

Порівняння з базовими підходом. Для оцінки ефективності проведено порівняння трьох методів: XGBoost з базовими ознаками, XGBoost з LLM-генерованими ознаками, XGBoost з LLM ознаками LLM критиком та контуром зворотного зв'язку. В таблиці 1 показані метрики ефективності для трьох методів.

Таблиця 1

Результати оцінки ефективності 3 методів

Метод	Accuracy	Precision	Recall	F1 score
XGBoost з базовими ознак	0.7596	0.3546	0.3993	0.3756
XGBoost з LLM ознак	0.6555	0.3012	0.6835	0.4181
XGBoost з LLM ознаками, LLM критиком та контуром зворотного зв'язку	0.6876	0.3236	0.6652	0.4354

Таким чином ми бачимо покращення F1 міри для незбалансованого датасету стрімінгової платформи з 0.3756 до 0.4354 за рахунок LLM-генерації ознак та LLM критика з контуром зворотного зв'язку.

Висновки

У роботі розроблено метод для ітераційного автоматизованого створення ознак у задачах прогнозування відтоку користувачів. Запропонований метод поєднує LLM-генератор і LLM-критик, що забезпечує багатокрокове вдосконалення ознак на основі попередніх результатів.

Такий механізм адаптивного зворотного зв'язку дозволив підвищити F1-score для незбалансованого датасету. Отримані результати підтверджують, що великі мовні моделі здатні не лише інтерпретувати структуру табличних даних, а й генерувати семантично значущі нові ознаки, які покращують точність прогнозування. Розроблений метод може бути застосований у будь-яких сферах з моделлю підписки, таких як SaaS, стрімінг, телекомунікаціях та e-commerce. Подальші дослідження варто спрямувати на врахування причинно-наслідкових зв'язків між ознаками та використання мультимодальних LLM-моделей у процесі створення ознак.

Список використаної літератури:

1. Suguna R., Suriya P., Pai H. A. et al. Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. *Scientific Reports*. 2025. Vol. 15. P. 16256. URL: https://doi.org/10.1038/s41598-025-01031-0.

2. Noviandy T. R., Idroes G. M., Hardi I. et al. A model-agnostic interpretability approach to predicting customer churn in the telecommunications industry. *Infolitika Journal of Data Science*. 2024. Vol. 2, no. 1. P. 34–44. URL: https://doi.org/10.60084/ijds.v2i1.199.

3. *ChurnKB: A Generative AI-Enriched Knowledge Base for Customer Churn Feature Engineering. Algorithms*. 2025. Vol. 18, no. 4. P. 238. URL: https://doi.org/10.3390/a18040238.

4. Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning. *arXiv preprint*. 2024. URL: https://arxiv.org/abs/2404.09491.

5. Imani M., Joudaki M., Beikmohammadi A., Arabnia H. R. Customer churn prediction: a systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Machine Learning and Knowledge Extraction*. 2025. Vol. 7, no. 3. P. 105. URL: https://doi.org/10.3390/make7030105.

6. Hegselmann S., Buendia A., Lang H., Agrawal M., Jiang X., Sontag D. TabLLM: Few-shot classification of tabular data with large language models. *arXiv preprint*. 2022. URL: https://doi.org/10.48550/arXiv.2210.10723.

7. Gong N., Wang X., Ying W., Bai H., Dong S., Chen H., Fu Y. Unsupervised feature transformation via in-context generation, generator-critic LLM agents, and duet-play teaming. *arXiv preprint*. 2025. URL: https://arxiv.org/abs/2504.21304.

8. Zhao H., Chen H., Yang F., Liu N., Deng H., Cai H., Wang S., Yin D., Du M. Explainability for large language models: a survey. *ACM Transactions on Intelligent Systems and Technology*. 2024. Vol. 15, no. 2. URL: https://doi.org/10.1145/3639372.

9. Zhang X., Zhang J., Rekabdar B., Zhou Y., Wang P., Liu K. Dynamic and adaptive feature generation with LLM. *arXiv preprint*. 2024. URL: https://arxiv.org/abs/2406.03505.

10. Kaggle dataset: Predictive Analytics for Customer Churn Dataset. URL: https://www.kaggle.com/datasets/safrin03/predictive-analytics-for-customer-churn-dataset.

Автори статті

Шаш Максим – аспірант, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.
ORCID: 0009-0009-3274-5318

Звенигородський Олександр – кандидат технічних наук, доцент, Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0008-6235-1638

Authors of the article

Shash Maksym – postgraduate, State University of Information and Communication Technologies, Kyiv, Ukraine.
ORCID: 0009-0009-3274-5318

Zvenyhorodskyi Oleksandr – Candidate of Sciences (technical), Associate Professor, State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0009-0008-6235-1638