Удосконалення автоматизованої генерації ознак мовними моделями для виявлення шахрайства у веб-застосунках

УДК 004.056:004.85

Вадим Яковець[1]

Ужгородський національний університет, [1]vadym.yakovets@uzhnu.edu.ua

Постановка проблеми. Виявлення шахрайства у веб-застосунках електронної комерції та банківських платіжних системах є практичною задачею прикладної кібербезпеки. Класифікатори цієї задачі працюють із сильно дисбалансованими класами, оскільки шахрайські транзакції зазвичай складають менше 0,1 % обсягу. Інженерія ознак потребує знань предметної галузі й переважно виконується вручну. Мовні моделі (CAAFE [1], LLM-FE [2]) дозволяють її автоматизувати, але у своєму дослідженні Küken J., Purucker L. та Hutter F. [3] показали важливу ваду: моделі майже завжди пропонують прості арифметичні оператори (додавання, множення) і дуже рідко пропонують операції групування з агрегацією. У задачах виявлення шахрайства саме операції групування дають доступ до агрегованих метрик користувача (частоти, обсягів, часових патернів), на яких будуються правила виявлення рідкісних подій. AML-датасети використано як наближений стенд, вони містять часові транзакції, контрагентів, суми та рідкісну позитивну мітку, але не покривають усіх веб-поведінкових ознак.

Метою роботи є удосконалення методології автоматизованої генерації ознак мовними моделями для класифікаторів виявлення шахрайства у веб-застосунках через зміну розподілу пропонованих операторів і запровадження вартісно-чутливого протоколу оцінювання.

Об'єктом дослідження є процеси автоматизованої генерації ознак для класифікаторів виявлення шахрайства; *предметом* є критерії відбору ознак з врахуванням вартості помилок та протокол часової вкладеної валідації для незалежної оцінки приросту якості.

Актуальність. Збитки від платіжного шахрайства в електронній комерції продовжують зростати, а ручна інженерія ознак для відповідних класифікаторів є затратною. Без виправлення зсуву моделей до простих операторів автоматична інженерія ознак мовними моделями не дає стабільного приросту якості в задачі виявлення платіжного шахрайства. У доступних українських публікаціях аналогічну задачу прицільно не розглядали.

Наукова новизна. Запропоновано протокол автоматизованої генерації ознак мовними моделями для виявлення шахрайства, що розвиває підхід CAAFE [1] та LLM-FE [2] трьома елементами: 1) шаблон системного запиту доповнено апріорними знаннями про оператори, що емпірично зсуває розподіл пропонованих моделлю операторів від простих арифметичних до групувань з агрегацією; 2) критерій прийняття кандидатної ознаки замінено: замість ROC-AUC використано очікувану вартість помилок із матрицею вартості, яка враховує робочий поріг; 3) фінальне оцінювання виконано на незалежному часовому тестовому розбитті з порогом, налаштованим лише на окремій калібрувальній вибірці.

Виклад розв'язку. Методологія включає класифікатор LightGBM на 27 базових ознаках, мовну модель як генератор кандидатних ознак із запитом, орієнтованим на предметну галузь і цикл прийому ознак з контролем на окремій калібрувальній вибірці, що схвалює нову ознаку за мінімумом очікуваної вартості. Часовий поділ 70/15/15. Оператори ознак навчаються лише на тренувальній вибірці, калібрувальна використовується для пошуку робочого порогу, а мітки тестової вибірки задіюються лише для фінального підрахунку метрик.

Результати пілотного експерименту. Публічні датасети IBM AML LI-Small [4] (6,92 млн транзакцій, частка шахрайства 0,0515 %) і LI-Medium (31,25 млн, 0,0513 %), стратифікована підвибірка 200 тис., по 20 ознак-кандидатів на модель. Протестовано сім моделей. Одну виключили за результатами перевірки запам'ятовування. До базової панелі увійшли шість моделей: Llama-3.1-8B, Mistral-v0.3, GPT-4o-mini, Claude-Haiku-4.5, Claude-Sonnet-4.6 та GPT-5.4. Усереднена частка простих операторів серед моделей, валідних в обох протоколах, наведена у табл. 1.

Таблиця 1.

Усереднена частка простих операторів за моделями, валідними в обох протоколах

Датасет	Моделей	CAAFE % простих	Новий % простих
LI-Small	4	82,2	1,3
LI-Medium	5	69,9	1,1

На LI-Small Mistral-7B-v0.3 і GPT-4o-mini виключено через невалідний формат відповіді, на LI-Medium виключено лише Mistral. У перевірці запам'ятовування за Kuken §3.4 (n = 100) усі шість моделей базової панелі залишаються нижче 50 %-го порогу. Відповідь моделі очікується у форматі JSON з полями {оператор, колонки, обгрунтування}, валідність визначається успішним парсингом і побудовою ознаки на тренувальній вибірці.

Обмеження: вартісний критерій прийому не залишив жодної ознаки (у калібрувальній вибірці було 17 шахрайських транзакцій на LI-Small і 12 на LI-Medium), тому ефект дав передусім шаблон запиту. На одному часовому відкладеному наборі зменшення вартості помилок не зафіксовано.

Висновки. У цьому пілоті підтверджено лише перший етап: запит справді змінює типи згенерованих ознак (табл. 1). Зменшення вартості помилок поки не показано, вартісний критерій прийому не залишив жодної кандидатної ознаки. Кількісне оцінювання приросту якості потребує повторних часових розбиттів.

1. Hollmann N., Müller S., Hutter F. LLMs for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. *NeurIPS*. 2023. arXiv:2305.03403.
2. Abhyankar N., Shojaee P., Reddy C.K. LLM-FE: Automated Feature Engineering for Tabular Data with LLMs as Evolutionary Optimizers. arXiv preprint. 2025. arXiv:2503.14434.
3. Küken J., Purucker L., Hutter F. Large Language Models Engineer Too Many Simple Features for Tabular Data. 3rd Workshop on Table

Representation Learning at NeurIPS. 2024. arXiv:2410.17787v2.
4. Altman E. et al. Realistic Synthetic Financial Transactions for Anti-Money Laundering Models. *NeurIPS*. 2023. arXiv:2306.16424.

Sustainable information technology for auditable financial anomaly prediction aligned with the EU AI Act, ESG and CSRD standards

UDK 004.8:502.131.1]:657.6:[341.171:061.1EU]

Mykola Zlobin[1]

[1]Chernihiv Polytechnic National University, [1]mykolay.zlobin@gmail.com

The digital transformation of EU financial institutions requires a shift from experimental AI models to industrial, auditable, and sustainable AI systems. This transition is central to the goals of the AIFEU project, which focuses on artificial intelligence in EU financial institutions. In banking, AI is increasingly used for credit scoring, fraud detection, anomaly monitoring, and risk assessment. Creditworthiness and credit-scoring AI systems are classified as high-risk under Annex III of the EU AI Act, while fraud-detection systems, although treated differently in the legal classification, still require strong governance because they affect operational risk, customer protection, and institutional accountability. This creates a direct need for financial AI systems that are transparent, stable, and resource-aware. The scientific contradiction is clear: models optimized solely for predictive accuracy may become fragile in real-world conditions, suffer from backtest overfitting, and incur unnecessary computational and environmental costs.

This paper presents Sentinel as a sustainable information technology for predicting financial anomalies. Sentinel is not defined as a single predictive model. It is a modular information technology designed to support the full analytical cycle. It connects raw data processing, model stability diagnostics, sustainability evaluation, resource-aware training, and automated reporting. The architecture follows a regulation-first logic. This means auditability, traceability, stability, and sustainability are not added after model training. They are embedded in the technical architecture from the beginning. Sentinel consists of 4 functional modules: Adaptive data engine, Diagnostic algorithmic core, Green AI guard, and reporting layer.

The first module is the Adaptive Data Engine. It implements the Data processing method for financial datasets with extreme class imbalance, noise, and leakage risk. In the experiment, the credit-card fraud dataset contained only 0.18% fraudulent cases. This creates a serious risk because a model can appear accurate while failing to detect rare anomalies. Sentinel addresses this through leakage-safe preprocessing. The method includes robust scaling, stratified sampling, under-sampling, outlier removal, and out-of-fold target encoding for categorical variables. These operations protect the integrity of the input data and support the logic of Article 10 of the EU AI Act, which emphasizes data quality, data governance, data preparation, and bias mitigation for high-risk AI systems, also supporting ESG and CSRD compliance.

The second module is the Diagnostic algorithmic core. It implements a stability diagnostic model based on dispersion indicators: variance, interquartile range, and 95% confidence interval. Unlike standard validation relying only on average accuracy or error, this module evaluates model stability across folds and complexity levels. It