Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40360413
Article received: 25.06.2026
Article accepted: 25.08.2026
Article published: 24.09.2026
Correspondence author: K. O. Tieliegin telegin.kirill@gmail.com

UDC 004.[85+94]

К. О. Тєлєгін, Я. В. Лавренюк
Житомирський державний університет імені Івана Франка, м. Житомир, Україна

КОНТРОЛЬОВАНА РЕКОНСТРУКЦІЯ LLM-ЗГЕНЕРОВАНИХ СИНТЕТИЧНИХ ПОВЕДІНКОВИХ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ ПОВТОРНОЇ ПОКУПКИ КЛІЄНТОМ У CRM-СИСТЕМАХ

Проаналізовано можливості використання синтетичних даних, згенерованих великими мовними моделями LLM (англ. *Large Language Model*), для вирішення завдань прогнозування повторної покупки клієнтом у CRM-системах (англ. *Customer Relationship Management* – управління взаємовідносинами із клієнтами). Встановлено, що безпосереднє використання сирих LLM-згенерованих транзакційних даних є неприйнятним через явище колапсу на рівні користувачів: у сирих наборах даних фіксується тільки декілька сотень унікальних користувачів замість очікуваних десятків тисяч. Запропоновано методику контрольованої реконструкції LLM-згенерованих синтетичних транзакційних даних, який відокремлює правдоподібність окремих транзакцій від реалістичності поведінки користувачів на агрегованому рівні. Розроблено процедуру реконструкції синтетичних даних, що перепризначає транзакції новим синтетичним користувачам відповідно до емпіричних розподілів кількості транзакцій, отриманих із реальних даних. Досліджено чотири напрями оцінювання синтетичних даних: REAL→REAL, LLM_REBUILT→REAL, REAL→LLM_REBUILT та REAL+LLM_REBUILT→REAL – з використанням часового поділу та без витоку даних. Застосовано модель Random Forest як основну модель та моделі XGBoost і LightGBM – для оцінювання робастності отриманих результатів та підтвердження відтворюваності висновків. З'ясовано, що реконструйовані синтетичні дані частково зберігають поведінковий сигнал, проте не відтворюють повністю реальну купівельну поведінку на рівні користувачів. Встановлено, що синтетичні дані краще зберігають монетарні закономірності, ніж часову поведінкову динаміку. Підтверджено, що гібридне навчання моделі на реальних і реконструйованих синтетичних даних покращує якість прогнозування поведінки користувачів порівняно з навчанням моделі тільки на синтетичних даних. Охарактеризовано контрольовану реконструкцію LLM-згенерованих синтетичних даних як перспективну стратегію доповнення реальних поведінкових даних у завданнях CRM-прогнозування, яку, водночас, не можна розглядати як їх повну заміну.

Ключові слова: машинне навчання; синтетичні дані; великі мовні моделі; управління взаємовідносинами із клієнтами; прогнозування поведінки користувачів; гібридне навчання моделі; модель Random Forest.

Вступ / Introduction

Прогнозування повторної покупки клієнтом є одним з ключових завдань систем управління взаємовідносинами із клієнтами CRM (англ. *Customer Relationship Management*). Практична цінність таких прогнозів зумовлена їх зв'язком із поведінкою клієнтів: повторна покупка є індикатором лояльності та ефективності маркетингових стратегій. У сучасній науковій літературі таке завдання досліджують переважно на підставі поведінкових ознак, похідних від RFM-моделі (англ. *Recency, Frequency, Monetary* – давність, частота, грошова цінність). Зокрема, автори в роботі [11] запропонували RFM-орієнтований підхід до сегментації клієнтів, водночас в праці [14] інтегрували RFM-аналіз із класифікаційними алгоритмами для прогнозування повторних покупок, а в роботі [13] дослідили особливості прогнозування постійної цінності клієнта засобами машинного навчання.

Побудова та оцінювання моделей прогнозування поведінки користувачів потребують повних поведінкових клієнтських даних, проте на практиці їх отримання ускладнюється вимогами конфіденційності, комерційною чутливістю та неповнотою. Це спонукає дослідників до розроблення методів генерування синтетичних даних. Серед сучасних підходів – дифузійні моделі TabDDPM [8, 15], а також генерування даних на підставі великих мовних моделей LLM (англ. *Large Language Model*) – GReaT [2]. Системні дослідження [4, 5, 10] вказують на те, що якість синтетичних даних потрібно оцінювати не тільки за статистичною подібністю, а й за придатністю для подальшого навчання моделей.

Попри широкий розвиток методів генерування синтетичних даних, питання придатності LLM-згенерованих транзакційних даних для завдань CRM-прогнозування на рівні користувачів залишається практично нерозглянутим. Жодна з проаналізованих робіт не описує явища колапсу унікальних користувачів у сирих LLM-виходах – коли десятки тисяч транзакцій генеруються

© Copyright 2026 by the author(s)

Тєлєгін Кирило Олександрович, аспірант, кафедра комп'ютерних наук та інформаційних технологій.
Email: telegin.kirill@gmail.com; https://orcid.org/0009-0004-0154-3952
Лавренюк Ярослав Васильович, д-р фіз.-мат. наук, ст. наук. співробітник, професор, кафедра комп'ютерних наук та інформаційних технологій. **Email:** ylavrenyuk@gmail.com; https://orcid.org/0000-0002-0266-0622

Цитування за ДСТУ: Тєлєгін К. О., Лавренюк Я. В. Контрольована реконструкція LLM-згенерованих синтетичних поведінкових даних для прогнозування повторної покупки клієнтом у CRM-системах. Науковий вісник НЛТУ України. 2026, т. 36, № 4. С. 114–120.
Citation APA: Tieliegin, K. O., & Lavreniuk, Ya. V. Controlled reconstruction of LLM-generated synthetic behavioral data for customer repeat-purchase prediction in CRM systems. *Scientific Bulletin of UNFU*, 36(4), 114–120. https://doi.org/10.36930/40360413

для кількох сотень синтетичних ідентифікаторів, що унеможливлює поведінкове моделювання на рівні клієнтів. Усунення цього структурного недоліку через контрольовану реконструкцію синтетичних даних і становить актуальність цього дослідження.

Об'єкт дослідження – генерування та реконструкція синтетичних поведінкових транзакційних даних для завдань CRM-систем.

Предмет дослідження – методи і засоби контрольованої реконструкції LLM-згенерованих синтетичних даних, які дадуть змогу оцінити їх придатність для прогнозування повторної покупки потенційним клієнтом.

Мета роботи – оцінити результати контрольованої реконструкції LLM-згенерованих синтетичних транзакційних даних стосовно їх придатності для прогнозування повторної покупки клієнтом у CRM-системах, що дасть змогу виправити нереалістичну концентрацію транзакцій порівняно з їх безпосереднім використанням.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

- проаналізувати структурні обмеження сирих LLM-згенерованих транзакційних даних на рівні користувачів, що дасть змогу обґрунтувати потребу їх реконструкції;
- розробити методику контрольованої реконструкції синтетичних даних на підставі емпіричних розподілів, що дасть можливість виправити нереалістичну концентрацію транзакцій;
- оцінити переносність реконструйованих синтетичних даних через чотири напрями експерименту, що забезпечить комплексне порівняння сценаріїв навчання моделі;
- дослідити вплив ознак RFM на якість прогнозування поведінки користувачів, які допоможуть визначити, які поведінкові закономірності зберігаються після реконструкції синтетичних даних;
- перевірити стійкість отриманих результатів на трьох незалежно реконструйованих наборах даних і моделях бустингу, які уможливлять підтвердження відтворюваності зроблених висновків.

Аналіз останніх досліджень та публікацій. У дослідженні [4] автори проаналізували методи генерування та оцінювання синтетичних даних, в т.ч. з GAN-підходами та критеріями якості, і показали, що придатність для подальшого навчання моделей є ключовим критерієм оцінювання синтетичних даних поряд зі статистичною подібністю до оригіналу.

У роботі [5] автори дослідили алгоритми генерування синтетичних даних у латентному просторі та вказали на те, що синтетичні дані мають оцінюватися через їхню корисність для подальшого навчання моделей, а не тільки через подібність до оригіналу.

У дослідженні [9] автори запропонували комплексний підхід до оцінювання синтетичних медичних табличних даних, що охоплює їхню якість, приватність, складність та придатність до використання – принцип, релевантний і для CRM-контексту, оскільки корисність є необхідною складовою оцінювання синтетичних даних.

У роботі [10] автори систематизували інструменти генеративного моделювання та метрики корисності для повністю синтетичних табличних даних, наголосивши на необхідності відтворюваності критеріїв їх оцінювання під час порівняння різних підходів до генерування наборів даних.

У дослідженні [8] автори представили TabDDPM – дифузійну модель генерування табличних даних, яка показала актуальність дифузійних підходів для гетерогенних наборів даних і визначила орієнтири їхньої якості для подальших досліджень у галузі синтетичних табличних даних.

У роботі [2] автори запропонували GReaT – підхід на підставі великих мовних моделей для генерування реалістичних табличних даних, і продемонстрували здатність LLM відтворювати правдоподібні записи на рівні окремих транзакцій, однак питання реалістичності поведінки на рівні користувачів автори не досліджували.

У дослідженні [1] автори запропонували підхід до створення синтетичних наборів даних для рекомендаційних систем з використанням GAN-моделей і показали, що такі дані можна застосовувати для тестування алгоритмів колаборативної фільтрації – аналогічний підхід є релевантним і для оцінювання CRM-алгоритмів.

У роботах [11, 13, 14] автори дослідили RFM-моделі та алгоритми машинного навчання моделі для завдань сегментації клієнтів, прогнозування лояльності користувачів і повторної покупки клієнтом, що обґрунтовує використання поведінкових змінних recency, frequency і monetary value як базових ознак у завданнях CRM-прогнозування.

Отже, аналіз наукових публікацій [1, 2, 3, 4, 5, 8, 9, 10, 11, 13, 14] показує, що попри широкий розвиток методів генерування синтетичних табличних даних, жодна з проаналізованих робіт не розглядає явища колапсу унікальних користувачів у сирих LLM-згенерованих транзакційних даних і не пропонує методологічного процесу його усунення. Саме це і визначає актуальність та наукову новизну нашого дослідження.

Матеріали та методи дослідження. У роботі використано казахстанський набір e-commerce транзакційних даних (Kaggle), що містить 562 149 валідних записів і 233 566 унікальних користувачів після очищення. Синтетичні дані згенеровано за допомогою моделі Qwen2.5-7B-Instruct-Turbo через хмарний API та реконструйовано на підставі емпіричних розподілів кількості транзакцій реальних даних. Для прогнозування повторної покупки клієнтом застосовано модель Random Forest (основна модель), модель XGBoost та модель LightGBM (оцінювання робастності отриманих результатів та підтвердження відтворюваності висновків) з ознаками на підставі RFM-моделі; оцінювання отриманих результатів проведено з часовим поділом і без витоку даних.

Результати дослідження та їх обговорення / Research results and their discussion

Підготовка реальних даних. Реальні транзакційні дані було сформовано на підставі відкритого набору e-commerce транзакцій, опублікованого на платформі Kaggle, який містить часові мітки транзакцій, ідентифікатори користувачів, категорії товарів, бренди та ціни покупок. Після фільтрації записів із відсутніми ідентифікаторами користувачів, некоректними цінами, некоректними часовими мітками та дубльованими записами очищений реальний набір даних містив 562,149 валідних транзакцій і 233,566 унікальних користувачів. Очищені реальні дані було розділено на calibration і holdout вибірки. Кожна вибірка містила 100,000 транзакцій без перетину між calibration і holdout. Усі валідні записи, використані у скоригованих експериментах, належали до 2020 року. Цей етап дослідження був критично важливим, оскільки попередні етапи експерименту, побу-

довані на неповних проміжних даних, давали надто оптимістичні результати. Скоригована методика використовує жорсткішу основу і не спирається на неповні дані.

Формування цільової змінної з часовим поділом. Щоб уникнути витоку даних, мітки повторної покупки було побудовано через часовий поділ history/future. Для кожного користувача історична поведінка використовувалася для обчислення ознак, а цільова змінна визначалася виходячи із майбутніх транзакцій. Було оцінено три сценарії формування цільової змінної: 0,6 weak, 0,7 weak, 0,6 strong. У назві сценарію число 0,6 або 0,7 означає частку індивідуальної часової історії користувача, яка використовується як історичний період для побудови ознак. Позначення weak означає, що позитивний клас визначається як наявність щонайменше однієї майбутньої покупки, а strong – як наявність щонайменше двох майбутніх покупок.

Статистику цільової змінної для реального набору даних наведено в табл. 1. Статистика цільової змінної показує, що скоригована задача є істотно незбалансованою. У сценарії 0,6 weak позитивний клас становить приблизно 5,5 %, у 0,7 weak – приблизно 3,1 %, а у 0,6 strong – приблизно 1,4-1,5 %. Це робить задачу реалістичнішою та складнішою.

Табл. 1. Статистика цільової змінної для реального набору даних / Target variable statistics for the real dataset

Сценарій	Вибірка	Кі-сть корист.	+ клас	– клас
0,6 weak	Calibration	44,238	2,433	41,805
0,6 weak	Holdout	44,390	$2{,}492^{+0{,}001}$	$41{,}897^{+0{,}001}$
0,7 weak	Calibration	50,919	1,560	49,359
0,7 weak	Holdout	50,978	1,601	49,377
0,6 strong	Calibration	44,238	617	43,621
0,6 strong	Holdout	44,390	652	43,738

Реконструкція синтетичних даних. Для фінального етапу експерименту було використано три незалежно згенеровані синтетичні набори даних V_1, V_2 і V_3, створені за допомогою моделі Qwen2.5-7B-Instruct-Turbo через хмарний API. Вибір моделі Qwen було зроблено після попереднього порівняння LLM-кандидатів у завданні генерування структурованих транзакційних записів. У ранніх тестах перевірялися Qwen2.5-7B, Gemma 7B і Llama 3.1 8B. Qwen показав найкраще співвідношення валідності JSON-відповідей, дотримання схеми та швидкості генерування: частка прийнятних відповідей становила приблизно 63,5 %, тоді як Gemma показала близько 55,1 %, а Llama 3.1 – близько 19,8 %. Саме тому Qwen було обрано як основну модель для побудови LLM-згенерованих транзакційних рядків у подальших експериментах. Сирі синтетичні рядки містили правдоподібні значення на рівні окремих транзакцій, але нереалістичну поведінку на рівні користувачів. Зокрема, сирі набори даних містили тільки декілька сотень унікальних користувачів і дуже високу середню кількість транзакцій на користувача. Це робило сирі синтетичні дані непридатними для прямого поведінкового моделювання. Саме на цьому етапі дослідження було застосовано контрольовану реконструкцію синтетичних даних.

Процедура реконструкції синтетичних даних ігнорувала сирі ідентифікатори користувачів, згенеровані LLM, і перепризначала транзакції новим синтетичним користувачам відповідно до емпіричних розподілів кількості транзакцій на користувача, отриманих із реальних даних. Було застосовано обмеження максимальної кількості транзакцій на користувача, щоб уникнути нереалістичних надмірно активних користувачів. Транзакції всередині кожного реконструйованого користувача було впорядковано за часом події.

Після реконструкції синтетичні набори даних мали приблизно 21,000 унікальних користувачів і в середньому близько 2,3 транзакції на користувача, що наближено до реального розподілу. Основні характеристики реконструйованих синтетичних наборів даних наведено в табл. 2. Етап реконструкції синтетичних даних перетворює сирий LLM-згенерований транзакційний матеріал на контрольований поведінковий набір даних. Це не гарантує повної його реалістичності, але робить отримані дані придатними для оцінювання отриманих результатів без витоку даних.

Табл. 2. Характеристики реконструйованих синтетичних наборів даних / Characteristics of reconstructed synthetic datasets

Набір даних	Сирі рядки	Після очищ.	Вилучено	К-сть корист.	Сер. tx/user	≥20 tx	≥50 tx
V_1	50,000	49,980	20	21,534	2,3210	81	0
V_2	50,001	49,997	4	21,192	2,3592	99	0
V_3	50,001	49,996	5	21,395	2,3368	103	0

Набори ознак. Для прогнозування повторних транзакцій та навчання моделей було використано метрики на підставі класичної моделі RFM: основний набір ознак R+M (recency_days, monetary) та діагностичний набір ознак R+M+F (recency_days, monetary, frequency). Ознака frequency розглядалася окремо, оскільки вона поведінково близька до цільової змінної repeat_purchase. Попередні етапи експерименту показували, що додавання frequency може створювати завищені результати. Тому frequency не використовувалася як основна методологічна ознака, але була врахована в ablation-дослідження та оцінювання робастності отриманих результатів.

Моделі. Основною моделлю був RandomForestClassifier. Додатково використовувалися моделі XGBoost і LightGBM. Їхня роль – перевірка незалежності висновків від конкретної моделі, а не оптимізація гіперпараметрів. Використані гіперпараметри моделей подано в табл. 3, тобто зовнішні конфігураційні параметри, які задаються користувачем (розробником) перед початком навчання моделі машинного навчання і не змінюються в процесі самого навчання.

Табл. 3. Гіперпараметри моделей / Model hyperparameters

Модель	Гіперпараметри
Random-Forest	n_estimators=100, class_weight=balanced, random_state=42
Модель XGBoost	n_estimators=100, max_depth=4, learning_rate=0,1, subsample=0,8, colsample_bytree=0,8, eval_metric=logloss, random_state=42
Модель LightGBM	n_estimators=100, max_depth=4, learning_rate=0,1, subsample=0,8, colsample_bytree=0,8, class_weight=balanced, random_state=42

Метрики оцінювання отриманих результатів. Використовувалися такі метрики оцінювання отриманих результатів: ROC-AUC; PR-AUC / Average Precision; Accuracy; Precision; Recall. Метрику PR-AUC було враховано через виражений дисбаланс класів. У таких умовах метрика ROC-AUC може переоцінювати практичну корисність моделі для виявлення позитивного

класу, а accuracy не є надійною через домінування негативного класу.

Напрями оцінювання отриманих результатів. Експерименти містили:

- REAL → REAL: модель навчається на реальній calibration-вибірці та тестується на реальній holdout-вибірці.
- LLM_REBUILT → REAL: модель навчається на реконструйованих синтетичних даних і тестується на реальній holdout-вибірці.
- REAL → LLM_REBUILT: модель навчається на реальній calibration-вибірці та тестується на реконструйованих синтетичних даних.
- REAL + LLM_REBUILT → REAL: модель навчається на суміші реальних і реконструйованих синтетичних даних із фіксованим навчальним бюджетом і тестується на реальній holdout-вибірці.

REAL → REAL визначає скоригований базовий рівень. LLM_REBUILT → REAL оцінює прикладну корисність синтетичних даних. REAL → LLM_REBUILT оцінює, чи можливе перенесення меж прийняття рішень, вивчених на реальних даних, у синтетичний простір. Гібридне навчання моделі дає змогу оцінити, чи можуть синтетичні дані доповнювати обмежений обсяг реальних даних.

Перевірка співвідношень моделей REAL/LLM у їх гібридному навчанні. Гібридне навчання моделей виконувалося за умови фіксованого загального обсягу навчальних даних:

- 10 % реальних даних / 90 % синтетичних даних;
- 20 % реальних даних / 80 % синтетичних даних;
- 30 % реальних даних / 70 % синтетичних даних;
- 40 % реальних даних / 60 % синтетичних даних;
- 50 % реальних даних / 50 % синтетичних даних;
- 60 % реальних даних / 40 % синтетичних даних;
- 70 % реальних даних / 30 % синтетичних даних;
- 80 % реальних даних / 20 % синтетичних даних;
- 90 % реальних даних / 10 % синтетичних даних.

Найкращий гібридний результат обирався за метрикою ROC-AUC. Фіксація загального обсягу навчальних даних була важливою, щоб гібридні моделі не демонстрували кращі результати через більшу кількість навчальних прикладів.

Стійкість отриманих результатів між синтетичними версіями. Стійкість отриманих результатів оцінювалася між наборами даних V_1, V_2 і V_3. Ці набори даних були незалежно згенеровані та реконструйовані. Стабільність результатів оцінювалася через середнє значення, стандартне відхилення, мінімум, максимум і діапазон метрик.

Оцінювання робастності отриманих результатів на моделях бустингу. Моделі XGBoost і LightGBM було застосовано після основних експериментів із моделлю Random Forest. Їхня роль була обмежена оцінюванням робастності отриманих результатів. Grid search, AutoML або вичерпне налаштування гіперпараметрів не виконувалися.

Базовий сценарій REAL→REAL. Скоригована базова модель Random Forest з ознаками R+M дала консервативні значення метрики ROC-AUC. Результати базового сценарію REAL→REAL наведено в табл. 4.

Табл. 4. Базовий сценарій REAL→REAL для моделі Random Forest / REAL→REAL baseline (Random Forest)

Сценарій	ROC-AUC	PR-AUC
0,6 weak	0,5603	0,0648–0,0694
0,7 weak	0,5633	0,0396–0,0518
0,6 strong	0,5512	0,0174

Ці результати істотно нижчі за попередні оцінки, чутливі до витоку даних. Це підтверджує, що скоригована постановка завдання з часовим поділом є складнішою та реалістичнішою. Низькі значення метрики PR-AUC узгоджуються зі значним дисбалансом класів.

Ablation-дослідження ознаки frequency. Додавання frequency покращило метрики, але не наблизило значення метрики ROC-AUC до 1.0. Підсумкові результати ablation-дослідження ознаки frequency наведено в табл. 5. Ознака frequency стабільно покращувала перенесення меж прийняття рішень із синтетичних даних на реальні. Водночас, оскільки frequency значно пов'язана з поведінкою повторної покупки клієнтом, вона залишається методологічно чутливою ознакою для основного оцінювання отриманих результатів.

Табл. 5. Підсумок ablation-дослідження ознаки frequency / Frequency ablation study summary

Ознака	Сценарій	REAL→REAL	LLM→REAL	Hybrid
R+M	0,6 weak	0,5603	0,5384	0,5806
R+M+F	0,6 weak	0,5744	0,5918	0,6037
R+M	0,7 weak	0,5633	0,5344	0,5926
R+M+F	0,7 weak	0,5899	0,5908	0,6137
R+M	0,6 strong	0,5512	0,5640	0,5870
R+M+F	0,6 strong	0,5701	0,6255	0,6294

TSTR-оцінювання: LLM_REBUILT→REAL. Етапи експерименту з перенесення меж прийняття рішень із синтетичних даних на реальні показали слабку, але невипадкову переносність реконструйованих синтетичних даних. Результати TSTR-оцінювання подано в табл. 6. Ці результати показують, що реконструйовані синтетичні дані містять певний переносний поведінковий сигнал. Проте цей сигнал недостатньо значний, щоб навчання моделі тільки на синтетичних даних могло замінити навчання моделі на реальних даних.

Табл. 6. Результати TSTR-оцінювання з ознаками R+M / TSTR results with R+M features

Сценарій	V_1	V_2	V_3
0,6 weak	0,5384	0,5304	0,5275
0,7 weak	0,5341	0,5344	0,5330
0,6 strong	0,5640	0,5389	0,5477
0,6 strong	0,5640	0,5389	0,5477

TRTS-оцінювання: REAL→LLM_REBUILT. Зворотний напрям перенесення меж прийняття рішень дав подібно консервативні результати. Результати TRTS-оцінювання подано в табл. 7. Подібність TSTR- і TRTS-оцінювання робастності отриманих результатів підтверджує, що синтетичні набори даних містять частковий поведінковий сигнал, але цей сигнал структурно відрізняється від реальної поведінкової динаміки.

Табл. 7. Результати TRTS-оцінювання з ознаками R+M / TRTS results with R+M features

Сценарій	V_1	V_2	V_3
0,6 weak	0,5273	0,5449	0,5276
0,7 weak	0,5633	0,5671	0,5591
0,6 strong	0,5575	0,5601	0,5528

Аналіз важливості ознак. Аналіз важливості ознак показав чітку різницю між моделями, навченими на реальних і синтетичних даних. Узагальнені патерни важливості ознак наведено в табл. 8. Головний висновок: реконструйовані синтетичні набори даних значно краще зберігають структуру, пов'язану з monetary, ніж часову їхню поведінку. Це пояснює, чому переносність

реконструйованих синтетичних наборів даних залишається обмеженою.

Табл. 8. Патерни важливості ознак / Feature importance patterns

Сценарій	recency_days	monetary	frequency
Навчання моделі на REAL, R+M	0,33–0,37	0,63–0,66	–
Навчання моделі на synthetic, R+M	0,04–0,05	≈0,95	–
Навчання моделі на REAL, R+M+F	≈0,38–0,42	≈0,51–0,57	0,05–0,08
Навчання моделі на synthetic, R+M+F	≈0,04–0,05	≈0,72–0,77	0,18–0,23
Найкращий hybrid, R+M+F	≈0,11	≈0,67	≈0,22

Стійкість отриманих результатів між наборами даних $V_1/V_2/V_3$**.** Аналіз робастності отриманих результатів показав низьку варіативність між трьома незалежно реконструйованими синтетичними наборами даних. Приклади: TRTS, 0,7 weak, R+M: mean ROC-AUC ≈ 0,563, range ≈ 0,008, std ≈ 0,003; TSTR, 0,7 weak, R+M: mean ROC-AUC ≈ 0,534, range ≈ 0,0014, std ≈ 0,0006; дослідження впливу ознаки frequency, 0,6 weak, R+M: mean ROC-AUC ≈ 0,532, range ≈ 0,011; дослідження впливу ознаки frequency, 0,6 weak, R+M+F: mean ROC-AUC ≈ 0,582, range ≈ 0,016.

Отже, виявлені обмеження переносності поведінкового сигналу не є артефактами одного генерування синтетичних даних. Синтетичні набори даних не є ідентичними, але основна закономірність їхньої структури є стабільною.

Оцінювання робастності отриманих результатів на моделях XGBoost і LightGBM. Моделі бустингу покращили абсолютну якість прогнозування поведінки користувачів, але не усунули розриву між синтетичним і реальним поведінковими просторами. Приклади оцінювання робастності отриманих результатів на моделях бустингу наведено в табл. 9. У цій таблиці наведено вибіркові результати оцінювання робастності для відповідних експериментальних конфігурацій, тому її потрібно інтерпретувати як додаткову перевірку стабільності висновків, а не як окремий повний блок ablation-дослідження ознак.

Табл. 9. Приклади оцінювання робастності отриманих результатів на моделях бустингу / Robustness check results on boosting models

Експеримент	Модель	Сценарій	ROC-AUC	PR-AUC
REAL→REAL	RF	0,6 weak	0,5603	0,0648
REAL→REAL	XGBoost	0,6 weak	0,6364	0,1066
REAL→REAL	LightGBM	0,6 weak	0,6340	0,1034
Best LLM→REAL	RF	0,6 weak	0,5384	0,0726
Best LLM→REAL	XGBoost	0,6 weak	0,5555	0,0757
Best LLM→REAL	LightGBM	0,6 weak	0,5296	0,0705
Best Hybrid	RF	0,6 weak	0,5806	0,0790
Best Hybrid	XGBoost	0,6 weak	0,6191	0,1050
Best Hybrid	LightGBM	0,6 weak	0,6166	0,1008
REAL→REAL	RF	0,6 strong	0,5701	0,0192
REAL→REAL	XGBoost	0,6 strong	0,6701	0,0566
REAL→REAL	LightGBM	0,6 strong	0,6613	0,0546
Best LLM→REAL	RF	0,6 strong	0,6255	0,0343
Best LLM→REAL	XGBoost	0,6 strong	0,6395	0,0507
Best LLM→REAL	LightGBM	0,6 strong	0,6424	0,0463
Best Hybrid	RF	0,6 strong	0,6294	0,0347
Best Hybrid	XGBoost	0,6 strong	0,6580	0,0569
Best Hybrid	LightGBM	0,6 strong	0,6277	0,0548

Ці результати показують, що значніші моделі покращують якість прогнозування поведінки користувачів, особливо у сценаріях REAL→REAL та гібридного навчання моделі. Проте вони не усувають структурного розриву між синтетичним і реальним поведінковими просторами.

Отже, отримані результати дослідження підтверджують ефективність запропонованої методики контрольованої реконструкції LLM-згенерованих синтетичних даних. Реконструкція істотно зменшила прояви явища колапсу на рівні користувачів (від 299-338 до ≈21 000 унікальних користувачів) і зробила синтетичні дані придатними для поведінкового моделювання: навчені на них моделі показали невипадкову переносність реконструйованих синтетичних даних на реальні дані (метрика ROC-AUC 0,53–0,63 у сценарії LLM_REBUILT→REAL), а гібридне навчання моделі на реальних і реконструйованих синтетичних даних забезпечило стабільний приріст якості прогнозування поведінки користувачів (до +0,076 для метрики ROC-AUC порівняно з базовим рівнем REAL→REAL). Стійкість цих результатів на трьох незалежно реконструйованих наборах даних V_1, V_2 і V_3 та на моделях XGBoost і LightGBM свідчить, що практичну корисність синтетичних даних у завданнях CRM-прогнозування забезпечує саме контрольована реконструкція, а не властивості конкретного генератора чи алгоритму навчання моделі.

Обговорення результатів дослідження. У дослідженні [3] автори встановили, що синтетичні дані, згенеровані для завдань незбалансованої класифікації, можуть ефективно покращувати баланс навчальної вибірки, проте їх корисність істотно залежить від структурної відповідності реальним даним і потребує обов'язкової перевірки на незалежному реальному тестовому наборі. Результати гібридного навчання моделі у цій роботі узгоджуються з таким висновком: поєднання реальних і реконструйованих синтетичних даних дає приріст метрики ROC-AUC до +0,076 над базовим рівнем REAL→REAL, тоді як навчання моделі винятково на синтетичних даних поступається базовому рівню.

У роботі [6] автори дослідили конкурентоспроможність моделей глибокого навчання моделі порівняно з деревоподібними моделями на табличних даних і встановили, що деревоподібні ансамблі стабільно перевершують нейронні мережі на типових структурованих наборах даних. Застосування таких моделей, як XGBoost і LightGBM, у ролі оцінювання робастності отриманих результатів у цій роботі підтверджує такий висновок: значніші моделі бустингу покращують абсолютні значення метрики ROC-AUC (до 0,67 у сценарії REAL→REAL проти 0,55 для моделі Random Forest), однак відносний розрив між синтетичним і реальним поведінковими просторами зберігається незалежно від обраного алгоритму.

У дослідженні [7] автори проаналізували причини, з яких деревоподібні моделі переважають методи глибокого навчання моделі на табличних даних, і показали, що ключову роль відіграє здатність дерев ефективно обробляти нерелевантні та нерівномірно розподілені ознаки. Це узгоджується з результатами аналізу важливості ознак у цій роботі: на синтетичних даних модель майже повністю спирається на монетарну ознаку (≈95 %), тоді як на реальних даних вага рівномірніше розподіляється між recency і monetary – що свідчить про

структурну різницю між двома поведінковими просторами даних.

У роботі [12] автори емпірично дослідили поведінку метрик accuracy, precision, recall, F1-score і MCC у завданнях із незбалансованими класами та встановили, що accuracy є ненадійним індикатором якості моделі за домінування негативного класу. Отримані у цій роботі результати підтверджують такий висновок: за частки позитивного класу 1,4–5,5 % значення accuracy перевищують 94 % навіть для тривіального класифікатора, тоді як метрики ROC-AUC і PR-AUC є більш інформативними для порівняння сценаріїв навчання моделі.

У дослідженні [15] автори запропонували модель TabSyn для синтезу табличних даних змішаних типів на підставі дифузії у латентному просторі і показали переваги цього підходу над GAN-базованими методами за метриками статистичної подібності та придатності для машинного навчання моделі. На відміну від TabSyn, запропонований у цій роботі підхід контрольованої реконструкції синтетичних даних не генерує нові транзакційні значення, а коригує структуру розподілу вже LLM-згенерованих даних на рівні користувачів – що є принципово відмінним і взаємодоповнювальним напрямом у галузі синтетичних транзакційних даних.

Отже, зіставлення отриманих наших результатів з даними проаналізованих досліджень [3, 6, 7, 12, 15] підтверджує доцільність застосування контрольованої реконструкції LLM-згенерованих синтетичних даних для завдань CRM-прогнозування. Виявлені структурні обмеження переносності поведінкового сигналу є характерними для синтетичних даних у завданнях із незбалансованими класами та поведінковим моделюванням на рівні користувачів, що вказує на необхідність подальшого вдосконалення методів реконструкції синтетичних даних.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – розроблено методику контрольованої реконструкції LLM-згенерованих синтетичних транзакційних даних для завдань прогнозування повторної покупки клієнтом у CRM-системах, яка відокремлює правдоподібність окремих транзакцій від реалістичності розподілу поведінки на рівні користувачів та істотно зменшує прояви явища колапсу унікальних користувачів у сирих LLM-згенерованих даних шляхом вибору кількості транзакцій із емпіричного розподілу реальних даних.

Практична значущість результатів дослідження – можна використати під час розроблення CRM-систем, де реальні поведінкові дані клієнтів є обмеженими, конфіденційними або недоступними для широкого використання: запропонований підхід контрольованої реконструкції синтетичних даних забезпечує доповнення реальних даних синтетичними з мінімальними витратами на їх генерування, що можна застосовувати в системах персоналізованих рекомендацій, програмах лояльності та автоматизованих CRM-агентах прийняття рішень.

Висновки / Conclusions

Оцінено результати контрольованої реконструкції LLM-згенерованих синтетичних транзакційних даних стосовно їх придатності для прогнозування повторної покупки клієнтом у CRM-системах, що дало змогу виправити нереалістичну концентрацію транзакцій порівняно з їх безпосереднім використанням. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Встановлено, що сирі LLM-згенеровані транзакційні дані є непридатними для прямого використання у завданнях поведінкового моделювання через явище колапсу на рівні користувачів – у сирих наборах зафіксовано тільки 299-338 унікальних користувачів замість очікуваних десятків тисяч.
2. Розроблено методику контрольованої реконструкції синтетичних даних, що перепризначає транзакції новим синтетичним користувачам відповідно до емпіричних розподілів реальних даних і дає змогу отримати реалістичні набори із ≈21 000 унікальних користувачів і середньою кількістю транзакцій ≈2,3 на користувача.
3. З'ясовано, що реконструйовані синтетичні дані частково зберігають поведінковий сигнал (метрики ROC-AUC 0,53-0,63 у сценарії LLM_REBUILT→REAL), проте істотно поступаються реальним даним у відтворенні часової поведінкової динаміки.
4. Досліджено важливість ознак і встановлено, що синтетичні дані зберігають монетарні закономірності (≈95 % ваги на monetary у синтетичних даних) значно краще, ніж часові поведінкові патерни.
5. Підтверджено, що гібридне навчання моделі на реальних і реконструйованих синтетичних даних забезпечує приріст метрики ROC-AUC до +0,076 порівняно з базовим рівнем REAL→REAL у межах фіксованого навчального бюджету.
6. Перевірено стійкість отриманих результатів на трьох незалежних синтетичних наборах (V_1, V_2, V_3) та моделях бустингу (XGBoost і LightGBM) і встановлено, що виявлені обмеження переносності поведінкового сигналу мають структурний характер і не залежать від конкретного алгоритму.

Заява про використання інструментів штучного інтелекту. Під час підготовки цього рукопису автори застосовували інструменти штучного інтелекту та програмне забезпечення для оброблення природної мови (ChatGPT, Copilot) винятково задля граматичного редагування, коригування орфографії та стилістичної адаптації складних синтаксичних конструкцій. Автори здійснювали критичний аналіз і подальше редагування згенерованого тексту та несуть повну відповідальність за достовірність, науковий зміст і остаточну редакцію публікації.

References

1. Bobadilla, J., Gutiérrez, A., Yera, R., & Martínez, L. (2023). Creating synthetic datasets for collaborative filtering recommender systems using generative adversarial networks. *Knowledge-Based Systems*, 280, article ID 111016. https://doi.org/10.1016/j.knosys.2023.111016
2. Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2023). Language models are realistic tabular data generators. *In Proceedings of the 11th International Conference on Learning Representations* (ICLR 2023). https://doi.org/10.48550/arXiv.2210.06280
3. DSouza, A., Swetha, M., & Sarawagi, S. (2025). Synthetic tabular data generation for imbalanced classification: The surprising effectiveness of an overlap class. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(15), 16127–16134. https://doi.org/10.1609/aaai.v39i15.33725
4. Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. Mathematics, *10*(15), article ID 2733. https://doi.org/10.3390/math10152733

5. Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data*, 10, article ID 115. https://doi.org/10.1186/s40537-023-00792-7
6. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932–18943. https://doi.org/10.48550/arXiv.2106.11959
7. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520. https://doi.org/10.48550/arXiv.2207.08815
8. Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. *In Proceedings of the 40th International Conference on Machine Learning* (ICML 2023). PMLR. https://doi.org/10.48550/arXiv.2209.15421
9. Kurakova, A., & Homayouni, H. (2025). A comprehensive evaluation framework for synthetic medical tabular data generation. *Journal of Biomedical Informatics*, 163, article ID 104939. https://doi.org/10.1016/j.jbi.2025.104939
10. Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Camp, P. (2024). Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data. *ACM Computing Surveys*, *57*(4), article ID 90. https://doi.org/10.1145/3704437
11. Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, 237, article ID 121449. https://doi.org/10.1016/j.eswa.2023.121449
12. Sujon, K. M., Islam, M. A., & Rahman, M. M. (2025). Accuracy, precision, recall, f1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, 12, article ID 268. https://doi.org/10.1186/s40537-025-01313-4
13. Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, *9*(2), article ID e13384. https://doi.org/10.1016/j.heliyon.2023.e13384
14. Verma, R., Rathor, D., Kumar, S., Mishra, M., & Baranwal, M. (2025). Enhancing customer repurchase prediction: Integrating classification algorithms with RFM analysis for precision and actionable insights. *IIMB Management Review*, *37*(2), article ID 100574. https://doi.org/10.1016/j.iimb.2025.100574
15. Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., & Karypis, G. (2024). Mixed-type tabular data synthesis with score-based diffusion in latent space. *In Proceedings of the 12th International Conference on Learning Representations* (ICLR 2024). https://doi.org/10.48550/arXiv.2310,09656

K. O. Tieliegin, Ya. V. Lavreniuk

Zhytomyr Ivan Franko State University, Zhytomyr, Ukraine

CONTROLLED RECONSTRUCTION OF LLM-GENERATED SYNTHETIC BEHAVIORAL DATA FOR CUSTOMER REPEAT-PURCHASE PREDICTION IN CRM SYSTEMS

Analyzed are the possibilities of using synthetic data generated by large language models (LLMs) to address repeat-purchase prediction tasks in customer relationship management (CRM) systems. Established is the fact that direct use of raw LLM-generated transactional data is inadmissible due to the user-level collapse phenomenon: raw synthetic datasets contain only a few hundred unique users instead of the expected tens of thousands, which renders them unsuitable for user-level behavioral modeling. Proposed is a methodological pipeline for the controlled reconstruction of LLM-generated synthetic transactional data, which separates the plausibility of individual transactions from the realism of user-level behavioral distributions by resampling transaction counts from the empirical distribution observed in real data, thereby correcting the structural deficiency of raw LLM output. Developed is a reconstruction procedure that reassigns transactions to new synthetic users according to empirical distributions derived from real data, producing datasets with approximately 21 000 unique users and a mean transaction count of 2.3 per user. Investigated are four evaluation directions – REAL→REAL, LLM_REBUILT→REAL, REAL→LLM_REBUILT, and REAL+LLM_REBUILT→REAL – employing time-based, leakage-free user-specific splits across three independently generated and rebuilt synthetic datasets. Applied are Random Forest as the primary model and XGBoost and LightGBM as robustness checks with fixed hyperparameters. Established is that rebuilt synthetic data partially preserve behavioral signal yet do not fully reproduce real purchasing dynamics at the user level. Determined is that synthetic data retain monetary-value patterns considerably better than temporal behavioral dynamics, as evidenced by feature importance analysis showing approximately 95 % weight on the monetary feature in models trained on synthetic data. Confirmed is that hybrid training on real and rebuilt synthetic data consistently outperforms training on synthetic data alone and yields ROC-AUC gains of up to +0,076 over the REAL→REAL baseline within a fixed training budget. Characterized is the controlled reconstruction of LLM-generated synthetic data as a viable strategy for augmenting real behavioral data in CRM prediction tasks, which at the same time cannot be regarded as a complete substitute for real customer behavioral records.

Keywords: machine learning; synthetic data; large language models; customer relationship management; user behavioral prediction; hybrid training; Radom Forest.