УДК 004.8:339.1

СИНТЕТИЧЕСКИЕ ДАННЫЕ ДЛЯ ОБУЧЕНИЯ РЕКЛАМНЫХ МОДЕЛЕЙ В ЭЛЕКТРОННОЙ КОММЕРЦИИ: МЕТОДЫ ГЕНЕРАЦИИ, ПРИВАТНОСТЬ И ОГРАНИЧЕНИЯ

Нуансенгси Д.В.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Пискун Г.А. – к.т.н., доцент, доцент кафедры ПИКС

Аннотация. Рассмотрены методы генерации синтетических данных (CTGAN, TVAE, диффузионные модели) для обучения рекламных CTR-моделей в электронной коммерции. Предложена методика оценки влияния синтетических данных на качество модели с использованием набора данных Criteo. Установлено, что аугментация выборки синтетическими примерами повышает AUC на 2,9%, тогда как полная замена реальных данных приводит к model collapse (–8,6% AUC). Обоснована необходимость гибридного подхода с применением дифференциальной приватности.

Ключевые слова: синтетические данные, CTGAN, CTR-прогнозирование, электронная коммерция, дифференциальная приватность, model collapse, аугментация данных.

Введение. Рекламные модели, прогнозирующие вероятность клика (CTR) и конверсии, составляют основу монетизации современных платформ электронной коммерции. Качество таких моделей напрямую зависит от объёма и разнообразия обучающих данных: исторических записей о показах, кликах, покупках и атрибутах пользователей [1]. Однако индустрия рекламных технологий сталкивается с нарастающим давлением трёх факторов, ограничивающих доступ к реальным данным.

Во-первых, ужесточение регуляторных требований. Вступление в силу GDPR, CCPA и аналогичных законов создало «trade-off между приватностью и полезностью данных»: чем строже анонимизация, тем ниже аналитическая ценность датасетов [2]. Во-вторых, технологическая трансформация: отказ браузеров от сторонних cookies (Safari, Firefox, Google Chrome через Privacy Sandbox) ликвидирует основной механизм индивидуального трекинга, на котором строились рекламные модели последнего десятилетия [3]. В-третьих, дефицит данных по редким, но коммерчески значимым событиям: конверсии составляют 1–3% от кликов, а «cold start» для новых товаров и аудиторий создаёт критический дисбаланс в обучающих выборках [4].

Синтетические данные – искусственно сгенерированные наборы, имитирующие статистические свойства реальных данных без привязки к конкретным пользователям – рассматриваются как перспективное решение этих проблем. По данным Gartner, синтетические данные будут составлять около 75% данных, используемых в AI-проектах, к 2026 году [5]. Рынок генерации синтетических данных оценивается в $267 млн (2023) с прогнозом роста до $4,6 млрд к 2032 году и CAGR 36,3% [6]. Всемирный экономический форум (WEF) в отчёте «Synthetic Data: The New Data Frontier» (2025) признал синтетические данные критически важной технологией для инноваций, подчеркнув необходимость гибридных подходов, комбинирующих синтетические и реальные данные [7].

В предшествующей работе авторов [8] было установлено, что reasoning-модели (Gemini 2.0 Flash Thinking, DeepSeek-R1) демонстрируют «коллапс креативности» при высокой сложности маркетинговых задач. Параллельно исследование Apple «The Illusion of Thinking» [9] выявило аналогичный «model collapse» в reasoning-моделях. В данной статье исследуется смежный, но отличный феномен – model collapse при обучении рекламных моделей на синтетических данных, – а также разрабатываются рекомендации по его предотвращению.

Основная часть. Синтетические данные – это искусственно созданная информация, генерируемая с помощью алгоритмов машинного обучения для имитации статистических распределений и корреляций реальных данных. В отличие от анонимизированных данных, синтетические данные не содержат персональной информации и не могут быть сопоставлены с конкретными пользователями, что выводит их за рамки действия GDPR [2]. Испанское агентство защиты данных (AEPD) классифицирует синтетические данные как технологию повышения приватности (PET), реализующую принцип «protection by design» [3].

В контексте рекламных моделей ключевыми методами генерации синтетических данных являются три класса архитектур: генеративно-состязательные сети (GAN), вариационные автокодировщики (VAE) и диффузионные модели. Их сравнительная характеристика представлена в таблице 1.

Таблица 1 – Методы генерации синтетических данных для рекламных моделей

Метод	**Принцип**	**Достоинства**	**Недостатки**
CTGAN	Состязательное обучение генератора и дискриминатора с условной генерацией	Лучшее качество для табличных данных; открытый код [6]	Нестабильность обучения; mode collapse
TVAE	Кодировщик сжимает данные в латентное пространство	Стабильное обучение; интерполяция	Менее чёткие границы распределений
Диффузионные модели	Последовательное зашумление и обратная денойзизация	Высокое качество; разнообразие	Высокая вычислительная стоимость

Применение синтетических данных в рекламе включает: аугментацию CTR-логов для балансировки классов, восполнение данных при cold start, privacy-preserving обучение без персональных данных. Для экспериментальной оценки использован набор Criteo [7] и три сценария обучения (таблица 2).

Таблица 2 – Результаты экспериментов

Сценарий	**Описание**	**AUC**	**Δ AUC**
Baseline	100% реальных данных	0,7892	–
Гибридный	80% реальных + 20% CTGAN	0,8124	+2,9%
Полностью синтетический	100% CTGAN-данных	0,7214	–8,6%

Гибридный подход (+2,9% AUC) подтверждает эффективность аугментации синтетическими негативными примерами [3]. Полная замена реальных данных приводит к model collapse (–8,6% AUC): «хвосты» распределения исчезают, модель «забывает» реальные паттерны [8]. Seddik et al. [9] доказали, что collapse предотвратим при смешивании реальных и синтетических данных. Данный эффект родственен «коллапсу креативности» reasoning-моделей [5] и «коллапсу рассуждений» LRM [10].

Для обеспечения приватности применяется дифференциальная приватность (DP-SGD): к градиентам генератора добавляется калиброванный шум с бюджетом ε [11]. Google использует данный подход в Privacy Sandbox Topics API [12]. Рекомендации: пропорция синтетических данных не более 20–30%; валидация на реальных тестовых выборках; применение DP-SGD при работе с персональными данными; цикл «генерация → фильтрация → обучение → оценка» с human-in-the-loop [13].

Заключение. Выполнен анализ методов генерации синтетических данных (CTGAN, TVAE, диффузионные модели) и их применимости для обучения рекламных моделей в электронной коммерции. Установлено, что синтетические данные решают три фундаментальные проблемы рекламных технологий: регуляторные ограничения (GDPR, CCPA), отказ от cookies и дефицит данных по редким событиям.

Экспериментально подтверждено, что гибридный подход (80% реальных + 20% синтетических данных) обеспечивает повышение AUC на 2,9% за счёт балансировки классов

и расширения покрытия редких комбинаций признаков. Вместе с тем полная замена реальных данных синтетическими приводит к model collapse (–8,6% AUC) – утрате «хвостов» распределения и деградации разнообразия предсказаний. Данный эффект представляет собой форму «информационного коллапса», родственную «коллапсу креативности» reasoning-моделей, выявленному авторами ранее [8], и «коллапсу рассуждений» LRM, описанному Apple [9].

Предложены практические рекомендации, включающие оптимальное соотношение реальных и синтетических данных, применение дифференциальной приватности, валидацию на реальных тестовых выборках и интеграцию человеческого контроля (human-in-the-loop) в цикл генерации. Результаты работы могут быть использованы при проектировании пайплайнов обучения рекламных моделей на маркетплейсах, при разработке privacy-preserving стратегий таргетирования и в учебном процессе подготовки специалистов по AI-маркетингу.

Список литературы

1. *Cairo, M. Synthetic Data and GDPR Compliance / M. Cairo // Journal of Technology Law & Policy. – 2023. – Vol. 28, Iss. 1.*
2. *From trace to model: how synthetic data is transforming digital privacy / ECIJA [Electronic resource]. – 2026. – Mode of access: https://www.ecija.com.*
3. *Taskin, F.C. Effects of Using Synthetic Data on Deep Recommender Models' Performance / F.C. Taskin [et al.] // arXiv:2406.18286. – 2024.*
4. *Cogent Infotech. Synthetic Data Explosion: How 2026 Reduces Data Costs by 70% [Electronic resource]. – 2025. – Mode of access: https://www.cogentinfo.com.*
5. *Пискун, Е.С. Ограничения масштабирования Reasoning-моделей в задачах автоматической генерации стратегий электронной коммерции / Е.С. Пискун, Д.В. Нуансенгси // Веснік ГрДУ. – 2025. – Т. 15, № 3.*
6. *Xu, L. Modeling Tabular Data using Conditional GAN / L. Xu [et al.] // NeurIPS 2019. – arXiv:1907.00503.*
7. *Shaped AI. Criteo Dataset: Tackling Large-Scale CTR Prediction [Electronic resource]. – 2025. – Mode of access: https://shaped.ai.*
8. *Shumailov, I. AI models collapse when trained on recursively generated data / I. Shumailov [et al.] // Nature. – 2024. – Vol. 631.*
9. *Seddik, M. How Bad is Training on Synthetic Data? / M. Seddik [et al.] // arXiv:2404.05090. – 2024.*
10. *Shojaee, P. The Illusion of Thinking / P. Shojaee [et al.] // Apple. – 2025.*
11. *Chivukula, V. Use of Differential Privacy to Enable Optimization of Ads / V. Chivukula // IJIRCT. – 2022. – Vol. 8, Iss. 5.*
12. *Differentially Private Synthetic Data Release for Topics API Outputs // arXiv:2506.23855. – 2025.*
13. *WEF. Synthetic Data: The New Data Frontier / World Economic Forum. – 2025.*

UDC 004.8:339.1

SYNTHETIC DATA FOR TRAINING ADVERTISING MODELS IN E-COMMERCE: GENERATION METHODS, PRIVACY, AND LIMITATIONS

Nuansengsy D.V.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Piskun G.A. – Cand. of Sci., associate professor, associate professor of the department of ICSD

Annotation. Methods of synthetic data generation (CTGAN, TVAE, diffusion models) for training advertising CTR models in e-commerce are investigated. It is established that augmentation with synthetic samples improves AUC by 2.9%, while complete replacement leads to model collapse (–8.6% AUC). The necessity of a hybrid approach with differential privacy is substantiated.

Keywords: synthetic data, CTGAN, CTR prediction, e-commerce, differential privacy, model collapse, data augmentation.