DOI 10.37882/2223-2966.2025.06-2.41

ВЛИЯНИЕ СИНТЕТИЧЕСКИХ ПОТОКОВЫХ ДАННЫХ, СОЗДАННЫХ ГЕНЕРАТИВНЫМИ МОДЕЛЯМИ МАШИННОГО ОБУЧЕНИЯ, НА АНАЛИТИКУ И МЕТОДЫ РАННЕГО ВЫЯВЛЕНИЯ ОТКЛОНЕНИЙ

THE IMPACT OF SYNTHETIC STREAMING DATA GENERATED BY GENERATIVE MACHINE LEARNING MODELS ON ANALYTICS AND EARLY DETECTION METHODS

K. Ulanov

Summary. The article explores how synthetic streaming data generated by modern generative models affects the operation of real-time analytical services and deviation detection systems. A formal intervention model with proportion–intensity–diversity parameters is proposed for the data flow processing stack. Experiments on real synthetic data streams have shown that as many as 5 % of synthetic messages can significantly impair the accuracy of forecasts and increase the delay in detecting anomalies. The work contributes to the theory of data quality control and formulates practical recommendations on the use of synthetic streams.

Keywords: generative streaming data, synthetic data, deviation detection, data quality control, streaming analytics.

Уланов Кирилл Анатольевич
аспирант, Московский государственный технологический университет «Станкин»
ulanovk08@gmail.com

Аннотация. Статья исследует, как синтетические потоковые данные, генерируемые современными генеративными моделями, влияют на работу аналитических сервисов реального времени и систем обнаружения отклонений. Предлагается формальная модель вмешивания с параметрами доля–интенсивность–разнообразие для стека обработки потоков данных. Эксперименты на реальных синтетических потоках данных показали, что уже 5 % синтетических сообщений могут существенно ухудшить точность прогнозов и увеличить задержку обнаружения аномалий. Работа вносит вклад в теорию контроля качества данных и формулирует практические рекомендации по использованию синтетических потоков.

Ключевые слова: генеративные потоковые данные, синтетические данные, обнаружение отклонений, контроль качества данных, потоковая аналитика.

Введение

С бурным развитием генеративных моделей машинного обучения организации всё чаще подмешивают синтетические события к реальным потокам для нагрузочного тестирования, устранения пробелов в данных и защиты приватности [1]. В сегменте потоковой обработки информации эта практика получила поддержку в популярных технологиях обработки потоковых данных Apache Kafka и Apache Flink, где специальные модули способны генерировать или подменять миллионы сообщений в секунду [2]. Новые модели уже способны воспроизводить сложные «стилизованные факты» финансовых рядов [3] и показывают корреляцию разнообразия синтетически сгенерированных данных с качеством последующего обучения [4].

Однако вместе с выгодами появляются риски искажения аналитики. В отличие от статичных наборов, синтетические потоки смешиваются с событиями реального времени, нарушая предположения стационарности и изменяя распределение признаков, по которым построены бизнес-аналитические панели и модели машинного обучения. Пользователи уже отмечают случаи, когда «правдоподобные» сгенерированные данные маскируют всплески аномалий, или создают их искусственно [5]. Система мониторинга, настроенная на реальные данные, начинает либо пропускать такие события, либо срабатывать слишком часто, генерируя ложные срабатывания [6]. Более того, стандартные детекторы дрейфа теряют чувствительность при локализованных изменениях, характерных для точечной вставки сгенерированной из синтетических данных [7].

Исходя из этого, выдвигается гипотеза: вставка синтетических событий, сгенерированных генеративными моделями машинного обучения, существенно изменяет скрытые распределения потоковых признаков и тем самым усложняет раннее выявление отклонений в аналитике.

Цель исследования — определить влияние доли и характера синтетических данных, генерируемых генеративными моделями машинного обучения на метрики качества аналитических сервисов (MAE, AUC, P@K) и задержку срабатываний систем мониторинга, а также

предложить адаптивные методы детекции, устойчивые к подобным искажениям.

Выявленный недостаток

Несмотря на бурный рост исследований о генерации данных, генеративных моделях и о детекции отклонений, систематических исследований влияния сгенерированных данных на панели бизнес-аналитики, модели машинного обучения и метрики качества данных практически нет. Лишь отдельные работы анализируют усреднённый «прирост» или «просадку» качества модели после подмешивания синтетических данных [9], но практически не затрагивают:

- как меняется латентное распределение признаков в реальном времени;
- какой объём/доля вставки приводит к деградации метрик;
- как адаптировать пороги раннего срабатывания мониторинга к смешанному потоку данных.

Формальное определение «синтетического потока данных»

Пусть бесконечная последовательность сообщений, поступающая из кафка-топика выражается формулой:

$$S = \{x_t | t \in \mathbb{N}\}, \quad (1)$$

Каждое событие представлено вектором признаков $x_t \in \mathbb{R}^d$ с меткой происхождения:

$$\ell_t = \begin{cases} \text{реальные}, x_t \sim \mathcal{D}_{real} \\ \text{синтетические}\, x_t \sim G_\psi(\mathcal{Z}_t)' \end{cases} \quad (2)$$

где G_ψ — генератор синтетических данных, $\mathcal{Z}_t$ — латентный шум [8]. Мы определяем синтетический поток как подпоследовательность $\{x_t : \ell_t = \text{синтетические}\}$.

Поток поступает в обработчик, после чего:

- Бизнес-аналитическая панель строит агрегаты — ошибку измеряем метрикой Mean Absolute Error (MAE);
- Модель машинного обучения выдаёт бинарный прогноз — оцениваем метрику Area under of curve (AUC);
- Сервис классифицирует тип события — считаем метрику F1.

Для каждого временного окна W_t размером w вычисляем метрики MAE, AUC, F1.

Модель влияния и основные задачи

Мы рассматриваем функцию качества:

$$Q(t,\theta) = g\big(MAE(t), AUC(t), F1(t)\big), \quad (3)$$

где g — монотонное свёртывание (например, взвешенная сумма).

Исследуемая величина — приращение качества, т.е. падение или рост метрик из-за синтетических данных:

$$\Delta Q(t;\theta) = Q_{mixed}(t;\theta) - Q_{real}(t), \quad (4)$$

где Q_{mixed} — качество данных со смешиванием с синтетическими данными, Q_{real} — качество реальных данных.

Вторая цель — оценить время обнаружения отклонения:

$$Delay_\alpha(\theta) = \mathbb{E}\big[t_{alert} - t_{inject}\big], \quad (5)$$

где t_{alert} — время срабатывания, t_{inject}— время подмешивания синтетических данных, α — порог ложных срабатываний [9].

Эксперимент

Для демонстрации различий между классами генераторов синтетических данных использованы три открытых реализации создания синтетических данных:

- TimeGAN (Time Generative adversarial network) — рекуррентный генератор для мультивариантных рядов [8];
- FM-TS (Flow-Matching for Time Series) — современная диффузионная архитектура, задающая обратимый поток шума [10];
- RCGAN-TS — условная генерация с управлением сезонностью и редкими выбросами, позволяющая строить целевые аномалии [11].

Таблица 1.

Методы оценки влияния

Метрика	Описание
P@K	Доля реальных ошибок среди первых K срабатываний
Задержка срабатывания мониторинга	Отставание срабатывания мониторинга от реального возникновения аномалии
MAE	Средняя абсолютная ошибка прогноза

В качестве набора данных использовались наборы данных о поездках такси (NYC Taxi) [12], набор медицинских данных интенсивной терапии (MIMIC-III Vital Streams) [13].

Используемые сценарии:

- Без синтетики — поток без добавления синтетических данных;
- Случайная синтетика — 5 % сообщений заменено случайными событиями, не учитывающими доменную корреляцию;
- Доменно-ориентированная синтетика — 5 % сообщений порождены FM-TS или TimeGAN, обученными на соответствующем домене.

Таблица 2.

Результаты эксперимента

Набор данных	Сценарий	MAE	P@1000	Задержка срабатывания мониторинга, с
NYX Taxi	Без синтетики	0	0,72	31
	Случайная синтетика	+0,013	0,59	78
	Доменно-ориентированная синтетика	+0,021	0,54	92
MIMIC	Без синтетики	0	0,70	34
	Случайная синтетика	+0,018	0,55	81
	Доменно-ориентированная синтетика	+0,029	0,49	95

Сгенерированные синтетические данные маскируют аномалии сильнее случайной реальной аномалии. Падение P@K на NYC Taxi составило — 18 п.п. против — 13 п.п. при случайной вставке. Медицинские потоки наиболее уязвимы. Из-за более строгих порогов безопасности метрики ΔMAE выросла на 2,9 % при той же доле синтетических данных. Задержка срабатывания мониторинга увеличивается нелинейно. При подмешивании всего 5 % сгенерированных синтетических данных среднее опоздание срабатывания почти утроилось. Адаптивные пороги мониторингов обязаны учитывать не только долю, но и качество синтетических сообщений; иначе рост метрик ошибок и задержек неизбежен.

Положительные аспекты внедрения синтетических потоков данных

Быстрый «холодный старт» моделей машинного обучения. Генерация дополнительных событий позволяет мгновенно заполнить редкие классы и поднять метрику AUC без дорогостоящего сбора реальных данных.

В нагрузочном тестировании синтетические данные воспроизводят экстремальные пики, не затрагивая продуктивные сервисы.

Безопасная интеграция. В медицине и финансах синтетические потоки данных позволяют отладить интеграцию с другими сервисами; Gartner прогнозирует, что к 2027 г. 70—% интеграций будут начинаться с синтетического набора данных [14].

Выводы

Количественная модель влияния сгенерированных данных. Впервые показано, что даже при доле 5 % «качественной» вставки прирост MAE аналитических прогнозов линейно увеличивается и сопровождается падением P@K до 18 п.п. и увеличением задержки срабатывания. Это уточняет понятие «наблюдаемости данных» для потоков данных и расширяет теорию контроля качества, вводя параметр «энтропия синтетики» как прямой фактор риска.

Заключение

Работа демонстрирует, что появление синтетических потоков данных, созданных генеративными моделями машинного обучения, кардинально меняет ландшафт контроля качества: «умные» синтетические данные столь же полезны, сколь и опасны. Дальнейшие исследования в области мультимодальности, федеративного обмена результатами и формализации контрактов данных представляются критически важными для эволюции практик контроля качества данных.

ЛИТЕРАТУРА

1. Netguru. Synthetic Data: Revolutionizing Modern AI Development in 2025 [Электронный ресурс]. — URL: https://www.netguru.com/blog/synthetic-data (дата обращения: 12.05.2025).
2. Waehner K. Real-Time Model Inference with Apache Kafka and Flink for Predictive AI and GenAI [Электронный ресурс]. — URL: https://kai-waehner.medium.com/real-time-model-inference-with-apache-kafka-and-flink-for-predictive-ai-and-genai-bf9459f66c13 (дата обращения: 12.05.2025).
3. Takahashi T., Mizuno T. Generation of Synthetic Financial Time Series by Diffusion Models // arXiv:2410.18897, 2024.
4. Li Y. и др. On the Diversity of Synthetic Data and Its Impact on Training Large Models // arXiv:2410.15226, 2024.
5. StateTech Magazine. Synthetic Data Supports State and Local Government AI Initiatives [Электронный ресурс]. — URL: https://statetechmagazine.com/article/2024/07/synthetic-data-supports-ai-initiatives-for-municipalities-perfcon (дата обращения: 12.05.2025).
6. Bates R. и др. Model Drift Monitoring: Continuously Tracking Model Performance Metrics to Detect Accuracy Degradation // ResearchGate, 2024.
7. Giobergia F. и др. A Synthetic Benchmark to Explore Limitations of Localized Drift Detections // arXiv:2408.14687, 2024.
8. Yoon J. et al. Time-Series Generative Adversarial Networks // Advances in Neural Information Processing Systems. — 2019.
9. Baena-García M. et al. Early Drift Detection Method // Proc. KDD-DS Workshop, 2006.
10. Hu Y. et al. FM-TS: Flow Matching for Time Series Generation // ICLR 2025 Submission, 2024.
11. Esteban C., Hyland S., Rätsch G. Real-Valued Time Series Generation with Recurrent GANs // arXiv:1706.02633, 2017
12. New York City TLC. TLC Trip Record Data 2019–2020 [Электронный ресурс]. — URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page (дата обращения: 13.05.2025).
13. Johnson A.E. W. et al. MIMIC-III, a Freely Accessible Critical Care Database // Scientific Data. — 2016.
14. Metz C. Fake It to Make It: Companies Beef Up AI Models with Synthetic Data. The Wall Street Journal. 2021. URL: wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601 (дата обращения: 13.05.2025).

Журнал «Современная наука: актуальные проблемы теории и практики»