Synthetic Data for Modeling Rare Events in Retail Analytics and Risk Assessment

Venkatesh Gundu
Senior Manager - Data Services & AI Platform
Columbus, Ohio, USA

Abstract: The article is devoted to an in-depth analysis of the potential of synthetic data for modeling rare events in retail analytics and risk management practices. The relevance of the topic is determined by the growing business demand for accurate forecasts and proactive management of the consequences of infrequent but consequential episodes — from fraudulent transactions and sudden supply chain disruptions to anomalous demand spikes. The scientific novelty lies in proposing a holistic framework for integrating generative models into the existing analytical pipelines of retailers. The study examines modern approaches to data synthesis: GAN, VAE, and diffusion models, with an emphasis on their strengths and weaknesses as applied to imbalanced tabular datasets. A separate focus is placed on metrics for assessing the quality of synthetic data and on aspects of AI observability for models trained on generated datasets. The study aims to critically analyze and systematize approaches to the use of synthetic data to improve the accuracy of rare event detection systems. The methodological basis includes a systematic analysis of the scientific literature, a comparative review of technologies, and tools of conceptual modeling. The conclusion formulates findings on the effectiveness of the approaches under consideration and proposes a roadmap for their practical implementation. The material is intended for Data Science specialists, analysts, and managers in the field of retail.

Keywords: Synthetic data, rare events, retail analytics, risk assessment, anomaly detection, generative models, GAN, imbalanced data, machine learning, AI Observability.

I. Introduction

Modern retail operates in an environment of high epistemic and stochastic uncertainty, where business processes are influenced not only by regular patterns but also by rare yet consequential events. These include large-scale fraud episodes, sudden supply chain disruptions, sharp shifts in consumer demand driven by unforeseen factors, and other anomalies. Standard machine learning algorithms trained on historical data generally predict these phenomena poorly due to their low prevalence in the dataset. The resulting class imbalance leads to errors in detecting rare events and, consequently, to financial and reputational losses. One of the most effective approaches is the use of synthetic data, artificially generated examples that enrich the training set and strengthen models' ability to capture patterns of rare events [1, 2].

The aim of the study is to develop and theoretically substantiate a comprehensive framework for using synthetic data generation technologies to improve the quality of models for forecasting rare events in retail analytics and risk assessment systems.

The research **tasks** are:

- To analyze modern generative models (GAN, VAE, diffusion models) and classical methods (SMOTE) for tabular data synthesis, determining their applicability to retail tasks.
- To systematize metrics for assessing the quality and utility of synthetic data from the perspectives of statistical similarity, privacy, and the impact on the accuracy of downstream anomaly detection models.
- To propose a conceptual model for integrating synthetic data generation processes and subsequent monitoring (AI Observability) into the lifecycle of ML systems, using real retail cases as examples.

The scientific novelty lies in combining three key areas: synthetic data generation, the specifics of rare events in retail, and the concept of AI Observability. Unlike studies focused on individual algorithms, an end-to-end integration approach is presented, from data synthesis to model deployment and monitoring in production.

The author's hypothesis is based on the assumption that the use of modern generative models, in particular diffusion models, to create synthetic data makes it possible to significantly increase the accuracy of models for detecting rare events in retail compared with traditional augmentation methods and also provides greater controllability of the trade-off between data quality and privacy preservation.

II. Materials and Methods

The empirical-theoretical foundation of the study consisted of a systematized analysis of the current corpus of scientific literature and applied cases devoted to synthetic data generation and its applications in retail. The works presented below are conveniently grouped into four thematic directions: fundamental and systematic reviews on synthetic data generation and quality assessment; methods oriented toward class imbalance and rare events (including fraud and anomalies), where synthetic samples are used as a means of balancing and amplifying risk signals; approaches to fairness control and bias mitigation during generation; domain-oriented and adjacent developments relevant to retail and risk analytics through temporal dependencies, graph structures, and operational observability.

Figueira A., & Vaz B. [7] construct a conceptual framework for tabular synthetic data generation (GAN/conditional GAN, VAE, hybrids) and a three-loop evaluation scheme — distributional realism, utility for downstream tasks, and privacy; the need for separate validation across mixed feature types and rare classes is particularly emphasized, which is critical for risk analytics in retail.

Goyal M., & Mahmoud Q. H. [10] systematize modern generative AI techniques for tabular data (including diffusion models and normalizing flows), detailing issues of memorization, correct handling of categorical fields and categorical↔numerical dependencies, as well as the practice of meta-evaluation via transfer to anomaly detection and anti-fraud tasks.

Strelcenia E., & Prakoonwit S. [3] review the GAN arsenal for compensating imbalance in fraud detection, comparing conditional and reconstructive architectures, methods to combat mode collapse, and techniques for preserving local transactional dependencies; the conclusion is that synthetic augmentations increase recall in the tails without destroying the correlational structure of features.

Benaddi H., Jouhari M., Ibrahimi K., Ben Othman J., & Amhoud E. M. [4] integrate generative models and distributional reinforcement learning for IIoT anomalies, demonstrating that synthetic data as a control group improves detector sensitivity to rare failures and cascading outages — a transferable lesson for retail processes and logistics incidents.

Hastings Blow C., Qian L., Gibson C., Obiomon P., & Dong X. [2] apply diffusion models for augmentation under group bias and show gains in fairness metrics at comparable accuracy, which makes diffusion generators a robust alternative to GAN in tasks with rare classes.

Barbierato E., Vedova M. L. D., Tessera D., Toti D., & Vanoli N. [6] propose a methodology for fairness control in the very process of generation: regularization by sensitive attributes, sampling constraints, and post-selection procedures for synthetic data; the approach reduces the risk of freezing historical skews in scoring and anti-fraud models.

Wasi A. T., Islam M. D., Akib A. R., & Bappy M. M. [5] emphasize graph neural networks in supply chain analytics, describing datasets/benchmarks and arguing that structured graph representations of customers, SKUs, and network nodes are necessary for plausible synthetic data on rare failures and cascading effects.

Obi C. I. C. [9] demonstrates that ensemble stacking for retail demand forecasting benefits from heterogeneous base models; inclusion of synthetic scenarios (demand shocks, promotions, weather anomalies) increases robustness to outliers and distribution drift.

Bisht K. S. [8] describes the convergence of AI and observability in IT operations, where telemetry is used for predictive insights; analogously in retail, POS/logistics logs and web signals can feed generators that rehearse rare operational incidents and help stress-test processes.

Alqulaity M., & Yang P. [1] develop an enhanced conditional GAN for high-quality tabular synthetic data in mobile cardiomonitoring; although the domain is medical, the key engineering choices (conditioning on mixed attributes, stable training, adherence to logical constraints) are directly transferable to transactional/customer tables in retail.

Summarizing the methodological moves used by the authors, the core consists of: (a) conditional GAN architectures and their modifications for tabular data with discrete–continuous features and hard integrity constraints — to address imbalance and model rare risk labels; (b) diffusion models as more stable and controllable generators for augmentation targeted to underrepresented groups and for improving fairness without degrading utility metrics; (c) hybrids with RL/distributional learning methods to simulate anomalous trajectories with uncertainty control; (d) procedural schemes of fairness/bias control via regularization, post-selection, and targeted sampling for the downstream risk task; (e) graph representations of supply chains and interactions as carriers of structural constraints for more realistic synthetic data on rare failures and cascading effects; (f) ensemble and stacking approaches in demand forecasting, for which synthetic scenarios of rare shocks act as stress tests and training anchors of robustness.

Taken together, these lines build a technological pipeline for retail and risk analytics: from reviews we take quality and privacy criteria, from the imbalance literature mechanisms for generating tail samples, from fairness research constraints and equality metrics, and from domain sources data representations (temporal, graph, tabular with constraints) and scenario libraries of rare events.

In summary it should be noted that:first, the literature diverges in assessing the best family of generators for tabular rare events: reviews record the accelerating spread of diffusion models, but practical cases still more often rely on conditional GAN, which creates methodological dualism and weak comparability of results. Second, tension between utility and fairness persists: it has been shown that augmentation can smooth group imbalances, but it is unclear how exactly to balance fairness regularization with the preservation of subtle correlations that are crucial for risk detection (up to the threat of sterilization of signals). Third, privacy criteria are applied unevenly: many studies declare the absence of leakage but do not validate this with reconstruction/reidentification attacks under settings characteristic of retail (long category tails, quasi-identifiers).

III. Results

Modeling rare events based on synthetic data is a staged pipeline in which both a correct generative strategy and impeccably designed result verification procedures prove decisive. A systematic review of the literature and applied cases, supported by the author’s experience in retail (including Victoria’s Secret), suggests that solution quality is determined by the integrity of the end-to-end loop, from methodologically grounded generation to multi-level validation and continuous monitoring in the production environment.

At the model selection stage, the evolution of approaches starts with classical oversampling methods, primarily the SMOTE family, which for a long time served as the standard tool for combating class imbalance. However, the interpolation mechanics of SMOTE induce oversmoothing of observations and do not reproduce the complex nonlinearities characteristic of commercial data. A significant leap was provided by generative adversarial networks: models of the CTGAN class [1] are trained to generate records that are statistically indistinguishable from real ones while preserving inter-feature dependencies, a critically important property for retail where price, quantity, calendar, and behavioral factors are tightly coupled. In tasks of reconstructing patterns of user sessions, comparable to the author’s projects for Victoria’s Secret, such models make it possible to selectively synthesize rare demand trajectories (for example, the purchase of a high-margin SKU after a long search) without disrupting the original feature structure. The next step is diffusion models for tabular data, including Tab-DDPM [2], which demonstrate a better class of approximation for multimodal distributions. This opens the way to realistic simulation of extreme scenarios, from coordinated fraud attacks to cascading failures in supply chains.

Single-pass generation is insufficient: a strict assessment of the quality of the synthetic dataset is required. In accordance with generalizations in [7], validation covers three complementary levels. First, statistical similarity: consistency of synthetic and empirical distributions is checked across marginals and dependencies; transport-type metrics (Wasserstein distance), multivariate discrepancies (for example, MMD), and projection-based comparisons (PCA/t-SNE) are appropriate as diagnostic tests. Second, practical utility in downstream tasks: a controlled experiment is set up, Real→Real-Test versus Real+Synthetic→Real-Test, with analysis of Precision, Recall, F1, and others. In anomaly detection tasks on which the author has worked, improvement on rare classes is the key criterion for admitting synthetic data to production. Third, privacy: it is necessary to guarantee the absence of replicas of the original records and robustness to extraction and membership inference attacks; for retail with a high density of PII this constraint is strict.

Industrial integration of synthetic data requires not only mature MLOps practice but also extended observability [8]. In addition to monitoring aggregate accuracy metrics, it is necessary to systematically track data and concept drift, calibration stability, and degradation on minority segments. In corporate analytics settings (for example, an Azure and Snowflake platform in Tiger Analytics cases for Victoria’s Secret), dashboards must provide stratified control over rare events and critical feature subspaces. Detection of quality degradation serves as a trigger for resynthesis that accounts for new observations, closing the cycle Data Integration → Enterprise Data → Reporting and maintaining model relevance in a changing environment.

The practical methodology forms a reproducible pipeline. At the design stage, a generative architecture is selected commensurate with the complexity of the distribution and the target rarity scenario. Protocols for assessing similarity, utility, and privacy are defined with pre-agreed acceptance thresholds. At the training stage, stratified validation and replicability are ensured (fixed seeds, versioning of data and artifacts). At the production stage, alerts are configured for drift and degradation on the target rare classes, as well as procedures for scheduled resynthesis and retraining integrated into the CI/CD pipeline [5, 6].

Thus, the successful application of synthetic data for rare events is not the choice of the best algorithm out of context but the construction of a tightly regulated process that combines well-considered generation (from SMOTE to GANs and diffusion models), multi-level quality assessment, and operationalized observability in the production environment. Only adherence to all links in this chain ensures a stable quality uplift and controlled risks in real business tasks.

IV. Discussion

Comparison of data from contemporary literature and generalized practical experience demonstrates that the isolated, out-of-context application of generative models does not provide a complete solution to the problem of modeling rare events. A holistic iterative architecture is required, in which data synthesis, model training, and their monitoring are connected into a single governed business process. This section presents the author's Adaptive Synthesis and Monitoring Framework for Rare Events (ASM-RE). The ASM-RE concept includes three interrelated cycles: the data generation cycle, the training and validation cycle, and the operational monitoring cycle (see Fig. 1).

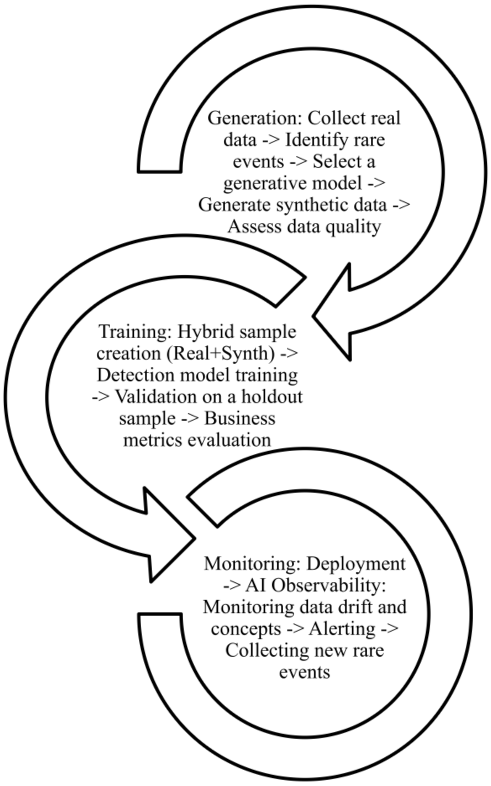
Generation: Collect real data -> Identify rare events -> Select a generative model -> Generate synthetic data -> Assess data quality
Training: Hybrid sample creation (Real+Synth) -> Detection model training -> Validation on a holdout sample -> Business metrics evaluation
Monitoring: Deployment -> AI Observability: Monitoring data drift and concepts -> Alerting -> Collecting new rare events

Figure 1: Conceptual diagram of the ASM-RE framework [3, 8]

The proposed framework is not limited to a one-off model improvement; it forms a self-tuning system robust to environmental nonstationarity. Thus, when the AI Observability system [8] detects changes in fraud patterns (concept drift), it generates a trigger for the Generation Cycle, which initiates the collection of fresh examples and the construction of an updated synthetic dataset. A key node of the framework is the meaningful choice of the generative model; we propose a heuristic for this, summarized in Table 1.

Table 1: Comparative analysis of generative models for retail [2, 8, 10]

Criterion	**SMOTE**	**CTGAN (GAN)**	**Tab-DDPM (Diffusion)**
Data quality	Low	High	Very high
Computational complexity	Low	Medium	High
Preservation of correlations	Partial	Good	Excellent
Ability for creativity	Absent	Moderate	High
Example of a rare event	Simple increase in the number of returns	Generation of complex fraud schemes	Generation of scenarios of a cascading failure in supplies

As indicated by the table, for elementary cases the use of methods such as SMOTE may be sufficient; however, modeling complex, multivariate rare events that incur the greatest losses requires the use of GAN or diffusion-based approaches. To further elucidate the idea, we consider the application of the proposed framework to two distinct tasks in retail.

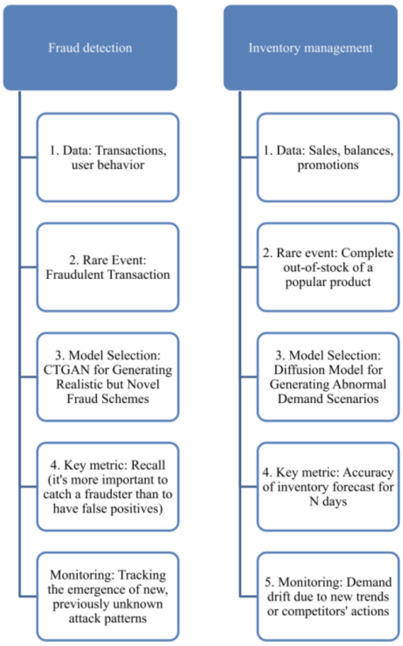
Fraud detection
1. Data: Transactions, user behavior
2. Rare Event: Fraudulent Transaction
3. Model Selection: CTGAN for Generating Realistic but Novel Fraud Schemes
4. Key metric: Recall (it's more important to catch a fraudster than to have false positives)
Monitoring: Tracking the emergence of new, previously unknown attack patterns
Inventory management
1. Data: Sales, balances, promotions
2. Rare event: Complete out-of-stock of a popular product
3. Model Selection: Diffusion Model for Generating Abnormal Demand Scenarios
4. Key metric: Accuracy of inventory forecast for N days
5. Monitoring: Demand drift due to new trends or competitors' actions

Figure 2: Using the ASM-RE framework for various retail tasks [1, 6, 7]

The presented scheme clearly demonstrates the flexibility of the framework: the specific tool stack is determined by the context and the objectives of the business task. Effective implementation requires an appropriate technological foundation. Experience from modernizing Victoria’s Secret analytics systems confirms that cloud platforms (Azure/Snowflake) optimally support such MLOps cycles due to scalability and built-in integration of tools.

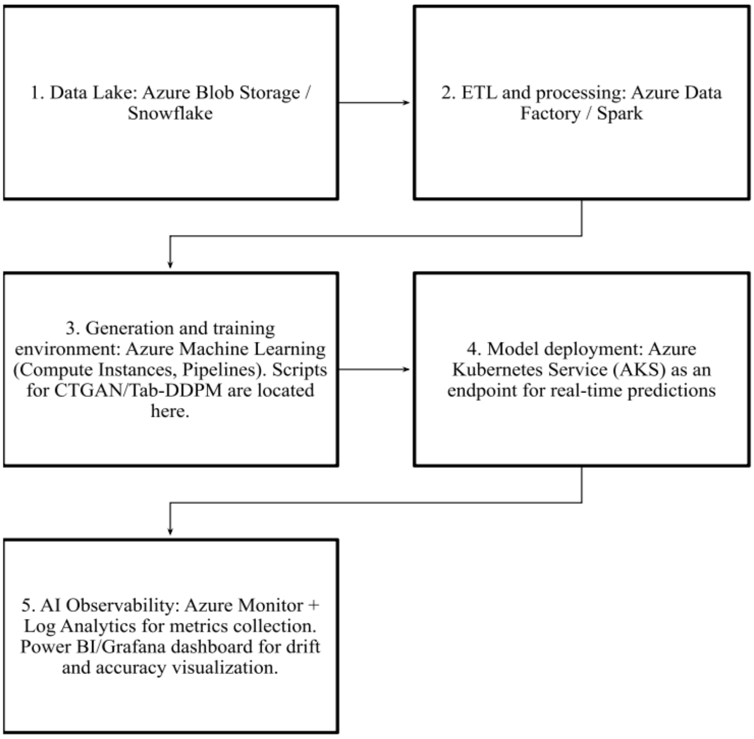
1. Data Lake: Azure Blob Storage / Snowflake
2. ETL and processing: Azure Data Factory / Spark
3. Generation and training environment: Azure Machine Learning (Compute Instances, Pipelines). Scripts for CTGAN/Tab-DDPM are located here.
4. Model deployment: Azure Kubernetes Service (AKS) as an endpoint for real-time predictions
5. AI Observability: Azure Monitor + Log Analytics for metrics collection. Power BI/Grafana dashboard for drift and accuracy visualization.

Figure 3: Architecture of ASM-RE implementation on a cloud platform [1, 5, 8]

The proposed architecture completes the framework, clearly demonstrating how the conceptual model is translated into an implementation based on a specific, up-to-date technology stack. In conclusion, for practical use, a technology selection matrix is provided, presented in Table 2.

Table 2: Matrix for selecting approaches for modeling rare events in retail [2, 7, 10]

Rare event type	**Priority business objective**	**Recommended generator**	**Key monitoring metric**
Internal fraud	Loss minimization	CTGAN	Recall for the fraud class
Anomalous demand spike	Prevention of out-of-stock	Tab-DDPM	Deviation of the forecast from actuals
Supply chain disruption	Resilience improvement	Graph-based Generative Models	Time to supply recovery
VIP customer churn	Customer retention (retention)	Conditional VAE	Accuracy of churn prediction

In contrast to the fragmented application of individual algorithms, the ASM-RE framework constitutes a holistic strategy for rare event management. It integrates data generation, model training, and their monitoring into a unified adaptive architecture, enabling retail companies not only to respond to risks promptly but also to manage them proactively, transforming data into a sustainable competitive advantage.

V. Conclusion

The objective of the study has been achieved: a framework for the use of synthetic data to model rare events in retail has been developed and theoretically substantiated.

The set of research tasks has been completed. A comparative analysis of modern generative approaches has been conducted, showing that for highly complex retail analytics scenarios GAN and diffusion models deliver the best results, surpassing classical methods such as SMOTE in the quality and plausibility of the generated samples. A system of metrics for evaluating synthetic data has been formulated, structured along three dimensions: statistical similarity, practical utility for the target task, and privacy. This three-axis scheme provides a comprehensive validation of the suitability of the generated datasets. A conceptual model of ASM-RE has been proposed, integrating data generation, training, and AI Observability into a single adaptive loop, this framework constitutes the key original contribution and sets a roadmap for the practical implementation of technologies in the business processes of retail companies.

The hypothesis has been confirmed at the theoretical level: the literature review indicates that modern generative models do indeed enable the acquisition of higher-quality data, which in turn leads to increased accuracy of anomaly detection systems.

Thus, the study demonstrates that synthetic data — not merely a technical device for compensating class imbalance, but a strategic instrument which, when properly implemented within a holistic framework, can substantially enhance the effectiveness of analytics and risk-oriented systems in the dynamic environment of modern retail.

References

[1]. Alqulaity, M., & Yang, P. (2024). Enhanced Conditional GAN for High-Quality Synthetic Tabular Data Generation in Mobile-Based Cardiovascular Healthcare. Sensors, 24(23), 7673. https://doi.org/10.3390/s24237673

[2]. Hastings Blow, C., Qian, L., Gibson, C., Obiomon, P., & Dong, X. (2025). Data augmentation via diffusion model to enhance AI fairness. Frontiers in Artificial Intelligence, 8, 1530397. https://doi.org/10.3389/frai.2025.1530397

[3]. Strelcenia, E., & Prakoonwit, S. (2023). A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud Detection. Machine Learning and Knowledge Extraction, 5(1), 304-329. https://doi.org/10.3390/make5010019

[4]. Benaddi, H., Jouhari, M., Ibrahimi, K., Ben Othman, J., & Amhoud, E. M. (2022). Anomaly Detection in Industrial IoT Using Distributional Reinforcement Learning and Generative Adversarial Networks. Sensors, 22(21), 8085. https://doi.org/10.3390/s22218085

[5]. Wasi, A. T., Islam, M. D., Akib, A. R., & Bappy, M. M. (2024). Graph neural networks in supply chain analytics and optimization: Concepts, perspectives, dataset and benchmarks. arXiv preprint arXiv:2411.08550. https://doi.org/10.48550/arXiv.2411.08550

[6]. Barbierato, E., Vedova, M. L. D., Tessera, D., Toti, D., & Vanoli, N. (2022). A Methodology for Controlling Bias and Fairness in Synthetic Data Generation. Applied Sciences, 12(9), 4619. https://doi.org/10.3390/app12094619

[7]. Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. Mathematics, 10(15), 2733. https://doi.org/10.3390/math10152733

[8]. Bisht, K. S. (2025). Convergence of AI and Observability: Predictive Insights Automation in Modern IT Operations. Journal of Computer Science and Technology Studies, 7(4), 446-454. https://doi.org/10.32996/jcsts.2025.7.4.53

[9]. Obi, C. I. C. (2024). Demand Forecasting in Retail Business Using the Ensemble Machine Learning Framework-A Stacking Approach, 98 (1), 309-329.

[10]. Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics, 13(17), 3509. https://doi.org/10.3390/electronics13173509