International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 4, Issue 1, 137-142, 2023
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V4I1P114

Original Article

Synthetic Data Generation Frameworks for Training Retail AI Models at Scale

Udit Agarwal[1], Aditya Gupta[2]
[1,2] Independent Researcher, USA.

Abstract - *The rapid expansion of Artificial Intelligence (AI) deployment in the retail sector necessitates robust, compliant, and scalable data infrastructure. Traditional reliance on raw, sensitive customer data poses significant legal, security, and operational challenges, severely impeding the training of large-scale predictive models. This paper provides an expert-level examination of modern synthetic data generation (SDG) frameworks designed to overcome these limitations. The analysis first categorizes SDG methodologies, emphasizing deep learning approaches such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Subsequently, the paper details advanced, domain-specific retail frameworks, including simulation platforms like RetailSynth, which fuse econometric discrete choice models with generative techniques to model realistic consumer behavior and operational constraints. For instance, literature reports that specialized GANs can generate realistic transactions by incorporating weighted stock constraints, a critical operational parameter often overlooked in general modeling. Finally, the paper articulates a comprehensive tripartite evaluation frameworkassessing Fidelity, Utility, and Privacywhich is essential for validating the analytical equivalence and trustworthiness of synthetic retail datasets. Fidelity metrics such as Wasserstein distance and Jensen-Shannon distance quantify statistical similarity, while Utility is assessed through predictive task performance (Accuracy, Lift, and Conviction). The successful implementation of these frameworks is critical for achieving competitive advantage through scaled, privacy-compliant AI applications like dynamic pricing and advanced demand forecasting. This paper reviews methodologies and frameworks reported in the literature, without presenting new experimental results.*

Keywords - *Synthetic Data, Retail AI, Generative Adversarial Networks, Scalability, Privacy Compliance, Consumer Behavior Modeling.*

1. Introduction

1.1. Context: The Imperative for Scalable AI in Retail

The modern retail landscape is fundamentally dependent on sophisticated AI models to maintain a competitive advantage. The application spectrum is broad and spans multiple machine learning paradigms. Supervised learning is commonly applied for crucial functions such as demand forecasting, inventory replenishment, and granular customer segmentation. Unsupervised learning methods are employed when data lacks clear labels, often used for clustering consumer behavior patterns and identifying anomalies within complex supply chains. Furthermore, Reinforcement Learning (RL) approaches enable models to improve over time by learning from feedback loops, proving particularly valuable in optimization domains like dynamic pricing and personalized product recommendations.

Recent advances in foundation models, a step-change evolution within deep learning, have introduced unprecedented capabilities. These models can process extremely large and varied sets of unstructured data, enabling new applications in retail, such as the creation of next-generation shopper interfaces. For example, Generative AI can enhance customer value management by delivering personalized marketing campaigns via interactive chatbots. Despite this technological potential, scaling AI model training presents significant hurdles. Training large-scale AI systems requires immense computational power and substantial storage capacity. High implementation costs and complexities involved in integrating modern AI tools with established legacy systems act as major organizational barriers. Moreover, effective deployment relies heavily on data management and quality assurance. This involves gathering vast amounts of data, cleaning it rigorously to eliminate errors, and ensuring sufficient diversity to prevent model biases that could lead to skewed outcomes. Crucially, the reliance on large volumes of real customer data introduces critical concerns regarding privacy and security, hindering the agility required for scalable model development.

1.2. Defining Synthetic Data and Foundational Value Propositions

Synthetic data is defined as artificially generated data that mathematically resembles real datasets but fundamentally lacks the personal, sensitive information contained in the original records. This artificial data captures the general patterns and statistical properties of the source data, but it is generated with enough algorithmic "noise" to mask the original data points while retaining

the properties necessary for training analytical models. Industry analysts have projected significant growth in synthetic data adoption for AI and analytics applications, underscoring its growing importance.

Synthetic data offers numerous value propositions for large enterprises seeking to scale their AI initiatives. Paramount among these is Privacy and Compliance. Since synthetic data is not linked to real individuals, it inherently eliminates the risk of exposing personal information, thereby ensuring privacy compliance. This capability transforms the compliance process from a constraint into an accelerator. Regulatory burdens traditionally slow down essential activities like data sharing; however, synthetic data facilitates compliant data sharing with external partners (e.g., fintechs or supply chain providers) for vendor performance evaluation and joint development while maintaining legal adherence. Internally, privacy regulations and access restrictions often delay data exchange for weeks; synthetic datasets can be shared freely and immediately across departments, such as marketing, product development, and operations, accelerating experimentation and innovation.

Another critical advantage is Scalability and Augmentation. Synthetic data can be produced efficiently in the large volumes required for robust machine learning applications. It allows researchers and developers to build and test machine learning models and software applications without compromising sensitive real data. Furthermore, it addresses data scarcity issues by filling gaps in real-world datasets and replacing historical data that is obsolete or otherwise unusable. This capability to generate data on demand ensures a sufficient, compliant input stream for continuous, large-scale AI model training.

2. Architecture of Synthetic Data Generation (SDG) Frameworks

2.1. Taxonomy of Generation Techniques

The field of synthetic data generation employs a hierarchy of techniques differentiated by their complexity and their ability to capture nuanced data distributions. The most straightforward approach is Rule-based approaches, which mimic real-world data by using predefined rules, constraints, and statistical distributions. While effective for simple, well-defined datasets, this method is limited in its ability to model high-dimensional, complex distributions found in modern transactional data. Statistical modeling represents a more advanced approach, relying on capturing explicit mathematical relationships between variables in real data to generate comparable characteristics in the synthetic output. The most sophisticated category is Machine learning-based techniques, which are considered state-of-the-art. These techniques, primarily based on deep neural networks, excel at probabilistic modeling and capturing complex, latent data distributions necessary for producing highly realistic data samples.

2.2. Deep Learning Methodologies: The Role of GANs and VAEs

Deep learning methodologies are crucial for generating the high-fidelity synthetic data required for training large-scale retail AI models. These models can generate diverse data modalities, including tabular data, images, radiomic data, and bio-signals. Generative Adversarial Networks (GANs) are one of the most widely utilized techniques today for artificial data generation. A GAN architecture operates by training two neural networks in an adversarial setting: a Generator and a Discriminator. The Generator attempts to create synthetic data instances, while the Discriminator attempts to distinguish between real and synthetic data. This min-max competition forces the Generator to produce output data that closely resembles the complex distributions of the real-world dataset, resulting in high-fidelity synthetic data.

Variational Autoencoders (VAEs) represent another foundational type of neural network used for synthetic data generation. VAEs capture complex data distributions using probabilistic modeling. By encoding the data into a latent space and subsequently decoding it, VAEs can produce realistic data samples that maintain the statistical characteristics comparable to the source data. The broad recognition of synthetic data's value has increasingly driven research toward building generative models that handle data in its raw tabular form, rather than relying solely on modeling features derived from aggregated or transformed data. This focus is essential in retail, where AI systems depend on high-granularity datasuch as individual transactions and customer trajectoriesto derive meaningful insights. Modeling raw tabular data directly ensures that the subtle interdependencies and relationships between features (e.g., correlations, temporal sequencing) are accurately preserved, which is vital for complex retail applications like fraud detection or precise customer journey simulation.

2.3. Classification of Synthetic Data: Fidelity vs. Privacy Trade-offs

The connection between the generated synthetic data and the real source data governs the resulting trade-off between privacy guarantees and analytical validity.

Fully synthetic data is fabricated data that has no actual connection to any real observations. This data is created solely through algorithms. It is utilized in scenarios where no real data is available, or crucially, when models require absolute confidentiality guarantees. Because there is no link to the original records, disclosure risk is eliminated.

Partially synthetic data is generated by combining real data values with fabricated ones. In these datasets, some true values remain, which results in higher analytical validitymeaning the dataset better mirrors the source's analytical properties. However, retaining true values increases the disclosure risk, necessitating careful management.

Table 1: Generative Machine Learning Frameworks for Synthetic Data

Framework	Architecture/Mechanism	Key Capability	Data Modalities Supported
Generative Adversarial Networks (GANs)	Generator/Discriminator trained in a min-max adversarial setting	Capturing complex data distributions for high-fidelity generation	Tabular data, MRI images, radiomic data, bio-signals
Variational Autoencoders (VAEs)	Encoder/Decoder capturing complex distributions via probabilistic modeling	Producing realistic data samples comparable to real data characteristics	Tabular data, images, bio-signals

3. Domain-Specific SDG Frameworks for Large-Scale Retail

While general-purpose deep learning models provide the foundational capability for data synthesis, large-scale retail AI requires frameworks that integrate domain-specific operational and behavioral constraints to ensure utility and realism.

3.1. Operational Realism: Integrating Constraints in Transaction Generation

Operational fidelity is critical; generative models must produce data that respects real-world limitations. Standard models often fail to incorporate logistical constraints, which compromises the practical applicability of the resulting AI models. To address this, specialized deep learning architectures have been developed. An innovative approach involves using a GAN framework specifically designed to generate synthetic transactions *under stock constraints*. This framework represents an advancement in modeling transactions within constrained systems, with potential implications for retail operations and strategy.

Some approaches diverge from conventional methodologies by integrating stock-keeping unit (SKU) data directly into the GAN architecture. The training process involves extracting real orders and, critically, stock embeddings. During the Discriminator training loop, both real orders and weighted stock information are utilized. The incorporation of this weighted stock information significantly enhances the realism of the generated data, as models must reflect the dynamic interplay between consumer demand and SKU availabilityan aspect frequently neglected in simulation.The source study reported that models incorporating weighted stock information demonstrated reduced distribution divergence (measured by lower Earth Mover's Distance (EMD) and Jensen-Shannon distance (JSD)), demonstrating a clear enhancement in generating realistic transactions relevant to inventory management and supply chain optimization.

3.2. Behavioral Realism: Simulation of the Full Customer Life-Cycle (RetailSynth)

Retail AI often requires understanding and predicting consumer response to policy changes (e.g., pricing, promotions). This necessitates frameworks that model complex consumer decision-making processes accurately. The RetailSynth framework is a specialized simulation environment that tackles these challenges by integrating econometric principles with generative models. Its foundation lies in fusing econometric-style generative models with Discrete Choice Models (DCM), which are rooted in utility theory. DCMs are recognized as essential tools in retail marketing for informing marketing-mix decisions. They are structured to jointly model purchase incidence, brand choice, and purchase quantity, which is required to fully capture the complexity of the consumer decision-making process.

RetailSynth addresses three key challenges specific to retail marketing:

1. Extending econometric-style generative models to cover the entire customer life-cycle, linking policies to individual decision-making from store visitation through to item purchase.
2. Creating realistic differences in price sensitivity across large numbers of customers and products, essential for precision pricing.
3. Generating highly scalable synthetic customer trajectories for vast numbers of products and consumers efficiently.

This framework, which was carefully calibrated on publicly available grocery data, is specifically designed for applied researchers to validate causal demand models for multi-category retail. It enables the incorporation of realistic price sensitivity into emerging benchmarking suites for sophisticated applications like personalized pricing, promotions, and product recommendations.

The complexity of dynamic pricing and promotion optimization often involves Reinforcement Learning (RL) algorithms. These systems must predict not just correlations in behavior but the *causal* impact of policy changes (e.g., how a price increase

causes a behavioral shift). By simulating causal models via DCMs within RetailSynth , the synthetic data supports the development and evaluation of AI systems capable of making prescriptive decisions, moving the retail sector beyond simple predictive modeling.

4. A Comprehensive Tripartite Evaluation Framework for Data Quality

To transition synthetic data from a research curiosity to a core production-scale asset, a rigorous and quantitative evaluation framework is indispensable. Academic literature proposes a comprehensive tripartite evaluation framework focused on assessing three critical dimensions: Fidelity, Utility, and Privacy. This framework ensures that the generated data provides the robust analytical equivalence required for training trusted AI models at scale.

4.1. Assessing Fidelity: Distribution Similarity Metrics

Fidelity measures how accurately the synthetic dataset replicates the known statistical properties and data distributions of the original data. This assessment must differentiate between continuous and discrete data attributes.

For measuring Marginal Distribution Similarity in numerical features (continuous data), the Wasserstein distance is reported. The Wasserstein distance, or Earth Mover's Distance, quantifies the difference between two probability distributions. A small value for this metric is a strong indicator that the synthetic dataset's distribution closely adheres to the real dataset's distribution. For categorical features (discrete data), the Jensen-Shannon distance is computed to quantify the degree of similarity between distributions. Similarly, a small Jensen-Shannon distance value is expected to indicate an excellent synthesizer capability.

Beyond marginal distributions, the framework assesses Joint Distribution Similarity to ensure that critical feature interdependencies and co-variances are preserved, which is vital for complex modeling. This is measured using the Pearson correlation matrix for number-to-number interactions, Theil’s U matrix for category-to-category interactions, and the correlation ratio for number-to-category interactions.

4.2. Assessing Utility: Performance in Predictive and Association Tasks

Utility is demonstrated by the synthetic data’s effectiveness in training models for real-world retail applications, thereby proving its value in predictive analytics and strategic planning. The evaluation framework formalizes two essential tasks to measure utility.

The first is a Classification Task, which scrutinizes the predictive power of models trained on synthetic data. An example task is identifying whether a customer will engage in a high-value purchase (e.g., purchasing more than 10 products) based on demographic and unit price data. The performance of the trained classification model (such as a Bagging Classifier) is validated using standard metrics, including Accuracy, F1 score, Receiver Operating Characteristic (ROC) area, precision, and recall.

The second task is Product Association Analysis, which is a form of Market Basket Analysis. This involves analyzing product affinities using algorithms like Apriori to ensure that significant product relationshipswhich underpin personalized recommendation and merchandising strategiesare maintained in the synthetic data. The key metrics reported for this analysis are Lift (which measures the likelihood that product B is bought when product A is purchased) and Conviction (which compares the probability of product A appearing without product B assuming independence, versus the actual frequency). These metrics confirm the preservation of associative patterns critical for retail optimization.

4.3. Assessing Privacy: Differential Privacy and Confidentiality

Privacy assessment is crucial for ensuring the widespread applicability and compliance of synthetic data frameworks. Privacy is safeguarded using Differential Privacy principles and related metrics. Differential Privacy ensures that the presence or absence of any single data record in the training set does not significantly alter the output of the model, providing strong confidentiality guarantees.

The evaluation framework mandates a comparison of the synthetic data’s proximity to both the original training data and an unseen holdout dataset. Robust privacy guarantees require the synthetic dataset to approximate both datasets equally well. This balanced approximation confirms strong privacy protection while ensuring adequate data utility. The fundamental security level against potential linkage attacks is assessed using the metric Distance to the closest record.

The combination of quantitative statistical measures (Wasserstein distance) and verifiable business metrics (Lift) provides quantitative proof of *analytical validity*. This comprehensive evaluation framework represents the essential bridge that allows synthetic data to transition from a conceptual solution to a trustworthy, production-scale asset.

Table 2: Evaluation Metrics for Synthetic Retail Data Fidelity and Utility

Assessment Dimension	Metric Category	Specific Metrics Used	Purpose in Retail Validation
Fidelity	Distribution Similarity (Numerical)	Wasserstein Distance	Measures closeness between synthetic and real continuous distributions.
Fidelity	Distribution Similarity (Categorical)	Jensen-Shannon Distance	Quantifies similarity of discrete feature distributions.
Utility	Classification Task Performance	Accuracy, F1, ROC, Precision, Recall	Validates model effectiveness in predictive tasks (e.g., identifying premium customers).
Utility	Product Association Analysis	Lift and Conviction	Assesses the preservation of significant product affinities (Market Basket Analysis).
Privacy	Confidentiality Guarantee	Differential Privacy, Distance to Closest Record	Ensures robust privacy protection and mitigates linkage risks.

5. Conclusion

The adoption of synthetic data generation frameworks represents a strategic priority for large-scale, compliant deployment of high-performing AI models within the retail industry. By providing data that accurately maintains the statistical characteristics of real-world transactions while eliminating direct privacy risks, synthetic data generation unlocks critical advantages, including the ability to scale computational resources efficiently, accelerate both internal and external data sharing, and adhere rigorously to confidentiality mandates.

State-of-the-art frameworks leverage advanced deep learning techniques, primarily Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to capture complex, high-dimensional data distributions required for retail modeling. Crucially, the implementation of these frameworks must integrate domain-specific operational and behavioral realism. Specialized implementations, such as the GAN model incorporating weighted stock constraints and the RetailSynth simulation environment leveraging discrete choice models (DCMs) , exemplify the necessity of fusing generative algorithmic power with econometric and logistical principles. These advanced frameworks directly address core retail challenges, specifically the modeling of the full customer life-cycle and realistic price sensitivity heterogeneity.

The adoption of a comprehensive tripartite evaluation framework focusing rigorously on Fidelity (using statistical measures like Wasserstein distance), Utility (validated via predictive and association tasks like Lift analysis), and Privacy (assessed via Differential Privacy) is indispensable for securing enterprise adoption. This rigorous validation provides quantitative evidence supporting the analytical equivalence of synthetic data establishing the necessary foundational trust required for scaling AI solutions that drive critical business outcomes such as personalized marketing, dynamic pricing, and optimized supply chains. Future research must continue to focus on enhancing the efficiency and realism of these specialized, constraint-aware simulation environments to ensure the frameworks keep pace with the increasing complexity and scale of global retail operations.

References

[1] Zhao, Z., Wu, H., Van Moorsel, A., & Chen, L. Y. (2023). *VT-GAN: Cooperative Tabular Data Synthesis using Vertical Federated Learning*. arXiv preprint arXiv:2302.01706. Introduces a GAN-based framework that synthesizes tabular data in a privacy-preserving federated learning setting for structured datasets.

[2] Hansen, L., Seedat, N., van der Schaar, M., & Petrovic, A. (2023). *Reimagining Synthetic Tabular Data Generation through Data-Centric AI: A Comprehensive Benchmark*. arXiv preprint arXiv:2310.16981. Proposes a synthetic data generation and evaluation framework guided by data profiling and benchmarking across multiple tabular datasets.

[3] Shen, X., Liu, Y., & Shen, R. (2023). *Boosting Data Analytics With Synthetic Volume Expansion*. arXiv preprint arXiv:2310.17848. Presents a Synthetic Data Generation for Analytics framework analyzing statistical performance and privacy trade-offs for synthetic datasets.

[4] Lampis, A., Lomurno, E., & Matteucci, M. (2023). *Bridging the Gap: Enhancing the Utility of Synthetic Data via Post-Processing Techniques*. arXiv preprint arXiv:2305.10118. Introduces post-processing pipelines to improve the representativeness of synthetic data generated by GANs for downstream model training.

[5] Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). *Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations*. In *Proceedings of EMNLP 2023* (pp. 10443–10461). Evaluates LLM-based synthetic text data generation and implications for performance, useful for task-specific AI model training.

[6] Sun, Z. (2023). *Query Aware Synthetic Data Generation*. University of California, Berkeley Technical Report UCB/EECS-2023-124. Proposes synthetic dataset generation sensitive to query distributions to improve representativeness for analytics tasks.

[7] Lu, Y., Wang, H., & Wei, W. (2023). *Machine Learning for Synthetic Data Generation: A Review*. arXiv preprint arXiv:2302.04062. Systematically surveys deep generative methods and frameworks for synthetic data across domains.

[8] Deng, H. (2023). *Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems*. UNIDIR report. Reviews synthetic data frameworks and their applicability to diverse AI systems.

[9] Aydore, S., Qian, Z., & van der Schaar, M. (2023). *Synthetic Data Generation with Generative AI*. NeurIPS 2023 Workshop. Discusses emerging frameworks and challenges in leveraging generative models for synthetic data across domains.

[10] Jadon, A., & Kumar, S. (2023). *Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy*. *SmartNets 2023*. Explores generative models (GANs, VAEs) for privacy-aware synthetic generation—framework insights relevant for structured data contexts like retail.