Computer Fraud and Security
ISSN (online): 1873-7056

Synthetic Data for Counterfactual Targeting in Regulated Industries

Arjun Sirangi

Advance Analytics Manager

Abstract:

Privacy constraints, data sparsity, and ethical considerations limit focused actions using consumer data in regulated businesses including healthcare, banking, and insurance. This study examines counterfactual targeting, which replicates "what-if" situations to evaluate policy or marketing decisions at the person level using synthetic data. Organisations may predict alternative outcomes without compromising user privacy or compliance by creating high-fidelity synthetic datasets that maintain statistical features and causal linkages of real-world data. The study uses synthetic data creation and causal inference algorithms to assess treatment effects across varied populations. Case studies show how this strategy promotes strategic decision-making while protecting data. The findings show that synthetic data may be used to innovate predictive modelling and individualized decision-making in high-stakes, regulated situations while protecting privacy.

Keywords: Synthetic Data, Counterfactual Targeting, Regulated Industries, Causal Inference, Data Privacy

1 Introduction:

Privacy rules, ethical considerations, and compliance requirements restrict the analysis and targeting of personal data in healthcare, finance, and insurance. This makes data-driven tactics, notably counterfactual targeting, difficult for organisations to implement. An effective option is synthetic data, which is intentionally manufactured to match real-world data structure and statistical trends. Synthetic data helps design fair, transparent, and legally acceptable targeting systems by enabling safe experimentation and analysis without disclosing sensitive information. It helps organisations to examine various methods and results while protecting privacy and complying with regulations when paired with causal inference techniques.

1.1 Problem Statement:

Legal, ethical, and privacy constraints hinder data modelling, experimentation, and decision-making in highly regulated businesses. GDPR, HIPAA, and financial compliance standards rarely incorporate individual-level data for counterfactual targeting. Modern data analysis and utilisation differ substantially. Risky targeting strategy testing can skew models, wasteful treatments, and unexpected policy outcomes. Data is needed for ethical, explainable, and regulatory-compliant targeting. We must study synthetic data creation for counterfactual reasoning without violating privacy or laws.

1.2 Research Objectives

1. Specifically, we want to look at how synthetic data might help regulated businesses with counterfactual targeting.

2. To create a system that integrates approaches for causal inference with those for generating synthetic data.

3. So that we can see how well synthetic data and real data perform in terms of counterfactual forecast accuracy.

4. In order to determine whether or not the use of synthetic data for targeting techniques complies with ethical and regulatory standards.

2. Theoretical Foundation and Context

Recent years have seen data-driven decision-making fundamental to operational efficiency and personalisation efforts across sectors. However, privacy restrictions and ethical concerns restrict the use of genuine consumer or patient data in regulated areas like healthcare, finance, and telecommunications. This hinders the use of advanced analytical methods like counterfactual targeting, which simulates alternative situations to optimise results.

2.1. Understanding Counterfactual Targeting

The term "counterfactual targeting" is the practice of analysing marketing activities, policy shifts, or risk assessments through the lens of hypothetical "what-if" situations. To find out what would have happened if a different group had been targeted, for example, counterfactual analysis looks at potential outcomes rather than just looking at past data. This method is vital for improving decision-making in fields where there are ethical, legal, or practical constraints on conducting direct experiments.

2.2. Challenges in Regulated Industries

Data governance and compliance regimes are quite stringent in industries including healthcare, banking, and telecoms. Privacy worries, legal hazards, and the possibility of abuse are the reasons why these frameworks limit the usage of actual user data for these purposes. Therefore, it becomes difficult to do counterfactual analysis without breaking restrictions, since it frequently depends on precise data at the individual level.

2.3. The Role of Synthetic Data

An alternative that does not compromise privacy is synthetic data, which is created using machine learning models like Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs). It conceals personally identifiable information (PII) while simulating real-world data statistical features. Safe experimentation, causal inference, and bias evaluation are all made possible by synthetic data when combined with counterfactual modelling tools. The combination of the two can aid regulated companies in improving algorithmic fairness, simulating policy consequences, and testing intervention strategies—all while maintaining compliance.

3. Privacy Concerns in Financial Data Sharing

At the outset of this part, we will address the financial domain-specific dangers associated with data sharing. After this, we will go over the literature on privacy attacks in machine learning. Next, we'll go over how privacy assaults relate to financial sector privacy risks. Lastly, we go over the many privacy defences that can be used on the original data or included in the process of creating synthetic data, and how they guard against privacy threats. These tiers are known as privacy levels.

3.1.Privacy risks in the finance domain

Financial institutions have regulations and internal guidelines for data sharing between lines of business and externally to protect clients' sensitive information and protect firms from MNPI, litigation, reputation, and competitive risks. We analyse major dangers and applicable policies in this section.

Fair Credit Reporting Act (FCRA) Consumer reporting companies (e.g., credit bureaus) cannot disclose information to anybody without a reason specified under the FCRA. In particular, removing identifying data fields is not adequate. Additionally, one must verify that additional data fields, algorithm results, and/or publicly available information cannot betray the identity.

Regulation on Unfair, UDAAP—Deception or Abuse If utilised or shared counter to consumer or client elections or representations, sharing and certain uses of identifiable data may pose UDAAP concerns. Consumer privacy preferences govern sharing identifiable data in numerous circumstances.

Litigation risks Inappropriate publication of data or functions of data (e.g., models trained on data, insights from data, or synthetic data matching these datasets) that reveal personally identifying information or global characteristics may invite litigation. Use of external vendor data is usually limited by contracts that specify its use.

Competitive risks Publishing customer data or data about industries and publicly traded firms a corporation is interested in may offer competitive, antitrust, and insider trading hazards. This applies to synthetic data published.

3.2. Privacy attacks in machine learning literature

Privacy attacks can target ML models and outputs [SZZ+23]. The assumption of a privacy assault is a malevolent adversary trying to obtain private information from model output. This research examines synthetic model output. Privacy attacks vary. Each attack assumes what information the enemy has, what has to be protected, what the purpose is, etc. Here, we briefly review some of the most important attacks on financial privacy.

Table 1 : Privacy attacks on synthetic data can violate financial restrictions.

	FCRA	UDAAP	Litigation Risk	Competitive Risk
Membership Inference Attack	Applicable	Applicable	Applicable	N/A
Attribute Inference Attack	Applicable	Applicable	Applicable	N/A
Property Inference Attack	N/A	N/A	Applicable	Applicable
Model Inference Attack	Applicable	Applicable	Applicable	Applicable

- **Membership inference attacks (MIAs)** In many circumstances, an individual's data alone can reveal sensitive information. MIA [SSS16] requires the adversary to detect an individual in the training dataset via a data processing approach like an ML classifier or synthetic data generator. An adversary who knows an individual is in the dataset can employ linkage attacks (reconstruction attacks) to identify sensitive attributes of that individual.

- **Reconstruction attacks (attribute inference attacks)** An adversary with partial knowledge of a set of features undertakes reconstruction attacks to retrieve sensitive features or the entire data sample. The attribute inference attack when an adversary uses a public set of non-sensitive characteristics to infer sensitive attribute values is an example [NS07].

- **Property inference attacks** Property inference extracts dataset properties from synthetic data. If synthetic data is available, property inference can learn any summary statistic of the original data (e.g. mean value, quantiles, histograms, etc.). Preventing property inference attack reduces synthetic data fidelity [LWSF23].

3.3. Defences against privacy attacks

This section discusses privacy defences. Privacy layers, a new hierarchy of privacy defences, are also proposed.

Anonymization/PII concealment Many methods use personal identifying information (PII) obscuration or anonymisation of sensitive sectors. These are vulnerable to linkage attacks, hence they do not ensure privacy.

Randomisation Data swapping via randomisation prevents adversaries from inferring data information with certainty, providing plausible deniability. Some or all data points are swapped between unique dataset people. Randomisation often protects privacy while maintaining the downstream task's usability or a dataset query's accuracy.

Differential privacy Differential privacy defence is a randomisation strategy. It theoretically assures that a possible adversary with algorithm output knowledge (e.g. fake data) cannot determine if a particular individual was in the input dataset. Let Xn represent the universe of n-entry datasets. D and D′ ∈ Xn are nearby if they differ in one data entry, i.e. individual. Let M be a mechanism that takes Xn datasets and outputs synthetic datasets. We define M as (ϵ,δ)-differentially private if any neighbouring datasets D and D′ and any set C of mechanism M results show

$$\mathbb{P}(\mathcal{M}(D) \in \mathcal{C}) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(D') \in \mathcal{C}) + \delta. \quad (1)$$

For δ = 0 and low ϵ values, ϵ-differential privacy ensures that every output from the mechanism is about equally likely to be observed on nearby databases. Alternatively, (ϵ,δ)-differential privacy ensures that privacy loss is limited by ϵ for surrounding datasets with probability at least 1−δ. We may now introduce our privacy framework.

3.4. Exploring Privacy Levels in Data Security

We now analyse privacy assaults, utility implications, and plausible privacy guarantees for each level of a six-level privacy defence hierarchy. Defence systems with better privacy protections increase with each level. These levels can advise firms on Synthetic Data security and use. They may enable internal exchange of Level 2 data from non-critical sources but demand Level 4 security for sensitive material. To balance corporate goals, security, speed of generation, and utility, the use case should define the privacy level. We discuss strategies that change data from the original dataset to Synthetic Data in the first four tiers. The original data is on the left, and the arrows show how it gets converted. These examples use tabular data, but the ideas apply to various data types.

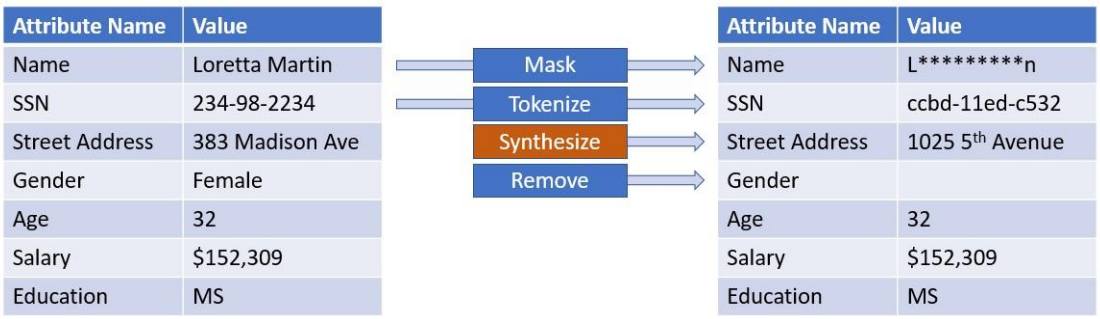

Figure 1 : Privacy Level 1: Obscure PII

3.4.1. Privacy

Level 1: Hidden PII These strategies include removing, replacing, masking, or anonymising PII attributes. This method does not degrade downstream task utility because it does not affect non-PII properties. This weakens privacy protection since reconstruction attacks can compromise data.

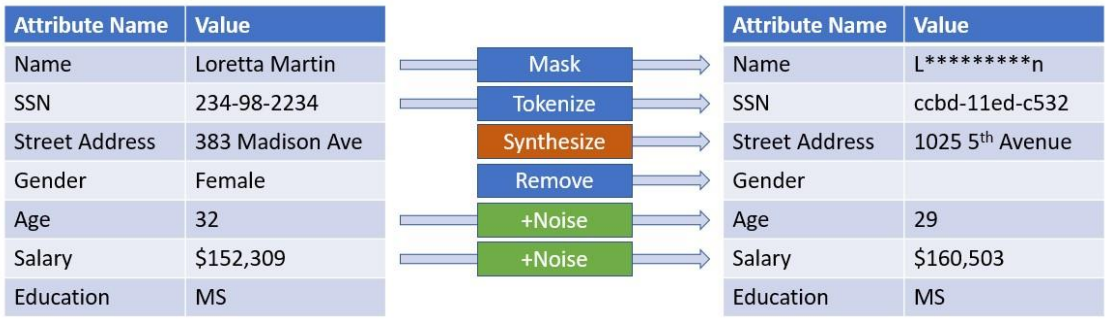

Figure 2 Privacy Level 2: Obscure PII + noise

3.4.2. Privacy Level 2:

PII obscured + noise We can also add noise to other properties to deter attacks in addition to hiding PII columns. Differential privacy approaches can assure MIA in writing. Randomly "swapping" data between entries is another method. For instance, a demographic dataset may randomly reshuffle individuals' ages. This method provides plausible yet randomised data by making it harder for an adversary to infer personal information. These methods improve privacy while keeping data utility for downstream tasks. Utility degradation depends on noise and downstream task. obfuscated or noisy row-by-row transcription of the source data. Thus, such datasets cannot exceed the original.

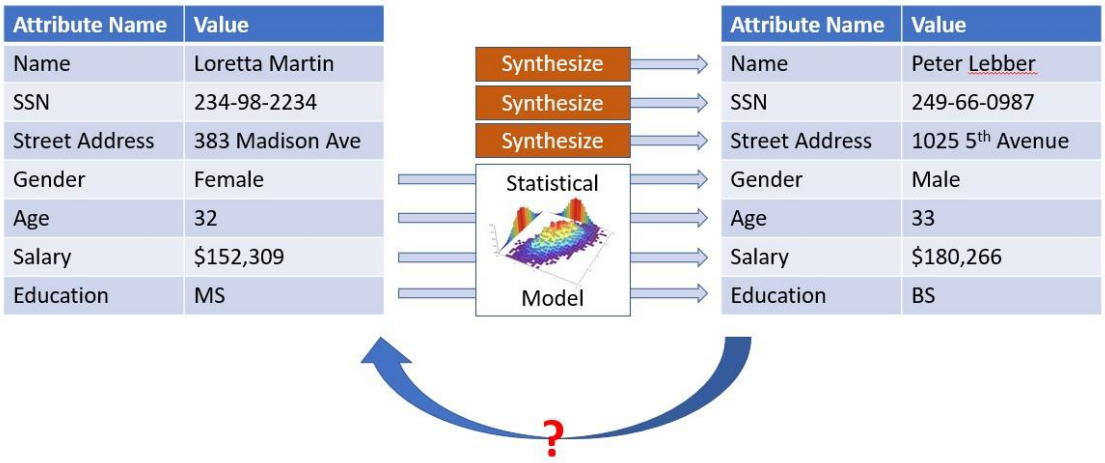

Figure 3 : Privacy Level 3: Generative modeling. The question mark suggests the possibility of reverse-engineering the data.

3.4.3. Privacy Level 3:

Generative modelling Note that Privacy Levels 1 and 2 involve Level 3 uses generative approaches to analyse original data and create a model that generates new data. Gaussian copula and generational adversarial networks (GAN) are examples. Other methods employ differential privacy to provide more guarantees. We model data using a KD-tree-based formulation that provides extra protections. All of these methods create new data components. They have more protection than Levels 1 and 2, but are still vulnerable. Large relative sizes of created and original data increase risk: If we generate one million samples from a 1,000-sample dataset, we anticipate them to cluster around the original data.

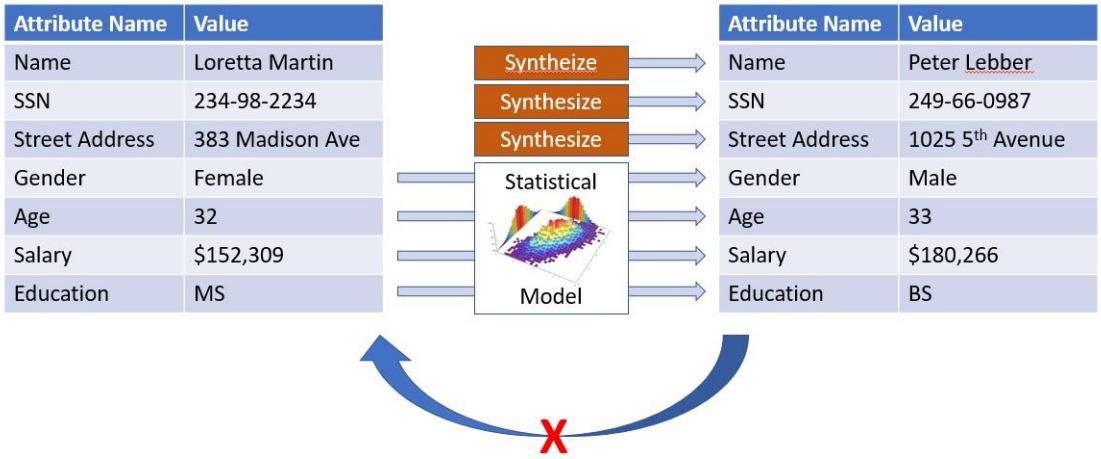

Figure 4 Privacy Level 4: Generative modeling + testing

3.4.4. Level 4:

Testing + generative modelling Level 4 adds explicit assessment of each created dataset for specific attack resistance. The data and application determine the exams and scores needed to "pass". Some data attributes may "leak" while others should not. To implement this, we use known attack techniques and score data based on attack success.

It is difficult to determine which test and score are needed to achieve Level 4 privacy in all circumstances, but the fact that the data is specifically examined is crucial. Each business must choose the use case test and score criteria. Resistance to membership inference, attribute reconstruction, and property attacks are scored. among others.

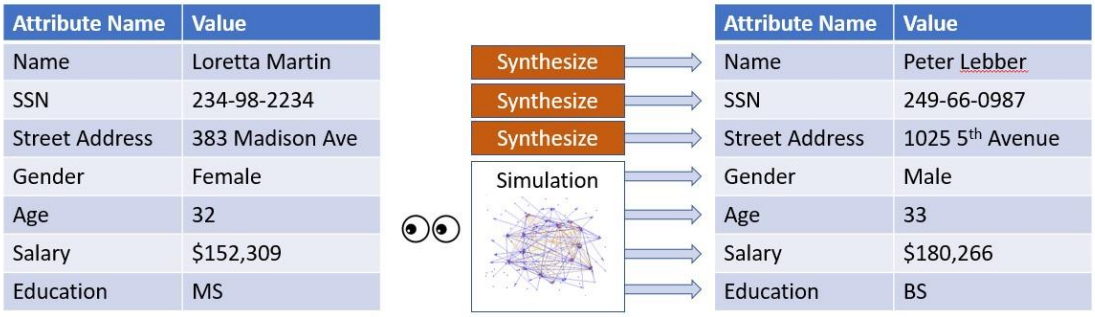

Figure 5 : Privacy Level 5: Calibrated simulation

3.4.5. Level 5:

Simulation calibrated Here, the generating algorithm is not trained on real data. Usually, this strategy does not teach. In place of real data, we use simulations regulated by rules or process expertise. These criteria are tuned to the real process such that the generated data follows specific statistical features of the real system. We could simulate the stock market to get stock price data. Our work includes calibrated equity market simulations at Level 5. privacy Downstream task and simulation framework determine utility deterioration. This defence is usually effective against adversaries. The simulator is calibrated to statistical properties of the real system, hence they may be vulnerable to Property Inference Attacks.

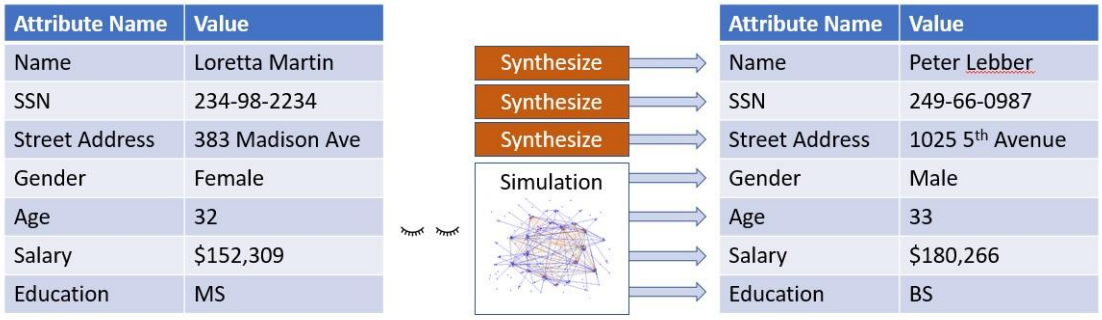

Figure 6 Privacy Level 6: Uncalibrated simulation

3.4.6. Level 6:

Uncalibrated simulation In this case, we may not know the statistical properties of the simulated system or may resist altering the simulation to match the real system. Even without high-fidelity data, a simulation can be useful. We could simulate all data field values to check if they "break" our downstream procedures during testing. Our systems may identify fraudulent transactions by integrating known cases. Other uses include creating what-if scenarios to see whether visualisation can show how one thing affects another. This approach usually protects privacy. It fixes a level 5 PIA defence issue: uncalibrated statistical attributes to the real dataset.

3.4.7. Privacy Levels to Guide Synthetic Data Use Cases

Corporations can employ these privacy levels. Levels one and two can exchange data after removing private data. Data and testing boost AI models in levels three and four. Levels five and six assist software engineers test apps and more. Section 6 of time-series will leverage ABIDES for level five synthetic data applications. Level 6 data can be used for software testing, proof of ideas, hackathons, app stress testing, and data migrations. Software Application Stress Testing Case Study Stress testing apps for performance is a common software engineering challenge. Financial systems may encounter unexpected trading volume increases owing to market instability like Brexit or Covid-19. Early testing ensures system stability and helps understand how supporting technology will perform during issues. Production data cannot be tested due to privacy concerns. We created level 5 calibrated simulation-based synthetic data with similar statistical properties and eliminated confidential data. Millions of rows of data were generated to test the algorithm for trade surges. It simplified stress testing.

4. Synthetic Data Generation for Financial Tabular Data

Here, we zero in on tabular data, a sort of financial data that is almost everywhere, and discuss topics including privacy, fairness, synthetic data generation, and the robustness of downstream classifiers.

4.1. Generation

Real-world domains are commonly described using tabular data. These datasets need synthetic data to address data shortage and quality issues while retaining their statistical properties and linkages across categories. Synthetic data integrates several sources, tests new ideas without changing real data, and makes sharing private. Produced data has major limitations. Synthetic data is better than training data, but is it free? Establishing a system to answer this query without performance degradation is necessary.

Model tabular data distribution with statistics, machine learning, Bayesian networks, and neural networks. Synthetic data creation approaches have distinct benefits. Many factors, including observed data distribution and synthetic data purpose, determine the best method. Knowing marginal distributions improves statistics. Everyone disagrees on how to solve new dataset and use case problems. Model selection via Bayesian optimisation was recently proposed.

We compare CTGAN, TVAE, and CopulaGAN neural network models. Since its launch, GAN research has presented novel optimisation approaches and improvements to overcome restrictions.

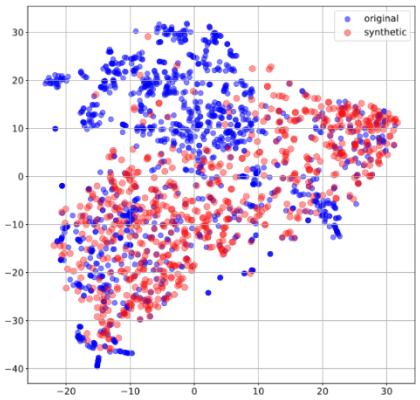

Figure 7 : TSNE plot showing similarity between original data and synthetic data.

GAN models like CTGAN improve on previous designs. Mode-specific normalisation captures multimodal and non-Gaussian distributions. This approach solves the problems of earlier GAN architectures like severely imbalanced category columns and sparse one-hotencoded vectors with a conditional generator and sampling. CopulaGAN modifies CTGAN using CDF-based transformation. TVAE uses variational autoencoders for neural networks. Analysing each table column as a random variable, constructing a multivariate probability distribution, and sampling statistically produces synthetic data.

4.1.1. Optimization Method

Because they disregard downstream tasks, most synthetic generation approaches are “unsupervised”. Label variable is treated like other factors in most methods. These methods create dataset-“similar” models. It conflicts with use cases that value downstream predictions over data similarity. Oversight and Composition To address the aforesaid issues, SC-GOAT is a new synthetic data generation method that optimises directly on the downstream loss function. This procedure is two-step. We start with a downstream-specific supervised component and apply Bayesian optimisation to fine-tune neural network hyperparameters. Meta-learning gives us the optimal mixture distribution of synthetic data generation methods in step two. The SC-GOAT method combines earlier synthetic data production technologies to generate synthetic data.

Table 1: Description of data sets.

Data set	Label	Observation	Continuous	Binary	Multi-class	Label = 0	Label = 1
Credit Balanced	“Class’	50,000	30	1	0	66.70%%	33.3%

4.2.Application: Credit Card Fraud

This section evaluates our machine learning-based fraud detection generative models. Synthetic tabular data is shown to be valuable using credit card fraud data3. Two-day European cardholder transactions from September 2013 are included. The dataset is skewed, with 492 frauds out of 284807 transactions, 0.172%. Card fraud datasets contain just numerical input variables with 31 features. V1–V28 are PCA fundamental components for privacy and secrecy. Only Time, Amount, and Class are unaffected by PCA. 'Time' contains the seconds between each transaction and the dataset's first. Goal variable 'Class' is 0 for no fraud, 1 for fraud. Transaction amount is

'Amount'. To balance the credit fraud dataset's class imbalance ratio, we oversampled the minority class and randomly undersampled the majority class. This needed repeating minority class instances to balance majority and minority classes. We used SMOTE.

Table 2 : Percentage class of synthetic datasets using each model.

Synthesizer	Class Frauds (1)	Class No Frauds (0)
Original	33.33%	66.67%
Gaussian Copula	41.9%	58.1%
Copula GAN	74.76%	25.24%
CTGAN	74.76%	25.24%
TVAE	40.07%	59.93%
Empirical Copula	33.81%	66.19%
DS 0	0.173%	99.83%
DS 0.1	45.47%	54.53%
SC-GOAT	36.10%	63.90%

Evaluation: Various benchmarking methods allow loss function modification for synthetic data production. This study will evaluate generative models later. Our generative models will train and test fraud detection algorithms, generate synthetic data from the real dataset, and evaluate using AUROC. Balanced model performance is shown. This study helps us identify the optimal synthetic data generator and fraud detection methods.

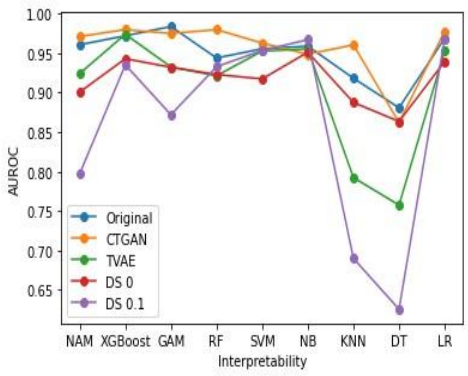

Figure 8 : Utility metrics for fraud detection classifiers. See text for details.

Homogeneous simulated datasets without conditional sampling from one trial. The original dataset is more balanced than CTGAN and Copula GAN databases. TVAE, Gaussian Copula, and Empirical Copula create class-balanced datasets that meet original and generated data performance. This picture shows the AUROC findings of multiple data models developed using different methodologies. GAM and decision tree models thrive on the original dataset.

Table 3: Average, standard deviation, and one-sided paired t-test for downstream test AUC score using XG Boost on 10 experiments for each approach.

		Untuned			Tuned	
Method	average	std	test statistic	p-value average	std	test statistic p-value
Gaussian Copula	94.45%	0.01	14.10	0 94.45%	0.01	14.31 0
CTGAN	95.34%	0.01	16.21	0 95.93%	0.01	13.15 0
Copula GAN	95.50%	0.01	14.18	0 96.41%	0.01	7.80 0

TVAE	98.52%	0.00	0.00	0.5	98.48%	0.00	0.00	0.5
SC-GOAT	98.52%	0.00	-	-	98.48%	0.00	-	-

With neural-network-based generative models CTGAN and TVAE, XGBoost runs better. When trained on neural network data, XGBoost detects fraud better. After the XGBoost model performed well, an XGBoost classifier was trained on the training data set and tested on a validation data set. AUROC results for various approaches. The results are from 10 experiments using 70% real data for training, 20% for validation, and 10% for testing. SCGOAT finds the optimal mixed distribution of synthetic data production methods.

4.3.Privacy

Maintaining financial data privacy is essential for compliance. Privacy-preserving synthetic tabular data generation is important in ML for finance because most financial data is tabular. Table privacy concerns imply rows represent people and columns reflect attributes. Here, we examine the most important privacy issues. Sometimes essential column values must be protected. Provide the source dataset and any output (including fake data) for sensitive property learning. Secure dataset global statistics like quantiles or correlations is another issue. Last, most tabular data privacy literature analyses its MIA, or person identification, capabilities. MIA is prevented by differential privacy. Later section discusses differentially private tabular synthetic data generation.

In statistical-based tabular data, Laplace mechanism perturbation often produces differential privacy. The PrivBayes and PrivSyn fit low-order marginals to graphical models. Privacy-preserving deep generative models using DP-SGD have been studied in synthetic data generation. Finance requires interpretability, which these methods lack. Recent interpretability studies use space partitioning and noise perturbation to create differentially private synthetic data. Privacy in tabular data diffusion models is examined.

Privacy in Credit Card Fraud Use case We wish to demonstrate the tradeoff between noisy disruption-measured privacy and synthetic data utility in a downstream activity. Deep learning mimics data-dependent space partitioning and noisy perturbation to generate DP synthetic data from the original data input.

Dataset: We use all features except time in our credit card fraud dataset. We create synthetic data using 80% of input data for varying privacy budgets ϵ. Test 12 classifiers on 20% of the input data after training on synthetic data.

Evaluation: We demonstrate average ROC degradation over 20 repeats. The [KNP+23] algorithm does not surpass DP-MERF in ROC values. This is expected since it doesn't use deep generative models. Unlike the DP-MERF, their performance falls slowly as privacy increases.

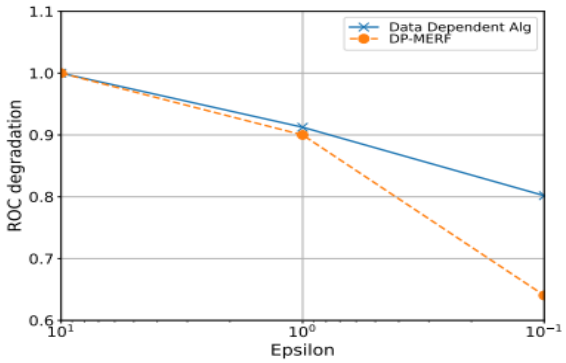

Figure 9 : Analysing [KNP+23] and [DP-MERF [HAP21] data algorithms for downstream categorisation. The degradation of ROC is measured as a ratio between the ROC for a certain budget and the ROC for ϵ=10.

5. Temporal Sequences and Event Data Analysis

Sequences, time, and features are recorded in event series data. Event series data usually contains numerous events of different sorts, and any event at time t may depend on preceding events. Unlike time series data, event series data are asynchronous since the time between events is frequently variable. Financial event series data are prevalent. Commercial banks track customer interactions via "customer journeys". Trading limit order books record buy and sell orders at a specific price or better. Customer impressions in marketing applications track ad views and purchases. Infectious illness patterns, earthquake logs, social media, and criminal models employ event series data. Modelling complicated consumer interactions using graphs complicates generation. Recent study created large-scale graphs using diffusion theories.

5.1.Using Automated Planning

One way to generate synthetic event series data is to model the environment and simulate situations. Simulation events become output dataset data. One way is classical automated planning. Banks likely acquire most customer data. Every client interaction is observable by the bank. Examples include account opening, wire transactions, and ATM withdrawals. A client's status can be described by their account balance, accounts opened, and regular payments at each time step. Each time step, clients have short-term financial goals like paying rent, buying a product, or receiving pay cheque. Since clients' interactions are defined by actions, states, and goals, we may use an automated planning framework to randomly generate client goals, build plans to achieve them from their current state, and execute those plans. Each action execution leaves a trace that can be included in an interaction dataset. Several studies have used this method's data for entanglement reasoning and goal recognition. This strategy yielded money laundering, fraud, and customer journey datasets.

Customer journey sample from planning-simulated trace. Client id, date-time tag, and event label are typical event descriptions. The event description includes the client channel—mobile in this case—and action.

Table 4: Small section of a generated trace of events using AI planning.

Date and time	Event	Customer ID
2021-05-24 21:22:14	mobile : logon	ID-22522
2021-05-24 21:25:14	mobile : transaction summary business	ID-22522
2021-05-24 21:26:14	mobile : transaction history prepaid account	ID-22522
2021-05-24 21:28:14	mobile : ultimate rewards info	ID-22522
2021-05-24 21:31:14	mobile : ultimate rewards activity	ID-22522
2021-05-24 21:36:14	mobile : log off	ID-22522

5.2. Application:

Artificial customer trip marketing spending Data with simulated customer journeys improves marketing campaigns. Marketing customer journeys includes search, social media, and TV commercials. Multichannel conversions affect how many consumers buy marketed goods, according to post-mortem expenditure research. MTA models attribute budget allocation to each channel, demonstrating its importance. MTA models consider the whole customer journey, unlike previous methodologies that concentrated on the last interaction. See examples. Temporal point procedures improve MTA model insights, customer journey dynamics, and training data.

Synthetic event series data is valuable in a public MTA dataset4. Monthly campaign data covers Facebook, Instagram, Paid Search, Online Display, and Search. We used a Markov chain MTA model to assess 10,000

customer journeys with a ~ 7% marketing conversion rate. XGBoost estimates client conversion using client trip data. Synthetic event series data has real data dynamics and improves prediction. Augmenting decay parameters and mixture number with 4 Hawkes processes optimises them independently. shows XGBoost model AUC versus data augmentation % in training (shaded bands indicate one standard deviation computed over 10 runs). After 50%, synthetic customer journeys enhance classifier performance but plateau. MTA credit assignments for 50% of fictional customer journeys. Synthetic travels boost internet video advertising's credit rating without modifying credit assignment.

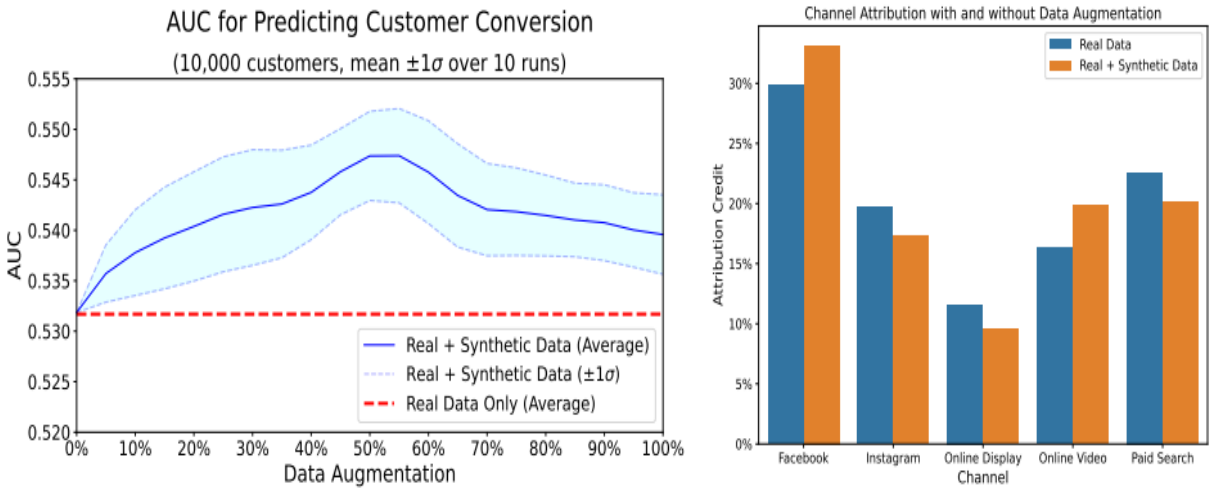

Figure 10 : Using actual and synthetic data to enhance event series data (left) and adjust MTA credit assignments (right).

Future Research

- Synthetic data can transform the financial industry by enabling privacy-preserving innovation in fraud detection, customer acquisition, distributional market shift modelling, constrained time-series generation, and OCR document automation. These use cases show synthetic data can bridge regulatory compliance with data-driven intelligence. However, major issues remain. Future research should strengthen the fidelity and generalizability of synthetic data across different financial contexts, while ensuring data utility and privacy, addressing bias in generated datasets, and improving downstream model robustness.

- Standardised privacy, fairness, and model robustness benchmarks should also be prioritised. Interdisciplinary initiatives combining finance, machine learning, and regulatory knowledge are needed to match technical breakthroughs with legal and ethical standards. Responsible and scalable deployment will unlock synthetic data generation's full potential once it is integrated into real-world financial institutions.

Conclusion

Synthetic data enables ethical counterfactual targeting in regulated businesses with data privacy, regulatory compliance, and limited data access issues. Synthetic data lets organisations model decision-making scenarios, evaluate initiatives, and refine targeting methods without violating regulations by duplicating real-world datasets' statistical features without disclosing sensitive information. This strategy enhances data-driven decision-making openness, fairness, and accountability while supporting advanced analytical models and causal inference. Synthetic data bridges innovation and regulation in healthcare, banking, and insurance as demand for personalised, compliant, and explainable systems rises.

References

1. Abhishek, V., Despotakis, S., & Ravi, R. (2017). Multi-channel attribution: The blind spot of online advertising. SSRN. https://doi.org/10.2139/ssrn.2959778

2. Abar, S., Theodoropoulos, G. K., Lemarinier, P., & O'Hare, G. M. P. (2017). Agent based modelling and simulation tools: A review of the state-of-art software. Computer Science Review, 24, 13–33. https://doi.org/10.1016/j.cosrev.2017.03.001

3. Alaa, A. M., Chan, A. J., & van der Schaar, M. (2020). Generative time-series modeling with Fourier flows. International Conference on Learning Representations (ICLR).

4. Alaa, A. M., van Breugel, B., Saveliev, E., & van der Schaar, M. (2021). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. arXiv preprint arXiv:2102.08921. https://arxiv.org/abs/2102.08921

5. Arava, S. K., Dong, C., Yan, Z., & Pani, A. (2018). Deep neural net with attention for multi-channel multi-touch attribution. arXiv preprint arXiv:1809.02230. https://arxiv.org/abs/1809.02230

6. Ardon, L., Vann, J., Garg, D., Spooner, T., & Ganesh, S. (2022). Phantom–An RL-driven framework for agent-based modeling of complex economic systems and markets. arXiv preprint arXiv:2210.06012. https://arxiv.org/abs/2210.06012

7. Arjovsky, M., Chintala, S., & Bottou, L. (2023). Wasserstein generative adversarial networks. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning (Vol. 70), 214–223. PMLR.

8. Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S., & Kaafar, M. A. (2019). Differentially private release of high-dimensional datasets using the Gaussian copula. arXiv preprint arXiv:1902.02938. https://arxiv.org/abs/1902.02938

9. Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Balch, T., Reddy, P., & Veloso, M. (2023). Generating synthetic data in finance: Opportunities, challenges and pitfalls. Proceedings of the First ACM International Conference on AI in Finance, 1–8. https://doi.org/10.1145/3383455.3422553

10. Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., Balch, T., & Veloso, M. (2020b). Generating synthetic data in finance: Opportunities, challenges and pitfalls. Proceedings of the First ACM International Conference on AI in Finance, 1–8. [Duplicate reference]

11. Bamford, T., Fons, E., El-Laham, Y., & Vyetrenko, S. (2023). MADS: Modulated autodecoding SIREN for time series imputation. arXiv preprint arXiv:2307.00868. https://arxiv.org/abs/2307.00868

12. Babkin, P., Watson, W., Ma, Z., Cecchi, L., Raman, N., Nourbakhsh, A., & Shah, S. (2023). BizGraphQA: A dataset for image-based inference over graph-structured diagrams from business domains. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3539618.3591712

13. Baum, L. E., & Petrie, T. (2023). Statistical inference for probabilistic functions of finite state Markov chains. The Annals of Mathematical Statistics, 37(6), 1554–1563. https://doi.org/10.1214/aoms/1177699147

14. Baüerle, N., & Rieder, U. (2020). Markov decision processes with applications to finance. Springer.

15. Belgodere, B., Dognin, P., Ivankay, A., Melnyk, I., Mroueh, Y., Mojsilovic, A., Navartil, J., Nitsure, A., Padhi, I., Rigotti, M., et al. (2023). Auditing and generating synthetic data with controllable trust trade-offs. arXiv preprint arXiv:2304.10819. https://arxiv.org/abs/2304.10819

16. Borovajo, D., & Veloso, M. (2020). Domain-independent generation and classification of behavior traces. arXiv e-prints, abs/2011.02918. https://arxiv.org/abs/2011.02918

17. Borovajo, D., Veloso, M., & Shah, S. (2020). Simulating and classifying behavior in adversarial environments based on action-state traces: An application to money laundering. In Proceedings of the First ACM International Conference on AI in Finance. https://arxiv.org/abs/2011.01826

18. Bouchaud, J.-P., Bonart, J., Donier, J., & Gould, M. (2018). Trades, quotes and prices: Financial markets under the microscope. Cambridge University Press.

19. Box, G. E. P., & Draper, N. R. (2020). Essentially, all models are wrong, but some are useful. Statistician, 3(28), 2013.

20. Brandimarte, P. (2014). Handbook in Monte Carlo simulation: Applications in financial engineering, risk management, and economics. John Wiley & Sons.

21. Byrd, D. (2019). Explaining agent-based financial market simulation. arXiv preprint arXiv:1909.11650. https://arxiv.org/abs/1909.11650

22. Byrd, D., Hybinette, M., & Balch, T. H. (2019). ABIDES: Towards high-fidelity market simulation for AI research. arXiv preprint arXiv:1904.12066. https://arxiv.org/abs/1904.12066