2023 S.T. Yau High School Science Award
Research Outline

Consumer Data Without Compromise: Integrating Differential Privacy and GANs for Privacy-Preserving Digital Marketing

In the rapidly evolving digital marketing landscape, the utilization of consumer data is essential for efficient targeting and personalization of marketing practices. However, the growing concerns regarding user privacy and stringent data protection regulations have created challenges in accessing and using consumer data for marketing purposes. **This paper introduces a novel approach that leverages Differential Privacy and Conditional Tabular Generative Adversarial Networks (CTGAN) to address these privacy concerns while maintaining the efficacy of data-driven digital marketing strategies.** Our approach amalgamates the strengths of Differential Privacy and CTGAN, applying differential privacy to the original dataset to ensure that extracted data cannot be tied back to individuals. We then train a CTGAN on several open marketing dataset, learning and generating synthetic data that closely resembles real-world consumer behavior. Through extensive empirical analysis, we evaluate the fidelity, utility, and trade-offs of our approach, demonstrating its effectiveness in synthesizing non-Gaussian and multi-modal distributions, and its applicability in real-world classification problems. The research also highlights the complexity of hyperparameter tuning and the importance of a balanced approach in model training. Our findings contribute valuable insights to both the theoretical understanding of generative models and practical guidance for digital marketing practitioners.

Keywords: Digital Marketing, Privacy, Consumer Data, Differential Privacy, Generative Adversarial Network

References:

"Quarterly Retail E-Commerce Sales Quarter 2023 - Census.Gov," U.S. Census Bureau News, August 17, 2023, https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf.

P.K. Kannan and Hongshuang "Alice" Li, "Digital Marketing: A Framework, Review and Research Agenda," *International Journal of Research in Marketing* 34, no. 1 (2017): 22–45, https://doi.org/10.1016/j.ijresmar.2016.11.006.

"Art. 5 GDPR – Principles Relating to Processing of Personal Data," General Data Protection Regulation (GDPR), October 22, 2021, https://gdpr-info.eu/art-5-gdpr/.

"California Consumer Privacy Act (CCPA)," State of California - Department of Justice - Office of the Attorney General, May 10, 2023, https://oag.ca.gov/privacy/ccpa.

Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-

customer-centric-approach-to-marketing-in-a-privacy-first-world.
"Privacy Noun - Definition, Pictures, Pronunciation and Usage Notes: Oxford Advanced American Dictionary at Oxfordlearnersdictionaries.Com," privacy noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced American Dictionary at OxfordLearnersDictionaries.com, accessed August 21, 2023, https://www.oxfordlearnersdictionaries.com/definition/american_english/privacy#:~:text=%2Fˈpra%C9%AAv%C9%99si%2F,I%20value%20my%20privacy.

1. INTRODUCTION

The marketplace today has undergone drastic reforms with the digital market emerging as a critical aspect of the U.S. economy. In the 2022 U.S. Census Bureau report, e-commerce accounts for an average of 14.65 percent of total U.S. retail sales and(adjusted for trading-day differences and moving holidays) are estimated to reach 15.1 percent by the first quarter of 2023 (adjusted for seasonal variation, but not for price changes).[i] Such trends are well reflected in the business landscape, where multinational technological corporations such as Google, Facebook, Amazon, Alibaba, and Uber are becoming key competitors.[ii] As a result, the prevalence of the digital market has motivated companies to highlight a more interactive and personalized customer experience through multi-channel digital marketing, in which consumer data play an essential role. However, digital marketing using big data has introduced a dilemmatic status quo. On one hand, data-based marketing campaigns largely enhance the consumer experience with tailored offerings and reduced informational asymmetries. On the other hand, consumers are growing alert against efforts made at the firm level to collect and leverage personally identifiable data, and such privacy concerns in turn diminish the utility of digital marketing.

In response, regulatory measures such as European Union's General Data Protection Regulation (GDPR)[iii] and California Consumer Privacy Act (CCPA)[iv] mandates strict restriction against the collection of consumer information and subsequent algorithm analysis. These regulations nonetheless operate at the cost of diminishing profit and marketing inefficiency, requiring 10-20 percent greater investment for companies to obtain the same level of return.[v] Moreover, certain privacy-unconcerned consumers are

turning toward third-party intermediaries in the private market, where they may exchange personal information for value in return.[vi] In general, there remains room for refinement in the study of privacy concerns and relevant solutions.

The purpose of this study is to propose an alternative measure of privacy preservation that preserves the fundamental utility of data-based digital marketing while in compliance with government and corporate privacy regulations. Primarily, we adopt the conditional tabular generative adversarial network (CTGAN) to synthesize high-quality consumer data for algorithmic analysis. The application of differential privacy in the training procedure of CTGAN act as a privacy buffer between the training dataset and the operator, ensuring that even an untrustworthy operator can not gain access to consumers' identifiable information. In this manner, the proposed methodology seeks to achieve the dual objective of profit maximization and privacy protection in data-based marketing.

This study presents several interdisciplinary contributions to the existing field of economics and machine learning. First, it improves on the existing commercial use of GAN learning using CTGAN as a suitable variant to generate real-world tabular data. Second, a comprehensive evaluation assessing the quality and trade-offs of different CTGAN customization is displayed over sample datasets, with modified metrics made accessible for further application. Lastly, a customized random forest classifier integrated with numerous models is used to analyze the predictive accuracy of the synthesized datasets to avoid possible distortion of overfitting and class imbalance.

This paper is structured as follows. Section 2 begins with a comprehensive review of existing literature in the study of privacy concerns and current measures of privacy

preservation, as well as GAN application in relevant areas. Section 3 introduces the principal structure of GAN and differential privacy. In addition, we show that CTGAN is suitable in the case of discussion and presents the set of procedure to implement CTGAN on Python Jupyter Notebook, along with customizations that are made specifically for the chosen dataset. Section 4 discusses the evaluation metrics and compares results yielded from differentially private synthesized datasets. Finally, Section 5 discusses possible implications and presents the general conclusion on this topic. Section 6 acknowledges contributions.

2. LITERATURE REVIEW

2.1. Defining the concept of privacy in the digital market

The semantic definition of privacy refers to the "state of being alone and not watched or disturbed by other people" in the Oxford Dictionary[vii], except this concept gains extended significance and entitlements in the contemporary digital market.

Stone and Stone presented a summarizing framework of prior definitions, concluding a necessary overlapping of three elements in an accurate definition of privacy[viii]. The first attribute pertains to "information control". Goffman encapsulates privacy as the regulation of "identity information" under different social situations.[ix] In more recent studies, this is specified as the selective disclosure of information and its subsequent dissemination. The importance of information control, therefore, rests in the capacity to limit others from extracting knowledge on the said individual, preventing the acquisition of past, present, and, possibly, deduced future intentions.[x] In a similar vein, various

works associated privacy with controlling the desirable amount of interaction one has with other people. Laufer and Wolfe argue that the amount of social interaction one receives is dependent on the role and respective socio-physical environment he/she partakes.[xi] In particular, this perspective has strong implications in the socio-physical environment of digital marketing as companies adopt calculative algorithms and multi-channel communications to interact with their customers; the former assigning specific role (sport-fan vs athlete) based on collected personal data and the latter sending tailored advertisements via various channels of media platform. Lastly, one may also generalize information and interaction control as determinants of an individual's degree of autonomy and freedom in the digital market. Initially, Goffman's seminal work in 1955 highlights that an individual is manipulated when others possess information about he/she.[xii] Privacy as autonomy and freedom in the digital market is therefore guaranteed when the individual can manage information and interaction disclosure to prevent external manipulation. Customers nowadays may not be entirely private as their online behaviors are constantly tracked by companies and their data are collected to generate targeted advertisements.

In essence, the comprehensive definition of privacy describes a consumer's rightful boundaries that safeguard the outflow of their information and social interaction to ensure an individual's freedom in the digitalized environment. The breaching of such boundaries is shown to raise considerable concerns.

In terms of machine learning, consumer privacy is guaranteed when their information is in a state of "differentially private". In particular, the concept of differential privacy is a promise that external attempts to extract personally related information from a

population of data can not be achieved and that individual consumers are protected from being personally identified.[xiii]

2.2. Privacy concerns and negative reactance

An early poll conducted by Equifax in 1992 reveals that 79% of American customers hold privacy concerns, while 55% believe that "protection of information about consumers will get worse by the year 2000" (Equifax).[xiv] Recent literature shows evidence that such distrust has been exacerbated over time. In their 2013 study, Pingitore et al found that a majority of surveyed subjects deem access to personal data through online cookies and social media as inappropriate, with 81% of customers believing that they do not have control over companies' usage of their personal data.[xv] In terms of organizational responsibility, Brodherson et al reported that only 33% of Americans trust that companies are using their information responsibly, whereas 25% hold a neutral response by unawareness.[xvi]

In the study of consumer psychology, privacy concerns are shown to be founded upon different impetuses. Notably, Smith et al provides a framework on the multidimensional nature of privacy concerns in 1996[xvii], and the taxonomy was further adapted to the domain of information privacy in the digital environment by Malhotra et al.[xviii] This literature review primarily focuses on Malhotra's modified construct of IUIPC.

One of the three dimensions that IUIPC integrates from Smith et al's traditional CFIP framework involves privacy concerns that arise from the excessive collection of personal data. In particular, Phelps et al found that 85.6% percent of respondents want to limit the

amount of personal data collected by marketers.[xix] A study by Cespedes and Smith reports an idiosyncratic level of "privacy threshold" that would raise considerable concern if trespassed, even if individuals may be willing to exchange personal data for benefits.[xx]

Another dimension of privacy concerns relates to consumer control. In the same study by Phelps et al, 84 percent of respondents expressed the desire to have more control over personal data to avoid commercial advertisement.[xxi] Research by Nowak and Phelps shows that consumers are less worried about data collection when they are explicitly given the option to opt out.[xxii] Conversely, an individual's inability to manifestly control how their personal data is being collected, leveraged, and subsequently disseminated at the firm level generates privacy concerns. Culnan, for instance, discovered that consumers who negatively view unauthorized secondary use of information are more likely to perceive their privacy as being invaded.[xxiii] Unfortunately, a study by Enonymous.com reports that only 3.5 percent of 30,000 investigated websites never shared personal information with a third party, with roughly 22,000 websites that do not provide privacy policies at all.[xxiv] Such massive redistribution of identifiable databases would likely give rise to more privacy concerns as technology renders data more accessible and easier to exchange.

Thirdly, an individual's understanding of organizational practices of personal data constitutes a crucial dimension in consumer privacy concerns. According to an analysis by Hoffman et al, 63 percent of consumers refuse to provide personal information because they do not trust cyber markets, whereas 69 percent of respondents who opt for mistrust do not provide information due to their lack of knowledge regarding its usage.[xxv]

Consistent with these findings, Phelps et al also report evidence that consumers seek more information and greater transparency regarding the organizational leverage of personal data.[xxvi]

As a result, privacy concern in the digital environment is highly heterogenous and context-specific. The specificity of the situation in which consumers' concerns toward their information are raised, therefore, leads to a variety of negative reactance. In the specific context of online advertising, for instance, Goldfarb and Tucker found that the combination of contextual (targeting) ads with obtrusive ads draws reduced purchase intent compared with when two respective types of ads are displayed independently.[xxvii] White et al found, aside from reduced purchase intent, consumer reaction to personal marketing communication can manifest in communication avoidance, information falsification, and derogatory word-of-mouth.[xxviii] Norberg and Horne also demonstrate that consumers with privacy concerns are more likely to submit falsified information.[xxix] Overall, these negative reactance uniformly leads to the reduced efficacy of digital marketing and data utility.

2.3. Privacy Preservation Measures

Regulations to ensure consumer privacy in the collection and analysis of personally identifiable data generally categorize into government enactments or corporation policies. The former is represented by numerous existing legislations, such as the EU's General Data Protection Regulation (GDPR),[xxx] California Customer Private Act (CCPA),[xxxi] Health Insurance Portability and Accountability Act,[xxxii] and other privacy proposals; The latter includes Google's recent removal of website cookies and Apple's built-in

features of minimizing database accessibility to third-party service providers.[xxxiii] These regulations present several downfalls. For one, prohibiting the collection and marketing operation of consumer data at a granular level negatively affects the efficacy of marketing campaigns. According to an investigation conducted under the implementation of GDPR by Goldfarb and Tucker, banner ads experienced a 65 percent reduction in ineffectiveness on display in European countries under GDPR, whereas this pattern is not discovered among ads released in non-European countries.[xxxiv] A study by Brodherson estimates that the phasing out of website cookies will likely cause marketers 10 to 20 percent more spending to generate the same returns.[xxxv] For another, regulatory measures do not pertain to all consumers. According to Westin, the consumer population can be subdivided into privacy fundamentalist, privacy unconcerned, and privacy pragmatist.[xxxvi] As such, while privacy fundamentalists may advocate the regulation of data-driven marketing, privacy unconcerned and privacy pragmatist may hold opposite attitudes toward such practice.

On the other hand, traditional techniques of anonymization also have inherent flaws that largely limit efficacy in privacy preservation. The procedure of k-anonymity, for instance, is highly dependent on large quantities of diverse quasi-identifiers. In the absence of database diversity, k-anonymity is inherently vulnerable to homogeneity and contextual (background knowledge) attacks, causing leakage of sensitive attributes.[xxxvii] Furthermore, de-anonymization techniques can effectively re-identify cloaked datasets even when only partial data are presented and background knowledge is insufficient. An evaluation of the effectiveness of de-anonymization attacks against high-dimensional micro-data released online by Narayanan and Shmatikov found that an adversary with

little contextual knowledge can successfully identify records of known users from limited databases, even uncovering sensitive information such as political and sexual preferences.[xxxviii]

Other studies suggest that the emerging private market is not a panacea to privacy concerns. A study by Awad and Krishnan reveals a paradox in terms of infomediary models; customers value these profitable outcomes while also maintaining feelings of vulnerability in the exchange of personal data.[xxxix] Moreover, the inherent mechanism of third-party intermediaries does not suffice the expected role of an institution in the social contract. The Power-Responsibility Equilibrium theory presented by Murphy et al indicates that the "more powerful partner in a relationship has the societal obligation to promote an environment of felt equality".[xl] Customers therefore do not have the responsibility to act on an initiative to protect their privacy. On the contrary, it is the duty of corporations to ensure privacy is guarded.

2.4. Differential Privacy and CTGAN in Privacy Preservation

In recent years, the Generative Adversarial Network has developed as a promising solution to reducing privacy concerns. In particular, the model's underpinning architecture can synthesize high-quality samples that are consistent with real-world conditions, which can be used to replace real consumer data in the process of analytic algorithms. The traditional DCGAN designed by Goodfellow et al, for instance, can be trained to synthesize high-quality pictures using batch norm in both its discriminative model and generative model.[xli] Further developments of table-GAN by Park et al[xlii] and CTGAN by Xu et al[xliii] are variations models used to generate tabular datasets consisting

of both discrete and numerical values. Other data synthesis variants of GAN, such as medGAN[xliv] and medBGAN,[xlv] have been applied to fields of medicine to generate statistically identify patient electronic health records (EHRs) while preserving the original sensitive information from being revealed.

On the other hand, differential privacy emerges as another influential technique in the preservation of privacy data. Introduced by Cynthia Dwork[xlvi] and Frank McSherry, et al. in 2006, the process of differential privacy ensures that the presence or absence of an individual's data in a dataset does not significantly affect the outcome of any computation or "mechanism" performed on the data by adding calibrated noise to the output of calculations and masking the contribution of any single individual while preserving the overall accuracy of the analysis. This noise addition is guided by parameters known as "epsilon and delta," which quantify the "privacy loss" or additional risk to an individual resulting from their data being used.

As a result, a number of literature have explored the combination of the two aforementioned privacy preserving techniques. Previous approaches conclude a two-layer algorithm, with which differential privacy is employed on the discriminative model to generate synthetic datasets that is below a designated epsilon, the common privacy budget. DPGAN by Xie et al[xlvii] and dp-GAN by Zhang, Ji, and Wang[xlviii] examine the performance of this algorithm in image datasets and electronic health record data (EHR). Torkzadehmahani, Kairouz, and Paten introduced the CGAN as a variant to GAN but applies the same framework.[xlix] Their analysis conducted on MNIST concludes promising preliminary results in the extension of the DP+GAN framework. Further

variation with DP-CTGAN was evaluated by Fang, Dhami, and Kersting,[1] with evaluation on numerous sets of medical tabular data.

Despite varying extensions to the DPGAN framework, existing work in this field has primarily focused on the privacy performance from visual and medical datasets, namely, MNIST and EHR. Studies on the effect of combined differential privacy and GAN variants in the field of digital marketing, therefore, remains under-explored. This motivate further analysis using datasets relevant to commercial marketing.

3. METHODOLOGY

3.1. Selected Dataset

This set of data describes the collection of consumer data incorporated into an online marketing campaign, consisting of 39 attributes/columns and 2205 instances/rows. Data values are arranged in two-dimensional attributes/columns and instances/rows, featuring both continuous and discrete values. Each row represents an array of different categorical information on one individual/user and reasonably resembles real-world data in the digital market. Evaluations based on this dataset therefore provide an empirical analysis of the success of CTGAN in the preservation of consumer privacy, specifically in the context of electronic marketing.

The metadata of this dataset can be summarized into five fields of measure: Accepted Campaign, Expenditure on Product purchasing, Number of Purchases, Shopping Behavior, and Other Identifiable Information. The column “Response” is decided as the class label for this dataset. A statistical model in the prediction problem will endeavor to predict the probability of "Response" based on all other inputs.

Table 1: Dataset *iFood* Dictionary

Columns	Description
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response (class label)	1 if customer accepted the offer in the last campaign, 0 otherwise
Complain	1 if customer complained in the offer in the last 2 years, 0 otherwise
Customer_Days	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spend on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on gold products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurhases	number of purchases made using catalog
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase
AcceptedCmpOverall	number of customer's accepted campaign from the company (1-4)

Table 1 shows the dictionary for the 39 attributes in a meta-table. A more detailed evaluation of the tabular traits that impair the learning process of the original GAN will be introduced in *section 3.3*.

3.2. Generative Adversarial Network

In this study, the underpinning framework of Generate Adversarial Network (GAN) is adopted as the architectural foundation for synthetic data modeling. Traditionally, the model of GAN is established on two neural networks competing in a min-max game, which features the Generative model (G) generating synthetic data from sampled distribution of real data and Discriminative model (D) subsequently distinguishing synthetic data from real input to optimize the generation of G. In mathematical terms, the generator G and discriminator D optimize the following objective value function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

where p-data is denoted as sampled distribution from real data instances and pz(z) as prior distribution placed on the randomly generated noise vector z. The denotation of x~pdata(x) and z~pz(z) each represent expected value of the total inputs into discriminator D and generator G respectively. We then refer function G(·) to probability output for generator G and function D(·) to probability output for discriminator D, with an output value span of [0,1]. When the discriminator D classifies an input data as authentic, as in D(x), a higher probability (close to 1) is produced, whereas when the discriminator D classifies input data as generated, as in D(G(z)), a lower probability is

produced. In broad terms, equation 2 min-max loss function can be seen as the sum of the discriminator D's average value prediction under input real data and the discriminator D's average value prediction under input synthetic data. The GAN model thus operate in a manner that the generator G attempts to minimize the value function V(D,G), while the discriminator D attempts to maximize it.

The mini-batch stochastic gradient descent is commonly adopted for its advantage in computational speedup. For per iteration, we input a mini-batch randomly sampled from real data and a mini-batch randomly sampled from its generated counterparts into the discriminative model. After multi-layer processing, discriminator D classifies input as either real or synthetic. Suppose that discriminator D misclassifies the correct authenticity of data input, it is penalized with a discriminator loss. After one round of training iteration, the discriminator D is updated via gradient adjustment.

The training iteration for the generator G is largely based on the classification of the discriminator D. Suppose the discriminator D correctly identified synthetic data, the generator G is penalized with a corresponding generator loss. By the end of the training iteration, generator G is updated via gradient adjustment. As such, training iterations of the generator and the discriminator are conducted simultaneously, with the generator G highly dependent on the classification results yielded by discriminator D. This cycle is repeated until both generator G and discriminator D achieve loss convergence, and the min-max loss function is optimized.

3.3. Challenges in the synthesis of tabular data using the original GAN model

This section presents several limitations of the original GAN algorithm in the generation

of high-quality tabular data. Data statistics from Dataset *iFood* are used as figurative illustrations.

Non-Gaussian Distribution:

The distribution of tabular data differs significantly from visual data that is traditionally adopted in the training of the GAN algorithm. Pixel values in image datasets typically adhere to Gaussian distribution, which can be normalized to the range [-1,1] using min-max transformation in the output layer of GAN'S multilayer perceptron. However, continuous values of tabular data are non-Gaussian. For example, Fig. 1 (a) and (b) show visualizations of non-Gaussian distribution from the continuous column "*NumDealsPurchases*" in tabular dataset 1. The visualization displays an evident right skewness, which is statically reflected by mode > median > mean (957 > 2 > 1.886107). Applying min-max transformation to non-gaussian numeric distribution would yield the vanishing gradient problem, which impairs the efficacy and quality in the learning process of the GAN model.

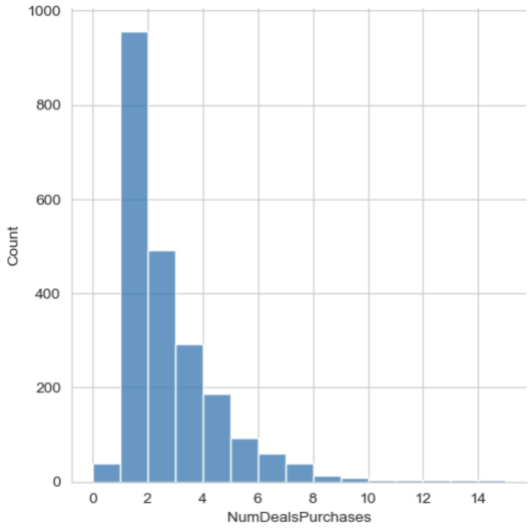

Fig. 1(a) Distribution of NumDealsPurchases"

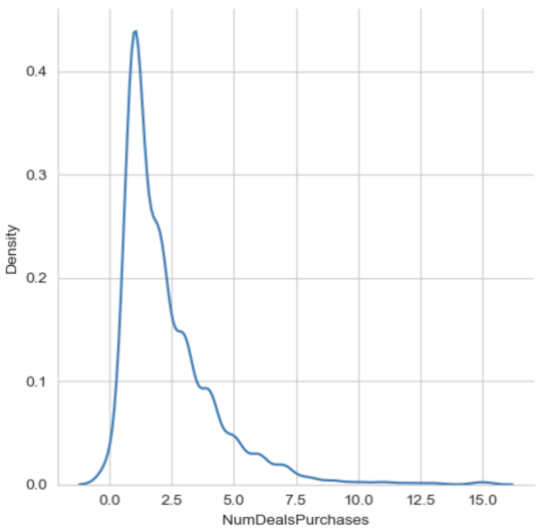

Fig. 1(b) Kernel Density Estimation of "NumDealsPurchases"

Multi-modal distribution:

Srivastava et al show that the original GAN modeling encounters difficulties when training a set of data with the multi-modal distribution. In particular, the multi-modal dataset presents a higher probability of mode collapse in GAN training, where modes of the training set are only partially captured for the generation of synthetic data. Such forms of underrepresentation largely affect the distribution resemblance of synthetic data in comparison to that of real data.

We use the Kernel Density Estimation to approximate multi-modal distribution in our continuous columns, specifically within dataset *iFood*. A Gaussian KDE is performed on the Python Jupyter Notebook using the default kernel width over all continuous columns. As a result, we discover 5/20 continuous columns feature multi-modal density. Fig. 2(a) and (b) show the multi-distribution of continuous column "Customer_Days".

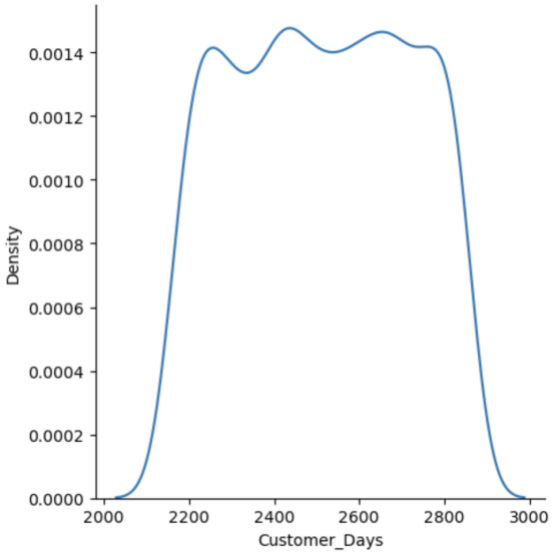

Fig. 2(a) Distribution of 'Customer_Days"

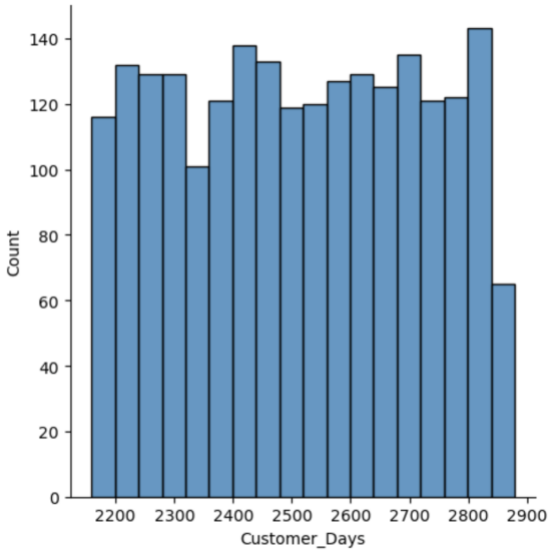

Fig. 2(b) Kernel Density Estimation of "NumDealsPurchases"

Imbalance of Categorical Columns:

Imbalanced categorical data introduces severe mode collapse and underrepresentation in GAN algorithm: The former impairs the process of deep learning, and the latter inadequately represent minor categories in the synthesized results. Specifically, this paper define the imbalance of categorical features as present when the minor category account for less than 10% of total instance. Data count reveals 10 out of 18 categories are associated with such imbalance. Table 2 shows the value counts with respect to each column data in dataset 1 corresponding to the categorical percentage.

Table 2: Dataset *iFood* Discrete Column Counts

Column Title	Value	Count	Percentage/Total instance
AcceptedCmp1	0	2063	0.0643
	1	142	
AcceptedCmp2	0	2175	0.0136
	1	30	
AcceptedCmp3	0	2042	0.0739
	1	163	
AcceptedCmp4	0	2041	0.0744
	1	164	
AcceptedCmp5	0	2044	0.0730
	1	161	
Complain	0	2185	0.0091
	1	20	
Response	0	1872	0.1510
	1	333	
education_2n Cycle	0	2007	0.0898
	1	198	
education_Basic	0	2151	0.0245
	1	54	
education_Graduation	1	1113	0.4952
	0	1092	
education_Master	0	1841	0.1651
	1	364	
education_PhD	0	1729	0.2159
	1	476	
marital_Divorced	0	1975	0.1043
	1	230	

marital_Married	0	1351	0.3873
	1	854	
marital_Single	0	1728	0.2163
	1	477	
marital_Together	0	1637	0.2576
	1	568	
marital_Widow	0	2129	0.0345
	1	76	

It deserves pointing out, for instance, that a majority of minor categories associated with categorical imbalance concern individuals who accepted the campaign or sufficed educational status. These individuals therefore provide valuable insights to the analysis of data, and their underrepresentation in training could significantly deviate the distribution of synthetic data from that of the real outcome.

Sparsity of One-hot-encoded vectors:

To synthesize categorical data, the GAN model creates a probability distribution over all categories. This form of output can be evidently distinguished from the distribution of real data transformed into one-hot encoded vectors, therefore enabling the discriminator to determine authenticity by simply comparing sparsity rather than evaluating realness.

3.4. Conditional Tabular Generative Adversarial Network

This study introduces the conditional tabular generative adversarial network as a variant of GAN to address the aforementioned issues in the synthesis of consumer tabular data. With the underpinning architecture remaining the same, CTGAN applies a mode-specific normalization featuring a variational Gaussian mixture model (VGM) to properly represent continuous data associated with Non-Gaussian and Multi-modal distribution. Additionally, a conditional generator is designed to evenly resample minor categories in the imbalanced discrete columns.

We import the CTGAN trainer from the Synthetic Data Vault library. The loaded source dataset is then trained under the following hyper-parameters for epochs 300, 500, 1000, 2000, and 5000 times to generate the respective quality of synthetic datasets.

```
{'enforce_min_max_values': True,
 'enforce_rounding': False,
 'locales': None,
 'embedding_dim': 128,
 'generator_dim': (256, 256),
 'discriminator_dim': (256, 256),
 'generator_lr': 0.0002,
 'generator_decay': 1e-06,
 'discriminator_lr': 0.0002,
 'discriminator_decay': 1e-06,
 'batch_size': 500,
 'discriminator_steps': 1,
 'log_frequency': True,
 'verbose': True,
 'epochs': 300,
 'pac': 10,
 'cuda': True}
```

Fig. 3 300 epochs synthesizer parameters

Before the CTGAN trainer is initiated, we add conditional constraints in the synthesis of categorical data to exclude potential distortion. For discrete columns of marital status and education level, we set constraint that only one instance can be synthesized per row. This will prevent situations in which an individual in the synthetic data is generated in a way that is contradictory to reality, for instance, being simultaneously “marital_Divorced” and “marital_Married”. The established constraint

parameters are as follows:

```
my_constraint_marital = {
    'constraint_class': 'OneHotEncoding',
    #'table_name': 'users', # for multi table synthesizers
    'constraint_parameters': {
        'column_names': ['marital_Divorced', 'marital_Married',
       'marital_Single', 'marital_Together', 'marital_Widow']
    }
}


my_constraint_education = {
    'constraint_class': 'OneHotEncoding',
    #'table_name': 'users', # for multi table synthesizers
    'constraint_parameters': {
        'column_names': ['education_2n Cycle', 'education_Basic', 'education_Graduation',
       'education_Master', 'education_PhD']
    }
}


synthesizer.add_constraints(constraints=[
    my_constraint_marital,
    my_constraint_education
])
```

Fig. 4 Adding Constraints

3.5. Differential Privacy Stochastic Gradient Descent

To ensure that the synthesized dataset is differentially private, we employ differential privacy stochastic gradient descent(DPSGD) from Tensorflow. A total of 1764 instances are screened from 2204 instances to train with stochastic gradient descent. Parameters is defined as follows:

```
l2_norm_clip = 1.5
noise_multiplier = 1.3
num_microbatches = 32
learning_rate = 0.25
```

Fig. 6

Gaussian noise is added for gradients per epoch perturb the data updates, thus protecting privacy of individuals. An inevitable decrease in utility performance, however, is followed.

4. Empirical Results

4.1. Analysis of Fidelity

Kolmogorov-Smirnov statistic and Total Variation Distance analysis

The Kolmogorov-Smirnov two-sample test is applied to evaluate the equality of numerical distributions between real and synthetic datasets with varying epochs for its sensitivity against slight binomial distribution and applicability with non-gaussian data (where a t-test would yield unreliable *p-value*). First, the metric computes the cumulative distribution functions (CDFs) of two corresponding univariate columns from two datasets. KS statistic is then obtained by quantifying the maximum difference between the real and synthetic column CDF (non-Gaussian). On the other hand, the Total Variation Distance is used to measure the difference between real and synthesized discrete columns. Similar to the KS statistic, the TVD score is obtained by calculating the difference between the probability frequency of two pairing discrete columns from real and synthetic datasets. Both metrics yield a difference score in the range of 0 and 1, where 0 denotes strong similarity between real and synthetic distributions, and thus better performance.

To offer a more intuitive understanding of the performance of CTGAN and the process of evaluation, we calculate the overall quality using complement scores of KS statistics and TVD by following simple mathematical procedures 1-(KS statistic) and $1-\delta(R, S)$. The modified metrics infer opposite effects compared to that of the original, and scores approaching 1 reflect higher similarity.

Lastly, the similarity of overall column shapes is calculated by averaging KS and TVD complements across all attributes.

Table 3: Column Shapes *iFood*

	Epochs = 300		Epochs = 500		Epochs = 1000		Epochs = 2000		Epochs = 5000	
	KS statistic/TVD	p-value	Mean KS statistic/TVD	p-value	Mean KS statistic/TVD	p-value	Mean KS statistic/TVD	p-value	Mean KS statistic/TVD	p-value
AcceptedCmp1	0.00408	/	0.05624	/	0.05714	/	0.02721	/	0.08435	/
AcceptedCmp2	0.04082	/	0.04671	/	0.03628	/	0.04399	/	0.03129	/
AcceptedCmp3	0.01723	/	0.06077	/	0.07256	/	0.00862	/	0.05578	/
AcceptedCmp4	0.01043	/	0.08753	/	0.00680	/	0.02812	/	0.04444	/
AcceptedCmp5	0.03628	/	0.01633	/	0.06168	/	0.03039	/	0.06984	/
Response	0.00499	/	0.03129	/	0.06712	/	0.12381	/	0.13651	/
Complain	0.03129	/	0.03991	/	0.03900	/	0.03537	/	0.02766	/
education_2n Cycle	0.32562	/	0.46304	/	0.31791	/	0.22086	/	0.10476	/
education_Basic	0.02721	/	0.02585	/	0.01361	/	0.02177	/	0.02041	/
education_Graduation	0.10249	/	0.21587	/	0.12200	/	0.12562	/	0.04717	/
education_Master	0.10930	/	0.11474	/	0.06576	/	0.02766	/	0.03628	/
education_PhD	0.14104	/	0.15828	/	0.14376	/	0.08934	/	0.04172	/
marital_Divorced	0.44354	/	0.45896	/	0.20590	/	0.13424	/	0.08345	/
marital_Married	0.20680	/	0.17098	/	0.19592	/	0.05533	/	0.04263	/
marital_Single	0.07710	/	0.15057	/	0.08435	/	0.01995	/	0.00181	/
marital_Together	0.15782	/	0.13107	/	0.07438	/	0.08571	/	0.03991	/
marital_Widow	0.00181	/	0.00635	/	0.01995	/	0.01315	/	0.00272	/
Age	0.07211	0.00002	0.14467	0.00000	0.13832	0.00000	0.16009	0.00000	0.12698	0.00000
Customer_Days	0.14739	0.00000	0.18730	0.00000	0.18095	0.00000	0.21859	0.00000	0.04172	0.04304
Kidhome	0.07347	0.00001	0.04490	0.02347	0.01497	0.96594	0.06848	0.00006	0.07120	0.00003
Teenhome	0.07075	0.00003	0.20907	0.00000	0.11429	0.00000	0.12245	0.00000	0.00091	0.99999
Income	0.34785	0.00000	0.17460	0.00000	0.13379	0.00000	0.07075	0.00003	0.09660	0.00000
MntFishProducts	0.15601	0.00000	0.29342	0.00000	0.19909	0.00000	0.11293	0.00000	0.22857	0.00000
MntMeatProducts	0.12472	0.00000	0.13333	0.00000	0.14150	0.00000	0.11474	0.00000	0.14649	0.00000
MntFruits	0.14739	0.00000	0.17823	0.00000	0.14966	0.00000	0.05397	0.00324	0.07664	0.00000
MntRegularProducts	0.38639	0.00000	0.15329	0.00000	0.13832	0.00000	0.17324	0.00000	0.06168	0.00045
MntSweetProducts	0.22902	0.00000	0.14739	0.00000	0.06395	0.00024	0.12245	0.00000	0.06168	0.00045
MntWines	0.16463	0.00000	0.24580	0.00000	0.14240	0.00000	0.10249	0.00000	0.13469	0.00000
MntGoldProds	0.15102	0.00000	0.14376	0.00000	0.04535	0.02144	0.11791	0.00000	0.10023	0.00000
MntTotal	0.16190	0.00000	0.24263	0.00000	0.13424	0.00000	0.09025	0.00000	0.10748	0.00000
NumDealsPurchases	0.04535	0.02144	0.09433	0.00000	0.08753	0.00000	0.07029	0.00004	0.01995	0.77241
NumCatalogPurhases	0.07800	0.00000	0.13152	0.00000	0.04717	0.01481	0.07937	0.00000	0.08662	0.00000
NumStorePurchases	0.06349	0.00027	0.14467	0.00000	0.05125	0.00610	0.05533	0.00234	0.04626	0.01785
NumWebPurchases	0.04762	0.01347	0.08571	0.00000	0.08163	0.00000	0.04308	0.03337	0.05850	0.00105
NumWebVisitsMonth	0.08435	0.00000	0.07574	0.00000	0.07937	0.00000	0.11383	0.00000	0.07755	0.00000
Recency	0.32426	0.00000	0.15420	0.00000	0.07800	0.00000	0.08254	0.00000	0.07937	0.00000
AcceptedCmpOverall	0.06213	0.00040	0.11519	0.00000	0.12562	0.00000	0.08345	0.00000	0.09388	0.00000
Mean KS Complement	**0.88010**	/	**0.86320**	/	**0.90430**	/	**0.91930**	/	**0.93360**	/

Table 3 reports the KS statistic and TVD score for columns of dataset *iFood*. A p-value is calculated for the KS statistic of numerical columns but displayed as "/" for categorical columns. An alpha-value of 0.05 is accepted, and a p-value > 0.05 concludes no statistical difference in the pair of numerical distributions.

The overall column shapes are improved from 0.8801 under 300 epochs to 0.9336 under 5000 epochs, reflecting a positive relationship between epoch quantity and column shapes similarity. An exception is found at 500 epochs, where the model underperformed compared to the results under 300 epochs. Additionally, the synthesis of discrete columns shows better performance across all numbers of epochs. Fig 5-14 shows a comparison between the numerical KS statistic and the discrete TVD score.

While the overall score has increased over augmented epochs, the CTGAN model does not guarantee the constant improvement of individual column similarity. To illustrate, the column shape similarity for the numerical column "*Age*" yields KS statistics of 0.07211, 0.14467, 0.13832, 0.16009, and 0.12698 across epochs 300, 500, 1000, 2000, and 5000. In this case, the highest similarity occurred in the synthetic data trained under 300 epochs, whereas higher epochs yielded less similarity.

Only three columns yielded a KS statistic that concludes no statistical difference between real and synthetic datasets. It deserves adding, however, that KS statistic calculates p-values on a strict sensitivity, and numerical distributions with p-values < 0.05 do not infer failure of CTGAN modeling. The KS statistic and Total Variation Distance remain the primary accessing tools for distributional similarity.

Pairwise correlation analysis

The Pearson Correlation coefficient is used to evaluate the pairwise similarity of real and synthetic discrete columns. We segment the Pearson correlation by five levels ranging from 0 to 1: [0-0.2] denoting weak correlation, [0.2-0.4] denoting moderate-weak correlation, [0.4-0.6] denoting moderate correlation, [0.6-0.8] denoting moderate-strong correlation, and [0.8-1.0] denoting strong correlation. Heat maps are used to compare pairwise correlation in varying levels of synthetic data, where light green and blue represent high and low correlation scores respectively, with darker colors as transitions (note that only half the total columns are displayed per axis due to limited space).

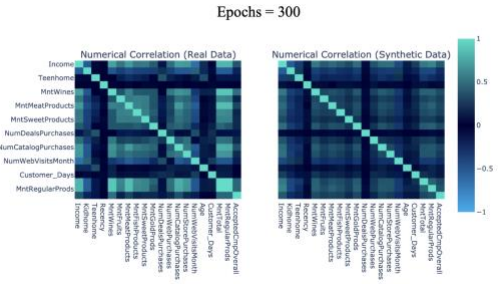

Fig. 5(a)

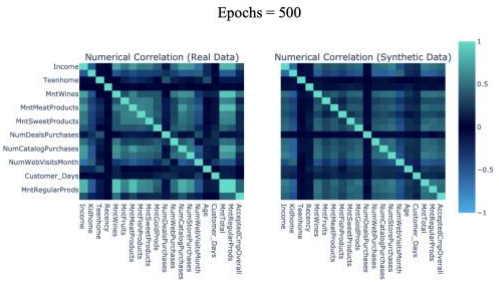

Fig. 5(b)

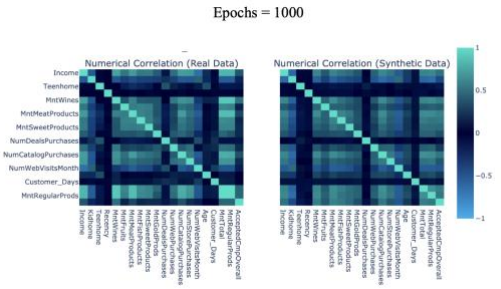

Fig. 6(c)

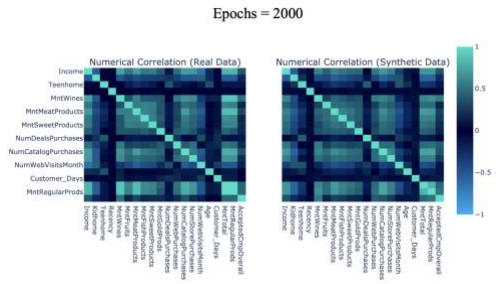

Fig. 5(d)

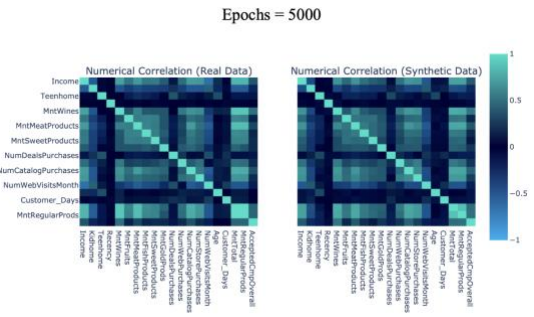

Fig. 6

Table 4: Column Pair Trends *iFood*

Epochs	CorrelationAccur
real data	1
300	0.8389
500	0.8280
1000	0.8813
2000	0.8981
5000	0.9183

Table 4 includes the accuracy score of the pairwise correlation similarity between real and synthetic datasets. We observe that CTGAN synthesized dataset largely preserves the inter-attributes connection of the source dataset, with all epochs yielding a correlation similarity score of [0.8-1.0]. The positive relationship between epochs quantity and correlation accuracy is maintained as in the report of column shapes and with epochs = 500 as an exception.

Fig 5(a-d) display pairwise trend correlation heat maps corresponding to varying epochs of CTGAN synthesized numerical columns.

4.2. Analysis of Utility

A random forest classifier is chosen to evaluate and compare the utility of synthetic datasets. First, an imported Synthetic Minority Oversampling Technique (SMOTE) is used to treat severe class imbalance in the dataset by synthesizing new instances of

minority data, after which the results are refined over a Recursive Feature Elimination (RFE) model. A cross-validation technique is then performed over this set of data, segmenting the data into subsets for training and testing. Lastly, data is imported into a random forest classifier to output a prediction accuracy score and *F*1 score, yielding insights on the performance of CTGAN-synthesized datasets.

There are several advantages to this set of modeling. To begin with, CTGAN training retains the patterns of class imbalance in the synthetic datasets as it represents the statistical properties of the source dataset. The application of SMOTE thus provides important functionality in alleviating distortion due to minority data in the evaluator by generating synthetic instances. On the basis of this dataset, the RFE model is used to select important features for the prediction of the target class, removing weakly correlated variables that would complicate the algorithmic training and lower prediction accuracy. Cross-validation is adopted for every training model to exclude selection bias and overfitting by separating the dataset into training set and testing set, otherwise yielded results can not be generalized over real-world data. Finally, the random forest classifier is used to predict the categorical output of the target class based on mix of numerical and discrete inputs, yielding an accuracy score and f1 score that implies dataset performance.

Table 5: Results of Random Forest Classifier *iFood* (n=100)

Epochs	PredictionAccur	F1	diff
real data	0.8808	0.8719	0
300	0.8317	0.8399	0.0409
500	0.8364	0.8107	0.0444
1000	0.8551	0.8387	0.0257
2000	0.8198	0.8122	0.0610
5000	0.8197	0.8098	0.0611

Table 6 reports the predictive accuracy and F1 score of real and synthetic datasets calculated using the customized random forest classifier. As all accuracy scores exceed a benchmark of 0.80 for accuracy and F1, the synthetic datasets are proved to be satisfactorily informative when applied to real-world classification problems. We observe, however, higher epochs do not necessarily yield high accuracy in the predictive ability or F1. Specifically, the prediction score of synthetic datasets reached the maximum output at epochs 1000, and higher epochs above this threshold display diminishing returns. This can be attributed to possible oversampling using the SMOTE augmentation, which similarly generates data as CTGAN (except the technique is set to generate minor category instances rather than the full-scale tabular data).

4.4. Trade-offs

In this dataset, the CTGAN models are trained without the aid of parallel platform CUDA and the use of a GPU, which would drastically accelerate the process of training. This subsection analyzes the change in CTGAN productivity as the number of epochs increases. Previous results are plotted as dependent variables and training time is plotted as the independent variable.

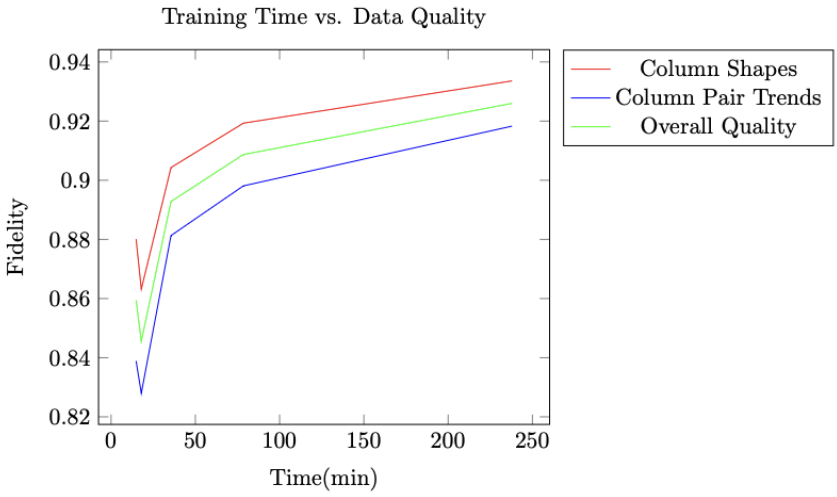

Fig. 7

Fig. 7 displays plotted graph on the relationship between training time and data fidelity for synthetic datasets of varying epochs. Graphing the execution time as input and data quality as output, we observe marginal diminishing productivity between CTGAN training time and performance as the number of epochs

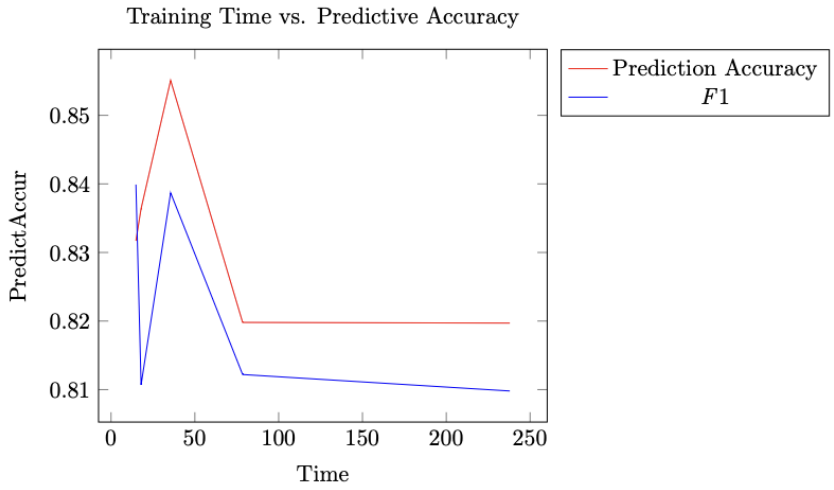

Fig. 8

Fig. 8 displays plotted graph on the relationship between training time and prediction accuracy for synthetic datasets of varying epochs, with execution time as input and predictive accuracy scores as output. The maximum predictive accuracy is obtained at epochs 1000, whereas the maximum F1 score is obtained at epochs 300. Previous to epochs 1000, we observe productivity increase for prediction accuracy but negative returns followed by productivity increase for F1 score. Both metrics experience negative returns after epoch 1000, and the difference in scores between epochs 2000 and epochs 5000 is shown to be insignificant. We predict that epochs 5000 onward are unlikely to yield statistically significant improvement in prediction accuracy and f1 score for this specific set of data.

4.3. Analysis of Privacy

The analysis of privacy uses epsilon (ϵ) benchmark at a Delta (δ) value of 10^-5. An epsilon is the privacy budget upper bound that reflects the degree of change in the modeling results from including or excluding an individual value. This intuitively suggest that a lower epsilon score, denoting small changes as additional data value is included, reflects strong privacy protection as an adversary is less likely to obtain useful information regarding the dataset.

For the synthesized dataset, we obtained an epsilon value of 8.5 and an average prediction accuracy of 82.77% after inputting the differentially private synthetic dataset into a logistic regression model. This leaves us with a reasonable privacy guarantee at the tier 2 level.

5. DISCUSSION

In this study, we have explored the application of the Conditional Tabular Generative Adversarial Network (CTGAN) in synthesizing tabular consumer data, with a particular focus on digital marketing. The empirical results have provided a nuanced understanding of the fidelity, utility, and trade-offs of the CTGAN model, revealing both its strengths and areas for improvement. In particular, the fidelity analysis, which employed the Kolmogorov-Smirnov statistic and Total Variation Distance, revealed that CTGAN could effectively model non-Gaussian and multi-modal distributions. This is a significant finding, as it addresses a fundamental challenge in the synthesis of tabular data, particularly in the context of digital marketing.

The positive relationship between the number of epochs and the similarity of column shapes was an important observation. However, the exceptions and diminishing returns observed in certain cases highlight the complexity of the model's behavior. This suggests that practitioners must approach hyperparameter tuning with caution, considering the specific characteristics of the dataset at hand. The intricate understanding of how the number of epochs influences the model's performance is a valuable contribution to the field.

The utility analysis using a random forest classifier further validated the credibility of the synthesized datasets. The observation that higher epochs do not necessarily yield higher accuracy in predictive ability or F1 score is a critical insight. It emphasizes the need for a balancing approach, considering both the quality of the synthetic data and the computational efficiency.

The trade-off analysis between training time and performance revealed marginal

diminishing productivity. This finding is particularly relevant for practitioners aiming to balance computational resources and model performance. It underscores the importance of selecting an optimal number of epochs, a consideration that may vary depending on the specific application and dataset.

The study's introduction of CTGAN as a novel approach to synthesizing consumer tabular data is a significant contribution. By addressing specific challenges such as non-Gaussian distribution, multi-modal distribution, and imbalanced categorical columns, the research offers a solution with broad applicability. This could extend beyond electronic marketing to domains such as healthcare, finance, or social sciences.

The comprehensive evaluation framework introduced in this research is another key contribution. By including fidelity, utility, and trade-off analyses, the framework offers a nuanced understanding of the CTGAN model's performance. This can serve as a guideline for evaluating other generative models, enhancing the rigor and robustness of future research in this area.

The ethical considerations surrounding synthetic data generation are complex and warrant further examination. The potential biases, privacy concerns, and consent issues related to synthetic data generation require careful consideration. Future research could delve into these aspects, developing guidelines and best practices to ensure that synthetic data generation aligns with societal values and norms.

Several avenues for future research emerge from this study. Investigating advanced optimization techniques to further enhance the performance of CTGAN, such as Federated Learning could be a valuable direction. Exploring the application of CTGAN in various domains and integrating it with other machine-learning models could broaden

the impact of this technology. Additionally, the development of user-friendly tools and platforms to facilitate the application of CTGAN by non-expert users could democratize access to this powerful technology.

In conclusion, this paper has provided a significant step forward in understanding the potential and limitations of CTGAN for synthesizing consumer tabular data. The insights gained not only contribute to the theoretical understanding of generative models but also offer practical guidance for researchers and practitioners working with synthetic data. The findings would allow us to further exploration and innovation in the rapidly evolving field of data synthesis and privacy preservation.

Endnotes:

---

[i] "Quarterly Retail E-Commerce Sales Quarter 2023 - Census.Gov," U.S. Census Bureau News, August 17, 2023, https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf.

[ii] P.K. Kannan and Hongshuang "Alice" Li, "Digital Marketing: A Framework, Review and Research Agenda," *International Journal of Research in Marketing* 34, no. 1 (2017): 22–45, https://doi.org/10.1016/j.ijresmar.2016.11.006.

[iii] "Art. 5 GDPR – Principles Relating to Processing of Personal Data," General Data Protection Regulation (GDPR), October 22, 2021, https://gdpr-info.eu/art-5-gdpr/.

[iv] "California Consumer Privacy Act (CCPA)," State of California - Department of Justice - Office of the Attorney General, May 10, 2023, https://oag.ca.gov/privacy/ccpa.

[v] Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

[vi] Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

[vii] "Privacy Noun - Definition, Pictures, Pronunciation and Usage Notes: Oxford Advanced American Dictionary at Oxfordlearnersdictionaries.Com," privacy noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced American Dictionary at OxfordLearnersDictionaries.com, accessed August 21, 2023, https://www.oxfordlearnersdictionaries.com/definition/american_english/privacy#:~:text=%2Fˈpra%C9%AAv%C9%99si%2F,I%20value%20my%20privacy.

[viii] Eugene F. Stone and Diana L. Stone, Stone, E. F., & stone, D. L. (1990). privacy in organizations ..., accessed August 20, 2023, https://www.researchgate.net/publication/259440817_Stone_E_F_Stone_D_L_1990_Privacy_in_organizations_Theoretical_issues_research_findings_and_protection_strategies_In_G_Ferris_K_Rowland_Eds_Research_in_personnel_and_human_resources_management_Vol_8_pp.

[ix] Erving Goffman, The presentation of self - monoskop, accessed August 20, 2023, https://monoskop.org/images/1/19/Goffman_Erving_The_Presentation_of_Self_in_Everyday_Life.pdf.

Erving L. Goffman, "Stigma: Notes on the Management of Spoiled Identity. by Erving Goffman. Englewood Cliffs, New Jersey: Prentice-Hall, 1963. 147 Pp. Cloth, $4.50; Paper, $1.95," *Social Forces* 43, no. 1 (October 1, 1964): 127–28, https://doi.org/10.1093/sf/43.1.127.

[x] Erving Goffman, The presentation of self - monoskop, accessed August 20, 2023, https://monoskop.org/images/1/19/Goffman_Erving_The_Presentation_of_Self_in_Everyday_Life.pdf.

Erving L. Goffman, "Stigma: Notes on the Management of Spoiled Identity. by Erving Goffman. Englewood Cliffs, New Jersey: Prentice-Hall, 1963. 147 Pp. Cloth, $4.50; Paper, $1.95," *Social Forces* 43, no. 1 (October 1, 1964): 127–28, https://doi.org/10.1093/sf/43.1.127.

Valerian J. Derlega and Alan L. Chaikin, "Privacy and Self-Disclosure in Social Relationships," *Journal of Social Issues* 33, no. 3 (1977): 102–15, https://doi.org/10.1111/j.1540-4560.1977.tb01885.x.

[xi] Robert S. Laufer and Maxine Wolfe, "Privacy as a Concept and a Social Issue: A Multidimensional Developmental Theory," *Journal of Social Issues* 33, no. 3 (1977): 22–42, https://doi.org/10.1111/j.1540-4560.1977.tb01880.x.

[xii] Erving Goffman, The presentation of self - monoskop, accessed August 20, 2023, https://monoskop.org/images/1/19/Goffman_Erving_The_Presentation_of_Self_in_Everyday_Life.pdf.

[xiii] Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy* (Boston (Mass.): Now publisher, 2014).

[xiv] "Equifax-Harris Consumer Privacy Survey | Worldcat.Org," WorldCat, 1991, https://worldcat.org/title/equifax-harris-consumer-privacy-survey/oclc/145391386.

[xv] Gina Pingitore et al., Consumer concerns about data privacy rising - future of Privacy Forum, October 29, 2013, https://fpf.org/wp-content/uploads/2013/12/JDPA-Data-Privacy-Research-Oct-20131.pdf.

[xvi] Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

[xvii] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke, "Information Privacy: Measuring Individuals' Concerns about Organizational Practices," *MIS Quarterly* 20, no. 2 (June 1996): 167, https://doi.org/10.2307/249477.

[xviii] Naresh K. Malhotra, Sung S. Kim, and James Agarwal, "Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model," *Information Systems Research* 15, no. 4 (December 2004): 336–55, https://doi.org/10.1287/isre.1040.0032.

[xix] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell, "Privacy Concerns and Consumer Willingness to Provide Personal Information," *Journal of Public Policy & Marketing* 19, no. 1 (Spring 2000): 27–41, https://doi.org/10.1509/jppm.19.1.27.16941.

[xx] Frank V. Cespedes and H. Jeff Smith, "Database Marketing: New Rules for Policy and Practice," MIT Sloan Management Review, July 15, 1993, https://sloanreview.mit.edu/article/database-marketing-new-rules-for-policy-and-practice/.

[xxi] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell, "Privacy Concerns and Consumer Willingness to Provide Personal Information," *Journal of Public Policy & Marketing* 19, no. 1 (Spring 2000): 27–41, https://doi.org/10.1509/jppm.19.1.27.16941.

[xxii] Glen J. Nowak and Joseph Phelps, "Direct Marketing and the Use of Individual-Level Consumer Information: Determining How and When 'Privacy' Matters," *Journal of Direct Marketing* 9, no. 3 (Summer 1995): 46–60, https://doi.org/10.1002/dir.4000090307.

[xxiii] Mary J. Culnan, "'How Did They Get My Name?': An Exploratory Investigation of Consumer Attitudes toward Secondary Information Use," *MIS Quarterly* 17, no. 3 (September 1993): 341, https://doi.org/10.2307/249775.

[xxiv] Roland T. Rust, P. K. Kannan, and Na Peng, "The Customer Economics of Internet Privacy - Journal of the Academy of Marketing Science," SpringerLink, 2002, https://link.springer.com/article/10.1177/009207002236917.

[xxv] Donna Hoffman, Marcos Peralta, and Thomas Novak, Building Consumer Trust Online - Researchgate, April 1999, https://www.researchgate.net/publication/220427207_Building_Consumer_Trust_Online.

[xxvi] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell, "Privacy Concerns and Consumer Willingness to Provide Personal Information," *Journal of Public Policy & Marketing* 19, no. 1 (Spring 2000): 27–41, https://doi.org/10.1509/jppm.19.1.27.16941.

[xxvii] Avi Goldfarb and Catherine Tucker, "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science* 30, no. 3 (2011): 389–404, https://doi.org/10.1287/mksc.1100.0583.

[xxviii] Tiffany Barnett White et al., "Getting Too Personal: Reactance to Highly Personalized Email Solicitations," *Marketing Letters* 19, no. 1 (September 4, 2006): 39–50, https://doi.org/10.1007/s11002-007-9027-9.

[xxix] Patricia A. Norberg and Daniel R. Horne, "Coping with Information Requests in Marketing Exchanges: An Examination of Pre-Post Affective Control and Behavioral Coping," *Journal of the Academy of Marketing Science* 42, no. 4 (2014): 415–29, https://doi.org/10.1007/s11747-013-0361-6.

[xxx] "Official Legal Text," General Data Protection Regulation (GDPR), September 27, 2022, https://gdpr-info.eu/.

[xxxi] "California Consumer Privacy Act (CCPA)," State of California - Department of Justice - Office of the Attorney General, May 10, 2023, https://oag.ca.gov/privacy/ccpa.

[xxxii] Health Insurance Portability and accountability act of 1996 (HIPAA), June 27, 2022, https://www.cdc.gov/phlp/publications/topic/hipaa.html.

[xxxiii] Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

[xxxiv] Avi Goldfarb and Catherine Tucker, "Privacy Regulation and Online Advertising," *SSRN Electronic Journal*, August 5, 2010, https://doi.org/10.2139/ssrn.1600259.

[xxxv] Marc Brodherson et al., "A Customer-Centric Approach to Marketing in a Privacy-First World," McKinsey & Company, May 20, 2021, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

[xxxvi] Ponnurangam Kumaraguru and Lorrie Faith Cranor, Privacy indexes: A survey of Westin's studies - CMU school of ..., December 2005, https://www.cs.cmu.edu/~ponguru/CMU-ISRI-05-138.pdf.

[xxxvii] Cansu Saatci and Efnan Sora Gunal, Preserving privacy in Personal Data Processing | IEEE Conference ..., accessed August 20, 2023, https://ieeexplore.ieee.org/document/8965432.

[xxxviii] Arvind Narayanan and Vitaly Shmatikov, "Robust De-Anonymization of Large Sparse Datasets," *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, February 5, 2008, https://doi.org/10.1109/sp.2008.33.

[xxxix] 1. Naveen Farag Awad and M. S. Krishnan, "The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization: MIS Quarterly: Vol 30, No 1," MIS Quarterly, March 1, 2006, https://dl.acm.org/doi/10.5555/2017284.2017287.

[xl] May Lwin, Jochen Wirtz, and Jerome D. Williams, "Consumer Online Privacy Concerns and Responses: A Power–Responsibility Equilibrium Perspective," *Journal of the Academy of Marketing Science* 35, no. 4 (February 8, 2007): 572–85, https://doi.org/10.1007/s11747-006-0003-3.

[xli] Ian J. Goodfellow et al., "Generative Adversarial Networks," arXiv.org, June 10, 2014, https://arxiv.org/abs/1406.2661.

[xlii] Noseong Park et al., "Data Synthesis Based on Generative Adversarial Networks," arXiv.org, July 2, 2018, https://arxiv.org/abs/1806.03384.

[xliii] Lei Xu et al., “Modeling Tabular Data Using Conditional Gan,” arXiv.org, October 28, 2019, https://arxiv.org/abs/1907.00503.

[xliv] Edward Choi et al., “Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks,” arXiv.org, January 11, 2018, https://arxiv.org/abs/1703.06490.

[xlv] Mrinal Kanti Baowaly et al., “Synthesizing Electronic Health Records Using Improved Generative Adversarial Networks,” Journal of the American Medical Informatics Association : JAMIA, March 1, 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647178/.

[xlvi] Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy* (Boston (Mass.): Now publisher, 2014).

[xlvii] Liyang Xie et al., “Differentially Private Generative Adversarial Network,” arXiv.org, February 19, 2018, https://arxiv.org/abs/1802.06739.

[xlviii] Xinyang Zhang, Shouling Ji, and Ting Wang, “Differentially Private Releasing via Deep Generative Model (Technical Report),” arXiv.org, March 25, 2018, https://arxiv.org/abs/1801.01594.

[xlix] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten, “DP-CGAN: Differentially Private Synthetic Data and Label Generation,” arXiv.org, January 27, 2020, https://arxiv.org/abs/2001.09700.

[l] Mei Ling Fang, Devendra Singh Dhami, and Kristian Kerstin, DP-CTGAN: Differentially private medical data generation using ctgans, accessed August 20, 2023, https://ml-research.github.io/papers/fang2022dpctgan.pdf.

Bibliography

"Art. 5 GDPR – Principles Relating to Processing of Personal Data." General Data Protection Regulation (GDPR), October 22, 2021. https://gdpr-info.eu/art-5-gdpr/.

Awad, Naveen Farag, and M. S. Krishnan. "The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization: MIS Quarterly: Vol 30, No 1." MIS Quarterly, March 1, 2006. https://dl.acm.org/doi/10.5555/2017284.2017287.

Baowaly, Mrinal Kanti, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. "Synthesizing Electronic Health Records Using Improved Generative Adversarial Networks." Journal of the American Medical Informatics Association : JAMIA, March 1, 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647178/.

Brodherson, Marc, Adam Broitman, Jason Cherok, and Kelsey Robinson. "A Customer-Centric Approach to Marketing in a Privacy-First World." McKinsey & Company, May 20, 2021. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/a-customer-centric-approach-to-marketing-in-a-privacy-first-world.

"California Consumer Privacy Act (CCPA)." State of California - Department of Justice - Office of the Attorney General, May 10, 2023. https://oag.ca.gov/privacy/ccpa.

Cespedes, Frank V., and H. Jeff Smith. "Database Marketing: New Rules for Policy and Practice." MIT Sloan Management Review, July 15, 1993. https://sloanreview.mit.edu/article/database-marketing-new-rules-for-policy-and-practice/.

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks." arXiv.org, January 11, 2018. https://arxiv.org/abs/1703.06490.

Culnan, Mary J. "'How Did They Get My Name?': An Exploratory Investigation of Consumer Attitudes toward Secondary Information Use." *MIS Quarterly* 17, no. 3 (September 1993): 341. https://doi.org/10.2307/249775.

Derlega, Valerian J., and Alan L. Chaikin. "Privacy and Self-Disclosure in Social Relationships." *Journal of Social Issues*33, no. 3 (1977): 102–15. https://doi.org/10.1111/j.1540-4560.1977.tb01885.x.

Dwork, Cynthia, and Aaron Roth. *The algorithmic foundations of Differential Privacy*. Boston (Mass.): Now publisher, 2014.

"Equifax-Harris Consumer Privacy Survey | Worldcat.Org." WorldCat, 1991. https://worldcat.org/title/equifax-harris-consumer-privacy-survey/oclc/145391386.

Fang, Mei Ling, Devendra Singh Dhami, and Kristian Kerstin. DP-CTGAN: Differentially private medical data generation using ctgans. Accessed August 20, 2023. https://ml-research.github.io/papers/fang2022dpctgan.pdf.

Goffman, Erving L. "Stigma: Notes on the Management of Spoiled Identity. by Erving Goffman. Englewood Cliffs, New Jersey: Prentice-Hall, 1963. 147 Pp. Cloth, $4.50; Paper, $1.95." *Social Forces* 43, no. 1 (October 1, 1964): 127–28. https://doi.org/10.1093/sf/43.1.127.

Goffman, Erving. The presentation of self - monoskop. Accessed August 20, 2023. https://monoskop.org/images/1/19/Goffman_Erving_The_Presentation_of_Self_in_Everyday_Life.pdf.

Goldfarb, Avi, and Catherine Tucker. "Online Display Advertising: Targeting and Obtrusiveness." *Marketing Science* 30, no. 3 (2011): 389–404. https://doi.org/10.1287/mksc.1100.0583.

Goldfarb, Avi, and Catherine Tucker. "Privacy Regulation and Online Advertising." *SSRN Electronic Journal*, August 5, 2010. https://doi.org/10.2139/ssrn.1600259.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." arXiv.org, June 10, 2014. https://arxiv.org/abs/1406.2661.

Health Insurance Portability and accountability act of 1996 (HIPAA), June 27, 2022. https://www.cdc.gov/phlp/publications/topic/hipaa.html.

Hoffman, Donna, Marcos Peralta, and Thomas Novak. Building Consumer Trust Online - Researchgate, April 1999. https://www.researchgate.net/publication/220427207_Building_Consumer_Trust_Online.

Kannan, P.K., and Hongshuang "Alice" Li. "Digital Marketing: A Framework, Review and Research Agenda." *International Journal of Research in Marketing* 34, no. 1 (2017): 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006.

Kumaraguru, Ponnurangam, and Lorrie Faith Cranor. Privacy indexes: A survey of Westin's studies - CMU school of ..., December 2005. https://www.cs.cmu.edu/~ponguru/CMU-ISRI-05-138.pdf.

Laufer, Robert S., and Maxine Wolfe. “Privacy as a Concept and a Social Issue: A Multidimensional Developmental Theory.” *Journal of Social Issues* 33, no. 3 (1977): 22–42. https://doi.org/10.1111/j.1540-4560.1977.tb01880.x.

Lwin, May, Jochen Wirtz, and Jerome D. Williams. “Consumer Online Privacy Concerns and Responses: A Power–Responsibility Equilibrium Perspective.” *Journal of the Academy of Marketing Science* 35, no. 4 (February 8, 2007): 572–85. https://doi.org/10.1007/s11747-006-0003-3.

Malhotra, Naresh K., Sung S. Kim, and James Agarwal. “Internet Users’ Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model.” *Information Systems Research* 15, no. 4 (December 2004): 336–55. https://doi.org/10.1287/isre.1040.0032.

Narayanan, Arvind, and Vitaly Shmatikov. “Robust De-Anonymization of Large Sparse Datasets.” *2008 IEEE Symposium on Security and Privacy (sp 2008)*, February 5, 2008. https://doi.org/10.1109/sp.2008.33.

Norberg, Patricia A., and Daniel R. Horne. “Coping with Information Requests in Marketing Exchanges: An Examination of Pre-Post Affective Control and Behavioral Coping.” *Journal of the Academy of Marketing Science* 42, no. 4 (2014): 415–29. https://doi.org/10.1007/s11747-013-0361-6.

Nowak, Glen J., and Joseph Phelps. “Direct Marketing and the Use of Individual-Level Consumer Information: Determining How and When ‘Privacy’ Matters.” *Journal of Direct Marketing* 9, no. 3 (Summer 1995): 46–60. https://doi.org/10.1002/dir.4000090307.

“Official Legal Text.” General Data Protection Regulation (GDPR), September 27, 2022. https://gdpr-info.eu/.

Park, Noseong, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. “Data Synthesis Based on Generative Adversarial Networks.” arXiv.org, July 2, 2018. https://arxiv.org/abs/1806.03384.

Phelps, Joseph, Glen Nowak, and Elizabeth Ferrell. “Privacy Concerns and Consumer Willingness to Provide Personal Information.” *Journal of Public Policy & Marketing* 19, no. 1 (Spring 2000): 27–41. https://doi.org/10.1509/jppm.19.1.27.16941.

Pingitore, Gina, Jay Meyers, Molly Clancy, and Kristin Cavallaro. Consumer concerns about data privacy rising - future of Privacy Forum, October 29, 2013. https://fpf.org/wp-content/uploads/2013/12/JDPA-Data-Privacy-Research-Oct-20131.pdf.

"Privacy Noun - Definition, Pictures, Pronunciation and Usage Notes: Oxford Advanced American Dictionary at Oxfordlearnersdictionaries.Com." privacy noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced American Dictionary at OxfordLearnersDictionaries.com. Accessed August 21, 2023. https://www.oxfordlearnersdictionaries.com/definition/american_english/privacy#:~:text=%2Fˈpra%C9%AAv%C9%99si%2F,I%20value%20my%20privacy.

"Quarterly Retail E-Commerce Sales Quarter 2023 - Census.Gov." U.S. Census Bureau News, August 17, 2023. https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf.

Rust, Roland T., P. K. Kannan, and Na Peng. "The Customer Economics of Internet Privacy - Journal of the Academy of Marketing Science." SpringerLink, 2002. https://link.springer.com/article/10.1177/009207002236917.

Saatci, Cansu, and Efnan Sora Gunal. Preserving privacy in Personal Data Processing | IEEE Conference ... Accessed August 20, 2023. https://ieeexplore.ieee.org/document/8965432.

Smith, H. Jeff, Sandra J. Milberg, and Sandra J. Burke. "Information Privacy: Measuring Individuals' Concerns about Organizational Practices." *MIS Quarterly* 20, no. 2 (June 1996): 167. https://doi.org/10.2307/249477.

Stone, Eugene F., and Diana L. Stone. Stone, E. F., & stone, D. L. (1990). privacy in organizations ... Accessed August 20, 2023. https://www.researchgate.net/publication/259440817_Stone_E_F_Stone_D_L_1990_Privacy_in_organizations_Theoretical_issues_research_findings_and_protection_strategies_In_G_Ferris_K_Rowland_Eds_Research_in_personnel_and_human_resources_management_Vol_8_pp.

Torkzadehmahani, Reihaneh, Peter Kairouz, and Benedict Paten. "DP-CGAN: Differentially Private Synthetic Data and Label Generation." arXiv.org, January 27, 2020. https://arxiv.org/abs/2001.09700.

White, Tiffany Barnett, Debra L. Zahay, Helge Thorbjørnsen, and Sharon Shavitt. "Getting Too Personal: Reactance to Highly Personalized Email Solicitations." *Marketing Letters* 19, no. 1 (September 4, 2006): 39–50. https://doi.org/10.1007/s11002-007-9027-9.

Xie, Liyang, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. "Differentially Private Generative Adversarial Network." arXiv.org, February 19, 2018. https://arxiv.org/abs/1802.06739.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling Tabular Data Using Conditional Gan." arXiv.org, October 28, 2019. https://arxiv.org/abs/1907.00503.

Zhang, Xinyang, Shouling Ji, and Ting Wang. “Differentially Private Releasing via Deep Generative Model (Technical Report).” arXiv.org, March 25, 2018. https://arxiv.org/abs/1801.01594.

Declaration of Academic Integrity

The participating team declares that the paper submitted is comprised of original research and results obtained under the guidance of the instructor. To the team's best knowledge, the paper does not contain research results, published or not, from a person who is not a team member, except for the content listed in the references and the acknowledgment. If there is any misinformation, we are willing to take all the related responsibilities.

Names of team members Zhuo Chen

Signatures of team members Zhuo Chen

Name of the instructor Qiwei Han

Signature of the instructor Qiwei Han

Date Aug/20 2023