Electronic Commerce Research
https://doi.org/10.1007/s10660-023-09713-5

Towards privacy-preserving digital marketing: an integrated framework for user modeling using deep learning on a data monetization platform

Qiwei Han[1] · **Carolina Lucas**[1] · **Emila Aguiar**[1] · **Patrícia Macedo**[1] · **Zhenze Wu**[1]

Accepted: 8 May 2023

Abstract
This paper presents a novel approach to privacy-preserving user modeling for digital marketing campaigns using deep learning techniques on a data monetization platform, which enables users to maintain control over their personal data while allowing marketers to identify suitable target audiences for their campaigns. The system comprises of several stages, starting with the use of representation learning on hyperbolic space to capture the latent user interests across multiple data sources with hierarchical structures. Next, Generative Adversarial Networks are employed to generate synthetic user interests from these embeddings. To ensure the privacy of user data, a Federated Learning technique is implemented for decentralized user modeling training, without sharing data with marketers. Lastly, a targeting strategy based on recommendation system is constructed to leverage the learned user interests for identifying the optimal target audience for digital marketing campaigns. Overall, the proposed approach provides a comprehensive solution for privacy-preserving user modeling for digital marketing.

Keywords Deep learning · Digital marketing · Data monetization · Data privacy · Hyperbolic embeddings · Federated learning

✉ Qiwei Han
qiwei.han@novasbe.pt

Carolina Lucas
44364@novasbe.pt

Emila Aguiar
44993@novasbe.pt

Patrícia Macedo
44359@novasbe.pt

Zhenze Wu
44524@novasbe.pt

[1] Nova School of Business and Economics, Carcavelos, Portugal

Published online: 12 June 2023

1 Introduction

Today, the use of personal data has become a critical aspect of digital marketing campaigns to both businesses and consumers [15, 26, 52]. For businesses, it enables them to create targeted campaigns that are more likely to result in conversions, by delivering the right message to the right people [9]. Personal data can also be used to gain valuable insights into consumer behavior, preferences, and trends, which can be used to inform business decisions and improve products and services [16, 30]. For consumers, personalized and relevant advertising can be more engaging and helpful, making it more likely that they will make a purchase or engage with a brand [26]. Additionally, the use of personal data can lead to more personalized experiences, such as tailored recommendations and promotions, which can enhance the overall customer experience [30]. However, consumers are becoming more privacy savvy and concerned about the lack of control over their personal information [1], while big tech companies achieve competitive advantage through getting empowered from the "free access" of user data [3]. Furthermore, consumers increasingly realize financial value of their data, and they may demand tangible returns. Anecdotal evidence shows that customer's personal email address could be worth about $90 to a brand [24]. In response, privacy regulations such as European Union's General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) have mandated digital marketing to protect consumer privacy by restricting companies' ability to collect personal data [47]. This has significant implications to digital marketing practices [36], e.g., influence the effectiveness of targeted marketing [20] and changes in market structure [47], and leads marketers to explore alternative avenues to collect and analyze consumer data in compliance to regulations.

A new type of data monetization platform has emerged to facilitate the trade of granular personal information between consumers and marketers for mutual benefits [28]. From two-sided market perspective, these platforms essentially allow consumers and marketers to perform data sharing activities as commercial transactions [7]. As such, this type of platform may not only add new data-based business model to the existing data monetization framework [44, 46], but also help build the path for tangible value co-creation [55]. [22] showcase several recent examples that offer cash or discounts to consumers in exchange for their demographic and behavioral data. To facilitate the data-based innovation and revenue generation, data monetization platforms need to 1) devise effective data governance strategies, such as pricing and information disclosure policies to ensure trusted transactions between consumer and marketers [40, 48]; 2) perform user modeling and generate meaningful representation from the collected data [13]; and provide technological solution to resolve privacy issues [23, 27].

This study was conducted in collaboration with a European data monetization platform aimed at helping European Union (EU) citizens better manage and monetize their data while complying with EU digital laws, particularly the General Data Protection Regulation (GDPR). The platform was developed as a mobile app, giving users complete control over how their data are collected, used, and

protected online. On the one hand, the app allows users to opt-in to share their data for fair compensation and receive marketing promotions. On the other hand, marketers can use the platform to target the right users by matching campaigns with user interests.

This research aims to help the platform improve its marketplace for both users and marketers while ensuring compliance with data privacy regulations. The proposed approach provides a comprehensive solution for privacy-preserving user modeling for digital marketing campaigns. By incorporating hierarchical user interests into hyperbolic space using hyperbolic embeddings, we can represent consumer behavior patterns across multiple data sources. To protect consumer privacy, synthesized user representations are generated using the Generative Adversarial Networks (GAN) technique, which approximates user interests while maintaining indistinguishability from the original ones. The training process is performed through Federated Learning (FL), a distributed learning method, which leverages data privacy and communication efficiency. The proposed solution enables the platform to identify the target audience that matches the right campaign while complying with data privacy regulations and giving users control over their data. By specifying consumer characteristics such as interests, demographics, or online behavior, marketers can create a list of users for each campaign, targeting not only the users that meet the campaign requirements but also those who are more likely to accept the offer.

The importance of data privacy and protection cannot be overstated in the context of digital marketing [36]. The proposed approach aims to protect users' personal data while allowing marketers to identify suitable target audiences for their campaigns. As a result, this study contributes to the field of e-commerce marketing by providing a privacy-preserving user modeling mechanism that benefits both consumers and marketers, because it balances the competing interests of personalization and data privacy. By utilizing advanced deep learning techniques, such as representation learning on hyperbolic space, GAN and FL, the proposed approach effectively captures latent user interests and generates synthetic representations, without compromising user privacy. This ensures compliance with data privacy regulations, while still allowing for accurate targeting and personalization of marketing campaigns.

The paper is organized as follows. Section 2 summarize the related work that covers a range of topics on representation learning for hierarchical user data, privacy-preserving machine learning approaches using GAN and FL and matching users with campaign with a recommender system. Section 3 provide step-by-step explanations in the system design. Section 4 describe the data used in this study from the collaborating data monetization platform and results from each step of the proposed approach. Section 5 discuss the business implications and conclude.

2 Related work

2.1 Representation learning for hierarchical user data

The primary objective of representation learning is to capture meaningful features from raw data to represent them in a more efficient and informative manner. In the

context of digital marketing campaigns, users often exhibit hierarchical structures in their interests and behaviors. The challenge lies in representing such hierarchical user data in a low-dimensional space while preserving their hierarchical relationships and similarities. For example, digital marketing accounts on popular platforms such as Facebook, Instagram, TikTok are typically assigned with multi-level categories to help users search and understand business [41]. Thus, extracting knowledge from hierarchical data from users' Facebook profiles that detail their interests in terms of likes to Facebook pages would require meaningful representations of hierarchical structure.

Hyperbolic embeddings have emerged as a powerful solution to address this challenge [37]. Unlike traditional Euclidean geometry, hyperbolic geometry is better suited for modeling complex networks with hierarchical data structures, as it can capture high-quality hierarchy information in a lower-dimensional space with minimal distortion [6]. This is due to the constant negative curvature of the hyperbolic space, which allows the area of a circle or volume of a sphere to expand exponentially with its radius. Consequently, the distances between objects in hyperbolic space are well preserved, providing a measure of their similarity and reflecting their semantic or functional relationships.

An example application of hyperbolic embeddings can be found in modeling users' hierarchical interests derived from their social media activities, such as Facebook profiles. By capturing users' likes and interactions with Facebook pages, a hierarchical representation of their interests can be constructed. The Poincaré ball model, a popular hyperbolic space, has been employed to learn hierarchical representations as it effectively preserves distances between categories and hierarchy in the data [37]. The utility of Poincaré embeddings is evident in their ability to preserve original graph distances, capturing the hierarchy of objects through their norm, as demonstrated in the Eq. 1 below:

$$d_H(x, y) = acosh\left(1 + 2\frac{\|x - y\|^2}{\left(1 - \|x\|^2\right)\left(1 - \|y\|^2\right)}\right) \tag{1}$$

Equation 1 Formula of distance on Hyperbolic Space.

2.2 Machine learning approaches for privacy-preserving with GAN

Privacy preservation is a crucial concern when dealing with sensitive user data, particularly in domains vulnerable to data leakage threats. Traditional privacy protection techniques, such as deidentification techniques (k-anonymization, l-diversity, and k-closeness), removal of personal identification values, and alteration of quasi-identifiers, have been found to be insufficient due to their inherent limitations [23]. These limitations include attackers' potential ability to retrieve private information when they possess background knowledge and the negative impact on the utility of the released data [2].

Generative Adversarial Networks (GANs) have emerged as a promising solution to address these challenges [21]. GANs have demonstrated a significant

breakthrough in synthesizing high-quality artificial samples that closely resemble the distribution of the original training data, making it difficult to distinguish between artificial and genuine data. For example, Deep Convolutional GAN (DCGAN) [42] generate high quality and realistic images using an adversarial process to train two models generator and discriminator simultaneously, where the generator learns to create realistic synthetic samples from random noise, while learns to differentiate between real and synthetic samples. Since then, several variants of GAN models have been proposed to generate synthetic data beyond images [61]. As an adaptation of the DCGAN, table-GAN [45] aims to synthesize relational tables that consist of different data types and are statistically identical to the original table, leveraging convolutional neural networks in the process. Another notable variation is CTGAN [56], which aims to synthesize tabular data, consisting of continuous and discrete columns, using a conditional generator. CTGAN faces challenges in handling imbalanced discrete columns and complex continuous columns with multiple nodes. Despite these challenges, CTGAN's performance has been favorable, outperforming Bayesian methods on most real datasets.

As a result, GANs have been widely adopted as a privacy-preserving mechanism across various domains subject to privacy regulations, such as healthcare, finance, and more. For instance, the healthcare industry has employed GANs including medGAN [12] and medBGAN [8] for generating artificial patient electronic health records (EHRs) that maintain the statistical properties of the original data while protecting sensitive patient information. This enables researchers and practitioners to develop and test algorithms, models, and applications without compromising patient privacy [60]. Similarly, the finance industry can utilize GANs to generate synthetic customer datasets that can be used for portfolio analysis [62], fraud detection [18], and risk assessment [33], without exposing the real customer data. In both cases, GANs offer an effective solution for preserving privacy while still enabling the use of realistic and statistically similar datasets for various purposes.

2.3 Machine learning approaches for privacy-preserving with federated learning

In recent years, the limitations of Centralized Learning have become increasingly apparent in the face of growing data complexity and volume. Centralized Learning can impose a significant burden on the network responsible for exchanging vast amounts of data and can challenge the server's ability to process large aggregates of information while ensuring data protection [14]. These issues have led to the development of alternative systems that distribute the machine learning burden across numerous computers or mobile devices, such as Federated Learning.

Federated Learning (FL) enables the training of data across multiple devices, coordinated by one or more central servers [11, 31, 57]. This approach can address the conflict between data privacy and data sharing for detached devices, as the data is not disclosed to a central server [38, 39]. Consequently, FL is well-suited for applications involving privacy-sensitive data. For example, FL has been employed in various industries such as banking for credit card fraud detection [59], image detection and representation [34], and healthcare for disease prediction and biomedical

imaging analysis using health records [50]. These applications demonstrate the potential of FL to protect privacy while maintaining the utility of shared data [11].

In particular, FL has shown great potential in privacy-preserving training for mobile devices, known as *on-device FL* [32]. In an era where mobile devices are ubiquitous and generate a plethora of personal data, ensuring the privacy and confidentiality of users becomes crucial. The decentralized nature of FL allows for the efficient utilization of data from mobile devices without compromising user privacy [11]. For example, FL can be applied to improve privacy-preserving mobile health applications. These applications require the collection and processing of sensitive health data from wearable devices, such as heart rate monitors and fitness trackers. By using on-device FL, mobile health applications can develop personalized models for users while maintaining their privacy [49].

3 Methods

3.1 Overview of the platform design

Building upon the concepts and methodologies discussed in Sect. 2, we present the design of the privacy-preserving platform for targeted marketing campaigns. The platform leverages hyperbolic embeddings, GANs, and FL to ensure privacy and efficiency while effectively targeting users based on their interests and preferences. Figure 1 illustrates the design of the platform with several key components. The core section demonstrates the modeling of users' representations using the Poincaré Embeddings, as described in Sect. 2.1. These embeddings capture the hierarchical nature of users' interests and preferences, which are then used to match them with the campaign that marketers initiate through a recommender system. To address

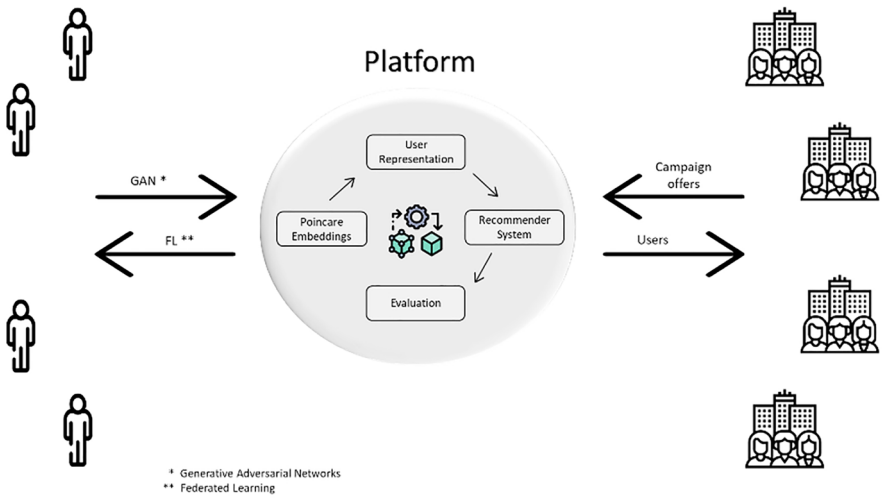

Fig. 1 Illustration of data monetization system design

privacy concerns, as discussed in Sects. 2.2 and 2.3, the platform incorporates GAN and FL techniques, shown on the left side of Fig. 1. GAN is implemented for generating synthetic user data to protect users' data privacy, while FL is implemented to allow the training process to perform decentralized on the user's own devices, aiming to leverage communication efficacy and defense techniques to ensure data privacy.

On the right side of Fig. 1, it is illustrated how the user representations are used to identify the target audience that matches the right campaign. When a marketer wants to send an offer through the platform's mobile app, they must identify the target audience by specifying user characteristics such as interests, demographics, or online behavior. In particular, for each campaign to be conducted on the platform, a list of users is created so that marketers target not only the users that meet the campaign requirements but also those who are more prone to accept the offer.

However, neither the marketers nor the platform would know which users may receive a specific campaign offer, as only the learned latent user embedding generated from synesthetic user data is used to create the list of users that matches the campaign offer. This is accomplished by modeling the user through a series of deep learning techniques detailed in the following.

3.2 User modelling from poincaré embeddings

We utilize the Poincaré embeddings to effectively model users in a hyperbolic space, enabling efficient representation of hierarchical data from user characteristics and interests. The outcome of the Poincaré model is a set of vectors or coordinates for each category on the Poincaré ball, which provides valuable insights into the position of each interest or characteristics within the space.

More specifically, each user was represented into the hyperbolic space by calculating an average of their interests and characteristics, creating one unique n-dimensional vector as Poincaré embedding [37]. Then, the Poincaré embeddings of the users can be used to create target groups of users with similar characteristics together. This was accomplished by the PoincareKMeans [5]. Notably, the implementation of this clustering method was not based on the original version of the K-Means using the Euclidean distance as a similarity measure. The context of the hyperbolic space prohibits the use of linear spaces to determine in its most accurate sense comparable features without exhausting the original graph distances in non-Euclidean space. Instead, we choose to partition users into K groups based on hyperbolic distances between their embeddings on the Poincaré ball. As such, we are able to capture the hierarchical relationship by placing users belong to the same category close to each other. Section 1 in Appendix provides more details on the training process and choice of hyperparameters of the Poincaré model.

With the target groups established, we create an avatar that represents the shared characteristics of users within each group. This process essentially offers several benefits. Firstly, representing and aggregating user information to a higher level reduces the risk for adversaries to re-identify individual users based on their interest and characteristics. Secondly, despite the anonymization, the use of avatars ensures

that the data remains useful for marketing purposes, because marketers can still identify and target specific audience segments effectively. Thirdly, hierarchical user characteristics is naturally equipped with the Poincaré embedding representation, thus making the target group creation more suitable than the one generated from similarity metrics in Euclidean space.

3.3 User representation synthesized using GAN

To synthesize user representations while preserving their privacy, we employ a customized GAN architecture. As Fig. 2 shows, the principal idea of GAN remains the same: training two neural networks simultaneously, a Generator (G) and a Discriminator (D), in a manner that G generates synthetic data resembling the original data distribution and D distinguishes fake data from authentic data to optimize the min–max loss function [32] shown in Eq. 2:

$$\min_{G}\max_{D} V(G,D) = \mathbb{E}_{x\sim pdata(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim pz(z)}\left[\log\left(1 - D(G(z))\right)\right] \tag{2}$$

Equation 2 min–max GAN Loss Function.where $p_{data(x)}$ denotes the original data distribution and $p_{z(z)}$ is the simple noise distribution (generally a normal distribution). $_{x\sim pdata(x)}$ represents the expected value in all instances of original data, and $_{z\sim pz(z)}$ represents the expected value over all random inputs to the Generator. Essentially, the goal of the generator G is to minimize the value function V(G,D), while the discriminator D aims to maximize it.

More specifically, before applying the embedding methods, GAN generates synthetic data that is indistinguishable from the original data to protect user privacy. The training process starts by feeding random noise created by the G and actual data from the original dataset into the D. The D then classifies synthetic data as real or fake. If the classification is incorrect, the D is penalized by the discriminator loss. Finally, hyperparameters are updated through backpropagation.

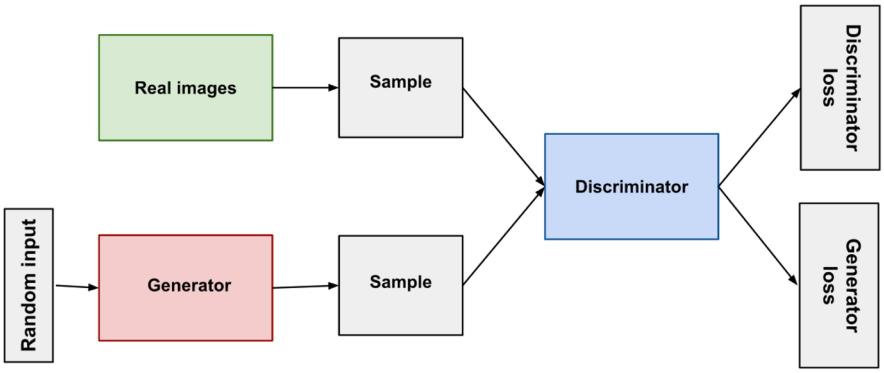

Fig. 2 Illustration of the complete system of GAN

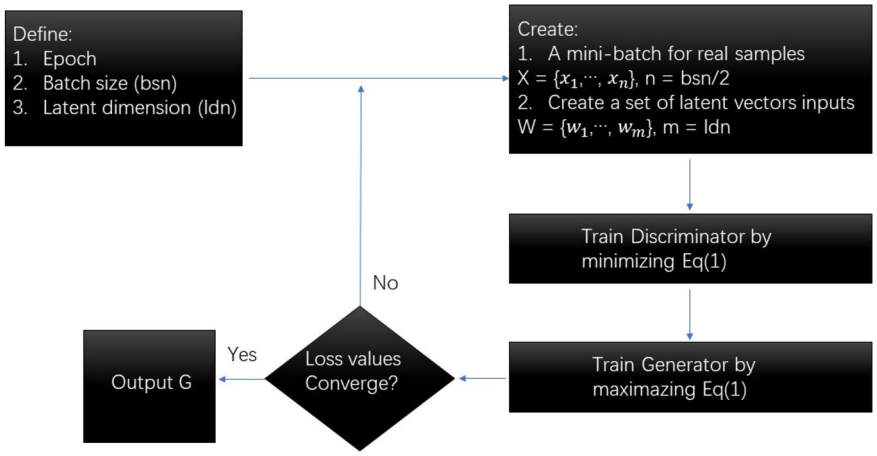

Fig. 3 Step-by-step illustration of GAN training algorithm

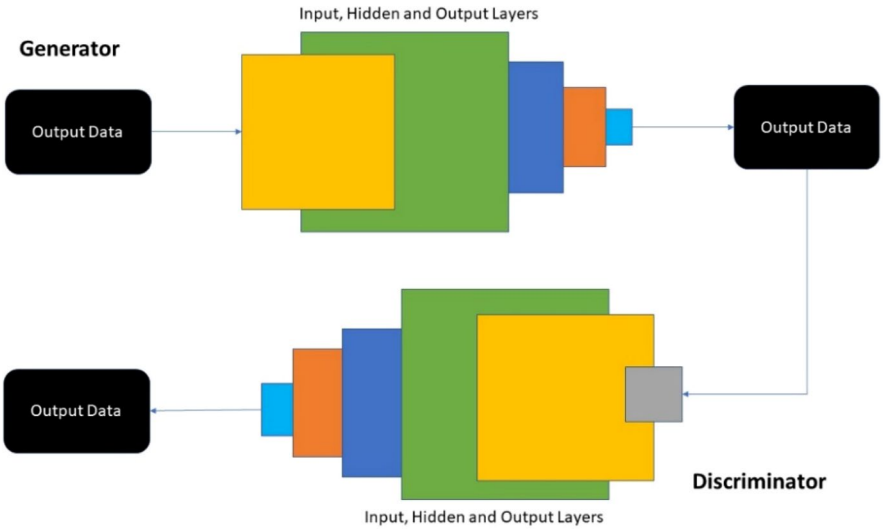

Fig. 4 Proposed GAN architecture for generator and discriminator

This process is repeated until the convergence of discriminator loss, minimizing the probability of making mistakes.

The G's capacity is highly dependent on the effectiveness of the D. It cannot be trained independently, and an evaluation from the D is necessary to update the G's hyperparameters. The G is penalized for providing a sample that the D classifies as artificial. Backpropagation starts at the output of the D and returns through it into the G. Figure 3 illustrates the GAN training algorithm:

Moreover, we have adapted the GAN architecture specifically for our use case. In particular, Fig. 4 shows the custom GAN architecture that consists of the following modifications. Firstly, our customized G comprises one input layer, three fully connected hidden layers, and one output layer. The hidden layers use the LeakyReLU activation function, while the output layer uses the tanh function. These adjustments help capture the possible correlations between variables in user data. Secondly, the D in our architecture has one input layer, four fully connected hidden layers, and one output layer. We use the LeakyReLU activation function for all hidden layers, the Sigmoid function for the output layer, and Dropout to avoid overfitting. Furthermore, we employ the Adam version of stochastic gradient descent with a learning rate of 0.0002 and a momentum of 0.5. Section 2 in Appendix provides more details on the training process of the custom GAN architecture.

The GAN implementation is a critical component of this study, as it enables the creation of synthetic data from the original user information while preserving user privacy. This prevents users from being identified and ensures the confidentiality of their information. Together with the user modeling using hyperbolic embeddings, GAN-generated synthetics data maintains the underlying structure and patterns of original user data, by adding an extra layer of privacy protection while ensuring the utility of data for digital marketing purposes.

3.4 Distributed training using federated learning

In this paper, the distributed training of the model is accomplished using a specific approach to FL, referred to as Partially Local Federated Learning [51]. This method is employed to handle situations where the model contains user-specific parameters, such as in matrix factorization [29]. Sending updates of user embeddings to the server when training a global federated model is undesirable in these cases, as it could expose potentially sensitive individual preferences. To address this issue, the model is partitioned into global and local parameters. More specifically, the matrix containing users' preferences is factorized into a user and an item matrix, generating a k-dimensional user-specific embedding for each user. This approach ensures that some parameters are not transferred to the server, although it does require clients to maintain their user embeddings across several rounds, which can be undesirable. In large-scale cross-device settings, users are unlikely to be sampled more than once during the training process, leading to performance degradation.

To tackle this challenge, a federated reconstruction framework is utilized in this paper, which eliminates the need for users to keep their local parameters across rounds by reconstructing them as necessary. A reconstruction algorithm is employed to restore the local parameters. The scheme in Fig. 5 demonstrates the process: first, for each round, the server retains and delivers the item matrix (global parameters) to the sampled users. Next, using one or more stages of SGD (Stochastic Gradient Descent), each user freezes the item matrix and trains their user embedding (local variables). Subsequently, each user freezes their user embedding and trains the item matrix using one or more steps of SGD. Finally, updates to the item matrix are

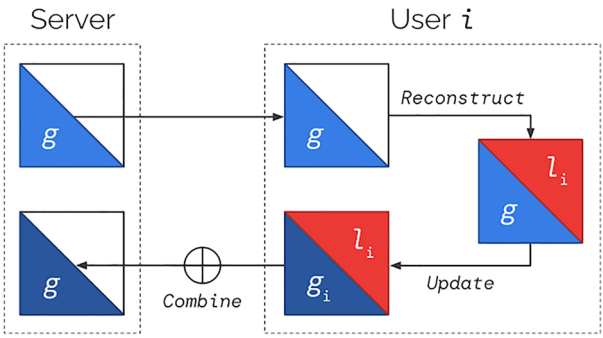

Fig. 5 Scheme of partial local federal learning adapted from [51]

collected across users, and the server copy of the item matrix is updated for use in the next round.

Notably, the training process in this study deviates from the typical federated learning process, which often employs federated averaging for federated aggregation. Instead, a reconstruction optimizer is passed to reconstruct the parameters that remain local, such as user embeddings. For both the server and user optimizer, the same SGD optimizer is used; however, the learning rates differ. Section 3 in Appendix illustrates the steps of the FL training framework.

3.5 Matching users with marketing campaigns using HyperML

The final step of the process involves leveraging the synthesized user embeddings and incorporating them into a recommender system model, to accurately identify the list of most suitable users for a given campaign offer. Given that user characteristics and interest are represented by Poincaré embeddings, we adopt a tailored approach Hyperbolic Metric Learning (HyperML) [53], a technique that is specifically designed for recommender systems in hyperbolic space such as the manifold Poincaré Ball for its computations of similarities. In particular, the training process with HyperML takes place on the Poincaré Ball manifold, which computes distances in the hyperbolic space. This allows for the learning of Poincaré embeddings of insights, master categories, and users, and ultimately, the objective is to predict new users who may be interested in a specific insight or category.

Selecting the appropriate users is crucial for the success of a campaign. As such, it is essential to identify a user base that ensures not only will marketers target their offer to those most likely to accept the campaign, but also that clients receive the offers they need and desire. The primary goal, therefore, is to recommend users for a specific campaign based on their interests and personal information, using this uniquely devised approach.

The benefits of employing this specialized approach for matching users with marketing campaigns are multifold. By utilizing HyperML and the Poincaré Ball manifold, the recommender system can better identify suitable users for specific

campaigns from their Poincaré embeddings, leading to improved targeting and more effective marketing strategies. This, in turn, can result in higher conversion rates and better return on investment (ROI) for marketers. Furthermore, this method allows for enhanced personalization of marketing campaigns, ensuring that users receive offers that are relevant and tailored to their needs and interests. This enhances user satisfaction and fosters loyalty to both brand and platform.

4 Results

4.1 Data description

The dataset used in this paper is generated from synthetic user data modeled on 10,000 users' real-world Facebook profiles. This synthetic data maintains the structure and distribution of the original data while protecting user privacy. The synthetic dataset details users' interests in terms of likes on Facebook pages. It is important to note that no actual Facebook profile data was accessed or used directly in this research, as the data utilized is entirely synthesized. The synthetic dataset can be obtained from the corresponding author upon reasonable request and subject to approval by the collaborating data monetization platform. Additionally, any supplementary materials, including code and models, can be accessed upon reasonable request and with the appropriate permissions from the authors.

Each Facebook page in the synthetic dataset belongs to an intrinsic category, which is hierarchically organized with a 3-level depth, necessitating the consideration of user interest hierarchies in user modeling. As shown in Fig. 6, an example is the category "Media" that contains child categories of "Music" and "Books & Magazines," with the Music category further containing child categories of "Song" and "Album." Consequently, user interests are represented as a set of hierarchical data containing different levels of categories. In total, there is a total of 1563 categories.

To match users with campaigns, it is crucial to identify users with specific interests and consider their preferences:

- If the user has a real interest in a category, the value is 1.
- If the user does not have an interest (explicitly), the value is -1.

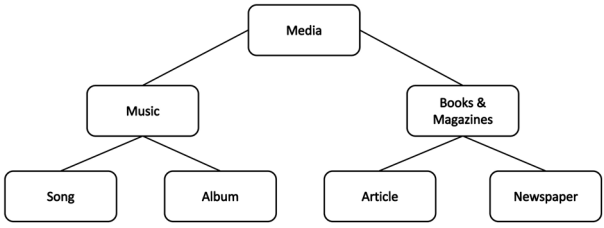

Fig. 6 Hierarchical facebook category example

For other categories that are not assigned a value of 1 or −1, they are considered unknown interests (value = 0), meaning we do not know whether the user has an interest or not.

4.2 Result from poincaré embeddings

Ideally, users with similar interests in the same parent category should be represented closely, which is challenging using traditional one-hot encoding in Euclidean space. To address this challenge, a Poincaré embedding model is adopted to learn user representations, transforming user interests into hyperbolic space, and calculating similarity using non-Euclidean distance.

We obtain Poincaré embeddings for all users as a set of n-dimensional vectors or coordinates of each category in the hyperbolic space, where n varies between 25 to 200. To evaluate the quality of Poincaré embeddings, the same approaches were used as described by [37]: **Reconstruction error** in relation to the embedding dimension, which is the reconstruction of a hierarchy from the embedding to evaluate the representation capacity; and **Link prediction** by splitting the data into a train, validation, and test set to evaluate generalization performance. We compare these tasks using Poincaré distance with traditional Euclidean distance $d(x, y) = \| x - y \|^2$.

These two approaches can be determined by two evaluation metrics:

- **Mean Rank**: the average of the ranks for all observations within each sample.
- **MAP**: refers to Mean Average Precision, a metric that captures how well each vertex's neighborhoods are preserved.

The evaluation results of the Poincaré embeddings are shown in Table 1. Examining the reconstruction task, we observe that Poincaré embeddings yield consistently better representation quality in comparison with Euclidean embeddings. Moreover, the quality largely increases with the dimensionality, while such improvement becomes marginal for 200 dimensions. Similarly, for Link Prediction task, we observe better representation quality for Poincaré embeddings.

Table 1 Evaluation of embedding quality obtained from reconstruction and link prediction tasks in hyperbolic space and Euclidean space

		Reconstruction				Link prediction			
		25D	50D	100D	200D	25D	50D	100D	200D
Euclidean	Mean Rank↓	4.56	4.28	4.19	4.12	3.95	3.85	3.73	3.51
	MAP↑	0.439	0.445	0.448	0.451	0.392	0.408	0.411	0.453
Poincaré	Mean Rank↓	2.59	2.55	**2.50**	2.53	3.19	2.98	3.07	**2.76**
	MAP↑	0.534	0.534	0.536	**0.54**	0.413	0.418	0.416	**0.538**

The bold indicates the optimal performance obtained from the model tasks and parameters. For Mean Rank, the lower the better. For MAP, the larger the better

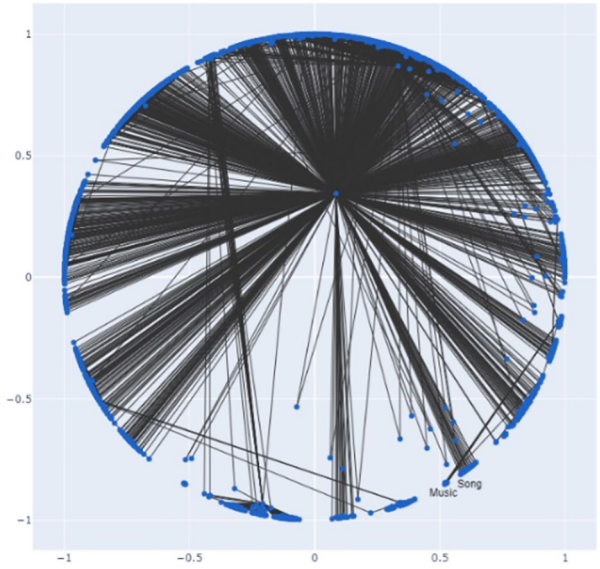

Fig. 7 Representation of the Poincaré embeddings in 2D

Moreover, the Mean Rank decreased from 3.2 to 2.76, and the MAP value rose from 0.41 to 0.538, indicating a significant improvement with the dimensionality. After fine-tuning our model, the best model has the following characteristics: 20 negative samples, 0 burn-in initialization, no regularization, 200 epochs in a 200-dimensional space. These results demonstrate that dataset is better represented on the hyperbolic space in terms of distances between interests and preserving the hierarchy of the 3 levels of categories.

Figure 7 displays the obtained embeddings of categories on the Poincaré Model as blue dots. The straight black lines represent the relations between the hierarchy levels of the categories. Additionally, the figure demonstrates that more similar categories should be situated closely, as in the case of "Music" and "Song."

Furthermore, to further demonstrate the representation quality of the Poincaré representations, we can partition all users into multiple groups with similar characteristics and interests using PoincaréKMeans model. Given that PoincaréKMeans is essentially an unsupervised learning algorithm, where the evaluation metrics for clustering analysis are still in development, the choice for the number of clusters is primarily based on domain knowledge and intuition. Figure 8 shows the representation of the partitions of user into 6 target groups using hyperbolic embeddings. Section A of Appendix shows the details of group characteristics from clustering analysis.

Such representations help our analysis and understanding of the characteristics of each group, for which it may increase the accuracy of representing categories for a user because the likelihood of sharing similar preferences to other users in the same group tend to be higher.

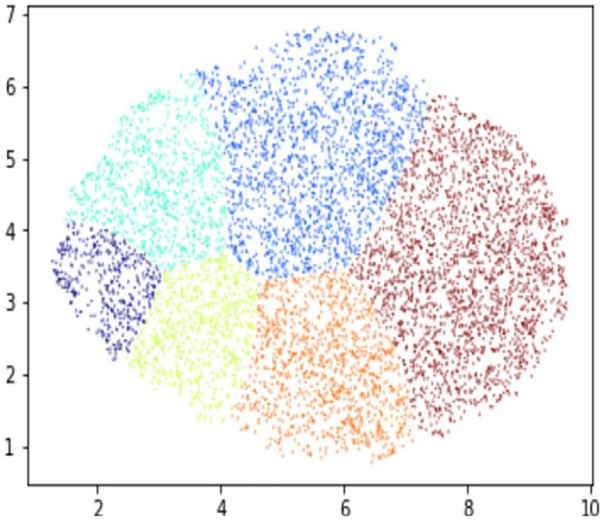

Fig. 8 Representation of the users into 6 clusters in Poincaré space

4.3 Results from GAN

We evaluate the performance of the GAN model by comparing the distribution of the original dataset with that of the synthesized dataset generated by the model. By analyzing the similarities between the two datasets, we can observe that the GAN-generated data effectively preserves the distributional characteristics of the original data, while ensuring user privacy and significantly reducing the risk of data leakage. Tables 5 and 6 in Sect. 2 of Appendix provide a comparison of the distribution for the original and the synthesized dataset, respectively.

The successful generation of a synthetic dataset that closely resembles the original data offers several advantages. Firstly, it allows for the safe application of the synthesized data in the subsequent steps of the process, with minimal risk of privacy breaches. Secondly, even if an attacker possesses some background information about a user's profile, they would be unable to extract any additional information from the synthetic dataset. This is because the synthesized data is generated by the GAN model and does not directly correspond to any real person. By achieving a balance between data utility and privacy preservation, the GAN model demonstrates its potential in addressing privacy concerns in sensitive data applications. Consequently, the synthesized dataset can be used for further analysis and modeling without compromising the privacy of users in the original dataset.

4.4 Results from federated learning

The performance of the federated learning model is assessed based on the Mean Squared Error loss and accuracy metrics computed for each sampled user on an unobserved portion of local data. This calculation uses the user-item matrix and

Table 2 Federated learning from the user-item matrix reconstruction evaluation

Metrics	Validation set	Test set
Loss	0.058	0.056
Accuracy	94%	94.2%

Table 3 Recommender system evaluation

	HR@10
Recommender system for bottom categories	0.904
Recommender system for master categories	0.965

the reconstructed user embedding by minimizing loss between observed ratings and predicted ratings as the dot product of item and user embeddings. By averaging the losses and accuracies across users, we obtain the total loss and accuracy for the model. Table 2 illustrates the loss and accuracy results for the user-item matrix reconstruction trained with federated learning on the validation and test sets.

Here, the model demonstrates satisfactory performance, with promising potential for successful implementation in real-world applications. However, it is worth noting that the high accuracy scores could also imply that the sparsity in the matrix may have resulted in an unbalanced dataset. Despite this potential issue, the federated learning model can effectively train the global and reconstructed local parameters without requiring direct access to users' specific information. This demonstrates the privacy-enhancing capabilities of the federated learning approach, ensuring that users' sensitive data remains secure while still allowing for accurate modeling and predictions.

4.5 Result from matching users using recommender system

Lastly, the synthesized user embeddings are incorporated into a recommender system model to match marketing campaigns using HyperML algorithm [53]. The primary goal is to recommend new users for a specific campaign based on their interests and personal information using their hyperbolic representations. The model identifies a list of users for a specific campaign offer and evaluates its performance using HR@10, a metric that measures the percentage of identified top 10 users interested in the particular campaign. Table 3 presents the evaluation results for matching campaigns with top 10 users at both bottom categories and master (root) categories:

The results show that the campaigns can be matched accurately, with over 90% of users likely to accept the offer. This performance enables marketers to confidently target their campaigns to the matched users without knowing their identities.

In summary, the GAN-generated synthetic dataset is trained to learn hyperbolic embeddings from hierarchical user data in a decentralized FL approach without

exposing sensitive user information. The hyperbolic embedding model, in turn, reconstructs the user-item matrix, which is then used in the recommender system to match marketing campaigns to potential users based on their interests and personal information. The combination of these methods effectively preserves user privacy while ensuring accurate marketing campaign targeting.

5 Discussions and conclusions

This research presented a novel approach to privacy-preserving user modeling for digital marketing campaigns using deep learning techniques on a data monetization platform. The primary goal of this study was to develop a solution that enables marketers to accurately identify suitable target audiences for their campaigns, while allowing users to maintain control over their personal data and ensuring compliance with data privacy regulations. The proposed approach integrated representation learning on hyperbolic space, Generative Adversarial Networks (GAN), and Federated Learning (FL) to capture latent user interests, generate synthetic representations, and perform decentralized training, all without sharing user data with marketers. By implementing a targeting strategy based on recommendation systems, the learned user interests were leveraged to identify the optimal target audience for digital marketing campaigns.

The contributions of this study include the 1) development of a privacy-preserving user modeling mechanism that balances personalization and data privacy, 2) the practical application of deep learning techniques for user modeling in digital marketing, and 3) the exploration of new business models that facilitate value co-creation between consumers and marketers on data monetization platforms.

However, this research has several limitations. First, the Poincaré Ball model cannot obtain the relations between categories and users, leading to potentially inaccurate user representations. Users' interests may change over time, which could affect the effectiveness of the model. Second, the GAN technique, while useful for generating synthetic user profiles, has limitations in generating categorical data and addressing imbalances in the dataset. Future research could investigate variations of GANs, such as medWGAN, medBGAN, and CTGAN, or experiment with different architectures for improved results. Third, the adoption of FL techniques for data synthesis and decentralized training offers several advantages, including reducing communication constraints and adding layers of privacy protection. However, this method is not infallible, and limitations exist in terms of the degree of protection it provides. Future research could explore applications of differential privacy or secure aggregation to further safeguard users' information from attacks. Fourth, the performance of the proposed approach was evaluated using a limited set of metrics, which may not capture all relevant aspects of user modeling and privacy preservation. Last, the study was conducted in collaboration with a single data monetization platform, which may limit the generalizability of the findings to other platforms or contexts.

This study represents an important step towards a more transparent, ethical, and sustainable digital marketing ecosystem that benefits both consumers and marketers alike. Future research directions include (1) exploring the applicability of the

proposed approach to different types of online platforms and E-commerce marketing contexts, (2) refine the proposed architectures with advances in deep learning techniques and evaluate the impact on performance, and (3) extending the evaluation framework to consider a broader range of metrics, such as user satisfaction, user privacy risk, and the potential impact on marketing outcomes. By addressing these limitations and building on the contributions of this study, future research can further enhance our understanding of privacy-preserving user modeling and its implications for the digital marketing landscape.

Appendix

Section 1: poincaré model training

In the code, a function called Model was created to train the Poincaré Model, based on some parameters that will be evaluated in order to get the best model possible.

In this case, fine tuning was as follows:

- Negative: Number of negative samples to use.
- Size: Number of dimensions of the trained model
- Burn-in: Number of epochs to use for burn-in initialization (0 means no burn-in)
- Regularization_coeff: Coefficient used for l2-regularization while training (0 effectively disables regularization)
- Epochs: hyperparameter that controls the number of complete passes through the training dataset.

To get the best model, the fine tuning of the parameters described above is needed. The values for the parameters are the following:

- Negatives: 10, 20;
- Dimensions: 5, 25, 50, 100, 200;
- Epochs: 50, 100, 150, 200;
- Burn-in: 0, 10;
- Regularization coeff: 0, 1;

Figure 9 illustrates the code that was built to apply these parameters. In Fig. 10, it is shown the best model hyperparameters (Fig. 11).

```
negatives_array=[10, 20]
dimensions = [5, 25, 50, 100, 200]
epochs_array = [50, 100, 150, 200]
burn_in_array = [0, 10]
regularization_coeff = [0, 1]

# Try the different values of the parameters;

for neg in negatives_array:
    for di in dimensions:
        for epo in epochs_array:
            for burn in burn_in_array:
                for reg in regularization_coeff:
                    model = Model(negative = neg, size = di, epochs = epo, burn_in=burn, regularization_coeff=reg )
```

Fig. 9 Fine-tuning of parameters on Poincaré Model

After analyzing, the choosen model was `200D_20N_0B_0R_200Ep` which has:

- `20 negatives`
- `0 burn-in`;
- `0 regularization`;
- `200 Epochs`;
- `200 Dimensions`;

Fig. 10 Best model parameters

The following tables identify the main characteristics of each cluster. The target groups are characterized as follows:

Cluster 1

Users are interested in Real Estate and Media, as well as arts and entertainment.

Businesses–Property
Businesses–Media/news company
Businesses–Arts & Entertainment

Cluster 2

Users are interested in commerce, sports centers, cinema, food and drink and religious places.

Businesses—Commercial & industrial
Businesses—Medical & health
Non-business places—religious
Businesses—Sports & Recreation

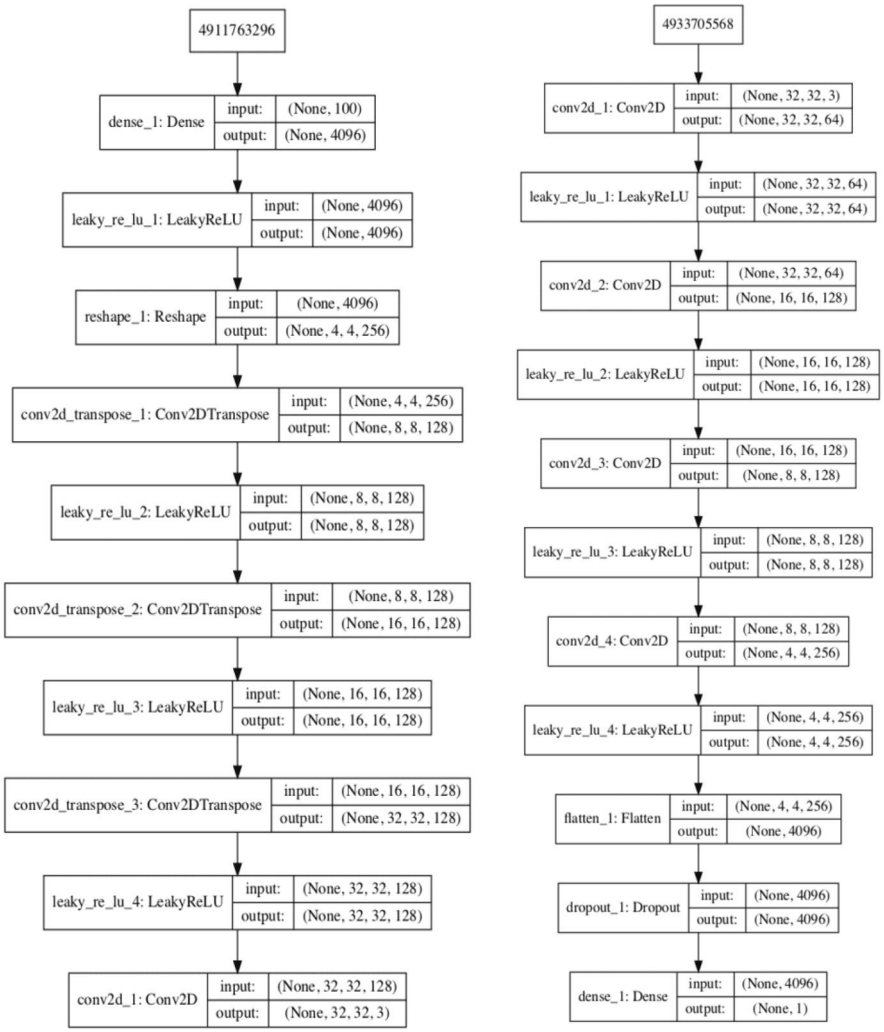

Fig. 11 Original architecture of the GAN model

Businesses—Commercial & industrial
Businesses—Food & Drink

Cluster 3

Users are interested in education, finance services and technology companies.

Businesses—Education
Businesses—Finance
Businesses—Science, technology & engineering

Cluster 4

Users like to visit Restaurants, and are interested in cars, motorcycles and boats. They also like sports.

Businesses—Food & Drink
Businesses—Vehicle, aircraft and boat
Businesses—Sport & recreation
Shopping & Retail

Cluster 5

Users like all types of brands, like to buy items and travel, and like books and music.

Other—Brand
Businesses—Food & Drink
Businesses—Travel & Transport
Media

Cluster 6

Users are interested in Medical and health; shopping and visiting restaurants.

Businesses—Medical & health
Shopping & Retail
Businesses—Food & Drink
Sports

Table 4 Generated data used for training the GAN model

	User_id	Insight_id	True	False	Birth_year	Gender	Civil_status	Degree	Professional_status	Annual_income	Practice sports
0	235	1149	1	0	1	1	0	4	2	0	1
1	5056	1149	1	0	0	1	2	5	6	3	0
2	5374	1149	1	0	7	0	0	0	4	0	1
3	735	1149	0	1	1	2	0	5	2	1	0
4	9719	1149	1	0	2	1	3	1	0	2	0

Table 5 Synthesized dataset

	User_id	Insight	True	False	Birth_year	Gender	Civil_status	Degree	Professional_ status	Annual_ income	Prac- tice sports
0	2105	1536	1	0	1	0	0	1	5	4	1
1	7864	1494	1	0	6	0	0	1	1	1	1

Model: "Discriminator"

Layer (type)	Output Shape	Param #
dense_7 (Dense)	(None, 11)	132
leaky_re_lu_6 (LeakyReLU)	(None, 11)	0
dropout_5 (Dropout)	(None, 11)	0
dense_8 (Dense)	(None, 100)	1200
leaky_re_lu_7 (LeakyReLU)	(None, 100)	0
dropout_6 (Dropout)	(None, 100)	0
dense_9 (Dense)	(None, 150)	15150
leaky_re_lu_8 (LeakyReLU)	(None, 150)	0
dropout_7 (Dropout)	(None, 150)	0
dense_10 (Dense)	(None, 50)	7550
leaky_re_lu_9 (LeakyReLU)	(None, 50)	0
dropout_8 (Dropout)	(None, 50)	0
dense_11 (Dense)	(None, 30)	1530
leaky_re_lu_10 (LeakyReLU)	(None, 30)	0
dropout_9 (Dropout)	(None, 30)	0
dense_12 (Dense)	(None, 10)	310
leaky_re_lu_11 (LeakyReLU)	(None, 10)	0
dense_13 (Dense)	(None, 1)	11

Total params: 25,883
Trainable params: 25,883
Non-trainable params: 0

Model: "Generator"

Layer (type)	Output Shape	Param #
dense_214 (Dense)	(None, 50)	5050
leaky_re_lu_177 (LeakyReLU)	(None, 50)	0
dense_215 (Dense)	(None, 150)	7650
leaky_re_lu_178 (LeakyReLU)	(None, 150)	0
dense_216 (Dense)	(None, 80)	12080
leaky_re_lu_179 (LeakyReLU)	(None, 80)	0
dense_217 (Dense)	(None, 15)	1215
leaky_re_lu_180 (LeakyReLU)	(None, 15)	0
dense_218 (Dense)	(None, 11)	176

Total params: 26,171
Trainable params: 26,171
Non-trainable params: 0

Fig. 12 New structure of GAN

Section 2: GAN training

The figure below demonstrates the original architecture of the GAN model (Tables 4 and 5).

In order to get reliable results on the Adversarial model, a new structure was implemented, as shown in Fig. 12.

The training of the GAN model encompassed the use a randomly generated dataset in which the one-hot encoding technique was applied, as shown in Table 6:

Before starting the training process, a standardization was performed to the original data set, using the function StandardScaler from Sklearn preprocessing module, due to the following reasons:

1. Standardizing the data would accelerate the training process.

Table 6 Original dataset

	User_id	Insight_id	True	False	Birth_year	Gender	Civil_status	Degree	Professional_status	Annual_income	Practice sports
0	235	1149	1	0	1	1	0	4	2	0	1
1	5056	1149	1	0	0	1	2	5	6	3	0

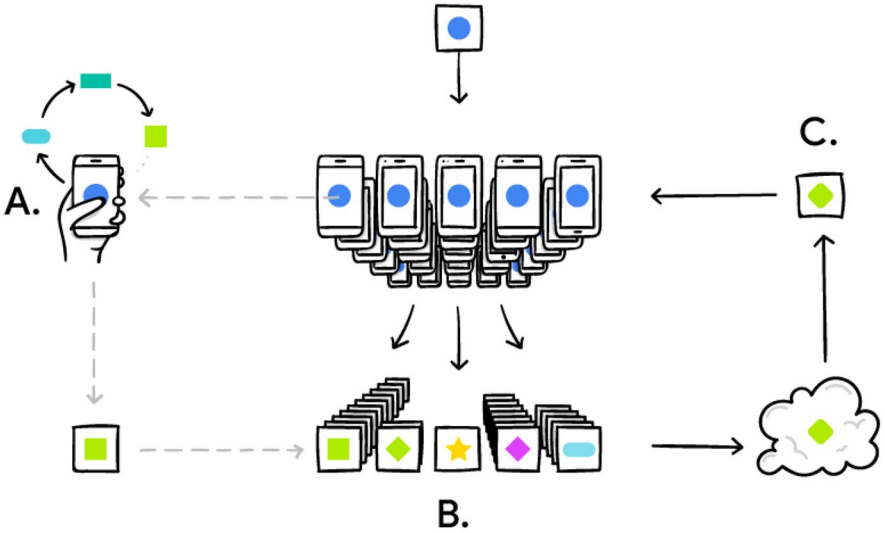

Fig. 13 Illustration of the federated learning framework

2. The output results range from [−1, 1] and [0, 1] for tanh and sigmoid functions, respectively. Thus, a standardization was performed to maintain all the parameters in the same order of magnitude.

After several experiments, the following values for the hyperparameters were chosen: n_epochs = 50, batch_size = 50 and latent_dim = 100 for the training process. The results are demonstrated in the next section.

Section 3: federated learning training

The underlying idea behind the figure above is the following (McMahan & Ramage, 2016):

- In step A, the local device uses their training data to personalize the model, based on usage;
- In step B, many of the updates to the model, from local usage, are aggregated to form a consensus change, normally performed with a Federated Averaging algorithm;
- In step C, the consensus change is share to the global model;
- The process repeats (Fig. 13).

Funding Open access funding provided by FCT|FCCN (b-on). This work was funded by Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020 and Social Sciences DataLab–PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016).

References

1. Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature, 54*(2), 442–492. https://doi.org/10.1257/jel.54.2.442
2. Agarwal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ACM SIGMOD Record, 29*(2), 439–450. https://doi.org/10.1145/335191.335438
3. Aiolfi, S., Bellini, S., & Pellegrini, D. (2021). Data-driven digital advertising: Benefits and risks of online behavioral advertising. *International Journal of Retail and Distribution Management, 49*(7), 1089–1110. https://doi.org/10.1108/IJRDM-10-2020-0410
4. Aïvodji, U. M., Gambs, S., & Martin, A. (2019). IOTFLA : A secured and privacy-preserving smart home architecture implementing federated learning. In *2019 IEEE Security and Privacy Workshops (SPW).* San Franscisco: IEEE.
5. AlexPof. (2021, 11). *PoincareKMeans: K-Means algorithm in the Poincare Disk Model*. Retrieved Feburary 7, 2023, from https://github.com/AlexPof/PoincareKMeans
6. Balazevic, I., Allen, C., & Hospedales, T. (2019). Multi-relational poincaré graph embeddings. Advances in Neural Information Processing Systems, 32.
7. Bataineh, A.S., Mizouni, R., Barachi, M., & Bentahar, J. Monetizing Personal Data: A Two-Sided Market Approach, *Procedia Computer Science*, 83, 472–479. https://doi.org/10.1016/j.procs.2016.04.211
8. Baowaly, M. K., Lin, C.-C., Liu, C.-L., & Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocy142
9. Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S., & Bala, P. K. (2020). Personalized digital marketing recommender engine. *Journal of Retailing and Consumer Services, 53*, 101799. https://doi.org/10.1016/j.jretconser.2019.03.026
10. Balazevic, I., Allen, C., & Hospedales, T. (2019). Multi-relational poincaré graph embeddings. *Advances in Neural Information Processing Systems, 32*, 4463–4473.
11. Cheng, Y., Liu, Y., Chen, T., & Yang, Q. (2020). Federated learning for privacy-preserving AI. *Communications of the ACM, 63*(12), 33–36. https://doi.org/10.1145/3387107
12. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2018). Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of 2nd Machine Learning Research for Healthcare Conference, 68:286–305*
13. Choi, H., & Mela, C. F. (2019). Monetizing online marketplaces. *Marketing Science, 38*(6), 948–972. https://doi.org/10.1287/mksc.2019.1197
14. Drainakis, G., Katsaros, K. V., Pantazopoulos, P., Sourlas, V., & Amditis, A. (2020). Federated vs. Centralized machine learning under privacy-elastic users: A comparative analysis. In *2020 IEEE 19th International Symposium on Network Computing and Applications (NCA)* (pp. 1–8). IEEE - Institute of Electrical and Electronics Engineers.
15. Evans, D. S. (2009). The online advertising industry: economics, evolution, and privacy. *Journal of Economic Perspectives, 23*(3), 37–60. https://doi.org/10.1257/jep.23.3.37
16. Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research, 2*(1), 28–32. https://doi.org/10.1016/j.bdr.2015.02.006

17. Faroukhi, A.Z., El Alaoui, I., Gahi, Y. et al. (2020). Big data monetization throughout Big Data Value Chain: a comprehensive review. *Journal of Big Data*, 7(3). https://doi.org/10.1186/s40537-019-0281-5.
18. Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences, 479*, 448–455. https://doi.org/10.1016/j.ins.2017.12.030
19. Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients – How easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems, 33*, 16937–16947.
20. Goldfarb, A., & Tucker, C. E. (2010). Privacy regulation and online advertising. *Management Science, 57*(1), 57–71. https://doi.org/10.1287/mnsc.1100.1246
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio Y. (2014). Generative Adversarial Nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems.*
22. Hunter, T., (2023) *These companies will pay you for your data. Is it a good deal?* Washington Post. Retrieved March 12, 2023, https://www.washingtonpost.com/technology/2023/02/06/consumers-paid-money-data/.
23. Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: A technological perspective and review. *Journal of Big Data, 3*(1), 1–25. https://doi.org/10.1186/s40537-016-0059-y
24. Jenkings, R., (2012). *How much is your email address worth?* The Drum. Rerieved March 12, 2023, from https://www.thedrum.com/opinion/2012/04/04/how-much-your-email-address-worth.
25. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N. & et al. (2021). Advances and Open Problems in Federated Learning, *Foundations and Trends® in Machine Learning:* 14(1–2), pp 1–210. https://doi.org/10.1561/2200000083.
26. Kannan, P. K., & Li, H. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing, 34*(1), 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006
27. Ke, T., & Sudhir, K. (2022). Privacy Rights and data security: GDPR and personal data markets. *Management Science*. https://doi.org/10.1287/mnsc.2022.4614
28. Kemppainen, L., Koivumaki, T., Pikkarainen, M., & Poikola, A. (2018). Emerging revenue models for personal data platform operators: When individuals are in control of their data. *Journal of Business Models, 6*(3), 79–105. https://doi.org/10.5278/ojs.jbm.v6i3.2053
29. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30–37. https://doi.org/10.1109/MC.2009.263
30. Lambrecht, A., Goldfarb, A., Bonatti, A., et al. (2014). How do firms make money selling digital goods online? *Marketing Letters, 25*, 331–341. https://doi.org/10.1007/s11002-014-9310-5
31. Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., & He, B. (2021). A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2021.3124599
32. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749
33. Lin, S.-Y., Liu, D.-R., & Huang, H. P. (2022). Credit default swap prediction based on generative adversarial networks. *Data Technologies and Applications, 56*(5), 720–740. https://doi.org/10.1108/DTA-09-2021-0260
34. Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2020). Federated Learning for Vision-and-Language Grounding Problems. *AAAI Conference on Artificial Intelligence.*
35. Liu, C-H. & Chen, C.L., "A Review Of Data Monetization: Strategic Use Of Big Data" (2015). *ICEB 2015 Proceedings (Hong Kong, SAR China).* 10.
36. Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science, 45*(2), 135–155. https://doi.org/10.1007/s11747-016-0495-4
37. Maximilian, N., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems, 30*, 6341–6350.
38. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* pp. 1273–1282.

39. McMahan, B., & Ramage, D. (2016). *Federated Learning: Collaborative Machine Learning without Centralized Training Data*. Google Research Blog, Retrieved Feburary 8, 2023, from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
40. Mehta, S., Dawande, M., Janakiraman, G., & Mookerjee, V. (2021). How to sell a data set? pricing policies for data monetization. *Information Systems Research, 32*(4), 1281–1297. https://doi.org/10.1287/isre.2021.1027
41. Meta Business Help Center. *Best practices to choose a category for your Page or profile on Facebook*. Rerieved March 12, 2023, from https://www.facebook.com/business/help/376650512904346?id=939256796236247.
42. Metz, L., Radford, A., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *International Conference on Learning Representations 2016.*
43. Najjar, M. S., & Kettinger, W. J. (2013). Data monetization: Insights from a technology-enabled literature review and research agenda. *MIS Quarterly Executive, 4*, 213–225.
44. Ofulue, J., & Benyoucef, M. (2022). Data monetization: Lesson from a Retailer's journey. *Management Review Quarterly*. https://doi.org/10.1007/s11301-022-00309-1
45. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data Synthesis based on Generative Adversarial Networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083. https://doi.org/10.14778/3231751.3231757
46. Parvinen, P., Pöyry, E., Gustafsson, R., Laitila, M., & Rossi, M. (2020). Advancing data monetization and the creation of data-based business models. *Communications of the Association for Information Systems*, 47, 25–49. https://doi.org/10.17705/1CAIS.04702
47. Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2022). Regulatory spillovers and data governance: Evidence from the GDPR. *Marketing Science, 41*(4), 746–768. https://doi.org/10.1287/mksc.2021.1339
48. Ray, J., Menon, S., & Mookerjee, V. (2020). Bargaining over data: When does making the buyer more informed help? *Information Systems Research, 31*(1), 1–15. https://doi.org/10.1287/isre.2019.0872
49. Rieke, N., Hancox, J., Li, W. et al. (2020). The future of digital health with federated learning. *npj Digital. Medicine.* 3, 119. https://doi.org/10.1038/s41746-020-00323-1
50. Silva, S., A. Gutman, B., Romero, E., M. Thompson, P., Altmann, A., & Lorenzi, M. (2019). Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).* IEEE.
51. Singhal, K., Sidahmed, H., Garrett, Z., Wu, S., Rush, K., & Prakash, S. (2021). Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems, 34*, 11220–11232.
52. Taylor, C. R. (2004). Consumer privacy and the market for customer information. *RAND Journal of Economics, 35*(4), 631–650. https://doi.org/10.2307/1593765
53. Vinh Tran, L., Tay, Y., Zhang, S., Cong, G., & Li, X. (2020). HyperML: A boosting metric learning approach in hyperbolic space for recommender systems. *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM'20)*, 609–617. https://doi.org/10.1145/3336191.3371850
54. Wu, J., Liu, Q., Huang, Z., Ning, Y., Wang, H., Chen, E., &, et al. (2021). Hierarchical personalized federated learning for user modeling. *In Proceedings of the Web Conference*. https://doi.org/10.1145/3442381.3449926
55. Visconti, R.M., Larocca, A., & Marconi, M. (2017). Big data-driven value chains and digital platforms: From value co-creation to monetization. In Arun K. Somani, Ganesh Chandra Deka (Eds.), *Big Data Analytics* (1st ed., pp.345–362), Chapman and Hall.
56. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems.*
57. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1–19. https://doi.org/10.1145/3298981
58. Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., & Beaufays, F. (2018). Applied federated learning: Improving google keyboard query suggestions.
59. Yang, W., Zhang, Y., Ye, K., Li, L., & Xu, C.-Z. (2019). FFD: A federated learning based method for credit card fraud detection. *Lecture Notes in Computer Science, 11514*, 18–32.

60. Yoon, J., Drumright, L., & van der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics, 24*(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262
61. Zhang, H., Sun, Y., Liu, L., Wang, X., Li, L., & Liu, W. (2020). ClothingOut: A category-supervised GAN model for clothing segmentation and retrieval. *Neural Computing and Applications, 32*(9), 4519–4530. https://doi.org/10.1007/s00521-018-3691-y
62. Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock market prediction based on generative adversarial network. *Procedia Computer Science, 147*, 400–406. https://doi.org/10.1016/j.procs.2019.01.256