JAISCR, 2024, Vol. 14, No. 4, pp. 309 – 319
10.2478/jaiscr-2024-0017

ACCELERATING USER PROFILING IN E-COMMERCE USING CONDITIONAL GAN NETWORKS FOR SYNTHETIC DATA GENERATION

Marcin Gabryel[1,2,*], Eliza Kocić[2], Milan Kocić[2], Zofia Patora-Wysocka[3], Min Xiao[4], and Mirosław Pawlak[5]

[1]*Department of Intelligent Computer Systems, Częstochowa University of Technology, 42-200 Częstochowa, Poland*

[2]*Spark Digitup, Plac Wolnica 13 lok. 10, 31-060 Kraków, Poland*

[3]*Management Department, University of Social Sciences, 90-113 Łódź, Poland*

[4]*College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

[5]*Information Technology Institute, University of Social Sciences, 90-113 Łódź, Poland*

[*]*E-mail: marcin.gabryel@pcz.pl*

Submitted: 13th February 2024; Accepted: 25th June 2024

Abstract

This paper presents the findings of a study on the profiling of online store users in terms of their likelihood of making a purchase. It also considers the possibility of implementing this solution in the short term. The paper describes the process of developing a profiling model based on data derived from monitoring user behaviour on a website. During the customer's subsequent visits, information is collected to identify the user, record their behaviour on the page and the fact that they made a purchase. The model requires a substantial amount of training data, primarily related to the purchase of products. This represents a small percentage of total website traffic and requires a considerable amount of time to monitor user behaviour. Therefore, we investigated the possibility of using the Conditional Generative Adversarial Network (CGAN) to generate synthetic data for training the profiling model. The application of GAN would facilitate a more expedient implementation of this model on an online store website. The findings of this study may also prove beneficial to webshop owners and managers, enabling them to gain a deeper insight into their customers and align their price offers or discounts with the profile of a particular user.

Keywords: price sensitivity, profiling users, synthetic data, conditional GAN

1 Introduction

The development of online technologies and the global pandemic of COVID-19 have led to a surge in the popularity of online stores as a sales channel. Retailers are increasingly turning to sophisticated technologies to monitor their customers' behaviour in order to better understand their shopping needs and preferences. One method that has gained traction in this context is user profiling.

User profiling is the process of creating a virtual representation of a user based on their behaviour and preferences. Profiling can be conducted at various levels, including customer satisfaction, customer loyalty, demographic or contextual data. Online store owners may be particularly interested in profiling users in terms of price sensitivity [1]. Determining the likelihood of an online shopper making a purchase opens up the possibility of giving them an adequate price reduction. While webshops offer discounts, it is important to note that the extent of these discounts or promotions may vary for different customers. In this context, the store owner can devise a strategy that is tailored to their specific needs. One approach is to assume that users who frequently make purchases may anticipate a larger discount. Another would be to incentivise hesitant customers to make a purchase by increasing the discount amount offered to them.

Pricing decisions are not an easy task. Consumers emphasise that price has the greatest impact on their buying choices. Typically, they make rational decisions being mindful of their constrained income and budget. The seller, on the other hand, will be able to stay in business if and only if they earn a profit, which is entirely price-driven. When a seller sets the price of a product, they must consider its quality, the availability of similar products on the market, and different types of the product. If the price is not set adequately, consumers will demonstrate their negative sensitivity to the product by either not buying it at all or buying the product in small volumes. Consumer pricing sensitivity is the most crucial aspect that sellers should prioritise [1]. By understanding the extent to which customers are sensitive to price, sellers can adjust their pricing strategies to influence customer decisions. This can be achieved by offering discounts and promotions at an appropriate level, which may encourage customers to make purchases.

This paper presents the findings of a study investigating the methodology for profiling online shoppers in terms of their price sensitivity. The customers are categorised into one of three groups based on their likelihood of making a subsequent purchase: low, medium, or high. The profile is selected through an analysis of the user's behaviour during visits to an online store. Information is collected on the number of visits made by the user, the duration of their visit, the time since the last visit, the type of device used by the customer, the number of clicks on the page, the time when the user switches the tab in the web browser, the way the page was scrolled and whether a purchase was made. Anonymous users are identified by the unique identifier stored in a cookie. The data thus gathered were employed in the construction of a model designed to predict the probability of a purchase and to profile the user. Furthermore, the efficacy of this methodology was validated during subsequent visits to the online store.

The development of a profiling model requires it to be properly trained and tested. It is necessary to have a sufficient quantity of data in order to carry out this process. Algorithms trained with insufficient data are prone to failure when implemented in an environment that differs from that in which the data were obtained [15]. On the other hand, the presence of class imbalance can diminish the efficacy of trained models in classification problems. To address this limitation, the use of synthetic data is becoming increasingly prevalent [6].

The speed at which data are collected from an online store website is largely dependent on the website's popularity. For major clients, with a significant volume of visitors, data can be obtained rapidly. However, in the case of small shops, where the volume of visitors is low, implementing a new solution may take a considerable amount of time. To reduce the time taken to collect data from the store's page, we proposed the use of synthetic data. The encouraging results of our experiments can substantially accelerate the process of preparing, testing, and implementing the profiling model on the online store's webpage.

The research presented in this article focuses mainly on the use of deep neural network models of the Generative Adversarial Network (GAN) type. A properly functioning GAN allows for the generation of artificial data similar to real data, obtained when tracking user behaviour on a website. This enables an increase in representative training and testing data. In our experiments, we selected the Conditional GAN network (CGAN) [20], which can generate any distribution of tabular data with both categorical and numerical values. Our results demonstrate that a small sample of data obtained from website users is sufficient to be multiplied and used to train the model.

The paper is divided into several chapters. The next chapter reviews the literature addressing the topic under discussion. Subsequently, the models employed and the data utilised for training the profiling model are outlined. The following chapter discusses the profiling model and the data generation algorithm. Finally, the achieved results are summarised.

2 Literature Review

Literature offers numerous analyses of customer behaviour and the practices that online shops must adhere to in order to persuade customers to make a purchasing decision. The paper [8] investigates the relationship between attributes and consumer price sensitivity. Price sensitivity is influenced by the level of purchase commitment, package discounts and brand loyalty. The paper [5] outlines the key features of online shops that impact customer satisfaction when shopping online, such as the responsiveness of complaint resolution, product imagery, and diverse payment methods. In [17] it is highlighted that the satisfaction of online shoppers hinges on factors such as the convenience of product delivery, perceived security, information quality and product diversity. Overview paper [10] delves into crucial aspects of user profiling, covering types of profiles (static and dynamic), methods of user modelling (behavioural, interests, intentions), data collection from various sources, and technical considerations in user modelling. The paper [18] presents an overview of user profiling techniques, exploring the utilisation of demographic, behavioural, and contextual data alongside considerations of data privacy. Additionally, the authors outline machine learning algorithms and data anonymisation techniques employed for this purpose. In the paper [1], the literature is reviewed with the aim of elucidating how consumers exhibit price sensitivity in their product purchase decisions. In [21], the authors analyse a one-month dataset comprising two million users' 67 million purchase and browsing logs. Their objective is to gain insight into how users browse and shop for products, as well as how these behaviours vary. A solution that is closely aligned with our proposed approach is presented in [16], where the authors introduce a neural network model to predict purchases during active user sessions on an online shopping platform. The training and testing data were sourced from server logs recording HTTP requests.

The utilisation of large, diverse, and representative datasets is of vital importance for the conduct of scientific research in a variety of fields and practical applications [6]. This is particularly crucial in the medical domain, where disease data or patient images are not readily available. In instances where access to authentic data is severely limited, synthetic data are employed to compensate. According to [23], the rationale behind the utilisation of synthetic data may be to enhance the performance of machine learning models. Models trained on such data tend to achieve higher accuracy in real-world tasks. This improvement is attributed to the synthetic data which, in the case of video footage, have no background and contain only the movement of objects or individuals.

There are many overview papers discussing the various options available for generating artificial data. The article [22] provides a comprehensive review of scientific work related to synthetic data generation, covering various aspects such as applications (e.g., computer vision, speech, natural language, healthcare, and business), machine learning techniques (especially neural network architectures and deep generative models), and considerations regarding privacy and fairness. In [9], the authors discuss various methods for generating synthetic data and evaluating the technique employed to check their quality. They also examine GAN applications in data synthesis in a variety of fields. Another overview paper [4] describes the application of GAN models to synthesise tabular data based on a specific example. Various GAN architectures are discussed, such as VanillaGAN, WGAN, WGAN-GP, CTGAN, CopulaGAN and TableGAN. The models are evaluated using quantitative and qualitative methods. In the literature, there are documented practical applications of the CTGAN model for generating synthetic data [20]. For instance, [2] presents a method leveraging machine learning to generate synthetic data, which aims to safeguard IoT-based smart environments.

Synthetic data appear to serve as a valuable supplement when an ample training dataset is unavailable. Correctly generated, the data should not impede the model training process but rather expedite the acquisition of sufficient data.

3 Materials and Methods

3.1 Description of the Data

The data utilised in the discussed experiments were derived from the monitoring of several Polish online shop websites and they represent user behaviour on the webpage. The monitoring was conducted using a JavaScript script that tracks events occurring on the page. The following is a list of parameters of the data collected during a user's visit to one website:

- `ip_info_id` – IP address identifier,
- `first_req_serv_time` – date and time of accessing the page,
- `first_req_st2_close_seconds` – time spent on the page (monitored for a maximum of 40 seconds),
- `mouse_clicks` – number of clicks on the page,
- `scroll_dist_diff` – number of pixels scrolled on the page,
- `scroll_num` – number of page scroll events,
- `card_changed` – whether a browser tab change occurred,
- `is_mobile` – type of device (mobile/desktop).
- `target_future_sale` – indication of a sale, obtained from a sales pixel placed on the sale confirmation page. It serves to predict the probability for two classes: "will buy" (1), "will not buy" (0).

The data are aggregated by creating additional columns containing information on the number of: visits by a single user on a given day, visits per week and visits per month. The data regarding completed purchases are the least frequent, as there are far more people browsing products than those actually making a purchase. This leads to a significant imbalance in the `target_future_sale` classes, thus affecting the model training results, which noticeably deteriorates the classification outcomes.

The data undergo preprocessing, during which any empty values are substituted with zeros, any duplicate records are eliminated, and any continuous values are standardised. Any categorical data are converted into zero-one format as a one-hot vector.

3.2 The Conditional Tabular Generative Adversarial Network

The Conditional Tabular Generative Adversarial Network (CTGAN) model [20] is employed to generate synthetic tabular data. Its design allows it to cope with the heterogeneity of discrete and continuous data and their non-uniform distribution. The network architecture employs the classical generative adversarial network GAN, which consists of a generator and a discriminator. The generator produces synthetic tabulated data, while the discriminator evaluates whether the data are real or synthetic. For continuous data, mode-specific normalisation is used, based on the Gaussian Mixture Model (VGM), with each column being normalised independently. In the case of categorical variables, training-by-sampling, conditional vector, and generator loss are implemented to address imbalance problems. In the training-by-sampling method, the discriminator estimates the output of the conditional generator, which assesses the distance between the learned conditional distribution and the conditional distribution on real data. The use of the vector enables control over the attributes of the output data. During training, CTGAN learns to map a variety of conditional vectors to the corresponding data distributions. The generator is penalised for differences between the conditional vector and the resulting discrete values. The discrete data are represented by one-hot vectors.

4 Experiment Protocol

4.1 Profiler Model

The monitoring of customer behaviour on a website enables the determination of the likelihood that a purchase will be made. A model based on an artificial neural network was developed, which assigns users to one of three groups based on their price sensitivity: low, medium, and high. The resulting profiles were validated during subsequent visits to the shop.

The model that was selected to determine the price sensitivity profiles of the users is a deep learning neural network. The network architecture comprises two dense hidden layers with an activation function 'relu', which consist of 50 and 20 neurons respectively, normalisation layers, a linear output layer and an embedding layer at the input of the network, which is used to convert the categorical data into a dense vector. The model is outlined as follows:

```
Model(
 (embeds): ModuleList(
  (0): Embedding(87255, 600)
 )
 (bn_cont): BatchNorm1d(9, eps=1e-05,
  momentum=0.1, affine=True,
  track_running_stats=True)
 (layers): Sequential(
  (0): LinBnDrop(
   (0): Linear(in_features=609, out_features=50,
    bias=False)
   (1): ReLU(inplace=True)
   (2): BatchNorm1d(50, eps=1e-05,
    momentum=0.1, affine=True,
    track_running_stats=True)
  )
  (1): LinBnDrop(
   (0): Linear(in_features=50, out_features=20,
    bias=False)
   (1): ReLU(inplace=True)
   (2): BatchNorm1d(20, eps=1e-05,
    momentum=0.1, affine=True,
    track_running_stats=True)
  )
  (2): LinBnDrop(
   (0): Linear(in_features=20, out_features=1,
    bias=True)
  )
 )
)
```

To evaluate the quality of the model, the F1-score metric with a 'macro' average was used, following the formula:

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} F1_i \tag{1}$$

where: $F1_i$ is the $F1$ score for the i-th class, N is the number of classes. In the $F1$-score with 'macro' average, each class contributes equally to the final result. This is particularly useful in the discussed case, where the datasets are imbalanced. Such averaging highlights how well the model performs in classifying each class independently without favouring larger classes. The $F1$ score for a single class is calculated as follows:

$$F1 = 2\frac{precision \times recall}{precision + recall} \tag{2}$$

where precision is the ratio of true positive predictions to all positive predictions, and recall is the ratio of true positive predictions to all cases that are in fact positive. The data collected in the first month were used to calculate aggregates of user visits over a week and a month.

The data, as described in section 3.1, were collected over a three-month period during which several online store websites were monitored. The data from the second month contain a total of 437,566 samples. Classes 0 (customer made no purchase) and 1 (purchase was made) contain 146,656 and 17,640 records, respectively. These were divided in an 80/20 ratio into training and testing sets. The data from the third month will be used for testing. They contain 41,557 records, with 40,080 samples in class 0 and 1,477 in class 1.

The results of training the discussed profiling model and a comparison to the results obtained from the random forest classifier [14] and XGBboost classifier [7] algorithms are presented in Table 1. The following columns contain information about the type of metric and the results for each model. As observed, the proposed model achieved the best results in both the training and testing processes. Figures 1 and 2 illustrate the ROC curves for both the training and validation data.

Table 1. Prediction results obtained by the individual models.

	Metric	Proposed model	Random forest classifier	XGBoost classifier
Validation	precision	0.9949	0.9577	0.7544
	recall	0.9949	0.8608	0.5475
	F1 score	0.9949	0.019	0.5613
Test	precision	0.6305	0.6272	0.6234
	recall	0.6203	0.5341	0.5359
	F1 score	0.6252	0.5499	0.5515

Table 2. Results of the effectiveness of the user profiling model for the test data.

Profile	Nr of customers	% of customers	Customers who made a purchase	% Customers who made a purchase
Low	37760	90.86	1002	2.65
Medium	3694	8.89	445	12.04
High	103	0.25	30	29.12

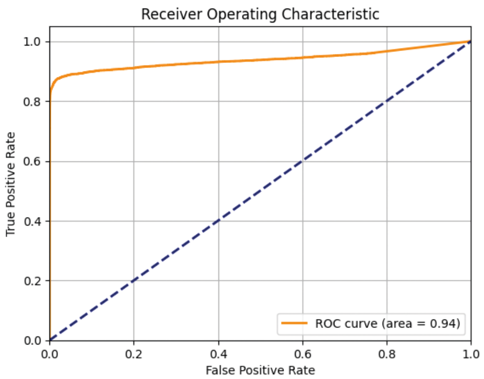


Figure 1. ROC curve obtained for the training data.

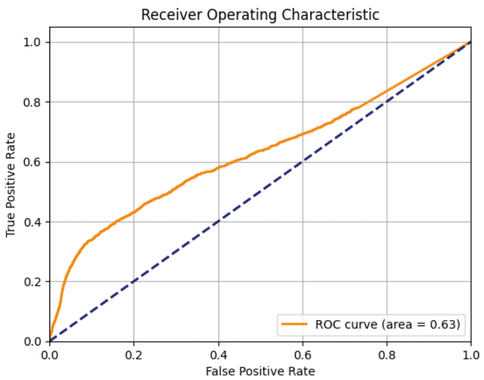


Figure 2. ROC curve obtained for the test data.

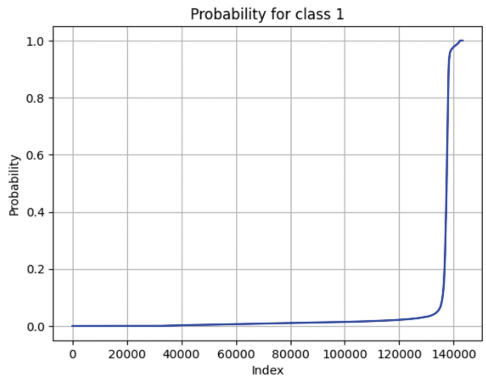


Figure 3. The probability obtained for the training data.

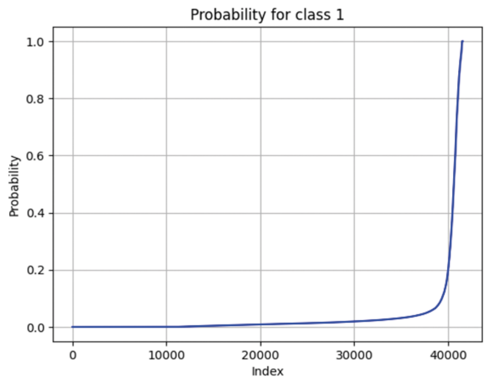


Figure 4. The probability obtained for the test data.

The subsequent study identified the values of the two thresholds that separate the probability in order to determine the appropriate profile. Figures 3 and 4 illustrate the changes in probability values obtained for the validation and test data. The horizontal axis represents the numbers of consecutive records from the dataset for which the prediction was made. The vertical axis shows the probability value determining whether the customer will buy again (class "1"). The data were sorted by probability value. Upon analysis of both plots, it was determined that the optimal thresholds were 0.05 and 0.97. Consequently, three profiles were defined:

– low (probability $<= 0.05$),

– medium ($0.05 <$ probability < 0.97),

– high (probability $>= 0.97$).

The efficacy of the model for the test data is demonstrated in Table 2. The columns illustrate, in turn, the number of users subjected to testing who were allocated to the respective group, the percentage of the total, the number of users who subsequently made a purchase and the percentage of these users from the respective group. As can be observed, the assigned profile aligns closely with those who made a subsequent purchase.

4.2 Synthetic Data

As previously stated, a sufficient amount of data is required to train a profiling model. However, data collection is a time-consuming process, particularly in online stores that are rarely visited. Therefore, we have proposed the generation of synthetic data from a small sample of real data. The combination of both records will serve to train the profiling model.

The algorithm for generating synthetic data is as follows:

1. Sample the real data while monitoring the website. The dataset should include information about the events occurring on the website and the date of sampling.

2. Generate synthetic samples using the GAN model and include the date.

3. Evaluate the effectiveness of the generated data using KSComplement and TVComplement metrics. If poor results are obtained, it is necessary to gather a larger amount of diverse data and to repeat the data generation process.

4. Combine the synthetic data with the real data. Use the date to calculate aggregates containing the number of user visits per month, week, and day.

The samples were generated using the CTGAN model, which is discussed in Section 3.2. The tools provided in the SDMetrics library [8] were used to assess the quality of the generated content, focusing on the degree of accuracy with which the synthetic data reproduced the properties of the real data. This process is referred to as synthetic data fidelity. Two metrics were employed *KSComplement* and *TVComplement*. The *KSComplement* metric computes the similarity between the real and synthetic columns in terms of marginal distribution. It is intended for continuous data. In this metric, a value of 1 indicates that the real data is identical to the synthetic data, while a value of 0 indicates that the real and synthetic data are as different as possible.

In contrast, *TVComplement* is a metric that calculates the similarity of a real column to a synthetic column for discrete (categorical) data and for logical data. Once again, a value of 1 indicates that the actual data is identical to the synthetic data, while a score of 0 indicates a complete difference.

The first three steps of the algorithm aim to conduct the training process of the CTGAN model. A total of 1,000 real samples were utilised for this purpose. The efficacy of the generated data was evaluated by calculating the similarity between the real and synthetic data for each parameter. The results of the sample generation for the profiling model are presented in Table 3. The columns display, respectively, the parameter name, the metric employed and the score value, which indicates the degree of similarity between the generated data and the real data. A review of the individual score values reveals that the similarity values oscillate around 80-90%, thereby confirming a high degree of similarity between the datasets.

Figures 5 to 9 illustrate the distribution of values for the sample parameters generated by the CTGAN model in comparison to the distribution of the real data. It can be observed that there is a substantial similarity in the distribution between the real and synthetic values, which indicates the effectiveness of this data generation method.

Table 3. The metric values assessing the quality of the generated synthetic data.

Parameter	Metric	Score
ip_info_id	TVComplement	0.908794
first_req_serv_time	KSComplement	0.868164
first_req_st2_close_second	KSComplement	0.939661
mouse_clicks	KSComplement	0.879492
scroll_dist_diff	KSComplement	0.883734
scroll_num	KSComplement	0.888328
card_changed	TVComplement	0.801155
is_mobile	TVComplement	0.853424
target_future_score	TVComplement	0.928200

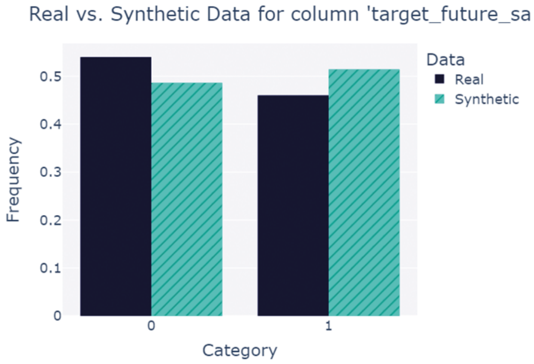


Figure 5. Comparison of the distribution of real and generated values for the example column `target_future_sales`.

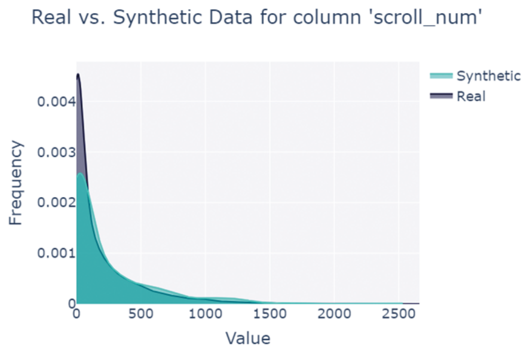


Figure 7. Comparison of the distribution of real and generated values for the example column `scroll_num`.

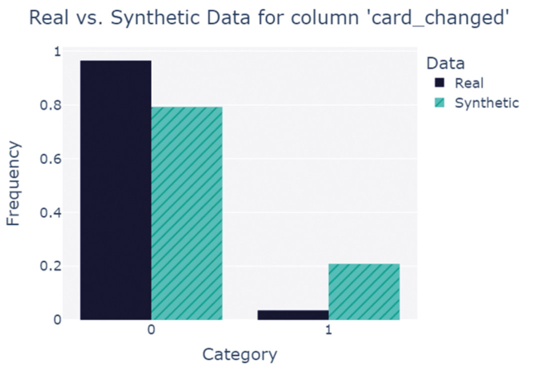


Figure 6. Comparison of the distribution of real and generated values for the example column `card_changed`.

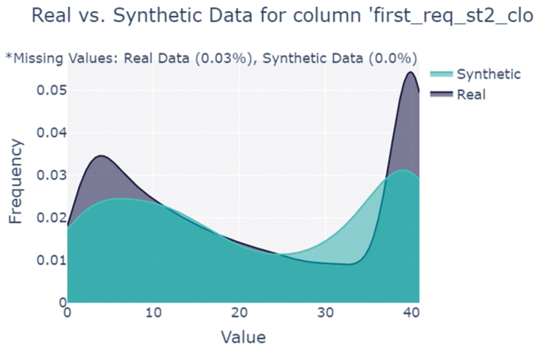


Figure 8. Comparison of the distribution of real and generated values for the example column `first_req_st2_close_seconds`.

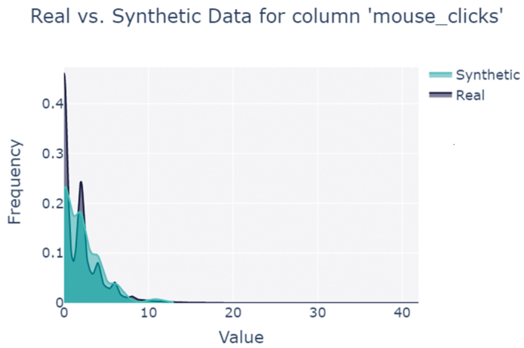


Figure 9. Comparison of the distribution of real and generated values for the example column mouse_clicks .

The generated data will be used to train the profiling model described in section 4.1. The experiment aims to simulate the model's readiness for implementation in an online shop after just one thousand user visits. The real training data were expanded with synthetic data. The results of the training and the effectiveness of the model in classifying the test data are presented in Table 4. For comparison, the results obtained by a model trained solely with real samples are also provided. The first column details the type of metrics for each stage of training and testing, while the results achieved by the models are shown in the next columns. Upon analysis of the metric values, it can be concluded that the data generated by the CTGAN network can effectively augment real data. The classification of 1,000 test samples performed by the model trained with augmented data is superior to the classification performed by the model trained with real data only. This experiment demonstrates the feasibility of implementing a profiling model in an online shop after just a few customer visits.

5 Summary

This paper presents the possibility of profiling e-commerce users in terms of price sensitivity. The proposed model allows for the determination of the likelihood of a subsequent purchase based on information about the customer's behaviour on the website. The results of the study confirm the effectiveness of the profiling model in predicting the purchase process. Moreover, the model can provide online shop owners, entrepreneurs, and marketers with valuable insights into e-commerce customer behaviour, which may then lead to better adjustment of strategies to meet customer needs and expectations. Further research on the model could involve the utilisation of fast neural networks for rapid model response [3] and the possibility of using a fingerprint device for better user identification [12]. Additionally, tracking mouse pointer movements on the website and subsequent analysis in the form of heatmaps [19] may also be explored.

The efficacy of using synthetic data to train a profiling model has been demonstrated. This approach reduces the time needed to acquire the necessary data, thereby enabling the rapid deployment of the profiling model to an online store. In the future, further work will be conducted on the system for generating synthetic samples. This will involve the use of autoencoder models [11], [13], which will reduce the dimensions of the data and combine information about multiple user visits into a single input vector. This will enhance the performance of the GAN model by removing categorical data.

Acknowledgments

The presented results are obtained within the realization of the project "Sales Bot 2.0 – development of an innovative sales system for e-commerce based on individual user price profiling with a dynamic product recommendation system based on machine learning and device fingerprint" financed by the National Centre for Research and Development; grant number POIR.01.01.01-00-0241/19-00.

References

[1] Abdullah-Al-Mamun, M. K. R., & Robel, S. D., A critical review of consumers' sensitivity to price: Managerial and theoretical issues, Journal of International Business and Economics, 2(2), 01-09, 2014.

[2] Alabdulwahab, S., Kim, Y. T., Seo, A., & Son, Y., Generating Synthetic Dataset for ML-Based IDS Using CTGAN and Feature Selection to Protect Smart IoT Environments, Applied Sciences, 13(19), 10951, 2023.

[3] Bilski, J., Kowalczyk, B., Kisiel-Dorohinicki, M., Siwocha, A., & Żurada, J., Towards a very fast feedforward multilayer neural networks training al-

Table 4. Prediction results obtained by the models trained with synthetic and real data and the model trained with real data.

	Metric	Model with synthetic + real data	Model with real data
Validation	precision	0.5624	0.4749
	recall	0.5046	0.4974
	F1 score	0.4985	0.4857
Test	precision	0.6243	0.6289
	recall	0.7434	0.5043
	F1 score	0.6522	0.4901

gorithm, Journal of Artificial Intelligence and Soft Computing Research, 12(3), 181-195, 2022.

[4] Bourou, S., El Saer, A., Velivassaki, T. H., Voulkidis, A., & Zahariadis, T., A Review of Tabular Data Synthesis Using GANs on an IDS Dataset, Information 2021, 12, 375, 2021.

[5] Bucko, J., Kakalejčík, L., & Ferencová, M., Online shopping: Factors that affect consumer purchasing behaviour, Cogent Business & Management, 5(1), 1535751, 2018.

[6] Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F., Synthetic data in machine learning for medicine and healthcare, Nature Biomedical Engineering, 5(6), 493-497, 2021.

[7] Chen, T., & Guestrin, C., Xgboost: A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794), 2021.

[8] Dominique-Ferreira, S., Vasconcelos, H., & Proença, J. F., Determinants of customer price sensitivity: an empirical analysis, Journal of Services Marketing, 30(3), 327-340, 2016.

[9] Figueira, A., & Vaz, B., Survey on synthetic data generation, evaluation methods and GANs, Mathematics, 10(15), 2733, 2022.

[10] Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F., A survey of user profiling: State-of-the-art, challenges, and solutions, IEEE Access, 7, 144907-144924, 2019.

[11] Grycuk, R., Scherer, R., Marchlewska, A., & Napoli, C., Semantic hashing for fast solar magnetogram retrieval, Journal of Artificial Intelligence and Soft Computing Research, 12(4), 299-306, 2022.

[12] Gabryel, M., Grzanek, K., & Hayashi, Y., Browser fingerprint coding methods increasing the effectiveness of user identification in the web traffic, Journal of Artificial Intelligence and Soft Computing Research, 10(4), 243-253, 2020.

[13] Gabryel, M., Lada, D., Filutowicz, Z., Patora-Wysocka, Z., Kisiel-Dorohinicki, M.,& Chen, G. Y., Detecting anomalies in advertising web traffic with the use of the variational autoencoder, Journal of Artificial Intelligence and Soft Computing Research, 12(4), 255-256, 2022.

[14] Ho, T. K., Random decision forests, In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE, 1995.

[15] Pencina, M. J., Goldstein, B. A., & D'Agostino, R. B., Prediction models—development, evaluation, and clinical application, New England Journal of Medicine, 382(17), 1583-1586, 2020.

[16] Suchacka, G., & Stemplewski, S., Application of neural network to predict purchases in online store, In Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part IV (pp. 221-231). Springer International Publishing, 2017.

[17] Mofokeng, T. E., The impact of online shopping attributes on customer satisfaction and loyalty: Moderating effects of e-commerce experience, Cogent Business & Management, 8(1), 1968206, 2021.

[18] Vakulenko, Y., Shams, P., Hellström, D., & Hjort, K., Online retail experience and customer satisfaction: the mediating role of last mile delivery, The International Review of Retail, Distribution and Consumer Research, 29(3), 306-320, 2019.

[19] Woldan, P., Duda, P., Cader, A., & Laktionov, I., A new approach to image-based recommender systems with the application of heatmaps maps, Journal of Artificial Intelligence and Soft Computing Research, 13(2), 63-72, 2023.

[20] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K., Modeling tabular data using conditional GAN, Advances in neural information processing systems, 32, 2019.

[21] Yan, H., Wang, Z., Lin, T. H., Li, Y., & Jin, D., Profiling users by online shopping behaviors, Mul-

timedia Tools and Applications, 77, 21935-21945, 2018.

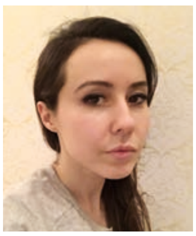
[22] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., Machine learning for synthetic data generation: a review, arXiv preprint arXiv:2302.04062, 2023.

[23] Kim, Y. W., Mishra, S., Jin, S., Panda, R., Kuehne, H., Karlinsky, L., ... & Feris, R., How transferable are video representations based on synthetic data?, Advances in Neural Information Processing Systems, 35, 35710-35723, 2022.

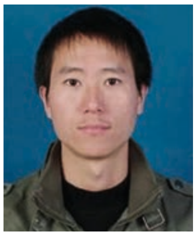
Marcin Gabryel earned his Ph.D degree in computer science at Czestochowa University of Technology, Poland, in 2007. He is an assistant professor in the Department of Computer Engineering at Czestochowa University of Technology. His research focuses on developing new methods in computational intelligence and data mining. He has published over 60 research papers. His present research interests include deep learning architectures and their applications in databases and security.
https://orcid.org/0000-0002-6701-0460

Eliza Kocić graduated with a BEng in Biomedical Engineering, specializing in Medical Informatics, from Gdańsk University of Technology. She furthered her studies at the same institution, earning an MSc in Biomedical Engineering with a focus on Artificial Intelligence and Machine Learning.
Since 2019 she has been working as a software engineer at Spark Digitup, dedicating her efforts to maintaining and enhancing systems designed to prevent fraudulent marketing practices and improving chatbots utilized in ecommerce. She is proficient in machine learning, specializing in training models and deploying them effectively for various applications. Her expertise includes clustering website visitors into profiles, as well as recognizing and collectively assessing patterns in website traffic.
https://orcid.org/0000-0001-7208-5099

Milan Kocić received his BEng Computer Science degree from Gdańsk University of Technology. He has been working as software engineer at Spark Digitup since 2016, mainly focusing on maintenance and development of marketing fraud protection systems and ecommerce chatbots. His areas of work include device fingerprinting, anomaly detection and integration with AI solutions.
https://orcid.org/0000-0002-6257-0936



Zofia Patora-Wysocka is a professor at the University of Social Science in Łódź, Poland. She received the Ph.D. degree from the Częstochowa University of Technology, Częstochowa, Poland in 2008, and the D.Sc. degree in economic sciences from the WSB University in Dąbrowa Górnicza in 2020. Her research interest includes change management, routine dynamics and strategy, practice theory, science and technology studies, and applications of data mining and artificial intelligence methods in management.
https://orcid.org/0000-0002-0429-0207



Min Xiao (Member, IEEE) received the B.S. degree in mathematics and the M.S. degree in fundamental mathematics from Nanjing Normal University, Nanjing, China, in 1998 and 2001, respectively, and the Ph.D. degree in applied mathematics from Southeast University, Nanjing, in 2007.,He was a Postdoctoral Researcher or a Visiting Researcher with Southeast University; The City University of Hong Kong, Hong Kong; and Western Sydney University (Sydney Campus), Sydney, NSW, Australia. He is currently a Professor with the College of Automation and the College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing. His current research interests include information security, anomalous diffusion systems, networked control systems, tipping and control, and cyber–physical systems.
https://orcid.org/0000-0002-8992-153X

Miroslaw Pawlak received the Ph.D. and D.Sc. degrees in computer engineering from Wrocław University of Technology, Wrocław, Poland. He is currently a Professor at the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada. He has held a number of visiting positions in North American, Australian, and European Universities. He was at the University of Ulm, University in Goettingen and Marburg University as an Alexander von Humboldt Foundation Fellow. His research interests include statistical signal processing, machine learning, and nonparametric modeling. Among his publications in these areas are the books Image Analysis by Moments (Wrocław Univ. Technol. Press, 2006), and Nonparametric System Identification (Cambridge Univ. Press, 2008), coauthored with Prof. Włodzimierz Greblicki. Dr. Pawlak has been an Associate Editor of the Journal of Pattern Recognition and Applications, Pattern Recognition, International Journal on Sampling Theory in Signal and Image Processing, Journal of Artificial Intelligence and Soft Computing Research, Opuscula Mathematica and Statistics in Transition-New Series.
https://orcid.org/0000-0003-2627-108X