Decentralized Segmentation and Prediction of E-Commerce Efficiency Using Machine Learning Methods

Oleh Havryliuk[1,†], Ihor Ponomarenko[2,*,†], Maryna Petchenko[3,†] and Oleksandr Yakushev[4,5,†]

[1] European University, 16B Akademika Vernads'koho blvd., 03115 Kyiv, Ukraine
[2] State University of Trade and Economics, 19 Kyoto str., 02156 Kyiv, Ukraine
[3] State University of Information and Communication Technologies, 7 Solomyanska str., 03110 Kyiv, Ukraine
[4] Cherkasy State Technological University, 460 Shevchenko blvd., 18006 Cherkasy, Ukraine
[5] Kherson National Technical University, 11 Instytutska str., 29016 Khmelnytskyi, Ukraine

Abstract

The feasibility of using various machine learning algorithms for processing heterogeneous data (structured, semi-structured, and unstructured) is substantiated, and the results obtained can be used to form effective management decisions in the field of marketing. The need to ensure information security is obvious due to the significant risks of unauthorized access to data by third parties for illegal purposes. The need to use decentralization technologies to secure confidential information based on high-performance machine learning algorithms is emphasized. The feasibility of ensuring decentralization by processing an aggregated data set and building a global model with subsequent implementation at the level of an individual company is argued, which minimizes the possibility of loss of commercial data. The study was conducted based on data on the activities of 81 online stores in the consumer electronics market in the Kyiv region for January-March 2025 using the specialized web resource Similarweb. The implementation of machine learning algorithms was based on 11 metrics. The importance of cluster analysis and various regression models for data processing and ensuring the efficiency of their integration into decentralized models is proven. The selection of optimal algorithms was based on special metrics and visualization methods. The effectiveness of using the obtained models for effective decentralized implementation at the level of individual online stores is highlighted.

Keywords

consumer electronics, digitization, decentralization, cluster analysis, machine learning, marketing, online stores.

1. Introduction

Various approaches are used for data processing, among which artificial intelligence has become particularly widespread. The presented field of knowledge encompasses machine learning, expert systems, natural language processing, evolutionary computing, and genetic algorithms, among others. In the fields of marketing and e-commerce, machine learning is widely used due to the presence of a large number of algorithms that are selected based on the specifics of the data (numerical expressions, text, photos, videos, audio) and the requirements for the quality of the results obtained. High-performance mathematical algorithms enable companies to process large datasets quickly, enhance the quality of models using self-learning principles, and deliver results that optimize business operations in the digital environment. The digital era is characterized by the possibility of collecting large amounts of private information that companies need in the process of forming personalized communications with consumers. However, at the same time, ethical and legal issues arise regarding the collection of personalized information, which may be perceived negatively by a large number of users and lead to an increase in the risks of illegal data appropriation by a third party for criminal activities. The active development of data leakage minimization technologies based

[1] SMICS'25: Workshop on Cryptology and Data Security, October 16-18, 2025, Lviv, Ukraine
* Corresponding author.
† These authors contributed equally.
o.havryliuk@e-u.edu.ua (O. Havryliuk); i.ponomarenko@knute.edu.ua (I. Ponomarenko); marinapetcenko@gmail.com (M. Petchenko); o.yakushev@chdtu.edu.ua (O. Yakushev)
0000-0001-6819-9296 (O. Havryliuk); 0000-0003-3532-8332 (I. Ponomarenko); 0000-0003-1104-571 (M. Petchenko); 0000-0002-0699-1795 (O. Yakushev)

on cybersecurity technologies and the implementation of regulatory acts (for example, the General Data Protection Regulation (EU) and the Personal Information Protection and Electronic Documents Act (Canada) stimulate consideration of the principles of decentralization [2, 3]. For Web3 ecosystems, decentralization has become the foundation, implemented through technologies such as edge computing and federated learning, utilizing mobile devices and the Internet of Things. The current level of development in digital technologies enables the successful integration of machine learning algorithms into decentralized systems, ensuring a high level of quality in big data processing [4].

2. The Aim

The research tasks include:

– outlining the possibilities of implementing machine learning algorithms while ensuring the principle of decentralization;

– conducting a cluster analysis on a set of online stores based on the optimization of mathematical approaches with the assessment of relationships using regression;

– substantiating the possibilities of minimizing the risks of data loss constituting a trade secret, as well as the consequences of cyberattacks, by adhering to the principle of decentralization.

3. Models and Methods

The active development of e-commerce stimulates the formation of modern information systems that enable the accumulation of relevant information in large sets. Data processing necessitates the development of effective analytical tools that lay the groundwork for implementing successful marketing strategies in the digital environment. The presence of a large number of modern machine learning algorithms allows building high-performance models in accordance with the specifics of available information and the strategic goals of each company. Ensuring decentralization based on machine learning algorithms can be achieved thanks to web analytics tools and other approaches to collecting data on the Internet. This study used information on the activities of 81 online stores in the consumer electronics market in the Kyiv region for January-March 2025 [5]; the data was collected using the Similarweb resource, and their set includes the following metrics: Name of store; Total visits; Monthly traffic; Mobile web share; Country rank; Visit duration; Pages per visit; Bounce rate; Organic traffic; Male share; Average age.

The study selected online stores in the consumer electronics market, since modern generations Y, Z, and Alpha - the largest buyers - regularly purchase new gadgets. The high level of digitalization has led to the development of e-commerce and the rapid dynamics of purchasing necessary innovative products online. The choice of the Kyiv region as the study's object was determined by the concentration of many online stores within the capital of Ukraine, as well as the presence of a large solvent population.

4. Experiment

Important practical areas studied based on a specific set of observations include classification carried out according to the existing system of metrics. After testing various algorithms, it was found that the best grouping results for this set of online stores can be achieved using hierarchical cluster analysis. To assess the quality of the latter, it is considered advisable to use a silhouette diagram (Fig. 1). The average silhouette value of 0.202 indicates a moderate quality of cluster formation. The best quality is inherent in Cluster 3, which demonstrates the highest silhouette value (approximately 0.350), indicating the homogeneity and clarity of the represented group. In general, this approach can also be applied to the process of updating data and increasing the sample size.

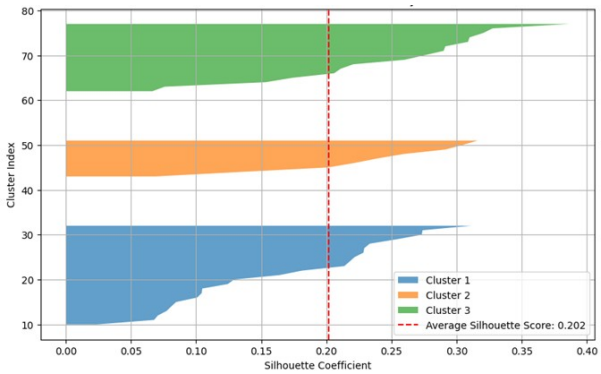

Figure 1: Silhouette plot for cluster quality. Source: calculated by the authors based on [5]

A comprehensive analysis of metrics in each selected cluster allows companies to identify the characteristics of online stores included in each group (Table 1). Cluster 1 includes small online stores that are not very popular among users. The presented group of companies has low visibility in search engines, which negatively affects user engagement on the Internet. To view the offers given by online stores, most visitors use mobile devices, and the level of retention on web resources remains relatively low compared to average market values.

Cluster 2 includes medium-sized online stores that are quite popular among modern users, although they do not belong to the group of leaders in the consumer electronics market in the Kyiv region at the beginning of 2025. Cluster 3 includes large online stores in the Kyiv region, which are characterized by significant popularity among the target audience (for example: FOXTROT, MOYO, CITRUS).

Table 1

Average values of online store metrics in January-March 2025 by clusters

Metrics	Cluster 1	Cluster 2	Cluster 3
Total visits	65776.87	241310.33	710302.25
Monthly traffic	21925.61	80436.89	236785.25
Mobile web share, %	68.11	58.91	68.97
Country rank	22118.17	9790.67	2695.94
Visit duration, seconds	103.17	87.08	168.17
Pages per visit	2.69	3.14	3.00
Bounce rate	49.63	39.20	46.80
Organic traffic, %	29.90	25.39	35.82
Male share, %	60.04	64.20	60.31
Average age, years	41.14	38.23	40.07

Source: calculated by the authors based on [5]

The next stage involves assessing the relationships between the available metrics for the studied set of online stores. Monthly traffic was utilized as an effective feature, and the duplicate metric “Total visits” was removed from the dataset. The obtained results indicate that gradient boosting provides the best description of the relationships for the presented set of online stores (R^2 = 0.9488, RMSE = 246174.3529, MAE = 133019.6945).

Figure 2 shows a fragment of the gradient boosting model, which describes the influence of factors on the monthly traffic metric. A comprehensive analysis for one of the decision trees used in the gradient boosting model allows to identify the following features:

• The most influential factor in the model is the country rating. According to the individual values of the national rating for each online store in the consumer electronics market in the Kyiv region, web resources can be divided into those with a large and a small number of visits.

• The greatest activity in e-commerce is inherent in the female audience, due to interest in specialized online stores with innovative electronics. Additionally, men tend to be less interested in spending a significant amount of time familiarizing themselves with gadget offers on websites.

• The given fragment of gradient boost trees demonstrates that the tree-like structure represents complex nonlinear relationships that cannot be identified based on the usual linear model.

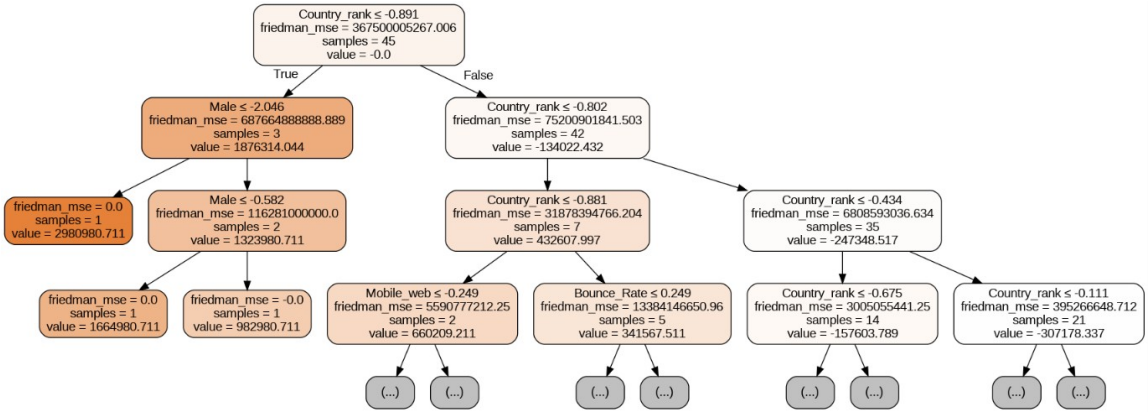

Figure 2: An example tree from the gradient boosting forest of online stores in the Kyiv region, from January to March, 2025. Source: calculated by the authors based on [5]

To assess the quality of predictive models, it is considered appropriate to use different approaches, in particular by constructing a cumulative growth diagram (Figure 3) [6]. Thanks to the latter, it is possible to assess the level of the model's effectiveness for identifying online stores with the highest potential and establish the accuracy of its predictive properties. These characteristics significantly enrich the presented approach by validating data for marketing purposes.

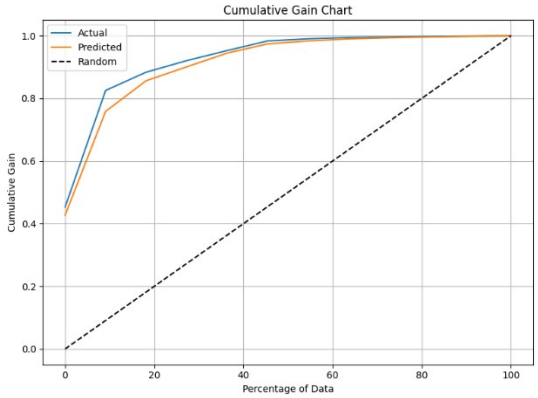

Figure 3: Cumulative gains chart comparing actual vs. predicted ranking performance against a random baseline. Source: calculated by the authors based on [5]

The construction of the cumulative increment diagram made it possible to draw the following conclusions:

• A significant excess of the actual and predicted values over the random level indicates a high predictive power, and therefore the undeniable benefit of gradient boosting.

• Minor deviations between the actual and predicted values of the resulting feature also characterize the high predictive power of gradient boosting.

• The specified model is advisable to use for a specific enterprise while adhering to the principles of decentralization, and the results obtained can be applied in the process of determining the most attractive segments in the field of e-commerce.

5. Further Research

The results obtained indicate the undeniable potential of using machine learning algorithms for the effective processing of big data in the field of e-commerce, as well as creating conditions for the protection of both personal and commercial data. The evolution of digital technologies, cloud

computing, and specialized algorithms has led to the emergence and application of new approaches in the field of marketing, enabling the identification of hidden connections in information arrays. In the future, this involves the accumulation of relevant information in the field of e-commerce from various web resources, primarily from social media. Conducting scientific research and making effective management decisions in the field of marketing should be based on high-performance machine learning algorithms and a diverse range of content. The combination of different information types will contribute to achieving more accurate results, which is especially relevant in conditions of intensive development in the digital environment.

6. Conclusions

To optimize business processes in a digital environment, it is advisable to utilize decentralized segmentation and forecasting of key e-commerce indicators. The implementation of machine learning algorithms for processing company data in compliance with the principles of decentralization holds significant prospects, as it enables the achievement of effective results while also securing personal and commercial data. The Internet is a valuable source of collecting various information, particularly on the functioning and key performance indicators of 81 online stores, which are used in the presented study. Securing data in a decentralized system of processing using machine learning methods requires implementation based on an aggregated dataset with direct implementation at the local level. Hierarchical cluster analysis and gradient boosting demonstrated the best results for the obtained data and can be implemented by a specific online store without the need to display information to third parties.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to: translate certain text fragments into English, perform grammar and spelling checks, and paraphrase or reword content. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Sabbagh, Digital Economy and Communication Technologies: Methods and Mechanisms of Promotion through E-Commerce and E-Marketing, Indian Journal of Data Communication and Networking, vol. 1, no. 3, 2021, pp. 10–22. doi: 10.54105/ijdcn.B5003.061321

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. [Online]. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[3] The Personal Information Protection and Electronic Documents Act (PIPEDA), 2025. URL: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

[4] A. Brecko, E. Kajati, J. Koziorek, and I. Zolotova, Federated Learning for Edge Computing: A Survey, Applied Sciences, vol. 12, no. 18, article 9124, 2022. doi:10.3390/app12189124.

[5] Similarweb, 2025. URL: https://pro.similarweb.com/

[6] Cumulative charts, 2025. URL: https://docs.datarobot.com/en/docs/modeling/analyze-models/evaluate/roc-curve-tab/cumulative-charts.html#:~:text=Cumulative%20Gain%20represents%20the%20sensitivity,to%2020%25%20of%20total%20customers