Journal of Innovation & Knowledge 10 (2025) 100828

Contents lists available at ScienceDirect

Journal of Innovation & Knowledge

journal homepage: www.elsevier.com/locate/jik

Applying knowledge transfer in data augmentation to improve online advertising performance of entrepreneurs

Ruben Huertas-Garcia [a], Laura Sáez-Ortuño [a], Santiago Forgas-Coll [a,*], Javier Sánchez-García [b]

[a] Universitat de Barcelona, Business Department, Avda. Diagonal, 690, 08034 Barcelona, Spain
[b] Universitat Jaume I, Department of Business Administration and Marketing, Avda. Vicent Sos Baynat, s/n, 12071 Castelló de la Plana, Spain

ARTICLE INFO

JEL classification:
M31 marketing
M37 advertising
C45 neural networks and related topics
L81 retail and wholesale trade • e-commerce
C53 forecasting and prediction methods • simulation methods

Keywords:
Data augmentation
Machine learning algorithms
Targeting
Market segmentation
Online advertising

ABSTRACT

Artificial intelligence (AI) is transforming the way businesses operate, enabling entrepreneurs to achieve diagnoses that were once only possible for large companies. This transformation is evident in digital advertising, where AI not only enables advanced analytics, but also offers the possibility of developing creative designs at low cost. However, this technological progress contrasts with predictions of a slowdown in online advertising in the coming years. Thus, entrepreneurs must change their strategies to overcome the defensive positions of competitors. This study proposes the combination of AI analytical algorithms (XGBoost) with data augmentation algorithms (SMOTE) to improve targeting accuracy when launching online communication campaigns. Specifically, a case study illustrates how a lead-gathering company uses these algorithms to profile five market segments (hearing aids, NGOs, energy distributors, telecommunications and finance). Subsequently, a field experiment was conducted with one of the products, solar panels, to assess external validity. The results reveal that the combination of both algorithms improves internal validity for four of the five products, and the field experiment confirms the external validity of the energy product. Finally, recommendations on the use of these tools are proposed to entrepreneurs.

Introduction

Artificial intelligence (AI) is revolutionising entrepreneurship in markets. It is not only transforming existing management models, but also creating opportunities for start-ups to generate disruptive innovations (Zhang et al., 2025). New entrepreneurs can now leverage the computational and analytical power of machine learning algorithms to make diagnoses, and the potential of generative AI to create content for their products, services or communication is bringing their operational capabilities on par with those of large companies (Buck et al., 2023), ultimately democratising the market (Choi et al., 2022). However, this technological transformation requires new approaches to training, combining education in science, technology and management (Blankesteijn et al., 2024; Juárez-Varón & Monreal, 2025; Li et al., 2024).

One of the areas where AI tools are most successfully serving these purposes is digital advertising, an area that is proving to be highly resilient to external pressures. Despite the effects of the pandemic, it has been estimated that advertising expenditure on retail media rose from USD 50 billion to USD 140 billion between 2019 and 2024. Over the same period, digital advertising increased its market share has grown from 18 % to 21.8 % (eMarketer, 2024a). However, the outlook for the next four years is less optimistic, as growth is expected to slow between 2024 and 2028, especially in Europe and China, sending signals that the market is entering a maturity phase (eMarketer, 2024b). Consequently, entrepreneurs must adapt their strategies to a market in which competitors are likely to adopt defensive strategies to preserve their market share and competitive advantage, often through the use of pricing and communication policies (Iacobucci, 2016).

However, a defensive position is countered by fine-tuning the precision of offensive actions. One way to achieve this is by improving the market segmentation process. A properly targeted communication campaign avoids wasting resources on uninterested consumers, concentrating instead on those with a higher probability of interaction,

* Corresponding author.
E-mail addresses: rhuertas@ub.edu (R. Huertas-Garcia), laurasaez@ub.edu (L. Sáez-Ortuño), Santiago.forgas@ub.edu (S. Forgas-Coll), jsanchez@uji.es (J. Sánchez-García).

https://doi.org/10.1016/j.jik.2025.100828
Received 11 June 2025; Accepted 22 September 2025
Available online 5 October 2025

conversions and long-term profitability (Choi & Mela, 2019; Johnson, 2013). Leading platforms such as Google and Facebook have responded to this need and offer clients, entrepreneurs and/or advertisers, on-demand targeting services, where each advertiser can profile their target audience by cross-referencing demographic and behavioural variables (Google, 2023). However, such processes have been questioned both in terms of cost-effectiveness and the risk of declining ad impact revenues (Celis et al., 2014). Moreover, comparative studies have found that less targeted campaigns may actually outperform more targeted ones due to their greater reach (Ahmadi et al., 2024).

A further issue is that the available data is not always appropriate for a given type of commercial offer, meaning that data often has to be collected from potential consumers. This requires the design of complex questionnaires, recruitment of participants and analysis of responses, all of which can be costly and logistically complex (Hair Jr et al., 2019). This difficulty is exacerbated by the growing concern among consumers about privacy, which has led to the introduction of regulatory standards that increasingly complicate the collection of information via the internet (Sáez-Ortuño et al., 2024; Tucker, 2014).

One way for start-ups to address these challenges is the use of AI algorithms capable of rapidly generating large volumes of synthetic data, as this facilitates comprehensive, granular analysis without incurring privacy concerns or the high costs of data acquisition.

The generation of artificial data to improve the performance of analytical models is not new in the business literature. In fields such as market research, it has been used to enhance forecasting by mimicking human consumer responses (Goli & Singh, 2024). But despite the growing research interest in the potential of synthetic data in market research and segmentation, together with its significant shortcomings (Wang et al., 2025), the application of AI in this regard has not been explored. Therefore, the central research question of this study is the following:

RQ: Can synthetic data improve the accuracy of market segmentation performed by AI algorithms, and help improve the effectiveness of online communication campaigns?

This study draws on the theoretical basis of transfer learning to explain the movement of knowledge from one area to another. Although it originated in educational psychology, this principle has been adapted in computer science to explain the ease of knowledge transfer from one connected domain to another, thereby reducing the need for information in the second domain to solve the problem (Zhuang et al., 2020). This paradigm explains, in situations of data scarcity, the process of knowledge transmission from real data to artificial data (Wang et al., 2025).

We address our research question by conducting: (1) a case study of a lead-gathering company that sought to segment a potential market using the eXtreme Gradient Boosting (XGBoost) algorithm with data augmentation (DA), generated by Synthetic Minority Over-sampling Technique (SMOTE), and estimate the willingness to purchase five generic products (hearing aids, NGOs, energy distributors, telecommunications and finance); and, (2) a split A/B test for the 'Energy' product segment to validate the robustness of DA-enhanced segmentation.

Background

Segmentation criteria in e-commerce

Market segmentation is a fundamental phase in the implementation of strategic marketing (Iacobucci, 2016) and is essential for entrepreneurs (Choi et al., 2022; Saez-Ortuño et al., 2023b). It is the first stage of the S(Segmentation) → T(Targeting) → P(Positioning) process, and consists of dividing a heterogeneous market into smaller segments, each containing a substantial number of consumers who share similar and stable preferences, and respond similarly to marketing actions (DeSarbo et al., 2017). However, achieving this is not easy, and as Verhoef et al. (2003) argue, even large companies have difficulty selecting segments that meet all the criteria, so marketers have to balance the various requirements.

There is little consensus in the literature regarding the criteria for segmenting the market, and they have evolved over time (Wedel & Kamakura, 2000). Initially, geo-demographic and socio-economic variables were considered. However, the results were not very satisfactory, so these were combined with purchase behaviour variables, such as frequency of use, repeat purchases, and loyalty (DeSarbo et al., 2017). More sophisticated approaches, such as the use of psychographic variables (personality, values or lifestyles) or the combination of product benefits with socio-demographic variables, have also been tried, but have generally been discarded due to application difficulties (Wedel & Kamakura, 2000).

In the digital ecosystem, new entrepreneurs combining technology and management continue to propose innovations (Lucarelli et al., 2025; Maziriri et al., 2024) that are being combined with classic criteria, derived from demographic and behavioural variables (Saez-Ortuño et al., 2023b). For example, new criteria have been added, such as the use of keywords linked to sales advertising strategies and those related to contextual factors (Ahmadi et al., 2024). However, in this study, we only consider classical segmentation criteria based on demographic and behavioural variables for the online market.

Machine learning segmentation algorithm

Traditionally, market segmentation processes have relied on clustering algorithms based on demographic data and responses to questionnaires completed by potential customers (Wedel & Kamakura, 2000). However, the digital ecosystem has replaced these survey-based methods with big data: matrices containing millions of data points from multiple sources, whose analysis requires AI algorithms capable of extracting behavioural patterns imperceptible to humans (Chang & Fan, 2023; Sáez-Ortuño et al., 2023b).

Numerous algorithms have been applied to segment markets depending on the nature of the data. Those proposed for analysing frequencies include the Naive Bayes algorithm, J48, OneR (Chang & Fan, 2023), K-Means, and XGBoost (Sáez-Ortuño et al., 2023b), the latter being one of the most successful algorithms for modelling consumers' purchasing patterns and segmenting them according to their willingness to buy (Chen & Guestrin, 2016; Sáez-Ortuño et al., 2023b).

Although XGBoost is widely used in computer science and engineering to estimate purchase intentions for e-commerce products (Li, 2022; Song & Liu, 2020), academic studies in the field of innovation and entrepreneurship are not abundant. Among the few exceptions, Song and Liu (2020) used the behavioural data of 12,330 users to compare the results generated by XGBoost with those of a random forest algorithm, showing the superior performance of the former. In an applied study, Li (2022) used web data on consumer behaviour (e.g. time spent browsing by product category, time spent shopping) and demographic data (e.g. gender, age, education) to predict purchase intent using XGBoost and compared the results with those of a logistic regression model. More recently, in the field of innovation and entrepreneurship, Sáez-Ortuño et al. (2023b) used XGBoost to segment, profile and locate potential buyers of seven products from a dataset of 5389,731 users, finding that it outperformed the unsupervised K-Means algorithm.

Our exploratory research compares the internal predictive ability of the XGBoost algorithm to segment consumers based on their willingness to buy five generic products (hearing aids, NGOs, energy distributors, telecommunications and finance) using real versus synthetically generated DA data.

The XGBoost algorithm is a refined version of gradient boosting that operates by forming models from the partitioning of full sample data into a hierarchical decision tree structure. A full sample consists of a dependent variable and several independent variables. The algorithm works in a similar way to a Bayesian probabilistic model: 1st) From the values of the independent variables, it forms a structure of decision trees descending from the probability distribution and makes an initial estimate of the objective value (dependent variable), which is compared with the real value, whereupon a residual or difference is obtained; 2nd) This residual becomes an objective function, and new tree structures are generated and their values are estimated until a turning point is reached and the residuals can no longer be reduced (Chen & Guestrin, 2016).

From a methodological perspective, the process is similar to Newton's impulse, which attempts to solve a loss function using the information provided by the Hessian matrix (the square matrix formed by the second-order partial derivatives). It uses a common trick in random forest algorithms: creating multiple decision trees (column subsampling) to reduce the correlation between posterior trees (Chen & Guestrin, 2016; XGBoost, 2025).

From an operational perspective, the XGBoost algorithm, which is available in Python, R, JVM and others (XGBoost, 2025), requires a three-phase process to build the predictive model: (1) training, (2) validation and (3) estimation.

1) In the training phase, the algorithm uses a complete database (consisting of a vector $\boldsymbol{y}$, with the values of the dependent variables, and a matrix $\boldsymbol{X}$, with those of the independent variables). By forming the tree, it de-clusters the $\boldsymbol{X}$ matrix and estimates the values of $\widehat{y}$ that minimises the residuals. In order to be as accurate as possible in the training phase, around 90 % of the complete data is usually used (Sáez-Ortuño et al., 2023b). The XGBoost algorithm uses a dual objective function, *Obj (J)* (Eq. (1)):

$$Obj(J) = \sum_{n}^{i=1} l(\widehat{y}_i, y_i) + \sum_{t}^{i=1} \Omega(f_i) \tag{1}$$

where the loss function, $l(\widehat{y}_i, y_i)$, measures the difference between the estimated and actual value, and $\Omega(f_i)$is a manager term that moderates the complexity of the model and mitigates problems arising from overfitting (Chen & Guestrin, 2016). Since the algorithm establishes a sequential iterative process, the estimated value at each step of the decision tree will be determined by the estimated value and the error made in the previous step. Thus, the decomposition of the loss function results in the following Eq. (2):

$$Obj(J) = \sum_{n}^{i=1} l\left(\widehat{y}_i^{t-1} + f_t(x_i), y_i\right) + \sum_{t}^{i=1} \Omega(f_i) \tag{2}$$

Furthermore, the regulation function, $\Omega(f_i)$, can also be decomposed into two terms: a variable measuring the number of branches forming the tree, T, and an estimator of the squared weight of each branch, w_j^2. That is, both terms measure the complexity of the tree structure, since the value increases as the number of branches increases. Eq. (3) describes these terms:

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda \sum_{T}^{j=1} w_j^2 \tag{3}$$

Since Eq. (1) is differentiable, it can be solved using Taylor's expansion. We consider the gradient, resulting from the first derivative, $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(\widehat{y}_i^{t-1} + f_t(x_i), y_i\right)$, and the Hessian, resulting from the second derivative, $h_i = \partial^2_{\hat{y}_i^{t-1}} l\left(\widehat{y}_i^{t-1} + f_t(x_i), y_i\right)$ to solve the algorithm. The development of the objective function after Taylor expansion is shown in Eq. (4):

$$Obj(J) = \frac{-1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda}\right] + 2\gamma \tag{4}$$

where the function is divided by the ratio of the squared gradient and relative Hessian of the leftward-expanding nodes (subscript L), with the first term inside the parenthesis, and the second term, the ratio of the gradient and Hessian of the rightward-expanding nodes (subscript R). The term outside the parenthesis is an estimator of the number of branches forming the tree, allowing it to be pruned to avoid overfitting.

2) The internal validation of the model is used as an estimator of the machine learning's predictive accuracy and consists of testing the model generated in the training phase with the rest of the full dataset (in some cases 10 %). This involves introducing the matrix $\boldsymbol{X}$ of independent variables into the algorithm and estimating the corresponding independent values. The difference between the estimated and actual values is used to assess the algorithm's predictive ability (Sáez-Ortuño et al., 2023b; Song & Liu, 2020).

3) The forecasting phase consists of applying the trained algorithm to a sample of incomplete data, represented by a matrix $\boldsymbol{X}$ of independent variables, to estimate consumers' willingness to buy and group them into segments. This information helps entrepreneurs to select the market segments that are most likely to become customers and to target them through online communication campaigns (Sáez-Ortuño et al., 2023b; Song & Liu, 2020).

However, the actual or external validity of these estimates in terms of click probability, conversion probability and (long-term) conversion margin can only be verified ex-post, when actual conversion results are compared with the data predicted by the algorithmic model (Ahmadi et al., 2024; Yan et al., 2009).

Generating synthetic data using AI algorithms

The generation of augmented data is based on the principle of transfer learning. This principle is inspired by the human ability to transfer knowledge from one field to another, as long as the two are related. For example, if someone has learned to ride a bicycle, it will be easier for them to learn to ride a motorbike than to play the piano, due to the similar skills required (Zhuang et al., 2020).

The theory of transfer learning has been applied to machine learning by designing algorithms capable of leveraging knowledge from one domain into another, meaning that the latter requires less data to generate a result (Zhuang et al., 2020). In other words, the knowledge gained in the previous task is carried over to the next, rather than starting from scratch each time. In our context of augmented data generation, the algorithm learns from real data and translates that knowledge into artificial data that replicates human behaviours (Wang et al., 2025).

DA is mainly used in the training of AI algorithms, as this is the critical phase for building an effective predictive model. In most cases, real data is scarce (Wang et al., 2025). In fact, 80–90 % of the available data is usually used to train the algorithm and the remain 20 % serves for validation. However, the literature recommends a better balance between training and validation sample sizes to improve generalisability and avoid overfitting, so that the learning error during training does not have to be sacrificed to reduce the validation error (Shorten & Khoshgoftaar, 2019).

However, additional data can be extremely difficult to obtain for entrepreneurs who are launching an e-commerce business (Sučić Funko et al., 2023), due to high acquisition costs (Hair Jr. et al., 2019), growing concerns about how online information is handled (such as disclosure to third parties), and low consumer confidence in the ability of organisations to ensure data security and privacy (Hernandez et al., 2022; Sáez-Ortuño et al., 2023a).

Synthetic data generation is offered as a possible solution to overcome all these limitations, since from a restricted amount of real data, algorithms are able to reproduce large amounts of synthetic data that replicates human behaviour. In fields such as market research (Goli & Singh, 2024) and computational sciences (Massaro et al., 2021), these algorithms are used to complement and/or balance data before training machine learning models. The literature has identified two approaches: (1) parametric models, in which the algorithm learns the characteristic parameters of the multivariate distribution of the original data and uses them to generate synthetic data; and (2) non-parametric models, where the algorithm learns the frequency distribution of the original data (histograms) and uses it to generate synthetic data (Massaro et al., 2021). Popular examples include SMOTE and Generative Adversarial

Networks (GAN) (Chawla et al., 2002; Fernández et al., 2018; Sliman et al., 2023).

SMOTE is an oversampling technique that generates artificial data with the aim of balancing unbalanced samples (Fernández et al., 2018). For example, if in a sample of results from a communication campaign the female population is underrepresented, SMOTE will create synthetic data for the minority class until the sample is balanced. Although one might think that a simple way to increase the sample would be to randomly duplicate the available data, this leads to an overlearning problem in AI algorithms as no new information is provided. In contrast, SMOTE generates artificial data by randomly selecting an element from the minority group, determining its nearest neighbours, and generating a new element by linear combination or interpolation with the nearest neighbour. In this way, the added data incorporates new information and diversity into the training dataset (Fernández et al., 2018). Fig. 1 presents an illustrative example. Balanced samples improve the ability of AI algorithms to analyse data features by reducing model bias, resulting in more accurate forecasts (Chawla et al., 2002). However, as Hernandez et al. (2022) point out, most of these algorithms still work well with data matrices containing few variables and limited variability, but become problematic as the dimensions and range of values for each attribute increase. In other words, more robust solutions need to be found.

GAN represents a breakthrough in synthetic data generation. This algorithm analyses a complete data set using two neural networks that compete with each other during the training phase: a generator that learns the probability distribution and uses game theory to create new data, and a discriminator that tries to distinguish between real data and generated data (Gui et al., 2021). As the algorithm is trained, the generated data becomes increasingly realistic, and in turn, the discriminator refines its ability to detect artificial data (Hernandez et al., 2022). GANs have shown promising results in the medical field, where they can be used to generate synthetic images (Sliman et al., 2023).

Although GAN represents an advance over SMOTE, the latter is preferable when the database consists of matrices with a small number of variables and the samples are highly unbalanced, as GAN tens to bias models that are over-influenced by the majority class (Gui et al., 2021). Moreover, SMOTE is much easier to interpret for practitioners, such as entrepreneurs who do not have deep technological knowledge of synthetic data generation algorithms (Fernández et al., 2018).

A previous application in the e-commerce domain is Massaro et al. (2021), who used XGBoost to analyse a sample of 7212,348 records (instances or rows) and 14 attribute parameters (variables or columns) of 30,159 products (average 240 data points per product) collected over two and a half years. Since for many products there was not enough data to train the algorithm effectively, the original dataset was multiplied by 10 useding a DA algorithm. The formula used to do so was a linear combination of sines and cosines with a given amplitude range (a, b), where the dependent variable (number of sales generated in a day) was randomly determined (Massaro et al., 2021).

Our study uses the XGBoost algorithm for data analysis, but proposes a literature-validated algorithm, SMOTE (Fernández et al., 2018), to generate augmented data, with the aim of achieving more accurate predictions than would be possible with the available data.

Method

Based on the existing literature and with the aim of illustrating the learning capacities of AI tools, this research proposes two empirical studies: an exploratory analysis of a real-world case, and a field experiment involving one of the analysed products (Caiado et al., 2021). Both studies outline the steps followed by a company in the e-commerce sector to address challenges in a business or entrepreneurial context. Fig. 2 summarises the research stages and design.

The case study describes the application of the XGBoost algorithm in a market segmentation process based on estimating willingness to buy each of five products. The hearing aid product includes the online promotion and sale of devices from different brands designed to correct mild to moderate hearing loss (Goman & Lin, 2016). The NGO product, which has been criticised for using a denial-based (non-governmental) definition that makes it difficult to determine which organisations may or may not be included, consists of all online memberships and donation agreements (Gray et al., 2006). Energy suppliers are companies that provide energy to end consumers, from distributors to solar panel installation companies, which, according to an EIA (2020) estimate, account for 21.69 % of global energy. The telecommunications product includes any online promotion and distribution of telephony, fibre optic and mobile internet, cloud services and similar (Wei & Chiu, 2002).

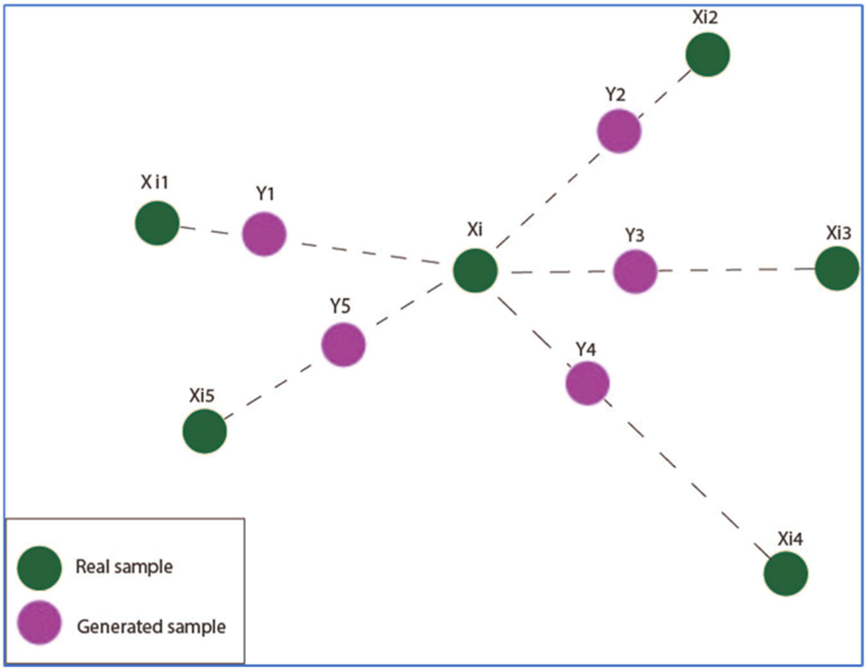

Fig. 1. Synthetic data creation.

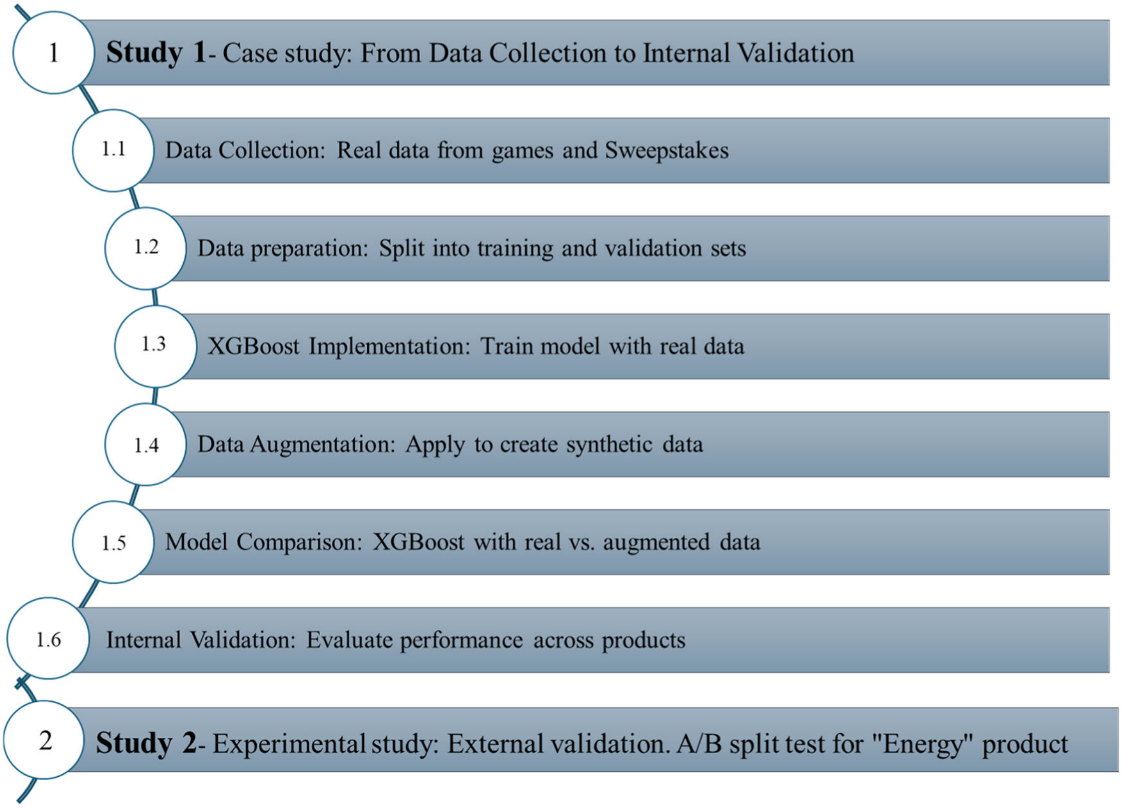
1
Study 1- Case study: From Data Collection to Internal Validation
1.1
Data Collection: Real data from games and Sweepstakes
1.2
Data preparation: Split into training and validation sets
1.3
XGBoost Implementation: Train model with real data
1.4
Data Augmentation: Apply to create synthetic data
1.5
Model Comparison: XGBoost with real vs. augmented data
1.6
Internal Validation: Evaluate performance across products
2
Study 2- Experimental study: External validation. A/B split test for "Energy" product

Fig. 2. Research steps and design.

Finally, finance includes the promotion, distribution and sale of savings, investment and financing products by utilities, financial institutions and banks (Ramaswami et al., 2000).

Study 1: case study

We collaborated with CoRegistros, S.L.U., a leading Spanish lead generation agency, to illustrate the process of segmenting and estimating the willingness to purchase the five generic products, and to explore whether DA can improve the performance of the XGBoost algorithm. To do this, we took a sample of 5389,731 users (36 % male and 64 % female), captured between 2010 and 2022, representing 11 % of the population of Spain. The sponsoring company collects information from internet users by advertising challenges and sweepstakes. The former test potential consumers with quizzes on history, geography, cooking, and other topics, and, once they have demonstrated their knowledge, they are entered into a prize draw to win attractive prizes (e.g. an iPhone or an Alexa Echo Dot) if they fill in their data on a form. The company then cleans this raw data to remove any false information and sells it to third parties for use in communication campaigns (Sáez-Ortuño et al., 2023a). Before using the data, we ensured that the company complies with the European General Data Protection Regulation (GDPR) and the corresponding Spanish LOPD-RGPD (*Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales*) (AEPD, 2018).

The database consists of demographic and behavioural information on the participants, provided by both consumer applicants and advertising companies and promoter clients. Specifically, it contains two matrices grouped into eight blocks, forming 37 columns. The first matrix contains five blocks of descriptive data: 1) Demographics of participants, 2) Type of communication campaign, challenge game or sweepstake through which the information was collected, 3) Users who received recommendations from sponsors, and converted to buyers in the past, 4) Types of advertisements used by client promoters to recommend their products, and 5) Specific to sweepstakes, how information was collected. The second matrix comprises three blocks with marketing specifications: 1) Advertising campaigns that each user received, 2) Type of prize for sweepstake winners (beauty, electronics, home, iPhone, leisure and travel) and 3) Users who purchased one of the promoted products in the most recent campaign. Since the data was of different types (string, Boolean, floating-point or interval), and combining binary variables with scales or ratios might result in larger variances dominating the model erroneously, all variables were transformed into Boolean binary format (Chakraborty et al., 2009). For more details on the origin, structure and nature of the data used in this study, see Sáez-Ortuño et al. (2023b).

Although the matrix contained over five million users, only 6852 were purchasers of any of the products promoted by the client companies. This sub-sample is referred to as complete data because it incorporates the dependent variable (conversion into a purchaser). The second columns in Table 1 shows the purchase frequencies for each product, indicating the data available for each product. Since most of the products have low frequencies, we opted to complement them with DA to explore whether this additional input improved the performance of the XGBoost algorithm in profiling the segments.

Table 1
Database with complete data.

Product	Total frequencies	Training data	Data for validation test	Training data (%)
Hearing Aids	5337	4828	509	90.46
NGOs	830	757	73	91.20
Energy	318	287	31	90.25
Telcos	251	214	37	85.25
Finance	116	100	16	86.20
Total	6852	6186	666	90.02

Results of the exploratory case study

To train the XGBoost algorithm, the complete data was split into training data (approximately 90 %, or 6186 records, see the 3rd column in Table 1), and data used to test the internal validity of the model (666 records, 4th column). Table 1 also shows the percentage represented by the training data (5th column).

Given the scarcity of data to train the algorithm, we applied the SMOTE algorithm to multiply its volume by 10, while simultaneously balancing the data across products. The result is shown in Table 2. Thus, the training data rose from 6852 to 59,627 records (3rd column in Table 2), and the validation test data from 666 to 6703 (4th column in Table 2). However, while the sample for Hearing Aids only had to be multiplied by 2.5, for Finance it had to be increased by a factor of 114 (5th column in Table 2).

The XGBoost algorithm was trained both with real (Table 1) and augmented (Table 2) data, and the results of the optimal model are shown in Table 3 for the former and Table 4 for the latter. On average, the variability explained by the tree structure model (4th column) was slightly higher for estimates made with the full data (82.14 %, Table 3) than for those made with DA (78.09 %, Table 4), which might be a cause for concern. However, it is worth noting that the fitting model went from analysing just over 6000 records (Table 3) to approximately 60,000 (Table 4), and despite the considerable manipulation involved in this tenfold increase, there was only a 4 % loss in accuracy.

In the next phase, the model was tested to assess internal validity, using the remaining 10 % of user data with conversions (Table 5) and that generated by DA (Table 6). The goal was to use the optimal model from the training phase to estimate the vector values of the dependent variable, both for real (666 items) and DA-generated (6703 items) subsamples. The results show that for real data the model achieved an average accuracy of 89.94 % (deviation of 10.6 % as shown in the 4th column in Table 5). However, a product-by-product analysis reveals considerable inaccuracies, with the sole exception of Hearing Aids, which showed a deviation of less than 15 %. In contrast, the results for the DA sample are significantly better (Table 6). Accuracy is 97.82 % (deviation of 2.18 %, as shown in the 4th column in Table 6). Four of the five products showed a marked improvement. A Chi-square test indicated that the distributions of the actual data by product (2nd column in Table 5) and the estimated data (3rd column in Table 5) are significantly different, 89.93 % (599/666; $X^2 = 91.59$, $p < 0.000$). By contrast, Table 6 shows that the distribution achieved with DA presented no significant difference: 97.86 % (6849/6703; $X^2 = 1.5924$, n.s.). A comparison of deviations by product between the real sample (4th column in Table 5) and the DA sample (4th column in Table 6) also reveals significant differences. For example, for the Energy product, the deviation decreased from over 70 % to less than 15 % ($X^2 = 15.60, p < 0.000$). The only product for which accuracy worsened with DA is Hearing Aids (14.34 % with real data vs. 34.15 with DA; $X^2 = 92.95$, $p < 0.000$).

The final step is demand forecasting, taking the incomplete dataset (5389,731 users without a dependent variable), and applying the tree-optimised model generated by XGBoost to estimate the willingness to buy of each potential consumer for the five products. These individuals are then grouped by similarity, and market segments are formed. These results are of little relevance to this research and, at the request of the sponsoring company, are not published here.

In conclusion, although the XGBoost algorithm is widely recognised in the literature for its considerable power (Chen & Guestrin, 2016), its performance depends on the amount of complete information fed to the algorithm in the training phase. In this study, for four of the five products the number of cases ranges from 100 to 757, resulting in very imprecise estimates of internal validity (deviations between 70 and 94 %).

To summarise the first part of the research question, the synthetic data generated by the SMOTE algorithm substantially improved the accuracy of market segmentation, as illustrated by the considerable reduction in the average error. However, in the per-product analysis, the XGBoost algorithm performed better for the four products with small samples, but not for the one with a large sample of real data, suggesting a positive but only partial effect.

Table 2
Database after SMOTE algorithm.

Product	Total frequencies	Training data generated by SMOTE	Data for validation test	Sample increase (%)
Hearing Aids	5337	11,963	1303	248.57
NGOs	830	11,938	1328	1598.31
Energy	318	11,902	1364	4171.70
Telcos	251	11,861	1405	5285.26
Finance	116	11,963	1303	11,436.21
Total	6852	59,627	6703	968.04

Table 3
Accuracy of the training model with complete data.

Product	Training data	Estimates by trained model	Accuracy (%)
Hearing Aids	4828	3735.91	77.38
NGOs	757	753.29	99.51
Energy	287	284.76	99.22
Telcos	214	208.59	97.47
Finance	100	98.99	98.99
Total	6186	5081.53	82.14

Table 4
Accuracy of the training model with SMOTE augmented data.

Product	Training data	Estimates by trained model	Accuracy (%)
Hearing Aids	11,963	7638.37	63.85
NGOs	11,938	8767.26	73.44
Energy	11,902	9885.80	83.06
Telcos	11,861	8541.10	72.01
Finance	11,963	11,732.11	98.07
Total	59,627	46,564.66	78.09

Table 5
Accuracy achieved with the validation test with complete data.

Product	Data for validation test	Estimates by trained model	Deviation (%)
Hearing Aids	509	582	14.34
NGOs	73	4	94.52
Energy	31	9	70.97
Telcos	37	1	97.30
Finance	16	3	81.25
Total	666	599	10.06

Table 6
Accuracy achieved with the validation test with SMOTE augmented data.

Product	Data for validation test	Estimates by trained model	Deviation (%)
Hearing Aids	1303	1748	34.15
NGOs	1328	936	29.52
Energy	1364	1168	14.37
Telcos	1405	1691	20.36
Finance	1303	1306	0.23
Total	6703	6849	2.18

Study 2: experimental study

To complement the study an answer the second part of the research question on whether DA can improve the effectiveness of online communication campaigns, an exploratory quasi-experiment was conducted, taking advantage of the fact that one of the advertisers, a solar panel installation company, planned to launch an email marketing campaign through its advertising agent. The company intended to use the willingness-to-buy data estimated in the case study, grouped under the Energy segment. This provided an opportunity to externally validate, if only with one of the five products, whether the predictive improvements achieved with the DA samples could be extended to a real environment.

Three basic factors are commonly cited to explain the success of digital advertisements over time: creative format, message content and targeting (Bruce et al., 2017). Researchers have also studied email advertising for both B2B and B2C marketing (Zhang et al., 2017), finding that its main advantages include fast, seamless and cost-effective interaction with customers in the short term (Ahmadi et al., 2024) and, in the medium to long term, the ability to forge a relational bond that fosters loyalty and future purchases (Kumar et al., 2014).

Several metrics can be used to measure the performance of email marketing. In the short term, common indicators include email opening, click-through and conversion rates, while in the medium term, conversion performance evolution, customer satisfaction and similar factors are considered (Kumar et al., 2014). However, sales are not always measurable in the short term, especially for installation services (solar panels) or complex products (cars), where conversion may simply mean requesting information, visiting a dealer or arranging an appointment with an agent (Ahmadi et al., 2024; Núñez-Cansado et al., 2024). Previous studies caution against the use of click-through rate instead of conversion rate, as click behaviour in email marketing is dynamic, tending to be high earlier on but diminishing as the consumer becomes more familiar with the product, pattern that is often independent of the actual conversion rate (Zhang et al., 2017).

For the design of the quasi-experiment, an A/B split model was used, similar to Ahmadi et al. (2024). Both the creative format and the content of the advertising message were constant, the sole variation being the segmentation criterion used to choose the two samples (Bruce et al., 2017). Given the complexity of the solar panel sales process and the fact that the analysis was conducted in the short term, conversion (defined as a request to meet a salesperson) was considered as the outcome variable.

The experiment consisted of sending four advertising e-mails (called “impacts”) to a sample of 20,000 potential customers over a six-month period, based on discussions with the sponsoring company’s management, from January to June 2024, and counting the number of consumers who requested a visit from an agent (dependent variable). Of the slightly more than 250,000 potential consumers in the Energy segment, two samples of 10,000 users were selected. All were over the age of 35, resided in a single-family house, and had an estimated willingness to convert of more than 60 %. One sample was selected using forecasts based on actual data and the other on the basis of DA, ensuring no overlap.

The results of the experiment are shown in Table 7, where the average conversion rate was 7.29 %. In line with expectations, the conversion rate for the DA subsample (7.79 %) was significantly higher than for the real data (6.80 %), as confirmed by a Chi-square test (779/10,000 vs. 680/10,000; $X^2 = 6.71, p < 0.01$). Thus, this study provides a positive response, albeit based on just one case.

Table 7
Results of the experiment in absolute values.

Sub-sample	Potential customers	Conversions	Conversion rates (%)
From augmented data	10,000	779	7.79
From real data	10,000	680	6.80
Total	20,000	1459	7.295

Discussion and conclusions

This study investigates the role that artificial data generation software can play in improving market segmentation and demand forecasting, and hence the performance of online advertising campaigns. Specifically, the SMOTE algorithm was used to improve the performance of XGBoost, and was subsequently tested in a field experiment promoting solar panels. The results show that synthetic data improved the training of the XGBoost algorithm and its predictive capacity for four of the five products analysed. The field experiment also shows that the sample of potential consumers selected from DA-based estimates generated a significantly higher conversion rate than the sample using real data. Given the descriptive and exploratory nature of the study, the results are not conclusive. Nevertheless, they suggest that combining real and synthetic data can improve the accuracy of AI algorithms and enhance the effectiveness of online communication campaigns.

The ability to augment samples with artificial data addresses the growing need for scalable and cost-effective methods in market research (Wang et al., 2025). For entrepreneurs, DA enables sophisticated analyses without the costs, time constraints and other restrictions of traditional market research (Hair Jr et al., 2019). Moreover, these AI tools democratise access to market information by enabling start-ups and smaller companies to perform sophisticated analyses that were previously only feasible for larger firms (Choi et al., 2022).

DA algorithms are based on the assumption that artificial consumer profile data elicits the same responses as if it came from real consumers (Brand et al., 2023). That is, DA algorithms may be able to simulate consumer behaviour and are therefore potentially valuable for developing a better understanding of it (Wang et al., 2025). However, these tools still have important shortcomings when it comes to replicating human behaviour. For example, real consumers often change their preferences based on factors such as prior experience, economic, technological and cultural changes (Iacobucci, 2016), while DA-generated consumers neither have prior experience nor evolve over time. They simply replicate data from the current situation, which leads to implicit, unintentional assumptions about the environment in which the purchase decision is made (Gui & Toubia, 2023).

Another controversial point is the amount of data needed for an algorithm to achieve optimal performance. The literature provides evidence that AI algorithms like XGBoost work best with large, balanced databases, but their performance decreases with smaller sample sizes (Chen & Guestrin, 2016; XGBoost, 2025). Our study shows that, with real samples of about 5000 cases, the XGBoost algorithm performs reasonably well in terms of both explained variability (77 %) and internal validity (14 % error). However, when the sample is increased by about 7000 cases generated by SMOTE, performance drops to 63 % explained variability and 34 % error. By contrast, when much smaller samples are considered (less than 1000), the addition of artificial data moderately reduces explained variability by between 1 and 26 %, but at the same time improves internal validity, with predictive deviations falling from 70–97 % to 0.2–29 %.

Finally, a field experiment was conducted using one of the products that required more DA support to improve its internal validity, namely the energy product. In this case, the sample of potential customers selected by the DA-configured model performed better than the sample configured with the model based on real data, supporting external validity (Ahmadi et al., 2024). However, this experiment was limited to a single product category, which also benefited considerably from the DA-boosted sample. Therefore, the validation is only partial.

Our findings underscore that the benefits of DA are not universally applicable, and prior analysis is recommended before using it in predictive models. A hybrid approach combining traditional data collection with advanced DA techniques should therefore be pursued, with efforts focused on identifying the optimal combination that maximises predictive power, especially for segmentation and purchase intention estimation (Chawla et al., 2002; Choi & Mela, 2019; Johnson, 2013).

The evolution of marketing analytics will be increasingly driven by AI, synthetic data generation, and other data augmentation strategies. These tools promise not greater segmentation and targeting accuracy but new possibilities for creativity, personalisation, and ethical consumer engagement. However, fully realising this potential requires closer cross-disciplinary collaboration. Scholars and practitioners can benefit from advances in data science, while computer scientists and AI researchers can gain valuable insights from the behavioural, managerial, and ethical perspectives of marketing and business. Interdisciplinary research at this intersection can accelerate the development of more robust, transparent, and impactful analytics frameworks that support both entrepreneurial innovation and societal trust.

Theoretical implications

Our study suggests that the application of DA algorithms to complete training data improves the performance of AI algorithms by making more accurate estimates and, at the same time, reduces errors in the selection of the target market segment. The theoretical framework, based on transfer learning, states that DA algorithms are able to learn the relationships established in real data and transfer them to the artificial data generation model. This approach is supported empirically by the fact that our estimator model not only reduces errors, but also achieves estimates with significant savings in real data. However, it is important to note that SMOTE was especially effective in this study because the data sample was Boolean-coded and distributed in an unbalanced manner. Therefore, we cannot assume that it would yield the same results with other types of data or distributions.

Moreover, the use of DA poses not only theoretical but also ethical challenges (Fernández et al., 2018). Market researchers tend to be highly scrupulous in their data handling, and the idea of 'making up data', even with AI, seems unethical and unprofessional. However, a review of the literature shows that these methods are increasingly accepted as researchers adapt to changing circumstances (Gui & Toubia, 2023; Wang et al., 2025). Experiments conducted with simulation models are often based on artificial data yet have been known to yield sufficient evidence to accept hypotheses (Hernandez et al., 2022). In addition, some increasingly widespread techniques for extending the original data, such as bootstrap resampling, now enjoy broad academic support (Chernick, 2011). However, the most pressing challenge today is the growing sensitivity of consumers to disclosing information to companies, due to increased privacy concerns (Acquisti et al., 2016), and tighter legal barriers against the use and distribution of consumer information (Choi & Mela, 2019; Tucker, 2014). Synthetic data offers a practical solution to these challenges (Tucker, 2014). However, its use needs to be managed transparently, informing clients about its use in model estimations, monitoring possible deviations, and complying with possible changes in data protection regulations (Acquisti et al., 2016; Sáez-Ortuño et al., 2024; Tucker, 2014).

Looking ahead, our study contributes to more efficient, economical and democratic market research, thanks to the potential for integrating real and artificial data. However, challenges remain for DA algorithms, and the results of this study can serve as a basis for future innovations, helping researchers and practitioners to improve the way they address trade-offs between artificial and human data.

Managerial implications

This research offers several recommendations for entrepreneurs. We show how synthetic data-generation algorithms can improve the performance of AI-based data analysis, thereby expanding the strategic possibilities for resource-constrained start-ups to perform analyses and obtain diagnostics similar to those of larger companies. Recent studies highlight the transformative role played by AI in corporate processes. However, integrating digital transformation with innovation is complex, requiring multi-level changes on both am organisational and technological level to drive competitive advantage (Saeedikiya et al., 2025; Salehe et al., 2024).

Our work contributes to the literature by showing how DA algorithms are a disruptive innovation in market research, as they enable more refined forecasts that, for example, lead to more effective communication campaigns. However, as Ludwig et al. (2025) point out in relation to large language models (LLM), the combination of artificial and real data is essential during the training phase to ensure accuracy. This must be complemented by external validation, without which the errors created by the automated data generation cannot be properly evaluated or explained. For entrepreneurs, both our findings and recent literature emphasise the need to combine artificial and real data to correct for certain deviations caused by artificial data, such as overemphasis on current conditions (Wang et al., 2025).

However, the most appropriate ratio of real to artificial data remains an open question. In our case, where all variables were transformed into Boolean format (Chakraborty et al., 2009), the Hearing Aids segment (2.5 times the sample) was the one where the greatest balance between real and augmented data generated the largest error in internal validity (34 %). This is in stark contrast to the other segments, which had between 15 and 114 times more synthetic data than real data. In other words, the best fits occurred when the dataset was heavily skewed toward synthetic data. This contradicts Wang et al. (2025), who recommend more balanced distributions, albeit in studies involving chain variables.

It is also important to note the wide range of DA algorithms, each of which is best suited to different types of databases. While the GAN algorithm is useful for medical imaging devices (Sliman et al., 2023), as is LLM for a wide variety of strings and responses (Ludwig et al., 2025), SMOTE is most effective with a small number of variables and significant imbalance (Chawla et al., 2002). However, as LLM research shows, a naive application of artificially generated data can distort both predictive capacity and parameter estimation. Therefore, it is important to test results against real data to establish external validity (Ludwig et al., 2025). This study addressed the need for external validation through an experimental test.

The combination of predictive algorithms like XGBoost with DA offers strong potential for refining consumer profiling and developing more personalised and effective marketing communication (Ahmadi et al., 2024). To realise this potential, entrepreneurs and market researchers must be trained in advanced analytics that combine real and synthetic data (Wang et al., 2025).

Limitations and future research directions

This study has certain limitations that could be addressed in future research. First, DA was used to improve the accuracy of market segmentation in online advertising, but these algorithms could also be applied to other marketing domains such as product design, customer relationship management or even offline marketing strategies (Choi & Mela, 2019; Johnson, 2013; Madzík et al., 2024).

Second, this study used a specific database with Boolean coding, the XGBoost algorithm and SMOTE. Based on the above discussion, it is important to examine applications with other algorithms and consider different levels of data augmentation.

Furthermore, although the literature recommends estimation of external validity (Ludwig et al., 2025), this was only assessed here for one product in the Energy category, which limits the generalisability of our findings. A logical extension would be to test the external validity of the other products.

Another extension would be replication of this cross-sectional study at various points in time, in the form of a longitudinal study. This would consider the long-term impacts of the use of synthetic data in marketing strategies and its effects on the sustainability of customer relationships, as well as brand loyalty.

Finally, although our original idea was to estimate conversion readiness, we ended up measuring not the probability of a click, but the probability of arranging a visit from a sales agent, which is the closest available proxy for an actual sale. Other measures such as first click, likelihood of conversion and (long-term) margin per conversion could also be investigated (Yan et al., 2009). Other objectives, such as increasing brand value or enhancing corporate image, were not considered.

CRediT authorship contribution statement

Ruben Huertas-Garcia: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Laura Sáez-Ortuño:** Writing – original draft, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Santiago Forgas-Coll:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Javier Sánchez-García:** Supervision, Project administration, Investigation.

References

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

AEPD. (2018). *Ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales*. Agencia Española de Protección de Datos. https://www.aepd.es/ Accessed 5 November, 2024.

Ahmadi, I., Abou Nabout, N., Skiera, B., Maleki, E., & Fladenhofer, J. (2024). Overwhelming targeting options: Selecting audience segments for online advertising. *International Journal of Research in Marketing, 41*(1), 24–40. https://doi.org/10.1016/j.ijresmar.2023.08.004

Blankesteijn, M. L., Houtkamp, J., & Bossink, B. A. G. (2024). Towards transformative experiential learning in science- and technology-based entrepreneurship education for sustainable technological innovation. *Journal of Innovation & Knowledge, 9*(3), Article 100544. https://doi.org/10.1016/j.jik.2024.100544. Article.

Brand, J., Israeli, A., & Ngwe, D. (2023). *Harvard Business School Marketing Unit Working Paper* (pp. 23–062). https://doi.org/10.2139/ssrn.4395751

Bruce, N. I., Murthi, B. P. S., & Rao, R. C. (2017). A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement. *Journal of Marketing Research, 54*(2), 202–218. https://doi.org/10.1509/jmr.14.0117

Buck, C., Clarke, J., Torres de Oliveira, R., Desouza, K. C., & Maroufkhani, P. (2023). Digital transformation in asset-intensive organisations: The light and the dark side. *Journal of Innovation & Knowledge, 8*(2), Article 100335. https://doi.org/10.1016/j.jik.2023.100335. Article.

Caiado, R. G. G., Scavarda, L. F., Gavião, L. O., Ivson, P., de Mattos Nascimento, D. L., & Garza-Reyes, J. A. (2021). A fuzzy rule-based industry 4.0 maturity model for operations and supply chain management. *International Journal of Production Economics, 231*, Article 107883. https://doi.org/10.1016/j.ijpe.2020.107883. Article.

Celis, L. E., Lewis, G., Mobius, M., & Nazerzadeh, H. (2014). Buy-it-now or take-a-chance: Price discrimination through randomized auctions. *Management Science, 60*(12), 2927–2948. https://doi.org/10.1287/mnsc.2014.2009

Chakraborty, H., Moore, J., Carlo, W. A., Hartwell, T. D., & Wright, L. L. (2009). A simulation-based technique to estimate intracluster correlation for a binary variable. *Contemporary Clinical Trials, 30*(1), 71–80. https://doi.org/10.1016/j.cct.2008.07.008

Chang, Y. T., & Fan, N. H. (2023). A novel approach to market segmentation selection using artificial intelligence techniques. *The Journal of Supercomputing, 79*(2), 1235–1262. https://doi.org/10.1007/s11227-022-04666-2

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). *Xgboost:* A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chernick, M. R. (2011). *Bootstrap methods: A guide for practitioners and researchers*. John Wiley & Sons.

Choi, H., & Mela, C. F. (2019). Monetizing online marketplaces. *Marketing Science, 38*(6), 948–972. https://doi.org/10.1287/mksc.2019.1197

Choi, T. M., Kumar, S., Yue, X., & Chan, H. L. (2022). Disruptive technologies and operations management in the Industry 4.0 era and beyond. *Production and Operations Management, 31*(1), 9–31. https://doi.org/10.1111/poms.13622

DeSarbo, W. S., Chen, Q., & Stadler Blank, A. (2017). A parametric constrained segmentation methodology for application in sport marketing. *Customer Needs and Solutions, 4*, 37–55. https://doi.org/10.1007/s40547-017-0086-7

EIA. (2020). Residential Energy Consumption Survey (RECS). https://www.eia.gov/consumption/residential/ Accessed 8 October, 2024.

eMarketer. (2024a). Worldwide Retail Media Ad Spending Forecast 2024: A Fifth of All Digital Ad Spending Will Go to the Fast-Growing Format. www.emarketer.com/content/worldwide-retail-media-ad-spending-forecast-2024 Accessed 24 September 2024.

eMarketer. (2024b). Worldwide ad spend will see steady growth through 2028. https://www.emarketer.com/content/worldwide-ad-spend-will-see-steady-growth-through-2028 Accessed 24 September 2024.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science, 43*(4), 709–722. https://doi.org/10.1287/mksc.2023.0306

Goman, A. M., & Lin, F. R. (2016). Prevalence of hearing loss by severity in the United States. *American Journal of Public Health, 106*(10), 1820–1822. https://doi.org/10.2105/AJPH.2016.303299

Google. (2023). Llega a una audiencia nueva o más amplia con la segmentación de la Red de Display de Google. https://ads.google.com/intl/es_es/home/resources/articles/reach-larger-new-audiences/ Accessed 16 September 2024.

Gray, R., Bebbington, J., & Collison, D. (2006). NGOs, civil society and accountability: Making the people accountable to capital. *Accounting, Auditing & Accountability Journal, 19*(3), 319–348. https://doi.org/10.1108/09513570610670325

Gui, G., & Toubia, O. (2023). The challenge of using LLMs to simulate human behavior: A causal inference perspective. arXiv https://arxiv.org/abs/2312.15524.

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2021). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering, 35*(4), 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191

Hair, J., Jr, Page, M., & Brunsveld, N. (2019). *Essentials of business research methods* (4th Ed.). Routledge. https://doi.org/10.4324/9780429203374

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing, 493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

Iacobucci, D. (2016). *Marketing management* (6th Ed.). South-Western.

Johnson, J. P. (2013). Targeted advertising and advertising avoidance. *The RAND Journal of Economics, 44*(1), 128–144. https://doi.org/10.1111/1756-2171.12014

Juárez-Varón, D., & Monreal, E. G. (2025). Market-oriented entrepreneurship and the impact of social media and crowdsourcing on continuous organizational learning. An empirical study. *Sustainable Technology and Entrepreneurship, 4*, Article 100088. https://doi.org/10.1016/j.stae.2024.100088

Kumar, V., Zhang, X., & Luo, A. (2014). Modeling customer opt-in and opt-out in a permission-based marketing context. *Journal of Marketing Research, 51*(4), 403–419. https://doi.org/10.1509/jmr.13.0169

Li, G., Sun, X., Ye, M., & Mardani, A. (2024). Assessment of digital transformation challenges for business model innovation in the context of higher education institutions using a decision support approach. *Journal of Innovation & Knowledge, 9*(3), Article 100527. https://doi.org/10.1016/j.jik.2024.100527. Article.

Li, L. (2022). Research on improved xgboost algorithm for big data analysis of e-commerce customer churn. *International Journal of Advanced Computer Science and Applications, 13*(12). https://doi.org/10.14569/IJACSA.2022.01312124

Lucarelli, C., Marinelli, N., Micozzi, A., Zifaro, M., Palazzo, M., & Ferri, A. (2025). Dynamic interactions between entrepreneurial domains and outcomes: The evolution of innovation ecosystems. *Journal of Innovation & Knowledge, 10*(3), Article 100696. https://doi.org/10.1016/j.jik.2025.100696

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). *Large language models: An applied econometric framework (NBER working paper no. 33344)*. National Bureau of Economic Research. https://doi.org/10.3386/w33344

Madzík, P., Falát, L., Yadav, N., Lizarelli, F. L., & Čarnogurský, K. (2024). Exploring uncharted territories of sustainable manufacturing: A cutting-edge AI approach to uncover hidden research avenues in green innovations. *Journal of Innovation & Knowledge, 9*(3), Article 100498. https://doi.org/10.1016/j.jik.2024.100498. Article.

Massaro, A., Panarese, A., Giannone, D., & Galiano, A. (2021). Augmented data and XGBoost improvement for sales forecasting in the large-scale retail sector. *Applied Sciences, 11*(17), 7793. https://doi.org/10.3390/app11177793. Article.

Maziriri, E. T., Dzingirai, M., Nyagadza, B., & Mabuyana, B. (2024). From perceived parental entrepreneurial passion to technopreneurship intention: The moderating role of perseverance and perceived parental entrepreneurial rewards. *Sustainable Technology and Entrepreneurship, 3*, Article 100051. https://doi.org/10.1016/j.stae.2023.100051. Article.

Núñez-Cansado, M., Carrascosa Mendez, G., & Juarez-Varon, D. (2024). Analysis of the residual effect using neuromarketing technology in audiovisual content entrepreneurship. *Sustainable Technology and Entrepreneurship, 3*, Article 100069. https://doi.org/10.1016/j.stae.2023.100069. Article.

Ramaswami, S. N., Strader, T. J., & Brett, K. (2000). Determinants of on-line channel use for purchasing financial products. *International Journal of Electronic Commerce, 5*(2), 95–118. https://doi.org/10.1080/10864415.2000.11044207

Saeedikiya, M., Salunke, S., & Kowalkiewicz, M. (2025). The nexus of digital transformation and innovation: A multilevel framework and research agenda. *Journal of Innovation & Knowledge, 10*(1), Article 100640. https://doi.org/10.1016/j.jik.2024.100640. Article.

Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., & Sánchez-García, J. (2023a). Online cheaters: Profiles and motivations of internet users who falsify their data online. *Journal of Innovation & Knowledge, 8*(2), Article 100349. https://doi.org/10.1016/j.jik.2023.100349. Article.

Sáez-Ortuño, L., Huertas-Garcia, R., Forgas-Coll, S., & Puertas-Prats, E. (2023b). How can entrepreneurs improve digital market segmentation? A comparative analysis of supervised and unsupervised learning algorithms. *International Entrepreneurship and Management Journal, 19*(4), 1893–1920. https://doi.org/10.1007/s11365-023-00882-1

Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., & Puertas-Prats, E. (2024). Chasing spammers: Using the Internet protocol address for detection. *Psychology & Marketing, 41*(6), 1363–1382. https://doi.org/10.1002/mar.21985

Salehe, M. A., Sesabo, J. K., Isaga, N., & Mkuna, E. (2024). Individual entrepreneurial orientation and firm performance: The mediating role of sustainable entrepreneurship practices. *Sustainable Technology and Entrepreneurship, 3*, Article 100079. https://doi.org/10.1016/j.stae.2024.100079. Article.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Sliman, H., Megdiche, I., Alajramy, L., Taweel, A., Yangui, S., Drira, A., & Lamine, E. (2023). MedWGAN based synthetic dataset generation for Uveitis pathology. *Intelligent Systems with Applications, 18*, Article 200223. https://doi.org/10.1016/j.iswa.2023.200223. Article.

Song, P., & Liu, Y. (2020). An XGBoost algorithm for predicting purchasing behaviour on E-commerce platforms. *Tehnički vjesnik, 27*(5), 1467–1471. https://doi.org/10.17559/TV-20200808113807

Sučić Funko, I., Vlačić, B., & Dabić, M. (2023). Corporate entrepreneurship in public sector: A systematic literature review and research agenda. *Journal of Innovation & Knowledge, 8*(2), Article 100343. https://doi.org/10.1016/j.jik.2023.100343. Article.

Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research, 51*(5), 546–562. https://doi.org/10.1509/jmr.10.0355

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. (2003). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems, 34*(4), 471–481. https://doi.org/10.1016/S0167-9236(02)00069-6

Wang, M., Zhang, D. J., & Zhang, H. (2025). Large language models for market research: A data-augmentation approach. *arXiv*. https://doi.org/10.48550/arXiv.2412.19363

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Kluwer Academic Publishers.

Wei, C. P., & Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*(2), 103–112. https://doi.org/10.1016/S0957-4174(02)00030-1

XGBoost. (2025). Documentation XGBoost. https://xgboost.readthedocs.io/en/latest/install.html Accessed July 4, 2025.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How much can behavioral targeting help online advertising?. In *Proceedings of the 18th international conference on World Wide Web* (pp. 261–270). https://doi.org/10.1145/1526709.1526745

Zhang, X., Kumar, V., & Cosguner, K. (2017). Dynamically managing a profitable email marketing program. *Journal of Marketing Research, 54*(6), 851–866. https://doi.org/10.1509/jmr.16.0210

Zhang, Y., Wang, H., Zheng, K., & Yang, W. (2025). Empowering women's entrepreneurship: The role of green knowledge, innovation, and family support. *Journal of Innovation & Knowledge, 10*(1), Article 100639. https://doi.org/10.1016/j.jik.2024.100639. Article.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., He, Q., & others. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555