A Predictive Information System: Harnessing Purchase Intent Signals Buried in Digital Data

Laura Sáez-Ortuño, Santiago Forgas-Coll, Ruben Huertas-Garcia, Javier Sánchez-Garcia

Abstract

E-commerce, whereby consumers are able to perform transactions easily online, has undergone rapid growth since its inception. Data mining with AI techniques, such as machine learning, can be a useful tool for market research in this field because it provides valuable information for the design of effective digital marketing strategies. This study examines how predictive information systems based on machine learning can improve the efficiency and competitiveness of e-commerce marketing. Using data from 5,389,731 users in Spain, we implement XGBoost to estimate purchase willingness across seven categories (insurance, hearing aids, NGOs, energy, gambling, telecommunications and finance). The supervised model is trained on historical conversions and behavioural–demographic variables, validated with holdout data, and deployed to score non-converted users. To complement algorithmic profiling with managerial insight, we conducted 63 qualitative interviews to uncover motivations by segment and category. XGBoost accurately predicts purchase likelihood when there is sufficient conversion depth, enabling granular segmentation and targeted promotion with reduced waste. The qualitative findings reveal distinct intrinsic and extrinsic drivers by product and declared sex, informing message framing and offer design. Taken together, the mixed-methods approach supports more efficient resource allocation, higher conversion rates and strengthened competitive positioning through data-driven segmentation and motivation-based personalisation. The contribution is twofold: a replicable, at-scale lead-scoring pipeline and actionable motivational maps for planning. Accordingly, this is an innovative, AI-enabled management method that increases efficiency, improves competitiveness and supports advantages in key processes, with multinational portability.

Keywords: *Data mining, willingness to buy, motivations, XGBoost, social media, Supervised algorithm, Lead generation*

JEL Classification: M31, M15, C89

Article history: Received: April 2024; Accepted: December 2025; Published: December 2025

1. INTRODUCTION

Internet and social networks have created a rapidly expanding digital social ecosystem in which users can search for, consume, and exchange information instantly with other users located anywhere in the world (Ahmed et al., 2022; Shu et al., 2017). In 2022, around 90% of the European population were Internet users and more than 67% had Facebook accounts. In the case of Spain, the figures are slightly above average, with internet penetration amounting to approximately 92% and more than 76% of the population being Facebook users (Internet World Stats, 2022). As a consequence, traditional business had no choice but to adapt to this new digital ecosystem, developing marketing and communication strategies through online platforms (Stone & Woodcock, 2013). By the end of the 2010s, e-commerce accounted for

https://doi.org/10.7441/joc.2025.04.04

more than 20% of global retail sales, thanks to the numerous advantages it offers, such as 24/7 availability, broad geographical reach, and relatively low investment costs (Pan & Zhou, 2020). Nevertheless, while global e-commerce sales exceeded USD 5.5 trillion in 2022, with an average growth rate of 20% over the previous three years, projections for 2025 are beginning to suggest signs of slowdown, with sales estimated at USD 7.3 trillion, and a growth rate of 10% for the next three years (Cramer-Flood, 2022). This trend is also evident in Asia. For example, sales among the most experienced e-commerce companies in China are following an "inverted U" pattern and are currently entering a phase of slowing growth and maturity (Ma & Liang, 2021).

In the deceleration and maturity phases, literature such as the life cycle model proposed by Quinn and Cameron (1983) recommend that communication and promotion actions should be intensified in an attempt to maintain market share and positioning. In the digital environment, this implies a greater presence on social networks and enhanced web positioning (Ahmadi et al., 2024; Szymański, 2021). In other words, in line with Varian's (2000, p.137) premise that "technology changes, economic laws do not", actions in e-commerce should pursue the same principles of incentivization and optimization of limited resources as offline commerce does. Therefore, whatever new channels and platforms might emerge, the need for companies to maximize profits and satisfy their consumers in a context of limited resources remains the same.

As the maturity phase requires greater efficiency of communication, there needs to be more awareness of consumers' needs and desires at all times, in order to align technological resources and adhere to economic principles. This can be achieved, for example, by making accurate product and service recommendations and providing links to purchase them (Tucker, 2014). Online recommendations can also be used to suggest alternative and/or complementary products, thus increasing cross-selling (Jin & Su, 2009). However, as Bolls, Muehling and Yoon (2003) argue, there is also the danger of overloading consumers with information, which complicates the decision-making process. To avoid such saturation, Sharma et al. (2021) suggest that online recommendations should be based more on consumer characteristics than on the type of product offered, to make the tailoring of incentives and promotions to consumer tastes more effective (Ahmadi et al., 2024; Jiang, Shang, Liu & May, 2015).

Big Data mining and Artificial Intelligence (AI) algorithms are often used to tailor recommendations to consumer profiles in the digital ecosystem, (Hartmann et al., 2019). Diebold (2003) defines Big Data as a phenomenon resulting from the boom of huge amounts of data generated by technological advances in recording and storage. Two decades later, the European Parliament proposed the following definition: "The term big data refers to collected data sets that are so large and complex that they require new technologies, such as artificial intelligence, to process them. The data comes from many different sources" (European Parliament, 2023). According to Kar and Dwivedi (2020), the use of large data sets, whether structured or unstructured, and whether from one or multiple platforms, and the need to use computational science methods to process them is changing the nature of research on information-generating systems. That is, AI applications are needed to perform data mining and to uncover hidden patterns of behaviour and relationships (Kannadath et al., 2018). Despite the aforesaid European Parliament definition of AI, we only consider AI applications, which are defined by Davenport et al. (2020) as the various algorithms, systems and devices that are capable of obtaining information, learning from data, performing analysis and generating results that assist in making informed decisions.

This study draws on previous research that has proposed the use of AI algorithms to analyse Big Data from the digital ecosystem to extract consumer segments or profiles, to which messages, offers and promotions personalised to the needs of the moment are sent (Dwivedi et al., 2023; Ismail et al., 2015; Kshetri et al., 2023). Fine-tuned segmentation is considered to improve marketing management by targeting actions at consumers who are interested in products rather than wasting money on those who are not (Johnson, 2013). For example, Joung & Kim (2023) propose the use of tools from the field of interpretive machine learning to segment the market based on the product attributes that have been rated highest in online reviews. They test this by means of a case study of a handheld device, and conclude that this methodology achieved more accurate segmentation of the market than previous procedures. In e-commerce, Song & Liu (2020) used an AI algorithm called XGBoost (Extreme Gradient Boosting) to predict consumer purchase behaviour with a sample of 12,330 users and compared the result with a random forest algorithm. Their findings indicate that XGBoost offered a better fit for predicting e-commerce platform users' shopping behaviour and performance. In the same line of comparative studies, Sáez-Ortuño et al. (2023) contrast two AI algorithms (XGBoost vs K-Means), the former supervised and the latter unsupervised, with a sample of more than five million users. Their results show that XGBoost produces better results with structured data.

This study has two objectives. The first is to illustrate the performance of XGBoost in profiling and locating potential buyers from a Big Data set of 5,389,731 unique users captured online. Specifically, the XGBoost algorithm will be used to estimate the willingness to buy seven generic online products (insurance, hearing aids, NGOs, energy distributors, gambling, telecommunications and finance). The second is to explore the motivations that influence purchase decisions for each of these products. In particular, although AI algorithms can be used to profile and locate potential buyers, an in-depth understanding of their motivations is not yet possible, hence the need to resort to traditional qualitative research (Sáez-Ortuño et al., 2023d).

To achieve these objectives, two methodologies are proposed: (1) a description of a real-life case study of the implementation of the XGBoost model, which was trained and tested with Big Data, to estimate the profiles of potential consumers according to declared sex and their willingness to buy the seven products; and (2) a qualitative study consisting of 63 in-depth interviews using a semi-structured questionnaire to find out the motivations to purchase the seven considered products. The author states the aim of the research topic, its focus, explains its originality, and introduces its structure.

2. THEORETICAL BACKGROUND

2.1 E-commerce Recommendation and Promotion Systems

E-commerce enables consumers to transact through digital platforms, which facilitate the exchange of information and payment, and courier services that handle the operational delivery of products (Pan & Zhou, 2020). The literature describes how internet users become consumers through the conversion funnel model (Hoban & Bucklin, 2015; Li & Kannan, 2013). This model considers that the user-to-customer path passes through four stages: non-visitor, visitor, authenticated user and converted customer. This model can also be used for promotion management at different points in time, such as for estimating and predicting conversion rates in each of the strata (Batra & Keller, 2016).

Analogous to offline commerce, as e-commerce progresses through the launch and growth stages, competition intensifies and the marketing focus shifts from the initial product-centric approach to a greater focus on customer experience to maintain the market position (Wigand et al., 2008). The maturity phase entails more intense competition and requires the modification of marketing strategies toward segmentation, targeting and positioning, placing greater emphasis on personalised communication and the promotion of offers, both to retain customers

https://doi.org/10.7441/joc.2025.04.04

and to attract the attention of potential ones (Quinn & Cameron, 1983; Ma & Liang, 2021; Szymański, 2021).

Meanwhile, online market research, unlike the offline environment, can use machine learning tools to study customer behaviour, propose target segments, and predict the best time for promotion (Ahmadi et al., 2024; Gupta et al., 2020). Once the target audience has been profiled and the time is right, personalised recommendations and promotions can be sent from different e-tailers that match their tastes (Li & Karahanna, 2015). Offer personalization facilitates user decision-making by reducing the costs of searching for information and making choices (Lo & Chien, 2024; Thirumalai & Sinha, 2009), and also improves attitudes towards the product (Blom & Monk, 2003). It has also been shown that messages accompanying recommendations are more effective when they refer to experiences with which the receiver can identify in some way (Ahmadi et al., 2024; Petty et al., 2000).

According to Adomavicius and Tuzhilin (2005), the agent's recommendation and promotion process to a potential customer follows three stages: 1) locating, identifying, and understanding the consumer to collect data and create profiles of their willingness to buy different products; 2) communication and personalized recommendation, which consists of adjusting the argument and recommended purchase based on their willingness to buy; and 3) results and feedback, which consists of estimating the impact of personalized recommendations and readjusting the strategy to improve the process. This research illustrates the first stage of the recommendation and promotion process based on implicit collection of data on the characteristics of previous buyers.

The first step of the first stage involves two basic methods: explicit and implicit (Lavie et al., 2010). In explicit methods, the market researcher approaches the consumer directly through an information collection protocol (e.g., questionnaire), while in implicit methods the researcher deduces preferences from the analysis of consumer behaviour (e.g., study of the characteristics of previous buyers). Both methods have their pros and cons. For example, the literature has gathered evidence that users are not always truthful when filling in questionnaires or providing their classification data (Sáez-Ortuño et al., 2023b), while others do not like providing personal information to third parties, in particular commercial organizations (Schiaffino & Amandi, 2004). The three most popular implicit methods are estimates derived from clickstreams, from the time spent observing the product on the web, and information from other past consumers (Ahmadi et al., 2024; Li & Karahanna, 2015). Some authors like Lavie et al. (2010) raise doubts about the feasibility of inferring purchase propensity from such indirect cues as clickstreams or the time spent observing an item on a website. Comparing implicit and explicit methods, Billsus et al. (2002), in a study on mobile applications, consider that although implicit methods are not the universal solution, it is much better to infer the user's profile from observation of their actions than from their answers to surveys due to the cognitive load it imposes on users. On the other hand, in a study on advertising via Short Messaging Services (SMS), it was observed that implicit methods were considered riskier than explicit ones in terms of safeguarding privacy (Xu & Du, 2011). Finally, Liang et al. (2006) recommend the use of both methods to ensure more accurate results.

Once the information has been collected, which is considered Big Data if there is a very large amount of it, the next step is to construct user profiles and then estimate their willingness to buy a series of products (Adomavicius & Tuzhilin, 2005). Since our study proposes an implicit methodology based on the behaviour of previous buyers, it is important to have demographic, geographic, social network usage and other information about previous customers in order to configure consumer profiles and extrapolate them to the rest of the population. The literature proposes a wide variety of algorithms to analyse structured data (i.e. from questionnaires), including genetic algorithms, Markov models, Bayesian networks, and others. (Massaro et al., 2021). In this study, the XGBoost algorithm is proposed (Sandoval, 2017).

2.2 The XGBoost Algorithm as a Tool for Data Mining

There is a growing consensus that greater efficiency can be achieved in marketing management if managers base their decisions on the results generated by data analytics (Li, 2022). For example, in the field of e-commerce, academia has proposed various criteria for segmenting consumers. Ahmadi et al. (2024) classified them into four: keywords in sponsored search advertising; location or timing; contextual factors (i.e. the degree of information disclosure); and factors related to behaviour and demographics.

Given the propensity of online environments to generate large volumes of data (Sandoval, 2017), the use of machine learning algorithms is recommended for their analysis (Athanasiou & Maragoudakis, 2017). Big Data uses these algorithms to search for similarities by following certain statistical laws, with the aim of establishing a model of relationships in the vector space that enables its optimisation (Li, 2020). Algorithms can be supervised or unsupervised. The former seek to optimise an objective by training the algorithm, while unsupervised algorithms can find clustering patterns, but without a specific objective (Sandoval, 2017). For example, Sáez-Ortuño et al. (2023c) used a structured database formed by behavioural and demographic variables to compare the degree of fit of supervised (XGBoost) and unsupervised (K-Means) algorithms when segmenting an e-commerce market to predict sales, with the former performing better (Sáez-Ortuño et al., 2023a). Several other comparative studies have found that XGBoost is more efficient than other algorithms, such as random forest or gradient boosting (Chen & Wan, 2023; Liang et al., 2019).

The XGBoost algorithm is based on decision trees and involves sequentially and recursively classifying a set of random and unordered data, following certain criteria set by the researcher (Kannadath et al., 2018). Commonly used with big data, the algorithm has good tolerance to a small number of missing values and automatically learns the optimal direction for splitting decision trees by means of a sparse distribution function (Chen & Wan, 2023). The main advantages are that (1) it uses a second-order loss function, which is more accurate than first-order fits; (2) it introduces a correction element, the regular term, to avoid over-fitting; and (3) it produces very accurate results. However, it has some disadvantages: (1) the algorithm needs to traverse all the data to organise the division of nodes, which consumes huge amounts of computer memory; (2) it only works with numeric data, so any non-numeric data must be converted to numbers; and (3) the way the data is labelled may condition the final result (Chen & Guestrin, 2016; Chen & Wan, 2023).

This is an algorithm that since its creation in 2014 has been widely used in different fields, such as in meteorology to make predictions about rainfall and in health to predict heart disease (Liu & Qiao, 2019), among others. It has also been used in the field of sales and e-commerce. For example, Song & Liu (2020) use XGBoost to predict consumers' purchase intentions on e-commerce platforms. They used a database of 12,330 users with information from 14 vectors (10 behavioural and 4 demographic), including behavioural data such as number of pages visited, total time spent accessing pages, etc. (Song & Liu, 2020). In another study, Li (2022) uses 13 indicators, the result of combining behavioural data (browsing times by categories, shopping times, browsing times by brands, viewing time of the last brand and viewing time of the last category) and demographic data (gender, age, marital status, education, occupation, etc.), to improve the effect of marketing communication. Our study also combines behavioural and demographic data, but adds the new feature of estimating the willingness to buy seven generic products, based on the declared profiles of shoppers, and analysing a data base of more than five million unique users.

XGBoost is a machine learning algorithm that generates a sequence of weak predictions made by decision trees. It also uses a loss function, which must be minimised at each iteration to reduce deviation and variance (Chen & Guestrin, 2016; Chen & Wan, 2023; Kannadath et al., 2018). Deviation measures the difference between the value estimated by the model and the actual value. Reduction in variance helps to prevent what is called overfitting, and is generally achieved by introducing regular terms to the model and reducing its complexity. Equation 1 shows how the objective function J is a grouping of additive functions formed by the loss function and the regular term:

$$J = \sum_{n}^{i=1} l(y_i, y_i) + \sum_{t}^{i=1} \Omega(f_i) \qquad (1)$$

The first term, function *l*, is the loss function that measures the deviation of the estimated value, $\hat{y}_i$, from the target value, y_i. Meanwhile, the second term, function Ω, is a regularization function that helps to moderate the value of the final estimators and avoid overfitting (Chen & Guestrin, 2016). Therefore, the second term helps to make the model more parsimonious, leading to a better balance between model complexity and predictive ability (Bagozzi & Yi, 1989).

Given the iterative nature of the algorithm, the estimated value at each step will be determined by the estimated value at the previous step and the error made. Consequently, the model defined in Equation 1 must incorporate this condition that modifies the second element of the first term (XGBoost, 2022). That is, by substituting $\hat{y}_i$ for $\hat{y}_i^{(t-1)} + f_t(xi)$. The final expression of the objective function is shown in Equation 2:

$$J = \sum_{n}^{i=1} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \qquad (2)$$

The regularization function Ω can be expanded, as illustrated by Equation 3, and presents an additive function formed by two terms: T is the number of leaves of the decision tree and w_j^2 is the squared weight of leaf j. Hence, this function estimates the complexity of the decision tree.

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

Due to the complexity of the objective function J of Equation 2 and since it is differentiable, the Taylor polynomial can be applied for its resolution. The XGBoost algorithm uses a second-degree polynomial approximation where *gi =∂ y(t−1)l(yi,ˆ y(t−1))* and *hi=∂2 ˆ y(t−1) l(yi,ˆ y(t−1)),* that is, the first and second derivative of the first term of Equation 1. Consequently, the objective function after the tree split is the following:

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

Data labelling is a crucial step in supervised learning, as training data must be manually classified into the desired categories. Consequently, if training data is not correctly labelled (i.e. partially or not in a manner representative of the actual class distribution) it can generate biased results (Dixon et al., 2018; Mohammed et al., 2022). To minimise bias, a meticulous process of documenting and verifying the quality of the generated labels must be followed. For example, Wang et al. (2017) claim that tagging by several human annotators will minimize individual subjectivity and that tag assignment guidelines should be agreed upon.

RQ1: How does data mining, using algorithms like XGBoost, estimate the inclination to buy different products online in a structured database consisting of behavioural and demographic data?

2.3. Motivations to Purchase Different Products Online

Motivation has been widely studied from different theoretical perspectives. Mitchell (1997) defined it as the key factor that drives an individual to act in a certain way and influences their intensity, direction, and persistence in achieving their goals. The theoretical foundations were laid in the 1950s. One of the pioneers, Hull (1952), proposed the existence of motivations of an instinctive or animal nature, describing the motivation process as a combination of drive and homeostasis. According to this author, homeostasis is a state of internal balance that is altered when the individual experiences need, at which point the drive (the motivator) acts by pushing the individual to restore that internal balance. Shortly afterwards, the question arose as to whether motivations arise only from objective reality or also depend on the individual's subjective interpretation of that reality. Rotter (1954) introduced the concept of goal attainment expectations, thus laying the foundations for subsequent theoretical developments. The next step was to clarify the nature and origin of motivations. Heider (1958) proposed the existence of internal and external motivational forces. He called the internal ones "dispositional," linking their emergence to the cognitive and physical effort expended by the individual to achieve their goals, and called the external ones "situational," relating them to the influence of the surrounding environment that may facilitate or hinder such achievement. Rotter (1954) and Heider (1958) significantly contributed to the progressive development of a more comprehensive theoretical framework on motivation. These concepts were the first to be applied to the study of consumer behaviour (Rossiter & Foxall, 2008), and although later authors expanded and improved upon them, the foundations established in the 1950s remain fundamental for our understanding of behavioural drivers.

Extensive literature has investigated specific online shopping motivations, analysing them either from a general, multi-sector perspective (Rohm & Swaminathan, 2004; Goldsmith & Horowitz, 2006) or in specific niches, for example, in sectors such as wine (Bruwer & Wood 2005) or food delivery services (Yeo, Goh & Rezaei, 2017). However, comparative studies analysing the purchase motivations for different types of products or sectors are very scarce. A valuable exception is the study by Rohm and Swaminathan (2004), in which the authors present a two-phase methodological approach. In the first phase, they classified consumer motivators into four broad categories: general convenience; preference for physical stores; use of information for planning and carrying out purchases; and variety seeking. In the second stage, they segmented the market by motivations and assigned the resulting segments to ten different product classes, ranging from books to financial services (Rohm & Swaminathan, 2004). Ultimately, there is still a lack of research that simultaneously compares the purchase motivations of various products purchased over the Internet.

An in-depth understanding of online consumer motivations is essential for marketers to formulate effective digital communication and advertising strategies that meet the needs and desires of shoppers in each sector. This knowledge is also basic for designing engaging online shopping experiences and building customer loyalty. Based on the evidence gathered, this study poses the following research question:

RQ2: Are there different purchasing motivations for each category of online product?

2.4. Differences by Declared Sex

The propensity to make online purchases of different products, and the motivations to do so, can vary significantly depending on the consumer's sex or gender identity. It is therefore critical for brands to understand these differences in order to design effective marketing strategies. According to Mayers-Levy and Loken (2015), sex is determined by the biological characteristics (hormonal and brain structures) that distinguish males from females, while gender identity depends on the degree to which someone perceives themselves as masculine or feminine. Although both sex and gender identity are used as criteria for market segmentation, Níquel et al. (2020) suggest an adaptation by product type. They propose the use of declared sex as a criterion to segment markets for convenience products (with little involvement) and gender identity for specialty or preference products, which are those that contribute to the configuration of identity or status signalling (Forgas-Coll et al., 2023).

In this study, given the characteristics of the promoted products, basically convenience products, the classification is based on declared sex. The main reason for doing this is the hypothesis of selectivity (Moss, 2017; Forgas-Coll et al., 2023), according to which consumers of different sexes demonstrate certain preferences and tastes, and find different types of images, shapes, and commercials attractive (Nickel, Orth & Kumar, 2020). For Darley & Smith (1995), the difference in preferences between males and females of the human species comes down to the fact that they use different strategies to gather and process information from the environment. While females process information captured by the senses holistically, interrelating the different elements of the environment to configure a complete picture, males tend to process pieces of information selectively, focusing their attention on a limited number of elements, and making an induction about the whole (Darley & Smith, 1995). This also affects purchase decision processes, different degrees of sociability and the ways in which males and females react to different group situations (Baumeister & Sommer, 1997). While females tend to focus on dyadic ties, on individual relationships, males tend to focus on group-based social structures. This suggests that female consumers are more loyal to specific sellers and extrapolate that loyalty to the brand, while males do the opposite, and are more loyal to the brand, which they perceive as loyalty to the seller (Melnyk, Van Osselaer & Bijmolt, 2009).

Based on the evidence collected on the different ways that males and females value experiences and establish social bonds, this study poses the following research questions:

RQ3: Does the propensity to make online purchases vary according to declared sex for different products?

RQ4: Are there differences in motivations by declared sex?

3. METHODOLOGY, RESULTS AND DISCUSSION

To illustrate how the XGBoost algorithm estimates the willingness to purchase seven generic products by declared sex, and to provide answers about the motivations for purchasing these products online, two studies are proposed: a case study and a qualitative one. Specifically, the seven products are: insurance, hearing aids, NGOs, energy suppliers, gambling, telecommunications, and finance. These are described in greater detail below.

The product classed as "insurance" includes life insurance, medical insurance, and so on. In other words, these products help people to address the possibility of poor results in the future (Wärneryd, 1999). NGO refers to financial donations made by consumers to non-governmental organizations. However, some authors have criticized how the definition of NGOs is based on something that they are not, as they include a wide variety of organizations grouped into a catch-all category that is difficult to define precisely (Gray et al., 2006). "Hearing aids" are devices to correct mild or moderate hearing loss, and a wide variety of technologies and brands are considered (Goman & Lin, 2016). The product "Telco" refers to a variety of telecommunications services, including telephone services, fibre optic and mobile internet, cloud services, and others (Wei & Chiu, 2002). "Gambling" refers to different products that are only authorized for adults (Lorains et al., 2011). The "energy" product refers to both energy distribution and energy trading companies targeting final consumers. However, according to an EIA estimate (2020), residential consumption accounts for 21.69% of global energy. The "finance" product is a generic term for saving, investment, and financing products provided by a variety of service companies, financial institutions, and banks (Ramaswami et al., 2000).

https://doi.org/10.7441/joc.2025.04.04

3.1. Study 1

With the assistance of a leading Spanish lead generation agency, CoRegistros, S.L.U., we were able to estimate willingness to buy seven generic products. We did so using a sample of 5,389,731 users (36% men and 64% women), captured between 2010 and 2022 in Spain. CoRegistros collects data online by advertising sweepstakes and challenges, including quizzes on such topics as history, geography, and cooking. The participants must provide their data to get the chance to win substantial prizes (for example, an iPhone or an Alexa Echo Dot). After filtering for false information, this data is provided to third parties for highly targeted advertising campaigns. Before using any of this data, we made sure that the company complies with the European Data Protection Regulation and the corresponding Spanish LOPD-RGPD, "Organic Law 3/2018, of December 5, on Personal Data Protection and Guarantee of Digital Rights" (AEPD, 2018).

3.1.1. Data set

The Big Data provided by the company contained structured demographic and behavioural data taken from the information provided by participants in order to register, as well as from client advertising companies. It consists of two data matrices grouped into eight blocks forming 37 columns. Matrix 1 contains five blocks of descriptive data: 1) Users, user identification data, 2) Marketing, communication and promotion actions through which the information was collected, 3) Conversions, captured users who, after receiving recommendations from the sponsors, became conversions, 4) Advertisements, campaigns carried out by the sponsors to recommend their products, and 5) Sweepstakes, method for collecting information by means of sweepstakes. Matrix 2 is made up of three blocks of marketing specifications: 1) Ad type, type of advertising campaign that specific users received, 2) Sweepstake type, the type of product that sweepstake entrants could potentially win (beauty, electronics, home, iPhone, leisure and travel) and 3) Conversions, users who ended up purchasing one of the promoted products (hearing aids, energy, finance, gambling, NGOs, insurance and telcos).

Matrices are made up of vector data of different types, such as string, Boolean, float and interval. String data represent sequences of characters (e.g. names of people) and are recorded in inverted commas. Booleans can only have two values (0, 1) and usually refer to whether or not a condition is met: 1 (true) and 0 (false). Floats are numerical data that allow positive as well as negative decimal values (e.g. geographic coordinates). Finally, intervals are an immutable list of sequentially ordered numbers (e.g. Likert scales).

Table I describes the five blocks identified in Matrix 1 and briefly explains the data contained in each. In turn, Table II shows the three blocks of commercial and conversion actions in Matrix 2, as well as a brief description of each.

Tab.1 - Description of the Blocks Contained in Matrix 1. Source: own search

Database Block Name	*Description of the Content*
1. Users	User master. Contains all fields with descriptive information about the user.
2. Marketing	Master of marketing campaigns by means of which users are registered. It relates to the users table through id_m.
3. Conversions	Conversions master. Contains the history of users who have converted in the past to some product. It relates to the users table through id_user.

4. Ads	Master of client campaigns. These campaigns are sent to users who are registered in the database to persuade them to convert to the offered product. It relates to the conversions table through id_ad.
5.Sweepstakes	Sweepstakes master. It relates to the sweepstakes column of the users table through the id_prom column. It also relates to the marketing table through id_prom, since each marketing campaign is assigned a sweepstake (the same sweepstake can be assigned to different marketing campaigns).

Tab. 2 - Description of the Auxiliary Tables for Matrix 2. Source: own search

Database Block Name	***Description of Block Content***
1. Ad type	Upon analysing the ads table, the need was identified to understand the description of the different types of campaign. To address this issue, the ads_type.csv file was created as a master list of campaign descriptions (with a tab as the separator). This file is related to the ads table through ad_type.
2. Sweepstakes	Upon analysing the sweepstakes table, the need was identified to classify the sweepstakes according to the product offered as a prize. To solve this problem, the clasificacion_sorteos.csv file was created (with a tab as the separator). This file is related to the sweepstakes table through id_prom. The created categories are: beauty, content, electronics, home, iPhone, leisure, test, and travel.
3. Conversions	Upon analysing the conversions table, the need was identified to classify the client campaigns (id_ad) that appear in said table (i.e., campaigns that have resulted in at least one conversion) according to the final product each user converted to. To solve this problem, the clasificacion_conversions.csv file was created. This file is related to the conversions and ads tables through id_ad. The created categories are: hearing aids, energy, finance, gambling, NGOs, insurance, and telcos.

Finally, the data were completed with information from official sources, such as the geographical longitude and latitude as determined from the postcode.

The data were received already clean and ready to be used (a description of the cleaning process can be found in Sáez-Ortuño et al., 2023b). However, of the 37 columns of data, the XGBoost algorithm was only applied to 25, namely 24 predictors and one of conversions to purchase, the latter being the target variable. Table III presents a list of the selected variables and the type of scale (string, Boolean, and interval).

Tab. 3 - List of Columns Used in the Study. Source: own search

Index	**Column**	**Type**

2	id_producto_conv	Int (*)
3	id_user	Int (*)
8	id_sexo	Bool
10	edad	Int
12	latitude	Float
13	longitude	Float
17	valido	Bool
18	finaliza	Bool
19	espactividad	Bool
20	estado_telf	Bool
23	dominio_email_gmail	Bool
24	dominio_email_hotmail	Bool
25	dominio_email_outlook	Bool
26	dominio_email_yahoo	Bool
27	dominio_email_live	Bool
28	dominio_email_msn	Bool
29	dominio_email_otros	Bool
30	cla_sorteo_belleza	Bool
31	cla_sorteo_contenido	Bool
32	cla_sorteo_electronica	Bool
33	cla_sorteo_hogar	Bool
34	cla_sorteo_iphone	Bool
35	cla_sorteo_ocio	Bool
36	cla_sorteo_test	Bool
37	cla_sorteo_viajes	Bool

3.1.2 Estimation Procedure of the Willingness to Purchase Using the Supervised XGBoost Algorithm

As shown in Table 3, of the twenty-five vectors, twenty consist of Boolean data, three of intervals and two of floats. The numerical data were normalized, and the database was divided into three parts:

1. The training database consists of a vector matrix formed by 90% of the users who ended up buying one of the promoted products, called conversions (23,050 records). This phase uses real data from past conversions on the products sponsored by the clients (marketing companies) (Bishop, 2006). With this information, the algorithm is trained to look for communalities in the distribution of decision trees on the independent variable data (behavioural and demographic) to predict Y_train (the vector of 23,050 records converted into buyers) from the matrix X_train (23 columns of 23,050 independent variable records). That is, the vector (Y_train) is set as the target, and the XGBoost algorithm builds a sequence of decision trees from the data in the matrix (X_train) until it finds the distribution that best fits the target vector (Goodfellow et al., 2016). The mean absolute percentage error is used as an estimator of goodness of fit. With this database, the XGBoost algorithm is trained to predict Y_train (1 vector of specific conversions of 23,050 records) from the X_train matrix (23 columns of 23,050 records).

2. The validation database. This database, consisting of the remaining 10% of users who became buyers, is used to test the predictive capacity of the communalities model. The objective is to use the model to estimate the vector y_test (the vector of dependent variables formed by the

specific conversions of 2,562 records) from the matrix X_test (23 columns of 2,562 records) and estimate its goodness of fit.

3. The estimation basis. Once the algorithm has been trained to form the communalities model, and subsequently tested, it is used to estimate the willingness to buy of the remaining users. That is, the model is applied to the matrix of independent variables, X_predict (23 columns of 5,364,119 non-conversion records), to estimate the probability of specific conversions, Y_predict (1 vector of estimates of specific conversions of 5,364,119 records). The resulting probabilities are considered an estimator of the willingness to buy the different products.

The operational procedure consisted of using the XGBClassifier library, from the XGBoost Python package, to train the algorithm (Chen & Guestrin, 2016; XGBoost, 2022).

3.1.3 Analysis of Results

The results of the training phase are shown in Table 4, which presents the frequencies of actual values by type of product, the frequencies estimated by the algorithm after running the decision trees many times (Y_train), and the error percentage. The estimated frequencies are fairly similar to the real ones, with the insurance product having the highest frequency, followed by gambling and hearing aids. The remaining products have lower frequencies, with finance being the most difficult to promote. The goodness of fit of the model, measured in percentage error, varies between 6-12%.

Tab. 4 - Frequencies of Conversions and Percentage of Correct Predictions by Product. Source: own search

Product	Total Frequencies	y_train Frequencies	Error (%)
Insurance	13,286	11,938	10.15
Gambling	5,447	4,885	10.32
Hearing Aids	5,363	4,848	9.60
NGOs	831	757	8.90
Energy	318	286	10.06
Telcos	251	234	6.77
Finance	116	102	12.07

Subsequently, once the model was "trained," the algorithm was applied to the validation basis to estimate its predictive capacity. However, at this phase it is possible to manually adjust two parameters of the algorithm (number of iterations and depth of the decision tree) to fine-tune the goodness of fit. In this case, different combinations of numbers of iterations (50, 100, 200, 500, 750, 1,000) and tree depth (4, 6, 8, 10, 15, 20 branches) were tested, and the combination of 100 iterations and 8 branches of tree depth was chosen. However, as shown in Table 5, this combination generates a good fit in the two highest frequency products (insurance and gambling), with errors below five percent, a somewhat higher error in the third product (hearing aids), and spikes in the lower frequency products. Consequently, the weight of the parameters of the NGO, Energy, Telecommunications, and Finance products had to be adjusted separately using the Python library class_weight. The calculation performed by this subroutine is as follows (XGBoost, 2022):

product_weight_i = (y_train) / (number_of_products · Product_frequency_i ∈ y_train)

Tab. 5 - Frequencies of Conversions and Accuracy Percentages by Product. Source: own search

Product	Estimation frequencies y_test	Frequencies y_test	Error (%)
(6) Insurance	1362	1348	1,03
(4) Gambling	588	562	4,42
(1) Hearing aids	599	515	14,02
(5) NGO	6	74	1133,33
(2) Energy	2	32	1500,00
(7) Telcos	1	17	1600,00
(3) Finance	4	14	250,00

Once the XGBoost algorithm had been trained, tested, and adjusted, it was applied to the database of non-converted users, X_predict, to estimate the willingness to buy the sponsored products, Y_predict. This is an estimate of the likelihood of each potential customer (who has not purchased any product) becoming a buyer of each of the said products. Table 6 shows the frequency of subjects with the highest probability of purchasing each of the products, as well as their division by declared sex.

Tab. 6 - Estimates of the Willingness to Purchase by Declared Sex. Source: own search

Product	Number of individuals inclined to acquire the product	Men	Women
Gambling	1,048,575	483,382	565,193
Insurance	1,045,429	472,888	572,541
Hearing aids	1,004,323	305,414	698,909
NGOs	587,199	133,555	453,644
Energy	363,922	76,429	287,493
Telecommunications	361,838	70,309	291,529
Finance	84,042	42,257	41,785
Total Base	**5,364,119**	**1,931,083**	**3,433,036**

Finally, we estimated whether there were significant differences by declared sex (36% men and 64% women) when comparing proportions using Chi-squared by type of product. The findings suggest that men predominate in gambling ($\chi^2(2)$ = 46.416, p = 0.000), insurance ($\chi^2(2)$ = 38.804, p = 0.000), and finance ($\chi^2(2)$ = 7.439, p = 0.000), whereas women predominate in hearing aids ($\chi^2(2)$ = 9.871, p = 0.000), NGOs ($\chi^2(2)$ = 44.781, p = 0.000), energy ($\chi^2(2)$ = 46.416, p = 0.000), and telecommunications ($\chi^2(2)$ = 43.114, p = 0.000).

3.1.4 Discussion

The application of the XGBoost algorithm to the study of consumer behaviour has proven useful for revealing behavioural patterns and predicting purchasing trends across different product categories. Data on previous conversions, types of engagement and reactions to promotions, supplemented with demographic information, were sufficient to set up a frequency distribution model with predictive capabilities. While the combination of the XGBoost algorithm and the database generated reasonable error rates in the training phase (mean: 9.69, standard deviation: 1.61), those rates soared in the validation phase (mean: 643.25, standard deviation: 737.19).

https://doi.org/10.7441/joc.2025.04.04

Reliable goodness-of-fit values of between 1.03 and 14.02% were only achieved in this phase for products with a substantial amount of available conversion data (insurance, gambling and hearing aids). Hence, the XGBoost algorithm only generates reliable results with validation samples above 500 observations, and cannot validate predictions with very small samples (Chen & Guestrin, 2016; Liang et al., 2019). Thus, although the estimates of willingness to buy insurance, games and hearing aids by reported sex insurance, games and hearing aids have been verified as reliable, those for NGOs, energy, telecommunications and finance have not.

Interestingly, although the estimation assigns the greatest willingness to buy to products with conversion rates over 1 million, the leading product (Insurance) is ranked in second position. In other words, although insurance has had the greatest promotional success in the past, the XGBoost algorithm estimates that the potential market will not purchase it in the same proportion as before. As Sáez-Ortuño et al. (2023c) point out, the degree of maturity of the product influences the ease of promoting it, making it easier to promote novel products in less saturated sectors. This is similar for the last three products: energy, telecommunications and finance. While the algorithm assigns a similar proportion to energy and telecommunications, it assigns a much lower proportion to finance than current consumption. However, in the latter cases, the likelihood of the forecast becoming reality is more uncertain.

Answering RQ1, this study shows how XGBoost is a useful tool to segment the market from structured data and with the usual limitations of any predictive model, particularly the lower predictive capacity when samples are small (Chen and Wan. 2023). These results are in line with previous studies that have highlighted the accuracy of XGBoost for classification and regression tasks (Liang et al., 2019).

Another relevant finding is that regarding RQ3, which inquired whether the propensity to make online purchases of different products varies according to declared sex, and our results confirm that it does, for men are more likely to acquire gambling products and women hearing aids. This agrees with previous studies that have found significant divergences in the propensity to buy online according to gender (Palan, 2001; Tifferet & Herstein, 2012). For example, the greater propensity of men to acquire gambling products and finance reflects gender roles and stereotypes, as argued by gender identity theory (Palan, 2001).

In conclusion, Study 1, a case study, illustrates the potential of XGBoost to analyse massive user data and generate accurate profiles of potential buyers, which has important applications for market segmentation and personalized digital marketing strategies.

3.2. Study 2

Study 2 aimed to explore online purchase motivations for seven generic products: gambling, insurance, hearing aids, NGOs, Energy, Telecommunications, and Finance. Defined as subjective drivers (Rotter, 1954), motivations can be intrinsic (driven by the consumers themselves) or extrinsic (generated by environmental stimuli) (Heider, 1958). Moreover, as suggested by Zaman et al. (2023), they can be multidimensional, and differ in terms of quantity and type.

To investigate the multidimensional and subjective nature of motivations, qualitative research was conducted by means of in-depth interviews based on a semi-structured questionnaire.

3.2.1. Data collection

Since the database used in the case study contains information on consumers who ended up buying the promoted products, we asked the lead capturing company for a sample of about 1400 users (roughly 200 per product). The target was to interview about seventy consumers, considering the acceptance rate of 5%, which is similar to Patton's (1990) intensity sampling. The recruiters were also asked to ensure that the sample had similar ratios in terms of sex (36%

men and 64% women) and Internet natives (under 40 years old) as well as people who have adapted to it (over 40 years old) (Ahuja, Gupta & Raman, 2003).

A recruitment team telephoned sample members to invite them to participate in a study of online purchasing motivations, and as an incentive they were offered a 50-euro gift voucher to spend in a local shopping centre. The sample size was determined according to semantic saturation criteria for gaming, insurance, hearing aids and NGO products, as recommended in the literature (e.g., Zaman et al., 2022; Sáez-Ortuño et al., 2023b). However, for energy, telecommunications and finance it was limited by the small number of buyers residing in Barcelona. In total, 1,117 phone calls were made, of which 475 were answered, and 144 candidates expressed their intention to participate in the study. Finally, 63 individuals (26 male (M) and 37 female (F)) between 21 and 70 years of age were selected (Table 7 shows the distribution by products and age cohorts). The interviews lasted between 25 and 40 minutes.

Tab. 7 - Distribution of the sample by products and age cohorts. Source: own search

	Gambling	Insurance	Hearing aids	NGO	Energy	Telcos	Finance	Total
21-30 years	2	0	0	2	0	2	0	6
31-40 years	4	3	0	4	2	4	2	19
41-50 years	2	4	0	3	3	2	3	17
51-60 years	1	3	5	1	2	1	1	14
61-70 years	0	0	6	0	1	0	0	7
Total	9	10	11	10	8	9	6	63

3.2.2. Data collection process

The interviews were conducted at a location in the centre of Barcelona that was easily accessible by public transport. As the interviewees arrived at the assigned times, they were welcomed and informed about the objectives of the study, as well as their rights when participating in a research study and were asked to sign a consent form. To contextualise the study, recall the purchase stimulus and avoid confusion, participants were shown various examples of advertisements of the product acquired and the platform from which it was purchased. The interview consisted of openly answering four questions, two contextualization questions and two questions about the objectives of the study, which were developed from previous studies (Ahuja et al., 2003; Zaman et al., 2022; Saez-Ortunño et al., 2023). The questions were:

1. How many times have you purchased products online in the last six months?
2. Do you remember when you purchased product X?
3. What were the reasons that led you to buy that product online?
4. Can you remember any other reason?

The interviews were conducted in Spanish and Catalan, and audio recordings were made.

Despite the recruitment team being asked for a balanced sample of non-native and native Internet users, a greater proportion were non-natives, 60% versus 40% (see Table 7).

The responses were transcribed and studied through thematic analysis, which consists of identifying and coding sentences or expressions that refer to the study object, in this case, the reasons that drive or justify a certain behaviour (Braun & Clarke, 2019). The coding followed a three-step process (Tuomi et al., 2021; Sáez-Ortuño et al., 2023b). First, the recorded responses were transcribed in the language used (Spanish or Catalan), read and highlighted. Second, the reasons were coded, resulting in between 8 and 24 codes depending on the product. Third, following the principle proposed by Bagozzi and Yi (1989) to maintain parsimony and operability, the codes were grouped based on the degree of similarity of meanings in a three-level pyramidal form. To promote internal validity, two different researchers read, highlighted, and coded the first three interviews separately (Zaman et al., 2023), and the degree of convergence was verified using Cohen's Kappa estimate, which produced values above 0.80 (Landis & Koch, 1977). Points of divergence were discussed until a consensus was reached. After the coding process, the reasons were divided into those common to all products and reasons specific to particular products. Finally, the researchers took into account the abundant literature on online purchasing motivations (Ahuja, Gupta & Raman, 2003; Rohm & Swaminathan, 2004; Goldsmith & Horowitz, 2006) to group the common reasons into: 1) convenience, linked to the ease of acquiring products or services "without leaving home"; 2) savings on transaction costs, such as the possibility of obtaining "better prices" and "saving time"; 3) availability of a wide offer, related to the possibility of accessing "a wide assortment"; or 4) problems with stock in physical stores, related to how "difficult it is to find" some products in local shops or supermarkets.

3.2.3. Analysis of Results and Discussion

The results on specific motivations by type of product are described below and some examples are shown in Table 8 to illustrate the labelling procedure.

The contextualisation questions helped the respondents to corroborate their recall of the purchase process. Regarding the frequency of online shopping in the last six months, the results differed between the two groups, with about three purchases in the older age group and more than six in the digital natives group. Regarding the recall question, given that the researchers have the specific dates, they serve as a means to check the degree of fit of their responses. In general, most respondents remembered the dates very closely.

Insurance. Since it was considered a general product, the respondents answered according to whether the insurance was life or health. For example, for health insurance, external reasons linked to the characteristics of the country where the interview was conducted were cited: insurance was acquired because of the "deterioration of public health in Spain" (Male, 58 years old), for the "possibility of obtaining a faster service" (Female, 34), and to be "more confident of being attended to in case of a medical emergency" (Female, 34). In addition to local reasons, efficiency and flexibility were also cited (Jadhav and Ramakrishna, 2023). With regard to life insurance, the main extrinsic argument was the existence of dependents (spouse, children, or other relatives), among both men and women, for example, to "protect their family" (Male, 45). On the other hand, men also consider an intrinsic motivation related to their "responsibility" (Male, 58) and "future planning" roles (Male, 37), while women talk about having "peace of mind" (Female, 39). Similar motives have been identified in the literature. For example, Mahdzan and Victorian (2013) talk about family protection, Guan et al. (2020) about responsibility, and Bhattacherjee (2002) about mental peace.

NGOs. Defining NGOs by what they are not creates operational complexity when selecting organizations (Gray et al., 2006). However, the participants in the study supported significant organizations (International Human Rights Foundation, Greenpeace Spain, Doctors Without Borders, etc.), so there was no confusion. In terms of motivations, intrinsic moral reasons based on the need to "do the right thing" (Male, 24), "religious beliefs" (Female, 40) and even "nationalist" motivations (Male, 21) were among those cited. Similar motivations were reported in the studies by Weaver et al. (2005) and Unerman and O'Dwyer (2006). Arguments linked to the interviewees' background and interests were also collected. For example, those who reported having a legal background expressed interest in organizations related to "human rights"

https://doi.org/10.7441/joc.2025.04.04

(Female, 32), while those with a background in natural sciences and biology were more interested in "environmental" causes and "the defence of the planet" (Female, 44). This convergence between backgrounds and interest in related causes has also been highlighted by Graafland and de Bakker (2021). Regarding declared sex, men participate due to a sense of "social responsibility" (Male, 33) and the "desire to change things" (Male, 24), while women more frequently mention "empathy" (Female, 32), the desire to "help others" (Female, 40) and a predisposition to support certain "social causes" (Female, 32) such as those linked to advocacy and women's rights. The argument of responsibility among men has also been pointed out by Einolf (2011). However, no extrinsic motivation has been expressed.

Hearing aids. This product is associated with hearing loss and ageing, whereby the selected participants were all over fifty years old (Dornhoffer et al., 2020). Although the interviewees acknowledge that their acquisition has improved their quality of life, such as being able to "understand my grandchildren" (Female, 62), which is an internal motivation, the stigma associated with their use linked to extrinsic causes often arises. Thus, they talk about the need to cope with shame and to "overcome challenges" (Female, 58). This stigmatization has been extensively addressed in the literature (references include Cienkowski & Pimentel, 2001; Gagné et al., 2011; Sebastian et al., 2015). For example, Cienkowski and Pimentel (2001) detect a stereotypical belief that people who wear hearing aids are less intelligent and have weak personalities. Even the users themselves express, consciously or unconsciously, a certain loss of self-esteem, embarrassment or lower self-efficacy (Gagné et al., 2011). Sebastian et al. (2015) find differences in the degree of acceptance of hearing aids among those who lost their hearing earlier compared to those who lost it later, with the former accepting them more than the latter. With regard to sex, in addition to "comfort" (Male, 59), men also express a greater interest in using the "most advanced technologies" (Male, 65), while women mention aesthetic arguments, such as the need "to fit me well" (Female, 67) and, above all, to be discrete (Female, 62). The literature indicates that women assign greater importance to the use of hearing aids to improve their social communication, and find all technical aspects effective as long as they do not negatively affect their physical appearance (Garstecki & Erler, 1999).

Telco. This product encompasses a wide variety of telecommunications services, telephone companies and brands, fibre optic and mobile Internet, cloud services, etc. (Wei & Chiu, 2002). Respondents for this product are mostly Internet natives, and the main reason for acquiring it is linked to external causes, such as a bad experience with the current service. For example, "shortcomings of the current company's service" (Male, 30), discrepancies with the price, "the monthly fee does not tally with the actual consumption" (Male, 37), and the "gradual increases in the fee without receiving any notice" (Male, 59). These reasons, disappointment and loss of trust, have been considered in the literature as the main reasons for switching companies (Buckinx & Van den Poel, 2005; Kim & Yoon, 2004). Although reasons linked to "price and coverage" (Female, 33) are proposed by both declared sexes, men report more motivations linked to service provision, such as "concern about data plans" (Male, 30), while women mention concepts such as "ease of connection with relatives" (Female, 33) and the need for systems to be "simple to use" (Female, 48). These results are in line with the classical literature on gender motivations regarding telecommunications products (Gefen & Straub, 1997).

Gambling. Gambling-related products, as well as their promotion or recommendation, arouse much controversy due to their connection with addiction. According to Lorains et al. (2011), gambling addition is characterized by the consumer's inability to resist the temptation to bet, which can lead to personal, family, and professional problems, as well as depression, anxiety, bankruptcy, and suicide. It is therefore a product that is intended for adults only and must include warnings about the dangers of addiction. The reasons given for purchasing this product are linked to intrinsic causes such as the thrill of gambling, the desire to "live the experience of winning" (Male, 32) or the "rush" (Male, 24) when this is achieved, and the "desire to win money" (Female, 36). Motivations linked with escapism are also mentioned, such as the "need to break the monotony" (Female, 36). These motivations are in line with those proposed by Lloyd et al. (2010), namely making money, excitement, intrinsic enjoyment of gambling, breaking routine and stress, and exploratory and competitive desires. Although both sexes view entertainment as an intrinsic motivation, some differences are perceived: men are motivated by adrenaline and quick money, while women mention breaking from routine and the possibility of winning big prizes.

Energy. The energy product refers to the contracting of a distribution service, which is an imperative need in developed societies (Jha et al., 2017). In Spain it is a mature and highly competitive market, and distributors make extensive use of communication and promotion to attract customers, mainly from their competitors. The purchase motives are broad, including intrinsic and extrinsic motivations, although they differ according to age. Younger respondents mention extrinsic environmental motivations, such as the company's commitment to "renewable energies" (Male, 32) or protection "of the environment" (Female, 41), while elders describe intrinsic factors linked to "price" and "energy saving" (Male, 62). There are also sex differences. Men express greater knowledge of tariff systems (hourly tariffs, difference between power and consumption, etc.) and therefore comment on their interest in obtaining "flexible options" (Male, 32). Women, on the other hand, cite service failures and disappointment with the current company as a reason for switching to another. For example, "I had a lot of problems with company E....." because "the breaker would trip when I was ironing" (Female, 47), and they pay a lot of attention to offers of a full service, as they are very interested in the company providing "electricity and heating for the home." These results present a high degree of consistency with those obtained by Neves and Oliveira (2021), who studied the motivations for purchasing heating equipment, finding that they are commonly related to energy savings, operational efficiency, maintenance and carbon footprint.

Finances. Financial products cover savings, investment and financing products offered by traditional and online banking and financial institutions. Only six interviewees had acquired savings and financing products, and none had purchased investment products. This is another mature product and the participants mainly cited extrinsic motivations for switching providers. There are differences between online and offline banking users. Younger users have already made the transition to digital banking, and therefore operate and work online. Those who used to work with offline banks made the switch to online banking due to failings of traditional banking services. For example, "I changed banks" because the "branches have closed" (Female, 31), "you waste the whole morning in huge queues" (Female, 34), "you need an appointment to solve the problem" and they ultimately "don’t solve it" (Female, 31). Differences also emerged between those who purchased savings and finance products. For those who purchased financial products from traditional banks, their main reason for switching service was "loss of trust in branch managers" (Male, 58). Although they acknowledge that the market has become "more insecure" (Male, 58), they still prefer "being attended to by a person" (Male, 58) and the "flexibility" of being able to adapt "to life's circumstances" (Male, 42). Those who acquire savings products mention "security" and "profitability" (Male, 40) as the most relevant motives. Some of these results are in line with those obtained by Ramaswami et al. (2000) on the intention to buy financial products online. Their study proposes three motivations: 1) conflicts with the current service; 2) familiarity with the online environment, and 3) experience of buying similar products online (Ramaswami et al., 2000). In terms of sex, both women and men claimed to have superficial knowledge, e.g., ‘I don't know how the financial world works’ (Female, 34) and highly value being informed about ‘the product they are going to buy’ (Male, 40). However, security, profitability and liquidity motivate men more, while women are more likely to value convenience and advice.

Tab. 8 - Examples of Reasons Cited by Men and Women by Product: Source: own search

Cluster	Main Motives - Men	Sample Quotation	Main Motives - Women	Sample Quotation
Insurance	Deterioration of public services, protect family, peace of mind, responsibility, foresight	"I want to ensure my family's future is secure, no matter what happens." - Male, 45	Efficient service, protect family, peace of mind	"I need to know that my family will be taken care of, even if I'm not around." - Female, 39
NGOs	Moral motivations, affinity of interest, nationalism, sense of social responsibility, change the world	"I want to contribute to a better world and help those less fortunate." - Male, 33	Religious motivations, affinity of interest, empathy, helping others, social sensitivity	"It's important for me to give back and support those who need it most." - Female, 40
Hearing Aids	Overcome stigma, improve impaired hearing, comfort, technology	"The latest technology can improve my hearing experience and make life easier." - Male, 65	Overcome stigma, improve impaired hearing, aesthetic reasons, discretion	"I want to hear better, but it's important for my hearing aid not to be noticeable." - Female, 62
Telecommunications	Price, mobile data, coverage	"I need a plan that offers good value for money, enough data, and wide coverage." Male, 30	Price, family plans, ease of use	"I'm looking for a family plan that's affordable and easy for everyone to use." - Male, 37
Gambling	Adrenaline, entertainment, make money fast	"The thrill of the game and the chance of a quick win is exhilarating." - Male, 50	Entertainment, socialization, big prizes	"I enjoy the social aspect of gambling and the chance of winning big." - Female, 45
Energy	Price, reliability, flexible options	"I want an energy provider that is affordable, reliable, and flexible." Male, 47	Price, sustainability, eco-friendly options	"I'm looking for an energy provider that's affordable and prioritizes sustainability." Female, 34
Finance	Security, returns, liquidity	"I need financial products that are secure, offer good returns, and provide easy access to my money." Male, 40	Security, convenience, advice	"I value financial security and appreciate convenient services and professional advice." Male, 42

https://doi.org/10.7441/joc.2025.04.04

Therefore, in response to RQ2 on the existence of different purchasing motivations for each category of online product, the findings suggest a complex picture that goes beyond traditional generalist propositions (Rohm & Swaminathan, 2004; Goldsmith & Horowitz, 2006). Specifically, they describe a complex system of interactions between intrinsic (personal values, needs and preferences) and extrinsic (social influence, market conditions and technological accessibility) motivations, as has been noted in previous studies (Bruwer & Wood 2005; Yeo et al., 2017).

Regarding RQ4, on different motivations by sex, we also observe common motivations, such as convenience and cost savings, and product-specific motivations, which also vary by reported gender (Rohm & Swaminathan, 2004; Melnyk, Van Osselaer & Bijmolt, 2009). These findings also align with earlier research (Choo & Mokhtarian, 2004).

In summary, qualitative research has revealed motivational patterns encompassing general, product-specific and gender-related factors, suggesting the need for a marketing approach that takes into account both demographic factors and individual motivations (Ahmadi et al., 2024; Dwivedi et al., 2021). However, given the exploratory nature of the research, this should be treated as a working hypothesis.

4. CONCLUSIONS

Although digital marketing is booming (Sáez-Ortuño et al., 2023d), e-commerce is beginning to show signs of maturity (Cramer-Flood, 2022) forcing retailers to shift their policies and strategies toward more focused objectives (Quinn & Cameron, 1983; Ma & Liang, 2021; Szymański, 2021). This requires more precise segmentation, consumer profiling and estimating the likelihood that they will purchase the promoted products (Malhotra, 2020). This is not an easy task, as despite the enormous effort to achieve accurate segments, marketers are never sure they have succeeded (Ahmadi et al., 2024). Data mining can help organisations discover hidden patterns and relationships in consumer profiles and do so quickly and efficiently (Pan & Zhou, 2020).

This article illustrates a real-world case of data mining in which an AI algorithm, XGBoost, is used to process large datasets and extract effective characteristics of buyers of seven products (insurance, hearing aids, NGOs, energy distributors, gambling, telecommunications, and finance). The big data consists of structured information on more than five million users from data registration templates on their demographic characteristics, as well as information on their online activities and transformation into buyers. The XGBoost algorithm uses decision trees to characterise users, estimate their predictive power and their willingness to buy the seven products. It also estimates whether there are significant differences between declared sexes, and a bias towards men is observed for gambling, insurance and financial products and towards women for headphones, NGOs, energy and telecommunications. All these analyses and calculations have been performed with minimal manual intervention by the researchers.

As the algorithm only works with numerical data, the study was complemented with a qualitative analysis to find out the specific purchase motivations for each of these products and compared by sex.

4.1. Theoretical Implications

The use of AI algorithms to analyse big data seems to be a highly promising avenue for conducting market research in e-commerce. However, studies of applications in the field of marketing are still scarce (Dorotic et al., 2024). Since the result of the XGBoost algorithm is not compared with another algorithm, its superiority could not be validated in this study. However, some of its advantages can be defended from the preceding literature. For example, Song and Liu (2020) used it to predict the intention to purchase an e-commerce product based on recorded user behaviour, and similarly, Li (2022) also uses the XGBoost algorithm to predict purchase intention from web-tracked data. In all these analyses, the advantages are its accuracy, since its tree structure can detect very complex patterns of non-linear relationships and interactions between variables (Chen & Guestrin, 2016), and its robustness to the presence of outliers, thanks to the optimisation of the objective function, and to heteroskedasticity in the variability (XGBoost, 2022). Turning to the disadvantages, this case illustrates how the algorithm only works with labelled data, be it keywords, categories or attributes. In this study, the data was already tagged by the lead-generating company, so we have no role in this process. It should also be noted that the XGBoost algorithm delivers results depending on the data incorporated, the criteria used for classification, the objective function and the reliability of the data used to train the algorithm (Chen & Guestrin, 2016). That is, with other types of information, different results could be obtained, so care must be taken with the information that is incorporated into the algorithm.

However, while the algorithm divides the sample into segments and estimates product purchase dispositions very accurately, it does not provide purchase motivations, which are essential for sales pitches and promotion (Ahmadi et al., 2024; Kshetri et al., 2023). A sample of 63 consumers participated in the qualitative research and online purchase motivations for seven products were separated into general (convenience, transaction cost savings, wide choice and stocking problems) and product-specific motivations.

This type of mixed-method research, which combines quantitative and qualitative studies, contributes to a holistic view of phenomena. According to Johnson et al. (2007), mixed methods are based on the combination of inductive (such as the discovery of patterns), deductive (such as hypothesis testing), and abductive (such as seeking the best explanations to understand the results) approaches in order to gain a broader and deeper understanding of the facts. Furthermore, this mixed approach uses pragmatism as a philosophy of science system (Harrison & Reilly, 2011).

In addition, mixed studies offer the possibility of redirecting one or both studies according to the results obtained. In other words, the findings of the qualitative study mean that the questions from the quantitative one can be revised (Harrison & Reilly, 2011). In this study, for example, it was explained earlier that the XGBoost algorithm provides results depending on how the data is classified (Chen & Guestrin, 2016), and qualitative research has provided clues to propose new classifications. That is, the information provided by qualitative research can help researchers and marketers consider factors that were not taken into account when feeding the algorithm, but may be relevant for both improving the accuracy and ranking of the data. For example, in the interviews on insurance purchase motivations, it was found that responses differed according to the type of insurance (life or health). This information can be used to relabel the variables in the databases and generate predictions adjusted to the type of insurance taken out.

However, while the proposal to micro-segment markets for personalised promotions is considered a panacea (Kshetri et al., 2023), recent research reveals that non-segmented campaigns can generate higher returns than micro-segmented ones (Ahmadi et al., 2024). These findings call into question the efficiency of segmenting the market very precisely, mainly because of the costs associated with message personalisation and performance per message (Ahmadi et al., 2024). Therefore, the key seems to lie in the costs of crafting micro-segment targeted promotional messages, and the solution could lie in the use of content-generative AI that would allow messages to be personalised with little human intervention. This could be seen as a possible extension of this work.

4.2. Managerial Implications

Data mining with AI algorithms, such as XGBoost, can help market researchers generate very precise segments of potential product buyers, as well as individualized estimates of their inclination to purchase each of the promoted products (for further details, see Sáez-Ortuño et al., 2023c). All of this information is considered extremely valuable for marketing managers to better target their strategies and make their communication campaigns and recommendations more effective (Malhotra, 2020). Moreover, if a mixed methodology is used, both feed back into each other to enable adjustment of the labelling of database variables and, in turn, boost the performance of AI algorithms. However, as noted above, an excessive approach can be counterproductive, mainly because of the costs involved in generating the content (Ahmadi et al., 2024).

On the other hand, in this study the data comes from a third-party lead capturing company, and the performance of the XGBoost algorithm depends on the quality of the data supplied to it. However, the current trend in policy is towards limiting or heavily restricting the use of data capture for third parties, which may further increase the cost per data by making limited audiences even less attractive (Ahmadi et al., 2024). In addition, data mining is relegating the researcher to a second role, shifting the spotlight towards assessing the algorithm's power. For example, Nguyen et al. (2021) illustrated how different approaches to analysing, organising and presenting raw data to a natural language processing (NLP) algorithm led to different results, interpretations and research directions. Therefore, the proposal to combine qualitative studies with algorithm applications, apart from enriching the results, helps to balance computational power with human knowledge (Nguyen et al., 2021).

4.3. Limitations and Future Research Directions

Although case studies have the advantage of describing real situations in depth and detail, there is the disadvantage that their results cannot be generalised. Nevertheless, publications dealing with the XGBoost algorithm typically illustrate real-life applications and often include comparisons with other methodologies that generate similar results (i.e. Li, 2022; Sáez-Ortuño et al., 2023; Song & Liu, 2020), thereby validating some of these approaches. Although the XGBoost algorithm is considered to be very efficient for segmentation and demand prediction, it tends to have difficulties forecasting small samples, as illustrated in this study (Chen & Guestrin, 2016; XGBoost, 2022). For example, although the training model showed commendable adaptability, with a margin of error of 12%, the group with the fewest observations (Finance) reached a margin of error of 250% in forecasting. This suggests the need for methods to improve the performance of these algorithms when using very small samples, for example, by creating synthetic data to enable better model fitting and improved accuracy (Massaro et al., 2021).

Another significant limitation is the danger of stereotyping, As the algorithm generalises consumers on the basis of just a few variables, some minority groups may be under-represented. Consequently, since each consumer is unique, special attention should be paid to these groups in the database (Lloyd, 1982). Finally, the mixed study revealed certain limitations of the XGBoost results, such as the consideration of generic products. If these had been separated into specific products, the segments and the inclination to purchase them would have been much better focused. Consequently, the study could be extended by including a wider product specification, although this is limited by the available sample size, or could use synthetic data generation procedures to improve the performance of the algorithms.

References

1. Adomavicius, G., & Tuzhilin, A. (2005). Personalization technologies: a process-oriented perspective. *Communications of the ACM*, 48(10), 83-90. https://doi.org/10.1145/1089107.1089109

2. 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. Retrieved October 10, 2024, from https://www.aepd.es/
3. Aguasca-Colomo, R., Castellanos-Nieves, D., & Méndez, M. (2019). Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife Island. *Applied Sciences*, 9(22), 4931. https://doi.org/10.3390/app9224931
4. 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
5. Ahmed, R., R., Streimikiene, D., Soomro, R., H. & Streimikis, J. (2022). Digital Transformation and Industry 4.0 Initiatives for Market Competitiveness: Business Integration Management Model in the Healthcare Industry. *Journal of Competitiveness*, 14(4), 6–24. https://doi.org/10.7441/joc.2022.04.01
6. Ahuja, M., Gupta, B., & Raman, P. (2003). An empirical investigation of online consumer purchasing behavior. *Communications of the ACM*, 46(12), 145-151. https://doi.org/10.1145/953460.953494
7. Athanasiou, V., & Maragoudakis, M. (2017). A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek. *Algorithms*, 10(1), 34. https://doi.org/10.3390/a10010034
8. Bagozzi, R. P., & Yi, Y. (1989). The degree of intention formation as a moderator of the attitude-behavior relationship. *Social Psychology Quarterly*, 52(4), 266-279. https://doi.org/10.2307/2786991
9. Batra, R., & Keller, K. L. (2016). Integrating marketing communications: New findings, new lessons, and new ideas. *Journal of Marketing*, 80(6), 122-145. https://doi.org/10.1509/jm.15.0419
10. Baumeister, R. F., & Sommer, K. L. (1997). What do men want? Gender differences and two spheres of belongingness: Comment on Cross and Madson (1997). *Psychological Bulletin*, 122(1), 38–44. https://doi.org/10.1037/0033-2909.122.1.38
11. Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems*, 19(1), 211-241. https://doi.org/10.1080/07421222.2002.11045715
12. Billsus, D., Brunk, C. A., Evans, C., Gladish, B., & Pazzani, M. (2002). Adaptive interfaces for ubiquitous web access. *Communications of the ACM*, 45(5), 34-38. https://doi.org/10.1145/506218.506240
13. Blom, J. O., & Monk, A. F. (2003). Theory of personalization of appearance: Why users personalize their PCs and mobile phones. *Human-computer interaction*, 18(3), 193-228. https://doi.org/10.1207/S15327051HCI1803_1
14. Bolls, P. D., Muehling, D. D., & Yoon, K. (2003). The effects of television commercial pacing on viewers' attention and memory. *Journal of Marketing Communications*, 9(1), 17-28. https://doi.org/10.1080/1352726032000068032
15. Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative *Research in Sport, Exercise and Health*, 11(4), 589-597. https://doi.org/10.1080/2159676X.2019.1628806
16. Bruwer, J., & Wood, G. (2005). The Australian online wine-buying consumer: motivational and behavioural perspectives. *Journal of Wine Research*, 16(3), 193-211. https://doi.org/10.1080/09571260600556666

17. Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European journal of operational research*, 164(1), 252-268.
18. 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
19. Chen, H., & Wan, W. (2023). Analysis of E-Commerce Marketing Strategy Based on Xgboost Algorithm. *Advances in Multimedia*, Article ID 1247890. https://doi.org/10.1155/2023/1247890
20. Choo, S., & Mokhtarian, P. L. (2004). What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice. *Transportation Research Part A: Policy and Practice*, 38(3), 201-222. https://doi.org/10.1016/j.tra.2003.10.005
21. Cienkowski, K. M., & Pimentel, V. (2001). The hearing aid 'effect'revisited in young adults. *British journal of audiology*, 35(5), 289-295. https://doi.org/10.1080/00305364.2001.11745247
22. Cramer-Flood, E. (2022). Global Ecommerce Forecast & Growth Projections As 2-Year Boom Subsides, Plenty of Bright Spots Remain. Insider Intelligence. Retrieved November 13, 2024, from https://www.insiderintelligence.com/content/global-ecommerce-forecast-2022
23. Darley, W. K., & Smith, R. E. (1995). Gender differences in information processing strategies: An empirical test of the selectivity model in advertising response. *Journal of Advertising*, 24(1), 41-56. https://doi.org/10.1080/00913367.1995.10673467
24. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0
25. Diebold, F. X. (2003, February). Big data dynamic factor models for macroeconomic measurement and forecasting. In M. Dewatripont, LP Hansen and S. Turnovsky (Eds), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, (Vol. 115, p. 22).
26. Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 67-73).
27. Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
28. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... Papagiannidis, S. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.
29. Dornhoffer, J. R., Meyer, T. A., Dubno, J. R., & McRackan, T. R. (2020). Assessment of hearing aid benefit using patient-reported outcomes and audiologic measures. *Audiology and Neurotology*, 25(4), 215-223. https://doi.org/10.1016/j.ijinfomgt.2023.102642
30. Dorotic, M., Stagno, E., & Warlop, L. (2024). AI on the street: Context-dependent responses to artificial intelligence. *International Journal of Research in Marketing*, 41(1), 113-137. https://doi.org/10.1016/j.ijresmar.2023.08.010

31. Eagly, A. H. (2013). Sex differences in social behavior: A social-role interpretation. Psychology Press.
32. Einolf, C. J. (2011). Gender differences in the correlates of volunteering and charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(6), 1092-1112. https://doi.org/10.1177/0899764010385949
33. European Parliament (2023) Big data: definition, benefits, challenges (infographics). Retrieved July 9, 2024, from https://www.europarl.europa.eu/topics/en/article/20210211STO97614/big-data-definition-benefits-challenges-infographics
34. Forgas-Coll, S., Huertas-Garcia, R., Andriella, A., & Alenyà, G. (2023). Gendered human–robot interactions in services. *International Journal of Social Robotics*, 5, 1791–1807. https://doi.org/10.1007/s12369-023-01035-8
35. Gagné, J. P., Southall, K., & Jennings, M. B. (2011). Stigma and self-stigma associated with acquired hearing loss in adults. *Hearing Review*, 18(8), 16-22.
36. Garstecki, D. C., & Erler, S. F. (1999). Older adult performance on the communication profile for the hearing impaired: Gender difference. *Journal of Speech, Language, and Hearing Research*, 42(4), 785-796. https://doi.org/10.1044/jslhr.4204.785
37. Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, 21(4), 389-400. https://doi.org/10.2307/249720
38. Goldsmith, R. E., & Horowitz, D. (2006). Measuring motivations for online opinion seeking. *Journal of Interactive Advertising*, 6(2), 2-14. https://doi.org/10.1080/15252019.2006.10722114
39. 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
40. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
41. Graafland, J., & de Bakker, F. G. (2021). Crowding in or crowding out? How non-governmental organizations and media influence intrinsic motivations toward corporate social and environmental responsibility. *Journal of Environmental Planning and Management*, 64(13), 2386-2409. https://doi.org/10.1080/09640568.2021.1873110
42. 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
43. Guan, L. P., Yusuf, D. H. M., & Ghani, M. R. A. (2020). Factors influencing customer purchase intention towards insurance products. *International Journal of Business and Management*, 4(5), 70-79. https://doi.org/10.26666/rmp.ijbm.2020.5.9
44. Gupta, S., Leszkiewicz, A., Kumar, V., Bijmolt, T., & Potapov, D. (2020). Digital analytics: Modeling for insights and new methods. *Journal of Interactive Marketing*, 51(1), 26-43. https://doi.org/10.1016/j.intmar.2020.04.003
45. Harrison, R. L., & Reilly, T. M. (2011). Mixed methods designs in marketing research. *Qualitative market research: an international journal*, 14(1), 7-26. https://doi.org/10.1108/13522751111099300
46. Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1), 20-38. https://doi.org/10.1016/j.ijresmar.2018.09.009
47. Heider, F. (1958). The psychology of interpersonal relations. New York: Wiley.

48. Hoban, P. R., & Bucklin, R. E. (2015). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*, 52(3), 375-393. https://doi.org/10.1509/jmr.13.0277
49. Hull, C. L. (1952). A behavior system; an introduction to behavior theory concerning the individual organism. Yale University Press
50. Internet World Stats (2022). World Internet Usage and Population Statistics. Retrieved December 22, 2024, from https://www.internetworldstats.com/
51. Ismail, M., Ibrahim, M. M., Sanusi, Z. M., & Nat, M. (2015). Data mining in electronic commerce: benefits and challenges. *International Journal of Communications, Network and System Sciences*, 8(12), 501. https://doi.org/10.4236/ijcns.2015.812045
52. Jadhav, V. & Ramakrishna, S. (2023). Motivations and barriers to purchase health insurance: A qualitative study. *Asia Pacific Journal of Health Management*, 18(1), 275-282. https://doi.org/10.24083/apjhm.v18i1.1689
53. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.
54. Jha, S. K., Bilalovic, J., Jha, A., Patel, N., & Zhang, H. (2017). Renewable energy: Present research and future scope of Artificial Intelligence. *Renewable and Sustainable Energy Reviews*, 77, 297-317. https://doi.org/10.1016/j.rser.2017.04.018
55. Jiang, Y., Shang, J., Liu, Y., & May, J. (2015). Redesigning promotion strategy for e-commerce competitiveness through pricing and recommendation. *International Journal of Production Economics*, 167, 257-270. https://doi.org/10.1016/j.ijpe.2015.02.028
56. Jin, Y., & Su, M. (2009). Recommendation and repurchase intention thresholds: A joint heterogeneity response estimation. *International Journal of Research in Marketing*, 26(3), 245-255. https://doi.org/10.1016/j.ijresmar.2009.06.004
57. Johnson, J. P. (2013). Targeted advertising and advertising avoidance. *The RAND Journal of Economics*, 44(1), 128-144. https://www.jstor.org/stable/43186411
58. Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112-133. https://doi.org/10.1177/1558689806298224
59. Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, 70, 102641. https://doi.org/10.1016/j.ijinfomgt.2023.102641
60. Kannadath, B. S., Cen, P., Rowe, J., Wray, C., Bynon, J., Rahimi, E. F., ... & Thosani, N. (2018). Decision tree analysis of pancreatic cyst fluid data for the detection of mucinous cysts: 73. *Official journal of the American College of Gastroenterology| ACG*, 113, S41-S42. https://doi.org/10.14309/00000434-201810001-00073
61. Kar, A.K., & Dwivedi, Y.K. (2020). Theory building with big data-driven research – Moving away from the "What" towards the "Why". *International Journal of Information Management*, 54, 102205. https://doi.org/10.1016/j.ijinfomgt.2020.102205
62. Kim, H. S., & Yoon, C. H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy*, 28(9-10), 751-765. https://doi.org/10.1016/j.telpol.2004.05.013Get rights and content
63. Kshetri, N., Dwivedi, Y.K., Davenport, T.H., & Panteli, N. (2023). Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *International Journal of Information Management*, In press,102716. https://doi.org/10.1016/j.ijinfomgt.2023.102716

64. Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33(2),363-374. https://doi.org/10.2307/2529786
65. Lavie, T., Sela, M., Oppenheim, I., Inbar, O., & Meyer, J. (2010). User attitudes towards news content personalization. *International journal of human-computer studies*, 68(8), 483-495. https://doi.org/10.1016/j.ijhcs.2009.09.011
66. Li, Z. (2022). Accurate digital marketing communication based on intelligent data analysis. *Scientific Programming*, 2022(1), 8294891. https://doi.org/10.1155/2022/8294891
67. Li, H. A., & Kannan, P. K. (2013). Attribution modeling: understanding the influence of channels in the online purchase funnel. *Marketing Science Institute Working Paper Series*, (12-115).
68. Li, S. S., & Karahanna, E. (2015). Online recommendation systems in a B2C E-commerce context: a review and future directions. *Journal of the Association for Information Systems*, 16(2), 72-107. https://doi.org/10.17705/1jais.00389
69. Liang, T. P., Lai, H. J., & Ku, Y. C. (2006). Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, 23(3), 45-70. https://doi.org/10.2753/MIS0742-1222230303
70. Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019, August). Product marketing prediction based on XGboost and LightGBM algorithm. In Proceedings of the 2nd international conference on artificial intelligence and pattern recognition (pp. 150-153).
71. Liu, Y., & Qiao, M. (2019). Heart disease prediction based on clustering and XGboost. *Computer Systems & Applications*, 28(01), 228-232. https://doi.org/10.1109/ICICICT54557.2022.9917678
72. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489
73. Lloyd, J., Doll, H., Hawton, K., Dutton, W. H., Geddes, J. R., Goodwin, G. M., & Rogers, R. D. (2010). How psychological symptoms relate to different motivations for gambling: An online study of internet gamblers. *Biological psychiatry*, 68(8), 733-740. https://doi.org/10.1016/j.biopsych.2010.03.038
74. Lo, F. Y., & Chien, Y. Y. (2024). Digital Internationalization of Small and Medium Sized Enterprises: online comments and ratings on Amazon platform. *Journal of Competitiveness*, 16(1), 146-166
75. Lorains, F. K., Cowlishaw, S., & Thomas, S. A. (2011). Prevalence of comorbid disorders in problem and pathological gambling: Systematic review and meta-analysis of population surveys. *Addiction*, 106(3), 490-498. https://doi.org/10.1111/j.1360-0443.2010.03300.x
76. Ma, S., & Liang, Q. (2021). Industry competition, life cycle and export performance of China's cross-border e-commerce enterprises. *International Journal of Technology Management*, 87(2-4), 171-204. https://doi.org/10.1504/IJTM.2021.120926
77. Mahdzan, N. S., & Victorian, S. M. P. (2013). The determinants of life insurance demand: A focus on saving motives and financial literacy. *Asian Social Science*, 9(5), 274. https://doi.org/10.5539/ass.v9n5p274
78. Malhotra, N. K. (2020). Marketing Research: An Applied Orientation. 7ed. Ed. Pearson.
79. 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

80. Melnyk, V., Van Osselaer, S. M., & Bijmolt, T. H. (2009). Are women more loyal customers than men? Gender differences in loyalty to firms and individual service providers. *Journal of Marketing*, 73(4), 82-96. https://doi.org/10.1509/jmkg.73.4.82
81. Meyers-Levy, J., & Loken, B. (2015). Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology*, 25(1), 129-149. https://doi.org/10.1016/j.jcps.2014.06.003
82. Mitchell, T.R. (1997), Matching motivational strategies with organizational contexts. *Research in Organizational Behavior*, 19, 57-94.
83. Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., Harmouch, H. (2022). The effects of data quality on machine learning performance. arXiv preprint arXiv:2207.14529. https://doi.org/10.48550/arXiv.2207.14529
84. Moss, G. (2017). Gender, design and marketing: How gender drives our perception of design and marketing. Routledge.
85. Neves, J., & Oliveira, T. (2021). Understanding energy-efficient heating appliance behavior change: The moderating impact of the green self-identity. *Energy*, 225, 120169. https://doi.org/10.1016/j.energy.2021.120169
86. Nguyen, H., Ahn, J., Belgrave, A., Lee, J., Cawelti, L., Kim, H. E., Prado, Y., Santagata, R., & Villavicencio, A. (2021). Establishing trustworthiness through algorithmic approaches to qualitative research. In Advances in Quantitative Ethnography: Second International Conference, ICQE 2020, Malibu, CA, USA, February 1-3, 2021, Proceedings 2 (pp. 47-61). Springer International Publishing.
87. Nickel, K., Orth, U. R., & Kumar, M. (2020). Designing for the genders: The role of visual harmony. *International Journal of Research in Marketing*, 37(4), 697-713. https://doi.org/10.1016/j.ijresmar.2020.02.006
88. Palan, K. M. (2001). Gender identity in consumer behavior research: A literature review and research agenda. *Academy of Marketing Science Review*, 10(2001), 1-31.
89. Pan, H., & Zhou, H. (2020). Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce. Electronic Commerce Research, 20, 297-320. https://doi.org/10.1007/s10660-020-09409-0
90. Patton M. Q. (1990). Qualitative Evaluation and Research Methods. 2nd ed. Sage, Newbury Park, California.
91. Petty, R. E., Wheeler, S. C., & Bizer, G. Y. (2000). Attitude functions and persuasion: An elaboration likelihood approach to matched versus mismatched messages. In G. Maio & J. Olson (Eds.), Why we evaluate: Functions of attitudes (pp.133–162). Lawrence Erlbaum Associates, Publishers, Mahwah, NJ (USA).
92. Quinn, R. E., & Cameron, K. (1983). Organizational life cycles and shifting criteria of effectiveness: Some preliminary evidence. *Management Science*, 29(1), 33–51. https://doi.org/10.1287/mnsc.29.1.33
93. 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
94. Rohm, A. J., & Swaminathan, V. (2004). A typology of online shoppers based on shopping motivations. *Journal of Business Research*, 57(7), 748-757. https://doi.org/10.1016/S0148-2963(02)00351-X
95. Rossiter, J. R., & Foxall, G. R. (2008). Hull-Spence behavior theory as a paradigm for consumer behavior. *Marketing Theory*, 8(2), 123-141. https://doi.org/10.1177/1470593108089201
96. Rotter, J. B. (1954). Social learning and clinical psychology. Prentice-Hall, Inc.

97. Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., & Sánchez-García, J. (2023a). What's on the horizon? A bibliometric analysis of personal data collection methods on social networks. *Journal of Business Research*, 158, 113702. https://doi.org/10.1016/j.jbusres.2023.113702
98. Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., & Sánchez-García, J. (2023b). Online cheaters: Profiles and motivations of internet users who falsify their data online. *Journal of Innovation & Knowledge*, 8(2), 100349. https://doi.org/10.1016/j.jik.2023.100349
99. Sáez-Ortuño, L., Huertas-Garcia, R., Forgas-Coll, S., & Puertas-Prats, E. (2023c). How can entrepreneurs improve digital market segmentation? A comparative analysis of supervised and unsupervised learning algorithms. *International Entrepreneurship and Management Journal*, 1-28. https://doi.org/10.1007/s11365-023-00882-1
100. Sáez-Ortuño, L., Sanchez-Garcia, J., Forgas-Coll, S., Huertas-García, R., & Puertas-Prat, E. (2023d). Impact of Artificial Intelligence on Marketing Research: Challenges and Ethical Considerations. In Philosophy of Artificial Intelligence and Its Place in Society (pp. 18-42). IGI Global.
101. Sandoval, L. L. (2017). Machine Learning algorythms for analysis and data prediction. 2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII), 1-5. Managua, Nicaragua. IEEE. http://doi.org/10.1109/CONCAPAN.2017.8278511
102. Schiaffino, S., & Amandi, A. (2004). User–interface agent interaction: personalization issues. *International Journal of Human-Computer Studies*, 60(1), 129-148. https://doi.org/10.1016/j.ijhcs.2003.09.003
103. Sebastian, S., Varghese, A., & Gowri, M. (2015). The impact of hearing loss in the life of adults: A comparison between congenital versus late onset hearing loss. *Indian Journal of Otology*, 21(1). https://doi.org/10.4103/0971-7749.152857
104. Sharma, D., Choudhury, T., Dewangan, B. K., Bhattacharya, A., & Dutta, S. (2021). A recommendation system for customizable items. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 1 (pp. 471-482). Springer Singapore.
105. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600
106. 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
107. Stone, M. D., & Woodcock, N. D. (2014). Interactive, direct and digital marketing: A future that depends on better use of business intelligence. *Journal of research in interactive marketing*, 8(1), 4-17. https://doi.org/10.1108/JRIM-07-2013-0046
108. Szymański, G. (2021). Marketing activities of local food producers in e-commerce. *Sustainability*, 13(16), 9406. https://doi.org/10.3390/su13169406
109. Thirumalai, S., & Sinha, K. K. (2009). Customization strategies in electronic retailing: Implications of customer purchase behavior. *Decision Sciences*, 40(1), 5-36. https://doi.org/10.1111/j.1540-5915.2008.00222.x
110. Tifferet, S., & Herstein, R. (2012). Gender differences in brand commitment, impulse buying, and hedonic consumption. *Journal of product & brand management*, 21(3), 176-182. https://doi.org/10.1108/10610421211228793
111. Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, 51(5), 546-562. https://doi.org/10.2139/ssrn.1694319

112. Tuomi, A., Tussyadiah, I. P., & Hanna, P. (2021). Spicing up hospitality service encounters: the case of Pepper™. *International Journal of Contemporary Hospitality Management*, 33(11), 3906-3925. https://doi.org/10.1108/IJCHM-07-2020-0739
113. Unerman, J., & O'Dwyer, B. (2006). On James Bond and the importance of NGO accountability. *Accounting, Auditing & Accountability Journal*, 19(3), 305-318. https://doi.org/10.1108/09513570610670316
114. Varian, H. R. (2002). Market structure in the network age. In E. Brynjolfsson & B. Kahin (Eds.). Understanding the Digital Economy: Data, Tools, and Research, 137-150. The MIT Press, Massachusetts.
115. Wang, J., Ipeirotis, P. G., & Provost, F. (2017). Cost-effective quality assurance in crowd labeling. *Information Systems Research*, 28(1), 137-158. https://doi.org/10.1287/isre.2016.0661
116. Wärneryd, K. E. (1999). The Psychology of Saving: A study on Economic Psychology. Glos: Edward Elgar.
117. Weaver, G. R., Treviño, L. K., & Agle, B. (2005). “Somebody I look up to: ethical role models in organizations”. *Organizational Dynamics*, 34(4), 313-330. https://doi.org/10.1016/j.orgdyn.2005.08.001
118. 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
119. Wigand, R. T., Benjamin, R. I., & Birkland, J. L. (2008, August). Web 2.0 and beyond: implications for electronic commerce. In Proceedings of the 10th international conference on Electronic commerce (pp. 1-5).
120. XGBoost (2022). XGBoost Documentation. Retrieved May 5, 2024, from https://xgboost.readthedocs.io/en/stable/
121. Xu, M., & Du, J. (2011, September). Design of SMS-based remote control system using TC35 and MCU. In 2011 International Conference on Internet Computing and Information Services (pp. 393-395). IEEE.
122. Yeo, V. C. S., Goh, S. K., & Rezaei, S. (2017). Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services. *Journal of Retailing and Consumer services*, 35, 150-162. https://doi.org/10.1016/j.jretconser.2016.12.013
123. Zaman, M., Vo-Thanh, T., Nguyen, C. T., Hasan, R., Akter, S., Mariani, M., & Hikkerova, L. (2023). Motives for posting fake reviews: Evidence from a cross-cultural comparison. *Journal of Business Research*, 154, 113359. https://doi.org/10.1016/j.jbusres.2022.113359

Contact information

Prof. Laura Sáez-Ortuño, Ph.D.
Universitat de Barcelona
Faculty of Economics and Business
Business Department
Barcelona, Spain
E-mail: laurasaez@ub.edu
ORCID: https://orcid.org/0000-0001-6660-9458

Prof. Santiago Forgas-Coll, Ph.D.
Universitat de Barcelona
Faculty of Economics and Business

Business Department
Barcelona, Spain
E-mail: santiago.forgas@ub.edu
ORCID: https://orcid.org/0000-0003-2288-3716

Prof. Ruben Huertas-Garcia, Ph.D.
Universitat de Barcelona
Faculty of Economics and Business
Business Department
Barcelona, Spain
E-mail: rhuertas@ub.edu
ORCID: https://orcid.org/0000-0001-6272-132X

Prof. Javier Sánchez-Garcia, Ph.D.
Universitat Jaume I
Faculty of Law and Economics
Department of Business Administration and Marketing
Castelló de la Plana, Spain
E-mail: jsanchez@uji.es
ORCID: https://orcid.org/0000-0002-7865-0076