Received: 12 May 2023 | Revised: 5 February 2024 | Accepted: 5 February 2024

DOI: 10.1111/ijcs.13030

ORIGINAL ARTICLE

International Journal of Consumer Studies WILEY

A data-driven approach to improve online consumer subscriptions by combining data visualization and machine learning methods

Elizabeth Fernandes [1] | Sérgio Moro [2,3] | Paulo Cortez [4]

[1]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

[2]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

[3]University of Jordan, Amman, Jordan

[4]ALGORITMI Research Center/LASI, University of Minho, Guimarães, Portugal

Correspondence
Elizabeth Fernandes, ISCTE—Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Avenida das Forças Armadas, Edifício II, D615, 1649-026 Lisboa, Portugal.
Email: elisabeth.ferna@gmail.com

Funding information
FCT—Fundação para a Ciência e Tecnologia, under the Projects, Grant/Award Numbers: UIDB/04466/2020, UIDP/04466/2020, UIDB/00319/2020.

Abstract

Effective online consumer research helps companies on defining a successful strategy to increase user loyalty and shape brand engagement. Digital innovation introduced a dramatic change in businesses, particularly in the online news industry. Content consumers have a wide offer across different channels which increase the digital challenge for online news media companies to retain their readers and convert them into online subscribers. Furthermore, digital news publishers often strive to balance revenue sources in online business models. Thus, this study fills a gap in the literature on media consumer research by proposing a data-driven approach that combines two machine learning (ML) models to allow managers dynamically improve their marketing and editorial strategies. Firstly, the authors present an online user profiling to identify consumer segments based on the interplay between several engagement' variables substantiated in the literature research. Second, as few studies have explored the factors influencing users' intention to pay for such services, the e*X*treme Gradient *B*oosting ML algorithm identifies the predictors of consumer's willingness to pay. Third, a dashboard presents the key performance indicators across the audience funnel. Thus, practical implications and business suggestions are presented in a two-fold strategy to maximize revenue from digital subscriptions and advertising. Findings provide new insights into an engagement approach and the relation to acquire a digital subscription in online content platforms. We believe that the provided recommendations are potentially useful to help marketing and editorial teams to manage their customer engagement process across the funnel in a more efficient way.

KEYWORDS
Cluster analysis, digital consumers, digital subscriptions, machine learning, online content platforms, user engagement

1 | INTRODUCTION

In the last decade, the decline of print advertising revenue, and the increase in digital subscription business models (DSBMs) changed the media business landscape (Arrese, 2016; Rios-Rodríguez et al., 2022). The challenges of digital transformation, the need to innovate, in the whole value chain, to ensure financial sustainability (Vara-Miguel et al., 2021) are some major concerns of the media industry. Thus, to improve consumer engagement and acquisition, online news publishers need to build strong digital strategies (Davoudi, 2018). Furthermore, effective engagement management analysis allows managers to understand consumer behavior by providing insights to conduct data-driven strategies (Barari et al., 2021). Moreover, the current highly competitive and nonlinear environment forces content platforms to guarantee that they are offering useful, informative, compelling, gratifying content, and more than ever, engaging content (Ksiazek et al., 2016; Rios-Rodríguez et al., 2022).

News media companies have begun to embrace other sources of revenue, such as, sponsored content, subscriptions or memberships, electronic commerce, and standard advertising (Vara-Miguel et al., 2021). However, the challenge is to define the right balance of all those revenue streams. Consequently, to define a marketing and/or editorial strategy, news publishers need to explore reader characteristics, for example, the effect of the content quality perception or the willingness to pay (WTP) for online news (Bomnüter et al., 2022). Hence, to diversify and innovate the business, it becomes fundamental to understand reader drivers to read and to subscribe (Vara-Miguel et al., 2021).

In the digital landscape, companies want to catch users' attention (Karampournioti & Wiedmann, 2021). Due to the speed with which user interacts, each second of time results in millions of user's interactions and huge volumes of data (Karampournioti & Wiedmann, 2021). This Big Data environment demands the development of analytical solutions. Thus, to define an effective strategy, e-commerce platforms are investing on customer segmentations (Koul & Philip, 2021; La Torre, 2020; Punhani et al., 2021). A wide range of segmentation methods have been explored, such as, behavioral, psychographic, geographic, demographic, value-based or propensity-based segmentation (Punhani et al., 2021). However, clustering techniques remain as an important and effective data mining solutions to identify users' clusters (Koul & Philip, 2021). Although consumer segmentations are popular across industries (Nasir et al., 2021; Vinothini & Priya, 2018), particularly in e-commerce (Koul & Philip, 2021; Punhani et al., 2021; Shen, 2021), or across B2C DSBM businesses, for example, online social games (Fu et al., 2017), movie industry (Tanuwijaya et al., 2021), education (Smit et al., 2019), and online news readers (La Torre, 2020), no study was found in the particular case of news publishers focused on engagement attributes.

Thus, based on the premise that the best predictor of future consumer behavior is past consumer behavior (Kamthania et al., 2018), this research addresses the problem of identifying different reader profiles based on reader engagement attributes (Fernandes et al., 2023) and behavioral features (Shen, 2021) collected from an online newspaper. The obtained insights are readable and potentially actionable, facilitating a data-driven decision making by the business and marketing teams. Moreover, while applied to a specific online newspaper, the proposed approach is replicable to other online news or even online business domains that depend on subscriptions. Indeed, by following the same engagement definition, the adopted attributes can be adapted and replaced for a particular online business context (e.g., digital subscriptions for games).

In this study, we identify online users' profiles based on the interplay between the engagement' attributes (EA) substantiated in the literature research, that are: recency, frequency, volume (Shen, 2021), volume of premium content, active days, interactivity, and scroll down (Ksiazek et al., 2016; Lajumoke et al., 2020). Besides, the derived results are actionable, interpretable (Whetten, 1989), and experimented in the Portuguese newspaper Público, that is, a B2C online news publisher with a DSBM. Furthermore, the web site of the Público online news is widely known in Portugal, obtaining a high number of visits between 2020 and 2022, a period where the Coronavirus disease 2019 (COVID-19) pandemic produced a global health crisis (Gordon-Wilson, 2022; Kursan Milaković, 2021; Purohit et al., 2022). In effect, Público web site reached more than 4 million users, reflected in more than 200 million user events per month. Thus, the analysed sample represents a substantial portion of the Portuguese online news readers.

The present study addresses three research questions and contributes to the existing literature in two ways: by proposing a bundle of reader engagement metrics to explain reader behavior into a B2C DSBM that can be applied in other industries; and by presenting a machine learning (ML) approach that combines k-means (Koul & Philip, 2021) and eXtreme Gradient Boosting (XGBoost) algorithms (Brownlee, 2016) to segment dynamically online readers while displaying into a dashboard the engagement features more related to the propensity to subscribe. Findings provide new insights into an engagement approach and the relation to acquire a digital subscription in online content platforms. Thus, the present original research has high utility (Corley & Gioia, 2011) in digital B2C subscription businesses.

Following this section, the paper continues with the literature review in Section 2. Section 3 describes the methodology and its respective application, including analysis and results. Findings are further discussed in Section 4 followed by the conclusion in Section 5.

2 | THEORETICAL BACKGROUND

2.1 | Segmenting online news readers

The explosive growth in available data from online businesses has induced a strong emphasis on Big Data analysis (BDA) (Mathew, 2021). Furthermore, BDA applications, such as algorithms in classification, clustering, and association, have been used in a disparate variety of industries like media, entertainment, and communication (Punhani et al., 2021; Vinothini & Priya, 2018). For instance, Spotify or Netflix are among the B2C companies with DSBMs that

already uses BDA (Mathew, 2021), as well as some widely known publishers like The New York Times (Rußell et al., 2020) or Aftenposten (Sjøvaag, 2016).

Cluster analysis is an important ML modeling technique to discover natural grouping of the observed data to create consumer profiles (Gonçalves & Carvalho, 2021; Silva et al., 2018; Vinothini & Priya, 2018). The concept of segmentation, firstly introduced by Smith (1956), consists in dividing a heterogeneous group into smaller homogeneous sub-groups in which consumers share something in common (Fu et al., 2017; Smith, 1956). As consumers are heterogeneous in their behaviors, a segmentation helps the company to develop customized strategies and suitable products for different segments (Fu et al., 2017; Joung & Kim, 2023; Lee et al., 2018; Liu et al., 2021; Vinothini & Priya, 2018).

In clustering, two learning problems are addressed that are supervised, also known as classification, and unsupervised (Rajput & Singh, 2023), known as clustering. The first referred to labeled data, the second one to unlabeled data (Gonçalves & Carvalho, 2021). One of the most popular and efficient clustering methods is the k-means algorithm (Bock, 2007; Koul & Philip, 2021; MacQueen, 1967; Vergani & Binaghi, 2018) which has been evolving since the first approaches proposed in the fifties of the twentieth century (Bock, 2007; MacQueen, 1967). In the digital landscape, many applications have been published, for example, to segment B2C e-commerce customers (Rajput & Singh, 2023) or online users (La Cruz et al., 2021). To explore previous research in online reader's segmentations, a set of keywords were used as inputs in the Scopus and Web of Science search engines. The queries used at Scopus and Web of Science were:

- TITLE-ABS-KEY ((“reader” OR “user” OR “customer”) AND (“segmentation” OR “clustering”) AND (“media” OR “publishers”)) AND PUBYEAR >2012 AND PUBYEAR <2024.
- (TI = (customer segmentation) OR TI = (reader segmentation) OR TI = (user segmentation) OR TI = (customer clustering) OR TI = (reader clustering) OR TI = (user clustering)) AND (AB = (customer segmentation) OR AB = (reader segmentation) OR AB = (user segmentation) OR AB = (customer clustering) OR AB = (reader clustering) OR AB = (user clustering)) AND PY = (2013 or 2014 or 2015 or 2016 or 2017 or 2018 or 2019 or 2020 or 2021 or 2022 or 2023).

The search resulted in 2806 and 1308 documents, respectively. More results were recovered from Scopus, as it is the largest abstract and citation database of peer-reviewed literature (Ballew, 2009) and highly recommended by scholars (Mongeon & Paul-Hus, 2016; Pranckutė, 2021).

In this sample, a wide range of successful applications across industries can be found, specially in social media (Sinha et al., 2020) or e-commerce (Ballestar et al., 2018; Koul & Philip, 2021; Shen, 2021). However, in the particular case of online news, research on segmenting online readers is quite scarce (Fernandes et al., 2023). Among media research articles, in the last 10 years, there are investigations on segmenting readers according to similarities in their preference patterns (Chakrabarty et al., 2019), or segmenting readers according to their perception of media website features such as, easy navigation, or content relevance or trustfullness (Cristobal-Fransi et al., 2017). Furthermore, we also found studies on the identification of content consumption patterns based on clicks behavior (Makhortykh et al., 2021); segmentation of library readers based on interviews that collected some reader engagement features, such as, the average number of books read in the 3 months (Paul, 2022); or online social media users segmentation based on posting volume (Sinha et al., 2020).

However, to the best of our knowledge, there is no approach that studies online news readers' segments and characteristics by exploring user engagement metrics. This is probably due to the complexity associated with the wide range of reader engagement levels, the high proportion of new monthly users (Jacob, 2021), and the high amount of data produced. Furthermore, behavioral segmentation is the most sophisticated method of segmentation that also has more potential to optimize the user experience as required in the digital landscape (Joshi & Garg, 2021).

In today's competitive Big Data environment, online news publishers need to segment and target readership to meet revenue goals. Furthermore, DSBMs became quite popular in the last decade (Mathew, 2021), in particular, across online publishers (Arrese, 2016; Fernandes et al., 2023). One fundamental step in the application of cluster analysis is to choose the variables along which to group individuals (Ketchen & Shook, 1996; Lee et al., 2018). Thus, seeking to define the segmentation attributes, Table 1 summarizes different segmentation models and attributes used to identify consumer’ clusters in B2C DSBMs.

For each approach, the table mentions the industry where the method was applied and the respective attributes used. The first approach consists of segmenting online players by using engagement, performance, and social interaction features (Fu et al., 2017). The second approach considers sociodemographic, user and behavior data (Tanuwijaya et al., 2021). Finally, the third approach segments consumers by using user and behavior data (Smit et al., 2019).

The present approach focuses on performing behavioral segmentation to influence future behavior. Thus, the selection of attributes is based on the literature presented in Table 1, coupled with the main features of the consumer engagement definitions found in the research literature, detailed further ahead.

2.2 | Reader engagement

In the 1990s of the 20th century, newspapers identified readers engagement by the hand-written letters to the editor (Risch & Krestel, 2020). With the establishment of online platforms, reader engagement definition evolved, and reader engagement has been measured in several different ways. From the number of comments (Risch & Krestel, 2020) to a wide range of digital metrics, such as visits or time spent reading (Moyo et al., 2019). Furthermore, consumer

TABLE 1 Literature review on consumers' segmentations in B2C DSBMs.

Author(s)	Business domain	Attributes/features	Method(s)
Fu et al. (2017)	Games	*Engagement*: playtime, login *Performance*: level, mission, quest, coin frequency, coin *Social interaction*: guild status, guild role type, common point, guild frequency, and friends	Fuzzy C-means clustering
Tanuwijaya et al. (2021)	Streaming	*Sociodemographic*: age, gender *User*: smartphone brand. *Behavior*: traffic, duration, sessions	Naïve Bayes Decision Tree Random Forest Logistics Regression LGBM- light gradient-boosting machine XGBoost Catboost
Smit et al. (2019)	Education	*User*: date and time at which the recorded event happened, country, time zone corresponding the geo location of the user *Behavior*: sessions, URL, type of device, type of browser type of operating system	DBSCAN - Density-based Spatial Clustering of Applications with Noise

Abbreviation: XGBoost, eXtreme Gradient Boosting.

engagement involves emotional attachment and rational loyalty (Barari et al., 2021; Lagun & Lalmas, 2016; Riskos et al., 2022; Wenzel et al., 2022).

In the media management literature, qualitative studies show that consumer brand engagement in the media industry can be expressed through the three dimensions of affection, cognition, and behavior (Riskos et al., 2022). Furthermore, authors agree that reader engagement is a multidimensional phenomenon (Steensen et al., 2020) related to the level of attention and involvement (emotional, cognitive, and behavioral) (Barari et al., 2021) with media product (Ksiazek et al., 2016; Mersey et al., 2010; Vreese & Neijens, 2016). Moreover, reader engagement measurement can be divided into three broad categories: self-report (such as surveys), physiological (such as observational methods), and web analytics methods (Davoudi, 2018). Nevertheless, three modes of news reader engagement can lead to different levels of engagement, depending on: routine news use (direct traffic), news use triggered by social media (social media traffic), and news use as part of a general search for information (organic traffic) (Möller et al., 2020).

Embracing a multidimensional approach and supported by the literature review performed by Fernandes et al. (2023), the present study adopts seven constructs referred to in the literature (Fernandes et al., 2023). Furthermore, to define the factors that explain user behavior, we considered comprehensive factors (Whetten, 1989) to evaluate reader engagement with the content: recency, frequency, volume, volume of premium content, active days, interactivity, and scroll down, which are explained as follows:

- *Recency, frequency, volume*: in 2016, the Financial Times marketing team adapted the renowned recency, frequency, monetary (RFM) analysis (Stone, 1989) to the media sector by defining the new engagement metric recency, frequency, and volume (RFV) (Goad, 2016). RFM is a simple but effective method used in marketing to analyse consumer purchasing behaviors (Coussement et al., 2014; Shen, 2021). Equivalently to RFM (Stone, 1989), the RFV relies on three dimensions: recency (*R*), frequency (*F*), and volume (*V*). Thus, *R* measures how recently the consumer has visited the website, *F* is the number of website visits within a time period, and *V* indicates the total number of articles read in a period. Those metrics measure the habit, that is, the habitualized behavior for online news readers that read daily, or inhabitualized readers that search for news in social media (Möller et al., 2020) and present low number of visits. Finally, the scores of all three variables are consolidated into a score (FTStrategies, 2022). As result of some successful cases around the media industry, the RFV became a strategic key performance indicator (KPI) across media companies to measure readers' engagement (Goad, 2016; Wenzel et al., 2022; Zontek, 2018). However, other approaches are emerging.
- *Volume of premium content, active days*: The APV engagement score developed in The Independent (UK), combines active days (*A*), volume of premium content read (*P*) and number of articles read (*V*) in a period (Lajumoke et al., 2020). Furthermore, the Wall Street Journal team found a direct correlation between the number of user active days and churn that led them to define *A* as the engagement north star (Seale, 2021). Regarding the importance of churn for newspaper (Peña et al., 2023), a churn model was computed with 1 year data of Público subscribers. Findings revealed that *A* is also a good predictor to predict Público subscriber's churn. It also consists of subscriptions studies in other industries, for example, the telecom market where usage affects churn and loyalty.
- *Interactivity*: Peterson and Carrabis (2008) presented a mathematical linear function that combines some widely used metrics, such as, the number of clicks (C_i) or visit duration (D_i) (Peterson &

Carrabis, 2008). However, the ability to act, interact, and co-create online are key characteristics of online media (Ksiazek et al., 2016). Thus, the involvement of the reader can be expressed as the interactivity that can be measured by the number of interactions with article features like number of comments or likes (Ksiazek et al., 2016).

Other engagement metrics can be found across the literature, such as *dwell* time (i.e., the time spend on a resource) (Davoudi et al., 2019; Grinberg, 2018; Lagun & Lalmas, 2016; Lehmann et al., 2012) or engaged time that measures the amount of time that users spend actively interacting with a page (Schwartz, 2013). However, each platform measures engaged time differently according to their definition of user active interaction. Furthermore, dwell time only provides partial information about reader activity in the page (Grinberg, 2018), and it does not capture reader engagement (Lagun & Lalmas, 2016). In the present research, we decided to measure how far users *scroll down* the article page, that is, the vertical scroll depth of 75%. It means that the reader visited the webpage at least at the end of the article text, providing information about article relevance (Grinberg, 2018) and reader involvement with the content (Barari et al., 2021).

Furthermore, to measure reader engagement, a wide range of metrics are proposed in the literature (Davoudi et al., 2019; Goad, 2016; Grinberg, 2018; Peterson & Carrabis, 2008; Seale, 2021). However, most of the metrics are used isolated and do not cover a broader perspective of the reader engagement as a multidimensional phenomenon (Steensen et al., 2020) related to the level of attention and involvement (emotional, cognitive, and behavioral) (Barari et al., 2021).

Our research analyses three questions. Firstly, we aim to understand the different groups of online news readers. The subscription models, also known as paywall mechanisms, offer a certain number of free articles and there is no direct relationship between the number of paywalls presented and the number of subscriptions (Davoudi et al., 2019). Thus, we aim to answer the question (RQ1): what are the main groups of readers and their characteristics? In fact, the same paywall mechanism for all users may turn away potential subscribers (Davoudi et al., 2019). Furthermore, as media companies can sell content, customer information, and advertising space (Lambrecht et al., 2014), a reader's segmentation allows managers to define different strategies according to the level of engagement and propensity to subscribe. Moreover, researchers acknowledged the relationship between consumer engagement and purchase intention in DSBMs. Thus, marketing and editorial teams should identify consumer's segments, and their main characteristics, across the conversion funnel to maximize consumer's engagement (Lagun & Lalmas, 2016), retention, and consequently conversion (Davoudi et al., 2018; Kotler et al., 2016; Villi & Picard, 2019).

A second aspect that we aim to analyse is differences between clusters that will potentiate editorial or marketing strategies to maximize revenue. Thus, the second research question (RQ2) is: what are the clusters to potentiate advertising revenue? As digital news publishers strive to balance revenue sources in the business models (Vara-Miguel et al., 2021), in this study, we aim to contribute with an operational solution that helps managers to optimize two sources of revenue. For that purpose, results were made available to the teams on a dashboard to improve the daily decision making process.

Lastly, we are interested in understanding the main drivers to subscribe that will help editorial, design, and marketing team to improve the product. We propose the following question (RQ3): What are the main subscription drivers? Thus, this study fills a gap in the literature on media consumer research by proposing a data-driven approach to allow DSBM managers dynamically improve their marketing and editorial strategies.

3 | METHOD

Our research assumes the workflow proposed in Figure 1 that is based on the known CRISP-DM methodology (Chapman et al., 2000; Moro et al., 2011).

The following subsections detail the remainder of the framework steps to analyse the data that was kindly provided by the Portuguese newspaper Público.

3.1 | Engagement window

Firstly, to collect Público consumer's data, and preprocess the raw data set, it was necessary to define the period to calculate the engagement metrics. Thus, an analysis was performed to define the "*engagement window*," that is, the period of significant consumer content involvement prior to subscription. All the user events (Google, 2022) in the 6 months before subscribe, of Público readers that subscribed for the first time between 1st September, 2021 and 15th June, 2022, were analysed. Thus, a total of 12,721,603 events were considered, from 10,294 new Público subscribers. An example of a frequent event is "*Ler Mais—click*" indicates that the reader clicked at the "Ler Mais" box, that is a recirculation element at the article page. The average number of events by day was computed to enable a time series analysis that shows the evolution of readers events on the website before subscribed.

The goal was to identify changes in the time series. Thus, to solve this changepoint detection problem, the standard deviation was used as a cost function (Katser et al., 2021). The content consumer behavior pattern has a cycle of a week, thus the standard deviation was calculated for 7 and 14 days. A sliding window through the 6-month time series, starting at the subscription day, presented maximums between the 34th and 39th day before subscribe (as presented in Figure 2). Both changes in mean and variance indicate a reader's behavior change through the increase in the number of events by day prior to subscription. Thus, the engagement window considered to calculate EA was 30 days, which is also a frequent period of reader analysis in the media industry (Blazejewski, 2019; Jacob, 2021).

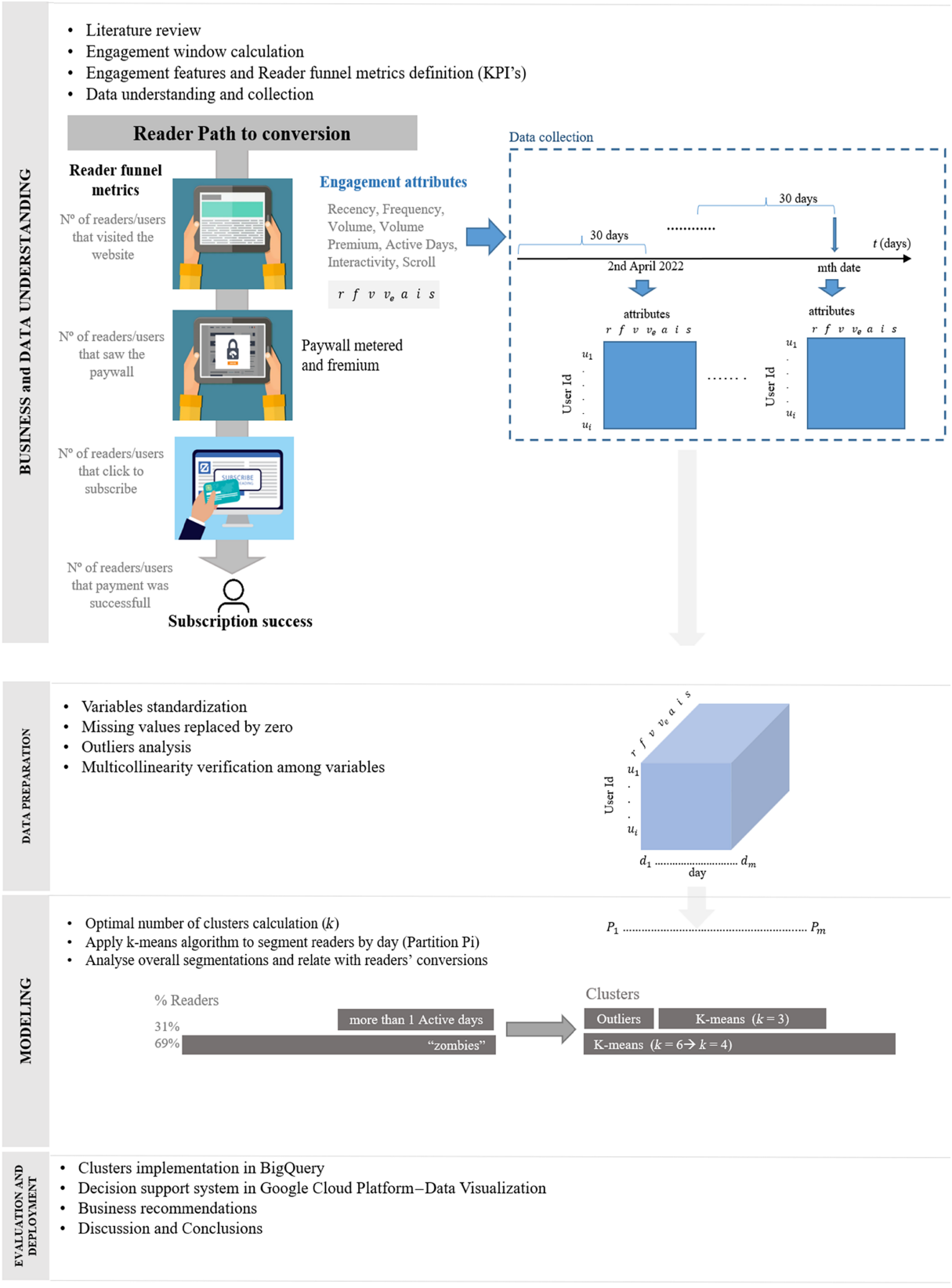

FIGURE 1 Research procedure. KPI, key performance indicator.

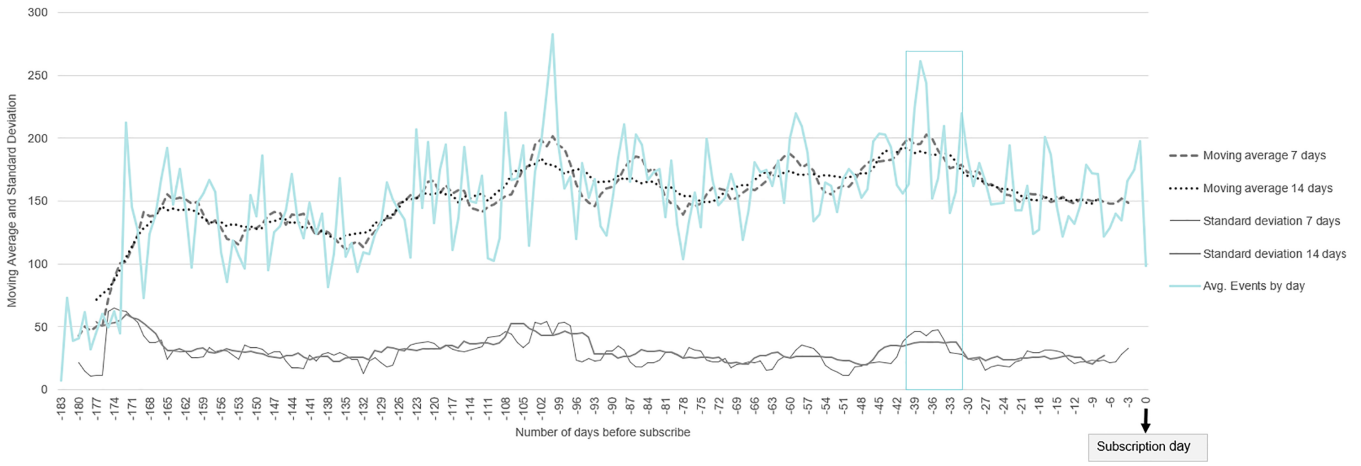

FIGURE 2 Changepoint detection across the time series.

3.2 | Data understanding and data preparation

A real dynamic data set was used in this research that is stored at Bigquery, a Google-managed data warehouse (Bisong, 2019; Mucchetti, 2020). Data were accessed through Colaboratory (see Figure 3), a product from Google Research that is a hosted Jupiter notebook service with access to Google hardware (Nelson & Hoover, 2020). BDA was performed by programming Python scripts in Colaboratory. ML models were automated, and results are presented into Looker dashboards (see Figure 3).

The data set contains the *date*, *user id*, and the engagement *features* calculated for a 30 days timeframe (see Table 3). Data values were stored on a daily basis, encompassing a period ranging from 2nd April to 1st June, 2022. A COVID-19 period (Chopdar et al., 2022; Kursan Milaković, 2021; Nayal et al., 2022) where the Strindency Index (Global Change Data Lab, 2021) was lower than 20, revealing a low level of government restrictions.

At BigQuery, the authors developed a table that calculates daily the recency, frequency, volume, volume of premium content, active days, interactivity, and scroll down by user. Across time, the main engagement metrics by user are saved to have historical data and monitor engagement evolution.

Missing values were replaced by zero as zero indicates that no user visit was registered. The zero-mean normalization method was used as attributes present different scales. Furthermore, as clustering methods are sensitive to outliers, those were considered as an individual cluster. For each column of the data set, the absolute *Z*-score was calculated, and readers with values greater than 3 were grouped into a cluster (i.e., data points that fall outside of 3 standard deviations were classified as outliers).

3.3 | Data modeling: clustering analysis

K-means is an efficient clustering method (Koul & Philip, 2021) that groups the data based on their closeness to each other (Antonio et al., 2022). The cluster's center point is the mean of that cluster, and the other points are the observations that are nearest to the mean value. Furthermore, the number of clusters is chosen in advance, and it uses an iterative procedure that minimizes the squared error of the following objective function:

$$J = \sum_{i=1}^{D} \sum_{K=1}^{K} \left\| X_{i(K)} - A_K \right\|^2,$$

where $\|...\|^2$ is a specific distance function, D are the data points, K are the clusters, $X_{i(K)}$ are the points in cluster K, A_K are the centroids of clusters K.

In a clustering analysis to determine the number of cluster, (K) is crucial (Fu et al., 2017). As no metric can guarantee optimal results, the present approach uses a combination of three widely used metrics (Gustriansyah et al., 2019; Xiao et al., 2017) presented in Table 2.

In the news domain, two cold start problems are frequently studied: the *item cold start*, that is, when a new article is published; and, *user cold start*, that is, when a new or unknown user visit the website (Delpisheh et al., 2016; Zihayat et al., 2019). This problem becomes more prominent when light users are dominant (Liu et al., 2022), approximately 70% of readers only visit the website once in 30 days (see Figure 1). Thus, the data set was divided into two samples. The first sample contains consumers that visited the website once in a month, frequently called "zombies" at the industry literature (Jacob, 2021; Lynes, 2021). The second group contains readers than present more than one active day in 30 days (Sample 1 at Table 3).

To define a segmentation of 30-day users (see Figure 1) at each day of the timespan, a dataframe was collected and a segmentation was performed. The main goal was to define the optimal number of clusters that daily need to be calculated. For both samples, we stored the data values daily, encompassing a period ranging from 2nd April to 1st June, 2022 (61 days).

Next steps present a 1-day users' segmentation of readers from Sample 1, at the 1st of June, that represents approximately 1.4 million

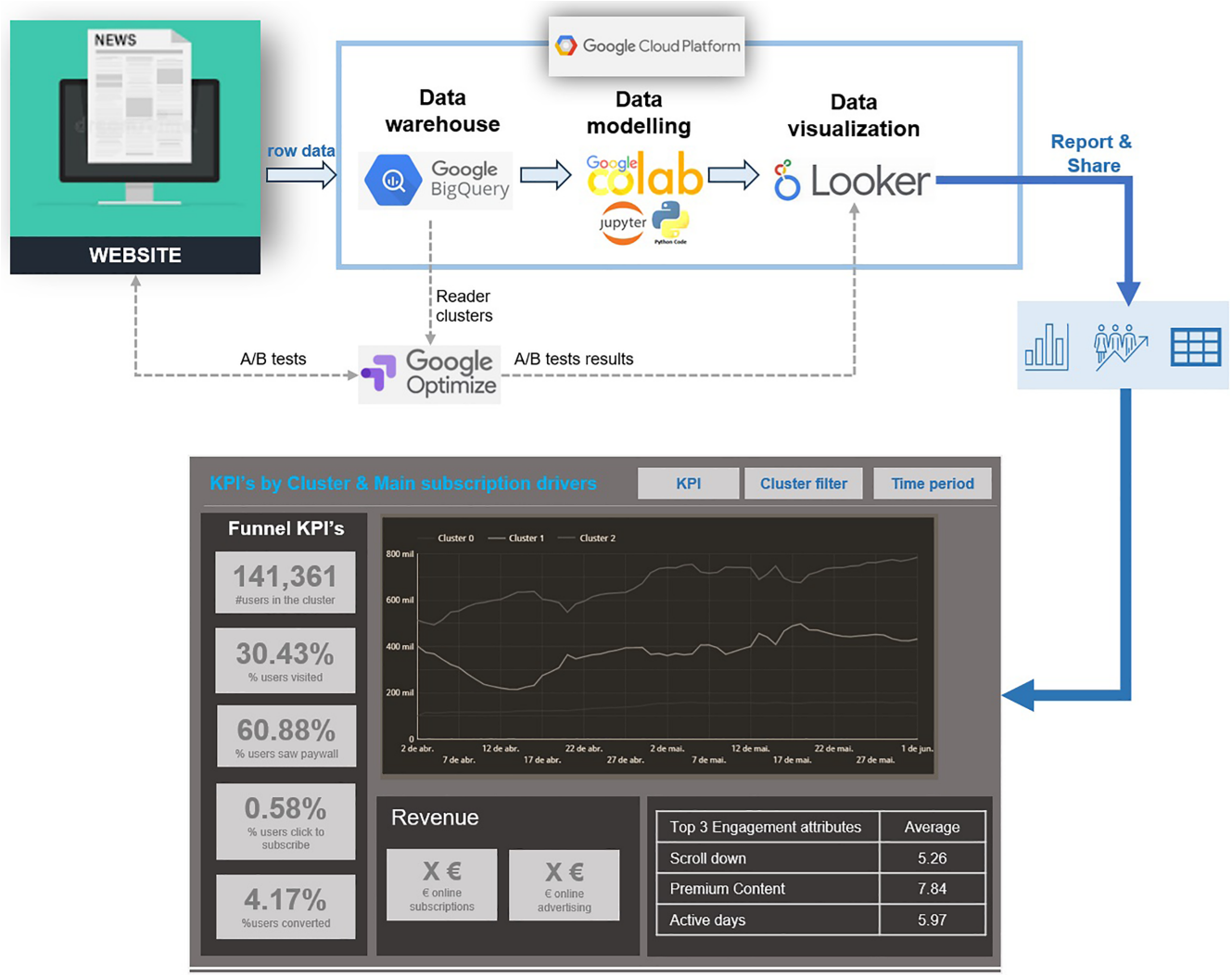

FIGURE 3 Decision support system designed in the Google Cloud Platform. Marketing and editorial teams can consult Looker dashboard to continuously monitor engagement and conversion key performance indicator (KPI), the main driver to subscribe, and the revenue.

users. Furthermore, Table 4 presents Pearson correlations (Hauke & Kossowski, 2011) among engagement variables. As expected, *V* and VP present a very high correlation. Despite the fact that a high correlation among clustering variables can be problematic (Ketchen & Shook, 1996), we decided to keep both because, in this business context, the content type is important to define strategies. Besides, Table 4 also indicates that an increase in *A* will increase *F*, as well as *V* and VP. Moreover, *I*, *R*, and *S* does not present strong linear correlations with other variables.

Then, to define the optimum value of clusters *K* (i.e., when inter-cluster dissimilarity and intra-cluster similarity were maximized), the three metrics mentioned in Section 2 were calculated. The number of clusters is optimal when the marginal gain of adding a cluster drops dramatically in the distortion score (DS) ("elbow criterion"), and when the CH score is maximum, and the DB score is minimum. On 1st of June, as presented in Figure 4, the optimal number of clusters (*K*) for Sample 1, is three.

Opposite to Sample 1, Pearson correlation is low between all EA except between *V* and VP that presents a moderate correlation (Table 5). Furthermore, Figure 5 presents the optimal number of clusters by metric. The DS score indicates six as the best number of clusters. However, CH score is maximum at 6 and 8, whereas DB score does not present an absolute minimum (Vergani & Binaghi, 2018). Thus, by combining the three measures, six seems to be the optimal number to partitionate "zombies."

4 | RESULTS, GENERAL DISCUSSION, AND IMPLICATIONS

tSeeking to answer the main research questions, this section presents: Section 4.1 presents the main clusters of readers obtained (RQ1) and the clusters were advertising revenue can be optimized (RQ2), followed by Section 4.2, which presents the main drivers to subscribe (RQ3).

TABLE 2 Metrics to determine the number of clusters (k).

Metric name	Definition and formula	Optimum point
Distortion score (DS) (Ketchen & Shook, 1996) or Within Cluster Sum of Squares (WCSS)	The sum of squares of distance between points (d) and cluster centers (C) $\mathrm{WCSS} = \sum_{C_n}^{C_K}\left(\sum_{d_i \text{ in } C_i}^{d_m} \text{distance}\,(d_i, C_K)^2\right)$	The value k when we increase the number of clusters the new is very near to some of the exiting and the WCSS decreases slowly
Calinski-Harabasz score (CH) (Calinski & Harabasz, 1974)	It consists in a ratio between inter-cluster covariance and intra-cluster divergence. It is an evaluation metric based on the degree of dispersion between clusters $\mathrm{CH}(K) = \frac{\left(\sum_{k=1}^{K} a_k \lVert \overline{x_k} - \overline{x} \rVert^2\right)(n-K)}{\left(\sum_{k=1}^{K}\sum_{c(j)=K} \lVert x_j - \overline{x_k} \rVert^2\right)(K-1)}$	The higher the ration is, the better the clustering effect is
Davies Bouldin score (DB) (Davies & Bouldin, 1979)	Calculates the average of the similarity measures of each cluster with a cluster most similar to it $\mathrm{DB}(K) = \frac{1}{K}\sum^{k} \max_{i \neq j}\left((s_i + s_j)/d_{ij}\right)$, where s_i is the cluster diameter and d_{ij} the distance between cluster centroids i and j	The value k where the average similarity is minimized

TABLE 3 Descriptive statistics of the engagement' attributes (EA) by sample from 2nd April to 1st June, 2022.

		Descriptive statistics				Descriptive statistics			
		Sample 2 "zombies" (n = 10,326,306 users)				Sample 1 "+1 active day" (n = 3,379,611 users)			
EA	Definition	Avg.	Std.	Min.	Max.	Avg.	Std.	Min.	Max.
R	Number of days since the last visit (high value means that reader made a visit recently)	−15.6	8.5	−30.0	−2.0	−10.5	7.7	−30.0	2.0
F	Number of different hours with at least on visit	1.2	0.7	1.0	664.0	6.9	14.1	1.0	696.0
A	Number of different days with at least one visit	1	0	1.0	1.0	4.4	4.3	2.0	30.0
V	Number of articles read	1.2	0.8	1.0	239.0	5.7	12.5	1.0	2484.0
VP	Number of premium articles read	1.1	0.6	1.0	55.0	3.4	7.2	1.0	735.0
S	Number of times that the reader achieved at least 75% of the webpage	1.4	2.5	1.0	768.0	5.1	31.3	1.0	13,935.0
I	Number of times that a reader shared an article by one social media options, or by email, or at least commnet the article.	1.6	1.8	1.0	139.0	4.9	29.9	1.0	3467.0

TABLE 4 Pearson correlation between variables after standardization.

	R	*F*	*V*	VP	*A*	*I*	*S*
R	1.00	0.16	0.14	0.12	0.29	0.01	0.04
F		1.00	0.65	0.61	0.74	0.13	0.36
V			1.00	0.95	0.62	0.23	0.30
VP				1.00	0.57	0.24	0.28
A					1.00	0.07	0.20
I						1.00	0.07
S							1.00

4.1 | Group profiling

In the period under analysis, all content consumers were impacted with the same marketing campaigns and communications. Thus, the click action happened under the same marketing *stimulus* (Stroud & van Duyn, 2020), despite different levels of content involvement. Seeking to answer RQ1, readers of Sample 1 were grouped into four clusters that are referred to Cluster_0, Cluster_1, Cluster_2, and Cluster_3. As presented in Table 6, larger values indicate high engagement levels. As we aim to identify the main characteristics of the detected clusters (RQ1), each cluster is detailed as follows:

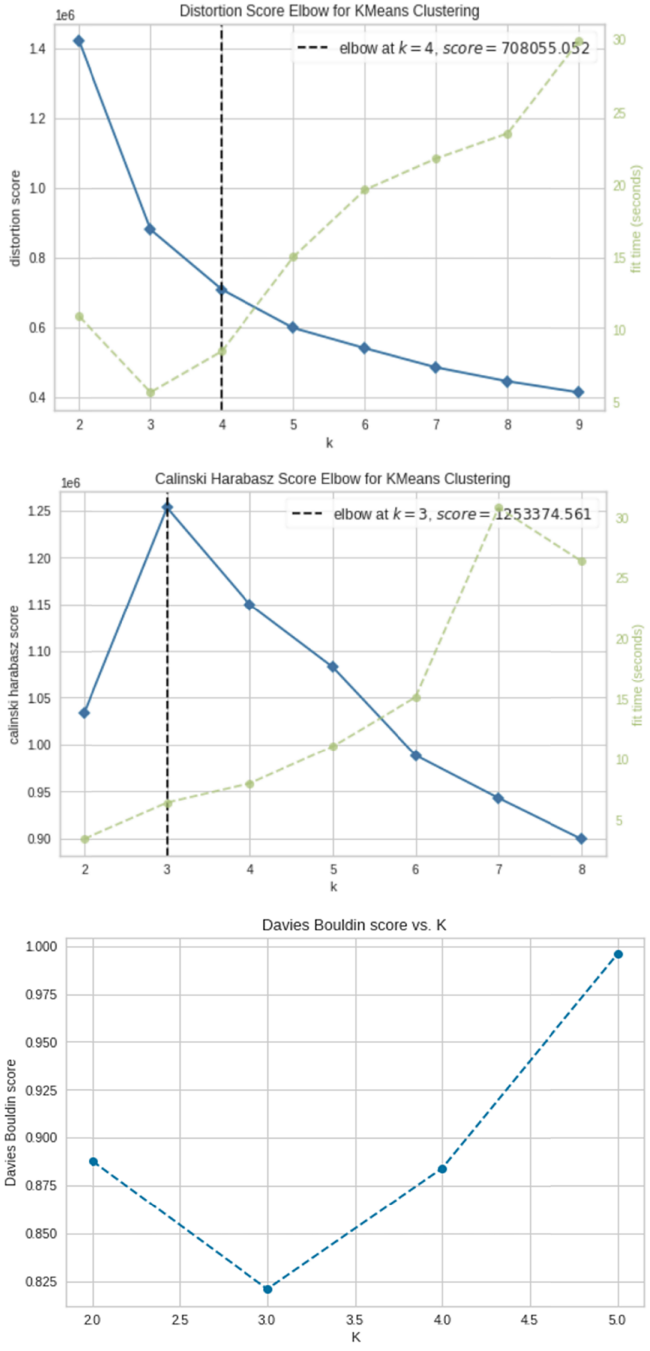

FIGURE 4 Evaluation metrics to find the optimal number of clusters for users with more than one active days at 1st June.

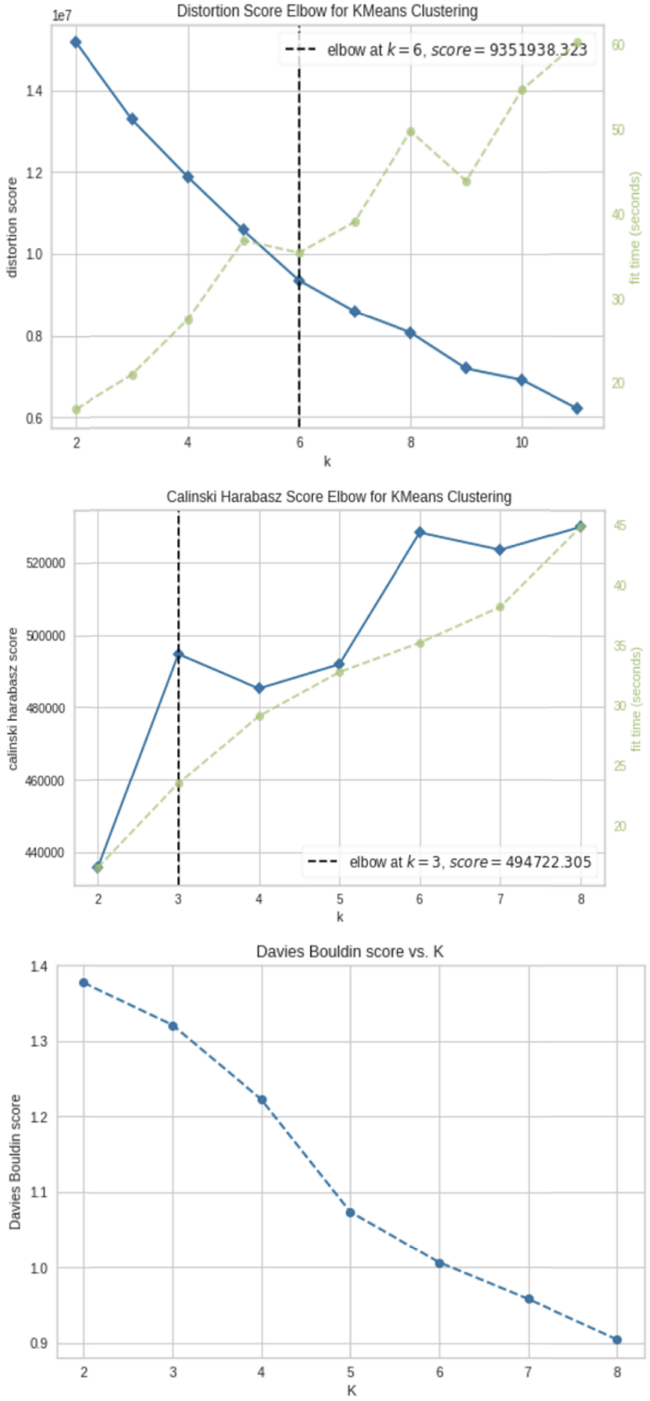

FIGURE 5 Evaluation metrics to find the optimal number of clusters for the "zombies" (sample 2).

TABLE 5 Pearson correlation between variables after standardization (Sample 2).

	R	*F*	*V*	VP	*I*	*S*
R	1.00	−0.004	0.017	0.014	0.000	0.002
F		1.00	0.181	0.116	0.048	0.149
V			1.00	0.577	0.142	0.240
VP				1.00	0.105	0.136
I					1.00	0.033
S						1.00

- *Cluster*$_0$ (*active*) contains approximately 11% of the sample, where around 30.43% visit the website in a day. From those readers, 61% see the paywall at least once, and then 0.58% click to subscribe. These are recent visitors, with high number of *A* and consequently high values of *F*, *V*, and VP. They are valuable readers with strong content involvement that increases ad revenue. However, they do not present the highest click and conversion rates.

TABLE 6 Engagement attributes average and reader funnel metrics by cluster (Sample 1).

	EA	Cluster 0 $n = 141{,}361$ 11%	Cluster 1 $n = 373{,}631$ 30%	Cluster 2 $n = 670{,}002$ 55%	Outliers $n = 49{,}356$ 4%
Engagement attributes	*R*	−6.01	−20.36	−6.51	−2.00
	F	9.17	4.39	7.25	9.71
	V	7.84	3.79	6.21	6.32
	VP	3.51	2.44	3.76	2.62
	A	5.97	2.97	4.62	6.77
	I	4.48	3.98	5.58	1.97
	S	5.29	3.62	5.53	4.66
Reader funnel metrics	Proportion of users that visited the website (%)	30.43	8.27	23.63	87.15
	Proportion of users that saw the paywall (%)	60.88	51.04	52.95	21.80
	Users that click to subscribe (%)	0.58	0.60	0.96	0.81
	Users that converted (%)	4.17	5.41	4.70	90.66

Abbreviation: EA, engagement' attributes.

TABLE 7 Engagement attributes average and reader funnel metrics by cluster (Sample 2).

	Variables	Zombies 0 $n = 276$ <0.5%	Zombies 0 $n = 1916$ <0.5%	Zombies 0 $n = 38{,}185$ 1%	Zombies 1 $n = 672{,}033$ 23%	Zombies 2 $n = 1{,}041{,}994$ 35%	Zombies 3 $n = 1{,}201{,}960$ 41%
EA	*R*	−16.36	−15.68	−16.04	−15.43	−23.67	−9.05
	F	9.13	2.04	3.13	1.12	1.10	1.10
	V	7.90	4.56	4.92	1.21	1.06	1.06
	VP	3.73	3.03	2.60	1.05	0.00	0.00
	I	1.00	5.05	1.37	1.16	1.13	1.13
	S	86.86	2.70	4.17	1.29	1.19	1.14
Reader funnel metrics	Proportion of users that visited the website (%)	2.54	0.63	1.36	1.62	0.71	1.57
	Proportion of users that saw the paywall (%)	42.86	41.67	36.68	39.67	36.09	34.70
	Users that click to subscribe (%)			0.53	0.46	0.30	0.50
	Users that converted (%)				10.00	12.50	12.12

Abbreviation: EA, engagement' attributes.

- $Cluster_1$ (*moderate*) corresponds to the readers that did not visit the website recently, they present less *A* and also less *S*, revealing low engagement with the content. In fact, only 8.27% of the cluster users return to the website, and of those, 51.04% see the paywall at least once. Only 0.6% click, but after clicking, they are more likely to convert than $Cluster_0$ and $Cluster_2$ (4.70%). Those results indicate that there is an opportunity to increase the conversion rate among readers who present higher engagement levels.
- $Cluster_2$ (*need attention*) constitutes the highest cluster (55%). They reflect a segment of readers that visited the website recently. But, despite not being active readers, after seeing the paywall they present the highest click rate. After manual inspection, we verified that such readers were attracted by premium content (from push notifications) and articles that induced button subscription click.
- *Outliers* (*super*) is the smaller cluster. However, 87.15% return to the website, indicating that the last visit was recent ($R = -2$) and *A* is the highest ($A = 6.77$). As a result, those readers present high values of *F*, *V*, and *S*. Nonetheless, *I* presents low value.

In what concerns to "*zombies*," clustering was performed considering the remaining six EA as *A* is constant, equal to 1.0. Furthermore, on average, those readers present low values of *F*, *V*, *I*, and *S* (see Table 3) indicating there is a need of recirculation strategies (Lioudis, 2019) to improve engagement levels (Barari et al., 2021). Also, they present an average of 1.4% website return that indicates an opportunity to a multichannel strategy improvement, for example, social media or notifications actions (Hullar, 2020; Loni et al., 2019). Moreover, the click rate is low (see Table 7) indicating inhabited

readers, that is, readers (Möller et al., 2020) with low levels of engagement are related to a low propensity to subscribe (Barari et al., 2021). Those readers are mostly mobile readers, highly related to high propensity to churn (Peña et al., 2023). However, those readers play an important role on advertising revenue (Arrese, 2016), as they represent 20% of website pageviews (RQ2). The characteristics of the clusters are detailed as follows:

- *Zombies* 0 (*casual*) consists of the first three clusters presented in Table 7 that represent less than 2% of the sample. They were grouped because they do not present significant click rate and conversion rates. Furthermore, the last visit was around 16 days ago and between all the zombies, they present better content consumption ($V > 4$).
- *Zombies* 1 (*sleepers*) account for 23% of total "zombies." This cluster contains readers who, on average, made their last visit 15 days ago. Furthermore, only 1.62% return to the website; from those, 39.67% see the paywall, where 0.46% click and 10% of those that clicked subscribe.
- *Zombies* 2 (*lost*) constitutes 35% of the sample. These readers do not perform well, the last visit was 23 days ago. Only 0.71% returned to the website and only 0.3% saw the paywall click to subscribe. However, 12.5% that click on the paywall subscribe, suggesting that there is an impulse related to breaking news.
- *Zombies* 3 (*recent visitors*) the larger cluster (41% of the sample) corresponding to readers that visited the website recently (9 days). However, they present lower *V* and only 1.57% return to the website.

For comparison of the results between clusters, the data were analysed with *t*-test (two-sample assuming unequal variances) and *p*-value. The independent samples *t*-test showed a significant difference between clusters ($p < 0.01$). Furthermore, to monitor the evolution of cluster profiles and to continuously improve the digital strategy, a business intelligence (Tripathi et al., 2023) solution was developed into Looker dashboards (see Figure 3). Moreover, EA and reader funnel metrics by cluster allow the team to see the impact of marketing and editorial actions. For instance, special events in news can make the daily editorial strategy vary according to the cluster's characteristics. As an example, zombies could be impacted by notifications while outliers can receive an editorial newsletter with more content about the special event.

4.2 | The relation between engagement metrics and intention to subscribe

To explore the main drivers to subscribe and answer our third research question, that is, what are the predictors of consumer's WTP (RQ3), we considered a data set with 189,000 users from Sample 1 that viewed the paywall. To start the payment flow, each reader may click or not click in the paywall, value 1 and 0 (see reader flow at Figure 1). Furthermore, we split the database into the training and testing subsets on a ratio of 0.75/0.25. The training set is used for establishing the models, while the testing set evaluates the prediction. XGBoost optimizes the four performance metrics (see Table 8). To avoid overfitting, the percentage of features used per tree was set to 50%. Moreover, we oversample the minority class to obtain a more balanced data distribution by using Synthetic Monority Oversampling Technique (Chawla et al., 2022), six neighbors were found that maximize the mean *A*rea *U*nder the receiver operating characteristic *C*urve (AUC). As XGBoost is a tree-based method, it does not require standardization (Wieland et al., 2021). Furthermore, to explain how much each EA contributed to the model's prediction when compared to the mean prediction *SH*apley *A*dditive ex*P*lanations (SHAP) values were calculated (Lundberg & Lee, 2017). In the particular case of a binary model, SHAP values give the difference between the predicted log odds and average predicted log odds (Wieland et al., 2021).

TABLE 8 Performance of classification methods (best values are highlighted by using a boldface font).

Classification method	Accuracy	F1 score	AUC	RMSE
LR	0.915	0.019	0.676	0.292
KNN	0.987	0.099	0.696	0.113
XGBoost	**0.993**	**0.126**	**0.752**	**0.088**

Abbreviation: AUC, *A*rea *U*nder the receiver operating characteristic *C*urve; KNN, K-Nearest Neighbor; LR, Logistic Regression; RMSE, *R*oot *M*ean *S*quare *E*rror; XGBoost, e*X*treme *G*radient *B*oosting.

Given the popularity and results achieved by three ML classifiers, the authors will apply the Logistic Regression (LR), K-Nearest Neighbor (KNN), and XGBoost (Chen & Guestin, 2016; Coussement et al., 2014). The obtained results are calculated and evaluated in terms of accuracy (Labatut & Cherifi, 2012), F1 score (Lipton et al., 2014), AUC (Gonçalves et al., 2014), and *R*oot *M*ean *S*quare *E*rror (see Table 8). LR and KNN were widely applied in DSBMs, such as churn modeling (de Caigny et al., 2018; Shahraki et al., 2017). Moreover, as XGBoost algorithm is an implementation of gradient boosted decision trees designed for speed and performance (Brownlee, 2016), it has achieved superior results in several ML challenges (Brownlee, 2016; Chen & Guestin, 2016; Wieland et al., 2021).

In Figure 6, the ranking indicates that *S* is the most important driver for the click action; thus experiments on design can improve the user experience to increase the scroll down. As brand experience plays a pivotal role in brand trust and brand love (Joshi & Garg, 2021), improving the user experience is also crucial to increasing the scroll down propensity.

Furthermore, VP was identified as the second major driver followed by *A*. Click on subscribe is more likely to happen through increasing *S*. In contrast, high values of *A* have a high negative contribution on subscribe. An explanation for this is derived from the high proportion of readers that only read the first paragraph of the article. Furthermore, *V* also presents a negative contribution, indicating that when a reader is impacted by the paywall, he is less likely to subscribe.

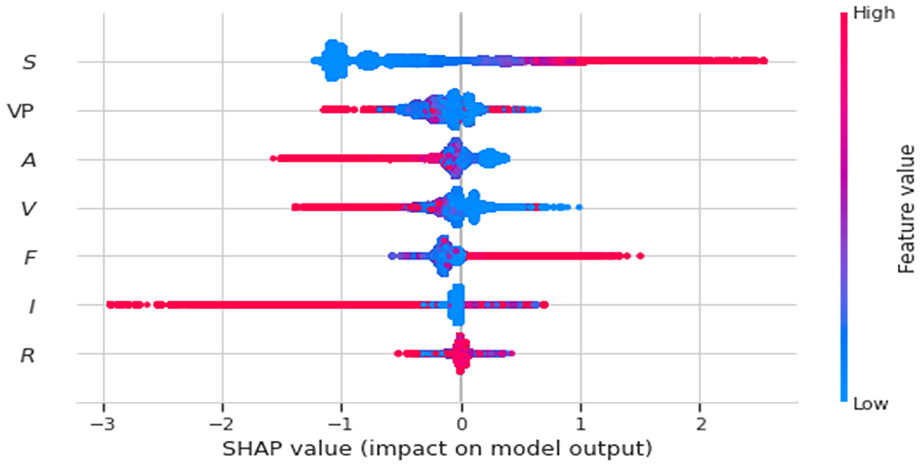

FIGURE 6 Engagement features ranking by *SH*apley *A*dditive ex*P*lanations (SHAP) values of the model (Lundberg & Lee, 2017). Each point in the cloud represents a row from the original data set. The color code denotes high (red) to low (blue) feature values.

These finding is in line with the study of (Davoudi et al., 2018) that argues the lack of assertively in the metered and freemium models. Thus, there is an opportunity to innovate the business paywall model to increase the propensity to subscribe by applying advanced analytics models that can optimize the revenue from subscriptions and/or advertising (B2C online business model).

The effect of *R* comes almost at the bottom, indicating that, in the sample under study, *R* seems to have little effect on the intention to subscribe. Surprisingly, *I* also has low effect, despite being a widely studied engagement feature. Moreover, the engagement drivers to subscribe could be daily monitored to understand the action's impact and analyse the SHAP ranking fluctuations. The fact that the segmentation will be running daily into bigquery (see Figure 3) will help the marketing team monitor engagement and funnel metrics by cluster, as presented in the dashboard in Figure 3.

4.3 | Discussion of managerial implications and contributions

In terms of managerial implications, a two-fold approach is proposed. Readers of Sample 1 have more intention to subscribe. Thus, by target marketing strategies the team can develop customized plans and schemes (Kotler et al., 2016). Furthermore, authors explored new paywall mechanisms (Davoudi et al., 2018) and the impact of paywall design (Aral & Dhillon, 2021) that could be adapted for tailoring strategies by cluster. Moreover, recommendation algorithms could be used to increase engagement, for example, in newsletters to provide a more personalized experience and increase perceived value.

Meanwhile, at Sample 2, the main challenge is to increase visits to the website, that is, invest in multichannel strategy focused on the three modes of news reader engagement that corresponds to different levels of engagement (Möller et al., 2020) and consequently maximize advertising revenue (RQ2).

Authors agree that push notifications drive more engagement than newsletters and it is a fast and effective way to inform users (Budiman & Akhlis, 2021; Gao et al., 2018; Loni et al., 2019). Thus, it could be an opportunity to develop a specific experience for "zombies" to induce them to receive notifications. Despite, too many notifications can create a poor experience, researchers developed optimization systems for notifications (Gao et al., 2018) and personalized systems for this business context (Loni et al., 2019). Furthermore, as those readers are mostly mobile readers, the user experience focused on mobile is also a point of concern for authors (Peña et al., 2023). Moreover, social media strategy can have an important role in increasing users and engagement in these clusters (Cage et al., 2020).

From an operational perspective, the results obtained allow online news editorial and marketing teams to define strategies to retain readers and/or acquire subscribers. Furthermore, the results are understandable for the business. By providing the main characteristics of the customers, the teams can apply business actions more assertively. User segmentation helps in increasing customer retention, loyalty, and to identify the value of each user (Koul & Philip, 2021). The eight clusters are labeled as *super*, *active*, *moderate*, *need attention*, *casual*, *sleepers*, *lost*, and *recent visitors*. These are simple terms that help the newsroom have a better understanding of reader's types (RQ1). The proposed approach was automated to present continuous consumer segmentation over time with an effective method that is simple to understand (Koul & Philip, 2021). Those clusters can be monitored by querying raw data from BigQuery and sending aggregated data to a Looker dashboard (see Figure 3). The dashboard can combine information about number of readers by clusters, average engagement attributes (RQ1), and average conversions across the reader funnel and advertising revenue (RQ2). Furthermore, Marketing and Editorial teams can implement A/B tests across the website, by cluster, using Google Optimize (an online split-testing tool that enables the user to experiment different ways of delivering the content). Moreover, the engagement drivers to subscribe could be daily monitored (RQ3) to understand the action's impact and also to analyse the SHAP ranking fluctuations. In addition, the proposed approach has the potential to be transferred to other online domains, for example, digital subscription businesses, such as those used by the music or games industries.

In terms of theoretical contributions, our research results are enlightening for model development into B2C online platforms in

customer relationship management. Firstly, we made some contributions to the existing literature on online user segmentation in the B2C digital subscriptions businesses. In this research, online engagement features were selected based on the literature and previous consumers' segmentations in DSBMs. A clustering model was proposed to identify the main drivers to subscribe. Although a wide variety of online DSBM exists, not all the variables have the same impact on subscription intention. This research shows that other engagement variables can be considered according to the type of content (visual and audio content). However, the method proposed clarifies how to identify the main online users' clusters and the users most likely to subscribe in an interpretable way, with a dashboard, to optimize marketing and editorial strategy, which is an achievement of this paper. Thus, we presented original research that has high utility (Corley & Gioia, 2011) for B2C digital subscriptions businesses.

5 | CONCLUSION AND LIMITATIONS

In this study, we provide a strategy to segment online users and identify the main drivers to subscribe. Each consumer presents different reading interests and diverse content consumption patterns. Furthermore, consumer engagement requires a long-term investment. Thus, in a Big Data context, it is impractical to apply the same strategy to engage all users. Reader segmentation is a challenge that combined with conversion data provides useful information to improve acquisition and retention strategies for online news publishers. Hence, the proposed method consists of attributes retrieval, clustering analysis, and business recommendations. Attributes retrieval is an important and time-consuming step that was supported by the literature research. Different engagement features can influence the results and lead to different clusters interpretations that can result in different marketing and editorial strategies.

Furthermore, the presented approach can be adapted and replicated to other online subscription businesses (DSBM) or B2C e-commerce businesses. Results obtained in previous sections can be studied in two complementary ways. Firstly, we can see the characteristics of each cluster in terms of engagement to implement multichannel and recirculation strategies, to improve the digital product, and to increase the perceived value. Then, we can study how engagement is related to intention to purchase to develop assertive strategies.

One way to personalize and optimize the digital experience is to separate users into groups and expose each group to the product experience that most fits the group's characteristics. In the present research, readers are divided into eight clusters, labeled as *super*, *active*, *moderate*, *need attention*, *casual*, *sleepers*, *lost*, and *recent visitors*. The new model identifies behavior patterns by cluster. From an operational perspective, the automated solution allows editorial and marketing team to define business strategies by monitoring the main KPI's across the funnel into Looker dashboards (see Figure 3).

Furthermore, the classification algorithm provides information about the main subscription drives in Sample 1. The derived results reveal that the number of times that the reader scrolled depth in the article page is a strong driver to increase the propensity to subscribe, followed by the volume, and active days. This information induced the team to improve the user experience to increase user loyalty, and to adapt the content distribution strategy to increase the number of visits to the website.

Digital news publishers strive to balance revenue sources in the business models and few studies have explored the factors influencing users' intention to pay for such services. Thus, the analysis of causal relationships between EA and consumer's intention to subscribe is an important contribution allowing managers to identify the main predictors of consumer's WTP. A two-fold strategy is proposed to guarantee engagement increase and conversion increase. Hence, a multichannel strategy improvement is recommended to motivate readers to subscribe newsletters or notifications to increase reach and engagement by device. Furthermore, the use of recommendation algorithms can play a pivotal role to increase website recirculation and consequently increase *V*, *I*, and *S*. Meanwhile, segmented marketing and design strategies can induce a *S* increase and click rate increase. Besides, the advertising strategy could be segmented to increase this source of revenue according to cluster goals.

This proposal is of interest of the news publisher, as it presents an useful and understandable model based on EA that are actionable, reliable, and readable. Moreover, the model was automated in the Google ecosystem allowing the teams to easily monitor values, implement website tests, and measure daily results impact into the Looker dashboards. As argued by (Tripathi et al., 2023), business intelligence tools help the company to rapidly generate insights to guide managers toward operational efficiencies, lead them to new opportunities. Thus, in this study the information is presented into a dashboard that will allow managers to improve user experience by improving the digital product and potentiate brand loyalty. The news publisher can ultimately improve the user experience, the content offer, and the multichannel content distribution to increase reader loyalty, engagement, and consequently maximize revenue.

The scope of future work lies in two main goals: to study recommendation algorithms to improve recirculation, and to improve multichannel strategy to face the user cold start problem. Both research lines fully related to the main consumer future research trends identified in the study of Paul and Bhukya (2021), that is, analyse consumer behavior and purchase intentions (Paul & Bhukya, 2021) to improve consumer satisfaction, loyalty awareness, and involvement.

Furthermore, research indicates that during the COVID-19 pandemic, journalism played a crucial role in society and democracy by increasing awareness and distributing accurate content (Perreault & Perreault, 2021). Moreover, COVID-19 has resulted in structural changes in consumer behavior (Gordon-Wilson, 2022; Kursan Milaković, 2021; Nayal et al., 2022; Purohit et al., 2022; Rayburn et al., 2021; Yap et al., 2022). For example, consumer behavior has undergone changes and online transactions have become part of people's life in many sectors such as healthcare, hospitality, financial services, food delivery (Chakraborty & Paul, 2022; Chopdar et al., 2022; Kim & Im, 2022; Siddiqi et al., 2022). As an outcome, we need new theories, scales, methods, and paradigms to carry research studies in

the post-pandemic era to analyse the new processes, patterns, and problems.

Despite its benefits and contributions, there are still limitations. This study models and analyses the behavior of different reader clusters in one news publisher. Furthermore, the proposed approach can be implemented in other content B2C subscription business models, such as e-books, digital publications, music, courses, or games. Those businesses also have as their main source of revenue advertising, affiliate marketing and sales, subscription, freemium, one-time purchase, and pay-per-use. However, future research can consider other segmentations from different media publishers that can lead to a strong benchmark for the media industry.

Moreover, k-means was selected as it is a widely used method. However, other segmentation algorithm or attributes could be applied. Besides, the 30-day period of metrics calculation is aligned to the most frequent practices across the industry; future work could be exploring the impact of engagement window variation.

ACKNOWLEDGEMENTS

This work was supported by the FCT—FCT—Fundação para a Ciência e Tecnologia, under the Projects. UIDB/04466/2020, UIDP/04466/2020, UID/CEC/00319/2019, and UIDB/50021/2020.

We would like to thank Público Comunicação Social S.A. for providing the data set used in this research.

CONFLICT OF INTEREST STATEMENT

No potential conflict of interest was reported by the author(s).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Elizabeth Fernandes https://orcid.org/0000-0002-2358-1229
Sérgio Moro https://orcid.org/0000-0002-4861-6686
Paulo Cortez https://orcid.org/0000-0002-7991-2090

REFERENCES

Antonio, N., Rita, P., & Saraiva, P. (2022). Effectiveness of COVID-19 vaccines: Evidence from the first-year rollout of vaccination programs. *Vaccine*, *10*(3), 409.

Aral, S., & Dhillon, P. S. (2021). Digital paywall design: Implications for content demand and subscriptions. *Management Science*, *67*(4), 2381–2402. https://doi.org/10.1287/mnsc.2020.3650

Arrese, Á. (2016). From gratis to paywalls: A brief history of a retro-innovation in the press's business. *Journalism Studies*, *17*(8), 1051–1067. https://doi.org/10.1080/1461670X.2015.1027788

Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. *Journal of Business Research*, *88*, 407–414. https://doi.org/10.1016/j.jbusres.2017.11.047

Ballew, B. (2009). Elsevier's Scopus® database. *Journal of Electronic Resources in Medical Libraries*, *6*(3), 245–252.

Barari, M., Ross, M., Thaichon, S., & Surachartkumtonkun, J. (2021). A meta-analysis of customer engagement behaviour. *International Journal of Consumer Studies*, *45*(4), 457–477. https://doi.org/10.1111/ijcs.12609

Bisong, E. (2019). Google bigquery. In *Building machine learning and deep learning models on Google cloud platform* (pp. 485–517). Apress.

Blazejewski, P. (2019). *User engagement drives subscriptions—RFV user engagement scores from deep.BI*. https://medium.com/deep-bi/user-engagement-drives-subscriptions-rfv-user-engagement-scores-from-deep-bi-8aa1ed23a923 Accessed 15th February 2024.

Bock, H. H. (2007). Clustering methods: A history of k-means algorithms. In *Selected contributions in data analysis and contribution* (pp. 161–172). Springer Science & Business Media.

Bomnüter, U., Hansen, N., Beuthner, M., König, U., & Hennig-Thurau, T. (2022). More important than ever before? Assessing readers' willingness to pay for local news as a constituent for sustainable business models. *Journal of Media Business Studies*, *20*, 1–24.

Brownlee, J. (2016). XGBoost with python: Gradient boosted trees with XGBoost and scikit-learn. In *Machine learning mastery*.

Budiman, K., & Akhlis, I. (2021). Changing user needs and motivation to visit a website through ad experience: A case study of a university website. *Journal of Physics: Conference Series*, *1918*(4), 042008.

Cage, J., Herve, N., & Mazoyer, B. (2020). Social media and newsroom production decisions. *SSRN Electronic Journal*, 3663899. https://doi.org/10.2139/ssrn.3663899

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27.

Chakrabarty, N., Rana, S., Chowdhury, S., & Maitra, R. (2019). RBM based joke recommendation system and joke reader segmentation. In *International Conference on Pattern Recognition and Machine Intelligence (Issue January 2020)*. Springer International Publishing. https://doi.org/10.1007/978-3-030-34872-4_26

Chakraborty, D., & Paul, J. (2022). Healthcare apps' purchase intention: A consumption values perspective. *Technovation*, *120*, 102481.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). *CRISP-DM 1.0*. CRISP-DM Consortium.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2022). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestin, C. (2016). XGBoost: A scalable tree boosting system. In A. Press (Ed.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD* (Vol. '16, pp. 785–794). ACM.

Chopdar, P. K., Paul, J., & Prodanova, J. (2022). Mobile shoppers' response to Covid-19 phobia, pessimism and smartphone addiction: Does social influence matter? *Technological Forecasting and Social Change*, *174*(121249), 121249.

Corley, K., & Gioia, D. (2011). Building theory about theory building: What constitutes a theoretical contribution? *The Academy of Management Review*, *36*(1), 12–32.

Coussement, K., Van den Bossche, F. A. M., & De Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, *67*(1), 2751–2758. https://doi.org/10.1016/j.jbusres.2012.09.024

Cristobal-Fransi, E., Hernández-Soriano, F., & Daries-Ramon, N. (2017). Nuevos lectores para nuevos medios: Segmentación de los e-lectores de un cibermedio. *Espacios*, *38*(39), 19.

Davies, D. L., & Bouldin,D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on PatternAnalysis and Machine Intelligence*, *PAMI-1*(2), 224–227.

Davoudi, H. (2018). *User acquisition and engagement in digital news media* (issue December). https://yorkspace.library.yorku.ca/server/api/core/bitstreams/185f044e-2f2d-410e-baee-85ffdf7d411f/content Accessed 15th February 2024.

Davoudi, H., An, A., & Edall, G. (2019). Content-based dwell time engagement prediction model for news articles. In *Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 2, pp. 226–233). Association for Computational Linguistics.

Davoudi, H., An, A., Zihayat, M., & Edall, G. (2018). Adaptive paywall mechanism for digital news media. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 205–214). Association for Computing Machinery.

de Caigny, A., Coussement, K., & de Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

Delpisheh, E., Davoudi, H., Boroujerdi, E. G., & An, A. (2016). Time aware topic based recommender system. *Big Data and Information Analytics*, *1*(2/3), 261–274. https://doi.org/10.3934/bdia.2016008

Fernandes, E., Moro, S., & Cortez, P. (2023). Data science, machine learning and big data in digital journalism: A survey of state-of-the-art, challenges and opportunities. *Expert Systems with Applications*, *221*, 119795. https://doi.org/10.1016/j.eswa.2023.119795

FTStrategies. (2022). *How the financial times brought data into the newsroom. FTStrategies*. https://www.ftstrategies.com/en-gb/insights/how-the-financial-times-brought-data-into-the-newsroom/

Fu, X., Chen, X., Shi, Y. T., Bose, I., & Cai, S. (2017). User segmentation for retention management in online social games. *Decision Support Systems*, *101*, 51–68. https://doi.org/10.1016/j.dss.2017.05.015

Gao, Y., Gupta, V., Yan, J., Shi, C., Tao, Z., Xiao, P. J., Wang, C., Yu, S., Rosales, R., Muralidharan, A., & Chatterjee, S. (2018). Near real-time optimization of activity-based notifications. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 283–292). Association for Computing Machinery. https://doi.org/10.1145/3219819.3219880

Global Change Data Lab. (2021). Our world data. https://ourworldindata.org/

Goad, R. (2016). Transforming a media organisation with big data. *EBU Big Data Conference*. https://www.slideshare.net/intotheminds/presentation-financial-times-big-data-at-ebu-big-data-conference

Gonçalves, J. N., Cortez, P., & Carvalho, M. S. (2021). K-means clustering combined with principal component analysis for material profiling in automotive supply chains. *European Journal of Industrial Engineering*, *15*(2), 273–294.

Gonçalves, L., Subtil, A., Oliveira, M. R., & De Zea Bermudez, P. (2014). ROC curve estimation: An overview. *Revstat Statistical Journal*, *12*(1), 1–20.

Google. (2022). *Event measurement*. https://developers.google.com/analytics/devguides/collection/analyticsjs/events

Gordon-Wilson, S. (2022). Consumption practices during the COVID-19 crisis. *International Journal of Consumer Studies*, *46*(2), 575–588.

Grinberg, N. (2018). Identifying modes of user engagement with online news and their relationship to information gain in text. In *The Web Conference 2018 – Proceedings of the World Wide Web Conference* (pp. 1745–1754). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186180

Gustriansyah, R., Suhandi, N., & Antony, F. (2019). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Hauke, J., & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Hullar, K. (2020). 3 social media tips from Chartbeat that will enhance your presence across channels. *Chartbeat Blog*. https://blog.chartbeat.com/2020/04/01/3-social-media-tips-across-channels/

Jacob, M. (2021). Nearly half of digital subscribers are '*zombies*,' Medill analysis finds. *Northwestern Local News Initiative*. https://localnewsinitiative.northwestern.edu/posts/2021/03/01/zombies/index.html

Joshi, R., & Garg, P. (2021). Role of brand experience in shaping brand love. *International Journal of Consumer Studies*, *45*(2), 259–272. https://doi.org/10.1111/ijcs.12618

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

Kamthania, D., Pahwa, A., & Madhavan, S. S. (2018). Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business. *Journal of Computing and Information Technology*, *26*(1), 57–68. https://doi.org/10.20532/cit.2018.1003863

Karampournioti, E., & Wiedmann, K. P. (2021). Storytelling in online shops: The impacts on explicit and implicit user experience, brand perceptions and behavioral intention. *Internet Research*, *32*(7), 228–259. https://doi.org/10.1108/INTR-09-2019-0377

Katser, I., Kozitsin, V., Lobachev, V., & Maksimov, I. (2021). Unsupervised offline changepoint detection ensembles. *Applied Sciences (Switzerland)*, *11*(9), 1–19. https://doi.org/10.3390/app11094280

Ketchen, D., & Shook, C. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441–458.

Kim, N. L., & Im, H. (2022). Do liberals want curbside pickup more than conservatives? Contactless shopping as protectionary action against the COVID-19 pandemic. *International Journal of Consumer Studies*, *46*(2), 589–600.

Kotler, P., Kartajaya, H., & Setiawan, I. (2016). *Marketing 4.0: Moving from traditional to digital*. John Wiley & Sons.

Koul, S., & Philip, T. M. (2021). Customer segmentation techniques on E-commerce. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021* (pp. 135–138). IEEE. https://doi.org/10.1109/ICACITE51222.2021.9404659

Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media & Society*, *18*(3), 502–520.

Kursan Milaković, I. (2021). Purchase experience during the COVID-19 pandemic and social cognitive theory: The relevance of consumer vulnerability, resilience, and adaptability for purchase satisfaction and repurchase. *International Journal of Consumer Studies*, *45*(6), 1425–1442.

La Cruz, A., Severeyn, E., Matute, R., & Estrada, J. (2021). Users segmentation based on Google analytics income using K-means. In *Information and Communication Technologies: 9th Conference of Equador, November 24-26* (pp. 225–235). Springer International Publishing.

La Torre, F. (2020). *Learning customer segmentation in the news media industry: From content and behavioral data to customer segments* [Master thesis]. Alto University.

Labatut, V., & Cherifi, H. (2012). Accuracy measures for the comparison of classifiers. arXiv preprint arXiv:1207.3790.

Lagun, D., & Lalmas, M. (2016). Understanding and measuring user engagement and attention in online news reading. In *WSDM 2016 – proceedings of the 9th ACM International Conference on Web Search and Data mining, 22–25* (pp. 113–122). Association for Computing Machinery. https://doi.org/10.1145/2835776.2835833

Lajumoke, T., Betts, T., Gautier, L., Part, T., Patel, U., & Meirinhos, L. (2020). Towards your North Star – Report on the outcomes of the european GNI subscriptions LAB 2020. https://www.ftstrategies.com/insights/towards-your-north-star-report-outcomes-european-gni-subscriptions-lab-2020/

Lambrecht, A., Goldfarb, A., Bonatti, A., Ghose, A., Goldstein, D. G., Lewis, R., Rao, A., Navdeep, S., & Yao, S. (2014). How do firms make money selling digital goods online? *Marketing Letters*, *25*, 331–341.

Lee, Y., Park, I., Cho, S., & Choi, J. (2018). Smartphone user segmentation based on app usage sequence with neural networks. *Telematics and*

Informatics, *35*(2), 329–339. https://doi.org/10.1016/j.tele.2017.12.007

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of user engagement. In *Lecture Notes in Computer Science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 164–175). Springer. https://doi.org/10.1007/978-3-642-31454-4_14

Lioudis, N. (2019). How recirculation builds engagement, supports reader acquisition efforts. *Chartbeat Blog*. https://blog.chartbeat.com/2019/08/21/recirculation-data-reader-acquisition/

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). *Thresholding classifiers to maximize F1 score*. stat, 1050, 14.

Liu, B., Bai, B., Xie, W., Guo, Y., & Chen, H. (2022). Task-optimized user clustering based on Mobile app usage for cold-start recommendations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*. Association for Computing Machinery. https://doi.org/10.1145/3534678.3539105

Liu, Y., Hsiao, A., & Ma, E. (2021). Segmenting tourism markets based on demand growth patterns: A longitudinal profile analysis approach. *Journal of Hospitality and Tourism Research*, *45*(6), 967–997. https://doi.org/10.1177/1096348020962564

Loni, B., Schuth, A., van de Hass, L., Jansze, J., & Visser, V. (2019). Personalized push notifications for news recommendation. *Proceedings of Machine Learning Research*, *109*, 36–45.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *4765–4774*, 552–564. https://doi.org/10.1016/j.ophtha.2018.11.016

Lynes, M. (2021). 3 types of newsreaders and strategies to engage them. *Twipe*. https://www.twipemobile.com/3-types-of-newsreaders-and-strategies-to-engage-them/

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *Volume 1*, *N°14: Statistics* (pp. 281–297). University of California Press.

Makhortykh, M., Vreese, C., Helberger, N., Harambam, J., & Bountouridis, D. (2021). We are what we click: Understanding time and content-based habits of online news readers. *News Media & Society*, *23*(9), 2773-2800.

Mathew, A. (2021). Role of big data analysis and machine learning in Ecommerce – Customer segmentation. In *Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2021* (Vol. 3(1), p.189). Amal Jyothi College of Engineering.

Mersey, R. D., Malthouse, E. C., & Calder, B. J. (2010). Engagement with online media. *Journal of Media Business Studies*, *7*(2), 39–56. https://doi.org/10.1080/16522354.2010.11073506

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2020). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, *38*(5), 616–632. https://doi.org/10.1177/0894439319828012

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Moro, S., Laureano, R. M. S., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011* (pp. 117–121). University of Minho.

Moyo, D., Mare, A., & Matsilele, T. (2019). Analytics-driven journalism? Editorial metrics and the reconfiguration of online news production practices in African newsrooms. *Digital Journalism*, *7*(4), 490–506. https://doi.org/10.1080/21670811.2018.1533788

Mucchetti, M. (2020). Google data studio. In *BigQuery for data warehousing* (pp. 401–416). Apress.

Nasir, V. A., Keserel, A. C., Surgit, O. E., & Nalbant, M. (2021). Segmenting consumers based on social media advertising perceptions: How does purchase intention differ across segments? *Telematics and Informatics*, *64*, 101687. https://doi.org/10.1016/j.tele.2021.101687

Nayal, P., Pandey, N., & Paul, J. (2022). Covid-19 pandemic and consumer-employee-organization wellbeing: A dynamic capability theory approach. *Journal of Consumer Affairs*, *56*(1), 359–390.

Nelson, M. J., & Hoover, A. K. (2020). Notes on using Google Colaboratory in AI education. In *Annual Conference on Innovation and Technology in Computer Science Education*, *ITiCSE* (pp. 533–534). Association for Computing Machinery. https://doi.org/10.1145/3341525.3393997

Paul, J., & Bhukya, R. (2021). Forty-five years of International Journal of Consumer Studies: A bibliometric review and directions for future research. *International Journal of Consumer Studies*, *45*(5), 937–963.

Paul, M. (2022). Segmentation of the reader and library user population in Poland. *The Library Quarterly*, *92*(3), 274–295.

Peña, V. C., Malthouse, E. C., & Mersey, R. D. (2023). Churning off the news: An analysis of newspaper subscriber churn across digital devices. *Newspaper Research Journal*, *44*(2), 190–205.

Perreault, M. F., & Perreault, G. P. (2021). Journalists on COVID-19 journalism: Communication ecology of pandemic reporting. *American Behavioral Scientist*, *65*(7), 976–991. https://doi.org/10.1177/0002764221992813

Peterson, E. T., & Carrabis, J. (2008). Measuring the immeasurable: Visitor engagement. In *Web Analytics Demystified*. Celilo Group Media.

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*, *9*(1), 12.

Punhani, R., Arora, V. P. S., Sabitha, S., & Kumar Shukla, V. (2021). Application of clustering algorithm for effective customer segmentation in E-commerce. In *2021 IEEE International Conference on Computational Intelligence and Knowledge Economy*, *ICCIKE 2021* (pp. 149–154). IEEE. https://doi.org/10.1109/ICCIKE51210.2021.9410713

Purohit, S., Arora, R., & Paul, J. (2022). The bright side of online consumer behavior: Continuance intention for mobile payments. *Journal of Consumer Behaviour*, *21*(3), 523–542.

Rajput, L., & Singh, S. N. (2023). Customer segmentation of E-commerce data using K-means clustering algorithm. In *2023 13th International Conference on Cloud Computing*, *Data Science & Engineering (Confluence)* (pp. 658–664). IEEE.

Rayburn, S. W., McGeorge, A., Anderson, S., & Sierra, J. J. (2021). Crisis-induced behavior: From fear and frugality to the familiar. *International Journal of Consumer Studies*, *46*(2), 524–539.

Rios-Rodríguez, R., Fernández-López, S., Dios-Vicente, A., & Rodeiro-Pazos, D. (2022). Reconversion in a declining market: The return to profitability of the print newspaper industry. *Journal of Media Business Studies*, *20*, 204–222. https://doi.org/10.1080/16522354.2022.2104556

Risch, J., & Krestel, R. (2020). Top comment or flop comment? Predicting and explaining user engagement in online news discussions. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)* (Vol. 2020, pp. 579–589). AAAI Press.

Riskos, K., Hatzithomas, L., Dekoulou, P., & Tsourvakas, G. (2022). The influence of entertainment, utility and pass time on consumer brand engagement for news media brands: A mediation model. *Journal of Media Business Studies*, *19*(1), 1–28. https://doi.org/10.1080/16522354.2021.1887439

Rußell, R., Berger, B., Stich, L., Hess, T., & Spann, M. (2020). Monetizing online content: Digital paywall design and configuration. *Business & Information Systems Engineering*, *1–8*, 253–260. https://doi.org/10.1007/s12599-020-00632-5

Schwartz, J. (2013). Using engaged time to understand your audience. *Chart*. https://blog.chartbeat.com/2013/03/18/using-engaged-time-to-understand-your-audience/

Seale, S. (2021). *How wall street journal uses metrics and engagement to drive digital subscriptions*. INMA International News Media Association. https://www.inma.org/blogs/conference/post.cfm/how-wall-street-journal-uses-metrics-and-enagagement-to-drive-digital-subscriptions

Shahraki, H., Pourahmad, S., & Zare, N. (2017). K important neighbors: A novel approach to binary classification in high dimensional data. *BioMed Research International*, *2017*, 7560807.

Shen, B. (2021). E-commerce customer segmentation via unsupervised machine learning. *ACM International Conference Proceeding Series*, *PartF168982*. https://doi.org/10.1145/3448734.3450775

Siddiqi, U. I., Akhtar, N., & Islam, T. (2022). Restaurant hygiene attributes and consumers' fear of COVID-19: Does psychological distress matter? *Journal of Retailing and Consumer Services*, *67*(102972), 102972.

Silva, S., Cortez, P., Mendes, R., Pereira, P. J., Matos, L. M., & Garcia, L. (2018). A categorical clustering of publishers for Mobile performance marketing. In *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications* (pp. 145–154). Springer International Publishing.

Sinha, P., Dey, L., Mitra, P., & Thomas, D. (2020). A hierarchical clustering algorithm for characterizing social media users. In *Companion Proceedings of the Web Conference* (Vol. 2020, pp. 353–362). Association for Computing Machinery.

Sjøvaag, H. (2016). Introducing the paywall: A case study of content changes in three online newspapers. *Journalism Practice*, *10*(3), 304–322. https://doi.org/10.1080/17512786.2015.1017595

Smit, G., Fahland, D., Dongen, B. F., & Farzami, T. (2019). *Customer segmentation using clickstream* [Bachelor Thesis]. Eindhoven University of Technology. https://pure.tue.nl/ws/portalfiles/portal/196163529/bsc_thesis_gijs_smit.pdf

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, *21*(1), 3–8.

Steensen, S., Ferrer-Conill, R., & Peters, C. (2020). (Against a) theory of audience engagement with news. *Journalism Studies*, *20*, 1–19. https://doi.org/10.1080/1461670X.2020.1788414

Stone, B. (1989). *Successful direct marketing methods* (3rd ed.). NTC Business Books.

Stroud, N. J., & van Duyn, E. (2020). Assessing the external validity of using news websites as experimental stimuli. *Communication Methods and Measures*, *14*(3), 212–218. https://doi.org/10.1080/19312458.2020.1718630

Tanuwijaya, S., Alamsyah, A., & Ariyanti, M. (2021). Mobile customer behaviour predictive analysis for targeting Netflix potential customer. In *9th international conference on information and communication Technology (ICoICT)* (pp. 348–352). IEEE. https://doi.org/10.1109/ICoICT52021.2021.9527487

Tripathi, M. A., Madhavi, K., Kandi, V. S. P., Nassa, V. K., Mallik, B., & Chakravarthi, M. K. (2023). Machine learning models for evaluating the benefits of business intelligence systems. *The Journal of High Technology Management Research*, *34*(2), 100470.

Vara-Miguel, A., Sánchez-Branco, C., Chalezquer, C. S., & Negredo, S. (2021). Funding sustainable online news: Sources of revenue in digital-native and traditional Media in Spain. *Sustainability*, *13*(20), 11328.

Vergani, A. A., & Binaghi, E. (2018). A soft Davies-Bouldin separation measure. In *2018 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE.

Villi, M., & Picard, R. G. (2019). Transformation and innovation of media business models. In *Making media: Production, practices, and professions* (pp. 121–132). Amsterdam University Press.

Vinothini, A., & Priya, S. B. (2018). Survey of machine learning methods for big data applications. In *ICCIDS 2017 – International Conference on Computational Intelligence in Data Science, Proceedings, 2018-Janua* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCIDS.2017.8272638

Vreese, C. H., & Neijens, P. (2016). Measuring media exposure in a changing communications environment. *Communication Methods and Measures*, *10*(2–3), 69–80. https://doi.org/10.1080/19312458.2016.1150441

Wenzel, S., Kleer, N., & Kunz, R. E. (2022). Customer engagement behavior in the media and technology industry: A quantitative content analysis of content types and COVID-19 context. *Journal of Media Business Studies.*, *20*, 241–263. https://doi.org/10.1080/16522354.2022.2139997

Whetten, D. (1989). What constitutes a theoretical contribution? *The Academy of Management Review*, *14*(4), 490–495.

Wieland, R., Lakes, T., & Nendel, C. (2021). Using SHAP to interpret XGBoost predictions of grassland degradation in Xilingol, China. *Geoscientific Model Development*, *14*(3), 1493–1510.

Xiao, J., Lu, J., & Li, X. (2017). Davies Bouldin index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327–1338.

Yap, S. F., Xu, Y., & Tan, L. (2022). Coping with crisis: The paradox of technology and consumer vulnerability. *International Journal of Consumer Studies*, *45*(6), 1239–1257.

Zihayat, M., Ayanso, A., Zhao, X., Davoudi, H., An, A., Rogers, T., & Technology, I. (2019). A utility-based news recommendation system. *Decision Support Systems*, *117*, 14–27. https://doi.org/10.1016/j.dss.2018.12.001

Zontek, S. (2018). *User engagement drives subscriptions. New RFV Engagement Scores from Deep.BI*. https://www.deep.bi/blog/user-engagement-drives-subscriptions-new-rfv-engagement-scores-from-deep-bi

How to cite this article: Fernandes, E., Moro, S., & Cortez, P. (2024). A data-driven approach to improve online consumer subscriptions by combining data visualization and machine learning methods. *International Journal of Consumer Studies*, *48*(2), e13030. https://doi.org/10.1111/ijcs.13030