Introduction

Recommendation systems are omnipresent in e-commerce stores and platforms. Their primary purpose is to assist customers in finding suitable offerings, thus increasing the likelihood and value of purchases from a store. Consumers can benefit from accurate recommendations by streamlining and shortening their purchase decisions and acquiring more satisfying and affordable products. However, the innate complexity of consumer behaviour makes accurate predictions difficult unless diversified data is involved (Lynch and Barnes, 2020; Patten and Ozuem, 2020).

Our study explores how recommendation performance is affected by different combinations of three data modalities (consumer behaviour data, product information, and visual inputs) in an online experiment within a real-life production setting operated by a large Internet footwear vendor. Making use of the records of consumers’ past interactions with the vendor, it was possible to complement the research design with constructs representing consumers’ loyalty to the vendor and their shopping involvement.
In this study, we account for information diversity in recommendation systems by deploying multiple combinations of data types (i.e., modalities). Such an approach presents three technological challenges: (1) preparing a balanced database in terms of modality diversity, (2) seamlessly combining variables with distinctly different properties into a uniform data representation (also “on the fly” in a production environment), and (3) selecting reliable performance metrics for online and offline tests.

A vast majority of existing recommendation systems considers multiple variables but only from a single modality; for example, webpage views, clicks, purchases, or text features are all parts of behavioural data. It seems evident in past research that adding different modalities of data should increase the performance of recommendation algorithms (Gao et al., 2020; Lahat et al., 2015; Liu et al., 2018). However, other studies show that certain modalities may have greater or lesser significance depending on the product type. For example, in a visual domain, like fashion, online product search is strongly driven by visual attributes (Saha et al., 2018; Wang et al., 2020). Moreover, past research suggests that, due to added sourcing and processing costs, it is technically challenging and economically inefficient to adjust existing recommendation systems to include additional data modalities (Chou et al., 2019; Yang et al., 2019; Naishe et al., 2020; Wang et al., 2020). Different modalities are characterized by diverse computational power requirements to train algorithms. The visual modality requires more computing power than behavioral or textual data. However, including the visual modality does not have to be better than omitting it, partially because of a trade-off between computing power costs and algorithm predictive performance (Cao et al., 2020). Finally, in practice we often experience a substantial discrepancy between the offline and online performance of recommendation models, with offline models showing a better fit to the data in their estimation samples that rarely carries over to practical applications (Yi et al., 2013). Considering that online experiments are thought to give better approximations of real-life performance, we adopt this research design to explore the accuracy of different configurations of recommenders in the fashion industry.

This paper relies on an experiment with data sourced from an online sales platform operated by a prominent footwear retailer in Eastern and Central Europe. The experiment was initially designed to investigate the impact of different combinations of data modalities used in generating recommendations on recommendation performance. Considering that additional consumer data were available, it was possible to extend the analysis to explore how recommendation performance was affected by various characteristics describing consumers’ attitudes towards both the vendor and the recorded shopping events. Scrutiny of the database of the vendor allowed us to identify a set of metrics that could be interpreted as proxies for the well-established marketing concepts of consumer loyalty and shopping involvement. Inspired by the Relationship Quality Model (Reichheld, 1996) and – to a lesser extent – the Theory of Planned Behaviour (Ajzen, 2020), these attitudinal variables gave another dimension and more nuance to the task of building a statistical model explaining recommendation system performance.

The methodological approach followed in this study is unique in that most past research on recommender effectiveness was carried out as offline experiments whereby recommendations were generated ex-post from preexisting datasets (Dacrema et al., 2019; Pradel et al., 2011). Offline experiments should not be considered actual experiments since researchers have no control over experimental conditions to establish treatment groups. Thus, they cannot observe and compare research subjects’ behaviour across the groups (Rossetti et al., 2016). In addition, the measures of consumer attitudes used in data modelling are based not on survey opinions, which could be easily biased, but on actual past behaviour, accurately recorded in the vendor databases.

In the following sections, we review relevant published works, develop research hypotheses, describe our research methods, discuss the results, and offer our thoughts on the limitations of this study and directions for further research.

**Literature review and research hypotheses**

**Theoretical framework of the study**

Considering that the focus of this research is consumer buying behaviour and loyalty is among its antecedents, the theoretical underpinnings could be sourced from conceptual models explaining purchase behaviour. Arguably, among the most popular conceptual frameworks in the literature are the Theory of Planned Behavior (TPB) and the Relationship Quality Model (RQM) (Buhais at al., 2020). The TPB predicts buying behaviour (as well as other non-mercantile behaviours of individuals) from a link with the construct of behavioural intention, which is itself formed by (1) the attitude towards the behaviour, (2) subjective norm (which reflects the impact of normative expectations of others), and (3) perceived control over resources needed for performing the behaviour under study (Ajzen, 2020). In a TPD-inspired model, loyalty tends to be positioned as the final endogenous variable preceded by purchase intention, like the approach adopted by Hsu et al. (2006) in their investigation of Internet retail shopping.
On the other hand, in the RQM (also known as the Satisfaction-Profit Chain) loyalty intentions are preceding the purchase and are developed over time through such intermediate effects as product/service satisfaction and overall relationship satisfaction, with additional influences of commitment and trust (Reichheld, 1996). Cannière et al. (2009) compare both models using a dataset consisting of both self-reported attitudinal measures from consumers as well as their recorded actual behaviour of buying (or not) from an online apparel store. They found that both approaches have statistical merits, but the TPD-driven model offers superior predictive power. They suggest that this is due to a closer temporal proximity between the self-reported TPD measures to the purchase behaviour than the attitudinal metrics in the RQM. This observation is crucial in the choice of the theoretical model for this study. Given that our dataset consists of only factual, recorded metrics describing behavioural patterns of consumers over a long period of time (depending on the moment of setting up an online account by the customer this could be longer than one year before the start of the experiment) there were no attitudinal metrics close in time to the occurrence of the observed behaviour. The proxy-variables that we were able to calculate using the raw data in the database were corresponding to concepts of trust in the vendor, engagement with the vendor and the experience with the vendor (its intensity and length). The trust variable corresponds directly to the trust construct in the RQM, while engagement and experience could be aspects of commitment with the vendor. Since it was impossible to identify in the database behavioural proxies for either behavioural intention or its antecedents in the TPD, the adoption of the TPD over the RQM would be misguided. One aspect of the TPD that we do implement in our research design (at least partially) is behavioural control. Online shopping creates a highly interactive environment with a high degree of control perceived by the consumer. Such a context, as postulated by Ajzen (2000) behavioural control exercised in online shopping is not only a function of the web-page interactivity and perceived ease-of-use but also – just like in off-line shopping is predicated on the individual’s purchasing power. We indirectly account for the consumer’s purchasing power by complementing our model with information about the price of the purchased product if lower prices coincide with higher perceived behavioural control.

## Data modalities and recommendation performance

Research on how different combinations of data modalities contribute to recommendation accuracy is scarce, which is surprising since the inclusion of some modalities in predictive models, such as the visual modality, raises resource requirements, leading to increased operating costs that could outweigh benefits. Thus, quantifying costs and benefits is critical to making informed decisions about deploying recommender algorithms relying on more diversified and comprehensive data representations. Of the few published works on the topic, Truong and Lauw (2019) investigated deep learning models for predicting restaurant ratings obtained from Yelp.com, a popular service built around crowd-sourced reviews of businesses. They found that including user-made pictures makes for more accurate predictions, but the effect is much weaker than contributions from reviews’ textual features, particularly in scarce datasets.

Another study by Chen et al., (2019) proposes a novel method for embedding visual data in the fashion domain by accounting for individual weights attached to different parts of a product image, indicated by consumers in their written reviews. Using an open-source dataset from Amazon.com, the authors tested several baseline models and conducted an ablation analysis of their method. They found that removing textual features from data input lowered the model’s accuracy only slightly, suggesting that the visual modality is the most potent factor in the model’s effectiveness. In comparison, models built around algorithms that relied primarily on textual data (e.g., a neural-rating regression model) showed weak accuracy losses from the removal of visual data. It indicates that adding or omitting the visual modality can affect differently algorithms that have an innate emphasis on text or images.

Many published works are dedicated to new recommender algorithms designed by their authors to integrate multi-modal data streams (Johansson, 2003; Koren et al., 2009; Parra and Amatriain, 2011; Rendle, 2010; Domingues et al., 2013; Da Costa & Manzato, 2016; Alayrac et al., 2020). For validation purposes, new recommenders were typically contrasted with older ones, employing fewer modalities, making it unclear if the superior accuracy was due to a new method, richer information, or both. Arguably, a more robust comparison would involve varying different modality combinations across experimental groups under the same recommender algorithms while retaining the possibility of comparison with a parallel baseline method, which is how we designed this study. In addition, if experimental groups are differentiated by the placing of banners with recommendations on the vendor’s website, this should also be accounted for in research design. Website placement was found to be a significant driver of click-through ratios for
advertisement banners (Bleier & Eisenbeiss, 2015), which makes plausible analogous effects to occur for on-line recommendation banners.

The current study tests the following set of hypotheses concerning the technicalities of the experiment:

H.1 Technical parameters of the experiment are associated with recommendation performance.
  H.1.1 Type of recommendation algorithm is associated with recommendation performance.
  H.1.2 Types of data modalities used in the recommendation algorithm are associated with recommendation performance.
  H.1.3 Recommendation placement is associated with recommendation performance.

Shopping involvement

Consumer involvement in shopping has a capacity to drive pre-purchase search activities, including information processing, which can alter responsiveness to recommendations (O’Cass, 2000). Involved consumers should pay more attention to absorb more information about a shopping situation which should enable them to arrive at more elaborate meanings and complex inferences (Kwon & Chung, 2010). Offering recommendations to involved customers poses a risk of failure because of the mechanism of psychological reactance, which is opposition against perceived attempts to control their behavior and limit their freedom of choice (Brehm, 1966). Empirical evidence of psychological reactance was found by Fitzsimons and Lehman (2004) in the case of unsolicited recommendations being ineffective. In another study, Kwon and Chun (2010) showed that highly involved customers pay attention to the core message of the recommendation, while largely ignoring the so-called peripheral cues, such as pictures, source characteristics, music, and message side-lines. In a similar vein, Whang and Hyunjoo (2018) discovered that highly involved shoppers reacted negatively to poorly personalized recommendations, while reactions were neutral to positive for low-involvement customers (those who were “only browsing”). This suggests that high-involvement consumers tend to have well-defined preferences that might not be satisfied by recommender systems based on a limited number of data points of only quantitative nature. Whang and Hyunjoo (2018) recommend including in recommender systems for high-involvement shoppers also qualitative inputs (such as social media interactions), however many recommender algorithms are technically constrained and unable to do so.

Fashion products are classified as high-involvement purchases with significant psychological risks due to their role as a means of self-expression and thus being instrumental in achieving personal goals (Kinley et al., 2010). However, individual consumers’ characteristics and particular shopping contexts can induce involvement variance not only among different consumers but also across separate shopping acts by the same consumers. Fashion interests, and hence purchase involvement, tends to be higher among female customers buying more expensive items for themselves rather than for other family members (Workman and Cho, 2012).

Considering the above discussion, the current study aims to test the following set of hypotheses:

H.2 Shopping involvement is negatively associated with recommendation performance.
  H.2.1 Product gender designation is associated with recommendation performance with female products linked to lower recommendation performance.
  H.2.2 Product price is negatively associated with recommendation performance.

Consumer loyalty

Loyalty is considered a critical factor in driving consumer behaviour in highly competitive markets such as fashion. As noted by Griffin (2002), loyalty is reflected in frequent repeat purchases by a customer, his or her willingness to use a wide range of products and services a company provides, proclivity for praising the company through word-of-mouth activities and weaker responsiveness to promotional activities by other companies. In online markets, where the switching costs are typically low and the competitive offer is rich, this makes customers less price sensitive and more willing to recommend the company to others (Srinivasan et al., 2002).

The ways to define and operationalize loyalty are somewhat inconsistent in the literature with different authors proposing diverse ways of understanding and measuring the concept (Diallo et al., 2020). In this study we adopt a definition used by Alonso-Dos-Santos et al. (2020) in their research of loyalty determinants in online banking. The definition states that loyalty is a commitment to repurchasing a product or service that is preferred over time, despite marketing efforts by competitors and their offers on the market.

Dick & Basu (1994) propose that loyalty is composed of three dimensions: commitment, trust, and satisfaction, leading to repeat patronage intentions and inducing loyal behaviour. This approach mirrors the three antecedents of relationship quality in the RQ model, which serves as the theoretical foundation to our research. Thus, we attempt to replicate this framework with recorded past behavioural data serving as proxies for trust, commitment, and satisfaction.

Besides its role as a determinant of purchase intention, loyalty is also a relevant influence on recommender system performance. Wan et al. (2018) found that accounting for
consumers’ product loyalty tends to improve recommendation success rates in grocery retailing. Ji et al. (2022) investigates if loyal customers are being offered more accurate recommendations through a series of experiments on several shopping platforms (including Movie Lens, Amazon, and Yelp). Their operationalization of loyalty was by necessity limited to behavioural data found in the studied datasets and was determined by the time a user has been active in a recommender system, or by the number of historical interactions a user has had. The findings are somewhat counterintuitive, since having many interactions or having a long active time was adversely affecting recommendation accuracy. The authors conclude that this could be because the currently popular recommendation systems utilize too few and only most recent datapoints, which works well enough for new users but seems to be counterproductive for experienced consumers.

Trust, one of the three components of loyalty and a key factor in consumer adoption behaviour, has been an object of research. As early as the year 2000, Gefen (2000) in the survey of customers of online bookstores found that the customers’ intention to buy was higher for greater levels of trust in the online bookstore. Later research in diverse retail setting corroborated this observation by providing evidence that consumers’ trust in an Internet vendor affected their purchase intention from the website (Chang & Chen, 2008; Chih et al., 2009; Kim et al., 2008; Salo and Karjaluoto, 2007; Wu and Chang, 2006). In addition to enhancing purchase intentions, trust in a vendor also tends to rise the credibility of information received from the vendor, including product recommendations. This effect was demonstrated by Hsiao et al. (2010) on survey data collected from 1219 Taiwanese fans of computer games who were interviewed on how their purchase new items for their gaming collections. They found positive correlations between trust in a website, trust in product recommendations received on the website and intention to purchase the recommended products. In a more recent study (Yang, 2021) uses on-line survey data from 161 users of the Chinese Internet service WeChat to probe the role of trust as a mediator between information appraisal metrics and continuance intention to use the recommendation system. The findings point to considerable correlations among trust, perceived information quality and continuance intention to use the recommendation system.

Building on the preceding discussion, the current study aims to test the following set of hypotheses pertaining to consumer loyalty and its dimensions:

H.3 Consumer loyalty is associated with recommendation performance.
H.3.1 Consumer engagement with the vendor is positively associated with recommendation performance.
H.3.2 Consumer trust in the vendor is positively associated with recommendation performance.
H.3.3 Consumer experience with the vendor is negatively associated with recommendation performance.
H.3.4 Consumer experience with the vendor negatively moderates the association of engagement, trust, and shopping involvement with recommendation performance.

Research methods

In the methods section, we first elaborate in more detail on the technicalities of the experiment: the EMDE algorithm and its CF-RS benchmark, to show how recommendations were generated with various combinations of data modalities. Next, we explain how proxy variables for pertinent marketing concepts were operationalized from the raw data obtained from the vendor’s computer systems. What concludes this section, is a short presentation of statistical methods that served as the means of testing the research hypotheses.

Employed recommendation algorithms

The EMDE algorithm is designed to take as its inputs multiple data modalities in different formats, such as text descriptions, user behavioural logs, audio files or images. Each modality of input data is transformed into a vector space. Depending on modality characteristics, a wide range of transformation procedures can be employed, yielding sparse representations, e.g., one-hot encoding of text (Cerda et al., 2018) or dense vector representations, e.g., word2vec (Mikolov et al., 2018), gloVe (Pennington et al., 2014), BERT (Devlin et al., 2019), node2vec results on graph representations (Grover and Leskovec, 2016), image deep neural networks results, e.g., VGG, ResNet embeddings (Kornblith et al., 2019).

All those data modalities and their representations are combined into a single numerical vector by the EMDE, using the locality-sensitive hashing (LSH) procedure (Slaney and Casey, 2008). Further, the EMDE representation (a so-called sketch) of a user’s history (a series of products each user viewed and their latest transactions) is fed into a neural network which then predict the product that the user would view or purchase. Subsequently, the results can be ranked or filtered depending on business rules for respective product categories. This way the EMDE representation is composed of the data gathered from all modalities about a single consumer. A conceptual understanding of the EMDE process is displayed in Fig. 1. A detailed description of the algorithm, including mathematical formulae, can be found in Basaj et al. (2020); Dabrowski et al. (2020); Rychalska
et al. (2021); Wieczorek et al. (2020a); Wieczorek et al. (2021), and Wróblewska et al. (2022).

The above-described EMDE algorithm was benchmarked against a standard industry method based on the Collaborative Filtering (CF) algorithm (Schafer et al., 2007). Here, the CF is modified by using as an input a graph-based structure. The algorithm serves as the baseline for the experiment, relying only on two data modalities: consumer product information and textual data. The input representation uses a bipartite graph whose vertices are divided into two disjointed and independent sets, i.e., one vertex set comprises users, and another - products. Edges in the graph are set among users and products to represent their mutual interactions, i.e., users viewing products. A textual network is generated similarly. An algorithm ranks a set of identified products based on cosine similarity and the history of their users to the input user who is the recommendation subject.

**Experimental design**

The market experiment took place between 24 August and 14 September 2020 on the Romanian site of one of the largest online footwear vendors in Central and Eastern Europe. This period was split into four parts, each one-week long, where different combinations of data modalities were used to create recommendations by the EMDE algorithm (the baseline CF-RS algorithm consistently used the same types of data), according to the following set-up:

- **Week 1**: textual and behavioural data,
- **Week 2**: textual, behavioural, and visual,
- **Week 3**: textual and visual,
- **Week 4**: visual and behavioural.

During the experiment, all participating customers were offered recommendations generated with the EMDE algorithm (with one out of four modality configurations) or CF-RS. The customers were assigned to each algorithm at random the moment they visited the vendor’s website.

**Data collection process**

The customers’ responses were recorded to establish how far in the purchasing process they progressed after receiving a set of recommendations. The stages of the consumer shopping process were as follows: (1) generating and receiving a set of recommendations, (2) viewing the recommendations, (3) clicking on one of the recommended products, (4) purchasing the product.

Each recommendation was composed of a set of 10 product propositions and new recommendations were set off by customers’ actions, such as refreshing a page or moving to other pages on the site. The new recommendations were yielded by the same algorithm and were typically similar but not identical to the previous ones for the same individuals. Consequently, customers prompted multiple recommendations during a single browsing session – typically between 8 and 10 sets per one visit, and each added a new data-point in the experiment database. Most recommendations were ignored by customers, but a small fraction was looked at more carefully (around 3%), which was reflected in a characteristic pattern of scrolling and cursor movements and led the script to classify them as viewed recommendations. When purchases were made, they usually involved a single pair of shoes (80%) or two pairs (14%).

The final database was obtained by merging 4 intermediate databases reflecting consecutive stages in the purchasing process:

1. generating and receiving a recommendation (26 677 253 records and 2 148 976 unique customers),
2. viewing a recommendation (721 356 records, 349 404 unique customers),
3. clicking one of the recommended products (296 496 records, 181 933 unique customers),
4. purchasing the product (96 495 records, 69 008 unique customers).

Databases 1 through 3 were merged on the same event number, with those recommendations that were not followed through by customers having missing values in the subsequent databases. Purchase transactions did not have matching event numbers, so the merging routine was different. For recall metrics of recommendation performance, which are central to the research presented in this paper, each purchase transaction was joined to only one entry for the same customer from the merged database involving shopping steps 1 to 3; it was the recommendation nearest in time before the purchase, but no earlier than 24 h. This approach can be deemed conservative regarding its impact on recommendation performance, since usually more than one recommendation came before a purchase event, which could also have some influence on the customer’s decision.
Dependent variable: recommendation system performance metrics

Successful recommendation

A consumer visit to the front page of the platforms triggered a recommendation set composed of 10 items. Further movements to other web pages on the site, as well as refreshing the same page, activated new sets of recommendations. Consequently, the customer received multiple recommendations during a single browsing event—typically between 8 and 10 sets per one visit. About 5–10% of recommendations were looked at, which was reflected in characteristic scrolling and cursor movement, which led to classify them as viewed recommendations.

A successful recommendation was considered the one which ended with the purchase of at least one of the recommended products no later than 24 h after issuing the recommendation. This definition of success was employed for precision and recall metrics. It is worth noting, that any of the precision and recall metrics described here can be used to represent recommendation system performance, because each one of them shows how accurately the system anticipates the shoppers’ actual preferences.

Precision metrics: PM1, PM2 and PM3

To investigate the effects of adding or removing data modalities on recommender performance, three precision metrics (PM1, PM2 and PM3) and one recall indicator were employed. They were consistent with common formulas for precision and recall as found—for example—in Cremonesi et al., (2010); Rawat and Dwivedi (2019); Rossetti et al. (2016).

During the online experiment, we used three different precision measures. PM1 refers to a standard precision at 10, which corresponds to the number of purchased products from the ten-generated recommendation set. PM2 and PM3 are the ratios of the number of successful recommendations sets to all received recommendations or only viewed recommendations, respectively.

Recall

The recall metric was defined as a ratio of sold products that complied with the directly preceding recommendations over all products sold in each experimental group. Here, each product purchased by a customer was compared with the most recent recommendation received by the customer in the preceding 24 h. If the product complied with the recommendation during this period, it was purchased due to this recommendation and increased the numerator in the recall metric by one.

To provide the reader with more context, it is worthwhile to concisely report on the outcomes of recommendation performance across different algorithms and modalities. To compare recommendation performance, several metrics were employed with three precision indicators (PM1, PM2, and PM3) and one recall measure. Accordingly, PM1 informs on the average number of recommendations in sets of ten that were matching the products purchased by the same customer within 24 h (it is equivalent to the precision at 10 metric frequently found in other publications on similar topics). PM2 and PM3 are ratios of the numbers of purchase-matching (i.e., successful) sets of recommendations over, respectively, all recommendations or only viewed recommendations. The recall metric was computed as the fraction of all individual products sold that were consistent with the directly preceding recommendation (i.e., no earlier than 24 h before the purchase).

Operationalization and distribution of independent variables

It is worthwhile to note, that this paper relies on data from an experiment initially designed to investigate the impact of using different combinations of data types (so-called modalities) in generating recommendations on recommendation performance. Considering that additional consumer data were available, it was possible to extend the analysis by developing new hypotheses to investigate how recommendation performance was affected by various attributes describing the whole of consumers’ relationship with the vendor and the characteristics of individual recorded shopping events.

The research hypotheses, laid out in preceding sections of the paper, call for three kinds of variables measuring: (H.1) technical attributes of experimental groups, (H.2) consumer shopping involvement, and (H.3) loyalty. Below, we describe how the data from the experiment and the records of customer interactions with the vendor were used to specify metrics for each concept. We also provide information on the distribution of these variables in our data set.

Technical characteristics of experimental groups were as follows:

1. Type of recommendation algorithm: a binary variable with 75.1% of 1s representing EMDE and 24.9% of 0s corresponding to CF-RS.
2. Three binary dummy variables reflecting which data types (modalities) were used to generate recommendations in EMDE, with 18.1% using textual data, 19.2% relying on behavioral inputs, and 17.8% on visual ones.
S. Sysko-Romańczuk et al.

3. A binary variable for recommendation placing on the main page (1 for 87.5%) or on a product page (0 for 12.5%).

An investigation of the vendor’s database allowed us to identify a set of variables that could be interpreted as proxies for well-established marketing concepts. We were able to link the construct of consumer loyalty and shopping involvement to the metrics from the vendor database.

Experience summarizes the length and intensity of the consumer’s relationship with the vendor and is based on four metrics:

1. Number of days since the account creation (mean 255.03, standard deviation 278.48).
2. Number of days since the first shopping transaction (mean 24.05, standard deviation 56.54).
3. Number of purchase transactions in the past year (mean 1.99, standard deviation 2.86).
4. Value of transactions in the past year (mean 188.20 Euro, standard deviation 822.39 Euro).

To reduce strong right-sided asymmetry, the four metrics of experience were logarithmized (with base 10 logarithm), standardized (to be measured on the same scales with the mean of 0 and the standard deviation of 1) and averaged to arrive at a single measure of experience.

Statistical methods

Considering the binary nature of the dependent variable, with 1s and 0s representing, respectively, purchases consistent and inconsistent with recommendations, a suitable statistical modelling approach was binary logistic regression.

A sequence of three progressively more complex models was estimated, using functions from Python libraries stats models and scikit. The first model (Model 1) involves only technical characteristics of the experiments, as specified in Hypothesis 1 (H.1), and it was designed as the starting point of the analysis to clearly separate consumer related variables from the more objective determinants of recommendation performance, which helps to avoid the risk of spurious correlations due to the endogeneity of independent variables.

The problem of endogeneity occurs when one or more of independent variables are correlated with the error term, typically due to other omitted independent variables Wooldridge (2020). As Wooldridge explains, the lack of critical, independent variables in a regression model leads to biased and inconsistent estimates of regression weights, frequently of too large magnitudes, leading to overestimating the effects of the retained predictors. Following this way of thinking, Model 2 augments Model 1 with the main effects of the variables describing aspects of shopping involvement and loyalty. The final regression equation (Model 3)
adds interaction effects of experience with other consumer related variables.

The reliability of the analysis is assessed with individual significance tests for all regressors and with overall goodness-of-fit metrics including the AUC metrics (giving the area under the ROC curve), and pseudo R-squared coefficients (measuring the amount of error reduction according to the Nagelkerke formula). In addition, likelihood ratio tests were performed to compare more complex models with their direct simpler predecessors (i.e., Model 2 vs. Model 1, and Model 3 vs. Model 2).

**Research results**

Performance metrics for the whole sample, with all data modalities and algorithms pooled together, were as follows: PM1 = 0.0174, PM2 = 0.0170, PM3 = 0.1707, Recall = 0.2283. More specific analyses across subgroups showed that the EMDE algorithm offered a considerably superior performance than CF-RS according to all metrics in all experimental groups except Week 1, when the visual modality was not included. In terms of data modalities, the best outcomes were achieved when both visual and behavioural modalities were used.

In the current analysis, recommendation performance is represented only by the recall metric which informs on the fraction of all products sold due to recommendations. In this dataset, each product purchased by a customer was compared with the most recent recommendation received by the customer in the preceding 24 h. If the product was consistent with the recommendation during this period, the product was purchased due to this recommendation and assigned a value of 1, indicating recommendation success, otherwise it was deemed recommendation failure with 0 as its value.

Precision metrics (PM1, PM2 and PM3) were not relevant in the current study since they considered all visitors to the vendor’s site, the majority of whom were not identified as registered customers. For such individuals no data were available about their history of interactions with the vendor and – clearly – since they were just browsing without buying anything, no information of the purchased product could be included in the analysis. On the other hand, recall can only be computed for sold products, and sales transactions can be described by several data points, both documenting the transaction itself as well as the customer. As such, the database that was used in this project counted 64,605 records, which reflects the number of completed transactions corrected for data points with missing values (corresponding to customers who made purchases without setting up an account in the vendor’s IT system).

Binary logistic regression developed in three stages illustrates which of the hypothesized antecedents of recommendation performance were found to significantly enhance the predictive capacity of the model. The results are depicted in Table 1.

As it was explained earlier, Model 1 is intended to account for experimental group differences due to various combinations of data modalities and recommendation algorithms deployed. Even though it was not part of deliberate design choices, the dataset contains recommendations displayed either on the main page of the vendor’s website or the product category page. Admittedly, one could expect systematic differences in effectiveness between recommendations placed in each of the two locations. The regression output for Model 1 shows that the experiment’s technical attributes indeed had some impact on recommendation performance, although it was relatively mild (Nagelkerke R Square = 0.07). It seems that a greater likelihood of successful recommendations could be expected when the EMDE algorithm was used (instead of the CF-RS) and when the recommendation was displayed on the product web page (rather than the main page). In terms of data modalities, the most significant positive contributions were observed for visual data, followed by behavioural records. This evidence validates hypothesis H.1.

Model 2 extends Model 1 by including the main effects of all pertinent consumer-related variables. All predictors in the new model are statistically significant, which translates into a considerable improvement in the goodness-of-fit, as evidenced by the increases in AUC (from 0.598 to 0.746) and pseudo-R-squared (from 0.07 to 0.216). The three shopping involvement measures appear to support the notion that lower involvement levels correspond to higher sensitivity to recommendations (H.2). Assuming that a vast majority of consumers in the database are females (over 90%), the negative link between purchasing women’s shoes and the logit of recommendation success can be ascribed to the customers buying for themselves and feeling more involved in the process. A somewhat weaker but still negative association for unisex offerings seems to corroborate this conclusion, as unisex fashion products tend to be worn more frequently by women than men.

The negative regression weight for product price implies that higher price tags tend to lower the odds for successful recommendations, which could also be explained by the negative effect of a deeper shopping involvement, as higher prices are typically linked with greater financial risk on the part of customers, thus prompting them to make more thoughtful (i.e., involved) decisions. Overall, the second hypothesis was fully supported by the outcomes of our analysis.
Among the three dimensions of loyalty, two of them – engagement with the vendor (measured as inbound and outbound communication) and trust (represented by a proxy variable for the amount of private information shared with the vendor) were increasing the odds of successful recommendations. On the other hand, experience (representing the length of time with the vendor and the number and value of past shopping activities) corresponded to lower recommendation performance. This discrepancy in the directions of impact provides some justification for not treating loyalty as a homogenous construct with only one score. Instead, it was introduced into the analysis as its component dimensions. In terms of our research hypotheses, these findings corroborate H.3.1, H.3.2, and H.3.3.

The final Model 3 complements Model 2 with interaction terms investigating how the length and intensity of experience with the vendor can moderate the regression weights between the metrics of shopping involvement and loyalty and the logit of recommendation success. Even though half of the interaction effects are significant, Model 3 does not improve much over Model 2, as demonstrated by small gains in AUC and R-squared. However, the likelihood ratio
test based on the difference in -2 log-likelihoods for the subsequent models is significant and shows that in a statistical sense, Model 3 is superior to Model 2. It should be noted that adding new terms to the analysis did not change the significance test outcomes or regression signs of the variables comprising Model 2.

Considering significant individual interactions, is evident that regression weights are all negative. This demonstrates that the impact of experience on recommendation performance is negative due to its primary effect (which is valid regardless of the other variables in the equation) and because of how it affects other regression coefficients in the model. Our findings imply that greater experience lowers the positive associations for outbound communication, personal information shared, and female product gender. For example, with all the other predictors held constant, one unit’s increase in experience (i.e., one standard deviation) lowers the regression weight for outbound communication by 0.185; thus, the total effect of outbound communication on the logit of regression success is 0.612-0.185=0.427. For consumers with exceptionally high experience of 2 standard deviations above the mean, the effect of outbound communication on the dependent variable is still positive but reduced by more than a half (0.612-2*0.185=0.242). Considering that not all interaction terms turned out to be significant (product price, inbound communication, and unisex product gender was the unaffected variables), the outcomes provide only partial support to hypothesis H.3.4.

**Discussion**

This paper presents the EMDE algorithm’s performance with different modalities on the CF-RS algorithm’s backdrop using data from a rigorous experimental set-up. Additionally, our research presents outcomes based on a unique data set compiled through an online experiment in a real-life fashion shopping environment.

Employing binary logistic regression while accounting for possible confounding variables, we managed to find statistical evidence on the relative importance of each data modality for recommendation accuracy in the fashion industry. Our findings suggest that the essential type of inputs for predicting footwear shoppers’ behaviour is the visual modality, followed by behavioural data. The least important in providing accurate recommendations seem to be textual data containing product descriptions. Somewhat unexpectedly, we found that the best performing algorithms included not three modalities but just two, which emphasizes the relative importance of visual data in the context of this type of market and the accompanying consumer behaviour.

An extensive set of 64,605 footwear purchase transactions was complemented by consumer records of previous interactions reaching back to the moment of setting up the account with the vendor. In contrast to many other consumer studies in marketing, we employed proxy variables for consumer attitudes derived from their real-life, well-documented behaviour instead of asking them directly through survey questions. This approach avoids familiar data-integrity problems due to common method bias and sometimes questionable validity of popular measurement scales employed in the probing opinions of individuals.

Our unique contribution is to demonstrate how shopping involvement and loyalty could be linked to automated recommendations’ success rate. If our interpretations of proxy variables are correct, it appears that consumers with more shopping commitment are less susceptible to recommendations. This effect could be driven by the higher perceived risk of shopping for more expensive articles. It should be noted that the prices registered in this research are final, including all applicable discounts and price promotions. However, the information about applying such discounts and price reductions was missing from the database, precluding us from a more in-depth investigation of the vendor’s pricing policies. Considering that most customers were female, the observation that women’s shoes were showing worse recommendation performance than either male or unisex offerings suggests that when buying for themselves, customers were more sceptical of the automated system’s suggestions. It could be due to their perceived expertise and proclivity for a more deliberate decision process if a fashion item was to be worn by themselves. Our construct of loyalty turned out to be heterogenous in how its elements correlated with recommendation performance. The best recall metrics were found for those customers who showed the highest levels of trust and engagement with the vendor. On the other hand, the more extensive was the experience with the vendor, the smaller number of products was bought owing to recommendations. The role of experience was even more profound due to its moderating effect with some of the other characteristics of consumer behaviour. More experienced customers were apt to have weaker links (though still positive) from trust and engagement metrics to recommendation performance. Interestingly, the negative interaction of experience with female product designation corroborates our earlier insight into the supposed perceived expertise as one reason behind less successful recommendations.

Overall, the outcomes reveal a rather complex relationship of loyalty with the effectiveness of automated advisory systems. There seem to be contradicting forces at play that call for a possibly more nuanced strategy on the part of designers of online recommenders and marketing managers. Adjusting parameters of recommendation algorithms
Limitations and directions of further research

In business literature, companies that frequently use large datasets are not fully satisfied with the currently available analysis systems and wish for solutions that will provide accurate predictions derived by automatic analysis of multiple customer records. This underutilized customer data includes direct inputs (comments, recommendations, etc.) and observations of how customers interact with the vendor’s website and mobile and offline ecosystems. Unlike existing solutions, meaningful progress can be achieved by combining real-time data from different channels and modalities. It offers a promise of significantly increasing the recommender algorithms’ performance and expanding their functionality. Improvements in predictive analytics of customer behaviour are a critical challenge for both business and academia.

This research is not without its constraints and points of weakness, which could indicate directions of future inquiry. First, this study relies on the dataset about only one online vendor, which, though extensive, might not accurately reflect the patterns of consumer behaviour found in other online stores. Thus, it calls for more research on shopping records sourced from other companies, especially in different retailing segments.

Second, we recognize that some other scholars might question our choice of proxy variables and transformations employed to create composite variables as not perfectly representing the constructs of loyalty and shopping involvement. We are fully aware of this limitation and make it clear by referring to those variables as proxies. At the same time, we do believe that our choice of variables is defensible, as they do seem to capture significant behavioural aspects of their underlying constructs, and are likely to be linked with them, if not in a directly causal manner, then at least through statistical correlations. Follow-up research could attempt to establish the level of affinity between proxy variables like ours with their corresponding constructs by combining our method of relying on online vendor databases with surveys using multivariate scales for measuring loyalty and shopping involvement.

Third, although the final model shows an acceptable level of explanatory power, there are still aspects of consumer behaviour and attributes and the vendor’s marketing policy that are unaccounted for. Thus, it is possible to extract more information from the consumer records to augment the model and improve its interpretability and goodness-of-fit. For example, the investigated vendor encouraged customers to leave opinions about their purchases, which could provide additional valuable data inputs into the model if not for the troubles with merging the two databases holding consumer shopping records and their expressed opinions.

Further conceptual and experimental work is required to develop next generation recommender algorithms. We addressed only partially the following crucial research areas: (1) multiple-input behavioural interaction types (e.g., clicks, purchases, add-to-cart, geo-locations), (2) multiple-input attribute modalities (e.g., text, image, video, numerical data, and other), (3) ease of adding new back-end algorithms, (4) specialized techniques to fuse multiple modalities, (5) effective deep learning models for visual search, (6) recommendations with and without session information, (7) high efficiency and scalability (services architecture), (8) convenient infrastructure for model evaluation and performance measurements.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-022-00579-3.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ajzen, I. (2020). The theory of planned behaviour: Frequently asked questions. Human Behaviour and Emerging Technologies, 2(4), 314–324

Alayrac, J. B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., & Zisserman, A. (2020). Self-supervised multimodal versatile networks. Neurips. https://github.com/deepmind/deepmind-research/tree/master/mmv

Alonso-Dos-Santos, M., Soto-Fuentes, Y., & Valderrama-Palma, V. A. (2020). Determinants of mobile banking users’ loyalty. Journal of Promotion Management, 26(5), 615–633

Basaj, D., Rychalska, B., Dabrowski, J., & Goluchawski, K. (2020). Synerise at Sigir Rakuten data challenge 2020: Efficient
Data modalities, consumer attributes and recommendation performance in the fashion industry

manifold density estimator for multimodal classification. Sigir-E-commerce Workshop. Retrieved from https://sigir-ecom.github.io/ecom20DCPapers/SIGIReeCom20DCPaper3.pdf

Beldad, A., De Jong, M., & Steehouder, M. (2010). How shall i trust the faceless and the intangible? a literature review on the antecedents of online trust. Computers in human behaviour, 26 (5), 857–869

Bleier, A., & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. Marketing Science, 34(5), 669–688

Brehm, J. W. (1966). A theory of psychological reactance. New York: Academic Press

Buháls, D., López, E. P., & Martínez-Gonzalez, J. A. (2020). Influence of young consumers’ external and internal variables on their e-loyalty to tourism sites. Journal of Destination Marketing & Management, 15, https://doi.org/10.1016/j.jdmm.2020.100409

Cao, W., Feng, W., Lin, Q., Cao, G., & He, Z. (2020). A review of hashing methods for multimodal retrieval. IEEE Access, 8, 15377-15391.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for domain understanding for accelerated solution development. IEEE Sensors Letters, 3(1), 1–4

Chang, H. H., & Chen, S. W. (2008). The impact of online store environment cues on purchase intention: Trust and perceived risk as a mediator. Online Information Review, 32(6), 818.

Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., & Zha, H. (2019). Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 765–774)

Chih, W. H., Huang, L. C., Yang, T. J., & Cho, H. P. (2009). Online relational bond, trust, and customer loyalty. International Conference on Information Management and Engineering (pp. 138–141). IEEE

Chou, C., Jin, X., Mueller, A., & Ostadabbas, S. (2019). Multimodal data fusion: moving from domain-specific algorithms to trans domain understanding for accelerated solution development. IEEE Sensors Letters, 3(1), 1–4

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. 4th ACM Conference on Recommender Systems (pp. 39–46)

Dabrowski, J., Rychalska, B., Daniluk, M., Basaj, D., Goluchowski, K., Babel, P., & Michalowski, A. (2021). An efficient manifold density estimator for all recommendation systems. Neuronal Information Processing: 28th International Conference ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021. Proceedings, Part IV Dec 2021 Pages 323–337. https://doi.org/10.1007/978-3-030-92273-3 27

Da Costa, A. F., & Manzato, M. G. (2016). Exploiting multimodal interactions in recommender systems with ensemble algorithms. Information Systems, 56, 120–132

Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. 13th ACM Conference on Recommender Systems (p. 101–109). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3298689.3347058

De Cannière, M. H., De Pelsmacker, P., & Geuens, M. (2009). Relationship quality and the theory of planned behavior models of behavioral intentions and purchase behavior. Journal of Business Research, 62(1), 82–92

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Diallo, M. F., Moulins, J. L., & Roux, E. (2020). Unpacking brand loyalty in retailing: a three-dimensional approach to customer-brand relationships. International Journal of Retail & Distribution Management, 49(2), 204.

Dick, A. S., & Basu, K. (1994). Customer loyalty: toward an integrated conceptual framework. Journal of the Academy of Marketing Science, 22(2), 99–113

Domínguez, M. A., Gouyon, F., Jorge, A. M., Leal, J. P., Vinagre, J., Lemos, L., & Sordo, M. (2013). Combining usage and content in an online recommendation system for music in the long tail. International Journal of Multimedia Information Retrieval, 2(1), 1–13

Fitzsimons, G. J., & Lehmann, D. R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. Marketing Science, 23(1), 82–94

Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5), 829-864.

Gefen, D. (2000). E-commerce: the role of familiarity and trust. Omega, 28(6), 725–737

Griffin, J. (2002). Customer loyalty: How to earn it, how to keep it. John Wiley & Sons

Hsu, M. H., Yen, C. H., Chiu, C. M., & Chang, C. M. (2006). A longitudinal investigation of continued online shopping behavior: An extension of the theory of planned behavior. International Journal of Human-Computer Studies, 64, 889–904

Ji, Y., Sun, A., Zhang, J., & Li, C. (2022). Recommender may not favor loyal users. arXiv preprint arXiv:2204.05927

Johansson, P. (2003, September). Madfilm-a multimodal approach to handle search and organization in a movie recommendation system. Proceedings of the 1st Nordic Symposium on Multimodal Communication (pp. 53-65).

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems, 44(2), 544–564

Kinley, T. R., Josiam, B. M., & Lockett, F. (2010). Shopping behaviour and the involvement construct. Journal of Fashion Marketing and Management, 14(4), 562–575

Kormbliith, S., Shiens, J., & Le, Q. V. (2019). Do better imagenet models transfer better? Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2661-2671).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

Kwon, S. J., & Chung, N. (2010). The moderating effects of psychological reactance and the involvement construct. Journal of Fashion Marketing and Management, 14(4), 562–575

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE, 103(9), 1449-1477.

Liu, K., Li, Y., Xu, N., & Natarajan, P. (2018). Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730

Lynch, S., & Barnes, L. (2020). Omnichannel fashion retailing: examining the customer decision-making journey. Journal of Fashion Marketing and Management: An International Journal.

O’Cass, A. (2000). An assessment of consumers product, purchase decision, advertising and consumption involvement in fashion clothing. Journal of Economic Psychology, 21(5), 545-576.
Patten, E., Ozuem, W., & Howell, K. (2020). Service quality in multichannel fashion retailing: An exploratory study. *Information Technology & People.*

Parra, D., & Amatriain, X. (2011, July). Walk the talk. *International Conference on User Modeling, Adaptation, and Personalization,* pp. 255-268. Springer, Berlin, Heidelberg.

Pradel, B., Sean, S., Delporte, J., Guérif, S., Rouveiro, C., Usunier, N., & Dufuia-joel, F. (2011, August). A case study in a recommender system based on purchase data. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 377-385.

Rawat, B., & Dwivedi, S.K. (2019). Selecting appropriate metrics for evaluation of recommender systems. *International Journal of Information Technology and Computer Science,* 1, 14–23.

Reichheld, F. F. (1996). *The loyalty effect: The hidden force behind growth, profits and lasting value.* 1996. Harvard Business School Press, Boston, MA.

Rendle, S. (2010, December). Factorization machines. *2010 IEEE International conference on data mining* (pp. 995-1000). IEEE.

Rossetti, M., Stella, F., & Zanker, M. (2016, September). Contrasting offline and online results when evaluating recommendation algorithms. *Proceedings of the 10th ACM conference on recommender systems* (pp. 31-34).

Saha, A., Nawhal, M., Khapra, M. M., & Raykar, V. C. (2018, March). Learning disentangled multimodal representations for the fashion domain. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 557-566). IEEE.

Salo, J., & Karjalisto, H. (2007). A conceptual model of trust in the online environment. *Online information review,* 31(5), 604.

Srinivasan, S. S., Anderson, R., & Ponnavolu, K. (2002). Customer loyalty in e-commerce: an exploration of its antecedents and consequences. *Journal of Retailing,* 78(1), 41–50.

Truong, Q. T., & Lauw, H. (2019, May). Multimodal review generation for recommender systems. *The World Wide Web Conference* (pp. 1864-1874).

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.