Applied Soft Computing 141 (2023) 110283

Contents lists available at ScienceDirect

Applied Soft Computing

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

Disaggregated retail forecasting: A gradient boosting approach

Luiz Augusto C.G. Andrade *, Claudio B. Cunha

Department of Transportation Engineering, Escola Politecnica, Universidade de São Paulo, Av. Prof. Almeida Prado, Travessa 2, n 83 Cidade Universitária ASO, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 8 May 2022
Received in revised form 30 March 2023
Accepted 4 April 2023
Available online 13 April 2023

Dataset link: https://www.kaggle.com/c/favorita-grocery-sales-forecasting

Keywords:
Forecasting
Machine learning
Retailing

ABSTRACT

Demand forecast is a relevant topic for retailers to manage effectively inventories comprising a wide range of Stock Keeping Units (SKUs) at store level. While this disaggregated forecast problem is central to ensure profitability as it supports accurate inventory decisions, thus avoiding either out-of-stock events or overstocks and inventory losses, it is also very complex due to aspects such as the large number of stores and products of modern retailers, the complex marketing and promotional strategies that impact customer demand together with cross-product effects that are all difficult to model. In this study, we propose more effective methods to handle these aspects. More specifically, we employ XGBoost, a non-linear non-parametric ensemble-based model, as the central learning algorithm and a structural change correction method to account for sudden changes in consumer behavior caused by external factors. Our approach also encompasses data cleansing procedures to correct sales observations during out-of-stock days as well as discrepancies between logical and physical inventory counts. Based on real data from a public dataset of a large retailer, we show that our methods outperform the Base-Lift model, a widely used benchmark model for retail forecasting, yielding significant improvements in accuracy metrics together with reductions in stockouts and in stock on hand. The proposed approach has also a high degree of automation, an important requirement for modern retailers.

Code metadata

Link to reproducible Capsule: https://doi.org/10.24433/CO.4410218.v1.

1. Introduction

The retail sector has a representative share in the global economy. It is the largest private employer and economic sector, comprising over 15% of the global GDP [1]. According to a market survey conducted by Deloitte Touche Tohmatsu Limited [2], the aggregate revenue of the top 250 retail companies globally was US$4.85 trillion in 2019, with an average size of US$19.4 billion per company; North America accounted for 47% of the revenues, followed by Europe and Asia-Pacific with 33% and 16%, respectively. Also, the global retail market reached a value of nearly US$20.33 trillion in 2020, having increased at a compound annual growth rate of 2.4% since 2015. This sector is expected to grow at a compound annual growth rate of 7.7% from 2020 to reach $29.45 trillion in 2025. Fast-moving consumer goods represent 66% of the retail market, and it is highly dependent on accurate business forecasting.

The process of business forecasting is paramount for inventory planning and operations profitability in the retail segment [3]. Inaccurate forecasts may result in undersupply or oversupply of inventory. In other words, forecasting lower than customer demand result in out-of-stock (OOS) events that lead to a lost revenue that oftentimes cannot be reversed as dissatisfied consumers may eventually switch to other retail chains; conversely, forecasting higher than customer demand result in overstocks and inventory losses. In the competitive environment of a globalized marketplace, companies' working capital must be as efficient as their operational processes. A study performed in 2015 showed that in 2014, retailers in North America made a loss of $634.1 billion due to products being out-of-stock and spent $471.9 billion on overstocking [4].

Tiacci and Saetta [5] report success cases in which improvements in demand forecasting accuracy result in lower inventories and increased customer service level. This specific problem of demand forecasting has received significant importance in business practice and in the specialized literature, as demonstrated by the authors. Fildes et al. [6] present a review

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: luiz.andrade@infracommerce.com.br (L.A.C.G. Andrade), cbcunha@usp.br (C.B. Cunha).

https://doi.org/10.1016/j.asoc.2023.110283
1568-4946/

of state-of-the-art applications of demand forecasting in the operations research (OR) context. They also highlight the significant interest of the OR scientific community in forecasting practices.

Some characteristics of retail operations make forecasting challenging: (i) the decentralized nature of retail operations require that forecasts are made for each store and for each Stock Keeping Unit (SKU), so that the forecasted values are used as inputs for inventory policies, (ii) sales are highly influenced by promotions and marketing activities, (iii) the impact of such promotions and marketing strategies is not constant over time and (iv) product life cycles have become shorter. Consequently, it is increasingly difficult to forecast sales for a specific SKU in a particular store since time series tend to be short as well [7].

Considering that modern retailers often sell tens of thousands of products in multiple locations, designing a forecasting method for each SKU in each store requires a high level of automation. Typical time series models, such as ARIMA [8] and State Space models [9] require expert intervention to identify model order or to define its structural matrix. Therefore, conventional time series methods do not comply with the automation level required in retail forecasting.

According to Ma et al. [10], aspects such as intra-category promotion schedules, inter-category promotion schedules can influence sales and, therefore, must be considered when performing sales forecasts. Not to mention that relationship between marketing actions and sales uplift cannot be assumed to be linear and hence linear models fail to fully capture its effects.

Huang et al. [11] address the forecasting problem at the SKU level in the presence of structural changes, which can be described as rapid changes in consumer behavior due to market disruptions (e.g., a new competitor entering the market, new product releases or disruptions in supply chains). The authors claim that recent studies focused on the SKU level forecasting consider marketing and promotion effects; however, such studies assume that these effects are constant over time, and do not consider the possibility of structural changes as well. Their models also consider constant marketing effects but deal with structural changes by simply correcting out-of-sample forecasts with the historical forecast bias, which results in less biased estimates. Another clear example of the need for such structural change corrections is the consumer behavior change due to the recent Coronavirus outbreak in 2020, in which hoarding, pent-up demand and online consumption have drastically changed what, when and how consumers buy their products [12].

Other important aspects when dealing with retail demand forecast at such disaggregated level is the data collection process and information quality. Boone et al. [13] present a study on the impacts of information availability and quality for demand forecasting in the supply chain. The authors claim that the popularization of information systems in the supply chain and the recent digitalization of the consumer has enabled the use of sophisticated data processing techniques. According to Gür Ali et al. [7] increasing information capability and data collection costs makes sense only if coupled with the use of more sophisticated models, such as machine learning models. One aspect to be considered in data collection processes is that, due to the dynamics of point of sales operations, the inventory data may not reflect the physical inventory, which can also harm the forecasting model accuracy.

From a mathematical modeling perspective, forecasting models comprise a research topic in itself. It encompasses different areas of applied math, from linear regression models to stochastic processes and state space modeling [14] to recent machine learning algorithms [11]. The choice of the best technique is specific for each case and depends on the characteristics of the process under analysis. Considering the retail setting, in which automation is an important factor [15], there are multiple marketing and promotion variables that influence sales and there are risks of structural change; non-parametric models trained with machine learning algorithms fit the requirements. Boone et al. [13] point that typical retail forecasting applications have large and sparse datasets, in which the challenge comes from choosing the most suitable variables, and machine learning models posit from this scenario, being more effective than traditional time series and multilinear regression models.

This research focuses on improving demand forecasting for each SKU in each store of a retail company to improve replenishment planning which in turn leads to stockout and loss reduction. More specifically, we propose a method for building forecasting models for retailers that builds upon the ideas of Huang et al. [11], Fildes et al. [16] and Gür Ali et al. [7] and tries to address the following challenges: (i) the need to automate the forecasting process given the large number of stores and products; (ii) inventory data inaccuracy, (iii) large and sparse datasets; and (iv) dealing with marketing, promotional, intra-category and structural changes effects. Our method brings new ideas regarding data cleansing, feature engineering and the use of non-parametric modeling methods for retail.

Our key contribution is to propose a combination of efficient methods in an effective manner to handle the problem of disaggregated forecasting sales for multiple SKUs at many stores in the presence of cross-product effects as well as marketing and promotional strategies that impact customer demand. This is accomplished in the following manner. First, we integrate XGBoost, a non-linear non-parametric ensemble-based model as the central learning algorithm with a statistical identification method and a data imputation method for improving low accuracy inventory information, especially to correct sales observations during out-of-stock days as well as discrepancies between logical and physical inventory. In this sense, our approach differs from Huang et al. [11] and Fields et al. (2018), as both rely on linear models to perform feature selection and forecasting. The closest application we found in the literature is proposed by Gür Ali et al. [7], in which linear regression models and regression trees are compared for the disaggregated retail forecasting problem. Second, we propose pooling time series of all products at all stores into a single dataset, once not only does it favor the use of non-parametric models that usually require larger datasets than probabilistic parametric models but also has the advantage of not assuming any specific structure about the model variables [17]. We also introduce a structural change correction method to account for sudden changes in consumer behavior caused by external factors that is embedded in a non-parametric model framework, thus allowing the automatic identification of relevant features. This is an important feature for modern retailers who often sell tens of thousands of products in their stores [11]. The results based on real-world data suggest that our approach outperforms other methods when compared to widely used benchmark model for retail forecasting, indicating that tree-based methods, such as GBMs, are promising candidates for solving the disaggregated forecasting problem.

The remainder of this paper is organized as follows: Section 2 provides a review of the relevant related literature, comprising the advancements of former retail forecasting techniques at SKU-store level as well as key aspects of gradient boosting machine theory and related applications. Section 3 presents a formal definition of the disaggregated retail forecasting problem. In Section 4 we describe our proposed approach in detail, from data cleansing and pre-processing to model training, forecasting and evaluation. The experimental setting used for validating the proposed approach is presented in Section 5. We compare our results with the Base-Lift model, which is the most commonly used model in practice [16]; the results are summarized and discussed in

Section 6. Finally, Section 7 comprises our concluding remarks, including a discussion of the limitations of this research and give future recommendations for researchers and retail forecasting practitioners as well.

2. Literature review

This section is organized into two distinct parts. We initially present a comprehensive review of the literature on retail forecasting at disaggregated level. Firstly, we discuss its importance; secondly, some key aspects such as stockouts, seasonality, promotions, special days and events that boost demand and cannibalization; we also revise the different methods and approaches for forecasting at this level. The second part focuses on the gradient boosting machine, the technique we propose to tackle this complex problem of retail forecast.

2.1. Retail forecasting at SKU level

2.1.1. Relevant aspects

According to Fildes et al. [16], retailers rely on forecasts to support strategic, tactical and operational decisions, and each level has a different goal. At the operational level, forecasts are performed in SKU per store level and support decisions, such as replenishment plans, inventory and space planning and store staff scheduling.

As highlighted by Fildes et al. [6], retailers' ability to accurately perform operational level forecasts over the short term directly increases customer satisfaction, minimizes waste and write-offs, and leads to increased sales and efficient distribution. They propose a definition for the forecast level of aggregation. Forecasts at market level, chain or store level are defined as aggregated forecasts and are usually performed in monetary units to support strategical and tactical decisions while forecasts at SKU-store level are defined as disaggregated forecasts and are often performed in units to support operational decisions, especially those related to replenishment and inventory optimization.

Aggregated forecasts often exhibit strong trends, seasonal cycles, and serial correlation [16]. Therefore, time series models provide useful solutions for the aggregated version of the retail forecasting problem. Some authors have found non-linear effects in aggregated retail sales series and applied non-linear methods, mainly artificial neural networks, to tackle the problem [18]. [19].

Another type of aggregated forecast is store-level forecast. In this case, forecasts are usually made in monetary units or even in consumer traffic for brick-and-mortar stores [20]. In such cases, univariate time series models or simple regression-based models provide better forecasts than expert judgement [21].

According to Syntetos et al. [22], the disaggregated retail forecasting problem can be characterized by three dimensions: time, product and supply chain dimensions (Fig. 1). The time dimension represents the time granularity of forecasts; the product dimension denotes the aggregation level in product hierarchy while the supply chain dimension represents the aggregation level geographically or channel-wise. In general, the higher the level of decision-making, from operational to strategical, the higher time aggregation is. Replenishment decisions require weekly or even daily forecasts, depending on a store delivery schedule. The SKU level is the smallest planning unit in product hierarchy used for planning daily inventory, to forecast store demand or even across the entire supply chain in case of promotion planning. In terms of supply chain hierarchy, store or even DC level forecasts are considered for inventory planning and replenishment decisions, which are valid for brick-and-mortar stores. For online retailers, regional forecasts are often more useful for the purposes of inventory positioning and planning.

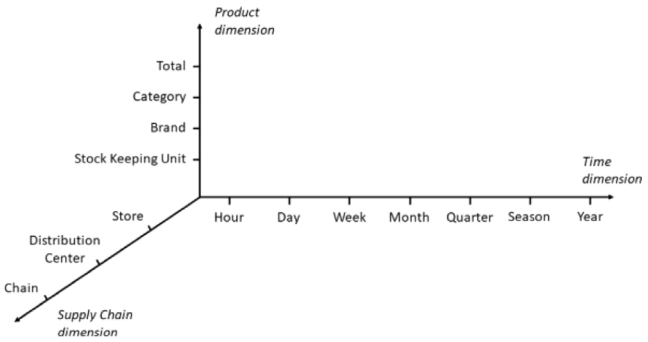

Fig. 1. Three dimensions of disaggregated forecasting problem. *Source:* Syntetos et al. [23].

Given the scope and objective of this research, the aggregation levels we consider are SKU, store and weekly levels in terms of product, supply chain and time dimensions, respectively.

Forecasting at SKU-store level requires taking into consideration that inventory and sales information are subject to inaccuracies. Sales data typically come from PoS systems transactions, which have biased sales information meaning that the actual demand differs from the observed sales [16]. When facing a stockout, customers may leave the store and not return later to purchase the product; alternatively, consumers may turn to substitute products. Both behaviors introduce bias in the observed sales. According to these authors, models that focus on stockout events are often explanatory rather than predictive and require more information than what is usually available. Therefore, models found in literature may fail to consider or accommodate these data inaccuracies that are inherent in retail data. In addition, any planning system for retailers needs to consider that inventory information is notoriously inaccurate, as highlighted by DeHoratius and Raman [24].

Retailer sales time series may also be influenced by several seasonal cycles. Hence, at any aggregation level, models that try to capture demand patterns should consider these cycles. One way of modeling such cycles is by introducing dummy variables as suggested by Huang et al. [25]; however, this may lead to an excess of variables and a dimensionality problem. Another more parsimonious way of modeling these cycles is by using trigonometric functions as in Huang et al. [11], which also offers the advantage of considering the time regularity of the cycles.

Another important factor in retail forecasting at disaggregated level is to consider seasonal or calendar events (e.g., Christmas, Black Friday, Valentines, Mother's Day), as well as include holidays and store events, such as sales promotions. Some holidays are cyclic and could be modeled as a seasonal cycle, but there are cases that they do not exhibit a cyclic pattern (e.g., holidays that do not have a fixed calendar date, such as Easter), and the standard way of considering such events in forecasting models is by using dummy variables [26].

For some products, sales may be strongly affected by weather conditions. There are three main factors that make the use of weather data difficult in retail forecast: (i) the fact that weather data, unlike control variables, such as price and promotion, must be forecasted, (ii) weather forecast at store level is hardly available in most cases, (iii) and medium-term weather forecast is typically not reliable [16]. To the best of our knowledge, no research in retail forecasting at SKU-store level has been capable of effectively using weather information to forecast future sales.

Marketing mix and promotions are other important aspects of retailers' operations, especially for inventory planning, since a marketing strategy may not accrue its desired outcome if the

Table 1
References and their impacts on the proposed method.

Year	Author	Related Contribution
2003	Alon, Qi and Sadowski	Neural networks capable of represent nonlinear effects in time series
2005	Zhang and Qi	Neural networks capable of represent nonlinear effects in time series
2005	Srinivasan, Ramakrishnan and Grasman	Use of cannibalization models in the process of business forecasting
2008	Fildes et al.	Taxonomy for different forecasting aggregation levels and its implications
2008	DeHoratius and Raman	Impact of inaccurate inventory in forecasting problems
2014	Huang, Fildes and Soopramanien	Algorithm to incorporate competitive information represented by product clusters
2015	Lang et al.	Data pooling recommendations
2016	Syntetos et al.	Taxonomy fo different dimensions of forecasting problem: time, product, sales channel
2016	Ma, Fildes and Huang	Use of intra product category promotion information
2018	Fildes, Ma and Kolassa	Review of state-of-the-art applications of demand forecasting in the operations research (OR) context
2019	Huang et al.	Incorporating structural change in the forecasting process

right amount of inventory is not transferred to stores to support the increase in demand. One aspect of retail operations that make marketing effects difficult to be accounted for in forecasting models is the pressure for novelty and innovation [27]. This makes retailers' marketers usually combine marketing actions with interaction effects, which translates into the need to predict the first instance of a campaign in most cases. As shown by Fildes et al. [16], it appears that the interaction effects of different marketing tactics have not yet been addressed in the forecasting literature.

Also, marketing tactics may produce cross-product effects: a sales incentive in one product may decrease sales in a substitute product (in other words, cannibalization) or increase sales in a complementary product (a.k.a.,complementarity). Incorporating cannibalization relationships may improve accuracy in aggregate forecasts but requires collecting additional information other than historical sales and prices as shown by Srinivasan et al. [28]. The same argument is valid for complementary products. Ma et al. [10] proposed a method for handling high dimensional forecasts with inter and intra product category promotion information.

Data pooling is also a factor to account in forecasting at SKU-store level. Whether to pool all data together or to group per store or per category remains an open question and is subject to judgement [16]. According to the authors, pooling datasets together assume homogeneity in the demand generation process and by pooling datasets that are heterogeneous the forecast equations of parametric models may become mis-specified, in other words, the equations of parametric models that assume constant parameters such as sales distribution mean and variance are not valid since these parameters are heterogeneous in the pooled dataset. Lang et al. [29] found that allowing heterogeneous data pooing while using flexible modeling functions leads to forecast improvement in the store level forecast.

Table 1 summarizes the references mentioned in this section and their relevance to our proposed methodology. It should be noted however, that the table is not exhaustive and contains only the references that had a direct influence in our model.

2.1.2. Forecasting models

The most basic approach for disaggregated demand forecasting comprises simple univariate methods, such as Exponential Smoothing [7], Autoregressive Integrated Moving Average methods (ARIMA) [8] and state space models [9]. Extensive literature has evidenced that these methods work well in aggregated forecasting without promotional influence but fail to produce accurate forecasts in cases the one specific product is subject to promotional drivers [7].

The most widely used method for SKU-store forecast in case of promotional drivers is the Base-Lift method [16]. This is a two-step method: first, a base forecast is generated by applying a standard time series model to a time series that has been previously cleansed from promotional effects; second, a lift is estimated based in the last similar promotion. The authors claim that the widespread use of this method is due to the installed base of software packages and not related to a thorough examination of its effectiveness. In fact, Fildes et al. [30] empirically showed that these lift adjustments increase forecasting bias. Also, considering the SKU-store level forecasting of a medium-sized retailer, manually adjusting the lift for every single SKU-store pair may be unfeasible in practical terms.

Aside from univariate or Base-Lift models, other streams of studies regarding SKU-store level forecasting comprise methods based on multiple linear regression [10,11,26,31]. These approaches can be deemed effective in the sense that they are capable of capturing demand and seasonal patterns, promotional drivers, and other exogenous factors; however, this is possible by assuming either a linear or a parametrized linear structure. Nonlinear effects and interactions cannot be considered by such models; consequently, they fail to incorporate such effects in SKU-store forecasting.

Nonlinear methods have also been briefly applied to SKU-store forecasting, which include nonlinear regressions, non-parametric regressions and machine learning algorithms that allow the use of generalized approximation functions that are learned from data. This less restrictive approach to model structure potentially leads to improved forecasting models. Ainscough and Aronson [32] compared the application of neural networks and linear regression and found forecasts are improved by using nonlinear methods. Pillo et al. [33] tested the application of support vector regression (SVR) for forecasting retail sales in the presence of promotions for a specific kind of pasta in two specific stores of a retailer. They concluded that the SVR achieved a superior result, but the comparison was made based on a small dataset.

Gür Ali et al. [7] compared the accuracy of different types of models, including stepwise linear regressions, SVRs and regression trees as well as the Base-Lift model that was used as benchmark. The authors considered a medium-sized full-service grocery retailer in Europe comprising 4 stores and one product category, totaling 168 pairs of SKU-store combinations. They also evaluated the effects of data pooling across stores and categories

in the overall accuracy. Their results revealed that the more complex a model is, the more it benefits from additional features; in addition, it only makes sense to add many features if the forecasting model has enough capacity to model the complexity of relationships between variables. They also showed that the regression tree model achieved the best result in terms of accuracy. Interestingly, all models were outperformed by the Base-Lift benchmark model in non-promotional periods; the benefits from employing more complex models and enhanced features were only noticeable during promotional periods.

Motivated by the influence of price and quality on demand of dairy products in Iran, in which a relevant aspect was to identify key factors that affect demand, Goli et al. [34] proposed a demand prediction model based on hybrid artificial intelligence tools in which the learning phase is improved by employing different meta-heuristic algorithms for selecting the effective features (e.g., population, inflation, price, etc.) that affect demand and should be used as inputs for the prediction network. In a related subsequent work, Goli et al. [35] also considered the Gray Wolf optimization algorithm for determining the features. Their results show that adaptive neuro-fuzzy interface system (ANFIS) had the most support for improvement with the help of the meta-heuristic algorithms compared to multi-layer perceptron neural network (MLP) and support vector regression (SVR). It should be noted, however, that their prediction motivation differs from ours as their main concern was to determine the key factors and not to provide forecast at disaggregated level (multiple SKUs, store level).

Another group of nonlinear models for forecasting comes from the deep learning literature. Flunkert et al. [36] proposed a recurrent autoregressive neural network architecture called DeepAR, which was designed to output forecast densities considering a negative binomial distribution. This approach is beneficial for SKU-store forecast since the safety inventory of replenishment plans is determined by the forecasting probability distribution dispersion. The authors present results for data from Amazon's e-commerce; however, no benchmark comparison was performed. Similarly, Wen et al. [37] proposed a sequence-to-sequence neural network that generates multi-horizon quantile forecasts. Both applications are centered in the e-commerce setting and have little evidence of outperforming benchmark methods, despite the presented results being promising.

More recently, Fildes et al. [16] presented a though examination of the SKU-store level forecasting literature. The authors highlighted that the evidence for nonlinearity generally leading to better forecasting accuracy is weak, with the positive evidence probably arising from a publication bias and lack of comparison with benchmark models. According to them, the approaches proposed in the literature have little in common in terms of methodology and are not generalized approaches; thus, it is not feasible to compare the relative performance between models without benchmark testing. They also indicate that, in general, multivariate models show substantial improvement over univariate benchmarks; however, the proposed multivariate models were proposed focusing on groceries and hence they need to have their applicability tested on other retail categories.

To summarize, there is a gap in the literature related to retail demand forecast at disaggregated SKU-store level that properly allows to consider several aspects, including inaccurate historical data, seasonality, non-linear promotional and marketing effects and structural changes in such a way that models can be automatically estimated and provide accurate, reliable and consistent forecasts. This may avoid the pitfalls of some recent and sophisticated models that fail to perform well for different conditions.

Table 2 summarizes the related research and their modeling approaches.

3. Problem definition

The problem that we address is the disaggregated retail demand forecasting problem to support operational decisions, such as store replenishment and optimizing inventory policies. Therefore, historical data is given in daily time buckets, and future demand should be estimated in weekly time buckets [10,11].

The problem is defined for a retailer $\mho$ that sells k different products through i different outlets, which may comprise both brick-and-mortar stores and the direct sales through, for instance, B2C e-commerce channel. Sales occur across time periods that comprise different consecutive days t. Given a set of disaggregated historical sales data for each pair (i, k) for a historical period of t days, $t \in \{0, 1, \ldots, T\}$, our aim is to estimate weekly future demand comprising the next H weeks ($w = W, W+1, \ldots, W+H$). Each product k should be interpreted as the SKU level in product hierarchy or the product inventory planning unit.

Let $y^t_{i,k}$, $I^t_{i,k}$ and $p^t_{i,k}$ be the observed (past) sales, opening inventory and average price of product k at store i on day t, respectively. The historical values are known from period 0 to period T.

To take into account data quality issues that may arise in retail, it is necessary to assume that the observed inventory may be different from the real inventory [24]. Let $I'^t_{i,k}$ be the corrected inventory of product k at store i during day t. The inventory correction method (Section 4.2), named $I_{\text{Correction}}$, aims to probabilistically identify days on which the time span since the last observed positive sale diverged from the historical distribution of periods between positive sales. For periods in which the probability is low, it is assumed that an inventory inaccuracy is present, and the observed inventory value is corrected to 0.

$$I_{\text{Correction}}(y^t_{i,k}, I^t_{i,k} | t = 1 \ldots T) \rightarrow I'^t_{i,k} \tag{1}$$

Let $d^t_{i,k}$ represent the potential demand for product k at store i on day t. The difference between variables y and d arise in the presence of stockouts. Stockouts so^t_{ik} are defined by days on which both variables y and I' are equal to 0.

$$so^t_{ik} = \begin{cases} 1 \; if \; I'^t_{ik} = 0 \; and \; y^t_{ik} = 0 \\ 0 \; otherwise \end{cases} \tag{2}$$

We also propose a method for correcting the observed sales and estimating the expected demand (Section 4.3). For a given store and product, if a stockout is identified in period t, the observed sales are substituted by a locally defined estimate $y'^t_{i,k}$, which is equivalent to the expected demand on the same day $d^t_{i,k}$.

$$y_{\text{Correction}}(y^t_{i,k}, so^t_{i,k} | t = 1 \ldots T) \rightarrow y'^t_{i,k} = d^t_{i,k} \tag{3}$$

The problem we address aims to estimate future demand in weekly time buckets. Therefore, historical corrected sales, inventory and price data are all aggregated into weekly historical buckets. Hence, for each pair (i, k), let $d^w_{i,k}$ be the aggregated demand for week w, $I^w_{i,k}$ the average inventory for week w and $p^w_{i,k}$ the average price in week w. Equivalently, these three variables are known for periods $(0, \ldots, W)$.

In addition to observed sales, inventory levels and prices, we also consider other demand drivers in retail (i.e., promotions, marketing strategies, calendar events and seasonality) as input features to explain demand patterns.

Promotions refer to discounts in products incumbent prices. As with Fildes et al. [16], we also assume that customers are more attracted by the magnitude of discounts than the final product price after discount. Thus, we define $promo^w_{ik}$ as the relative discount given at store i for product k in planning period w over its incumbent price $p^w_{i,k}$.

Marketing strategies may have diverse formats, such as displays, featuring in different media, announcements, and discounts

Table 2
Main retail forecasting modeling approaches.

Year	Authors	Modeling approach
1970	Box and Jenkins	Autoregressive Integrated Moving Average methods
1999	Ainscough and Aronson	Neural Networks
2009	Gür Ali et al.	Exponential Smoothing and Regression Trees
2012	Durbin and Koopman	State space models
2016	Ma et al.	Least absolute shrinkage and selection operator - Multiple linear Regression
2016	Pillo et al.	Support Vector Regression
2017	Flunkert, Salinas and Gasthaus	Deep Neural Networks
2017	Wen et al.	Sequence to Sequence Neural Network
2018	Fildes et al.	Least absolute shrinkage and selection operator - Multiple linear Regression
2019	Fildes et al.	Least absolute shrinkage and selection operator - Multiple linear Regression
2019	Huang et al.	Least absolute shrinkage and selection operator - Multiple linear Regression
2021	Goli et al.	Hydrid Artificial Intelligence – Multilayer Perceptron, Support Vector Regression and Adaptive Neuro-Fuzzy Interface System

strategies. In order to properly consider their impact on sales, it is necessary to classify each type of strategy. Hence, we define $dummy\_mkt^{w}_{cik}$ as a dummy variable indicating the presence of marketing strategy c applied to product k at store i during week w. In order to differentiate different intensities of instances of the same marketing strategy, we define $value\_mkt^{w}_{cik}$ as the budget of marketing strategy c applied to product k at store i during week w.

Regarding seasonality, we consider the recommended trigonometric functions proposed in Harvey [38] to represent the demand cyclic patterns. We consider a monthly and a yearly cycle represented by pairs of *sin* and *cos* functions.

Calendar events affect customer demand not only on the day of the event itself but also some weeks in advance [11]. We define dummy variables Cal_e^{w-v} that indicate that period $w-v$ is in the week of calendar event e. For example, if v equals 0, this variable indicates the week of the calendar event and, if v equals 1, it indicates the week before the event.

Competition also plays an important role in forecasting [25]. Cross-product pricing strategy may affect the customer willingness to buy products from one brand or another, or may even stimulate the customer to purchase complementary products. In this research, we assume that there is no competition between different stores. Considering a specific product k, we consider that all variables from all other products in a specific store influence the aggregated demand for week w ($d^{w}_{i,k}$).

In addition, we propose introducing information regarding product and store categorization. Products are often organized in families or categories. Stores are commonly segregated in clusters defined by geography and sales format (e.g., cash and carry, convenience and supermarket). We propose using dummy variables to represent such categorizations. Assuming that product category is a categorical variable with R different levels, we define R dummy variables Cat_{kr} that are equal to 1 if a product k belongs to category r and 0 otherwise. Analogously, we define Cat_{id} dummy variables for store categorization.

Following the findings by Gür Ali et al. [7], we also propose the introduction of historical sales statistics of the past weeks as explanatory variables. The utilization of average, median and standard deviation of past observed sales introduce contextual knowledge to the models and may improve forecast accuracy. For a given pair (i, k) in week w, we consider the average, median and standard deviation of demand in the past p weeks.

Considering the aforementioned groups of variables, we can define the disaggregated forecasting problem for each pair (i, k) as depicted in expression (4). It is worth noting that we pool all stores and products in the same problem. Also, we introduce variable *trend*, representing time periods to capture trends in each time series and as input for the trigonometric seasonal variables. Finally, we assume that historical variables affect $d^{w}_{i,k}$ up to lag l.

$$f\left(\begin{array}{c} d^{w}_{i,k}, p^{w}_{i,k}, promo^{w}_{ik}, dummy_{mkt}{}^{w}_{cik}, value_{mkt}{}^{w}_{cik}, d^{w}_{i,q}, p^{w}_{i,q}, \\ trend, Cal_e^{w-v}, Cat_{kr}, Cat_{id} \\ i \in I, k \in K, w = W-1, \ldots, W-l, q \in K, q \neq k, r \in R, d \in D \end{array}\right) \rightarrow d^{w}_{i,k}\{w = W+1, \ldots, H\} \quad (4)$$

4. Methodology

In this section, we describe the proposed methodology to solve the problem presented in Section 3. It relies on an inventory correction method, a sales correction method and the application of a specific Gradient Boosting Machine (GBM) implementation called XGBoost [39] to extract the demand pattern from input variables to output variables and a forecast correction method to account for the possibility of structural change in consumer purchasing behavior. We initially present some main concepts of GBM, followed by the specific details of our approach.

4.1. Gradient boosting machines

According to Natekin and Knoll [40], many machine learning problems can be summarized as building a single model based on a collected dataset of a specific process or phenomenon without having any particular domain theory or expert knowledge as assumptions. The procedure usually applied to such problems is to fit a non-parametric model, such as a neural network or a support vector machine [41]. Following this approach, one tries to build a single model to represent all possible relations among the problem variables. Another alternative approach is to build a collection of simpler models aiming to capture simpler patterns in the dataset, that together form a more robust model for the whole dataset. This approach is often called ensemble modeling and the resulting collection of models is called ensemble.

Gradient Boosting Machines are one type of ensemble in which weak learners are sequentially adjusted to the data and stacked together to compose a single robust model. The methodology was first proposed by Friedman [42] and is posed as a gradient descent method, in which each step consists in fitting a non-parametric model to the residues of a previous model. Natekin and Knoll [40] state that, in practice, the most common type of weak learner used are low depth decision trees. In this subsection, we thus describe the GBM theory focusing on this type of weak learner as a building block.

The GBM theory considers the function estimation problem in the classical supervised learning setting the goal of which is to find a function $F(x)$ that minimizes a loss function L over a set of N pairs of examples $\{x, y\}_{i=0}^{N}$. To make the problem tractable, it is common to restrict the function search space to $F(x; \theta)$, a function class parametrized by θ. In this case, the problem becomes a parameter estimation problem. For the sake of clarity, it is worth noting that despite F being parametrized by θ, what makes F parametric or non-parametric is the dimensionality of θ. In the parametric case, θ has a predefined dimension space and, in the non-parametric case, θ has an undefined dimension space.

Closed form solutions for θ^* are usually not available and, therefore, iterative numerical procedures are often applied to this problem [40]. The most commonly applied method is the steepest gradient descent, based on consecutive improvements in parameter set θ along the direction of the gradient of the loss function with respect to θ. In such case, the estimate of θ after M iterations, is given by a sum of consecutive improvements θ_i.

The difference between the classical steepest descent and the GBM theory is that the optimization is carried out in the function space rather than in the parameter space. Function $F(x)$ is defined as an additive function as depicted in expression (5), where M is the number of iterations, F_0 is the initial estimate and each F_i is called a step or a "boost".

$$F(x) = F^M(x) = \sum_{i=0}^{M} F_i(x) \tag{5}$$

As in the classical function estimation problem, it is common to use a parametrized function class $h(x; \theta)$, which corresponds to the aforementioned weak learner. At each iteration, the gradient boosting algorithm applies a greedy stage wise approach of incrementing $F(x)$ with the next step or boost that minimizes expression (6).

$$F^m(x) = F^{m-1}(x) + \rho_m . h(x; \theta_m) \tag{6}$$

$$(\rho_m, \theta_m) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} L\left(y_i, F^{m-1}(x_i)\right) + \rho . h(x_i; \theta) \tag{7}$$

The exact solution of expression (7) is equivalent to the solution of a line search optimization problem [43] which in turn equates to finding the gradient descent direction with respect to $F(x)$. Instead of solving the line search problem defined by (7), the gradient boosting method converts the problem into finding $h(x; \theta)$ which is most correlated with the sum of gradients $g_m(x_i)$ (expression (8)). This approach results in a classical least square minimization problem defined by expression (9).

$$g_m(x_i) = \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{\boldsymbol{F}=\boldsymbol{F}_{\boldsymbol{m}-1}} \tag{8}$$

$$(\rho_m, \theta_m) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} [g_m(x_i) + \rho . h(x_i; \theta)]^2 \tag{9}$$

To summarize, the steps of the gradient boosting methodology are presented in Fig. 2. More details about the methodology can be found in Hastie et al. [44]. There are many variations of the basic GBM varying the choice of the loss function and base learner.

In this research, we consider the XGBoost [39], which can be described as a scalable end-to-end tree boosting system that applies the GBM model framework.

4.2. Inventory correction method

Inventory inaccuracy is a common data quality issue in retail that may have harmful consequences to store management, such as fraud and inventory losses [45]. Taking into consideration solely the demand forecasting problem, this issue is only relevant in cases in which the observed inventory is not null, but the physical stock is equal to zero. This results in missing actual stockouts and observing zero sales in daily data. The identification of such events can only be estimated, since otherwise it could only be verified by comparing the inventory information and the physical inventory, which is impractical in most retail operations. We propose an adaptation of the technique developed by Karabati et al. [45] that is designed to identify product purchase substitution patterns based on stockouts and PoS data.

For each product k at each store i, let TBS_{ik} be a random variable representing the time between consecutive positive (i.e., non-null) sales. For each period t, we can calculate $last\_sale_{ik}^{t}$ as the number of periods since the last observed positive sale. We identify an inventory inaccuracy in period t for product k at store i if the single-sided probability of TBS_{ik} being greater than $last\_sale_{ik}^{t}$ is less than a constant ξ as in expression (10). If period t is identified as an inventory inaccuracy event, the observed inventory value is assumed as 0. More precisely, if the probability of observing a group of l consecutive zero sales is less than a threshold value, then the group of l consecutive periods is labeled as "inventory inaccuracy events".

$$P\left(TBS_{ik} > last\_sale_{ik}^{t}\right) = \xi \tag{10}$$

One question discussed by Karabati et al. [45] is how many observations of sales data should be considered to effectively characterize the distribution of TBS_{ik}. They suggest that the number of observations should be enough to characterize a stable probability distribution. If the linear combination of two random variables of a given probability distribution results in another random variable with the same distribution dislocated by the mean and variance then, this distribution is considered stable. In the case of retail forecasting, this parameter should be estimated based on expert knowledge or experimentation.

Another aspect to consider is the probability of false negatives or false positives and its relation to parameter ξ. Considering expression (10), a false negative is defined as missing the identification of an actual inventory inaccuracy and a false positive is defined as wrongly classifying a period as inventory inaccuracy. Larger values of ξ increase the probability of labeling a false negative and vice-versa. We consider the same value as Karabati et al. [45] who observed through numerical analysis that values close to 10^{-4} are adequate for the case of product substitution.

4.3. Sales correction method

A sales correction method is required to correct the sales information in periods of stockouts and periods in which inventory inaccuracy is identified, as discussed in Section 3. There are different correction methods in the literature for time series data, such as moving average filters, Kalman filters, linear interpolation or even substitution by the mean [46]. The choice of what method to apply depends on the series dimensionality, the relative number of observations to be corrected and the origin of information inaccuracy.

As time series are unidimensional, we assume, in the case of disaggregated forecasting, that the number of incorrect observations is small comparatively to the number of observations. We also consider that data inaccuracies occur at random, making the missing observations to be corrected randomly distributed; this enables us not to address the data collection process but rather to focus on correcting the collected data considering its probability distribution. Taking into account these two assumptions, we propose using a moving average filter to replace inaccurate sales

Inputs:
Dataset $\{x, y\}_{i=0}^{N}$; Number of iterations M; Choice of loss function $L(y, F(x))$; Choice of base learner $h(x; \theta)$

Algorithm

1. Initialize $F^0(x)$ as a constant that minimizes $L(y, F(x))$
2. **for** t = 1 to M
3. Compute the negative gradient $g_t(x)$
4. Fit a new base learner function $h(x; \theta)$ to $g_t(x)$
5. Find the best step size ρ by solving $\rho_t = \arg\min_{\rho} \sum_1^N L\left(y_i, F^{t-1}(x_i) + \rho h(x_i; \theta_t)\right)$
6. Update the function estimate $F^t(x) = F^{t-1}(x) + \rho_t . h(x_i; \theta_t)$
7. **end for**

Fig. 2. Friedman's gradient boosting algorithm.
Source: Natekin and Knoll [40].

information whose aim is to replace the inaccurate information by a locally smoothed estimate of the value.

Expression (11) represents the local estimate for a single-sided window of J periods. For each time series, we recommend that period parameter J be defined as equal to the largest number of consecutive periods identified with inaccurate information, either due to stockouts or to inventory inaccuracies. This can be explained as a large number of consecutive periods of zero observed sales leads to an underestimation of the substitution value, which can be mitigated by considering a number of periods in the moving average filter that is large enough to accommodate these consecutive periods of inaccurate information.

$$d_{ik}^{t} = \frac{\sum_{i=0}^{J} y_{ik}^{t-i-1}}{J} \tag{11}$$

4.4. Learning algorithm

We propose a single-stage algorithm to solve the problem defined in Section 3. The algorithm is applied after the methods for inventory correction and sales correction are applied to the observed sales and inventory data and then aggregated in weekly time windows. Differently from Huang et al. [11] and Ma et al. [10] that first identify groups of products that interact together, we rely on the XGBoost model [39] to find the relevant input variables and forecast future sales at the SKU-store level. We consider data from all products (SKUs) in a specific store to define the input variables and then pool data from all stores to compose the dataset.

We chose XGBoost as it has shown high effectiveness in different machine learning problems as well as for its efficient training time as described in Bentéjac et al. [47]. The comparison with other methods goes beyond the purpose of this research and is addressed in the conclusions section. We compare the XGBoost with the Baseline method as it is reported by Huang et al. [11] as the most used method in practical applications.

We apply a log transformation to the corrected sales and prices variables as in Huang et al. [11]. Observed sales and prices are non-negative and usually exhibit a log-normal shaped distribution; hence, the log transformation is appropriate to make the distribution of these variables symmetric.

For a given product k at store i, we consider past observed weekly sales and prices from weeks 1 to 4 (as in [11]) as well as week 52. This enables the model to search for relations in previous month sales and last year sales and prices in the same week. Promotions, and marketing variables are considered up to lag 4 [11]. We also consider three rolling statistics for the latest 4 periods: average, median and standard deviation of observed sales in the last 4 and 12 periods as in Gür Ali et al. [7] corresponding to the last month and last quarter statistics.

Calendar events should be defined considering the main holidays in each location. For example, in the United States, the main calendar events are Halloween, Thanksgiving, Christmas, New Year's Day, President's Day, Easter, Memorial Day, the 4th of July, and Labor Day. As in Huang et al. [11], we consider that calendar events affect demand one week prior and along its incumbent week.

To consider weekly and yearly seasonality, we consider pairs of trigonometric functions, one pair for each seasonality type, as suggested by Harvey [38]. We also include variable w to represent possible trends.

Data on store and product categories are also considered in the model. We use two sets of dummy variables to represent product categories and store categories.

As mentioned above, we consider that demand and prices of every product at a given store interact. Therefore, for a given product k at store i, we consider sales, price variables for every other product at store i up to lag 4 as inputs. Similarly, marketing strategies at a given store interact and affect all its products at the same time and, therefore, we consider marketing variables of every product at store i up to lag 4 as inputs.

Expression (12) denotes the inputs and output of the learning algorithms. Each variable between braces represents a set of variables with their respective indexes in subscripts. This expression can be interpreted as an instance of the general problem definition given by Expression (4).

It is worth noting that the model defined by (12) has a high dimensionality, and the determination of relevant variables is carried out by the gradient boosting methodology. Also, a model with such high dimensionality is prone to overfitting [17] and regularization techniques play an important role in the model training to avoid this. XGBoost has the advantage of dealing with high dimensional problems, as well as allowing the application of regularization techniques, such as subsampling during the training phase [39] (see Eq. (12) that is given in Box I).

4.5. Structural change correction method

As highlighted by Huang et al. [11], if a time series is subject to structural change, forecast models will generate biased results. To account for this, we apply the Intercept Correction (IC) method proposed by them to calculate the correction bias. If the time series is subject to structural change, we can estimate the forecast bias by taking the average of the past residuals as given by expression (13).

$$BIAS = 1/\lambda \sum_{i=1}^{\lambda} \left(d_{i,k}^{w} - \hat{d}_{i,k}^{w}\right) \tag{13}$$

$$\hat{d}^{w}_{i,k} = f\left(\begin{array}{c} \left\{\ln\left(d^{w-s}_{i,k}\right)\right\}_{s=1,2,3,4,52}, \left\{\ln\left(p^{w-s}_{i,k}\right)\right\}_{s=0,1,2,3,4,52}, \\ \left\{\ln\left(d^{w-s}_{i,q}\right)\right\}_{s=1,2,3,4;q\neq k}, \left\{\ln\left(p^{w-s}_{i,q}\right)\right\}_{s=0,1,2,3,4;q\neq k}, \\ \left\{promo^{w-s}_{i,k}\right\}_{s=0,1,2,3,4}, \left\{promo^{w-s}_{i,q}\right\}_{s=0,1,2,3,4;q\neq k}, \\ \left\{dummy_{mkt}{}^{w-s}_{cik}\right\}_{s=0,1,2,3,4;c\in C}, \left\{dummy_{mkt}{}^{w-s}_{ciq}\right\}_{s=0,1,2,3,4;c\in C;q\neq k}, \\ \left\{value_{mkt}{}^{w-s}_{cik}\right\}_{s=0,1,2,3,4;c\in C}, \left\{value_{mkt}{}^{w-s}_{ciq}\right\}_{s=0,1,2,3,4;c\in C;q\neq k}, \\ \sin\left(\frac{2\pi w}{52}\right), \cos\left(\frac{2\pi w}{52}\right), \sin\left(\frac{2\pi w}{4}\right), \cos\left(\frac{2\pi w}{4}\right) \\ \left\{Cal^{w-v}_{e}\right\}_{v=0,1;e\in E}, w \\ \left\{Cat_{kq}\right\}_{r=1,\ldots,R}, \left\{Cat_{id}\right\}_{d=1,\ldots,D} \\ \left\{\ln\left(d_{avg}{}^{w,p}_{i,k}\right)\right\}_{p=4,12}, \left\{\ln\left(d_{median}{}^{w,p}_{i,k}\right)\right\}_{p=4,12}, \left\{\ln\left(d_{std}{}^{w,p}_{i,k}\right)\right\}_{p=4,12} \\ \left\{\ln\left(d_{avg}{}^{w,p}_{i,q}\right)\right\}_{p=4,12;q\neq k}, \left\{\ln\left(d_{median}{}^{w,p}_{i,q}\right)\right\}_{p=4,12;q\neq k}, \left\{\ln\left(d_{std}{}^{w,p}_{i,q}\right)\right\}_{p=4,12;q\neq k} \end{array}\right) \tag{12}$$

Box I.

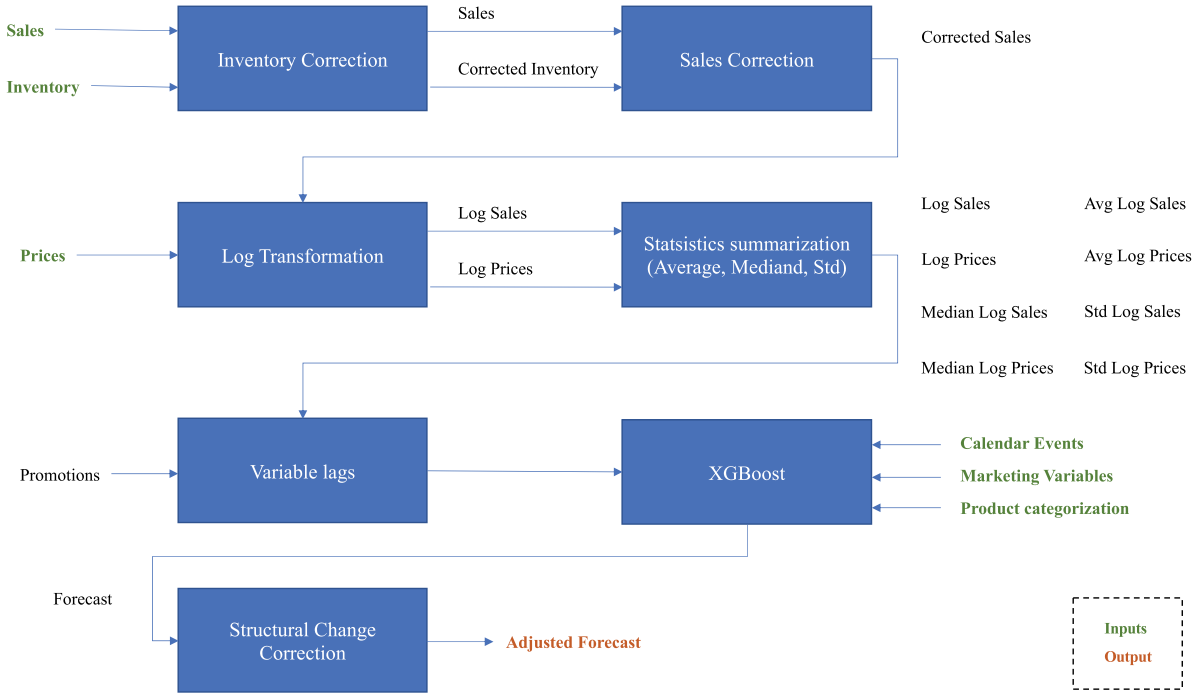

Fig. 3. Proposed method flowchart.

In addition to the structural change correction proposed by Huang et al. [11], we also apply a statistical method to identify the need for such a correction. We propose using the Chow test [48] that indicates if the regression coefficients of a time series with respect to the time variable are different for different splits of a time series dataset; the null hypothesis of the test is that there is no breaking point of structural change between the split time series. Test splitting is employed at each point of the time series and the null hypothesis is verified for each forecasting period; if the null hypothesis cannot be verified, we apply the intercept correction.

4.6. Method summary

Fig. 3 summarizes how the different corrections methods and the learning algorithm interact. The figure contains all the inputs and respective transformations. The final output is represented as the corrected forecast after the structural change correction.

5. Experiment design

In this section, we describe the experiment conducted to validate our proposed method and compare its results with benchmark methods. We briefly describe the dataset and the reasons why it was selected. We define our benchmark reference model as the Base-Lift method and detail the error measures used for comparison.

5.1. Dataset

To validate our proposed method, it is necessary to consider a dataset that contains all the characteristics that make disaggregated retail a forecasting challenge: (i) a large number of

stores and products to validate the automation of the method application, (ii) data that has been collected from a real retail operation subject to data inaccuracies to validate the proposed correction methods and (iii) available marketing information to validate marketing impacts and cross products effects.

In this research, we consider empirical data collected from a large grocery retailer in Ecuador called Corporación Favorita (Corporación [49]). The dataset contains information from 54 supermarkets and 4,036 different items from 2013 to 2017. Historical sales contain the daily number of units sold of each item in each store, and an integer value indicating if the item was being promoted in the store or not. The dataset also contains store information: a unique identifier number, the location, type and cluster allocation of each store. Data also contains detailed product information: family, category and class of each product, and an integer value indicating if the item is perishable or not. In addition to the historical sales, store and product data, the dataset also contains a definition of relevant holidays, which are important according to the model definition.

Table 3 shows descriptive statistics for each product category in the dataset. The product categories are the same as in the detailed product information that represents Corporación Favorita product structure as defined by the company. The table contains the weekly median units sold for each product category, the weekly standard deviation of units sold, the weekly average number of products in promotion and the average number of unique products sold. In general, it can be observed that the standard deviation of units sold is close to the median indicating highly volatile sales pattern. Also, it is possible to identify categories usually promoted, such as DELI, GROCERY I, MEATS and PRODUCE and categories that are not, such as BABY CARE, BEUTY, HARDWARE and PET SUPPLIES. It is also possible to differentiate categories that have a high number of distinct products being sold; for example, CLEANING and GROCERY I categories that have less unique products being sold, as for example, BABY CARE and BEAUTY. This dataset encompasses products with distinct behaviors in terms of sales; we thus believe that it is adequate to test and to validate the proposed methodology to solve the disaggregated forecasting problem.

We consider that this dataset exhibits all the characteristics required to validate our proposed method: it comprises a real dataset with many stores and products; in addition, the dataset comes from a real data collection process and is therefore subject to data inaccuracies, as well as having promotion and marketing information and a diverse number of product categories.

Fig. 4 shows the daily sales time series for the top 10 categories in terms of total sales. The data shows positive trend and a few structural changes, for example the beverage sales in the first quarter of 2015.

In our experiment, we removed from the dataset products with the lowest sales volume. We included in the dataset only the products that comprise 80% of the historical sales (Pareto Principle), leading to a total of 1623 products. By doing this, we discarded very specific products sold for a brief period and products with extremely low sale representativeness.

Also, to validate the proposed methodology for each product in each store, we use the proposed methodology to forecast the next 8 weeks. By doing this, we are able to evaluate the model accuracy as the forecasting horizon extends.

5.2. Benchmark model

The Base-Lift method was selected as the benchmark model, as it is the most commonly used method in practice [16]; it has been used in previous studies as the benchmark models [7,10,11,26].

In the Base-Lift method, the forecasted value for product k at store i in week w is given by expressions (14) and (15), in which M_{ik}^{w} represents the baseline forecast in week w by a simple exponential smoothing model, S_{ik}^{w} represents the actual sales, α is the smoothing parameter and the adjustment is calculated as the increased sales of product k during its most recent promotion compared to the corresponding baseline sales.

$$\hat{d}_{ik}^{w} = \begin{cases} M_{ik}^{w}, & \text{if product } k \text{ at store } i \textbf{ is not being} \text{ promoted} \\ M_{ik}^{w} + adjustment, & \text{if product } k \text{ at store } i \textbf{ is being} \text{ promoted} \end{cases} \tag{14}$$

$$M_{ik}^{w} = (1-\alpha)\, M_{ik}^{w-1} + \alpha S_{ik}^{w-1} \tag{15}$$

5.3. Error measures

We evaluate the model forecasting performance using four different error metrics. We consider the following traditional error measurements: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), the Symmetric Mean Absolute Percentage Error (sMAPE), which is a scale-free and symmetric error measure [50]. We also consider the Mean Absolute Scaled Error (MASE) proposed by Hyndman and Koehler [51] that represents the mean absolute error of the forecast values, divided by the mean absolute error of the in-sample one-step naive forecast. The MASE is recommended to evaluate improvements of one particular forecast method. Lastly, we consider the Coefficient of Determination (R-squared) which measures the proportion of the variation captured by the forecasting method; it is worth mentioning that unlike the other error measures, for the R-squared we do not calculate a deviation metric because it is, by definition, a quotient of variances.

Given as set of forecast estimates $\hat{d}_{ik}^{w}$ and their actual values d_{ik}^{w}, the error measures are defined by expressions (16), (17), (18) and (19).

$$MAE = \frac{1}{W}\sum_{i=0}^{W}\left|\hat{d}_{ik}^{w} - d_{ik}^{w}\right| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{W}\sum_{i=0}^{W}\left(\hat{d}_{ik}^{w} - d_{ik}^{w}\right)^{2}} \tag{17}$$

$$sMAPE = \frac{1}{W}\sum_{i=0}^{W}\frac{\left|\hat{d}_{ik}^{w} - d_{ik}^{w}\right|}{\left|\hat{d}_{ik}^{w}\right| + \left|d_{ik}^{w}\right|} \tag{18}$$

$$MASE = \frac{1}{H}\sum_{h=1}^{H}\left|\frac{\hat{d}_{ik}^{h} - d_{ik}^{h}}{\frac{1}{W-1}\sum_{q=2}^{W}\left|d_{ik}^{q} - d_{ik}^{q-1}\right|}\right| \tag{19}$$

$$R^{2} = 1 - \frac{\sum_{i=1}^{W}\left(d_{ik}^{w} - \hat{d}_{ik}^{w}\right)}{\sum_{i=1}^{W}\left(d_{ik}^{w} - \overline{d_{ik}^{w}}\right)} \tag{20}$$

6. Results and discussions

This section discusses the results of our proposed methodology and compares it with the benchmark model. Throughout this section, the proposed methodology and all its steps are collectively referred to as XGBoostModel. The benchmark model described above in Section 5.2 is referred to as Base-Lift model.

6.1. Forecasting results

Table 4 presents the summarized error metrics for all products in all stores. The results show that XGBoostModel significantly

Table 3
Descriptive statistics for each product category.

Product Category	Median Units Sold	Standard deviation of units sold	Average number of products in promotion	Average number of unique products sold
AUTOMOTIVE	30.00	31.66	0.12	8.16
BABY CARE	6.00	8.70	0.00	1.00
BEAUTY	15.00	15.64	0.01	3.20
BEVERAGES	7875.00	8618.26	3.53	147.79
BREAD/BAKERY	2809.13	1854.41	0.77	54.54
CELEBRATION	57.00	55.05	0.57	12.43
CLEANING	6380.50	3887.73	3.13	228.69
DAIRY	3833.50	4033.93	2.21	110.63
DELI	1471.96	1180.39	17.29	53.17
EGGS	1031.00	920.77	6.69	21.62
FROZEN FOODS	465.19	828.35	1.23	21.96
GROCERY I	21589.00	14232.23	10.58	554.97
GROCERY II	101.00	159.83	0.01	4.88
HARDWARE	6.00	5.38	0.01	2.36
HOME AND KITCHEN I	132.50	131.69	0.23	17.46
HOME AND KITCHEN II	55.00	56.62	0.09	9.75
HOME APPLIANCES	3.00	4.61	0.01	1.00
HOME CARE	1741.00	1365.78	3.00	54.27
LADIESWEAR	82.00	97.81	0.07	15.18
LAWN AND GARDEN	43.00	35.44	0.19	4.24
LINGERIE	44.00	47.86	0.03	9.04
LIQUOR/WINE/BEER	471.00	504.38	0.01	18.97
MAGAZINES	5.00	11.92	0.00	1.54
MEATS	1718.28	2227.83	10.35	32.43
PERSONAL CARE	1610.50	1189.25	0.36	60.39
PET SUPPLIES	17.00	44.92	0.00	4.83
PLAYERS AND ELECTRONICS	32.00	53.15	0.45	7.13
POULTRY	1659.06	2625.07	8.15	25.45
PREPARED FOODS	420.23	749.63	0.25	9.71
PRODUCE	78.00	7176.22	15.59	21.41
SCHOOL AND OFFICE SUPPLIES	5.00	12.73	0.00	3.00
SEAFOOD	123.11	236.96	1.04	4.40

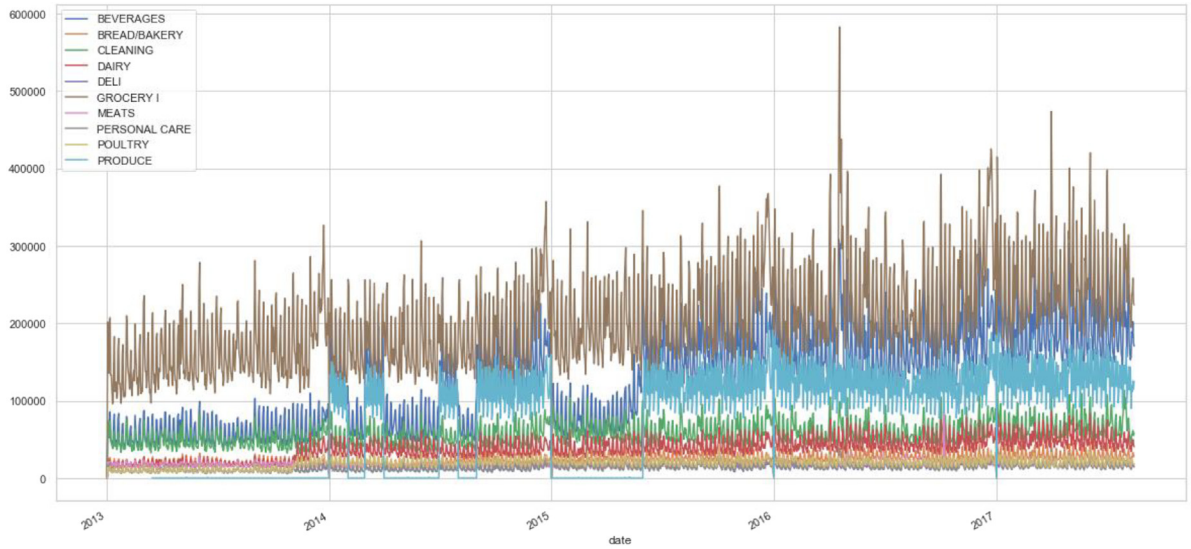

Fig. 4. Daily sales for the top 10 categories in the experiment dataset.

Table 4
Compared results.

	MAE	Std MAE	RMSE	Std RMSE	sMAPE (%)	Std sMAPE (%)	MASE	Std MASE	R^2
XGBoostModel	18.18	77.40	79.51	665.04	29.00	30.42	0.856	0.673	0.856
Base-Lift	24.81	90.59	93.93	774.40	35.84	36.80	1.168	0.883	0.425
Difference (%)	26.72%	14.56%	15.35%	14.12%	19.08%	17.34%	26.71%	23.78%	101.41%

outperforms the Base-Lift model for all error measures, reaching a relative improvement of 26.72% in the case of MAE. Also, the standard deviation of the error measures is lower for the XGBoostModel.

To statistically compare model accuracies, we conducted multiple Diebold–Mariano tests, which are model-free comparisons to evaluate time-series forecast accuracy [52]. The null hypothesis of the test is that the models are equally accurate. The test is

Table 5
Diebold–Mariano test results.

Horizon (week)	MAD		MAPE		RMSE	
	DM-Stat	*p*-value	DM-Stat	*p*-value	DM-Stat	*p*-value
1	−33.375	0.000	−27.345	0.000	−3.365	0.001
2	−36.198	0.000	−35.312	0.000	−4.018	0.000
3	−36.514	0.000	−40.180	0.000	−3.137	0.002
4	−34.260	0.000	−34.390	0.000	−2.594	0.009
5	−32.887	0.000	−38.164	0.000	−2.048	0.041
6	−29.578	0.000	−37.234	0.000	−1.849	0.064
7	−31.046	0.000	−39.548	0.000	−1.969	0.049
8	−41.782	0.000	−66.965	0.000	−2.442	0.015

* p_values equal to 0.000 indicate p_values lower than 0.001.

Table 6
Compared results per forecast horizon — Metrics MAE and RMSE.

Horizon	MAE			RMSE		
(week)	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)
1	**14.39**	20.15	28.60%	**52.59**	69.92	24.78%
2	**15.69**	21.43	26.78%	**81.48**	89.56	9.02%
3	**16.70**	23.30	28.32%	**82.44**	102.93	19.91%
4	**17.78**	23.50	24.35%	**80.09**	100.63	20.42%
5	**19.47**	24.56	20.74%	**104.61**	111.93	6.54%
6	**20.57**	24.60	16.35%	**95.40**	107.37	11.15%
7	**17.45**	22.83	23.58%	**70.65**	77.96	9.38%
8	**23.39**	38.07	38.56%	**54.06**	82.54	34.50%

Table 7
Compared results per forecast horizon — Metrics MASE and R^2.

Horizon	MASE			R^2		
(week)	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)
1	**24.66%**	29.84%	17.35%	0.893455	0.675650	24,38%
2	**25.07%**	30.44%	17.63%	0.898645	0.566461	36,96%
3	**26.09%**	33.11%	21.20%	0.893264	0.380812	57,37%
4	**26.92%**	33.18%	18.86%	0.892901	0.378707	57,59%
5	**27.83%**	32.88%	15.34%	0.888109	0.401184	54,83%
6	**29.08%**	30.19%	3.68%	0.878279	0.405180	53,87%
7	**27.76%**	33.18%	16.32%	0.888485	0.526270	40,77%
8	**44.58%**	63.93%	30.27%	0.845362	0.554235	34,44%

designed to compare error metrics; therefore, it is not applicable to the R-squared metric. Table 5 presents the test statistics and the p-values for different error measures and for different forecasting horizons. The low p-values indicate that we can reject the null hypothesis for all error measures and all forecasting horizons indicating that the proposed method is more accurate than the benchmark model.

In Tables 6 and 7, we investigate the effect of the forecasting horizon in the accuracy metrics. We present the accuracy metrics for the proposed methodology and the benchmark model for different forecasting horizons, except the MASE metric, which, by definition, depends upon the previous value to be calculated. Except for the RMSE measure, the proposed methodology yielded better results than the benchmark model for all the forecasting horizons. It is also evident that as the forecasting horizon increases, the forecasting accuracy decreases for all the metrics considered; however, for the proposed model, the decrease in forecasting accuracy as the horizon grows is lower than the Base-Lift model, meaning that the advantage of the proposed methodology increases as the forecasting horizon is extended.

Table 8 shows the standard deviation of the corresponding measures for different forecasting horizons. Results show that except for one instance (the standard deviation of the RMSE for the week 2) the standard deviation of the error measures for the XGBoostModel is lower than the benchmark model.

Table 8 shows the standard deviation of the corresponding measures for different forecasting horizons. Results show that except for one instance (the standard deviation of the RMSE for the week 2) the standard deviation of the error measures for the XGBoostModel is lower than the benchmark model. We also analyze the effects of promotions on a specific product. Tables 9 and 10 contains the accuracy metrics for periods with and without promotions. Table 11 contains the standard deviation of the corresponding metrics. The row with promotion equal to one indicates periods with promotional effects on a specific product and the row with promotion equal to zero denotes periods without it. Interestingly, the largest improvements occur in periods without promotions. One way to interpret this result is that the drivers that affect demand go beyond just considering promotions in one focal product but rather considering all the promotions of all products, calendar effects and many other variables as the methodology proposes. Additionally, standard deviation of metrics is also lower for the XGBoostModel.

Fig. 5 shows the absolute error distribution (y-axis) for the top 95% product families in terms of sales. The graph shows boxplot comparisons of the absolute error distribution for the XGBoostModel and the Base-Lift model. Results indicate that the proposed model not only has a lower error median but also a lower dispersion of errors, which is also a desirable quality of a forecasting model.

Table 8
Compared results per forecast horizon — Standard Deviation of Metrics.

Horizon	MAE			RMSE			sMAPE		
(week)	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)
1	**50,59**	66,96	24,45%	**491,21**	554,01	11,33%	**25,67%**	31,66%	18,90%
2	**79,96**	86,96	8,05%	897,26	**877,54**	−2,25%	**25,93%**	32,08%	19,18%
3	**80,73**	100,26	19,48%	**704,29**	871,91	19,22%	**27,16%**	34,12%	20,39%
4	**78,09**	97,85	20,19%	**548,97**	830,28	33,88%	**27,63%**	34,29%	19,43%
5	**102,78**	109,20	5,88%	**711,90**	861,16	17,33%	**27,85%**	34,19%	18,57%
6	**93,15**	104,51	10,87%	**670,83**	842,08	20,34%	**28,07%**	32,00%	12,28%
7	**68,46**	74,54	8,16%	**497,76**	525,44	5,27%	**28,62%**	34,65%	17,40%
8	**48,74**	73,24	33,45%	**174,48**	295,84	41,02%	**43,56%**	46,40%	6,11%

Table 9
Compared results for promotional periods — Metrics MAE and R^2.

Promotion	MAE			RMSE		
	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)
0	13.8	21.89	58.62%	46.08	65.31	41.73%
1	35.2	36.15	2.70%	150.49	162.92	8.26%

Table 10
Compared results for promotional periods — Metrics MASE and R^2.

Promotion	MASE			R^2		
	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)
0	27.48%	36.00%	29.29%	0.855	0.404	111.63%
1	34.91%	37.06%	6.16%	0.617	0.396	55.80%

Table 11
Compared results for promotional periods — Standard Deviation of Metrics.

Promotion	MAE			RMSE			sMAPE		
	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)	XGBoost Model	Base-Lift	Diff (%)
0	43.96	61.53	28.56%	381.34	461.22	17.32%	31.14%	38.22%	18.52%
1	146.32	158.86	7.89%	966.82	1121.47	13.79%	26.64%	30.60%	12.94%

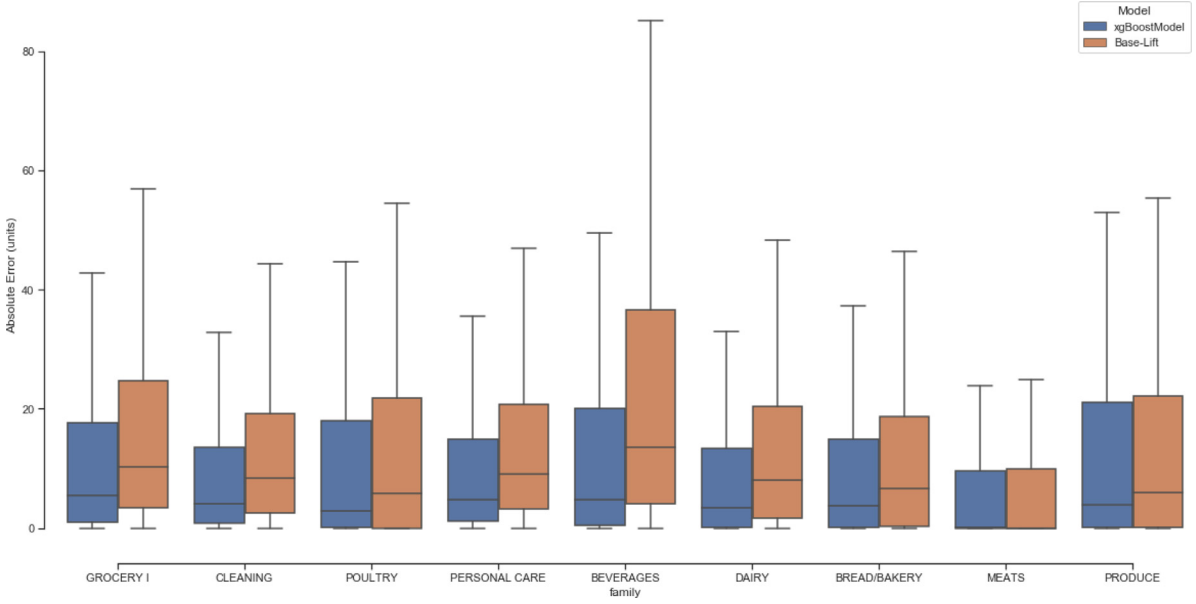

Fig. 5. Absolute error distribution for the proposed and benchmark models.

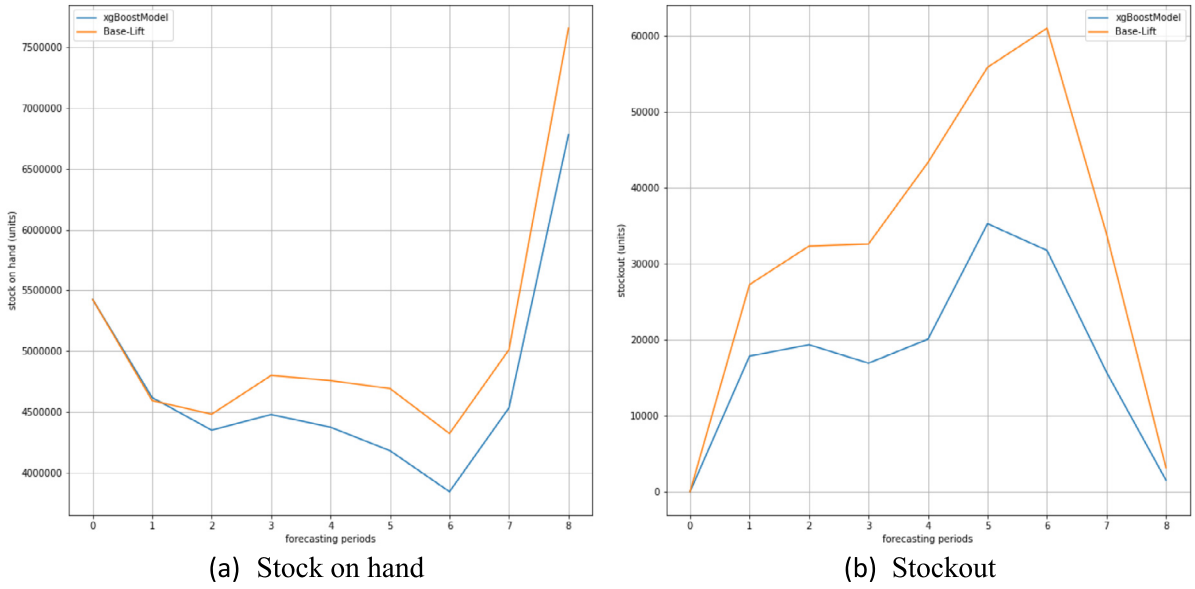

Fig. 6. Simulated stockouts and stock on hand curves.

Table 12
Simulated stockouts and stock on hand values.

Horizon	Stockout (units)			Stock on hand (units)		
	xgBoostModel	Base-Lift	Diff (%)	xgBoostModel	Base-Lift	Diff (%)
1	17827	27234	−34.54%	4617924	4592362	0.56%
2	19334	32302	−40.15%	4349266	4480514	−2.93%
3	16901	32577	−48.12%	4477817	4799438	−6.70%
4	20063	43302	−53.67%	4373374	4757017	−8.06%
5	35245	55808	−36.85%	4179985	4691595	−10.90%
6	31731	60949	−47.94%	3842600	4321459	−11.08%
7	15731	33907	−53.61%	4533251	5011804	−9.55%
8	1518	3149	−51.81%	6780623	7656952	−11.44%
		Average	**−45,83%**		Average	**−7,51%**

6.2. Inventory results

In order to quantify the advantage in terms of inventory metrics we conducted a simulation of the replenishment process considering both the proposed XGBoostModel and the benchmark Base-Lift model. The simulation considers an order-up-to inventory policy with a weekly replenishment cycle in which the order-up-to level is calculated by the forecasted demand during the lead time plus the coverage provided by the safety stock. The safety stock is determined by the estimated root mean squared forecasting error multiplied by the cumulative inverse distribution function of a standard normal distribution at a specific service level. We consider a 0.975 service level as in Taleizadeh et al. [53], which is a commonly adopted value in retail supply chains.

The aim is to investigate the effects of the proposed methodology in stockout and stock on hand across the forecasting periods. Fig. 6 depicts the sum of stockouts and stock on hand for the proposed methodology and the benchmark model. Results indicate that XGBoostModel yields to significantly lower stockouts (right diagram) and lower stock on hand (left diagram), thus providing a more efficient inventory level. Detailed stockout and stock on hand levels for the simulation are given in Table 12. The average stockout and stock on hand reductions are 45.83% and 7.51%, respectively.

Fig. 7 depicts the inventory efficiency curves for both models. The horizontal axis represents the stockout levels and the vertical axis represents the stock on hand levels for different forecasting periods. Each point represents a forecasting period and the correspondent stockout and stock on hand for both models. The graph also shows a fitted second-degree polynomial that illustrates the inventory efficiency boundary for both approaches. The results show that the proposed methodology together with a simple inventory policy produces an improved efficiency curve with lower stockouts and stock on hand.

7. Concluding remarks

As retailers supply a wide range of SKUs, disaggregated forecast is key to operations profitability once it supports accurate inventory decisions on point of sales in order to avoid either lost sales due to out-of-stock or excessive inventory costs due to overstocks. Recent works on SKU level forecasting found in the literature consider marketing and promotion effects in a simplified manner, as they are assumed to be constant over time, and structural changes are either disregarded and out-of-sample forecasts are simply corrected with the historical forecast bias, which is not adequate, particularly when consumption drastically changes, as is the case of the recent coronavirus outbreak.

In this paper, a methodology for the disaggregated (SKU-store level) forecasting problem, characterized by a large number of stores and products, as well as different marketing and promotion effects and cross-category effects, was proposed. It comprises information correction models that are able to correct sales observations during out-of-stock days in addition to inventory information that is different from physical inventory.

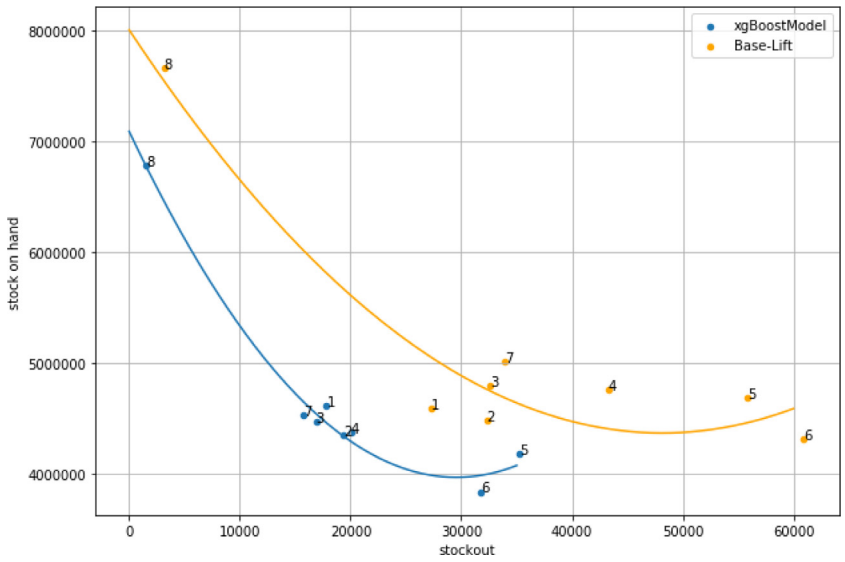

Fig. 7. Inventory efficiency curve.

It also embeds XGBoost, a non-linear non-parametric ensemble-based model as the central learning algorithm, which is based on an implementation of GBM framework, and a structural change correction method to account for sudden changes in consumer behavior caused by external factors (i.e., natural disasters, a new competitor entering the market or technology disruptions). The advantage of such approach comes from the fact that it does not assume a functional form for the forecasting model but rather learns it from the available data, and is capable of dealing with a large number of variables and a sparse input matrix without incurring in overfitting issues.

Our model can effectively improve accuracies in diverse error metrics and is consistent across different forecasting horizons and product families as evidenced by the results using a real-world public dataset. A 26.72% MAE improvement was also achieved in comparison with Base-Lift model, widely used as benchmark. The results also show that the improvement increases as the forecasting horizon expands, which also motivates the use of our proposed methodology for medium-term forecasting by retailers. We also compared forecasting error distributions of top-sales product categories, and the results indicate that our proposed methodology not only yields better average accuracy but also produces an error distribution with less dispersion when compared to the benchmark model.

Based on a simulation considering a simple order-up-to inventory policy in conjunction with forecasts generated by our model show that it produces a more efficient inventory curve with significantly lower stockouts (45.83% reduction) and lower stock on hand (7.51% reduction) than the Base-Lift model. This evidences that employed with simple inventory policies may ensure more efficient retail supply chain operations.

We believe that the approach that we propose can help retailers to improve their disaggregated demand forecasting processes. It can be easily embedded in computational packages, and it allows a high level of automation, which is required for modern retailers. Practitioners may use the methodology to produce accurate disaggregated forecasts that support optimized inventory policies, reducing unnecessary working capital, stockouts and lost sales, and ultimately leading to more profitable retail operations with improved customer service levels as the results of our simulation have evidenced.

Our primary research focus was on the accuracy capability of the forecasting models. However, in some scenarios, forecasts must be coupled with the rationale behind it (e.g., medium-term forecasts to subsidize large purchases, campaign forecasts for deciding on a marketing strategy). The analysis and explanation of specific forecasts is an active research field, and the investigation of the relevant variables and forecasts explanations of our proposed model can be seen as a next step in disaggregated forecast research.

Other non-linear and non-parametric models, such as Neural Networks, also exist and the investigation of their effectiveness instead of XGBoost as the main learning model may also be explored as a next step of this research. In addition, marketing strategies and promotions may have complex and diverse structures and the exploration of different, more detailed representations of marketing and promotional variables in place of a straightforward representation using dummy variables may also lead to improvements in accuracy of the proposed approach.

CRediT authorship contribution statement

Luiz Augusto C.G. Andrade: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Claudio B. Cunha:** Conceptualization, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is publicly available - https://www.kaggle.com/c/favorita-grocery-sales-forecasting.

References

[1] World Economic Forum, How turning retail stores into e-commerce centres can avoid massive emissions, 2022, https://www.weforum.org/agenda/2022/09/retail-stores-into-ecommerce-centres-avoid-carbon-emissions/. (Accessed 23 2022).
[2] Deloitte Touche Tohmatsu Limited, Global powers of retailing, 2021.

[3] D. Corsten, T. Gruen, Desperately seeking shelf availability: An examination of the extent, the causes, and the efforts to address retail out of stocks, Int. J. Retail Distrib. Manage 31 (2003) 605–617.

[4] OrderDynamics, Retailers and the ghost economy: The haunting of returns, 2015, http://engage.dynamicaction.com/WS-2015-06-IHL-Ghost-Economy-Haunting-of-Returns-AR_LP.html.

[5] L. Tiacci, S. Saetta, An approach to evaluate the impact of interaction between demand forecasting method and stock control policy on the inventory system performances, Int. J. Prod. Econ. 118 (2009) 63–71.

[6] R. Fildes, K. Nikolopoulos, S.F. Crone, A.A. Syntetos, Forecasting and operational research: a review, J. Oper. Res. Soc. 59 (9) (2008) 1150–1172, http://dx.doi.org/10.1057/palgrave.jors.2602597.

[7] Ö. Gür Ali, S. Sayin, T. van Woensel, J. Fransoo, SKU demand forecasting in the presence of promotions, Expert Syst. Appl. 36 (2009) 12340–12348.

[8] G. Box, G. Jenkins, Time Series Analysis Forecasting and Control, Holden-Day, San Francisco, 1970.

[9] J. Durbin, S.J. Koopman, Time Series Analysis By State Space Methods, second ed., in: Oxford Statistical Science Series, OUP Oxford, 2012.

[10] S. Ma, R. Fildes, T. Huang, Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information, European J. Oper. Res. 249 (2016) 245–257.

[11] T. Huang, R. Fildes, D. Soopramanien, Forecasting retailer product sales in the presence of structural change, European J. Oper. Res. 279 (2019) 459–470.

[12] J. Sheth, Impact of Covid-19 on consumer behavior: Will the old habits return or die? J. Bus. Res. 117 (2020) (2020) 280–283.

[13] T. Boone, R. Ganeshan, A. Jain, N.R. Sanders, Forecasting sales in the supply chain: Consumer analytics in the big data era, Int. J. Forecast. 35 (2019) 170–180.

[14] J.G. De Gooijer, R.J. Hyndman, 25 years of time series forecasting, Int. J. Forecast. 22 (2006) 443–473.

[15] R. Fildes, S. Ma, S. Kolassa, Retail forecasting: Research and practice, Int. J. Forecast. 38 (2009) 1283–1318.

[16] R. Fildes, S. Ma, S. Kolassa, Retail Forecasting: Research and Practice, Lancaster University Working paper, Lancaster University Management School, 2018.

[17] E. Alpaydin, Introduction to Machine Learning, The MIT press, 2010.

[18] A. Alon, M. Qi, R.J. Sadoswski, Forecasting aggregate reail sales: A comparison of artificial neural networks and traditional methods, J. Retail. Consum. Serv. 8 (2001) 147–156.

[19] G.P. Zhang, M. Qi, Neural network forecasting for seasonal and trend time series, European J. Oper. Res. 160 (2005) 501–514.

[20] S. Lam, M. Vandenbosch, M. Pearce, Retail sales force scheduling based on store traffic forecasting, J. Retail. 74 (1998) 61–88.

[21] M.D. Geurts, J.P. Kelly, Forecasting retail sales using alternative models, Int. J. Forecast. 2 (1986) 261–272.

[22] A.A. Syntetos, Z. Babai, J.E. Boylan, S. Kolassa, K. Nikolopoulos, Supply chain forecasting: Theory, practice, their gap and the future, European J. Oper. Res. 252 (2016) 1–26.

[23] A.A. Syntetos, Z. Babai, J.E. Boylan, S. Kolassa, Supply chain forecasting: Theory, practice, their gap and the future, European J. Oper. Res. 252 (2016) 1–26.

[24] N. DeHoratius, A. Raman, Inventory record inaccuracy: An empirical analysis, Manage. Sci. 54 (2008) 627–641.

[25] T. Huang, R. Fildes, D. Soopramanien, The value of competitive information in forecasting FMCG retail product sales and the variable selection problem, European J. Oper. Res. 237 (2014) 738–748.

[26] L.G. Cooper, P. Baron, W. Levy, M. Swisher, P. Gogos, Promocast (tm): A new forecasting method for promotion planning, Mark. Sci. 18 (1999) 301–316.

[27] K.L. Ailawadi, B.A. Harlam, J. César, D. Trounce, Promotion profitability for a retailer: The role of promotion, brand, category, and store characteristics, J. Mar. Res. 43 (2006) 518–535.

[28] S.R. Srinivasan, S. Ramakrishnan, S.E. Grasman, Incorporating cannibalization models into demand forecasting, Mark. Intell. Plan. 23 (2005) 470–485.

[29] S. Lang, W.J. Steiner, A. Weber, P. Wechselberger, Accommodating heterogeneity and nonlinearity in price effects for predicting brand sales and profits, European J. Oper. Res. 246 (2015) 232–241.

[30] R. Fildes, P. Goodwin, D. Önkal, Use and misuse of informa- tion in supply chain forecasting of promotion effects, Int. J. Forecast. 35 (2019) 144–156.

[31] S. Divakar, B.T. Ratchford, V. Shankar, CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer-packaged goods, Mark. Sci. 24 (2005) 334–350.

[32] T.L. Ainscough, J.E. Aronson, An empirical investigation and comparison of neural networks and regression for scanner data analysis, J. Retail. Consum. Serv. 6 (1999) 205–217.

[33] G.D. Pillo, V. Latorre, S. Lucidi, E. Procacci, An application of support vector machines to sales forecasting under promotions, 4OR 14 (2016) 309–325.

[34] A. Goli, H. Khademi Zareh, R. Tavakkoli-Moghaddam, A. Sadeghieh, A comprehensive model of demand prediction based on hybrid artificial intelligence and metaheuristic algorithms: A case study in dairy industry, J. Ind. Syst. Eng. 11 (4) (2018).

[35] A. Goli, H. Khademi-Zare, R. Tavakkoli-Moghaddam, A. Sadeghieh, A. Sasanian, R.M. Kordestanizadeh, An integrated approach based on artificial intelligence and novel meta-heuristic algorithms to predict demand for dairy products: a case study, Network: Comput. Neural Syst. (2021) http://dx.doi.org/10.1080/0954898X.2020.184984.

[36] V. Flunkert, D. Salinas, J. Gasthaus, DeepAR: Probabilistic forecasting with autoregressive recurrent networks, 2017, arXiv preprint.

[37] R. Wen, K. Torkkola, B. Narayanaswamy, D. Madeka, A multi-horizon quantile recurrent forecaster, 2017, arXiv:1711.11053.

[38] A. Harvey, Seasonality and unobserved components models: An overview, in: Proceedings of the Eurostat Conference on Seasonality, Seasonal Adjustment and their Implications for Short-Term Analysis and Forecasting, 2006.

[39] C. Tianqi, C. Guestrin, XGBoost: A scalable tree boosting system, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[40] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobotics (2013).

[41] V.N. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1995.

[42] J. Friedman, Greedy boosting approximation: a gradient boosting machine, Ann. Statist. 29 (2001) 1189–1232, http://dx.doi.org/10.1214/aos/1013203451.

[43] R. Fletcher, Practical Methods of Optimization, John Wiley & Sons, 2013.

[44] T. Hastie, R. Tibshirani, J.H. Friedman, Elements of Statistical Learning, second ed., Springer 2008, 2008.

[45] S. Karabati, B. Tan, O.-C. Öztürk, A method for estimating stock-out-based substitution rates by using point-of-sale data, IIE Trans. 41 (5) (2009) 408–420, http://dx.doi.org/10.1080/07408170802512578.

[46] A. Andiojaya, H. Demirhan, A bagging algorithm for the imputation of missing values in time series, Expert Syst. Appl. 129 (2019) 10–26.

[47] C. Bentéjac, A. Csörgő, G. Gonzalo Martínez-Muñoz, A comparative analysis of XGBoost, 2019, arXiv:1911.01914.

[48] G.C. Chow, Tests of equality between sets of coefficients in two linear regressions, Econometrica 28 (1960) 591–605.

[49] Corporación Favorita, Corporación Favorita grocery sales forecasting, Version 1, 2018. https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data. (Retrieved 4 May 2020).

[50] S. Kolassa, W. Schütz, Advantages of the MAD/Mean ra- tio over the MAPE, Foresight: Int. J. Appl. Forecast. 6 (2007) 40–43.

[51] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688.

[52] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econom. Statist. 13 (1995) 253–263.

[53] A.A. Taleizadeh, I. Shokr, I. Konstantaras, M. VafaeiNejad, Stock replenishment policies for a vendor-managed inventory in a retailing system, J. Retail. Consum. Serv. 55 (2020).