International Journal of Forecasting 38 (2022) 209–223

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Classification-based model selection in retail demand forecasting

Matthias Ulrich [a,*], Hermann Jahnke [a], Roland Langrock [a], Robert Pesch [b], Robin Senge [b]

[a] Department of Business Administration and Economics, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany
[b] inovex GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, Germany

ARTICLE INFO

Keywords:
Forecasting
Inventory
e-commerce
Retailing
Model selection

ABSTRACT

Retailers supply a wide range of stock keeping units (SKUs), which may differ for example in terms of demand quantity, demand frequency, demand regularity, and demand variation. Given this diversity in demand patterns, it is unlikely that any single model for demand forecasting can yield the highest forecasting accuracy across all SKUs. To save costs through improved forecasting, there is thus a need to match any given demand pattern to its most appropriate prediction model. To this end, we propose an automated model selection framework for retail demand forecasting. Specifically, we consider model selection as a classification problem, where classes correspond to the different models available for forecasting. We first build labeled training data based on the models' performances in previous demand periods with similar demand characteristics. For future data, we then automatically select the most promising model via classification based on the labeled training data. The performance is measured by economic profitability, taking into account asymmetric shortage and inventory costs. In an exploratory case study using data from an e-grocery retailer, we compare our approach to established benchmarks. We find promising results, but also that no single approach clearly outperforms its competitors, underlying the need for case-specific solutions.

1. Introduction

In retailing, any mismatch between inventory and customer demand can cause monetary consequences for the retailer. Stock-out situations result in immediate shortage costs due to missed sales, and long-run costs due to customer churn. Excess inventory generates costs as a result of operational inefficiencies and, in some cases, spoilage (Anderson et al., 2006). In retail practice, shortage costs are usually specified to be higher than inventory costs. This asymmetry in costs corresponds to a cost-optimal service level (in-stock probability) in excess of 50%, for each stock keeping unit (SKU) (Teunter et al., 2010; Rădăşanu, 2016). When forecasting demand in order to optimize the inventory level, retailers thus need to identify a suitable approach specifically with respect to quantifying the right tail of the distribution of the demand, rather than the mean (Nahmias, 1994; Agrawal & Smith, 1996; Fildes et al., 2019; Ulrich et al., 2021).

A key challenge in statistical forecasting of retail demand results from the heterogeneity of demand patterns, which across SKUs may differ in terms of demand frequency, demand quantity, demand variation, and demand regularity. Time series often exhibit strong trends, seasonal variations, or irregular demand peaks, rendering demand forecasting very challenging (Fildes et al., 2019).

* Corresponding author.
E-mail addresses: matthias.ulrich@uni-bielefeld.de (M. Ulrich), hjahnke@uni-bielefeld.de (H. Jahnke), roland.langrock@uni-bielefeld.de (R. Langrock), robert.pesch@inovex.de (R. Pesch), robin.senge@inovex.de (R. Senge).

https://doi.org/10.1016/j.ijforecast.2021.05.010
0169-2070/

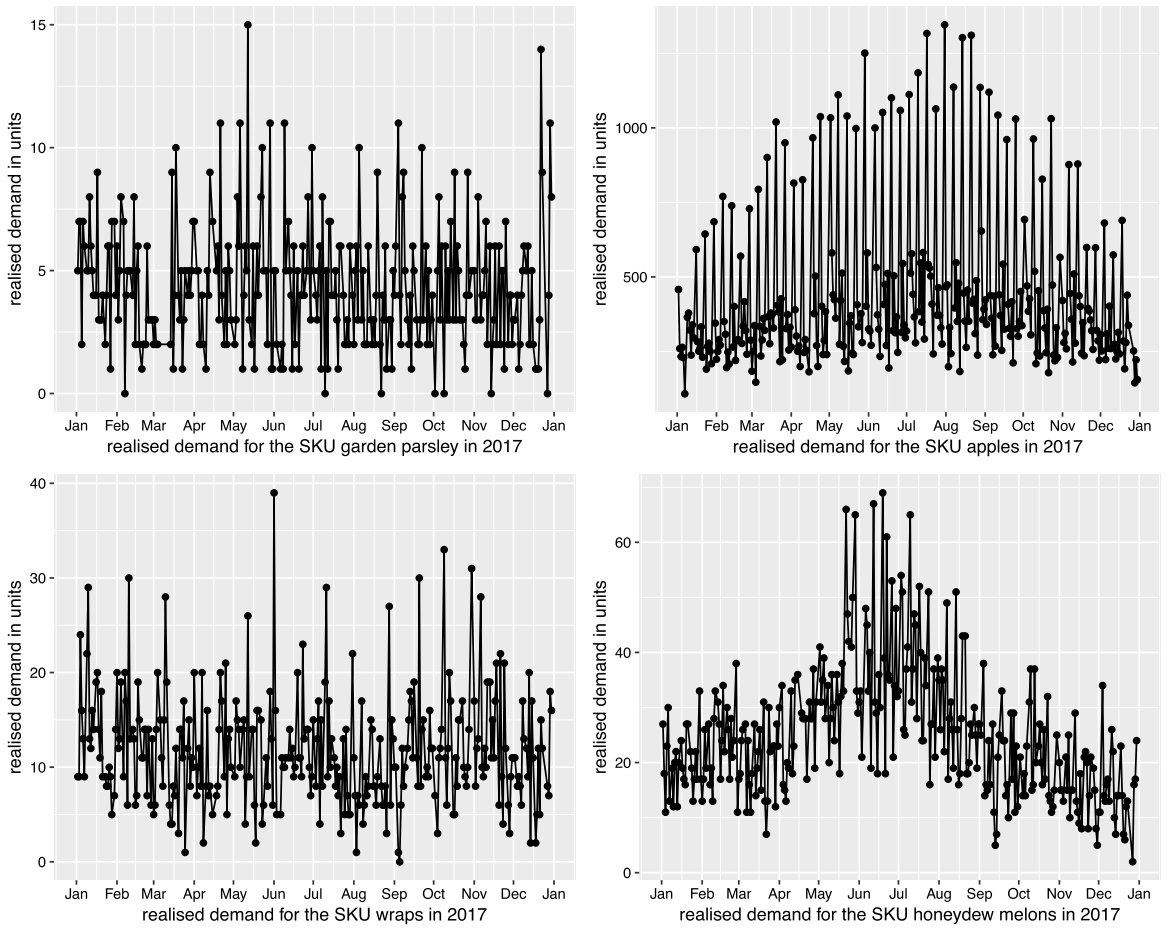

Fig. 1. Demand patterns for the SKUs of garden parsley, apples, wraps, and honeydew melons in 2017 for one of five fulfillment centers considered.

As an illustrative example, Fig. 1 displays the demand patterns of four SKUs considered in this study, namely garden parsley, apples, wraps, and honeydew melons. Each observation here corresponds to one demand period, i.e. one day of delivery. First, it can be seen that apples and honeydew melons show a higher *demand frequency* than wraps and garden parsley. For the latter, in fact, there was no demand in 2% of the demand periods. Second, the average *demand quantity* varies from ∼5 units per demand period for garden parsley to ∼700 units for apples. Third, garden parsley and wraps show *demand variations* below 100%, whereas the same figure exceeds 100% for apples and honeydew melons. Finally, while the demand variation for apples is largely due to regular weekly peaks on Mondays, there is no such *demand regularity* for garden parsley, wraps, and honeydew melons. Given this immense diversity in demand patterns, it comes as no surprise that there is consensus in the literature that no single model can universally outperform all other natural candidate models for demand forecasting (Fildes, 1989; Ulrich et al., 2021). Both the class of statistical models (e.g. linear vs. quantile regression) and also any associated distributional assumption (e.g. normal vs. gamma) may have to be adapted to suit any particular demand pattern at hand (Ulrich et al., 2021). For retailers, there is thus a need to identify, for all relevant SKUs, which type of forecasting model is most promising with regard to cost minimization.

A natural strategy for identifying a suitable forecasting model is to assess any candidate model's historical performance, i.e. considering its accuracy in previous demand periods. To this end, based on an extensive literature review, Fildes (1989) distinguishes *individual selection*, *aggregate selection*, and *combination forecasts* as general strategies for assessing past performance and subsequently producing predictions. Individual selection allocates a specific model to each SKU, while aggregate selection uses the model that performs best on average, across all SKUs. The former strategy is more flexible but less robust than the latter; in other words, aggregate selection is more stable (low variance) in terms of model selection, but involves a higher risk of misspecification (high bias). Combination forecasts build the prediction as a weighted average of each candidate model's forecast, usually trying to optimize the models' weights, although unweighted combinations are surprisingly competitive (referred to as the "forecast combination puzzle"; Claeskens et al., 2016). Rather than matching an individual model to each single SKU, as in individual selection, SKUs can also be clustered *a priori* based on similarities in demand patterns, such that subsequently,

forecasting models are matched on the cluster level (see, e.g., Williams, 1984; Eaves & Kingsman, 2004).

In this contribution, we suggest the use of a classification algorithm for automatically allocating any given demand pattern to a suitable model. In contrast to most existing model selection approaches in the literature on demand forecasting—and in particular Armstrong (2001), who uses manually generated decision trees—our approach to model selection is data-driven. That is, the classifier learns from historical data how to match demand patterns to models. Our classification-based model selection (CMS) proceeds by first constructing a labeled training data set, comprising past demand periods with information on the candidate models' performances with regard to predicting demand in those periods. Subsequently, a classifier is trained to match demand patterns to candidate models. For future data, given the information available (e.g. the day of the week, known demand by the time of the replenishment decision, etc.), the most promising model is then selected by applying the classifier. Thus, within our approach the classes to be identified correspond to the different candidate models for forecasting demand (e.g. linear regression, quantile regression, etc.). The CMS framework allows us to address the ever more complex demand patterns encountered in retail practice. Automated model selection allows for daily SKU-specific demand forecasting decisions for 10,000–20,000 SKUs in each fulfillment center.

We evaluate CMS in an exploratory case study using data from a European e-grocery retailer. The data cover three years and 111 SKUs from five different fulfillment centers. As benchmarks for the new approach proposed here, we consider individual selection, aggregate selection, combination forecasts (with equal weights and model pre-selection, as proposed by Kourentzes et al., 2018), and all individual models (without selection). Considering the cost values for shortage and excess inventory as specified by the retailer, we compare the out-of-sample costs for one year of test data. The case study provides a proof of concept of the proposed CMS approach, and more generally a comparison of several conceptually different approaches to model selection in an e-grocery setting.

2. Problem statement and related literature

2.1. Business problem

Given the data set available to us, our analysis focuses on perishable SKUs in the categories of fruits, vegetables, and bake-off products. For these SKUs, we assume that internal quality requirements restrict the sales period to roughly one demand period. Excess inventory cannot be sold in the following demand period and generates spoilage. Thus, the demand forecast defines the replenishment order quantity and the inventory level at the beginning of the corresponding demand period. As shortage costs typically exceed inventory costs in retail practice, the application of classical statistical performance measures such as the mean average error or the mean average percentage error may be inadequate for model choice (see, e.g., Makridakis & Winkler, 1983; and Carbonneau et al., 2008).

Table 1
Notation used to quantify the costs.

Parameter	Definition
b	Costs for each unit of shortage
h	Costs for each unit of excess inventory
t	Demand period
q	Inventory level at the beginning of the demand period
D	Stochastic customer demand
C	Total costs

Table 1 provides an overview of the notation we use to describe the business problem. To capture the asymmetric economic impact of shortage and excess inventory for each demand period t, we introduce the total cost C_t resulting from any potential mismatch between the inventory level and realized demand. Assuming single and independent demand periods, we apply the newsvendor problem to determine the optimal inventory level under uncertain demand (Zipkin, 2000). Hence, for each SKU and each demand period t, the retailer aims at minimizing the expected total cost C_t,

$$E[C_t(q_t)] = hE(q_t - D_t)^+ + bE(D_t - q_t)^+,$$

as realized for the stochastic demand D_t with respect to the inventory level at the beginning of the demand period, q_t. Excess inventory costs are denoted by h, while shortage costs are given by b.

The optimal inventory level q_t^* is obtained as

$$q_t^* = \underset{q_t}{\operatorname{argmin}}\, E[C_t(q_t)] = F_t^{-1}\left(\frac{b}{b+h}\right), \tag{1}$$

where F_t is the true cumulative distribution function of the demand distribution in period t, and $b/(b+h)$ is the optimal demand quantile given b and h. The ratio $b/(b+h)$ corresponds to the service level selected by the retailer. In practice, the true cumulative distribution function F_t describing the stochastic demand is unknown. That is, the optimal solution in Eq. (1) is not available. For estimating q_t^*, we statistically model the stochastic demand D_t as a function of features. Applying the resulting estimated distribution function $\hat{F}_t$, we are subsequently able to derive an estimated optimal inventory level $\hat{q}_t^*$ via Eq. (1).

For the evaluation of these forecasts' quality, we aim to specify cost parameters b for shortage costs and h for costs of excess inventory in a way that represents the cost relationship found in retail practice as closely as possible. Retailers are often able to calculate inventory costs from purchasing and operational handling costs without much difficulty. Contrary to that, due to the economic effects of losing sales in the medium or long term, it is much harder to come up with numerical values for shortage costs, especially in innovative and markedly growing business environments as, e.g., in e-grocery. Therefore, as an alternative to such a specification of shortage costs, retailers often define a strategic service-level target. For example, 97% would currently be a realistic service level for an e-grocery retailer (cf. Ulrich et al., 2021). Despite a higher risk of spoilage, these high service-level targets are also often valid for fruits, vegetables, and bake-off products, due to limited options for substitution, and hence the risk

Table 2
Shortage costs for constant holding costs and an increase in the service level selected by applying the ratio $\alpha = b/(b+h)$.

α	h	b
50%	1	1
75%	1	3
90%	1	9
95%	1	19
97%	1	32

of customer churn in case of a stock-out (cf. Ulrich et al., 2021).

To derive b, we then make use of the relation $\alpha = b/(b+h)$ from the newsvendor problem, where α equals the service-level target selected. Given the pre-specified service-level target and an estimation of h from purchasing and operational handling costs, we obtain the associated b as

$$b = \frac{h\alpha}{1-\alpha}. \tag{2}$$

According to this ratio, shortage costs increase faster than by a linear rate with an increase in the service level (see Table 2). Thus, the selected service level substantially affects the cost evaluation for a given forecasting error distribution.

For the purpose of evaluation, we calculate the total costs that occur under the estimated value $\hat{q}_t^*$ obtained from $\hat{F}_t$, for each SKU and each demand period t by

$$C_t(\hat{q}_t^*) = h(\hat{q}^* - d_t)^+ + b(d_t - \hat{q}^*)^+. \tag{3}$$

For each SKU, we sum up the costs for all demand periods t to obtain the total realized costs in the evaluation period.

2.2. Retail demand forecasting in the literature

Demand forecasts provide essential information for decision making in retailing on strategic, tactical, and operational levels. The strategic level includes long-term decisions such as the distribution network, whereas the tactical level aims at maximizing category profits within the distributional channel, e.g. decisions on pricing and the listing and delisting of SKUs. At the operational level we examine here, demand forecasts are crucial for inventory management decisions and impact the corresponding costs, as discussed in Section 2.1. Demand forecasting at the SKU level for individual stores is the lowest hierarchical level in retail forecasting. This level of detail renders demand forecasting very challenging, as various SKU-specific characteristics may affect customer behavior (Fildes et al., 2019).

Time series methods, regression models, and, more recently, techniques from machine learning are the main approaches for quantitative forecasting. In what follows, we briefly review the use of these frameworks in the context of demand forecasting, noting that our approach, as developed below, can easily take into account additional candidate models if desired. For a much more comprehensive recent review of the various existing approaches for estimating retail demand, see Fildes et al. (2019).

Classical time series analysis methods—e.g. exponential smoothing, ARIMA-type models, and state-space models—forecast demand based on the past sales history, and are thus appropriate only if historical temporal patterns are expected to continue during future demand periods. This often does not hold in retail practice, as exogenous effects such as irregular price promotions can have a significant impact on customer demand. Time series models that allow the incorporation of explanatory variables, such as ARIMAX models, attempt to combine the strengths of time series and regression modeling. For example, Arunraj and Ahrens (2015) showed that ARIMAX outperformed an ARIMA model for predicting daily sales of bananas.

Regression models aim to explain customer demand based on explanatory variables, potentially including time itself and, more generally, time-varying variables. Multiple linear regression is the most established approach for modeling directed relationships between explanatory variables (features) and demand. These features may include exogenous effects, such as price and marketing activities. However, the assumptions of linearity of the predictor, homoscedasticity, and normality of the response made in linear regression are often violated in the case of complex customer demand patterns (Ramaekers & Janssens, 2008; Ulrich et al., 2021). In retail applications, regression models with different specifications, such as multiplicative (log–log) and exponential (semi-log), have been used to model nonlinearities (Fildes et al., 2019). Steiner et al. (2007) and Weber et al. (2017) in particular arrived at improved forecasts when using nonlinear regression models. Regarding the response variable, i.e. demand, various distributional assumptions have been proposed in the literature, such as normal (Nahmias, 1994), gamma (Burgin, 1975), Poisson (Conrad, 1976), and negative binomial (Agrawal & Smith, 1996) distributions. Generalized additive models for location, scale, and shape (GAMLSS) allow us to build corresponding flexible distributional regression models with any distributional assumption for the response, and for potentially nonlinear effects of explanatory variables (Ulrich et al., 2021). Depending on the selected distribution, the mean, the variance, and the shape of the distribution can be modeled as a function of features. Quantile regression focuses on forecasting selected quantiles rather than the mean (Koenker & Hallock, 2001; Maciejowska et al., 2016), without making any distributional assumption for the response. Taylor (2007) combines time series and regression methods by applying an exponentially weighted quantile regression model for daily supermarket sales.

Machine learning methods have also attracted considerable attention in demand forecasting (cf. Fildes et al., 2019), and recent results of the M5 competition imply an exceptional performance of machine learning methods versus statistical models (Makridakis et al., 2020b; Makridakis et al., 2020c). For example, quantile regression, as described above, has been embedded within a random forest framework, yielding so-called quantile regression forests (Meinshausen, 2006), which have been applied in the context of load forecasting on the electricity market (Zhang et al., 2018). Other machine learning methods

explored in the demand forecasting literature include, *inter alia*, various types of (artificial) neural networks, such as back propagation, fuzzy, and recurrent (Ainscough & Aronson, 1999; Kuo, 2001; Zhang & Qi, 2005; Aburto & Weber, 2007; Salinas et al., 2020), support vector machines (Gür Ali & Yaman, 2013; Di Pillo et al., 2016), and generally many different classification methods (Ahmed et al., 2010; Ferreira et al., 2016; Baardman et al., 2018; Huber & Stuckenschmidt, 2020).

2.3. Model selection in the retailing literature

The literature on demand forecasting suggests that no single model universally outperforms other widely applied candidate models (Fildes, 1989; Wolpert & Macready, 1997; Ulrich et al., 2021). Across different SKUs, the associated best-performing model may differ, for example, in the model class (e.g. linear vs. quantile regression), or in their distributional assumption (e.g. normal vs. gamma). For any given retailer, there is thus a need to find a mapping from the demand pattern of any given SKU to the associated most promising combination of forecasting model and (potentially) associated distribution, with the aim of minimizing the overall costs.

A common strategy for selecting the best forecasting model is to assess the candidate models' historical performances, i.e. considering their accuracy in previous demand periods. This is most commonly done using either *individual selection*, *aggregate selection*, or *combination forecasts* (Fildes, 1989). Under individual selection, a model is chosen separately for each SKU under consideration, by assessing, for any given SKU, the candidate models' past performances exclusively for that SKU. Aggregate selection is used if the decision maker selects a single model to be applied to forecasting demand for all SKUs of interest, by evaluating the past performances across all these SKUs. In either of the two approaches, one further needs to determine the length of the temporal resolution, e.g. one day, one week, or one month, for both the evaluation and the actual application (before reassessment). In principle, the potential to minimize costs increases with shorter periods. However, this increased potential comes at the cost of increased variability in the model selection. Irrespective of the choice of the length of the temporal resolution, i.e. the time resolution, stability is the main advantage of aggregate selection over individual selection. Indeed, in ex-ante model selections, individual selection only rarely generates lower costs than the simpler aggregate selection (Fildes, 1989). Combination forecasts combine a set of forecasting models by building the prediction as a weighted average, either with equal weights (Fildes, 1989) or by optimizing each model's weights (Claeskens et al., 2016). Claeskens et al. (2016) identify that an unweighted combination forecast outperforms an optimally weighted one surprisingly often, as the estimation of the weights introduces a source of variance.

With respect to the selection of candidate models to be included in a combination forecast, Kourentzes et al. (2018) argue that the construction of a pool of reasonable models for forecasting is of crucial importance, and propose a heuristic for forecast pooling. Clemen (1989), Genre et al. (2013), and Claeskens et al. (2016) show that simple weighting schemes, e.g. the arithmetic mean, often yield equally good or better forecasts than more complex weighting schemes; individual model weightings are more promising only if there is ex-ante evidence that the different models generate notably different forecasting accuracies (Fildes & Petropoulos, 2015). The combination forecast reduces the risk of SKU-specific model misspecification, i.e. bias, compared to when using aggregate selection, while at the same time offering higher stability than individual selection. Thus, depending on the demand structure of the SKU, the benefits of robustness and stability from using combination forecasts may exceed the potential gains through model selection (Barrow & Kourentzes, 2016). In any given application, any of the three approaches may yield the highest forecasting accuracy (Fildes, 1989).

An alternative strategy that more explicitly addresses the heterogeneity in demand patterns is to cluster items with similar demand patterns, and then assign each cluster to an appropriate forecasting model (Eaves & Kingsman, 2004; Williams, 1984). Tanaka (2010), Ferreira et al. (2016), and Schneider and Gupta (2016) apply clustering approaches for new SKUs without historical demand. Existing SKUs are clustered based on similarity measures of data set characteristics, e.g. demand patterns, brand, price, and customer reviews, and new SKUs are assigned to the most adequate of these clusters. Boylan et al. (2008) propose a categorization framework for non-normal demand patterns, including intermittent demand (infrequent demand occurrences), slow-moving demand (low average demand), erratic demand (highly variable demand), lumpy demand (intermittent and highly variable demand), and clumped demand (intermittent, but constant demand). Potential problems with this type of clustering of seemingly homogeneous SKUs include the selection of break points between clusters, e.g. between low and high demand variability, and the selection of the number of clusters. Both decisions are made based on the historical data available, and as such may not always generalize well to new data sets. This is particularly problematic for demand patterns that may vary over time, for which partitioning the SKUs into clusters requires constant monitoring and regular adaptations.

At the SKU level in retailing, demand forecasts aim to generate predictions for a large number of SKUs over a short forecasting horizon (Fildes et al., 2019). Instead of building clusters manually, meta-learning methods allow the algorithm to learn the structure of the problem in an automated way based on the given data. The meta-learner gains meta-knowledge by linking the features to the forecasting accuracy of the forecasting methods available, and applies the developed knowledge to select the best expected forecasting method for a given feature set. Thus, a meta-learner usually consists of a feature set, base forecasters, and learned knowledge of how to select the best forecasting method given the features and base forecasters (see, e.g., Lemke et al., 2015; and Ma & Fildes, 2020). Contributions in the literature have demonstrated improved forecasting accuracy using meta-learning frameworks compared to the application of base forecasters (see, e.g., Prudêncio & Ludermir,

2004; Wang et al., 2009; Lemke & Gabrys, 2010; Monteromanso et al., 2020; and Ma & Fildes, 2020).

However, none of the existing contributions derives predictions for high service-level targets, accounting for the asymmetric costs for shortage and excess inventory that we find in retail practice. This limitation will be overcome with the approach developed below.

3. Classification-based model selection

We frame the model selection problem as a classification task, where the classes correspond to the different forecasting models available. Classification is a supervised learning approach where an algorithm learns the optimal mapping of input variables onto a target variable (corresponding to a class). After training the algorithm based on historical input variables with known class labels, the mapping can be applied to new input variables associated with previously unseen data. Referring to the definition of meta-learning by Ma and Fildes (2020), our approach includes all three components of meta-learning: features that explain customer demand, a pool of base forecasters, and a meta-learner that links the features to a forecasting method. In our approach, the decision tree corresponds to the meta-learner, as it selects the best expected model from labeled data.

In our setting, p input variables relevant for demand forecasting—e.g. price, known demand, or day of the week—are combined into a p-dimensional feature vector $\mathbf{x}$, and any class $m \in M = \{\text{linear regression, quantile regression, } \ldots\}$ corresponds to a candidate model for forecasting demand. In the first stage of our classification-based model selection (CMS), we label data from past demand periods according to an evaluation of the hypothetical performance of forecasting models within those demand periods. Specifically, in the training data, each demand period t comprises the associated feature vector $\mathbf{x}_t$ and a label $m_t \in M$ resulting from hypothetical model performance. Based on the training data, we then set up a classifier, i.e. a mapping

$$\begin{aligned} \phi &: \Omega \to M; \\ \mathbf{x} &\mapsto \phi(\mathbf{x}) = m, \end{aligned} \tag{4}$$

with Ω denoting the p-dimensional space of all possible feature vectors. For future data, in the second stage we apply ϕ to automatically select the most promising forecasting model from M based on the feature vector $\mathbf{x}$. In essence, the classification algorithm thus selects the demand forecasting model to apply by identifying which models worked best in past demand periods similar to the one currently considered.

3.1. Generating labeled data

In the first stage indicated above, we need to complement each feature vector in the training data, corresponding to certain characteristics of the associated demand period, by a label indicating the most adequate model. At first sight, it might appear natural to simply choose as the label, for each past demand period, the model that would have (ex-post) produced the lowest total costs from (3) if it had been applied to forecast demand in that period. However, at a very high service level, say 97%, a model that generally tends to underestimate demand variance, e.g. Poisson regression (Ulrich et al., 2021), might be the best-performing model in the majority of demand periods—simply because its competitors' predictions in most instances will more drastically overshoot the actual demand—but would be extremely costly in those demand periods where it is not the best-performing model, due to undershooting the actual demand, which comes at a much higher price. As a consequence, it would be myopic to ex-post deem a model to be the most adequate choice for any given demand period just because it *happened* to have produced the lowest costs on that particular occasion, when in fact the *expected* costs under that model were very high.

It is worth remembering that labeling the feature vector of some past demand periods is only an intermediate step towards solving the underlying economic problem. When a decision on q is to be made at some time t, the decision maker seeks to choose the best forecasting model m for predicting the probability distribution of the uncertain demand at this period t. The resulting cumulative distribution function, $\hat{F}_t^m$, would subsequently be used to determine the optimal inventory level $\hat{q}_t^m$ with respect to the newsvendor problem, as stated in Eq. (1). Hence, a decision maker would choose model m if the inventory level $\hat{q}_t^m$ resulting from the application of that model yields minimal expected total costs,

$$E[C_t(\hat{q}_t^m)] = hE(\hat{q}_t^m - D_t)^+ + bE(D_t - \hat{q}_t^m)^+. \tag{5}$$

Within the labeling process, model performance thus ought to be measured according to (5), rather than based only on the actual costs realized at time t. Since the true probability distribution that is needed to calculate (5) is unknown, even for past demand periods, we need to estimate expected total costs using information from our data set. We propose to do so by evaluating the overall performance of model m, for any given past demand period t, on a *set* Ψ_t of demand periods similar to the one under consideration. This idea is similar in spirit to nonparametric estimation techniques (e.g. kernel-based nonparametric regression), where in order to inform the estimator of a quantity of interest (e.g. the value of the regression function at a given covariate value), a neighborhood of sufficiently similar data points is considered (defined, e.g., using bandwidth and kernel functions).

Specifically, for the data from the training set, the labeling process in our approach involves the following steps:

1. estimation of each candidate forecasting model m based on historical data;
2. for each demand period t and each m, we obtain
   - an estimate $\hat{F}_t^m$ of the demand distribution F_t, which we apply to derive the cost-optimal $\hat{q}_t^m$ according to the newsvendor problem,
   - or, alternatively, a direct prediction of the cost-optimal $\hat{q}_t$ for the nonparametric quantile regression models;

3. from those forecasts, and given cost parameter values for shortage b and excess inventory h, the model-specific costs for each demand period are calculated according to Eq. (3);
4. for each demand period t, we identify a set Ψ_t of demand periods similar to the one demand period considered, and we select as the label m_t the model that realized the lowest overall costs across all elements from Ψ_t:

$$m_t = \underset{m \in M}{\operatorname{argmin}} \sum_{t' \in \Psi_t} c_{t'}(\hat{q}_{t'}^{m})$$

The main choice to be made is clearly how to select the demand periods to be included in Ψ_t. To this end, we define a quantile threshold $v \in [0, 1]$ (e.g. 10% of all observations for $v = 0.1$), and from all available demand periods, we then include those in Ψ_t that are among the $100 \cdot v\%$ with the smallest distances to the feature vector associated with the demand period at time t. With this approach, the size of Ψ_t depends on the total number of feature vectors available in the training data, which is desirable: for larger training data sets, there will be more feature vectors in the vicinity of the feature vector at time t, in which case it makes sense to build a larger Ψ_t to increase robustness.

The quantile threshold v can be optimized via hyperparameter tuning, i.e. by validating each v from a set of possible quantile thresholds using out-of-sample data. Selecting the most adequate value of v via hyperparameter tuning thus requires an additional partition of the training data into a calibration and a validation set. Specifically, we propose to split the training data available such that 90% of the data are used to train the classifier, for any given quantile threshold v, with the remaining 10% of the data used to evaluate the performance of the classifier across the different values of v. The parameter v is then fixed at the value that led to the best performance in the validation data set—as measured by the total costs resulting from applying the classifier for the given v—and subsequently used to build the final classification algorithm based on the complete training set.

To measure distances between feature vectors, various metrics can be used, such as the Manhattan, Minkowski, or Mahalanobis distance; see Prasath et al. (2017) for a review, including a performance comparison in multiple applications. We chose to use the Euclidean distance metric, as it is most widely used, due to its simplicity and intuitive interpretation as the straight-line distance between two points (Hu et al., 2016). To avoid the dominance of features with values that are relatively large in magnitude (e.g. mean demand), and to generally avoid any scale dependence, we normalize features to values between 0 and 1, corresponding to the lowest value and the highest value of the feature in the training set. Categorical features such as the "day of the week" are rewritten using binary features (e.g. Monday vs. not Monday, and likewise for the other days), and are identified as numerical using the values 0 for "no" or 1 for "yes". To avoid the relative dominance of the binary features (e.g. five specifications for the feature "day of the week"), we divide the corresponding feature values by the number of dummy variables used to specify the feature (see the Appendix for an example).

3.2. Selection of a classifier

After building the labeled training data, as described in the previous section, with a fixed v, chosen for example via hyperparameter tuning, the CMS approach we suggest then proceeds by training a classifier ϕ, which in our context is a mapping from the set of potential feature vectors $\mathbf{x} \in \Omega$ to the set of candidate forecasting models $m \in M$; cf. Eq. (4). In principle, various classifiers can be used for this purpose, e.g. logistic regression, linear discriminant analysis, k-nearest-neighbors, or support vector machines (Press & Wilson, 1978; Kotsiantis et al., 2006). In retail practice, an important requirement of operational management is transparency regarding how forecasts are obtained, as this allows the inclusion of expert knowledge in case of exceptional circumstances (cf. Trusov et al., 2006; Fildes et al., 2009; Davydenko & Fildes, 2013). Decision trees thus constitute an obvious choice, as they are particularly intuitive and easy to interpret (Goodwin & Wright, 2014). Unlike, say, k-nearest-neighbors or support vector machines, they do not only produce a class as output, but also allow insights into the drivers of the model choice (Armstrong, 2001; Schwartz et al., 2014). In this contribution, we thus focus on decision trees, noting that the performance of the CMS approach might be further improved by additionally optimizing with respect to the choice of the classifier applied.

A decision tree is an algorithm comprising a sequence of decisions to build a tree structure, which, when applied to a new feature vector, ultimately leads to its classification. Decisions are commonly binary, and in what follows we restrict our presentation to this type of tree. Fig. 2 shows an example of a decision tree involving binary decisions. The *root node* at the top represents the full data set. For any new feature vector to be classified, the classification is performed by traversing through the nodes from top to bottom, making binary decisions at every node until arriving at a *leaf node*, which represents the final class decision, in our case the forecast model to be applied for the given feature characteristics. Every node between the root node at the top and the leaf nodes at the bottom is called a *decision node*. Each decision node comprises information on (i) the split dimension (i.e. based on which feature the decision is made) and (ii) the split point (i.e. which value defines the threshold, either side of which corresponds to a different decision; for binary variables this naturally reduces to a yes-or-no decision). In the example shown in Fig. 2, a feature vector including known orders greater than 100 units, on a Saturday, leads to the selection of the quantile regression as the forecasting model.

When training the decision tree, each node is optimized with respect to (i) and (ii) using some criterion, e.g. the Gini impurity,

$$\mathrm{GI} = \sum_{m=1}^{M} p_m(1 - p_m),$$

where p_m denotes the proportion of class m observations within a given set of points, hence the probability that a randomly selected unit will be from class m. For each

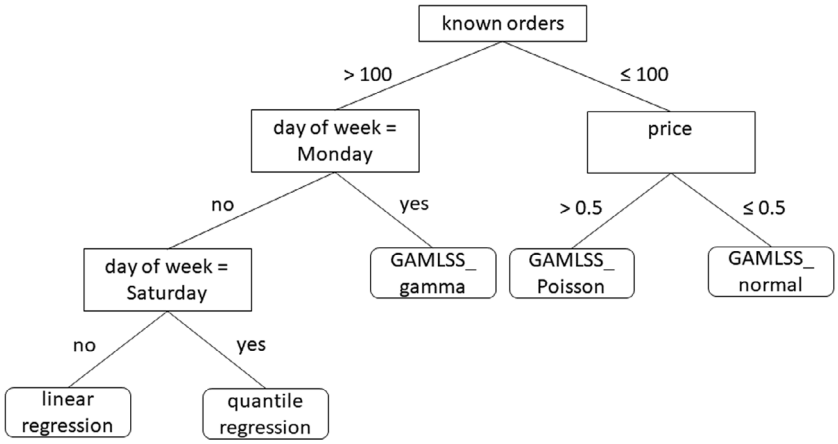

Fig. 2. Example of a general decision tree.

class resulting from a decision, this criterion measures its degree of impurity. If the class is pure, then $p_m \in \{0, 1\}$ for all m, such that GI $= 0$. Optimization with respect to (i) and (ii) then seeks to find, for each decision node (starting at the top), the split dimension and the associated split point such that the weighted sum of the Gini impurities of the resulting classes is minimized (with the weights equal to the overall proportion of the classes). As decision trees are prone to overfitting, trees are usually restricted in their growth, or pruned after maximum growth (Thomassey & Fiordaliso, 2006).

4. Case study

4.1. E-grocery data

The data set available to us covers 111 SKUs within the categories of fruits, vegetables, and bake-off products from five regional fulfillment centers. As most SKUs are supplied in each fulfillment center, the data include 546 time series. The business processes of the e-grocery retailer considered here comprise new types of customer demand data that are not as such available in store retailing. In particular, during the online shopping process, no inventory information is shown to the customer. Customers are thus able to add any SKU with infinite stock availability to the shopping basket. This allows for a comprehensive monitoring of true customer preferences. As a consequence, in contrast to store retailing, where point-of-sale data are distorted by stock-outs, the customer demand information available to the e-grocery retailer is effectively uncensored. In our case study, we consider demand as monitored during the online shopping process of the customer before any restricted availability of SKUs is revealed during check-out. Compared to sales data, demand is thus higher than sales for SKUs affected by stock-outs, and lower than sales for SKUs selected as substitutes. In addition, customers can order up to 14 days in advance. As a consequence, the e-grocery retailer can include the known demand information at the replenishment decision time in the final replenishment order. Assuming that most customers do not cancel a placed order, known customer demand for an SKU can be considered as a minimum demand value. The e-grocery retailer considered here replenishes daily, with a lead time of three days, which includes the processes of stowing and picking within the warehouse. We thus generate three-day-ahead forecasts on a daily basis, e.g. on Wednesday to forecast Saturday demand.

4.2. Feature selection

We generally use features (i.e. explanatory variables) to predict q_t. In the CMS approach, we additionally use the features to identify a set of similar demand periods. The following six features were pre-selected based on business considerations:

(1) price (a continuous variable);
(2) seasonality as captured by the trigonometric variables $\sin\left(\frac{2\pi t}{365}\right)$, $\cos\left(\frac{2\pi t}{365}\right)$, $\sin\left(\frac{4\pi t}{365}\right)$, and $\cos\left(\frac{4\pi t}{365}\right)$;
(3) day of the week (five dummies);
(4) ID of the FC (four dummies);
(5) marketing activities (one dummy);
(6) known orders (a count variable).

We cover the supply side of seasonality (price), the demand side of seasonality (as per (2)), the demand profile of the week (the day of the week), geographical differences (the ID of the FC), and marketing activities (a binary flag for whether or not marketing activities were in place). With respect to marketing activities, using a single dummy variable is a rather simple parameterization that might be improved by including additional variables indicating possible marketing measures available in online retailing, such as newsletter marketing or promoting price discounts at the landing page. However, for our case, we simply do not have more specific information in the data set available. We further include the known orders as an additional feature that is not available in traditional store retailing. As customers are able to place orders up to 14 days in advance, known orders represent the corresponding customer demand information available to the retailer at the time of the replenishment decision, i.e. three days prior to the demand period considered. Between the replenishment decision and the

demand period, additional customer demand may occur. Consequently, known orders are usually smaller than realized demand.

For each individual SKU, we select the subset of the features above by a stepwise forward selection such that the Akaike information criterion (AIC) is minimized. We minimize the AIC by a stepwise forward selection in a Gaussian linear model and—for simplicity—we use the selected features within all models. By minimizing the AIC, we include only those features that notably improve the goodness of fit, thus reducing the risk of overfitting. As opposed in particular to the Bayesian information criterion (BIC), we used the AIC, since this criterion aims at selecting the model with the highest predictive accuracy, whereas the BIC aims at selecting the model most likely to be true (see, e.g., Shmueli, 2010), which is not of primary interest in our setting.

4.3. Sliding-window approach to model fitting and classifier training

The complete data set we consider covers the time period from 11/2015 to 10/2018, i.e. 36 months. To account for trends in the fast-growing e-grocery business, the different statistical models considered—e.g. linear regression and quantile regression (details below)—are updated dynamically using the following sliding-window approach:

- fit model to data from months 1–12 (i.e. 11/2015–10/2016) to forecast demand in month 13;
- fit model to data from months 2–13 to forecast demand in month 14;
- . . .;
- fit model to data from months 24–35 to forecast demand in month 36.

We thus obtain, for each of the statistical models taken into account, predictions for each demand period in months 13–36 (i.e. in the time period 11/2016–10/2018).

Specifically for the CMS approach to model selection, we use an analogous sliding-window approach, partitioning any given 13 months of data into 12 months of labeled data for classifier training and the remaining month for the actual application of the classifier:

- train classifier to labeled training data from months 13–24 to automatically allocate a forecasting model to each demand period in month 25;
- . . .;
- train classifier to labeled training data from months 24–35 to automatically allocate a forecasting model to each demand period in month 36.

Note in particular that the labeling step is conducted based on out-of-sample forecasts: for example, labels in months 13–24 are generated based on out-of-sample forecasts in this period, obtained from the models fitted to data from months 1–12.

4.4. Candidate forecasting models

Based on our literature review, we consider linear regression assuming normally distributed errors (Reg), generalized additive models for location, scale, and shape (GAMLSS), quantile regression (QuantReg), quantile regression forests (QuantRegForest), and ARIMAX as our set of candidate models for demand forecasting. For the GAMLSS class, we consider the normal, gamma, Poisson, and negative binomial distribution, as proposed in the literature. For each demand period t from November 2016 to October 2018 we obtain an estimate $\hat{F}_t$ for the demand distribution F_t. We apply this estimate to derive the cost-optimal $\hat{q}_t^*$ according to the newsvendor problem, for each of the parametric models, and a direct prediction of the cost-optimal $\hat{q}_t^*$, for the nonparametric quantile regression models. From those forecasts, and given the cost values for shortage b and excess inventory h, the model-specific costs for each demand period are calculated according to Eq. (3). For the selected service-level target of 97% and an estimation of the inventory cost parameter h based on the retailer's margin and operational costs, we derive the associated shortage cost according to Eq. (2). After the application and evaluation of the forecasting models, each demand period from November 2016 to October 2018 is thus complemented by a demand prediction and a corresponding cost value for each model.

Table 3
Demand forecasting models that realize the lowest total costs for four SKUs analyzed for all demand periods in 2017 for one selected fulfillment center.

SKU	Year	Best model
Garden parsley	2017	GAMLSS.gamma
Apples	2017	QuantRegForest
Wraps	2017	GAMLSS.negBin
Honeydew melons	2017	QuantRegForest

Comparing the realized costs of the applied models, we observe that there is indeed no single model that universally realizes the lowest total costs, as was to be expected based on the literature (Fildes, 1989; Ulrich et al., 2021). To illustrate this, Table 3 displays the forecasting model that realizes the lowest total costs for the SKUs of garden parsley, apples, wraps, and honeydew melons in 2017 for one selected fulfillment center. For this small subset of four SKUs, three different forecasting models yield the lowest total costs.

For a corresponding comparison at a finer temporal resolution, Table 4 displays the demand forecasting models that lead to the lowest total costs for honeydew melons for each month in 2017 for FC 1, and the percentage savings in costs compared to QuantRegForest, which performed best for 2017 in total. However, even the overall best-performing model, QuantRegForest, gives the best results only for two months in 2017. For the other ten months, different model classes (ARIMAX, Reg, QuantReg, GAMLSS) realize lower costs.

When applying an even finer temporal resolution, we observe that for each week in January, a different forecasting model minimizes total costs. The same holds true if we look for the models minimizing total costs for each day of the week (results not shown). This exercise demonstrates the potential to save costs by matching any given demand pattern to its most appropriate forecasting model. However, our results also demonstrate the challenge of selecting the most appropriate model.

Table 4
Demand forecasting model that realizes the lowest total costs for the SKU of honeydew melons for each month in 2017 for one selected fulfillment center and the percentage savings in costs compared to QuantRegForest.

SKU	Month–year	Best model	Savings
Honeydew melons	01–17	QuantRegForest	0%
Honeydew melons	02–17	GAMLSS.gamma	−49%
Honeydew melons	03–17	Reg	−13%
Honeydew melons	04–17	QuantRegForest	0%
Honeydew melons	05–17	ARIMAX	−42%
Honeydew melons	06–17	ARIMAX	−17%
Honeydew melons	07–17	QuantReg	−22%
Honeydew melons	08–17	GAMLSS.negBin	−32%
Honeydew melons	09–17	GAMLSS.Poisson	−39%
Honeydew melons	10–17	GAMLSS.Poisson	−10%
Honeydew melons	11–17	GAMLSS.Poisson	−34%
Honeydew melons	12–17	GAMLSS.Poisson	−16%
	2017		−18%

4.5. Building labeled data

For the CMS approach to be applicable, we need labeled training data (cf. Section 3.1). To that end, for any 13 months of data, with the sliding-window approach to classifier training and subsequent classification (see Section 4.3), the demand periods from the first 12 months are labeled. This is achieved by identifying, for each of the demand periods to be labeled, a set Ψ_t of demand periods similar to the one considered (from the same 12 months of data). Specifically, we include all those demand periods in Ψ_t whose feature vectors are among the $100 \cdot v\%$ nearest to the feature vector of the demand period at time t. The individual v for each SKU and month was chosen by validating, separately for each SKU, each v from a set of possible quantile thresholds (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%) using out-of-sample data. Each feature vector is then labeled with the model that realizes the lowest total costs overall for Ψ_t.

4.6. Training the decision tree

As described in Section 4.3, for each month from 11/2017–10/2018, the decision tree is trained based on the preceding 12 months of labeled data, and then applied to automatically select the forecasting model for each demand period in the month considered. Thus, all demand periods from 11/2017–10/2018 are used as out-of-sample test data. To tailor the decision tree to the training data, we use the R function `rpart` (Therneau & Atkinson, 2018) using the default options. Thus, in particular, we use the Gini impurity measure to determine the split points. These default options also reduce the risk for overfitting when training the tree.

4.7. Benchmark approaches

Based on our literature review in Section 2.3, we consider individual selection, aggregate selection, and an equally weighted combination forecast as benchmark methods. In principle, individual selection and aggregate selection can be applied for any temporal resolution, both for evaluation and for the actual application (before reassessment); we consider daily, weekly, and monthly model selection. Considering the concept of pooling by Kourentzes et al. (2018)—i.e. to pre-select only reasonable models for forecasting—we exclude GAMLSS.Poisson from the combination forecast, because for this model the realized costs are significantly higher compared to the benchmarks, and the service level strongly deviates from the service-level target (Ulrich et al., 2021). We further use all base forecasters (i.e. fixed individual models, without selection) as additional benchmarks. Overall, we apply 15 benchmarks to assess the performance of our CMS approach: daily/weekly/monthly individual selection, daily/weekly/monthly aggregate selection, combination forecast, ARIMAX, GAMLSS.gamma, GAMLSS.negBin, GAMLSS.normal, GAMLSS.Poisson, Reg, QuantReg, and QuantRegForest.

4.8. Results

The boxplots shown in Fig. 3 display the *SKU-specific* percentage differences between the costs obtained under the seven selection and combination benchmarks on the one hand, and the CMS approach on the other hand, for the data from the test period 11/2017–10/2018 under the 97% service-level target.

Overall, the CMS approach and the combination forecast performed about equally well. In particular, they achieved the same median total costs across SKUs. Individual and aggregate selection yielded 2%–13% higher median total costs than CMS, depending on the time resolution considered. Selection at the daily level performed particularly poorly, due to the relatively strong negative impact of the Poisson model: for high service-level targets, the Poisson performs best in many demand periods but is extremely costly in those demand periods where it is undershooting. To test our results for formal significance, we additionally applied the multiple comparison with the best (MCB) test (Koning et al., 2005; Makridakis et al., 2020a). The performance of CMS was deemed to be significantly better than those of all benchmarks except for weekly aggregate selection and the combination forecast. Within the CMS approach, QuantReg was most often selected as the forecasting model (28% of all cases), followed by the GAMLSS.negBin (23%), and GAMLSS.gamma (17%).

Table 5 displays the complete results, including the performance of the base forecasters, i.e. the individual models. QuantReg and GAMLSS.negBin, as fixed models, did indeed perform as well as CMS and the combination forecast in terms of median costs. The combination forecast, QuantReg, weekly individual selection, and CMS generated the lowest costs for 18%, 15%, 10%, and 9% of all SKUs, respectively. CMS, the combination forecast, GAMLSS.gamma, and GAMLSS.negBin achieved service levels of 96%. That is, they deviated from the service-level target by only 1%. As expected, GAMLSS.Poisson substantially undershot the desired service level.

Overall, the results obtained in this case study indicate that CMS generates modest improvements in terms of median costs and service-level realization compared to

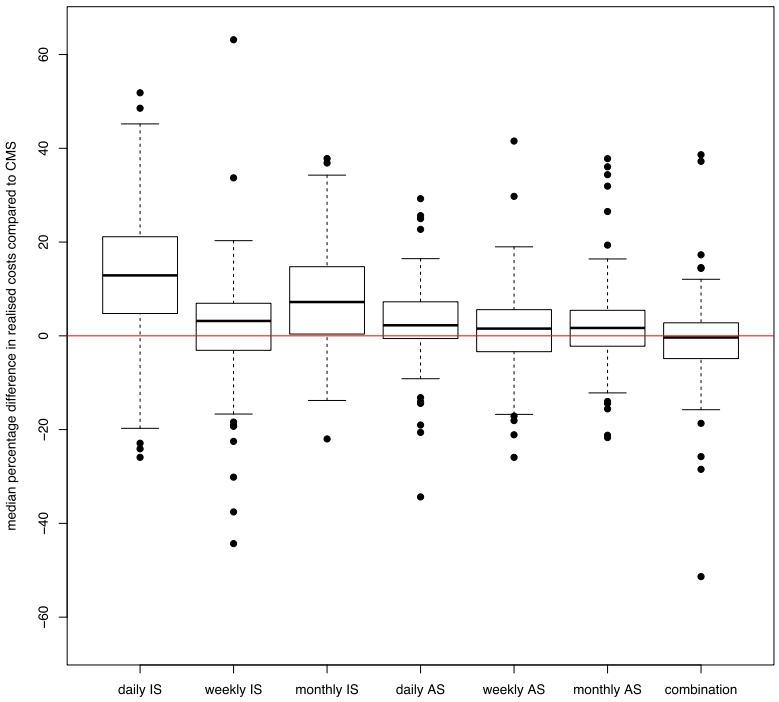

Fig. 3. Boxplots of the SKU-specific percentage differences in realized costs for individual selection (IS), aggregate selection (AS), and combination forecast (combination) with the different temporal resolutions considered (day, week, and month) compared to the CMS method. Across the seven benchmarks, there are ten percentage differences that lie outside the range depicted here.

Table 5
Results of the CMS approach and the benchmark methods, for 111 SKUs analyzed in the test period 11/2017–10/2018. *Share best method* gives the percentages of SKUs for which a given method realized the lowest costs. *Diff. median costs* are the percentage differences between the median total costs (over all SKUs) under a method compared to those obtained when using CMS. *Realized SL* indicates the service level achieved.

Method	Diff. median costs	Share best method	Realized SL
Model selection approaches			
CMS	0%	9%	96%
Daily individual	13%	4%	92%
Weekly individual	3%	10%	95%
Monthly individual	7%	4%	95%
Daily aggregate	2%	1%	95%
Weekly aggregate	2%	8%	95%
Monthly aggregate	2%	5%	95%
Combination forecast			
Combination (equal weights)	0%	18%	96%
Base forecasters			
ARIMAX	27%	1%	95%
GAMLSS.gamma	4%	5%	96%
GAMLSS.negBin	0%	9%	96%
GAMLSS.normal	2%	2%	95%
GAMLSS.Poisson	25%	5%	88%
Reg	2%	4%	95%
QuantReg	0%	15%	95%
QuantRegForest	17%	2%	94%

the established selection methods. Whether a cost reduction of about 2%–3% compared to the best of these benchmark approaches is practically relevant, and in particular whether this can offset the additional effort involved in implementing and continually running the much more complex CMS approach, clearly depends on the overall magnitude of the sales. The combination forecast and the two base forecasters, GAMLSS.negBin and QuantReg, performed about equally well in our case study. Thus, our case study does not (and clearly cannot) ultimately determine which of these approaches is superior. More empirical comparisons are required, and it should be

emphasized here that both the CMS approach (see the discussion below) and the combination forecast (via optimization of the weights) can be further refined and improved. Our case study thus merely provides a first impression of the potential of the CMS approach relative to existing benchmarks, under the conditions present in our setting.

The setting in our case study involves a fairly high service-level target of 97%. When re-running the analysis for the much lower (and in our case economically disadvantageous) service-level target of 75%, the differences in performance between the selection and combination methods considered were much lower (0%–3% difference in median costs for the 75% service level vs. 0%–13% for the 97% service level).[1] This is to be expected, because in such a case the different forecasting models applied are not as sensitive as when predicting the extreme right tail of the demand distribution, due to the lower asymmetry in the cost parameters.

5. Discussion

In this paper, we presented classification-based model selection (CMS) as a general new approach for automated model selection in retail demand forecasting. CMS offers a flexible framework for addressing the practical challenges associated with the ever more complex demand patterns encountered in retail practice. Specifically, we proposed leveraging the large amounts of demand data being collected, using a data-driven (algorithmic) approach to the model selection exercise. Compared to the application of more complex meta-learning methods that pool SKUs and aim to find the best combination of base forecasters (see, e.g., Ma & Fildes, 2020; Montero-manso et al., 2020), CMS benefits from better interpretability and, as a consequence, potentially higher acceptance by the decision makers. The approach allows expert adjustments for information not captured by the data—a common practice in retail forecasting (Fildes et al., 2019). Compared to the established individual selection approach, the key idea of our framework is that model performance is evaluated not based on the most *recent* demand periods, but rather the most *similar* past demand periods. However, as general limitations, considering each SKU individually does not allow us to capture cross-sectional patterns, and limited data may result in noisy elasticities (Ma & Fildes, 2020).

Our exploratory case study constitutes a proof of concept of the proposed CMS approach, and more generally a comparison of several conceptually different approaches to model selection in an e-grocery setting. At this point, however, there is no clear evidence indicating that our approach is superior to existing ones, and in particular to combination forecasts. While in our case study we used e-grocery data, the approach is also applicable to traditional store retailing data, or in fact to other fields of forecasting. The main objective of the present paper was to lay out the general idea of using classifiers to match demand periods to forecasting models. As a consequence, for clarity of presentation we abstained from optimizing each individual analysis step. In particular, we made rather ad hoc choices regarding (i) the distance measure used in the labeling process, (ii) which features to include in the forecasting models, and—perhaps most importantly—(iii) which classification algorithm to apply for selecting models. To some extent these choices were made based on business considerations, e.g. when pre-selecting features known to be relevant in retail practice, or when using decision trees for interpretability. Other choices concerning the implementation were driven primarily by the need for conciseness of presentation. For example, much effort could go into feature selection: rather than using the rather simple AIC-based selection (and based only on the linear model), LASSO regularization could have been used (see, e.g., Uniejewski et al., 2019). The inclusion of additional numerical features capturing the characteristics of the time series data, e.g. concerning stability, lumpiness, correlation, etc., may potentially further improve the performance (Ma & Fildes, 2020). However, it is difficult to capture such properties embedded in time series data by hand, as the space of possible features is large. A machine learning framework for automatically learning a suitable feature representation from raw time series, as proposed by Ma and Fildes (2020), could in principle be implemented within our CMS approach, but would further increase the computational complexity. Finally, including information on feature relevance might improve the selection of demand periods similar to the one being considered.

The given contribution should thus be regarded as a starting point for future research into automated model selection in demand forecasting, as it seems highly likely that the optimization of steps (i)–(iii), above, will further improve the performance of the approach. We also envisage possible extensions on a higher conceptual level. First, there may be alternative strategies for labeling the training data, e.g. that more explicitly accommodate the misclassification costs (thus replacing the usage of v, and hence a neighborhood, by a cost function within the classification). Second, many of the classifiers that could potentially be applied within the CMS approach do in fact provide *probabilistic* information on class membership (here corresponding to a pointer to the most promising model for demand forecasting). This information can easily be translated into weights of the associated candidate models within a combination forecast. Overall, given its conceptual appeal and some promising first results, we are confident that the CMS approach will prove to be a useful addition to the suite of demand forecasting approaches. However, our empirical comparison of various model selection approaches also provides further evidence that there is no one-size-fits-all method, indicating that demand forecasting in practice will generally need to be tailored to the specific business setting.

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.

[1] Detailed results for the 75% service level are reported in the Supplementary Material.

Acknowledgments

We thank two anonymous referees and the associate editor for their insightful comments, which were tremendously useful.

Appendix A

As an example of dividing feature values by the number of dummy variables (see the corresponding remarks at the end of Section 3.1), we consider three hypothetical feature vectors that differ in the feature "day of the week" (Monday or Tuesday, as indicated by the first two components), as well as in the feature price (0.5 EUR, 0.49 EUR, or 0.1 EUR, as indicated by the last component):

$$(1, 0, 0, 0, 0, 0.50)$$
$$(0, 1, 0, 0, 0, 0.49)$$
$$(1, 0, 0, 0, 0, 0.10)$$

Normalizing these feature vectors, we obtain:

$$(1, 0, 0, 0, 0, 1.00)$$
$$(0, 1, 0, 0, 0, 0.98)$$
$$(1, 0, 0, 0, 0, 0.00)$$

The Euclidean distance then equals 1.41 between the feature vectors (1) and (2), and 1 between the feature vectors (1) and (3), although the price between the feature vectors (1) and (3) is very similar, and the feature "day of the week" differs by only one day. Thus, to account for the relative importance of features, we reduce the high weight of the binary feature, dividing each day by five. We obtain the following feature vectors:

$$(\tfrac{1}{5}, 0, 0, 0, 0, 1.00)$$
$$(0, \tfrac{1}{5}, 0, 0, 0, 0.98)$$
$$(\tfrac{1}{5}, 0, 0, 0, 0, 0.00)$$

After the manipulation of the binary features, the Euclidean distance equals 0.42 between the feature vectors (1) and (2), and 1 between the feature vectors (1) and (3), indicating that the feature vectors (1) and (2) are more similar to each other than the feature vectors (1) and (3).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.05.010.

References

Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, *7*(1), 136–144. http://dx.doi.org/10.1016/j.asoc.2005.06.001.

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics*, *43*(6), 839–861. http://dx.doi.org/10.1002/(SICI)1520-6750(199609)43:6<839::AID-NAV4>3.0.CO;2-5.

Ahmed, N. K., Atiya, A. F., El Gayar, N., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5), 594–621. http://dx.doi.org/10.1080/07474938.2010.481556.

Ainscough, T. L., & Aronson, J. E. (1999). An empirical investigation and comparison of neural networks and regression for scanner data analysis. *Journal of Retailing and Consumer Services*, *6*(4), 205–217. http://dx.doi.org/10.1016/s0969-6989(98)00007-1.

Anderson, E. T., Fitzsimons, G. J., & Simester, D. (2006). Measuring and mitigating the costs of stockouts. *Management Science*, *52*(11), 1751–1763. http://dx.doi.org/10.1287/mnsc.1060.0577.

Armstrong, J. S. (2001). Selecting forecasting methods. In *Principles of forecasting: A handbook for researchers and practitioners (Ed. J. Scott Armstrong)* (pp. 417–439). Kluwer, https://repository.upenn.edu/cgi/viewcontent.cgi?article=1181&context=marketing_papers.

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321–335. http://dx.doi.org/10.1016/j.ijpe.2015.09.039.

Baardman, L., Levin, I., Perakis, G., & Singhvi, D. (2018). Leveraging comparables for new product sales forecasting. *Production and Operations Management*, *27*(12), 2340–2343. http://dx.doi.org/10.1111/poms.12963.

Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. http://dx.doi.org/10.1016/j.ijpe.2016.03.017.

Boylan, J. E., Syntetos, A. A., & Karakostas, G. C. (2008). Classification for forecasting and stock control: a case study. *Journal of the Operational Research Society*, *59*(4), 473–481. http://dx.doi.org/10.1057/palgrave.jors.2602312.

Burgin, T. A. (1975). The gamma distribution and inventory control. *Operational Research Quarterly*, *26*(3), 507–525. http://dx.doi.org/10.2307/3008211.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140–1154. http://dx.doi.org/10.1016/j.ejor.2006.12.004.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762. http://dx.doi.org/10.1016/j.ijforecast.2015.12.005.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. http://dx.doi.org/10.1016/0169-2070(89)90012-5.

Conrad, S. (1976). Sales data and the estimation of demand. *Operational Research Quarterly*, *27*(1), 123–127. http://dx.doi.org/10.2307/3009217.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*(3), 510–522. http://dx.doi.org/10.1016/j.ijforecast.2012.09.002.

Di Pillo, G., Latorre, V., Lucidi, S., & Procacci, E. (2016). An application of support vector machines to sales forecasting under promotions. *4OR*, *14*(3), 309–325. http://dx.doi.org/10.1007/s10288-016-0316-0.

Eaves, A. H. C., & Kingsman, B. G. (2004). Forecasting for the ordering and stock-holding of spare parts. *Journal of the Operational Research Society*, *55*(4), 431–437. http://dx.doi.org/10.1057/palgrave.jors.2601697.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, *18*(1), 69–88. http://dx.doi.org/10.1287/msom.2015.0561.

Fildes, R. (1989). Evaluation of aggregate and individual forecast method selection rules. *Management Science*, *35*(9), 1056–1065. http://dx.doi.org/10.1287/mnsc.35.9.1056.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23. http://dx.doi.org/10.1016/j.ijforecast.2008.11.010.

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2019.06.004.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701. http://dx.doi.org/10.1016/j.jbusres.2015.03.028.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121. http://dx.doi.org/10.1016/j.ijforecast.2012.06.004.

Goodwin, P., & Wright, G. (2014). *Decision analysis for management judgment* (5th ed.). Wiley.

Gür Ali, Ö., & Yaman, K. (2013). Selecting rows and columns for training support vector regression models with large retail datasets. *European Journal of Operational Research*, *226*(3), 471–480. http://dx.doi.org/10.1016/j.ejor.2012.11.013.

Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k - nearest neighbor classification for medical datasets. *Springer Plus*, *5*(1304), http://dx.doi.org/10.1186/s40064-016-2941-7.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420–1438. http://dx.doi.org/10.1016/j.ijforecast.2020.02.005.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*(4), 143–156. http://dx.doi.org/10.1257/jep.15.4.143.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*(3), 397–409. http://dx.doi.org/10.1016/j.ijforecast.2004.10.003.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Supervised machine learning: A review of classification techniques. *Artificial Intelligence Review*, *26*(3), 159–190. http://dx.doi.org/10.1007/s10462-007-9052-33.

Kourentzes, N., Barrow, D., & Petropoulos, F. (2018). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226–235. http://dx.doi.org/10.1016/j.ijpe.2018.05.019.

Kuo, R. J. (2001). Sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research*, *129*(3), 496–517. http://dx.doi.org/10.1016/S0377-2217(99)00463-4.

Lemke, C., Budka, M., & Gabrys, B. (2015). Metalearning: a survey of trends and technologies. *Artificial Intelligence Review*, *44*(1), 117–130. http://dx.doi.org/10.1007/s10462-013-9406-y.

Lemke, C., & Gabrys, B. (2010). Neurocomputing meta-learning for time series forecasting and forecast combination. *Neurocomputing*, *73*(10), 2006–2016. http://dx.doi.org/10.1016/j.neucom.2009.09.020.

Ma, S., & Fildes, R. (2020). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111–128. http://dx.doi.org/10.1016/j.ejor.2020.05.038.

Maciejowska, K., Nowotarski, J., & Weron, R. (2016). Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, *32*(3), 957–965. http://dx.doi.org/10.1016/j.ijforecast.2014.12.004.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M5 accuracy competition: Results, findings and conclusions. In *Working Paper*.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M5 uncertainty competition: Results, findings and conclusions. In *Working Paper*.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987–996. http://dx.doi.org/10.1287/mnsc.29.9.987.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999, http://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf.

Montero-manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. http://dx.doi.org/10.1016/j.ijforecast.2019.02.011.

Nahmias, S. (1994). Demand estimation in lost sales inventory systems. *Naval Research Logistics*, *41*(6), 739–757. http://dx.doi.org/10.1002/1520-6750(199410)41:6<739::AID-NAV3220410605>3.0.CO;2-A.

Prasath, V. B. S., Alfeilat, H. A. A., Lasassmeh, O., Hassanat, A. B. A., & Tarawneh, A. S. (2017). Distance and similarity measures effect on the performance of k-nearest neighbor classifier - a review. (pp. 1–50). ArXiv Preprint https://arxiv.org/abs/1708.04321.

Press, J. S., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*(364), 699–705. http://dx.doi.org/10.1080/01621459.1978.10480080.

Prudêncio, R. B., & Ludermir, T. B. (2004). Meta-learning approaches to selecting time series models. *Neurocomputing*, *61*(1–4), 121–137. http://dx.doi.org/10.1016/j.neucom.2004.03.008.

Ramaekers, K., & Janssens, G. K. (2008). On the choice of a demand distribution for inventory management models. *European Journal of Industrial Engineering*, *2*(4), 479–491. http://dx.doi.org/10.1504/EJIE.2008.018441.

Rădăşanu, A. C. (2016). Inventory management, service level and safety stock. *Journal of Public Administration*, *9*(9), 145–153, https://pdfs.semanticscholar.org/98b7/5ecc0cd653b5620180aab79f4024f9726016.pdf.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191. http://dx.doi.org/10.1016/j.ijforecast.2019.07.001.

Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, *32*(2), 243–256. http://dx.doi.org/10.1016/j.ijforecast.2015.08.005.

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2014). Model selection using database characteristics: developing a classification tree for longitudinal incidence data. *Marketing Science*, *33*(2), 188–205. http://dx.doi.org/10.1287/mksc.2013.0825.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. http://dx.doi.org/10.1214/10-STS330.

Steiner, W. J., Brezger, A., & Belitz, C. (2007). Flexible estimation of price response functions using retail scanner data. *Journal of Retailing and Consumer Services*, *14*(6), 383–393. http://dx.doi.org/10.1016/j.jretconser.2007.02.008.

Tanaka, K. (2010). A sales forecasting model for new-released and nonlinear sales trend products. *Expert Systems with Applications*, *37*(11), 7387–7393. http://dx.doi.org/10.1016/j.eswa.2010.04.032.

Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, *178*(1), 154–167. http://dx.doi.org/10.1016/j.ejor.2006.02.006.

Teunter, R. H., Babai, M. Z., & Syntetos, A. A. (2010). ABC classification: Service levels and inventory costs. *Production and Operations Management*, *19*(3), 343–352. http://dx.doi.org/10.1111/j.1937-5956.2009.01098.x.

Therneau, T. M., & Atkinson, E. J. (2018). Rpart: Recursive partitioning and regression trees (version 4.1-13). https://cran.r-project.org/package=rpart.

Thomassey, S., & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems*, *42*(1), 408–421. http://dx.doi.org/10.1016/j.dss.2005.01.008.

Trusov, M., Bodapati, A. V., & Cooper, L. G. (2006). Retailer promotion planning: Improving forecast accuracy and interpretability. *Journal of Interactive Marketing*, *20*(3–4), 71–81. http://dx.doi.org/10.1002/dir.20068.

Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R. (2021). Distributional regression for demand forecasting in e-grocery. *European Journal of Operational Research*, *294*(3), 831–842. http://dx.doi.org/10.1016/j.ejor.2019.11.029.

Uniejewski, B., Marcjasz, G., & Weron, R. (2019). Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. *International Journal of Forecasting*, *35*(4), 1533–1547. http://dx.doi.org/10.1016/j.ijforecast.2019.02.001.

Wang, X., Smith-Miles, K., & Hyndman, R. (2009). Neurocomputing rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing*, *72*(10), 2581–2594. http://dx.doi.org/10.1016/j.neucom.2008.10.017.

Weber, A., Steiner, W. J., & Lang, S. (2017). A comparison of semiparametric and heterogeneous store sales models for optimal category pricing. *OR Spectrum*, *39*(2), 403–445. http://dx.doi.org/10.1007/s00291-016-0459-6.

Williams, T. M. (1984). Stock control with sporadic and slow-moving demand. *Journal of the Operational Research Society*, *35*(10), 939–948, https://www.jstor.org/stable/2582137.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. http://dx.doi.org/10.1109/4235.585893.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, *160*(2), 501–514. http://dx.doi.org/10.1016/j.ejor.2003.08.037.

Zhang, W., Quan, H., & Srinivasan, D. (2018). Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination. *Energy*, *160*, 810–819. http://dx.doi.org/10.1016/j.energy.2018.07.019.

Zipkin, P. H. (2000). *Foundations of inventory management* (first ed.). Boston: McGraw-Hill.