Contents lists available at ScienceDirect

International Journal of Forecasting

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

Personalized choice model for forecasting demand under pricing scenarios with observational data—The case of attended home delivery

Özden Gür Ali [a,*], Pedro Amorim [b]

[a] Koc University, College of Administrative Sciences and Economics, Rumeli Feneri Yolu, Sariyer, 34450, Istanbul, Turkey
[b] INESC TEC, Faculty of Engineering of University of Porto, Rua Dr Roberto Frias s/n, 4200-465 Porto, Portugal

ARTICLE INFO

Keywords:
Personalization
Choice model
Orthogonalization
Causal effect
Price sensitivity
Substitution pattern
Attended home delivery
Regularization
Consideration set
Interpretable model

ABSTRACT

Discrete choice models can forecast market shares and individual choice probabilities with different price and alternative set scenarios. This work introduces a method to personalize choice models involving causal variables, such as price, using rich observational data. The model provides interpretable customer- and context-specific preferences, and price sensitivity, with an estimation procedure that uses orthogonalization. We caution against the naïve use of regularization to deal with the high-dimensional observational data challenge. We experiment with the attended home delivery (AHD) slot choice problem using data from a European online retailer. Our results indicate that while the popular non-personalized multinomial logit (MNL) model does very well at the aggregate (day–slot) level, personalization provides significantly and substantially more accurate predictions at the individual–context level. But the "naïve" personalization approach using regularization without orthogonalization wrongly predicts that the choice probability will *increase* if the slot price increases, rendering it unfit for forecasting demand with pricing scenarios. The proposed method avoids this problem. Further, we introduce features based on potential consideration sets in the AHD slot choice context that increase accuracy and allow for more realistic substitution patterns than the proportional substitution implied by MNL.

© 2023 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

Discrete choice models forecast demand when a decision maker faces a choice among a finite set of alternatives (Goodwin, Meeran, & Dyussekeneva, 2014; Wan, Zhang, & Wang, 2014). They can be used to forecast demand at the aggregate and segment levels, e.g., to foresee the effect of a new pricing policy on the market share of different alternatives (Allenby, 2017). With the emergence of digital channels, there is an opportunity to perform forecasts at a customer and contextual level, and to optimize related decisions, e.g., by offering a personalized assortment or discount to steer customer choices.

An example of this digitalization trend is online retailing, which increased significantly during the Covid-19 pandemic, bringing rich data sources—including customer purchasing and choice behavior tracking—and providing an opportunity for more accurate predictions about distinct customer preferences (Fildes, Kolassa, & Ma, 2021). While traditional, non-personalized, multinomial logit (MNL) choice models help predict the aggregate demand, average preferences, and price elasticities, a personalized choice model can provide distinct choice probability and price sensitivity at the individual–context level. This granularity in the prediction can be achieved based on information about the customer's historical behavior and

[*] Corresponding author.
E-mail addresses: oali@ku.edu.tr (O. Gur Ali), pamorim@fe.up.pt (P. Amorim).

https://doi.org/10.1016/j.ijforecast.2023.04.008

0169-2070/© 2023 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

the context within which the choice is made, such as the shopping basket size or content, feeding more precise decision-making (Chen, Owen, Pixton, & Simchi-Levi, 2022).

Retailers often use attended home delivery (AHD) to fulfill the rising number of online orders, where the customer must be present to receive the goods. In this fulfillment strategy, customers are typically given a choice among several alternative time slots that span multiple days (Amorim, DeHoratius, Eng-Larsson, & Martins, 2020). Retailers need forecasts of delivery volume by time slot and location to manage the logistics efficiently and effectively. Literature on optimizing these decisions (e.g., Agatz, Campbell, Fleischmann, & Savelsbergh, 2011) frequently uses the MNL model. One of the advantages of the MNL model is its link to random utility theory (McFadden, 2001). However, it also comes with a significant drawback—the independence of irrelevant alternatives (IIA) property, or the proportional substitution assumption—which regularly requires adjustments. Take the example of a customer buying online groceries to cook in the evening on the day of the purchase. This property implies that introducing a new time slot with a lead time of a week has a proportional impact across the time slots of the different days, hence proportionally reducing the utility of slots for same-day delivery (which, in this case, should not be impacted at all).

With a more proactive stance, retailers can go beyond managing logistics and start influencing customer choices in AHD by adjusting the delivery slots' availability and/or pricing. Customers have different preferences regarding, for example, days of the week to receive deliveries and distinct price sensitivities to the several time slot alternatives (Amorim et al., 2020). Moreover, both preferences and sensitivities vary across customers and depending on the context (e.g., with more or less perishables in the shopping basket, customers may change their slot choices and willingness to pay). Equipped with personalized choice models, retailers may control costs while keeping customers satisfied by exploiting the differences in customer preferences and contexts. For example, they can shift demand away from a slot that is close to exhausted by making it unavailable when a customer has a high probability of choosing alternatives with higher available capacity. Alternatively, when shifting the demand to underutilized slots, the retailer can provide a discount to those customers whose choice probabilities for a given slot within the context can be nudged with a small discount. These are essential value levers in a business that struggles with profitability (McKinsey, 2022). However, this and other use cases need a personalized forecast approach that can provide good accuracy for individual choices. This performance implies, at least, that the actual choice is among the top alternatives based on predicted probability.

The development of accurate, personalized models requires gathering data with information about the alternatives, the context, and the customer's characteristics and behavior over time to identify generalizable differences in preferences and sensitivities. Therefore, personalization involves the complications of high-dimensional and frequently sparse data with *potentially* valuable features. Further, a usual difficulty with observational data is biased estimates for the effect of causal variables (e.g., prices) due to confounding. In the AHD context, the price of the alternative time slots is often set based on the cost of delivery or the demand that the retailer expects, e.g., high prices for evening time slots that are highly desirable and costly to deliver in, resulting in a clear correlation between price and demand. The literature offers a set of common approaches to deal with these data challenges: confounding bias for causal variables, and high dimensionality.

For the first challenge, data from controlled experiments with random assignments can be used to correctly estimate the effect of causal factors such as price. Alternatively, with observational data, which is most often the case in practice, a carefully and correctly specified model that includes all confounders may be used (Greene, 2017). For example, controlling for the time slot would show that evening slots are highly preferred given the same price, and increasing the price for the same slot would decrease the choice probability. For the second challenge, a common approach to selecting the relevant features among the many potentially useful ones is L1 norm (lasso) regularization in generalized linear models, due to its favorable statistical properties (Hastie, Tibshirani, Friedman, & Friedman, 2009). Unfortunately, regularization may cause confounding and bias in the parameter estimates (Hahn, Carvalho, Puelz, & He, 2018). Particularly, price sensitivity estimation is notorious, as factors that drive desirability and high demand are typically associated with high prices (as in the example above). The regularization that penalizes complexity often results in a model that reduces or removes the control variables, such as an indicator for the evening time slot, resulting in positive price sensitivity estimates. In retail forecasting with promotions, Gür Ali and Gürlek (2020) show that regularization-induced confounding causes lower forecast accuracy, and misleading price and promotion effects.

These two challenges are thus intertwined and call for an alternative approach to personalized choice models. Therefore, we aim to answer the following three research questions in this work.

1. How can choices be forecasted accurately at the aggregate and individual–context levels?
2. How can we prevent biased price sensitivity estimates when using regularization to deal with high dimensionality?
3. How do we escape the proportional substitution probability implication of the IIA assumption?

In answering these research questions, we offer the following contributions to the literature.

Using an extensive dataset from a European online grocery retailer using AHD, we show that the causal estimates in personalized choice models can be biased to the point of having the *wrong* sign most of the time for the top alternatives. These results render the models unfit for forecasting with counterfactual pricing scenarios. Furthermore, they hold even when the predicted individual choice probabilities are highly accurate and all confounders are included in the training dataset.

We propose a method for estimating a personalized choice model with a causal (price) variable using high-dimensional observational data containing all confounders. Its random utility theory-driven structure and domain knowledge-driven feature group specification support causal estimation and guard against overfitting. L1 norm (lasso) regularization for feature selection is followed by unregularized estimation to achieve unbiased efficient estimates. Orthogonalization concerning confounders before estimating the causal effects avoids regularization-induced confounding. To the best of our knowledge, this is the first use of orthogonalization in choice models to avoid regularization-induced confounding.

We empirically show that orthogonalization is an effective remedy for causal estimates with the wrong sign in the AHD setting. We additionally find that the non-personalized MNL model does a very good job of predicting choice at the aggregate (day–slot) level and provides a meaningful price sensitivity estimate at the population level. While these findings justify its pervasive use in AHD pricing and slotting optimization models, its accuracy in predicting individual-level choices is significantly and substantially lower than that of personalized models. Personalized models predict the top alternative twice as accurately as the non-personalized MNL.

Finally, we propose features relating an alternative to potential consideration sets to allow for more realistic substitution probabilities for the AHD slot choice problem. We empirically show that including these features significantly improves the forecasting accuracy of the proposed personalized choice model. All five replications in our experiments selected the same feature, suggesting that customers consider alternatives from earlier days with a time slot similar to the one they are interested in. The feature based on the popular hypothesis that customers consider time slots within a chosen day was not selected (Yang, Strauss, Currie, & Eglese, 2016).

The rest of the paper is organized as follows. The next section briefly summarizes related literature on the MNL choice model, personalization of choice models, confounding, and choice models in attended home delivery. Section 3 introduces the empirical setting and the proposed consideration-set features in the AHD context. Section 4 describes the proposed method, including the design elements, the model, and the estimation procedure. Next, in Section 5, we describe the performance measures and the benchmark methods used in the experiment. Section 6 provides the results. We conclude, in Section 7, by discussing the implications of the experimental results for personalization in choice models in general and in the AHD setting in particular. We further discuss the limitations of the proposed method.

2. Related literature

Our work is related to several lines of research. First, we start by reviewing the most commonly used choice model, i.e., the MNL. Second, we summarize the work regarding the personalization of choice models. Third, we deep-dive into the confounding issue and approaches to tackle it. Finally, because our work is instantiated in AHD, we review choice models in this retail context.

2.1. MNL choice model

Discrete choice models explain and predict the choice probability among discrete alternatives. As we discuss below, their connection to behavioral theory and interpretability makes them valuable compared to black-box models. Random utility theory suggests that individuals select the alternative that provides the highest utility (Tversky, 1969). Representing the utility of each option by a systematic and a random component, McFadden (1973) showed that assuming that the random component of alternative utility is i.i.d. and Gumbell-distributed implies a logit formula for choice probabilities. One of the critical assumptions for the resulting MNL is that the unobserved factors are uncorrelated over alternatives and have the same covariance. This assumption implies proportional substitution across alternatives, the so-called IIA property. In the context of assortment optimization, Kök and Fisher (2007) point out that the MNL model does not allow two alternatives to have the same market share but different substitution rates. In this model, it is possible to control the substitution rate, but that directly determines the initial market share and limits its applicability.

Nested logit models relax the IIA assumption by assuming a nesting structure where the alternatives have some correlation within the nest and no correlation across nests. These models use a generalized extreme value distribution for the random component (Train, 2009). The choice probability here can be interpreted as the product of the probability of choosing the nest and the probability of selecting the alternative given the nest. A disadvantage of these models is that the nested structure is fixed across individuals and defined *a priori*.

In the transportation route choice context, where overlapping road segments among the alternative routes make the proportional substitution unrealistic, the deterministic part of the utility function is extended with a commonality factor to correct the MNL model (Prato, 2009). The C-logit, path-size logit (Ben-Akiva & Bierlaire, 1999), and path size correction logit models use different functional forms to express the commonality/overlap factor that measures the degree of similarity of each route with other routes in the set of alternatives.

2.2. Personalization of choice models

Digitalizing every aspect of life has made data more readily available for analytics (Hübner et al., 2021). The online channel is particularly rich in data about the choice situation, including customer information and context. In this environment, personalized choice models may provide customer- and context-specific choice probabilities by modeling heterogeneity in the systematic component as a function of the observable characteristics. For example, Feldman, Zhang, Liu, and Zhang (2022) use personalized choice models for assortment optimization with Alibaba marketplace data, creating static and dynamic product features, such as category, seller IDs, number of reviews, and price; and customer features, including demographics and past behavior, e.g., number of products viewed, collected, purchased and returned; as well as

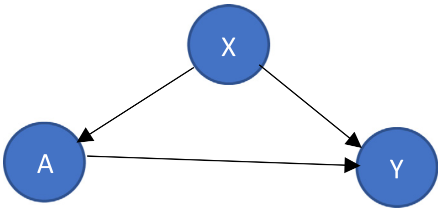


Fig. 1. X is a confounder variable, since it affects the outcome Y and the causal variable A.

joint features linking customers and products. Chen et al. (2022) provide finite sample convergence properties for fitting a personalized MNL model with a lasso penalty to select features in a high-dimensional setting in the context of assortment and pricing personalization. The algorithm assumes that prices are not a function of observed features (i.e., no personalized pricing in the data), and it was applied in a case with experimental data where prices were assigned randomly to customers.

These papers illustrate the challenges and advantages of the personalization of choice models. On the one hand, personalization allows for improved business metrics, such as click-through, conversion rate, and profitability. On the other hand, it requires dealing with high-dimensional data in estimation.

2.3. Confounding, orthogonalization, and machine learning

While controlled random experiments make estimation easier by eliminating selection bias (see, e.g., Chen et al., 2022), causal effects can also be estimated with observational data with non-random prices, provided that the model is specified correctly and contains all confounders (Greene, 2017). Confounding variables affect both the causal variables and the outcome, as seen in Fig. 1. As in ordinary regression, a model that does not control for confounding variables results in biased estimates for the coefficient of the causal variable in logistic regression (Hosmer Jr, Lemeshow, & Sturdivant, 2013). Suppose that historically the price is set high for the desired scarce products or time slots to discourage their demand. In that case, a choice model will provide a positive coefficient for the price, but controlling for the characteristics that drive price and preferences will give the *true* effect of price while keeping these characteristics constant.

When personalizing a choice model, there is a need to identify the relevant features to include in the model out of the many potential features that describe the decision maker, the alternatives, the context, and their interactions. Lasso regularization is a frequently used feature selection method for generalized linear models, like logistic regression, that penalizes model complexity (Hastie et al., 2009). Regularization-induced confounding bias occurs when the confounding features are eliminated or their coefficients are inappropriately shrunk due to regularization, preventing appropriate control (Hahn et al., 2018).

Assuming that there are no unobserved confounders, orthogonalization of the causal variable and the outcome with respect to confounders can be used as a remedy for regularization-induced confounding. In forecasting retail sales with inter- and intra-category promotion effects and high-dimensional input data, Gür Ali and Gürlek (2020) show that orthogonalization before elastic net regression significantly increases the forecasting accuracy performance. The Frisch–Waugh–Lovell theorem (Lovell, 1963) states that for ordinary least squares regression (OLS), the same regression coefficient will be obtained for a x_1 by regressing y on x_1 and x_2, as regressing Δy on Δx$_1$, where Δy and Δx$_1$ are the residuals of regressing y on $x_{2,}$ and x_1 on x_2, respectively. Regressing y on x_2, and x_1 on x_2, are referred to as orthogonalization with respect to $x_{2.}$ We can think of x_1 as the causal variable, and x_2 as the confounder. Robinson (1988) replaces the orthogonalizing linear regressions with a non-parametric regression and shows the root n-consistency of the causal parameter estimate. Chernozhukov et al. (2018) provide the root n-consistent estimator with confidence intervals of the causal parameter using machine learning estimators, and coin the term "double machine learning". They use split sample estimation, which refers to splitting the training dataset into folds for orthogonalization and causal effect estimation. Athey and Imbens (2016) show that sample splitting for heterogeneous treatment effect estimation with machine learning increases the coverage of the confidence intervals but reduces the accuracy, due to a smaller sample size for estimation.

Machine learning methods are handy in the presence of high-dimensional data, as they automatically select features and model non-linear and arbitrary interaction effects. They frequently provide higher accuracy in predicting choices compared to MNL but often lead to worse results when used to optimize the offered alternatives. The study mentioned above by Feldman et al. (2022) found that the MNL-based approach generates significantly higher revenue per visit than a machine learning algorithm even though its predictions are less accurate. On predicting travel choices, Zhao, Yan, Yu, and Van Hentenryck (2020) note that “the random forest model produces behaviorally unreasonable arc elasticities and marginal effects”, even though it has significantly better prediction accuracy than MNL models. Wang, Wang, and Zhao (2020) investigate to what extent deep neural nets (DNNs) can be relied upon to provide economic information in choice models, such as substitution patterns and elasticities. They find that DNNs can be unreliable when the sample size is small, due to sensitivity to hyperparameters and model complexity, and they suggest using aggregate data for robust performance. Surveying efforts to combine machine learning with choice models, van Cranenburgh, Wang, Vij, Pereira, and Walker (2022) find that theory-driven choice models are still superior in economically meaningful outputs, in line with the findings above. Establishing causality versus correlation requires domain knowledge beyond data (Athey, 2017). Hence, while the choice modeling community recognizes machine learning tools such as regularization and cross-validation for model selection, the interpretability of choice models remains of paramount importance (Aboutaleb, Danaf, Xie, & Ben-Akiva, 2021).

2.4. Choice models in attended home delivery

When managing attended home delivery logistics, retailers have two important levers influencing customer choices: pricing and available time slots. Models supporting these decisions require predictions of customer choice probabilities when different sets of time slots are offered with different price scenarios.

In the literature on dynamic time slot pricing and slotting for attended home delivery, many papers model customer choice as an MNL (Asdemir, Jacob, & Krishnan, 2009; Lang, Cleophas, & Ehmke, 2021; Strauss, Gülpınar, & Zheng, 2021; Yang & Strauss, 2017; Yang et al., 2016). Mackert (2019) uses the generalized attraction model proposed by Gallego, Ratliff, and Shebalov (2015) to express the customer's context-dependent dissatisfaction with the absence of each missing alternative in the choice set. Koch and Klein (2020) note issues with the IIA property and the inability of MNL models to deal with different customer segments and overlapping time windows. Therefore, they use the more flexible reservation price model, where the estimates can come from different choice models.

With a focus on inference, Amorim et al. (2020) explore the drivers of customers' choices for delivery slots with conditional MNL and mixed logit models and find that customers value speed, precision, and timing. They also find substantial heterogeneity in preferences, including more price-insensitive customer segments.

To illustrate the implications that MNL properties yield, consider the introduction of an alternative slot in a distant horizon (e.g., one week from the purchasing day). In this case, the predicted choice probability for a given preferred slot decreases by the same ratio for slots on the same day as for the newly introduced slot. However, one would expect the choice probability of the preferred slot not to be affected as much by the new alternative as by slots on the same day.

3. Empirical setting

Before introducing the proposed method, let us describe the empirical setting that motivated our research and which is used in our empirical evaluation.

We work with the delivery slot choice data from a large European grocery chain, where the customers can shop online and choose an attended delivery slot to receive their groceries after finalizing their shopping basket. For every order, a customer can choose from a maximum of 53 possible delivery slots: four slots on the day of the order and seven slots per day on the following seven days. The customer is offered a set of these alternatives with associated delivery fees (prices). This set depends on the customer's address, the number of delivery slots that have their capacity exhausted, and the cut-off time of the delivery slots for same-day delivery. More importantly, for each transaction, it is possible to identify the set of alternatives, from which the customer picks one option.

Besides prices, customers observe several other descriptors of the alternatives: the slot width, the time of the day, the day of delivery, and, consequently, the days to delivery. On the other hand, the retailer may observe current and historical customer characteristics based on past orders and infer shopping basket attributes and descriptors of activity, such as recency, frequency, spending, and previous slot choices.

In the time frame of our dataset (October 2016 to September 2017), the retailer did not engage in dynamic slotting or pricing. That is, the slots and the respective prices were not adjusted continuously, nor were they personalized. Hence, customers purchasing from similar coordinates around the same time would have the same alternatives available. From a tactical perspective, the delivery cost is considered when determining the (static) delivery fees. After finalizing the shopping basket, customer abandonment of the cart at the slot choice stage is reported to be negligible.

The retailer may adjust slot availability and/or pricing for each delivery in this empirical setting to steer customers to more profitable alternatives. The optimization would require demand forecasts for each slot and the predicted choice probabilities under different slot assortment and pricing scenarios for each delivery based on individual and context characteristics. However, several challenges must be tackled for a personalized choice model that can provide these forecasts/predictions. Firstly, the related dataset has high dimensionality. Secondly, the observational nature of the dataset may bias the estimates of the causal effect of price on choice. Thirdly, applying standard choice models (e.g., MNL), even if they predict individual preferences well, would result in counterintuitive substitution patterns (cf. the IIA property).

Next, we propose a model that could be applied in this setting to predict individual-level choices for different scenarios.

4. Method

Our proposed method combines data-driven and theory-driven approaches to achieve high accuracy for choice prediction and meaningful causal estimates (price sensitivity and substitution behavior among alternatives). The method uses typical high-dimensional data available in e-commerce applications. These data include static and dynamic descriptors of the customer behavior, the context, and the alternative, including the price (that we assume, without loss of generality, as the causal variable), to model the utility of the alternative to the customer within the specific context. We assume that the data include all confounders.

The method addresses the three research questions introduced in Section 1: first, forecasting accurately at the aggregate and individual–context levels; second, preventing biased causal estimates; and third, correcting for the proportional substitution probability implication of the IIA property.

4.1. Design elements

Our method relies on the following four design elements:

Theory-driven model structure and domain knowledge-driven feature set specification: Our personalized choice model has two kinds of personalization: individual–context-level preferences and individual–context-level price sensitivities. We use the behavioral choice theory framework and express the systematic component of the utility as a function of the observed characteristics and the causal variable. We incorporate domain knowledge about causal relationships by specifying appropriate groups of features as main effects and two-way interaction terms in our choice model to ensure interpretability and reduce its capacity to overfit.

Feature selection and unbiased estimation: We use L1 norm (lasso) regularization in the logit model to select features from the set of pre-specified explanatory variables (Lokhorst, 1999; Park & Hastie, 2007), determining the regularization parameter based on the cross-validation results within the training data (Hastie et al., 2009). The coefficient estimates of the selected terms are determined with unregularized models to achieve lower bias and good convergence (Belloni & Chernozhukov, 2013; Meinshausen, 2007).

*Orthogonalization with respect to confounders***:** To prevent regularization-induced bias of the causal estimates, we orthogonalize the outcome and the causal variable with respect to potential confounding variables before estimating the causal effects.

Features allowing more realistic substitution probabilities: We introduce features that describe the alternative relative to potential consideration sets, to model more realistic substitution probabilities. These features enable alternatives within the same consideration set to have a higher substitution probability than one that is proportional to their initial choice probability. This design element is inspired by: (i) the empirical finding that consideration sets are simply indicators of preferences and do not imply a two-step process where consideration precedes choice (Horowitz & Louviere, 1995), and (ii) models in the context of route choice that use a commonality feature with other alternatives in the choice set (Prato, 2009). As with additional features, the lasso regularization identifies which, if any, of the hypothesized consideration sets improve the model given the historical choices.

4.2. The model

We consider the choice situation s, where an individual chooses among alternatives $j \in \mathbb{C}_s$. The number of alternatives $|\mathbb{C}_s|$ can vary by situation, but there is a fixed set of alternatives: $\mathbb{C}_s \subset \{1, \ldots, J\}$. Let us then define the decision y_{sj}:

$$y_{sj} = \begin{cases} 1 \; if \;\; alternative \; j \; is \; chosen \; for \; situation \; s \\ 0 \; otherwise \end{cases} \text{ and}$$

$$\sum_{k \in \mathbb{C}_s} y_{sk} = 1.$$

and $\sum_{k \in \mathbb{C}_s} y_{sk} = 1$.

Each choice situation s is characterized by a set of alternatives $\mathbb{C}_s$, descriptors of the alternatives X_{sj} and the decision y_{sj} for all $j \in \mathbb{C}_s$. X_{sj} includes descriptors of the alternative j, X_j^{alt}; the causal variable value (price), p_{sj}; summaries of the individual's historical choices, X_{sj}^{beh}, other descriptors of the individual and the context, X_s^{cont}; and consideration-set features, X_{sj}^{Cset}. See Table 1 for the notation.

Our goal is to model U_{sj}, the utility of alternative j in choice situation s. To that end, it is relevant first to introduce the row vector Z_{sj}, which includes all determinants of U_{sj}, excluding the price: X_j^{alt}, X_{sj}^{beh}, X_{sj}^{Cset}, and the multiplicative interaction of each descriptor in X_s^{cont} with each alternative characteristic in X_j^{alt}. The interaction of the context and alternative characteristics personalizes the preferences for the context, for example by expressing higher attractiveness for shorter delivery lead time alternatives when the basket contains fresh goods. X_{sj}^{beh} personalizes for the individual, for example by expressing higher attractiveness for slots that this customer has chosen before. The model of U_{sj} is then as follows:

$$U_{sj} = f_1\left(Z_{sj}\right) + f_3\left(\Delta p_{sj} W_{sj}\right) + \varepsilon_{sj} \tag{1}$$

$$\Delta p_{sj} = p_{sj} - \mathrm{E}\left[p_{sj} | X_{sj}\right] = p_{sj} - \mathrm{E}\left[p_{sj} | V_{sj}\right] = p_{sj} - f_2\left(V_{sj}\right) \tag{2}$$

$$\mathrm{E}\left[\Delta p_{sj}\right] = 0 \tag{3}$$

$$\mathrm{E}\left[\varepsilon_{sj}\right] = 0 \tag{4}$$

The first term in Eq. (1), $f_1\left(Z_{sj}\right)$, represents the utility of alternative j in situation s, at the expected price given in X_{sj}. The second additive term in Eq. (1) expresses the utility component due to the deviation from the expected price given in X_{sj}, which is denoted by Δp_{sj}. This term, $f_3\left(\Delta p_{sj} W_{sj}\right)$, models the personalized price sensitivity. For example, price sensitivity can differ by basket content, the day of delivery, or the customer's favorite time slot. We model the heterogeneity of price sensitivity similar to Gür Ali (2013) by forming interaction effects of the causal variable with many potential moderators and relying on regularization to select the appropriate terms. W_{sj} is a row vector that contains a subset of the features in X_{sj}, which are potential moderators of price sensitivity. ε_{sj} represents the random component of the utility, independently and identically distributed with mean 0.

In Eq. (2), the row vector V_{sj} contains the confounders, which are determinants of the prices observed in the training data and the utility of the alternative. For example, as explained above, the time of the delivery slot affects both the price and utility of the alternative. We do not include determinants of the causal variable that do not affect the outcome, as they would increase the estimator's variance (Cinelli, Forney, & Pearl, 2021).

Note that in contrast to choice models that set the price coefficient to -1 and estimate the effect of other factors relative to price (e.g., Danaf, Becker, Song, Atasoy, & Ben-Akiva, 2019), this specification can discover situations where changing the price will be most influential on the choice probability as well as cases where the customer is not sensitive to price changes within the observed range.

While the functions f_1, f_2, and f_3 can be more flexible machine learning algorithms, such as boosting or splines, as in generalized additive models, for this paper, we specify them as linear additive functions to allow for a direct

Table 1
List of variables and parameters.

Term	Definition
$\mathbb{C}_s$	Set of alternatives for the choice situation s
J	Number of all possible alternatives
y_{sj}	Binary variable indicating whether alternative j is chosen in choice situation s
X_{sj}	Vector of all descriptors of alternative j in choice situation s
X_j^{alt}	Vector of static descriptors the alternative j
p_{sj}	Price of alternative j in choice situation s
X_{sj}^{beh}	Vector of descriptors of the individual's historical choices regarding alternative j in choice situation s
X_s^{cont}	Vector of descriptors of the individual and the context in choice situation s
X_{sj}^{Cset}	Vector of consideration-set features for alternative j in choice situation s
U_{sj}	Utility of alternative j in choice situation s
Z_{sj}	Row vector with X_j^{alt}, X_{sj}^{beh}, X_{sj}^{Cset}, and the multiplicative interaction of each descriptor in X_s^{cont} with each alternative characteristic in X_j^{alt}
ε_{sj}	Random component of the utility of alternative j in choice situation s
Δp_{sj}	Deviation from the expected price given X_{sj}
V_{sj}	Vector containing the confounders, which is a subset of the features in X_{sj}.
W_{sj}	Vector containing the potential moderators of price sensitivity for alternative j in choice situation s and the identity (1)
$\boldsymbol{\beta}_1$	Coefficient vector for f_1 that yields the expected utility of alternative j in situation s at the expected price, given X_{sj}
$\boldsymbol{\beta}_2$	Coefficient vector for f_2 that yields the expected price, given V_{sj}
$\boldsymbol{\theta}$	Coefficient vector for f_3 that yields the expected personalized price sensitivity contribution to the expected utility for alternative j in choice situation s, given $\Delta p_{sj} W_{sj}$
λ_1, λ_2, λ_3	The regularization constants for the L1 norm (lasso) penalty in the estimation of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\theta}$ coefficient vectors, respectively
M_1, M_2, M_3	Set of indices corresponding to selected features after the lasso estimation of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\theta}$ coefficient vectors, respectively
α_j	Alternative specific constant for alternative j in the B0A and B0B MNL models
β_{price}	Coefficient for price in the B0A MNL model
$\beta_{price,j}$	Alternative specific coefficient for price in the B0B MNL model

performance comparison with the non-personalized MNL:

$$f_1\left(Z_{sj}\right) = Z_{sj}\boldsymbol{\beta}_1, \qquad f_2\left(V_{sj}\right) = V_{sj}\boldsymbol{\beta}_2, \qquad f_3\left(\Delta p_{sj} W_{sj}\right) = \Delta p_{sj} W_{sj}\boldsymbol{\theta} \tag{5}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\theta}$ are parameter column vectors with length dimensions that match Z_{sj}, V_{sj}, and W_{sj}, respectively.

The probability that alternative j is chosen among the alternatives in the choice set $\mathbb{C}_s$ is given by Eq. (6), as in the random utility framework:

$$P\left(y_{sj} = 1\right) = exp\left(U_{sj}\right) \Big/ \sum_{k\in\mathbb{C}_s} exp\left(U_{sk}\right) \tag{6}$$

4.3. Estimation procedure

First, we estimate the expected price given the confounders and the utility of alternatives at the expected price. Then, we estimate the heterogeneous price sensitivity to deviations from the expected price using these estimates as a given.

Our estimation process consists of three steps: the first two steps estimate the functions f_1 and f_2 using first a (random) partition of the training data; and given these estimates, f_3 is calculated with the second partition.

An intuitive justification of the procedure is as follows. As we have seen in the discussion on confounding, estimated price coefficients can be biased (positive) when the preference drivers are not adequately accounted for. This can happen, for example, when a desirable but costly attribute (such as evening delivery), increases both the demand and the price. In this example, when regularization eliminates variables that increase both the price and the demand for the alternative and uses the price variable to express the association with a simpler model, we have regularization-induced confounding. Therefore, this estimation procedure with orthogonalization makes sure that, first, the preferences at expected prices are explained with a parsimonious model with f_1, and then the effect of deviation from the expected prices is estimated with f_3.

Each step has two parts. First, a lasso logistic regression selects terms to be included in the model. Then, a second unregularized logistic regression estimates the parameters of the chosen terms to achieve lower bias and good convergence (Belloni & Chernozhukov, 2013). We determine the regularization parameter lambda based on cross-validation results within the same partition.

Step 1. Estimate f_1:

$$\widetilde{\beta_1}\left(\lambda_1\right) = \underset{\beta_1}{argmin} - \frac{1}{\boldsymbol{N}} \sum_s \sum_j y_{sj}\left(Z_{sj}\boldsymbol{\beta}_1 - log\left(\sum_{k\in\mathbb{C}_s} Z_{sk}\boldsymbol{\beta}_1\right)\right) + \lambda_1 \left\|\boldsymbol{\beta}_1\right\|_1 \tag{7}$$

$$\hat{\boldsymbol{\beta}}_{1,\boldsymbol{M}1} = \underset{\boldsymbol{\beta}_{1,\boldsymbol{M}1}}{argmin} - \frac{1}{\boldsymbol{N}} \sum_{s} \sum_{j} y_{sj} \left(\sum_{l\epsilon M_1} Z_{sj[l]} \boldsymbol{\beta}_{1[\boldsymbol{l}]} - log \left(\sum_{k \in \mathbb{C}_s} \sum_{l\epsilon M_1} Z_{sk[l]} \boldsymbol{\beta}_{1[\boldsymbol{l}]} \right) \right) \tag{8}$$

where N is the number of observations. The lasso multinomial logistic regression selects features by minimizing the sum of the negative log-likelihood (Train, 2009) and the L1 norm of $\boldsymbol{\beta}_1$ weighted by the penalty λ_1, which are the first and second terms in Eq. (7), respectively. The coefficients of selected features are estimated with the non-regularized model, as in Eq. (8), where $\hat{\boldsymbol{\beta}}_{1,\boldsymbol{M}1}$ refers to coefficients with indices in the set M_1. $M_1 = \left\{ l : \widetilde{\beta}_{1[l]} \neq 0 \right\}$. The subscript $[l]$ refers to the l^{th} element of the vector. The coefficients of unselected features are set to 0.

Step 2. Estimate f_2:

$$\widetilde{\beta_2} (\lambda_2) = \underset{\boldsymbol{\beta}_2}{argmin} - \frac{1}{\boldsymbol{N}} \sum_{s} \sum_{j} \left(p_{sj} - V_{sj} \boldsymbol{\beta}_2 \right)^2 + \lambda_2 \left\| \boldsymbol{\beta}_2 \right\|_1 \tag{9}$$

$$\hat{\boldsymbol{\beta}}_{2,\boldsymbol{M}2} = \underset{\boldsymbol{\beta}_{2,\boldsymbol{M}2}}{argmin} - \frac{1}{\boldsymbol{N}} \sum_{s} \sum_{j} \sum_{l\epsilon M_2} \left(p_{sj} - V_{sj[l]} \boldsymbol{\beta}_{2[\boldsymbol{l}]} \right)^2 \tag{10}$$

The set $M_2 = \left\{ l : \widetilde{\boldsymbol{\beta}}_{\boldsymbol{2}[\boldsymbol{l}]} \neq 0 \right\}$ contains the indices of features selected by the lasso regressing observed prices p_{sj} against V_{sj}. The coefficients of unselected features are set to 0. The subsequent OLS step (i.e., Eq. (10)) with selected features ensures $E\left[\Delta p_{sj} \right] = 0$.

Step 3. Estimate f_3:

$$\tilde{\boldsymbol{\theta}} (\lambda_3) = \underset{\boldsymbol{\theta}}{argmin} - \frac{1}{\boldsymbol{N}} \sum_{s} \sum_{j} y_{sj} \left(Z_{sj} \hat{\boldsymbol{\beta}}_1 + \left(p_{sj} - V_{sj} \hat{\boldsymbol{\beta}}_2 \right) W_{sj} \boldsymbol{\theta} - log \left(\sum_{k \in \mathbb{C}_s} Z_{sk} \hat{\boldsymbol{\beta}}_1 + \left(p_{sj} - V_{sj} \hat{\boldsymbol{\beta}}_2 \right) W_{sj} \boldsymbol{\theta} \right) \right) + \lambda_3 \left\| \boldsymbol{\theta} \right\|_1 \tag{11}$$

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{M}3} = \underset{\boldsymbol{\theta}_{\boldsymbol{M}3}}{argmin} - \frac{1}{\boldsymbol{N}} \sum_{s} \sum_{j} y_{sj} \left(Z_{sj} \hat{\boldsymbol{\beta}}_1 + \sum_{l\epsilon M3} \left(p_{sj} - V_{sj} \hat{\boldsymbol{\beta}}_2 \right) W_{sj[l]} \boldsymbol{\theta}_{[\boldsymbol{l}]} - log \left(\sum_{k \in \mathbb{C}_s} Z_{sk} \hat{\boldsymbol{\beta}}_1 + \sum_{l\epsilon M3} \left(p_{sj} - V_{sk} \hat{\boldsymbol{\beta}}_2 \right) W_{sk[l]} \boldsymbol{\theta}_{[\boldsymbol{l}]} \right) \right) \tag{12}$$

$\hat{\boldsymbol{\theta}}$ is estimated by minimizing the negative log-likelihood of the multinomial logit model, where the utilities are defined as in Eq. (1), substituting the functions in Eq. (5), given coefficient estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ from Step 1 and Step 2. $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}3}$ refers to coefficients of the selected features, where $M_3 = \left\{ l : \tilde{\boldsymbol{\theta}}_{[\boldsymbol{l}]} \neq 0 \right\}$. The coefficients of unselected features are set to 0.

4.4. Consideration-set features

The problem we are trying to address with the consideration-set features is the IIA issue, independent of the personalization of choice probabilities or price sensitivity estimates. As explained above, the IIA property of the MNL model requires that the ratio of the choice probabilities of two alternatives remains constant for any choice set. The nested logit model addresses this problem by imposing a nesting structure based on theoretical knowledge, such that the IIA assumption holds within each non-overlapping nest but not between different nests. However, based on theory or domain knowledge, we do not always know which nesting structure is appropriate. Further, the decision-making process may involve multiple overlapping consideration sets, for example when the set of slots in the favorite delivery days and favorite delivery hours have some slots that overlap. The proposed approach directly adjusts the utility of the alternative for the number of relevant alternatives in the choice set: as the number of relevant alternatives in the choice set increases, the choice probability for the item should decrease beyond what is implied by the proportional substitution assumption. A consideration set is a subset of alternatives that the decision maker screens based on their criteria, which may involve constraints such as awareness or prior experience, attitudes, and perceptions (Shocker, Ben-Akiva, Boccara, & Nedungadi, 1991). For example, a customer's work schedule in the AHD case may limit the feasible slots for him or her. Horowitz and Louviere (1995) posit that the utility of each alternative in the individual's consideration set is greater than the utility of every alternative not in this set. Consideration sets are relevant alternative sets. Hence, with our approach, the estimation procedure identifies the best consideration-set features to adjust the utility for the number of relevant alternatives in the choice set.

5. The experimental setup

Our objective with the experiments is to evaluate the proposed method and the contribution of its main design components, namely personalization (which includes a theory-driven model structure and domain knowledge-driven feature set specification, as well as feature selection and unbiased estimation), orthogonalization, and consideration-set features. Below, we detail the experiment design and the features that are used. Next, we provide the performance measures for aggregate (day–slot)-level accuracy, individual-level accuracy, and meaningful causal (price sensitivity) estimates. Finally, we introduce the benchmark methods.

5.1. Experimental design and features

We use randomly chosen customer choice situations from our dataset for the experiments. We use the orders in the last two months (August and September 2017) as holdout (test) data. There are 871 customers in the overall dataset. The mean number of cases for a customer in the dataset is 15.

For implementing five replications, we generate 10 datasets by randomly drawing orders without replacement. As seen in Table 2, the number of available alternatives in a choice situation varies between seven (the minimum number of alternatives) and 51 (the maximum number of alternatives) out of a possible 53 alternatives, with a mean of around 41 (the mean number of

Table 2
Dataset statistics for each sample.

Replication	Number of cases	Mean number of alternatives	Minimum number of alternatives	Maximum number of alternatives	Mean price (euros)	Standard deviation of price (euros)
1	1018	41.22	9	49	6.33	0.59
2	1079	41.55	11	51	6.34	0.62
3	1118	41.56	7	51	6.35	0.61
4	1073	41.50	7	51	6.34	0.61
5	1129	41.31	9	51	6.34	0.61
Holdout	2204	41.19	10	51	6.65	0.78

Table 3
Variable statistics for training and holdout data, excluding indicators for slots, day of week, and alternatives.

Variable description		Training data		Holdout data	
		Mean	Std	Mean	Std
X_j^{alt}	Days between purchase and delivery	3.98	2.03	3.99	2.03
	Indicator for weekend delivery	0.28	0.45	0.28	0.45
	Minutes in the slot window (slot_width)	144.73	11.42	144.56	11.55
X_{sj}^{beh}	Indicator for first online purchase	0.06	0.24	0.03	0.17
	% of customer's previous choices for this slot	0.02	0.08	0.02	0.07
	% of customer's previous choices of similar slots	0.06	0.13	0.06	0.11
	% of customer's previous choices of day of week	0.13	0.18	0.14	0.15
p_{sj}	Price of delivery	6.34	0.61	6.65	0.78
X_s^{cont}	Discount percentage on the purchases	18.34	10.97	18.88	10.46
	Fresh produce percentage of the basket	11.14	10.05	10.78	9.99
	Days since first purchase	131.47	88.78	267.62	104.75
	Customer's # of purchases	13.98	13.76	26.34	21.08
	Average days between customer's purchases	10.53	10.01	14.71	16.1
	Average amount purchased	105.18	66.1	105.29	64.52
X_{sj}^{Cset}	# of alternatives on the same slot and earlier day	2.96	2.03	2.94	2.03
	# of alternatives in the same day as this alternative	5.91	0.67	6.24	0.92

alternatives). Due to the many alternatives, in Steps 1 and 3, we approximate the conditional logit model with an unconditional logit with fixed effects for each case (Katz, 2001). The mean price is around 6.34 euros (mean price) with a standard deviation of 0.61 euros (standard deviation of price) in the training data, while the holdout period has slightly higher prices and variability, with a mean of 6.65 euros and a standard deviation of 0.78 euros.

Descriptors of the alternatives, X_j^{alt}, include 25 features coding the slot time on the day of delivery and days to delivery, as well as weekend delivery and slot width. We denote the alternative dummies as X_j^A. Since the relative prices of alternatives depend on alternative characteristics, whereas the context affects all prices and no customer-specific pricing is practiced, $V_{sj} = [X_j^A, X_j^{alt}]$. X_{sj}^{beh} contains four features summarizing the customers' historical frequency of choosing the alternative, the day of the week, and days to delivery.

Other descriptors of the individual and the context, X_s^{cont}, include six cart content features and customer RFM (recency, frequency, monetary value)-type features. X_{sj}^{Cset} includes two consideration-set features based on the literature, as explained below. Table 3 provides an overview of the statistics of the main variables for training and holdout data.

First, based on the common assumption of AHD optimization models that customers first choose a day based on their needs and then select a slot on that day (see, e.g., Yang et al., 2016), we define *n_same_day* as the number of offered alternatives on the same day of the focal alternative. Second, based on the hypothesis that each customer has a favorite slot in the day and a preference for the delivery to arrive sooner rather than later, we define *n_sooner_same_slot* as the number of offered alternatives in the same slot (time of the day) but on an earlier day. If selected, we would expect these features to have a negative coefficient. Table 4 illustrates these features in an example with four possible daily delivery slots over five days. In this example, the customer is offered nine alternatives. For slot C3, *n_same_day* is 3 (because there are three slots on that day), and *n_sooner_same_slot* is 2 (because there are two similar slots on previous days). Assuming that these features are selected in the model and have the expected negative sign, if the alternative C2 is not offered, n_*sooner_same_slot* for C3 would decrease from 3 to 2, and its choice probability among the reduced set of alternatives would be higher than what is prescribed under the proportional substitution property (IIA). Removal of C2 would also affect the n_same_day feature for A2, decreasing it from 2 to 1, and similarly increasing its choice probability disproportionately.

Finally, as heterogeneity dimensions of price sensitivity, we use the alternative characteristics, the context, the customer's past choice behavior, and the choice-set characteristics, as well as the average price of the alternatives ($AvgPrice_s$). We include the average price of the alternatives as a context feature, to capture whether the customer is more (or less) price sensitive when the

Table 4
Illustration of choice-set features. The first numbers in the unshaded cells provide (n_same_day/n_sooner_same_slot) feature values, where the shaded cells indicate unavailable alternatives.

(n_same_day/ n_sooner_day_same_slot)

	Days to delivery	**0**	**1**	**2**	**3**	**4**
Slots	**A**	▨	▨	2/0	3/1	▨
	B	▨	2/0	▨	▨	1/1
	C	1/0	▨	2/1	3/2	▨
	D	▨	2/0	▨	3/1	▨

overall price level is high. $W_{sj} = \left[1, X_j^{alt}, X_s^{cont}, X_{sj}^{beh}, X_{sj}^{Cset}, AvgPrice_s\right]$.

5.2. Performance measures

Aggregate-level accuracy is important to forecast the demand in different day–slot combinations and to help with more tactical time slot management decisions (e.g., fleet sizing). We measure the aggregate-level accuracy based on the choices made in the holdout period T_h (8/1/2017 to 9/30/2017) by delivery day and slot, for days in T_d (8/7/2017 to 9/30/2017). This split avoids the ramp-up and tail periods, as the demand for *DeliveryDate* d and slot t can come from orders placed on the same day or during the previous seven days. This accuracy is formalized as follows:

$$RMSE = \sqrt{\sum_{d \in T_d} \sum_{t} \left(Actual_{dt} - Predicted_{dt}\right)^2}$$

$Actual_{dt} = \sum_{s:\, OrderDate(s) \in T_h} \sum_j y_{sj} \mathrm{I}(DeliveryDate\,(sj) = d) \mathrm{I}\,(slot\,(sj) = t)$, and

$Predicted_{dt} = \sum_{s:\, OrderDate(s) \in T_h} \sum_j \hat{P}\left(y_{sj} = 1\right) \mathrm{I}\,(DeliveryDate(sj) = d)\,\mathrm{I}\,(slot(sj) = t)$.

Here, $I(condition)$ takes value of 1 if the condition holds, and 0 otherwise. We use the root mean square error (RMSE) for aggregate-level accuracy and provide the mean absolute error (MAE) for robustness checks.

Individual-level accuracy is essential if the retailer wants to influence choices with individualized actions (e.g., pricing). We measure the individual-level accuracy with the top-1, top-2, and top-3 accuracy metrics for choices made in the holdout period. The top-k accuracy is a popular measure for multi-class classification tasks, and it is defined as the proportion of choice situations where the real choice was among the top k alternatives based on the predicted probability.

Top_k accuracy

$$= \sum_s \sum_{j \in \mathbb{C}_s} \mathrm{I}\left(rank\left(\hat{P}\left(y_{sj} = 1\right), \mathbb{C}_s\right) \le k\right) \Big/ \sum_s \sum_j y_{sj}$$

Here the *rank* (x, set) function returns the rank of the x value within the set when values are sorted in decreasing order. The top-1 accuracy is the classification accuracy.

Meaningful causal (price sensitivity) estimates come from the agreement in the literature that positive price sensitivity is an indicator of a biased estimation process, unless we are dealing with status consumption (Goldsmith, Flynn, & Kim, 2010).[1] Hence, we use the percentage of positive price sensitivity predictions to measure misleading price sensitivity estimates.

PositivePriceSensitivityRatio

$$= \sum_s \sum_{j \in \mathfrak{M}} \mathrm{I}\left(W_{sj}\hat{\boldsymbol{\theta}} > 0\right) \Big/ \sum_s \sum_j 1$$

We define two versions of this estimate: Overall Positive Price Sensitivity, where $\mathfrak{M} = \mathbb{C}_s$; and Top-3 Positive Price Sensitivity, where $\mathfrak{M} = \left\{j : rank\left(\hat{P}(y_{sj} = 1), \mathbb{C}_s\right) \le 3\right\}$ and focuses on the top three alternatives with the highest predicted probability by the method being evaluated. This performance indicator is particularly interesting when looking to steer customers to time slots that are more operationally beneficial by pricing them appropriately. In that use case, it is fundamental to have a good prediction of the impact of price increases (and decreases).

5.3. Benchmark methods

We evaluated our proposed *orthogonalized personalized choice model* against three main approaches.

B0A. Non-personalized MNL: The MNL model with alternative specific constants and a common price coefficient has the following utility specification:

$$U_{sj} = \alpha_j + \beta_{price} p_{sj} + \varepsilon_{sj}$$

B0B. Non-personalized MNL with alternative specific price coefficient: This MNL model utility allows price sensitivity to differ across alternatives:

$$U_{sj} = \alpha_j + \beta_{price,j} p_{sj} + \varepsilon_{sj}$$

We estimate the MNL models with the state-of-the-art Biogeme open-source Python package (Bierlaire 2020), available at http://biogeme.epfl.ch.

B1. Personalized choice model without orthogonalization: The *B1* benchmark uses the same features as the basic orthogonalized personalized choice model, except for being a function of price rather than using a deviation from the expected price. *B1* is estimated with a multinomial logit model, first selecting the terms with lasso and then

[1] One exception pointed out via personal communication is that customers who are entitled to free delivery privileges are more likely to choose more expensive slots, which they do not have to pay for, presumably because they feel that they are getting more value.

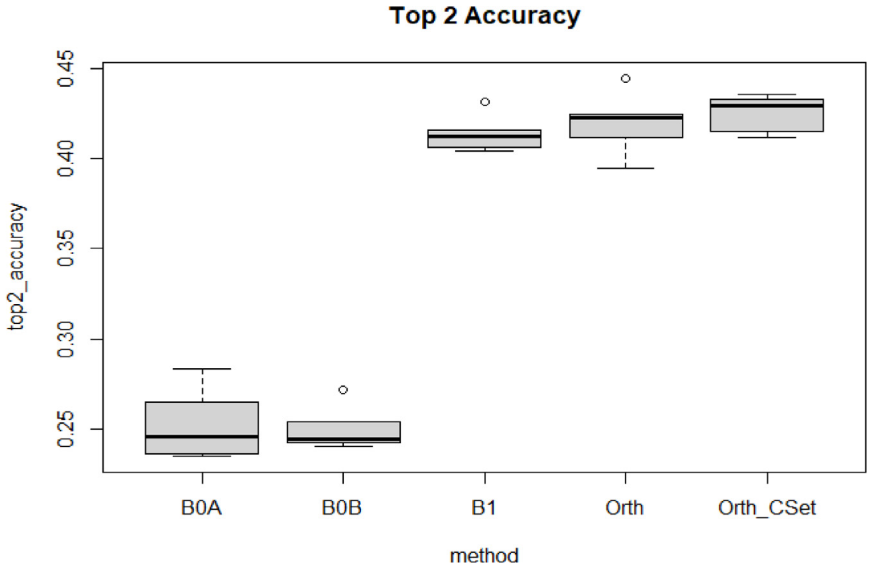


Fig. 2. Box plot of sample replication results – Individual-level accuracy (Top-2).

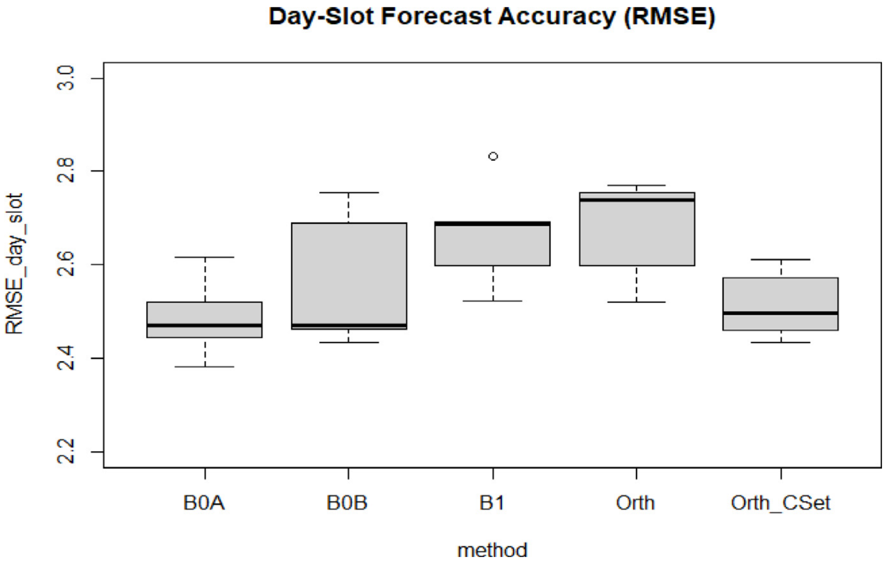


Fig. 3. Box plot of sample replication results – Day–slot-level accuracy (RMSE).

estimating the parameters of the selected terms without regularization. However, *B1* does not entail orthogonalization with respect to confounders, even though it uses regularization to deal with high-dimensional data and is, therefore, named naïve personalization. This model is fit with the glmnet package (Hastie, Qian, & Tay, 2021) and the glm base R routine.

$$U_{sj} = Z_{sj}\boldsymbol{\beta}_{\boldsymbol{B}1} + p_{sj}W_{sj}\boldsymbol{\theta}_{\boldsymbol{B}1} + \varepsilon_{sj}$$

Orthogonalized personalized choice model We evaluate two versions of the proposed model to assess the effect of orthogonalization and consideration-set features on performance, as summarized in Table 5. All versions are fit in R using the glmnet package (Hastie et al., 2021) and the glm and lm base R routines.

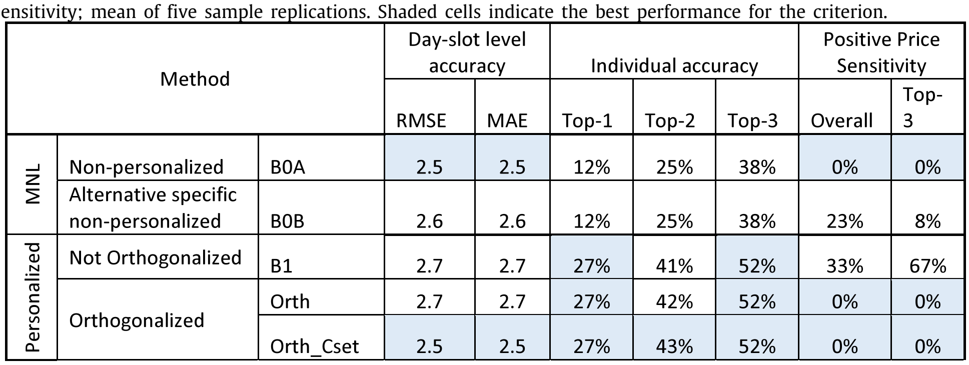
Table 5
Personalized choice models.

Personalized choice model	Orthogonalization	Consideration-set features
B1	N	N
Orth	Y	N
Orth_CSet	Y	Y

6. Results

Table 6 provides the mean values for the key performance indicators across the five sample replications. Further, Figs. 2, 3, and 4 visualize the variability across replications with boxplots of top-2 accuracy, day–slot-level

Table 6
Experimental results for day–slot-level accuracy, individual accuracy, and percentage of predictions with positive price sensitivity; mean of five sample replications. Shaded cells indicate the best performance for the criterion.



Method			Day-slot level accuracy		Individual accuracy			Positive Price Sensitivity	
			RMSE	MAE	Top-1	Top-2	Top-3	Overall	Top-3
MNL	Non-personalized	B0A	2.5	2.5	12%	25%	38%	0%	0%
	Alternative specific non-personalized	B0B	2.6	2.6	12%	25%	38%	23%	8%
Personalized	Not Orthogonalized	B1	2.7	2.7	27%	41%	52%	33%	67%
	Orthogonalized	Orth	2.7	2.7	27%	42%	52%	0%	0%
		Orth_Cset	2.5	2.5	27%	43%	52%	0%	0%

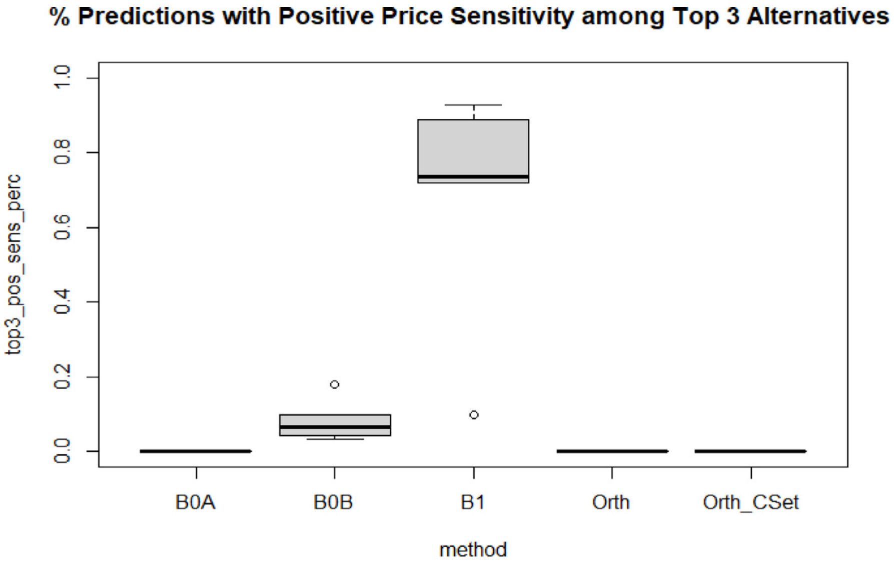


Fig. 4. Box plot of sample replication results – Percentage of predictions with positive price sensitivity among the top three highest-probability alternatives.

forecast accuracy (RMSE), and the percentage of predictions with positive price sensitivity among the top three alternatives, respectively. We observe that personalization significantly improves the accuracy of the choice probability predictions at the individual level. As seen under the Individual accuracy – Top-2 column in Table 6, for the non-personalized MNL, the actual choice is one of the top two predictions 25% of the time. In comparison, the top two predictions of the personalized models capture it at least 41% of the time. Both values are well above the 5% top-2 accuracy value we would expect for random guessing with 40 alternatives, but personalization increases individual accuracy further. Similarly, the top-1 and top-3 accuracies for the non-personalized MNL are 12% and 38%, respectively, far below the lowest values among the personalized models (27% and 52%, respectively). These differences are statistically significant at the 0.001 level for all three top-k measures.[2]

All personalized models improve over the non-personalized MNL in terms of individual-level accuracy. But the personalized model without orthogonalization (B1) provides the wrong sign (positive) in price sensitivity more than two-thirds of the time for its top three alternatives, as seen in the right-most column under Positive price sensitivity – Top-3. This percentage is 0 for the orthogonalized models. There is a statistically significant and substantial difference in the percentage of price sensitivity predictions that are positive between each of the orthogonalized personalized models (Orth and Orth_CSet) and the non-orthogonalized B1 at the 0.001 level among all predictions and the top three alternatives, (see Table 6).

It must be noted that despite its poor individual-level performance, the non-personalized MNL provides very good aggregate accuracy at the day–slot level. Interestingly, the alternative-specific MNL (B0B) model does not improve over the non-personalized MNL (B0A) in any performance measure and frequently provides positive price sensitivity predictions.

[2] Based on the post hoc multiple comparison of means (Tukey's HSD test), following an ANOVA that controls for the replication samples.

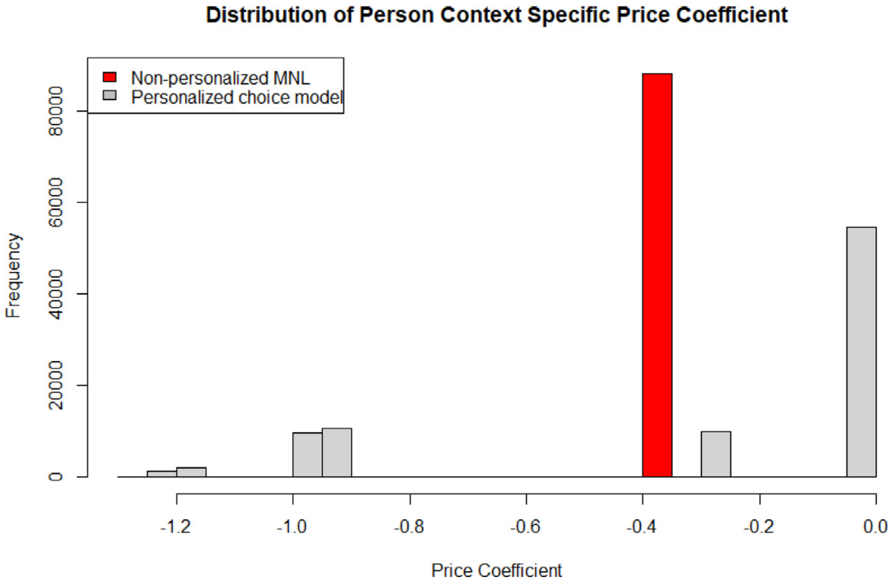


Fig. 5. – Distribution of the individual–context-level price coefficient for B0A and Orth_Cset models.

To further detail the power of personalization models in discriminating different price sensitivities, we plot, in Fig. 5, the distribution of the price coefficient at the individual–context level, as estimated by our personalized choice model, and we compare it with the estimate from the non-personalized MNL (B0) model. The graph shows that while the non-personalized MNL assumes the same price sensitivity in every choice situation and alternative, the personalized model captures the heterogeneity. Interestingly, some customers in given contexts are not price-sensitive. By using the personalized model, the retailer would be better equipped to provide discounts where they help change behavior and avoid unnecessary losses.

The ANOVA controlling for sample replicates showed that the consideration-set features significantly improve the aggregate-level accuracy (at the 0.001 level) in terms of both the RMSE and MAE. They also decrease positive price sensitivity, but not statistically significantly. All of the five replications of the orthogonalized personalized choice model with consideration-set features (Orth_Cset) selected the *n_sooner_same_slot* feature and have the expected negative sign, while none selected the *n_alts_same_day* feature.

We also investigated whether sample splitting, which reduces overfitting in double machine learning (Chernozhukov et al., 2018), is beneficial in this more structured setting. We conclude that sample splitting for orthogonalization does not improve causal estimates but rather hurts the accuracy of the predictions at both the aggregate and individual levels. This is most likely due to our structured model having a lower propensity to overfit than the complex black-box machine learning models with which sample splitting is primarily used. The effect of the smaller sample size increasing the variance of the estimates exceeded the overfitting bias that sample splitting was intended to prevent.

In summary, the proposed orthogonalized personalized choice model with choice-set features (Orth_Cset) is as good as the non-personalized MNL model in aggregate forecasting, has the best individual-level accuracy, and does not result in (biased) positive price sensitivity estimates as it selects the relevant features for personalization.

7. Conclusions

We proposed a new method for predicting customer- and context-specific (i.e., personalized) preferences and price sensitivity in a discrete choice model. The method is equipped to deal with the complications of high-dimensional data and regularization-induced confounding.

Our approach was motivated and evaluated by the customer choice problem in AHD. However, the method should prove applicable in other settings where predicting personalized preferences is relevant. That is the case for other decision problems faced by online retailers, such as deciding the assortment to show to each customer (Bernstein, Modaresi, & Sauré, 2019) or the pricing of fulfillment options at the checkout (e.g., click-and-collect and home delivery).

The experimental results for the AHD context showed that personalized choice models with features describing the alternatives, the context, and the individual's past behavior offer a large (two-fold) improvement over the non-personalized MNL model for predicting individual-level choice probabilities in delivery slot choice modeling. This result is consistent with the commercial success of personalized models.

The results also showed the danger of using high-dimensional data with regularization for personalizing choice models. Despite their high predictive accuracy, such models inappropriately predicted positive price sensitivity in more than two-thirds of the relevant choice alternatives. This result renders them inadequate for evaluating pricing scenarios or optimizing the set of offered alternatives.

Further, the results showed that orthogonalization with respect to the confounders can provide meaningful causal estimates by avoiding regularization bias, provided that there are no unobserved confounders. In the case of endogenous pricing, experimental data would be needed.

The consideration-set features that we introduced to the delivery slot choice problem allowed us to model more realistic substitution probabilities and significantly improve the predictive accuracy. Interestingly, the selected features indicated that customers are more likely to choose a slot when there are fewer alternatives in the same daytime slot on earlier days.

Given the large amount of data surrounding the problems that online retailers face, it may be tempting for practitioners and researchers to rely solely on data to solve business problems and be less rigorous at model development. Our approach and results indicate that the data generation process regarding potential confounders and decision-maker considerations are essential to achieve better outputs.

For future research, it would be interesting to assess the impact of the proposed forecasting models, which can improve accuracy, on the prescriptive side of the problem. In our AHD context, this research gap translates into measuring the bottom-line impact of better predicting how customers react to changes in time slot attributes.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ozden Gur Ali reports a relatonship with Tazi Bilisim Teknolojileri that includes: consultng or advisory.

Acknowledgments

This project received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under grant agreement No. 952060.

References

Aboutaleb, Y. M., Danaf, M., Xie, Y., & Ben-Akiva, M. (2021). Discrete choice analysis with machine learning capabilities. arXiv preprint arXiv:2101.10261.

Agatz, N., Campbell, A., Fleischmann, M., & Savelsbergh, M. (2011). Time slot management in attended home delivery. transportation science. *45*(3), 435–449.

Allenby, G. M. (2017). Structural forecasts for marketing data. *International Journal of Forecasting*, *33*(2), 433–441.

Amorim, P., DeHoratius, N., Eng-Larsson, F., & Martins, S. (2020). Customer preferences for delivery service attributes in attended home delivery. *Chicago booth research paper (20-07)*.

Asdemir, K., Jacob, V. S., & Krishnan, R. (2009). Dynamic pricing of multiple home delivery options. *European Journal of Operational Research*, *196*(1), 246–257.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Ben-Akiva, M., & Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, (pp. 5–33). Springer.

Bernstein, F., Modaresi, S., & Sauré, D. (2019). A dynamic clustering approach to data-driven assortment personalization. *Management Science*, *65*(5), 2095–2115.

Chen, X., Owen, Z., Pixton, C., & Simchi-Levi, D. (2022). A statistical learning approach to personalization in revenue management. *Management Science*, *68*(3), 1923–1937.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., et al. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Cinelli, C., Forney, A., & Pearl, J. (2021). A crash course in good and bad controls. *Sociological Methods & Research*, 00491241221099552.

Danaf, M., Becker, F., Song, X., Atasoy, B., & Ben-Akiva, M. (2019). Online discrete choice models: Applications in personalized recommendations. *Decision Support Systems*, *119*, 35–45.

Feldman, J., Zhang, D. J., Liu, X., & Zhang, N. (2022). Customer choice models vs. machine learning: Finding optimal product displays on alibaba. *Operations Research*, *70*(1), 309–328.

Fildes, R., Kolassa, S., & Ma, S. (2021). Post-script—Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1319–1324.

Gallego, G., Ratliff, R., & Shebalov, S. (2015). A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research*, *63*(1), 212–232.

Goldsmith, R. E., Flynn, L. R., & Kim, D. (2010). Status consumption and price sensitivity. *Journal of Marketing Theory and Practice*, *18*(4), 323–338.

Goodwin, P., Meeran, S., & Dyussekeneva, K. (2014). The challenges of pre-launch forecasting of adoption time series for new durable products. *International Journal of Forecasting*, *30*(4), 1082–1097.

Greene, W. H. (2017). *Econometric analysis: Pearson*.

Gür Ali, Ö (2013). Driver moderator method for retail sales prediction. *International Journal of Information Technology and Decision Making*, *12*(6), 1261–1286.

Gür Ali, Ö., & Gürlek, R. (2020). Automatic interpretable retail forecasting with promotional scenarios. *International Journal of Forecasting*, *36*(4), 1389–1406.

Hahn, P. R., Carvalho, C. M., Puelz, D., & He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, *13*(1), 163–182.

Hastie, T., Qian, J., & Tay, K. (2021). An introduction to glmnet. *CRAN R Repositary*.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *Vol. 2, The elements of statistical learning: data mining, inference, and prediction*. Springer.

Horowitz, J. L., & Louviere, J. J. (1995). What is the role of consideration sets in choice modeling? *International Journal of Research in Marketing*, *12*(1), 39–54.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Hübner, A., Amorim, P., Fransoo, J., Honhon, D., Kuhn, H., de Albeniz, V. M., et al. (2021). Digitalization and omnichannel retailing: Innovative OR approaches for retail operations. *European Journal of Operational Research*, *294*(3), 817–819.

Katz, E. (2001). Bias in conditional and unconditional fixed effects logit estimation. *Political Analysis*, *9*(4), 379–384.

Koch, S., & Klein, R. (2020). Route-based approximate dynamic programming for dynamic pricing in attended home delivery. *European Journal of Operational Research*, *287*(2), 633–652.

Kök, A. G., & Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, *55*(6), 1001–1021.

Lang, M. A., Cleophas, C., & Ehmke, J. F. (2021). Multi-criteria decision making in dynamic slotting for attended home deliveries. *Omega*, *102*, Article 102305.

Lokhorst, J. (1999). The lasso and generalised linear models. *Honors project*, Australia: The University of Adelaide.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, *58*(304), 993–1010.

Mackert, J. (2019). Choice-based dynamic time slot management in attended home delivery. *Computers & Industrial Engineering*, *129*, 333–345.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (2001). Economic choices. *American Economic Review*, *91*(3), 351–378.

McKinsey (2022). Achieving profitable online grocery order fulfillment. https://www.mckinsey.com/industries/retail/our-insights/achieving-profitable-online-grocery-order-fulfillment.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393.

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*(4), 659–677.

Prato, C. G. (2009). Route choice modeling: Past, present and future research directions. *Journal of Choice Modelling*, *2*(1), 65–100.

Robinson, P. M. (1988). Root-*N*-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954.

Shocker, A. D., Ben-Akiva, M., Boccara, B., & Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*, *2*(3), 181–197.

Strauss, A., Gülpınar, N., & Zheng, Y. (2021). Dynamic pricing of flexible time slots for attended home delivery. *European Journal of Operational Research*, *294*(3), 1022–1041.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning – discussion paper. *Journal of Choice Modelling*, *42*, Article 100340.

Wan, A. T., Zhang, X., & Wang, S. (2014). Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting*, *30*(1), 118–128.

Wang, S., Wang, Q., & Zhao, J. (2020). Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C (Emerging Technologies)*, *118*, Article 102701.

Yang, X., & Strauss, A. K. (2017). An approximate dynamic programming approach to attended home delivery management. *European Journal of Operational Research*, *263*(3), 935–945.

Yang, X., Strauss, A. K., Currie, C. S., & Eglese, R. (2016). Choice-based demand management and vehicle routing in e-fulfillment. *Transportation Science*, *50*(2), 473–488.

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, *20*, 22–35. http://dx.doi.org/10.1016/j.tbs.2020.02.003.