# The Customer Journey as a Source of Information

Nicolas Padilla,[1*] Eva Ascarza,[2] Oded Netzer,[3]

[1] London Business School, University of London

[2] Harvard Business School, Harvard University

[3] Columbia Business School, Columbia University

October 25, 2023

## Abstract

In the face of heightened data privacy concerns and diminishing third-party data access, firms are placing increased emphasis on first-party data (1PD) for marketing decisions. However, in environments with infrequent purchases, reliance on past purchases 1PD can pose a challenge in fully harnessing this data. We address these challenges by introducing a probabilistic machine learning model that fuses customer click-stream data, and, when available, purchase data, within and across journeys. Combining data across journeys can be challenging if customers have different needs in different journeys. We account for such "context heterogeneity", using a Bayesian non-parametric Pitman-Yor process. Drawing from within-journey, past journeys, and cross-customer behaviors, our model offers a solution to the "cold start problem," enabling firms to predict customer preferences even in the absence of prior interactions. Notably, the model allows for continuous updating of the inferred preferences as the customer keeps interacting with the firm. We apply the model to data from an online travel platform, revealing the substantial benefits of consolidating 1PD from both current and previous customer journeys. In scenarios with restricted access to customers' historical data collected via cookies, our model suggests that within-journey data can help offset the loss of historical customer data.

Keywords: Customer Journey, Probabilistic Machine Learning, Bayesian Nonparametrics, First-party Data, Privacy, Clickstream Data, Customer Search

[*] Nicolas Padilla is the corresponding author. He is an Assistant Professor of Marketing at London Business School. Eva Ascarza is the Jakurski Family Associate Professor of Business Administration at Harvard Business School. Oded Netzer is the Arthur J. Samberg Professor of Business at Columbia Business School. The authors thank Asim Ansari, Bruce Hardie, Anja Lambrecht, Vicki Morwitz, and Eric Schwartz for their comments, and the participants of the 2019 Marketing Science Conference, the 2020 European Quantitative Marketing Seminar (EQMS), the 2021 European Bayesians in Marketing Summit, the 2022 Customer Journeys In a Digital World Conference at Bocconi University, the 2023 Workshop in Management Science at Universidad de Chile, the 2023 London Quantitative Marketing Conference, the 2023 Choice Symposium, seminars at the University of Colorado Boulder, the University of Texas at Austin, London Business School, Imperial College London, Oxford University, ESADE, IESE, Georgetown University, Dartmouth College, John Hopkins University, Northwestern University, The Pennsylvania State University, University of Washington, National University of Singapore, University of Southern California, McGill University, Harvard Business School, University of Delaware, Pontificia Universidad Catolica de Chile, and Temple University. In writing the manuscript, the authors used ChatGPT (version 3.5) to correct grammatical mistakes and to refine (prompted) sentences or paragraphs, but never to create new context.

# 1 Introduction

The increasing awareness of data privacy and the rise of data privacy laws have led marketers to rely more on their first-party data (1PD) for marketing strategies. With diminished access to third-party sources like cookies and information from data brokers, companies are intensifying their reliance on insights from their proprietary first-party data, encompassing information from a company's internal channels (Murphy, 2022). The implications of limited third-party data access are more pronounced in situations with infrequent purchases. To compensate for the limited history of transactions from the same customer, firms often rely on digital footprints from customer-firm interactions on the company's own channels, such as search queries, clicks, and filtering of alternatives. This collection of behaviors is what we refer to as the *first-party customer journey*, with each journey encapsulating the entirety of behaviors collected by the company as customers navigate towards fulfilling distinct needs. This wealth of data can be monitored by firms across multiple journeys undertaken by the same customer and across multiple customers. Despite this rich source of information, however, many firms are still not leveraging 1PD to its full potential (Huberman, 2021). The challenge often lies in thinking through how to combine different types of customer interactions, such as textual queries, clicks, and actual purchases. What should be the weight of clicks versus actual purchases? How should one weigh historical journey or purchase data versus current journey activity? How should one combine information across journeys if customer needs may change across journeys?

The objective of this research is to propose a modeling approach for firms to extract relevant information from first-party customer journeys, integrating diverse data sources originating from the multitude of interactions customers engage in with the focal firm as they aim to satisfy different needs. Our approach leverages the (potentially valuable) information across journeys and across customers, with the aim to provide a better understanding of customer needs, even in situations characterized by limited or absent historical data about customers. Several challenges arise from combining information from multiple customer interactions across multiple journeys. The first pertains to the amount of valuable information in prior-to-purchase interactions. While most users engage with a company's website or apps repeatedly, more often than not, those interactions do not end up in a purchase. Hence, while the information prior to purchase should, in theory, be informative of users' needs and preferences, it also has the potential to be irrelevant or misleading, given that on many occasions, the customer did not purchase the product that they were engaging with. Second, even if pre-purchase information is relevant, how should it be "weighted" and combined with actual past purchase information? Finally, another challenge emerges from combining multiple journeys, as even the same customer may not always be seeking to satisfy the same need. For example, a user ordering dinner on Grubhub may sometimes be looking for a sophisticated meal to celebrate a special dinner occasion with a friend, while at other times, they may be seeking a quick lunch bite on their own. (Similar phenomena occur in other purchase settings, such as booking a flight or a hotel, or selecting a movie to watch.) As a result, combining signals across customer journeys must take such cross-journey differences into account.

To address these issues, we introduce a probabilistic machine learning model that links the customer click-stream data over the course of a journey and integrates that information across journeys, and across the customer's history of purchases. The model accounts for what we call *context heterogeneity*, which are journey-specific preferences that capture customers' journey-specific needs and allow us to combine information across journeys with possibly varying needs. We model the journey decisions on what to search, what to click, and what to buy to be both a function of customers' stable preferences and the unique needs of the context of the journey. Intuitively, contexts are unobserved segments that capture need-specific preferences (e.g., a special dinner vs. a quick lunch) that are shared across customers. We uncover those contexts non-parametrically using a Pitman-Yor process, which allows for the creation of new contexts that have not been previously observed as new journey observations arrive.

The model leverages three main sources of information, all collected from the company's own channels: (1) within journey's behavior (e.g., the customer's search query and what the customer searched for and clicked on in the focal journey); (2) past journeys' behavior (e.g., what the customer clicked on and purchased in past journeys); and (3) across customers' behavior (e.g., what other customers with similar search behavior clicked on and purchased).

The within-journey information, particularly click information, allows us to identify the unique journey-specific preferences. The past journeys' information (including not only past purchases but also searches, filters, and clicks) allows us to infer the customer's stable preferences. Finally, the information across customers augments the (possibly thin) information from the first two sources with data from other customers with similar preferences. Thus, in situations where customers have not yet been identified (e.g., have not logged in yet or it is their first purchase) and when past behaviors cannot be linked to a particular customer (e.g., where cookies were not enabled or customers used incognito), our model can augment the information collected along the focal journey with across-customer information from other customers' journeys of similar contexts. This presents a novel solution to the so-called "cold start problem" (Padilla and Ascarza, 2021) by enabling firms to infer preferences from customers even before they make their first purchase, or before they identify themselves.

We apply the model to consumer journeys on one of the largest online travel platforms. We find substantial value in consolidating the various signals derived from customers' journey traces in 1PD. The model demonstrates the capacity to infer crucial aspects of what customers are looking for, such as airline selection, number of stops, or price preferences. For example, incorporating information from the first two clicks in the current journey enhances the predictive ability of the customer's preferred airline alliance by 25% compared to the prediction when the user just inserted their travel query. Further, by incorporating information from five clicks, the predictive ability soars by 73%, showcasing the model's adaptability by continuously updating these inferences as customers progress through their journeys. This model's adaptability in inference stands in stark contrast to traditional panel-data models, which rely solely on purchase data (e.g., Rossi et al., 1996) and, therefore, cannot adapt inference as the company collects 1PD along the focal journey. Indeed, compared to a model that only leverages historical purchase data, our model can predict the actual product chosen 10 times more accurately, highlighting the value of prior-to-purchase data, which in this setting includes clicks and filters. We find that combining historical journeys (even from different customers) can alleviate the cold-start problem, as they are particularly valuable at the beginning of a new journey, improving alliance choice predictions by 38% (when compared to predictions that do not leverage past journeys).

In addition to providing valuable insights into customers' current journey preferences, our model allows us to shed light on the different types of customer journeys and how to leverage them over time. For example, we account for the fact that data from a customer's past searches for a business trip may be highly informative for the current customer's journey if it is a business trip, but less so if it is a family vacation. The model uncovers a total of 22 distinct contexts, each representing different "needs" that customers seek to fulfill, along with their corresponding preferences. For instance, one context identified by the model encompasses "one-way solo domestic trips," characterized by last-minute flight searches, unaccompanied passengers, and with a strong aversion to smaller airlines and multiple alliances. On the other hand, another context, which we term "Family vacations in the Caribbeans," involves customers traveling with other adults and children, seeking roundtrip flights between the US and other destinations in North America or the Caribbean. In this context, flight searches typically commence approximately three months prior to departure and customers have a strong preference for non-stop and shorter routes.

Finally, the proposed modeling framework enables quantifying the impact of data loss under alternative data privacy scenarios, for example, one in which cookies are eliminated and past journeys cannot be identified to a specific customer. Conducting this examination with our data, we find that the model's ability to leverage information along the current journey can partially compensate for the loss in the firm's ability to identify the individual customer — e.g., five clicks in the current journeys can compensate for unobserved past journeys when predicting product choice. In other words, the value of integrating data across first-party journeys proves particularly advantageous when attempting to understand the preferences of customers who lack historical data or have yet to identify themselves. This aspect becomes particularly relevant due to the rise of platforms that enable potential customers to conduct searches without the need for logging in.

Overall, this paper contributes to the literature by proposing a method to infer customer preferences in settings where there is thin purchase history — i.e., most customers have not purchased multiple times, some customers have not even purchased yet — and even when customers' needs might change from one purchase occasion to another. We do so by jointly modeling all pieces of information from the 1PD customer journey: search queries, filter selections, clicks and purchases; both within journeys and across journeys. Managerially, this work is valuable to data-driven organizations as it enables them to best leverage their 1PD, an objective that is increasingly important due to changes in the data regulatory environment and the enhanced focus on customer privacy. Furthermore, by illuminating distinct contexts, the model provides valuable insights into the specific preferences and characteristics that shape customer behavior within each empirical setting. This comprehensive understanding enhances the company's ability to tailor marketing strategies and meet the unique needs of diverse customer segments.

The rest of the paper is organized as follows. Section 2 discusses previous work that attempts to capture customers' preferences from thin and richer data. In Section 3, we develop our modeling framework to fuse the multiple facets of 1PD and use customer journeys as a source of information. We describe our empirical setting and modeling results in Section 4. Using our modeling framework, we quantify the value of 1PD in Section 6 Finally, we conclude by discussing the generalizability of our modeling approach, as well as potential limitations and future directions.

# 2 Using transactional data and the customer journey to capture consumers' preferences

The current work contributes to the rich literature in marketing on using transactional data to estimate consumers' preferences at the individual level (Rossi et al., 1996; Allenby and Rossi, 1998; Duvvuri et al., 2007; Fiebig et al., 2010). Most notably, (Rossi et al., 1996) has demonstrated that even one or a few purchases can be informative in understanding consumers' preferences and future choices. These models have been widely used by researchers and practitioners in settings where individual transactions are available. However, there are many business contexts in which observing multiple, or even one, purchases by the same customer is rare, making it difficult to reliably estimate individual-level preferences. To offset that limitation, the e-commerce environment allows researchers to observe the customer journey, which can include both clickstream activity (e.g., search, filter, and purchase) within the focal journey and possibly across past journeys that may or may not have ended up in a purchase. The challenge we address in this paper is how to combine such clickstream data with past purchase data, and particularly when different journeys may have different purposes and underlying preferences.

There is a rich literature on consumer search and the use of clickstream data (e.g., Montgomery et al., 2004; Kim et al., 2010; Ghose et al., 2019; Seiler, 2013; Bronnenberg et al., 2016; Honka and Chintagunta, 2017; Chen and Yao, 2017; Ursu, 2018; Donnelly et al., 2023) that uses within-journey information to infer customer preferences and to predict purchase. For example, Montgomery et al. (2004) find that customers' browsing behavior can predict future steps in the browsing process as well as conversion. De los Santos and Koulayev (2017) shows how firms can use data on the current visit to optimize click-through rates. Donnelly et al. (2023) leverages co-occurrences of products clicked during the same journey and borrows information cross-sectionally using latent factorization. These studies analyze only the focal journey, ignoring the information provided by previous journeys (for an exception see Morozov et al., 2021), limiting the model's ability to capture rich heterogeneity in customer preferences. We extend this literature by looking both within and across journeys while accounting for, and leveraging, the fact that preferences may vary across journeys.

Substantively, this paper relates to previous work that has incorporated other sources of information when data on the main behavior of interest is thin — for example, by leveraging preferences from other product categories (Iyengar et al., 2003), by combining individual-level usage in digital channels with aggregate consumption summaries in other (more traditional) channels (Feit et al., 2013), by semantically linking web pages content and clicking to text-based search queries in search engines (Liu and Toubia, 2018), or by leveraging detailed acquisition data to infer future purchase behavior (Padilla and Ascarza, 2021). The present research contributes to this stream of literature by highlighting the value of extracting information from current and previous customer's journeys to infer current customer preferences.

Our work also relates to the literature on context-dependent product recommendations (e.g., Sarwar et al., 2001; Hidasi et al., 2016; Jacobs et al., 2016; Yoganarasimhan, 2020). This growing literature in the areas of computer science as well as in marketing has proposed diverse machine learning approaches — including item-to-item recommendation approaches using similarities across products, Recursive Neural Networks, or topic modeling — to recommend products when there is lack of historical individual-level data. Most of these methods require the observation of several individuals interacting with the same set of products, as well as each individual interacting with several products. Our approach relaxes this stringent requirement by extracting preferences for attributes and not only for "entire" products. As a result, our model can be applicable when the product space is large, is growing over time, and even if it includes products that have not been purchased in the past as in our travel platform application.

Managerially, this paper contributes to the growing literature on data privacy regulation (Schneider et al., 2017; Goldberg et al., 2019; Aridor et al., 2020; Sun et al., 2023; Tian et al., 2023; Korganbekova and Zuber, 2023). Due to increasing restrictions in the firms' ability to share and collect third-party data, organizations are looking for ways to keep relying on data in making decisions without relying on third-party data (Latvala et al., 2022). We contribute to that literature by proposing a modeling framework that enables firms to better leverage their 1PD to counterbalance potential information loss when customers opt for non-tracking.

# 3 Fusing different facets of the first-party journey data

## 3.1 Model overview

We propose a modeling framework that allows firms to extract the information contained in the various first-party interactions customers have with the focal firm. To that end, we define the *first-party customer journey* (1PJ) as the set of all the interactions observed by a firm during the course of the customer journey aimed at satisfying the customer journey-specific needs. We highlight several components of this definition. First, a 1PJ only includes the interactions that the firm collects in its own platforms (e.g., web, apps, call center). We exclude interactions with competing firms aimed to satisfy the same need, as this information is unobserved to the focal firm. Note that the customer may start the journey in the focal platform and end it on another or vice versa. Second, these interactions do not necessarily need to occur within a simple session but might comprise several sessions occurring at different times or days. Third, the 1PJ comprises the interactions that a customer has with the focal firm aimed at satisfying a *specific need* related to a particular consumption opportunity and/or purchase in a specific category. For example, a customer that interacts with a food delivery platform to get lunch on a Tuesday and dinner on a Friday night is aiming at satisfying two distinct needs on different consumption occasions. Hence, we separate these interactions into two separate journeys.

Importantly, as customer interactions within a 1PJ are aimed at satisfying a specific need, the same customer might be looking for very different things from one journey to another, which generates a challenge when combining information across journeys. For example, in the food delivery context, a customer may be looking for a sophisticated dinner to celebrate a special occasion with a friend on Friday night, while looking for a quick bite for Tuesday lunch. As a result, it is important that our modeling framework takes into account that some of customers' preferences will be transferable across journeys (e.g., the customer does not like spicy food) and others will not (e.g., price sensitivity might be lower for Friday dinner than for Tuesday lunch) when combining information across journeys. To overcome that methodological challenge, we introduce the idea of *context* of a journey (e.g., "a quick bite alone", "a fancy dinner with partner"), which enables the model to capture differences in what customers are looking for across different journeys in a meaningful way. The historical data of a customer may be comprised of a single or multiple 1PJs. Additionally, different privacy regulations may limit the observation of 1PJs over time, or what can be stored about the customer from historical 1PJs.

With this conceptualization of a journey, we develop a probabilistic machine learning modeling framework that allows firms to combine several sources of 1PD across multiple journeys — both present and past journeys — and across many customers. The model flexibly combines multiple types of behaviors, namely search queries, clicks, filters, purchases, and can accommodate other behaviors collected by the focal company via their website/app or other channels. Using this framework, a focal firm can leverage its 1PD to (dynamically) infer what a particular customer is looking for as they advance in their focal journey.

We index customers by $i \in \{1,...,I\}$, and their journeys by $j \in \{1,...,J_i\}$, where $J_i$ is the number of purchase journeys customer $i$ has undertaken. We use $t \in \{1,...,T_{ij}\}$ as the cardinal order of actions (hereafter, step) of customer $i$ in journey $j$. Our model is informed, mainly, by three types of actions, all captured by the firm's website/app: queries ($\mathbf{q}_{ij}$), clicks ($y^c_{ijt}$), filters ($\mathbf{f}_{ij}$), and purchase ($y^p_{ij}$). We specify each component of the model next.

## 3.2 Model components

### 3.2.1 Query

We leverage the information in the search query through multiple query variables that capture the context of a journey. These variables help capture journey-specific needs, even for different journeys of the same customer. For example, day-of-the-week and time-of-the-day may be relevant to infer whether a search query in a food delivery platform relates to a single-person weekday lunch vs. a romantic Friday night dinner. In a travel setting, the number of passengers, dates of travel, and destination, among others, can be informative about the context for what the user might be looking for.

We denote by $\mathbf{q}_{ij}$ the vector of query variables that describe journey $j$ for customer $i$,

$$\mathbf{q}_{ij} = \begin{bmatrix} q_{ij1} & ... & q_{ijM} \end{bmatrix}',$$

where each component indexed by $m \in \{1,...,M\}$ describes a different type of query variable (e.g., length of the stay, traveling with kids). Because these pieces of information are provided by the customer to obtain a set of product results that match their preferences, we treat each query variable as an outcome that depends on some unobserved component that captures the customer's true need in the focal journey.[1] We model $\mathbf{q}_{ij}$ as a function of a vector of parameters $\boldsymbol{\omega}_j = \begin{bmatrix} \omega_{j1} & ... & \omega_{jM} \end{bmatrix}'$. Note that the model treats query variables as an outcome, not a feature or covariate. Doing so allows the model to easily account for missing query variables, or query

[1] Potentially, customers could slightly modify the query along the journey while searching for a product to satisfy the same need (e.g., changing the departing date when customers search for flight tickets). We model only the first query by a customer in each journey due to the minimal additional information these often provide. That being said, the model can easily be extended to learn from multiple query instances.

variables that are not present in certain journeys, increasing the amount of 1PD that can be leveraged by the firm.

We assume that given $\boldsymbol{\omega}_j$, the components of $\mathbf{q}_{ij}$ are conditionally independent, that is:

$$p(\mathbf{q}_{ij}|\boldsymbol{\omega}_j) = \prod_{m=1}^{M} p(q_{ijm}|\omega_{jm}). \tag{1}$$

Each type of query variable $m$ could be of multiple types: (1) binary, (2) categorical, (3) continuous real-valued, or (4) continuous positive-valued. We flexibly model $q_{ijm}$ using a different distribution $p_m(q_{ijm}|\omega_{jm})$ for each type of variable $m$,

$$q_{ijm} \sim \begin{cases} \text{Bernoulli}(\omega_{jm}) & \text{if } q_{ijm} \text{ is binary} \\ \text{Categorical}(\omega_{jm}) & \text{if } q_{ijm} \text{ is categorical} \\ \exp(\omega_{jm}) & \text{if } q_{ijm} \text{ is continuous positive-valued} \\ \mathcal{N}(\omega_{jm}, \sigma_m^2) & \text{if } q_{ijm} \text{ is continuous,} \end{cases} \tag{2}$$

where each parameter $\omega_{jm}$ has the appropriate support given the distribution it governs.[2] Our model can accommodate other distributions such as Poisson or Binomial for count variables, and Student's t-distribution or Cauchy for long-tailed continuous variables.

### 3.2.2 Joint model of clicks and purchase

We structure the modeling of clicks and purchase decisions in two phases. First, customers explore products through clicks and potential filtering to form a consideration set. Following this, customers proceed to the purchase decision stage, where they either choose an item from their considered set or decide not to make a purchase. All of these decisions are guided by a shared set of customer preferences, denoted as $\boldsymbol{\beta}_{ij}$.

#### 3.2.2.1 Click decisions

Along the journey, the customer clicks through pages of product results. The customer can navigate back and forth between clicking on products and refining their searches. In each step, the customer decides among: (1) clicking on one of the products shown on the page to consider

[2] We choose to define $\sigma_m$ fixed across all journeys, to avoid singularity issue. That is analogous to approaches that prevent regularity issues commonly found when estimating Gaussian mixtures with component-specific variances.

it for purchase, (2) continuing to search to receive a new set of results (e.g., by adjusting the query or filtering the results), or (3) ending the search and moving to the purchase decision among those considered.

We model the click decision of alternative $k$ at step $t$ of the journey using a discrete choice model. We define $\text{Page}_{ijt}$ as the set of products displayed to customer $i$ in journey $j$ at step $t$. The customer faces a decision between: clicking on one of the displayed products $k \in \text{Page}_{ijt}$, continue searching to get a new set of products ($k=s$), or finish the search process and move to the purchase decision ($k=e$), which could mean either purchasing a considered product or deciding not to buy. We denote the choice consumer $i$ makes at step $t$ of journey $j$ by $y^c_{ijt} \in \text{Page}_{ijt} \cup \{s,e\}$, which we model using a multinomial probit specification with latent propensities $u^c_{ijtk}$, such that

$$y^c_{ijt} = \underset{k \in \text{Page}_{ijt} \cup \{s,e\}}{\text{argmax}} \left\{u^c_{ijtk}\right\}, \text{ with}$$

$$u^c_{ijtk} = \begin{cases} \beta^{0c}_{ij} + \mathbf{x}^c_{ijtk}{}' \cdot \boldsymbol{\beta}^x_{ij} + \text{log-rank}_{ijtk} \cdot \eta + \varepsilon_{ijtk} & \text{if } k \in \text{Page}_{ijt}, \\ \beta^{0s}_{ij} + \varepsilon_{ijts} & \text{if } k = s, \\ \beta^{0e}_{ij} + \varepsilon_{ijte} & \text{if } k = e, \end{cases} \tag{3}$$

where $\varepsilon_{ijtk} - \varepsilon_{ijte} \sim \mathcal{N}(0,\sigma^2)$, $\mathbf{x}^c_{ijtk}$ is the vector of attributes of product $k$, $\boldsymbol{\beta}^x_{ij}$ is the vector of customer- and journey-specific product-attribute preferences, $\beta^{0c}_{ij}$ is the intercept for clicking on a product, $\beta^s_{ij}$ is the intercept for the decision to continue searching, and $\beta^{0e}_{ij}$ is the intercept for finishing the search process, normalized to 0 for identification purposes. Note that by definition, customers stop searching in the last observed step and move to the purchase decision (i.e., $y^c_{ijT_{ij}} = e$).

We control for ranking effects on search (Ursu, 2018) by incorporating the log of the position of product $k$ within the results page into the search in $u^c_{ijtk}$ and using $\eta$ to capture such ranking effects.[3] Such a term also captures search costs within a page, along with the intercepts in (3) that capture users' propensity to keep searching and are related to search costs across pages. We denote the vector of product-attributes $\mathbf{x}^c_{ijtk}$ to be $t$-specific to allow for a subset of all

[3] Following Ursu (2018), we include the log-position rank in the click decision but not in the purchase-given-clicks decision.

attributes $\mathbf{x}_{ijk}$ to be shown differently in different types of pages. For example, while customers observe all attributes at the moment of purchase, they may not observe all of them on certain pages while searching. Similarly, the observability of certain attributes may even differ among different types of pages (e.g., departing and returning results pages for flights in online travel).

### 3.2.2.2 Filter decisions

Websites and apps usually collect other types of interactions — e.g., whether a user filters results based on some attributes — information that can be used to further inform about customer preferences in that particular journey. Unlike clicks, filters are not frequently observed in the data — many journeys do not have filters, and when they do, it generally occurs only once along the entire journey. We avoid computational burden by modeling the filtering decision at the overall journey level (rather than at the step level $t$); that is, we model whether the customer uses a particular filter *at any time* during the journey.

We denote by $\ell \in \{1,...,L_{ij}\}$ the level customer $i$ in journey $j$ can filter on and define $\mathbf{f}_{ij}$ the vector of summarized filter decisions for customer $i$ in journey $j$,

$$\mathbf{f}_{ij} = \begin{bmatrix} f_{ij1} & ... & f_{ijL_{ij}} \end{bmatrix}',$$

where each component $f_{ij\ell}$ is defined by

$$f_{ij\ell} = \begin{cases} 1 & \text{if customer } i \text{ filters on level } \ell \text{ within journey } j \\ 0 & \text{otherwise.} \end{cases}$$

We model each component $\mathbf{f}_{ij}$ using a binary probit specification such that

$$f_{ij\ell} \sim \text{Bernoulli}\Big(\Phi\Big(\alpha^0_\ell + \mathbf{w}_{ij\ell}{}' \cdot \boldsymbol{\alpha}^w_\ell + \boldsymbol{\beta}^{x}_{ij}{}' \cdot \boldsymbol{\alpha}^\beta_\ell\Big)\Big), \tag{4}$$

where $\alpha^0_\ell$ is the intercept of filtering on level $\ell$, $\boldsymbol{\beta}^x_{ij}$ is the same set of preferences that drive clicks and purchases, and $\boldsymbol{\alpha}^\beta_\ell$ is the vector that relates those preferences to the filtering decision. It is this term that allows the model to learn preferences for attributes by fusing filtering decisions about those attributes. To make sure that we only pick the signals from filtering decisions that pertain to customer preferences (and not to other contextual factors that might affect filtering decisions), we include $\mathbf{w}_{ij\ell}$ capturing a set of controls that summarize the set of (unfiltered) results. In particular, we control in $\mathbf{w}_{ij\ell}$ for the number of total products, the percentage of products with level $\ell$, and the number of top 5 products with level $\ell$ in the unfiltered results.[4]

#### 3.2.2.3 Purchase given consideration

After clicking and possibly filtering through product results, customers make the purchase decision, which we model as a discrete choice among the alternatives in a consideration set $\mathcal{C}_{ij}$. Specifically, we define the consideration set as the set of products that have been clicked on at least once during the course of the journey, plus the outside option of not purchasing

$$\mathcal{C}_{ij} = \{k : k \in \text{Page}_{ijt}, y^c_{ijt} = k, t \in \{1,...,T_{ij}\}\}. \tag{5}$$

We model the purchase decision using a multinomial probit specification with latent propensities $u^p_{ijk}$. That is,

$$y^p_{ij} = \underset{k \in \mathcal{C}_{ij} \cup \{\text{NoPurchase}\}}{\text{argmax}} \{u^p_{ijk}\}\text{ , where}$$

$$u^p_{ijk} = \begin{cases} \beta^{0p}_{ij} + \mathbf{x}_{ijk}{}' \cdot \boldsymbol{\beta}^x_{ij} + \epsilon_{ijk} & \text{if } k \in \mathcal{C}_{ij} \\ \beta^{0o}_{ij} + \epsilon_{ijo} & \text{if } k = \text{NoPurchase}, \end{cases} \tag{6}$$

with $\epsilon_{ijk} - \epsilon_{ijo} \sim \mathcal{N}(0, \sigma^2_p)$, and where $\mathbf{x}_{ijk}$ is the vector of attributes of product $k$, $\boldsymbol{\beta}^x_{ij}$ is the same vector of customer- and journey-specific product-attribute preferences shown in (3), $\beta^{0p}_{ij}$ is the intercept for purchasing a product, and $\beta^{0o}_{ij}$ is the intercept for not buying, normalized to 0 for identification purposes.

Finally, we define

$$\boldsymbol{\beta}_{ij} = (\beta^{0c}_{ij}, \beta^{0s}_{ij}, \beta^{0p}_{ij}, \boldsymbol{\beta}^{x}_{ij}{}')', \tag{7}$$

as the vector of all clicks and purchase preferences.

## 3.3 Heterogeneity: The Role of Contexts

In models of repeat purchase of frequently purchased products such as consumer packaged goods, a common assumption is that all past journeys share the same or similar preferences (e.g., Rossi et al., 1996; Allenby and Rossi, 1998).[5] However, in many settings such as the one

[4] As journeys may contain multiple unfiltered results due to multi-session journeys, we average these controls across the unfiltered results pages of all sessions within a journey.

[5] A notable exception is Jacobs et al. (2016) that assume customers have motivations that drive different basket selections.

we consider here, customers might exhibit *journey*-specific preferences. That is, a customer may exhibit different behavior when looking for a flight domestically versus internationally, or when flying for leisure versus for business. This variation poses a challenge when we try to fuse information across journeys, especially when the historical data for each individual customer is relatively sparse. To capture this behavior, we assume that journeys belong to one of many *journey contexts*, which we model via latent components that capture need-specific preferences that are shared across customers.

Incorporating context-specific preferences adds several methodological complexities. First, the journey contexts are unobserved and, therefore, need to be inferred from the data. Second, it is neither the case that customers systematically "transition" from one context to another (like, for example, in hidden Markov models), challenging identification because individual behavior in the previous journey does not necessarily inform the context in the current journey. Third, we do not know how many contexts there are — and this number is likely to be different across settings — so ideally, we would like to learn the number of contexts from the data, without the need to run the model for each number of contexts. Finally, we want to provide enough flexibility to the model such that it will be able to capture *meaningful* contexts as informed by both the query and the customer clickstream behavior (e.g., a "summer family trip" context that bundles together journeys that are more likely to be international trips, with more than one adults and with children, which may involve strong preferences for non-stop destinations and moderate price sensitivity).

To overcome these challenges, we model the journey context as a non-parametric latent segmentation over journeys across customers, using information from the query variables ($\boldsymbol{\omega}_j$) as well as the preferences of these journeys that drive clicks and purchases ($\boldsymbol{\beta}_{ij}$). Specifically, we separate $\boldsymbol{\beta}_{ij}$ from Equation (7) into an individual customer heterogeneity component and a context heterogeneity component that varies from one journey to another as

$$\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\rho}_j, \tag{8}$$

where $\boldsymbol{\mu}_i$ is an individual-specific vector that is shared across all journeys of a particular customer, and $\boldsymbol{\rho}_j$ is a context-specific vector that "shifts" customers' preferences depending on the need of the specific journey but is shared by all customers with the same need. Note that because most customers are observed for a few journeys, often a single journey, we cannot directly include a customer-journey specific term.

We account for *customer heterogeneity* by modeling the individual specific vector of parameters following a multivariate normal distribution:

$$\boldsymbol{\mu}_i \sim \mathcal{N}(0,\Sigma). \tag{9}$$

We further define $\boldsymbol{\gamma}_j$ as the vector of all context-specific parameters,

$$\boldsymbol{\gamma}_j = \left[ \underbrace{\boldsymbol{\omega}_j}_{\text{Query}} \quad \underbrace{\boldsymbol{\rho}_j}_{\text{Clicks and purchase}} \right] \tag{10}$$

which includes the set of parameters that determine queries, clicks, and purchases.

We model *context heterogeneity* non-parametrically assuming that the context-specific component of a journey, $\boldsymbol{\gamma}_j$, is drawn from an unknown discrete distribution $F$, which we call the context distribution (e.g., a histogram of contexts). We assume that this histogram $F$ is drawn using a Pitman-Yor Process. The Pitman-Yor Process (Pitman and Yor, 1997) is a distribution over infinite almost surely discrete measures used in non-parametric Bayesian models. Thus, we draw the context-specific parameters $\boldsymbol{\gamma}_j$ from the context distribution $F$, and we place a Pitman-Yor process prior on the context distribution $F$. That is,

$$\boldsymbol{\gamma}_j \sim F \tag{11}$$

$$F \sim \mathrm{PY}(d,a,F_0), \tag{12}$$

where $0 \leqslant d < 1$ is a discount parameter, $a > -d$ is a strength parameter, and $F_0$ a base distribution over the same space as $\boldsymbol{\gamma}_j$, such that $F_0$ is the mean distribution of $F$.

Note that when $d = 0$, the Pitman-Yor process reduces to a Dirichlet process with concentration parameter $a$ and base distribution $F_0$. The addition of the parameter $d$ allows the drawn distributions from a Pitman-Yor process to exhibit a more flexible long-tail distribution of weights for the mass points, as opposed to the weights decaying exponentially when drawn from Dirichlet processes. This means that the Pitman-Yor process allows for more distinct mass points in the drawn histogram to appear as new observations come in. This feature of the Pitman-Yor process allows the model to capture new contexts that may not have been observed before or contexts that may happen rather infrequently. In Figure 1, we show that as more observations come in, the expected unique number of clusters grows for both the Dirichlet process (left most figure in Figure 1) and the Pitman-Yor process with varying values of the discount parameter $d$. In contrast to the Dirichlet process, the Pitman-Yor process allows for more flexible patterns of how these unique clusters appear in the data. Moreover, using a Pitman-Yor process as a prior for our context distribution, similar to the Dirichlet Process, allows our model to infer the number of contexts directly from the data.

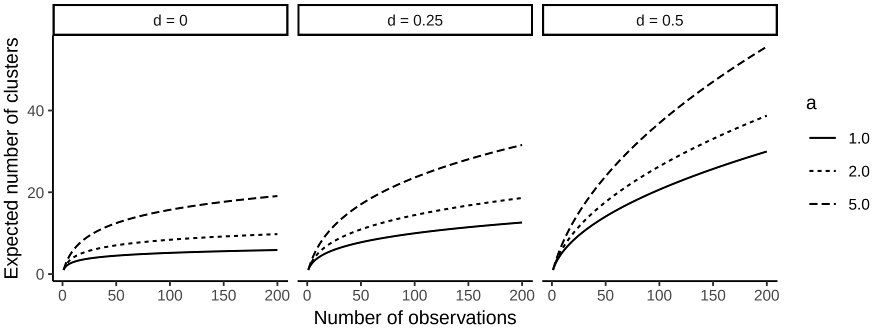


**Figure 1:** Expected number of clusters from a Dirichlet Process ($d=0$, left) vs. a Pitman-Yor process ($d=0.25$, middle; and $d=0.5$, right)

We express the context distribution $F$ in terms of the stick-breaking representation of the Pitman-Yor process (Ishwaran and James, 2001), $F=\sum_{c=1}^{\infty}\pi_c\delta_{\theta_c}(\cdot)$ where

$$\theta_c \sim F_0, \tag{13}$$

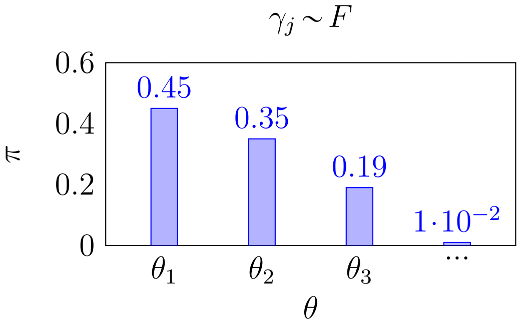
$$\pi_c = V_c\prod_{h=1}^{c-1}(1-V_c), \qquad V_c \sim \text{Beta}(1-d, a+c\cdot d). \tag{14}$$

We can then rewrite (12), using the location vectors $\theta_c$ and a context assignment variable $z_j \in \{1,2,...\}$ such that

$$\begin{aligned}\boldsymbol{\gamma}_j &= \theta_{z_j}\\ p(z_j=c) &= \pi_c.\end{aligned} \tag{15}$$

These context assignment variables will become useful in understanding the journey-specific need of each journey.

We provide some intuition on how the Pitman-Yor process prior in this model captures the contexts non-parametrically, which we illustrate using Figure 2. The distribution $F$ acts as a histogram of contexts, where each location $c=1,2,...$ of the histogram represents a different context (e.g., the summer family vacation, an east-coast business week trip). For any new journey that a customer undertakes, the model would draw its journey specific parameter $\boldsymbol{\gamma}_j$ from this histogram of contexts $F$. The histogram has two main set of parameters, the location $\theta_c$ and the context size $\pi_c$. The locations $\theta_c$ indicate the set of query, click, and purchase preferences that are associated with context $c$. The context size $\pi_c$ represent how likely is context $c$ to be drawn.



- $\theta_1$: summer family vacation
- $\theta_2$: east-coast business week trip
- $\theta_3$: honeymoon
- ...

**Figure 2:** Example of a context distribution drawn from a Pitman-Yor process prior

Methodologically, our work builds on and contributes to the literature on Bayesian non-parametric models in marketing (Ansari and Mela, 2003; Kim et al., 2004; Braun and Bonfrer, 2011; Bruce, 2019), and particularly, on models to capture multiple sources of heterogeneity (Dew et al., 2020; Boughanmi and Ansari, 2021). We extend that literature by proposing a Pitman-Yor process for inferring heterogeneous discrete distributions with an unknown number of components, in our case, contexts, which are key elements in our modeling framework to fuse information across different journeys in a meaningful way.

## 3.4 Model summary

In sum, our model combines multiple flows of information (queries, clicks, filters and purchases) from different sources (past and current journeys) to learn what customers might be looking for in different purchase occasions. Figure 3 depicts the model and the main assumptions about the data-generating process.

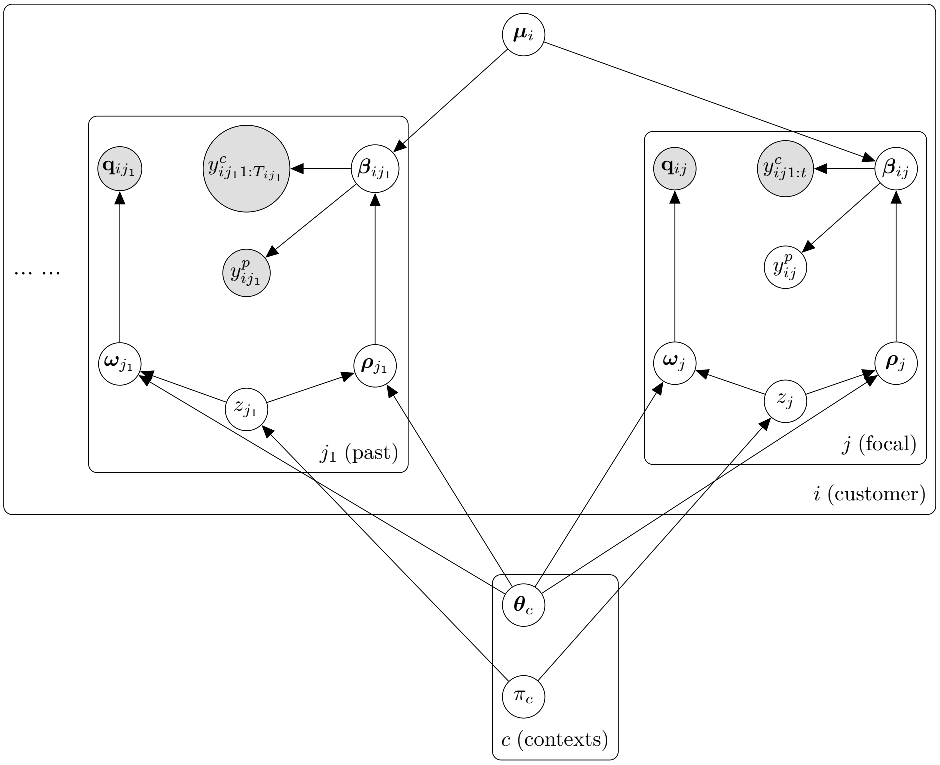


**Figure 3:** Simplified directed acyclical graph of the data generating process. Variables in shaded (white) circles are observed (unobserved). We omit from the graph the dispersion across customers ($\Sigma$), the consideration set ($\mathcal{C}_{ij}$), parameters $a$ and $d$ from the Pitman-Yor process, and all the priors. To simplify the notation, we also omit filters $\mathbf{f}_{ij}$, but for all purposes of this figure, we can consider them as part of the clicks $y^c_{ij1:t}$. Parameters $\boldsymbol{\omega}_j$ and $\boldsymbol{\rho}_j$ are known (deterministic) given the context assignment $z_j$ and the context locations $\{\boldsymbol{\theta}_c\}$ (analogous for $\boldsymbol{\omega}_{j_1}$ and $\boldsymbol{\rho}_{j_1}$). Similarly, the vector of preferences $\boldsymbol{\beta}_{ij}$ is also deterministic given $\boldsymbol{\rho}_j$ and $\boldsymbol{\mu}_i$ ($\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\rho}_j$).

Figure 3 also helps to understand how the model uses the information within and across journeys to infer what customers are looking for. Consider a customer $i$ in a journey $j$ (focal journey) where we have observed $t$ steps that include clicks (and filters), if any. As an ongoing journey, purchase $y^p_{ij}$ has not been unobserved yet. How does the model leverage multiple sources of information to understand the customer need in this particular journey? First, the model learns from the information provided by customer $i$ during the course of the focal journey $j$. The queries inform to which context this focal journey most likely belongs to. Queries are a function of query parameters $\omega_j$; which, in turn, are determined by $z_j$, the context the journey belongs to. The context indirectly informs preferences $\beta_{ij}$ via $\rho_j$. Clicks (and filters) provide a strong signal of what the customer wants, as both outcomes are a direct function of preferences

$\boldsymbol{\beta}_{ij}$ (Equations 3 and 4). Moreover, clicks in the focal journey also inform about the products being currently considered in the journey.

Second, the model learns from queries, clicks, filters and purchases made in past journeys, with journey $j_1$ being one of them. Those clicks, filters and purchases are all driven by the corresponding preferences of those journeys, all of which share a customer-specific component, $\boldsymbol{\mu}_i$, with the preferences for each particular journey. In addition, past queries are related to the contexts on those journeys because the context shares explanatory power with the customer stable preferences $\boldsymbol{\mu}_i$ to describe clicks, filters, and purchases in those past journeys. That is, even though past contexts $z_{j_1}$ and customer stable preferences $\boldsymbol{\mu}_i$ are unconditionally independent, they are not independent when conditioning on past clicks and purchases, as these are co-determined by past context and stable preferences jointly. Consequently, past queries through past contexts are related to the preferences of the focal journey.

Third, the model learns from other customers whose journeys belong to the same context as the focal journey. As contexts are unobserved, the assignment of contexts is probabilistic and the information in other journeys will be shared greatly for those journeys that are more likely to belong to the same context as the focal journey. This information allows the model to learn the context location parameters $\boldsymbol{\theta}_c$.

## 3.5 Model estimation and inference: Predicting future activity

The model is estimated on historical data of finished journeys (which may or may not have ended up in a purchase). Once the model has converged, we can easily draw from the posterior to summarize the relevant parameters, including contexts and customer preferences. Importantly, we also want to predict purchase behavior for "ongoing" journeys, defined as those in which the customer might still be exploring/considering to purchase. The challenge in these cases is that consideration sets are not fully observed (as the customer can continue clicking for a while), and thus predicting purchases requires us to augment products that *would be considered* before we can predict what the eventual purchase would be. We describe such a calculation next.

### 3.5.1 Integrating over future unobservables

Consider an "ongoing" journey $j$ where we observe the query ($\mathbf{q}_{ij}$), clicks ($y^c_{ij1:t}$) and filters ($\mathcal{L}_{ijt} \subseteq \{1,...,L_{ij}\}$) up to step $t$. The posterior purchase probability of journey $j$ can be written as

$$p(y^p_{ij}|\mathbf{q}_{ij},y^c_{ij1:t},\mathcal{L}_{ijt}) = \int_{\boldsymbol{\beta}_{ij}} p(y^p_{ij}|y^c_{ij1:t},\boldsymbol{\beta}_{ij}) \cdot p(\boldsymbol{\beta}_{ij}|\mathbf{q}_{ij},y^c_{ij1:t},\mathcal{L}_{ijt}) \cdot d\boldsymbol{\beta}_{ij} \tag{16}$$

where $p(y^p_{ij}|y^c_{ij1:t},\boldsymbol{\beta}_{ij})$ is a conditional purchase probability given preferences and clicks up to step $t$, and $p(\boldsymbol{\beta}_{ij}|\mathbf{q}_{ij},y^c_{ij1:t},\mathcal{L}_{ijt})$ is the posterior distribution of the preferences given the partial information.

There are two probabilities inside the integral. Computing the latter term is straightforward as it comes directly from drawing from the posterior distributions of our model (see detail in Appendix C). However, computing the former quantity is more difficult because, to predict a purchase, we need to condition on the customer's consideration set, which might evolve *after* step $t$. We overcome this challenge by constructing an *augmented* (probabilistic) consideration set that includes products that may be clicked on for the remainder of the journey (i.e., after $t$). Specifically, we compute

$$p(y^p_{ij}|y^c_{ij1:t},\boldsymbol{\beta}_{ij}) = \int_{\mathcal{C}_{ij}} p(y^p_{ij}|\boldsymbol{\beta}_{ij},\mathcal{C}_{ij}) \cdot p(\mathcal{C}_{ij}|y^c_{ij1:t},\boldsymbol{\beta}_{ij}) \cdot d\mathcal{C}_{ij}, \tag{17}$$

where the first term is computed directly from the model specification as in (6), and the second term can be approximated by drawing the consideration set probabilistically, as we describe next.

### 3.5.2 Approximation of consideration probabilities

To compute the second term in (17), consider the probability that a product $k$ belongs to the consideration set at the end of the journey (i.e., right before purchase), given the clicks observed up to step $t$, and given preferences $\boldsymbol{\beta}_{ij}$ inferred before then, i.e., $p(k \in \mathcal{C}_{ij}|y^c_{ij1:t},\boldsymbol{\beta}_{ij})$. If the product has been clicked before step $t$, the probability to be considered is one. If the product has not been clicked before $t$, the probability of being considered afterwards involves an infinite sum because we do not know how many more steps the journey will have. To make such a forecast tractable and scalable, we approximate this probability by imputing the predictions from a flexible (reduced-form) model that, leveraging finished journeys in the training data, estimates the probability that a product will be added to the consideration set.[6] Specifically, we infer consideration given product characteristics $\mathbf{x}_{ijk}$ and preferences $\boldsymbol{\beta}_{ij}$ through a reduced-form predictor function $\hat{g}_{\mathcal{C}}(\mathbf{x},\boldsymbol{\beta})$ such that,

$$p(k\in\mathcal{C}_{ij}|y^c_{ij1:t},\boldsymbol{\beta}_{ij})\approx\begin{cases}1 & \text{if clicked on before, i.e., } \exists t'\leqslant t, y^c_{ijt'}=k,\\ \\ \hat{g}_{\mathcal{C}}(\mathbf{x}_{ijk},\beta_{ij}) & \sim .\end{cases} \tag{18}$$

Such a prediction function can be estimated using standard machine learning (ML) models trained using all displayed products in finished journeys of the training data, for which we precisely observe whether each product was added to the consideration set. In our application, we use a binary XGBoost model to estimate the function $\hat{g}_{\mathcal{C}}(\mathbf{x},\boldsymbol{\beta})$. As features for the XGBoost model, we use the exact observed product attributes used in the purchase model ($\mathbf{x}_{ijk}$) as well as draws from the posterior distribution of (individual-level) customer preferences ($\boldsymbol{\beta}_{ij}$), obtained from the main model when estimated using the same journeys in the training data. Note that adding customer preferences as features enriches the ML model predictions as those capture the unobserved individual-level preferences. We provide further details on the specification and performance of this reduced-form model in Web Appendix E.

Following (18), we can now approximate the conditional purchase probabilities in (17) using a Monte Carlo approximation where we draw consideration sets. In each iteration of our MC simulation, we form each consideration set by first including all products that have been clicked on up to that point. Subsequently, for all remaining products, we add them to the consideration set with a probability given by $\hat{g}_{\mathcal{C}}(\mathbf{x},\boldsymbol{\beta})$. Finally, once we have drawn the consideration set in each iteration of our simulation, we compute the purchase probabilities given each consideration set,[7] while all other products that do not belong to the consideration have a null probability of purchase. We outline such procedure in Web Appendix D, where we compute the purchase probabilities given a draw from the posterior distribution $p(\boldsymbol{\beta}_{ij}|\mathbf{q}_{ij},y^c_{ij1:t},\boldsymbol{\mathcal{L}}_{ijt})$.

[6] Theoretically, one could build here a forward-looking search model to estimate the consideration set probabilities. Given how non-linear our journeys are and the large number of possible items to query, filter, and click on, such a model would be intractable.

[7] Specifically, we use the GHK-algorithmGeweke (1991) to approximate purchase probabilities from a multinomial probit model given a consideration set.

# 4 Empirical setting

We apply our model to the context of airline ticket purchases using data from one of the largest worldwide online travel platforms. The dataset contains clickstream data on the focal platform of 4,500 customers who searched for flight tickets between May 2017 and November 2017. [8] For each web page shown to those customers, we observe the customer ID, the timestamp of when the customer accessed the page, the parameters of the search query associated with that page, and the list of results (including the flight attributes such as price, length, or airline carrier) observed by the customer after entering the query. The data also contains clicks on specific flights and the confirmation page in cases where the customer purchases a flight itinerary. We observe a total of 5,285,770 flight offers displayed in 120,614 results pages, which resulted in 3,718 flight itineraries purchased.

## 4.1 The first-party journey of airline travel

Consistent with our conceptualization of a first-party journey, the journey starts when the customer lands at the website's homepage to search for a flight. There are two types of trips that the customer can choose from: (1) Roundtrip, and (2) One-way.[9] For roundtrips, the customer includes an origin and a destination; a departure date for the portion of the trip from the origin to the destination, known as the *outbound leg*; and a returning date for the portion of the trip from the destination back to the origin, known as the *inbound leg.* Each leg of the trip is composed by either one non-stop flight or multiple connecting flights. One-way itineraries have only one direction of travel.

Note that, as is the case in numerous business settings, a customer journey can be highly non-linear (Grewal and Roggeveen, 2020). That is, the customer may go back from each step to enter a new/revised query, to click on alternative outbound or inbound results, etc. Moreover, this process does not need to occur during the same internet session, but can occur over the course of multiple days (Lee et al., 2018). Accordingly, we use the queries in our data to construct

[8] The focal firm randomly selected 4,500 customers among those that they define as "active" users.

[9] We drop from our analysis the third type of trip, multi-cities trips, as they constitute a very small portion of the trips. Moreover, our data does not contain searches on packages (e.g., flight + hotel), and therefore we focus on journeys over flights only.

a flexible definition of the customer journey by combining pages/sessions that belong to the same customer with the same "trip need". Specifically, a customer journey comprises all sessions that, while they might have occurred at different points in time, (1) have departing or arrival dates within up to 4 days; and (2) have origin or destination to close-by airports and cities within a 140 miles range (approx. 225 km.). Once all these pages are combined, we sort them by timestamp and remove subsequent searches of that same journey after a purchase is made, to remove the infrequent behavior of customers checking prices of the same itinerary after purchase. To avoid mislabeling censored and potentially unfinished journeys as no-purchase journeys, we remove all journeys (where the purchase has not been observed) in which the itinerary's first flight departs after our observation window's end. This process resulted in a total of 25,402 journeys, corresponding to an average of 5.6 journeys per customer. The conceptualization of journeys as described above, rather than simply using individual search queries as sessions, allows us to seamlessly integrate behavior across sessions that are aimed at covering the same need. Next, we describe the flow of the roundtrip purchase journey as one-way is a nested version of the roundtrip purchase journey.

After a customer lands on the homepage, they start specifying the search query (see Figure F.2a) by selecting the type of trip to search for (e.g., *roundtrip*) and filling multiple fields (all of them required): origin and destination cities/airports, outbound and inbound departing dates (i.e., "departing" and "returning dates" in Figure F.2a, respectively), and travelers (number of adults and children). The customer then clicks on the "Search" button, which triggers the platform to search the flight results that match the information from the query. The website displays the set of results, sorted increasingly by price, for the outbound itineraries (see Figure F.2b in Web Appendix F.1). Each of these itineraries are fully described by a path of flights that start at the origin airport and finish at the destination airport. The website clearly displays all relevant information of the outbound legs of the product search results, including price, the total duration of leg, the airline carrier, the number of stops, departing and arrival times. Note that for the outbound leg, the price displayed corresponds to the price of the complete roundtrip itinerary, including the price of the outbound leg and the cheapest inbound leg that corresponds to the outbound leg. At this point, the customer might want to explore some of the presented options (see “Select” below), click “More” to get exposed to more flights, “Filter” to restrict the characteristics of the flights to be shown, or abandon the search.

If the customer clicks on the “Select” button of one of the outbound offers, the website displays the set of corresponding inbound results that corresponds to the clicked outbound leg (see Figure F.2c). For those resulting inbound offers, the website displays the same level of information displayed for the outbound offers (see Figure F.2c), including the extra price of each alternative compared to the minimum price (i.e., the price displayed in the outbound page of results). Once the customer clicks on the “Select” button of one of the inbound offers, the website shows a page with the details of all the information mentioned before from both the outbound and the inbound legs (see Figure F.2d), as well as the full breakdown of the price (taxes and fees clearly displayed). After the customer clicks on “Continue Booking”, the customer fills in information about the passengers and proceeds with the payment steps. Finally, after finalizing the purchase, the customer is shown a confirmation page.[10] The one-way purchase journey is very similar, with the exception that instead of clicking through two sets of results (outbound and inbound), the customer is displayed only one page of results, “One-way results”.

## 4.2 Data preparation

### 4.2.1 Search queries

We construct several variables that aim to capture in more details the context of a journey in our model. While some pieces of information are directly provided by the customer (e.g., destination), others can be indirectly determined; e.g., whether the trip includes weekends can be extracted from the dates, or the trip distance (inferred from the origin and destination airports). We combine these variables into a vector of query variables, $\mathbf{q}_{ij}$ in (1), that aim to capture information about the journey in four different dimensions: (1) who is traveling, (2) which market this flight belongs to (origin-destination), (3) when is the trip, (4) when was the search made. Table 1 shows these variables and their corresponding summary statistics.

[10] While we do not observe the customer’s activity on the checkout page, we can observe if they were shown a confirmation page.

| **Query variable** | **Mean** | **SD** | **Quantiles** | | |
|---|---|---|---|---|---|
| | | | 5% | 50% | 95% |
| **Continuous** | | | | | |
| Trip distance (km.) | 3,584.16 | 3,465.07 | 448 | 2,269 | 11,529 |
| Time in advance to buy (days) | 50.73 | 59.82 | 1 | 29 | 182 |
| Length of stay (only RT) (days) | 11.80 | 21.25 | 2 | 6 | 37 |
| **Binary** | | | | | |
| Is it roundtrip? | 0.66 | . | 0 | 1 | 1 |
| Traveling with kids? | 0.08 | . | 0 | 0 | 1 |
| More than one adult? | 0.28 | . | 0 | 0 | 1 |
| Is it domestic?[a] | 0.59 | . | 0 | 1 | 1 |
| Is it summer season? | 0.37 | . | 0 | 0 | 1 |
| Holiday season? | 0.03 | . | 0 | 0 | 0 |
| Does stay include a weekend? | 0.66 | . | 0 | 1 | 1 |
| Flying from international airport? | 0.74 | . | 0 | 1 | 1 |
| Searching on weekend? | 0.21 | . | 0 | 0 | 1 |
| Searching during work hours? | 0.49 | . | 0 | 0 | 1 |
| **Categorical** | | | | | |
| Market | | | | | |
| US Only | 0.51 | . | 0 | 1 | 1 |
| US Overseas | 0.18 | . | 0 | 0 | 1 |
| Within North America[b] | 0.15 | . | 0 | 0 | 1 |
| Non-US within continent | 0.10 | . | 0 | 0 | 1 |
| Non-US across continent | 0.06 | . | 0 | 0 | 1 |
| Type of departure location | | | | | |
| Airport | 0.92 | . | 0 | 1 | 1 |
| Multi-airport City | 0.08 | . | 0 | 0 | 1 |

[a] We define domestic as flights between the US and Canada, as well as flights within the European Union (EU).

[b] This category includes Canada and Mexico and excludes US-only trips.

**Table 1:** Summary statistics of query variables

We observe a great variety of trip characteristics in the data: 66% of journeys are roundtrip (vs. one-ways); 28% include more than one adult and 8% include kids. The average stay for roundtrips is 11.80 days, 37% of journeys are for flights during the summer season and 3% for the holiday season,[11] and 66% of flight searches include stays over weekends. The average trip distance is 3,548 kilometers or 2,205 miles (e.g., approx. New York to Las Vegas); 59% of journeys are domestic (including US-Canada, within-EU, or within-country flights) and 51% are US Only. Purchase journeys occur, on average, 50.73 days prior to the departing date; 92% introduce a departing location code for an airport (e.g., JFK), 8% a departing code of a city (e.g., NYC), and the rest include a departing code that refers to both city and airport (e.g., MIA).

[11] We define the summer season from June 30th to September 4th, and holiday season stays that include either Thanksgiving, Christmas, or New Year's holidays.

### 4.2.2 Click and purchase occasions

Once the query information is captured, we build a set of "click occasions" faced by the customer. These click occasions are composed of a set of alternatives to click on and the outcome of what was actually clicked on (or not). There are two types of click occasions: (1) those on an outbound results page (where clicking on a product leads to an inbound results page) and (2) those on an inbound results page (where clicking on a product leads to flight detail page). We observe and allow in our model customers to click on multiple flights from each results page by adding a click opportunity with the same page results for each additional clicked product. By default, results within a page are sorted increasingly in price. Once the customer sees the results, they can sort the flights differently by clicking on sorting options, which include leg duration (hours) and departure/arrival time. While these alternative sorting actions could be valuable pieces of information for inferring preferences, unfortunately, we do not observe them explicitly because the firm records these actions as if the customer starts the search again, which is how we model them.

We create the consideration set, $\mathcal{C}_{ij}$ in (6), by including all products that were clicked before purchase (e.g., Bronnenberg et al., 2016). For roundtrip flights, only products where both the outbound and inbound legs of the itinerary were clicked on are added to the consideration set. We then register the outcome of the purchase occasion as a purchase for the product that was purchased (a purchase confirmation page), if any, or as a non-purchase in case no product was purchased.

### 4.2.3 Product attributes

Customers observe multiple product attributes when making a click and purchase decision. For a roundtrip journey, all attributes, except price, are specific to *each leg* of the trip. That is, there is a set of attributes that describe the outbound leg of the trip, and there is the same set of attributes that describe the inbound (returning) leg of the trip. Our model allows users to look for different attributes in a flight, depending on the leg.

A subset of these attributes is summarized in Table 2 (see Web Appendix F for full set of summary statistics). Prices are measured at the whole trip level. The average offer displayed is priced at \$1,547; but offers vary significantly in their price, with a standard deviation of \$3,249.

Furthermore, journeys have different price level that depends on origin-destination and the dates. This variation in price becomes clearer when looking at the price of the cheapest offer per journey. The cheapest price displayed per journey has an average of $698 across all customer journeys, with a standard deviation of $1,526. This indicates that raw prices may not be a good proxy to capture price sensitivity as prices are only compared within a journey. For example, a New York - Chicago roundtrip ticket for $600 may be considered expensive, whereas a roundtrip flight from New York to Buenos Aires for $800 may be considered a good deal. Therefore, we transform prices by computing the log difference between the focal flight and the lowest price per journey. Log transformation accounts for the long-tail dispersion in price differences.

| **Product attribute** | **Mean** | **SD** | **Quantiles** | | |
|---|---|---|---|---|---|
| | | | 5% | 50% | 95% |
| **Product level attributes** | | | | | |
| Price | 1,547 | 3,269 | 196 | 751 | 5,320 |
| Cheapest price per journey | 698 | 1,526 | 98 | 401 | 2,117 |
| **Outbound level attributes** | | | | | |
| Length of trip (hours) | 11.28 | 8.49 | 2.05 | 8.42 | 28.60 |
| Shortest length of trip per journey (hours) | 5.86 | 5.05 | 1.25 | 4.07 | 17.08 |
| Number of stops: Non stop | 0.20 | . | 0 | 0 | 1 |
| Alliance: OneWorld (American) | 0.27 | . | 0 | 0 | 1 |
| Alliance: Skyteam (Delta) | 0.27 | . | 0 | 0 | 1 |
| Alliance: Star Alliance (United) | 0.23 | . | 0 | 0 | 1 |
| Dep. time: Early morning (0:00am - 4:59am) | 0.04 | . | 0 | 0 | 0 |
| **Inbound level attributes** | | | | | |
| Length of trip (hours) | 11.08 | 9.02 | 1.83 | 7.92 | 29.50 |
| Shortest length of trip per journey (hours) | 6.17 | 5.31 | 1.25 | 4.27 | 17.75 |
| Number of stops: Non stop | 0.19 | . | 0 | 0 | 1 |
| Alliance: OneWorld (American) | 0.51 | . | 0 | 1 | 1 |
| Alliance: Skyteam (Delta) | 0.13 | . | 0 | 0 | 1 |
| Alliance: Star Alliance (United) | 0.15 | . | 0 | 0 | 1 |
| Dep. time: Early morning (0:00am - 4:59am) | 0.03 | . | 0 | 0 | 0 |

**Table 2:** Summary statistics of a subset of product attributes in page results

Offers also differ in terms of how long each leg of the trip is. The average outbound leg of a displayed trip takes 11.28 hours, with a large variation within and across journeys. The shortest flight per journey takes, on average, 5.86 hours for the outbound leg. Most displayed flights are one-stop flights (59%-70% for inbound or outbound legs), whereas nonstop flights account for 19%-20% of offers. As most of the variation in length is driven by the number of stops a leg has, we subtract from the length of a leg the length of the shortest itinerary with the same number of stops to isolate the effect of longer connections from more connections.

Airline data is sparse, so we aggregate airlines into alliances. Alliances are groups of airlines that share benefits and usually operate code-shared flights. For example, a JFK to Madrid flight operated by Iberia might also be sold by American Airlines, British Airways, and Finnair, all belonging to the same alliance. The three biggest alliances are Oneworld, SkyTeam, and Star Alliance, accounting for 77%-79% of all offers. Individual airlines not in an alliance but representing a significant proportion of offers are categorized as their own airline, while smaller airlines not in an alliance are grouped under "Other-No alliance. Finally, we label as "Multiple alliances" offers that have connecting flights of different alliances in the same leg of the trip.

### 4.2.4 Filters

Customers can filter the product results on a page by clicking on one (or multiple) product attribute levels: airline, number of stops, and departure and arrival times. In the data, we do not directly observe the use of filters. Instead, we infer the application of filters from the data by contrasting the product results shown to the customer at each step of the journey with the set of products available at the beginning of the journey. Number of stops is the attribute with the largest number of filter occasions in our data, with 13.8% of journeys having the customer filtering for non-stop flights at some point during the course of the journey. For all other attributes and levels, filters are applied very infrequently (less than 4% of journeys), supporting the model simplification of modeling filter outcomes once for the entire course of a journey as opposed to at the page-level. Web Appendix F.3 discusses the construction process of the filters and summary statistics of the filtering actions.

## 4.3 Summary statistics

Table 3 shows the the total number of customers, journeys, purchases, click steps and clicked products. We observe a total of 25,402 journeys, for which we aim to estimate individual-level preferences. The data indeed exhibit thin past purchase history at the individual level—while, on average, each customer undertakes 5.645 purchase journeys, the average number of purchases per customer is only 0.83. With low historical purchase rate at the customer level, the use of traditional models that rely on long individual purchase histories, such as scanner panel data, is limited. In many settings such as ours, the 1PD on purchases is thin, but because customers often use the platform for search, even when a purchase eventually does not occur, these searches can still provide valuable information. The amount of data collected during the journey is rich — on average, customers in our sample clicked on 1.14 products per journey, having a total of 6.45 clicks in total. These 1PD can become very valuable to the firm.

| **Variable** | **Total** | **Average per...** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Customer** | | **Journey** | | **Purchased journey** | | **Non-purchased journey** | |
| | | Mean | s.e. | Mean | s.e. | Mean | s.e. | Mean | s.e. |
| Customers | 4,500 | . | . | . | . | . | . | . | . |
| Journeys | 25,402 | 5.645 | 0.076 | . | . | . | . | . | . |
| One-way | 8,692 | 1.932 | 0.047 | . | . | . | . | . | . |
| Roundtrip | 16,710 | 3.713 | 0.056 | . | . | . | . | . | . |
| Purchases | 3,718 | 0.826 | 0.016 | 0.146 | 0.002 | 1.000 | . | 0.000 | . |
| Steps | 120,614 | 26.803 | 0.373 | 4.748 | 0.041 | 8.675 | 0.133 | 4.075 | 0.040 |
| ... in OW search | 40,184 | 8.930 | 0.262 | 4.623 | 0.068 | 6.127 | 0.155 | 4.174 | 0.074 |
| ... in RT outbound | 51,393 | 11.421 | 0.208 | 3.076 | 0.036 | 5.270 | 0.141 | 2.824 | 0.036 |
| ... in RT inbound | 14,441 | 3.209 | 0.046 | 0.864 | 0.012 | 2.547 | 0.053 | 0.671 | 0.011 |
| Clicked products | 29,037 | 6.453 | 0.072 | 1.143 | 0.014 | 2.945 | 0.048 | 0.834 | 0.013 |
| ... in OW search | 5,884 | 1.308 | 0.035 | 0.677 | 0.013 | 1.676 | 0.031 | 0.379 | 0.011 |
| ... in RT outbound | 14,441 | 3.209 | 0.046 | 0.864 | 0.012 | 2.547 | 0.053 | 0.671 | 0.011 |
| ... in RT inbound | 8,712 | 1.936 | 0.029 | 0.521 | 0.008 | 1.872 | 0.036 | 0.366 | 0.007 |
| Filtered attributes | 10,121 | 2.249 | 0.056 | 0.398 | 0.006 | 0.639 | 0.019 | 0.357 | 0.006 |
| ... in OW search | 4,654 | 1.034 | 0.039 | 0.183 | 0.004 | 0.381 | 0.016 | 0.149 | 0.004 |
| ... in RT | 5,467 | 1.215 | 0.035 | 0.215 | 0.005 | 0.258 | 0.013 | 0.208 | 0.005 |

**Table 3:** Data summaries, per customer and per journey.

On average, 14.6% of journeys end with a purchase. This number may seem high when compared to standard metrics of conversion for an online retailer, but note there are two caveats to this quantity. First, our data correspond to a sample of customers defined as “active” by the focal firm, and therefore this figure would be lower for the average customer of the firm. Second, in this paper, we adopt a broader definition of a journey, the 1PJ, which not only includes multiple sessions for the same customer but also combines searches that include nearby airports on similar dates, as the customer is trying to satisfy the same need. In contrast, traditional conversion rates tend to treat different search queries, with different variations of airports or dates, as different and independent purchase funnels.

## 4.4 Estimation and implementation

We split the data into training and validation at the customer and journey level: for each customer, we use some of their journeys for training and leave the last journey (or last few journeys) as a hold-out, such that we can explore the model performance in new journeys for existing customers. We leverage the data split in different ways. The training data is used to estimate the model and to summarize overall preferences and contexts in this market (Sections 5.1 and 5.2). We use held-out journeys to illustrate model inferences at different stages of the journey (Section 5.3), to evaluate the model's overall predictive ability (Section 5.4), and finally, to quantify the value of leveraging first-party journeys both in situations when customers are traceable and when they are not (Section 6).

We estimate the model parameters in a fully Bayesian framework using MCMC. Specifically, we use a blocked Gibbs sampler implemented in Julia to draw from the context posterior distribution with Pitman-Yor process priors following its stick-breaking representation (Ishwaran and James, 2001). This approach allows us to implement a fast sampler that is able to draw the context assignment $z_{ij}$ in parallel across journeys, given context locations and context size parameters; as opposed to using marginal samplers that marginalize the context distribution $F$ but sample sequentially for each journey conditioning on context assignments for all other journeys (Neal, 2000). We use adaptive Metropolis-within-Gibbs steps to draw the Pitman-Yor parameters $a$ and $d$, and we use Gibbs steps to draw the rest of the parameters using their full conditionals. We specify each of these steps in Appendix B. We estimate our model for 100,000 warm-up iterations and we use a sample of 1,000 draws from the posterior distribution (5,000 iterations saving a draw every 5 iterations). We assess convergence by monitoring traceplots over the model parameters.

# 5 Results

## 5.1 Average preferences for flight attributes

We start by describing mean level preferences ($\boldsymbol{\beta}_{ij}$). As these preferences vary across journeys, we compute the population mean estimates averaging $\boldsymbol{\beta}_{ij}$ across all in-sample journeys (Figure 4). As expected, customers prefer lower prices and flights of shorter lengths, they significantly prefer non-stops over one-stop, and tend to prefer the OneWorld alliance over all other alternatives.[12] They favor departure times either in the morning or in the afternoon for the outbound leg, while they prefer to depart in the afternoon or in the evening for the return leg.

[12] For categorical attributes, the base levels are: one-stop, OneWorld, and morning times.

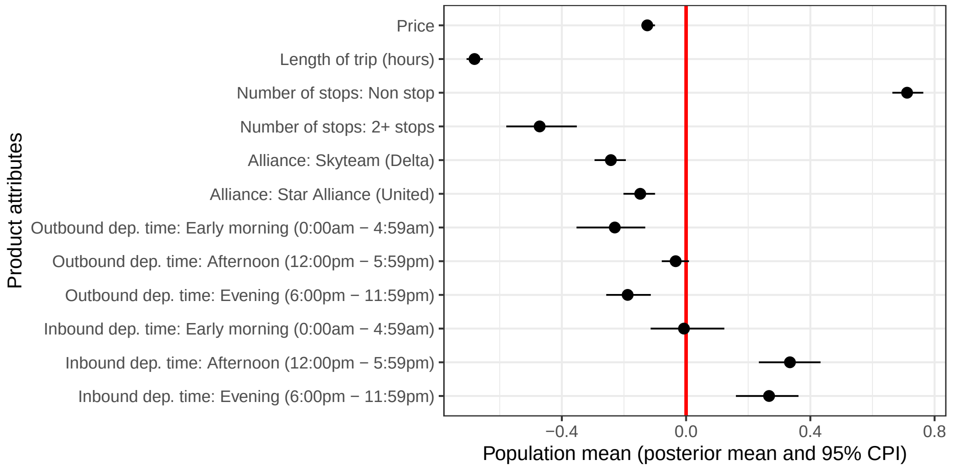
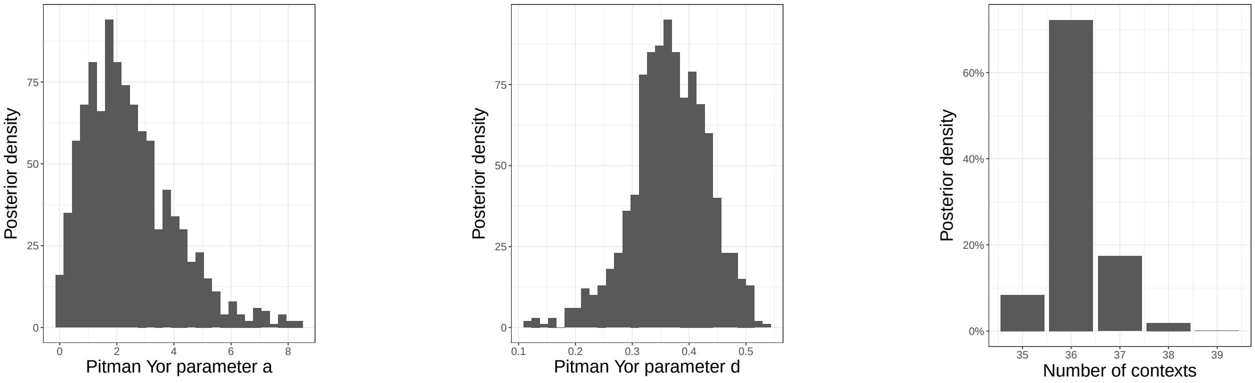


**Figure 4:** Population level estimates: Posterior Mean and 95% Credible Posterior Interval (CPI).

## 5.2 Contexts uncovered from the data

While the mean parameters are informative to explore the overall preferences for airline tickets, customers book flights to satisfy a wide variety of needs (i.e., a family vacation is not the same as a 2-day conference trip), for which preferences might also vary. We explore those differences by examining the rich set of contexts uncovered by our model. As explained in Section 3 and illustrated in Figure 1, we recover the number of contexts non-parametrically. Figure 5 shows the posterior distribution of the parameters of the Pitman Yor parameters as well as of the number of contexts (i.e., those with at least one journey assigned to them across the posterior distribution). While Figure 5 shows a maximum of 39 contexts in the data, when looking at the relative size of each context (Figure 6), we find that only 22 of these contexts appear in a journey with a probability higher than 1% (dotted line).

We focus on those 22 contexts next. The model parameters allow us to explore the type of trip (or "need") that each context represents. To interpret the model parameters we need to normalize the location parameters $\theta_c$ to compare both the query parameters and the flight preferences across the 22 contexts (see posterior statistics for all parameters of the 22 contexts in Web Appendix G.1). We normalize the location parameters to compare contexts with



**(a)** Posterior density of $a$

**(b)** Posterior density of the number of contexts

**(c)** Posterior density of the number of contexts

**Figure 5:** Posterior density of Pitman Yor parameters and resulting number of contexts

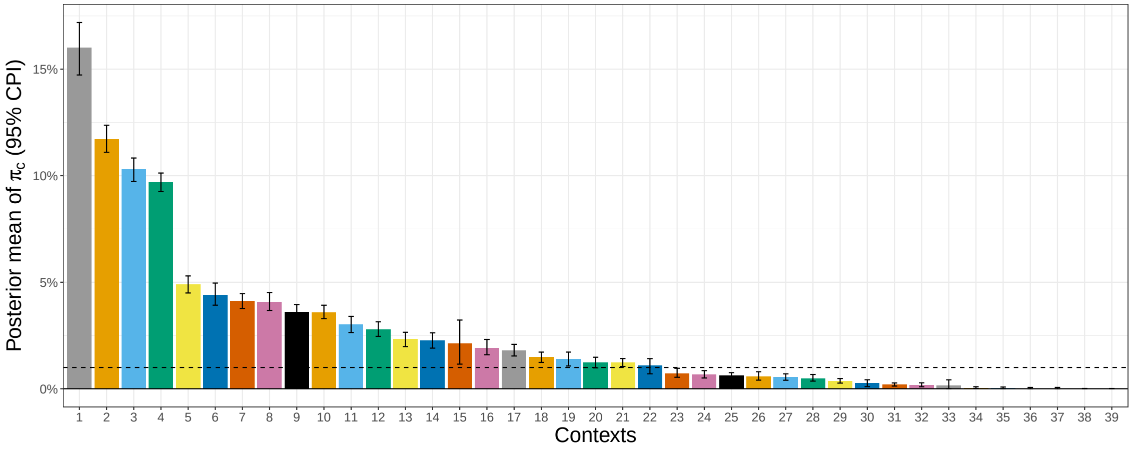


**Figure 6:** Posterior mean (and 95% CPI) of context probabilities $\pi_c$

respect to whether they score higher or lower than average on each of the query parameters and preferences. Figure 7 shows these relative scores for a subset of variables.[13]

Finally, using airport information, we identify the top most frequent 50 routes per context (Figure 8). Note that the exact destination is not included in the model and therefore not used to draw contexts from the data. However, we summarize this information for two main reasons. First, to explore the external validity of the contexts uncovered by the model. We find that destinations are largely congruent with the queries and preferences of each context (e.g., Hawaii is a common destination for week-length family trips with a strong preference for non-stop flights).

[13] For simplicity, we present the results for a subset of query parameters and product-attribute preferences. Details on the normalizing procedure and the full set of results are shown in Web Appendix G.2.

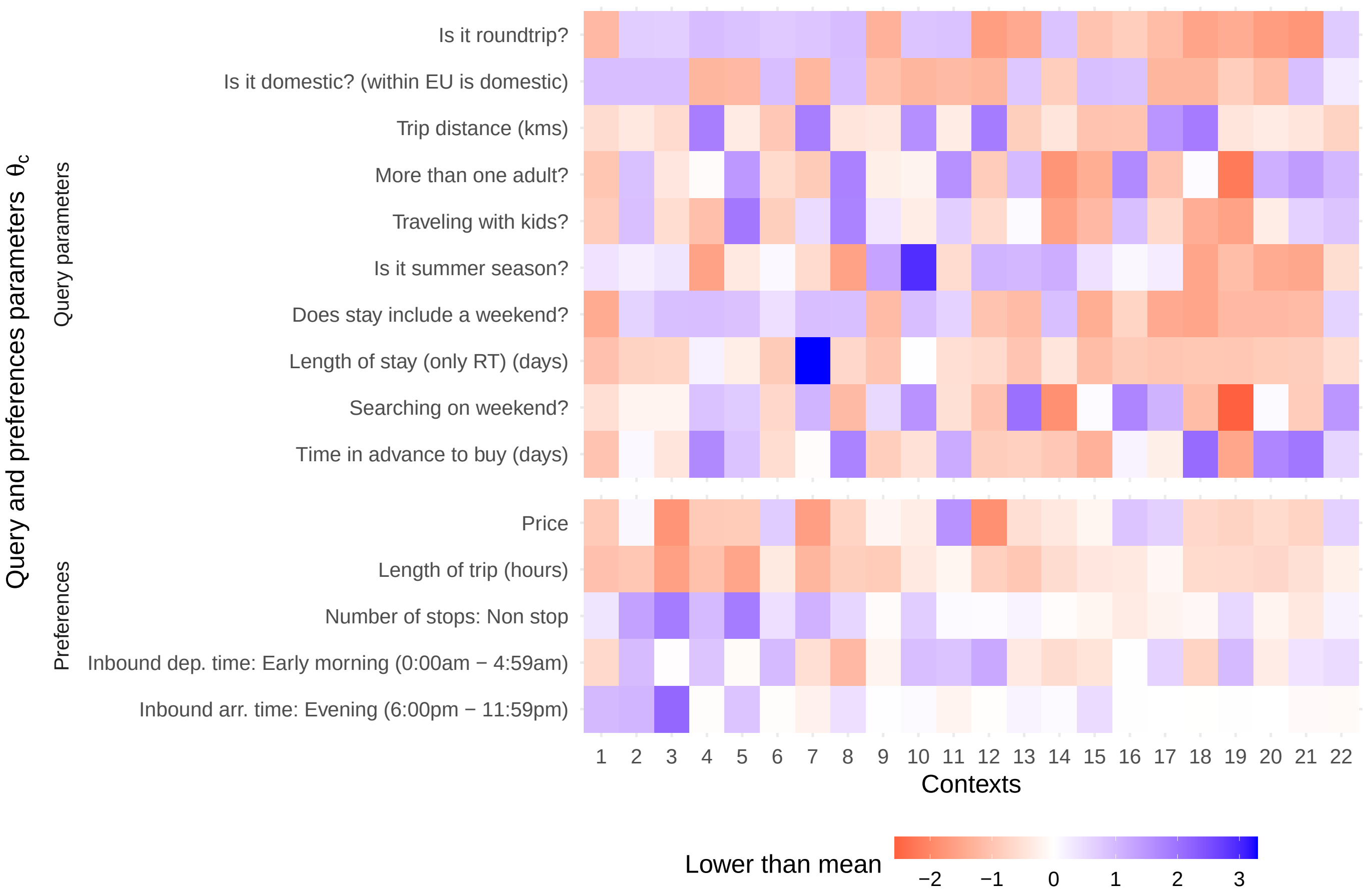
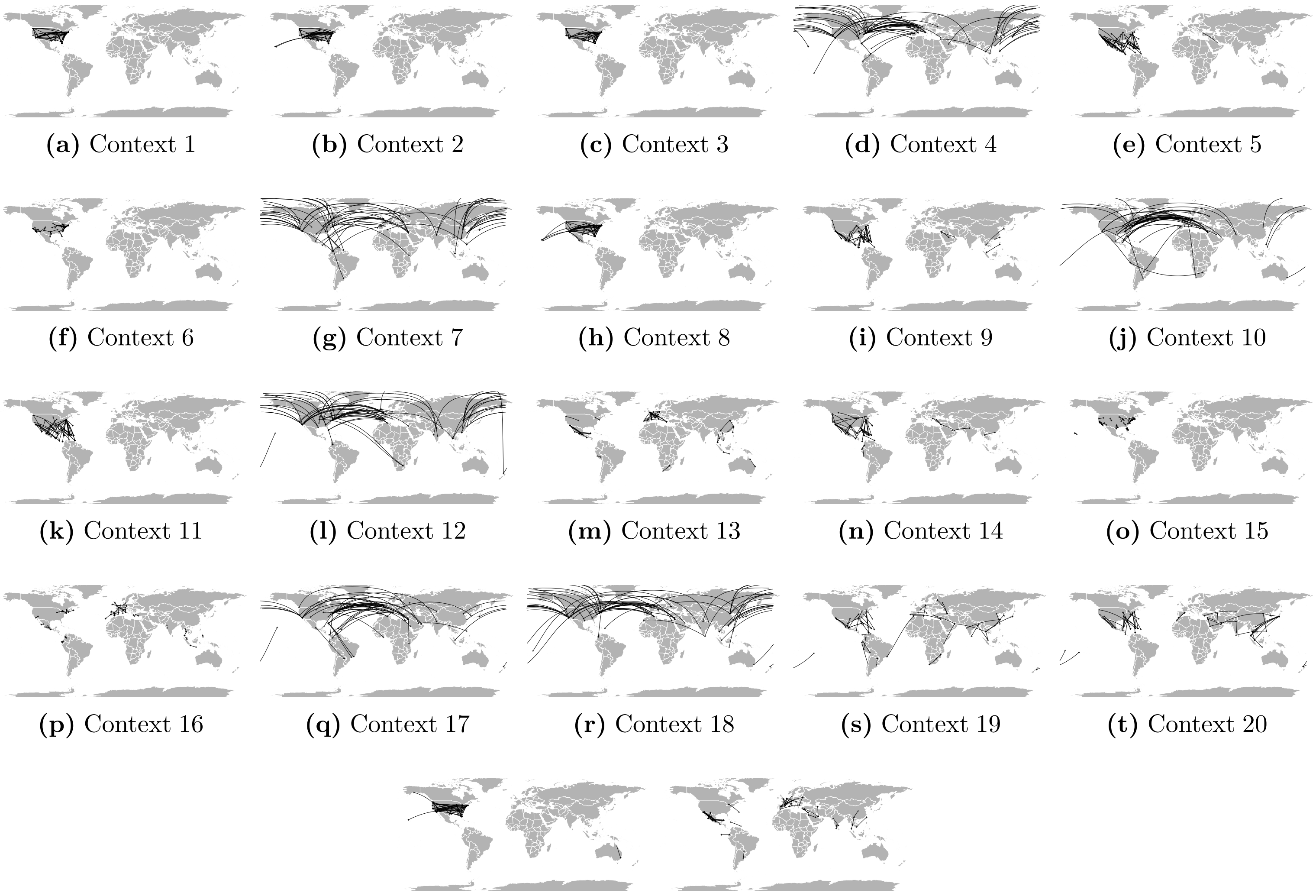


**Figure 7:** Posterior mean of context location parameters $\theta_c$, relative to the average in the population. The top figure shows how each context deviates from the average with respect to the query variables. The bottom figure shows deviations with respect to the preference parameters. Blue (red) boxes mean positive (negative) deviation from the average in the population.



**(a)** Context 1

**(b)** Context 2

**(c)** Context 3

**(d)** Context 4

**(e)** Context 5

**(f)** Context 6

**(g)** Context 7

**(h)** Context 8

**(i)** Context 9

**(j)** Context 10

**(k)** Context 11

**(l)** Context 12

**(m)** Context 13

**(n)** Context 14

**(o)** Context 15

**(p)** Context 16

**(q)** Context 17

**(r)** Context 18

**(s)** Context 19

**(t)** Context 20

**(u)** Context 21

**(v)** Context 22

**Figure 8:** Top 50 routes per context

Second, flight destinations might be useful to further characterize the contexts uncovered by the model. Accordingly, combining the insights from the model parameters and top destinations, we describe some of these contexts next:

- **Context 1 - One-way solo domestic trips (mostly US):** These journeys tend to be one-way and domestic, primarily involving no other adults or children. Typically, searches for these journeys initiate approximately 24 days before the departure date. Relative to the population, customers in these journeys are slightly more price sensitive, they have a stronger dislike for longer flights and routes with two or more stops, and have weaker preferences for routes with Spirit, Frontier, and multiple alliance flights.
- **Context 5 - Family vacations in the Caribbeans:** A prototypical journey in this context is a roundtrip route between the US and other North American or Caribbean countries. These journeys often involve additional adults and children, with an average stay of 9.8 days and searches beginning around 82 days before departure. There's a pronounced preference for Delta (Star Alliance), non-stop flights, and shorter routes.
- **Context 10 - 2-weeks long-haul summer trip:** This context captures customers searching for long-haul roundtrips, around 39 days prior to departure, with stays of approximately 13 days and almost exclusively over the summer season. Customers in this context are somewhat less price sensitive than the population, and also less sensitive to longer layovers.
- **Context 13 - Short Non-US Domestic flights:** This context seems to capture one-way domestic flights outside of the US. Most of these routes are within EU countries, in Mexico, or in east Asia. Customers in this context are as price sensitive as the average of the population and have stronger preferences for routes combining multiple alliances, but a strong dislike for carriers without any alliance (possibly avoiding independent low-cost airlines).

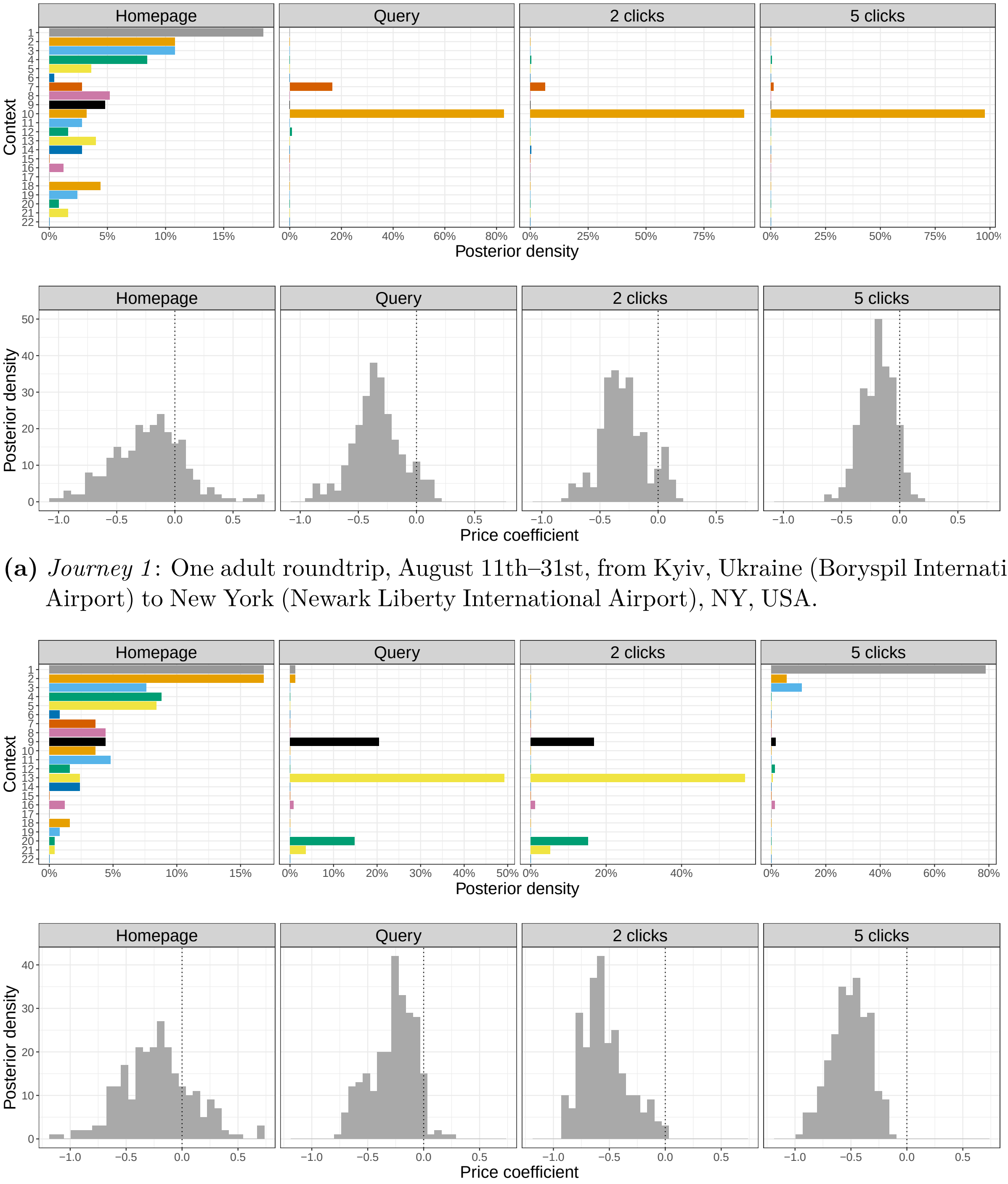
## 5.3 Learning *along* the journey

Another desirable characteristic of the model is that it enables the focal firm to update customer insights as users advance in their journey. To illustrate this process, we select a customer with two journeys in the holdout data and examine some of the model inferences at different stages of each journey. Figure 9 shows the model posterior distribution of context (top row) and price sensitivity (bottom row) for each journey, given the data available at four different stages: (1) at the homepage, (2) after the query was inserted, (3) after two clicks, and (4) after five clicks.

At the homepage, the model does not have any information about the focal journey. Hence, the inference for context corresponds to the average propensities across the population. The small differences between the first and second journeys are simply due to sampling error. Then, the model incorporates the query information and updates its inference about the context (first row, second column). For *Journey 1* ("One adult roundtrip from Kyiv, Ukraine to New York, USA"), the probability for contexts 7 and 10 notably increases whereas for *Journey 2* ("One adult oneway, Kyiv, Ukraine to Lisbon, Portugal") contexts 9, 13, and 20 become more likely to occur. After 2 click steps, the likelihood that *Journey 1* belongs to context 10 increases further, whereas the probabilities for *Journey 2* remain largely unchanged. Finally, after 5 steps, the model has more information about how the customer is clicking. Notably, *Journey 1* is clearly identified as belonging to context 10 (which we describe as the 2-week long-haul summer trip), whereas the inference for *Journey 2* changes drastically, with context 1 having the highest probability.

Similarly, we examine how the model infers price sensitivity as the customer moves across the journey (vertical histograms in second row). At homepage, both distributions are similar as both journeys belong to the same customer. After query, both distributions tighten, with a lower price coefficient for *Journey 2* than *Journey 1.* Interestingly, while both these journeys belong to the same user, their inferred price sensitivity differs between journeys, especially after observing some clicks. In particular, after 5 click steps, the model identifies that the customer is more price-sensitive for the flight to Lisbon than for the flight to New York. This is consistent with the fact that, across the population, journeys in context 1 exhibit stronger price sensitivity.

This illustrative example underscores the ability of the proposed model to seamlessly integrate information from diverse customer interactions *throughout* the 1PJ. By doing so, it extracts contexts that represent different customers' needs, and generates insights into



**(a)** *Journey 1*: One adult roundtrip, August 11th–31st, from Kyiv, Ukraine (Boryspil International Airport) to New York (Newark Liberty International Airport), NY, USA.

**(b)** *Journey 2*: One adult oneway, September 12th, from Kyiv, Ukraine (Boryspil International Airport) to Lisbon, Portugal (Humberto Delgado Airport).

**Figure 9:** Posterior context ($p(z_{ij}|\text{Data}_{ijt})$) and posterior price sensitivity ($p(\beta_{ij,\text{price}}|\text{Data}_{ijt})$) for two journeys of the same (sampled) customer at different stages of the journey, $t$: (1) homepage, (2) after query, (3) after 2 click steps, (4) after 5 click steps.

customers' preferences, empowering firms to continuously refine their understanding of what customers seek as they move along the journey.

## 5.4 Model predictive validity

Although Figure 9 demonstrates the model's inference update process, it doesn't provide insight into the quality or accuracy of these inferences. We address this aspect next by assessing the model's predictive efficacy. While the primary objectives of the model extend beyond predictions, confirming the model's precision in forecasting customer behavior serves as an indicator of its ability to harness and meaningfully leverage the data collected throughout the first-party journey.

For this exercise, we focus on predicting whether a transaction occurs at the end of each (held-out) journey, and if so, which product is chosen. We compare our model with other established methodologies that have been extensively tested for such predictive tasks. We evaluate predictive validity of our model on two distinct occasions: immediately after the customer submits the query (i.e., before any clicks are observed) and once the customer has clicked on five occasions. These time points (i.e., after query and after clicks) dictate the available information for making predictions, both for our model and the benchmark models.[14]

*Benchmarks of comparison*

We compare our model to two commonly used machine learning classification methods, Random Forest (RF) and XGBoost. As these benchmarks can only be used to predict outcomes given a set of features, we create a comprehensive set of features that capture the information available (to the firm) at each prediction point. In addition to the attributes of the corresponding product, our set of features includes summary statistics from: (1) the query of the Focal journey (capturing the attributes of the search), (2) the attributes of the products shown in the first page of the Focal journey (capturing the attributes of the products presented to the consumer), (3) the clicks and filters (during the focal journey) up to the moment when the prediction is made (capturing what the customer has been clicking), and (4) the queries,

[14] When a (held-out) journey has less than 5 clicks occasions, we use all clicks prior to purchase. Doing so not only provides a more conservative measure of the predictive improvements but also avoids selection biases due to shorter and longer journeys. Also, results are qualitatively similar when using different numbers of click occasions; see Web Appendix G.4 for results when using fewer clicks.

product attributes, clicks, filters, and purchases from past journeys (capturing past behavior). (See details about the features and the estimation of benchmarks in Web Appendix G.3)

Machine learning classification models are often built for binary tasks. In our case, not only is the classification not binary, but the choice set size (the number of possible available flights to choose from) varies from one classification task (journey) to another. To address this complication, we create a series of binary classifications and use normalization to convert these to a multinomial choice task. Specifically, we train each benchmark model as a binary predictive model where each observation represents choosing one product in a particular journey (for each journey, we include an additional no-purchase "product" for the outside option). Before making predictions, we normalize the prediction scores per customer per journey such that the scores for all alternatives (in each journey) sum to one. To predict if the customer buys any product or does not buy, we label it as a purchase if the normalized score for the no-purchase alternative is lower than 0.5. For the product choice task, we normalize the predictive scores of each product by dividing by the sum of the scores of all products except the outside option (i.e., choice conditional on purchase) and label as the chosen product, the product with the highest score. For this evaluation, we only consider (held-out) journeys that ended up in a purchase, for which we observe the actual product choice.

*Measures of predictive performance*

Due to the predominant occurrence of non-purchase outcomes in the majority of the journeys, we evaluate predictive ability in purchase incidence based on *balanced accuracy* (Brodersen et al., 2010) as it provides a more reliable measure of model performance when classes are imbalanced. It is calculated as the average of sensitivity/recall (true positive rate) and specificity (true negative rate) and therefore ranges from 0 to 1, where a value of 1 indicates perfect prediction performance. Additionally, we report the commonly used precision and recall measures. For product choice given incidence, we report hitrate (proportion of journeys that were correctly predicted) and balanced accuracy at the product level. (Results are shown in Table 4, please refer to Web Appendix G.4 for further details and other measures of predictive ability.)

| Model | Incidence |  |
|---|---|---|
|  | After query | After 5 steps |
| **Balanced accuracy** |  |  |
| Proposed model | 0.62 | 0.65 |
| Random forest | 0.60 | 0.70 |
| XGBoost | 0.50 | 0.59 |
| **Precision** |  |  |
| Proposed model | 0.21 | 0.22 |
| Random forest | 0.28 | 0.33 |
| XGBoost | n/a* | 0.60 |
| **Recall** |  |  |
| Proposed model | 0.83 | 0.91 |
| Random forest | 0.40 | 0.67 |
| XGBoost | 0.00 | 0.20 |

*: Model predictions were zero for all units.

**(a)** Purchase incidence of proposed vs. benchmark models.

| Model | Product choice given purchase |  |
|---|---|---|
|  | After query | After 5 steps |
| **Hitrate** |  |  |
| Proposed model | 0.16 | 0.62 |
| Random forest | 0.16 | 0.19 |
| XGBoost | 0.03 | 0.62 |
| **Balanced accuracy** |  |  |
| Proposed model | 0.58 | 0.81 |
| Random forest | 0.58 | 0.59 |
| XGBoost | 0.51 | 0.81 |

**(b)** Choice given purchase of proposed vs. benchmark models.

**Table 4:** Prediction of proposed vs. benchmark models.

While there is no clear dominant model in the predictive ability across the two tasks (incidence and product choice) and the two journey steps (after queries and after 5 steps), the proposed model is either the most predictive or close to the most predictive across all four "cells". Specifically, while the RF model predicts well the incidence task, particularly after 5 steps, it does quite poorly in predicting the product chosen after 5 steps. The XGBoost does as well as the proposed model in predicting the product chosen after 5 steps, but it performs no better than chance (balanced accuracy of 50%) in predicting both incidence and the product chosen after query.

Taken together, these results indicate that our model performance at forecasting journey outcomes is on par, and overall more robust across prediction task, relative to the machine learning models commonly used for these predictive tasks. The performance of the proposed model relative to the machine learning models is impressive because the machine learning models were specifically trained to predict each of the task, whereas the proposed model was trained to piece together sources of information to customer needs and the context of the focal customer journey. This finding suggests that our proposed model effectively captures and incorporates relevant information from the first-party data. It does so by incorporating in a unified framework the multiple pieces of information that customers leave behind as they aim to satisfy different needs. This modeling framework enables firms to uncover hidden contexts from the data as well as infer, and update, what customers might be looking for as they evolve in their journey. Such inference will be lost with a "black box" machine learning predictive model. In the next section, we take a broader view of the model predictions to quantify the overall value of the first-party journey along different dimensions.

# 6 The value of first-party data

We quantify the value of 1PD at three different levels. First, we assess the value of the *current journey.* This analysis highlights the benefits of incorporating customer information "live" into a firm's systems. Second, we evaluate the incremental value of integrating *prior-to-purchase data* with the current journey data. This quantification is important because traditional approaches that use sales or scanner-panel data models only rely on actual transactions, which can become very thin when customers purchase the product infrequently. 1PD collected along the customer journey (including both current and past journeys) can provide a much richer understanding of customer behavior that can potentially help firms optimize their marketing strategies to better meet customer needs. Third, we analyze the value of *tracking customers* across visits. As customer privacy is becoming a key priority, it is crucial for firms to quantify the value of logins and cookies, which enable firms to identify customers across different journeys.

## 6.1 Empirical strategy

Before presenting our findings, we discuss how we conduct this examination. The primary approach is to utilize our full model, along with nested versions that utilize subsets of the 1PD to predict customer choices. We use a set of held-out journeys to mimic assessing customer needs in unforeseen journeys. We predict customer choices at three "moments" in the journey: (1) when they insert the query, (2) after they make two clicks, and (3) after they make five clicks. This examination expands the analyses presented in Section 5.4 not only by considering more moments along the journey, but also by exploring more broadly the types of products (or attributes) customers are looking for.

For this analysis, we select the (held-out) journeys for which we observe a purchase and use the characteristics of the flight chosen as the "ground truth" of what the customer was looking for. Consistent with the analysis presented in Section 5.4, we compute the probability that the customer would select the (chosen) product based on the model's parameters at each point in time. We also explore the model's ability to predict the main attributes of the chosen flight (namely alliance, number of stops, price, and outbound length), which might represent a more realistic goal to the focal company. We use hitrates for product choice, alliance and the number of stops (as they are categorical variables); and RMSE (Root Mean Squared Error) for price and length, which are continuous. Web Appendix G.5 provides further details about predictions' calculations.

| | **After...** | | |
|---|---|---|---|
| | **Query** | **2 clicks** | **5 clicks** |
| **Product choice** (hitrate, ↑ better) | | | |
| Full model | 0.16 | 0.27 | 0.62 |
| % dif. vs. query | | (+69.88%) | (+289.16%) |
| Only purchase | 0.07 | 0.07 | 0.07 |
| No customer tracking | 0.10 | 0.29 | 0.62 |
| **Alliance** (hitrate, ↑ better) | | | |
| Full model | 0.44 | 0.55 | 0.76 |
| % dif. vs. query | | (+25.33%) | (+73.36%) |
| Only purchase | 0.39 | 0.39 | 0.39 |
| No customer tracking | 0.32 | 0.46 | 0.71 |
| **Number of Stops** (hitrate, ↑ better) | | | |
| Full model | 0.59 | 0.62 | 0.78 |
| % dif. vs. query | | (+5.84%) | (+32.14%) |
| Only purchase | 0.43 | 0.43 | 0.43 |
| No customer tracking | 0.39 | 0.50 | 0.72 |
| **Price** (RMSE, ↓ better) | | | |
| Full model | 0.96 | 0.92 | 0.81 |
| % dif. vs. query | | (-4.13%) | (-15.30%) |
| Only purchase | 1.16 | 1.16 | 1.16 |
| No customer tracking | 1.15 | 1.07 | 0.91 |
| **Length** (RMSE, ↓ better) | | | |
| Full model | 0.88 | 0.84 | 0.73 |
| % dif. vs. query | | (-4.16%) | (-16.52%) |
| Only purchase | 1.11 | 1.11 | 1.11 |
| No customer tracking | 1.20 | 1.10 | 0.93 |

**Table 5:** Model's accuracy at predicting the exact product chosen as well as the main attributes of the (chosen) product, evaluated at different stages of the journey. Prediction measures: Hitrate (product, alliance, and number of stops) and RMSE (price and length). Performance of `Full` vs. `Only purchase` vs. `No customer tracking` models.

## 6.2 The value of data from the current journey

We start by measuring the value of the data collected during the *current journey*. Table 5 shows how the model significantly improves its accuracy at inferring what the customer is looking for as it collects information from the current journey. While this finding should not be surprising — as customers click, they often get closer to purchase — it is reassuring that the model is able to integrate such information in an effective manner. Across all cases, the model's ability to predict what customers are looking for increases notably after observing a few clicks (e.g., the hit rate for the product chosen increases by 69.88% just after two clicks, and by 289.16% after five clicks) relative to the query step. Exploring the main product attributes in this setting, we find that the (incremental) value of current journey clicks seems particularly strong when inferring what alliance the customer may choose, with hit rates increasing from 0.44 at query and 0.76 after five clicks.

## 6.3 The value of prior-to-purchase data

To quantify the incremental value of prior-to-purchase data, we compare the `Full` model against a traditional hierarchical model of purchase, labeled `Only purchase` model, that uses purchase data exclusively while accounting for customer heterogeneity ($\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i$). Such a benchmark not only ignores the click information collected in the past, but also fails to update information along the current journey. This model is similar to models often used for scanner panel data (e.g., Allenby and Rossi, 1998). With such an analysis we want to highlight the importance of fusing the multiple behaviors observed along the journey; behaviors that can be easily collected by firms but are sometimes not stored for past or even the current journey.

The results (presented in Table 5) suggest that there is significant value in the information collected prior to purchase throughout all stages of the customer journey. Specifically, the hit rate for the product chosen after query is more than twice as large for the `Full` model (0.16) than for the `Only purchase` model (0.07). Additionally, the `Only purchase` model consistently predicts 15%-20% worse than the `Full` model for flight attributes after query. These findings indicate that queries, clicks, and filters from *prior journeys* allow the model to make better predictions even before leveraging the click and filter information from the current journey.

Looking across columns, we observe that the benefit of utilizing prior-to-purchase data is even more pronounced when incorporating query and clicks from the current journey. For example, the hit rate for product chosen after five clicks becomes as high as 0.62, which is about 9 times larger than when using only purchase data.

Overall, this analysis underscores the significance of storing and effectively utilizing customer clickstream data beyond purchases not only in search models of the current journey but also across journeys. Prior work has often failed to incorporate such cross-journey data. As exemplified in our empirical context, the integration of prior-to-purchase data within a modeling framework can substantially augment the predictive capacity of the model and improve firms' understanding of customer preferences and needs, particularly when customer purchase history is thin.

## 6.4 The value of tracking customers

Finally, to measure the value of tracking users, we contrast the performance of the `Full` model against a nested version labeled `No customer tracking`, where the customer identity is unknown. This (nested) model considers every journey as if it belongs to a "new" customer (for whom we don't know individual preferences) as journeys cannot be attributed to a specific customer,[15] and can either represent a situation of customers using the platform without logging in (and no cookie tracking) or a privacy scenario where the firm is not allowed to store customer historical data. This comparison helps us quantify the losses incurred when/if the company loses its ability to identify customers. Importantly, we compare both models at different moments of the journey, enabling us to measure the trade-off between the information lost due to the inability to track users and the information gained from the ability to collect and integrate data along the current journey.

Not surprisingly, not being able to track users hinders the models' ability to infer what they might be looking for (i.e., `No customer tracking` underperforms `Full` in all cases). The model particularly benefits from identifying a customer when making inferences early in the journey. Across all attributes, the full model performs substantially better than a model with no tracking information at query (e.g., a 37.5% increase on predicting Alliance: from 32% hit rate when using the `No customer tracking` model to 44% when using the `Full` model). The disparity diminishes significantly as more clicks are observed in the current journey. After five clicks, even when the firm is unable to identify consumers, the model performs comparably to the model

[15] For this model, we keep the specification of the model, $\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\rho}_j$, where the stable preferences term $\boldsymbol{\mu}_i$, in this case, captures journey-specific idiosyncratic preferences not explained by context. This allows the model to still learn, beyond context, from each click in the current journey.

with complete information, achieving a 62% hit rate in product choice. This underscores the importance of incorporating customer interactions throughout the entire journey, particularly in cold-start scenarios where the company lacks prior information about the customer. However, even when the model cannot precisely predict the chosen product, leveraging past journeys enhances its ability to discern the type of product the customer seeks (e.g., for Alliance 76% hitrate for the `Full` model vs. 71% for the `No customer tracking` model; and for stops 78% hitrate for the `Full` model vs. 72% for the `No customer tracking` model.)

Taken together, these results show great value on collecting and integrating real-time 1PD, in combining those data sources across journeys, even when customer cannot be tracked. In turn, the results underscore that, in settings where cookies are no longer available to track individual users, the model's ability to leverage information along the current journey (combined with past journeys from "anonymous" customers) can compensate for the loss in firm's ability to identify the individual customer.

# 7 Conclusion

We propose a probabilistic non-parametric Bayesian machine learning model to integrate the information collected along the first-party journey and to combine such information both across customers and across journeys. The model capitalizes on within-journey information to identify distinct journey-specific preferences. Past journeys' information, including purchases, searches, and clicks, aids in inferring stable customer preferences. Behavior across customers enriches information by integrating data from customers with analogous context-specific preferences. This approach not only helps firms to infer what customers are looking for, even in settings where purchases occur infrequently, but also addresses the so-called "cold start problem," offering a solution to compensate the lack of historical data with real-time data collected along the 1PJ.

Applying the model to data from an online travel platform highlights its ability to extract essential aspects of customer preferences, like airline choices, flight stop preferences, and price considerations. The model demonstrates remarkable adaptability, continuously updating inferences as customers progress in their journeys, a contrast to traditional panel-data models that rely solely on purchase data. Compared to collecting only purchase or sales data, our model enhances predictive accuracy in product selection by tenfold, underlining the significance of pre-purchase data, which includes clicks and filters. Furthermore, our modeling framework gauges data loss value under heightened data privacy scenarios, illustrating how leveraging information across first-party journeys can compensate for reduced customer identification capabilities. This becomes increasingly relevant as platforms allow searches without requiring login credentials, or are required to limit the storing of customers' data.

There are several limitations of our research. First, the proposed model's complexity might pose challenges in terms of computational efficiency and scalability, especially when applied to large datasets. Computationally faster approximation strategies to the posterior distribution of customer preferences (e.g., point estimates or variational approximations) may be necessary to enhance its practicality for real-time applications or extensive datasets. Additionally, modern inference approaches, like amortized variational inference (Kingma and Welling, 2013; Agrawal and Domke, 2021), could be used to rapidly update the posterior of journey-specific preferences. Second, while our research showcases the advantages of incorporating 1PD, it does not account for external factors that might impact customer behavior, such as broader economic trends or unexpected events. This could limit the model's ability to leverage past journeys to augment the limited data available early in the customer journey (e.g., at query or just after a few clicks). A promising avenue for future research would be to explore ways of incorporating broader dynamics (or shocks) by enabling contexts to evolve in their prototypical behaviors. This adaptation would enable the model to leverage past journeys while accommodating significant shifts in preferences.

Third, due to infrequent purchase behavior in our data, we assume common context distribution across customers. In settings with more frequent journeys, there may be long enough journey histories that allow firms to infer customer-specific contexts. Finally, alternative empirical settings, where journeys on multiple categories are jointly observed, could allow for the development of models that can not only combine information across journeys from a customer and across customers within a category but also integrate information across categories. We hope that these limitations will stimulate further studies into leveraging first party customer journeys and its implications for marketing strategies.

In conclusion, we aim to bring to the forefront of customer relationship management and choice modeling research the opportunity to fuse historical purchase and click-stream data with current first-party journey data. Taking such a point of view is particularly needed with the evolving challenges posed by data privacy concerns and the dynamic nature of customer journeys.

# References

Agrawal, A. and Domke, J. (2021). Amortized variational inference for simple hierarchical models. *Advances in Neural Information Processing Systems*, 34:21388–21399.

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1-2):57–78.

Ansari, A. and Mela, C. F. (2003). E-Customization. *Journal of Marketing Research*, 40(2):131–145.

Aridor, G., Che, Y.-K., and Salz, T. (2020). *The economic consequences of data privacy regulation: Empirical evidence from gdpr*. National Bureau of Economic Research Cambridge, MA, USA.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828.

Boughanmi, K. and Ansari, A. (2021). Dynamics of musical success: A machine learning approach for multimedia data fusion. *Journal of Marketing Research*, 58(6):1034–1057.

Braun, M. and Bonfrer, A. (2011). Scalable Inference of Customer Similarities from Interactions Data Using Dirichlet Processes. *Marketing Science*, 30(3):513–531.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Bronnenberg, B. J., Kim, J. B., and Mela, C. F. (2016). Zooming In on Choice: How Do Consumers Search for Cameras Online? *Marketing Science*, 35(5):693–712.

Bruce, N. I. (2019). Bayesian nonparametric dynamic methods: Applications to linear and nonlinear advertising models. *Journal of Marketing Research*, 56(2):211–229.

Chen, Y. and Yao, S. (2017). Sequential Search with Refinement: Model and Application with Click-Stream Data. *Management Science*, 63(12):4345–4365.

De los Santos, B. and Koulayev, S. (2017). Optimizing Click-Through in Online Rankings with Endogenous Search Refinement. *Marketing Science*, 36(4):542–564.

Dew, R., Ansari, A., and Li, Y. (2020). Modeling dynamic heterogeneity using gaussian processes. *Journal of Marketing Research*, 57(1):55–77.

Donnelly, R., Kanodia, A., and Morozov, I. (2023). Welfare effects of personalized rankings. *Marketing Science*.

Duvvuri, S. D., Ansari, A., and Gupta, S. (2007). Consumers' price sensitivities across complementary categories. *Management Science*, 53(12):1933–1945.

Feit, E. M., Wang, P., Bradlow, E. T., and Fader, P. S. (2013). Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *Journal of Marketing Research*, 50(3):348–364.

Fiebig, D. G., Keane, M. P., Louviere, J., and Wasi, N. (2010). The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science*, 29(3):393–421.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, volume 571, page 578. Fairfax, Virginia: Interface Foundation of North America, Inc.

Ghose, A., Ipeirotis, P. G., and Li, B. (2019). Modeling Consumer Footprints on Search Engines: An Interplay with Social Media. *Management Science*, 65(3):1363–1385.

Goldberg, S., Johnson, G., and Shriver, S. (2019). Regulating privacy online: An economic evaluation of the gdpr. *Available at SSRN 3421731*.

Grewal, D. and Roggeveen, A. L. (2020). Understanding retail experiences and customer journey management. *Journal of Retailing*, 96(1):3–8.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. (2016). Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Honka, E. and Chintagunta, P. (2017). Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry. *Marketing Science*, 36(1).

Huberman, E. (2021). First-party data collection is more crucial than ever. *Forbes*, December 21.

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173.

Iyengar, R., Ansari, A., and Gupta, S. (2003). Leveraging information across categories. *Quantitative Marketing and Economics*, 1(4):425–465.

Jacobs, B. J., Donkers, B., and Fok, D. (2016). Model-Based Purchase Predictions for Large Assortments. *Marketing Science*, 35(3):389–404.

Kim, J. B., Albuquerque, P., and Bronnenberg, B. J. (2010). Online Demand Under Limited Consumer Search. *Marketing Science*, 29(6):1001–1023.

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2004). Assessing Heterogeneity in Discrete Choice Models Using a Dirichlet Process Prior. *Review of Marketing Science*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Korganbekova, M. and Zuber, C. (2023). Balancing User Privacy and Personalization. Working paper.

Latvala, L., Horn, J., and Bruno, B. (2022). Thriving in the age of privacy regulation: A first-party data strategy. *Applied Marketing Analytics*, 7(3):211–220.

Lee, L., Inman, J. J., Argo, J. J., Böttger, T., Dholakia, U., Gilbride, T., Van Ittersum, K., Kahn, B., Kalra, A., Lehmann, D. R., et al. (2018). From browsing to buying and beyond: The needs-adaptive shopper journey model. *Journal of the Association for Consumer Research*, 3(3):277–293.

Liu, J. and Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):930–952.

Montgomery, A. L., Li, S., Srinivasan, K., and Liechty, J. C. (2004). Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science*, 23(4):579–595.

Morozov, I., Seiler, S., Dong, X., and Hou, L. (2021). Estimation of preference heterogeneity in markets with costly search. *Marketing Science*, 40(5):871–899.

Murphy, K. (2022). With first-party data, marketers are finally in the driver's seat. *Harvard Business Review Online*, September 27.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Padilla, N. and Ascarza, E. (2021). Overcoming the cold start problem of customer relationship management using a probabilistic machine learning approach. *Journal of Marketing Research*, 58(5):981–1006.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Rossi, P. E., McCulloch, R. E., and Allenby, G. M. (1996). The Value of Purchase History Data in Target Marketing. *Marketing Science*, 15(4):321–340.

Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the tenth international conference on World Wide Web - WWW '01*, volume 3, pages 285–295, New York, New York, USA. ACM Press.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., and Yu, Y. (2017). Protecting customer privacy when marketing with second-party data. *International Journal of Research in Marketing*, 34(3):593–603.

Seiler, S. (2013). The impact of search costs on consumer behavior: A dynamic approach. *Quantitative Marketing and Economics*, 11(2):155–203.

Sun, T., Yuan, Z., Li, C., Zhang, K., and Xu, J. (2023). The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy. *Management Science*.

Tian, L., Turjeman, D., and Levy, S. (2023). Privacy Preserving Data Fusion. Available at SSRN: https://ssrn.com/abstract=4451656.

Ursu, R. M. (2018). The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science*, 37(4):530–552.

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science*, 66(3):1045–1070.