Bridging Choice Theory and Recommender Systems: A Unified Framework for Latent Preference Discovery and Algorithmic Diagnosis in Recommender Systems

Aniket Sakpal
Expedia Group
USA
asakpal@expedia.com

Adam Domanski
Expedia Group
USA
adomanski@expedia.com

Jean-Louis Guillemardet
Expedia Group
USA
jguillemardet@expedia.com

Abstract

Modern search and recommendation systems achieve high predictive accuracy yet remain largely opaque: they optimize aggregate click or conversion metrics without revealing why different users choose different items or which user segments the current ranking systematically under-serves. We present a principled framework that addresses this gap by applying Discrete Choice Models (DCM)— tools from structural econometrics — to the search ranking and recommendation context. Grounded in Information Foraging Theory, our framework decomposes search failure into information problems (surfacing the wrong items) and friction problems (surfacing the right items but in the wrong way), and uses DCM to quantify the marginal utility each item attribute contributes to user choice. An extension to DCM - the latent class discrete choice model (LCDCM) - then uncovers latent preference segments within the user population then uncovers latent preference segments within the user population — groups that the population-average model obscures — and connects each segment's preference profile to a set of interpretable gap metrics that audit how well the current ranking serves each segment. We validate the framework on a large-scale lodging search dataset from a online travel marketplace, spanning 500k search sessions. The population-wide DCM identifies identifies 4 core preference drivers with statistically significant trade-off estimates; LCDCM recovers behaviorally distinct user segments with meaningfully different preference profiles. Segment-stratified gap metrics reveal that the ML derived ranking model under-serves a segment of smaller-party (i.e., ≤ 2 individuals) and longer-duration (i.e., ≥ 7 nights) travelers by surfacing hotel properties with nightly prices approximately \$60 higher than their preferred price range, motivating targeted improvements to underlying search ranking systems. The framework is model-and data-agnostic: it requires only observed choices and item features, making it applicable to any two-sided marketplace search or recommendation system.

Keywords

search ranking, recommendation, explainability, discrete choice model, latent class analysis, preference heterogeneity, behavioral segmentation, algorithm audit, information foraging theory

RecSys '26, Minneapolis, MN, USA

ACM ISBN 978-X-XXXX-XXXX-X/XX/XX
https://doi.org/10.1145/XXXXXXX.XXXXXXX

ACM Reference Format:
Aniket Sakpal, Adam Domanski, and Jean-Louis Guillemardet. 2026. Bridging Choice Theory and Recommender Systems: A Unified Framework for Latent Preference Discovery and Algorithmic Diagnosis in Recommender Systems. In *Proceedings of the Twentieth ACM Conference on Recommender Systems (RecSys '26), Minneapolis, MN, USA, September 29–October 1, 2026*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 Introduction

1.1 Motivation

The dominant paradigm in search and recommendation research is predictive: given historical interaction data, learn a model that maximises future clicks, conversions, or engagement. This approach has delivered substantial improvements in ranking quality [12, 13] but carries a structural blind spot — it tells us *what* users choose, not *why*, and it collapses the diversity of user preferences into a single aggregate signal.

The consequences are twofold. First, a ranking optimised for aggregate conversion will systematically under-serve preference minority segments [4]. Second, when a ranking model underperforms, practitioners lack actionable diagnostics — post-hoc tools such as SHAP [17] and LIME [21] describe the model's internal logic but not the gap between what the algorithm surfaces and what users actually want [25, 28].

This paper proposes a different lens: rather than asking *how well does the ranking model predict clicks?*, we ask *how well does the ranking reflect each user segment's revealed preferences?* The distinction matters on two-sided platforms, where ranking decisions simultaneously affect user welfare (demand side) and item provider exposure (supply side) [1, 22], and where a mismatch between rankings and preferences has a measurable welfare cost [7, 25].

1.2 Theoretical Grounding

We ground our framework in **Information Foraging Theory (IFT)** [20], which models search behavior as a cost-benefit optimization: users forage for information by navigating a system that imposes cognitive and navigational costs, abandoning it when the rate of information gain falls below a threshold. IFT implies any search system failure can be classified into one of two root causes:

- **Information problems**: the system is not surfacing the items the user needs (*what we show*);
- **Friction/cost problems**: the system makes it difficult for the user to reach items that do meet their needs (*how we show it*).

This taxonomy organizes both our diagnostic metrics and our intervention design.

1.3 Contributions

(1) **A behavioral search decomposition framework** grounded in IFT that classifies search failures into information vs. friction problems.
(2) **A DCM-based preference estimation methodology** recovering interpretable, cardinal attribute weights from observed choices — including Marginal Rates of Substitution (MRS) that quantify trade-offs in user-interpretable units.
(3) **An LCDCM extension** that uncovers distinct preference segments hidden beneath the population-average model, revealing which segments the current ranking systematically mis-serves.
(4) **Segment-aware gap metrics and an algorithm audit procedure** connecting behavioral segment profiles to ranking evaluation metrics, enabling targeted diagnosis and improvement.
(5) **A large-scale empirical validation** on approximately 500k lodging search sessions from a production search and ranking system, demonstrating measurable segment-level improvements from a targeted ranking intervention.

The framework is *model-agnostic*: it sits alongside any existing ranking model as a behavioral auditing layer, requiring only a log of observed choices and item features. It is applicable to e-commerce, streaming, job boards, and any marketplace where users choose among a ranked list of alternatives.

1.4 Paper Organization

Section 2 reviews related work. Section 3 presents the theoretical framework. Section 4 describes the DCM. Section 5 presents the LCDCM extension. Section 6 introduces the explainability framework and gap metrics. Section 7 presents the case study. Section 8 discusses generalizability, limitations, and ethics. Section 9 concludes.

2 Related Work

2.1 Ranking Evaluation and Unbiased Learning-to-Rank

Standard NDCG [12] assumes a single ground-truth relevance ordering — an assumption that breaks down under preference heterogeneity. Joachims et al. [13] demonstrate that LTR models trained on click logs are statistically inconsistent due to position bias, proposing IPS correction. Niu et al. [19] show that user-level heterogeneity in examination propensity introduces a systematic second-order bias in standard IPS estimators. This directly parallels our LCDCM finding: the same phenomenon framed as a preference heterogeneity problem with a complementary structural remedy. Our DCM separates the preference signal from the position effect structurally, rather than through propensity reweighting [11].

2.2 Discrete Choice Models in Platform Ranking

Ursu [25] is the most directly relevant prior work: a DCM on OTA data showing that preference-aligned ranking nearly doubles consumer welfare. We extend this by adding the LCDCM dimension: segment-level misalignment rather than a population-average estimate. Koulayev [15, 16] develops the identification strategy for DCM on search logs. De los Santos and Koulayev [4] prove theoretically that preference heterogeneity makes a single ranking suboptimal; our LCDCM is the empirical quantification of that result. Ghose et al. [7] share our logical structure but do not apply latent class analysis; Cavenaghi et al. [3] show aggregate click models underweight minority preference segments, directly motivating our approach.

2.3 Latent Class Discrete Choice Models and Fairness

Train [24] provides the canonical LCDCM treatment; the approach originates with Swait [23]. Greene and Hensher [9] argue discrete segments are more interpretable than continuously distributed heterogeneity when variation is type-based. Wedel and Kamakura [27] establish that joint LCDCM estimation via EM is preferable to post-hoc clustering. El Zarwi et al. [6] validate LCDCM for travel demand. For scalable estimation with large choice sets, we adopt the EM approach of von Haefen and Domanski [26]. On fairness, Burke [2] and Karimi et al. [14] characterise the Pareto frontier between recommendation accuracy and provider-side fairness — the same tension our segment-aware interventions navigate.

3 Theoretical Framework

3.1 Information Foraging Theory as a Design Lens

IFT [20] models an information-seeking agent as optimizing the rate of information gain:

$$\text{Rate of Gain} = \frac{\text{Information Value}}{\text{Cost of Obtaining Information}}. \tag{1}$$

A system failure occurs when this rate falls below the user's threshold, causing abandonment or unsuccessful search.

3.2 A Taxonomy of Search Failures

Information problems arise when the system does not surface the items a user needs (*what we show*). The root cause lies in the ranking model's relevance estimation.

Friction problems arise when relevant items are surfaced but made difficult to access (*how we show it*). The root cause lies in presentation, position, or UI.

This decomposition is operationally important: the same aggregate metric decline can be caused by either failure type, but the remediation is completely different. Our DCM and LCDCM framework directly addresses information misalignment.

3.3 Observable and Unobservable Preference Factors

We model user choice using a standard random utility model (RUM) specification [18], in which the utility user n derives from item i is decomposed as:

$$U_{ni} = V_{ni} + \varepsilon_{ni}, \tag{2}$$

where $V_{ni} = \boldsymbol{\beta}_n \mathbf{x}_{ni}$ represents utility derived from observable item attributes, and ε_{ni} captures unobservable idiosyncratic factors (picture quality, personal history, unstated requirements). Under the RUM framwork, user n selects item i if and only if $U_{ni} > U_{nj}$ for all $j \neq i$ in the choice set.

4 Discrete Choice Model of Search Behavior

4.1 Choice Situation and Choice Set Definition

The RUM approach is operationalized by defining a choice occasion as a single search session in which a user selects one property from choice set produced by an ML ranking algorithm. This choice set C_n consists of the **top 20 properties** displayed to user n in that session, satisfying the mutual exclusivity and finiteness assumptions of the RUM framework. While in reality, a consumer's choice set is not limited by those alternatives presented in a single search session, this approach is consistent with recent research evaluating the validity of evaluating consumer choice among consideration sets that are limited by information search costs or non-trivial probabilities of selection, both of which are achieved by presenting a set of results optimized by an ML search algorithm. [8] For search sessions with multiple searches, resulting in very large consideration sets, computational tractability is achieved by sampling from alternatives, as demonstrated by von Haefen and Domanski. [26]

4.2 Model Specification

Assuming ε_{ni} is i.i.d. Type-I Extreme Value, the probability that user n selects item i takes the conditional logit form [18]:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j \in C_n} \exp(V_{nj})}. \tag{3}$$

Using the full set of observable features that vary across alternatives, the deterministic utility component decomposes into four conceptual feature groups common across search and recommendation domains:

$$\begin{aligned} V_{ni} = {} & \beta_1 \cdot \text{Cost}_{ni} + \beta_2 \cdot \text{SocialProof}_{ni} \\ & + \beta_3 \cdot \text{Proximity}_{ni} + \beta_4 \cdot \text{Quality}_{ni} + \boldsymbol{\beta}_5^\top \mathbf{A}_{ni} + \gamma \cdot \log(\text{rank}_{ni}), \end{aligned} \tag{4}$$

where **Cost** captures monetary cost to the consumer, **SocialProof** captures credibility signals, **Proximity** captures relevance to the user's stated context, **Quality** captures intrinsic item quality, and $\mathbf{A}_{ni}$ is a vector of domain-specific attributes. The term $\gamma \cdot \log(\text{rank}_{ni})$ explicitly models residual position effects generated by the ML ranking algorithm: including it prevents position-driven choices from biasing the estimated attribute coefficients [16]. A negative $\hat{\gamma}$ indicates residual position preference beyond what item attributes explain.

In our lodging search case study (Section 7), Cost, SocialProof, Proximity, and Quality instantiate as nightly price, review count, distance from search centroid, and aggregate rating, respectively.

4.3 Identification and Estimation

Parameters are estimated by maximum likelihood. To address the endogeneity of consideration sets [15], we condition on the displayed choice set C_n, treating the algorithm's ranking as exogenous variation in what is shown [16]. The log-rank term absorbs residual systematic position preference - reflecting a combination of information-search costs and user-trust in the ranked search results, further separating attribute-driven utility from position-driven selection. This specification allows direct recovery of marginal utilities and economically interpretable trade-offs.

4.4 Marginal Rates of Substitution

A key advantage of this structural approach is the ability to compute the Marginal Rate of Substitution (MRS), which quantifies how users trade off attributes:

$$\text{MRS}_{kl} = -\frac{\beta_k}{\beta_l}. \tag{5}$$

This is particularly relevant for the MRS between product attributes and price, providing the rate at which a consumer is willing to tradeoff product attributes with cost, while keeping utility constant. For example, if $\hat{\beta}_{\text{proximity}} = -0.30$ per unit of distance and $\hat{\beta}_{\text{price}} = -0.15$ per dollar, then $\text{MRS}_{\text{proximity,price}} = 2.0$ — users require a \$2 price reduction to remain indifferent about an item one unit further away. MRS estimates from our case study are reported in Section 7.

5 Latent Class Discrete Choice Model

5.1 Motivation: Preference Heterogeneity

The standard DCM in Equation (3) recovers a single preference structure for all users. De los Santos and Koulayev [4] prove that when preferences are in fact heterogeneous, no single ranking is optimal for all users. The question is not *whether* heterogeneity exists, but how to characterize it in a way that is actionable for ranking system design. This can be achieved by applying a mixed-logit specification of the discrete choice model. While either a continuous or discrete mixing distribution can be applied, the discrete mixing distribution (aka, the Latent Class Discrete Choice Model) provides additional structure that elucidates the hetergenous preference structure of consumers.

5.2 Model Formulation

The Latent Class Discrete Choice Model (LCDCM) assumes the population consists of C unobserved classes, each with its own preference vector $\boldsymbol{\beta}_c$. The unconditional probability that user n chooses item i is:

$$P_{ni} = \sum_{c=1}^{C} S_{nc}(\mathbf{z}_n, \boldsymbol{\delta}) \cdot P_{ni}(\boldsymbol{\beta}_c), \tag{6}$$

where $S_{nc}(\mathbf{z}_n, \boldsymbol{\delta})$ is the probability that user n belongs to class c, modeled as a multinomial logit over observable user characteristics $\mathbf{z}_n$:

$$S_{nc}(\mathbf{z}_n, \boldsymbol{\delta}) = \frac{\exp(\boldsymbol{\delta}_c^\top \mathbf{z}_n)}{\sum_{l=1}^{C} \exp(\boldsymbol{\delta}_l^\top \mathbf{z}_n)}. \tag{7}$$

5.3 Estimation via the EM Algorithm

We estimate the LCDCM using the EM algorithm [5, 24, 26], which recursively estimates two steps until model stability is achieved via some pre-defined convergence criteria:

E-step. Compute posterior class membership probabilities:

$$h_{nc}(\phi^t) = \frac{S_{nc}(\boldsymbol{\delta}^t) \cdot L_n(\boldsymbol{\beta}_c^t)}{\sum_{l=1}^{C} S_{nl}(\boldsymbol{\delta}^t) \cdot L_n(\boldsymbol{\beta}_l^t)}. \tag{8}$$

M-step. Update preference parameters for each class, using the class probabilities as weights:

$$\phi^{t+1} = \arg\max_{\boldsymbol{\delta},\boldsymbol{\beta}} \sum_n \sum_c h_{nc}(\phi^t) \ln\left[S_{nc}(\boldsymbol{\delta}) \cdot L_n(\boldsymbol{\beta}_c)\right]. \tag{9}$$

The separability of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ enables efficient estimation via two independent conditional logit optimizations per iteration [26].

5.4 Class Selection

We select the number of classes using the corrected Akaike Information Criterion (crAIC) [10], which penalizes unnecessary model complexity and guards against overfitting. The optimal number of classes and the resulting segment profiles are reported in Section 7.

6 Explainability Framework for Search Ranking

6.1 Framework Overview

The complete explainability pipeline proceeds as follows: observed search sessions feed into the DCM to recover cardinal preference weights; the LCDCM identifies behavioral segments; segment profiles are mapped to system-controllable ranking signals; gap metrics quantify the misalignment; and the IFT taxonomy classifies each gap as an information or friction problem, directing the intervention.

6.2 Gap Metrics: Preference-Aligned Ranking Evaluation

For each behavioral segment c and ranking attribute k, we define a *gap metric*:

$$\text{Gap}_{c,k} = \mathbb{E}\left[\hat{x}_{ki}^{\text{rank}}\right]_c - \mathbb{E}\left[x_{ki}^*\right]_c, \tag{10}$$

where $\hat{x}_{ki}^{\text{rank}}$ is the value of attribute k in items surfaced at the top of the ranking for segment-c users, and x_{ki}^* is the preference-optimal value implied by the segment's preference profile. A positive gap indicates the system over-exposes the attribute relative to the segment's preference; a negative gap indicates under-exposure.

For a price-sensitive segment, a positive price gap means the algorithm surfaces items more expensive than the segment's revealed willingness-to-pay — a direct information problem under IFT.

6.3 Segment-Stratified Offline Evaluation

Standard NDCG collapses across all users. We propose computing NDCG separately for each latent segment, using the segment's DCM preference weights to construct relevance labels:

$$\text{Relevance}_{ni}^{(c)} = \frac{\exp(\hat{\boldsymbol{\beta}}_c^\top \mathbf{x}_{ni})}{\sum_{j \in C_n} \exp(\hat{\boldsymbol{\beta}}_c^\top \mathbf{x}_{nj})}. \tag{11}$$

This segment-aware $\text{NDCG}^{(c)}$ reveals disparities invisible to the aggregate metric.

6.4 Diagnosing Ranking Model Failures

Gap metrics and segment-stratified NDCG jointly enable a structured diagnostic:

(1) Compute gap metrics for each segment and attribute.
(2) Identify segment–attribute combinations with the largest absolute gaps.
(3) Classify each gap using IFT (information vs. friction problem).
(4) Prioritize interventions by segment size × gap magnitude.

This transforms ranking explainability from an abstract post-hoc analysis into an actionable audit with clear intervention targets.

7 Case Study

7.1 Data

We evaluate the framework on approximately **500,000 converted search sessions** from a production lodging search and ranking system at a major online travel marketplace. The dataset was constructed as follows:

- *Booking filter*: Only sessions resulting in a property booking are retained, ensuring each observation provides a reliable revealed preference signal; sessions without conversion are excluded to avoid noisy implicit signals.
- *Temporal undersampling*: Sessions are drawn uniformly across a full calendar year to ensure adequate seasonal coverage and avoid over-representation of peak travel periods.
- *Choice set*: The top 20 displayed properties per converted session, satisfying the mutual exclusivity and exhaustiveness assumptions of the RUM framework.

Item features include: nightly price, aggregate rating, review count, distance from search centroid, property type, and amenity indicators. In the notation of Section 4, these instantiate as: Cost → nightly price, SocialProof → review count, Proximity → distance from search centroid, Quality → aggregate rating.

User context features include: device type (mobile/tablet/desktop), trip duration (number of nights), party size (number of guests), lead time (days between search and intended check-in), and booking window (days between booking and check-in). Exploratory analysis identified *trip duration* and *party size* as having the strongest association with latent class membership, measured by mutual information with posterior class assignments from the LCDCM. These two features serve as the primary class membership covariates $\mathbf{z}_n$ in Equation (6). Device type, lead time, and booking window showed weaker discriminative signal and were retained as auxiliary covariates.

Results are reported without disclosure of platform-identifying metrics, consistent with the internal research agreement under which the data was made available.

7.2 DCM Results

Table 1 reports the population-average parameter estimates. All three core attributes are statistically significant ($p < 0.05$) and directionally consistent with economic theory. Distance has the largest per-unit effect ($\hat{\beta}_{\text{distance}} = -0.41\%$ per mile), followed by review count (+0.44% per review) and nightly price (−0.22% per dollar).

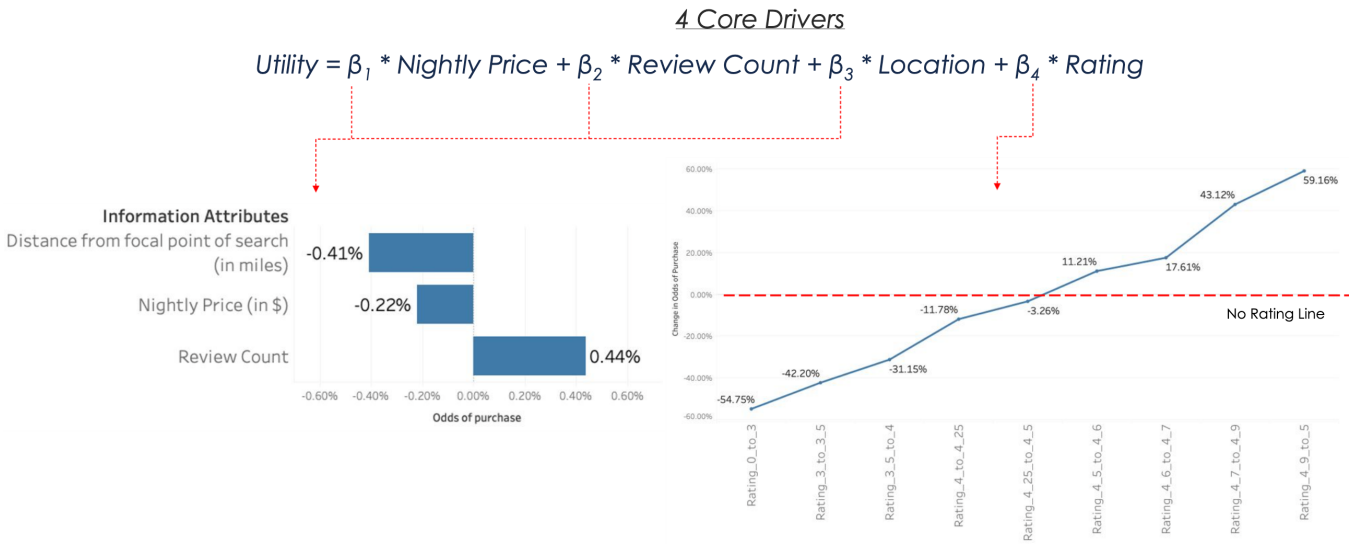

Figure 1: DCM population-average results. *(a)* Odds-of-purchase coefficients for Distance, Nightly Price, and Review Count. *(b)* Non-linear rating effect — large penalties for unrated properties and increasing premiums above 4.5 stars.

Table 1: DCM parameter estimates — population average model. Standard errors in parentheses. $^{*}p < 0.05$.

General Feature	Lodging Instantiation	$\hat{\beta}$	Std. Error
Cost	Nightly Price (per $1)	−0.22%*	(0.021)
SocialProof	Review Count (per review)	+0.44%*	(0.038)
Proximity	Distance from Center (mi)	−0.41%*	(0.035)
Quality	Rating	non-linear — see Figure 1	

Rating exhibits strong non-linearity (Figure 1): unrated properties carry a severe penalty (−54.75%), while properties in the highest rating bucket command a +59.16% premium. The effect is negligible between 3.5 and 4.25 stars, indicating meaningful quality signals only translate to higher purchase probability above the 4.25-star threshold.

The MRS between distance and price implies the average user is willing to pay $1.86 more per night for each mile of additional proximity — concretely, a property 15 miles from the search centroid must be priced approximately $28 lower per night than an identical city-center property to achieve equal expected utility. Figure 2 illustrates this across three reference properties.

7.3 LCDCM Results

crAIC model selection identifies $C = 2$ latent classes as optimal. Table 2 and Figure 3 report the class-specific estimates.

Segment 1 (70% of sessions). Distance coefficient (−0.61% per mile) is 50% larger than the population average, while price sensitivity (−0.09% per dollar) is less than half the average. These users forage primarily for proximity and are comparatively tolerant of higher prices. Figure 4 confirms this segment concentrates in the region of short stays (<7 nights) and larger party sizes (above 2 guests), consistent with couples or individuals on brief city trips.

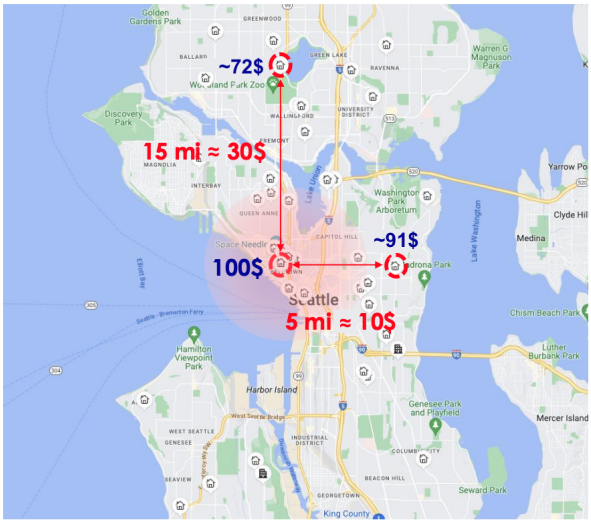

Figure 2: Three-property comparison illustrating the MRS between Proximity and Cost. A reference property at the city center ($100/night) is utility-equivalent to one 5 miles away at $91/night, and one 15 miles away at $72/night.

Segment 2 (30% of sessions). Price coefficient (−0.85% per dollar) is approximately 4× the population average, making cost the dominant driver. Distance sensitivity is low (−0.16%). Figure 4 confirms this segment spans longer stays (≥7 nights) and smaller party sizes (≤2 guests), consistent with solo or couple travelers on extended trips where total accommodation cost over a longer duration is a binding budget constraint.

Table 2: LCDCM segment profiles — class-specific preference coefficients vs. population average. $^*p < 0.05$.

Attribute	Average $\hat{\beta}$	Segment 1 $\hat{\beta}_1$	Segment 2 $\hat{\beta}_2$
Cost (Nightly Price)	$-0.22\%^*$	$-0.09\%^*$	$-0.85\%^*$
SocialProof (Review Count)	$+0.44\%^*$	$+0.37\%^*$	$+0.56\%^*$
Proximity (Distance, mi)	$-0.41\%^*$	$-0.61\%^*$	$-0.16\%^*$
Quality (Rating)	non-linear — see Figure 1		
Share of sessions	100%	70%	30%

Table 3: Price gap metric by segment: mean nightly price of shown vs. clicked properties. A positive gap indicates the ranking over-exposes the segment to higher-priced properties than their revealed willingness-to-pay supports ($^*p < 0.05$).

	Mean Nightly Price ($)		
	Shown	Clicked	Gap
Population Average	\$200	\$180	$+\$20^*$
Segment 1	\$200	\$195	+\$5
Segment 2	\$200	\$140	$+\$60^*$

The population-average model *misrepresents* both segments: the average price coefficient is less than a quarter of Segment 2's true sensitivity, meaning a ranking calibrated on aggregate data will systematically surface over-priced results for the most cost-constrained users [4].

The class membership model shows that trip duration and party size are the strongest predictors of segment assignment, providing an interpretable basis for targeting ranking interventions at session creation time.

7.4 Gap Metrics and Ranking Audit

Figure 5 and Table 3 reveal the central diagnostic finding. At the population-average level, shown prices are shifted modestly above clicked prices (gap ≈ \$20), indicating a mild but consistent tendency to surface above-willingness-to-pay results.

The signal sharpens dramatically once users are stratified by latent segment. For Segment 1, the price gap is small (\$5): low price sensitivity means the ranking's occasional surfacing of higher-priced properties does not meaningfully conflict with their preferences. For Segment 2, however, the gap is \$60 — the shown price distribution has a pronounced rightward shift relative to the clicked distribution, despite this segment's 4×-elevated price sensitivity.

Using the IFT taxonomy, this constitutes an *information problem*: the system is surfacing the wrong items for Segment 2 users. The fix lies in ranking logic, not interface or presentation changes.

7.5 Ranking Intervention and Results

The gap metric analysis directly motivates a targeted ranking intervention for Segment 2. We implement a heuristic price rule that promotes lower-priced properties for sessions with high Segment 2 membership probability $h_{n,2}$: properties priced above the DCM-derived willingness-to-pay ceiling $\overline{\text{WTP}}_2$ are soft-demoted via a re-scoring penalty proportional to $h_{n,2} \cdot (p_i - \overline{\text{WTP}}_2)$, and a slot-injection fallback promotes in-band properties when the top-K set contains too few candidates meeting the price criterion.

Evaluated using segment-stratified NDCG (Equation (11)), this intervention produces improvements of approximately 2–4% in NDCG@10 for Segment 2.

Pathways to feature engineering. The heuristic intervention demonstrates the framework's diagnostic value, but the same segment insights translate naturally into durable ranking model improvements. The LCDCM segment membership probabilities $h_{n,c}$ can be joined directly as ranking features, enabling the learned model to internalize segment-varying price weights without hard-coded rules. Complementary features include: a session-relative price score $(p_i - \bar{p}_s)/\sigma_{p,s}$ capturing within-set affordability; a value index (review count per dollar) encoding Segment 2's joint price–social-proof sensitivity; and predicted trip-type indicators (party size, stay duration) as proxy signals for segment membership when explicit probabilities are unavailable. This feature engineering pathway converts the behavioral audit into a closed loop: gap metrics identify the misalignment, the heuristic validates the hypothesis, and the resulting features embed the preference signal permanently into the ranking model.

8 Discussion

8.1 Generalizability

The framework requires only: (1) observed discrete choices, (2) item feature vectors, and (3) an optional user characteristic vector. It is therefore applicable to any search or recommendation context where users choose from a displayed set: e-commerce product search, streaming, job boards, real estate, and restaurant discovery. The IFT decomposition provides a consistent vocabulary for cross-domain ranking diagnostics.

8.2 Limitations

IIA within classes. The conditional logit kernel assumes Independence of Irrelevant Alternatives within each class. The LCDCM relaxes IIA across classes but not within. Nested logit kernels could be used for richer within-class substitution at the cost of identifiability. Notably IIA is a necessary condition for sampling of alternatives within classes. Should computationally tractability

Static preference estimates. DCM and LCDCM are estimated on a historical snapshot and implicitly assume preference stationarity over the estimation period.

Consideration set endogeneity. We condition on the displayed choice set, which is itself a function of the algorithm. Unobserved factors driving prior ranking can correlate with unobserved user preferences [15]. In fully observational settings, this is a maintained assumption.

Segment label interpretability. LCDCM segment labels are data-driven and should be validated with domain experts before deployment.

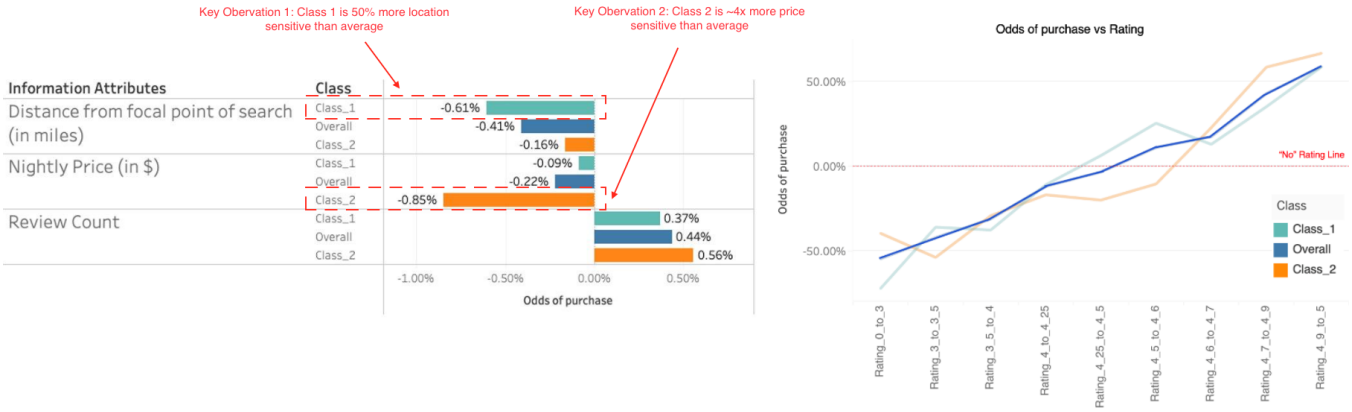

Figure 3: LCDCM segment-level preference coefficients vs. the population average. *(a)* Segment 1 is 50% more sensitive to distance; Segment 2 is approximately 4× more sensitive to nightly price. *(b)* Rating response curves by segment.

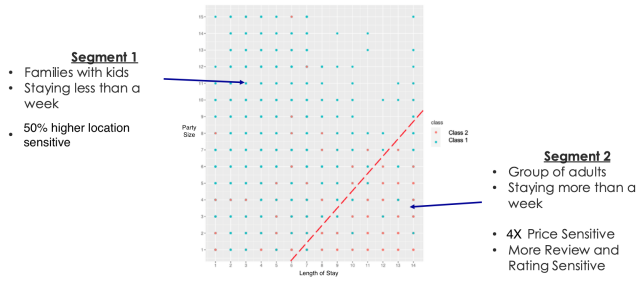

Figure 4: Segment profiling by trip characteristics (party size vs. length of stay). Segment 1 (teal) concentrates in the short-stay, small-party quadrant; Segment 2 (red) spans longer stays and larger parties.

8.3 Supply-Side Insights

The DCM and LCDCM produce not only rankings but also interpretable utility estimates that the platform can surface to item providers — a dimension typically absent from black-box ranking systems.

Market-level pricing guidance. For a given market, the estimated price coefficient $\hat{\beta}_{\text{price}}$ and the segment-level WTP ceiling $\overline{\text{WTP}}_c$ provide a data-driven reference for property owners. A property can compare its nightly price against the DCM-implied willingness-to-pay for its location–review profile: if $p_i > \overline{\text{WTP}}_c$ for the dominant local segment, a price adjustment is likely to improve both click-through and conversion independent of any ranking intervention. This is qualitatively different from naïve competitor benchmarking because it anchors pricing to revealed traveler utility rather than supply-side price distribution.

Attribute-level improvement signals. Beyond price, the marginal utility estimates $\hat{\beta}_{\text{SocialProof}}$ and the non-linear rating curve (Figure 1) identify rating thresholds at which utility increases discontinuously. A property with a rating just below such a threshold benefits more from targeted review quality improvement than from price reduction; conversely, a property with abundant reviews relative to competitors may find that location or amenity improvements yield higher marginal utility gain. The framework thus converts segment preference profiles into actionable, attribute-specific improvement signals for property owners.

Platform-assisted onboarding for new properties. New listings lack the review history needed to compete on social proof signals. The platform can use the LCDCM utility estimates for comparable established properties (matched on location, property type, and price tier) to identify which segments are most likely to convert, and direct initial exposure toward those segments. This acts as a cold-start mitigation: rather than exposing new properties to all segments uniformly (where they will lose on social proof), the platform can seed traction in the segments where the new property's non-review attributes are most competitive according to the estimated utility function.

8.4 Ethical Considerations

Segment-aware ranking raises fairness concerns: if segments correlate with protected attributes (age, income, nationality), segment-targeted optimization could constitute discriminatory ranking. Practitioners should audit segment correlations with protected attributes before deploying segment-targeted interventions. Additionally, transparency obligations to item providers require disclosure when ranking policy changes materially affect their exposure.

9 Conclusion

We have presented a behavioral explainability framework for search ranking and recommendation combining Information Foraging Theory, Discrete Choice Modelling, and the Latent Class Discrete Choice Model. The framework diagnoses ranking failures at the segment level — identifying which users the current ranking underserves and why — and connects these findings to actionable gap metrics and preference-aligned ranking policies.

Validated on approximately 500,000 lodging search sessions from a production ranking system, the framework reveals that

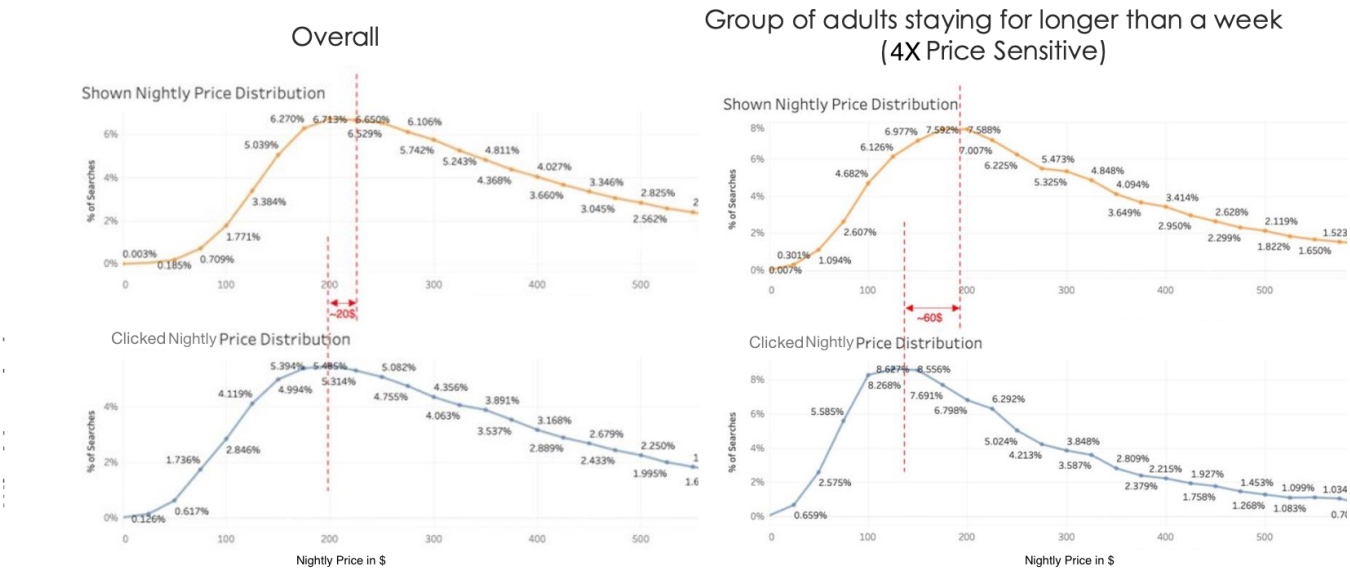

Figure 5: Price gap metric. Each panel compares shown (orange) vs. clicked (blue) nightly price distributions. *(a)* Population-average: gap ≈ $20. *(b)* Segment 2: gap ≈ $60 — a direct information problem under IFT.

the population-average model obscures significant preference heterogeneity: users in Segment 2 (price-sensitive, 30% of sessions) face a price gap of approximately $60 per night, meaning the ranking systematically surfaces properties priced well above their revealed willingness-to-pay. A targeted heuristic intervention closes this gap with indicative improvements of 2–4% in segment-stratified NDCG@10, and the segment utility estimates open a supply-side channel through which the platform can guide property owners on pricing, attribute improvement, and new-listing traction — value that aggregate ranking metrics do not capture.

The framework is model-agnostic and domain-agnostic, requiring only observed choices and item features, and is broadly applicable to any two-sided marketplace search or recommendation system.

Acknowledgments

[Suppressed for double-blind review.]

Data Availability

The dataset used in this study is proprietary and was made available under an internal research agreement. It cannot be publicly released due to commercial confidentiality and user privacy constraints. The methodology, model specifications, and evaluation procedures are described in full detail to enable replication on comparable datasets.

Ethics Statement

This study was conducted entirely on anonymized, aggregate behavioral logs. No personally identifiable information was accessed, stored, or processed at any stage. The behavioral segmentation produced by the LCDCM is used solely for improving search ranking quality and is not linked to any individual user profile. Potential fairness implications are discussed in Section 8.

References

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multi-stakeholder Recommendation: Survey and Research Directions. *User Modeling and User-Adapted Interaction* 30, 127–158. doi:10.1007/s11257-019-09256-1

[2] Robin Burke. 2017. Multisided Fairness for Recommendation. In *Proceedings of the FATML Workshop, co-located with KDD.*

[3] Emanuele Cavenaghi, Luigi Camaione, Pietro Minasi, Gabriele Sottocornola, Fabio Stella, and Markus Zanker. 2022. An Analysis of User Click Behaviour in Online Hotel Search. In *Workshop on Recommenders in Tourism (RecTour), co-located with ACM RecSys.*

[4] Babur De los Santos and Sergei Koulayev. 2017. Optimizing Click-Through in Online Rankings with Endogenous Search Refinement. *Marketing Science* 36, 3 (2017), 434–452. doi:10.1287/mksc.2016.1020

[5] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B* 39, 1 (1977), 1–38. doi:10.1111/j.2517-6161.1977.tb01600.x

[6] Feras El Zarwi, Akshay Vij, and Joan L. Walker. 2017. A Discrete Choice Framework for Modeling and Forecasting the Adoption and Diffusion of New Transportation Services. *Transportation Research Part B: Methodological* 100 (2017), 101–119. doi:10.1016/j.trb.2017.01.013

[7] Anindya Ghose, Panagiotis G. Ipeirotis, and Beibei Li. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science* 31, 3 (2012), 493–520. doi:10.1287/mksc.1120.0724

[8] Peter Gibbard. 2021. Disentangling Preferences and Limited Attention: Random-utility models with consideration sets. *Journal of Mathematical Economics* 94, 102468 (2021). doi:10.1016/j.jmateco.2020.102468

[9] William H. Greene and David A. Hensher. 2003. A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit. *Transportation Research Part B: Methodological* 37, 8 (2003), 681–698. doi:10.1016/S0191-2615(02)00046-2

[10] Stephen Hynes, Nick Hanley, and Riccardo Scarpa. 2008. Effects on Welfare Measures of Alternative Means of Accounting for Preference Heterogeneity in Recreational Demand Models. *American Journal of Agricultural Economics* 90, 4 (2008), 1011–1027. doi:10.1111/j.1467-8276.2008.01148.x

[11] Md. Ariful Islam, Kathryn Vasilaky, and Elena Zheleva. 2025. Correcting for Position Bias in Learning to Rank: A Control Function Approach. arXiv:2502.xxxxx

[12] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulative Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446. doi:10.1145/582415.582418

[13] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM).* ACM, 781–789. doi:10.1145/3018661.3018699

[14] Shahrzad Karimi, Hossein A. Rahmani, Mohammadmehdi Naghiaei, and Leila Safari. 2023. Provider Fairness and Beyond-Accuracy Trade-offs in Recommender Systems. arXiv:2306.xxxxx

[15] Sergei Koulayev. 2009. *Estimating Demand in Search Markets: The Case of Online Hotel Bookings.* Technical Report 09-16. Federal Reserve Bank of Boston.

[16] Sergei Koulayev. 2014. Search for Differentiated Products: Identification and Estimation. *RAND Journal of Economics* 45, 3 (2014), 553–575. doi:10.1111/1756-2171.12061

[17] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. 4765–4774.

[18] Daniel McFadden. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, Paul Zarembka (Ed.). Academic Press, 105–142.

[19] Zhengyi Niu, Lang Mei, Liyi Yang, Zhuoran Zhao, Qifan Yan, Jiaxin Mao, and Ji-Rong Wen. 2025. Addressing Personalized Bias for Unbiased Learning to Rank. arXiv:2502.xxxxx

[20] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (1999), 643–675. doi:10.1037/0033-295X.106.4.643

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1135–1144. doi:10.1145/2939672.2939778

[22] Jean-Charles Rochet and Jean Tirole. 2003. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association* 1, 4 (2003), 990–1029. doi:10.1162/154247603322493212

[23] Joffre Swait. 1994. A Structural Equation Model of Latent Segmentation and Product Choice for Cross-Sectional Revealed Preference Choice Data. *Journal of Retailing and Consumer Services* 1, 2 (1994), 77–89. doi:10.1016/0969-6989(94)90002-7

[24] Kenneth E. Train. 2009. *Discrete Choice Methods with Simulation* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511805271

[25] Raluca M. Ursu. 2018. The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science* 37, 4 (2018), 530–552. doi:10.1287/mksc.2017.1072

[26] Roger H. von Haefen and Adam Domanski. 2018. Estimation and Welfare Analysis from Mixed Logit Models with Large Choice Sets. *Journal of Environmental Economics and Management* 87 (2018), 72–87. doi:10.1016/j.jeem.2017.08.003

[27] Michel Wedel and Wagner A. Kamakura. 2000. *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). Kluwer Academic Publishers. doi:10.1007/978-1-4615-4651-1

[28] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends in Information Retrieval* 14, 1 (2020), 1–101. doi:10.1561/1500000066