Next Best Action: Causal Constrained Offline RL for Business Action Targeting

Surya Ramachandiran
Lyft, Inc.
San Francisco, CA, USA
surya.ramachandiran18@gmail.com

Yiming Li
Lyft, Inc.
San Francisco, CA, USA
yli@lyft.com

Juan Mancilla-Caceres
Lyft, Inc.
San Francisco, CA, USA
juanmc@lyft.com

Tao Ye
Lyft, Inc.
San Francisco, CA, USA
taoye@lyft.com

Jun Liu
Lyft, Inc.
San Francisco, CA, USA
junliu@lyft.com

Abstract

Business action targeting systems must select from a variety of interventions (e.g., discounts, promotions, and lifecycle actions), optimize for long-term economic value, and satisfy complex operational constraints, yet lack a rigorous framework for doing so jointly. To address this gap, we introduce an auditable action-selection module that adapts decisions to user state, encodes explicit causal and long-term rationales, and provides constraint verifiability. We integrate off-policy causal correction of the estimated Markov Decision Process (MDP) directly into the dual Linear Programming (LP) formulation of the Constrained MDP (CMDP). This enables dual-based constraint verifiability while selecting actions that are both causally grounded and optimized for long-term value, a capability not previously demonstrated in this setting. Notably, the LP formulation yields shadow prices that make the economic trade-offs of operational constraints (e.g., fairness and user welfare) explicit, strengthening policy explainability. We validate the proposed offline Reinforcement Learning (RL) framework on an industry case study involving coupon targeting in rideshare, with customer states represented by Recency, Frequency, and Monetary (RFM) stages. Using model-based off-policy evaluation, with causal dynamics estimated from a temporal holdout, we find that the optimized policy improves customer lifetime value (LTV) by 19% over the estimated behavior policy and 8% over a myopic baseline, when all policies operate on the RFM stages and under the observed coupon budget. Although the demonstrated action space is binary, the underlying causal, long-term, and constraint primitives are action-agnostic, yielding a general-purpose decision module for targeting systems that manage multiple interventions.

CCS Concepts

• **Computing methodologies** → **Sequential decision making**; *Causal reasoning and diagnostics*; • **Mathematics of computing** → **Linear programming**; • **Information systems** → *Recommender systems.*

Keywords

Constrained Markov Decision Process, Offline Reinforcement Learning, Causal Inference, Linear Programming, Recommender Systems, Customer Lifetime Value

ACM Reference Format:
Surya Ramachandiran, Yiming Li, Juan Mancilla-Caceres, Tao Ye, and Jun Liu. 2026. Next Best Action: Causal Constrained Offline RL for Business Action Targeting. In *Proceedings of ACM RecSys Workshop on Online & Adaptive Recommender System (OARS @ RecSys'26).* ACM, New York, NY, USA, 6 pages.

1 Introduction

Consumer platforms increasingly employ *business action targeting* systems to determine the optimal timing and type of intervention (e.g., discounts, promotions, or lifecycle actions) for each customer to maximize long-term engagement [6, 13, 27, 29, 31]. Developing such systems raises several coupled challenges illustrated in Figure 1, yet most existing research addresses them in isolation.

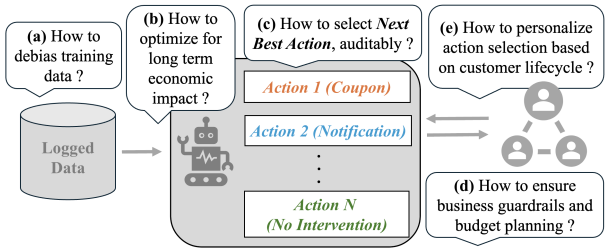

Figure 1: Challenges for business action targeting systems

The first challenge is *training data bias.* Because logged interactions are generated by a historical targeting policy rather than a randomized experiment, naive estimates of action effects are heavily confounded [4, 22]. To correct for this, large-scale recommenders commonly employ off-policy estimators such as propensity-based, doubly robust, and self-normalized importance sampling methods [6, 7, 10, 14, 21, 24, 25]. These approaches, however, predominantly target single-step (bandit) outcomes rather than multi-step customer dynamics.

A second theme of research addresses this gap by framing recommendation and targeting as sequential decision-making problems. Early MDP recommenders [23] and lifetime value targeting models [26] have since matured into industrial deep RL systems that optimize long-term engagement [6, 13, 27, 29, 31]. Typically trained offline [17], these methods rely on pessimism [15] or stationary distribution corrections (e.g., DICE frameworks [16, 18]) to mitigate out-of-distribution extrapolation error. However, standard

offline RL rarely guarantees compliance with complex business constraints (e.g., budget caps, exposure limits, fairness), while DICE methods leave observational confounding unaddressed and remain unstudied in recommendation and targeting contexts.

A third theme, popular in both research and industry, tackles business constraints directly by casting targeting as a two-step incentive allocation problem that pairs an uplift model [11] with mathematical optimization [1, 9, 28, 30]. While effective at enforcing budgets and business guardrails, this approach is myopic. It maximizes immediate conversion or short-term uplift while ignoring how current actions shape long-term customer value.

In short, each existing approach addresses only a subset of the challenges shown in Figure 1 (e.g., correcting for causal bias without optimizing for the long horizon, maximizing long-term value without exact constraints or business guardrails, or allocating under constraints without long-term objectives). To our knowledge, this is the first framework that unifies causal off-policy correction with constrained RL to enable auditable, constraint-verifiable business intervention targeting in an industry setting.

2 Methodology

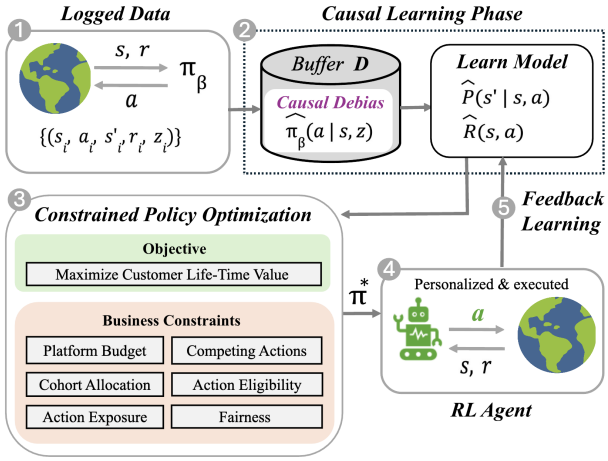

Figure 2: Next Best Action (NBA) framework

Our Next Best Action (NBA) framework (see Figure 2) addresses these challenges (see Figure 1): (a) because the logged training data reflect a biased prior policy, we propose to use propensity weighting [21, 25] in model-based offline RL [17] to causally correct the learned transition and reward estimates; (b) leveraging these corrected estimates, we formulate action selection as a CMDP to directly optimize for long-term value rather than immediate rewards or clicks; (c) we solve the CMDP using its dual Linear Program (LP), which allows us to choose the right next action verifiably [2, 20]; (d) the LP formulation allows us to model business guardrails (like budget caps, exposure limits, and fairness) as explicit constraints, while also allowing for budget planning and scenario analysis; (e) the resulting optimal policy is explainable as the decisions are adapted to the user lifecycle stages [5]. Because every action carries a transparent, causal, and constraint-verifiable rationale, the system is auditable by design.

While highly interpretable, the current formulation is restricted to discrete states, leaving continuous extensions for future work. Offline RL [17] is also vulnerable to extrapolation errors caused by limited data coverage. Because the initial case study uses discrete user lifecycle states, every state-action pair is well-supported by the industrial dataset. This strong coverage mitigates the out-of-distribution extrapolation errors [8, 15] that typically plague offline RL. However, as we generalize to richer states and several actions, data coverage becomes sparse, requiring pessimism to be modeled explicitly [14, 15, 17].

2.1 Constrained RL Problem Notation

We model user-agent interactions as a CMDP $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma, \mu, \omega, B \rangle$ [2, 20]. Here $\mathcal{S}$ is the set of discrete customer lifecycle states and $\mathcal{A}$ is the set of actions, where $a_0 \in \mathcal{A}$ denotes the *No Intervention* action and $\mathcal{A}_{\mathrm{I}} = \mathcal{A} \setminus \{a_0\}$ is the set of *Intervention* actions. The dynamics are given by the transition probabilities $P(s' \mid s, a)$ and the immediate rewards $R(s, a)$. The time discount factor is denoted by $\gamma \in (0, 1)$, and initial state distribution is given by $\mu(s)$. To enforce business guardrails, $\omega \in \mathbb{R}^{|\mathcal{A}|}$ specifies the state-invariant action costs, with $\omega(a_0) = 0$. Finally, $B \in \mathbb{R}$ is the scalar platform budget cap. The remaining operational thresholds (cohort, exposure, competition, and fairness) are introduced alongside their respective constraints.

2.2 Learn Causal Model Dynamics

To account for logging policy bias, we first estimate the propensity scores $\hat{\pi}_b(a \mid s, z)$ [21] via multinomial logistic regression on the historical data, conditioning on observed states s and confounders z that span historical behavior and context. Using $\hat{\pi}_b$, we compute Inverse Propensity Weights (IPW) $\hat{w}_i = 1/\hat{\pi}_b(a_i \mid s_i, z_i)$ for each observed one-step transition i. To mitigate high variance associated with the standard IPW, we use Self-Normalized Inverse Propensity Weights (SNIPW) and estimate causally corrected transition probabilities $\hat{P}(s' \mid s, a)$ and rewards $\hat{R}(s, a)$ as shown in Equation (1).

$$\hat{P}(s' \mid s, a) = \frac{\sum_{i \in \mathcal{D}_{s,a,s'}} \hat{w}_i}{\sum_{i \in \mathcal{D}_{s,a}} \hat{w}_i}, \qquad \hat{R}(s, a) = \frac{\sum_{i \in \mathcal{D}_{s,a}} \hat{w}_i \, r_i}{\sum_{i \in \mathcal{D}_{s,a}} \hat{w}_i} \tag{1}$$

The causal validity of these re-weighted estimates rests on the *strong ignorability* [21] assumption. This comprises two conditions: *unconfoundedness*, meaning (s, z) captures all common causes of the action and outcomes, and *positivity*, meaning $\pi_b(a \mid s, z) > 0$ for every action a available in state s and all (s, z) in the data. Under this assumption, the re-weighted estimators $\hat{P}(s' \mid s, a)$ and $\hat{R}(s, a)$ recover the true interventional transition distribution $P(s' \mid \mathrm{do}(a), s)$ and the true interventional reward $R(s, \mathrm{do}(a))$, respectively, rather than their confounded observational counterparts. Any unobserved confounding violates this assumption, introducing residual bias in $\hat{P}$ and $\hat{R}$ that propensity weighting alone cannot eliminate. To mitigate any numerical instability that may arise from extreme propensity weights, we consider percentile-based clipping to the weights.

2.3 Linear Program

The CMDP is formulated through its dual LP using the state-action occupancy measure ($\zeta(s,a)$) [2, 20]. The formulation maximizes the expected discounted cumulative reward, subject to model dynamics (see Equation (1)) and comprehensive business constraints (see Table 1). The optimal state-action occupancy measure $\zeta^\star(s,a)$, obtained from solving the LP, is used to extract the optimal policy $\pi^\star(a \mid s)$ [20].

Table 1: Supported classes of business constraints

Constraint Class	Operational Definition / Business Goal
Platform Budget	Caps overall system level incentive/spend.
Cohort Allocation	Restricts or guarantees intervention spend within target segments.
Action Exposure	Caps frequency per intervention type to prevent user fatigue.
Competing Actions	Allocates a shared, finite budget across mutually exclusive actions.
Action Eligibility	Excludes invalid state-action pairs (e.g., subscription targeting for already subscribed).
Fairness Guardrails	Guarantees minimum action exposure to prevent any systematic segment neglect.

$$\max_{\zeta} \quad \sum_{s\in\mathcal{S},\, a\in\mathcal{A}} \zeta(s,a)\,\hat{R}(s,a) \tag{2}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} \zeta(s,a) - \gamma \sum_{\substack{s'\in\mathcal{S}\\ a'\in\mathcal{A}}} \hat{P}(s \mid s',a')\,\zeta(s',a') = \mu(s), \quad \forall s \in \mathcal{S} \tag{3}$$

$$(1-\gamma) \sum_{s\in\mathcal{S},\, a\in\mathcal{A}} \omega(a)\,\zeta(s,a) \le B, \tag{4}$$

$$\underline{B}_g \le (1-\gamma) \sum_{s\in\mathcal{S}_g,\, a\in\mathcal{A}} \omega(a)\,\zeta(s,a) \le \overline{B}_g, \quad \forall g \in \mathcal{G} \tag{5}$$

$$(1-\gamma) \sum_{s\in\mathcal{S}} \zeta(s,a) \le E_a, \quad \forall a \in \mathcal{A}_{\mathrm{I}} \tag{6}$$

$$(1-\gamma) \sum_{s\in\mathcal{S},\, a\in\mathcal{A}_j} \omega(a)\,\zeta(s,a) \le C_j, \quad \forall j \in \mathcal{J} \tag{7}$$

$$\zeta(s,a) = 0, \quad \forall (s,a) \in \mathcal{E} \tag{8}$$

$$(1-\gamma) \sum_{s\in\mathcal{S}_g} \zeta(s,a) \ge F_{g,a}, \quad \forall g \in \mathcal{G},\ a \in \mathcal{A}_{\mathrm{I}} \tag{9}$$

$$\zeta(s,a) \ge 0, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{10}$$

The objective (2) maximizes the expected discounted cumulative reward, which is linear in the occupancy measure $\zeta(s,a)$. Here $\zeta(s,a)$ denotes the total expected discounted number of times the agent visits state s and takes action a under a given policy. Because ζ is unnormalized, every rate constraint carries a $(1-\gamma)$ factor that converts occupancy mass into a normalized frequency [2].

The constraints jointly define the feasible policy space. Equation (3) imposes *Bellman flow conservation* under the causally corrected dynamics $\hat{P}$, ensuring that ζ is a valid occupancy measure; the remaining constraints encode the business guardrails and operational constraints mentioned in Table 1. Specifically, Equation (4) caps the system-level intervention cost at B, while Equation (5) confines per cohort spend to $[\underline{B}_g, \overline{B}_g]$. Equation (6) bounds each intervention's long-run frequency (e.g., $E_a = 0.05$ limits exposure to promotions to at most 5% of interactions). Equation (7) caps the shared budget of mutually exclusive and competing action groups. Equation (8) prohibits ineligible state-action decisions. Equation (9) guarantees a minimum per cohort exposure to various interventions to ensure fairness. Finally, Equation (10) enforces non-negativity so that ζ remains a valid measure.

On solving the above LP, we directly obtain the optimal state-action occupancy measure $\zeta^\star(s,a)$. The corresponding optimal policy $\pi^\star(a \mid s)$ is then extracted by normalizing over actions (see Equation (11)). For states with $\sum_a \zeta^\star(s,a) = 0$, the normalization is undefined; such states carry zero occupancy mass, and to preserve the budget constraint under a plausible distribution shift at deployment, we default to the *No Intervention* action ($\pi^\star(a_0 \mid s) = 1$ and $\pi^\star(a \mid s) = 0$ for all $a \in \mathcal{A}_{\mathrm{I}}$). In rare cases, the LP solution may have multiple optima $\zeta^\star$ (e.g., two actions with identical long-term value at a state); the LP solver breaks such ties arbitrarily, with no effect on the objective. Finally, because the problem is constrained, the resulting optimal policy is in general *stochastic*.

$$\pi^\star(a \mid s) = \frac{\zeta^\star(s,a)}{\sum_{a'\in\mathcal{A}} \zeta^\star(s,a')}. \tag{11}$$

Solving the LP also provides us with shadow prices associated with each constraint, which serve as an auditable rationale for the resulting allocation and support budget planning. In particular, the dual variable $\lambda_B^\star$ of the platform-level budget constraint (Equation (4)) gives the marginal gain in expected long-term value per additional unit of spend, i.e., the return on the next budget dollar. Similarly, the duals of the cohort, exposure, competing actions, and fairness constraints are the shadow prices of the respective business guardrails. Each provides the marginal long-term value forgone by enforcing that guardrail, or equivalently, the value to gain from relaxing it by one unit. Finally, nonzero duals identify *binding* constraints actively shaping the policy, while zero duals indicate *slack*, making the trade-offs of every decision transparent and verifiable.

3 Case Study

To evaluate the NBA framework, we conduct an offline coupon-targeting case study using 68 weeks of proprietary historical data (consisting of a 52 week training period and a subsequent 16 week temporal holdout) from the rideshare platform with millions of active users. We map historical weekly rider interactions into discrete states, actions, rewards, and observed confounders. States are defined via RFM (Recency, Frequency, and Monetary) segmentation [5], classifying customers into eight lifecycle stages (e.g., *Champion*, *Hibernating*) based on the RFM values of their rides in the

past 28 days (see Table 3). The *Coupon Action* is a ride discount offered to a customer in a given week, incentivizing them to ride; *No Action* does not provide this incentive. The immediate reward is a blended business value metric generated by that customer in that week. NBA maximizes the expected discounted cumulative reward a customer generates, which maps to customer LTV [3], i.e., the long-run business value of a customer. The observed confounders are comprehensive and span historical behavior and ride context.

Following extant literature [5, 19], we assume RFM-based representations approximately satisfy the Markov property. We also independently validated that RFM-based features are the strongest predictors of rider LTV. Furthermore, given the few, finite states ($|\mathcal{S}| = 8$) and binary action ($|\mathcal{A}| = 2$), our industrial dataset provides high data support across all 16 state-action pairs.

We benchmark candidate policies using model-based off-policy evaluation on a temporal holdout. The optimal policy $\pi^\star$ is trained on the first 52 weeks of data (D_{train}) and evaluated on SNIPW-corrected transition dynamics $\hat{P}_{\text{eval}}$ and reward dynamics $\hat{R}_{\text{eval}}$, estimated from the subsequent 16 week holdout (D_{eval}). To prevent instability in the SNIPW estimator, we clip propensity weights at the 99th percentile, following standard practice in causal inference.

Table 2: Case study formulation

Components	Description
Learning Episode (τ)	Sequence of a rider's states, actions, rewards, and observed confounders over 52 weeks
Time step (t)	Weekly decision stages
States ($s_t \in S$)	Rider lifecycle stages, based on RFM
Action ($a_t \in A$)	(1) No Action (2) Coupon Action
Rewards (r_t)	Blended business metric
Discount factor (γ)	Weights on future rewards, $\gamma = 0.9$
Optimal Policy ($\pi^\star$)	$\arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t)\right]$ subject to constraints (i.e. maximizes constrained rider LTV)

Table 3: RFM segmentation states (H=High, L=Low)

States	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Narrative	*Champion*	*Loyalist*	*Promising*	*New*	*At Risk*	*Slipping*	*Hibernating*	*Dormant*
Recency	H	H	H	H	L	L	L	L
Frequency	H	H	L	L	H	H	L	L
Monetary	H	L	H	L	H	L	H	L

Note: The High/Low classification is determined by a static median split.

4 Results and Discussion

In this section, we analyze the estimated causal transition and reward dynamics, benchmark the obtained optimal policy in terms of long-term value, and examine the optimal coupon policy and its returns across varying budget constraints.

4.1 Causal Transition and Reward Dynamics

Figure 3 presents the expected immediate rewards $\hat{R}(s, a)$ after causal correction. Under the coupon action, the *Hibernating* segment yields nearly three times its baseline reward, while the *Champion* segment shows only a marginal gain. Similarly, Figure 4 illustrates the causal effect of the coupon action on state transitions $\hat{P}(s' \mid s, a)$. Exposing the *Promising* segment to a coupon more than doubles its probability of transitioning to the *Champion* state relative to the baseline.

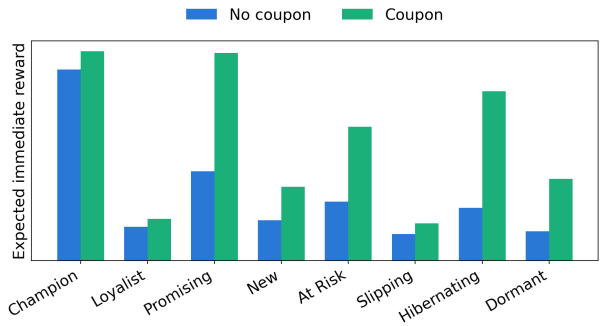

Figure 3: Learned causal reward estimates

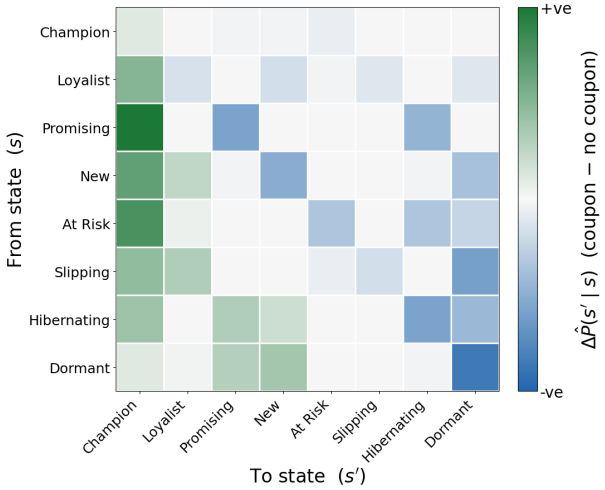

Figure 4: Causal effect of coupon on state transitions

4.2 Policy Performance Benchmarking

We benchmark the (a) *Optimal Policy* ($\pi^\star$) obtained from solving the LP against the (b) *Behavior Policy* ($\hat{\pi}_b$), i.e., the historical coupon allocation pattern across RFM segments, and the (c) *Myopic Policy*, which greedily maximizes the immediate reward $\hat{R}(s, a)$. All three are obtained and evaluated offline under the observed coupon rate cap. While the behavior policy distributes relatively more coupons to champions and loyalists, the optimal policy redirects some of these coupons toward lapsed high-value customers (at risk, hibernating) and recent low-frequency customers (promising, new) who need the incentive to return or form habits (see Figure 5). The optimal and myopic policies diverge most notably on dormant versus

new customers. While the myopic policy prioritizes recently dormant customers for their larger immediate uplift, the optimal policy invests in new customers whose habit formation compounds into higher lifetime value. Under offline evaluation, with causal dynamics estimated on a temporal holdout, the optimal policy yields an estimated 19% and 8% improvement in expected customer LTV over the behavior and myopic baselines, respectively. The offline experiments could help select a winning candidate for online validation.

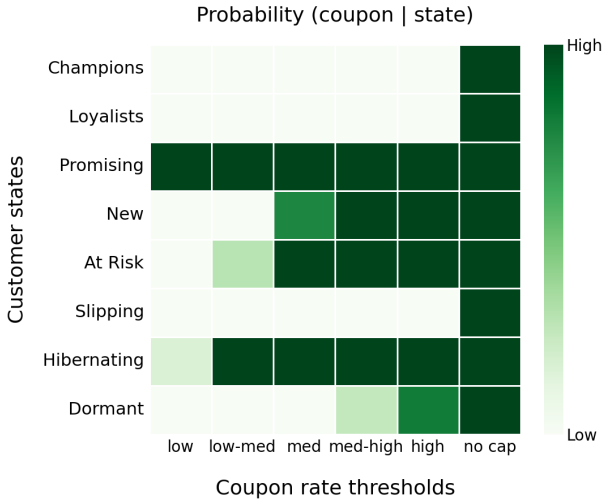

Figure 5: Optimal policies under coupon caps

4.3 Coupon Budget Planning

A key strength of the LP formulation is its ability to model budget constraints and business guardrails. Figure 5 visualizes the optimal policy $\pi^\star(a \mid s)$ as the coupon budget scales from highly restrictive to unconstrained. At the historical coupon rate, the optimal policy is highly selective regarding which rider segments receive coupons. However, when the budget constraint is removed, coupons are distributed across all stages of the customer lifecycle. Notably, the *Promising* and *Champion* segments yield the highest state values or expected LTV.

Beyond policy optimization, the LP formulation provides a practical tool for budget planning and scenario analysis. We evaluate returns across varying budget levels and fit an exponential saturation function to trace the budget-performance trade-off curve in Figure 6. This curve shows diminishing marginal returns on coupon expenditures, helping managers identify effective spending thresholds.

5 Conclusion and Future Work

We propose a general-purpose, auditable action-selection module for business action targeting. The NBA framework unifies causal inference, counterfactual evaluation, model-based offline RL, and constrained optimization to maximize long-term value under six classes of operational constraints. We validate the framework on a coupon targeting case study in rideshare. Under offline evaluation, with causal dynamics estimated on a temporal holdout, the optimal policy provides an estimated 19% and 8% improvement in expected customer LTV over the behavior and myopic baselines, respectively. The framework additionally supports robust budget planning and what-if scenario analysis.

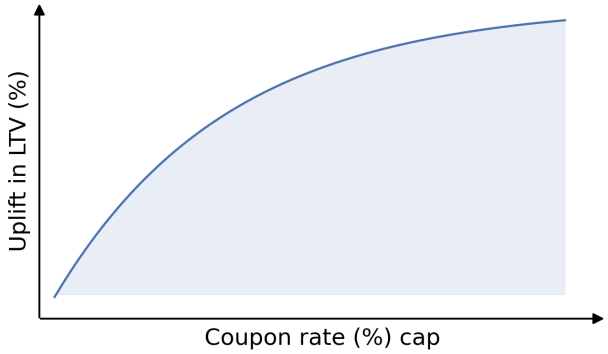

Figure 6: Fitted uplift in LTV under coupon caps

The results are based on offline policy evaluation and can help identify a winning candidate policy for potential online validation. Specifically, an online A/B test against the behavior policy may be conducted to further validate these findings. Future work could expand the state representation and incorporate multiple actions to test the framework's action-agnostic design. However, if we generalize to richer states and multi-actions, sparse data coverage may arise, requiring pessimism to be modeled explicitly [14, 15, 17]. Finally, additional static user features could be modeled as context, similar to contextual MDPs [12], allowing for personalized transition and reward dynamics without exploding the state space.

References

[1] Javier Albert and Dmitri Goldenberg. 2022. E-Commerce Promotions Personalization via Online Multiple-Choice Knapsack with Uplift Modeling. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*.

[2] Eitan Altman. 2021. *Constrained Markov decision processes*. Routledge.

[3] Paul D Berger and Nada I Nasr. 1998. Customer lifetime value: Marketing models and applications. *Journal of interactive marketing* 12, 1 (1998), 17–30.

[4] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14 (2013), 3207–3260.

[5] Daqing Chen, Sai Laing Sain, and Kun Guo. 2012. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management* 19, 3 (2012), 197–208.

[6] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM)*.

[7] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*.

[8] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

[9] Dmitri Goldenberg, Javier Albert, Lucas Bernardi, and Pablo Estevez. 2020. Free lunch! retrospective uplift modeling for dynamic promotions recommendation within roi constraints. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 486–491.

[10] Shashank Gupta, Olivier Jeunen, Harrie Oosterhuis, and Maarten de Rijke. 2024. Optimal baseline corrections for off-policy contextual bandits. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 722–732.

[11] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal Inference and Uplift Modelling: A Review of the Literature. In *Proceedings of the 3rd International Conference on Predictive Applications and APIs (Proceedings of Machine Learning Research,*

Vol. 67). 1–13.

[12] Assaf Hallak, Dotan Di Castro, and Shie Mannor. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259* (2015).

[13] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. 2019. SlateQ: A Tractable Decomposition for Reinforcement Learning with Recommendation Sets. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*.

[14] Olivier Jeunen and Bart Goethals. 2021. Pessimistic Reward Models for Off-Policy Learning in Recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys)*.

[15] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems* 33 (2020), 1179–1191.

[16] Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. 2022. COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation. In *International Conference on Learning Representations (ICLR)*.

[17] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[18] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. 2019. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems* 32 (2019).

[19] Phillip E Pfeifer and Robert L Carraway. 2000. Modeling customer relationships as Markov chains. *Journal of interactive marketing* 14, 2 (2000), 43–55.

[20] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[21] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[22] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*.

[23] Guy Shani, David Heckerman, and Ronen I. Brafman. 2005. An MDP-Based Recommender System. *Journal of Machine Learning Research* 6 (2005), 1265–1295.

[24] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

[25] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems* 28 (2015).

[26] Georgios Theocharous, Philip S. Thomas, and Mohammad Ghavamzadeh. 2015. Personalized Ad Recommendation Systems for Life-Time Value Optimization with Guarantees. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

[27] Ruiyang Xu, Jalaj Bhandari, Dmytro Korenkevych, Fan Liu, Yuchen He, Alex Nikulkov, and Zheqing Zhu. 2023. Optimizing long-term value for auction-based recommender systems via on-policy reinforcement learning. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 955–962.

[28] Kui Zhao, Junhao Hua, Ling Yan, Qi Zhang, Huan Xu, and Cheng Yang. 2019. A Unified Framework for Marketing Budget Allocation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. 1820–1830.

[29] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference (WWW)*.

[30] Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. 2023. Direct Heterogeneous Causal Learning for Resource Allocation Problems in Marketing. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. 5446–5454.

[31] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*.