Geographical and Behavioral Information: Complements or Substitutes in Ad Personalization?

Mohammad Mosaffa* Cornell University

Omid Rafieian* Cornell University

Abstract

Firms collect vast amounts of behavioral and geographical data on individuals. While behavioral data captures an individual's digital footprint, geographical data reflects their physical footprint. Given the significant privacy risks associated with combining these data sources, it is crucial to understand their respective value and whether they act as complements or substitutes in achieving firms' business objectives. In this paper, we combine economic theory, machine learning, and causal inference to quantify the value of geographical data, the extent to which behavioral data can substitute it, and the mechanisms through which it benefits firms. Using data from a leading in-app advertising platform in a large Asian country, we document that geographical data is most valuable in the early "cold-start" stage, when behavioral histories are limited. In this stage, geographical data *complements* behavioral data, improving targeting performance by almost 20%. As users accumulate richer behavioral histories, however, the role of geographical data shifts: it becomes largely *substitutable*, as behavioral data alone captures the relevant heterogeneity. These results highlight a central privacy–utility trade-off: location data provides short-run targeting gains, yet it imposes substantial privacy risks and delivers little long-term value.

*Please address all correspondence to: mm3322@cornell.edu, or83@cornell.edu.

1 Introduction

Mobile devices now dominate the digital economy, reshaping how consumers interact with media, commerce, and advertising. Nearly all Americans own a mobile phone, and nine in ten use a smartphone (Pew-Center, 2024), creating an ideal environment for targeted marketing. In response to this shift, mobile advertising now accounts for more than two-thirds of all digital ad spending in the United States, surpassing $200 billion in 2024 (eMarketer, 2024). Much of this growth stems from advances in tracking technologies that allow advertisers to monitor users' digital and physical activities and tailor messages to individual consumers. While these capabilities have made mobile advertising a central channel for marketers, they have also intensified privacy concerns, raising questions about whether the economic value generated by such data practices justifies the privacy risks involved.

Privacy concerns arise most directly from two types of data collected by advertisers and advertising platforms: *behavioral data*, such as click histories, search queries, or in-app purchases, and *geographical data*, such as GPS coordinates, IP-based locations, or broader regional indicators (Ghose et al., 2019). Behavioral data capture a user's digital footprint within apps, while geographical data capture their physical footprint—their movements and proximity in the real world. Geographic information is particularly sensitive, as even a few location points can be sufficient to re-identify individuals (De Montjoye et al., 2013). When combined with behavioral data, these risks amplify: the two data types complement each other in enabling more precise inference of personal attributes and increasing the likelihood of re-identification, even when datasets are anonymized (De Montjoye et al., 2018).

Given the complementarity of geographical and behavioral information in amplifying privacy risks, it is crucial to examine whether a similar complementarity exists in the factor that motivates firms to continue collecting them—the economic value they create. In particular, it remains unclear whether behavioral and geographical data act as *substitutes* or *complements* in generating economic value for advertisers. Addressing this question requires first establishing how to measure the value of each data type, then comparing their relative contributions and interactions, and finally identifying the mechanisms through which geographical data may provide distinct informational value. These considerations motivate the following research questions:

1. How can we measure the value of geographical data in advertising relative to behavioral data?
2. To what extent do geographical data generate value once behavioral data are available, and how does this value change as behavioral data accumulate?
3. Through which mechanisms do geographical data provide distinct information, and does it persist after accounting for behavioral data?

We face several challenges in answering these questions. The first challenge is conceptual: how should we define the value of information? While one common approach is to focus on prediction accuracy, such as testing whether adding geographical data improves click-through forecasts, this metric alone does not guarantee better advertising outcomes (Ascarza, 2018). What matters is whether available data improve the quality of decisions. In our framework, we draw on economic theory to define the value of geographical

and behavioral data relative to one another through their impact on decision outcomes. This perspective allows us to test whether the two sources act as *complements*, where their joint use delivers gains larger than the sum of their individual effects, or as *substitutes*, where the contribution of one source diminishes once the other is available. This definition matters because it directly links information value to real trade-offs: if geographical data merely duplicate what behavioral data already capture, their collection imposes privacy costs without improving outcomes, whereas if they provide complementary value, both firms and policymakers can justify their use.

The second challenge is empirical: to assess the value of different information sets in terms of decision quality, we must first be able to prescribe policies conditional on each information set. Doing so requires models that capture the underlying structure of the data. Behavioral data are temporal, reflecting evolving histories of impressions, while geographical data are high-dimensional, spanning thousands of regions with spatial correlations rather than simple coordinates. If these features are ignored, the prescribed policies may be misspecified, leading to biased conclusions about the contribution of data. To address this challenge, we design machine learning architectures that flexibly capture the temporal and spatial structure of the information sets, leveraging long short-term memory (LSTM) networks with attention mechanisms.

The third challenge is methodological: policy evaluation requires counterfactual outcomes. Observed data only show how users responded under the targeting policies they actually received; they do not reveal how the same users would have responded under alternative policies, such as targeting with or without geographical data. To address this, we draw on the causal inference literature and apply inverse propensity scoring (IPS) (Horvitz and Thompson, 1952), which estimates counterfactual outcomes by reweighting observed responses with assignment probabilities. A key advantage of our setting is the platform's proportional auction mechanism, which allocates ads in proportion to bid-weighted scores. This mechanism generates plausibly exogenous variation that supports credible propensity score estimation and the validity of our counterfactual evaluation.

Together, these three challenges motivate a unified framework. We combine economic theory to define information value through comparison, machine learning to estimate response functions from complex behavioral and geographic data, and causal inference to leverage IPS for counterfactual evaluation. This framework formally quantifies the value of two data sources and provides a statistical test of whether they operate as complements or substitutes. While our application focuses on geographical and behavioral data, the approach generalizes to other settings where multiple information sets guide decision-making.

We apply our framework to data from the leading mobile ad network in a major Asian country, covering 10.5 million impressions from 439,344 users over a 10-day period. To implement the framework, we define four targeting scenarios: the benchmark $X^{\varnothing}$ uses only contextual information (e.g., app category or time of day); X^{G} augments contextual features with geographical data such as city or province; X^{B} augments contextual features with behavioral data from users' prior impression and click histories; and X^{GB} combines both. Because the data structure differs across scenarios, we adopt models appropriate to each case: gradient-boosted trees (XGBoost) for $X^{\varnothing}$ and X^{G}, where features are static or cross-sectional, and recurrent neural networks (LSTM with an attention mechanism) for X^{B} and X^{GB}, where sequential behavioral histories must be captured. Leveraging the quasi-proportional auction rule and consistent propensity

score estimation, we use IPS to recover the counterfactual click-through rates for each scenario. We then apply our framework, which combines these counterfactual estimates with our decision-based definition of information value, to quantify the contribution of behavioral and geographical data and to test whether they act as complements or substitutes.

At the aggregate level, we first assess all targeting regimes relative to the raw baseline CTR in the data. Targeting without user-level data ($X^{\varnothing}$), which serves as our benchmark, improves CTR by 20.44%. Adding geographical data (X^{G}) raises CTR by 28.7%, behavioral data (X^{B}) by 30.3%, and leveraging both together (X^{GB}) delivers an improvement of 41.5%. Although the joint regime achieves the highest overall performance, our aggregate test does not provide systematic evidence that geographical and behavioral data act as either complements or substitutes. This motivates turning to heterogeneity in user histories to examine how their relationship evolves as behavioral data accumulate.

As users are exposed to more impressions, advertisers gain a richer behavioral record, changing both the value of behavioral data and how it interacts with geographical data. In the earliest stage (1–2 impressions), behavioral data provide almost no information, so performance is driven almost entirely by geographical data, and the combined regime performs nearly the same as geography alone. In the intermediate stage (about 5–25 impressions), behavioral data begin to accumulate but remain noisy. In this range, behavioral data and geographical data act as complements, with the combined model delivering gains substantially larger than either source in isolation. Once users exceed 25 impressions, behavioral data becomes sufficiently rich to capture preferences on its own, and geographical data shifts into a substitute, offering little or no incremental value. Together, these results show that the role of geographical data is dynamic: initially dominant when behavioral data are absent, complementary when behavioral data are sparse, and ultimately substitutable once behavioral data are rich.

The shift from complementarity to substitutability naturally raises the question of *why* geography matters in the first place. While our framework establishes whether geographical data adds incremental value beyond behavioral data, it does not identify how this occurs. Conceptually, geography could improve targeting through three channels: homophily, where nearby users share preferences; social influence, where engagement diffuses locally; and confounding, where contextual events or local shocks affect specific regions. As behavioral histories grow, repeated actions increasingly reveal individual preferences and absorb much of the homophily channel, diminishing the role of geography. Influence and confounding, however, may persist beyond what behavior alone captures. To test whether geographical data provides independent information once behavior is accounted for, we develop a residualized spatial autocorrelation (RSA) test that measures whether ad responses continue to cluster spatially after conditioning on behavioral data.

Applying this test, we document three main results. First, raw ad responses exhibit strong spatial autocorrelation, indicating meaningful geographic structure in engagement. Second, once outcomes are residualized on behavioral data, most of this correlation disappears, consistent with repeated user actions absorbing underlying similarities. Third, residual clustering remains only when behavioral histories are sparse but vanishes entirely when histories are rich. Together, these findings suggest that geographical data contribution derives primarily from homophily that behavioral data eventually capture, leaving little independent role for geography once user-level histories are sufficiently informative.

Our findings carry direct implications for both managers and policymakers. Geographical data create value only in the short run, acting as a temporary complement when behavioral data are sparse, but quickly become substitutable as user histories accumulate. For advertisers, this implies that geo-targeting can be useful during the cold-start phase but delivers little incremental benefit once behavioral models are established. For regulators, the results highlight a critical trade-off: location data poses substantial privacy risks while offering limited long-term benefits in digital advertising. Firms should therefore reconsider the strategic role of geographical data, employing it selectively when behavioral information is unavailable and phasing it out as richer user profiles emerge.

In summary, our paper offers several contributions to the literature. First, while prior research has emphasized the value of either behavioral or geographical data in isolation, we provide the first systematic comparison of the two. Second, we develop a unified framework that combines economic theory to define information value, machine learning to model advertising responses, and causal inference to evaluate counterfactual outcomes. This framework enables managers and policymakers to assess whether different types of information operate as complements or substitutes, offering a practical tool to evaluate the value of information in targeting decisions while considering potential privacy costs. Substantively, we show that both geographical and behavioral data generate value, but their roles evolve as behavioral information accumulates: geography complements behavior when histories are sparse but becomes substitutable once behavioral information is rich. Mechanistically, we demonstrate that the early value of geography reflects spatial clustering consistent with homophily not yet captured by behavioral models, and that this effect fades as behavioral data accumulate. These results imply that geo-targeting offers short-term benefits but limited long-term usefulness, underscoring the importance for firms and regulators of weighing its contribution against potential privacy concerns.

The remainder of the paper is organized as follows. §2 reviews related work on the value of information, advertising and privacy, and spatial economics. §3 introduces the institutional setting of mobile advertising and describes our large-scale dataset. §4 formalizes the problem within a decision-under-uncertainty framework to define the value of information and highlight key challenges, while §5 presents our identification strategy, beginning with the definitions of complementarity and substitutability and leveraging machine learning for click prediction and inverse propensity scoring for counterfactual evaluation. §6 reports empirical findings on the value of behavioral and geographical data and their heterogeneity, and §7 examines the mechanisms behind these patterns using a residualized spatial autocorrelation test. §8 discusses implications for data use, privacy, and targeting efficiency, and §9 concludes with a summary of contributions and directions for future research.

2 Related Work

First, our work relates to research on how the value of information is quantified for decision-making under uncertainty. Theoretically, Blackwell (1953) provides the benchmark by ordering signals according to whether they increase expected utility across all decision problems (the Blackwell order). Extending to multiple signals, Börgers et al. (2013) characterize conditions under which signals act as complements or substitutes by examining whether one signal's marginal value rises or falls in the presence of another. Kamenica and Gentzkow (2011) study persuasion games, valuing information through the sender's induced

equilibrium payoff and the actions it elicits. From a market-design perspective, Bergemann and Bonatti (2011) represent informativeness with a targeting-precision parameter and trace how greater precision shifts matches, ad prices, and revenues.

Empirically, related work evaluates how data access and targeting policies affect outcomes using causal strategies. Randomized experiments include Ascarza (2018), who compare targeting rules based on churn risk versus treatment-effect lift, and Wernerfelt et al. (2025), who conduct a large-scale RCT that withholds offsite tracking data to measure its incremental value. Natural experiments such as Aridor et al. (2024) exploit Apple's App Tracking Transparency as an exogenous shock to behavioral data access and document associated performance losses. For off-policy evaluation (OPE), Rafieian and Yoganarasimhan (2023) document techniques such as inverse propensity scoring and doubly robust estimation to estimate policy value from logged or experimental data, and Rafieian (2023) apply OPE within an offline reinforcement-learning framework to optimize dynamic ad sequencing. While these studies credibly measure the impact of specific information set or targeting strategies, they do not provide a unified framework for systematically comparing multiple information sets in high-dimensional settings. Our study builds on this literature by integrating economic theory, causal inference, and machine learning to estimate policy value across alternative information regimes and formally test whether information sets act as complements or substitutes.

Second, our study connects to the literature on privacy and personalized advertising, particularly in user tracking and engagement modeling. A key challenge in digital advertising is balancing effective targeting with privacy concerns. Acquisti et al. (2016) discuss the trade-offs firms face between personalization benefits and consumer resistance to data collection, while Tucker (2014) show that personalized ads in social networks can increase relevance but also drive demand for stronger privacy controls. In display advertising, Goldfarb and Tucker (2011) find that ad intrusiveness and privacy sensitivity significantly impact engagement, with well-targeted yet subtle ads performing best. Rafieian and Yoganarasimhan (2021) explore how privacy restrictions affect targeting efficiency, showing that detailed tracking improves targeting but may reduce market competition. Johnson et al. (2020) quantify the cost of consumer opt-outs, finding that although opt-outs are rare, they lead to substantially lower ad revenues. A major but often overlooked issue is the role of geographical data in this trade-off. While De Montjoye et al. (2013) demonstrate that even anonymized mobility traces can be re-identified with high accuracy, exposing vulnerabilities in location-based tracking, prior research has not examined how geographical data contribute to targeting effectiveness and the privacy risks they introduce. Our study fills this gap by investigating to what extent geographical data can provide value beyond behavioral data.

Third, empirical work in marketing and economics applies econometric and spatial diagnostics to detect and interpret geographic dependence in outcomes. Foundational tests such as Moran's I and Geary's C provide global measures of spatial autocorrelation in outcomes and regression residuals (Moran, 1950; Geary, 1954). Building on these diagnostics, Bronnenberg and Mahajan (2001) model spatial dependence in market shares and promotions across neighboring markets to capture correlated, unobserved retailer actions, and Bronnenberg et al. (2009) document persistent geographic structure in brand demand, showing that brands retain higher shares near their historical origins using cross-city spatial patterns. In studies of diffusion and contagion, Manchanda et al. (2008) and Iyengar et al. (2011) exploit adjacency in physician

and social networks to separate targeted communication from peer influence, and Bollinger and Gillingham (2012) use zip-code–level exposure to quantify neighborhood effects in solar adoption. Although the literature leverages spatial correlation extensively, existing approaches cannot determine whether observed geographic structure reflects distinct information or is subsumed by other information sets, such as behavioral data. We develop a residualized spatial autocorrelation test that identifies whether geography explains systematic spatial structure in outcomes beyond what behavioral histories account for.

3 Setting & Data

This section begins by outlining the institutional setting of the mobile advertising platform (§3.1), followed by a description of the dataset of impressions, clicks, and contextual variables (§3.2). We then present our sampling strategy and summary statistics (§3.3), and conclude with the train–test split design used for model evaluation (§3.4).

3.1 Setting

Our data are sourced from a leading mobile in-app advertising platform in a large Asian country, which held over 85% of the mobile advertising market during the time of our study. The platform serves as an intermediary between advertisers and mobile app publishers and is responsible for delivering over 50 million ad impressions daily. This marketplace consists of four primary players:

- **Users** are mobile app consumers who generate impressions and may choose to click on displayed ads.
- **Publishers** are app developers that integrate ads into their apps and monetize based on ad clicks.
- **Advertisers** design banner ads and specify per-click bids. They can target users based on variables such as province, smartphone brand, app category, internet service provider (ISP), hour of the day, and connectivity type. The ad network does not support detailed personalized targeting.
- **Platform or ad network** manages real-time auctions to match impressions with ads. Ads are placed as bottom banners and are refreshed every minute. Only clicks result in payment under a cost-per-click (CPC) scheme.

Figure 1 illustrates the structure of the in-app advertising marketplace. When a user opens an app, the platform initiates a real-time auction among eligible ads, allocating impressions through a quasi-proportional rule: ads with higher bid–quality products are more likely to be shown, though selection remains stochastic. Quality scores are fixed and not customized at the user level, meaning there is no user-level personalization. If the user remains active beyond one minute, a new impression is generated and another auction is run.

Given this environment, we are able to observe ad impressions at a granular level and, more importantly, track individual users over time. This allows us to study the evolution of user behavior from the moment they enter the system, capturing both their initial exposure and how their responsiveness changes as more behavioral data accumulates. The repeated nature of ad exposures, combined with detailed contextual and user-level information, creates a unique opportunity to analyze how geographical and behavioral information contribute to predicting user responses in a real-world, high-frequency setting.

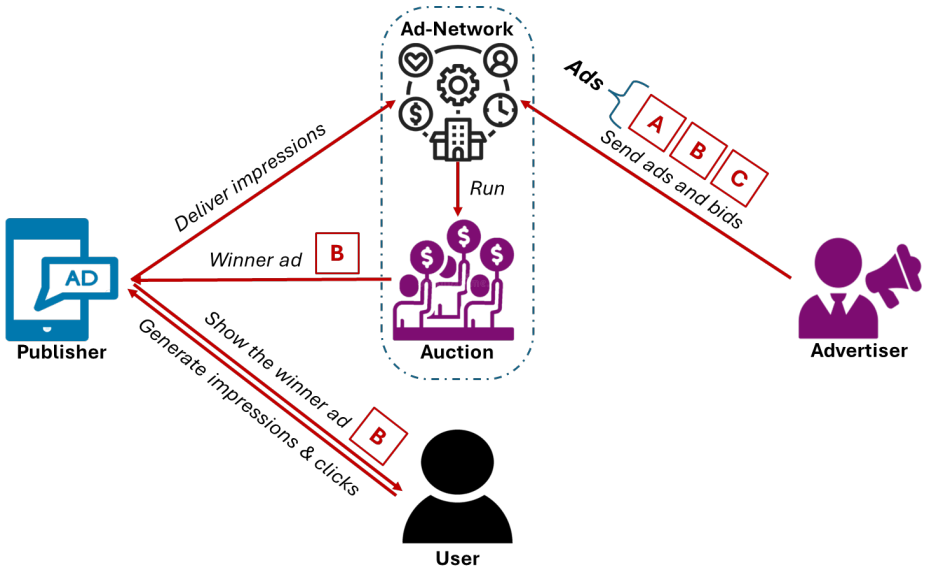


Figure 1: Schematic of the mobile in-app advertising marketplace.

3.2 Data

We have data on all impressions and click outcomes observed over a 30-day period from September 30, 2015, to October 30, 2015. During this period, we observe a total of 1,594,831,699 ad impressions and 14,373,293 clicks, resulting in an overall click-through rate (CTR) of approximately 0.90%. For each impression, the dataset includes detailed information across several dimensions:

For each impression, the dataset records a comprehensive set of variables. First, we observe *(1) Timestamp*, capturing the exact time of the impression, and *(2) AAID*, the Android Advertising ID, a user-resettable, unique device identifier that enables anonymous tracking across applications. At the ad-delivery level, we record *(3) AppID*, the identifier of the app displaying the ad, and *(4) CreativeID*, the identifier of the specific ad creative shown. We also observe *(5) Bid*, the advertiser's submitted bid amount (fixed throughout the sample period), and *(6) CPC*, the cost-per-click charged in the event of a click. Geographic attributes include *(7) Latitude*, *(8) Longitude*, and *(9) Province*, which are available for most impressions. Further contextual information comprises *(10) Connectivity*, indicating whether the user was on Wi-Fi or cellular data, *(11) Brand*, the smartphone manufacturer, *(12) MSP*, the mobile service provider, and *(13) ISP*, the internet service provider. Finally, *(14) Click* is a binary indicator equal to one if the user clicked on the ad during that impression and zero otherwise.

A key aspect of this dataset is that it is sourced directly from the ad platform and includes the complete set of variables that advertisers could potentially use for targeting. As such, we observe the same information available to both the platform and the advertisers at the time of impression delivery. This mitigates common concerns in observational studies around hidden targeting mechanisms or unobserved confounding. In our context, the completeness of the data supports modeling assumptions such as conditional ignorability, which are crucial for later analyses involving counterfactual analysis and policy evaluation.

3.3 Sampling & Summary Statistics

To enable user-level analysis while maintaining computational tractability, we restrict attention to impressions from the top 10 ads served between October 20 and October 30, yielding an initial sample of 16,662,783 impressions. We address missing values, particularly in variables that serve as key features in our analysis, by excluding records without geographic information, specifically latitude and longitude. After this filtering, the working dataset contains 10,537,179 impressions and has no remaining missing values in any other columns. From this dataset, we identify new users who joined the platform during the first three days of the window, October 20 to October 22, and follow their subsequent activity through October 30. This design allows us to observe user responsiveness from the start of their platform interaction, yielding a cohort of 439,344 unique users.

We now present some summary statistics for key categorical variables in the dataset. Table 1 reports, for each variable, the number of unique categories, the share of impressions associated with the top three values, and the number of non-missing observations.

Table 1: Summary Statistics for the Categorical Variables

Variable	**Number of categories**	**Share of top categories**			**Number of impressions**
		1st	**2nd**	**3rd**	
App	9,515	25.69%	9.44%	4.46%	10,537,179
Ad	10	22.16%	13.59%	12.64%	10,537,179
Unique User	439,344	0.26%	0.18%	0.09%	10,537,179
Smartphone brand	7	41.87%	30.57%	9.39%	10,537,179
Connectivity Type	2	55.14%	44.86%	0.00%	10,537,179
ISP	8	61.40%	26.73%	5.16%	10,537,179
Province	828	11.06%	8.96%	7.26%	10,537,179

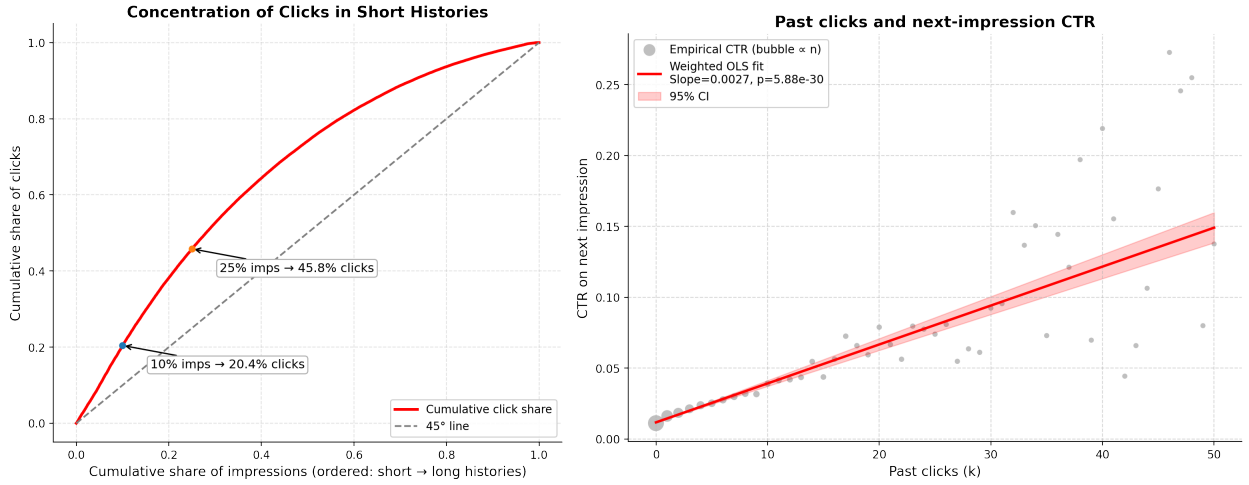
We observe a total of 9,515 unique apps, with the top three accounting for a sizable share of total impressions. Among the ten ads included by design, exposure is uneven, with the most frequently shown ad representing over 22% of impressions. The distribution of user identifiers, based on Android Advertising IDs, is highly diffuse, with no single user accounting for more than 0.26% of total impressions. Other variables, such as smartphone brand, connectivity type, ISP, and province, exhibit varying degrees of concentration, reflecting both common patterns and localized variation in usage across the population.

While Table 1 highlights variation in exposure across contextual features, it does not reveal how user responsiveness may differ across behavioral or geographical information. To explore this further, we present descriptive evidence on how CTR varies with user history and spatial location. These patterns help motivate the relevance of behavioral and geographical data for downstream modeling tasks.

3.3.1 Behavioral Heterogeneity

We present two descriptive results that illustrate how behavioral patterns shape user CTR. First, we examine how the length of a user's exposure history relates to CTR. Second, we analyze the relationship between past click behavior and the likelihood of future clicks.

Figure 2a plots the cumulative share of impressions against the cumulative share of clicks, where users are ordered by the number of prior impressions they have seen. The curve lies well above the 45° line: a disproportionate share of clicks comes from early exposures (e.g., the first 25% of impressions, corresponding to short histories, generate nearly half of all clicks). Figure 2b shows the probability of a click on the next impression, $\Pr(\text{click}_{t+1} = 1 \mid \text{prior clicks} = k)$, as a function of the number of past clicks k. The weighted linear fit has a positive and statistically significant slope: users who have accumulated more clicks are more likely to click again on the next impression.[1]



(a) Concentration of clicks in short histories

(b) Persistence in click behavior

Figure 2: Behavioral heterogeneity: (a) Clicks are concentrated in early exposure histories; (b) users with prior clicks are more likely to click again.

Taken together, these descriptive results reveal two dimensions of behavioral heterogeneity. First, Figure 2a shows that CTR declines as exposure histories lengthen, indicating the need to account for user trajectories. Second, Figure 2b shows persistence in click behavior: users with more prior clicks are more likely to click again, reflecting heterogeneity in underlying propensities and highlighting the predictive value of behavioral history.

3.3.2 Geographical Heterogeneity

We now turn to geographical patterns of responsiveness to explore how user response varies across space. To investigate this, we group impressions by county and compute the CTR within each spatial unit. Specifically, for a given county c, we calculate $\text{CTR}_c = \frac{1}{N_c} \sum_{i \in c} y_i$, where $y_i \in \{0,1\}$ indicates whether impression i resulted in a click, and N_c is the number of impressions observed in county c. Figure 3 visualizes the resulting CTR values as a choropleth map.

The figure shows that responsiveness varies meaningfully across counties. Some regions exhibit higher average CTRs, while others show lower responsiveness, even after aggregating over large numbers of impressions. These differences may reflect underlying geographic heterogeneity, shaped by socioeconomic

[1] Final impressions (with no $t+1$) are excluded. These patterns are descriptive and do not imply causal effects.

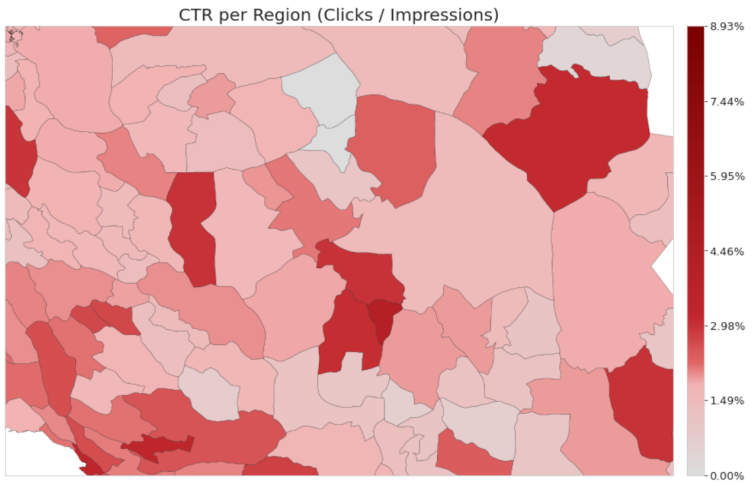


Figure 3: Average CTR by County

context, that influences users' likelihood of clicking on ads.

3.4 Train-Test Split Strategy

As part of our analysis evaluates how well predictive models perform out-of-sample (§5.2), we partition the data into training and test sets. To avoid information leakage, the split is performed at the user level rather than the impression level: each user's full history is contained entirely within either the training or the test partition. This design allows us to assess how models generalize to previously unseen users, which is central to evaluating the value of different information sets.

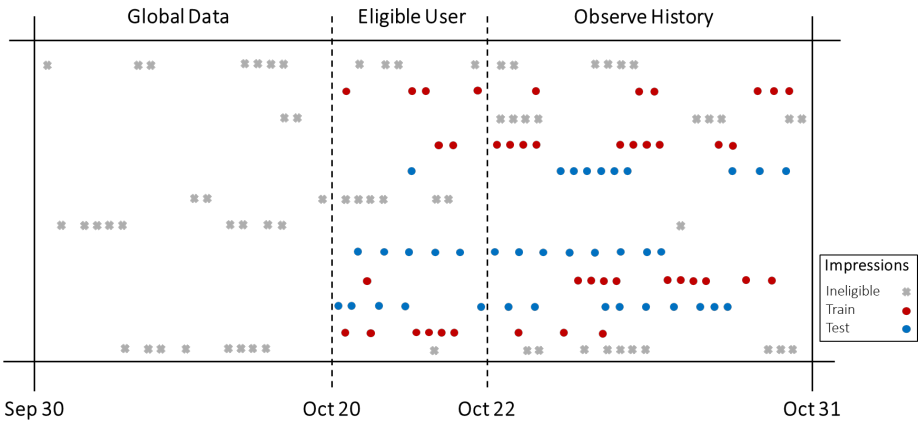


Figure 4: Schema for Data Generation and User-Based Train-Test Split

Figure 4 illustrates the procedure. A user becomes eligible once they first appear in the platform's logs, after which we track their impression history. Each user is then randomly assigned to either the training set (60%, 297,749 users; 6,356,413 impressions) or the test set (40%, 141,595 users; 4,180,766 impressions). Gray dots in the figure mark ineligible impressions that precede a user's first appearance, while red and blue dots denote impressions assigned to the training and test sets, respectively. The resulting split maintains

similar distributions of activity and click behavior across both partitions.

4 Problem Definition:

Firms operate in environments where advertising decisions must be made under uncertainty about user preferences. Consider the case of an ad network deciding which ad to show to a particular user. The central goal of the platform is to maximize some measure of effectiveness, such as the probability that the user clicks on the ad, despite not fully knowing who the user is or how they will respond. This problem is fundamentally one of decision-making under uncertainty, where the challenge lies in selecting the right action based on partial information about the user.

Let $Y \in \{0,1\}$ denote the binary outcome of user engagement (e.g., a click), and let $a \in \mathscr{A}$ denote the action taken by the platform (i.e., which ad is shown), where $\mathscr{A}$ is the set of available ads. The outcome Y is determined by a combination of the action a and the user's latent type $\omega \in \Omega$, which encodes unobserved characteristics such as preferences, motivations, or behavioral tendencies. This latent type is not directly observable, but the platform may have access to informative proxy data. We categorize available user-level information into two main types:

- **Behavioral Information** ($X^B \in \mathscr{X}^B$) captures a user's past interactions in the digital environment, including prior app usage patterns, ad exposure sequences, click history, and other engagement-based indicators that reflect individual preferences and interests over time.
- **Geographical Information** ($X^G \in \mathscr{X}^G$) describes the user's spatial context, encompassing attributes such as precise location (e.g., latitude/longitude) and broader administrative regions such as city/province that may correlate with demographic or regional characteristics.

We also treat *contextual information*, including device characteristics (e.g., brand, operating system) and network conditions (e.g., connectivity type), as part of the standard metadata attached to each ad impression. These variables, like behavioral and geographical information, reflect aspects of the user's latent type $\omega \in \Omega$, but they differ in two important respects. First, they are usually available to platforms for every impression, making them a natural baseline input. Second, consumers generally view them as substantially less privacy-sensitive than detailed behavioral or location-based data (Jerath and Miller, 2024). Accordingly, we include contextual information in all targeting regimes X^B and X^G. The specific variables included in each regime are described in §5.2.1.

One common approach to evaluating the value of information is through predictive performance. The goal is to learn a function $\hat{y} = f(X,a)$ that estimates $\mathbb{P}(Y = 1 \mid X,a)$. The quality of information is assessed via predictive metrics such as AUC or log-loss. In information-theoretic terms, better prediction corresponds to lower conditional entropy $H(Y \mid X,a)$, indicating that the information sharpens beliefs about the user. However, better predictive fit does not necessarily imply better decisions or outcomes when policies are deployed (Ascarza, 2018; Rafieian and Yoganarasimhan, 2023).

We therefore adopt a policy-based framework that evaluates information by its contribution to *decision quality*. Here, the focus is not predictive fit but the expected utility of policies guided by different information sets X. The central object is the *value function*, which quantifies the performance of a policy across the population of users. Formally, the setup is as follows:

- **Environment.** The platform must choose an action from a finite set $\mathscr{A}$, where each $a \in \mathscr{A}$ represents a feasible option (e.g., an ad that can be shown). Users differ in unobserved characteristics captured by a latent type $\omega \in \Omega$, with Ω denoting the space of possible user types. The utility of taking action a for a user of type ω is represented by a function:

$$u : \mathscr{A} \times \Omega \to \mathbb{R}, \qquad u(a, \omega) \text{ (e.g., expected CTR).}$$

  This utility function encodes the effectiveness or payoff of each action for each user type and serves as the platform's objective.

- **Information set and user types.** The platform does not observe the user's type ω directly. Instead, they observe an information set $X \in \mathscr{X}$ that provides partial knowledge about ω. The information set X may include behavioral variables, device attributes, geographic indicators, or other covariates. We assume the pair (X, ω) is drawn from a joint distribution p over $\mathscr{X} \times \Omega$, which factorizes as which factorizes as $p(x, \omega) = p(\omega)\, p(x \mid \omega)$, where $p(\omega) \in \Delta(\Omega)$ represents the population-level distribution of user types. Upon observing $X = x$, the platform forms a posterior belief over types:

$$q_x(\omega) := p(\omega \mid x) \in \Delta(\Omega).$$

  For notational clarity, we also define the *random posterior* $q_X := p(\cdot \mid X)$, a random variable that maps each realization $X = x$ to its corresponding posterior q_x. This posterior encodes the platform's updated belief about the user's type after observing x.

- **Policies.** A policy specifies how the platform selects actions based on the observed information set. Formally, a policy is a measurable function:

$$\pi : \mathscr{X} \to \mathscr{A},$$

  which maps each realization $x \in \mathscr{X}$ to an action $\pi(x) \in \mathscr{A}$. We focus on deterministic policies because, under linear expected utility objectives, randomization offers no additional value (by an extreme-point argument; see Kamenica, 2019; Smith et al., 2023).[2]

- **Expected utility and optimal actions.** Given a belief $q \in \Delta(\Omega)$ over user types, such as the posterior q_x induced by observing x, the expected utility of choosing action $a \in \mathscr{A}$ is

$$U(a \mid x) = \mathbb{E}_{\omega \sim q_x}[u(a, \omega)].$$

  Following, among all possible policies $\pi \in \Pi$ that map observed information to actions, the optimal policy $\pi^*(x)$ is the one that maximizes expected utility for each observation:

$$\pi^*(x) \in \arg\max_{a \in \mathscr{A}} \mathbb{E}_{\omega \sim q_x}[u(a, \omega)]. \tag{1}$$

[2] When $\mathscr{A}$ is finite and the objective is linear, any mixed policy can be written as a convex combination of deterministic ones, and the maximum is achieved at an extreme point. Ties are resolved by an arbitrary but fixed rule.

- **Decision value.** Let X denote the information set observed prior to action selection, and let $q_X = p(\cdot \mid X)$ be the induced posterior over user types. For any deterministic policy $\pi : \mathscr{X} \to \mathscr{A}$, its ex-ante value is

$$V(\pi) = \mathbb{E}_{(X,\omega)\sim p}[u(\pi(X),\omega)] = \mathbb{E}_X\left[\mathbb{E}_{\omega\sim q_X}[u(\pi(X),\omega)]\right].$$

  The value of information from X is the maximum achievable value over all such policies:

$$V_X = \sup_{\pi:\mathscr{X}\to\mathscr{A}} V(\pi) = \sup_{\pi:\mathscr{X}\to\mathscr{A}} \mathbb{E}_X\left[\mathbb{E}_{\omega\sim q_X}[u(\pi(X),\omega)]\right]. \tag{2}$$

  The optimal policy $\pi^*(x)$ defined in Equation 1 attains this supremum by maximizing expected utility for each observation.

Throughout, X denotes the available information set: baseline contextual data $X^{\varnothing}$, behavioral information X^B added to the baseline, geographical information X^G added to the baseline, and their combination X^{GB}.

4.1 Challenges

The central objective of this framework is to assess the value of information by quantifying and comparing how behavioral data (X^B) and geographical data (X^G) contribute to improved decision-making. The value function in Equation 2 provides the formal basis for this assessment by defining the maximum expected utility attainable under each information set. Implementing this framework, however, raises several challenges:

4.1.1 Challenge 1: Information Value, Complementarity, and Substitutability

The first challenge is to formalize how to compare the value of two distinct sources of information. When multiple data sources, such as behavioral (X^B) and geographical (X^G), are available, their value is inherently *relative*: each may provide incremental benefit beyond the other, or they may overlap in what they convey. The marginal value of one source depends on whether the other is already observed, raising the possibility of complementarity or substitutability. Without precise definitions, it is impossible to separate the individual contribution of each source from its joint interaction. A formal framework is therefore essential to isolate the standalone value of each source, its incremental value conditional on the other, and their combined effect. We address this challenge in §5.1.

4.1.2 Challenge 2: Utility Function Estimation for Decision Policies

The second challenge is that the utility function $u(a,\omega)$, which captures how effective each action is for each user type, is not observed and must be estimated from data. In practice, we only observe realized outcomes under specific actions, without direct access to the underlying payoff structure. This step is critical because accurate utility estimation is the foundation for policy evaluation: without it, the counterfactual performance of alternative targeting rules cannot be assessed. The task is further complicated by the high dimensionality of the information space: behavioral data form sequences of user histories, and geographical data span a wide spatial domain with correlated features, making it difficult to obtain stable utility estimates across actions and user types. We address this challenge in §5.2.

4.1.3 Challenge 3: Information Value Estimation and Policy Evaluation

The third challenge arises in estimating the value of information V_X once the utility function has been recovered. By definition, V_X represents the maximum expected utility attainable under a given information set, which requires evaluating the performance of alternative policies. This evaluation must consider not only the actions observed in the data but also counterfactual choices that were never taken. The difficulty is fundamental: each user reveals an outcome under only one action, while the value function aggregates over all possible actions. As a result, estimating V_X requires inferring outcomes for unobserved actions, which is especially challenging when some actions are rarely or never chosen. Our strategy for addressing this challenge is presented in §5.3.

5 Empirical Strategy

To address the challenges outlined in §4.1, we develop a unified framework that integrates economic theory, machine learning, and causal inference. The framework consists of three components: (i) a formal definition for comparing the value of distinct information sets (§5.1), (ii) a method for estimating the utility function from high-dimensional observational data (§5.2), and (iii) a solution for evaluating the value function in the presence of unobserved counterfactuals (§5.3). Taken together, these components yield a consistent framework for quantifying the informational contribution of behavioral and geographical data.

5.1 Substitutability and Complementarity of Information Sets

To address Challenge 4.1.1, we use formal definitions from the economic theory literature to compare the value of different information sets in a decision problem. In our context, we assess how behavioral data X^B and geographical data X^G contribute to guiding optimal actions. Each information set may improve decision quality on its own, but its joint value depends on whether they provide overlapping or distinct insights about the user's latent type ω.

We adopt the decision-theoretic framework of substitutability and complementarity introduced by Börgers et al. (2013) [3]. Recall that X^B and X^G denote behavioral and geographical data, respectively. We define the combined information set as $X^{GB} := (X^B, X^G) \in \mathscr{X}^B \times \mathscr{X}^G$, which represents joint access to both sources. As a baseline, we consider $X^\varnothing$, corresponding to the absence of user-level information. In this case, decisions rely only on *contextual information* that informs the prior belief about user types, so the posterior reduces to the prior, $q_{X^\varnothing} \equiv p(\omega)$. The associated value, $V_{X^\varnothing}$, therefore represents decision-making under prior uncertainty without conditioning on individual-level signals. Substitutability and complementarity are defined by the marginal value of one information set conditional on access to the other, using the value function V_X introduced in Equation 2.

Definition 1 (Substitutes)**.** *Behavioral information X^B is a* substitute *for geographical information X^G* if,

[3] Börgers et al. (2013) use the term *signal* to denote an information source. In this paper, we use the terms *data* and *information set* instead to emphasize that we work with observed variables available to the platform rather than abstract signals.

for every decision problem $(\mathscr{A}, u)$[4] ,

$$V_{X^G} - V_{X^\varnothing} \;\geq\; V_{X^{GB}} - V_{X^B}.$$

This inequality compares the marginal value of geographical data when used alone versus when combined with behavioral data. The left-hand side, $V_{X^G} - V_{X^\varnothing}$, reflects the value of geographical data on its own, while the right-hand side, $V_{X^{GB}} - V_{X^B}$, reflects its incremental contribution when behavioral data are already available. If the inequality holds, behavioral data reduces the marginal usefulness of geographical data, making the two substitutes.

Definition 2 (Complements)**.** *Behavioral information X^B is a* complement *to geographical information X^G* if, for every decision problem $(\mathscr{A}, u)$,

$$V_{X^G} - V_{X^\varnothing} \;\leq\; V_{X^{GB}} - V_{X^B}.$$

Here, the marginal value of geographical data is greater when combined with behavioral data than when used alone. The left-hand side, $V_{X^{GB}} - V_{X^B}$, measures the incremental contribution of geographical data given that behavioral data are already available, while the right-hand side, $V_{X^G} - V_{X^\varnothing}$, captures their standalone value. If the inequality holds, behavioral and geographical data enhance each other's usefulness and are thus considered complements.

5.2 Machine Learning Framework for Utility Function Estimation

To address Challenge 4.1.2, we develop a machine learning framework for estimating expected utility when the utility function $u(a, \omega)$ is unobserved. The advertiser cannot observe latent types ω, nor directly measure the utility associated with each action. Instead, we re-express utility in terms of observable data: user characteristics X, chosen actions A, and realized engagement outcomes $Y \in \{0, 1\}$.

We specify utility through the binary click outcome Y and the reward value of engagement. Suppose that showing ad $a \in \mathscr{A}$ to a user of type $\omega \in \Omega$ yields utility only when the user clicks, and that the advertiser receives a known per-click reward $v(a) \in \mathbb{R}_{\geq 0}$. In this case, the expected utility from taking action a for a user of type ω is given by

$$u(a, \omega) := \mathbb{E}[Y \mid A = a, \omega] \cdot v(a) = \Pr(Y = 1 \mid A = a, \omega) \cdot v(a).$$

This formulation captures both the probabilistic nature of engagement and the economic payoff per click, and it serves as the foundational definition of utility in our analysis. Although user types ω are unobserved, conditioning on observed characteristics $X = x$ induces a posterior belief $q_x(\omega) := p(\omega \mid x)$ about the user's latent type. The expected utility of action a given these characteristics is:

$$U(a \mid X = x) = \mathbb{E}_{\omega \sim q_x}[u(a, \omega)] = v(a) \cdot \mathbb{E}_{\omega \sim q_x}[\Pr(Y = 1 \mid A = a, \omega)].$$

[4]Following Börgers et al. (2013), "every" refers to the inequalities holding for *all* decision problems (A, u). Our empirical analysis evaluates these same inequalities for a fixed problem (A, u) (CTR objective and action set), yielding complement/substitute conclusions *for this problem*.

By the law of iterated expectations, this simplifies to:

$$U(a \mid X = x) = v(a) \cdot \Pr(Y = 1 \mid X = x, A = a),$$

since averaging $\Pr(Y = 1 \mid A = a, \omega)$ over the posterior q_x yields the click probability conditional on $X = x$. This equivalence is crucial: it shows that the expected utility of an action, conditional on observed characteristics, can be recovered by estimating the conditional probability of a click. Accordingly, we focus on estimating the function:

$$f_X(a,x) := \Pr(Y = 1 \mid X = x, A = a),$$

which maps each action–characteristics pair to the probability of engagement. This function is the central object in our machine learning framework. Once f_X is estimated, expected utility can be computed as:

$$\hat{U}(a \mid x) = v(a) \cdot \hat{f}_X(a,x),$$

and the advertiser can select actions by maximizing this estimated utility. This approach can be operationalized using flexible prediction models trained on observational data, provided that the model class is sufficiently expressive and the estimation procedure is properly regularized and calibrated.

We model f_X by specifying both the feature set X and the functional form $f(\cdot)$. §5.2.1 classifies X into the four informational regimes $\{X^{\varnothing}, X^{G}, X^{B}, X^{GB}\}$. Conditional on a choice of X, §5.2.2 describes how we select a learning algorithm f from a flexible class $\mathscr{F}$ that can approximate the conditional probability $\Pr(Y = 1 \mid X, A)$. We estimate these models on observational data with regularization and calibration, as detailed in §5.2.3, to ensure reliable prediction. The estimated function $\hat{f}_X$ delivers action-specific utilities and induces policies $\pi(x)$. These policies provide the foundation for our policy evaluation across informational regimes.

5.2.1 Structure of Information Sets X

For our machine learning model, each targeting regime $X^{(\cdot)}$ corresponds to a specific set of observable features. We classify these features into three dimensions, *contextual*, *behavioral*, and *geographical*, which form the building blocks of the different regimes:

Contextual Features (X^{C}): Contextual features characterize the *setting* in which each ad impression occurs, independent of user identity or location. These include time of day, day of week, app, device, and network provider. Engagement often follows systematic patterns along these dimensions; for instance, click rates in gaming apps may peak in the evening when many users log in.

Behavioral Features (X^{Behave}): Behavioral features capture a user's *online footprint*, such as prior ad exposures, cumulative clicks, time since last click, and category-specific CTRs. These variables reveal individual preferences and engagement; for example, a user who frequently clicks on sports ads is much more likely to respond to a new sports campaign.

Geographic Features (X^{Geo}): Geographical features reveal the user's *physical footprint*, their location at the time of the impression, which may be represented at various levels of granularity, such as city or precise latitude and longitude. They reflect spatial variation in preferences and engagement, for example,

ride-sharing ads see higher click rates in dense urban centers during rush hour compared to rural areas.

While we have conceptually defined the targeting regimes, here we map them to their feature representations. Each $X^{(\cdot)}$ corresponds to a specific subset of contextual, behavioral, and geographic features used in estimation:

- **($X^{\varnothing} = X^{C}$):** The advertiser observes only contextual features (e.g., device, app category, time, or network conditions). No user-level information, such as behavioral or geographic data, is available, so actions are chosen based on context-dependent averages rather than latent user types ω. Such information, typically available from campaign reporting or platform analytics.

- **($X^{B} = X^{\textbf{Behave}} \cup X^{C}$):** The advertiser observes both the context of each impression and features that summarize the user's prior behavior. Together, these variables provide insight into the user's likely preferences and engagement, revealing information about the latent type ω through observed patterns.

- **($X^{G} = X^{\textbf{Geo}} \cup X^{C}$):** The advertiser observes both contextual information and the user's geographic location. Geographic features can reveal spatial heterogeneity in preferences, such as differences due to local events, regional trends, or time zones, and thus offer indirect information about a user's type ω.

- **($X^{GB} = X^{\textbf{Behave}} \cup X^{\textbf{Geo}} \cup X^{C}$):** The advertiser has access to the full set of behavioral, geographic, and contextual features. This regime enables the model to capture the most complete picture of user heterogeneity to infer latent user types ω and guide targeting.

These categories mirror the types of data advertisers realistically access in practice. The no-information case serves as a benchmark for non-personalized targeting, while the other regimes reveal distinct dimensions of latent user preferences. Details on feature construction, variable definitions, and preprocessing appear in the Web Appendix (§A).

5.2.2 Learning Algorithm Selection Across Information Regimes

We estimate the conditional click function $f_X(a,x)$ using observational data $\mathscr{D} = \{(X_i, A_i, Y_i)\}_{i=1}^{n}$, where each record contains observed features X_i, the chosen action $A_i \in \mathscr{A}$, and the engagement outcome $Y_i \in \{0,1\}$. For each information regime $X \in \{X^{\varnothing}, X^{G}, X^{B}, X^{GB}\}$, we fit a model $f_X \in \mathscr{F}$. The function class $\mathscr{F}$ is chosen to match the complexity and structure of the corresponding feature set.

For the contextual-only ($X^{\varnothing}$) and geography-only (X^{G}) regimes, where inputs are cross-sectional, we use XGBoost, a gradient-boosted tree algorithm widely adopted in advertising and personalization research (Rafieian and Yoganarasimhan, 2021). XGBoost performs well on heterogeneous, tabular data with high-cardinality categorical variables such as device type, app category, ISP, time of day, and geographic indicators. Hence, the use of XGBoost follows prior empirical evidence and reflects a deliberate model selection consistent with the static nature of the data.

In contrast, the behavioral (X^{B}) and combined (X^{GB}) regimes involve sequential user histories that require capturing temporal dependencies. To model these dynamics, we design a Long Short-Term Memory

(LSTM) network with an attention mechanism. The LSTM component captures the evolution of user engagement over time, while attention identifies the most relevant parts of each user's exposure–click history. This architecture has been shown to perform effectively in dynamic, history-dependent engagement prediction tasks (Quadrana et al., 2018; Yin et al., 2025). In the combined regime X^{GB}, static geographic features are concatenated with each behavioral step to allow spatial–temporal interactions. The same network is used for both X^B and X^{GB} to maintain comparability across regimes and ensure that $\hat{f}_X(a,x)$ remains action-conditional.

To preserve the temporal structure of behavior, we split the data at the user level, as described in §3.4. Let j index a user and t the t-th impression for that user. At each step t, the model predicts the probability of a click y for an observed action a using only information revealed in the sequence $\{1,\dots,t\}$. This forward-looking setup ensures that predictions rely solely on data available to the advertiser at the time of decision, eliminating information leakage.

As t grows, the LSTM model accumulates a richer behavioral history for user j, allowing a better estimate of preferences and latent types. We therefore hypothesize that estimation quality improves with t in any specification that incorporates behavioral data, whether behavioral-only (X^B) or combined (X^{GB}):

Hypothesis 1 (Behavioral Accumulation and Convergence of $\hat{f}$)**.** *Let $X^B_{jt} = \{X^B_{j1}, X^B_{j2}, \dots, X^B_{jt}\}$ denote the behavioral history of user j up to their t-th impression. Suppose there exists a true utility-generating function $f : \mathscr{X} \times \mathscr{A} \to [0,1]$, representing the probability of user engagement given state and action. Then, for any estimator $\hat{f}$ based on an information set that includes behavioral information (e.g., X^B or X^{GB}), we expect:*

$$\hat{f}(X^B_{jt}, a) \xrightarrow{p} f(X^B_{jt}, a) \quad \text{as } t \to \infty \quad \text{for all } a \in \mathscr{A},$$

and the joint mapping from behavioral histories and actions to expected outcomes is consistently learned. That is, as the behavioral sequence grows, the model increasingly recovers the structure of the true action-outcome relationship, enabling improved inference and decision quality.

This hypothesis outlines the idea that longer behavioral sequences yield richer information about user preferences. In §6.3, we provide empirical evidence consistent with this hypothesis.

5.2.3 LSTM-Based Modeling of User Behavior

This section describes the sequence model used for regimes that include behavioral histories. We design an LSTM architecture equipped with a causal multi-head attention mechanism. The LSTM captures short- and long-range temporal dependencies in user interactions, while attention highlights the most informative parts of a user's exposure–click history. Together, these components allow the model to represent both persistent behavioral trends and short-term recency effects, which are central to engagement modeling (Grbovic and Cheng, 2018; Quadrana et al., 2018).

We train the model on user interaction histories using a sliding window of 150 impressions. At each time step, the input combines categorical and numerical features. We map high-cardinality categorical variables (e.g., device model, app ID, network identifiers) into dense vectors with embedding layers (Guo and Berkhahn, 2016) and concatenate them with continuous covariates. To encode order, we add absolute

positional embeddings (Vaswani et al., 2017). We also process the log-transformed inter-arrival time through a time-gap projection, which helps the model detect irregular timing patterns that often indicate behavioral shifts.

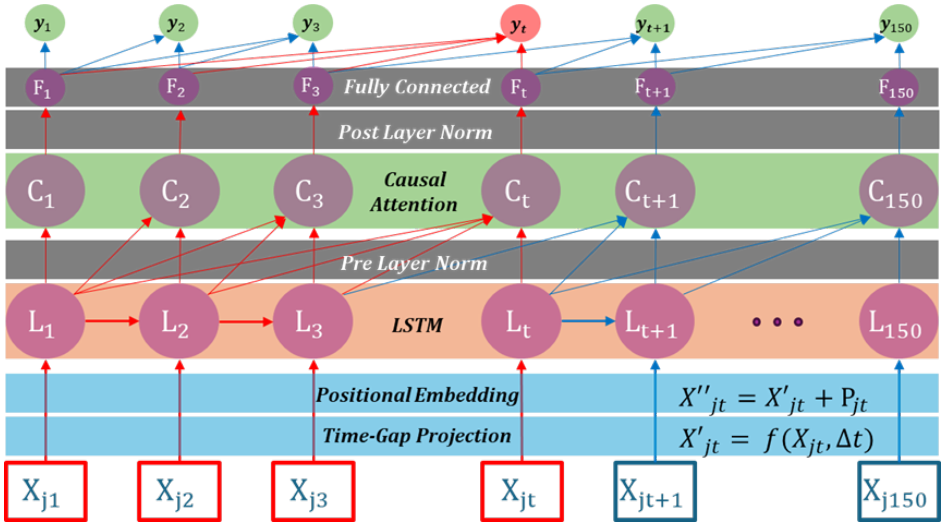


Figure 5: LSTM with causal attention architecture

We feed the enriched sequence into a four-layer LSTM with hidden size 512. On top of the LSTM outputs, we leverage causal multi-head self-attention with four heads. The attention module uses a future-masked matrix to block access to unseen impressions, which preserves temporal causality and prevents information leakage (Vaswani et al., 2017). We apply layer normalization before and after the attention block to stabilize training. Next, we add a gated projection head that runs parallel sigmoid and tanh transformations, combines them elementwise, and applies dropout for regularization. A fully connected output layer finally maps the hidden states to click probabilities at each step.

Figure 5 illustrates the designed sequence model architecture and its key components. The design directly leverages sequential order, irregular timing, and varying relevance of past impressions. As users generate longer histories, the LSTM–attention framework provides richer information about preferences and latent types, which supports the Behavioral Accumulation Hypothesis 1. We provide additional implementation details in Web Appendix §A.4.

5.3 Estimating the Value of Information via Inverse Propensity Scoring

Challenge 4.1.3 highlights the core difficulty in estimating the value of information V_X (Equation 2):

$$V_X = \sup_{\pi:\mathscr{X}\to\mathscr{A}} \mathbb{E}_X\left[\mathbb{E}_{\omega\sim q_X}[u(\pi(X),\omega)]\right],$$

The data only reveal outcomes under the action chosen by the platform's logging policy, while V_X requires utilities aggregated over all possible actions. Each impression, therefore, provides information on one action, but evaluating a policy involves counterfactual outcomes for actions that were not taken.

We address this problem with Inverse Propensity Scoring (IPS), a standard method in causal inference

for adjusting observational data (Hirano et al., 2003; Rafieian and Yoganarasimhan, 2023). IPS identifies the value of a fixed target policy by reweighting observed outcomes according to how likely the logging policy was to select the same action as the target policy. By doing so, we approximate the utility that would have been realized under alternative policies, even though those policies were never deployed in practice.

Formally, for each information set X, we define a target policy π^{f_X} that deterministically selects the action maximizing the estimated expected utility based on our predictive model:

$$\pi^{f_X}(a \mid x) \;=\; \mathbb{I}\Big\{ a \;=\; \arg\max_{a' \in \mathscr{A}} \hat{U}(a' \mid x) \Big\}, \qquad \hat{U}(a \mid x) \;=\; v(a)\, \hat{f}_X(a,x),$$

where $\hat{f}_X(a,x)$ is the estimated click probability given information set x and candidate action a. Thus, π^{f_X} prescribes, for each instance, the action expected to maximize utility under the model. The value of this policy, consistent with our decision-theoretic objective, is

$$V(\pi^{f_X}) \;=\; \mathbb{E}_X\big[\mathbb{E}_{\omega \sim q_X}\big[u(\pi^{f_X}(X), \omega)\big]\big]\,.$$

However, only outcomes corresponding to the logging policy's actions are observed in the data. To estimate the value of the induced policy $V(\pi^{f_X})$, we use the IPS estimator, which reweights the observed utility outcomes for each user-impression pair by the inverse of the probability that the logging policy would have selected the action recommended by the target policy:

$$\hat{V}(\pi^{f_X}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{I}\big[A_i = \pi^{f_X}(X_i)\big]}{\pi^{\mathscr{D}}(A_i \mid X_i)}\, v(A_i)\, Y_i,$$

where A_i is the action taken in the historical data for impression i, $\pi^{\mathscr{D}}(A_i \mid X_i)$ is the probability that the logging policy selected A_i given information set X_i, and Y_i is the observed binary engagement outcome. Each term thus adjusts the observed utility for representativeness under the target policy. For the IPS estimator to recover the true value V_X, three key conditions must hold regarding the data-generating process and the structure of the logging policy. We state these assumptions formally below.

Assumption 1 (Overlap)**.** *For all $x \in \mathscr{X}$ and $a \in \mathscr{A}$, if $\pi^{f_X}(a \mid x) > 0$, then $\pi^{\mathscr{D}}(a \mid x) > 0$. In other words, the logging policy must assign positive probability to every action that the target policy might select.*

Assumption 2 (Unconfoundedness)**.** *Potential outcomes are independent of the action actually taken, conditional on observed covariates X; that is, $Y(a) \perp A \mid X$ for all $a \in \mathscr{A}$. This ensures that all confounding factors are captured in X.*

Assumption 3 (Policy optimality under X)**.** *If $\hat{f}_X$ consistently estimates $\Pr(Y = 1 \mid A = a, X = x)$ and the maximizer of $v(a)\Pr(Y = 1 \mid A = a, X = x)$ is unique a.e., then $\pi^{f_X}(x) = \arg\max_a v(a)\hat{f}_X(a,x)$ is optimal under X a.e.*

Proposition 1 (Standard IPS identification and recovery of V_X)**.** *Under Assumptions 1–3, and when π^{f_X} is evaluated on an independent sample,*

$$\mathbb{E}\big[\hat{V}(\pi^{f_X})\big] \;=\; V_X, \qquad \hat{V}(\pi^{f_X}) \xrightarrow{p} V_X.$$

That is, the IPS estimator is unbiased and consistent for V_X under the stated conditions. A formal proof is provided in Web Appendix §B.1.

The IPS approach addresses Challenge 4.1.3 by enabling estimation of the ex-ante value of any information set X through the policy it induces, even when counterfactual outcomes are unobserved. We apply this method consistently across models using geographical, behavioral, or combined information, which allows us to compare their value in terms of the utility gains they generate in policy evaluation.

5.3.1 Estimating and Validating Propensity Scores

To implement the IPS estimator, we estimate the logging policy probabilities $\pi^{\mathscr{D}}(a \mid x)$, defined as the probability that the historical platform selected action a for a given information set x. First, we describe the ad allocation mechanism that generates these probabilities, then examine empirical evidence supporting the IPS identification assumptions, and finally outline our procedure for estimating and validating propensity scores used in the policy evaluation.

Quasi-Proportional Allocation Mechanism. In our setting, the mobile in-app advertising network allocates impressions using a *quasi-proportional auction* (Mirrokni et al., 2010). For each impression i, let $\mathscr{A}_i \subseteq \mathscr{A}$ denote the set of ads eligible to participate in the auction. Each ad $a \in \mathscr{A}_i$ submits a bid b_a and has a platform-assigned quality score q_a, both of which are fixed and observed during our sample period. The probability that ad a wins the auction and is shown in impression i, conditional on covariates x_i, is given by:

$$\pi^{\mathscr{D}}(a \mid x_i) = \frac{b_a q_a}{\sum_{j \in \mathscr{A}_i} b_j q_j}.$$

This allocation rule induces a probabilistic assignment over eligible ads, with probabilities that are fully determined by observed features. In contrast to deterministic formats (e.g., second-price auctions) where only the top-ranked ad is observed, the quasi-proportional mechanism introduces randomization across all eligible ads. This variation is central to our identification strategy, as it enables estimation of counterfactual outcomes and supports off-policy evaluation.

Remark 1 (Overlap)**.** *Any ad participating in the auction for impression i (i.e., $\forall a \in \mathscr{A}_i$) has a nonzero propensity of being shown in impression i.*

This follows directly from the quasi-proportional rule: every ad with a positive $b_a q_a$ has a strictly positive probability of being selected. Hence, the overlap assumption 1 is satisfied by construction for all participating ads in each auction.

Remark 2 (Unconfoundedness)**.** *For any impression i, ad allocation is independent of the set of potential outcomes for participating ads ($a \in \mathscr{A}_i$), after controlling for the observed covariates. Thus,*

$$\{Y_i(a)\}_{a \in \mathscr{A}_i} \perp A_i \mid x_i.$$

Therefore, the unconfoundedness assumption 2 is satisfied by the transparent structure of the auction: all inputs that determine allocation, namely, bids b_a, quality scores q_a, and eligibility constraints, are fully

observed and fixed over the sample period. For each impression i, we observe the complete covariate vector x_i. Advertiser bids are not dynamically adjusted, and the platform does not personalize quality scores across users. As a result, the assignment rule is fully determined conditional on x_i.

This structure implies that for any impression i, we can compute not only the probability $\pi^{\mathscr{D}}(A_i \mid x_i)$ of the ad that was actually shown, but also the probabilities of all counterfactual ads that were eligible in the same auction. That is, even if a particular ad a was not displayed in impression i, as long as it was eligible (i.e., $a \in \mathscr{A}_i$), we can recover the counterfactual allocation probability $\pi^{\mathscr{D}}(a \mid x_i)$.

Eligibility Filtering. Although the quasi-proportional rule defines probabilities $\pi^{\mathscr{D}}(a \mid x_i)$ for all ads in the auction, not all ads in the global set $\mathscr{A}$ are eligible in every impression. To ensure valid counterfactual estimation, we restrict attention to ads with nonzero probability of participation in each impression. Two factors determine eligibility: (i) *Contextual Targeting*: ads may be restricted to specific provinces, times, or app categories, and are excluded when their targeting criteria are not met; (ii) *Campaign availability*: some ads may be inactive due to budget exhaustion or campaign timing. In practice, this is rare among top ads, as we select the top 10 ads for our analysis.

We construct an eligibility matrix $E \in \{0,1\}^{N \times A}$, where $e_{i,a} = 1$ indicates that ad a was eligible to compete in impression i, based on observed targeting and availability constraints. In practice, this requires that the impression's metadata match the ad's targeting filters on province, hour-of-day, and app. To avoid misclassification, we drop any impressions with missing targeting variables and restrict attention to a filtered sample where eligibility can be verified.

While this filtering step identifies the support of the logging policy, it does not tell us how likely each eligible ad is to be selected. For unbiased off-policy evaluation, we must account for the non-random assignment probabilities across eligible ads. Next, we describe how we estimate these propensities and assess the validity of the unconfoundedness assumption via covariate balance.

Propensity Score Estimation and Covariate Balance To correct for unequal selection probabilities inherent in the auction, we estimate the *propensity scores* $\pi^{\mathscr{D}}(a \mid x_i)$, defined as the probability that the logging policy assigns impression i to ad a given the observed features x_i. These propensities quantify the exposure pattern generated by the platform's allocation mechanism and form the basis for IPS in our policy evaluation.

Although the quasi-proportional rule maps bids and quality scores into theoretical probabilities, we adopt a data-driven estimation strategy to capture the realized assignment process in practice. This approach flexibly accommodates deviations from the theoretical rule, nonlinear effects, and high-order interactions among features. We use XGBoost to estimate $\pi^{\mathscr{D}}(a \mid x_i)$, motivated by its empirical performance in high-dimensional classification problems (Rafieian and Yoganarasimhan, 2023). To avoid overfitting and ensure out-of-sample validity, we implement a 5-fold cross-fitting procedure, so that each estimated propensity score is computed on a model trained without the corresponding observation.

While observed inputs fully determine the allocation rule, we validate our identifying assumption by testing whether inverse-propensity weighting balances the distribution of observed covariates across treatment. The variables examined include province, app context, time of day, device brand, network type, and mobile service provider, features that advertisers can directly target. For each covariate and treatment group, we compute the standardized mean difference (SMD) before and after applying the estimated propensity

weights, considering absolute values below 0.2 as indicative of acceptable balance (McCaffrey et al., 2013). The diagnostics show substantial improvements in covariate alignment across ads after weighting, providing empirical support for our research design. Details of propensity score estimation and balance statistics are reported in Web Appendix §B.2.1.

6 Empirical Results

We now present the empirical results derived from our proposed framework. §6.1 reports the predictive performance of machine learning models trained on different information sets. We then turn to the core question of how behavioral and geographical information contribute to decision quality, examining it at two levels of analysis. First, at the *aggregate level* (§6.2), we quantify the overall value of each information set and test whether the two act as substitutes or complements in improving targeting performance. Second, we extend the analysis to the *user level* (§6.3), examining how the value and interaction of these information types change with the amount of behavioral history observed by the platform.

6.1 Predictive Performance Across Information Sets

Predicting user engagement is inherently challenging due to the extreme class imbalance; only about 2% of impressions result in a click, while the remaining 98% do not. In such cases, accuracy is an unreliable metric, as even a trivial classifier that always predicts "no click" would achieve high accuracy while offering no actionable insight. We therefore evaluate performance using metrics that better capture probabilistic and ranking quality:

- **Log Loss:** For each impression with observed outcome $y \in \{0,1\}$, the model predicts $\hat{f}_X(a,x)$, and the log loss is $\ell(\hat{f}_X(a,x),y) = -\left[y\log\left(\hat{f}_X(a,x)\right) + (1-y)\log\left(1-\hat{f}_X(a,x)\right)\right]$. Averaging over impressions gives the total loss $\mathscr{L}_{\text{model}}$. Lower values indicate better-calibrated probabilities and align with our model's training objective.

- **Relative Information Gain (RIG):** RIG is defined as $\left(1-\frac{\mathscr{L}_{\text{model}}}{\mathscr{L}_{\text{baseline}}}\right)\times 100$, where $\mathscr{L}_{\text{baseline}}$ is the log loss of a model that predicts the average click-through rate. Higher RIG reflects greater predictive value relative to this naive benchmark.

- **Area Under the ROC Curve (AUC):** AUC measures how well the model ranks clicked above non-clicked impressions. It is threshold-independent and robust to class imbalance, making it well-suited for rare events like clicks.

We evaluate four models trained under different informational regimes: (i) XGBoost for the contextual-only baseline ($X^{\varnothing}$) and the geography-only regime (X^{G}), (ii) and LSTM for the behavioral regime (X^{B}) and the full information set (X^{GB}). The performance of each model on the held-out test set is summarized in Table 2.

The results reveal substantial variation in predictive performance across information structures. The contextual-only baseline performs weakest, with an RIG of 18.90%. Adding geographic features raises performance only slightly, with an RIG of only 19.22% and an AUC of 0.722. In contrast, the behavioral model delivers a substantial improvement, achieving an RIG of 84.18% and an AUC of 0.809, indicating

Table 2: Predictive performance of models using contextual data ($X^{\varnothing}$), geographical data (X^{G}), behavioral data (X^{B}), and both (X^{GB}).

Model	Log Loss	AUC	Relative Information Gain (%)
($X^{\varnothing}$)	0.072	0.720	18.900
(X^{G})	0.072	0.722	19.220
(X^{B})	0.014	0.809	84.180
(X^{GB})	0.014	0.812	84.410

Notes: The table reports Log Loss, AUC, and RIG on the test set. RIG is computed relative to the baseline click-through rate $p = 0.017592$. Sample size is $N = 3{,}162{,}376$ impressions.

both richer information and superior ability to rank click outcomes. These gains underscore the richness of behavioral data for predicting user engagement. Adding geographic information to the behavioral model yields only marginal additional improvements, suggesting that behavioral features already account for most of the predictive variation.

Taken together, these results demonstrate the dominant predictive value of behavioral information in estimating click likelihood and are closely aligned with findings in the literature that examine similar prediction tasks in advertising settings. Our results show that while the AUC of the behavioral model remains comparable when using XGBoost (See Rafieian and Yoganarasimhan (2021)) instead of LSTM, the RIG improves substantially under LSTM, indicating better probabilistic calibration. These gains suggest that incorporating temporal structure into behavioral modeling meaningfully enhances predictive quality relative to tree-based methods. These levels of performance provide empirical support for Assumption 3. We provide further details and a direct comparison between LSTM and XGBoost models in Web Appendix §C.1, highlighting the predictive gains from temporal modeling.

That said, it is important to emphasize again that predictive performance does not always translate into decision value. A model may be well-calibrated or rank instances correctly, yet offer limited benefit when used to guide actions under realistic constraints. We therefore turn next to assessing the actual targeting value generated by each model in terms of the value function.

6.2 Behavioral vs. Geographical Information: Aggregate Level

Having established model performance in §6.1, we now assess how behavioral and geographical information improve decision quality. As described in §5.3, we evaluate each targeting policy $\pi^{(f_X)}$ using IPS. We proceed in two steps. First, in §6.2.1, we quantify the *aggregate value* of each information set by comparing the expected utility of policies based on different data sources. Then, in §6.2.2, we test whether behavioral and geographical information act as *substitutes or complements* in shaping decision quality.

6.2.1 Aggregate Value of Behavioral and Geographic Information

We consider a set of *deterministic greedy policies* indexed by the information set $X \in \{X^{\varnothing}, X^{G}, X^{B}, X^{GB}\}$ available at the time of decision. For each impression i, let $\mathscr{A}_i \subseteq \mathscr{A}$ denote the set of eligible actions under the platform's allocation rule (see §5.3.1). Each policy selects the ad with the highest predicted

click probability among supported actions, where support is defined by $\hat{\pi}^{\varnothing}(a \mid X_i) > 0$. This restriction ensures that the selected action lies within the empirical support of the logging policy, satisfying the overlap condition in Assumption 1.

Each policy represents a distinct counterfactual scenario in which the advertiser optimizes targeting using only the corresponding information set X, and we estimate its value $\hat{V}_X$ using IPS. While IPS is widely used for offline policy evaluation, it is also known to suffer from high variance, especially when logging propensities are small or highly uneven. This variance can undermine estimator stability and complicate inference, particularly in high-dimensional or sparsely supported action spaces (Swaminathan and Joachims, 2015).

To assess the reliability of our IPS estimates and quantify estimation uncertainty, we report two diagnostics. First, we construct 95% confidence intervals based on *cluster-robust* standard errors (clustered at the user level) calculated from the IPS contributions aggregated within clusters. These intervals allow us to quantify sampling variability and evaluate whether each policy yields statistically significant gains relative to the observed click-through rate under the historical logging policy. Second, we report the *effective sample size* (ESS), a diagnostic commonly used in causal inference and off-policy evaluation to measure the degree of variance inflation caused by skewed importance weights (Kallus, 2019). ESS reflects how many equally weighted samples would provide the same estimator precision as the current reweighted sample. When ESS is low, the IPS estimate is effectively driven by a small subset of data points with large weights, raising concerns about stability.

The results are reported in Table 3. Across all policies, we find statistically significant improvements over the historical logging baseline, with tight confidence intervals and high ESS. In particular, McCaffrey et al. (2013) advocates trimming weights until ESS $\geq 0.10N$, a threshold all policies in our setting easily exceed. This suggests that the estimated policy values are stable and inference is well-powered.

Table 3: Estimated Policy Value Using IPS

Policy	IPS Estimate	95% CI	t-stat	SE	Lift (%)	ESS
$\hat{V}_{X^{GB}}$	0.025^{***}	[0.024, 0.025]	27.611	0.000	41.450	760316
$\hat{V}_{X^{B}}$	0.023^{***}	[0.022, 0.023]	19.675	0.000	30.280	555871
$\hat{V}_{X^{G}}$	0.023^{***}	[0.022, 0.023]	16.572	0.000	28.710	592427
$\hat{V}_{X^{\varnothing}}$	0.021^{***}	[0.021, 0.022]	18.782	0.000	20.440	542747

Notes: All estimates are computed using IPS. The baseline CTR under the logging policy is 0.017592. Lift is computed relative to this baseline. SE denotes the standard error of the estimate. Effective sample size (ESS) measures the number of equally weighted observations that would yield equivalent precision. Number of observations: 3,162,376. Cluster-robust standard errors are computed using 141,595 user clusters. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Starting from the contextual information policy $X^{\varnothing}$, we observe a 20.4% improvement over the baseline CTR. Although the advertiser lacks access to user-level data, they can still optimize by leveraging contextual variation and selecting the ad with the highest average CTR. Importantly, because the original data were generated under a quasi-proportional allocation rule without targeting, this policy functions as a quasi-random benchmark. That it nonetheless yields substantial gains highlights the value of leveraging even aggregate performance data in non-personalized settings.

Building on this benchmark, access to richer information sets produces clear incremental improvements in policy value. Geographical data (X^G) raises performance by an additional 8.3% over the contextual policy, suggesting that user location captures latent demand patterns or regional preferences relevant to ad effectiveness. Behavioral data (X^B) yield a marginally larger improvement, approximately 9.0% over the contextual benchmark, consistent with the idea that past engagement behavior provides a more direct signal of ad relevance.

Notably, unlike the predictive results in §6.1, where behavioral information explained most of the performance gain, here, geographical data contribute comparably to overall decision value. Combining both information sets (X^{GB}) yields the highest policy value, a 41.5% improvement over the logging baseline, indicating that the two types of information together provide the most effective targeting strategy. This pattern underscores that improvements in predictive accuracy do not necessarily lead to better decision quality. In Web Appendix §C.2, we compare the value generated by each information regime under two different learning algorithms, XGBoost and LSTM.

6.2.2 Complement or Substitute? Aggregate Level

We now investigate whether behavioral and geographical data act as *substitutes* or *complements* in informing ad targeting decisions, starting with the aggregate level. Formally, let:

$$\Delta = \left(\hat{V}_{X^{GB}} - \hat{V}_{X^{B}}\right) - \left(\hat{V}_{X^{G}} - \hat{V}_{X^{\varnothing}}\right),$$

where $\hat{V}_X$ denotes the IPS-based value estimate for policy X. The first term in parentheses measures the incremental value of adding geographic information when behavioral data are already available, while the second term measures the incremental value of adding geographic information when no user-level data are observed.

If $\Delta > 0$, behavioral and geographical information are *complements* (Definition 2), meaning the value of combining them exceeds the sum of their stand-alone gains. If $\Delta < 0$, they are *substitutes* (Definition 1), meaning the combined gain is less than additive and the two information sets overlap in the information they provide. We estimate Δ using per-impression differences and conduct a one-sample t-test on the mean, clustering standard errors at the user level to account for correlation within users. We report the two-sided test, as one-sided results cannot be significant when the two-sided test fails to reject the null.

Table 4: Aggregate Complementarity Test for Behavioral and Geographic Information

	Estimate ($\hat{\Delta}$)	Std. Error (clustered)	t-stat	95% Confidence Interval	p-value
Value	0.000509	0.000334	1.526	[-0.000144, 0.001164]	0.127

Notes: The table reports the aggregate test of complementarity between behavioral and geographic information. Standard errors are clustered at the user level. Number of observations = 3,162,376; number of user clusters = 141,595.

As shown in Table 4, the aggregate interaction estimate is close to zero and statistically insignificant. This implies that, on average, combining behavioral and geographical data produces nearly additive gains, with no systematic evidence of complementarity or substitutability at the market level. While each information set captures distinct user heterogeneity, their joint effect does not amplify or diminish targeting value in

the aggregate. Given this null aggregate finding, we next examine whether the nature of interaction varies with user impression depth, when complementarity or substitutability may emerge.

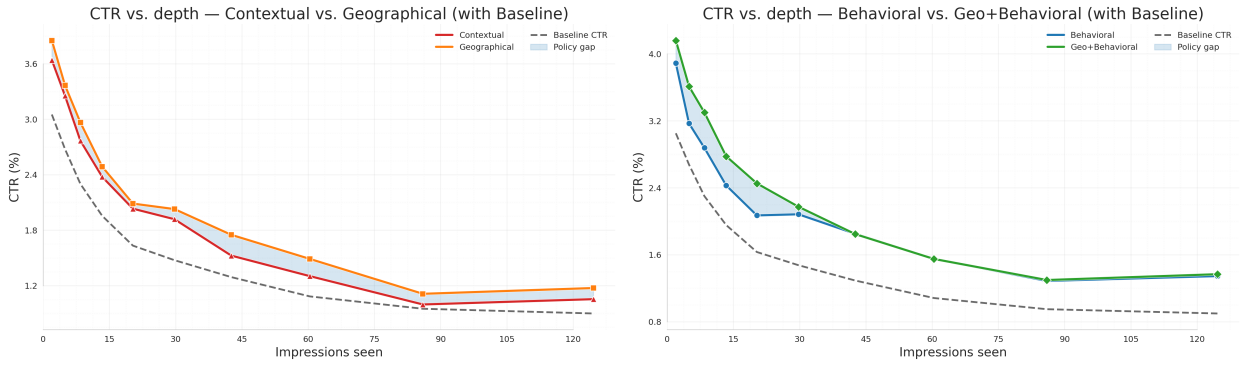
6.3 Behavioral vs. Geographical Information: Heterogeneity by User Exposure

The null result at the aggregate level motivates a closer examination of how the value and interaction of behavioral and geographical information vary across users. We focus on heterogeneity by *impression depth*, the number of impressions observed from each user, which proxies how much the platform has learned about each user. We proceed in two steps. First, in §6.3.1, we assess how the decision value of each information set changes with user exposure. Then, in §6.3.2, we test whether the relationship between behavioral and geographical information shifts from substitutive to complementary as more behavioral data accumulate.

6.3.1 Heterogeneous Value of Behavioral and Geographical Information by User Exposure

Building on the aggregate results presented in §6.2.1, we now explore how the value of information varies with the user's impression history. Specifically, we investigate whether the gain from targeting differs depending on how many impressions a user has previously seen. This analysis is motivated by Hypothesis 1, which posits that the utility-relevant information $f(X_t^B, a)$ converges as more behavioral data is observed. If true, the marginal value of behavioral or combined data may evolve with impression depth.

To operationalize this, we first sort all impressions for each user j by their timestamp t and assign a depth index accordingly. We then group observations into bins such that each bin contains the same number of impressions (i.e., quantile-based binning). Within each bin, we compute the absolute CTR levels for each targeting policy using the IPS-estimated click probabilities, alongside the empirical baseline CTR. These per-bin averages are then plotted against impressions have seen by user to visualize how click-through performance evolves as users receive additional exposures. Figure 6a and Figure 6b present these results for the main targeting comparisons.



(a) Contextual vs. Geographical Targeting.

(b) Behavioral vs. Geo + Behavioral Targeting.

Figure 6: Heterogeneity in Policy Value by Impression Depth. Each plot shows the *absolute CTR levels* (in %) for two targeting policies and the empirical baseline. Shaded areas indicate the policy gap, while the dashed line marks the baseline CTR.

In the left panel of Figure 6a, the X^G (Geographical) policy consistently delivers higher click-through rates than the $X^\varnothing$ (Contextual) policy across all impression depths. This persistent advantage demonstrates

that location-based segmentation contributes a stable and meaningful improvement in predictive accuracy beyond what contextual cues alone can capture. Both strategies rely on static features, yet geographical information introduces cross-regional heterogeneity that enhances targeting precision. The visible and sustained policy gap indicates that spatial variation systematically enriches model performance, yielding higher engagement probabilities even when no adaptive or personalized data are used.

In the right panel of Figure 6b, both the X^B (Behavioral) and X^{GB} (Combined) policies achieve substantially higher click-through rates than the baseline, underscoring the value of behavioral information in driving personalization. The X^{GB} policy consistently dominates X^B across the number of impressions seen by each user, indicating that leveraging both behavioral and geographical data yields a persistent performance advantage. However, the gap between the two narrows over time as user histories become richer, suggesting that behavioral data alone eventually become sufficient for accurate targeting as more interactions are observed. From a marketing perspective, this pattern implies that while geography initially enhances personalization in early exposures, its marginal contribution diminishes once rich behavioral histories accumulate.

6.3.2 Complement or Substitute? Heterogeneity by User Exposure

Building on the aggregate results in Table 4, we now examine how the relationship between X^B and X^G varies with the number of impressions a user has seen. We group impressions into bins of equal size, each containing the same number of observations, defined as the number of prior ads shown to the same user. For each bin, we compute two policy-value differences: (i) the value difference between the X^{GB} and X^B policies, and (ii) the value difference between the X^G and $X^\varnothing$ policies.

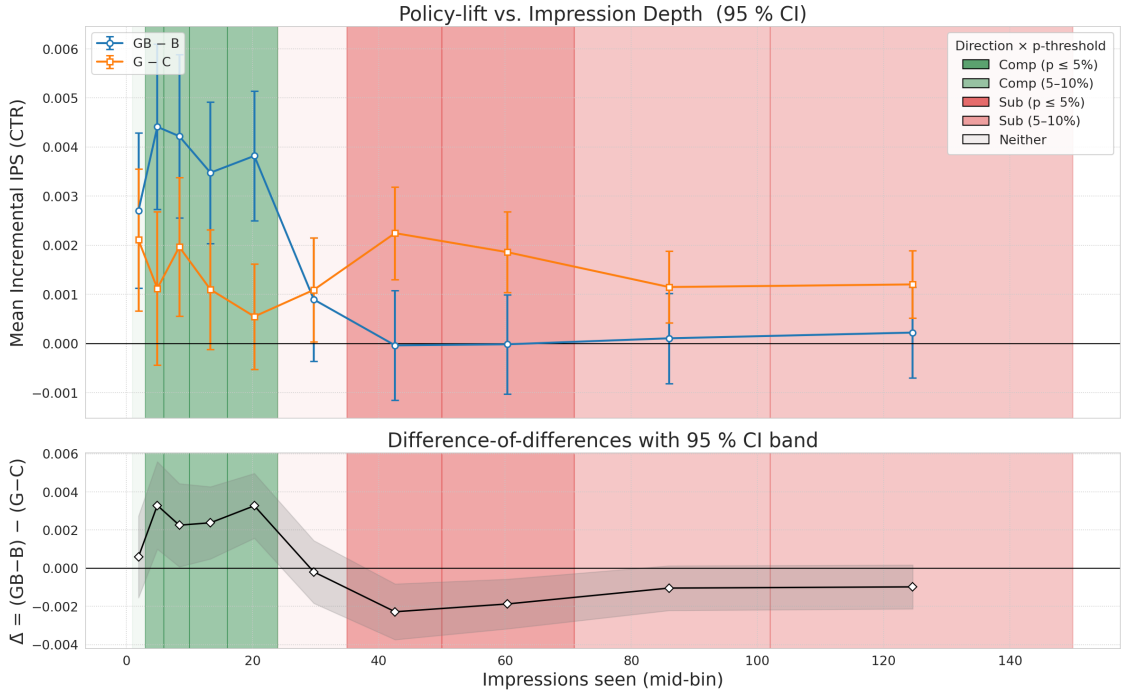


Figure 7: Complementarity and substitutability by impression depth. Top: $X^{GB} - X^B$ (blue) and $X^G - X^\varnothing$ (orange) with 95% CIs. Bottom: difference-in-differences $\bar{\Delta}$ with 95% CI; positive indicates complementarity, negative indicates substitutability.

The top panel of Figure 7 shows the value difference $X^{GB} - X^{B}$ (blue) and $X^{G} - X^{\varnothing}$ (orange) across impression depth, with vertical bars indicating 95% confidence intervals. The bottom panel summarizes their relationship through the difference-in-differences $\bar{\Delta} = \left(X^{GB} - X^{B}\right) - \left(X^{G} - X^{\varnothing}\right)$, where positive values indicate complementarity (joint value exceeds additivity) and negative values indicate substitutability (information sets overlap in value). Green and red shading mark depth ranges with statistically significant complementarity or substitutability, respectively.

The pattern in Figure 7 reveals three phases. In the *minimal behavioral history* stage (about 1–2 impressions), X^{B} contains basically no behavioral data, so the X^{B} policy has little ability to distinguish among users' latent types ω. At this point, the stand-alone contribution of X^{G} relative to $X^{\varnothing}$ is at its highest, because static geographic segmentation can immediately capture broad, location-related differences in preferences and demand that are otherwise unavailable in the absence of behavior. In contrast, adding X^{B} to the X^{G}-based policy (i.e., moving from X^{G} to X^{GB}) yields little incremental value, as there is no substantive behavioral data to interact with geography. The value of X^{G} in this stage is therefore realized almost entirely in both stand-alone X^{G} and X^{GB} policies.

In the *sparse behavioral history* stage (up to roughly 25 impressions), X^{B} begins to accumulate meaningful cues about ω, such as early engagement patterns and emerging ad preferences, that are partially orthogonal to the information in X^{G}. Here, X^{G} transitions from being primarily a stand-alone differentiator to acting as a complement to X^{B}: the value difference $X^{GB} - X^{B}$ overtakes $X^{G} - X^{\varnothing}$, indicating that the joint use of geographic information and emerging behavioral data now increases targeting value beyond the sum of their separate gains. From a marketing perspective, this stage marks the point where the cold-start problem is mitigated and where multi-information sets targeting produces the greatest synergy: behavioral history is informative enough to support personalization, while geography still contributes distinct, non-overlapping variation in ω that sharpens audience matching.

Finally, in the *rich behavioral history* stage (from the low 20s into the 70s and beyond), X^{B} becomes substantially more informative about ω. By this point, accumulated behavioral history captures not only individual-specific preferences but also many dimensions correlated with X^{G}, such as regional patterns of interest inferred indirectly from repeated engagement behavior. As a result, the marginal benefit of adding X^{G} to X^{B} relative to $X^{\varnothing}$ turns negative, indicating substitutability. In this regime, X^{B} already encodes much of the value that X^{G} would provide, so the X^{GB} policy yields no more value than the sum of the stand-alone contributions. From a marketing perspective, this stage marks a strategic pivot: once behavioral profiles are rich, geographic segmentation no longer delivers independent value, and targeting can be streamlined to X^{B} alone. For the full set of statistical test results underlying this analysis, we refer readers to Web Appendix §C.3.

7 Mechanisms Underlying the Role of Geographical Data

We now turn to the mechanisms that explain why geographical data (X^{G}) contributes to targeting outcomes in relation to behavioral data (X^{B}). Our unified framework has shown how the two information sets act as complements or substitutes in the decision value they generate. What remains unresolved is *why* geography improves performance in certain cases: does X^{G} capture an independent channel of information about user responsiveness, or does it primarily serve as a proxy for preference patterns that behavioral data eventually

reveal? To address this question, we investigate the channels through which a user's location can shape the probability of clicking on an ad. Prior work in marketing and consumer networks highlights three broad sources of spatially correlated responses: homophily, localized influence, and contextual confounding (e.g., Lovett et al., 2013; Aral and Walker, 2012). Building on this classification, we distinguish three channels through which geographical data can affect ad responsiveness and consider, for each, the extent to which behavioral data may eventually capture the same information:

1. *Spatial homophily.* People who live in the same area tend to share traits that matter for advertising outcomes, such as cultural norms, income levels, shopping patterns, or media habits. Geographical data captures this by using location as a proxy for these shared preferences, which vary smoothly across space. Behavioral data, however, can eventually recover much of the same information, since repeated user actions reveal the underlying preferences that geography initially stands in for.

2. *Social influence.* When individuals are geographically close, their choices may affect one another: if neighbors or friends engage with an ad, others nearby may become more likely to click as well. Geographical data captures this channel indirectly, since proximity makes peer effects more likely even if they are not observed directly. Behavioral data may partially pick up such influence when peer-driven patterns leave traces in individual activity, but because influence is relational rather than individual, it is harder for behavioral histories to fully capture it.

3. *Contextual confounding.* Locations often differ in external conditions that shape exposure and responsiveness, such as local events, promotions, or retail availability. Geographical data captures these shocks by linking users to the regions where such conditions occur. Behavioral data can sometimes absorb these effects if they translate into distinctive usage patterns or response histories, but geography remains useful when the drivers are external and not directly recorded in user behavior.

In all three cases, geographical data improves targeting when ad responses exhibit spatially structured variation that is not yet explained by behavioral data. To formalize this, we express click propensity as:

$$y = f(X^B) + \varepsilon, \tag{3}$$

where $f(X^B)$ represents the component of y explained by behavioral data, and ε captures the residual variation unexplained by X^B. Geographical data X^G is informative precisely when it explains systematic spatial structure in ε, whether that structure arises from homophily, peer influence, or contextual confounding. The key question, therefore, is whether this residual spatial structure persists as behavioral histories become richer. If geography primarily proxies for homophily, then richer X^B should absorb the clustering and leave little residual role for X^G. If, instead, geography reflects independent channels such as persistent local context or influence effects, spatial correlation should remain even after controlling for behavioral data. In what follows, we evaluate this distinction directly.

7.1 Empirical Strategy: Residualized Spatial Autocorrelation Test

To evaluate these mechanisms, we test whether the residual component ε in Equation 3 contains systematic spatial structure that could be captured by X^G. If such structures remain after controlling for behavioral data,

this would suggest that geography conveys independent information; if not, it would indicate that geography primarily proxies for patterns that X^B already absorbs. We implement this idea through a residualized spatial autocorrelation (RSA) test.

Our empirical strategy proceeds by aggregating outcomes at the regional level. Let Y_c denote the total number of clicks and I_c the total number of impressions in region c. Each impression can be viewed as a Bernoulli trial with region-specific click probability p_c, but because click events are rare, the Binomial distribution is well approximated by a Poisson model:

$$Y_c \mid p_c \sim \text{Poisson}(I_c\, p_c).$$

This specification is standard in spatial count-data analysis. The offset term $\log(I_c)$ adjusts for heterogeneous exposure across regions, while the residuals from the fitted model provide a natural basis for testing whether unexplained variation in click-through rates is spatially correlated.

Step 1: Baseline spatial structure in raw CTR. As a starting point, we examine whether CTRs exhibit geographic clustering in the absence of any behavioral controls. We estimate a Poisson count model that includes only an exposure offset and define the residual:

$$Y_c \sim \text{Poisson}(\mu_c), \quad \log \mu_c = \alpha + \log(I_c), \quad \varepsilon_c^{(0)} = \log(Y_c) - \alpha - \log(I_c),$$

Here, $\log(I_c)$ ensures that μ_c/I_c corresponds to the expected CTR in region c, while $\varepsilon_c^{(0)}$ measures deviations from the global mean CTR after adjusting for exposure volume. Evidence of spatial autocorrelation in $\varepsilon_c^{(0)}$ would indicate that raw CTRs contain location-based structure, consistent with homophily, localized influence, or contextual confounding.

Step 2: Residual spatial structure after controlling for behavior. Next, we ask whether this spatial structure remains once behavioral data are taken into account. To do so, we augment the model with the fitted behavioral component $f(X_c^B)$ and compute the adjusted residual:

$$Y_c \sim \text{Poisson}(\mu_c), \quad \log \mu_c = \alpha + f(X_c^B) + \log(I_c), \quad \varepsilon_c^{(B)} = \log(Y_c) - \alpha - f(X_c^B) - \log(I_c),$$

This residual captures variation in CTR unexplained by either exposure or behavioral features. If $f(X_c^B)$ successfully absorbs location-linked patterns, spatial dependence in $\varepsilon_c^{(B)}$ should be weaker than in $\varepsilon_c^{(0)}$.

Step 3: Testing for spatial dependence. Finally, we test whether $\varepsilon_c^{(0)}$ and $\varepsilon_c^{(B)}$ exhibit spatial autocorrelation. We apply two standard measures: Moran's I (Moran, 1950), which captures global correlation across regions, and Geary's C (Geary, 1954), which is more sensitive to local clustering. For a generic set of residuals ε_c indexed by region c, these are defined as:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(\varepsilon_i - \bar{\varepsilon})(\varepsilon_j - \bar{\varepsilon})}{\sum_i (\varepsilon_i - \bar{\varepsilon})^2}, \quad C = \frac{(N-1)\sum_i \sum_j w_{ij}(\varepsilon_i - \varepsilon_j)^2}{2\sum_i (\varepsilon_i - \bar{\varepsilon})^2},$$

where N is the number of regions, w_{ij} denotes the (i, j) element of the spatial weight matrix W, and $\bar{\varepsilon}$ is the mean of the residuals. Significant positive values of Moran's I (or values of Geary's C below one) indicate that geographically proximate regions have similar residuals. By comparing the statistics for $\varepsilon_c^{(0)}$ and $\varepsilon_c^{(B)}$, we can assess whether behavioral features reduce spatial dependence, thereby clarifying whether geography provides independent information beyond what behavior explains.

7.2 Results: Spatial Correlation in CTRs

We present two sets of results. First, in §7.2.1, we examine spatial correlation in baseline versus behavior-adjusted CTR residuals. Second, in §7.2.2, we compare spatial dependence under sparse versus rich behavioral histories. Together, these analyses show how the role of geography changes once behavioral information is taken into account.

7.2.1 Residual Spatial Dependence in Baseline and Behavior-Adjusted Models

We begin by applying the RSA framework introduced in §7.1. Specifically, we compare spatial correlation in the baseline residuals $\varepsilon_c^{(0)}$, obtained without behavioral controls, to the behavior-adjusted residuals $\varepsilon_c^{(B)}$. This analysis is conducted at both the county and city levels to assess how much of the observed spatial dependence can be explained by behavioral data.

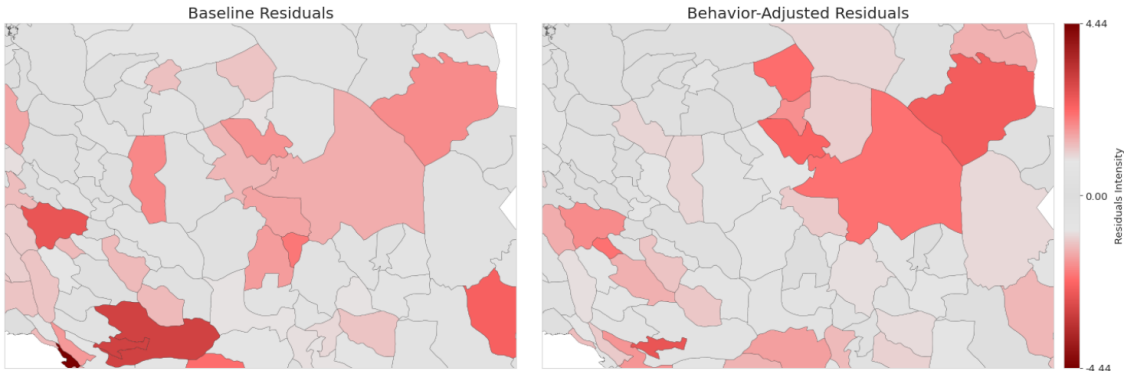


Figure 8: Spatial distribution of baseline (left) and behavior-adjusted (right) residuals at the county level. Darker red shades indicate higher residual intensity. Behavioral adjustment visibly reduces, but does not fully eliminate, spatial clustering.

Figure 8 provides an intuitive visualization for the county-level case. The left panel displays the baseline residuals $\varepsilon_c^{(0)}$ from the model without behavioral controls. Several contiguous regions show clusters of similarly high or low residual values, indicating clear spatial autocorrelation. Such clustering is consistent with the presence of spatial homophily, locally clustered influence, or contextual shocks. In contrast, the right panel shows the behavior-adjusted residuals $\varepsilon_c^{(B)}$. Incorporating $f(X_c^B)$ visibly reduces the strength and extent of spatial clusters, particularly in the southwest, although notable pockets remain. This pattern suggests that behavioral data absorb a substantial share of the location-linked variation, but not all of it.

To quantify the patterns observed in Figure 8, Table 5 reports Moran's I and Geary's C at both county and city levels. At the county level, the baseline residuals $\varepsilon_c^{(0)}$ exhibit strong and significant spatial dependence: regions with unusually high (or low) CTR residuals tend to be geographically proximate. After

controlling for behavioral information, both statistics decline markedly, indicating that much of this clustering is explained by differences in user behavior. However, the remaining positive spatial correlation implies that geography continues to capture additional variation not yet absorbed by behavior.

At the city level, the attenuation is sharper. Baseline residuals still display detectable clustering, but once behavioral features are included, both Moran's I and Geary's C fall to levels that are statistically indistinguishable from zero. This result suggests that at finer geographic resolution, most of the spatial variation in CTR can be accounted for by behavioral histories, leaving little independent role for geography. The contrast between county and city levels thus highlights how geography's incremental contribution diminishes as spatial units become more granular, consistent with the idea that geography is largely serving as a proxy for preference similarities that behavioral data can eventually capture.

	County-level, $\varepsilon_c^{(0)}$	**County-level, $\varepsilon_c^{(B)}$**	**City-level, $\varepsilon_c^{(0)}$**	**City-level, $\varepsilon_c^{(B)}$**
Moran's I	0.122	0.084	0.055	-0.001
p-value	0.00006	0.00318	0.00242	0.49156
Geary's C	0.891	0.900	0.948	0.996
p-value	0.00094	0.00162	0.01646	0.41863

Table 5: Moran's I and Geary's C statistics for spatial autocorrelation of residuals before (baseline) and after (behavior-adjusted) controlling for behavioral features, at county and city levels.

Taken together, these results indicate that geography provides clear predictive power when behavior is excluded, that a substantial portion of this power is absorbed once behavioral data are included, and that the unexplained residual role of geography is smaller at finer spatial resolution. To probe whether this attenuation also depends on the depth of behavioral histories observed for each user, we next split impressions into sparse and rich subsets and repeat the analysis.

7.2.2 Residual Spatial Dependence under Sparse versus Rich Behavioral Histories

As discussed earlier, the ability of the behavioral model $f(X^B)$ to account for variation in CTR depends on the amount of behavioral history available for each user. To examine this effect, we split the sample into two groups: (i) a *sparse-history* group, consisting of the first 50% of impressions observed for each user, and (ii) a *rich-history* group, consisting of an equal number of later impressions for the same users. This design allows us to compare scenarios where targeting models rely on limited versus extensive behavioral histories.

Figure 9 illustrates the spatial distribution of behavior-adjusted residuals at the county level for the two groups. In the sparse-history case (left panel), several large contiguous clusters of high residuals remain even after controlling for $f(X^B)$, indicating that location continues to explain a meaningful share of CTR variation. By contrast, in the rich-history case (right panel), residual clustering is visibly weaker and more fragmented, suggesting that much of the spatial dependence disappears once users accumulate longer behavioral records.

Table 6 formalizes these visual patterns. Under sparse histories, Moran's I is positive and statistically significant at the county level, while Geary's C also indicates non-random clustering. These results confirm that when behavior is limited, geography captures residual variation that remains spatially structured. In the rich-history case, however, both statistics decline substantially in magnitude and lose statistical significance across county and city levels. This indicates that once behavioral data are rich, $f(X^B)$ already accounts for

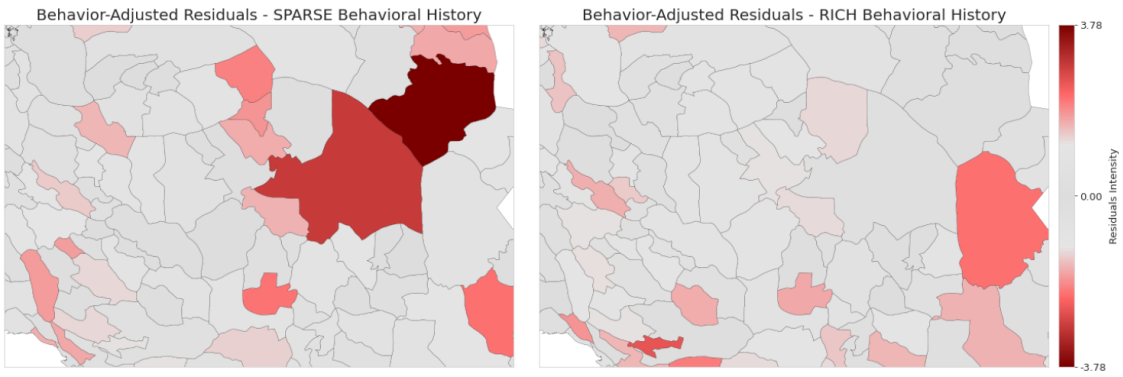


Figure 9: Behavior-adjusted residuals at the county level for sparse (left) and rich (right) behavioral histories. Darker shades indicate higher residual intensity.

	County-level, Sparse	**County-level, Rich**	**City-level, Sparse**	**City-level, Rich**
Moran's I	0.069	0.025	-0.005	0.0005
p-value	0.011	0.183	0.412	0.459
Geary's C	0.952	0.976	1.012	1.006
p-value	0.072	0.225	0.285	0.403

Table 6: Moran's I and Geary's C statistics for spatial autocorrelation of behavior-adjusted residuals under sparse and rich behavioral histories, at county and city levels.

nearly all of the systematic spatial correlation in CTR.

The sparse–rich comparison also clarifies the mechanisms through which geography matters. With sparse histories, geography captures (i) *spatial homophily*, since user preferences have not yet been revealed through repeated actions; (ii) *localized influence*, as early behaviors may not fully reflect peer-driven effects; and (iii) *contextual confounding*, when local shocks affect responsiveness before leaving behavioral traces. As histories grow richer, however, behavioral data progressively absorb these channels. Homophily is revealed directly through observed choices, contextual shocks leave footprints in repeated usage or response patterns, and even influence effects can be partially proxied when sustained peer-driven behaviors manifest in individual records. What remains most difficult for X^B to capture are influence effects that are inherently relational rather than individual.

The sparse–rich comparison also sheds light on which mechanisms drive the value of geographical data. The fact that spatial correlation is strong when behavioral histories are short but vanishes once histories become rich points most clearly to *spatial homophily* as the dominant channel. In the early stage, location acts as a proxy for shared preferences that are not yet visible in limited behavioral traces. As histories accumulate, these latent similarities are revealed directly through user actions, allowing $f(X^B)$ to absorb nearly all of the geographical data.

By contrast, channels such as localized influence or contextual confounding appear to play at most a minor role. If peer effects or persistent local shocks were central, we would expect geography to retain explanatory power even after conditioning on rich behavioral data. Instead, their fading significance suggests that such effects are either weak or leave footprints that behavior eventually captures. Taken together, the evidence indicates that geography's value stems primarily from homophily, which behavioral data progres-

sively subsume as user histories grow richer.

8 Managerial and Policy Implications

Our findings highlight a critical trade-off for advertisers and policymakers: while geographic data introduces privacy risks in combination with behavioral data, its value in engagement modeling is limited in addition with behavioral data. Although geographic proximity may serve as a temporary complement when behavioral data is sparse, our analysis suggests that it can be substituted by behavioral information in the long run. This raises important questions about whether the continued collection of location data is justified, given both its limited utility and growing regulatory concerns. In this section, we offer a series of managerial and policy implications in light of our results.

8.1 Geographic Data as a Complement in the Absence of Behavioral History

When behavioral information is sparse, geographical data acts as a short-term complement, providing incremental value that compensates for the lack of rich behavioral histories. In these early stages, spatial correlations in engagement patterns allow geographic data to capture similarities among users in the same location, revealing latent preference structures that behavioral models cannot yet detect. This complementary role enables advertisers to improve targeting decisions during the cold-start phase, before sufficient behavioral interactions have accumulated to support robust personalization.

8.2 Behavioral Data as a Substitute for Geographic Targeting

When behavioral information becomes sufficiently rich, it substitutes for the value previously provided by geographic information. In this regime, robust behavioral models capture much of the latent heterogeneity that geography once explained, causing the underlying spatial correlations in residual engagement to vanish once behavioral factors are accounted for. This shift implies that geographic targeting no longer delivers independent incremental value, and continued reliance on location-based segmentation serves only as a redundant proxy. Given increasing regulatory constraints on location tracking and heightened consumer privacy concerns, the diminishing marginal value of geographic information in the presence of rich behavioral histories suggests that firms should re-evaluate the strategic role of geo-targeting in their advertising portfolios. At the same time, firms can boost consumer trust and brand reputation by adopting transparent, user-controlled data practices—e.g., clear opt-out options once behavioral data become sufficiently rich—since transparency and control reduce perceived vulnerability (Martin et al., 2017).

8.3 When Does Geographic Data Remain a Meaningful Compelment?

Our study focuses on online in-app advertising, where engagement is primarily shaped by behavioral patterns rather than physical location. However, in other industries, geographic data may still play a more substantial role. Sectors such as ride-hailing, food delivery, local event promotions, and brick-and-mortar retail discounts rely heavily on user location to optimize their outcomes of interest, making geographic data potentially more valuable in these contexts. While our findings indicate that behavioral information can substitute for geographic targeting in digital advertising, we cannot generalize this conclusion to industries where location may carry intrinsic value. Further research is needed to assess whether behavioral data alone is sufficient across different sectors or if geographic insights remain a critical component of engagement

strategies.

8.4 Reassessing the Trade-Off Between Privacy Risks and Geographic Targeting

As privacy regulations become stricter and user concerns over data collection grow, advertisers must reconsider whether the limited benefits of geographic data justify its privacy risks. When behavioral data is available, geographic information adds little value, making its continued collection difficult to justify. However, for users with little engagement history, geographic data may still serve a temporary role in engagement modeling. A more privacy-conscious approach would involve using geographic data as a complement in the short term but transitioning toward behavioral substitutes as user profiles become richer.

9 Conclusion

In this paper, we develop a unified framework to evaluate the value of information in digital advertising, integrating economic theory, machine learning, and causal inference. Conceptually, we extend decision-based definitions of information value to test whether behavioral and geographical data act as complements or substitutes. Methodologically, we combine LSTM architectures with attention mechanisms to capture temporal and spatial dynamics and use inverse propensity scoring to recover counterfactual performance from observational data. Together, these elements provide a generalizable approach for comparing multiple information sets in high-dimensional environments, with applications extending beyond advertising.

Our results show that both geographical and behavioral data enhance targeting, but their roles change as behavioral histories become richer. At the aggregate level, complementarities and substitutions offset one another, yielding no clear overall interaction. As users accumulate impressions, geography explains most variation when behavioral data contain almost no information, complements behavior when histories are sparse, and becomes a substitute once histories are rich. This progression indicates that the value of geography is temporary, concentrated in the early stages before behavioral information alone suffices for accurate targeting.

To explore the underlying mechanism, we develop a Residualized Spatial Autocorrelation test. We first document strong spatial clustering in ad responses, consistent with nearby users exhibiting similar preferences. Once outcomes are residualized on behavioral data, much of this correlation vanishes, indicating that repeated user actions absorb the underlying similarity. Comparing sparse and rich histories, residual clustering persists only when behavioral information is limited but disappears once histories are extensive. These results suggest that the apparent value of geography is primarily driven by homophily that behavioral data ultimately capture. We do not find evidence of an independent role for geography beyond this channel and make no claims regarding other mechanisms such as influence or local shocks.

Future research could extend our analysis in several ways. First, beyond in-app advertising, the value of geographic data may differ in sectors where location remains central, such as transportation, local retail, or urban planning. Second, future work should trace the privacy–utility frontier by evaluating policy value under explicit data-use constraints, including coarse geolocation, k-anonymity, differential privacy, or federated learning. Third, while our results suggest homophily as the main driver of geographic value, further research could disentangle the roles of social influence and local shocks through experimental or network-based identification strategies.

References

Acquisti, A., Taylor, C., and Wagman, L. (2016). The economics of privacy. *Journal of economic Literature*, 54(2):442–492.

Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Aridor, G., Che, Y.-K., Hollenbeck, B., McCarthy, D., and Kaiser, M. (2024). Evaluating the impact of privacy regulation on e-commerce firms: Evidence from apple's app tracking transparency.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of marketing Research*, 55(1):80–98.

Bergemann, D. and Bonatti, A. (2011). Targeting in advertising markets: implications for offline versus online media. *The RAND Journal of Economics*, 42(3):417–443.

Blackwell, D. (1953). Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24(2):265–272.

Bollinger, B. and Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31(6):900–912.

Börgers, T., Hernando-Veciana, A., and Krähmer, D. (2013). When are signals complements or substitutes? *Journal of Economic Theory*, 148(1):165–195.

Bronnenberg, B. J., Dhar, S. K., and Dubé, J.-P. H. (2009). Brand history, geography, and the persistence of brand shares. *Journal of political Economy*, 117(1):87–115.

Bronnenberg, B. J. and Mahajan, V. (2001). Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Science*, 20(3):284–299.

De Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., De Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., et al. (2018). On the privacy-conscientious use of mobile phone data. *Scientific data*, 5(1):1–6.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5.

eMarketer (2024). Mobile advertising 2024: In-app ads drive growth past $200 billion. https://www.emarketer.com/content/mobile-advertising-2024. Accessed August 2025.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146.

Ghose, A., Li, B., and Liu, S. (2019). Mobile targeting using customer trajectory patterns. *Management Science*, 65(11):5027–5049.

Goldfarb, A. and Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404.

Grbovic, M. and Cheng, H. (2018). Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 311–320.

Guo, C. and Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hochreiter, S. (1997). Long short-term memory. *Neural Computation MIT-Press*.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Iyengar, R., Van den Bulte, C., and Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing science*, 30(2):195–212.

Jerath, K. and Miller, K. M. (2024). Consumers' perceived privacy violations in online advertising. *arXiv preprint arXiv:2403.03612*.

Johnson, G. A., Shriver, S. K., and Du, S. (2020). Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, 39(1):33–51.

Kallus, N. (2019). Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems*.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Lovett, M. J., Peres, R., and Shachar, R. (2013). On brands and word of mouth. *Journal of marketing research*, 50(4):427–444.

Manchanda, P., Xie, Y., and Youn, N. (2008). The role of targeted communication and contagion in product adoption. *Marketing Science*, 27(6):961–976.

Martin, K. D., Borah, A., and Palmatier, R. W. (2017). Data privacy: Effects on customer and firm performance. *Journal of marketing*, 81(1):36–58.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

Mirrokni, V., Muthukrishnan, S., and Nadav, U. (2010). Quasi-proportional mechanisms: Prior-free revenue maximization. In *Latin American Symposium on Theoretical Informatics*, pages 565–576. Springer.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Pew-Center (2024). Mobile fact sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/. Accessed August 2025.

Quadrana, M., Cremonesi, P., and Jannach, D. (2018). Sequence-aware recommender systems. *ACM computing surveys (CSUR)*, 51(4):1–36.

Rafieian, O. (2023). Optimizing user engagement through adaptive ad sequencing. *Marketing Science*, 42(5):910–933.

Rafieian, O. and Yoganarasimhan, H. (2021). Targeting and privacy in mobile advertising. *Marketing Science*, 40(2):193–218.

Rafieian, O. and Yoganarasimhan, H. (2023). Ai and personalization. *Artificial Intelligence in Marketing*, pages 77–102.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Smith, A. N., Seiler, S., and Aggarwal, I. (2023). Optimal price targeting. *Marketing Science*, 42(3):476–499.

Swaminathan, A. and Joachims, T. (2015). The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28.

Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of marketing research*, 51(5):546–562.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wernerfelt, N., Tuchman, A., Shapiro, B. T., and Moakler, R. (2025). Estimating the value of offsite tracking data to advertisers: Evidence from meta. *Marketing Science*, 44(2):268–286.

Yin, J., Feng, Y., and Liu, Y. (2025). Modeling behavioral dynamics in digital content consumption: An attention-based neural point process approach with applications in video games. *Marketing Science*, 44(1):220–239.

Web Appendix

A Feature Generations X and Learning Algorithm Implementation

The features X comprise **contextual** (X^C), **geographical** (X^{Geo}), and **behavioral** (X^{Behave}) components. Contextual features include static metadata (e.g., time, device, network), geographical features encode spatial attributes (latitude/longitude, province, city) using low-dimensional embeddings, and behavioral features capture dynamic engagement patterns. We detail feature generation, preprocessing, and the variables associated with each information set in the following subsections.

A.1 Contextual features X^C

To capture the cyclical nature of time, the hour and minute of the day are encoded using sine and cosine transformations. For hour h and minute m, these features are defined as:

$$\text{Hour\_SIN} = \sin\left(2\pi\frac{h}{24}\right), \quad \text{Hour\_COS} = \cos\left(2\pi\frac{h}{24}\right),$$

$$\text{Minute\_SIN} = \sin\left(2\pi\frac{m}{60}\right), \quad \text{Minute\_COS} = \cos\left(2\pi\frac{m}{60}\right).$$

These transformations ensure that $23:00$ and $00:00$ are numerically close, accurately reflecting the cyclical nature of time. For each categorical variable—media package (app) name (M), device model (D), brand ID (B), operator ID (O), ISP ID (I), and advertisement ID (A)—embeddings are generated to convert them into numerical representations. While M, D, B, O, and I focus on the top 50 categories, A is restricted to 5 predefined categories:

$$f_C(c) = \begin{cases} k, & \text{if } c \in \text{Top 50 categories}, \ k \in \{1,\ldots,50\}, \\ 0, & \text{otherwise}, \quad \text{for } C \in \{M,D,B,O,I\}, \\ k, & \text{if } c \in 5 \text{ Categories}, \ k \in \{1,\ldots,5\}, \\ 0, & \text{otherwise}, \quad \text{for } C = A. \end{cases}$$

Here, C represents each feature set, and k corresponds to the embedded category index. This encoding reduces complexity by focusing on the most common categories while ensuring advertisement ID (A) maintains its specific 5-category structure.

A.2 Geographical features X^{Geo}

Geographical information captures the user's spatial context at varying levels of granularity. In our setting, these features include precise coordinates, *longitude* (λ) and *latitude* (ϕ), as well as administrative divisions such as *province* (P) and *city* (C). The coordinates (λ, ϕ) provide fine-grained locational information that can be used to compute spatial proximity between users or match impressions to regional market conditions. Province and city identifiers serve as coarser geographic categories that can be linked to aggregated demographic, socioeconomic, or cultural attributes. Because province and city are discrete categorical variables with many unique values, we represent them using learnable embeddings. This transformation assigns each

category to a dense, low-dimensional vector in a continuous space, enabling the model to capture similarity patterns between locations without losing category-specific information. Formally:

$$f_P(p) = \mathbf{e}_p \in \mathbb{R}^{d_P}, \quad f_C(c) = \mathbf{e}_c \in \mathbb{R}^{d_C},$$

where $\mathbf{e}_p$ and $\mathbf{e}_c$ are embedding vectors of dimensions d_P and d_C, respectively, learned jointly with the predictive model.

Integrating X^G with contextual features X^C can enhance targeting performance in multiple ways. First, location variables may interact with device and network attributes, for example, certain smartphone brands or ISPs may dominate in specific regions, allowing the model to capture joint patterns that are not evident from either source alone. Second, combining spatial context with temporal features can improve the detection of region-specific engagement cycles, such as peak usage hours in different cities. Finally, location information, when paired with contextual metadata such as operator ID or ISP ID, can serve as a proxy for unobserved market segmentation factors, thereby enriching the representation of the user's latent type ω without substantially increasing privacy risks. Models such as gradient-boosted decision trees (e.g., XGBoost) are particularly well-suited for this integration, as they can flexibly capture nonlinearities and high-order interactions between heterogeneous feature types, including continuous coordinates, and categorical embeddings, without requiring strong parametric assumptions.

A.3 Behavioral features X^{Behave}

The behavioral features capture user interactions and engagement dynamics at different levels. These features are categorized into three main types: *user-level features*, which track overall user exposure, clicks, and recency of interactions; *ad-level features*, which measure ad performance, frequency, and user engagement with specific ads; and *app-level features*, which focus on app usage, preferences, and their influence on user clicks. Together, these behavioral features provide a comprehensive view of user activity, enabling the model to identify patterns and relationships that drive engagement and response.

A.3.1 User Behavioral Features

Exposure Count (EC): Tracks the cumulative number of ad exposures for user i up to but not including the current impression n:

$$E_{i,n} = n - 1.$$

This feature provides a measure of user exposure history, which helps evaluate the frequency of ad interactions over time.

Click History (CH): Represents the cumulative number of clicks by user i up to but not including the current impression n:

$$C_{i,n} = \sum_{k=1}^{n-1} \text{CLICK}_{i,k}.$$

By excluding the current click, this feature highlights past user engagement and serves as a strong indicator of historical interaction behavior.

Session Click-Through Rate (SCTR): Measures the user's CTR within the session, calculated as the ratio of past clicks to past exposures for user i at impression n:

$$S_{i,n} = \frac{C_{i,n}}{E_{i,n}}.$$

This feature captures the effectiveness of ads in driving clicks during a session and reflects user engagement intensity.

Time Since Last Exposure (TSE): Represents the time gap between the current and previous ad exposures for user i:

$$TSE_{i,n} = \text{TIME}_{i,n} - \text{TIME}_{i,n-1}.$$

This feature helps identify the recency of ad interactions, which can influence user responsiveness to repeated exposures.

Time Since Last Click (TCE): Tracks the time elapsed since the last click by user i, up to the current impression n:

$$TCE_{i,n} = \text{TIME}_{i,n} - \text{LAST\_CLICK\_TIME}_i.$$

By measuring click recency, this feature provides insight into how recent engagement influences future click behavior.

A.3.2 Ad-Level Behavioral Features

Ad Frequency (F): Represents the cumulative number of times user i has been exposed to ad j up to the current impression n:

$$F_{i,j,n} = n.$$

This feature helps track ad repetition for a specific user, which is critical for understanding the impact of ad frequency on engagement.

Ad Click-Through Rate (Ad CTR): Measures the click-through rate for ad j by user i, based on clicks up to but not including the current impression n:

$$\text{CTR}_{i,j,n} = \frac{\sum_{k=1}^{n-1} \text{CLICK}_{i,j,k}}{n-1}.$$

By excluding the current click, this feature captures past user engagement with a specific ad, offering insights into its historical performance.

Overall Ad Click-Through Rate (Overall Ad CTR): Represents the overall click-through rate for ad j across all users, based on clicks up to but not including the current impression n:

$$\text{CTR}_j^{\text{Overall}} = \frac{\sum_i \sum_{k=1}^{n_{i,j}-1} \text{CLICK}_{i,j,k}}{\sum_i (n_{i,j}-1)}.$$

This feature benchmarks the performance of an ad across the entire user base, providing a global measure of its effectiveness.

A.3.3 App-Level Behavioral Features

Usage Share of App (U): Represents the proportion of exposures for app a relative to all exposures for user i up to but not including the current impression n:

$$U_{i,a,n} = \frac{\text{Exposures}_{i,a,n}}{\text{Total\_Exposures}_{i,n}}.$$

This feature highlights user preferences for specific apps by tracking their exposure distribution across all apps.

Effect of App (E): Measures the contribution of app a to cumulative user clicks, excluding the current click at impression n:

$$E_{i,a,n} = \frac{\text{Clicks}_{i,a,n}}{\text{Total\_Clicks}_{i,n}}.$$

By isolating historical click behavior, this feature evaluates the influence of a specific app on user engagement.

Preference for App (P): Represents the user's preference for app a within the context of ad j, based on exposures up to but not including the current impression n:

$$P_{i,a,j,n} = \frac{\text{Exposures}_{i,a,j,n}}{\text{Total\_Exposures}_{i,j,n}}.$$

This feature captures how user interactions with apps vary within specific ad contexts, providing insight into app-specific ad performance.

Influence of App (I): Quantifies the impact of app a on user clicks within the context of ad j, excluding the current click at impression n:

$$I_{i,a,j,n} = \frac{\text{Clicks}_{i,a,j,n}}{\text{Total\_Clicks}_{i,j,n}}.$$

This feature measures the effectiveness of an app in driving clicks for a specific ad, offering app-level insights for targeted strategies.

Overall App Usage (U^{Overall}): Represents the share of exposures for app a across all users, based on cumulative exposures up to but not including the current impression n:

$$U_{a,n}^{\text{Overall}} = \frac{\text{Exposures}_{a,n}}{\text{Total\_Exposures}_{n}}.$$

This feature tracks the overall popularity of an app by monitoring its relative exposure share across the dataset.

Overall App Effect (E^{Overall}): Measures the contribution of app a to overall clicks across all users, excluding the current click at impression n:

$$E_{a,n}^{\text{Overall}} = \frac{\text{Clicks}_{a,n}}{\text{Total\_Clicks}_n}.$$

This feature evaluates the global impact of an app on user engagement, providing a benchmark for app performance.

A.4 Behavioral Learning Algorithm Implementation: LSTM

To effectively model the sequential and temporal dynamics of user engagement, we employ the Long Short-Term Memory (LSTM) predictive model, a specialized class of Recurrent Neural Networks (RNNs) (Rumelhart et al., 1986). LSTMs, first introduced by Hochreiter (1997), address the limitations of traditional RNNs, such as the vanishing gradient problem, by incorporating a memory cell and gating mechanisms that regulate the flow of information. These features make LSTMs highly effective for tasks involving long-term dependencies and sequential patterns. LSTMs have been widely adopted in user engagement modeling due to their ability to handle sequential dependencies and dynamic user behavior (Grbovic and Cheng, 2018; Quadrana et al., 2018).

An LSTM unit maintains two key components at each time step t: the cell state C_t, which acts as a long-term memory reservoir, and the hidden state h_t, which encapsulates information relevant for the current time step. The evolution of these components is controlled by three gating mechanisms: the forget gate, the input gate, and the output gate, as defined:

- **Forget Gate.** The forget gate determines how much of the previous cell state C_{t-1} should be retained:

  $$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f),$$

  where f_t is the forget gate vector, h_{t-1} is the hidden state from the previous time step, x_t is the input at the current time step, W_f and b_f are learnable parameters (weights and biases), and $\sigma(\cdot)$ is the sigmoid activation function, mapping values to the range $[0, 1]$.

- **Input Gate.** The input gate determines the extent of new information to be incorporated into the memory cell. It is defined as:

  $$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i),$$

  $$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C),$$

  where i_t is the input gate vector, $\tilde{C}_t$ is the candidate cell state, and W_i, W_C, b_i, b_C are trainable parameters. The hyperbolic tangent (tanh) maps the candidate state values to $[-1, 1]$.

- **Cell State Update.** The cell state C_t is updated by combining the contributions of the forget and input gates:

  $$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t,$$

  where $\odot$ denotes element-wise multiplication.

- **Output Gate.** The output gate determines the current hidden state h_t, which is used to generate the output and influence subsequent computations:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o),$$

$$h_t = o_t \odot \tanh(C_t),$$

where o_t is the output gate vector and W_o, b_o are learnable parameters.

These equations collectively enable the LSTM to selectively retain, update, and output information across time steps, making it robust to long-term dependencies in sequential data.

In addition to the recurrent backbone, our predictive framework integrates several complementary components, each designed to address a specific challenge in sequential engagement modeling. First, *embedding layers* transform high-cardinality categorical inputs, such as device type, province, city, or network provider, into low-dimensional dense vectors. This approach not only reduces dimensionality but also enables the model to learn latent similarity structures among categories, improving generalization and mitigating the curse of dimensionality (Guo and Berkhahn, 2016). Second, temporal structure is enriched through two mechanisms: (i) *absolute positional embeddings*, which provide the model with awareness of an event's position within the sequence, and (ii) *time–gap embeddings*, which capture irregular intervals between events, allowing differentiation between dense bursts of activity and prolonged inactivity. Third, a *stacked LSTM* serves as the primary sequence encoder, effectively modeling order-sensitive dependencies and evolving behavioral patterns over time. Fourth, a *causal multi-head self-attention* layer enables the model to focus adaptively on past events that are most relevant to the current prediction, capturing long-range dependencies while enforcing strict causality constraints (Vaswani et al., 2017). Finally, a *gated projection head* selectively filters and transforms the attended sequence representations before a linear readout layer maps them to click probabilities, enhancing the interpretability and robustness of the final decision stage. Table A1 reports the architecture and parameter counts.

Table A1: Summary of the Sequence Model Architecture and Parameters

Layer (type)	**Output Shape**	**Param #**
Input Features	[-1, 150, 240]	0
Time-Gap Projection	[-1, 150, 240]	57,840
Absolute Positional Embedding	[1, 150, 240]	36,000
LSTM (4 layers)	[-1, 150, 512]	4,269,568
Layer Norm (pre-attention)	[-1, 150, 512]	1,024
Causal Multi-Head Attention	[-1, 150, 512]	1,049,600
Layer Norm (post-attention)	[-1, 150, 512]	1,024
Gate Projection (sigmoid)	[-1, 150, 512]	262,656
Body Projection (tanh)	[-1, 150, 512]	262,656
Dropout	[-1, 150, 512]	0
Final FC Output Layer	[-1, 150, 1]	513
Total Parameters		**5,941,905**

In the following, we provide a detailed, step-by-step description of the architecture, formally defining its components and explaining how they interact to transform raw sequential inputs into engagement probability forecasts.

We model user engagement as a sequential decision prediction problem, where the state at each time step incorporates both numerical and categorical information observed up to that point. Let j index users and $t \in \{1,\ldots,T\}$ denote the position in the sequence of impressions observed for user j, with a maximum sequence length $T = 150$.

Input Representation. At each time step t, we observe:

$$X_{jt}^{\text{num}} \in \mathbb{R}^n, \quad \{X_{jt}^{\text{cat},k}\}_{k=1}^{c},$$

where n is the number of numerical features and c is the number of categorical features. Numerical features capture continuous signals (e.g., time-based statistics, historical engagement rates, contextual metrics). Categorical features correspond to high-cardinality identifiers such as device model, province, city, or advertisement ID. Direct one-hot encoding of categorical features would be prohibitively high-dimensional and sparse. Instead, we map each categorical variable $X_{jt}^{\text{cat},k}$ to a dense, learnable embedding vector:

$$\mathbf{e}_{jt}^{(k)} \in \mathbb{R}^{d_k},$$

where d_k is the embedding dimension for category k. Embeddings enable the model to learn latent similarity structures, e.g., provinces with similar user behavior or devices with similar performance profiles, without manual feature engineering. Concatenating all categorical embeddings yields:

$$\mathbf{e}_{jt} = [\mathbf{e}_{jt}^{(1)},\ldots,\mathbf{e}_{jt}^{(c)}] \in \mathbb{R}^{E_{\text{cat}}}, \quad E_{\text{cat}} = \sum_{k=1}^{c} d_k.$$

The full feature vector at time t is:

$$X_{jt} = [X_{jt}^{\text{num}}, \mathbf{e}_{jt}] \in \mathbb{R}^{d_{\text{in}}}, \quad d_{\text{in}} = n + E_{\text{cat}}.$$

Temporal Augmentation. To model sequence position and irregular arrival times, we augment X_{jt} with two temporal components:

1. *Absolute positional embedding* $\mathbf{P}_t \in \mathbb{R}^{d_{\text{in}}}$, a learnable vector for each position t, allows the model to differentiate between early- and late-stage interactions even if other features are identical.

   While LSTMs can track order, they do so implicitly through hidden state transitions. Positional embeddings make temporal position an explicit feature, which is helpful in cases where position itself carries meaning (e.g., "first impressions" vs. "later impressions" behave differently). LSTM gates learn relative dependencies; positional embeddings give a direct absolute time index, which is crucial for long sequences where LSTM memory may fade.

2. *Time–gap embedding* models the elapsed time since the previous impression:

$$\Delta t_{jt} = \text{time}(j,t) - \text{time}(j,t-1),$$

transformed as $\log(1+\Delta t_{jt})$ and projected to $\mathbb{R}^{d_{\text{in}}}$ through a learned linear layer. This captures behavioral intensity, distinguishing between dense bursts of activity and long idle periods.

The spacing between events often changes behavior (e.g., if a user sees an ad after 5 seconds vs. after 5 days). Neither positional embeddings nor LSTM hidden states naturally capture this irregular spacing. Positional embeddings tell us "this is the 10th impression," but not whether it came an hour or a week after the previous one. Time–gap embeddings fill that gap.

The resulting time-aware token is:

$$\tilde{X}_{jt} = X_{jt} + \mathbf{P}_t + \text{TimeGapProj}\big(\log(1+\Delta t_{jt})\big).$$

Sequence Encoding. The augmented sequence $\{\tilde{X}_{j1},\ldots,\tilde{X}_{jT}\}$ is processed by a stacked LSTM with hidden size $H = 512$ and $L_{\text{LSTM}} = 4$ layers:

$$h_{jt}^{(\ell)} = \text{LSTM}_\ell\big(h_{jt}^{(\ell-1)}\big),$$

where $h_{jt}^{(0)} = \tilde{X}_{jt}$. The LSTM captures local and medium-range temporal dependencies, preserving order-sensitive dynamics such as gradual preference shifts or decaying effects of prior interactions. Attention alone doesn't model sequential dependencies as naturally as LSTMs, and attention without recurrence can sometimes struggle with shorter-term patterns when data is noisy. The LSTM provides a strong temporal backbone before attention refines it.

Causal Multi-Head Attention. To model longer-range dependencies and non-local feature interactions, we apply a causal multi-head self-attention layer with $H_{\text{attn}} = 4$ heads to the LSTM outputs. Attention reweights historical states based on their relevance to the current prediction. We enforce causality with a triangular mask:

$$\text{mask}(t',t) = \begin{cases} 0, & t' \le t, \\ -\infty, & t' > t, \end{cases}$$

ensuring that predictions at time t use only past and present information. A key–padding mask derived from $m_{jt} \in \{0,1\}$ ignores padded steps. Layer normalization before and after attention stabilizes optimization.

LSTMs compress history into a single hidden state, which can lose fine-grained details over long sequences. Attention re-opens the past and selectively retrieves important past events directly. It complements the LSTM by providing content-based memory access useful when long-range dependencies matter, e.g., a behavior 100 impressions ago is relevant now.

Gated Projection Head. The attention-enhanced representation is passed through a gated projection mechanism, which adaptively filters and transforms features:

$$g_{jt} = \sigma(W_g h_{jt}), \quad b_{jt} = \tanh(W_b h_{jt}),$$

$$z_{jt} = \text{Dropout}(g_{jt} \odot b_{jt}).$$

Here, the gate vector g_{jt} acts as a learnable feature selector, while b_{jt} generates candidate transformations. Their elementwise product modulates information flow, suppressing irrelevant history and emphasizing predictive patterns.

It lets the model selectively emphasize or suppress certain dimensions in the representation before making predictions, essentially a learned attention mechanism at the feature level. Attention looks across time, but this gating looks within the feature vector, so it's a different form of selectivity.

Prediction Layer. The gated features are mapped to logits:

$$\tilde{y}_{jt} = w^\top z_{jt}, \quad \hat{y}_{jt} = \sigma(\tilde{y}_{jt}),$$

where $\hat{y}_{jt}$ is the predicted click probability at time t.

Training Objective. We minimize the binary cross-entropy loss with logits:

$$\mathscr{L}(\Theta) = \sum_{j=1}^{n} \sum_{t=1}^{T} m_{jt}\, \ell_{\text{BCE-logits}}(y_{jt}, \tilde{y}_{jt}),$$

where m_{jt} masks out padded steps, and Θ includes all learnable parameters (embeddings, positional encodings, LSTM, attention, gating, and readout). Optimization uses AdamW (Loshchilov and Hutter, 2019), which combines adaptive learning rates with decoupled weight decay. Dropout is applied both in attention and projection layers for regularization.

This architecture integrates heterogeneous data types, numerical and categorical, into a unified, dense representation. Embeddings allow the model to exploit latent similarities among categorical entities without incurring the cost of high-dimensional sparse vectors. The positional and time-gap embeddings encode absolute sequence position and irregular temporal dynamics, both of which are crucial for interpreting user engagement behavior. The LSTM provides strong modeling of local order-sensitive dependencies, while the causal attention layer captures global relationships, enabling the model to focus selectively on distant but influential events. Finally, the gated projection head acts as a content-aware filter, passing forward only the most relevant transformed features for prediction. Together, these design elements yield a model capable of accurately forecasting engagement while respecting strict causality constraints.

B IPS and Propensity Score Evaluations

B.1 Proof of Proposion 1 (deterministic π^{f_X})

Define the deterministic IPS estimator

$$\hat{V}(\pi^{f_X}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{1}\{A_i = \pi^{f_X}(X_i)\}}{\pi^{\mathscr{D}}(A_i \mid X_i)}\, v(A_i)\, Y_i,$$

and treat π^{f_X} as fixed w.r.t. the evaluation sample (by using a holdout or cross-fitting).

- *Step 1:* By i.i.d. sampling of (X_i, A_i, Y_i) from the logging DGP $\big(X \sim p_X, A \mid X \sim \pi^{\mathscr{D}}(\cdot \mid X), Y \mid X, A \sim \Pr(\cdot \mid X, A)\big)$ and boundedness (Overlap Assumption 1 ensures the denominator is nonzero whenever $\mathbf{1}\{A = \pi^f(X)\} = 1$; $Y \in \{0,1\}$; $v(\cdot)$ finite), the strong law of large numbers (SLLN) yields

$$\hat{V}(\pi^{f_X}) \xrightarrow{a.s.} \mathbb{E}\left[\frac{\mathbf{1}\{A = \pi^{f_X}(X)\}}{\pi^{\mathscr{D}}(A \mid X)}\, v(A)\, Y\right].$$

- *Step 2:* First law of iterated expectations condition on X yields:

$$\mathbb{E}\left[\frac{\mathbf{1}\{A = \pi^{f}(X)\}}{\pi^{\mathscr{D}}(A \mid X)}\, v(A)\, Y\right] = \mathbb{E}_X\left[\mathbb{E}\left[\frac{\mathbf{1}\{A = \pi^{f}(X)\}}{\pi^{\mathscr{D}}(A \mid X)}\, v(A)\, Y \,\middle|\, X\right]\right].$$

- *Step 3:* Second law of iterated expectations condition on $A \mid X$ and expand the sum yields:

$$\mathbb{E}\left[\frac{\mathbf{1}\{A = \pi^{f_X}(X)\}}{\pi^{\mathscr{D}}(A \mid X)}\, v(A)\, Y \,\middle|\, X\right] = \sum_{a \in \mathscr{A}} \pi^{\mathscr{D}}(a \mid X)\, \frac{\mathbf{1}\{a = \pi^{f_X}(X)\}}{\pi^{\mathscr{D}}(a \mid X)}\, v(a)\, \mathbb{E}[Y \mid X, a].$$

- *Step 4:* The logging propensity cancels given Overlap Assumption 1:

$$= \sum_{a \in \mathscr{A}} \mathbf{1}\{a = \pi^{f_X}(X)\}\, v(a)\, \mathbb{E}[Y \mid X, a] = v(\pi^{f_X}(X))\, \mathbb{E}[Y \mid X, A = \pi^{f_X}(X)].$$

  Therefore,

$$\mathbb{E}\left[\frac{\mathbf{1}\{A = \pi^{f}(X)\}}{\pi^{\mathscr{D}}(A \mid X)}\, v(A)\, Y\right] = \mathbb{E}_X\left[v(\pi^{f}(X))\, \mathbb{E}[Y \mid X, A = \pi^{f}(X)]\right].$$

- *Step 5:* By Unconfoundedness Assumption 2 and the usual consistency ($Y = Y(A)$),

$$\mathbb{E}[Y \mid X, a] \;=\; \mathbb{E}[Y(a) \mid X].$$

  Hence

$$\mathbb{E}_X\left[v(\pi^{f_X}(X))\, \mathbb{E}[Y \mid X, A = \pi^{f_X}(X)]\right] = \mathbb{E}_X\left[v(\pi^{f_X}(X))\, \mathbb{E}[Y(\pi^{f_X}(X)) \mid X]\right].$$

- *Step 6:* By the definition, $u(a, \omega) = v(a) \Pr(Y = 1 \mid a, \omega)$ and $q_X(\omega) = \Pr(\omega \mid X)$, so for any x and a,

$$v(a)\, \mathbb{E}[Y(a) \mid X = x] = v(a) \sum_{\omega} \Pr(Y = 1 \mid a, \omega)\, q_x(\omega) = \sum_{\omega} u(a, \omega)\, q_x(\omega).$$

Plugging $a = \pi^{f_X}(x)$ gives:

$$\mathbb{E}_X\left[v(\pi^{f_X}(X))\,\mathbb{E}[Y(\pi^{f_X}(X)) \mid X]\right] = \mathbb{E}_X\left[\sum_{\omega} u(\pi^{f_X}(X),\omega)\,q_X(\omega)\right] = V(\pi^{f_X}).$$

- *Step 7:* Combining Steps 1–6,

$$\mathbb{E}\big[\hat{V}(\pi^{f_X})\big] \;=\; V(\pi^{f_X}), \qquad \hat{V}(\pi^{f_X}) \xrightarrow{a.s.} V(\pi^{f_X}).$$

- *Step 8:* If Policy optimality under X Assumption 3 holds, then

$$V(\pi^{f_X}) = \sup_{\pi} V(\pi) \;=:\; V_X.$$

Therefore,

$$\mathbb{E}\big[\hat{V}(\pi^{f_X})\big] = V_X, \qquad \hat{V}(\pi^{f_X}) \xrightarrow{p} V_X.$$

□

B.2 Propensity Score Estimation and Validation

Our IPS implementation requires estimates of the platform's logging policy probabilities (*propensity scores*) and verification of covariate balance. Appendix §B.2.1 details the estimation procedure, and Appendix §B.2.2 reports balance diagnostics confirming small post-weighting differences.

B.2.1 Propensity Score Estimation

Support and Eligibility Filtering. In our setting, not all ads are eligible to appear in every impression due to platform-level targeting constraints such as location, time-of-day, device type, or campaign-specific restrictions. To encode these constraints, we construct an *eligibility matrix* following Rafieian and Yoganarasimhan (2021):

$$E \in \{0,1\}^{N\times A},$$

where $e_{i,a} = 1$ if and only if ad a is eligible to be shown in impression i. The matrix E is obtained by intersecting eligibility indicators across multiple categorical dimensions, province, hour-of-day, app category, smartphone brand, network type, and media package name, where each dimension's support is derived from the set of observed ad–context combinations in the historical data.

The eligibility matrix E serves two roles in our estimation procedure. First, the columns $\{e_{\cdot,a}\}_{a=1}^{A}$ are included as binary features in the propensity model, allowing the learner to incorporate eligibility information directly into its predictions. Second, after model estimation, we enforce the *support condition* by setting $\hat{\pi}^{\mathscr{D}}_{i,a} = 0$ whenever $e_{i,a} = 0$, followed by renormalization across eligible ads. This guarantees that

$$\hat{\pi}^{\mathscr{D}}(a \mid x_i) > 0 \quad \text{only if} \quad e_{i,a} = 1,$$

while still permitting the model to flexibly learn variation in serving probabilities within the support.

Outcome and Covariates. The propensity score model predicts, for each impression $i \in \{1,\ldots,N\}$, which ad from the candidate set $\mathscr{A} = \{a^{(1)}, a^{(2)}, \ldots, a^{(K)}\}$ was displayed. The outcome variable is the multi-class indicator $\mathbf{1}\{A_i = a^{(j)}\}$, where A_i denotes the ad shown in impression i.

The covariate vector x_i is constructed to capture all observed determinants of ad allocation and includes the following components:

1. **Eligibility indicators:** For each ad a, we include the binary feature $e_{i,a} \in \{0,1\}$ from the eligibility matrix E described above. This ensures that the model incorporates ad-specific availability and targeting rules directly into its predictions.

2. **Targeting variables:** Province, app category, hour-of-day, smartphone brand, network connection type, and mobile service provider, each factorized into integer-encoded indices to capture categorical variation.

3. **Fine-grained location:** Latitude and longitude, enabling the model to detect geographic heterogeneity beyond province-level effects.

4. **Temporal controls:** Time-of-day features measured at the minute level, absorbing short-run fluctuations in bids or quality scores due to diurnal cycles or transient platform shocks.

5. **Device and context characteristics:** Device identifiers, fraud detection codes, and media package identifiers, capturing residual variation in serving probabilities not explained by higher-level targeting variables.

All categorical covariates are transformed into numeric indices using factorization, preserving a consistent mapping across impressions. The eligibility indicators $e_{i,a}$ for all $a \in \mathscr{A}$ are appended to the base covariate set, so the final feature matrix combines both contextual features and ad-specific availability indicators.

Estimation Procedure. The platform's quasi-proportional allocation rule implies that the probability of displaying ad $a \in \mathscr{A}$ in impression i can be expressed as

$$\pi^{\mathscr{D}}(a \mid x_i) \;=\; \frac{b_a(x_i)\, q_a(x_i)}{\sum_{a' \in \mathscr{A}_i} b_{a'}(x_i)\, q_{a'}(x_i)},$$

where $b_a(x_i)$ and $q_a(x_i)$ denote, respectively, the bid and quality score functions conditional on the observed covariates x_i, and $\mathscr{A}_i \subseteq \mathscr{A}$ is the set of ads eligible in context x_i.

In practice, these bid and quality components are not fully observed, and the realized allocation process may deviate from the theoretical form due to unobserved adjustments, platform-level interventions, and complex interactions among covariates. We therefore estimate the realized propensities $\pi^{\mathscr{D}}$ directly from the historical allocation data using a flexible, nonparametric method. Specifically, we adopt a binary one-vs-all approach: for each $a^{(j)} \in \mathscr{A}$, we estimate the conditional probability

$$\pi_j^{\mathscr{D}}(x_i) \;=\; \Pr\big(A_i = a^{(j)} \,\big|\, x_i\big),$$

By estimating a separate binary classifier for each ad $a^{(j)}$, we capture ad-specific allocation patterns without imposing a single functional form across all ads—an important flexibility given heterogeneous eligibility rules, competition, and campaign objectives. This one-vs-all formulation also mitigates the severe class imbalance of large-scale ad data: rare ads, which would exert little influence in a multiclass model's joint decision boundary, are given dedicated models in which their appearances are positive events. Moreover, it avoids the restrictive multiclass assumption that the same covariate effects and interactions govern all alternatives equally, an implausible condition in our setting.

For each ad-specific classifier, the training set includes all N impressions, even those in which the ad was ineligible ($e_{i,a^{(j)}} = 0$), allowing the model to learn to assign zero probability in contexts where the ad cannot be served. Each classifier is implemented as an XGBoost with a logistic loss for binary outcomes. To balance predictive accuracy and computational efficiency in a high-dimensional feature space, we adopt hyperparameters for the learning algorithm XGBosst motivated by Rafieian and Yoganarasimhan (2021).

To ensure that estimated propensities are free from in-sample overfitting, we adopt a K-fold cross-fitting procedure with $K = 5$. Let the set of impressions be indexed by $\{1,\ldots,N\}$, and let $\mathscr{I}_1,\ldots,\mathscr{I}_K$ denote a partition of this set into K disjoint folds of (approximately) equal size. For each fold $\mathscr{I}_k$, the ad-specific binary classifiers are trained on the complement set $\bigcup_{m\neq k}\mathscr{I}_m$, and predictions $\hat{\pi}^{\mathscr{D}}_{i,a}$ are generated only for impressions $i \in \mathscr{I}_k$. By construction, every propensity score is obtained from a model that did not observe the corresponding impression during training. Concatenating these out-of-fold predictions across all $k = 1,\ldots,K$ yields a complete $N \times A$ matrix of out-of-sample estimates. This procedure preserves the predictive structure learned by the models while eliminating the upward bias in inverse propensity estimators that arises from reusing training predictions.

Post Processing. Following propensity estimation, we align the predictions with the platform's targeting constraints by imposing the *support condition*. Given $E \in \{0,1\}^{N\times A}$ be the eligibility matrix, where $e_{i,a} = 1$ iff ad a is eligible in impression i. We set

$$\hat{\pi}^{\mathscr{D}}_{i,a} = 0 \quad \text{for all} \quad a \notin \mathscr{A}_i,$$

where $\mathscr{A}_i = \{a \in \mathscr{A} : e_{i,a} = 1\}$ is the eligible set for impression i. The remaining probabilities are renormalized:

$$\hat{\pi}^{\mathscr{D}}_{i,a} \leftarrow \frac{\hat{\pi}^{\mathscr{D}}_{i,a}}{\sum_{a'\in\mathscr{A}_i}\hat{\pi}^{\mathscr{D}}_{i,a'}} \quad \forall a \in \mathscr{A}_i,$$

so that $\sum_{a\in\mathscr{A}_i}\hat{\pi}^{\mathscr{D}}_{i,a} = 1$ and

$$\hat{\pi}^{\mathscr{D}}(a \mid x_i) > 0 \iff e_{i,a} = 1.$$

This adjustment preserves the relative allocation probabilities learned by the model among eligible ads while ensuring no probability mass is assigned to ineligible alternatives.

Results from Propensity Score Estimation. The estimated propensities capture systematic variation in ad allocation without overfitting to deterministic rules. The macro-averaged AUC is 0.6335 and the micro-averaged AUC is 0.6276, indicating moderate discriminatory power above random guessing but far from perfect classification. This aligns with expectations: if the platform's allocation contains substantial quasi-

randomization (subject to eligibility and proportional bidding), a very high AUC would be undesirable, as it would imply near-deterministic serving and reduced overlap. Likewise, the macro and micro log-loss values (0.4862 and 0.4388) are low enough to indicate well-calibrated predictions but not so low as to suggest overconfidence.

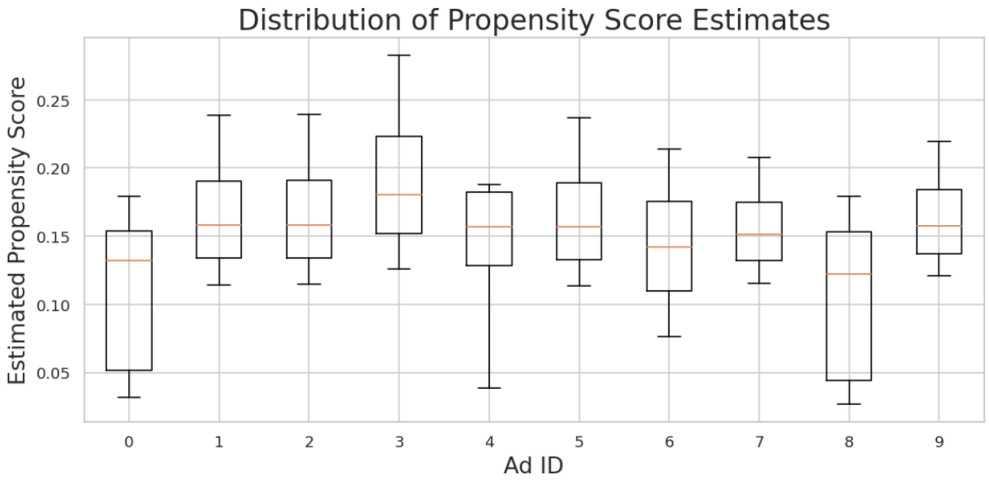


Figure A.1: Distribution of estimated propensity scores across ads. Dispersion and non-degeneracy indicate adequate common support.

Figure A.1 plots the distribution of estimated propensities for each ad. For most ads, the median serving probability lies between 0.12 and 0.18, with interquartile ranges spanning several percentage points. This dispersion indicates that the model assigns a range of probabilities across impressions rather than concentrating mass at extreme values, which is essential for maintaining overlap. Crucially, no ad exhibits degenerate propensities concentrated near zero (indicating near-never assignment) or near one (indicating near-deterministic assignment). Even the rarest ads display a nontrivial upper tail, implying that all ads have a positive probability of being shown in multiple contexts. To further assess whether reweighting achieves covariate balance across treatment arms, we next compute the standardized bias (SB) before and after applying the estimated inverse propensity weights in §B.2.2.

B.2.2 Covariate Balance Diagnostics

After estimating the logging policy propensities $\hat{\pi}^{\mathscr{D}}_{i,a}$ and enforcing support, we evaluate whether inverse-propensity weighting balances the distribution of each covariate across treatment arms using the *standardized bias* (SB) metric. The goal is to check whether, after weighting, the covariate distribution for impressions assigned to any given ad matches that of the eligible population, thereby providing evidence consistent with the unconfoundedness assumption under overlap. Let X denote a generic covariate, X_i its value for impression i, and A_i the ad shown. For ad $a \in \mathscr{A}$, the *unweighted* mean of X in impressions where $A_i = a$ is

$$\bar{X}_a \;=\; \frac{\sum_{i=1}^{N} \mathbf{1}(A_i = a)\, X_i}{\sum_{i=1}^{N} \mathbf{1}(A_i = a)},$$

while the mean in the *eligible* population is

$$\bar{X} \;=\; \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Let σ_X be the standard deviation of X in the eligible population. The *unweighted* standardized bias for X when assigned to a is then

$$SB_a(X) \;=\; \frac{|\bar{X}_a - \bar{X}|}{\sigma_X},$$

and the worst-case imbalance for X across all ads is

$$SB(X) \;=\; \max_{a\in\mathscr{A}} SB_a(X).$$

Intuitively, $SB_a(X)$ measures the relative shift (in standard deviation units) between the covariate mean for impressions shown ad a and the overall eligible population mean; $SB(X)$ takes the largest such shift across ads. Following McCaffrey et al. (2013), we adopt the benchmark $|SB(X)| < 0.20$ as the criterion for acceptable balance. Values above this threshold indicate potentially meaningful imbalance in X between treatment arms. After weighting by the inverse of the estimated propensities, the weighted mean of X for impressions shown ad a is

$$\bar{X}_a^{\hat{\pi}} \;=\; \frac{\sum_{i=1}^{N} \mathbf{1}(A_i = a)\,\frac{X_i}{\hat{\pi}_{i,a}^{\mathscr{D}}}}{\sum_{i=1}^{N} \mathbf{1}(A_i = a)\,\frac{1}{\hat{\pi}_{i,a}^{\mathscr{D}}}}.$$

Let $\bar{X}^{\mathrm{elig}(a)} = \frac{1}{|\mathscr{S}_a|}\sum_{i\in\mathscr{S}_a} X_i$ and $\sigma_X^{\mathrm{elig}(a)}$ denote, respectively, the mean and standard deviation of X in the ad-specific eligible set $\mathscr{S}_a$ (i.e., rows with support for ad a). The *post-weighting* standardized bias for X when assigned to a is

$$SB_a^{\hat{\pi}}(X) \;=\; \frac{\left|\bar{X}_a^{\hat{\pi}} - \bar{X}^{\mathrm{elig}(a)}\right|}{\sigma_X^{\mathrm{elig}(a)}},$$

and the worst-case post-weighting imbalance is

$$SB^{\hat{\pi}}(X) \;=\; \max_{a\in\mathscr{A}} SB_a^{\hat{\pi}}(X).$$

If $SB^{\hat{\pi}}(X)$ satisfies the balance threshold for every X, then the inverse-propensity weighting has successfully aligned the covariate distributions across all treatment arms (conditional on the enforced support). Comparing $SB(X)$ (pre-weighting) and $SB^{\hat{\pi}}(X)$ (post-weighting) quantifies the extent to which our weighting procedure mitigates initial selection bias. Figure A.2 reports the worst-case standardized bias $SB(X)$ and its post-weighting counterpart $SB^{\hat{\pi}}(X)$ for each covariate, as defined.

Before weighting (blue), several covariates, most notably HOUR, exhibit large imbalances, with $SB(X) > 0.20$, indicating substantial divergence between the covariate distribution for impressions assigned to certain ads and that of the eligible population. After inverse-propensity weighting (orange), all covariates satisfy the balance criterion, i.e., $SB^{\hat{\pi}}(X) < 0.20$. This demonstrates that reweighting has effectively aligned the covariate distributions across treatment arms, substantially mitigating the selection bias present in the unweighted

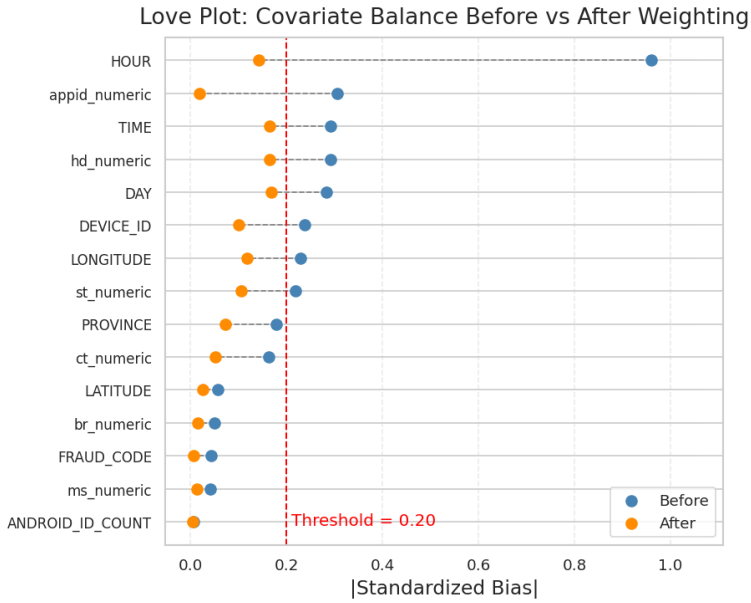


Figure A.2: Standardized bias for each covariate before and after inverse-propensity weighting. The red dashed line marks the 0.20 balance threshold from McCaffrey et al. (2013).

data. The improvement is consistent across both high-variance contextual features (e.g., PROVINCE, LONGITUDE) and stable identifiers (e.g., DEVICE_ID), providing empirical support for the overlap assumption in our setting.

C Additional Empirical Results

C.1 XGBoost vs. LSTM in Predictive Performance

To benchmark our models against approaches used in the literature, we implement an XGBoost model following the specifications in Rafieian and Yoganarasimhan (2021). This allows us to directly compare the predictive performance of our LSTM model with a widely-used tree-based method in ad-targeting research.

For both behavioral (X^B) and combined (X^{GB}) feature sets, we train the XGBoost model using the same training/test splits, hyperparameter tuning protocol, and feature preprocessing pipeline applied to our LSTM models. Hyperparameters are selected via grid search to maximize AUC on the validation set, and class weights are adjusted to account for click sparsity. Table A2 reports performance metrics for the two methods.

Model	Log Loss	AUC	Relative Information Gain
XGBoost (X^B)	0.0552	0.8235	37.64%
XGBoost (X^{GB})	0.0549	0.8333	37.98%
LSTM (X^B)	0.0140	0.8088	84.18%
LSTM (X^{GB})	0.0138	0.8115	84.41%

Table A2: Comparison of model performance for behavioral (X^B) and combined (X^{GB}) features using XGBoost and LSTM model. RIG computed as baseline CTR of test set is $p = 0.017592$.

The results reveal three key patterns. First, consistent with our earlier findings, behavioral features provide the dominant source of predictive power. Second, adding geographical information to behavioral features yields only marginal improvements, confirming that behavioral data capture most of the variation in click propensity. Third, the LSTM model substantially outperforms XGBoost in log loss and RIG for the same feature set, despite almost similar AUC values. This indicates that the LSTM's ability to model temporal dependencies in behavioral sequences translates into better probabilistic calibration, even if its rank-ordering performance (AUC) is comparable to the tree-based benchmark.

It is important to emphasize that high predictive performance does not necessarily translate into higher decision value in targeting. A model may accurately rank or calibrate predicted click probabilities but still yield limited improvements in actual campaign outcomes under realistic budget and exposure constraints. In the next section, we therefore evaluate these models in terms of their realized *decision value* as measured by the targeting value function.

C.2 XGBoost vs. LSTM in Value of Information Performance

To complement the predictive performance comparison in Appendix §C.1, we benchmark the *decision value* delivered by LSTM and XGBoost models. Decision value is measured using the IPS-based targeting value function, which quantifies the incremental CTR lift relative to the logging-policy baseline under realistic targeting constraints. This analysis addresses the fact that models with similar predictive accuracy can differ in their ability to generate value when deployed in practice.

For both behavioral (X^B) and combined (X^{GB}) feature sets, we use the same training/test partitions and preprocessing as in the predictive exercise, ensuring that differences in decision value are attributable to the learning algorithm rather than data handling. IPS estimates are computed following the same procedure described in §5.3, and standard errors are clustered at the user level. Table A3 reports the estimated policy values, confidence intervals, t-statistics, standard errors, percentage lifts over the baseline CTR of $p = 0.017592$, and effective sample sizes.

Table A3: Estimated Policy Value Using IPS

Policy	IPS Estimate	95% CI	t-stat	SE	Lift (%)	ESS
$\hat{V}_{X^{GB}}$ LSTM	0.024883^{***}	[0.024366, 0.025401]	27.611	0.000264	41.45%	760316
$\hat{V}_{X^{GB}}$ XGBoost	0.023531^{***}	[0.023056, 0.024006]	24.521	0.000242	33.76%	803214
$\hat{V}_{X^{B}}$ LSTM	0.022918^{***}	[0.022387, 0.023448]	19.675	0.000271	30.28%	555871
$\hat{V}_{X^{B}}$ XGBoost	0.022738^{***}	[0.022261, 0.023214]	21.183	0.000243	29.25%	735365

Notes: All estimates are computed using IPS. The baseline CTR under the logging policy is 0.017592. Lift is computed relative to this baseline. SE denotes the standard error of the estimate. Effective sample size (ESS) measures the number of equally weighted observations that would yield equivalent precision. Number of observations: 3,162,376. Cluster-robust standard errors are computed using 141,595 user clusters. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

The results yield two main insights. First, the LSTM model consistently achieves higher IPS estimates than XGBoost for the same feature set. This gap is particularly pronounced for the combined feature set, where LSTM delivers a 41.45% lift versus 33.76% for XGBoost. These differences align with the log-loss and RIG advantages observed in the predictive exercise, suggesting that the LSTM's temporal-sequence

modeling translates into superior decision value in deployment. Notably, while XGBoost achieves comparable, and in the case of AUC, slightly higher rank-ordering performance in the predictive task, these apparent advantages do not translate into higher realized value. This underscores that strong predictive metrics do not guarantee superior targeting outcomes. Second, the divergence between predictive and decision-value rankings highlights the importance of evaluating targeting algorithms using realized performance metrics, not solely predictive scores. While AUC and calibration remain useful for model assessment, the ultimate measure of a targeting model's utility is the incremental value it generates in live allocation settings.

C.3 Statistical Test of Complementarity vs. Substitutability

Here we present the detailed statistical results underlying Figure 7. The figure shows mean incremental policy lifts across impression depth, while Table A4 reports the corresponding formal test of whether geography complements or substitutes behavioral data at different stages of exposure. The test logic follows a difference-in-differences (DoD) design. Specifically, we define

$$\hat{\Delta} = (V^{GB} - V^{B}) - (V^{G} - V^{\varnothing}),$$

where V^{GB} is the value of the combined geography + behavior policy, V^{B} is behavior only, V^{G} is geography + context, and $V^{\varnothing}$ is context only. Intuitively, $\hat{\Delta}$ compares the incremental contribution of geography when behavioral information is present versus when it is absent. A positive $\hat{\Delta}$ indicates that geography acts as a *complement* to behavior (its marginal value is higher once behavior is available), while a negative $\hat{\Delta}$ implies *substitutability* (its marginal value declines once behavior is available).

Inference is based on cluster-robust standard errors at the user level. This adjustment ensures that repeated impressions from the same individual do not bias the test statistics. For each impression-depth bin, the table reports the point estimate $\hat{\Delta}$, its 95% confidence interval, the t-statistic, a two-sided p-value, and one-sided bootstrap probabilities $p(\Delta > 0)$ and $p(\Delta < 0)$. The final two columns summarize the decision rule (Complement, Substitute, or Inconclusive) and the corresponding significance tier.

The results confirm the visual evidence from Figure 7: during early exposures (up to roughly 20–25 impressions), $\hat{\Delta}$ is positive and statistically significant, indicating that geographical data provide complementary value when behavioral histories are sparse. Beyond 50 impressions, $\hat{\Delta}$ becomes negative and significant, showing that geography turns into a substitute once behavioral data are sufficiently rich. Mid-range bins (around 30–40 impressions) produce inconclusive or transitional estimates. Together, the figure and table establish a dynamic pattern: geography is initially complementary, but ultimately substitutable as behavioral histories accumulate.

Table A4: Complement vs. Substitute by Impression Depth (Clustered DoD Test)

Last-Impr	$\hat{\Delta}$	95% CI	t-stat	p (two)	$p(\Delta > 0)$	$p(\Delta < 0)$	Decision	Signif.
3	0.0006	[–0.0017, 0.0029]	0.50	0.616	0.308	0.692	Inconclusive	none
6	0.0033	[0.0008, 0.0057]	2.63	0.009	0.004	0.996	Complement	strong
10	0.0023	[–0.0001, 0.0046]	1.86	0.063	0.032	0.968	Complement	mod
16	0.0024	[0.0003, 0.0045]	2.23	0.026	0.013	0.987	Complement	mod
24	0.0033	[0.0014, 0.0052]	3.37	0.001	0.000	1.000	Complement	strong
35	–0.0002	[–0.0021, 0.0017]	–0.20	0.839	0.581	0.419	Inconclusive	none
50	–0.0023	[–0.0039, –0.0006]	–2.69	0.007	0.996	0.004	Substitute	strong
71	–0.0019	[–0.0034, –0.0004]	–2.49	0.013	0.994	0.006	Substitute	strong
102	–0.0010	[–0.0024, 0.0003]	–1.54	0.123	0.938	0.062	Substitute	weak
150	–0.0010	[–0.0024, 0.0005]	–1.31	0.189	0.905	0.095	Substitute	weak

Notes: $\hat{\Delta} = (V^{GB} - V^{B}) - (V^{G} - V^{\varnothing})$ reports the difference-in-differences estimate of complementarity versus substitutability. t-statistics and confidence intervals are based on cluster-robust standard errors at the user level. $p(\Delta > 0)$ and $p(\Delta < 0)$ are one-sided bootstrap probabilities.