Marketing Science Institute Working Paper Series 2025

Report No. 25-148

Modeling Aggregate Consumer Journeys in a Cookie-Free Environment: A Two-Stage Framework for Marketing Mix Models

Victor Churchill, H. Alice Li and Dongbin Xiu

Modeling Aggregate Consumer Journeys in a Cookie-Free Environment: A Two-Stage Framework for Marketing Mix Models

by

Victor Churchill
Trinity College
victor.churchill@trincoll.edu

H. Alice Li
The Ohio State University
li.815@osu.edu

Dongbin Xiu
The Ohio State University
xiu.16@osu.edu

October 2, 2025

Marketing Science Institute Working Paper Series

Modeling Aggregate Consumer Journeys in a Cookie-Free Environment: A Two-Stage Framework for Marketing Mix Models

Abstract

This paper addresses the challenge of measuring marketing effectiveness when only aggregate data are available, a growing concern in an era of heightened privacy regulations and declining access to individual-level tracking. We propose a two-stage marketing mix modeling (MMM) framework that combines modern machine learning with classical econometric approaches. In the first stage, the framework augments classical MMM with a machine-learning classifier to capture directional changes in sales, followed by a calibration stage that maps these signals into continuous sales forecasts. This design extracts robust patterns from noisy, aggregate time-series data while retaining flexibility across modeling contexts. Applied to a proprietary dataset of monthly sales and counts of marketing touchpoints at the channel level from a global software firm, the framework outperforms conventional baselines such as linear regression and Gaussian process regression, and also outperforms ensemble methods like AdaBoost and neural network baselines, while remaining robust across data-constrained settings. We further illustrate the interpretability of the framework by applying Shapley value analysis, which highlights the touchpoints that consistently contribute to sales while filtering out redundant activity. In summary, our framework can extend the MMM paradigm, helping firms evaluate campaign effectiveness and allocate budgets with confidence in a cookie-free, aggregate data environment.

Keywords: Marketing mix model (MMM), Machine learning, Aggregate data, Ensemble methods, Privacy-first analytics, Campaign effectiveness.

1 Introduction

Marketing mix modeling (MMM) has long provided firms and researchers with tools to evaluate the effectiveness of marketing activities using aggregate data. Rooted in time-series econometrics, classic MMM studies often employ regression frameworks to relate advertising expenditures, promotions, and other marketing inputs to sales outcomes. Early contributions by Hanssens et al. (2001) established the econometric tradition of marketing mix modeling, highlighting how advertising and other marketing activities affect sales over time. More recent work by Pauwels and colleagues advanced this tradition by developing methods that capture persistence and feedback among marketing variables, often using vector autoregressive (VAR) models (Pauwels et al. 2002, 2004). These models have proven useful for quantifying how shocks to marketing investments propagate over time, but they are typically limited to a small number of aggregate inputs and assume largely linear, stationary relationships.

The environment in which firms operate today presents new challenges for MMM. The marketing ecosystem now involves dozens of channels and hundreds of campaign types, generating high-dimensional data that strains the assumptions of classical econometric models. At the same time, growing privacy restrictions and the deprecation of third-party cookies have curtailed the availability of individual-level journey data. In prior years, attribution models that relied on detailed clickstream tracking, such as those developed by Li and Kannan (2014), Ghose and Todri-Adamopoulos (2016), and Kireyev et al. (2016), offered insights into the contributions of specific touchpoints along the consumer path to purchase. Yet in the current context, attribution modeling is becoming increasingly infeasible: as third-party cookies and cross-site identifiers are blocked by browsers (such as Safari and Firefox) and legislated against (e.g., under GDPR, CCPA), granular user-level sequences lose coverage, making traditional last- or multi-touch attribution models unreliable or unusable going forward. At the same time, firms often only have access to short, noisy periods of aggregate

data, which further complicates measurement and forecasting. As a result, managers must rely on aggregate metrics of impressions, clicks, and conversions within these short horizons, echoing the classical MMM tradition but under far more constrained conditions.

While classical VAR and regression-based MMM methods have offered important insights, their performance can become strained when applied to dozens of correlated touchpoint variables, often requiring simplifying assumptions or aggregation that may come at the cost of managerial detail. Moreover, aggregate data are noisy, and the signal-to-noise ratio can be low when predicting monthly or weekly sales. Recent advances in machine learning (ML) provide a pragmatic way forward. While not conceptually new, ensemble methods such as gradient boosted trees and neural networks can flexibly handle large input spaces and nonlinear interactions, offering predictive accuracy beyond that of traditional econometrics. Their use in MMM is not about novelty for its own sake, but about meeting the practical need for tools that can extract meaningful signals from aggregate yet complex marketing data.

In this paper, we leverage recent advances in ML to address the need for effective marketing measurement using only aggregate data. We propose a two-stage framework that combines flexible machine learning with classical MMM approaches. In the first stage, an adaptable ensemble-based classifier (e.g., boosting methods or random forests) can be selected and tuned by the firm to fit its specific data context and predict the directional change in sales. In the second stage, a classical model (such as regression, VAR, or other time-series techniques) calibrates these directional signals into continuous sales forecasts. Designed to accommodate a broad class of established models, the framework extracts meaningful patterns from noisy, aggregated time-series data while remaining adaptable to diverse empirical settings.

Applied to a proprietary dataset of monthly marketing touchpoints and sales from a global software firm, the two-stage MMM framework demonstrates performance gains relative to traditional econometric baselines. In our implementation, the first stage uses XGBoost to

predict sales direction, followed by a regression calibration in the second stage. While the model is presented for illustrative purposes and does not exhaust the many alternative specifications available in either stage, it nonetheless outperforms standard benchmarks such as linear regression and Gaussian process regression, and even outperforms some ensemble and neural network baselines. In addition to predictive accuracy, interpretability remains important for managerial use. We incorporate Shapley value analysis (Shapley 1953, Lundberg and Lee 2017) as an illustrative diagnostic tool, allowing us to decompose predictions into channel-level contributions. While not intended to provide precise elasticities, this approach highlights which touchpoints consistently influence outcomes and makes the model's reasoning more transparent.

Overall, our work aims to extend the marketing mix modeling paradigm into a new era, where individual-level tracking is no longer feasible. Our findings highlight how a directional-first approach can provide managers with reliable guidance on campaign effectiveness and budget allocation in a cookie-free environment. In doing so, the framework helps bridge the gap created by the decline of granular consumer data and demonstrates how marketing analytics can remain informative even under increasing data constraints.

2 Literature Review

2.1 Marketing Mix Modeling with Aggregate Data

MMM has long served as a primary tool for linking marketing inputs to sales outcomes at the aggregate level. A large body of research has demonstrated the dynamic effects of advertising and promotions on both short- and long-term outcomes. For example, Hanssens et al. (2001) documented persistence, lagged response, and equilibrium adjustment in sales and brand performance. Neslin and Shoemaker (1989) analyzed the effects of sales promotions, showing both immediate lift and carryover implications. Mela et al. (1997) highlighted how advertising and promotions jointly influence brand equity over time. Pauwels et al. (2002)

emphasized carryover and competitive dynamics, illustrating how marketing investments affect consumer incidence, brand choice, and purchase quantity.

The expansion of digital channels has further increased the complexity of MMM. Academic work has responded with more flexible estimation strategies, including shrinkage and regularization, to improve stability when many channels are observed simultaneously. Field experiments also provide causal benchmarks for advertising effects (Hoban and Bucklin 2015, Gordon et al. 2019), offering important validation for observational models.

At the same time, industry has provided large-scale MMM platforms tailored for privacy-compliant settings, such as Google's Meridian and Meta's Robyn, which are accessible to practitioners. These developments reflect the growing emphasis on aggregate, privacy-compliant measurement frameworks. In practice, however, many managers lack the long and stable time series that traditional MMM requires to estimate dynamic effects. Campaign cycles, reporting changes, and privacy restrictions often leave managers with only short and noisy aggregate windows, further complicating estimation. This limitation highlights the importance of developing methods that remain informative even when only short aggregate windows are available.

2.2 Attribution Modeling

Another stream of research focuses on multichannel attribution, which studies how sequences of touchpoints contribute to conversion. Early models include logistic regression and hidden Markov models to capture transitions along the funnel (Montgomery et al. 2004, Li and Kannan 2014), as well as approaches that account for cross-channel dynamics (Kireyev et al. 2016). These studies underscore the importance of order, timing, and interaction effects in understanding the path to purchase.

More recent work brings machine learning methods into attribution modeling. Topic models and related probabilistic approaches have been used to uncover latent themes in consumer search and clickstream behavior (Trusov et al. 2016, Li and Ma 2020). Poisson

factorization has been applied to interpret search queries and click-throughs at scale (Liu et al. 2021). Beyond probabilistic methods, newer implementations adapt flexible algorithms such as neural networks (Churchill et al. 2024) and transformers (Lu and Kannan 2025) to attribution settings, aiming to extract predictive signals from increasingly fragmented or high dimensional data (Ma and Sun 2020). These advances extend attribution modeling beyond rule-based heuristics and classical probability models, providing richer ways to evaluate channel effectiveness under privacy and data sparsity constraints.

2.3 Interpretability and Model Diagnostics

As marketing analytics increasingly incorporate machine learning methods, interpretability has become an important complement to predictive accuracy. Managers not only need reliable forecasts but also transparency into which channels or activities drive results. A growing set of diagnostic tools provides this visibility, ranging from impulse response analyses in time-series models to marginal effects in regression-based models.

More recently, approaches based on Shapley values (Shapley 1953, Lundberg and Lee 2017) have gained popularity because they allocate predictive contributions to individual inputs in a consistent and theoretically grounded manner. In marketing, attribution models, such as Li and Kannan (2014), highlight the value of decomposing multichannel effects, and manually calculate Shapley values when estimating the marginal impact of individual channels within a multichannel setting. More recent studies extend this line by applying Shapley methods directly. For example, Churchill et al. (2024) leverage the interpretable image features and Shapley plots to visualize how various digital touchpoints are associated with sales, demonstrating the flexibility of Shapley analysis for unpacking complex models and communicating insights to both academic and managerial audiences.

For contexts where user-level data are unavailable or privacy constraints require aggregation, interpretability becomes especially critical. Shapley analysis provides a way to illustrate which touchpoints are meaningful in a given data window, helping firms distinguish recurring

signals from fluctuations. In this sense, interpretability tools connect advances in predictive modeling with the managerial need for transparent and actionable guidance.

Building on these streams, our study proposes an aggregate-data, privacy-aligned two-stage approach. By focusing first on directional prediction through pairwise ranking and then connecting these rankings to sales levels, the framework addresses the challenge of forecasting absolute outcomes in short aggregate windows while still producing actionable measures. The inclusion of Shapley analysis offers a qualitative and illustrative view of channel contributions, complementing the predictive core and situating our work at the intersection of MMM, attribution modeling, and interpretable machine learning.

3 Data

We analyze a proprietary dataset from a U.S.-based multinational software service provider that records monthly marketing activity and sales from 2018 to 2021. The firm markets a portfolio of more than 20 subscription-based software applications through diverse digital channels. In total, the dataset covers 20,556 unique users, of whom 2,425 (11.8%) subscribed at least once during the observation window. Each month, we observe the counts of 31 distinct marketing touchpoints, including display impressions and clicks, email sends, opens, and clicks, as well as organic and paid search interactions. These counts are aggregated across all users, and the corresponding number of sales in the following month is recorded[1]. Thus, the data is utilized as input/output pairs:

$$\{\mathbf{x}(t), y(t+1)\}_{t=1}^{N(T)-1}, \tag{1}$$

[1]Many companies make marketing decisions on a monthly basis, rather than weekly or daily, which motivates our use of months as the time scale here. Finer granularity (daily or weekly) offers more observations but is meanwhile highly volatile, whereas coarser granularity, e.g., multi-month, smoothes volatility but leaves too few observations for generalizable estimation. Nonetheless, the technique presented is malleable and can accommodate shorter or longer aggregation periods as needed.

where $N(T)$ is the total number of months in the data period, $\mathbf{x}(t) \in \mathbb{R}^{31}$ are touchpoint counts in month t, and $y(t+1) \in \mathbb{R}$ is the sales count in month $t+1$.

We adopt this lagged specification because marketing actions in the current month are designed to influence future outcomes, not contemporaneous sales. Additionally, it mitigates simultaneity concerns given that sales and touchpoints often move together within the same month. Meanwhile, in our exhaustive empirical testing, the previous month's touchpoint counts were a superior predictor of the current month's sales than the current month's touchpoints, where supplementing the previous month's touchpoints with those of the current month did not improve predictive accuracy.

Although the raw dataset spans more than three years, the counts of each type of touchpoint are heavily imbalanced. For example, one category of touchpoints remained at several thousand per month for a time, then dropped to zero for many months before spiking again. It is unclear whether such zeros represent a genuine absence of marketing activity due to strategic reallocation, or instead arise from missing data caused by vendor changes or reporting issues. Including these irregular months would confound interpretation and risk attributing results to data artifacts rather than true marketing effects. This degree of imbalance also reflects the difficulty of drawing reliable inferences from the full series, and later motivates our focus on two shorter continuous periods of consistent coverage.

With increasing restrictions on individual-level tracking, managers often lack access to detailed clickstream or journey data. In addition, certain marketing channels, such as mass media and sponsorships, report only at the aggregate level. Aggregate counts, therefore, remain one of the few feasible data sources for evaluating marketing effectiveness. By working with short windows of aggregated monthly series, our study provides a framework for extracting predictive insights when only limited aggregate-level data are available, offering practical solutions to the more constrained data collection opportunities.

The two continuous windows for study are Period 1 (13 months) and Period 2 (10 months). Table 1 reports the means and standard deviations of monthly touchpoint counts across these

two analysis windows. The allocation of touchpoints differs noticeably between Periods 1 and 2. Some channels were entirely inactive in one period but present in the other. For example, Display Impression Type 4, Display Click Types 2 and 4, and Paid Affiliate Click appear only in Period 2, while Paid Social Impression and Paid Social Click are present only in Period 1. Even when touchpoints are active in both periods, their intensity varies substantially.

Figures 1 and 2 show how the differences in marketing allocations across periods translate into different sales trajectories. In Figure 1, the sales in Period 1 fluctuate with more frequent ups and downs, but overall within a relatively narrow band. In contrast, Period 2 in Figure 2 shows greater volatility.

Although the duration of the two periods is different, within each period, we use the first five months for training and the remaining months for testing. This design achieves two goals: it eliminates ambiguity created by scattered missing values, and it provides a realistic test of whether meaningful predictive signals can be extracted from short period of aggregate marketing data.

With this dataset, we would like to investigate how aggregate data, rather than user-level tracking, can inform marketing decisions when individual journeys cannot be observed. Our purpose is not to claim definitively that aggregate data are sufficient for prediction, but rather to illustrate the possibilities and limitations. In doing so, we aim to initiate a discussion on what marketing analytics looks like when the field shifts back to aggregate-level data after decades of relying on detailed user-level records, and how more recent methods, such as neural networks and boosting, may or may not help in this transition.

4 Model

Given the data described in Section 3, our objective is to learn a regression model $\mathbf{R} : \mathbb{R}^{31} \rightarrow \mathbb{R}^{+}$ that accurately maps monthly touchpoint counts to subsequent sales, i.e., $\mathbf{R}(\mathbf{x}(t)) \approx$

Table 1: Means (μ) and Standard Deviations (σ) of Touchpoint Counts in Periods 1 and 2

Touchpoint Type	μ_1	σ_1	μ_2	σ_2
Display Impression Type 1	34151	20195	3359.2	2542.2
Display Impression Type 2	109.7	195.77	10523	5410.5
Display Impression Type 3	87430	73361	325912	64341
Display Impression Type 4	0	0	10820	8353.1
Display Impression Type 5	2434	1555.6	17553	5236.4
Display Click Type 1	22.5	9.0093	30.923	20.694
Display Click Type 2	0	0	10.692	4.3471
Display Click Type 3	14.8	12.173	75.308	18.154
Display Click Type 4	0	0	14.154	7.1513
Display Click Type 5	1.3	1.567	5.2308	3.7893
Email Sent Type 1	46040	12740	73200	16288
Email Sent Type 2	1904.6	2245.9	2993	1854.3
Email Sent Type 3	28112	12708	22334	4065
Email Sent Type 4	8330.2	2574.3	6441.9	2370.8
Email Open Type 1	38864	8486.2	51336	11264
Email Open Type 2	1408.9	1703	1239.8	967.05
Email Open Type 3	28806	10382	17892	3482.7
Email Open Type 4	6561.8	2325.8	3864.9	1171.3
Email Click Type 1	2993.4	866.47	3073.8	1020.1
Email Click Type 2	57.5	95.328	40.154	30.315
Email Click Type 3	2147.3	646.92	1001.3	301.66
Email Click Type 4	395.1	178.52	143	67.012
Paid Search Click Type 1	857.8	259.66	987.92	188.41
Paid Search Click Type 2	428.2	249.81	196.38	47.1
Paid Search Click Type 3	316	30.203	656.15	82.248
Paid Search Click Type 4	123	45.736	79.923	20.87
Paid Social Impression	3001.7	3792.8	0	0
Paid Social Click	7.8	7.2234	0	0
Owned Social Click	145	42.716	101.31	54.116
Earned Social Click	466.2	91.576	209.46	72.294
Paid Affiliate Click	0	0	142.62	126.49

Figure 1: Month-by-month sales counts for Period 1 (blue).

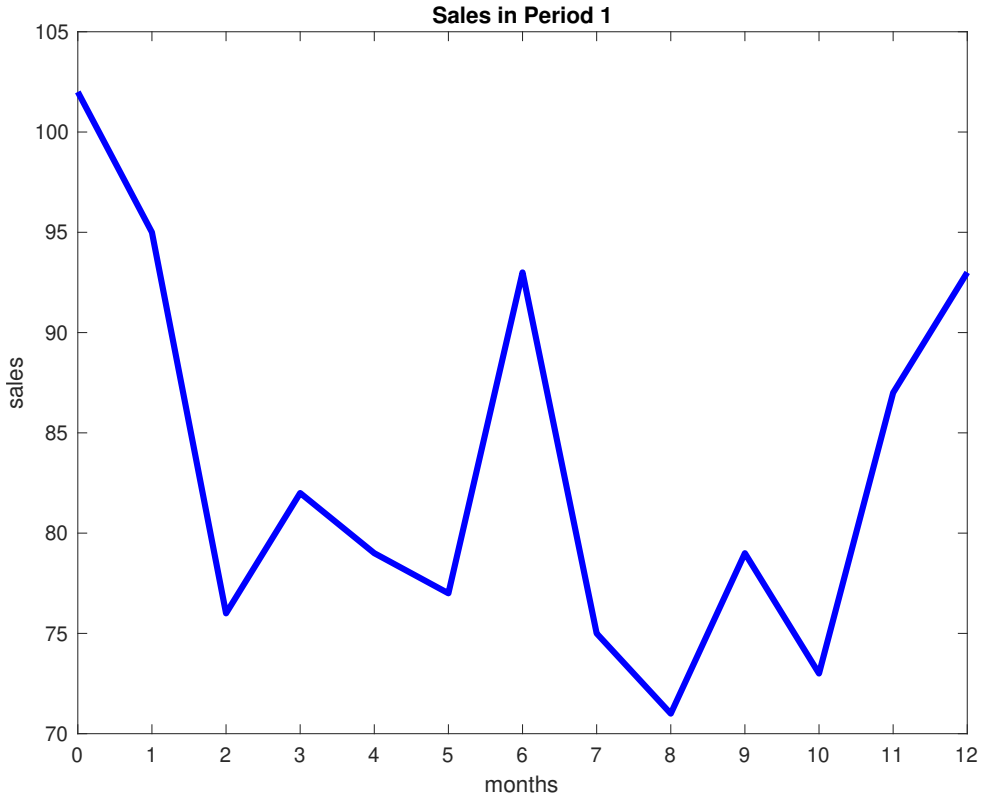
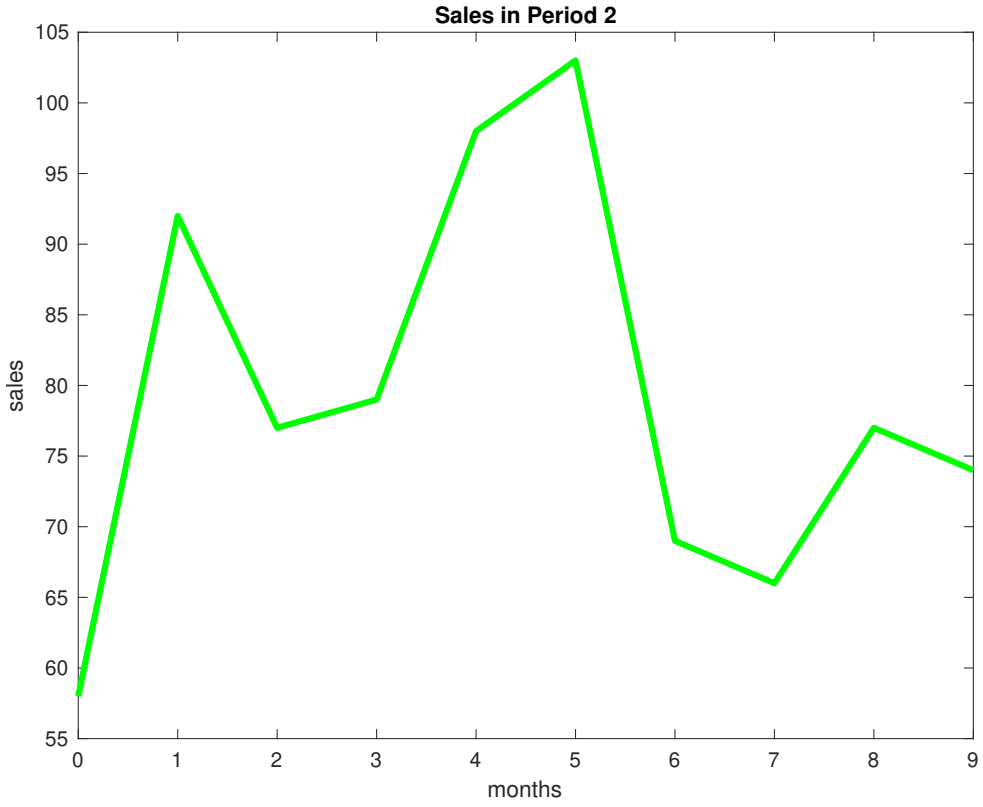


Figure 2: Month-by-month sales counts for Period 2 (green).



$y(t+1)$ for all months t in the data period. For both managerial and empirical reasons, as detailed in Section 3, we consider the lagged specification the most appropriate to capture the effect of marketing efforts on sales in this study.

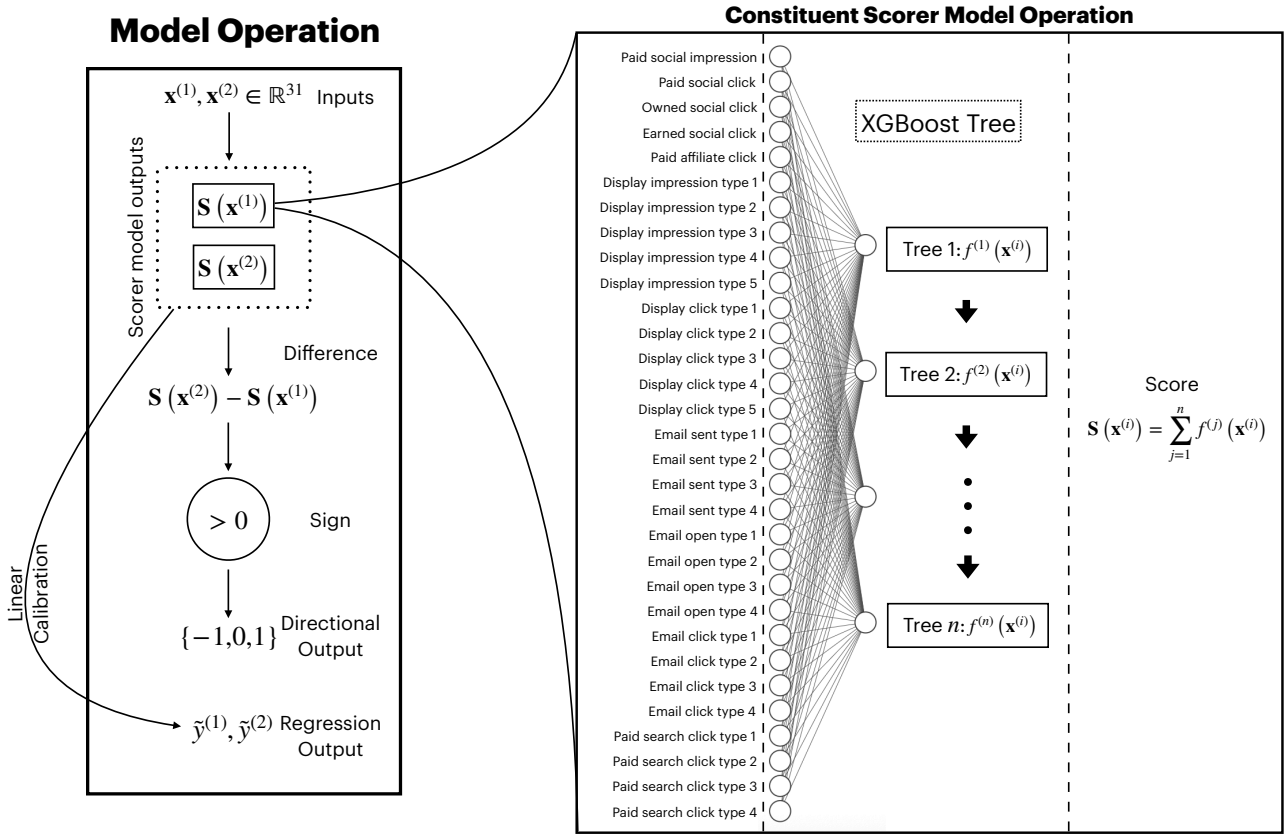
Directly predicting exact sales levels is challenging with short, aggregated, and noisy data. As we will later demonstrate, single-stage methods, such as classical regressions or more flexible machine learning models, struggle to capture stable relationships in this setting. To address this challenge, we formulate the prediction task in two stages. First, we focus on an intermediate goal of predicting sales directionality, i.e., whether next month's sales will be higher or lower than this month's, before calibrating those predictions into actual sales. This step is particularly useful in a cookie-free environment, where marketing managers can use our model to assess whether their campaigns are moving performance in the right direction. Second, the directional accuracy model is linearly calibrated to predict sales, to recover sales forecasts while preserving directional accuracy.

This hybrid directionality-regression model is illustrated in Figure 3 and detailed in the rest of this section. More specifically, Section 4.1 describes the model configuration and Section 4.2 describes the training procedure.

4.1 Model Configuration

The foundation of the two-step model is a pairwise ranking model that predicts sales directionality from touchpoint counts. Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ and $y^{(1)}, y^{(2)}$ denote two arbitrary input/output pairs, where $\mathbf{x}^{(i)} \in \mathbb{R}^{31}$ are touchpoint counts in month i and $\mathbf{y}^{(i)} \in \mathbb{R}^{+}$ are the corresponding sales in month $i+1$. The model compares these two inputs and predicts whether sales are expected to rise, fall, or remain the same. Formally, we model the

Figure 3: Diagram of the Proposed Hybrid Directionality-Regression Model



directionality with a function $\mathbf{D} : \mathbb{R}^{31\times 2} \to \{-1, 0, 1\}$ given by

$$\mathbf{D}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \operatorname{sgn}\left(\mathbf{S}\left(\mathbf{x}^{(2)}\right) - \mathbf{S}\left(\mathbf{x}^{(1)}\right)\right) \approx \begin{cases} -1 & \text{if } y^{(2)} < y^{(1)} \\ 0 & \text{if } y^{(2)} = y^{(1)} \\ 1 & \text{if } y^{(2)} > y^{(1)} \end{cases} \tag{2}$$

where $\mathbf{S} : \mathbb{R}^{31} \to \mathbb{R}$ is a constituent scorer model that assigns a real-valued score to each input. Given real-valued scores for the two inputs being compared, the sign of their difference indicates whether sales following $\mathbf{x}^{(2)}$ are predicted to be lower, the same, or higher than those following $\mathbf{x}^{(1)}$.

For the scorer model $\mathbf{S}$, we use gradient boosted trees (GBTs), in particular XGBoost, (Chen and Guestrin 2016), which build a sequence of decision trees to score touchpoint counts. GBTs are well suited for this setting because they are robust with respect to small training data and have deterministic and therefore reproducible training. In general, choices for $\mathbf{S}$ range from linear models to deep neural networks and in between, and ultimately depend on the dataset and application in question.

Additionally, the scorer model $\mathbf{S}$ can also serve as the basis for our ultimate regression goal. The outputs of $\mathbf{S}$ are real-valued and represent whether touchpoint patterns are associated with higher or lower sales in the next month (e.g., if $\mathbf{S}(\mathbf{x}^{(2)}) > \mathbf{S}(\mathbf{x}^{(1)})$ then the model predicts $y^{(2)} > y^{(1)}$). However, these scores from $\mathbf{S}$ are only ordinal and are not directly related to sales magnitudes. Therefore, an additional step must be taken to build such a relationship while maintaining the directionality of the prediction. In this study, we use a linear calibration step, learning scalar coefficients a and b such that

$$y^{(i)} \approx a \cdot \mathbf{S}\left(\mathbf{x}^{(i)}\right) + b \tag{3}$$

for all pairs. Together, the pairwise ranking stage in Equation (2) and the linear calibration in Equation (3) form the hybrid directionality-regression model in Figure 3 and the training procedure is described in Section 4.2.

4.2 Training

Training the hybrid model proceeds in two stages that mirror its configuration. We first estimate the pairwise ranking component to capture sales directionality, and then calibrate the resulting scores to observed sales counts.

The data described in Section 3 is first divided into training and testing sets. Data Period 1 contains 13 input-output pairs, and Period 2 contains 10 pairs. In each of the two data periods, we choose the first 5 input/output pairs to form the training set, with the remainder reserved for testing. Specifically, in Period 1, although 13 months are available, the lagged setup yields 12 usable input–output pairs. With 5 pairs allocated to training, the model is evaluated on the remaining 7 test months. Likewise, Period 2 spans 10 months, yielding 9 input-output pairs - 5 pairs for training and 4 pairs for testing.

These training data elements are then combined into all possible unordered pairs (i.e., input/output pair 1 with input/output pair 2, input/output pair 1 with input/output pair 3, and so on). With 5 training pairs per period, this yields $\binom{5}{2} = 10$ training samples for each period. It is worth reiterating that this is an extremely limited amount of data with which to train a generalizable model.

The directionality scorer model $\mathbf{S}(\cdot)$ is then trained to assign higher scores to inputs associated with higher sales. Formally, this is achieved by minimizing the pairwise logistic loss (Burges et al. 2005):

$$\arg\min_{\Theta} \left\{ \sum_{y^{(i)} > y^{(j)}} \log \left[1 + e^{\mathbf{D}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \Theta)} \right] \right\} \tag{4}$$

over the training set, where Θ represents the parameterization (in this case, tree structure)

of the constituent scorer model $\mathbf{S}$. This pairwise logistic loss function smoothly penalizes the model whenever it ranks items in the wrong order, while rewarding large score differences in the correct direction. The deterministic LambdaMART optimization algorithm (Burges 2010), which is well-suited for ranking tasks, is used to minimize the loss.

Once the directionality model has been trained, we proceed to the calibration stage. Here, the scores produced by Equation (2) are mapped to sales using a linear regression model in Equation (3). Specifically, we minimize:

$$\arg\min_{a,b} \left\{ \sum_{i} \left[y^{(i)} - \left(a \cdot \mathbf{S}(\mathbf{x}^{(i)}) + b \right) \right]^2 \right\} \tag{5}$$

over all training input/output pairs. This step anchors the ordinal information from the score-ranking stage onto the sales scale, ensuring that predictions capture both relative changes and absolute magnitudes.

Together, these two stages produce a hybrid model that leverages the relative predictability of sales directionality while still delivering actionable forecasts in terms of sales levels.

5 Predictive Performance

To demonstrate the enhanced predictive performance of our proposed hybrid directional-regression model, it undergoes comparative analysis against a broad set of established regression models: linear regression, Gaussian process regression, and multilayer perceptron (neural networks). Additionally, it was tested against state-of-the-art ensemble tree-based classifiers such as bagging, random forest, AdaBoost, and gradient boosting. Details of these standard methods can be found in Hastie et al. (2009). The implementation of these competing models was performed according to Pedregosa et al. (2011). Identical training and test data are used to ensure comparability.

Because our framework emphasizes directional accuracy as an intermediate predictive goal, we note that this metric is not commonly reported in the marketing modeling literature.

While analogous measures have been used in forecasting and financial prediction contexts, they have not typically been applied in studies of consumer purchase journeys or marketing mix models. To provide a fair comparison, we compute the directional accuracy for all benchmark regression models in addition to the conventional regression metrics, such as R^2, mean absolute error (MAE), and mean squared error (MSE). This allows us to evaluate both whether models capture the direction of sales movements and whether they accurately predict magnitudes.

Tables 2 and 3 present the accuracy metrics for our proposed model versus the benchmark models in Periods 1 and 2, respectively. Meanwhile, Figures 4 and 5 display sales predictions for our proposed model and the aforementioned benchmark models, using Period 1 and Period 2, respectively.

Table 2: Accuracy Metrics for All Models For Period 1

Period 1	Directional Accuracy	R^2	MAE	MSE
Our Model	**0.857143**	**0.165042**	**6.256816**	**56.77715**
Linear Regression	0.428571	-8.302473	21.159179	632.5682
Gaussian Process	0.000000	-0.494706	8.950000	101.640000
Neural Network	0.142857	-0.473888	8.919137	100.224353
Random Forest	0.714286	-0.502312	9.083750	102.157237
Bagging	0.285714	-2.287629	12.462500	223.558750
AdaBoost	0.000000	-0.702206	9.750000	115.750000
Gradient Boost	0.571429	0.005039	7.130799	67.657340

Note: Bold indicates the best performance in each column.

Table 3: Accuracy Metrics for All Models For Period 2

Period 2	Directional Accuracy	R^2	MAE	MSE
Our Model	**0.75**	-0.725445	15.318771	299.123108
Linear Regression	**0.75**	-16.643235	51.821152	3058.631143
Gaussian Process	0.00	-0.051915	11.880000	182.360000
Neural Network	**0.75**	-43.516697	83.533825	7717.414634
Random Forest	**0.75**	**0.107104**	11.266000	**154.792380**
Bagging	0.25	-0.648419	16.180000	285.770000
AdaBoost	0.25	0.021689	**10.800000**	169.600000
Gradient Boost	0.25	-0.556795	14.722078	269.886024

Note: Bold indicates the best performance in each column.

Beginning with Period 1 in Table 2, our hybrid model achieves the strongest performance across all evaluation criteria. It attains the highest directional accuracy, correctly predicting 6 out of 7 sales transitions. Furthermore, the linear calibration works quite well in pulling values toward the training sales trend in this case, yielding the highest R^2 value (0.165) among all models, as well as the lowest MAE and MSE values. By contrast, most benchmark models fail to capture the underlying dynamics. Their predictions in Figure 4 are largely flat around the sample mean, with the exception of linear regression and bagging, yielding at best $R^2 \approx 0$, despite small fluctuations. In particular, linear regression and bagging generate higher-variance predictions that deviate substantially from observed sales without improving accuracy, resulting in highly negative R^2 values.

These comparisons underscore the value of our two-step design, which not only detects the correct direction of sales changes but also translates those signals into accurate magnitude forecasts when sufficient data are available. In contrast, directly fitting sales levels from aggregate touchpoints with a single-stage model either collapses toward the mean or overfits noise.

In Period 2, since there are only 5 months in the prediction period, there are only 4 "transitions." Our hybrid model achieves a directional accuracy of 3 out of 4, or 0.75, tied for the highest directional accuracy with linear regression, multilayer perceptron (or neural networks), and random forest models. By contrast, other benchmark models, such as Gaussian Process, Bagging, AdaBoost, and Gradient Boosting, perform poorly, with only 0 or 1 correct predictions out of 4.

Despite being the only method that clearly predicts the correct "shape" of the sales trajectory, as shown in Figure 5, the linear calibration is not sufficient here, and our model does not achieve a positive R^2. This outcome reflects a structural mismatch: the slope and intercept estimated from the training months differ significantly from the sales trend in the testing months, limiting the ability of a simple linear mapping to recover magnitudes. In terms of regression outputs, the statistically best comparative model was the random

Figure 4: Sales Predictions for All Models – Period 1

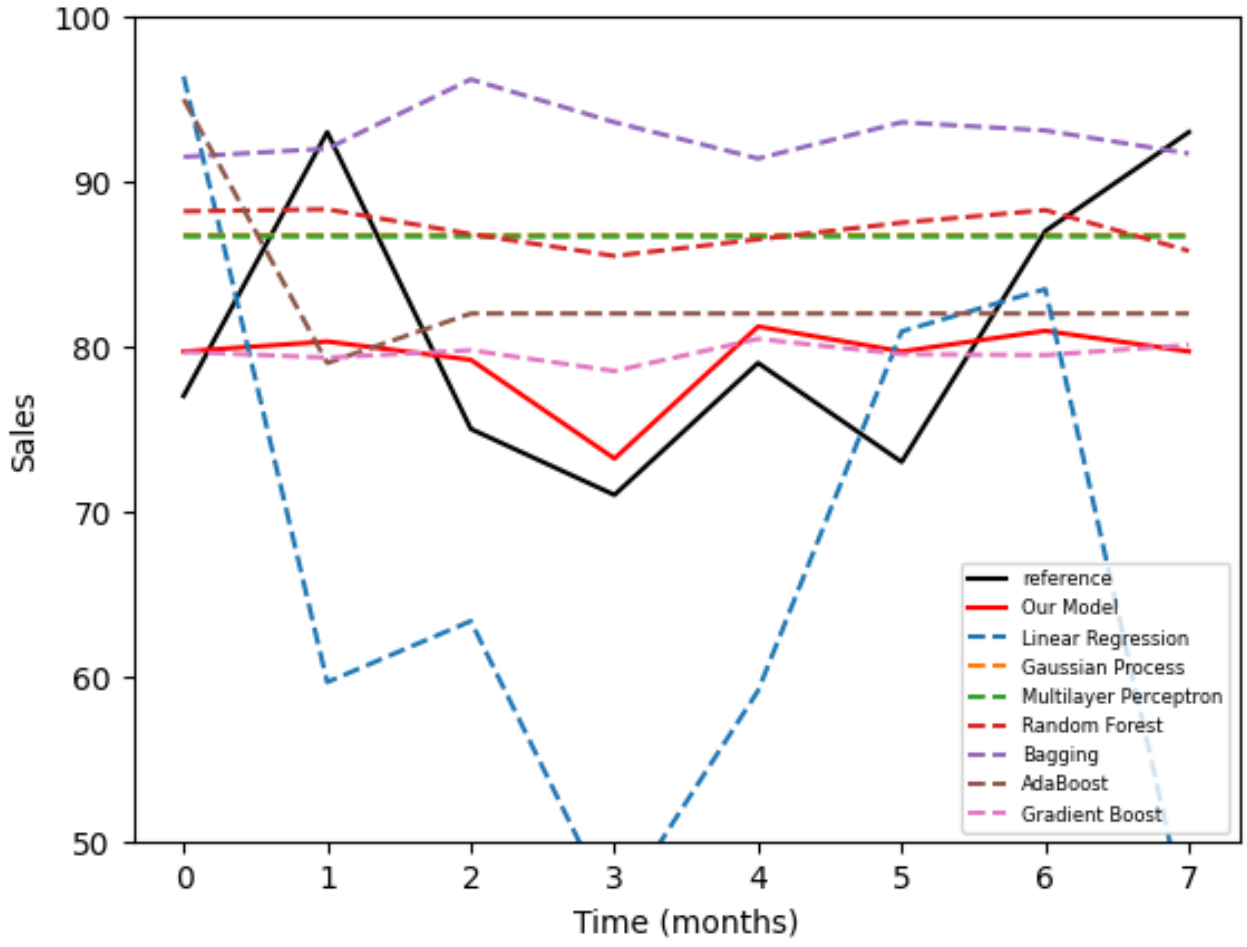
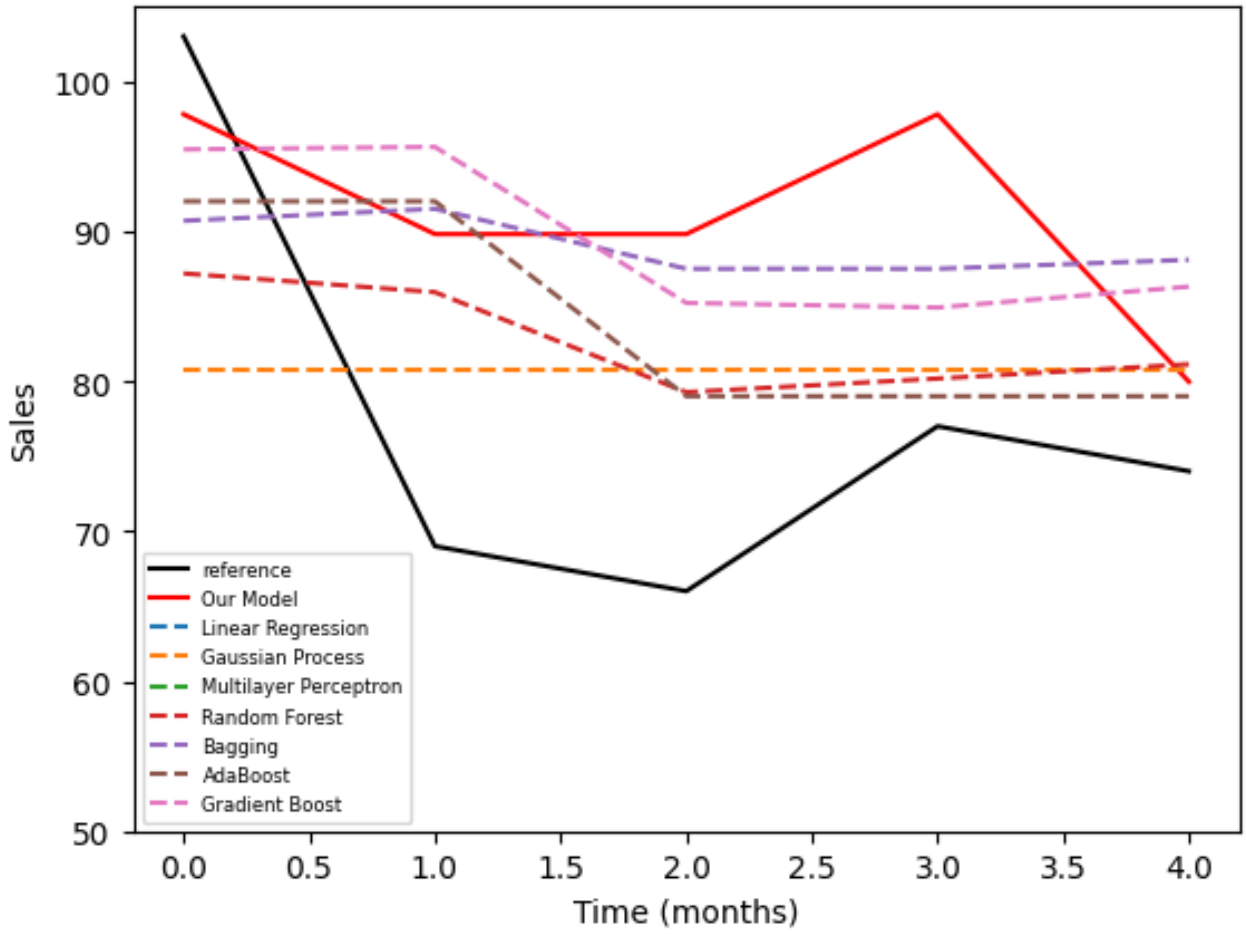


Figure 5: Sales Predictions for All Models – Period 2



forest, which had $R^2 = 0.107$, suggesting alternative scoring or calibration functions may offer advantages in contexts where underlying sales trends shift. Our main takeaway is that while directional signals remain reliable even with very limited data, translating them into precise sales forecasts may require more flexible calibration techniques.

Taken together, the results across the two evaluation periods highlight both the strengths and the boundaries of our approach. When a richer set of training data is available, as in Period 1, the hybrid model clearly outperforms established alternatives in terms of both directionality and regression fit. In shorter and noisier windows such as Period 2, the model continues to provide robust directional signals, but calibration to sales magnitudes becomes more challenging. This is not surprising given that the analysis is conducted on very short periods of aggregate monthly data, where trends may shift between training and testing periods and individual-level variation is unobserved. These findings highlight our central claim: relative changes in sales are more predictable than absolute levels, and a two-step framework that first models directionality provides a robust foundation for marketing prediction in data-constrained, cookie-free environments.

6 Discussion

The linear calibration step of our framework rescales the raw ranking scores into the sales scale, producing estimated coefficients for both slopes and intercepts. However, these coefficients should not be over-interpreted. Since the raw ranking scores have no natural units, the slope and intercept serve only to map relative rankings onto the observed sales levels. For reference, the estimated slopes were 5.994 in Period 1 and 8.339 in Period 2, with corresponding intercepts around 86.9 and 80.6, respectively. These values confirm that our calibration effectively aligns model outputs with sales but do not carry behavioral meaning in the way coefficients from a classical MMM regression might. The interpretive value of our framework lies instead in the directional predictions and the attribution diagnostics provided

by Shapley analysis, as discussed in this section.

Figures 6 and 7 illustrate these diagnostics for Period 1 and Period 2, respectively. Each beeswarm plot shows how the framework attributes predictive contributions across touchpoint types, with each dot representing one month in the testing period. The horizontal position reflects the Shapley value for a given touchpoint type, with points to the right indicating positive contributions to predicted sales and points to the left indicating negative contributions. Touchpoints are ordered vertically by overall importance, so those at the top have the largest average impact across months. The color of each dot corresponds to the relative magnitude of that touchpoint in a given month, with higher counts colored red and lower counts in blue. Interpreting these plots, therefore, involves examining both which touchpoints consistently show influence and whether higher or lower intensities of activity are associated with stronger predicted outcomes.

In our aggregate setting, many touchpoints appear relatively uninformative, with Shapley values clustering near zero. This outcome reflects both the limited data available within each period (only five training months) and the fact that not all touchpoints carry incremental predictive power in every window. By contrast, prior work using disaggregated data (Churchill et al. 2024) shows that Shapley analyses can recover richer and more granular patterns of touchpoint influence when longer histories and user-level variation are available. Our results thus offer a complementary view: in environments where only aggregate data can be used, the model highlights the subset of touchpoints that consistently contribute, while existing disaggregate models can be applied when user-level records are accessible. In either case, the Shapley plots make the filtering process explicit and transparent, offering managers a clear view of which activities emerge as meaningful within the data at hand.

It is not surprising that the two periods yield somewhat different attribution patterns. As Figures 1 and 2 already demonstrate, the firm deployed distinct marketing mixes in these windows, and user populations or market conditions may also have shifted. Consequently, some discrepancies are expected.

Figure 6: Beeswarm Plot of Shapley Values in Period 1

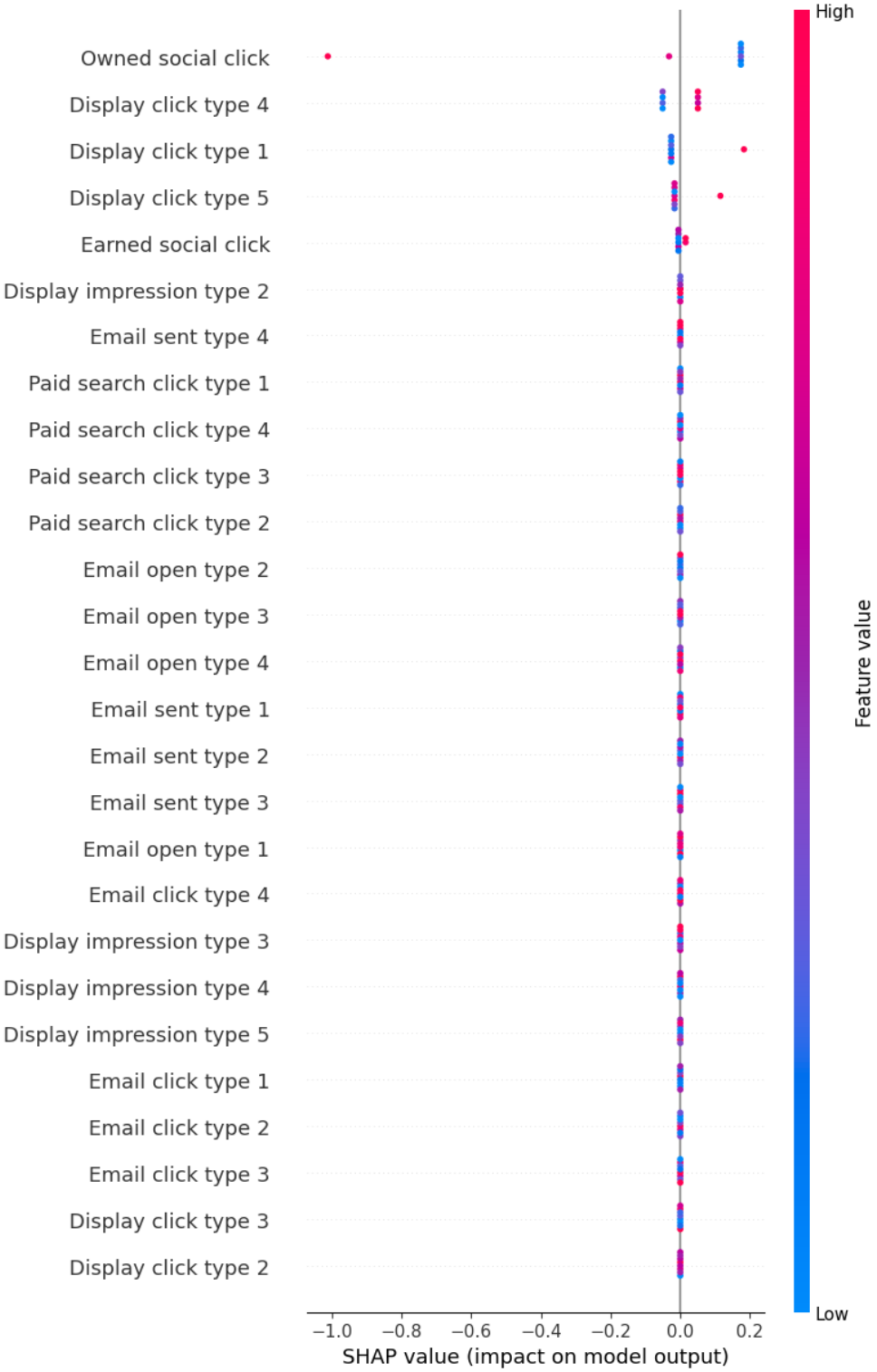
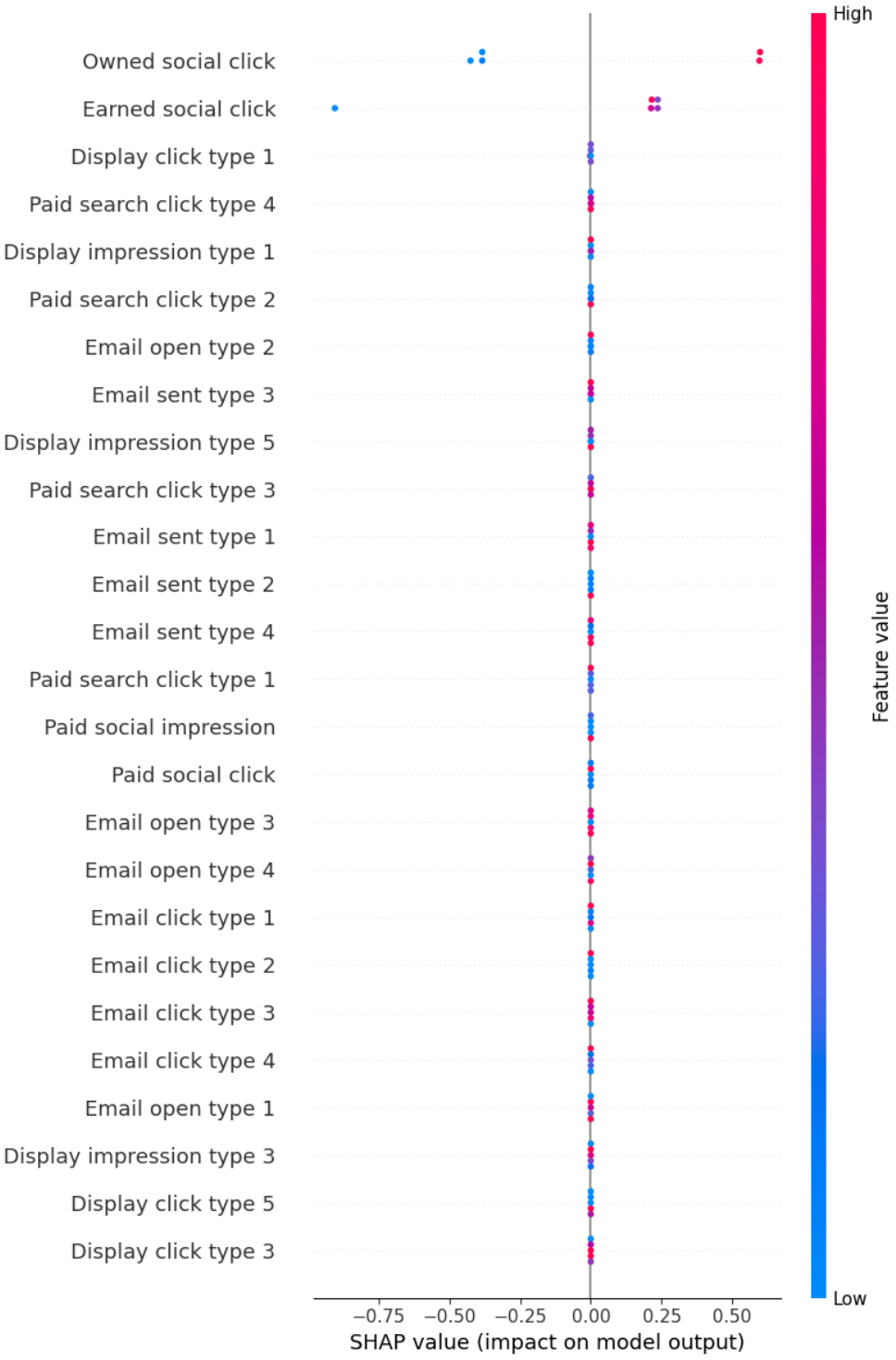


Figure 7: Beeswarm Plot of Shapley Values in Period 2



In terms of vertical order, the top five touchpoints in Period 1 are Owned Social Click, Display Click Types 1, 4, and 5, and Earned Social Click. In period 2, three of these touchpoint types remain among the top 5 most influential. However, Display Click Type 4 is not observed in Period 2, while Display Click Type 5 drops to the bottom of the ranking.

Regarding horizontal distribution, Owned Social Click appears to be negatively associated with sales in Period 1 but positively associated in Period 2. Meanwhile, Display Click Types 1, 4 and 5 are all positively associated with sales in Period 1, but not in Period 2. Earned Social Click remains positive in both periods. However, it also shows different horizontal patterns. In Period 1, its Shapley values are more tightly clustered near zero, while in Period 2, more dots shift to the right, suggesting a stronger positive association with predicted sales in that window. Such shifts should not be interpreted as contradictions, but rather as reflections of changing strategic emphasis and consumer response. In this aggregate, short-horizon setting, Shapley values are best interpreted qualitatively: they highlight recurring directional patterns rather than providing stable elasticities of the kind sometimes extracted from long-horizon MMMs. Importantly, the results remind managers not to over-interpret short-run fluctuations but to focus on signals that recur across campaigns and contexts.

Overall, the Shapley analysis should be seen as illustrative. It shows how the proposed two-stage framework not only enhances directional prediction in noisy, aggregate data but also provides transparency into which touchpoints matter most in a given period. Even when individual signs fluctuate, the model reliably separates signal from noise, offering managers a practical tool for prioritizing channels and avoiding overinvestment in those with little incremental impact.

7 Conclusion and Future Research

This paper develops a two-stage framework for marketing mix modeling under aggregate data constraints, motivated by the deprecation of user-level tracking and the growing reliance on

privacy-first analytics. The framework combines a pairwise ranking stage to capture sales directionality with a calibration step that maps these ordinal signals to the sales scale. By emphasizing directionality first, the model provides managers with a reliable diagnostic of whether marketing activities are moving sales in the right direction, while also yielding forecasts that can inform planning and resource allocation.

The empirical application uses proprietary data from a global software firm, comprising 31 distinct touchpoint types tracked monthly across two continuous periods. This setting reflects the challenges managers face in practice: data are high-dimensional, imbalanced across channels, and available only in aggregate form due to privacy and reporting constraints. The two-stage model is well suited to this environment. In Period 1, it achieved the highest directional accuracy (6 of 7 transitions correctly predicted) and produced the best calibration to sales magnitudes. In Period 2, where sales trends shifted more sharply, the model continued to deliver reliable directional signals, though calibration to absolute levels was more difficult. These findings highlight both the promise and the boundary conditions of the framework: relative changes in sales are more predictable than absolute levels when working with short, noisy aggregate data.

Building on this predictive core, the Shapley analyses illustrate how the framework can be made interpretable. Even with limited training horizons, the plots highlight which touchpoints consistently contribute and which appear redundant, allowing managers to filter through complexity and focus on channels that matter most. In this way, the framework provides not only stronger forecasts but also transparent diagnostics, helping firms better understand the role of different touchpoints in shaping outcomes. Together, these results suggest that the two-stage model offers a flexible measurement tool that adapts to different levels of data availability while remaining predictive and interpretable.

Several avenues remain open for future research. First, alternative calibration functions, such as nonlinear mappings or Bayesian shrinkage, could improve the translation of ranking signals into sales magnitudes, especially when underlying trends shift across training and

testing periods. Second, future work might integrate additional firm-owned (first-party) data sources, which could enhance model performance while remaining consistent with privacy-first principles. Third, researchers could investigate how aggregate-level modeling might be combined with selective user-level experiments to balance privacy protection with predictive precision. Finally, although we focused on monthly aggregation to align with managerial decision cycles, testing the framework at weekly or quarterly resolutions could provide insights into temporal granularity and stability.

Funding and Competing Interests

This research was supported by the Marketing Science Institute (MSI). The views expressed are those of the authors and do not necessarily reflect the positions of MSI or the anonymous firm that provided the data.

References

Burges, Christopher JC. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* **11**(23-581) 81.

Burges, Christopher JC, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, Greg Hullender. 2005. Learning to rank using gradient descent. *Proceedings of the 22nd international conference on Machine learning*. 89–96.

Chen, Tianqi, Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

Churchill, Victor, H Alice Li, Dongbin Xiu. 2024. Unraveling consumer purchase journey using neural network models. *Journal of Machine Learning for Modeling and Computing* **5**(1).

Ghose, Anindya, Vilma Todri-Adamopoulos. 2016. Toward a digital attribution model. *MIS quarterly* **40**(4) 889–910.

Gordon, Brett R, Florian Zettelmeyer, Neha Bhargava, Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science* **38**(2) 193–225.

Hanssens, Dominique M, Leonard J Parsons, Randall L Schultz. 2001. *Market response models: Econometric and time series analysis*. Springer.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer.

Hoban, Paul R, Randolph E Bucklin. 2015. Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research* **52**(3) 375–393.

Kireyev, Pavel, Koen Pauwels, Sunil Gupta. 2016. Do display ads influence search? attribution and dynamics in online advertising. *International Journal of Research in Marketing* **33**(3) 475–490.

Li, Hongshuang, PK Kannan. 2014. Attributing conversions in a multichannel online marketing

environment: An empirical model and a field experiment. *Journal of marketing research* **51**(1) 40–56.

Li, Hongshuang, Liye Ma. 2020. Charting the path to purchase using topic models. *Journal of Marketing Research* **57**(6) 1019–1036.

Liu, Jia, Olivier Toubia, Shawndra Hill. 2021. Content-based model of web search behavior: An application to tv show search. *Management Science* **67**(10) 6378–6398.

Lu, Zipei, P. K. Kannan. 2025. Express: Ai for customer journeys: A transformer approach. *Journal of Marketing Research* **forthcoming**.

Lundberg, Scott M, Su-In Lee. 2017. A unified approach to interpreting model predictions. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4765–4774. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Ma, Liye, Baohong Sun. 2020. Machine learning and ai in marketing–connecting computing power to human insights. *International Journal of Research in Marketing* **37**(3) 481–504.

Mela, Carl F, Sunil Gupta, Donald R Lehmann. 1997. The long-term impact of promotion and advertising on consumer brand choice. *Journal of Marketing research* **34**(2) 248–261.

Montgomery, Alan L, Shibo Li, Kannan Srinivasan, John C Liechty. 2004. Modeling online browsing and path analysis using clickstream data. *Marketing science* **23**(4) 579–595.

Neslin, Scott A, Robert W Shoemaker. 1989. An alternative explanation for lower repeat rates after promotion purchases. *Journal of Marketing Research* **26**(2) 205–213.

Pauwels, Koen, Dominique M Hanssens, Sivaramakrishnan Siddarth. 2002. The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. *Journal of marketing research* **39**(4) 421–439.

Pauwels, Koen, Jorge Silva-Risso, Shuba Srinivasan, Dominique M Hanssens. 2004. New products, sales promotions, and firm value: The case of the automobile industry. *Journal of marketing* **68**(4) 142–156.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pretten-

hofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** 2825–2830.

Shapley, Lloyd S. 1953. *A value for n-person games*. Princeton University Press, 307–317.

Trusov, Michael, Liye Ma, Zainab Jamal. 2016. Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science* **35**(3) 405–426.