Causal impact of digital display ads on advertiser performance

Koen Pauwels
Amazon Ads
Manuele Caddeo
Amazon Ads
German Schnaidt
Amazon Ads

Cite as:
Pauwels Koen, Caddeo Manuele, Schnaidt German (2022), Causal impact of digital display ads on advertiser performance. *Proceedings of the European Marketing Academy*, 51st, (108183)

Paper from the 51st Annual EMAC Conference, Budapest, May 24-27, 2022

Causal impact of digital display ads on advertiser performance

Abstract: Brands are searching for innovative ways to reach customers online. Sponsored Display (SD) by Amazon Ads is a new way to do so, and allows targeting by category, product and audience. However, advertisers are uncertain how much SD improves their performance over different time horizons. This paper studies more than 40,000 brands with two different methods: a diffusion-regression state-space time-series analysis that predicts response counterfactuals during a 20-weeks period post SD adoption of audience targeting, and a newly developed Two-stage Gaussian Process algorithm that generates probabilistically-equivalent twins for causal inference in a shorter time frame of 1 month post SD adoption of category and product targeting. The performance variables include impressions, page views, sales, new-to-brand consumers and Return on Advertising Spend. The results are consistent and quantify how much adding SD to the ad mix increases performance.

Keywords: display advertising, machine learning, causal inference

Track: Digital Marketing and Social Media

1. Objectives and Overview of the Research

Digital marketplaces have continued to grow over recent years, with the Amazon marketplace bringing together over 1.9 million independent suppliers with over 100 million Amazon Prime customers in addition to non-Prime customers (Rangaswamy et al. 2020). Likewise, digital advertising effectiveness and its causal attribution have seen strong research attention. However, advertisers are regularly offered new ways to reach audiences online and are uncertain how much such digital ads can improve their different performance measures over different time horizons.

Display ads are online ads that combine copy and visual elements with a call-to-action message that links to a landing page. Consumers typically see display ads along the top or sides of a website—or sometimes, in the middle of the content they are reading. Recently, Amazon.com introduced Sponsored Display (SD) based on three potential audience strategies: the category the consumer is browsing in, the specific product and the audience remarketing to customers who browsed the product in the past. These three strategies are not mutually exclusive, instead they can be combined to engage with customers who are exploring products within a category, or evaluating a specific product, or re-engage those who have browsed specific products without making a purchase to help ensure missed sales opportunities. Digital advertisers are uncertain to what extent these SD strategies drive performance variables such as impressions, Buy Box detailed page views (i.e., visits to the product pages), sales and new-to-brand consumers (i.e., consumer who have not bought the brand in the previous year on Amazon.com). In particular, they want to know how they can use SD to add to their current portfolio of Sponsored Products (SP), i.e., ads showing individual products to Amazon shoppers in related shopping results and product pages, and Sponsored Brands (SB), ads showcasing the advertiser's brand to Amazon shoppers in related shopping results and product pages (Amazon Learning Console 2021). While SP is widely seen as a bottom-funnel tactic and SB as a mid-funnel tactic, SD is considered an upper-funnel tactic, which may increase the customer base and sales over the longer run, but can decrease efficiency, typically measured as Return on Advertising Spend (Robb 2021).

To measure the performance impact of SD strategies over different time horizons, we use two different, complementary methodologies. First, we estimate a diffusion-regression state-space model that predicts the counterfactual response that would have occurred over a 20-week look forward horizon had the advertiser not adopted audience targeting SD. Second, we apply a 2-stage Gaussian Process algorithm that generates probabilistically-equivalent counterfactuals to evaluate

the impact of SD adoption among advertisers using category and product SD strategies over a 1-month time horizon after enabling SD.

The results are consistent. The counterfactual time series analysis shows that advertisers activating SD with audience targeting for the first-time increased sales by +14% on average within the first 20 weeks after adopting SD as compared to not enabling SD. The 2-stage Gaussian Process shows that brands that began using category targeting in SD strategies for the first time saw, on average, positive impacts across different metrics during the next month after adoption as compared to advertisers that didn't: +33.9% more impressions, a +3.6% increase in Buy Box Detailed Page Views (DPV) and a +2.6% increase in New-To-Brand (NTB) customers. Similarly, Brands that created an SD product targeting campaign for the first time saw, on average, increases of +28.8% in sales, +12.4% in DPV, +3.2% in NTB customers' awareness, and +4.2% in NTB customers' consideration the following month, compared to brands that did not.

2. Methodology

2.1 Diffusion-regression state space model

In contrast to difference-in-differences schemes (Lechner 2011), state-space models allow inferences about the temporal evolutions of attributable impact, and flexibly accommodate multiple sources of variation, including the time-varying influence of contemporaneous covariates, local and linear trends, and seasonality components. In addition, these models can adopt a fully Bayesian nature by incorporating empirical priors on the model parameters which adds extra flexibility and robustness to the analysis. In this context, for the first part of our analysis we applied a Bayesian Structural Time Series Model (Brodersen et al. 2015). We selected 284 advertisers that satisfied the following conditions within a 50-week timeframe: (1) advertisers were active SP and/or SB for at least 30 weeks prior to SD Audience Targeting activation, (2) in the last 20 weeks of the analyzed time period, the only advertising-specific action they took was launching an SD Audience Targeting campaign, and (3) advertisers should be similar in term of business size and ad-campaign activity. Finally, we selected advertisers with ad-support sales higher than 5-th percentile in SD to prevent skewing our results towards those advertisers that are still in a test-and-learn phase for these ads.

For the selected advertisers, we calculated the impact on sales during the 20 weeks following their SD-Audience activation by predicting how their sales would have evolved if the SD-Audience activation had not occurred. We trained our model the first 30 weeks (pre-SD adoption period) to

predict the counterfactual response for the following 20 weeks (post-SD adoption), using 10 additional covariates including sales, units, glance views, etc. at a vertical aggregation level and compares the same covariates for 600 advertisers that never activated SD within the same timeframe. Finally, we measured the lift in sales by subtracting the observed sales (real value) from the counterfactual sales (predicted sales).

2.2 Two-stage Gaussian Process Machine Learning Model

Second, to measure a shorter-term causal impact (i.e., 1-month post-activation) on advertisers who adopted SD product targeting and SD category targeting for the first time, we took inspiration from the causal inference machine learning domain (Alaa and Van der Schaar 2018), and selected 43,720 advertisers in the US marketplace. Our analysis is based on a method recently developed by Amazon Ads Scientists called 2-stage GP (2-stage Gaussian Process) that shows improved performance on various metrics (e.g., placebo test, RMSE, etc.) when applied within the context of advertising as compared to existing state-of-the-art methodologies such as Double Machine Learning (Chernozhukov et al 2018) and Causal Forests (Wager and Athey 2018). This proprietary Amazon Ads algorithm generates adaptive weights that are used to construct counterfactuals for each advertiser that adopted an ad product for the first time (e.g., SD) and then uses the pairs {observed, counterfactual} to estimate the causal impact this intervention. These adaptive weights result from statistical similarities between treatment and control populations spanned by the 50+ features we used to account for confounding (retail and advertising related). For every treated advertiser, the algorithm generates its counterfactual as an adaptive linear combination of the un-treated units.

Our 2-stage GP method builds upon the idea that treatment effects can be modeled as non-linear functions of factors $x_i \in R^d$ (i.e., attributes that in our problem describe the advertiser characteristics), motivated by the flexibility and estimation properties of GPs as seen in different applications in Machine Learning (ML) such as Regression, Smoothing, and Experiments.

2.3 Two-stage GP Algorithm specification

Let $\mathrm{Y_i(x_i)^{(w_i)}}$ represent the target response (e.g., sales) for which we want to measure the impact of a specific action (e.g., SD adoption), conditional on a vector of control variables $\boldsymbol{x_i} \in R^d$. Also, let $w_i \in \{0,1\}$ be an indicator variable for a binary treatment assignment ($w_i = 1$ denotes sample i

has taken the action, whereas $w_i = 0$ denotes the opposite). Suppose the relation between $Y_i(\mathbf{x_i})^{(w_i)}, w_i$ and $\boldsymbol{x_i}$ has the following form:

$$Y_i(x_i)^{(w_i)} = f_0(x_i) + f_1(x_i) \cdot w_i + \epsilon_i, \quad (1)$$

where ϵ_i are random errors satisfying $Cov(\epsilon_i, \epsilon_j | \boldsymbol{x_i}, w_i) = 0, \forall i \neq j, E[\epsilon_i | x_i, w_i] = 0,$ $and\ Var[\epsilon_i | x_i, w_i] = \sigma^2$; $f_0(\cdot), f_1(\cdot)$ are (possibly) non-linear functions that depend on the control variables $\boldsymbol{x_i} \in R^d$. Under strong-ignorability assumptions, we can show that $f_1(\cdot)$ represents the effect of the action as a function of the advertiser characteristics (i.e., the Conditional Average Treatment Effect or CATE). Once we have estimated $\widehat{f_0}(\cdot)$ and $\widehat{f_1}(\cdot)$ from an independent, identically distributed (iid) dataset $\{(Y_i, X_i, W_i)\}_{i=1}^N$, we can obtain the ATE as $\hat{\gamma} = \frac{1}{N}\sum_{i=1}^N \widehat{f_1}(\boldsymbol{x_i})$.

Now we introduce Gaussian Processes (GPs) into the previously described setting. Suppose that for w $\in \{0,1\}$ we can model each function in Equation (1) as $f_w(\boldsymbol{x}) \sim GP\left(0\,,\, k_{\beta_w}(\boldsymbol{x}, \boldsymbol{x}')\right)$. Then, we can describe each observed target outcome $Y_i(\mathbf{x_i})^{(w_i)}$ by:

$$Y_i(x_i)^{(w_i=0)} \sim GP\left(0\,,\, \sigma^2 + k_{\beta_0}(x_i, x_i)\right) \text{ and } Y_i(\boldsymbol{x_i})^{(w_i=1)} - \mathrm{f_0(x)} \sim GP\left(0\,,\, \tau^2 + k_{\beta_1}(\boldsymbol{x_i}, \boldsymbol{x_i})\right) \quad (2)$$

Where $GP(\mu, \Sigma)$ denotes a Gaussian Process with mean μ and Variance Σ, and $k_{\beta_w}(x, x')$: $R^d \times R^d \to R_+$ represents a kernel with parameters specified by $\beta_w \in \beta \subset R^m$, and σ, τ denote additional noise parameters in the model that account for unexplained variability in the data. From the above definition, and by splitting the sample according to the treatment indicator w_i, $i = 1, \ldots, N$, we can structure our data as follows: $Y^{(W=0)}$, $Y^{(W=1)}$, $X^{(W=0)}$, and $X^{(W=1)}$. These represent the response vectors for untreated and treated with dimensions N_c and N_t, and the control variables matrices for untreated and treated with dimensions $N_c \times N_c$ and $N_t \times N_t$, respectively. Here $N = N_c + N_t$. Using this structure, for a new advertiser with characteristics $\mathbf{x_{new}}$ the CATE, conditional on the observed sample $\{(Y_i, \boldsymbol{X_i}, W_i)\}_{i=1}^N$, is given by:

$$\widehat{\mathrm{f_1}}(\mathrm{x_{new}}) = \mathrm{k}_{\beta_1^*}^{\mathrm{T}}\left(\mathrm{x_{new}}, \mathrm{X}^{(\mathrm{W=1})}\right)\left(\mathrm{K}_{\beta_1^*}\left(\mathrm{X}^{(\mathrm{W=1})}\right) + \widehat{\tau^{*2}}\mathbf{I_{N_t}}\right)^{-1}\mathbf{R}^{(\mathbf{W=1})}, \quad (3)$$

where $\boldsymbol{R}^{(\boldsymbol{W=1})} = \boldsymbol{Y}^{(\boldsymbol{W=1})} - \hat{\boldsymbol{f}}_{\boldsymbol{0}}\left(\boldsymbol{X}^{(\boldsymbol{W=1})}\right)$. Here, $k_{\beta_w^*}^T\left(x_{new}, X^{(W=w)}\right)_i = k_w\left(x_{new}, x_i^{(w_i=w)}\right), i = 1, \ldots, N_w$ and $\boldsymbol{I_{N_w}}$ denotes the identity matrix. Note that the model is completely specified by the

optimal parameters τ^* and β_w^* obtained by maximizing the likelihood of $Y^{(W)}$ as a function of (β_w, τ), conditional on the observed sample.

2.4 Intuition About the Bias Correction Induced by the Two-stage GP Method

As seen in Equation (3), the expression for the treatment effect function $\widehat{f_1}(\cdot)$ depends on the residual vector $R^{(W=1)}$, that in turn, depends on $\hat{f}_0(\cdot)$. This translates into the following representation for the CATE function $\hat{f}_1(\cdot)$:

$$\widehat{f_1}(\boldsymbol{x_{new}}) = \widetilde{f_1}(\boldsymbol{x_{new}}; \beta^*) - \widehat{f_{0,1}}(\boldsymbol{x_{new}}; \beta^*). \quad (4)$$

where $\widetilde{f_1}(\cdot)$ corresponds to a function that depends only the treated samples, and $\beta^* = [\beta_1^* \, \beta_0^*]$ denotes the stacked vector of kernel parameters for $f_1(\cdot)$ and $f_0(\cdot)$, respectively. Writing down the term $\widehat{f_{0,1}}(\boldsymbol{x_{new}}; \boldsymbol{\beta}^*)$ in Equation (4) and using results from Equation (3), it follows:

$$\widehat{f_{0,1}}(x_{new}; \beta^*)$$
$$= k_{\beta_1^*}^T\left(x_{new}, X^{(W=1)}\right)\widetilde{K_{\beta_1^*}}\left(X^{(W=1)}\right)^{-1}\boldsymbol{K}_{\beta_0^*}\left(X^{(W=1)}, X^{(W=0)}\right)\widetilde{K_{\beta_0^*}}\left(X^{(W=0)}\right)^{-1}\boldsymbol{Y}^{(\boldsymbol{W=0})}. \quad (5)$$

Here, $\widetilde{K_{\beta_1^*}}\left(X^{(W=1)}\right) = K_{\beta_1^*}\left(X^{(W=1)}\right) + \widehat{\tau^{*2}}\boldsymbol{I}_{\boldsymbol{N_t}}$. As Equation (5) shows, $\widehat{f_{0,1}}(\boldsymbol{x_{new}}; \boldsymbol{\beta}^*)$ is a linear combination of the observed target responses for the non-treated population (i.e., $\boldsymbol{Y}^{(\boldsymbol{W=0})}$), with weights determined by three components: (1) The kernel-based similarity between the new advertiser characteristics x_{new} and the observed treated samples $X^{(W=1)}$; (2) The kernel-based similarity between the observed treated sample $\boldsymbol{X}^{(\boldsymbol{W=1})}$ and the non-treated population $\boldsymbol{X}^{(\boldsymbol{W=0})}$ spanned by the cross-covariance term $K_{\beta_0^*}\left(X^{(W=1)}, X^{(W=0)}\right)$; finally, (3) The intra-population dependence for treated and untreated groups that is contained in the corresponding precision matrices $\widetilde{K_{\beta_w^*}}\left(X^{(W=w)}\right)^{-1}$, $w \in \{0,1\}$. This representation can be interpreted as an adaptive matching that takes place on the high-dimensional space induced by the kernel functions.

As an alternative, we can represent the estimator in Equation (4) as:

$$\widehat{f_1}(x_{new}) = \sum_{i \in \mathcal{S}_\mathcal{T}} \alpha\left(x_{new}; X_i^{(1)}\right)\left(Y_i^{(1)} - \widetilde{Y_\iota^{(0)}}\right), \quad (6)$$

where $\mathcal{S}_\mathcal{T}$ represents the treated sample, $\alpha\left(x_{new}; X_i^{(1)}\right) \geq 0$ are weights that account for the similarity between the new observed sample $\boldsymbol{x_{new}}$ and observed treated samples in $\mathcal{S}_\mathcal{T}$, and $\widetilde{Y_\iota^{(0)}}$

denotes the $i-th$ treated sample counterfactual. This counterfactual is constructed as a feature-dependent linear combination of the observed non-treated responses as $\widetilde{Y_i^{(0)}} = \sum_{j \in \mathcal{S}_C} k_{\beta_0^*}\left(X_i^{(1)}, X_j^{(0)}\right) \cdot Y_j^{(0)}$, where $Y_j^{(0)}$ are the observed non-treated responses and $k_{\beta_0^*}\left(\boldsymbol{X_i^{(1)}}, \boldsymbol{X_j^{(0)}}\right)$ corresponds to the $i-th$ row of the translation matrix $H_{\beta_0^{**}}\left(X^{(1)}, X^{(0)}\right) = K_{\beta_0^*}\left(X^{(1)}, X^{(0)}\right)\widetilde{K_{\beta_0^*}^{-1}}\left(X^{(0)}\right)$ that projects the observed treated samples into the non-treated space by adaptively allocating adaptive weights. This process resembles a synthetic control but with the difference that the weights are constructed by leveraging the probabilistic relationship between Y and X given by the GP.

2.5 Benchmarking of the Two-stage GP Method Using Synthetic Data

We compare Two-stage GP with competing algorithms using the Infant Health and Development Program (IHDP) dataset. This dataset has been used for evaluating Causal Models (Alaa and Van der Schaar 2018). In all the evaluations, CMGP and 2-GP are fitted using a "Radial Basis Function" (RBF) Kernel with Automatic Relevance Determination (ARD). We refer to each model as: Causal Forests (CF), Double Machine Learning with Partially Linear Model (DML, rf_dml), Causal Multitask Gaussian Processes (CMGP), and Linear Regression with Propensity Scores and Bootstrapping (lin_boots). We estimate the Average Treatment Effect (ATE) to identify an approach that achieves low Root Mean Squared Error (RMSE) and is robust to small sample sizes.

Figure 1: Boxplots of Error in 50th Percentile Estimated DSI across 100 replications for IHDP dataset, N=700.

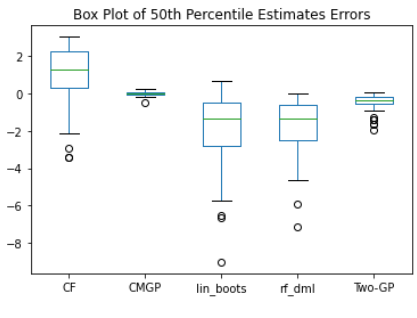


Figure 2: RMSE plots for 50th Percentile DSI estimation across 100 replications for IHDP dataset.

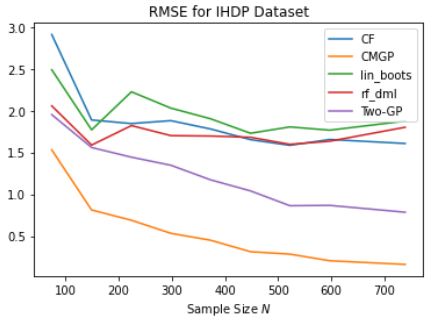
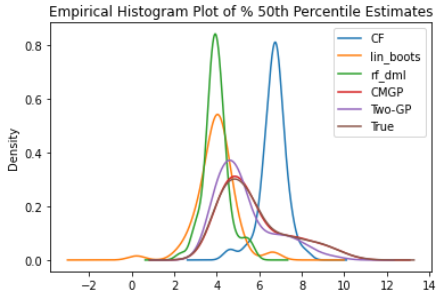


Figure 3: Histogram plots for 50th percentile DSI estimation using IHDP dataset (100 replications), N=739.



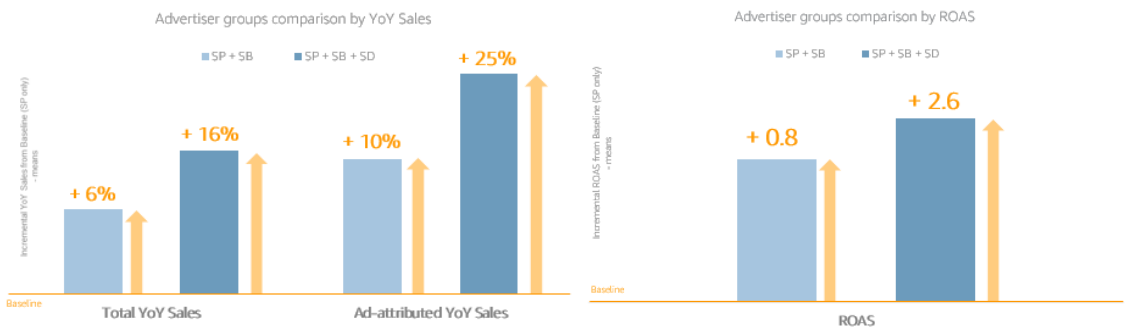
Our results show that Gaussian Processes-based methods perform best with respect to IHDP (and also in Amazon-based synthetic data, and the back-shifted placebo test) when sample sizes are below 1,000. Although CMGP stands out as the best method in those scenarios, it comes at a high computational cost due to its Kernel structure. The fact that Two-Stage GP offers computational

savings by using a simpler Kernel structure, and is flexible to be adapted to other use cases with little performance tradeoffs makes it a better candidate to be used at scale.

3. Major Results

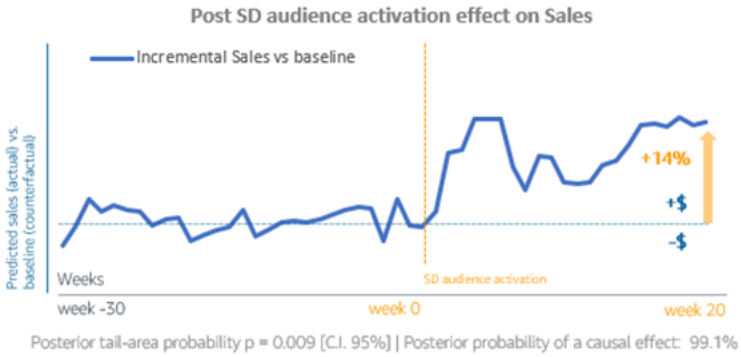
In our analysis we used data between April 2020 and June 2021. As to model-free evidence, ***Figure 4*** below compares the Year-over-Year (YoY) growth of brands combining SD with SP and SB vs. those using SP + SB (which we considered the baseline). The left panel shows: (a) Total Sales, and (b) Total Ad-attributed Sales Year-over-Year (YoY) growth. The right panel shows the Return on Advertising Spending (ROAS), i.e., the $ sales revenue earned for each $1 spend on advertising. Brands combining SP + SB + SD generated +16% more in total sales and +25% more in ad-attributed sales, respectively, vs. brands using SP only, and they achieved +2.6 higher ROAS. Thus, adding Sponsored Display audience targeting (an upper-funnel tactic) to the existing lower- and mid-funnel tactics, is associated with higher sales growth (effectiveness) without decreasing efficiency (as measured by ROAS), answering a key advertiser question.

Figure 4: Higher performance for advertisers who added Sponsored Display (Model-free)



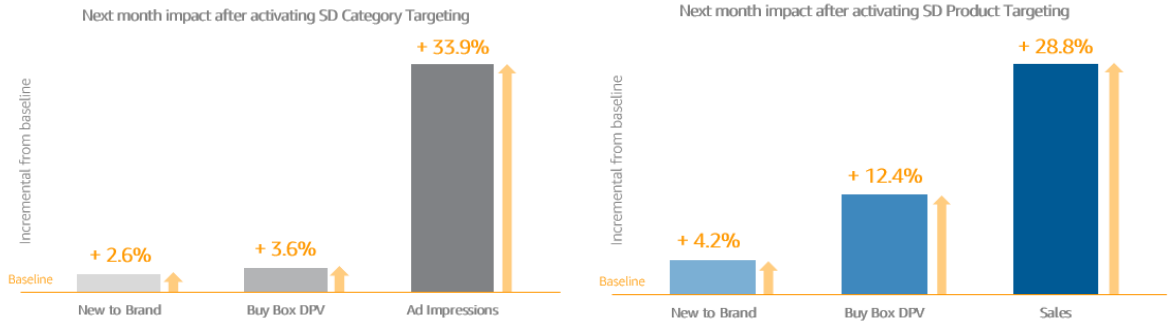
From a medium term impact perspective, in our *counterfactual causal analysis*, the 284 advertisers that adopted SD with *audience targeting* saw, on average, a +14% increase in total sales in the following 20 weeks, compared to their estimated sales without SD adoption. Figure 5 shows on the left panel that the model works well predicting the average baseline of actual sales (i.e., weeks -30 to 0 in the horizontal dotted line) and on the right side the impact of adding SD to SP+SB during the 20 weeks post SD adoption. The posterior probability of a causal effect is 99.1%. We performed validation separately with posterior tail area probability of p=0.009.

Figure 5: Incremental sales from adopting Sponsored Display Audience vs. counterfactual baseline



Finally, using our Two-stage GP method we estimated the impact of adding SD over the next month post SD adoption. Brands that began using *category targeting* within SD strategies for the first time saw, on average, +33.9% more impressions, a +3.6% increase in Detailed Page Views and a +2.6% increase in New-To-Brand customers the following month compared to those that didn't. Similarly, brands that created an SD *product targeting* campaign for the first time saw, on average, a +28.8% sales increase, +4.2% DPV increase and a +2.6% NTB increase the next month compared to those that didn't. ***Figure 6*** summarizes these results We measured the statistical significance with a 5% significance level of these estimates using a bootstrapping procedure.

Figure 6: Two-stage Gaussian Process estimates of Sponsored Display Targeting's impact.



4. Implications: Incorporate Sponsored Display strategies in digital advertising

Using a multi-method approach, we conclude that brands that incorporated Sponsored Display experienced increases in total sales ranging from +10% to +29%, as well as increases in impressions, Detailed Page Views, New-to-Brand customers, ad-attributed sales, and Return on Advertising Spend (ROAS), compared to brands that only use Sponsored Products, or Sponsored Products and Sponsored Brands on Amazon.com. Based on these results, we recommend that advertisers incorporate Sponsored Display to their media plans. We also recommend that brands consider using multiple SD tactics, such as audience, category and product targeting. Future research is needed to explicitly compare audience with category and product targeting, and to quantify how combinations of these approaches are best for advertisers under different conditions, such as category and brand characteristics (e.g. consumer involvement and brand strength).

References

Alaa, A. M. and van der Schaar, M. (2018) "Bayesian nonparametric causal inference: Information rates and learning algorithms." *IEEE Journal of Selected Topics in Signal Processing*, 12(5),1031–1046.

Amazon Learning Console (2021), https://advertising.amazon.com/library/courses/reach-shoppers-with-sponsored-display/?ref_=a20m_us_rfy_lbr_crs_1 .

Brodersen, K. H., F. Galluser, J. Koehler, N. Remy and S.L. Scott (2015), "Inferring Causal Impact using Bayesian Structural Time-Series Models." *The Annals of Applied Statistics*, 9 (1), Institute of Mathematical Statistics, 247–274, http://www.jstor.org/stable/24522418.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J, Robins (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal*, 21(1):C1–C68 February, https://doi.org/10.1111/ectj.12097

Lechner, M. (2011), "The Estimation of Causal Effects by Difference-in-Difference Methods", *Foundations and Trends® in Econometrics*: 4 (3), 165-224.

Rangaswamy, A., Moch, N., Felten, C., van Bruggen, G., Wieringa, J. E., & Wirtz, J. (2020). The role of marketing in digital business platforms. *Journal of Interactive Marketing*, *51*, 72-90.

Robb, K. (2021), Amazon Sponsored Display Ads: A Full-Funnel Approach, *Teikametrics*, https://www.teikametrics.com/blog/amazon-sponsored-display-ads-a-full-funnel-approach/

Wager, S. and Athey, S. (2018), "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113 (523),1228–1242.