Contextual Multi-Armed Bandits for Causal Marketing

Neela Sawant [1] **Chitti Babu Namballa** [1] **Narayanan Sadagopan** [1] **Houssam Nassif** [1]

Abstract

This work explores the idea of a *causal contextual multi-armed bandit* approach to automated marketing, where we estimate and optimize the causal (incremental) effects. Focusing on causal effect leads to better return on investment (ROI) by targeting only the persuadable customers who wouldn't have taken the action organically. Our approach draws on strengths of causal inference, uplift modeling, and multi-armed bandits. It optimizes on causal treatment effects rather than pure outcome, and incorporates counterfactual generation within data collection. Following uplift modeling results, we optimize over the incremental business metric. Multi-armed bandit methods allow us to scale to multiple treatments and to perform off-policy policy evaluation on logged data. The Thompson sampling strategy in particular enables exploration of treatments on similar customer contexts and materialization of counterfactual outcomes. Preliminary offline experiments on a retail Fashion marketing dataset show merits of our proposal.

1. Introduction

Personalizing marketing campaigns is a focal problem in advertising (Chen et al., 2009). Conventional advertising and recommendation approaches optimize on observed outcomes such as clicks and transactional revenue that tend to reinforce known behaviors (Teo et al., 2016; Nassif et al., 2016). For instance, if a customer is already a heavy purchaser in mobile apps, a model optimizing on observed conversions may show more mobile app marketing campaigns to the customer, even if the customer would have made the purchase anyway. In fact, a recent study found that only about a quarter of the clicks typically attributed to the recommender system on Amazon are actually caused by it (Sharma et al., 2015).

To improve the return on investment (ROI), there is a need to optimize for incremental (or causal) effects (Rubin & Waterman, 2015) in personalizing marketing campaigns for customers[1]. From the advertiser's perspective, this leads to efficient utilization of marketing budget. From the customer's perspective, this potentially introduces new programs instead of the same recommendations time and time again. There are three key challenges in addressing the causal effects of marketing:

- Counterfactual estimation - While we know the outcome of a campaign targeted at a particular customer, we also need to know the counterfactual outcome (what would have been) if a different campaign was presented instead. As it is impossible to get the ground truth for counterfactuals, the true incremental effect cannot be measured directly.
- Unknown total effect size - Marketing campaigns may drive multiple actions or induce repeat behavior. For example a Fashion ad may lead to a halo effect inducing a customer to purchase multiple accessories in addition to the advertised product.
- Fast experimentation - Most marketing campaigns are short-lived. We need to quickly estimate the incremental effect and optimize campaign allocation.

We explore a *causal bandit* idea that optimizes campaign allocation based on estimated incremental effects. This approach builds on strengths of causal inference, uplift modeling, and multi-arm bandits. We use the probabilistic exploration of Thompson sampling for materializing counterfactual outcomes which are used to estimate and optimize context level incremental effect of campaigns.

2. Related Work

2.1. Causal Treatment Effect Estimation

We first specify the causal treatment effect estimation problem. Let $i = 1, \ldots, N$ denote a customer. Suppose we

[1]Amazon.com, USA. Correspondence to: Neela Sawant <nsawant@amazon.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

[1]We use following words interchangeably - (a) campaign, treatment; (b) incremental effect, causal effect.

have only one treatment and one control. Let $W_i = \{0, 1\}$ be a binary treatment indicator, with $W_i = 1$ denoting that customer i received the treatment and $W_i = 0$ denoting that customer i received the control. Let $Y_i^{(obs)}$ be the observed customer action outcome (e.g. click, revenue, other business metric). Let X_i represent the context, potentially including customer features. Each user has a pair of potential outcomes $(Y_i(0), Y_i(1))$. Following the Rubin Causal Model (Rubin, 1974; Imbens & Rubin, 2015), the customer-level treatment effect τ_i is defined as the difference in potential outcomes:

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

Dropping suffix i for brevity, let Δ denote the random variable corresponding to the difference in potential outcomes. Following Athey & Imbens we define the Conditional Average Treatment Effect (CATE) for a given context $\mathbf{x}$ as:

$$\tau(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0)|X = \mathbf{x}] = \int \Delta Pr(Y(1) - Y(0) = \Delta|X = \mathbf{x})\mathrm{d}\Delta. \tag{2}$$

Say unconfoundedness (selection on observables) assumption holds, which is satisfied in randomized experiments or when the treatment selection is completely determined through observed X. Then, treatment assignment W_i is independent of the potential outcomes given X:

$$W \perp\!\!\!\perp (Y(0), Y(1))|X. \tag{3}$$

Prior work in machine learning based estimation of binary CATE can be categorized in three model types (Athey & Imbens, 2015; Radcliffe, 2007)[2]:

- Single-model - As the name suggests, this approach trains one model to predict Y as a function of W and X covariates. Let $\mu(w, \mathbf{x}) = \mathbb{E}[Y^{(obs)}|W = w, X = \mathbf{x}]$, and $\hat{\mu}$ its estimator. Given context vector $\mathbf{x}$, the estimator is scored twice using potential values for treatment $(0, 1)$, and CATE estimated as their difference:

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}(1, \mathbf{x}) - \hat{\mu}(0, \mathbf{x}). \tag{4}$$

- Two-model - This approach trains two models $\hat{\mu_T}$ and $\hat{\mu_C}$ to separately estimate Y on treatment and control customers. Given context vector $\mathbf{x}$, both estimators are scored once and CATE estimated as the difference:

$$\hat{\tau}(\mathbf{x}) = \hat{\mu_T}(\mathbf{x}) - \hat{\mu_C}(\mathbf{x}). \tag{5}$$

[2] Athey & Imbens specify decision-tree models. Since the core ideas apply to general supervised learning approaches, we take the liberty of paraphrasing.

- Transformed outcome model - This approach uses the insight that if the observed outcome is transformed as:

$$\tilde{Y} = Y(1) \times \frac{W}{Pr(W = 1|\mathbf{x})} - Y(0) \times \frac{1 - W}{1 - Pr(W = 1|\mathbf{x})} \tag{6}$$

then $\mathbb{E}[\tilde{Y}|\mathbf{x}]$ is an unbiased estimator of CATE $\tau(\mathbf{x})$. One transforms all outcomes, and uses conventional regressors to directly predict CATE.

An important shortcoming of single and two model approaches is the lack of an appropriate goodness-of-fit or cross-validation measure for model selection, as there is no ground truth for the conditional average treatment effect $\tau(\mathbf{x})$ (Sugiyama et al., 2007). The transformed outcome method discards information, and it may be more efficient to estimate $\tau(\mathbf{x})$ using triplet $(Y^{(obs)}, W, X)$ than only $(\tilde{Y}, X)$.

Athey & Imbens finally present a new algorithm: Causal Tree. They refine the tree construction algorithm to obtain an unbiased estimator of $\tau(\mathbf{x})$. In particular the average treatment effect is estimated by averaging the treatment effect from tree leaves. Units clustered into decision tree leaves are considered synthetic twins of each other, using which CATE within a leaf is estimated. Shalit et al. present theoretical analysis and a family of algorithms based on representation learning that estimates CATE (or individual treatment effects) from observational data. Their algorithm learns a balanced representation such that the induced treatment and control distributions look similar.

2.2. Uplift Modeling

Uplift modeling is a marketing technique to measure the effectiveness of a marketing action and predict its incremental response (Victor & Lo, 2002; Hansotia & Rukstales, 2002). Similar to causal methods, the simple two-model uplift approach models two populations separately and subtracts their lift (Radcliffe, 2007). For continuous models, two-model uplift is defined as:

$$\mathbb{E}(outcome|treatment) - \mathbb{E}(outcome|control). \tag{7}$$

The quality of an uplift model is often evaluated by computing an uplift curve (Radcliffe & Surry, 2011; Kuusisto et al., 2014), similarly to how an ROC curve is constructed (Nassif et al., 2013b). First, we defines a model lift measure over the ranked customers. A lift example is the cumulative outcome amongst the model's top ρ fraction of ranked examples. We then plot the uplift curve by variying $0 \leq \rho \leq 1$ in:

$$\text{Uplift}(\rho) = \text{Lift}(\rho, \text{treatment}) - \text{Lift}(\rho, \text{control}). \tag{8}$$

Uplift curves capture the additional outcome obtained due to the treatment, and is sensitive to variations in coverage (Nassif et al., 2013a). The higher the uplift curve, the more profitable a marketing model/intervention is.

2.3. Multi-Armed Bandits

We need to predict the incremental effect of future campaigns when presented to potentially new customers. Marketing campaigns are transient, share common features, and need to be tested quickly (Hill et al., 2017). Multi-armed bandit algorithms are effective for rapid experimentation because they concentrate testing on treatments that have the greatest potential reward (Bubeck & Cesa-Bianchi, 2012; Hill et al., 2017). They optimally balance exploration and exploitation to achieve minimal regret, and incorporate customer context to personalize prediction (Agrawal & Goyal, 2013; Dani et al., 2008).

Thompson sampling is a common bandit algorithm where a campaign is selected proportionally to the probability of that campaign being optimal, conditioned on previous observations. In practice, this probability is not sampled directly. Instead one samples model parameters from their posterior and picks the campaign that maximizes the reward (Chapelle & Li, 2011).

The standard bandit setting assumes unconfoundedness, although non-contextual bandits with unobserved confounders have been developed (Bareinboim et al., 2015). In the case of known confounders, one may resort to inverse propensity weighing and other off-policy policy evaluation methods to unbias the estimates (Li et al., 2015; Swaminathan & Joachims, 2015b).

2.4. Multi campaign setting

In this study we focus on general multi campaign settings where $k > 2$ marketing campaigns can be active at any time. Let $W = \{0, 1, \ldots, K\}$ be the campaign indicator with $W = k$ if customer received campaign k. Let $W = 0$ be a special case (customer held out of marketing or receives control). Let H denote the binary conversion indicator for some action that we want to drive through these marketing campaigns, and Y its business metric. Then the estimated CATE of campaign k becomes (see proof in Appendix A):

$$\begin{aligned}\hat{\tau}_k(\mathbf{x}) =& (Pr(H = 1|X = \mathbf{x}, W = k) \\ &- Pr(H = 1|X = \mathbf{x}, W = 0)) \\ &\times (\mathbb{E}[Y|X = \mathbf{x}, H = 1] - \mathbb{E}[Y|X = \mathbf{x}, H = 0]).\end{aligned} \tag{9}$$

One can estimate CATE's first term (incremental propensity) in three steps: a) build a conversion model on customers exposed to a given campaign; (b) build a conversion model on customers not exposed to that campaign; (c) score both models on a new customer and take their score difference. One can also use predictive models to approximate Eqn. 9's second term, business metric difference. This two-model approach has limitations:

- Requiring a M:1 mapping from campaigns to actions is limiting. A campaign may drive more than one action and its total incremental effect should include all its actions. Attribution can be challenging if multiple campaigns are involved. If we allow a M:N mapping between campaigns and actions, a campaign's incremental effect becomes more comprehensive but computationally cumbersome:

$$\begin{aligned}&\hat{\tau}_k(\mathbf{x}) = \\ &\sum_j (Pr(H^j = 1|X = \mathbf{x}, W = k) \\ &\qquad - Pr(H^j = 1|X = \mathbf{x}, W = 0)) \\ &\times (\mathbb{E}[Y|X = \mathbf{x}, H^j = 1] - \mathbb{E}[Y|X = \mathbf{x}, H^j = 0]).\end{aligned} \tag{10}$$

- Customer response to marketing is impacted by content quality and environmental factors (time of day, day of week, device). It would be ideal to compute campaign or context-specific incremental propensity.
- Getting a pure control group ($W_i = 0$) is challenging. A control customer may receive substitute campaigns via other marketing channels. It is difficult to account for such exogenous factors.
- Uplift modeling approaches that model uplift directly tend to outperform their two-model counterparts (Radcliffe & Surry, 2011; Nassif et al., 2012).

3. Causal Effect Prediction Proposal

Our approach draws on the strengths of causal inference, uplift modeling, and multi-armed bandits. Akin to causal trees, we optimize on causal treatment effect rather than pure outcome, and incorporate counterfactual matching within data collection. Following uplift modeling results, we directly optimize the incremental business metric. We use contextual bayesian multi-armed bandits to scale to multiple treatments and to perform off-policy evaluation on logged data. In particular, Thompson sampling enables exploration of treatments on similar customers and materialization of counterfactual outcomes.

3.1. Problem Formalization

As we are dealing with multiple treatments, we generalize the customer-level treatment effect to:

$$\tau_k = Y(W = k) - Y(W \neq k), \tag{11}$$

where $(Y(W = k), Y(W \neq k))$ are the postulated outcomes of receiving treatment k and not receiving treatment k, respectively. We define:

$$Y(W \neq k) = \int_y y \sum_{m \neq k} Pr(Y = y|X, W = m)Pr(W = m)\mathrm{d}y. \quad (12)$$

Similarly, we define the treatment-specific Conditional Average Treatment Effect (CATE) for campaign k as:

$$\tau_k(\mathbf{x}) = \mathbb{E}[Y(W = k) - Y(W \neq k)|X = \mathbf{x}]. \quad (13)$$

The unconfoundedness assumption becomes:

$$W \perp\!\!\!\perp (Y(0), \ldots, Y(K))|X. \quad (14)$$

Let us define the customer-level treatment-specific incremental effect for customer i as:

$$Y^*_{i,k} = Y_i(W_i = k) - Y_i(W_i \neq k). \quad (15)$$

We now present a treatment-specific CATE based on expectation of the difference, rather than the difference in expectations of outcomes from two separate models:

$$\hat{\tau}_k(\mathbf{x}) = \mathbb{E}[Y^*_{i,k}|X_i = \mathbf{x}]. \quad (16)$$

The choice of $W \neq k$ baseline in Eqn. 11, which includes the possibility of not showing any treatment, is in the same spirit as one-versus-all multi-class classification. Such an approach is appropriate from a business-perspective, and our experiments show benefit over using non-causal approaches. However, since each arm has a different baseline, one may argue that using a common baseline across all arms may result in a better CATE estimate. For example, one can drop customers who didn't see any treatment, and compare against a constant control arm, or the average performance of all arms. We note that the latter methods conserve the same arm ranking as a non-causal bandit, and argue that preserving non-treatment customers is closer to the production logic. We are exploring these different types of baseline and their trade-offs as part of our future work.

3.2. Training Data Generation with Incremental Target Calculation

To measure and predict the incremental effect of treatments $Y^*_{i,k}$, we leverage off-policy evaluation techniques (Swaminathan & Joachims, 2015b) and require that historical treatments are logged properly. Let logs be $L = \{X_i, W_i, P_i, Y_i^{(obs)})\}$ where user i with context X_i was shown treatment $W_i = k$ with probability $P_i > 0$ and outcome $Y_i^{(obs)}$. The nature of observed outcome is application-specific, such as clicks, repeat visits, difference in customer spending before and after the experiment, etc. We estimate the counterfactual $Y_i(W_i \neq k)$ based on historical records whose context matches X_i.

More formally, let $\Phi(X)$ be a context matching algorithm that finds similar customers to X, like K-Nearest Neighbor, locality sensitive hashing, or propensity matching. Given customer X_i with observed outcome $Y_i^{(obs)}$, we use $\Phi(X_i)$ to identify similar customers who were not shown treatment k, and use their logged outcomes to estimate a counterfactual outcome $Y_i(W_i \neq k)$ for customer X_i. Algorithm 1 shows our training data generation approach where we leverage similar historical logs to build the incremental effect training data.

Algorithm 1 Training Data Generation for Incremental Outcome Prediction Model

Inputs: Size of training data $M > 0$; logged events L; number of matching observations M'; context matching algorithm Φ.
$T_0 \leftarrow \emptyset$ {An initially empty training data}
for $m = 1$ **to** M **do**
  Sample record $(\mathbf{x}_m, w_m, p_m, y_m^{(obs)})$ from L.
  $y_m^{(tr)} \leftarrow y_m^{(obs)}/p_m$ {Apply policy bias correction to outcome}
  $y_m^{(cf)} \leftarrow 0$ {Initialize counterfactual outcome to zero}
  Select M' similar records using $\Phi(\mathbf{x}_m, \mathbf{x}_{m'})$
  for $m' = 1$ **to** M' **do**
    Pick record $(\mathbf{x}_{m'}, w_{m'}, p_{m'}, y_{m'}^{(obs)})$ with $w_m \neq w_{m'}$ from L
    $y_m^{(cf)} \leftarrow y_m^{(cf)} + y_{m'}^{(obs)}/p_{m'}$
  end for
  $y_m^{(cf)} \leftarrow y_m^{(cf)}/M'$ {Final counterfactual estimate}
  $y^*_{m,w_m} \leftarrow y_m^{(tr)} - y_m^{(cf)}$ {Unbiased incremental target}
  $T_m \leftarrow$ CONCATENATE$(T_{m-1}, (\mathbf{x}_m, w_m, y^*_{m,w_m}))$
end for

Suppose we have historical data logs such that $P(k, X_i) > 0, \forall k, X_i$. There are no hidden confounders because W_i is chosen with probability P_i and that probability is computed depending only on X_i. The outcome Y_i and covariates outside X_i are unknown to the software that evaluates P_i, so Y_i must be independent of these. We then use inverse propensity estimation to unbias Y^* (Li et al., 2015). One may further refine the inverse propensity estimator to control for variance (Swaminathan & Joachims, 2015a).

3.3. Using Thompson Sampling for Counterfactual Materialization

To ensure online-optimization and a balanced explore-exploit mix, we use multi-armed bandits trained on Algorithm 1 data to estimate Eqn. 16. Let $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ denote the model parameters for the K-armed bandit model.

Let $W^*(\theta_t, \mathbf{x})$ denote the optimal treatment given context $\mathbf{x}$ and model parameters θ at time t. At any given time $t+1$, Thompson sampling selects a treatment $W = k$ proportionally to the probability of that treatment being optimal:

$$W_{t+1} \sim Pr(W = W^*(\theta_t, \mathbf{x})|X = \mathbf{x}). \quad (17)$$

When there is no strong evidence that any one treatment is optimal, the bandit explores multiple treatments on similar contexts. As the bandit improves its treatment success estimates, it is more likely to exploit the winning one (Chapelle & Li, 2011). As such, Thompson sampling bandits are a perfect fit for the materialization of counterfactual outcomes. The observed contexts and outcomes determine θ_t and $Pr(W = W^*)$. We log $P_i = Pr(W_i = W^*)$, making the proposed framework a closed-loop system (Agarwal et al., 2016).

Algorithm 2 specifies the overall contextual bayesian bandit algorithm. We suppose that at time t, multiple contexts are received. The increment in t can be understood to correspond to one day where predictions are made on multiple contexts, but model update happens asynchronously after delayed incremental outcomes are computed.

Algorithm 2 Thompson Sampling based Contextual Multi-Armed Bandits with Online Scoring and Batch Training

Initialization: Time $t = 0$; event log $L = \{\}$; d-dimensional bandit arm contextual distribution parameters $\theta^k \sim \mathcal{N}_d(0, 1), \forall k$ arms.
for $t = 1, 2, \ldots$ **do**
  {Online Scoring}
  for $m = 1, 2, \ldots$ **do**
    Receive d-dimensional context $\mathbf{x}_m$.
    for Arm $k = 1, 2, \ldots, K$ **do**
      Sample $\theta \sim \theta_k$
      Estimate $y^*_{m,k} = \theta^T \mathbf{x}_m$.
    end for
    Play arm $w_m = \arg\max_k y^*_{m,k}$.
    Log (x_m, w_m, p_m).
  end for
  {Batch Training}
  Asynchronously update log L with delayed reception of new rewards $Y_m^{(obs)}$ for $m = 1, 2, ...$
  Update $\theta_k, \forall k$ using training data generated by Algorithm 1.
end for

4. Experiments

Experiments in this paper are based on offline evaluation on marketing logs. We collected an Amazon Fashion marketing dataset where 410K randomly sampled treatment customers were randomly targeted with one of 16 marketing campaigns upon visiting a retail app. These campaigns spanned across men's and women's fashion in clothing, shoes, jewelry, and watches (examples shown in Figure 1). Separately, we randomly sampled a control hold-out set of 100K customers that did not receive any campaign from our experiment (customers may have seen other business-as-usual messages). The resulting dataset was split into 70:30 training and testing. Following the concepts of off-policy policy evaluation, we only consider targeted customers who were shown the same campaign as the new model prediction (Li et al., 2015).

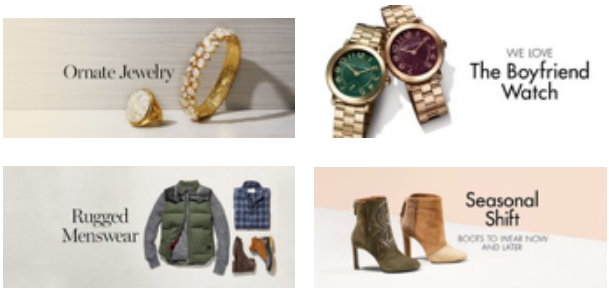


Figure 1. Examples of Fashion campaigns used in data collection.

As incrementality measures are often continuous, our Algorithm 2 needs to be a contextual bayesian bandit that optimizes a continuous target metric. We use a Bayesian Linear Probit regression extension of (Graepel et al., 2010) that has a continuous support. We use this regression as the basis of a Bayesian Generalized Linear Bandit, as done in (Teo et al., 2016) and (Hill et al., 2017).

We featurized campaign content and customer behaviors as well as their interactions. We winsorized continuous features to reduce the effect of outliers, applied log-transformation to spend and pricing related features, and an overall min-max scaling to limit features to a [0,1] range. Given the large number of features and campaigns compared to the size of available training data, we performed feature selection prior to building predictive models. We chose the S best features ($S \in \{100, 200, 300, \ldots\}$) based on the regression feature-target F-value between features and the continuous target variable. We then trained a bandit model on each feature set S, and measured its cross-validated performance to identify the best model to test.

Incremental outcomes (Y^*) are generated using Algorithm 1, as the difference in the business metric between the targeted customer and M' similar customers who were not shown that specific campaign (or any campaign). As a context matching algorithm Φ, we used a GPU-based nearest-neighbor search service that uses the *hierarchically navigable small worlds* similarity algorithm (Malkov & Yashunin, 2016) to match normalized customer feature vectors. We

tried two different values of $M' \in \{10, 50\}$ both of which had similar results (Pearson correlation 0.92, p-value 0.0). We fix $M' = 10$ in further experiments.

5. Results

5.1. Causal Bandit Performance

We first compare two types of models:

Non-incremental model uses the bayesian probit regression bandit to optimize the total value of a fashion business metric of the targeted customers.

Incremental model is the the proposed causal bandit model. It uses the bayesian probit regression bandit to optimize the incremental value of the same fashion business metric of the targeted customers, wherein increment is computed using Eqn.15 and Algorithm 1.

For each customer, each bandit model scores all campaigns, and recommends the campaign with the highest score. We train each model on the treatment training set, and use the trained model to score and recommend campaigns for both the treatment testing set, as well as the control hold-out set.

To visualize the treatment effect, we rank the testing and hold-out customers by their recommended campaign predictive score, and divide the ranked list into ten deciles. Decile 10 represents the top 10% campaign-susceptible customers and decile 1 represents the bottom 10%. Each decile contains a segment of targeted customers shown the same campaign as recommended by the scoring model and a segment of hold-out customers. We refer to these segments as the decile-level treatment and control groups. The decile treatment effect is the differrence between the observed business metric for treament and control decile-level segments (recall customers were randomly assigned to treatment and control).

Figure 2 uses the decile ranking produced by the incremental model, and plots the actual business metric average per decile [3]. For the top decile, incremental-model targeting results in 1.2 business metric units, compared to 0.46 units for control. The difference of 0.74 between these two values is precisely the conditional average treatment effect (CATE) in that decile. Analyzing all deciles, we observe a good correlation between predictive incremental bandit score and the actual causal effect, thus highlighting the model's utility in making useful advance predictions for business decisions.

Readers may wonder about the oddity that decile 2 has a smaller score than decile 1. We believe this is caused by new-to-site customers who are placed in decile 1 due to lack of historical behaviors (e.g. model confuses them for inactive customers and assigns a low score). However, these customers are actually responsive and raise the per-customer performance in decile 1 compared to decile 2.

[3]The business metric is observed in a future time period post targeting.

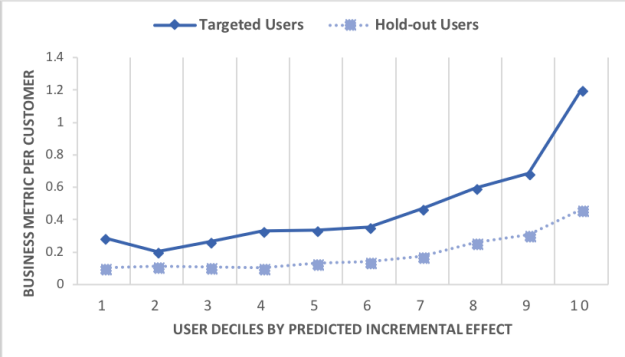


Figure 2. Per-customer business metric per decile, measured on the treatment and control sets. Deciles are ranked using the incremental model. The difference between curves is the causal effect of marketing in that decile.

Budget limitations typically force business to limit the number of customers exposed to different marketing activities. This suggests that business teams will place more importance in top-ranked customers and desire a segment of top-ranked customer which yields better cumulative return on investment. This can be represented as an uplift curve (Eqn. 8). A point $\rho \in [0, 1]$ on the X-axis of Figure 3 corresponds to the top ρ fraction of customers as sorted by decreasing order of bandit score. We normalize lift to represent the per-customer average outcome amongst the model's top ρ fraction of ranked examples.

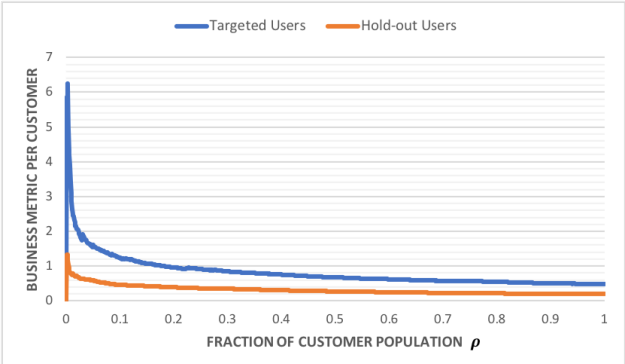
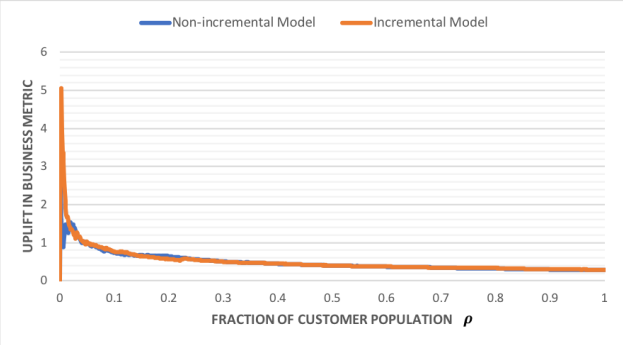
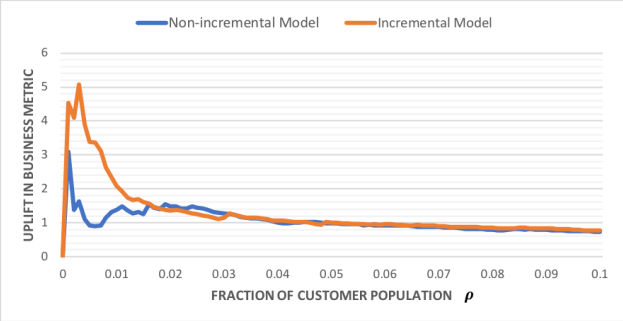


Figure 3. Causal model lift curves for targeted and hold-out customers. The difference between the curves at point ρ on X-axis is the uplift in business metric at that point.

The difference between the two normalized lift curves for treatment (targeted) and control (hold-out) customers is the



(a) Entire population



(b) Top 10% ranked population

Figure 4. Comparing uplift in business metric by incremental and non-incremental approaches.

uplift curve, plotted in Figure 4(a). The incremental model dominates almost entirely, and uplift differences can be better distinguished in the top 10% targetable population (Figure 4(b)), stressing the importance of modeling incrementality. The dip at zero is natural because there is zero total reward to be obtained without selecting any customers.

To understand the relationship between Figures 2 and 3, one can compare the performance on targeted customers in both figures. We see that the point estimate at $\rho = 0.1$ of Figure 3 is the same as the per customer business metric in decile 10 of Figure 2, as both measure the average value of the business metric over the targeted 10% customers with highest predictive scores. The point estimate at $\rho = 0.2$ in Figure 3 is the average of estimates in deciles 9 and 10 of Figure 2 (top 20% customers) and so on. In other words, Figures 2 and 3 represent the non-cumulative and cumulative versions of the same business metric, corresponding to CATE and uplift respectively.

At $\rho = 0.01$, the incremental model has an uplift of 209 units of business metric, compared to 108 units for the non-incremental model. At $\rho = 0.1$, the incremental model has an uplift of 74 units of business metric compared to 70 units for the non-incremental model. For the whole population $\rho = 1$, the uplift values are 28.2 and 26.7 units respectively. Such high values, specifically for the top percentile, showcase the potential of our method.

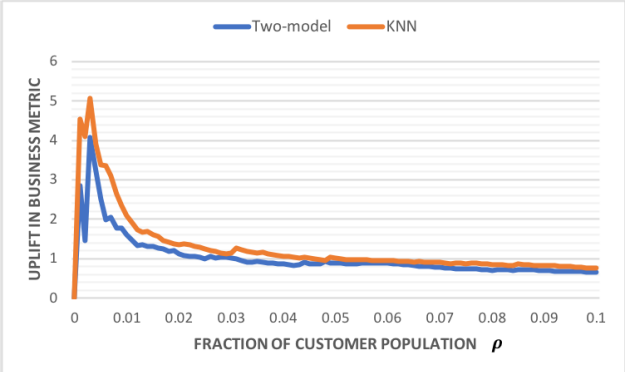


Figure 5. KNN-based approach to CATE estimation yields better uplift than the two-model approach overall, particularly in the 10% of scored population.

5.2. KNN Context Matching

To test the effect of our counterfactual estimation, we replace the KNN-based CATE estimator of incremental model by a two-model alternative extended to multiple campaigns. For the latter, we train two separate linear regression models to predict the business metric for treatment and control customers ($\hat{\mu_T}$ and $\hat{\mu_C}$ in Eqn. 5). After scoring each customer with both models, we determine the campaign with the highest difference $\hat{\mu_T}(\mathbf{x}) - \hat{\mu_C}(\mathbf{x})$. Using the same offline evaluation technique as before, we determine and plot the uplift curve.

The KNN uplift curve tends to dominate its two-model counterpart especially in the top 10% (Fig. 5), where the two-model performance fluctuates. We hypothesize two reasons as to why KNN leads to superior incremental effects. First, the KNN approach determines causal effect with respect to other marketing campaigns, while the two-model approach compares only to a no-targeting control option. Thus, KNN approach may be more comprehensive in counterfactual estimation. Second, KNN can represent complex non-linear relationships while the current two-model approach is limited to linear.

6. Conclusion and Future Work

We present a multi-armed bandit approach that optimizes advertising campaign targeting based on incremental or causal outcomes. We present proof-of-concept results using an offline fashion marketing dataset. We compare our causal approach to non-causal alternatives, and observe that our approach dominates in terms of incremental outcomes in targeted customers over a random hold-out group.

We are currently investigating a few improvements. First, we plan to utilize a consistent baseline across all marketing campaigns instead of varying baselines ($W_i \neq k$) in Section 3. We are also interested in explicitly modeling the trade-off between short-term and long-term objectives as a composite objective based on a weighted average, like: $\beta \times P(click) \times \mathbb{E}(metric|click) + (1-\beta) \times \mathbb{E}(incremental\_metric|impression)$. This is important in developing generic campaign management frameworks where the notion and importance of short-term and long-term objectives may be different. Finally, we plan on deploying our system at a larger scale to enable further experimentation, assess impact and fine tune our methodology.

References

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., Sen, S., and Slivkins, A. A multiworld testing decision service. *CoRR*, abs/1606.03966, 2016.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 127–135, Atlanta, Georgia, 2013. JMLR.

Athey, S. and Imbens, G. Machine learning methods for estimating heterogeneous causal effects. In *arXiv preprint arXiv:1504.01132*, 2015.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. In *Proceedings of the 28th Neural Information Processing Systems Conference (NIPS)*, pp. 1342–1350, 2015.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine learning*, 5 (1):1–122, 2012.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Chen, Y., Pavlov, D., and Canny, J. Large-scale behavioral targeting. In *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 209–218, 2009.

Dani, V., Hayes, T.P., and Kakade, S. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 355–366, Helsinki, Finland, 2008.

Graepel, T., Candela, J.Q., Borchert, T., and Herbrich, R. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 13–20, 2010.

Hansotia, B. and Rukstales, B. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46, 2002.

Hill, D.N., Nassif, H., Liu, Y., Iyer, A., and Vishwanathan, S.V.N. An efficient bandit algorithm for realtime multivariate optimization. In *Proceedings of the 23rd International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1813–1821, 2017.

Imbens, G. and Rubin, D. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA, 2015.

Kuusisto, F., Santos Costa, V., Nassif, H., Burnside, E., Page, D., and Shavlik, J. Support Vector Machines for Differential Prediction. In *European Conference on Machine Learning (ECML-PKDD)*, pp. 50–65, 2014.

Li, L., Chen, S., Kleban, J., and Gupta, A. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 929–934, 2015.

Malkov, Y. and Yashunin, D. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, abs/1603.09320, 2016.

Nassif, H., Santos Costa, V., Burnside, E., and Page, D. Relational differential prediction. In *European Conference on Machine Learning (ECML-PKDD)*, pp. 617–632, 2012.

Nassif, H., Kuusisto, F., Burnside, E., Page, D., Shavlik, J., and Santos Costa, V. Score as you lift (SAYL): A statistical relational learning approach to uplift modeling. In *European Conference on Machine Learning (ECML-PKDD)*, pp. 595–611, 2013a.

Nassif, H., Kuusisto, F., Burnside, E.S., and Shavlik, .J. Uplift modeling with ROC: An SRL case study. In *Proceedings of the International Conference on Inductive Logic Programmin (ILP)*, pp. 40–45, 2013b.

Nassif, H., Cansizlar, K.O., Goodman, M., and Vishwanathan, S.V.N. Diversifying music recommendations. In *Proceedings of the Machine Learning for Music Discovery Workshop at the 33rd International Conference on Machine Learning (ICML)*, 2016.

Radcliffe, N.J. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Journal*, 1:14–21, 2007.

Radcliffe, N.J. and Surry, P. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions, 2011.

Rubin, D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Rubin, D. and Waterman, R. Estimating the causal effects of marketing interventions using propensity score methodology. In *Statistical Science 2006, Vol. 21, No. 2, 206-222*, 2015.

Shalit, U., Johansson, F., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

Sharma, A., Hofman, J., and Watts, D. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the 16th ACM Conference on Economics and Computation*, pp. 453–470, 2015.

Sugiyama, M., Nakajima, S., Kashima, H., Bünau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS)*, pp. 1433–1440, 2007.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16: 1731–1755, 2015a.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML'15, pp. 814–823, 2015b.

Teo, C.H., Nassif, H., Hill, D., Srinavasan, S., Goodman, M., Mohan, V., and Vishwanathan, S.V.N. Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, pp. 35–38, 2016.

Victor, S. and Lo, Y. The true lift model - a novel data mining approach to response modeling in database marketing. *SIGKDD Explorations*, 4(2):78–86, 2002.

A. Appendix

Proof. We now prove Eqn. 9. Let H denote the binary conversion indicator for the action we want to drive, then:

$$\begin{aligned}
&\mathbb{E}[Y|X=\mathbf{x},W=k] \\
&\quad= \int y \times Pr(y|X=\mathbf{x},W=k)\,\mathrm{d}y \\
&\quad= \int y \times [\sum_{h\in\{0,1\}} Pr(y,H=h|X=\mathbf{x},W=k)]\mathrm{d}y \\
&\quad= \int y \times [\sum_{h\in\{0,1\}} Pr(y|H=h,X=\mathbf{x},W=k) \\
&\qquad\times Pr(H=h|X=\mathbf{x},W=k)]\mathrm{d}y \\
&\quad= \sum_{h\in\{0,1\}} [Pr(H=h|X=\mathbf{x},W=k) \\
&\qquad\times \int y \times Pr(y|H=h,X=\mathbf{x},W=k)\,\mathrm{d}y] \\
&\quad= \sum_{h\in\{0,1\}} [Pr(H=h|X=\mathbf{x},W=k) \\
&\qquad\times \int y \times Pr(y|H=h,X=\mathbf{x})\,\mathrm{d}y].
\end{aligned} \tag{18}$$

The last step is simplified based on the assumption that Y is conditionally independent of W given H and X. For $W=0$, we can similarly show that:

$$\begin{aligned}
&\mathbb{E}[Y|X=\mathbf{x},W=0] = \\
&\qquad\sum_{h\in\{0,1\}} [Pr(H=h|X=\mathbf{x},W=0) \\
&\qquad\times \int y \times Pr(y|H=h,X=\mathbf{x})\,\mathrm{d}y].
\end{aligned} \tag{19}$$

Recall binary CATE: $\hat{\tau}_k(\mathbf{x}) = E(Y|X=\mathbf{x},W=k) - E(Y|X=\mathbf{x},W=0)$. Subtract Eqn. 19 from Eqn. 18:

$$\begin{aligned}
&\hat{\tau}_k(\mathbf{x}) = \int y \times Pr(y|H=1,X=\mathbf{x})\,\mathrm{d}y \\
&\times [Pr(H=1|X=\mathbf{x},W=k) - Pr(H=1|X=\mathbf{x},W=0)] \\
&+ \int y \times Pr(y|H=0,X=\mathbf{x})\,\mathrm{d}y \\
&\times [Pr(H=0|X=\mathbf{x},W=k) - Pr(H=0|X=\mathbf{x},W=0)]
\end{aligned} \tag{20}$$

Simplifying and re-arranging,

$$\begin{aligned}
&\hat{\tau}_k(\mathbf{x}) = \\
&(Pr(H=1|X=\mathbf{x},W=k) - Pr(H=1|X=\mathbf{x},W=0)) \\
&\times (\mathbb{E}[Y|X=\mathbf{x},H=1] - \mathbb{E}[Y|X=\mathbf{x},H=0]).
\end{aligned} \tag{21}$$

□