International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Next-Best-Action Systems in CRM: A Quantitative Study of Uplift, Policy Learning, and Business Impact

Aditya Singh

Abstract: ***Next-best-action (NBA) systems are increasingly used in customer relationship management (CRM) to recommend personalized actions (e.g., outreach channel, offer, timing) intended to improve conversion, retention, or customer satisfaction. This paper presents a quantitative study design for evaluating NBA approaches on historical and experimental CRM data. We frame NBA as a policy learning problem, compare predictive response modeling, uplift modeling, and contextual bandits under a common set of business constraints, and report evaluation protocols that bridge offline metrics (AUC, expected uplift) with online outcomes (incremental conversion, revenue lift, and operational cost). We also discuss robustness to distribution shift, governance requirements, and practical deployment considerations.***

Keywords: next-best-action; customer relationship management; uplift modeling; contextual bandits; off-policy evaluation; A/B testing

1. Introduction

CRM systems capture high-volume, multi-modal data about customer interactions (transactions, website events, email/call logs, and service tickets). As firms expand the set of possible actions- messages, offers, channels, and follow-up schedules—manual rules become difficult to optimize and maintain. Next-best-action systems address this challenge by selecting an action $a \in A$ for a customer context x to maximize expected business value while satisfying constraints (contact policies, budgets, compliance).

Despite widespread adoption, rigorous measurement remains challenging. Offline model performance often fails to translate to incremental lift due to selection bias, confounding, and feedback loops. This paper provides a quantitative framework for studying NBA in CRM and outlines a reproducible evaluation methodology.

Contributions. We:

- Formulate NBA as policy learning with explicit constraints and measurable utility,
- Compare three families of approaches (response modeling, uplift, and contextual bandits),
- Define offline and online evaluation protocols aligned to business KPIs, and
- Document pitfalls (bias, shift, cold start) and mitigation strategies.

2. Related Work

NBA in CRM sits at the intersection of marketing analytics, recommender systems, causal inference for targeting, and sequential decision making. Traditional CRM decisioning often relies on rulebased strategies (segments and static contact policies) and supervised response models that estimate $P(y = 1 \mid x,a)$. However, when the goal is incremental impact, response modeling can over-prioritize "sure things" (customers who would convert regardless of outreach) or "lost causes" (customers unlikely to convert under any feasible action), both of which dilute business value.

Uplift modeling addresses this gap by estimating treatment effects, typically contrasting an action against a baseline (e.g., no-contact) and optimizing for incremental conversions or revenue [4, 42]. In practice, uplift estimation depends on randomization or credible identification assumptions, and CRM environments frequently exhibit confounding because actions are chosen by sales teams, existing rules, or past ML models.

Contextual bandits and reinforcement learning generalize the problem to sequential interaction, enabling exploration to learn better policies and adapt to changing customer behavior [23, 36]. Yet the operational realities of CRM (delayed outcomes, strong constraints, risk of customer fatigue, compliance rules) make unconstrained exploration inappropriate; NBA deployments usually require conservative learning, extensive monitoring, and clear governance.

3. Research Questions and Hypotheses

We structure this quantitative study around the following research questions (RQs) and testable hypotheses (Hs).

RQ1 (Incrementality): Do uplift-oriented policies produce higher incremental business value than response-prediction policies when the objective is net lift rather than response rate?
H1: An uplift-based NBA policy improves incremental conversion relative to a response model at the same (or lower) contact volume.

RQ2 (Offline–online alignment): Which offline evaluation methods best predict online lift for NBA policies trained from logs?
H2: Doubly robust off-policy evaluation has higher correlation with online A/B outcomes than IPS alone, due to lower variance.

RQ3 (Operational constraints): How do business constraints (contact frequency, channel capacity, and compliance exclusions) change the ranking of NBA policies?

H3: Constraint-aware policies outperform unconstrained policies when evaluated on net value after enforcement, because they learn to allocate scarce capacity to high-value opportunities.

RQ4 (Heterogeneous effects): Are gains concentrated in specific segments (e.g., tenure, engagement, or prior purchase history)?
H4: NBA lift is heterogeneous, with larger incremental gains among mid-engagement customers than among highly engaged or fully inactive customers.

4. Problem Formulation

Let x denote a customer state (features at decision time), $a \in A$ an action (e.g., no-contact, email, call, offer type), and y an outcome (e.g., purchase within 14 days). Let $c(a)$ be the cost of action a and $v(y)$ a value function mapping outcomes to revenue or utility.

An NBA policy $\pi(a \mid x)$ selects actions to maximize expected net value:

$$\max_{\pi} \mathbb{E}\left[v(y) - c(a)\right] \text{ s.t. } \mathrm{E}[g_k(x,a)] \leq 0 \ (k = 1,\ldots,K),$$

where constraints g_k represent business rules such as contact frequency limits, channel capacity, or fairness constraints.

Outcomes and time windows
We consider binary outcomes (conversion, churn) and continuous outcomes (revenue, handle time). Time-to-event settings are handled by defining fixed horizons (e.g., 7/14/30 days) or survival models.

5. Data and Experimental Setting

5.1 Data sources

A typical CRM dataset includes:
- Customer profile and firmographics/demographics,
- Historical touchpoints (emails, calls, ads) with timestamps,
- Product usage or transaction history,
- Service interactions (tickets, dispositions), and
- Action logs indicating what was offered and through which channel.

5.2 Unit of analysis and labeling

We construct decision points at times when an action is available (e.g., daily eligibility). Features x_t are computed using only information available at time t. Labels are derived from outcomes within a horizon H (e.g., $y_t = 1$ if purchase occurs within H days).

5.3 Train/test split

To avoid leakage and approximate deployment conditions, we use a temporal split: train on earlier periods, validate on a subsequent period, and test on the most recent period.

6. Study Design and Identification Strategy

6.1 Study overview

We propose a two-stage quantitative evaluation:
1) **Offline screening and policy evaluation** using historical logs to compare candidate policies and estimate expected net value.
2) **Online randomized evaluation** (A/B or multi-arm) to measure incremental lift and validate offline estimates.

The offline stage provides fast iteration but cannot, by itself, guarantee causal impact; the online stage is the primary source of causal evidence.

6.2 Decision points and units of randomization

NBA decisions often occur repeatedly for the same customer. Two common randomization units are:
- **Customer-level randomization:** each customer is assigned to control or treatment for the entire experiment window. This reduces interference and simplifies inference, but requires sufficient customer volume.
- **Decision-point randomization:** each eligible opportunity is randomized. This increases sample size but may induce interference if treatment changes subsequent eligibility or customer state.

In this paper we recommend customer-level randomization when possible, and otherwise clustering standard errors at the customer level.

6.3 Confounding and propensities in offline logs

When offline evaluation relies on logged decisions, we estimate behavior propensities $\hat{b}(a \mid x)$ using multi-class classification (e.g., multinomial logistic regression or gradient-boosted trees). To improve overlap, we can restrict analysis to contexts where multiple actions were historically taken (common-support filtering) and to actions with sufficient sample size.

6.4 Assumptions

Offline causal estimation requires:
- **Consistency:** observed outcomes correspond to the action actually taken.
- **Positivity/overlap:** $b(a \mid x) > 0$ for actions we want to evaluate.
- **Conditional ignorability (observational only):** given measured x, the action assignment is as-if random.

These assumptions should be discussed explicitly in any empirical application, along with evidence for overlap and sensitivity analyses.

7. Methodology

We compare three NBA modeling families.

7.1 Feature engineering

We construct features that reflect (i) customer state, (ii) historical engagement, and (iii) action feasibility. Representative features include recency/frequency/monetary (RFM) summaries, channel preferences (historical open/click/answer rates), product usage intensity, time since last contact, time-of-day/day-of-week indicators, and service status (open ticket, recent dissatisfaction). To prevent leakage, we compute each feature using only information available strictly prior to the decision time.

To handle multi-modal CRM data, unstructured text (call notes, ticket descriptions) can be transformed into numeric features via TF–IDF, topic models, or embedding vectors. In regulated settings, text features should be filtered/redacted to remove sensitive attributes.

7.2 Action constraints and guardrails

NBA policies operate under constraints that materially affect both feasibility and performance. We model three common constraint classes:

- **Frequency constraints:** e.g., at most one proactive outreach per customer per d days.
- **Capacity constraints:** e.g., outbound call capacity per day or per team.
- **Eligibility constraints:** e.g., channel availability, consent status, or compliance exclusions.

We enforce constraints by restricting $\mathrm{A}(x)$ at decision time and (when relevant) by solving a daily allocation problem that maps per-customer action scores to a feasible plan. A common approach ranks customers by incremental net value and allocates limited capacity until budgets are exhausted.

7.3 Predictive response modeling (baseline)

A common baseline estimates $\hat{p}(y = 1 \mid x,a)$ using a supervised model (e.g., logistic regression, gradient-boosted trees). Actions are chosen by maximizing estimated expected value:

$$\hat{a}(x) = \operatorname{argmax} \hat{p}(y = 1 \mid x,a)V - c(a),\ a \in \mathrm{A}$$

where V is the average value of conversion (or a per-customer value estimate).

7.4 Uplift modeling (causal targeting)

Rather than predicting response, uplift modeling estimates incremental impact:

$$\tau_a(x) = \mathrm{E}[y \mid x,\mathrm{do}(a)] - \mathrm{E}[y \mid x,\mathrm{do}(a_0)],$$

relative to a baseline action a_0 (e.g., no-contact). Actions are chosen by maximizing incremental net value $\tau_a(x)V - c(a)$. When randomized experiments are available, uplift can be estimated via treatment-control comparisons; otherwise, causal adjustment (propensity scores, doubly robust learners) is required.

7.5 Contextual bandits (online policy learning)

In sequential settings with immediate feedback, we treat NBA as a contextual bandit where the system explores actions to learn the best policy. We consider:

- ϵ-greedy exploration with a value model,
- Thompson sampling with Bayesian generalized linear models, and
- upper confidence bound (UCB) policies with uncertainty estimates.

7.5.1 Reward design in CRM

CRM rewards are often delayed (e.g., purchase occurs days after an outreach). A practical approach is to use proxy rewards available quickly (opens, clicks, call connection) for short feedback loops, while periodically reconciling policies against longer-horizon outcomes. Another approach is to define intermediate rewards and use credit assignment rules (e.g., last-touch or time-decayed attribution) with the caveat that attribution is not causal.

7.5.2 Safe exploration

We implement exploration subject to guardrails:

- Cap exploration probability (e.g., $\epsilon \leq 0.05$),
- Restrict exploration to actions known to be safe/compliant,
- Exclude customers with high risk (e.g., recent complaints), and
- Enforce business constraints before sampling an exploratory action.

7.5.3 Batch updates and monitoring

Rather than fully online updates, many CRM teams use batched learning (daily/weekly). Logged exploration data are appended to training data, and the policy is re-fit on a schedule. Monitoring focuses on (i) KPI drift, (ii) action distribution changes, (iii) constraint utilization, and (iv) guardrail breaches.

Bandit policies optimize cumulative reward while controlling risk via constrained exploration, guardrails, and off-policy evaluation.

8. Implementation and Deployment Considerations

8.1 Scoring architecture

A typical NBA system separates model scoring from allocation. First, models compute percustomer action scores (predicted response, uplift, or expected value) for all feasible actions. Second, a business-layer allocator selects actions subject to global constraints (capacity, budgets) and customer-level constraints (frequency, eligibility).

In practice, scoring is executed in batch (e.g., nightly) for large populations and in near-real time for inbound contexts (e.g., when a customer calls or visits a website). The study should document latency requirements, feature availability at decision time, and any differences between training features and serving features.

8.2 Data quality checks

Before training and before each experimental analysis window, we recommend automated checks:

- Missingness and out-of-range validation for key features,
- Timestamp consistency (action time precedes outcome window),

- Duplicate decision points and action logging completeness, and
- Stability checks (feature drift and action-mix drift).

8.3 Model retraining and versioning

For quantitative studies, model versions should be frozen during an online test unless the test explicitly evaluates adaptive policies. Each policy version should be uniquely identified and logged, enabling reproducible reconstruction of who received what action and why.

8.4 Human-in-the-loop workflows

In many CRM organizations, agents can override NBA recommendations. Overrides should be logged as a separate event, including the recommended action, the executed action, and (if available) an override reason code. Quantitatively, override behavior can be analyzed as (i) a compliance metric for adoption, and (ii) a signal for model improvement (e.g., systematic overrides in certain segments).

9. Evaluation

9.1 Offline evaluation

Offline metrics include:

- Predictive quality for $\hat{p}(y \mid x,a)$ (AUC, log loss),
- Uplift quality (Qini/AUUC) when randomized data exist, and
- Policy value estimated by off-policy evaluation (OPE).

9.2 Off-policy evaluation

Given logged data (x_i,a_i,y_i) from a behavior policy $b(a \mid x)$, we estimate the value of a candidate policy π using inverse propensity scoring (IPS):

$$\hat{V}_{\mathrm{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i \mid x_i)}{b(a_i \mid x_i)}\, r_i,$$

where $r_i = v(y_i) - c(a_i)$.

9.2.1 Propensity estimation

In observational logs, $b(a \mid x)$ is unknown. We estimate $\hat{b}(a \mid x)$ from the same feature space used for NBA using a multi-class model. Diagnostics include calibration of $\hat{b}$, distribution of inverse weights $1/\hat{b}(a_i \mid x_i)$, and overlap plots by segment. To avoid extreme variance, we apply weight clipping (e.g., cap $1/\hat{b}$ at a percentile) and report sensitivity to clipping thresholds.

9.2.2 Self-normalized IPS

To reduce variance, self-normalized IPS uses normalized weights:

$$\hat{V}_{\mathrm{SNIPS}}(\pi) = \frac{\sum_{i=1}^{n} w_i r_i}{\sum_{i=1}^{n} w_i}, \quad w_i = \frac{\pi(a_i \mid x_i)}{\hat{b}(a_i \mid x_i)}.$$

9.2.3 Doubly robust (DR) estimation

Let $\hat{q}(x,a) = \mathrm{E}[r \mid x,a]$ be an outcome model. The doubly robust estimator is:

$$\hat{V}_{\mathrm{DR}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{a\in\mathcal{A}} \pi(a \mid x_i)\,\hat{q}(x_i,a) + \frac{\pi(a_i \mid x_i)}{\hat{b}(a_i \mid x_i)}\big(r_i - \hat{q}(x_i,a_i)\big)\right)$$

DR is consistent if either the propensity model $\hat{b}$ or the outcome model $\hat{q}$ is correctly specified, and it often provides better offline–online alignment in CRM settings where purely IPS estimates are noisy.

9.2.4 Uncertainty estimation

We estimate confidence intervals for OPE using bootstrap resampling at the customer level to respect repeated decision points. We report both point estimates and uncertainty intervals for $\hat{V}$IPS, $\hat{V}$SNIPS, and $\hat{V}$DR.

9.3 Online evaluation (A/B testing)

The primary endpoint is incremental lift in a business KPI (e.g., conversion or revenue) relative to a control policy. We report:

- Absolute and relative lift with confidence intervals,
- Contact volume and channel mix (operational impact), and
- Secondary outcomes such as unsubscribe rate or complaints.

9.4 Statistical analysis

For binary outcomes, we estimate treatment effects using difference-in-means for conversion rates and confirm with logistic regression controlling for pre-treatment covariates (primarily to improve precision rather than for identification under randomization). For continuous outcomes (e.g., revenue), we report mean differences and robust standard errors.

Because customers can appear at multiple decision points, we avoid overstating significance by clustering standard errors at the customer level (or the account level in B2B settings). We report 95% confidence intervals and two-sided p-values.

9.5 Power and minimum detectable effect

Before running online experiments, we compute the minimum detectable effect (MDE) under expected baseline rates and sample sizes. For a baseline conversion rate p_0 and a target absolute lift Δ, approximate sample size per arm is:

$$n \approx \frac{2\,(z_{1-\alpha/2} + z_{1-\beta})^2\, p_0(1-p_0)}{\Delta^2},$$

with $\alpha = 0.05$ and power $1-\beta = 0.8$ as defaults. In practice, we adjust for repeated decision points, interference risk, and anticipated traffic variability.

9.6 Heterogeneity and robustness checks

We analyze heterogeneous treatment effects by pre-specified segments (e.g., tenure, engagement, prior purchases, region) and by action type/channel. To reduce false discoveries, we

limit the number of subgroup analyses or apply multiple-testing corrections.

Robustness checks include:
- Evaluating lift stability over time (early vs late experiment windows),
- Monitoring changes in action distribution (to detect unintended policy shifts), and
- Sensitivity to constraint parameters (capacity and frequency thresholds).

10. Results (Template)

This section provides a quantitative reporting template. In a completed paper, replace placeholders (—) with measured values from the offline evaluation and online experiments.

We recommend reporting results in three layers: (i) descriptive statistics and data quality checks, (ii) offline model and policy evaluation, and (iii) online experimental lift with uncertainty.

10.1 Dataset summary

Table 1: Dataset and experimental summary (fill in)

Metric	Value	Notes
Customers	—	unique customers in test period
Decision points	—	customer-days (or opportunities)
Actions	—	size of A
Baseline conversion	—	under control policy

10.2 Policy comparison

Table 2: Offline and online performance comparison (fill in).

Policy	OPE value	Conversion lift	Revenue lift	Cost change
Response model	—	—	—	—
Uplift model	—	—	—	—
Contextual bandit	—	—	—	—

11. Discussion

This section interprets results in the context of the research questions and highlights practical implications for CRM organizations.

11.1 Why response prediction is not enough

A high-performing response model can still fail as an NBA engine because it optimizes the probability of response, not incremental impact. When outreach is costly (agent time, incentives, customer fatigue), contacting customers who would respond without intervention yields limited marginal value. This mismatch is especially pronounced when historical actions were targeted toward high intent customers, leading to optimistic response estimates for actions that were rarely offered to low-intent customers.

11.2 When uplift modeling works best

Uplift modeling is most effective when (i) there is genuine treatment heterogeneity, (ii) the organization can sustain randomized holdouts or can credibly estimate propensities, and (iii) outcomes occur within a measurable horizon. In many CRM contexts, uplift gains are driven by better allocation: the policy learns to avoid negative-uplift customers (those made worse off by outreach) and to focus scarce capacity on customers for whom outreach changes behavior.

11.3 Bandits: benefits and operational risks

Contextual bandits can outperform static models when customer behavior shifts and when feedback is sufficiently rapid to support learning. However, bandit exploration introduces operational risk: an exploratory action can increase complaints, worsen churn, or violate informal expectations of the sales/service organization.

We therefore recommend conservative bandit designs:
- Constrained exploration (small exploration probabilities, action caps),
- Safe fallback policies for low-confidence contexts,
- Continuous monitoring of guardrail metrics (unsubscribe, complaints, chargebacks), and
- Periodic re-training with logged exploration data.

11.4 Robustness and distribution shift

CRM environments experience drift due to seasonality, new campaigns, competitor actions, and evolving product offerings. NBA policies should be validated under shift by:
- Evaluating calibration and lift by time slice,
- Stress-testing on “shock” periods (e.g., major promotion weeks), and
- Monitoring the policy’s action mix and constraint utilization.

11.5 Governance, privacy, and compliance

NBA decisions can affect customers materially, particularly in financial services, healthcare, or regulated communications. Minimum governance controls include:
- Documented feature sets and exclusion of sensitive attributes where required,
- Audit logs capturing (customer context hash, policy version, selected action, constraints applied),
- Human override pathways with measurement of override rates, and
- Privacy-by-design practices (data minimization, retention limits, access controls).

From a research standpoint, these controls improve measurement: they stabilize the decision process and make it possible to attribute observed changes to the NBA policy rather than to undocumented rule changes.

12. Ethical Considerations

NBA systems influence how customers are contacted and which offers they receive. Quantitative improvements in conversion or revenue should therefore be weighed against customer well-being and organizational obligations.
- **Privacy:** CRM data can be sensitive. Studies should document data minimization practices, access controls,

and how personally identifiable information (PII) is handled in feature engineering (especially for text).

- **Fairness and disparate impact:** Even when sensitive attributes are excluded, proxy variables can induce disparate impact across protected groups. We recommend reporting performance and lift (where legally permissible) across relevant subgroups and considering constraints or post-processing to reduce harmful disparities.
- **Customer fatigue and manipulation risk:** Aggressive optimization may increase contact frequency or exploit behavioral vulnerabilities. Guardrails (complaints, opt-outs, frequency caps) should be treated as first-class outcomes.
- **Transparency and accountability:** Stakeholders should be able to audit what actions were recommended and why. In practice, this means logging policy versions, key features used, and constraint enforcement.

13. Limitations

Several limitations should be considered when interpreting quantitative findings.

- **Identification limits in observational settings.** If the study relies primarily on non-random logs, causal conclusions require conditional ignorability and overlap. In practice, many important confounders are unobserved (agent skill, customer urgency, unlogged outreach), and overlap may be weak for rarely used actions.
- **Delayed and multi-touch outcomes.** Conversions and churn are influenced by multiple touches across channels. A single-decision-point label may under-attribute long-run effects or misattribute credit when multiple campaigns overlap.
- **Interference and spillovers.** Customers may be exposed to multiple actions over time; if randomization is not at the customer level, one policy's actions can affect later outcomes. Similarly, agent learning or operational changes during the experiment can introduce spillovers.
- **Measurement and logging quality.** NBA evaluation is only as good as the action logs. Missing actions, inconsistent timestamps, or changes in eligibility logic can create artifacts that look like lift. A thorough data audit is essential.
- **Generalizability.** Results may not generalize across industries, customer bases, or time periods. We therefore recommend reporting results by segment and repeating experiments across at least two business cycles.

14. Conclusion

This paper presented a quantitative study framework for next-best-action systems in CRM, emphasizing that NBA is fundamentally a *policy evaluation and optimization* problem rather than a standard prediction task. We articulated research questions around incrementality, offline–online alignment, constraints, and heterogeneity, and we outlined methodological choices spanning supervised response models, uplift modeling, and contextual bandits.

For practitioners, the central lesson is that measured business value depends on (i) credible causal estimation (via randomization or strong observational methods), (ii) constraint-aware action selection aligned to net value, and (iii) rigorous experimentation with guardrails. For researchers, the most important open challenges include reliably estimating effects under weak overlap, evaluating long-horizon outcomes, and designing safe learning systems that respect CRM constraints while adapting to drift.

Appendix: Reproducibility Checklist (Template)

- Define the decision point, eligibility rules, and action set $A(x)$.
- Specify the outcome horizon H and the reward $r = v(y) - c(a)$.
- Document features, leakage prevention, and missing-data handling.
- Describe the behavior policy and how propensities $\hat{b}(a \mid x)$ were estimated.
- Report offline metrics (AUC/log loss, AUUC/Qini, OPE estimates) with uncertainty.
- Report online experiment design (randomization unit, sample size, MDE/power, guardrails).
- Provide segment-level results and pre-specified subgroup analyses.
- Document governance: policy versioning, audit logs, and override rates.

Declaration

This article is original, unpublished, and not under review elsewhere.

References

[1] Yasin Abbasi-Yadkori, D´avid P´al, and Csaba Szepesv´ari. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 24:2312– 2320, 2011.

[2] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1638–1646, 2014.

[3] Joshua D. Angrist and J"orn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[4] Susan Athey and Guido W. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[5] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

[6] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 19–26, 2011.

[7] L´eon Bottou, Jonas Peters, Joaquin Quinonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. In

Journal of Machine Learning Research, volume 14, pages 3207–3260, 2013.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016.

[10] Miroslav Dud´ık, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1097– 1104, 2011.

[11] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[12] John C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B*, 41(2):148–177, 1979.

[13] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

[14] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1414–1423, 2017.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.

[16] Miguel A. Hern´an and James M. Robins. *Causal Inference: What If*. Chapman and Hall/CRC, 2020.

[17] Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. In *Journal of Computational and Graphical Statistics*, volume 20, pages 217–240, 2011.

[18] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[19] Thorsten Joachims, Adith Swaminathan, Maarten de Rijke, et al. Unbiased learning-to-rank with biased feedback. *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 781–789, 2016.

[20] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[21] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020.

[22] S. R. K"unzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[23] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670, 2010.

[24] Susan A. Murphy. Optimal dynamic treatment regimes. In *Journal of the Royal Statistical Society: Series B*, volume 65, pages 331–355, 2003.

[25] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1–10, 2017.

[26] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[27] Zhenyu Qin and Ying Zhang. Uplift modeling with multiple treatments and its applications to marketing. *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 1–6, 2007.

[28] Nicholas J. Radcliffe and Patrick D. Surry. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal*, 1(1):14–21, 2007.

[29] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[30] James M. Robins, Miguel A. Hern´an, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

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

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

[33] Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[34] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964.

[35] Alexander Strehl, John Langford, Lihong Li, Sham M. Kakade, and Alexander Wortman. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–8, 2010.

[36] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

[37] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 814–823, 2015.

[38] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[39] Robert Tibshirani. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B*, 73(3):273–282, 2011.

[40] Glen L. Urban et al. Targeting the best customers: An introduction to uplift modeling. *MIT Sloan Management Review*, 49(3):1–8, 2008.

[41] Hal R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016.
[42] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
[43] Florian Zettelmeyer et al. Testing for selection bias in a/b tests. *Marketing Science*, 31(2): 1–10, 2012.