Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Predictive Analytics to Intelligent Intervention: Integrating Causal Machine Learning, Treatment Optimization, and Decision Utility Assessment

Sai Srujan Reddy Tokala

saisrujanreddytokala@creatrox.info

Clark University https://orcid.org/0009-0004-1924-8417

Research Article

Keywords: Causal Machine Learning, Decision-Centric AI, Heterogeneous Treatment Effects, Policy Learning, Uplift Modeling

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11767181/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

From Predictive Analytics to Intelligent Intervention: Integrating Causal Machine Learning, Treatment Optimization, and Decision Utility Assessment

*Sai Srujan Reddy Tokala

Clark University, Worcester, United States, MA 01610

E-mail: saisrujanreddytokala@creatrox.info

ORCID: 0009-0004-1924-8417

Abstract

Predictive accuracy does not establish that an intervention is effective. This study proposes a decision-centric artificial intelligence workflow combining predictive modelling, SHAP explainability, heterogeneous treatment-effect estimation, policy learning, and intervention-utility evaluation. Four predictive models (Logistic Regression, Random Forest, XGBoost, and LightGBM) and six causal and uplift estimators (S-Learner, T-Learner, X-Learner, LinearDML, Causal Forest, and Random Forest Two-Model Uplift) were evaluated using a treatment-balanced sample of 500,000 observations from the Criteo benchmark. Binary visit was the primary outcome, while conversion was used descriptively. XGBoost and LightGBM achieved validation AUC-ROC values of 0.94467 and 0.94455, respectively. The Spearman rank correlation between XGBoost predicted visit probabilities and Causal Forest estimated treatment effects was 0.3942 (95% bootstrap CI: 0.3873–0.4020), indicating a moderate association between predictive response and estimated incremental treatment benefit. At a 30% intervention budget, full-test-set policy values were 0.047333 for XGBoost and 0.046827 for Causal Forest, with corresponding net benefits of 0.045833 and 0.045327. In a separate 20,000-observation treatment-stratified evaluation subset, paired-bootstrap comparisons produced a Causal Forest–XGBoost policy-value difference of 0.001200 (95% CI: −0.000903 to 0.003600) and a net-benefit difference of 0.001177 (95% CI: −0.000945 to 0.003562). Neither

difference was statistically distinguishable from zero. These findings demonstrate that predictive discrimination, treatment-effect ranking, and intervention utility represent distinct evaluation objectives. The results provide empirical evidence of their complementary roles within the Criteo benchmark but do not establish universal causal-policy superiority or cross-domain generalizability.

Keywords: Causal Machine Learning; Decision-Centric AI; Heterogeneous Treatment Effects; Policy Learning; Uplift Modeling

1 INTRODUCTION

AI has made remarkable progress in many sectors, such as healthcare, finance, marketing, manufacturing, and public administration. The typical measures for assessing the effectiveness of AI systems include predictive performance metrics like accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). These are the most common measures used in machine learning and practice for measuring the predictive power of a model. While prediction is helpful, there are many things in the real world where more than prediction is needed. To understand how to take action to enhance the outcomes instead of simply predicting what is likely to happen. More recently, there has been a move from a prediction-driven to a decision-driven and prescriptive AI paradigm, which explicitly supports intervention planning and decision-making. Research has emphasized that beyond predictive analytics, it is crucial to shift to actionable intelligence where machine learning systems are assessed for their impact on decision-making and the measurable value of interventions (Lo and Pachamanova, 2023, Sun et al., 2024, Ridwan, 2025). Recent research involving human-AI collaboration has shown that AI systems are increasingly expected to be involved in decision-making processes and not only as prediction engines (Hao et al., 2024).

Recent developments in machine learning have further demonstrated the importance of evaluating predictive systems in relation to transparency, contextual adaptation, and their practical decision-support capabilities. Nelli et al. (2026) developed a mathematical model for the joint optimization of predictive accuracy and interpretability by using explanation techniques like SHAP, integrated gradients and LIME. Their findings highlight the importance of considering interpretability alongside predictive effectiveness when developing trustworthy AI systems. Similarly, Deshpande et al. (2026a) examined comparative performance and computational efficiency across Transformer-based language models and multiple NLP benchmarks, illustrating the importance of evaluating machine-learning architectures using complementary performance and efficiency criteria. These developments reinforce the need to assess AI systems through multiple dimensions rather than relying exclusively on predictive discrimination. In the context of intervention-oriented AI, however, interpretability and predictive performance must also be distinguished from the estimation of treatment effects and the empirical utility of intervention decisions.

Although considerable progress has been made in the field of predictive modelling, there is one major constraint. Traditional machine learning algorithms are all correlational approaches. They look for statistical relationships between characteristics and results, but they do not imply that one particular intervention is responsible for a change in behavior or performance. Therefore, those who are likely to convert, respond or fail are not necessarily the most likely to benefit from an intervention. In applied settings such as targeted marketing, healthcare treatment allocation, risk management, and public policy, decision makers may need to distinguish between individuals who are likely to experience an outcome and those who are expected to benefit from an intervention. However, the applicability of decision-centric causal AI approaches in these domains depends on domain-specific data, treatment-assignment mechanisms, causal identification assumptions, outcome definitions, and operational

constraints. The present study evaluates these concepts exclusively within a randomized digital advertising benchmark and does not empirically validate their effectiveness in healthcare or public-policy settings. Recent studies have thus suggested that predictive accuracy is not enough for making decisions, but causal reasoning is needed when the goal is to predict the impact of actions rather than the results (Doutreligne and Varoquaux, 2025, Feuerriegel et al., 2024). This change is part of a larger trend in the application of AI from predicting to intervening, moving toward personalized and evidence-based decisions (Vallee, 2026).

The distinction between outcome prediction and incremental effect of an intervention is a difference between different types of uplift modelling (Wang et al., 2025). More recently, there have been several studies on the Criteo benchmark which demonstrated that treatment-effect evaluation in advertising data is randomized and, in addition, found that ranking was not necessarily correlated with intervention value (Diemert et al., 2018, Devriendt et al., 2018). Actionable machine learning is illustrated through context-aware recommendation research, which includes the research by Deshpande et al. (2025a) that leveraged mobile learning for recommendations, and Deshpande and Mangalwede (2021) using location- and time-aware content delivery, and Deshpande et al. (2026a) adding facial emotion recognition, speech transcription, and natural language processing for customer-satisfaction analysis. These studies illustrate personalized decision support, but forecasted preferences are not causal effects. Benchmarking is a way to separate treatment-effect estimation from policy performance. Rößler and Schoder (2022) evaluated 15 uplift methods, and the Treatment Dummy Approach yielded a Criteo unscaled Qini of 1.7923, Two-Model Random Forest Uplift had an unscaled Qini of 1.7154, and LinearDML had an unscaled Qini of 0.9462. Diemert et al. (2018) reported normalized Qini coefficients of 0.1500 and 0.3120 for Two-Model and Revert-Label approaches. These values are not directly comparable because Qini formulations and experimental configurations differ. Further studies emphasize base-learner selection, sample

splitting, and cross-fitting, alongside operational considerations in personalized uplift (Moraes et al., 2023, Salditt et al., 2024, Betlei et al., 2021). Related research has also examined meta-learners for estimating heterogeneous treatment effects under multi-valued treatment settings, extending the methodological literature beyond binary treatment designs (Acharki et al., 2023). Policy learning translates effects into assignments, Athey and Wager (2021) proposed a doubly robust approach with assignment constraints, while Gubela and Lessmann (2021) showed that treatment-effect rankings need not maximize business value. Related causal machine-learning research has also investigated Double Machine Learning for feature selection in dynamic gesture recognition (Yan et al., 2025), illustrating its application beyond conventional predictive modelling. However, this application differs from uplift estimation and intervention-policy evaluation, which remain the focus of the present study.

Educational predictive studies demonstrate the importance of architecture and context. Recent research has demonstrated the application of Bi-LSTM architectures and SHAP-based interpretation to student-performance prediction using integrated institutional and synthetic educational data (Deshpande et al., 2026b). Deshpande et al. (2026c) investigated Bi-LSTM student-performance prediction using OULAD–AISHE integration and synthetic institutional mapping; a related study examined Indian higher education using synthetic data. Deshpande and Mangalwede (2021) compared adaptive neuro-fuzzy prediction across offline and online settings. They illustrate diverse predictive approaches across contexts, but do not advertise causal effects. Similarly, machine-learning-based sentiment analysis of faculty feedback demonstrates how predictive models can extract actionable educational insights from textual data (Deshpande et al., 2025b). Accurate outcome prediction does not necessarily identify who benefits from an intervention.

Predictive discrimination, treatment-effect ranking, and intervention utility are distinct objectives. This study compares Logistic Regression, Random Forest, XGBoost, and

LightGBM with S-Learner, T-Learner, X-Learner, LinearDML, Causal Forest, and Random Forest-based uplift estimation (Hu, 2023). It uses the 13,979,592-observation Criteo dataset and a treatment-balanced sample of 500,000; binary visit is the primary outcome, while conversion is descriptive. Predicted visit probabilities and estimated effects define policies under a 30% allocation budget. Evaluation includes empirical uplift, normalized Qini and AUUC, inverse-propensity-weighted policy value, simulated net benefit, and bootstrap uncertainty. Analyses examine budget sensitivity, costs, allocation parity across four anonymized-feature groups, and SHAP-assisted targeting. SHAP explains predictive outputs, not causal effects; groups are not demographic or protected categories.

The decision problem in this study is to prioritize individuals for intervention under limited allocation capacity. The primary objective is to maximize expected intervention benefit while accounting for intervention costs and a predefined allocation budget. Predictive policies rank individuals using estimated visit probabilities, whereas causal policies use estimated treatment effects. The primary evaluation considers a 30% intervention budget and a simulated cost of 0.005 successful-visit-equivalent units per intervention. Policy performance is assessed using IPW policy value, net benefit, uplift gain, and Qini measures.

At the 30% allocation budget, Causal Forest achieved an uplift gain of 0.026372 and policy value of 0.046827, compared with 0.026420 and 0.047173, respectively, for LightGBM. Paired-bootstrap comparisons did not establish statistically distinguishable differences in the principal utility measures. The study therefore examines the relationship between predictive and treatment-effect rankings and the influence of allocation budgets, costs, and parity constraints. External validity remains limited because the evaluation uses only the Criteo benchmark.

2 METHODOLOGY

This study employed a quantitative empirical machine-learning design to evaluate the effectiveness of causal machine learning for decision-centric artificial intelligence. The proposed framework builds on predictive modelling, explainable AI, treatment-effect estimation, policy learning, and intervention utility evaluation in a single analytical pipeline. The goal was to evaluate the performance of traditional prediction-centric methods and to estimate the performance of the causal machine learning methods in their ability to inform intervention decision-making and provide actionable feedback.

2.1 Dataset Description and Sampling Procedure

This study used the Criteo Uplift Prediction Dataset derived from randomized digital advertising experiments Diemert et al. (2018). The original dataset had 13,979,592 cases with 2,096,937 controls. There were twelve anonymized numerical features (f0–f11), an exposure indicator, and treatment, visit and conversion outcomes in each observation.

To minimize the number of computations required, a treatment-balanced sample of 500,000 observations was created by stratified random sampling without replacement, with each treatment group containing 250,000 observations. Sampling was done with NumPy RandomState = 42; no outcome or covariates were used for sampling. No weights were applied to the samples, so the results are not population-weighted but are for the analytical sample. No missing values were detected, but records were excluded if required variables were missing. No further exclusions, outlier or duplicate removal occurred.

The sample was divided into three sets: A training set of 70% (n = 350,000), a validation set of 15% (n = 75,000) and a testing set of 15% (n = 75,000) obtained by joint stratification by treatment and visit outcome. Equal treatment representation for each partition was observed. The validation set was used to develop the model, and the test set was used to finalize the

model. Only the visit outcome (binary) was measured for prediction, causal estimation, and policy evaluation, while conversion was only used for descriptive analysis. The treatment was defined on the basis of the randomization (T = 1 versus T = 0), not on the basis of the separate exposure indicator. Therefore, all causal and policy effects reported are for the probability of a visit. For each of the 12 anonymized features, there is no semantic definition, and the features are used as statistical inputs only, not giving any behavioural, demographic or contextual meaning. Analytical sample was stored in the Parquet format with a manifest of type SHA-256.

In this context, the treatment effect refers to the difference in outcomes between the treatment and control groups, whereas the CATE refers to this difference within the groups of students who are similar with respect to the observed covariates. Uplift is the treatment response score, or the assessment of treatment response score that is incrementally greater than the score of the no-treatment group. The policy value represents the expected outcomes of an intervention policy estimated by inverse-propensity weighting (IPW). In addition, for Utility, assumed intervention costs are included, and for targeting efficiency, it refers to the percentage of positive-estimated-CATE observations that are selected by a policy.

The analytical sample consisted of 500,000 observations which were chosen to be a representative randomized sample of both treatment groups. All observations were independently sampled within each treatment arm, with 250,000 observations in each treatment arm. The 50:50 treatment allocation resulting from this differs from the approximate 85:15 distribution in the original data, so the reported estimates are not a population-weighted reconstruction of the original data. Robustness to different sample sizes for the training sample was tested with 100,000, 200,000, and 350,000 observations, and robustness to different random seeds was tested using random-seed numbers 21, 42, and 84. These experiments assess the stability of the given causal-model configurations under different sampling settings but not on the entire original data set. Although treatment-specific sampling preserved equal

representation of both experimental groups, the balanced sample does not reproduce the original treatment prevalence or necessarily preserve all marginal covariate distributions of the complete dataset. No sampling weights were applied; therefore, the findings are interpreted within the constructed analytical sample rather than as population-representative estimates. The available balance diagnostics support measured-covariate comparability across experimental partitions but do not establish full representativeness of the original dataset.

2.2 Research Design

Following the construction of the treatment-balanced analytical sample, the observations were partitioned into training (70%; n = 350,000), validation (15%; n = 75,000), and held-out test (15%; n = 75,000) subsets using a two-stage stratified random-sampling procedure. Stratification was conducted on both assignment to treatment and the binary outcome of visit, resulting in four strata: control without a visit, control with a visit, treatment without a visit, and treatment with a visit. In the first phase, 70% of observations were assigned to the training sample and 30% to an intermediate sample. The intermediate subset was split into validation and test sets of equal size in the second stage. A fixed seed (42) was used for both operations. The twelve anonymized covariates were not included as stratification variables, nor were there any test observations for model fitting and hyperparameter tuning.

The three partitions were treated separately to ensure treatment representation. A total of 175,000 treatment observations and 175,000 control observations were used in the training set; a total of 37,500 observations in each of the treatment conditions were used in the validation set and 37,500 in each of the treatment conditions in the test set. As a result, the ratio of empirical treatment allocation was 50:50 in each partition. It was also found that the frequency of visits to the treatment group was very similar to that of the training control group (0.038166 and 0.047634, respectively), as was also the case for the validation groups (0.038160 and

0.047627, respectively) and test groups (0.038187 and 0.047627, respectively). The data indicate that the treatment–outcome composition was very well maintained within the experimental partitions.

Standardized mean differences (SMDs) were computed for each partition, between treatment and control observations, for each of the twelve anonymized features, to evaluate covariate balance that was not accounted for by the stratification on treatment and outcome. Two-sample Kolmogorov–Smirnov statistic was also used to explore distributional differences. The highest SMDs in terms of absolute value were 0.045263 for the training set, 0.059421 for the validation set, and 0.040116 for the test set. An absolute SMD of 0.10 or less was not exceeded in any of the partitions for any feature. The maximum Kolmogorov–Smirnov statistics were 0.013263, 0.020080, and 0.016960, respectively. The maximum absolute SMDs within each treatment arm were 0.007412 to 0.019572, and the maximum Kolmogorov–Smirnov statistics were 0.004615 to 0.011093. The diagnostics suggest that the measured covariates are comparable within the partitions. They are descriptive checks and do not provide complete multivariate overlap, nor do they provide independent evidence of causal identification. The test partition was held out for final evaluation, while the validation partition was used for model assessment and development of the predefined analytical procedures. This separation was preserved to minimize the threat of information leakage between model development and evaluation of model performance.

2.3 Prediction Models

Four supervised classifiers-Logistic Regression, Random Forest, XGBoost, and LightGBM-were implemented to predict the binary visit outcome using twelve anonymized features (f0–f11). The dataset was partitioned into training (350,000), validation (75,000), and testing (75,000) observations. Logistic Regression provided a linear baseline, while the tree-based

models captured potentially nonlinear relationships (Doutreligne and Varoquaux, 2025, Feuerriegel et al., 2024).

Logistic Regression used L2 regularization (C = 1.0), SAGA solver, and 500 maximum iterations, with training-fitted standardization. Random Forest employed 200 trees, maximum depth 8, and minimum leaf size 50. XGBoost used 300 estimators, depth 6, learning rate 0.05, and row and feature subsampling of 0.8. LightGBM used 300 estimators, depth 6, learning rate 0.05, 63 leaves, minimum leaf size 50, and subsampling parameters of 0.8. LightGBM employed validation-based early stopping (patience 30); XGBoost used validation monitoring without automatic early stopping. Parameters not specified retained library defaults. All models used random seed 42, and hyperparameters were manually specified rather than optimized through automated search.

Uncertainty in predictive performance was assessed using 1,000 stratified bootstrap resamples of the validation observations, with resampling conducted separately within the positive and negative outcome classes. Percentile-based 95% confidence intervals were calculated for AUC-ROC, AUC-PR, and F1-score using fixed model predictions. These intervals quantify evaluation-sample uncertainty conditional on the fitted models and do not incorporate model-retraining uncertainty.

Performance was evaluated using AUC-ROC, average precision (AUC-PR), accuracy, precision, recall, F1-score, and Brier score. Given outcome imbalance, AUC-PR, precision, recall, and F1-score received particular emphasis. Classification metrics used default prediction thresholds without additional optimization. Model configurations are provided in Table 1.

Insert Table 1 Here

2.4 Explainable AI Analysis

The SHapley Additive exPlanations (SHAP) were employed to interpret the predictions of the XGBoost and LightGBM and quantify the influence of twelve anonymous features (f0 - f11) on the predicted probability of visits (Fryer et al., 2021, Ponce-Bobadilla et al., 2024, Wang et al., 2022). Mean absolute SHAP value was used to quantify the importance of each feature in the model across the entire dataset, and summary plots were used to visualize the magnitude and direction of each feature's attribution (Wang et al., 2024).

An identical model-agnostic PermutationExplainer was used for both models. The prediction functions were provided as probabilities for the positive class instead of margins or log-odds. For 2000 randomly selected validation observations, a common background sample of 100 training observations was used, with 25 permutation evaluations per observation, and explanations were generated for these observations. To measure the consistency of the explanations, a consistency test using additive reconstruction checks was performed on the same observations, while mean absolute SHAP values were calculated. This means that magnitudes of attributions are also on the same probability scale, but that differences in magnitudes are not necessarily differences in effects of the causal features but differences in the predictive behaviour of the models.

SHAP also gave an explanation for a predictive targeting heuristic. Absolute standardized deviations of each feature value from the value derived from the reference set of training data were aggregated with the normalized global XGBoost mean absolute SHAP values. Deviations were limited to ten standard deviations, normalised by the 95th percentile of the validation scores and clipped to the range [0,1]. The resulting attribution-informed score was equally weighted (0.5:0.5) with XGBoost predicted visit probabilities. The highest-ranked 30% of observations were selected for intervention, without optimizing weights using held-out outcomes.

Importantly, SHAP represents predictive attribution, not causality. Because the original meanings of f0–f11 are unavailable, their importance is interpreted exclusively as model-dependent statistical contributions. For example, the high attribution of f8 indicates predictive influence under the evaluated configuration but does not establish its behavioural meaning or causal effect. SHAP-assisted targeting was therefore evaluated as a predictive allocation strategy, while heterogeneous treatment effects were estimated separately using causal and uplift models.

2.5 Causal Machine Learning Models

Six causal and uplift estimators were implemented using predefined configurations (Zhang et al., 2025). S-Learner, T-Learner and X-Learner employed Random Forest-based outcome and treatment-effect models, while LinearDML and CausalForestDML used Random Forest nuisance models with three-fold cross-fitting. Two-Model Random Forest Uplift used separate treatment and control classifiers. All estimators used the same twelve anonymized covariates and binary visit outcome. Hyperparameters were predefined without automated optimization or post-fitting calibration. Detailed estimator configurations, including base learners, tree parameters, minimum leaf sizes and cross-fitting settings, are reported in Appendix Table A5.

The primary estimand was the conditional average treatment effect (CATE):

$$\tau(x) = E[Y(1) - Y(0) \mid X = x] \quad (1)$$

where Y (1) and Y (0) denote potential binary visit outcomes under treatment and control, respectively, and X represents the twelve anonymized features. Treatment was defined by randomized assignment T=1 versus T=0; therefore, estimated effects represent absolute differences in visit probabilities and are interpreted as intention-to-treat effects.

Models were fitted using 80,000 training observations (40,000 per treatment group), selected with seed 42. A separate 20,000-observation test subset supported CATE diagnostics and

bootstrap analysis, while the complete 75,000-observation test set was used for designated policy evaluations.

S-Learner used a single Random Forest incorporating treatment and covariates; T-Learner employed separate treatment and control models; and X-Learner additionally used pseudo-effect regressors and a propensity classifier. Their Random Forest components used 100 trees, depth six, minimum leaf size 50, and seed 42. LinearDML and CausalForestDML used Random Forest nuisance models with three-fold cross-fitting. CausalForestDML additionally employed 200 trees and minimum causal leaf size of 20. Two-Model Uplift used separate Random Forest classifiers with 150 trees, depth eight, and minimum leaf size 50.

All configurations were predefined without automated hyperparameter optimization or post-fitting calibration. Evaluation included CATE distributions, positive-effect proportions, Qini and uplift metrics, and bootstrap uncertainty. These analyses evaluate estimated effects and policy ranking accuracy but not the accuracy of individual treatment effects as they happen in the real world. The full set of causal machine-learning estimators' configurations is summarized in Appendix Table A5, consisting of base learners, tree parameters, minimum leaf sizes, and cross-fitting settings.

In order to be consistent, the observed outcomes must reflect the treatment given, and SUTVA implies that there is no interference between people. The randomized treatment assignment supports conditional exchangeability, and positivity will be a property of the representation both within and across treatment conditions in the covariate space. Empirical diagnostics confirmed the propensity scores ranged from 0.323751 to 0.697938, and all of the observations were within the interval [0.05, 0.95]. Training, validation, and test partitions had standardized mean differences < 0.10. These diagnostics can be used to facilitate measured-covariate balance

and treatment overlap, but cannot be used to establish SUTVA or to prove all identification assumptions.

2.6 Policy Learning and Intervention Decision Rules

Intervention policies had to be built from model-specific ranking and threshold definitions, enabling common resource constraints for predictive and causal targeting. The main allocation budget comprised 30% of the held-out test population (N = 75,000; K = 22,500). The models were ranked according to the probabilities predicted by them in the case of the predictive model, while the other models (X-Learner, Causal Forest and Two-Model Uplift) were ranked based on the treatment effect or uplift score. The baseline was set by a random allocation policy. The probabilities generated by XGBoost and the feature-deviation scores were normalized and then combined with equal weights (0.5:0.5) in SHAP-assisted targeting. A cost-sensitive Causal Forest policy selected observations with estimated CATE exceeding 0.005, subject to the maximum allocation budget. This is a simulated cost, not a cost to the individual.

Additional sensitivity experiments evaluated the stability of Causal Forest estimation and intervention-policy performance under alternative training-sample sizes, random seeds, and model specifications. Training samples of 100,000, 200,000, and 350,000 observations were considered, with random seeds 21, 42, and 84. Two specifications were compared: the primary Causal Forest configuration (200 trees and a minimum causal leaf size of 20) and a more regularized configuration (300 trees and a minimum leaf size of 50). Each fitted model was evaluated on the same held-out test population of 75,000 observations. Intervention budgets of 10%, 20%, 30%, 40%, and 50%, together with CATE thresholds of −0.005, 0, 0.0025, 0.005, and 0.01, were examined. Simulated intervention costs of 0, 0.0025, 0.005, 0.01, and 0.02 were also considered. These experiments assessed sensitivity to analytical sample size, model regularization, allocation capacity, intervention eligibility thresholds, and assumed intervention

costs. Mean performance and standard deviations across random seeds were used to summarize the results.

Allocation parity was investigated by splitting the observations into four equal groups by their empirical quartiles of f0 (the feature following the nonparticipation) and allocating the same number of observations according to the Causal Forest rankings. These groups are not based on demographic or protected categories, but rather on a statistical partition and thus allocation parity is not the same as feature-based allocation parity. The results are purely related to the Criteo advertising benchmark and cannot be generalized to other application areas.

The fairness-aware allocation policy used allocation-rate parity between four groups roughly equal in size based on the empirical quartiles of an anonymized feature f0. The relationship between fairness and explainability is also recognized as an important consideration in supervised learning, motivating their complementary examination in decision-support systems (Grabowicz et al., 2022). The observations were ranked in each group by their estimated Causal Forest treatment effects, and a similar number of observations in each group was allocated under the prespecified allocation budget. The goal of the fairness objective, therefore, was not to seek to achieve demographic or protected-group fairness, but to seek to minimize differences in rates of intervention for feature-defined groups. The groups cannot be interpreted as demographic, socioeconomic or behavioural groups since the semantic meaning of f0 is not available. The policy is a limited allocation-parity experiment rather than an overall guarantee of benefits for equitable treatment.

2.7 Intervention utility Evaluation

Since there are no observable individual treatment effects, the evaluation of the causal models was based on the treatment-effect ranking and intervention utility instead of on prediction errors at the individual level. The following 6 complementary measures were employed.

Empirical Uplift Gain and Qini Coefficient. For the highest-ranked proportion q of observations, empirical uplift was calculated as the difference between treated and control visit rates:

$$U(q) = \frac{\sum_{i \in S_q} T_i Y_i}{\sum_{i \in S_q} T_i} - \frac{\sum_{i \in S_q} (1-T_i) Y_i}{\sum_{i \in S_q} (1-T_i)} \quad (2)$$

The cumulative Qini curve was defined as:

$$Q(q) = n_1(q)[\mu_1(q) - \mu_0(q)] \quad (3)$$

where n_1(q) denotes selected treated observations and mu_1(q) and mu_0(q) represent treated and control visit rates. The normalized Qini coefficient compares cumulative uplift against random and ideal reference curves.

Policy Value and Net Benefit. Inverse-propensity-weighted (IPW) policy value was estimated as:

$$\hat{V}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\pi_i T_i Y_i}{e} + \frac{(1-\pi_i)(1-T_i) Y_i}{1-e} \right] \quad (4)$$

where pi_i represents the intervention decision, and e=0.50 is the treatment-assignment probability in the balanced sample. Cost-adjusted net benefit was calculated as:

$$\widehat{NB}(\pi; c) = \hat{V}(\pi) - c \frac{\sum_i \pi_i}{N} \quad (5)$$

The primary simulated intervention cost was c=0.005 successful-visit-equivalent units per intervention, representing an assumed utility penalty rather than actual monetary expenditure.

Model-Relative Regret and Targeting Efficiency. Regret was calculated relative to a Causal Forest reference allocation selecting the same number of observations:

$$\hat{R}(\pi) = \frac{1}{N} \sum_i \left(\pi_i^{CF,K} - \pi_i \right) \hat{\tau}_i^{CF} \quad (6)$$

Lower regret indicates closer agreement with the reference allocation; zero indicates identical allocation. This represents model-relative regret. Targeting efficiency was defined as:

$$TE(\pi) = \frac{\sum_i \pi_i \mathbb{1}(\hat{\tau}_i^{CF} > 0)}{\sum_i \mathbb{1}(\hat{\tau}_i^{CF} > 0)} \quad (7)$$

It measures the proportion of observations with positive estimated CATE selected by a policy, rather than verified identification of actual beneficiaries.

Statistical uncertainty in intervention utility was assessed using 1,000 paired, treatment-stratified bootstrap resamples of a 20,000-observation subset drawn from the 75,000-observation held-out test partition. The subset contained 10,000 observations from each treatment arm. All competing policies were evaluated on identical bootstrap resamples, preserving paired comparisons. For Causal Forest relative to XGBoost, the estimated differences were 0.003965 in uplift gain (95% CI: −0.004648 to 0.013333), 0.001720 in Qini coefficient (95% CI: −0.043788 to 0.050864), 0.001200 in policy value (95% CI: −0.000903 to 0.003600), and 0.001177 in net benefit (95% CI: −0.000945 to 0.003562). All four intervals included zero. These results do not establish a statistically distinguishable advantage in intervention utility. They should be interpreted separately from the full-test-set policy estimates reported in Table 5.

Direct individual-level CATE prediction errors can't be calculated since the actual outcomes for individual students can't be observed, making it impossible to compare them with predictions. Thus, the treatment-effect estimators were assessed based on cumulative uplift, Qini performance, uplift AUC, empirical uplift at the predefined allocation budget, and intervention-policy value. The CATE-quartile analysis also compared the estimated treatment effects to the observed treatment–control visit rate differences in the ranked treatment subgroups. These comparisons are descriptive evidence of group-level treatment-effect ranking behaviour, but do not formally validate individual CATE accuracy. It is therefore important to

remember that the treatment effects estimated herein are conditional, and that they are not necessarily proven treatment benefits.

Development of the models and final policy assessment were separated to reduce information leakage. The training dataset was used for model fitting, and the validation dataset was used to aid in the predictive assessment and model development decisions. The test results were not used for model fitting or hyperparameter selection, but were only used for final evaluation. The primary intervention budget was specified and not selected to achieve maximum performance on the observed test sets, as were the simulated intervention-cost assumptions, the CATE eligibility thresholds, and the feature-based allocation-parity rules. For policy ranking, the held-out observations were treated with estimated treatment effects, and for policy evaluation, only the observed test outcomes were used.

The experimental framework provides details of the analytical sample construction, treatment-balanced sampling procedure, random seeds, partitioning of training, validation and test sets, feature preprocessing, model hyperparameters, causal estimator configurations, intervention-policy rules, and statistical evaluation procedures, to facilitate reproducibility. The software libraries and versions used in the implementation of the predictive models, SHAP explanations, causal estimators, and policy evaluations are reported in Appendix Table A6. The experimental setup was primarily set with random seed 42, and sensitivity experiments were conducted with random seeds 21 and 84. The following materials are publicly archived on Zenodo, as detailed in the Code Availability section below: computational workflow and associated implementation materials.

3 CONCEPTUAL FRAMEWORK

The conceptual approach of this research is grounded in the new paradigm of decision-based AI, which goes beyond prediction and focuses on the effectiveness of interventions and

decision usefulness. Conventional machine learning systems are primarily designed to predict future outcomes using historical data; however, accurate prediction alone does not necessarily provide guidance regarding which actions should be taken to improve those outcomes. Recent studies have thus argued for changing the trajectory of AI from prediction toward decision-making, which explicitly incorporates intervention planning, policy optimization, and utility maximization (Lo and Pachamanova, 2023, Vallee, 2026).

The proposed framework assumes that individuals are characterized by a set of observable features that influence both treatment assignment and subsequent outcomes. In cases of exposure to treatment, outcomes are assessed after the treatment and analyzed to estimate variability in treatment effects between individuals. The treatment-effect estimates are used for policy learning, which involves optimizing intervention strategies to reach those who are likely to benefit most from the intervention. The role of intervention utility evaluation is to determine if the learned policies result in measurable gains in outcomes, taking into account efficiency, cost, fairness and overall decision value. The process highlights the main goal of causal machine learning: to not only predict what is likely to happen but also what actions can be followed to achieve a better result (Feuerriegel et al., 2024).

To implement this framework, the study implements a sequential, analytical pipeline, starting with predictive modelling and explainability analysis, followed by causal treatment-effect estimation, policy learning, and policy evaluation. Prediction models set the reference for performance and detect patterns of response, and explainable AI techniques give clarity on the features contributing to the prediction. Causal machine learning techniques then estimate the individual treatment effects, which lead to the formulation of an intervention policy that is assessed based on a utility function. These are all interconnected elements that run alongside one another in a single analytical process to create an integrated decision-centric AI framework that connects prediction, causation, policy optimization, and treatment effectiveness.

Figure 1 represents a sequence of user features, treatment assignment, observed outcome, estimation of treatment effect, policy learning and evaluation of intervention utility. It also shows the analytical pipeline, which includes predictive modeling, explainability analysis, causal inference, policy optimization and decision evaluation.

Insert Figure 1 Here

The figure 2 workflow illustrates dataset preparation, treatment-balanced sampling, stratified data partitioning, predictive modelling and SHAP explainability, followed by a separate causal-estimation branch for heterogeneous treatment-effect estimation. Predictive attribution and estimated treatment effects inform intervention-policy construction, including prediction-based, CATE-based, cost-sensitive and fairness-constrained allocation strategies. Policies are evaluated on held-out observations using policy value, uplift, Qini, net benefit and allocation-parity measures. Robustness and reproducibility analyses complete the experimental pipeline.

Insert Figure 2 Here

4 RESULTS

4.1 Dataset Characteristics

The dataset characteristics were examined to describe treatment allocation, outcome frequencies, and feature distributions in the analytical sample. Both visit and conversion rates are reported to characterize the dataset, but visit is the primary outcome for all predictive modelling, treatment-effect estimation, and policy-learning analyses. The conversion is given for descriptive purposes only. The original Criteo benchmark is a randomized design, which gives the basis for estimating effects of the treatment, under the following causal assumptions. The analytical sample, however, was purposefully split between treatment and control groups,

whereas the original data set was not; therefore, the results should only be considered to reflect policy performance on the sampled population and not be generalized to all target populations.

In table 2, the visit rate represents the proportion of observations with a positive binary visit outcome (visit = 1) among all 500,000 observations in the treatment-balanced analytical sample. The conversion rate represents the corresponding proportion with a positive conversion outcome (conversion = 1) in the same sample. Both rates are calculated using the full analytical sample as the denominator. Visit is the primary outcome used in all predictive modelling, CATE estimation, intervention-policy construction, and policy-utility evaluation. Conversion is reported exclusively for descriptive characterization and is not used as a modelling or policy-learning outcome. The 12 input features (f0–f11) are anonymized numerical variables without established semantic interpretations.

Insert Table 2 Here

Table 3 compares the treatment allocation rate and the rate for treatment-specific outcomes of the treatment-balanced analytical sample with the rates for the original Criteo dataset to assess the distributional characteristics of the treatment-balanced analytical sample. The original data sample was comprised of 85% treated and 15% control observations, while the analytical data sample was comprised of 50% and 50%, respectively. Even though the treatment allocation was a conscious change, the conditional visit rates were fairly consistent. The visit rates of treated and control groups were 4.8543% and 3.8201%, respectively, in the original data, and 4.7632% and 3.8168%, respectively, in the analytical sample. The numbers of conversions in the treated observations dropped from 0.3089% to 0.2816%, while the values for the control observations fell from 0.1938% to 0.1760%. The results suggest that the proportions of the outcomes of treatment remained broadly similar across the sample, with some significant differences, especially in the less frequent conversion outcome. The similarity observed thus

lends support to the use of the analytical sample for the intended experimental comparisons, but does not demonstrate absolute distributional equivalence or population-level representativeness. Estimated treatment effects should therefore be interpreted with the deliberate change in treatment allocation in mind, and the performance of the policy should be extrapolated to the original population.

Insert Table 3 Here

A descriptive sensitivity analysis was performed to assess the sensitivity of the observed treatment–control visit-rate difference to random subsampling, with five random seeds (11, 21, 31, 41, and 51). 25,000 observations per treatment group, in the training partition, were sampled without replacement for each seed. The percentages of visits for the treated groups ranged from 4.504% to 5.000%, whereas the percentages of visits for the control groups ranged from 3.752% to 3.996%. The difference in visit rate between the observed and control groups was positive in each of the 5 resamples, between 0.00748 and 0.01248 or 0.748–1.248 percentage points. Within this range of differences, the difference that was observed in the primary analytical sample was 0.009464 (0.9464 percentage points).

These results suggest that the sign of the difference in visit rate of the aggregate remained constant over the five descriptive resamples. This analysis, however, only examines the difference observed in the group; it does not assess the robustness of predictions, CATE estimates, model rankings, or intervention utility measures to alternative groups. The entire model set was not fitted to the entire dataset of 13,979,592 observations.

Figure 3 shows the treatment allocation, the distributions of outcomes, the selected distributions of features, and the structure of the correlation between the features and the outcome in the analytical dataset. The visualization verifies the balance of treatment allocation and shows that the outcomes that were observed were very skewed. The marginal linear

associations for the considered variables are generally moderate with regard to correlation analysis. Weak pairwise correlations are not a reason to exclude nonlinear relationships and higher-order interactions or dependency of the relationship, nor are they sufficient to call for a specific modelling method. These descriptive patterns, therefore, serve as background to the subsequent discussion of the viability and utility of alternative approaches to predictive and causal estimation, rather than as evidence of the superiority and necessity of the nonlinear machine learning and causal inference methods.

Sensitivity analysis was performed to explore how these factors affect the utility of the policy and how it may be allocated. The simulated costs ranged from 0 to 0.02, the allocation budgets from 10% to 50%, and the CATE thresholds from -0.005 to 0.01. The comparisons were made to assess if the decision parameters affected the rankings of intervention, allocation rates, policy value and cost-adjusted net benefit. The results are subject to the cost assumptions used and should not be viewed as providing evidence of optimal economic spending or feature-based allocation parity describing f0 groups.

Insert Figure 3 Here

4.2 Prediction Performance Analysis

The four classification models were assessed according to their predictive performance by the AUC-ROC, area under the precision–recall curve (AUC-PR), precision, recall, F1-Score and accuracy. Since the classification is positive in a relatively small percentage of the analytical sample, accuracy cannot be used as a sole measure of classification performance (around 4.29%). Only 95.71% accuracy would be obtained using a majority-classifier that always classifies the negative class. It is important to note that such a classifier, a majority-classifier, achieves about 95.71% accuracy alone, and should not be interpreted without accounting for positive-class detection metrics.

The models performed between 0.4591 and 0.4901 in terms of AUC-PR, showing good performance in discriminating positive visit observations in the presence of class imbalance as reported in Table 4. The best result was obtained by LightGBM with AUC-PR (0.4901), and Random Forest with precision (0.6474). The best result obtained in the validation was the highest recall (0.2860) and F1-score (0.3869) values obtained by Logistic Regression. The differences show the compromise in attempts to capture positive observations and to reduce false positive predictions. The models' accuracy was over 96%, but the recall was much lower, so that a figure of over 96% accuracy would overestimate the ability of the models to recall positive class detections. For this reason, along with AUC-ROC, the accuracy is retained as a secondary descriptive measure, and AUC-PR, precision, and recall are emphasized.

Insert Table 4 Here

All metrics are calculated on the validation set using binary visit as the positive outcome. AUC-PR represents average precision. Precision measures the proportion of predicted positive observations that are actually positive, recall measures the proportion of actual positive observations correctly identified, and F1-score is the harmonic mean of precision and recall. Accuracy is reported as a supplementary metric because of the low prevalence of positive visit outcomes. Conversion is not used as a predictive modelling target. Bootstrap-based confidence intervals for the predictive performance metrics are provided in Appendix Table A1.

LightGBM achieved the highest AUC-PR (0.4901), followed by XGBoost (0.4873), Random Forest (0.4865), and Logistic Regression (0.4591). Precision ranged from 0.5978 to 0.6474, while recall ranged from 0.2157 to 0.2860. These results indicate stronger positive-class ranking performance among the ensemble models, although threshold-dependent classification performance varied across algorithms.

Figure 4 compares the predictive performance of Logistic Regression, Random Forest, XGBoost, and LightGBM. Statistical differences were assessed using 1,000 paired, stratified bootstrap resamples of the common validation dataset. Pairwise differences in AUC-ROC, AUC-PR, and F1-score were evaluated using 95% percentile confidence intervals and two-sided tests, with Holm correction across 12 comparisons.

Insert Figure 4 Here

Random Forest, XGBoost, and LightGBM achieved significantly higher AUC-ROC and AUC-PR than Logistic Regression (Holm-adjusted $p = .012$). XGBoost improved AUC-ROC by 0.01686 (95% CI [0.01377, 0.02015]) and AUC-PR by 0.02818 (95% CI [0.02061, 0.03685]). LightGBM achieved corresponding improvements of 0.01674 (95% CI [0.01372, 0.01999]) and 0.03102 (95% CI [0.02386, 0.03889]). Conversely, Random Forest and LightGBM produced significantly lower F1-scores than Logistic Regression (adjusted $p = .012$ and .045, respectively), whereas the XGBoost difference was insignificant ($p = .541$). No significant differences were observed between XGBoost and LightGBM across the three metrics (adjusted $p > .05$). These findings demonstrate that improved discrimination does not necessarily translate into better threshold-dependent classification performance. Complete comparisons are provided in Table 5.

Insert Table 5 Here

A component ablation analysis was conducted to investigate the individual and combined contributions of predictive modelling, SHAP-informed feature attribution, causal treatment-effect estimation, intervention-cost sensitivity, and allocation fairness. The analysis was designed to distinguish the contribution of each component from the effects of their integration within the decision-centric framework.

The predictive component used XGBoost to predict visit probabilities as the baseline prioritization score. The SHAP component incorporated global probability-scale feature-importance information into an attribution-informed ranking score. The causal component used Causal Forest to estimate conditional average treatment effects to prioritize individuals according to their predicted incremental response. Cost sensitivity imposed an eligibility constraint requiring estimated treatment effects to exceed the prespecified simulated intervention cost. The fairness component introduced equal allocation quotas across four analytical groups defined by quartiles of the anonymized feature f0.

The factorial comparison evaluated individual components, pairwise combinations, and progressively integrated configurations. All configurations were evaluated using the same held-out test observations, intervention budget, cost assumptions, and policy-evaluation procedures. The assessment considered policy value, net benefit, empirical uplift, Qini coefficient, and treatment-allocation rates. Allocation differences across the four analytical groups were additionally examined to characterize the effect of the fairness constraint.

The 22 ablation configurations demonstrated different trade-offs between predictive ranking, SHAP-assisted targeting, causal estimation, intervention costs, and allocation parity. The standalone parity constraint eliminated allocation-rate disparities across the four feature-defined groups, whereas combining cost eligibility with parity constraints reduced the achievable allocation rate. The results show that each policy component has a different impact on the utility of the intervention and the allocation patterns, but they do not establish universal superiority of the integrated configuration of the policies.

4.3 Explainability Analysis

A preliminary comparison revealed that there was a mismatch between the scale of the output and the scale of the visit probability; therefore, SHAP analyses were recalculated on a

consistent visit-probability scale. The same validation observations and training-data background were given to both XGBoost and LightGBM, and described with PermutationExplainer. The updated MAE SHAP values (Table 6) indicate that f8 is the top predictor for both models, with XGBoost (0.035269) and LightGBM (0.027531), making it the most important anonymized predictor. The new values have been corrected and indicate no large cross-scale discrepancy as noted earlier.

Insert Table 6 Here

SHAP Analysis and X-Learner Correction. The mean absolute SHAP values were computed using a training background sample size of 100 and 2,000 validation observations. Confidence intervals were identical in both models (seed 42, same configuration of PermutationExplainer using a probability scale). The top features f8 are also the same for both models: XGBoost (0.035269) and LightGBM (0.027531), suggesting that the two models have similar predictive attribution patterns, but not evidence of causal influence or treatment-effect modification.

The CATE estimates were degenerate all zeros in the preliminary X-Learner, because the estimator configuration was not appropriate. It was corrected with the use of random forest regressors to obtain treatment-specific outcomes and pseudo-treatment effects, while keeping a classifier for the estimation of the propensity. The corrected model resulted in a nondegenerate CATE distribution (mean = 0.005562; SD = 0.016917) with 95.47% of positive estimates. At the 30% allocation budget, normalized Qini AUC was 0.08941 (95% CI: 0.03327–0.14832), and empirical uplift was 0.02051 (95% CI: 0.00532–0.03428). . These results show implementation behaviour and ranking of informativeness, not the accuracy of the individuals who use CATE. Complete paired-bootstrap comparisons of causal and prediction-based ranking performance are reported in Appendix Table A3a.

The SHAP summary plots and global feature importance comparisons of XGBoost and LightGBM visit-probability predictions are shown in Figure 5. The summary plots show the distribution and direction of attribution values for each of twelve anonymized features (f0 – f11), and the mean absolute SHAP represents the predictive contributions of features globally. The most important features are f8, f9 and f2 in both models. These attributions are for the prediction of behavior and not causal effects or treatment-effect heterogeneity, and conversion is not a prediction target.

Insert Figure 5 Here

4.4 Treatment Effect Estimation Results

All six causal and uplift estimators were tested on a standard held-out test sample of 20,000 observations. All estimates in Table 7 reflect conditional differences in treatment probabilities of visits between the treatment and control conditions. Positive Effects (%) refers to the percentage of observations where the estimated CATE is strongly positive, and not just statistically significant individual treatment effects.

Insert Table 7 Here

An implementation audit identified specification errors in the preliminary S-Learner, T-Learner, X-Learner, and Two-Model Random Forest uplift estimators. The S-Learner and T-Learner were corrected by using Random Forest regressors for continuous outcome-probability estimation. The X-Learner was reconfigured using Random Forest regressors for treatment-specific outcomes and pseudo-treatment-effect estimation, while retaining a classifier for propensity estimation. The Two-Model Random Forest uplift implementation was rectified to use the right ordering of treatment and outcome variables. Table 7 shows summary statistics for the revised estimators, which yielded nondegenerate conditional treatment-effect estimates. Such implementation changes enhance computational consistency but do not demonstrate any

accuracy for individual CATE or any universal superiority of the estimator. Model-based ATE inference is reported for LinearDML and Causal Forest, the two estimators for which the implemented EconML workflow provides the corresponding inference procedures. The remaining meta-learners and uplift estimators are evaluated using their estimated CATE distributions and held-out policy-performance measures rather than equivalent model-based ATE confidence intervals. Average treatment-effect estimates and confidence intervals are presented in Appendix Table A4 based on the model. Model-based ATE inference is reported for LinearDML and Causal Forest, the two estimators for which the implemented EconML workflow provides the corresponding inference procedures. The remaining meta-learners and uplift estimators are evaluated using their estimated CATE distributions and held-out policy-performance measures rather than equivalent model-based ATE confidence intervals.

CATE distributions for all six estimators with common axis limits and histogram bins are shown in Figure 6. The distributions show differences in the central tendency, variability, and estimated heterogeneity of treatment effects. Further, the highest quartile of CATE had empirical incremental visit response of 0.030386, while the lowest quartile had 0.002760.

Insert Figure 6 Here

The CATE-quartile analysis offers descriptive evidence of the differential observed treatment effects between treatment groups sorted by estimated treatment effects. Although the highest-ranked groups exhibited positive empirical treatment–control visit-rate differences, intermediate-group discrepancies indicate that estimated CATE values should not be interpreted as perfectly calibrated individual treatment effects. Consequently, Qini, uplift, policy-value, and bootstrap analyses are considered complementary evaluation measures rather than direct measurements of individual CATE accuracy.

4.5 Policy Learning and Intervention Utility Results

Treatment-effect estimates were converted into intervention policies and evaluated using decision-centric metrics. Because true individual treatment effects are unobserved, Qini and uplift curves were interpreted as measures of ranking-level targeting performance rather than direct measures of individual CATE accuracy. Performance varied across models and metrics. The reported normalized Qini AUC was 0.05585 for XGBoost prediction, 0.05543 for LightGBM prediction, 0.06514 for LinearDML, 0.05757 for Causal Forest, and 0.08941 for X-Learner. The corresponding 95% bootstrap confidence intervals were [0.00488, 0.10933], [0.00524, 0.10988], [0.01063, 0.11731], [0.00467, 0.11695], and [0.03327, 0.14832], respectively. These results indicate variation in treatment-effect ranking performance, but the paired comparisons did not establish statistically distinguishable differences between Causal Forest and the prediction-based approaches. These findings do not establish comparative advantage under the evaluated experimental conditions of causal estimators over prediction-based approaches. Instead, they indicate that the effectiveness of treatment targeting depends on the ranking score, estimator, and evaluation metric. Detailed causal ranking and targeting performance estimates, together with their bootstrap confidence intervals, are reported in Appendix Table A2.

Table 8 compares intervention utility and allocation behaviour under different intervention-cost assumptions. At zero intervention cost, Causal Forest selected 30.00% of observations, achieving an empirical uplift of 0.026372 and a net benefit of 0.045327. At a simulated cost threshold of 0.005, the allocation rate was 29.78%, empirical uplift was 0.026526, and net benefit was 0.045364. At a threshold of 0.02, the allocation rate decreased to 11.345%, while empirical uplift increased to 0.05392 and net benefit decreased to 0.04360. These results illustrate the trade-off between targeting fewer observations with higher estimated treatment effects and maximizing aggregate intervention utility. Costs represent simulated penalties rather than actual campaign expenditure.

Insert Table 8 Here

Statistical uncertainty was assessed using 1,000 paired, treatment-stratified bootstrap resamples of the 20,000-observation held-out evaluation subset. Competing policies were evaluated on identical resamples, generating 95% percentile confidence intervals for policy value, net benefit, uplift gain, and Qini performance. However, the paired Causal Forest–XGBoost policy-value difference was 0.001200 (95% CI: -0.000903 to 0.003600), indicating no statistically distinguishable difference. As intervals included zero, consistent causal-policy superiority was not established. Comparisons were exploratory without multiple-testing adjustment. Detailed paired-bootstrap comparisons of intervention-policy utility, including uplift gain, policy value, and net benefit, are reported in Appendix Table A3b.

Figure 7 Comparative evaluation of intervention policies. (a) Qini curves comparing prediction-based, SHAP-assisted, and causal targeting strategies; (b) cumulative uplift across alternative ranking approaches; (c) treatment-allocation rates across four anonymized feature-defined subgroups; (d) treated and control visit rates across Causal Forest-estimated CATE quartiles; and (e) decision gaps across alternative intervention-allocation budgets. The panels illustrate differences in ranking performance, observed treatment response, subgroup allocation, and budget sensitivity. Results should be interpreted alongside policy-value, net-benefit, and uncertainty estimates and do not establish universal causal-policy superiority or demographic fairness.

Insert Figure 7 Here

The results demonstrate that high predicted response probabilities do not necessarily identify individuals with the greatest estimated incremental treatment benefits. Predictive models and causal estimators therefore address complementary prioritization objectives. Causal targeting yielded comparable intervention-utility estimates, but the paired-bootstrap comparisons did not

provide a consistently statistically superior intervention-utility compared to prediction-based targeting. The results then offer an argument for the consideration of predictive discrimination and treatment-effect ranking and held-out intervention utility in addition to the assumption that one modelling approach is universally superior. XGBoost prediction rankings and Causal Forest treatment-effect rankings had a correlation of 0.3942, meaning that people with higher predicted visit probabilities did not necessarily have higher estimated incremental treatment benefits.

At the primary 30% intervention budget, XGBoost achieved a policy value of 0.047333 and net benefit of 0.045833, compared with 0.046827 and 0.045327, respectively, for Causal Forest. The X-Learner produced estimates of 0.047733 and 0.046233, respectively. Paired-bootstrap comparisons using 1,000 treatment-stratified resamples showed no statistically distinguishable differences in uplift gain, Qini coefficient, policy value, or net benefit. For example, the paired Causal Forest–XGBoost policy-value difference was 0.001200 (95% CI: −0.000903 to 0.003600). Because the confidence interval included zero, the analysis did not establish a statistically distinguishable policy-value advantage. These findings do not establish consistent causal-policy superiority.

The policy sensitivity analysis showed that relative performance depended on the intervention budget. Causal Forest achieved higher empirical uplift than XGBoost at allocation budgets between 5% and 25%, whereas XGBoost performed marginally better at 30%, 40%, and 50%. Increasing the simulated intervention-cost threshold from zero to 0.02 reduced the Causal Forest allocation rate from 30.00% to 11.345%. Over the same comparison, empirical uplift increased from 0.026372 to 0.05392, while net benefit decreased from 0.045327 to 0.04360. These results demonstrate the trade-off between intervention eligibility, targeting intensity, and aggregate utility.

The model-relative regret analysis provides an additional diagnostic of targeting efficiency. Using the Causal Forest estimated treatment effects as the reference ranking, the full-test-set regret values were 0.001728 for XGBoost, 0.001723 for LightGBM, 0.001246 for the X-Learner, 0.001141 for Two-Model Random Forest Uplift, and 0.000000 for the Causal Forest policy. The zero regret of the Causal Forest policy is expected because it defines the reference allocation under this metric. These estimates therefore quantify disagreement with the selected model-based reference ranking rather than deviation from an observable optimal intervention policy. They should not be interpreted as evidence of true individual-level causal regret or verified beneficiary identification.

The expanded sensitivity analysis demonstrated that the relative performance of the two Causal Forest configurations depended on training-sample size and model specification. Based on average policy-net-benefit rankings across the evaluated budgets, thresholds, and seeds, the regularized configuration achieved a better average rank at training-sample sizes of 100,000 (1.373 versus 1.627) and 200,000 (1.147 versus 1.853). At 350,000 observations, the two configurations had nearly identical average ranks (1.493 and 1.507, respectively). These findings indicate that model regularization can influence intervention-policy performance, particularly under smaller training samples, but do not establish universal superiority of either configuration. Changes in allocation budget and CATE threshold also affected the number of selected observations and the resulting intervention utility. The complete sensitivity results are provided in the supplementary computational outputs.

The allocation-parity experiment studied four groups based on the empirical quartiles of anonymized feature f0. The standalone parity-constrained policy had zero allocation-rate disparity across the four f0-defined groups, while the unconstrained policy had a maximum allocation-rate disparity of 0.420373. Policy value decreased from 0.046827 to 0.046507. The results illustrate the trade-off of imposing equal rates of intervention allocation across

anonymized feature-defined groups. They do not establish fairness across identifiable demographic or protected groups.

The SHAP-assisted policy changed 31.59% of conventional XGBoost intervention selections but did not improve the complete-test-set results. Its policy value was 0.046933, compared with 0.047333 for XGBoost, while net benefit was 0.045433 versus 0.045833. Bootstrap comparisons did not establish statistically distinguishable differences. SHAP therefore provided predictive explanations rather than evidence of causal effect modification. Overall, the findings demonstrate that predictive ranking, treatment-effect estimation, intervention utility, and allocation constraints capture complementary aspects of decision-making. The results remain specific to the treatment-balanced Criteo benchmark and do not establish comparative advantage under the evaluated experimental conditions or cross-domain generalizability.

5 DISCUSSION

Results from this study show the importance of embedding causal machine learning in decision-focused AI systems. The results show that predictive accuracy is not always a reliable indicator of effective decisions to intervene, although prediction models like XGBoost and LightGBM performed well on traditional classification metrics. The latter distinction underlines the fundamental difference between prediction-centric and decision-centric AI. While traditional machine learning models predict future outcomes, AI-driven decision-making models aim to determine the actions that can be taken to achieve the greatest positive outcome by intervening with specific actions. Recent work has suggested that the performance of causal policies on uplift generation, policy value and treatment-targeting efficiency should be increasingly emphasized in the evaluation of AI systems, moving beyond just predictive performance (Lo and Pachamanova, 2023, Vallee, 2026).

This study reveals that there is significant variability in treatment between people. The estimated conditional treatment effects were heterogeneous across observations made by the evaluated causal estimators, suggesting that the treatment had varying effects on these observations. These are, however, model-dependent estimates, and they do not form an independent basis for estimating whether there is genuine individual-level heterogeneity of treatment effects. The results confirm the separation between the estimation of the predictive outcome and the estimation of the treatment effect. It is important to remember that people with a high predicted response probability do not necessarily mean that they will gain the most from an intervention. This means that prediction-based targeting might miss patients with lower response rates at baseline but higher treatment benefit. Causal machine learning thus offers complementary information for detecting treatment-effect heterogeneity and guiding resource allocation for limited budgets of interventions (Feuerriegel et al., 2024). The results, however, do not constitute a proof of superiority of causal targeting on an overall basis, and highlight the need to measure predictive discrimination and intervention utility.

The SHAP analysis was used to find the relative contributions of anonymised features to the output of the predictive model. These attributions, however, are predictive relevance in nature and not causal influence or treatment effect modification (Fryer et al., 2021, Ponce-Bobadilla et al., 2024). Likewise, targeting with SHAP did not demonstrate greater intervention value than would traditional prediction-based targeting. Predictive explainability and causal estimation are therefore complementary, but methodologically distinct.

There are a number of methodological limitations that should be taken into account. First, the study has been run on a treatment-balanced sample of 500,000 observations from the Criteo original dataset with 13,979,592 observations, rather than the previous treatment allocation ratio of about 85:15. No sampling weights were used, and policy estimates are based on the analytical sample and not on a reconstructed sample. Secondly, the twelve anonymized features

have no substantive definitions, meaning that no behavioral or demographic mechanism can be interpreted. Third, the analysis is restricted to a single intervention (advertising) and a binary measure of visit (visit/no visit), rather than to longitudinal treatment histories and long-term outcomes.

Randomized assignment provides support for conditional exchangeability, but causal interpretation requires consistency, positivity, and the Stable Unit Treatment Value Assumption (SUTVA). Not all of these assumptions can be independently substantiated. Alternative algorithms and hyperparameters might not recover the true treatment-effect structure, which can have a significant impact on the predictive and causal estimates. Moreover, CATE estimates can be affected by finite-sample and model-specification errors, and should be understood as being on a conditional model basis and in the sense of model-specific conditional estimates rather than being causal estimates of the individual. Allocation budgets, decision thresholds and assumptions on intervention costs also can affect policy performance. While certain configurations have been explored, robustness has not been proven for all configurations. Although the expanded sensitivity analysis examined model configurations, training-sample sizes, and random seeds, it did not establish robustness across all possible hyperparameter settings, data-generating processes, or alternative datasets. The experiments therefore provide evidence of sensitivity under the configurations examined rather than universal stability of treatment-effect estimates and intervention-policy rankings. Further, no quantitative sensitivity analysis for hypothetical unmeasured confounding was performed.

Another limitation is the external validity, as the experiments were conducted using only Criteo Uplift Prediction Dataset. This randomization advertising benchmark is used for evaluation of predictive modelling, estimation of treatment effects, and allocation of interventions, but cannot be directly applied to other populations, treatment mechanisms, or outcome settings.

Performance and utility of the models and interventions might be significantly different across digital advertising, e-mail marketing, healthcare, finance, and public administration.

Thus, there are cautious implications for healthcare and public policy. While the proposed workflow might provide methodological lessons for allocation of resources and for planning interventions, such applications are not necessarily experimentally proven results, but potential research directions. Clinical data, clinically relevant endpoints, and adequate identification of causal relationships, patient safety evaluation and regulatory clearance would be required for implementation in healthcare (Cross et al., 2024). For public-policy applications, representative populations would also need to be available, as would the costs of interventions, measures of social outcomes, and distributional and ethical analyses. The current results do not, therefore, demonstrate effectiveness, safety, or equity in these areas.

Future studies should analyze the workflow with independent randomized data that have not been used here, such as the Hillstrom email-marketing experiment (Diemert et al., 2018, Rößler and Schoder, 2022), which introduces another intervention and population setup. This study does not validate such; hence, no empirical claim is made with regard to cross-dataset generalizability. Other research should explore methods of alternative treatment assignment, longitudinal results, model specification, policy thresholds, and the ethical and operational needs of the policy in the specific domain of the problem. The proposed framework should thus be treated as an analytical process that may be transferred to other domains in the future, not as a cross-domain decision-making process proven by empirical evidence.

6 CONCLUSION

This study introduced and tested an integrated workflow of an AI system that combines predictive modelling, explainability using SHAP, heterogeneous treatment-effect estimation, policy learning, and intervention utility evaluation, in a decision-centric perspective. The study

explored the correlation between predictive performance, estimated treatment heterogeneity and resource-constrained intervention allocation in randomized digital advertising with a treatment-balanced sample of 500,000 observations from the Criteo Uplift Prediction Dataset.

The results showed high prediction power of XGBoost and LightGBM with AUC-ROC around 0.945. Predictive model behaviour was explained on a probability scale by SHAP, and causal estimators produced different estimates of treatment effects and alternative ranks of interventions. This correlation was 0.3942 between XGBoost prediction rankings and Causal Forest treatment-effect rankings, suggesting that predictive response and estimated incremental treatment benefit are different but related prioritization goals. Causal Forest achieved competitive results and higher point estimates for selected intervention metrics, but was not statistically superior to prediction-based targeting for principal policy-value and net-benefit measures, as determined through paired-bootstrap comparisons.

A key result is a decision-centric comparative framework for established approaches to predictive, explainability, and causal modelling in relation to intervention-oriented decision-making. The analysis focuses on allocation budgets, simulated costs of intervention, the use of SHAP to guide allocation targeting, and subgroup allocation restrictions, as well as the distinctions between predictive discrimination and the estimation of average treatment effects on groups (conditional average treatment effects, or CATE) and intervention-policy evaluation. The study does not propose a new causal estimator or the theoretically optimal policy-learning algorithm.

The results are only applicable to the advertising benchmark evaluated by Criteo. Limited external validity due to the treatment-balanced sampling procedure, anonymized explanatory variables, binary treatment, and no longitudinal outcomes. Therefore, the results cannot be generalized to the healthcare, finance, or public administration sectors. Future studies are

needed to examine independent datasets, other treatment-assignment designs, domain-specific causal assumptions, longitudinal outcomes, operational costs, and ethical and regulatory considerations.

Acknowledgements

The author gratefully acknowledges the developers and maintainers of the publicly available Criteo Uplift Prediction Dataset for providing a valuable benchmark resource for this research. The author also acknowledges the open-source software communities whose machine-learning and causal-inference libraries facilitated the implementation and evaluation of the research framework. No additional institutional or individual assistance requiring formal acknowledgment was received.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data used in this study are publicly available through the official Criteo AI Lab distribution platform: https://ailab.criteo.com/criteo-uplift-prediction-dataset/. The processed analytical sample and derived results were generated from this publicly accessible benchmark dataset.

Code and software availability

The computational workflow and associated research materials are publicly archived on Zenodo: https://doi.org/10.5281/zenodo.23940302. The archive provides the notebook, implementation materials, reproducibility instructions, and supporting computational resources. The original Criteo Uplift Prediction Dataset is obtained separately from its original provider and is not redistributed in this archive.

Ethical Approval

This study used a publicly available, anonymized benchmark dataset and did not involve human participants, human biological material, or animals. Therefore, ethical approval was not required.

Consent to Participate

Not applicable. This study did not involve direct participation of human subjects or the collection of personally identifiable information.

Consent to Publish

Not applicable. This study does not report identifiable individual-level information requiring consent for publication.

CRediT Author Statement

Sai Srujan Reddy Tokala: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing.

REFERENCES

ACHARKI, N., LUGO, R., BERTONCELLO, A. & GARNIER, J. 2023. Comparison of meta-learners for estimating multi-valued treatment heterogeneous effects. *Proceedings of the 40th International Conference on Machine Learning.* Proceedings of Machine Learning Research, Vol. 202: PMLR.

ATHEY, S. & WAGER, S. 2021. Policy Learning With Observational Data. *Econometrica,* 89**,** 133-161.

BETLEI, A., DIEMERT, E. & AMINI, M.-R. Uplift Modeling with Generalization Guarantees. 2021.

CROSS, J. L., CHOMA, M. A. & ONOFREY, J. A. 2024. Bias in medical AI: Implications for clinical decision-making. *PLOS Digit Health,* 3**,** e0000651.

DESHPANDE, S. B., DESAI, P., GOH, K. O. M., AJAWAN, P., MATHAD, K. & DESHPANDE, U. U. 2026a. A comprehensive comparative analysis of transformer based large language models across natural language processing benchmarks. *Discover Artificial Intelligence,* 6**,** 1326.

DESHPANDE, S. B., KALKUNDRI, P. U., AJAWAN, P., JILO, J. A., KALKUNDRI, R. U. & JADHAV, R. 2026b. Analysis of Student Performance Prediction Using the Bi-LSTM Technique Focused on the Indian Higher Education Context Using a Synthetic Dataset. *Engineering Reports,* 8**,** e71084.

DESHPANDE, S. B., KALKUNDRI, R. U., AJAWAN, P., JILO, J. A., KALKUNDRI, P. U. & JADHAV, R. 2026c. Predicting Student Performance Using BiLSTM With Synthetic Institutional Data Mapping: A Case Study Contextual Integration of OULAD and AISHE in Educational Context. *Engineering Reports,* 8**,** e71025.

DESHPANDE, S. B. & MANGALWEDE, S. R. 2021. Artificial nero fuzzy inference model for location and time aware m-learning system - an empirical investigation. *International Journal of Intelligence and Sustainable Computing,* 1**,** 115-127.

DESHPANDE, S. B., MICHAEL, G. K. O., KADKOL, R. J., KAREKAR, N. V., DESHPANDE, U., LAMANI, D. & EJBALI, R. 2025a. Optimising Content Recommendations in Context-Aware Mobile Learning Platform Through Machine Learning. *Applied Computational Intelligence and Soft Computing,* 2025**,** 6982455.

DESHPANDE, S. B., TANGOD, K. K., SRINIVASAIAH, S. H., ALAHMADI, A. A., ALWETAISHI, M., ONG MICHAEL, G. K. & RAJENDRAN, S. 2025b. Elevating educational insights: sentiment analysis of faculty feedback using advanced machine learning models. *Advances in Continuous and Discrete Models,* 2025**,** 89.

DEVRIENDT, F., MOLDOVAN, D. & VERBEKE, W. 2018. A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. *Big Data,* 6**,** 13-41.

DIEMERT, E., BETLEI, A., RENAUDIN, C. & AMINI, M. R. A large-scale benchmark for uplift modeling. Proceedings of the AdKDD & TargetAd (ADKDD '18) Workshop, in conjunction with the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. ACM.

DOUTRELIGNE, M. & VAROQUAUX, G. 2025. How to select predictive models for decision-making or causal inference. *Gigascience,* 14.

FEUERRIEGEL, S., FRAUEN, D., MELNYCHUK, V., SCHWEISTHAL, J., HESS, K., CURTH, A., BAUER, S., KILBERTUS, N., KOHANE, I. S. & VAN DER SCHAAR, M. 2024. Causal machine learning for predicting treatment outcomes. *Nat Med,* 30**,** 958-968.

FRYER, D., STRUMKE, I. & NGUYEN, H. 2021. Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. *IEEE Access,* 9**,** 144352-144360.

GRABOWICZ, P., PERELLO, N. & MISHRA, A. 2022. *Marrying Fairness and Explainability in Supervised Learning*.

GUBELA, R. M. & LESSMANN, S. 2021. Uplift modeling with value-driven evaluation metrics. *Decision Support Systems,* 150**,** 113648.

HAO, X., DEMIR, E. & EYERS, D. 2024. Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction. *Technology in Society,* 78.

HU, A. 2023. Heterogeneous treatment effects analysis for social scientists: A review. *Soc Sci Res,* 109**,** 102810.

LO, V. S. Y. & PACHAMANOVA, D. A. 2023. From Meaningful Data Science to Impactful Decisions: The Importance of Being Causally Prescriptive. *Data Science Journal,* 22.

MORAES, F., MANUEL PROENÇA, H., KORNILOVA, A., ALBERT, J. & GOLDENBERG, D. 2023. Uplift Modeling: From Causal Inference to Personalization. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.* Birmingham, United Kingdom: Association for Computing Machinery.

NELLI, A. V., BATTUR, R. B., MANGASULI, S. M., TANGOD, K. K., DESHPANDE, S. B., JILO, J. A. & MISHRA, P. 2026. Balancing Trust and Performance: A Mathematical Framework for Explainable Natural Language Processing. *Applied Computational Intelligence and Soft Computing,* 2026**,** 6216386.

PONCE-BOBADILLA, A. V., SCHMITT, V., MAIER, C. S., MENSING, S. & STODTMANN, S. 2024. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin Transl Sci,* 17**,** e70056.

RIDWAN, I. B. 2025. Optimizing Enterprise Decision-Making through Causal Machine Learning Models and Real-Time Business Intelligence Integration. *International Journal of Research Publication and Reviews,* 6**,** 67-88.

RÖßLER, J. & SCHODER, D. 2022. Bridging the Gap: A Systematic Benchmarking of Uplift Modeling and Heterogeneous Treatment Effects Methods. *Journal of Interactive Marketing,* 57**,** 629-650.

SALDITT, M., ECKES, T. & NESTLER, S. 2024. A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. *Adm Policy Ment Health,* 51**,** 650-673.

SUN, W., MCFADDIN, S., TRAN, L. H., SUBRAMANIAN, S., GREENEWALD, K., TENZIN, Y., XUE, Z., DRISSI, Y. & ETTL, M. 2024. PresAIse, a prescriptive AI solution for enterprise. *INFOR: Information Systems and Operational Research,* 62**,** 629-645.

VALLEE, A. 2026. From prediction to intervention: causal digital twins for personalized clinical decision support. *J Transl Med,* 24.

WANG, D., THUNELL, S., LINDBERG, U., JIANG, L., TRYGG, J. & TYSKLIND, M. 2022. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J Environ Manage,* 301**,** 113941.

WANG, H., LIANG, Q., HANCOCK, J. T. & KHOSHGOFTAAR, T. M. 2024. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data,* 11**,** 1-16.

WANG, J., TAN, Y., JIANG, B., WU, B. & LIU, W. 2025. Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. *Symmetry,* 17**,** 610.

YAN, K., LAM, C. F., FONG, S., MARQUES, J. A. L., MILLHAM, R. C. & MOHAMMED, S. 2025. A Novel Improvement of Feature Selection for Dynamic Hand Gesture Identification Based on Double Machine Learning. *Sensors (Basel),* 25**,** 1126.

ZHANG, J., JIN, Y. & WANG, X. 2025. Comparing Meta-Learners for Estimating Heterogeneous Treatment Effects and Conducting Sensitivity Analyses. *Mathematical and Computational Applications,* 30.

Table 1: Hyperparameter Configurations of Prediction Models

Model	Hyperparameter	Configuration
Logistic Regression	Regularization strength (C)	1.0
	Solver	SAGA
	Maximum iterations	500
	Parallel jobs	-1
	Random seed	42
Random Forest	Number of trees	200
	Maximum depth	8
	Minimum samples per leaf	50
	Parallel jobs	-1
	Random seed	42
XGBoost	Number of estimators	300
	Maximum depth	6
	Learning rate	0.05
	Row subsampling	0.8
	Column subsampling	0.8
	Evaluation metric	Log-loss
	Tree construction	Histogram
	Parallel jobs	-1
	Random seed	42
LightGBM	Number of estimators	300
	Maximum depth	6
	Learning rate	0.05
	Maximum leaves	63
	Minimum child samples	50
	Subsampling parameter	0.8
	Column subsampling	0.8
	Early-stopping patience	30 rounds
	Parallel jobs	-1
	Random seed	42
	Subsample frequency	0 (library default; row bagging disabled)
	Random seed	42

Table 2. Dataset Summary Statistics

Metric	Value
Total observations	500,000
Number of features	12
Treatment ratio	0.5000
Control ratio	0.5000
Visit rate	0.0429
Conversion rate	0.002288
Missing values	0

Training samples	350,000
Validation samples	75,000
Test samples	75,000

Table 3: Comparison of Original and Analytical Sample Distributions

Indicator	Original dataset	Analytical sample
Total observations	13,979,592	500,000
Treated observations	11,882,655	250,000
Control observations	2,096,937	250,000
Treatment proportion	85.00%	50.00%
Control proportion	15.00%	50.00%
Visit rate - treated	4.8543%	4.7632%
Visit rate - control	3.8201%	3.8168%
Conversion rate - treated	0.3089%	0.2816%
Conversion rate - control	0.1938%	0.1760%
Treated sampling fraction	-	2.104%
Control sampling fraction	-	11.922%

Table 4: Predictive Model Performance on the Validation Dataset

Model	AUC-ROC	AUC-PR	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.92781	0.45911	0.5978	0.2860	0.38688	0.9611
Random Forest	0.94356	0.48646	0.6474	0.2157	0.32362	0.9613
XGBoost	0.94467	0.48729	0.6229	0.2726	0.37924	0.9617
LightGBM	0.94455	0.49013	0.6270	0.2624	0.36993	0.9617

Table 5: Statistical Significance Tests

Model comparison	Metric	Difference	95% CI	Adjusted p
RF vs LR	AUC-ROC	+0.01575	[0.01284, 0.01890]	.012
RF vs LR	AUC-PR	+0.02735	[0.02018, 0.03470]	.012
RF vs LR	F1	-0.06326	[-0.07698, -0.05001]	.012
XGBoost vs LR	AUC-ROC	+0.01686	[0.01377, 0.02015]	.012
XGBoost vs LR	AUC-PR	+0.02818	[0.02061, 0.03685]	.012
XGBoost vs LR	F1	-0.00764	[-0.02078, 0.00617]	.541
LightGBM vs LR	AUC-ROC	+0.01674	[0.01372, 0.01999]	.012

LightGBM vs LR	AUC-PR	+0.03102	[0.02386, 0.03889]	.012
LightGBM vs LR	F1	-0.01695	[-0.02885, -0.00443]	.045
LightGBM vs XGBoost	AUC-ROC	-0.00012	[-0.00052, 0.00029]	.590
LightGBM vs XGBoost	AUC-PR	+0.00284	[-0.00096, 0.00658]	.515
LightGBM vs XGBoost	F1	-0.00931	[-0.01843, -0.00039]	.184

Note. Confidence intervals are unadjusted 95% paired-bootstrap percentile intervals. P-values are adjusted using the Holm procedure across the 12 pairwise metric comparisons. Statistical significance is determined using the Holm-adjusted p-values at the 0.05 level.

Table 6: Mean Absolute SHAP Feature Importance (Visit-Probability Scale)

Rank	Feature	XGBoost Mean |SHAP|	LightGBM Mean |SHAP|
1	f8	0.035269	0.027531
2	f9	0.017085	0.015640
3	f2	0.014812	0.016447
4	f6	0.005651	0.004973
5	f0	0.003142	0.001949
6	f3	0.001037	0.000927
7	f7	0.000808	0.000621
8	f10	0.000541	0.000344
9	f5	0.000465	0.000188
10	f4	0.000259	0.000173
11	f11	0.000224	0.000178
12	f1	0.000093	0.000092

Table 7: Treatment Effect Estimation Results

Estimator	Mean CATE	SD	Minimum	Maximum	Positive CATE (%)
S-Learner	0.00112	0.00539	-0.00366	0.06291	20.06
T-Learner	0.00552	0.02470	-0.19334	0.30012	82.70
X-Learner	0.00556	0.01692	-0.11337	0.17229	95.47
LinearDML	0.00570	0.01455	-0.24824	0.13290	52.11
Causal Forest	0.00552	0.01587	-0.10294	0.11990	61.09
Two-Model Random Forest Uplift	0.00576	0.02022	-0.13839	0.18622	83.03

Table 8: Intervention utility, fairness, and Decision Evaluation

Policy	Uplift gain	Policy value	Net benefit

Random	0.009966	0.041040	0.039540
Logistic Regression	0.024466	0.046907	0.045407
Random Forest	0.025662	0.046933	0.045433
XGBoost	0.027201	0.047333	0.045833
LightGBM	0.026420	0.047173	0.045673
SHAP-assisted	0.025315	0.046933	0.045433
X-Learner	0.027167	0.047733	0.046233
Causal Forest	0.026372	0.046827	0.045327
Two-Model Random Forest Uplift	0.026803	0.047307	0.045807
Cost-sensitive causal	0.026526	0.046853	0.045364
Fairness-aware causal	0.027984	0.046507	0.045007

Figure 1: Conceptual Decision-Centric AI Framework

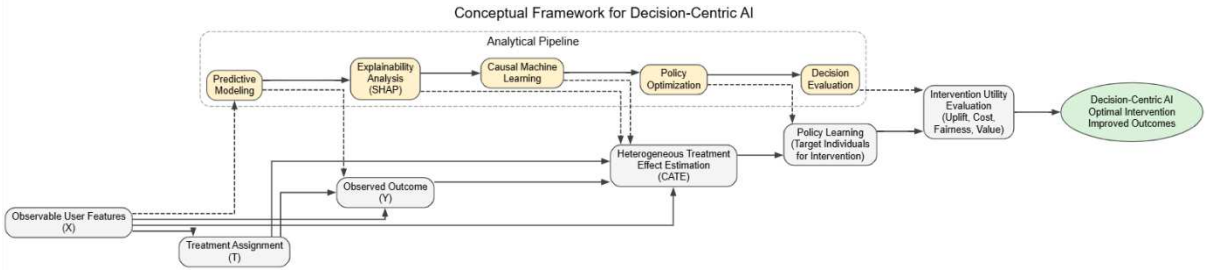


Figure 2: Experimental Workflow

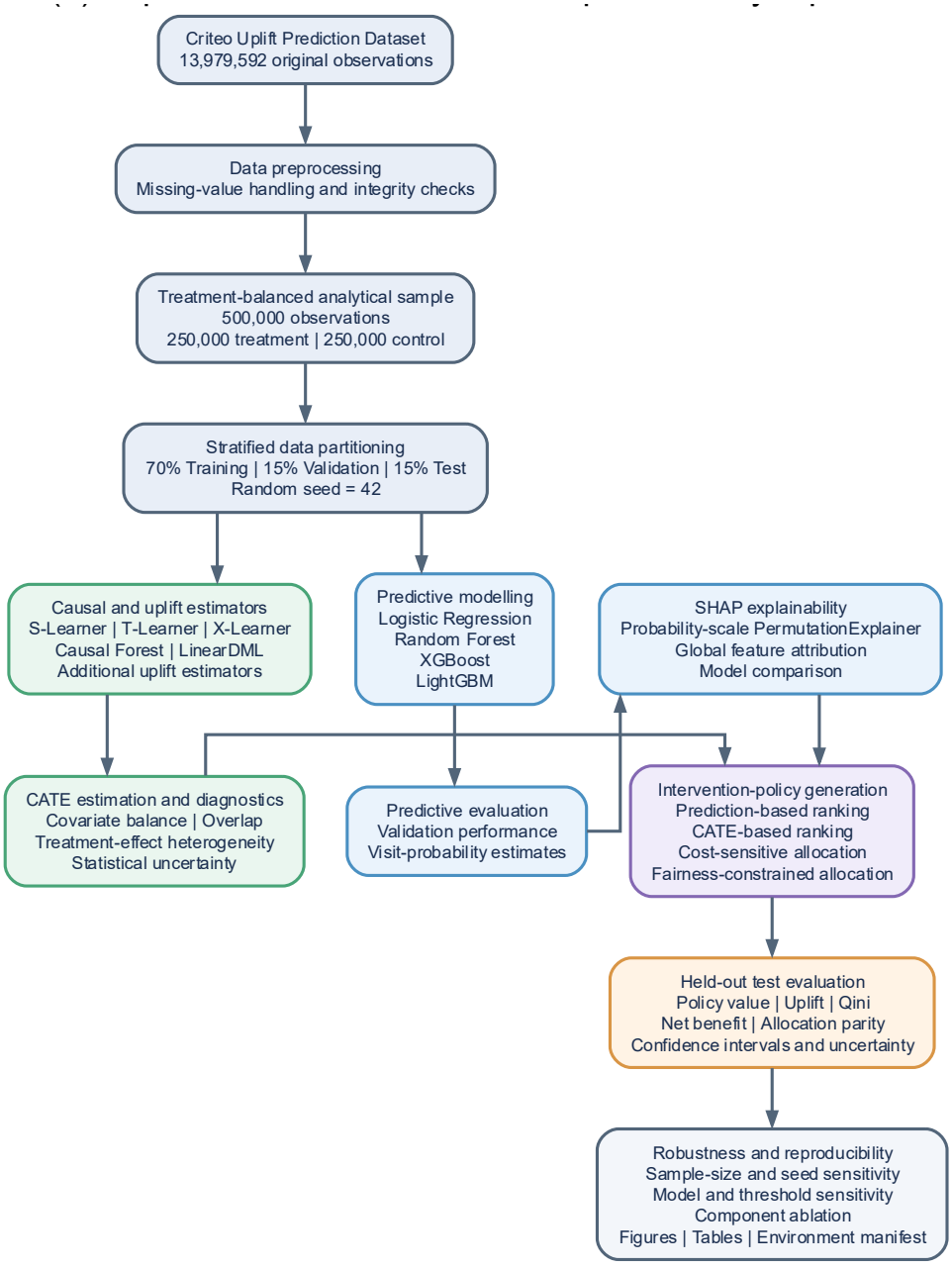


Figure 3: Distribution of treatment groups, outcome variables, feature characteristics, and correlations among study variables. (a) Distribution of treatment classifications, visit rates, and conversion outcomes; (b) feature distributions stratified by treatment group; and (c) correlation matrix illustrating relationships among the features, outcomes, and treatment assignment.

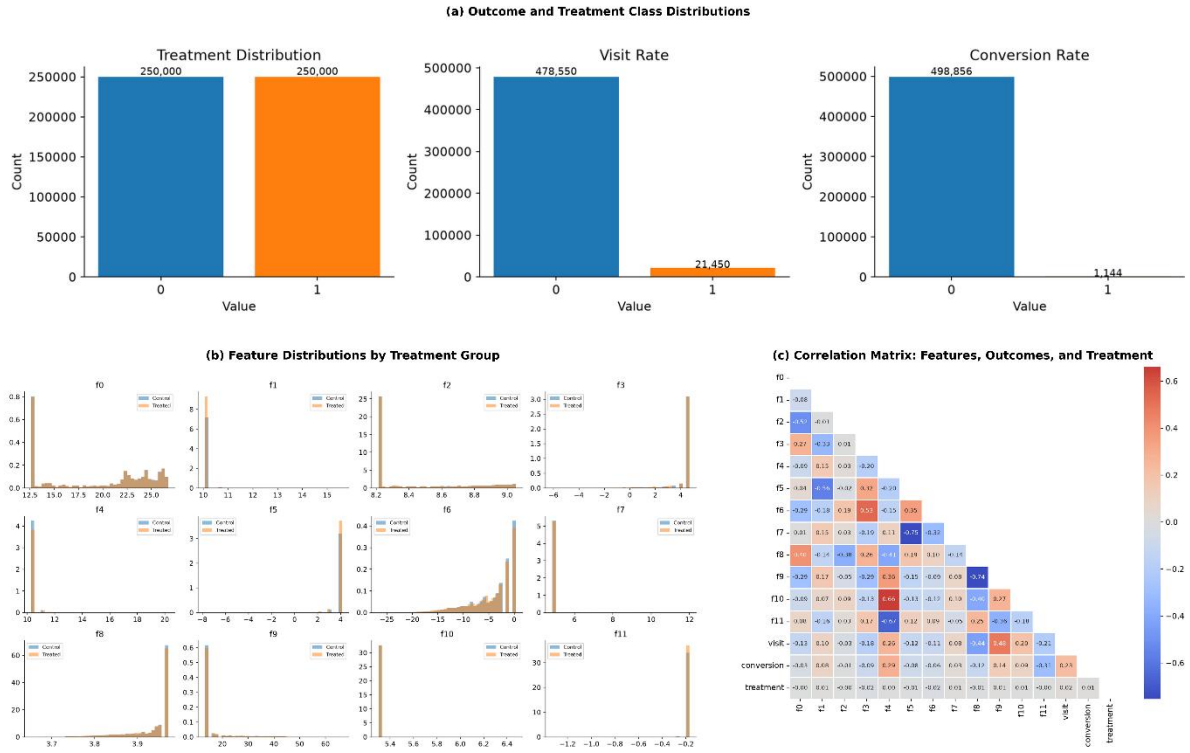


Figure 4: Predictive Model Performance

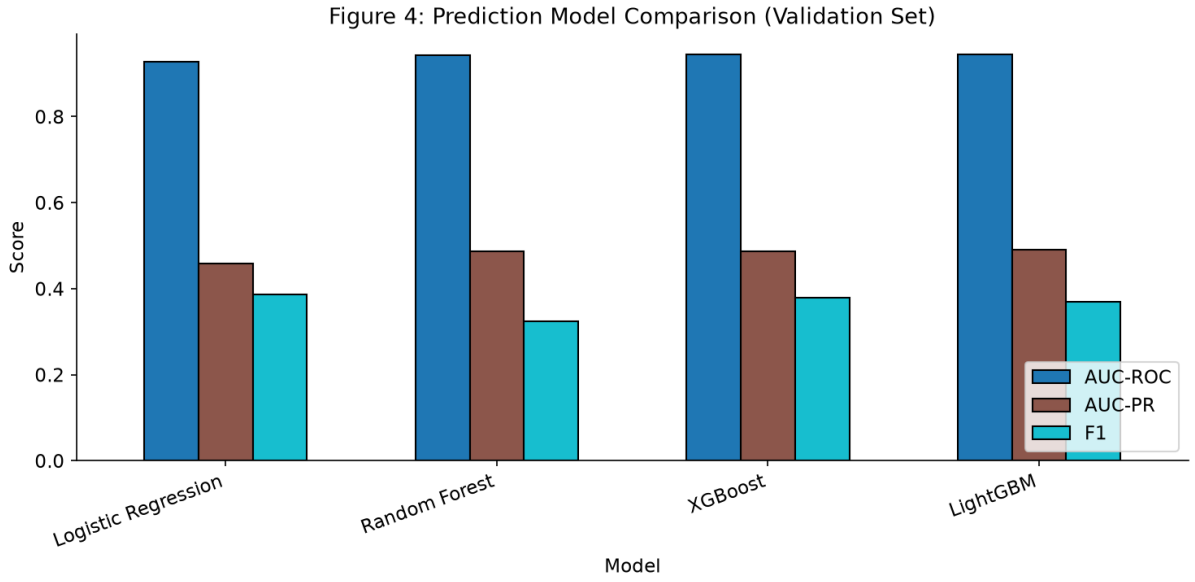


Figure 5: SHAP-Based Explainability a) Comparable SHAP feature importance-probability scale b) LightGBM: Mean (SHAP) C) XGBOOST SHAP summary-visit probability d) LightGBM SHAP summary-visit probability

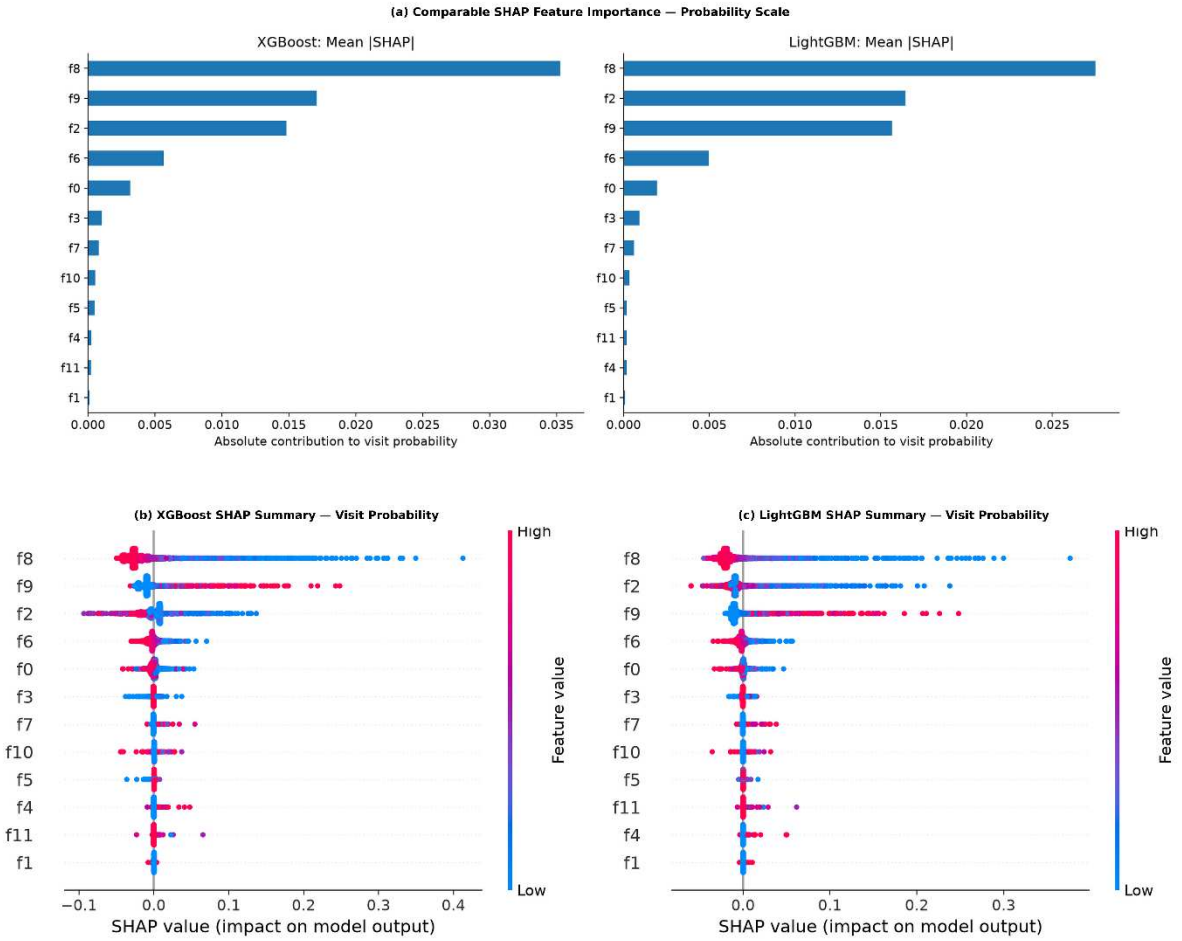


Figure 6: CATE Distributions Across Six Estimators

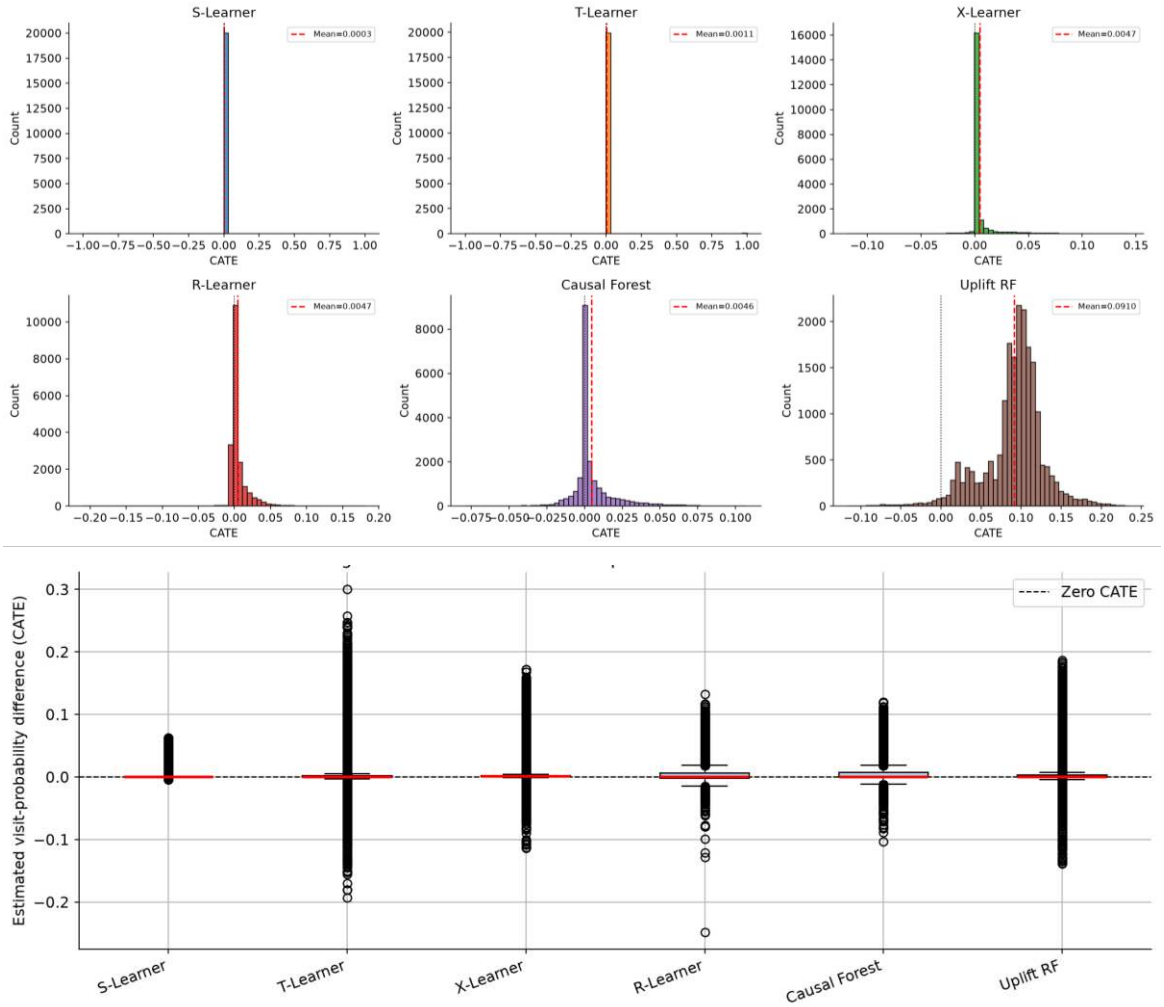
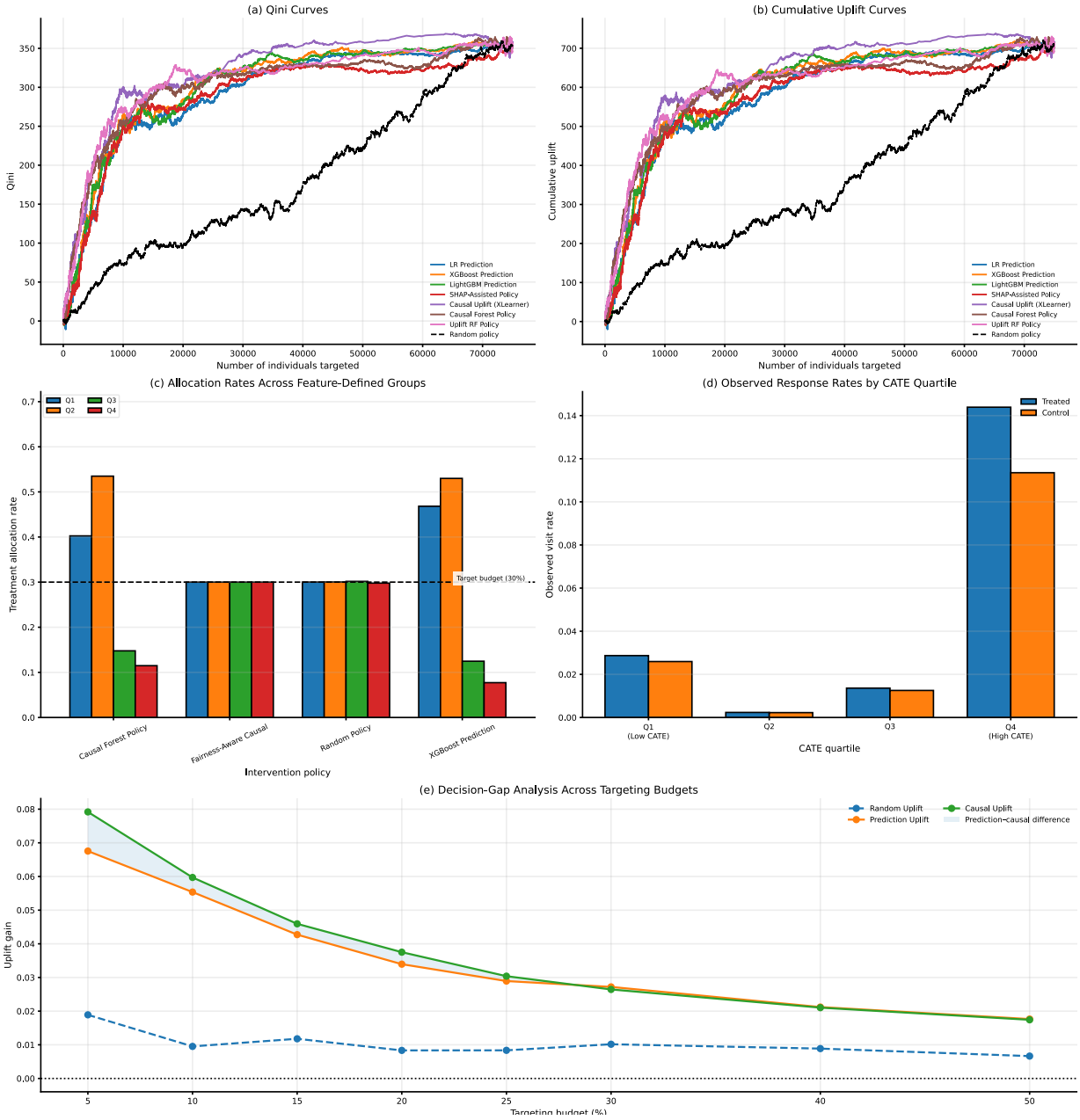


Figure 7: Intervention Policy Evaluation



Appendices

Appendix A. Additional Statistical Results and Reproducibility Details

Table A1. Predictive Model Performance with 95% Confidence Intervals

Model	Metric	Estimate	95% CI Lower	95% CI Upper
Logistic Regression	AUC-ROC	0.92781	0.92232	0.93296
Logistic Regression	AUC-PR	0.45911	0.44251	0.47686
Logistic Regression	F1-score	0.38688	0.36996	0.40317
Random Forest	AUC-ROC	0.94356	0.94041	0.94721
Random Forest	AUC-PR	0.48646	0.46939	0.50550
Random Forest	F1-score	0.32362	0.30625	0.34152
XGBoost	AUC-ROC	0.94467	0.94144	0.94814
XGBoost	AUC-PR	0.48729	0.47143	0.50530
XGBoost	F1-score	0.37924	0.36201	0.39693
LightGBM	AUC-ROC	0.94455	0.94133	0.94810
LightGBM	AUC-PR	0.49013	0.47354	0.50886
LightGBM	F1-score	0.36993	0.35317	0.38723

Note. CI = confidence interval; AUC-ROC = area under the receiver operating characteristic curve; AUC-PR = average precision. Estimates were calculated on the validation dataset (n = 75,000). Confidence intervals were obtained using 1,000 stratified bootstrap resamples and the percentile method. Intervals represent evaluation-sample uncertainty conditional on the fitted models.

Table A2. Causal Ranking and Targeting Performance with 95% Confidence Intervals

Model	Metric	Estimate	95% CI Lower	95% CI Upper
XGBoost Prediction	Normalized Qini AUC	0.05585	0.00488	0.10933
XGBoost Prediction	Normalized Uplift AUC	0.00827	0.00076	0.01643
XGBoost Prediction	Empirical Uplift at 30%	0.01599	0.00065	0.03219
LightGBM Prediction	Normalized Qini AUC	0.05543	0.00524	0.10988
LightGBM Prediction	Normalized Uplift AUC	0.00820	0.00075	0.01629
LightGBM Prediction	Empirical Uplift at 30%	0.01653	0.00078	0.03242
X-Learner	Normalized Qini AUC	0.08941	0.03327	0.14832
X-Learner	Normalized Uplift AUC	0.01315	0.00490	0.02212
X-Learner	Empirical Uplift at 30%	0.02051	0.00532	0.03428
S-Learner	Normalized Qini AUC	0.06530	0.00868	0.12179
S-Learner	Normalized Uplift AUC	0.00967	0.00135	0.01807
S-Learner	Empirical Uplift at 30%	0.01832	0.00374	0.03358
T-Learner	Normalized Qini AUC	0.08893	0.03035	0.14505
T-Learner	Normalized Uplift AUC	0.01309	0.00440	0.02129
T-Learner	Empirical Uplift at 30%	0.01992	0.00625	0.03364
Uplift RF	Normalized Qini AUC	0.08372	0.02499	0.14093
Uplift RF	Normalized Uplift AUC	0.01229	0.00367	0.02079

Uplift RF	Empirical Uplift at 30%	0.01775	0.00383	0.03161
LinearDML	Normalized Qini AUC	0.06514	0.01063	0.11731
LinearDML	Normalized Uplift AUC	0.00958	0.00159	0.01741
LinearDML	Empirical Uplift at 30%	0.01557	0.00051	0.03067
Causal Forest	Normalized Qini AUC	0.05757	0.00467	0.11695
Causal Forest	Normalized Uplift AUC	0.00851	0.00062	0.01707
Causal Forest	Empirical Uplift at 30%	0.02070	0.00649	0.03504

Note. Results are based on the held-out causal evaluation subset. Confidence intervals were calculated using 1,000 treatment-stratified bootstrap resamples. Empirical uplift at 30% represents the treatment-control outcome difference among observations selected under the corresponding ranking policy.

Table A3a. Paired Comparisons of Causal and Prediction-Based Targeting Performance

Comparison	Metric	Difference	95% CI
CF vs XGBoost	Normalized Qini AUC	0.00172	[-0.04339, 0.05396]
CF vs XGBoost	Normalized Uplift AUC	0.00024	[-0.00674, 0.00795]
CF vs XGBoost	30% Empirical Uplift	0.00471	[-0.00385, 0.01313]
CF vs LightGBM	Normalized Qini AUC	0.00214	[-0.04227, 0.05369]
CF vs LightGBM	Normalized Uplift AUC	0.00031	[-0.00655, 0.00800]
CF vs LightGBM	30% Empirical Uplift	0.00417	[-0.00416, 0.01282]
X-Learner vs XGBoost	Normalized Qini AUC	0.03356	[-0.01650, 0.08786]
X-Learner vs XGBoost	Normalized Uplift AUC	0.00488	[-0.00259, 0.01311]
X-Learner vs XGBoost	30% Empirical Uplift	0.00452	[-0.00384, 0.01232]
LinearDML vs XGBoost	Normalized Qini AUC	0.00929	[-0.02366, 0.04333]
LinearDML vs XGBoost	Normalized Uplift AUC	0.00131	[-0.00359, 0.00656]
LinearDML vs XGBoost	30% Empirical Uplift	-0.00042	[-0.00693, 0.00615]
Uplift RF vs XGBoost	Normalized Qini AUC	0.02787	[-0.03233, 0.08536]
Uplift RF vs XGBoost	Normalized Uplift AUC	0.00402	[-0.00485, 0.01248]
Uplift RF vs XGBoost	30% Empirical Uplift	0.00177	[-0.00778, 0.01046]

Note. CF = Causal Forest. Differences were calculated as the first estimator's performance minus the reference estimator's performance using 1,000 paired, treatment-stratified bootstrap resamples of identical held-out evaluation observations. Confidence intervals were obtained using the percentile method. None of the reported intervals excluded zero. These comparisons do not establish estimator equivalence or universal performance superiority.

Table A3b. Paired Bootstrap Comparisons of Intervention Policy Utility

Causal Forest minus XGBoost	Difference	95% CI

Uplift gain	0.003965	[-0.004648, 0.013333]
Qini coefficient	0.001720	[-0.043788, 0.050864]
Policy value	0.001200	[-0.000903, 0.003600]
Net benefit	0.001177	[-0.000945, 0.003562]

Note. Differences represent Causal Forest minus XGBoost. Estimates were calculated using 1,000 paired, treatment-stratified bootstrap resamples of the same 20,000-observation held-out evaluation subset. The subset was sampled from the 75,000-observation test partition with 10,000 observations from each treatment arm. Policies were fixed before resampling, and all competing methods were evaluated using identical resampled observations. Confidence intervals were calculated using the percentile method. These intervals quantify evaluation-sample uncertainty conditional on fitted models and do not incorporate model-refitting uncertainty. None of the four policy-utility confidence intervals excluded zero.

Table A4. Model-Based Average Treatment Effect Estimates and 95% Confidence Intervals

Model	Estimated ATE	95% CI Lower	95% CI Upper
LinearDML	0.005699	0.003476	0.007921
Causal Forest	0.005524	-0.026337	0.037386

Note. ATE = average treatment effect; CI = confidence interval. Estimates represent average model-estimated CATEs over the held-out evaluation observations. Confidence intervals were obtained using the fitted EconML estimators' model-based inference procedures at the 95% confidence level. These intervals quantify uncertainty in the estimated average treatment effects and should not be interpreted as individual-level CATE confidence intervals.

Table A5. Causal Machine-Learning Estimator Configurations

Estimator	Base learners	Trees	Maximum depth	Minimum leaf size	Cross-fitting
S-Learner	One RF regressor	100	6	50	Not configured
T-Learner	Two RF regressors	100 each	6	50	Not configured

X-Learner	RF outcome, CATE and propensity models	100 each	6	50	Not configured
LinearDML	RF outcome and treatment nuisance models; linear final stage	100 each	6	50	3 folds
Causal Forest DML	RF outcome and treatment nuisance models; causal forest final stage	100 nuisances; 200 causals	6 nuisances	50 nuisances; 20 causals	3 folds
Two-Model Uplift RF	Two RF classifiers	150 each	8	50	Not configured

Note. RF = Random Forest; CATE = Conditional Average Treatment Effect; DML = Double Machine Learning. All estimators used twelve anonymized covariates, binary randomized treatment assignment, and binary visit outcome. Causal models were fitted on a common treatment-balanced training subset of 80,000 observations using random seed 42. Random Forest nuisance and outcome models used 100 trees, maximum depth six, and minimum leaf size 50. Causal Forest DML used 200 final-stage causal trees and a minimum causal leaf size of 20. The two-model uplift estimator used 150 trees per treatment-specific classifier. Three-fold cross-fitting was configured for LinearDML and CausalForestDML. No automated hyperparameter search was conducted. Parameters not explicitly specified retained their respective software-library defaults.

Table A6. Computational Environment and Software Versions

Component	**Version**
Python	3.11.9
NumPy	1.26.4
pandas	2.3.3
SciPy	1.15.3
scikit-learn	1.6.1
XGBoost	3.2.0
LightGBM	4.7.0
SHAP	0.48.0
EconML	0.16.0
scikit-uplift	0.5.1
statsmodels	0.14.6
Hugging Face datasets	4.8.4
PyArrow	24.0.0
Matplotlib	3.11.1

Seaborn	0.13.2
Pillow	12.2.0
huggingface-hub	1.12.0
Joblib	1.5.3

Note. Primary random seed = 42; analytical sample size = 500,000; training/validation/test allocation = 70:15:15; primary outcome = binary visit; treatment = randomized advertising assignment. Package versions correspond to the recorded notebook execution environment.