Machine Learning for Predicting and Optimizing User Behavior in Digital Marketing

Stanela Gacova, Cveta Martinovska Bande

Faculty of Computer Science, Goce Delcev University, Stip, North Macedonia
stanela.210231@student.ugd.edu.mk
cveta.martinovska@ugd.edu.mk

DOI: 10.29322/IJSRP.16.09.2026.p17719
https://dx.doi.org/10.29322/IJSRP.16.09.2026.p17719

Paper Received Date: 20th August 2026
Paper Acceptance Date: 11th September 2026
Paper Publication Date: 20th September 2026

Abstract- The increasing use of customer data in digital marketing has created new opportunities for applying machine learning to marketing decisions. This study uses the Predict–Estimate–Explain–Optimize (PEEO) framework. The empirical analysis is conducted on a random sample of 500,000 observations from the Criteo Uplift Prediction Dataset, a large-scale benchmark derived from randomized controlled experiments in digital advertising. Logistic Regression, Random Forest, and XGBoost are used for conversion prediction, while a T-Learner is used to estimate the individual treatment effect. SHAP analysis is applied to compare the features that influence conversion and uplift. Random, Predictive, and Uplift Targeting are then compared under different targeting levels and economic scenarios. The results show that the best targeting strategy depends on the share of users being targeted. Predictive Targeting performed better at the 1%, 5%, and 20% targeting levels, while Uplift Targeting performed better at 10%. This shows that a model with good predictive performance is not always the best choice for allocating a marketing budget. The study combines predictive and causal machine learning in one framework and shows the importance of considering both conversion probability and the estimated effect of a marketing intervention when making targeting decisions.

Index Terms- causal machine learning, digital marketing, machine learning, targeting optimization, uplift modeling.

I. INTRODUCTION

Digital marketing has become a data-intensive discipline in which nearly every user interaction — a click, a visit, a purchase — is recorded and can, in principle, be used to guide targeting decisions. Machine learning has driven a shift from descriptive reporting toward predictive modeling of clicks, conversions, churn and customer value, and is now embedded in the operational infrastructure of advertising platforms rather than being used only as an offline analytical tool. Yet a model that accurately predicts who is likely to convert answers a narrower question than the one marketers actually need answered: who will convert because of the marketing intervention. A user with a high baseline probability of purchasing may convert regardless of whether an advertisement is shown, while a user with a lower baseline probability may convert only as a direct result of that exposure. Treating these two users as equivalent, simply because a model assigns them a similar conversion probability, risks directing budget toward customers Radcliffe **[1]** termed ‘sure things’ rather than toward ‘persuadables’ — the users whose behavior the intervention can actually change.

This difference has led to increased interest in uplift modeling and causal machine learning in marketing. These approaches focus on estimating the individual treatment effect (ITE) of a marketing intervention rather than only predicting the probability of an outcome. As the models used for targeting become more complex, their interpretability becomes a practical concern, since marketing teams need to understand which factors drive both predicted conversion and predicted uplift rather than receive an opaque score. Explainable AI techniques such as SHAP make this possible by attributing each prediction to the contribution of individual features. Previous studies have mainly examined predictive modeling, uplift estimation, explainability, and budget-constrained targeting as separate research areas. However, only a limited number of studies have combined all four aspects using the same dataset and targeting setting.

To address this gap, the study proposes the Predict–Estimate–Explain–Optimize (PEEO) framework, which combines four main stages. First, conversion probability is predicted, followed by the estimation of individual treatment effects using a T-Learner. SHAP is then used to explain the model results, while the final stage focuses on optimizing the targeting strategy under a limited marketing budget. Using the Criteo Uplift Prediction Dataset — a randomized-controlled-trial benchmark widely used in causal machine learning research — the study investigates a central question: does the model that most accurately predicts conversion also provide the best basis for allocating a constrained marketing budget? Specifically, the paper (1) compares Logistic Regression, Random Forest and XGBoost for

conversion prediction; (2) applies a T-Learner to estimate heterogeneous treatment effects and evaluates it through uplift deciles and the Qini curve; (3) compares the SHAP-based drivers of conversion and of uplift; and (4) evaluates Random, Predictive and Uplift targeting strategies across a range of budget, cost and conversion-value scenarios, including a bootstrap and sensitivity analysis of their relative stability.

II. RELATED WORK

Machine learning is widely used in several areas of digital marketing, such as click-through rate (CTR) and conversion prediction, recommendation systems, and personalization. In recent years, causal and uplift modeling have also received increasing attention. Xie et al. **[2]** map the broader research landscape of AI-driven marketing and identify personalization, customer experience and service automation as the dominant application areas, while also flagging privacy, algorithmic bias and ethical constraints as recurring concerns. Kaponis et al. **[3]** similarly review how machine learning improves user experience through personalization and behavioral analysis, emphasizing that the value of these models lies not only in higher sales but in the relevance of the experience delivered to the user.

Many studies on CTR prediction use the Criteo Display Advertising Challenge dataset. Recent research has also focused on improving the performance of deep learning models such as xDeepFM, which are commonly evaluated using AUC and Log Loss **[4].** These studies confirm that even small metric improvements can be economically meaningful at the scale of programmatic advertising, but they generally stop at predictive performance and do not examine whether a more accurate ranking of users translates into a larger causal effect of the campaign.

Uplift modeling estimates the individual treatment effect $\tau(x) = P(Y=1|T=1,X=x) - P(Y=1|T=0,X=x)$ rather than $P(Y=1|X)$, with meta-learner approaches such as the T-, S- and X-Learner **[5]** and tree- or forest-based causal estimators **[6]–[8]** among the most widely used implementations. Gutierrez and Gérardy **[9]** review this literature and note that classical A/B testing estimates only the average treatment effect, whereas uplift methods allow the effect to vary across users, which is precisely the information a budget-constrained campaign needs. Devriendt et al. **[10]** provide a systematic evaluation of uplift methods and highlight the Qini curve and related cumulative-gain measures as the standard tools for assessing whether a model ranks users correctly by expected incremental response, an approach followed in the present study.

Model explainability has also become an important topic in recent research. Lundberg and Lee **[11]** introduced SHAP as a unified, game-theoretic approach to attributing model output to individual features, and Ribeiro et al. **[12]** motivated the broader need for local explanations of otherwise opaque classifiers. In marketing specifically, explainability has been discussed mainly as a trust and governance requirement rather than as a tool for comparing what drives conversion against what drives incremental response — a comparison this paper makes explicit.

Table 1 summarizes the main findings from the related studies and shows how the present study differs from previous research. Existing studies tend to advance either predictive accuracy, or uplift estimation, or model interpretability, or campaign-level optimization, but rarely combine all four elements within one experimental design evaluated on identical data. To address this research gap, this study compares random, predictive, and uplift targeting using the same dataset. It also compares the main features that influence conversion and uplift and examines how the best targeting strategy changes depending on the available marketing budget.

Research direction	Representative source	Main focus	Gap relative to this study
AI-marketing landscape	Xie et al. [2]	Trend mapping across AI-marketing applications	No experimental comparison of predictive vs. causal targeting
Personalization / UX	Kaponis et al. [3]	Role of ML in personalization and user experience	Limited comparison of concrete targeting strategies
CTR / conversion prediction	Alotaibi & Alotaibi [4]	Deep feature-interaction models for CTR	Focus on predictive accuracy only, no incremental-effect analysis
Uplift / causal ML	Gutierrez & Gérardy [9]	Review of causal inference for marketing response	Conceptual review, no budget-constrained empirical evaluation
This study	—	Predictive ML, uplift modeling (T-Learner) and SHAP compared jointly under budget constraints	Integrates prediction, causal estimation, explanation and economic optimization on one dataset

Table 1. Comparative overview of related research and the identified gap.

III. MATERIALS AND METHODS

The empirical part of this study is based on the Predict–Estimate–Explain–Optimize (PEEO) framework, which consists of four main phases. In the Predict phase, the conversion probability $P(Y=1|X)$ is estimated. The Estimate phase focuses on calculating the individual treatment effect, $\tau(X) = P(Y=1|T=1,X) - P(Y=1|T=0,X)$. In the Explain phase, SHAP is used to identify and compare the factors that influence these estimates. The Optimize phase uses the obtained results to compare targeting strategies under a limited marketing budget. All analysis was carried out in Python, using Pandas and NumPy for data handling, scikit-learn for the predictive models, XGBoost for the gradient-boosting and T-Learner components, and SHAP for interpretability.

A. Dataset

The study uses the Criteo Uplift Prediction Dataset, a benchmark developed by Criteo AI Lab from randomized controlled experiments in digital advertising **[13], [14]**. Because users are randomly assigned to a treatment group (exposed to advertising) or a control group (not exposed), the dataset supports an unbiased estimate of the incremental effect of the intervention, which is not possible with observational click or purchase data alone. The original dataset contains 13,979,592 records and 16 variables. To make the analysis more manageable, a random sample of 500,000 records was used. The data were first shuffled using a fixed random seed of 42, and the first 500,000 rows were then selected. This sample was used to reduce the computational requirements while maintaining the statistical structure of the original dataset. The sample uses twelve anonymized numerical features (f0–f11), a binary treatment indicator T, and a binary conversion outcome Y. Feature semantics are not disclosed by Criteo in order to protect commercial and user information, so the analysis in this paper focuses on the statistical and relative importance of the features rather than on their substantive marketing meaning.

The sample is strongly imbalanced: of 500,000 users, only 1,457 (0.29%) recorded a conversion, against 498,543 (99.71%) who did not, which is representative of real digital-advertising campaigns where only a small share of exposed users act on an intervention. Treatment assignment is also unbalanced by design, with 425,026 users (85.0%) in the treatment group and 74,974 (15.0%) in the control group. No missing values were found in any of the 16 variables, so no imputation was required.

B. Data preparation

The twelve anonymized features (f0–f11) were used as predictors, conversion (Y) as the target variable for the predictive models, and treatment (T) as the group indicator for the causal analysis. The dataset was split into training (80%, 400,000 records) and test (20%, 100,000 records) sets using stratified sampling on the conversion label, so that both subsets preserve the same, highly imbalanced conversion rate; all models were trained and evaluated on identical splits to ensure a fair comparison. Preprocessing was model-specific: features were standardized (zero mean, unit variance, parameters fit on the training set only) for Logistic Regression, which is sensitive to feature scale, while Random Forest and XGBoost were trained on the raw feature values, consistent with their tree-based decision mechanism, which is scale-invariant.

C. Predictive modeling (predict)

Three different machine learning algorithms were compared for estimating $P(Y=1|X)$: Logistic Regression, Random Forest, and XGBoost. Logistic Regression was used as a linear baseline, while Random Forest represents an ensemble of decision trees **[15]**. XGBoost uses gradient boosting and is well suited for tabular data **[16]**. It was also used later as the base learner for the causal model. Because of the pronounced class imbalance, Accuracy was not used for evaluation, since a trivial classifier that always predicts the majority class would score close to 100% without any practical value. Instead, models were compared using ROC-AUC (overall ranking ability), PR-AUC (performance on the rare positive class), Log Loss (calibration of the predicted probabilities), and, as secondary indicators, Precision, Recall and F1-score at the default 0.5 threshold. In addition, a Top-k targeting analysis ranked users by predicted conversion probability and measured how many of the actually observed conversions fell within the top 1%, 5%, 10% and 20% of the ranking, which reflects how a limited marketing budget would realistically be allocated.

D. Uplift modeling with a t-learner (estimate)

Before causal modeling, an aggregate comparison of the treatment and control groups was performed. A two-proportion Z-test was used to check whether the difference between the treatment and control conversion rates was statistically significant, providing an aggregate justification for pursuing individual-level effect estimation. Individual treatment effects were then estimated with a T-Learner, one of the most widely used meta-learning approaches for heterogeneous treatment effect estimation **[5]**. The method trains two independent XGBoost models on the same feature set: one on the treatment-group data, estimating $P(Y=1|T=1,X)$, and one on the control-group data, estimating $P(Y=1|T=0,X)$. The individual treatment effect (uplift score) for each user is then $\tau(x) = P(Y=1|T=1,X=x) - P(Y=1|T=0,X=x)$. Given the class imbalance within each group, two variants were compared: an Unweighted T-Learner, trained without additional class rebalancing, and a Weighted T-Learner, in which the scale_pos_weight parameter of XGBoost was set separately for the treatment and control models. The variant whose average predicted effect was closer to the observed treatment effect was selected for further analysis. It was chosen because its estimate of the incremental effect was more conservative and better calibrated.

The resulting uplift scores were evaluated by ranking the test-set users into ten equally sized deciles from highest to lowest predicted uplift, and comparing the average predicted uplift in each decile with the corresponding observed uplift (the difference in conversion rates between treated and control users within that decile). The cumulative Qini curve was used to evaluate the performance of the uplift

model **[10]**. It shows the cumulative incremental conversions obtained when users are targeted from the highest to the lowest predicted uplift and compares them with random targeting. The Normalized Area Above Random (NAAR) was also calculated to measure the difference between the model's cumulative gain curve and the random-targeting curve.

E. Explainability (explain)

SHAP (SHapley Additive exPlanations) values were computed to identify which of the twelve anonymized features contribute most to (a) the predicted conversion probability and (b) the predicted uplift score, based on a sample of 5,000 test-set users **[11]**. Global importance was summarized by the mean absolute SHAP value for each feature, and the resulting rankings for the conversion model and the uplift model were compared directly, including the rank shift of each feature between the two rankings and the Spearman rank correlation between them. This comparison tests directly whether the features that best explain who is likely to convert are the same features that best explain who is likely to convert because of the intervention.

F. Targeting optimization (optimize)

In the final phase, three targeting strategies were compared: Random Targeting, in which users are selected without any model; Predictive Targeting, in which users are ranked and selected by predicted conversion probability; and Uplift Targeting, in which users are ranked and selected by predicted individual treatment effect. Each strategy was evaluated under four targeting shares (1%, 5%, 10% and 20% of users) by computing the incremental number of conversions attributable to targeting that share of the ranked population.

To assess statistical robustness, a bootstrap resampling analysis (with resampling of the test set) compared the average incremental effect and the 95% confidence interval of the difference between Predictive and Uplift Targeting at each budget level, together with the empirical probability that one strategy outperforms the other in a given resample. An economic scenario and sensitivity analysis translated incremental conversions into a simulated return on investment (ROI), combining three assumed values of a realized conversion (€50, €100, €200), four assumed per-user targeting costs (€0.05, €0.10, €0.50, €1.00) and the four targeting shares described above, for a total of 48 economic scenarios. For each scenario the optimal strategy (the one yielding the highest simulated ROI) was recorded, allowing an assessment of how consistently the ranking of strategies holds as economic assumptions change. As stated throughout the source research, these ROI figures are simulated illustrative values for methodological validation rather than the real financial results of an operating company.

IV. RESULTS

This section reports the empirical results of the four PEEO phases in sequence: predictive model comparison, individual treatment effect estimation, SHAP-based explanation, and targeting optimization.

A. Predict: conversion prediction

Table 2 reports the test-set performance of the three predictive models. All three achieved high discriminative ability, with ROC-AUC above 0.94, but the ranking of models changes depending on the metric used. Logistic Regression obtained the highest ROC-AUC (0.9617), ahead of XGBoost (0.9488) and Random Forest (0.9430), indicating the strongest overall ranking of users by conversion probability. On PR-AUC — the more informative metric for this strongly imbalanced dataset — XGBoost performed best (0.2237), narrowly ahead of Logistic Regression (0.2189) and Random Forest (0.1923), showing the strongest ability to isolate the rare positive class. Random Forest achieved the lowest (best) Log Loss (0.0638), reflecting the best-calibrated probability estimates, while Logistic Regression combined the highest recall (0.8866) with the lowest precision (0.0288), and Random Forest combined the highest precision (0.0976) and F1-score (0.1685) with the lowest recall (0.6186).

Model	ROC-AUC	PR-AUC	Log Loss	Precision	Recall	F1-score
Logistic Regression	0.9617	0.2189	0.2713	0.0288	0.8866	0.0558
Random Forest	0.9430	0.1923	0.0638	0.0976	0.6186	0.1685
XGBoost	0.9488	0.2237	0.1258	0.0470	0.7354	0.0883

Table 2. Predictive performance of the conversion models on the held-out test set.

A Top-k targeting analysis provides a more budget-relevant comparison. XGBoost captured the most true conversions within the top 1% of ranked users (154, versus 148 for Logistic Regression and 146 for Random Forest), whereas Logistic Regression captured the most conversions once the targeted share widened to 5%, 10% or 20% (230, 264 and 280 conversions, respectively). This pattern indicates that XGBoost concentrates its correct predictions more tightly at the very top of the ranking, which is valuable under a very small budget, while Logistic Regression spreads correct predictions more evenly across a wider share of the ranked population, which is preferable once a larger share of users can be targeted. Because XGBoost combined the strongest PR-AUC with the best top-1% precision — the setting most representative of a constrained marketing budget — it was retained as the base learner for the causal analysis that follows.

B. Estimate: individual treatment effect

The aggregate comparison of the two experimental groups showed a conversion rate of 0.3070% in the treatment group (1,305 conversions among 425,026 users) versus 0.2027% in the control group (152 conversions among 74,974 users), a difference of 0.1043

percentage points, equivalent to a 51.5% relative increase. A two-proportion Z-test confirmed that this difference is statistically significant ($Z = 2.2421$, $p = 0.0250$, $\alpha = 0.05$), justifying the move from an aggregate treatment effect to an individually estimated one.

Between the two T-Learner variants, the Unweighted T-Learner produced an average predicted individual treatment effect of 0.001556, close to the observed aggregate effect of 0.001070, whereas the Weighted T-Learner produced a substantially higher and more dispersed average effect of 0.065992, indicating a tendency to over-estimate uplift once class rebalancing is applied. The Unweighted T-Learner was therefore selected for subsequent evaluation. Its individual treatment-effect estimates were right-skewed (mean 0.001556, median 0.000065, standard deviation 0.019472), meaning that most users are predicted to have a small incremental response while a smaller subset is predicted to respond strongly, consistent with the heterogeneous nature of marketing response found in the literature.

Decile analysis confirmed that the model ranks users meaningfully: the top uplift decile showed by far the highest average predicted uplift (0.018933) and the highest observed uplift (0.007504) among all ten deciles, while the remaining deciles clustered close to zero, with small negative values appearing in a few middle and lower deciles as expected given the rarity of conversions. The cumulative Qini curve for the Unweighted T-Learner lay clearly above the random-targeting benchmark for the higher-ranked share of users, converging toward it as lower-ranked users were progressively included. Quantitatively, the area under the model's cumulative gain curve was 68.09 versus 45.47 for random targeting, an improvement ratio of about 1.50, and the Normalized Area Above Random (NAAR) was 0.332, confirming that the T-Learner captures a substantial share of the achievable incremental gain relative to random selection, without achieving perfect separation between deciles.

C. Explain: shap-based comparison of drivers

SHAP analysis identified f8, f2, f0 and f9 as the four most influential features for the conversion model (mean absolute SHAP values of 1.273, 0.952, 0.780 and 0.683, respectively), with f1, f5, f7 and f11 contributing least. For the uplift model, the ranking changed: f2 became the most influential feature (2.811), ahead of f0 (1.105), f8 (0.950) and f6 (0.905). Feature f2, which ranked second for conversion, became the single most important driver of uplift, while f8, the leading driver of conversion, dropped to third place for uplift — the largest negative rank shift observed (−2 positions). Feature f6 rose from fifth to fourth place, and f1, though still a comparatively minor feature overall, showed the largest positive rank shift (from twelfth to ninth), suggesting it carries information more relevant to treatment sensitivity than to baseline conversion propensity; f10 was the only feature to hold an identical rank (sixth) in both models.

Feature	SHAP (conversion)	Rank (conv.)	SHAP (uplift)	Rank (uplift)	Rank shift
f2	0.9520	2	2.8113	1	+1
f0	0.7801	3	1.1054	2	+1
f8	1.2732	1	0.9500	3	−2
f6	0.4625	5	0.9052	4	+1
f9	0.6827	4	0.6476	5	−1
f10	0.3345	6	0.3608	6	0
f1	0.0091	12	0.0989	9	+3

Table 3. Comparison of mean absolute SHAP importance and ranking between the conversion model and the uplift (T-Learner) model, selected features.

Despite these shifts, the overall similarity between the two rankings was high: the Spearman rank correlation between the conversion-SHAP ranking and the uplift-SHAP ranking was $\rho = 0.923$ ($p = 1.9\times10^{-5}$), indicating substantial agreement on which features matter overall, even though their relative importance, and in some cases their rank, differs between the two objectives. In practical terms, the model that best predicts who is likely to convert and the model that best predicts who is likely to convert because of the campaign draw on a similar but not identical information set, which supports treating conversion prediction and uplift estimation as complementary rather than interchangeable analytical tools.

D. Optimize: targeting-strategy comparison

Table 4 compares the incremental number of conversions achieved by Random, Predictive and Uplift Targeting at four targeting shares. Random Targeting was consistently and substantially weaker than the other two strategies at every share, confirming the practical value of model-based targeting over an unguided selection of users. Between Predictive and Uplift Targeting, no single strategy dominated across all budget levels. Predictive Targeting produced the larger incremental effect at 1% (30.35 vs. 17.31), 5% (55.76 vs. 50.68) and 20% (80.14 vs. 63.50) of targeted users, while Uplift Targeting produced the larger incremental effect at 10% (64.61 vs. 59.77).

Targeted users (%)	Random Targeting	Predictive Targeting	Uplift Targeting	Best strategy
1	0.56	30.35	17.31	Predictive
5	4.65	55.76	50.68	Predictive
10	8.41	59.77	64.61	Uplift
20	16.97	80.14	63.50	Predictive

Table 4. Expected incremental conversions by targeting strategy and targeted user share.

This crossover pattern was corroborated by bootstrap resampling: at the 10% targeting level, Uplift Targeting outperformed Predictive Targeting in 57.8% of resamples — the highest such probability among all four budget levels — while at 1%, 5% and 20% the average advantage favored Predictive Targeting, most clearly at 20%, where its average advantage was 16.55 incremental conversions. At every targeting level, the 95% confidence interval for the difference between the two strategies included zero. This means that the observed differences between Predictive and Uplift Targeting were not statistically significant. This does not remove the practical value of the pattern; it indicates that the choice between the two strategies should be informed by both the expected direction of the effect and by the surrounding statistical uncertainty.

The economic scenario and sensitivity analysis reinforced this budget-dependent pattern. A total of 48 scenarios were analyzed using different conversion values (€50, €100, and €200), targeting costs (€0.05, €0.10, €0.50, and €1.00), and targeting shares (1%, 5%, 10%, and 20%). The results showed that the optimal strategy depended on the targeting share, while changes in conversion value and targeting cost did not affect which strategy performed best. Predictive Targeting was the best strategy in 36 of the 48 scenarios (75%), while Uplift Targeting was better in the remaining 12 scenarios (25%). All 12 of these scenarios were at the 10% targeting level. This means that Uplift Targeting was the best strategy in all scenarios at the 10% targeting level, while Predictive Targeting performed better at the other three targeting levels. This stability across a wide range of economic assumptions indicates that the crossover between Predictive and Uplift Targeting is a robust feature of this dataset and campaign structure, rather than an artifact of a single choice of conversion value or targeting cost.

V. DISCUSSION

The results of this study show that a model that performs well in predicting conversions is not always the best choice for targeting users. Logistic Regression achieved the highest ROC-AUC, while XGBoost had the highest PR-AUC and performed best when targeting the top 1% of users. Conversion prediction and uplift modeling have different objectives. A predictive model identifies users who are likely to convert, while an uplift model focuses on users who are more likely to change their behavior because of the marketing intervention.

The T-Learner results showed that some users had a higher predicted and observed uplift than others. The Qini analysis also showed better results than random targeting, with a NAAR value of 0.332. This indicates that the uplift model was able to identify users who could generate additional conversions as a result of the campaign.

The comparison of targeting strategies showed that the best strategy depends on the percentage of users being targeted. Predictive Targeting performed better at the 1%, 5%, and 20% targeting levels, while Uplift Targeting performed better at 10%. The same pattern was observed across all 48 economic scenarios. This can be related to the distinction between “sure things” and “persuadables” discussed **[1]**. Predictive Targeting can select users who are already likely to convert, while Uplift Targeting gives more attention to users whose decision may be influenced by the campaign.

These results should be interpreted carefully. The overall treatment effect was statistically significant ($Z = 2.24$, $p = 0.025$), which shows that the campaign had an effect on conversion. At the same time, the 95% confidence intervals from the bootstrap analysis included zero at every targeting level. Therefore, although Uplift Targeting performed better at the 10% level, this difference was not statistically significant.

The SHAP results showed that similar features were important for both conversion and uplift. The Spearman correlation between the two feature rankings was high ($\rho = 0.92$), but there were still some differences. For example, f8 was the most important feature for conversion but ranked third for uplift, while f2 was the most important feature for uplift. This shows that the factors related to conversion and the factors related to the effect of a marketing intervention are similar, but their importance can differ.

The results suggest that there is no targeting strategy that is always the best. The choice between Predictive and Uplift Targeting should depend on the available marketing budget and the share of users that can be targeted. Since the features in the Criteo dataset are anonymized, it is not possible to explain their specific business meaning. Future research could apply the same approach to datasets with interpretable customer features and compare the T-Learner with other uplift modeling methods.

VI. CONCLUSION

This study examined whether the model that best predicts conversion is also the best choice for marketing targeting. For this purpose, the Predict–Estimate–Explain–Optimize (PEEO) framework was applied to the Criteo Uplift Prediction Dataset. The results showed that conversion prediction and uplift modeling are related, but they have different objectives.

The predictive models showed different strengths depending on the evaluation metric. XGBoost achieved the highest PR-AUC and the best top-1% capture rate, while the T-Learner based on XGBoost was able to identify users with higher predicted uplift. The Qini analysis also showed that the uplift model performed better than random targeting, with a NAAR value of 0.332.

One of the main findings is that the best targeting strategy depends on the targeting level. Predictive Targeting performed better at 1%, 5%, and 20%, while Uplift Targeting performed better at 10%. The same pattern was observed across all 48 economic scenarios. The

bootstrap analysis showed that the differences between the strategies were not statistically significant. Therefore, neither strategy can be considered the best choice in every situation.

The SHAP analysis showed a high correlation between the feature rankings for conversion and uplift ($\rho = 0.92$), but some differences in feature importance were also observed. This further shows that predicting conversion and estimating the effect of a marketing intervention are not the same task.

The main contribution of this study is the integration of prediction, causal effect estimation, model explanation, and targeting optimization within a unified decision-oriented PEEO framework. Rather than proposing a new machine learning algorithm, the study demonstrates that superior predictive performance does not necessarily translate into superior targeting performance under budget constraints. The findings therefore highlight the importance of evaluating machine learning models not only by their predictive accuracy, but also by their ability to support effective marketing resource allocation.

REFERENCES

[1] N. J. Radcliffe, "Using control groups to target on predicted lift: Building and assessing uplift model," *Direct Mark. Anal. J.*, no. 3, pp. 14–21, 2007.
[2] Xie, Y. Nie, L. Liu, and Y. Chen, "Machine learning-based research of AI marketing: Topic analysis and model construction," *Future Bus. J.*, vol. 11, Art. no. 272, 2025.
[3] A. Kaponis, M. Maragoudakis, and K. C. Sofianos, "Enhancing user experiences in digital marketing through machine learning: Cases, trends, and challenges," *Computers*, vol. 14, no. 6, Art. no. 211, 2025.
[4] S. Alotaibi and B. Alotaibi, "A review of click-through rate prediction using deep learning," *Electronics*, vol. 14, no. 18, Art. no. 3734, 2025.
[5] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 10, pp. 4156–4165, 2019.
[6] P. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling with single and multiple treatments," *Knowl. Inf. Syst.*, vol. 32, pp. 303–327, 2012.
[7] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. Am. Stat. Assoc.*, vol. 113, no. 523, pp. 1228–1242, 2018.
[8] S. Athey and G. W. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 27, pp. 7353–7360, 2016.
[9] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modelling: A review of the literature," *Proc. Mach. Learn. Res.*, vol. 67, pp. 1–13, 2017.
[10] Devriendt, D. Moldovan, and W. Verbeke, "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*, vol. 6, no. 1, pp. 13–41, 2018.
[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1135–1144.
[13] Criteo AI Lab, *Criteo Uplift Prediction Dataset (v2.1)*, Criteo Research, 2018. [Data set].
[14] Diemert, A. Betlei, C. Renaudin, and M.-R. Amini, "A large scale benchmark for uplift modeling," in *Proc. AdKDD TargetAd Workshop, KDD*, 2018.
[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 785–794.

AUTHORS

First Author – Stanela Gacova, Master's Student, Faculty of Computer Science, Goce Delcev University, Stip, North Macedonia, stanela.210231@student.ugd.edu.mk
Second Author – Cveta Martinovska Bande, PhD, Full Professor, Faculty of Computer Science, Goce Delcev University, Stip, North Macedonia, cveta.martinovska@ugd.edu.mk

Correspondence Author – Stanela Gacova, stanela.210231@student.ugd.edu.mk, (+389) 78427900