IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# A Dynamic Framework for Causal User Profiling and Treatment Segmentation via Uplift Modeling in Internet Lending

**JIANQING JIANG[1], PROF. MADYA DR. NOR ASILAH WATI ABDUL HAMID[12], DR. NG KENG YAP[12], PROF. MADYA DR. CHOO WEI CHONG[3*]**

[1]Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia (UPM), Serdang, Malaysia (e-mail: jiangjianqingupm@gmail.com)
[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang, Malaysia
[3]School of Business and Economics (SBE), Universiti Putra Malaysia (UPM), Serdang, Malaysia (e-mail: wcchoo@upm.edu.my)

Corresponding author: Prof. Madya Dr. Choo Wei Chong (e-mail: wcchoo@upm.edu.my).

This work received no specific grant from any funding agency.

**ABSTRACT** The growth of internet lending has created a need for decision frameworks based on models that are both personalized and causally interpretable. Conventional uplift models detect treatment responsiveness without recognizing user heterogeneity, the temporal consistency of user behavior, or the upstream design choices that carry important causal implications. This paper proposes an integrated and reproducible Causal User Profiling (CUP) framework that combines causal inference, uplift modeling, and response-based segmentation within a single pipeline. CUP realizes treatment-effect heterogeneity through a four-type response taxonomy (Persuadable, Sure Thing, Lost Cause, Do-Not-Disturb) and embeds it in a multi-stage pipeline involving hybrid feature selection (Information Value (IV), Causal Forest importance, Population Stability Index (PSI) stability, and Stepwise refinement), stratified clustering with a "C2 replacement strategy," and meta-learning via both the X-Learner and the Doubly Robust (DR) Learner using Logistic Regression (LR). A component-wise ablation analysis finds that feature selection increases AUUC by 25–30%, C2 clustering by 10–12%, and the DR-Learner + LR by another 5–8%. Overall, the integrated CUP framework yields 45–50% higher AUUC than the baseline ("all features + no clustering + standard learner") while retaining behaviorally coherent and temporally stable insights. Methodologically, we provide: (i) an end-to-end causal user profiling framework that interoperates profiling, causal estimation, clustering, and uplift evaluation; (ii) a behaviorally and causally consistent response segmentation mechanism grounded in the potential-outcomes model; and (iii) a reproducible experimental design that quantifies pipeline-level uplift gains through systematic ablation. Applied to large-scale internet-lending data, CUP reveals opportunities for treatment-aware personalization, enabling financial institutions to target Persuadables, support Sure Things, and avoid disturbing Do-Not-Disturbs based on causal evidence.

**INDEX TERMS** C2 Clustering Strategy; Causal Precision; Causal User Profiling; Decision Support Systems; DR-Learner; Feature Selection; Heterogeneous Treatment Effects; Internet Lending; Meta-Learners; Response Segmentation; Uplift Modeling; X-Learner.

## I. INTRODUCTION

THE rapid evolution of digital platforms has heightened the importance of personalization and targeting, shifting attention toward core aspects of data-driven decision-making. In this tussle, user profiling [1]–[3]—the process of acquiring, analyzing, and organizing multi-dimensional user data to create static and/or dynamic user profiles or models of the user's behaviors, tastes, preferences, and other demographics—is foundational to data-driven systems. Profiling helps design interpretable user representations that drive downstream applications such as recommender systems, targeted marketing, and risk assessment [4], [5] but conventional profiling pipelines are observational in nature, concerned with who the user is instead of how the user would react if acted upon [6], and almost entirely ignore causation [7], relegating them to use-case-specific black-box components [8].

During the same period, a shift in the way the field thought about the phrase "learning from data" influenced how data—used for the targeted deployment of algorithms—came to be conceptualized. A widely used formulation characterizes learning from data as the process in which "a program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." While predictive power in analytics is impressive, machine learning (ML) algorithms are limited by the fact that making decisions can change the very distribution of the outcomes one wishes to predict [9]. In many situations, prediction is insufficient, and one needs to understand the causal structure of the world because interventions change the distribution of data [10]. It is this limitation that led to additional research on causal inference, defined as "the study of the relationship between cause and effect" [9] , which also extensively informs decision-making since ML necessarily restricts its power by learning only patterns instead of, for instance, generating causal relations for its predictions. Causal inference has two paradigms which can inform solutions to these problems. The first is the potential outcomes framework, where "the causal effect of a treatment on a unit is the difference between the outcome when the unit receives the treatment and the outcome when the unit does not" [11]. The second is "the process of using data together with causal assumptions to answer questions about causal relations—such as predicting the effect of interventions or explaining observed dependencies" [9]. The first provides a coherent underpinning for counterfactual reasoning [12], and the second provides models for reasoning qualitatively about the data-generating process and is critical for the transportability of causal knowledge [9]. The synthesis of the two is clear: "by leveraging causal inference, you go beyond description and association," being able to ask what alternative actions would do under differing situations [10]. These advances form the basis of heterogeneous treatment effect (HTE) estimation, which directly focuses on individual-level responsiveness to interventions [6], [13], [14]. In this line of work, Causal Trees reveal treatment heterogeneity using recursive partitioning [6], and Causal Forests extend this methodology, resulting in consistent Conditional Average Treatment Effects (CATEs) [13]. Meta-learners (e.g., S-, T-, X-, and R-learners) [15] reformulate the causal estimation task into modular supervised-learning settings that allow for flexibility and scalability across data environments [15]. Beyond estimation, policy learning integrates causal inference with decision-making contexts, creating decision rules from estimates of treatment effects [16].

In practical internet lending systems, interventions such as interest coupons, fee reductions, credit line adjustments, and targeted reminders are routinely deployed to influence user borrowing behavior. When such actions are guided solely by predictive models or raw treatment-effect estimates, platforms may repeatedly allocate incentives to users who would borrow regardless of intervention, while failing to activate users who are truly responsive to targeted actions. Moreover, these interventions are often applied repeatedly under budget and risk constraints, making it difficult to translate heterogeneous treatment effect estimates into stable, interpretable, and decision-aligned user representations.

Concurrently with the emerging HTE literature, uplift modeling first arose in applied domains such as marketing, healthcare, and finance to directly estimate incremental impact: the difference between the probability of response from a group exposed to treatment and that of a comparable group not exposed [17]. In practical terms, this quantity answers how much more likely a user is to respond as a direct result of an intervention, rather than due to baseline propensity. Uplift models [18], [19]that focus on modeling the "treatment effect induced," rather than overall predictive accuracy, can employ tree-based approaches. The "four-type" consumer classes— Persuadables, Sure Things, Lost Causes, and Do-Not-Disturbs—serve as the conceptual framework for studying individual causal response [20]–[22]. This taxonomy is widely used to align incremental-effect estimation with operational targeting, because it distinguishes true incremental responders (Persuadables) from always-responders (Sure Things), never-responders (Lost Causes), and users for whom treatment may be harmful (Do-Not-Disturbs). [20]–[22] Uplift modeling evaluates targeting strategies against metrics such as Area Under the Uplift Curve (AUUC) and Qini coefficient, which capture the incremental gain produced [17], [19]. Recently conducted reviews warn that uplift performance is very sensitive to upstream design choices—feature selection, clustering, and labeling—and that the value of an integrated and transparent pipeline is greater than isolated model and algorithm comparisons [8], [22].

Yet, "traditional" user profiling continues to follow the predictable, descriptive steps of data collection, normalization and cleansing, feature extraction, clustering, and performance evaluation. The primary goal remains predictive segmentation and operational classification [4], [23], [24]. Current profiling systems, effective as they are for prediction, are not designed to estimate how users would respond to interventions, nor do they derive properties from causal heterogeneity [7]. From the perspective of profiling, causal reasoning has not yet been embedded into an end-to-end analysis pipeline, and we ask three broad methodology questions: (i) how to design new AAUC-driven response segmentation, where labeling both informs and determines evaluation; (ii) how to integrate feature selection, stratified clustering, bias adjustment, and treatment-effect estimation into a new unified causal user profiling workflow; and (iii) how to measure the marginal contribution of each pipeline component through component-wise ablation analysis [22].

This work intends to close these gaps by introducing a unified method called Causal User Profiling (CUP), integrating user profiling, causal inference, and uplift modeling into a single analysis process, building on the previous descriptive roadmap but embedding causal estimation and uplift-based evaluation into its core. Conceptually, it allows causal rea-

soning to take form inside profiling methods, stating not only who users are, but how they respond to actions performed on them [25].

We summarize three contributions:

(1) Causal User Profiling Framework. We propose an integrated methodological framework that connects feature selection, clustering, confounding adjustment, and causal effect estimation into a coherent causal user profiling pipeline.

(2) AAUC-Driven Post-Evaluation Response Segmentation. We develop a performance-based segmentation mechanism that classifies users into the four causal response types (Persuadables, Sure Things, Lost Causes, and Do-Not-Disturbs) based on AAUC results, bridging model evaluation with actionable user interpretation. For example, separating Persuadables from Sure Things clarifies whether a high-response segment reflects true incremental impact or merely high baseline propensity.

(3) Component-Wise Ablation and Performance Analysis. We quantify the marginal contribution of each pipeline stage—feature selection, clustering, bias adjustment, and causal estimation—to overall uplift performance, providing reproducible methodological insights for practitioners.

Validating on an internet-lending dataset, results illustrate how embedding causal reasoning within user profiling provides a means to deliver additional value to customers and businesses alike, ultimately leading to better personalization, more precise targeting, and more effective data-driven decision-making under real-world constraints [22], [23].

#### a: Organization of the Paper

The remainder of this paper is organized as follows. Section II reviews related work on heterogeneous treatment effect estimation, uplift modeling, causal inference in recommender systems, and user profiling, and identifies the methodological gaps addressed in this study. Section III introduces the Causal User Profiling (CUP) framework, detailing its core modules including feature selection, causal estimation, clustering, and response-type segmentation. Section IV describes the data source, preprocessing procedures, and experimental design. Section V presents and discusses the empirical results, focusing on model performance, stability, and interpretability under repeated interventions. Section VI outlines the limitations of the proposed framework. Finally, Section VII concludes the paper and discusses directions for future research.

## II. RELATED WORK

### A. UPLIFT MODELING AND EVALUATION

Uplift modeling—also referred to as incremental response modeling—reconceptualizes prediction as the estimation of differential treatment response, emphasizing the causal effect of an intervention rather than its absolute outcome level. In contrast to conventional predictive modeling, which estimates the likelihood of an outcome, uplift modeling explicitly focuses on the change in outcome probability attributable to an action. As Radcliffe and Surry [17] state, uplift defines the notion of "the difference in response rates attributable to a treatment" that "shifts analytics from descriptive prediction into the prescriptive space." Early approaches adopted a two-model strategy, in which separate predictive models are trained for treated and untreated groups, and the difference is interpreted as the incremental effect [20]. Although simple to conceptualize, two-model approaches can be unstable and lead to biased estimates when treatment allocation is imbalanced or when covariate distributions differ substantially across groups.

A significant methodological advance arrived with tree-based uplift models, which introduced recursive partitioning to seek maximum treatment–control heterogeneity within subgroups [18]. These Uplift Decision Trees offered interpretable segmentation rules and provided groundwork for subsequent ensemble extensions. Uplift random forests and causal forests improved robustness and consistency through aggregation, although at some cost to interpretability [6], [13]. This line of work reflects a broader methodological transition toward explicitly modeling heterogeneous treatment effects (HTE) to inform intervention decisions. As summarized by Devriendt, Moldovan, and Verbeke [22], this evolution represents a broader shift from purely predictive response models to prescriptive analytics that conceptually situates uplift modeling within modern causal inference.

Parallel advances occurred in meta-learning approaches that reinterpret uplift estimation as superimposed supervised learning tasks. Frameworks like the S-, T-, X-, R-, and DR-learners unify heterogeneous treatment effect (HTE) estimation and uplift prediction under flexible templates [14], [15]. These approaches decouple the estimation of nuisance components, such as outcome and treatment assignment models, from the final treatment-effect estimator, enabling flexible combinations with different base learners. In practice, meta-learners differ mainly in how they reuse outcome models and propensity information under imbalance and limited overlap, and their stability is therefore strongly shaped by base-learner choice and nuisance-model specification. [14], [15], [19] These methods yield additional generalization across settings but remain sensitive to base-learner selection, sample size, and hyperparameter tuning. Until now, tabular models and representation-learning–based causal networks, such as TARNet, CFRNet, DragonNet, and GANITE, have adopted deep architectures to mitigate covariate imbalance, reduce the burden of counterfactuals, or model nonlinear treatment effects on the response [26]–[28]. While such models improve expressive capacity, prior studies note trade-offs in interpretability, stability, and reproducibility, particularly in profiling-oriented applications.

As models improved, so did evaluation. Uplift modeling concerns incremental gain; thus standard accuracy metrics are not useful. Uplift-specific ranking measures such as the "Qini coefficient(s)" and the "Area Under the Uplift Curve (AUUC)" are now standard [17], [19]. Both metrics evaluate how effectively a model ranks individuals by incremental response rather than by absolute outcome likelihood. The Qini

coefficient sums cumulative differences across incremental uplift-ranked segments between treated and control groups. AUUC measures the total incremental effect via the area between the uplift curve and a diagonal baseline. The latter can be sensitive to treatment imbalance or sparse samples, and recent advances have introduced multi-treatment AUUC and cross-treatment gain surfaces to cover multi-arm and dose-response settings [29], [30]. Intuitively, gain-surface style evaluations summarize how incremental ranking performance varies across intervention arms, making cross-arm trade-offs and sensitivity to treatment choice explicit. [29], [30] Reviews emphasize that the ability to generate strong uplift curves is attributable largely to upstream pipeline design choices, particularly feature selection, user segmentation, and response-type/target labeling [8], [22]. Other comparative studies consider uplift algorithms to be "black boxes" and provide limited detail on implementation or sensitivity analyses, which undermines reproducibility and reduces practical interpretability [17], [18]. In this study, reproducibility refers to reporting and structuring the full pipeline—feature construction/selection, clustering settings, propensity modeling, learner configuration, and labeling rules—so that an independent team can rerun the workflow and obtain consistent uplift curves and response-type assignments under the same data and protocol. These concerns motivate a methodological shift from isolated model comparison toward workflow-level optimization, focusing on transparent design choices, pipeline configuration, and component-wise diagnostics—principles that underpin the Causal User Profiling (CUP) framework introduced in this study.

### B. HETEROGENEOUS TREATMENT EFFECTS (HTE) AND CATE ESTIMATION

Although uplift modeling is used mostly in marketing and intervention targeting, the concept of Heterogeneous Treatment Effects (HTE) provides the theoretical foundation for uplift modeling. HTE methods aim to estimate the Conditional Average Treatment Effect (CATE) for each individual or subgroup in the population—that is, the expected causal effect conditional on observed features [6], [13]. From this perspective, uplift modeling can be viewed as an operationalization of HTE estimation that emphasizes ranking and targeting decisions rather than pointwise effect estimation alone.

Tree-based methods such as Causal Trees start with the full covariate space and recursively partition it to identify regions with distinct treatment effects. Causal Forests, for several well-founded reasons, instead employ ensemble aggregation, yielding more consistent estimators and supporting valid statistical inference across regions [6], [13]. Generalized Random Forests (GRF) extend this local/posterior forest framework, unifying a large class of forest-based estimators into a general nonparametric framework [16]. Beyond tree ensembles, Bayesian and nonparametric approaches introduce uncertainty quantification through credible intervals and yield more robust estimates in small-sample or high-variance settings [31], [32] via methods such as Bayesian Additive Regression Trees (BART), Bayesian Causal Forests (BCF), and Gaussian Process models. These approaches are often preferred in settings where variance control and uncertainty assessment are critical to downstream decision-making.

While lighter-weight estimators relax assumptions required for CATE estimation, meta-learning approaches—S-, T-, X-, R-, and DR-learners—decompose the CATE estimation task into modular supervised learning problems and offer flexibility in base-learner choice and treatment-variable specification [14], [15]. The combination of meta-learners with different base learners is motivated by the need to balance bias, variance, and robustness under heterogeneous data-generating conditions, rather than by any universally optimal configuration. Empirical analyses show no universally dominant learner, highlighting that pipeline-level optimization is preferred over naive model substitution [22]. Representation-learning–based causal models such as TARNet, CFRNet, DragonNet, and GANITE build deep latent representations to reduce covariate imbalance and improve counterfactual estimation [26]–[28]. While these models increase expressive capacity, prior studies note trade-offs in transparency, stability, and reproducibility, particularly when interpretability is required for profiling-oriented analysis [7].

Evaluation frameworks in the HTE literature closely parallel those used in uplift modeling. Metrics such as Incremental AUUC and Qini measure incremental ranking performance, while Precision in Estimation of Heterogeneous Effects (PEHE) and Mean Squared Error of Individual Treatment Effects (MSE(ITE)) are common in semi-synthetic benchmarks [15]. For joint learning of treatment policies, policy value and doubly robust off-policy evaluation (OPE) assess expected reward of policies derived from estimated treatment effects [33]. When multiple interventions are available, multi-treatment AUUC and consistency-based metrics further characterize the stability of ranking and policy decisions across treatment arms [29], [30].

Recent reviews provide empirical insight into the adoption and implementation of HTE methods across domains. A forthcoming 2024 methodological review of Causal Forest applications analyzes 133 peer-reviewed studies across areas from health to marketing, documenting widespread reliance on the `grf` package but limited reporting of identification assumptions and tuning parameters [34]. A PRISMA-guided scoping review of HTE estimation in randomized controlled trials (RCTs) using machine learning reports predominance of forest-based (60%) and Bayesian (53%) models in domains such as health and education, while again noting incomplete documentation of generalizability checks and identification strategies [35]. In RCT-emulation pipelines that benchmark recent methodologies—including Ding et al.'s RIF—reviews report frequent failures in confounding adjustment or validation, reinforcing the importance of reproducibility, variance control, and coherent pipeline design [36].

**TABLE 1.** Representative Methods and Evaluation Frameworks in HTE and Uplift Modeling

| Category / Metric | Representative Methods / Metrics | Key References | Purpose / Contribution |
|---|---|---|---|
| **Tree-based Models** | Causal Tree, Causal Forest, Generalized Random Forest (GRF) | [6], [13], [16] | Nonparametric estimation of Conditional Average Treatment Effects (CATE) |
| **Meta-learners** | T-, S-, X-, R-, DR-learners | [14], [15] | Modular frameworks decomposing CATE into supervised subproblems |
| **Representation Learning** | TARNet, CFRNet, DragonNet, GANITE, CEVAE | [26]–[28] | Deep causal models capturing nonlinear treatment–response patterns |
| **Bayesian / Nonparametric** | BART, Bayesian Causal Forest, Gaussian Process Models | [31], [32] | Probabilistic inference and uncertainty quantification |
| **Uplift-Specific Metrics** | AUUC, Qini coefficient | [17], [22] | Measurement of incremental gain and ranking quality |
| **Causal Evaluation Metrics** | PEHE, MSE(ITE), Policy Value, Doubly Robust OPE, Multi-Treatment AUUC | [15], [30], [33] | Evaluation of counterfactual accuracy, policy performance, and robustness |

Note: These evaluation measures complement modeling advances by quantifying incremental gain (AUUC/Qini) and counterfactual accuracy (PEHE/MSE), forming a foundation for workflow-level optimization in causal user profiling.

### C. CAUSAL INFERENCE IN RECOMMENDER SYSTEMS AND THE USER-PROFILING PIPELINE

In recommender systems, causal inference has been applied to address exposure bias, selection bias, and to conduct policy evaluation [23]. Rather than focusing solely on predictive accuracy, this stream of work explicitly treats recommendation actions as interventions and evaluates their causal effects on user behavior. Recent reviews identify core causal objectives along three interrelated dimensions:

(1) causal objectives, such as de-biasing item exposure and estimating the treatment effects of recommendation actions;

(2) identification strategies, including inverse propensity scoring (IPS), doubly robust estimation (DR), and instrumental-variable (IV) approaches;

(3) evaluation paradigms, covering offline policy learning, counterfactual simulation, and online contextual bandits.

Together, these components form an integrated end-to-end causal pipeline spanning data collection, identification, policy optimization, and evaluation. This pipeline perspective emphasizes decision-oriented evaluation under explicit interventions rather than static prediction.

Conversely, traditional user-profiling research remains largely predictive. As summarized by Wu et al. [8], profiling pipelines typically consist of five sequential components—data collection, data preprocessing, feature extraction, modeling, and evaluation—each originally designed for descriptive segmentation or predictive accuracy rather than interpretability. Similarly, Maraj et al. [37] argue that most profiling systems emphasize data enrichment, privacy protection, and governance considerations rather than response-driven causal mechanisms. Purificato et al. [7] observe that many existing profiling approaches "focus on correlations rather than causal mechanisms," limiting their ability to support responsive interventions or predict user-level treatment responsiveness.

Thus, although causal recommender modeling has seen significant advances, operational techniques that integrate causal inference into user-profiling pipelines remain comparatively scarce. In particular, conventional profiling workflows typically lack explicit components for treatment-effect identification, response-type labeling, and uplift-based evaluation. A conceptual gap therefore persists between profiling workflows and causal-inference pipelines. Addressing this gap is essential for moving beyond static or purely predictive user profiles toward representations that capture how users are likely to respond under alternative interventions.

### D. GAPS OUR WORK ADDRESSES

Across the surveyed literatures, two practice-oriented gaps arise repeatedly.

First, applications of HTE estimation and Causal Forest modeling often under-report critical design and tuning decisions, making it difficult to reproduce results uniformly or determine which components of the pipeline actually contribute to uplift or CATE performance. Rehill [34], Inoue et al. [35], and Ling et al. [36] show that studies frequently rely on heavy default hyperparameters, rarely justify identification assumptions, and often omit reporting clustering or feature-selection strategies. As a result, it remains unclear whether observed performance differences stem from causal estimators themselves or from upstream design choices such as feature selection, clustering, or labeling. This leaves unclear whether feature selection, clustering, labeling, or other design choices are responsible for observed uplift or CATE performance differences [22].

Second, mainstream user-profiling frameworks are comprehensive in data preparation and feature engineering, but remain largely descriptive and correlation-based [7], [8]. They generally lack components for causal estimation, uplift-based evaluation, or assignment of causal response types. Consequently, existing profiling systems are not designed to represent how users are expected to respond under alternative interventions, limiting their suitability for decision-oriented personalization. This disconnect between predictive profiling and causal reasoning inhibits existing systems from explaining how users will respond to user-level interventions. Yet such capabilities are core to adaptive personalization, broad targeting optimization, and prescriptive analytics.

To address these gaps, we offer two methodological contributions.

First, we introduce a module called Four-Type Response Segmentation, aligned with uplift theory, which uses high-

confidence uplift thresholds, model-assisted inference, and post-hoc label refinement to operationalize the dependence between AAUC-based evaluation metrics and response-type labeling [18], [20]. This design explicitly links model evaluation outcomes to interpretable response categories, addressing the ambiguity between ranking performance and user-level interpretation noted in prior studies.

Second, we propose an new Improved Causal User Profiling (CUP) Roadmap, which embeds feature selection, clustering, confounding adjustment, causal estimation, and response-type labeling into a unified, reproducible, and resource-aware workflow. By structuring these components as an integrated pipeline rather than isolated modeling steps, CUP directly responds to reproducibility and transparency concerns highlighted in the HTE and uplift literature. By combining causal inference, uplift modeling, and user profiling into a single analytic pipeline, CUP addresses the reproducibility issues and methodological silos highlighted in earlier work, providing a principled foundation for causal user profiling—a next-generation framework for data-driven personalization, targeting, and intervention design.

## III. RESEARCH FRAMEWORK

This paper proposes an integrated methodological framework, Causal User Profiling (CUP), that connects three previously disparate domains—user profiling, causal inference, and uplift modeling—into a common analytic pipeline for personalized treatment analysis. This comes from our observation that user-profiling studies typically investigate who the user is, identifying demographic and behavioral segments [4], [8], while "for practical intervention it is important to first understand how a user would react if we act" [6]. "Current user-modeling methods … focus on correlations rather than causal mechanisms," as noted by Purificato et al. [7], which limits their interpretability for strategic targeted interventions. The CUP framework addresses this issue by embedding causal estimation and response-based labeling into the profiling pipeline, turning the de-facto descriptive workflow into a causally interpretable and response-aware analytic system [22].

The CUP pipeline (Figure 1) builds on the conventional user-modeling workflow. It starts with data collection: behavioral and contextual data are captured from production platforms, followed by data preprocessing to ensure quality, consistency, and readiness for analysis [8]. Next comes feature extraction, converting behavioral logs and demographic information into representations suitable for modeling.

Although not shown in Figure 1 as a red-arrow component of its own, feature selection is an essential overhead to this step, ensuring that only variables with both predictive and causal relevance move down the causal analysis pipeline. Eke et al. [4] emphasize that profiling based on user behaviors requires appropriate choice of representative variables, while Wager and Athey [13] warn that it is easy to accept predictive but non-causal features with varying predictive noise (which tends to dilute estimated treatment effects). CUP therefore uses a two-pronged approach: information value for predictive importance, and causal importance from a Causal Forests approach [13] to stabilize the interventions.

The first truly new component of CUP is clustering for stratification, grouping users into behaviorally similar and causally comparable groups. As Wu et al. [8] remark, "most prior clustering methods for user profiling are mainly descriptive and static; causal analysis benefits from grouping such that the treatment and control users in a group are balanced in terms of their distributions." This aligns with findings from Devriendt et al. [22], who show that uplift models are sensitive to sample imbalance and perform better in reliable uplift segments. Thus, CUP's clustering step trades off descriptive interpretability for causal validity by pairing comparability with segmentation.

Causal effects are estimated—after treatment-control selection—in the next submodule, confounding and bias adjustment. This step helps ensure that treatment and control groups are comparable to permit causal attribution in non-experimental settings. Bias-adjustment methods include inverse probability weighting (IPW) and stratified reweighting (see [6] and [14]). Through these techniques, CUP mitigates selection bias and strengthens internal validity. To ensure comparability between treated and control groups in the empirical analysis, treatment assignment probabilities are estimated using observed covariates and incorporated through inverse propensity weighting during uplift evaluation. This adjustment mitigates treatment imbalance arising from non-random intervention assignment and reduces bias when comparing incremental outcomes between treatment arms.

Next, within the potential-outcomes framework [11], is the causal-effect estimation module. Causal Trees [6] track treatment heterogeneity using recursive partitioning, and this logic is extended in Causal Forests [14]. CATEs (conditional average treatment effects) are the resulting estimands from such hierarchical estimators. Meta-learning algorithms such as T-, S-, X-, and R-learners [15] restate causal estimation as modular (still supervised-learning) tasks that may be adjusted for cross-environment flexibility. These estimators output each user's uplift value (the effect an intervention has on user behavior).

Following causal estimation, the performance-evaluation submodule measures the captured impact. The Area Under the Uplift Curve (AUUC) is, according to Devriendt et al. [22] and Gutierrez and Gérardy [19], the preferred metric for uplift performance—measuring not predictive accuracy, but impact captured (see [17]). To interpret user-response behavior and to disentangle module contributions, CUP performs component-wise ablation analysis, isolating the per-stage effect of the four modules: feature selection, clustering, bias adjustment, and causal estimation. This follows the principle that "uplift performance depends strongly on upstream design choices" [22]. The purpose of the ablation analysis in this study is not to conduct formal hypothesis testing, but to assess the relative contribution and stability of individual pipeline components under repeated real-world deploy-

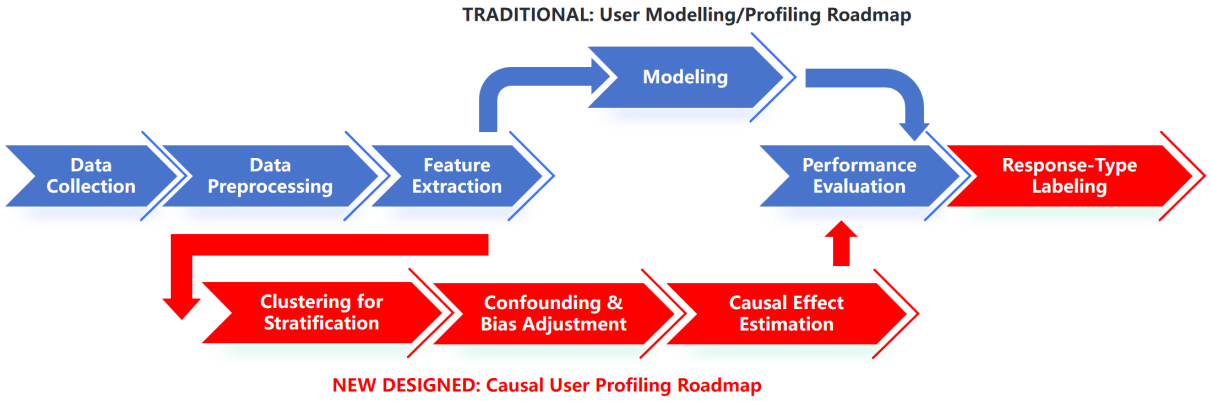



**FIGURE 1.** Traditional versus newly designed Causal User Profiling (CUP) roadmap. Blue arrows represent the traditional user profiling process—Data Collection, Data Preprocessing, Feature Extraction, Modeling, and Performance Evaluation—while red arrows represent CUP extensions: Clustering for Stratification, Confounding and Bias Adjustment, Causal Effect Estimation, and Response-Type Labeling.

ments. Each ablation experiment removes one component from the CUP workflow while keeping all others fixed, and performance differences are evaluated consistently across six consecutive monthly datasets. By examining whether performance changes persist across time periods rather than relying on a single snapshot, the analysis provides empirical evidence on whether observed gains are systematic rather than incidental. Given the operational nature of the study and the use of large-scale observational data, we focus on temporal consistency and magnitude of performance differences rather than formal statistical significance testing.

Response-type labeling occurs after performance evaluation. Using AUUC-based results, users are tagged to the four classic causal-response categories—Persuadables, Sure Things, Lost Causes, and Do-Not-Disturbs—as defined by Radcliffe and Surry [17] and extended by Jaskowski and Jaroszewicz [21]. The result of this post-evaluation process is interpreting user-model outputs into actionable causal profiles, completing the causal-profiling loop.

CUP brings descriptive profiling and causal inference closer together in a single sequence. Preservation of interpretability from user profiling [4], [8], combined with the structure of causal-effect estimation formalism [6], [15], is central to the method, as is adopting the evaluative rigor of uplift modeling [17], [22]. With utility demonstrated on an Internet-lending dataset, the CUP framework shows how causal "laws" can be operationalized within profiling to support personalized and data-driven decision-making under real-world constraints.

### A. FOUR-TYPE RESPONSE SEGMENTATION MODULE

#### 1) Concept and Taxonomy

Based on the potential outcomes framework [11], the Four-Type Response module assigns users to four canonical causal-response types—Persuadable (A), Sure Thing (B), Lost Cause (C), and Do-Not-Disturb (D). These types describe how an individual's outcome would change were the intervention present or absent, capturing causal responsiveness rather than behavioral similarity [17], [20].

In contrast to typical segmentation methods that cluster users based on demographic or behavioral factors, uplift-based profiling emphasizes incremental impact estimation—the assessment of how the probability that a user takes the desired action changes when the treatment is applied. Framed in this way, user profiling becomes not merely descriptive, but a treatment-aware decision process, informing the design and evaluation of personalized interventions in marketing, healthcare, and credit-analytics domains [18], [19], [22].

#### 2) Behavioral Evidence Partition

To provide empirically grounded baselines for the theoretical response types above, users are partitioned by treatment assignment $T$ and outcome $Y$, resulting in four cells of evidence based on observed behavioral congruency expected across the categories(Table 2). This integrates causal estimation and response-based labeling into the profiling pipeline, converting the conventional descriptive workflow into an analytically causally interpretable and response-aware system [22].

**TABLE 2.** Behavioral evidence partition linking theoretical response categories with observed treatment–outcome patterns.

| Cell | Condition | Plausible Response Types |
|---|---|---|
| Set 1 | $T=1, Y=1$ | A / B |
| Set 2 | $T=1, Y=0$ | C / D |
| Set 3 | $T=0, Y=1$ | B / D |
| Set 4 | $T=0, Y=0$ | A / C |

This partition restricts theoretical labeling to behavioral realizability: a user assigned label "A" but found in $T = 1, Y = 0$(Set 2) would exhibit an anchored violation of behavioral consistency and would therefore be de-provisioned

of the "A" label (referencing "refinement" from the previous subsection). Such anchoring makes the Four-Type taxonomy both causally interpretable and empirically grounded.

### 3) Evolving the Segmentation Designs for Four-Type Response (V1-3)

Designs for Four-Type Response (V1–V3) The Four-Type Response Segmentation module evolved through three generations of increasing interpretability and robustness as we addressed limitations in earlier designs. The key methodological distinctions are summarized concisely in Table 3.

#### V1

The initial version (V1) adopts a simple triage approach that partitions users into High, Medium, or Low responsiveness groups solely based on the uplift score:

$$\begin{aligned} \hat{p}_1 &= P(Y=1 \mid T=1, X), \\ \hat{p}_0 &= P(Y=1 \mid T=0, X), \\ u &= \hat{p}_1 - \hat{p}_0. \\ \text{label}_i &= \begin{cases} \text{High}, & u_i \geq \delta, \\ \text{Low}, & u_i \leq -\delta, \\ \text{Medium}, & \text{otherwise}. \end{cases} \end{aligned} \tag{1}$$

This design provides a lightweight way to register individual sensitivity to intervention and offers a computationally inexpensive method for identifying users likely to be affected. However, it cannot tell us why some people are responsive and others are not (nor can any known design), so it remains a descriptive rather than a causative segmentation proxy, not a fully causal segmentation method.

**Algorithm 1** Version 1 – Uplift-only Triage (High / Medium / Low)

**Input:** Dataset $D$ with column `uplift_score` $u$ (or $u = \hat{p}_1 - \hat{p}_0$ if available)
**Output:** Triage label $\in$ {High, Medium, Low} for each sample
**Protocol:** Split $D$ into outer train/valid/test; select $\delta$ on valid to maximize AUC; freeze $\delta$ for test (no leakage)

**for** each $i \in D$ **do**
  **if** $u_i > \delta$ **then**
    $label \leftarrow$ High
  **else if** $|u_i| < \delta$ **then**
    $label \leftarrow$ Medium
  **else**
    $label \leftarrow$ Low
  **end if**
**end for**

**Optional mapping:**

- if $u_i \geq \delta \rightarrow$ candidate for A (Persuadable)
- if $u_i \leq -\delta \rightarrow$ candidate for D (Do Not Disturb)
- if $|u_i| < \delta \rightarrow$ candidate region for B/C (resolved by V2/V3)

#### V2

The second version (V2) extends the framework by introducing a counterfactual mapping between treatment and control outcome probabilities, constructing a two-dimensional prediction space:

$$\begin{aligned} \hat{p}_1 &= P(Y=1 \mid T=1, X), \\ \hat{p}_0 &= P(Y=1 \mid T=0, X), \\ u &= \hat{p}_1 - \hat{p}_0. \end{aligned} \tag{2}$$

Based on calibrated thresholds $(y_1^{th}, y_0^{th}, \delta)$ users are assigned to the four canonical response types as follows:

This design corresponds closely to the uplift-modeling literature [6], [18] and enables interpretability as it visualizes causal responsiveness across probability quadrants. In reality, however, only a fraction of users are cleanly segmented into these quadrants; many lie close to decision boundaries, resulting in ambiguous labels or oscillating between them. V2 therefore has better interpretability than V1, but at the cost of robustness, as demonstrated during evaluation [22].

**Algorithm 2** Version 2 – Counterfactual Four-Quadrant Segmentation

**Input:** Dataset $D$ with columns $y1\_prob(p_1), y0\_prob(p_0)$ (arm-wise calibrated); uplift $u = p_1 - p_0$
**Output:** Response type $\in \{A, B, C, D\}$ for each sample
**Protocol:** Outer train/validation/test split; grid-search $(h, \ell, \delta)$ on validation to maximize AUUC; fix $(h^*, \ell^*, \delta^*)$ on test; drop samples violating overlap.

**for** each sample $i \in D$ **do**
  **if** $p_{1,i} \geq h^*$ and $p_{0,i} \leq \ell^*$ and $u_i \geq \delta^*$ **then**
    $\text{label}_i \leftarrow A$ (Persuadable)
  **else if** $p_{1,i} \leq \ell^*$ and $p_{0,i} \geq h^*$ and $u_i \leq -\delta^*$ **then**
    $\text{label}_i \leftarrow D$ (Do Not Disturb)
  **else if** $p_{1,i} \geq h^*$ and $p_{0,i} \geq h^*$ and $|u_i| < \delta^*$ **then**
    $\text{label}_i \leftarrow B$ (Sure Thing)
  **else if** $p_{1,i} \leq \ell^*$ and $p_{0,i} \leq \ell^*$ and $|u_i| < \delta^*$ **then**
    $\text{label}_i \leftarrow C$ (Lost Cause)
  **else**
    $\text{label}_i \leftarrow \text{NearestCorner}(p_{1,i}, p_{0,i}; h^*, \ell^*, \text{sign}(u_i))$
  **end if**
**end for**
**return** $\{\text{label}_i\}$

**Note:** Figure 2b presents the conceptual diagram of counterfactual segmentation under V2.

#### V3

V3 introduces a hybrid causal–behavioral architecture that uses uplift estimation, behavioral validation, and model reassessment to produce a joint labeling framework. V3 begins with high-confidence lift-based labeling and reconciles theoretical label assignments with empirical behavior, followed by classifier-based refinement to re-label ambiguous or conflicting cases.

**TABLE 3.** Comparison of the three segmentation designs (V1–V3) by input dependencies and methodological focus.

| Version | Use of $(T, Y)$ | Use of $(\hat{p} * 1, \hat{p} * 0, u)$ | Output Labels | Core Characteristics |
|---|---|---|---|---|
| **V1 - Uplift-Only Triage** | Not used | Uplift score only | High / Medium / Low | Coarse sensitivity ranking; lacks causal interpretability |
| **V2 - Counterfactual Mapping** | Not used | Yes | A / B / C / D | Counterfactual quadrant assignment; interpretable but unstable near decision boundaries |
| **V3 - Hybrid Refinement** | Incorporated from Step 2 onward | Yes | A / B / C / D | Integrates causal estimation with behavioral validation; empirically consistent and robust |

**TABLE 4.** Counterfactual quadrant mapping for four-type labeling.

| Condition | Assigned Type |
|---|---|
| $\hat{p}_1 \geq y_1^{th}$ and $\hat{p}_0 \leq y_0^{th}$ and $\mu \geq \delta$ | A(Persuadable) |
| $\hat{p}_1 \geq y_1^{th}$ and $\hat{p}_0 \geq y_1^{th}$ and $\lvert\mu\rvert < \delta$ | B(Sure Thing) |
| $\hat{p}_1 \leq y_0^{th}$ and $\hat{p}_0 \leq y_0^{th}$ and $\lvert\mu\rvert < \delta$ | C(Lost Cause) |
| $\hat{p}_1 \leq y_0^{th}$ and $\hat{p}_0 \geq y_1^{th}$ and $\mu \leq \delta$ | D(Do-Not-Disturb) |

**Workflow Detail: Hybrid Labeling Pipeline**

**1. High-Confidence A/D Labeling (Uplift-Only).**
Samples with $|n| \geq n_{\text{high}}$ are directly assigned as **A** (Persuadable) or **D** (Do-Not-Disturb), forming an initial high-confidence set solely based on uplift ranking.

**2. Behavioral Evidence Matching.**
Remaining samples are validated against their behavioral evidence cells (Table 1). Inconsistencies between observed behavior $(T, Y)$ and assigned labels trigger conflict flags for subsequent correction.

**3. Preliminary B/C Attribution.**
For unlabeled cases, users with $Y = 1$ are provisionally assigned to **B** (Sure Thing), and those with $Y = 0$ to **C** (Lost Cause). Samples whose uplift contradicts these behavioral expectations are marked as conflicts (**B_conflict**, **C_conflict**).

**4. Model-Assisted Refinement.**
To resolve conflicts, binary classifiers are trained on appropriate subsets:

- **A/D discrimination**: $(S_1 \cup S_3)$ and $(S_2 \cup S_4)$
- **B/C discrimination**: $(S_2 \cup S_3)$, $(S_1 \cup S_4)$, and combined sets with A/D treated as noise.

Model predictions above calibrated thresholds overwrite conflicting labels, ensuring logical consistency.

The V3 hybrid causal–behavioral procedure consists of four stages, illustrated below:

Using this multi-stage infusion of uplift prediction, counterfactual reasoning, and behavioral correction, V3 produces causally interpretable and empirically consistent response-type segmentation, achieving significant AUUC stability and behavioral-consistency improvements over prior versions. By anchoring response-type assignments to both uplift estimates and observed $(T, Y)$ behavioral evidence, the V3 design promotes label consistency across time windows while allowing individual users to transition between response states as their behavior evolves.

This segmentation logic of the V3 hybrid labeling framework shows:

What is the uplift distribution?

It is partitioned into five pieces by adjustable thresholds, which correspond to the four stages of the labeling pipeline:

(1) uplift-based extreme segmentation;

**Algorithm 3** Hybrid Labeling (Compact Layout)

**Input:** Dataset $D$, $T \in \{0, 1\}$, $Y \in \{0, 1\}$
**Output:** Type $\in \{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\}$, Flags
**Protocol:** Train/Valid/Test split; maximize AUUC on Valid; freeze on Test.

(A) **Arm-wise models & Calibration**
Train $p_1(x)$ & $p_0(x)$; Calibrate per arm.
$u(x) = p_1 - p_0$; drop overlap violators.

(B) **Behavioral evidence partition**

$$S_1 = \{T_1, Y_1\}, S_2 = \{T_1, Y_0\},$$

$$S_3 = \{T_0, Y_1\}, S_4 = \{T_0, Y_0\}.$$

(C) **High/Mid-confidence A/D rules**
Tune $(\mu_{\text{high}}, \mu_{\text{mid}})$ on Valid. For $i \in D$:
a) If $\mu_i \geq \mu_{\text{high}} \rightarrow \lambda$; Else if $\leq -\mu_{\text{high}} \rightarrow D$
b) Else check $\mu_{\text{mid}} + b$-behavior consistency:
- $\mu_i \geq \mu_{\text{mid}}$ & $(S_1$ or $S_4) \rightarrow \lambda$
- $\mu_i \leq -\mu_{\text{mid}}$ & $(S_2$ or $S_3) \rightarrow D$
- Else conflict $\rightarrow$ `AD_conflict`

(D) **Initial B/C with conflict flags**
For remaining $i$:
- If $Y = 1$: $(\mu < -\mu_{\text{conf}}$?`B_conflt:B`$)$
- If $Y = 0$: $(\mu > \mu_{\text{conf}}$?`C_conflt:C`$)$

(E) **Model-assisted A↔D refinement**
Train Classifier on Step (C) labels (stratified).
Reassign if Prob $\geq \tau_{AD}$ & consistent quadrant.

(F) **Model-assisted B↔C refinement**
Use clean B/C from (D) as supervision.
Train/Calibrate B vs C models.
Refine if consensus $\geq \tau_{BC}$.

(G) **Output**
Return final labels, conflict types & diagnostics.

**Note:** Figure 2c presents the conceptual diagram of the hybrid labeling workflow (P3) combining confidence rules and model refinement.

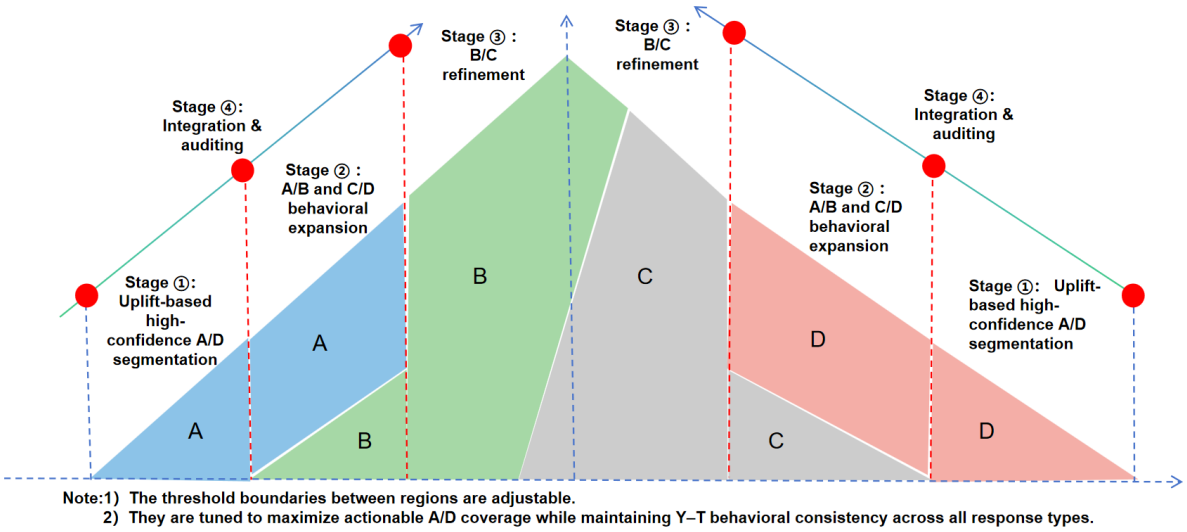



**FIGURE 2.** V3-Based Four-Stage Causal Labeling Framework.

(2) behavioral expansion of A/D boundaries;
(3) model-assisted tagging of ambiguous cases; and
(4) final integration and auditing.

The net result: maximum coverage of A and D users while keeping $(T, Y)$ behavior consistent across all four response types.

**TABLE 5.** Stages and Functional Roles in the V3 Causal Labeling Framework

| Position in Figure | Stage & Description |
|---|---|
| **Left Tail (A/D Extremes)** | **Stage 1 Uplift-Based Extreme Segmentation.** Identify high-confidence A/D users at the extreme ends of the uplift distribution, based on validated threshold cutoffs. |
| **Intermediate Zones (A/B and C/D Boundaries)** | **Stage 2 Behavioral Extension.** Expand A/D boundaries by incorporating behaviorally similar users whose $(Y, T)$ patterns align with the respective response logic. |
| **Central Region (B/C Area)** | **Stage 3 Model-Assisted Refinement.** Refine ambiguous users in the middle uplift region via supervised classifiers trained on non-conflict samples. |
| **Overall Framework (Final Output)** | **Stage 4 Final Integration & Auditing.** Integrate refined labels, perform $Y$–$T$ behavioral consistency validation, and maximize actionable A/D coverage. |

#### 4) Summary and Discussion

The three versions of user segmentation represent a journey from the very light-handed uplift-based stratification (V1), through the more complex counterfactual segmentation (V2), to the causal–behavioral hybrid refinement (V3) with which the final design aligns. The addition of counterfactual estimation and user profiling into the approach of V2 makes the return to descriptive profiling reach toward a causal user-description framework that is interpretable and grounded in the data. This journey reflects how the principles of causal inference can help operationalize user profiling, bridging incremental-impact modeling with behavioral realism to support decision-making within systems that rely on interventions.

### B. IMPROVED EXPERIMENTAL DESIGN

Re-approaching uplift modeling as an evaluation of user profiling, the improved experimental design strengthens the reasoning behind how components of uplift modeling work together. Rather than iteratively optimizing a single predictive model, we reframe the assessment of intervention approaches and refine all methodological components (feature selection, clustering, causal estimation, response labeling), each of which shapes the overall Area Under the Uplift Curve (AUUC). With a structured modular approach in place, the pipeline evaluation allows us to identify how components work together transparently, ensuring that an uplift-based user-profiling pipeline is reproducible and as interpretable as possible.

#### 1) Pipeline Overview

The framework for user profiling is based on causal principles. Following from causal user profiling, we add to the pipeline (Figure 3) a user-profiling and "de-causalizing" methodology based on feature uniqueness and response. In particular, clustering, causal estimation, user profiling, and component-wise ablation are incorporated to quantify contributions.

Operating on a monthly rolling basis—across both cross-sectional and temporal vigilance—the pipeline allows examination of how feature selection, clustering, response-interaction information, and causal adjustment work together as individual components of a Causal User Profiling

(CUP) pipeline. All experiments share a common time-based train/validation/test protocol with shared random seeds and fixed preprocessing across months, highlighting that these choices drive stability and practical reliability in modeling user profiling under the causal framework.

### 2) Core Modules: From Feature Selection to Causal Estimation

The pipeline begins by identifying features that are both stable and causally relevant, allowing us to draw causal insights generalizable to unseen treatment groups. A multi-stage selection process combines statistical relevance (Information Value, IV), temporal stability (Population Stability Index, PSI), and causal importance (Causal Forest variable importance and stepwise regression) [13]. This produces multiple feature sets from the base DataFrame: IV-only/PSI subsets, CF-only/PSI, and hybrid types (IV + CF, IV + CF + STEP). The aim is to obtain a parsimonious set of features representative of the population, standardized and imputed on the same folds to ensure fairness and causal interpretability.

Before uplift estimation, we must address any remaining treatment-control imbalances, often done using Propensity Score Matching (PSM) and Inverse Probability Weighting (IPW) [10], [38]. These covariate-adjustment techniques reweight samples to render treated and control groups comparable so that the uplift reflects the genuine treatment effect rather than asymmetries in treatment allocation (i.e., selection bias). For randomized subsets, they also serve as a sanity check for stability: estimated uplift effects should confirm confounding control.

Uplift estimation is conducted using two families of models: meta-learners and Causal Forests. Meta-learners such as T-, S-, X-, R-, and DR-learners factor CATE estimation into modular supervised-learning tasks [14], [15], while Causal Forests provide nonparametric and asymptotically consistent estimates of heterogeneous treatment effects [16]. To preserve causal integrity, all learners share the same preprocessing, data, and evaluation strategy. Logistic Regression often serves as the baseline learner for computational efficiency, while Random Forests, GBDT, and XGBoost are employed as sanity checks. Each learner's outputs are then passed to the Four-Type Response Segmentation Module (Section 3.2) to yield CATE/uplift scores tagged with behavioral names or interpretable types.

In a final step to facilitate causal interpretability, we cluster users into causally homogeneous strata. Recognizing that real-world behavioral heterogeneity drives treatment-effect variation, CUP employs K-Means clustering on standardized features to form stable subpopulations, each serving as a contextual unit for uplift estimation [8], [22]. The optimal number of clusters K is chosen via elbow and silhouette criteria. Clusters below 2% of total samples or yielding AUUC lower than the baseline are merged into others. Notably, clustering enhances interpretability and reveals behavioral regimes that would otherwise remain hidden.

### 3) Evaluation and Ablation Design

This evaluation procedure quantifies how each decision either promoted or detracted from ultimate uplift performance. We use area under the uplift curve (AUUC) as the core performance measure, as well as the Qini coefficient and uplift@k [17], [19], [22]. We separate marginal effects through component-wise ablation (e.g., removing the clustering, causal feature selection, or label refinement module). Performance is evaluated both globally and by cluster, using a weighted metric:

$$\text{Weighted AUUC} = \sum_{k=1}^{K} w_k \cdot \text{AUUC}_k, \tag{3}$$

where $w_k$ denotes the relative proportion of valid samples in cluster $k$.

**Algorithm 4** Core Experimental Loop (Baseline and Clustering-enabled)

**Input:** Dataset $D$; FeatureSets $\{F_1, \ldots, F_6\}$; Metalearners $\{T, S, X, \text{DR, CF}\}$; Baselearners $\{\text{LR, RF, GBDT, XGB}\}$; Months $\{M_1, \ldots, M_6\}$
**Output:** AUUC scores; cluster statistics; variable importance

**for** each month $m \subseteq \{M_1, \ldots, M_6\}$ **do**
  $D_m \leftarrow \text{subset}(D, \text{month} = m)$
  **for** each feature set $F$ in FeatureSets **do**
    $X \leftarrow$ select_features$(D_m, F)$; $T \leftarrow$ treatment; $Y \leftarrow$ outcome
    **for** each $ML \in$ Metalearners **do**
      **for** each $BL \in$ Baselearners **do**
        **if** clustering_enabled **then**
          $clusters \leftarrow \text{KMeans}(X, K)$
          **for** each cluster $c$ in $clusters$ **do**
            **if** valid_cluster($c$) **then**
              $model_c \leftarrow$ train$(ML, BL, X_c, T_c, Y_c)$
              $score_c \leftarrow$ evaluate_AUUC$(model_c, X_c, T_c, Y_c)$
            **else**
              skip
            **end if**
          **end for**
          $clusters \leftarrow$ merge_low_AUUC$(clusters)$
        **else**
          $model \leftarrow$ train$(ML, BL, X, T, Y)$
          $score \leftarrow$ evaluate_AUUC$(model, X, T, Y)$
        **end if**
      **end for**
    **end for**
  **end for**
**end for**

Other diagnostics include:

- Label adherence rates (band, quadrant, and behavioral-cell consistency)
- Type-wise outcome balance between treatment and control groups

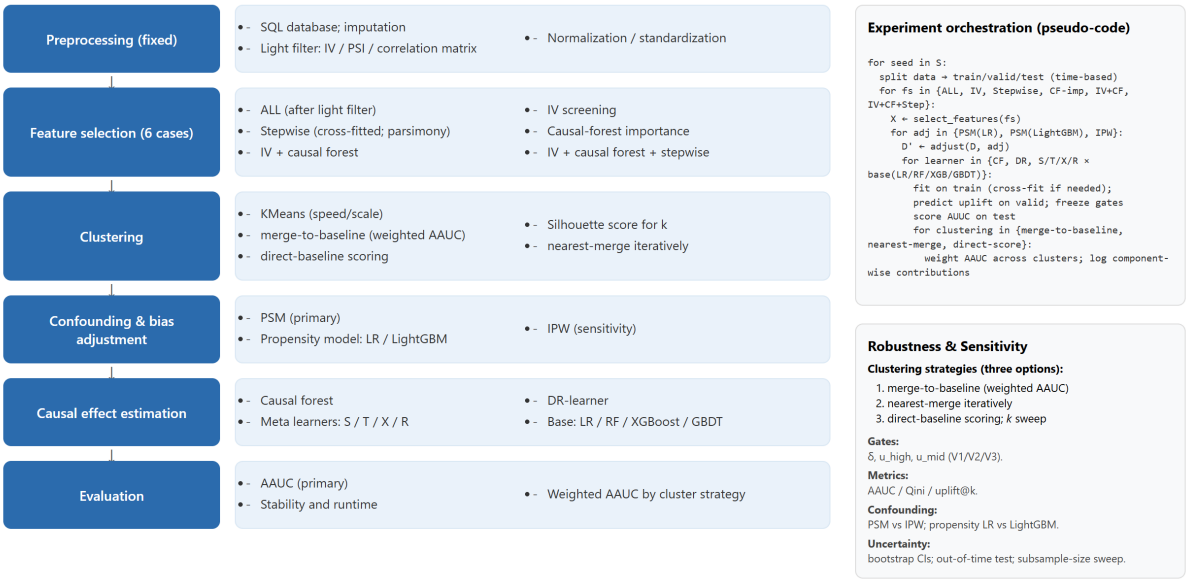



**FIGURE 3.** Improved experimental pipeline of the Causal User Profiling (CUP) framework. The pipeline integrates feature selection, clustering, confounding adjustment, causal effect estimation, and evaluation into a reproducible workflow executed in monthly rolling loops.

- Temporal stability measured via Cohen's $\kappa$ across months
- Computational efficiency (runtime and convergence)

We computed robustness via bootstrap confidence intervals, out-of-time validation, and selective sweeps through subsamples when the uplift they contained appeared erratic. We performed all experiments with fixed random seeds, version-controlled datasets, and standardized preprocessing pipelines. We took comprehensive logs of AUUC, variable importance, and cluster diagnostics from each run, and then stitched together the pieces. This provided a transparent "causal audit trail" we could use to replicate ourselves.

To alleviate deliberate decisions "standing apart" from our evaluation framework, we made reproducibility a foundation within our evaluation rather than separate it out as its own stage. This smooth trade-off between methodological rigor and real-world reliability was seen empirically: this evaluation strategy stabilized the AUUC variance across monthly slices. It reduced the damage from confounding-driven fluctuations while giving us a reliable measure of true treatment heterogeneity.

### C. SUMMARY

In this chapter we presented an enhanced experimental design that integrates both causal inference and user profiling into a single uplift-modeling approach. The Four-Type Response Segmentation Module guarantees causal labeling robustness, while component-wise ablation analysis gauges the contribution of each methodological component—feature selection, clustering, and causal estimation—to the uplift score. All of this is implemented in one consolidated process that weaves evaluation and reproducibility together, thereby laying the groundwork for a transparent, interpretable, and empirically validated basis for dynamic causal user profiling.

## IV. DATA DESCRIPTION

In our empirical analysis we employ proprietary data collected from users of a leading Chinese internet finance platform's mobile app. To enable causal interpretability, we consider only active borrowers, i.e., users with outstanding loan balances in the observation window. Compared to non-borrowers (who are not indebted or have satisfied prior loan demands), active borrowers are more homogeneous: they seek credit more consistently, reducing the number of possible confounders arising from variation in borrowing motives. By ensuring comparability between treatment users' behaviors, this focus improves the power of causal identification.

The dataset consists of an intermingled trove of information covering demographic behavior as well as credit status—including both static user attributes and dynamic financial indicators—comprising: gender, age, and city tier; borrowing frequency; repayment performance; credit-line utilization limits; overdue history and repayment discipline; and consumption-related activity. These variables capture treatment heterogeneity and user-specific behavior in general.

Prior to model fitting, several procedures were conducted to improve comparability and robustness within and across potential covariates. Following business conventions, missing or invalid values were set according to domain logic; numerous continuous features were scaled for numeric comparability; and feature selection was subsequently applied through a diagnostic three-stage procedure consistent with previously documented theory [22]. This screening stage

is performed before constructing the downstream feature-set configurations (e.g., IV-only, Causal, and hybrid sets), ensuring that all reported feature sets are derived from the same filtered and stability-checked candidate pool:

1. Information Value (IV) was computed to quantify predictive relevance for the target outcome.

2. Population Stability Index (PSI) was used to evaluate temporal stability and detect distributional shifts across months.

3. Pairwise correlation analysis identified redundant variables and mitigated multicollinearity.

Variables having IV < 0.05 or PSI > 0.25 were dropped, and features were merged when their pairwise correlation exceeded 0.8 in absolute terms. This process produced a balanced set of features encompassing interpretability, robustness, and predictive capability—in line with recent methodological standards in heterogeneous treatment-effect research for similar use cases [41].

Descriptive statistics for our dataset are shown in Table 4-1. The dataset contains approximately 720,000 user-months generated from six calendar months. The six-month window reflects a practical trade-off between behavioral stability and sample coverage in a real-world lending system, providing repeated exposure to interventions for temporal robustness checks while limiting structural drift in user composition and platform policy. All loans have a mean utilization rate of 62.8% (SD = 24.5%), indicating behavioral heterogeneity. An average repayment-timeliness index of 0.91 reflects a disciplined borrower population. Approximately 78% of users live in a tier-2 or lower-tier city.

Users in the treatment group appear to exhibit somewhat higher utilization and higher-frequency engagement than users in the control group, implying heterogeneous intervention responses relevant for uplift analyses. Feature stability and relevance diagnostics further affirm the robustness of this variable set: the average IV across retained features is 0.23, and the average PSI is 0.07, both within accepted stability bounds [42]. Behavioral and repayment indicators have the strongest predictive power, mirroring previous research showing that dynamic behavioral attributes are highly predictive of treatment responsiveness. Correlation diagnostics confirm that we effectively controlled for multicollinearity.

These procedures ensure that the empirical model identifies true behavioral heterogeneity rather than confounding arising from unstable or superfluous predictors. In summary, the dataset underwent a systematic and judicious filtering and refinement process consistent with best practices in causal inference and comparable recent works in the literature. The resulting analysis environment is stable, representative, and amenable to estimating heterogeneous treatment effects, as well as to implementing causal models of users in subsequent sections.

### A. SEGMENTATION AND DISTRIBUTION OF FOUR RESPONSE TYPES

This section presents the empirical results of the Four-Type Response Segmentation Module introduced in Chapter 3, illustrating how uplift-based causal labeling emerges in the dataset. Within the causal user-profiling framework, users are segmented based on their Individual Treatment Effect (ITE) under the potential-outcomes model. In particular, we define four canonical response types as:

- **Persuadables (Type A):** perform the target action only if treated
$$(y_1 = 1, y_0 = 0)$$
- **Sure Things (Type B):** perform the action regardless of treatment
$$(y_1 = 1, y_0 = 1)$$
- **Lost Causes (Type C):** never perform the action
$$(y_1 = 0, y_0 = 0)$$
- **Do Not Disturb (Type D):** perform the action only if not treated
$$(y_1 = 0, y_0 = 1)$$

This typology lends intelligibility to a varied desert of behaviour and enables causal modelling, tailored intervention strategies, and marketing resource allocation to begin conceptually. In practice, the frequency of types across users is summarized in Figure 5a. There are significantly more Persuadables and Lost Causes relative to Sure Things in the user population, meaning that there are slightly more people who will change behaviour under some treatment than will remain the same across any treatment; but the most observations are of users who are resistant to treatment. Sure Things should have stable demand, although—as with everything else—they should be given equal rates of exposure in the broadest possible sense of the term. Do-Not-Disturb users are a hindrance, but also function as a warning against the disadvantages of treatment, serving as a caution against over-targeting and treatment fatigue, even if they constitute only a small portion of the total cohort.

To make even clearer how this segmentation might typically be done, Figure 5b plots the empirical joint distribution of $(y_0, y_1)$ and notes the four decision regions as dictated by the potential-outcomes formulation.

### B. EMPIRICAL EVALUATION OF THE CUP FRAMEWORK

This section describes the empirical evaluation of the Causal User Profiling (CUP) framework. We evaluated each of the core components of CUP—feature selection, clustering, and meta-learner configuration—individually and as a combined system to assess their influence on uplift performance, as measured by the Area Under the Uplift Curve (AUUC).


**FIGURE 4.** Workflow of Data Preprocessing and Variable Selection. Note. The workflow depicts sequential data preparation steps: Raw Data Collection → Data Cleaning and Missing Value Treatment → Normalization and Standardization → IV Computation → PSI Evaluation → Correlation Filtering → Feature Retention for Modeling.

**TABLE 6.** Descriptive Statistics of the Sample

| Variable Category | Representative Variables | Mean | Std. Dev. | Interpretation |
|---|---|---|---|---|
| Demographics | Age, City Tier | 33.7 | 7.9 | Stable, low-variance |
| Borrowing Behavior | Loan Utilization, Application Frequency | 0.63 | 0.25 | High behavioral heterogeneity |
| Repayment Performance | Repayment Ratio, Overdue Days | 0.91 | 0.10 | Strong repayment discipline |
| Credit Status | Credit Limit, Credit Score | 0.58 | 0.22 | Moderate variability |
| Consumption Behavior | Transaction Count, Avg. Spend | 0.47 | 0.18 | Medium transactional diversity |

### 1) Feature Selection and Uplift Performance

To explore how variable screening affects heterogeneous treatment-effect estimation, six feature sets were compared:

ALL — all variables;

IV — selected by information value;

Causal — chosen by causal-forest importance and PSI stability;

Stepwise — retained through cross-fitted regression;

IV + Causal — intersection of the first two;

IV + Causal + Stepwise — the hybrid refinement.

Findings show that appropriate variable selection is critical to uplift modeling. Using ALL resulted in the lowest mean AUUC and the highest variance—it appears that when the models must "dig through" this noise, it severely weakens identification with respect to the heterogeneous effects themselves.

Both IV and Causal selections produce meaningfully better results than the baseline, but in different ways:

IV features push the AUUC higher on the simpler meta-learners (T- and X-Learners),

whereas Causal features assist with overall model stability, particularly for the larger DR-Learner.

Because benefits come from different sources, it is not obvious a priori how the responses will sort themselves out when combining them. The IV + Causal set produces the overall highest mean AUUC—validating that "picking dimensions gives you information gain, and picking good dimensions gives you causal relevance." Once this outer-join set of dimensions is added, Stepwise refinement can then be incorporated. These hybrid sets exhibit slightly lower mean AUUC but are less volatile month-to-month, yielding smaller uplift-consistency (UC) indices. Although their mean AUUC is marginally lower, the ability to depend on the model producing similar results across months is worth the trade-off in mean values.

In summary:

IV + Causal accelerates accuracy at the cost of some stability.

IV + Causal + Stepwise favours stability at the cost of some accuracy.

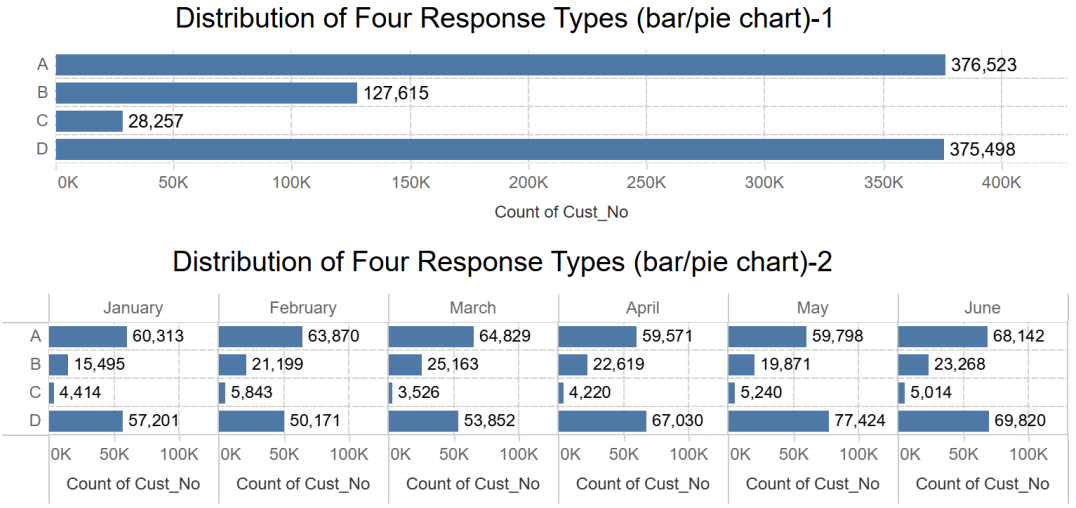
The second approach is more deployment-ready and is therefore used as "the features" moving forward. Across learners as well, the meta- and base-learners perform relatively consistently. Logistic Regression (LR) stands out as the strongest base learner under the hybrid feature sets, while tree-based learners show weakness when noise remains in the embeddings. Therefore, feature selection forms the skeletal structure for CUP.

### 2) Clustering Strategies and AUUC Enhancement

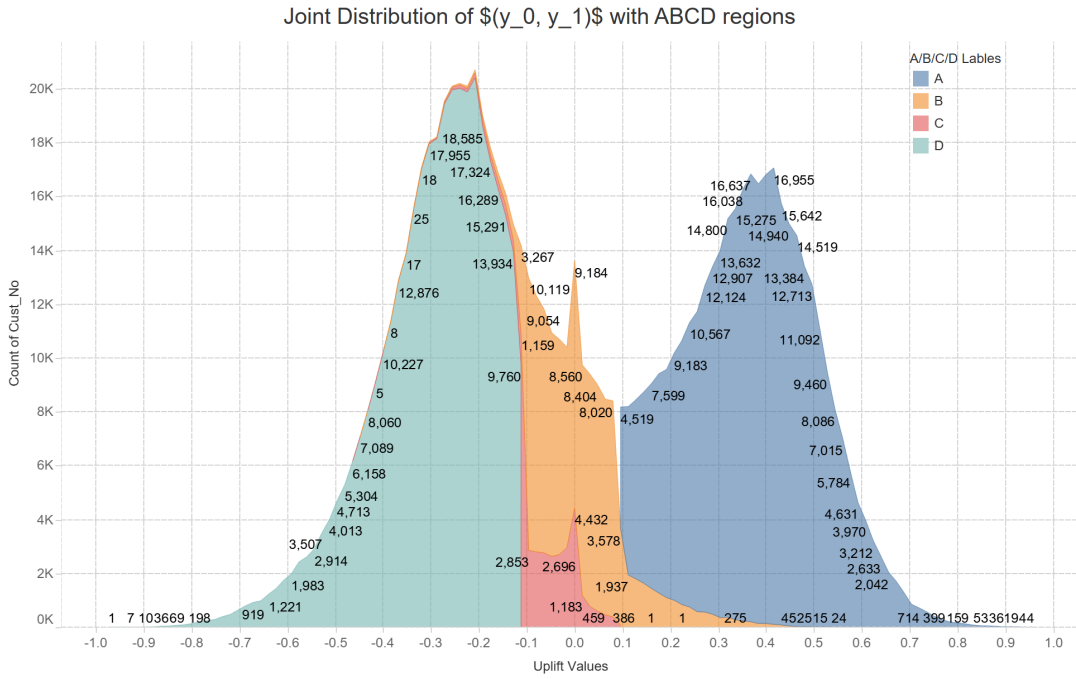
Clustering was introduced to investigate local treatment heterogeneity independent of an explicit association with outcomes, consistent with the notion that subgroups with different causal effects can be mapped out by recursive partitioning [6]. The empirical results suggest that clustering is helpful but is conditional rather than absolute, and as a consequence should be used cautiously.

The contrast between Direct, C1 (Merging), and C2 (Replacement) shows a clear trade-off. The Elbow Method suggests optimal clusters around K=6, at which point the silhouette score levels off (see Figure 7a). This is parsimonious and easy to interpret, and we derive clusters following this throughout our analysis.

Across the three methods, the C2 step is consistently the strongest, yielding the highest uplift effect: mean AUUC ($\approx 0.09$) with the lowest variance ($\pm 0.01$). By "replacing" weak clusters with global predictions, C2 has a regularizing effect, balancing local adaptivity and global stability. This variance-reducing characteristic maps onto the findings of Devriendt et al. [22], who show that "ensemble-style" uplift models produce lower variance (and greater reliability) than "isolated" two-model structures. In our case, Direct identifies




(a) Distribution of Four Response Types (bar/pie chart)

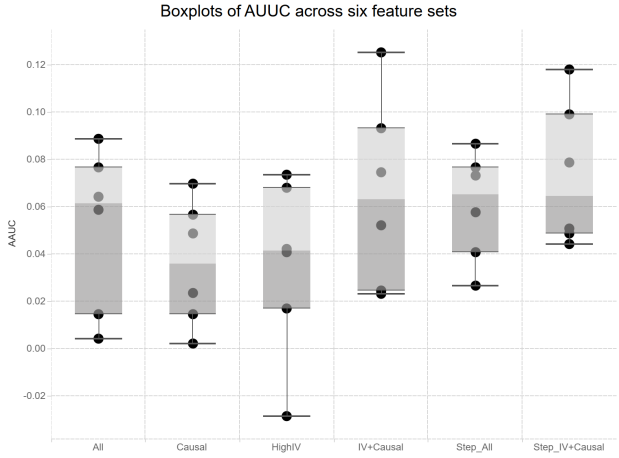



(b) Joint Distribution of $(y_0, y_1)$ with ABCD regions

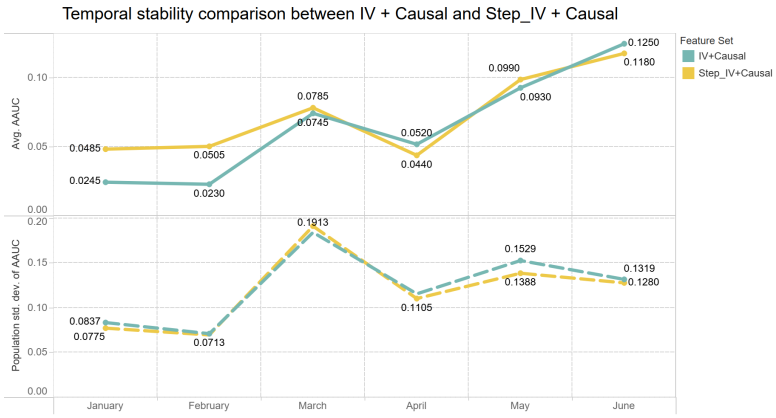
**FIGURE 5.**

behavioral heterogeneity but does not produce stable models. C1 reduces this variance somewhat, but part of it remains. Empirically, the clustering order is C2 > C1 > Direct (see Figure 7b), placing clustering as a refinement rather than a requisite.

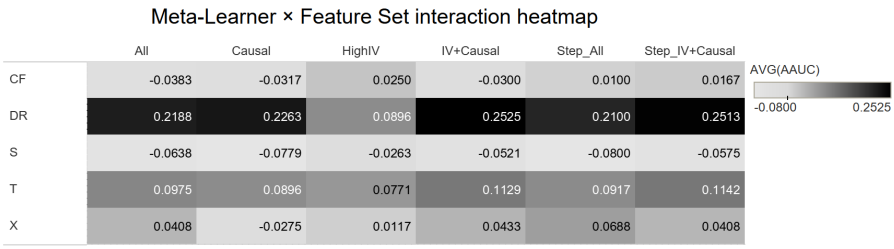
Clustering interacts differently with the various meta-learners. As reflected in Figure 7c, the X-Learner shows the most robust and consistent improvement, reflecting its theoretical gain in segmented or unbalanced samples [15]. In contrast, the DR-Learner and T-Learner show AUUC drops or inconsistencies after clustering, and the Causal Forest (CF) even experiences extreme performance deterioration. This is consistent with earlier findings that tree-based causal estimators lose efficiency when subgroup sample sizes are small or unbalanced [13], [35] —meta-learners capable of "cross-arm information sharing" (such as the X-Learner) remain stable under segmentation, whereas others are sensitive to fragmentation. Figure 6d illustrates cluster-level performance under the X-Learner. As expected, clustering amplifies both signal and noise: some cluster-level AUUC values improve substantially, while others drop sharply. This follows the familiar "variance-amplifying effect," where smaller subsamples increase both estimation bias and variance [13]. Thus, clustering uncovers latent heterogeneity but can also amplify



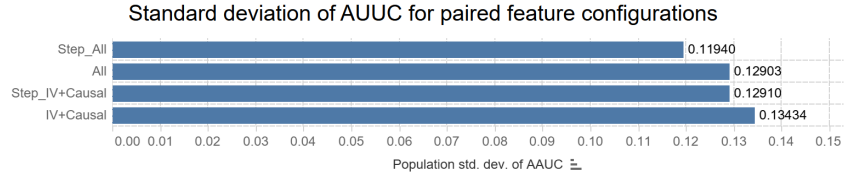

(a) Boxplots of AUUC across six feature sets.




(b) Temporal mean AUUC over six months comparing IV + Causal and IV + Causal + Stepwise.




(c) Heatmap showing AUUC values across Meta-Learner × Feature Set combinations.




(d) Standard deviation of AUUC for four paired feature configurations, illustrating the variance-reduction effect of Stepwise refinement.

**FIGURE 6.**

random noise, helping to explain why its performance varies across datasets and time periods.

The effectiveness of clustering depends critically on the feature space used. In Figure 7e, clustering derived from the Causal feature set yields the most consistent gains, outperforming clustering based on IV-only or ALL-variable feature sets. This indicates that clustering is effective only when the feature space encodes causally relevant information: causal features yield a more coherent and interpretable map of the underlying terrain, enabling clustering to identify more precise treatment heterogeneity [43], [44]. To recap: causal features structure the space, and clustering amplifies heterogeneity. Clustering based solely on predictive or correlation-driven variables tends to amplify noise.

Clustering does not universally improve outcomes. It produces a wide spread in AUUC across clusters; improvement in some clusters may be accompanied by deterioration in others. This dual-edged nature makes clustering powerful but dangerous. Clustering should be viewed as a "positive calibration mechanism," effective only when strongly rooted in causal features and paired with robust learners (such as the X-Learner) that remain stable under segmentation. Its feedback on the Four-Type causal segmentation (Persuadable, Sure Thing, Lost Cause, Do-Not-Disturb) must also be monitored, as behavioral balance can be distorted through subgroup redistribution.

In summary, clustering improves uplift estimation only "under reasonably complete causal feature structures" and with "meta-learners that remain stable under segmentation." Clustering performs best with the C2 replacement strategy. Given that clustering can amplify both gains and volatility, it should be treated as a calibrated instrument, not as a mandatory component of every causal user-profiling pipeline.

#### 3) Meta-Learner and Base-Learner Configurations

A total of sixteen meta/base combinations (four meta-learners × four base learners) were explored using the optimal feature set. Among the meta-learners, the DR-Learner consistently produced the best mean AUUC and the worst variance, showing that this model is robust to model misspecification. The X-Learner offers strong competition, particularly under treatment-imbalance situations, whereas the T- and S-Learners are far more unreliable in complex, high-variance environments. Among the base learners, Logistic Regression (LR) performs best in terms of stability and interpretability, while tree-based learners exhibit substantially more variability overall.

#### 4) Optimal Pathway and Overall Effect

Putting all these pieces together, the final CUP workflow uses a hybrid IV + (Causal ∩ Stepwise) feature-selection design, employs the C2 "replacement strategy" with a moderate number of clusters $K$, and adopts a DR-Learner with Logistic Regression as the modeling configuration. This integrated pathway yields a large and stable improvement in AUUC relative to the baseline ("all features + no clustering + standard learner"). The main performance gains come from variables that are informative (high IV) and causally relevant, and the C2 approach provides a form of structural regularization that prevents overfitting and stabilizes heterogeneous treatment estimation. Together, these components lead to a balanced pipeline that improves both accuracy and interpretability.

Quantitatively, each module of the CUP framework contributes a distinct and measurable improvement to uplift performance. Feature selection alone increases AUUC by approximately 25–30%, reflecting the value of filtering out noisy predictors and emphasizing causally important variables. Incorporating C2 clustering provides an additional 10–12% gain by stabilizing weak clusters and harmonizing local heterogeneity with global patterns. Optimizing the meta/base configuration (DR-Learner + Logistic Regression) yields a further 5–8% improvement, enhancing robustness while maintaining interpretability. Cumulatively, the integrated CUP workflow achieves roughly 45–50% higher AUUC than the standard uplift-modeling baseline, demonstrating that each component contributes a clear and persistent increment across the six monthly slices.

## V. DISCUSSION

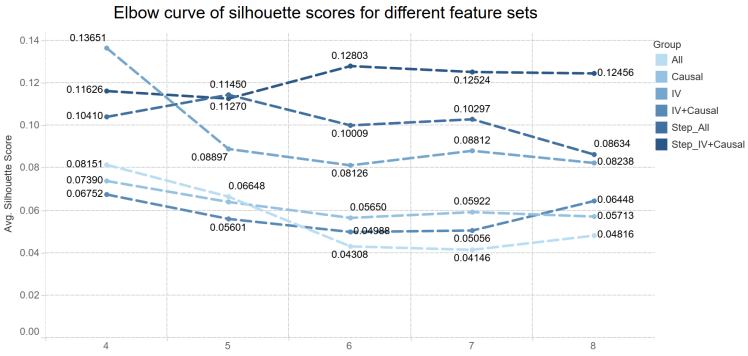
The empirical validation of the framework demonstrates that causal inference and uplift modelling can be rigorously brought to bear on user analytics, closing a classic gap in the analytics space between prediction and intervention. This section discusses the empirical findings in relation to existing uplift modeling and user profiling approaches, with emphasis on stability, interpretability, and operational relevance.

Causal User Profiling moves beyond descriptive or purely predictive profiling by centering treatment responsiveness as its primary analytical dimension. This reframing abstracts user modelling from the question of who users are, to how users behave when acted on [6], [22]. This move from correlation to causation represents a transformation in the conceptual underpinnings of personalization science, toward a framework that predicts behaviour while also explaining why effects transpire. The empirical findings confirm that causal interpretability is distinctive in its potential to redefine user segmentation. Rooted in the potential outcomes framework [11], later augmented into the metalevel framework of meta-learning [15], CUP quantifies heterogeneous treatment effects through the four-type response taxonomy of Persuadables, Sure Things, Lost Causes, and Do-Not-Disturbs [17].

This treatment-aware representation of user behaviour indicates that the heterogeneity observed in behaviour is not an artefact of random variation but rather a marker of differential responsiveness to intervention. In the digital lending case-study context, these distinctions illustrate how interventions may activate engagement, reinforce inevitable outcomes, protect users from unnecessary actions, or respect non-responsiveness. Compared with conventional uplift modeling pipelines that emphasize ranking accuracy alone, CUP aligns estimation, evaluation, and response interpretation within a unified analytical framework, yielding more stable gains

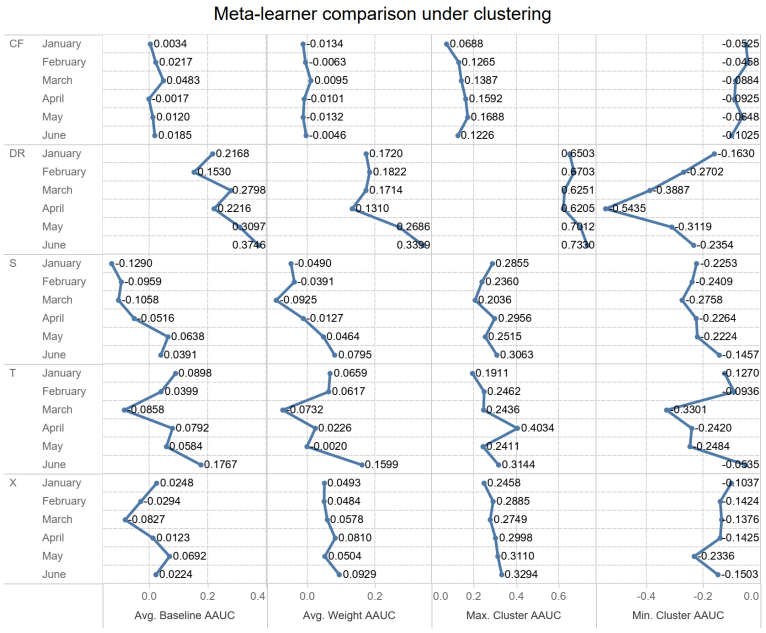
This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3670857



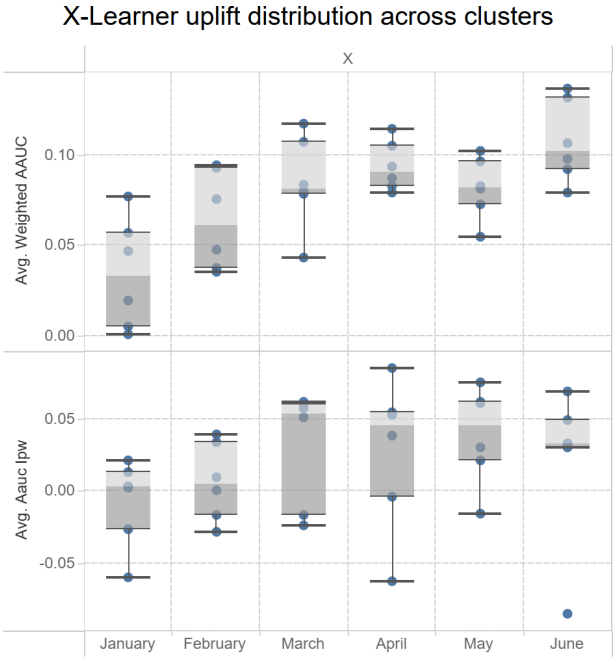



(a) Elbow curve for cluster number selection. The silhouette score flattens near K=[4,8], suggesting the optimal trade-off between within-cluster cohesion and between-cluster separation.

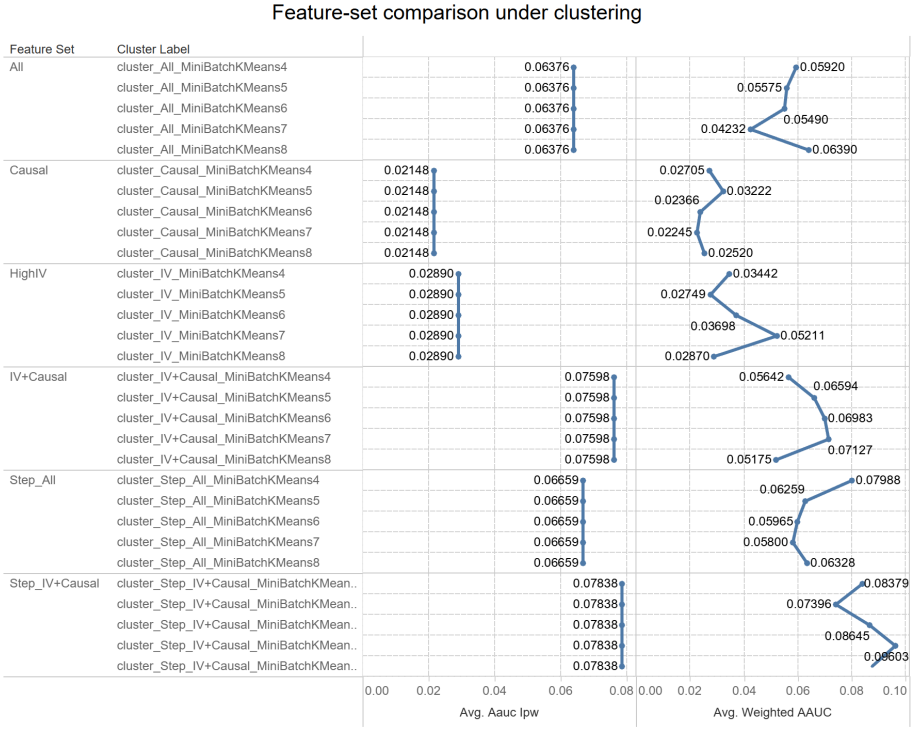

(b) Comparison of clustering strategies (Direct, C1, C2). Weighted AUUC distributions indicate the performance order C2 > C1 > Direct, with C2 achieving the highest stability and lowest variance.




(c) Meta-learner comparison under clustering. X-Learner shows consistent improvement, while DR-, T-, and CF-Learners fluctuate or decline, highlighting X-Learner's robustness.




(d) X-Learner uplift distribution across clusters. Boxplots show that clustering magnifies both signal and noise, increasing the variance in AUUC between subgroups.




(e) Feature-set comparison under clustering. Clustering based on causal features yields the highest and most stable AUUC improvements, confirming the advantage of causal-based representation.

**FIGURE 7.**

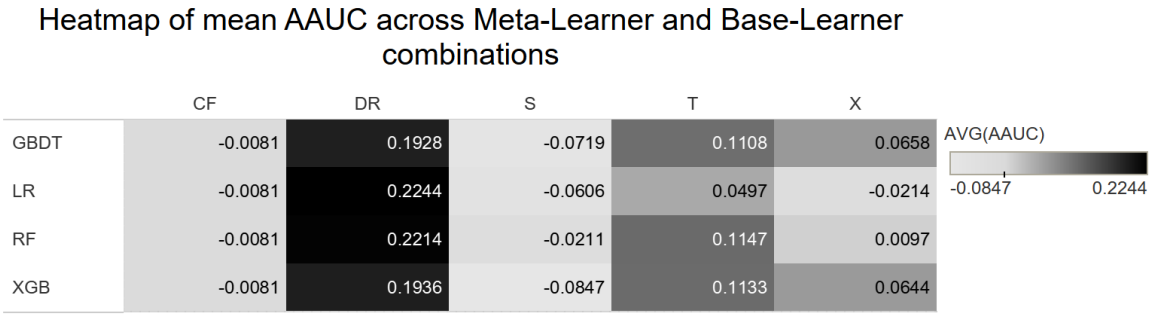



**FIGURE 8.** Heatmap showing the mean AAUC values for all Meta-Learner × Base-Learner configurations. Darker cells indicate higher uplift performance. The DR-Learner combined with Logistic Regression (LR) achieves the highest mean AAUC and lowest variance, demonstrating the most balanced trade-off between accuracy, stability, and interpretability.

under repeated interventions.

A first salient insight from the results concerns the interrelationship between feature quality, model stability, and causal validity. The hybrid feature selection strategy—combining Information Value (IV), Causal Forest importance, and Stepwise refinement—produced the most stable uplift performance across monthly samples. This observation confirms the argument that causal estimation depends as much on data stability and regime design as on algorithmic sophistication [6], [40]. High-IV variables ensure that the model learns from information-rich dimensions; causal importance anchors relevance to treatment effects; and Stepwise selection serves as a form of variance regularization. Together, these components yield a "causal feature space" that balances predictive strength with generalizable structural characteristics. This finding reinforces the notion of "data refinement" [4], consistent with evidence in causal machine learning that emphasizes disciplined feature design over increasing algorithmic complexity.

A second insight relates to methodological integration, which contributes directly to the robustness of CUP. Rather than operating as isolated modules, feature selection, clustering, and causal estimation work synergistically. The C2 replacement strategy compensates for unstable clusters by substituting their predictions with those of the global model, producing a hierarchical regularization effect that maintains local sensitivity while ensuring global consistency [4]. Within this structure, the Doubly Robust (DR) Learner combined with Logistic Regression (LR) delivers a favourable balance between robustness and interpretability. While alternative configurations such as the X-Learner may perform well under certain imbalance conditions, the DR–LR configuration demonstrated greater temporal stability across repeated deployments, which is essential in operational environments.

From a practical perspective, CUP grounds causal reasoning in consequences experienced by decision-makers. By estimating conditional treatment effects rather than predictive probabilities, the framework enables intervention design based on causal evidence rather than intuition or correlation. Empirically, the sequential stacking of methodological components yields meaningful and interpretable gains in performance:

- Hybrid feature selection improves AUUC by approximately 25–30% over the baseline;
- Clustering contributes an additional 10–12% through the C2 refinement strategy;
- The DR-Learner + LR configuration adds a further 5–8% uplift.

Cumulatively, the full CUP pipeline achieves an approximately 45–50% improvement in model performance relative to conventional profiling approaches. More importantly, these gains persist across multiple time windows, indicating systematic rather than incidental improvements. From a computational standpoint, the runtime of CUP is dominated by base learner training and clustering stages and introduces no additional asymptotic complexity beyond standard uplift modeling pipelines, making it tractable for large-scale tabular datasets.

## VI. LIMITATIONS OF THE STUDY

Despite its empirical strengths, this study has several limitations. First, the analysis is based on data from a single digital lending platform. While this enables controlled evaluation under realistic operational conditions, it limits external generalizability. Future research should assess CUP across additional domains such as e-commerce, insurance, and public finance to evaluate cross-context robustness.

Second, the treatment variable aggregates heterogeneous interventions (e.g., coupons, credit-line increases, outbound calls), which may obscure intervention-specific behavioural mechanisms. Extending CUP to explicit multi-treatment or dynamic intervention settings would enable finer-grained analysis of intervention-specific causal effects.

Third, the evaluation emphasizes temporal consistency across six monthly deployments rather than formal statistical hypothesis testing. This design choice reflects an operational focus on stability and reproducibility but may limit inference in settings that require formal significance testing.

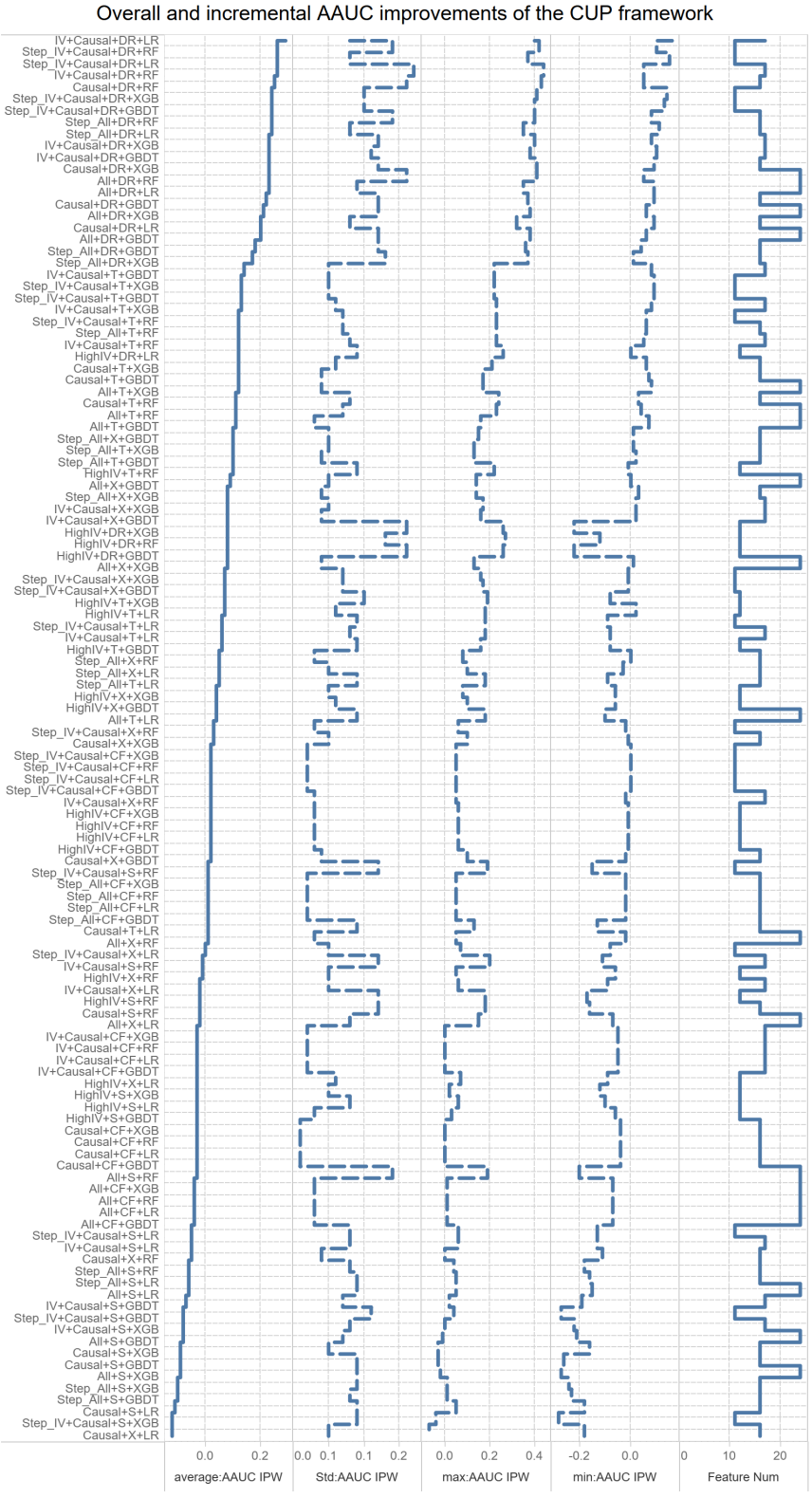



**FIGURE 9.** Cumulative gain curves comparing the optimized CUP pathway with the baseline; the CUP curve uniformly dominates.).

Finally, issues of fairness, transparency, and ethical deployment warrant further attention. Incorporating fairness-

aware learning and causal explainability into CUP represents an important direction for future work, particularly in sensitive financial applications [1], [4].

## VII. CONCLUSIONS AND FUTURE WORK

This study contributes to the growing literature on causal analytics by positioning causality as the organizing principle of user profiling. We propose Causal User Profiling (CUP), an integrated pipeline that combines feature selection, clustering, and meta-learning into a reproducible and interpretable workflow that connects causal estimation with actionable decision-making.

Empirically, CUP captures heterogeneous treatment effects with temporal stability; conceptually, it reframes user profiling as the causal understanding of behavioural response; and practically, it provides a scalable foundation for treatment-aware strategies in diverse digital ecosystems. Rather than competing with predictive machine learning, CUP complements it by explaining why interventions work and for whom, advancing personalization from outcome prediction toward causal understanding and decision optimization.

## ACKNOWLEDGMENT

This study benefited from the Haier Group Digital Finance Innovation Initiative (which provided the access to data and computer resources for empirical validation of our proposed model), and we are especially grateful to them for the implementation of the CUP.

## REFERENCES

[1] Rich, E. (1979). User modeling via stereotypes. Cognitive Science, 3(4), 329–354.

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749. https://doi.org/10.1109/TKDE.2005.99

[3] Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), The Adaptive Web (pp. 3–53). Springer. https://doi.org/10.1007/978-3-540-72079-9_1

[4] Eke, C. I., Norman, A. A., & Ozuem, W. (2019). User profiling in personalized recommender systems: A systematic review. IEEE Access, 7, 146923–146940. https://doi.org/10.1109/ACCESS.2018.2887321

[5] Mirylenka, D., Ricci, F., & Rokach, L. (2019). User modeling and personalization. In Recommender Systems Handbook. https://doi.org/10.1145/3357384.3357818

[6] Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

[7] Purificato, F., Rago, A., Belkhir, A., Lanzini, P., & Cirillo, P. (2024). Deep causal models: A survey. Information Processing & Management, 61(3), 103579. https://doi.org/10.1016/j.ipm.2023.103579

[8] Wu, W., Yuan, F., Huang, J., Yu, X., & Zhang, M. (2024). Social-network-based user profiling: A survey. Information Sciences, 648, 119021. https://doi.org/10.1016/j.ins.2024.119021

[9] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803478

[10] Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. Chapman & Hall/CRC. https://doi.org/10.1201/9780429259654

[11] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701. https://doi.org/10.1037/h0037350

[12] Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

[13] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Biometrika, 105(2), 287–301. https://doi.org/10.1093/biomet/asx045

[14] Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. Annals of Statistics, 49(6), 3935–3963. https://doi.org/10.1214/20-AOS1964

[15] Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects. PNAS, 116(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

[16] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

[17] Radcliffe, N. J., & Surry, P. D. (2011). Uplift modelling with significance-based trees. Stochastic Solutions Technical Report.

[18] Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems, 32, 303–327. https://doi.org/10.1007/s10115-011-0434-0

[19] Gutierrez, P., & Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. Information Sciences, 420, 590–598. https://doi.org/10.1016/j.ins.2017.02.002

[20] Radcliffe, N. J. (2007). Using control groups to target on predicted lift. Stochastic Solutions.

[21] Jaskowski, M., & Jaroszewicz, S. (2012). Uplift modeling for clinical trial data. ICDM Workshops, 17–23. https://doi.org/10.1109/ICDMW.2012.103

[22] Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A literature survey and experimental evaluation of uplift modeling: A stepping stone toward prescriptive analytics. Big Data, 6(1), 13–41. https://doi.org/10.1089/big.2017.0104

[23] Zhang, Z., Zhao, P., Li, X., & Liu, Y. (2024). Deep causal models: Taxonomy and roadmap. ACM Computing Surveys, 56(3).

[24] Chen, J., Wang, Y., & Li, X. (2024). A survey of user profiling: State-of-the-art, challenges and solutions. Information Processing & Management, 61(2), 103676. https://doi.org/10.1016/j.ipm.2023.103676

[25] Devriendt, F., Moldovan, D., & Verbeke, W. (2021). Prescriptive analytics through uplift modeling: A review. Information Fusion, 73, 67–86. https://doi.org/10.1016/j.inffus.2021.02.003

[26] Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Representation learning approaches. ICML Proceedings, 3076–3085. https://doi.org/10.48550/arXiv.1606.03976

[27] Shi, C., Blei, D. M., & Veitch, V. (2019). Adapting neural networks for causal inference. arXiv preprint. https://doi.org/10.48550/arXiv.1905.12776

[28] Yoon, J., Jordon, J., & van der Schaar, M. (2018). GANITE: Estimating individualized treatment effects. arXiv preprint. https://doi.org/10.48550/arXiv.1806.04968

[29] Olaya, D., Ponce, H., Gutiérrez-Andrade, M. A., & Martínez-Velázquez, O. (2020). Multi-treatment uplift modeling. KDD Proceedings, 106533. https://doi.org/10.1145/3394486.3403196

[30] Lee, K., & Berger, J. (2024). Cross-treatment gain surface and multi-treatment uplift. Information Sciences, 694, 119240. https://doi.org/10.1016/j.ins.2024.119240

[31] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Statistical Science, 26(1), 1–27. https://doi.org/10.1214/11-STS367

[32] Hahn, P. R., Murray, J. S., & Carvalho, C. (2020). Bayesian regression tree models for causal inference. Bayesian Analysis, 15(3), 965–1056. https://doi.org/10.1214/19-BA1195

[33] Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. arXiv preprint. https://doi.org/10.48550/arXiv.1103.4601

[34] Rehill, J. (2024). A gentle introduction to uplift modelling. arXiv preprint. https://doi.org/10.48550/arXiv.2403.03822

[35] Inoue, T., Yamamoto, K., & Okuno, T. (2024). Machine-learning-based heterogeneous treatment effect estimation in randomized trials: A PRISMA review. Trials, 25(134). https://doi.org/10.1186/s13063-024-07943-0

[36] Ling, C., Sutherland, D., Johansson, F., & Wiens, J. (2023). Causal inference pipelines for RCT emulation. arXiv preprint. https://doi.org/10.48550/arXiv.2302.03070

[37] Maraj, A., Vuković, M., & Hotovec, Ž. (2024). A systematic review of uplift modeling. Information Processing & Management, 61(2), 103692. https://doi.org/10.1016/j.ipm.2023.103692

[38] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score. Biometrika, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

[39] Caron, A., Baio, G., & Manolopoulou, I. (2022). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*(3), 1115–1149. https://doi.org/10.1111/rssa.12824

[40] Hu, N. (2023). Heterogeneous treatment effects analysis for social scientists: A review. *Social Science Research, 109*, 102810. https://doi.org/10.1016/j.ssresearch.2022.102810

[41] Zhang, Z., Zhao, P., Li, X., & Liu, Y. (2021). Causal representation learning. KDD, 3467381. https://doi.org/10.1145/3447548.3467381

[42] Alaa, A. M., & van der Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. IEEE Transactions on Neural Networks and Learning Systems, 29(12), 5547–5561. https://doi.org/10.1109/TNNLS.2018.2817009

[43] Guo, X., Yu, K., Liu, L., Cao, F., & Li, J. (2024). Causal representation learning: A survey. Artificial Intelligence, 320, 104072. https://doi.org/10.1016/j.artint.2024.104072

[44] Zhang, Z., Zhao, P., Li, X., & Liu, Y. (2023). Deep causal models for ITE estimation: A survey. ACM Computing Surveys, 55(12). https://doi.org/10.1145/3527154

## DATA AVAILABILITY STATEMENT

The dataset used in this study is subject to institutional and commercial restrictions and therefore cannot be publicly released at this time. Aggregated statistics and derived experimental results are reported in the manuscript.

The authors plan to release a reproducibility package, including synthetic data examples and representative code implementations, subject to data-sharing approval in future work.

## APPENDIX A. RESPONSE-TYPE LABEL CONSTRUCTION AND CONSISTENCY

Response-type labels in CUP are constructed to reflect both estimated treatment effects and observed behavioral outcomes, ensuring causal interpretability while maintaining temporal stability.

For each deployment period, uplift scores are first estimated using the selected meta-learner configuration. High-confidence uplift thresholds are applied to identify users with strong positive or negative estimated treatment effects. These uplift-based signals are then cross-validated against observed $(T, Y)$ realizations to refine response-type assignments and to prevent logically inconsistent labels.

Specifically, Persuadables and Do-Not-Disturb users are identified through a combination of uplift magnitude and treatment-outcome alignment, while Sure Things and Lost Causes are distinguished based on behavioral invariance with respect to treatment exposure. This two-stage procedure mitigates label noise arising from estimation uncertainty and ensures that response categories remain behaviorally meaningful.

To promote consistency across time windows, response-type definitions are held fixed, while individual users are allowed to transition between response states as their behavior evolves. This design yields stable population-level semantics while preserving individual-level dynamics under repeated interventions.

## APPENDIX B. C2 REPLACEMENT STRATEGY

In the CUP framework, clustering is treated as a flexible and corrective component rather than a hard segmentation step. Clustering is performed at the individual level, while model evaluation and stability assessment are conducted across monthly deployment windows to reflect system-level performance under repeated interventions.

After initial clustering, uplift performance is evaluated at the cluster level and compared against a non-clustered (Direct) baseline using the same monthly evaluation protocol. Cluster-based prediction paths that exhibit unstable or inferior uplift performance relative to the Direct global model are not propagated to downstream response-type assignment.

This comparison gives rise to three evaluation paths:

- **Direct**: uplift estimation without clustering;
- **C1**: uplift estimation within clusters;
- **C2**: cluster-level evaluation followed by fallback to Direct predictions when clustering degrades stability or performance.

Rather than enforcing cluster-specific predictions at the individual level, the C2 strategy operates as a path-level regularization mechanism. It preserves cluster-based heterogeneity when beneficial, while reverting to the global model when clustering introduces noise or instability.

## APPENDIX C. ABLATION ANALYSIS DESIGN

The purpose of the ablation analysis in this study is not to conduct formal hypothesis testing, but to assess the relative contribution and stability of individual components within the CUP pipeline under repeated real-world deployments.

Each ablation experiment removes or modifies one component of the framework while keeping all others fixed. Performance differences are evaluated consistently across six consecutive monthly datasets, allowing assessment of whether observed effects persist over time rather than arising from a single snapshot.

Given the operational nature of the study and the use of large-scale observational data, emphasis is placed on temporal consistency and the magnitude of performance differences rather than formal statistical significance testing. This design provides empirical evidence on whether observed gains are systematic and reproducible under repeated intervention settings.

## APPENDIX D. FEATURE SCREENING AND SCALING STRATEGY

Feature preprocessing in CUP is designed to enhance numerical comparability, stability, and causal relevance prior to model estimation, rather than to optimize predictive accuracy through aggressive normalization or transformation.

Before model fitting, all candidate covariates undergo a unified screening process to ensure consistency across downstream feature configurations. Continuous variables are rescaled to a common numerical range to improve comparability across features with heterogeneous magnitudes and to facilitate stable optimization in subsequent modeling stages.

This rescaling is applied uniformly and does not alter the relative ordering or distributional shape of individual features.

Feature selection proceeds through a diagnostic multi-stage procedure that combines complementary criteria: (i) information-based screening to retain variables with sufficient outcome relevance, (ii) causal importance assessment to identify features consistently associated with treatment effects across time windows, and (iii) stepwise refinement to control redundancy and variance inflation.

Importantly, this screening stage is performed prior to constructing downstream feature-set configurations (e.g., information-based, causal-based, or hybrid sets), ensuring that all reported models draw from a common pool of filtered and stability-checked covariates. This design avoids feature-induced confounding when comparing alternative model specifications and promotes reproducibility under repeated deployment.

## APPENDIX E. COMPUTATIONAL CONSIDERATIONS

The empirical study was conducted using large-scale observational data from a real-world digital lending platform. Due to the involvement of sensitive customer-level financial information, the underlying dataset cannot be publicly released at this stage. Data access was granted under institutional collaboration and confidentiality agreements, and results are reported in aggregated and anonymized form. Subject to future approval and appropriate de-identification protocols, selected data summaries or code components may be made available.

From a computational perspective, the CUP framework is designed to remain tractable for large tabular datasets. The dominant computational costs arise from repeated uplift model estimation, causal feature screening, and clustering-based analyses across multiple monthly deployments. While no additional asymptotic complexity is introduced beyond standard uplift modeling pipelines, the cumulative experimental workload is substantial due to the breadth of model configurations and ablation settings evaluated.

In this study, extensive ablation experiments and stability checks were prioritized to assess robustness under repeated deployment. As a result, certain algorithmic choices—such as the selection of clustering methods and base learners—reflect a deliberate trade-off between methodological coverage and feasible computational execution under single-machine constraints.

Future work will extend the CUP framework to incorporate additional feature selection strategies, uplift estimators, and clustering algorithms as computational resources permit. These extensions are expected to further enrich comparative analysis without altering the core methodological principles established in the present study.

JIANQING JIANG is currently a Ph.D. candidate with the Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia. His research lies at the intersection of user profiling, causal machine learning, uplift modeling, and heterogeneous treatment effect estimation, with a particular focus on dynamic user profiling and personalized intervention design. He has more than seven years of industry experience in data science and business intelligence, holding professional roles in China and Singapore in credit analytics, customer modeling, and enterprise data systems. Prior to his doctoral studies, he worked as a senior data scientist developing credit scoring models, customer segmentation frameworks, and large-scale data governance platforms. His current research integrates causal inference with behavioral modeling to improve decision-making in internet lending and other high-stakes operational environments.

NOR ASILAH WATI ABDUL HAMID is currently the Deputy Director of the Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia. She also serves as an Associate Professor in the Department of Communication Technology and Network, Faculty of Computer Science and Information Technology.

She received her Ph.D. in Computer Science from the University of Adelaide, Australia, in 2008. From 2013 to 2015, she was a Visiting Scholar at the High Performance Computing Lab, The George Washington University, USA. In 2015, she was awarded the CUDA Teaching Centre recognition by NVIDIA and subsequently established the CUDA Lab at her faculty.

Dr. Nor Asilah has authored or co-authored more than 80 journal articles and conference papers. Her research has been supported by both government and industry funding, with interests focused on parallel and distributed high-performance computing, cloud computing, and data-intensive computing. She is currently the Editor-in-Chief of Malaysian Journal of Mathematical Sciences and serves as a reviewer for several well-regarded journals and international conference proceedings.

NG KENG YAP received the B.Sc. and M.Sc. degrees in computer science from Universiti Putra Malaysia, in 2001 and 2005, respectively, and the Ph.D. degree in computer science from The University of Manchester, U.K., in 2015. He is currently a Senior Lecturer with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. He has authored articles in IEEE Access and other indexed journals. He has been involved in multiple research projects, including studies on palm oil production analytics, traffic flow analysis, and disruptive technology in construction project management. His research interests include software components, business analytics, and software engineering for artificial intelligence (SE4AI) systems.

ASSOC. PROF. DR. CHOO WEI CHONG is an Associate Professor at the School of Business and Economics, Universiti Putra Malaysia (UPM). He has earned his Bachelor's degree in Science (Statistics) and Master of Science (Business Statistics) from UPM. He has also completed his PhD and Post-Doctoral in Management Studies/Decision Science at the University of Oxford, United Kingdom. His research focuses on volatility modeling, high-frequency financial data, machine learning–econometrics hybrid forecasting, text-based analytics, and AI applications in healthcare and tourism.