Decision Analytics Journal 17 (2025) 100639

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A machine learning framework for uplift modeling through customer segmentation

Paulo Pinheiro [a], Luís Cavique [b,*]

[a] Universidade Aberta and Cedis, Lisboa, Portugal
[b] Universidade Aberta and Lasige-FCUL, Lisboa, Portugal

ARTICLE INFO

Keywords:
Machine learning
Decision trees
Customer segmentation
Causal inference
Marketing analytics

ABSTRACT

In uplift modeling, the goal is to identify high-value customers based on persuadable customers, those who make a purchase only if contacted. To achieve this, uplift modeling combines machine learning techniques with causal inference, allowing businesses to refine their customer targeting strategies and focus efforts where they are most profitable. This study proposes a practical and reproducible two-phase procedure for identifying high-value customers. In the first phase, customers are segmented using decision trees, which offer a transparent and data-driven approach to grouping individuals with similar characteristics. This segmentation lays the groundwork for a meaningful interpretation of customer behavior. In the second phase, uplift is calculated for each customer segment by comparing the outcomes of the treatment and control groups. This enables the identification of customer groups with the highest uplift. A real-world use case further illustrates the value and applicability of the proposed method. To validate model performance, the procedure employs established metrics such as the Qini index and Cohen's kappa, which provide insights into both the effectiveness and reliability of the uplift estimates. This work presents a decoupled procedure for uplift modeling that leverages well-established libraries, fostering transparency and a clear understanding of the analytical process. A key contribution to uplift modeling and causal inference is the use of decision trees for stratification, which enables the creation of meaningful segments and their evaluation through the average treatment effect. By integrating theory with practical implementation, this work offers a comprehensive framework for uplift modeling that bridges academic rigor and business usability.

1. Introduction

Depending on the context, uplift modeling may also be referred to as treatment effect modeling, persuasion modeling, or incremental response modeling. Uplift modeling is increasingly utilized in various fields, including marketing, healthcare, and education, to enhance decision-making by targeting individuals who are most likely to respond positively to an intervention [1].

In marketing, uplift models help to identify customers who are more likely to purchase when exposed to a particular advertisement, targeting those individuals whose behavior can be positively influenced by the campaign. In healthcare, uplift models can predict which patients are most likely to benefit from a specific treatment, thereby improving patient outcomes and resource allocation by prioritizing treatments for patients who show the highest likelihood of responding effectively and thereby maximizing the impact of healthcare interventions [2].

Uplift identifies the net positive impact of an intervention, ensuring resources target those who can be influenced effectively. Persuadable customers, or compliers, are crucial as they represent the segment and are most likely to respond positively, maximizing investment and strategic impact. Uplift modeling is valuable because it provides more actionable insights than traditional models, enabling organizations to optimize their return on investment by concentrating efforts where they will have the most impact [3].

Recent studies demonstrate the growing importance of statistics and machine learning methods in enhancing strategic analytics across various industries.

Shrestha et al. [4] developed and validated a Customer-Based Brand Equity framework for the smartphone industry through exploratory factor analysis, confirmatory factor analysis, and structural equation modeling. Drawing on specific brand models, this study examines how product features, brand awareness, and perceived quality influence the formation of brand image and preference, ultimately shaping customer loyalty and repurchase intentions. The findings offer strategic insights for managing brand equity in a highly competitive smartphone market.

* Corresponding author.
E-mail addresses: ppinheiro@cedis.pt (P. Pinheiro), luis.cavique@uab.pt (L. Cavique).

https://doi.org/10.1016/j.dajour.2025.100639
Received 17 January 2025; Received in revised form 6 September 2025; Accepted 16 September 2025
Available online 18 September 2025

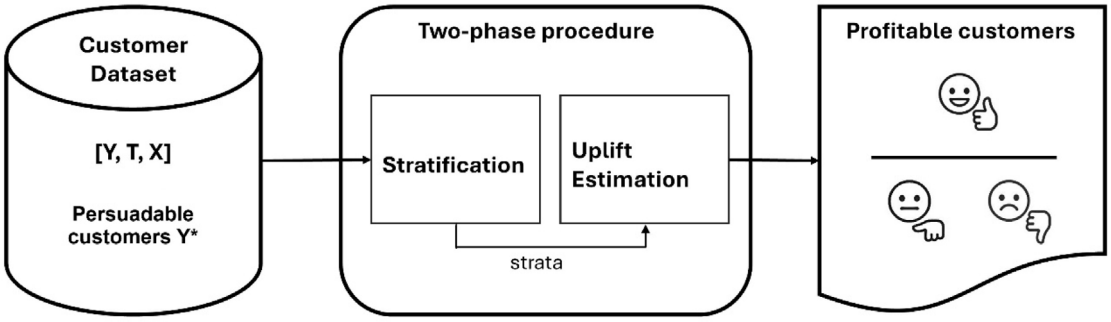

Fig. 1. Two-phase procedure to find profitable customers.

Akhavan and Hassannayebi [5] present a hybrid machine learning framework to predict customer satisfaction by analyzing customer experience and behavior in online services. Using DBSCAN clustering and multi-class decision trees, it addresses challenges like intra-class variance and data imbalance. Tested on real-life data, the model accurately identifies dissatisfied customers, aiding service improvement through predictive process monitoring.

Rajendran et al. [6] employ machine learning to investigate consumer intention and willingness to pay for energy-efficient food products within the food supply chain. The Random Forest algorithm achieved the highest accuracy, highlighting its potential in guiding sustainability strategies. Results support data-driven decisions to enhance energy efficiency and promote the adoption of energy-efficient technologies throughout the food supply chain.

1.1. Problem, objectives, and contributions

A Randomized Controlled Trial (RCT) is an experimental study design where participants are randomly assigned to treatment and control groups. It is considered the gold standard for measuring causal effects because randomization minimizes bias and confounding. Outcomes between groups are compared to estimate the effect, or lift, of an intervention [7].

In traditional lift measurement, participants are also divided into treated and control groups, and the lift is calculated as the difference in response rates between these groups [8]. Uplift modeling is an extension of lift analysis using observational data. In uplift modeling, the customer's decision-making process is explicitly considered: a contacted customer may choose to buy or not buy, and similarly, a non-contacted customer may also decide independently to buy or not buy [9].

The customer characteristics in lift analysis and uplift modeling are also different. In lift analysis, the most profitable customers are those who respond positively to the campaign. In contrast, in uplift modeling, the most profitable customers tend to be the persuadable ones, those who purchase only when contacted [10].

In uplift modeling, a varied nomenclature and a complex combination of prediction methods with causality are employed [11]. Despite various implementations of uplift modeling, challenges remain, including software transparency, the inability to directly identify valuable customers, and a lack of meaningful rules for extracting these customers.

The objective is to create a procedure that is easy to implement in companies, utilizing standard tools to identify meaningful patterns and extract profitable customers. The research questions can be formulated as follows:

Q1: Considering the customers' decision-making process, how can you find the most profitable customers from a dataset?

Q2: How can meaningful patterns/rules be found to extract profitable customers?

This work extends the previous works [12]. Fig. 1 illustrates the workflow of the proposed two-phase procedure for identifying profitable customers, utilizing a straightforward and reproducible approach.

The proposed procedure uses a dataset with a potential outcome variable Y, a treatment/intervention T, and the covariates X. The proposed procedure consists of two stages:

(i) Stratification: Targeting persuadable customers, using a machine learning algorithm, such as logistic regression or decision trees, to create homogeneous customer segments.
(ii) Uplift estimation: The uplift is calculated for each stratum or segment, and customers with the highest uplift are identified as the most profitable.

The main contribution is to answer the research question by identifying profitable customers and patterns that can be extracted from future datasets, thereby enhancing the company's knowledge.

1.2. Organization

The subsequent sections of this document are organized as follows. Section 2 presents the uplift modeling concepts and concerns. Section 3 reviews the work related to causality and uplift modeling. Section 4 proposes a new two-phase procedure to estimate uplift. Sections 5–7 present a use case based on the Telco dataset to illustrate the proposed procedure. Section 5 presents the dataset information and transformations required to execute the use case. Section 6 shows the Telco use case's two-phase procedure for finding profitable customers. Section 7 presents the computational results of the use case, utilizing measures such as Cohen's kappa statistics, Qini index, and cluster purity. Finally, Section 8 offers concluding remarks and summarizes the findings presented in this paper.

This work combines terminology from machine learning and causality. The terms 'segment' and 'stratum' are used interchangeably. Additionally, the terms 'treatment' and 'intervention' are often used interchangeably.

2. Uplift modeling

Uplift modeling involves estimating the incremental effect, or uplift, that a specific treatment will have compared to a control group. This section presents the uplift modeling concepts, concerns, and strategies for addressing the knowledge gap identified by Devriendt et al. [10].

2.1. Lift analysis

In marketing, decile analysis, lift, and gain measures are pivotal in understanding and optimizing customer response rates. These methods categorize customers into ten equal groups based on their response probability, allowing marketers to focus their efforts on the most promising segments. By targeting the top deciles, businesses can achieve higher response rates and maximize their campaign return on investment [8].

Given a dataset with customer responses (e.g., 1 for a purchase and 0 for no purchase), the procedure for decile analysis begins by

running a supervised algorithm, such as logistic regression, to estimate the probability that a customer will make a purchase. Once scores are generated, they are sorted in descending order. Based on these, the dataset is divided into ten equal-sized groups (or deciles). For each decile, the response rate is calculated by dividing the number of responders by the total number of customers. Note that the dataset is not split into training and testing sub-datasets, as is usual in prediction. By using the supervised algorithm, the objective is to score the customers [9].

Lift analysis is performed to measure the improvement in targeting efficiency compared to random selection, using the ratio of cumulative responder percentage to cumulative customer percentage. A gain chart visually assesses how well the top deciles perform relative to the rest. Insights from the analysis guide business decisions, with marketing efforts typically focused on higher deciles, where response rates are most remarkable.

2.2. Uplift modeling concepts

Traditional lift analysis solutions aim to identify customers most likely to respond and prevent churn in marketing. In a different scientific context, a typical randomized controlled trial (RCT) divides the population into a treatment group, which receives the intervention, and a control group, which does not, aiming to assess the causal effect of the intervention. Uplift modeling integrates this causal inference framework with observational data to identify individuals most likely to respond positively to targeted actions. However, unlike RCTs, uplift modeling relies on observational data to evaluate causal effects, aiming to identify the most profitable customers.

Uplift modeling seeks to determine if interventions directly cause customer purchases. However, not all individuals targeted for intervention attended the treatment, and some who were not called showed up to undergo the intervention. Given this second split, every customer is categorized into one of four quadrants, as shown in Fig. 2.

The resulting matrix is based on the customer's purchase decisions, depending on whether a marketing campaign targeted them [11]. The four quadrants include:

- **Persuadable**: Customers are likely to stay only if they receive an intervention. Persuadable customers, also known as compliers in healthcare, buy only if they are exposed to a marketing campaign.
- **Sure Things**: Customers who will stay regardless of the intervention. Secure customers (or sure things) who make purchases, irrespective of whether they are the treatment target.
- **Lost Causes**: Customers who will leave regardless of the intervention. Lost causes represent customers who, regardless of being the treatment target, will not make a purchase.
- **Do-not-disturbs:** Customers who might leave if they receive an intervention. These customers, also classified as defiers or sleeping dogs, should not be treated in a way that would lead to the risk of them becoming dropouts.

Uplift modeling is a problem with numerous practical applications, but its resolution is complex. Devriendt et al. [10] distinguish between 'what we know' and 'what we want to know'.

To clarify some causality concepts, the following notation is used:

- Y∈{1,0} to indicate the potential outcome of the intervention (e.g., buy or respond);
- T∈{1,0} indicates whether an individual is treated or not, respectively, belongs to the treatment group (T = 1, e.g., receive an offer) or control group (T = 0);
- Moreover, X corresponds to the covariates that describe the characteristics of a group of individuals. The three variables constitute the causal inference triple (T, Y, X).

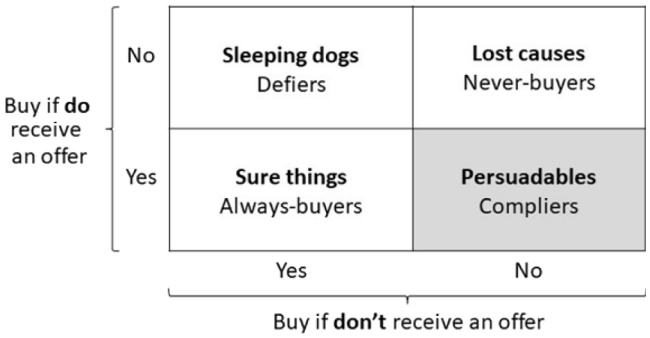

Fig. 2. Conceptual response segments.
Source: Adapted from [11].

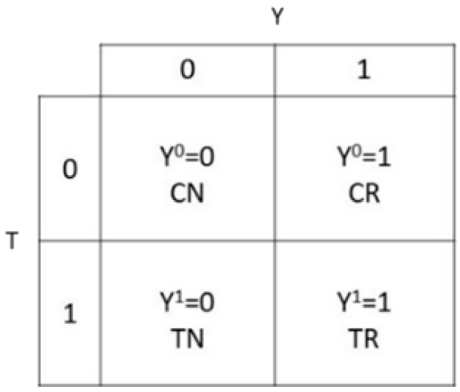

Fig. 3. Conceptual table of 'what we know'.
Source: Adapted from [10]

In causal inference, the effect of the treatment (T) on the potential outcome (Y) can be expressed as the difference between the potential outcomes when the treatment is applied ($Y^{T=1}$ or Y^1) and when it is not applied ($Y^{T=0}$ or Y^0).

Uplift modeling aims to find persuadable customers and not intervene with do-not-disturb ones. However, the data we have available for building models only contains information on whether the customer intervened and whether they responded. Considering Y as the binary potential outcome and T as the binary treatment, in reality, 'what we know' is illustrated in Fig. 3.

- Those assigned to treatment and response (TR): It remains uncertain whether their response is solely due to the intervention, making them potentially persuadable or sure things.
- Those assigned to treatment but do not respond (TN): Their lack of response is undetermined, raising the possibility that they could be lost causes or do-not-disturb.
- Those not assigned to treatment but responding (CR): The nature of their response and whether it would have occurred with intervention are unclear, making them potential sure things or do-not-disturb.
- Those not assigned to treatment and do not respond (CN): It is uncertain if intervention would have prompted a response, categorizing them as potentially persuadable or lost causes.

On the other hand, 'what we want to know' does not match 'what we know'. Fig. 4 presents, in shaded form, the subsets that must be considered for each segment of response customers (persuadable, sure things, lost causes, do-not-disturb).

Firstly, uplift modeling identifies persuadable customers and avoids treating those classified as 'do-not-disturb'. In other words, the goal

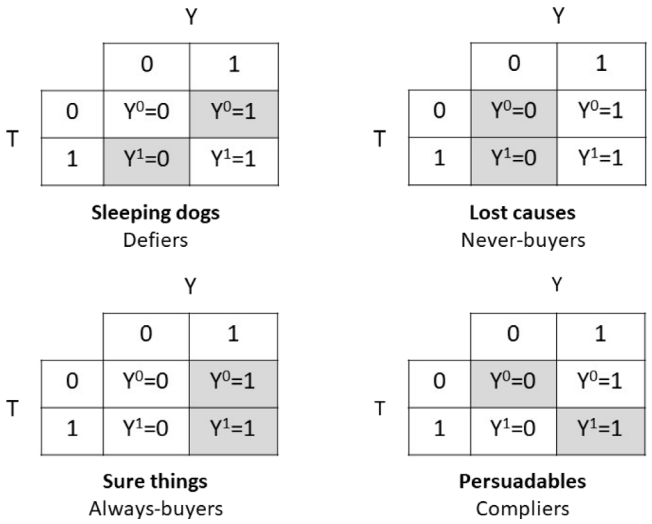

Fig. 4. Conceptual table of 'what we want to know'.
Source: Adapted from [10].

Table 1
Variables T and Y turn into Y-star for the two response segments.

T	Y	$Y^T = \{0,1\}$	Y*	Response segment
1	1	$Y^1 = 1$	1	persuadable
1	0	$Y^1 = 0$	0	defiers
0	1	$Y^0 = 1$	0	defiers
0	0	$Y^0 = 0$	1	persuadable

is to select individuals who respond to a campaign when they are called, and not respond to a campaign when they are not called. These individuals comply with the campaign's objectives. On the other hand, the non-compliers, or defiers, are against the campaign's objectives.

It is not feasible to isolate only persuadable customers, as they are located in two quadrants of the 'what we know' table: those who responded to the intervention (TR or $Y^1 = 1$) and those who did not respond without intervention (CN or $Y^0 = 0$). These two quadrants are also used by the 'lost causes' and 'sure things' segments, but the 'sleeping dogs' segment uses them.

2.3. Transformed Outcome Approach (TOA)

The knowledge gap between 'what we want to know' and 'what we know' reveals a clear discrepancy between desired knowledge and current knowledge.

To overcome the knowledge gap, Jaskowski and Jaroszewicz [13] propose extracting persuadable customers by creating a new variable, Y* (read as 'Y star'). The transformation of the variables Y and T into Y* is known as the Transformed Outcome Approach (TOA). The transformation of the outcome Y* is as follows: Y* = 1 in the cases when (T = 1 and Y = 1) or when (T = 0 and Y = 0), and Y* = 0 in all other cases, as shown by Eq. (1):

$$Y^* = \begin{cases} 1 & if\ Y^1 = 1 \\ 1 & if\ Y^0 = 0 \\ 0 & otherwise \end{cases} \quad (1)$$

Table 1 details the transformation of variables T and Y into Y* for persuadable customers and defiers (or sleeping dogs) response segments. Y* = 1 corresponds to persuadable customers, who are the target of this study. On the other hand, Y* = 0 matches the defiers that should be avoided under any circumstances.

Note that in the TOA transformation, we simplify the initial problem from four response segments to just two response segments. The other response segments, always-buyers, and never-buyers are neglected. In the proposed problem, the key focus is identifying persuadable customers and defiers.

This approach, in addition to being presented by Jaskowski and Jaroszewicz [13], is also explored in the work of Athey and Imbens [14]. It involves transforming the observed outcome Y into a modified outcome Y*, such that the uplift corresponds to the conditional expectation of Y*. As noted in several reviews on uplift modeling, such as [15], [10], and [16], the Transformed Outcome Approach is conceptually simple and often outperforms alternative supervised methods. One of its key advantages is its flexibility: it can be combined with any supervised learning technique to estimate uplift directly. However, these reviews also point out its limitations. Specifically, when dealing with continuous outcomes, the method heavily depends on an accurate estimation of the propensity score. For binary outcomes, achieving balance between the treatment and control groups becomes essential.

For summarizing the meaning of the potential outcomes of Y and Y-star given by TOA:

- $Y \in \{1,0\}$ is the potential outcome of the intervention, where Y = 1 is a customer who buys or responds to the contact, and Y = 0 is a non-responder customer;
- $Y^* \in \{1,0\}$ is the new potential outcome, where Y* = 1 is a persuadable customer and Y* = 0 is a defier or sleeping dog.

3. Related work

This section presents concepts of causality, uplift estimation, and a literature review of uplift modeling, highlighting some gaps in the field of uplift modeling.

3.1. Causality concepts

A central question in the field of causal inference is clarifying the relationship between variables, specifically, how the value of one variable, the treatment T, influences the outcome Y. The primary focus should be on understanding the significance and extent of this impact.

With the potential outcome framework [17], each subject i can potentially present two outcomes where Y_i^1 represents the potential outcome if the subject received the treatment T and Y_i^0 represents the potential outcome if the subject did not receive the same treatment. The effect of treatment T regarding subject i, known as the individual treatment effect (ITE), is then defined by Eq. (2):

$$ITE_i = Y_i^1 - Y_i^0 \quad (2)$$

However, in the real world, it is only possible to observe one of the potential results since an individual can be or cannot be, at the same time, the target of treatment. We thus have, for subject i, an observed potential result and a potential result that is not observed as counterfactual, resulting in the Causal Inference Fundamental Problem [18–20]. The Fundamental Problem of Causal Inference states that we cannot observe all potential outcomes for the same unit simultaneously. For instance, we cannot see a patient's health with and without treatment. One observation is factual, and the other is counterfactual. Counterfactuals are hypothetical scenarios used to understand what could have happened if a different action had been taken.

To overcome the Fundamental Problem of Causal Inference, researchers estimate average treatment effects (ATE) over a sample or subpopulation. Since individual treatment effects cannot be observed, the Conditional Independence Assumption (CIA) allows for identifying the ATE by assuming treatment assignment is independent of potential outcomes, given observed covariates, as shown by Eq. (3):

$$ATE = E[Y^1 - Y^0] \quad (3)$$

Stratification offers an extra benefit when managing multiple groups. Given the Average Treatment Effect (ATE) concept, a treatment that

works on average may still be ineffective or harmful for specific subgroups, necessitating the need for differentiated policies. Identifying sources of heterogeneity can lead to a deeper understanding of the underlying mechanisms. Heterogeneous treatment effects (HTE) refer to the idea that the causal effect of a treatment may not be uniform across all individuals in a population. In other words, different individuals might experience different effects from the same treatment. Therefore, the population should be divided into distinct strata for more tailored analysis and intervention to address HTE. The conditional average treatment effect (CATE) is represented by Eq. (4):

$$CATE(x) = \mathrm{E}\left[Y^1 - Y^0 | X = x\right] \quad (4)$$

Zhang et al. [16] noted that since uplift modeling techniques assume data are obtained from experiments with randomized treatment assignments or observational data, the uplift calculation is equivalent to calculating the causal average treatment effect.

3.2. Causal inference approaches

A Randomized Controlled Trial (RCT) is widely regarded as the gold standard for establishing causal relationships; however, it also carries significant drawbacks and limitations. RCTs are often costly, time-consuming, and resource-intensive, making them impractical in many real-world contexts. Ethical challenges may also arise, especially when withholding potentially beneficial treatments from a control group is not acceptable. An alternative approach is to rely on observational data, which is information already collected and owned by organizations, providing valuable insights without the constraints of experimental design.

To draw causal claims from observational data, economists employ natural experiments, utilizing approaches such as regression discontinuity, difference-in-differences, and methods that incorporate additional variables, including instrumental variables [21]. These quasi-experimental approaches, such as natural experiments and related econometric methods, provide credible alternatives for identifying causal effects.

Other approaches are more common in areas such as machine learning and biostatistics. In machine learning, causal inference often relies on counterfactual reasoning and employs meta-learner algorithms, such as the T-learner, S-learner, or X-learner [22]. These algorithms enable the estimation of heterogeneous treatment effects through counterfactual reasoning, complementing traditional econometric approaches. An extended discussion of these methods can be found in [23].

On the other hand, biostatistics employs propensity score methods, particularly matching [24], as well as stratification. The matching technique helps create a comparison pair of units approximating the counterfactual by matching treated and untreated units with similar characteristics. Another strategy, like matching, involves stratifying or subclassifying the population, creating strata with similar characteristics. Matching and stratification adhere to the concept of Ceteris Paribus (all other things being equal), as referred to in causal inference [25,26]. Defining pairs or strata helps approximate the condition where all other things are equal.

Matching and stratification are standard techniques for reducing bias in observational studies. Matching pairs treated and control units with similar covariates, aiming to mimic randomization. Stratification divides the sample into subgroups based on covariates, ensuring comparisons are made within more homogeneous groups. Austin [27], citing [28], demonstrated that dividing subjects into quintiles based on the propensity score can eliminate approximately 90% of the bias from measured confounders when estimating a linear treatment effect.

In this work, instead of applying the usual machine learning approach to counterfactuals, we employ decision trees for stratification.

3.3. Uplift modeling estimation

In causal inference, the average treatment effect is given by ATE = E [Y^1-Y^0], assuming that the customer dataset is divided into two groups. Conditioning X = x, the conditional average treatment effect is given by CATE(x) = E[Y^1-Y^0| X = x]. Similarly, the uplift represents the increase in response rates resulting from the campaign, calculated as the difference in response rates between the two groups. Formally, uplift is expressed as a difference of probabilities, as shown by Eq. (5):

$$Uplift(x) = Pr(Y^1 = 1|X = x) - Pr(Y^0 = 1|X = x) \quad (5)$$

The estimation of the uplift using the transformed target variable Y* proposed by Jaskowski and Jaroszewicz [13], where modeling the conditional uplift of Y is equivalent to modeling the conditional distribution of the new variable Y*, as shown by Eq. (6):

$$Uplift^{\mathrm{JJ}}(x) = Pr(Y^1 = 1|x) - Pr(Y^0 = 1|x) = 2.Pr(Y * = 1|x) - 1 \quad (6)$$

Alternatively, using the 'what we know' classification, which includes the CR, CN, TR, or TN groups, the estimation of the uplift proposed by Kane et al. [29], shown by Eq. (7):

$$Uplift^{\mathrm{Kane}}(x) = [Pr(TR|x) + Pr(CN|x)] - [Pr(TN|x) + Pr(CR|x)] \quad (7)$$

Instead of using differences, ratios can also be used. Relative uplift is a metric in uplift modeling that quantifies the proportional increase in response rate due to treatment compared to a control group, as shown by Eq. (8):

$$Uplift^{\mathrm{relative}}(x) = P(Y^1 = 1 \mid X = x) \div P(Y^0 = 1 \mid X = x) \quad (8)$$

Jaroszewicz [30] provides an encyclopedic view of specialized models, such as uplift decision trees and ensembles, for estimating treatment effects, as well as evaluation methods like Qini curves.

3.4. Uplift modeling literature review

The literature review presents various approaches for structuring uplift modeling. This work presents the contributions of Gutierrez and Gérardy [15], Devriendt et al. [10], Gubela et al. [2], Olaya et al. [1], and Zhang et al. [16].

Gutierrez and Gérardy [15] review uplift modeling techniques used to estimate the causal impact of interventions on customer outcomes, merging machine learning with causal inference. The paper categorizes uplift methods into three fundamental approaches: (i) the Two-Model Approach [31], which creates separate models for treatment and control groups; (ii) the Class Transformation Approach [13], transforming the outcome variable to simplify uplift estimation, and (iii) Direct Uplift Modeling using tree-based methods [14], modifying models like decision trees to infer treatment effects directly. Traditional uplift metrics, such as decile and cumulative gain charts, are used in conjunction with Qini curves [31] to evaluate model performance.

Devriendt et al. [10] provide a comprehensive review of uplift modeling, a method central to prescriptive analytics that estimates the causal impact of an action, such as a marketing campaign, on customer behavior. They classify customers as persuadable, sure things, lost causes, and do-not-disturb [11,29], targeting only those likely to respond positively to treatment. Uplift modeling approaches reviewed include transformation techniques that modify the target variable and direct modeling methods, such as random forests, which consistently perform well across datasets. Evaluation methods, such as Qini curves [31], help assess model performance but fall short of intuitively aligning with business goals, highlighting a gap in actionable metrics. The authors emphasize the instability and sensitivity of uplift models to dataset characteristics, highlighting opportunities to enhance their stability and practicality for real-world business applications. A new metric, promoted cumulative gain, is introduced and optimized, showing competitive or superior performance to standard uplift modeling methods [32].

Gubela et al. [2] benchmark uplift modeling strategies in e-commerce, comparing methods such as the Two-Model Approach [31], interaction-based models [33], and class transformation [13]

across algorithms, including random forests and SVMs. They find that random forests, paired with suitable uplift strategies, often yield strong performance, especially in real-time targeting. The study evaluates business impact using Qini curves and incremental purchase metrics, showing that the Two-Model and interaction methods effectively boost incremental conversions.

Olaya et al. [1] examine multi-treatment uplift modeling, extending traditional single-treatment approaches to scenarios with multiple interventions. They introduced two new techniques and benchmarked these against existing models using marketing, healthcare, and public policy datasets. Leveraging frameworks like propensity score matching for bias reduction, they assess models with metrics such as Qini curves and expected response [34].

Zhang et al. [16] unify treatment effect heterogeneity modeling and uplift modeling, which are used to estimate causal impacts in healthcare and marketing. They review vital methods, such as the Two-Model Approach and Causal Forests, and evaluation tools like Qini curves. Highlighting that uplift models rely on randomized experiments, while heterogeneity models can also handle observational data with certain assumptions, they establish a common framework to integrate these methods.

In conclusion, the reviewed studies collectively advance the field of uplift modeling. Gutierrez and Gérardy [15] and Devriendt et al. [10] highlight foundational methods and customer segmentation techniques, while Gubela et al. [2] demonstrate real-time applications in e-commerce. Olaya et al. [1] introduce multi-treatment frameworks, and [16] bridge treatment effect heterogeneity with uplift modeling under a unified approach.

4. Proposed model

Despite the multiple methods, we believe there is a lack of an integrated framework that can combine the TOA with customer segmentation, ultimately finalizing with uplift estimation, which could be easily implemented in companies using standard machine learning packages.

To improve interpretability, the integrated framework could be developed in two phases: segmentation using machine learning, followed by uplift estimation. The two-phase procedure belongs to the category of sequential systems and avoids the complexity associated with coupled systems, a characteristic effectively addressed by methodologies such as Axiomatic Design [35,36] and Design Structure Matrix [37], both of which focus on the design and analysis of complex systems. Axiomatic Design emphasizes minimizing complexity by avoiding coupled systems, where components interact in a manner that creates design challenges. Uncoupled systems are preferred as they ensure simplicity and independence between elements. Sequential systems are less ideal than uncoupled systems, but are acceptable if dependencies follow a predictable order. The two-phase approach offers a comprehensive and decoupled process for uplift modeling that is practical for companies' implementation. At each phase, the data can be reexamined and validated, where decision trees are preferred because they are intuitive and easy to interpret.

The framework prioritizes transparency, customization, and minimal dependency by relying on self-written code built on standard libraries. This approach ensures complete visibility into data handling and analysis, avoiding 'black box' behavior, and is especially valuable in academic and regulatory settings where rigor, accountability, and reproducibility are crucial.

In this study, we employ machine learning for stratification, utilizing decision trees to form segments rather than the conventional approach of addressing counterfactuals by imputing missing values. To the best of our knowledge, this represents a novel application within causal inference. However, the use of decision trees for segmentation has been previously examined, for example, by Venkatasubramaniam et al. [38]. Decision trees are effective for identifying homogeneous subgroups defined by combinations of individual characteristics, where each path from the root to a leaf represents a segmentation rule. They enhance simplicity and interpretability by providing clear rules that define each segment, highlighting the most important variables for segmentation, and offering visualizations that can be easily explained to decision makers.

Again, in this study, we propose the new estimator SATE, the stratum average treatment effect. SATE differs from CATE, as CATE usually includes a short condition, e.g., (internetService = 1). In contrast, SATE supports a rule extracted by the decision trees, e.g., (tenure = [34, 46] and streamingMovies = 0 and internetService = 0).

This work aims to provide a comprehensive and decoupled procedure for uplift modeling that is easy to implement in companies using machine learning. The reason for using decision trees is that they are often considered self-explanatory. Moreover, to the best of our knowledge, literature does not provide a straightforward framework for estimating uplift modeling using classic decision trees for stratification purposes.

4.1. The two-phase procedure

Uplift modeling is a particular case of a causal inference method that enables researchers to estimate the impact of an intervention. Stratified data plays a critical role in mitigating bias and confounding by dividing the population into homogeneous groups, thereby ensuring that subgroup-specific characteristics are adequately represented in the analysis.

Fig. 5 Shows the Direct Acyclic Graph (DAG) using the original observational data (i) and the stratified data, as everything is equal within each stratum (ii). In Fig. 5 (ii), eliminating the direct path between treatment T and a covariate variable X mitigates the influence of confounding factors.

This work proposes a comprehensive two-phase uplift modeling procedure utilizing machine learning, with a focus on decision tree algorithms for stratification.

To implement the two-phase procedure, the input dataset must identify the potential outcome Y, the treatment T, and the covariates X, collectively referred to as the triple (Y, T, X). The objective of uplift modeling is to identify profitable customers. The previous step transforms outcome Y into a new variable Y*, as defined by the Transformed Outcome Approach (TOA), to identify persuadable customers, those who respond to a campaign, if and only if they receive an offer.

In phase 1, a machine learning model is applied. The model groups customers into distinct strata that share similar characteristics. In phase 2, the uplift effect is calculated for each stratum. The uplift represents the estimated impact of the treatment within each segment. Finally, the procedure returns profitable customers with an uplift greater than zero, indicating an adequate response.

Previously, in the two-phase procedure, phase 0 (Fig. 6) was calculated for each observation, Y*, as a function of the outcome Y and the treatment T, i.e., Y* = f0(Y, T), and used as the target attribute for the supervised algorithm.

4.2. Phase 1: Stratification

In the first phase, the stratification utilizes a machine learning approach that operates at the individual level. According to Y* and covariates X, Y* = f1 (X), a machine learning algorithm is applied to the available data. The machine learning algorithm uses Y* as the target attribute. The objective is to find a model that allows us to separate TR and CN customers: those who respond when called for an intervention and those who do not respond when they are not called for an intervention (Y* = 1).

Different types of models stratify data in distinct ways. Depending on the algorithm, there are two main models for stratifying data. Regression-based algorithms return a numeric score for each individual

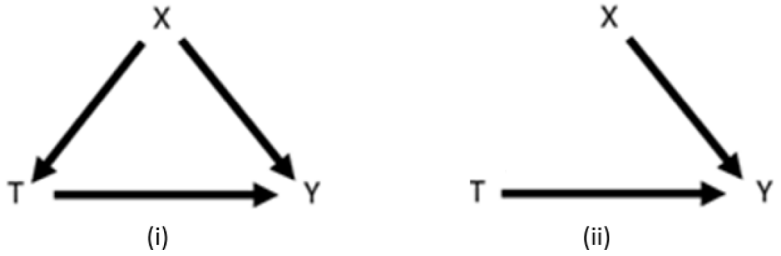

Fig. 5. (i) Observational data and (ii) stratified data.

Two-phase Procedure:
Input: dataset (Y, T, X)
Output: profitable customers
0. variable transformation, TOA: Y* = f0(Y, T)
1. stratification:
- machine learning model Y* = f1(X)
- the model generates strata/segments, S
2. estimation:
- for each stratum estimates Uplift(s) = f2(Y, T, S=s)
Return: the profitable customers with Uplift > 0

Fig. 6. Two-phase Procedure.

or group, while Classification-based algorithms assign individuals to distinct groups based on specific rules. The scoring models, such as logistic regression and linear discriminant analysis (LDA), assign scores to observations, and partitioning models, like decision trees, divide the input space into distinct groups.

The scoring models retrieve individual scores. Uniform strata are created after the observations are sorted by their score and divided into the number of groups defined by the analyst. For example, if (0.428 < Score ≤ 0.836), then stratum = 2. On the other hand, rules can be extracted from partitioning models. Each rule corresponds to a uniform stratum given by the algorithm. For example, if (Tenure>13.0 and Charges = 67.5), then stratum = 2.

Table 2 illustrates that two types of strata can be identified, depending on the algorithm type (scoring or partitioning), in terms of the number of strata, the number of elements in each stratum, and their impact on the charts. When the number of elements in each stratum is equal, the strata are called iso-cardinality strata. When the number of elements in each stratum differs, they are called hetero-cardinality strata.

Scoring models create iso-cardinality strata; the user can define the number of strata, and the resulting bar chart has the same bin size. Otherwise, partitioning models create hetero-cardinality strata; the algorithm generates the number of strata, and the resulting bar chart has a variable bin size. In the rule-based approach, the hyperparameter tuning is essentially related to the number of strata.

As in lift analysis, the dataset is not split into train and test subdatasets, as is usual in predictive modeling. The goal here is to use the supervised algorithm to score customers and categorize them into strata based on scoring methodologies, or to partition customers using the information of the leaves of the decision trees.

Table 2
Two types of strata.

	Score-based	Rule-based
Algorithm	Regression-based: Log regression, Linear discriminant analysis	Classification-based or Partitioning-based: Decision trees
# Strata	Defined by the analyst (ex: 10)	Defined by the algorithm (ex: leaves of the decision tree)
# Elements	Equal # elements in each stratum	Different # elements in each stratum
Cardinality	Iso-cardinality	Hetero-cardinality
Bar char	Bar chart with the same bin size	Bar chart with variable bin size

4.3. Phase 2: Uplift estimation

In this second phase, variable Y is used to calculate the uplift. The reason for using Y* in stratification and Y for calculating uplift is as follows: Y* = 1 indicates that the customer is persuadable, while Y* = 0 indicates the customer is a defier. While Y = 1 denotes that the customer makes a purchase, Y = 0 means the customer does not make a purchase. Thus, Y* classifies customers by their responsiveness (persuadable vs. defier) to identify segments for targeted interventions, whereas Y directly measures actual purchasing behavior to quantify the uplift effect.

Given the generation of the set of strata S, the uplift can be calculated for each stratum. The estimation includes treatment T and the outcome Y, Uplift(s) = f2(Y, T, S = s). In a marketing campaign, the uplift of a stratum 's' is the difference in the response rate between two groups, the treated group Y^1 and the control group Y^0. In other words, uplift or SATE is the difference of the conditional probabilities given a stratum S = s, as shown in Eq. (9):

$$SATE(s) = Uplift(s) = Pr(Y^1 = 1 \mid S = s) - Pr(Y^0 = 1 \mid S = s) \quad (9)$$

Where:

- Pr (Y^1 = 1 | S = s) is the probability of a positive outcome in the treated group within stratum s.
- Pr (Y^0 = 1 | S = s) is the probability of a positive outcome in the control group within stratum s.

In the rule-based approach, each stratum is defined by a rule extracted from the corresponding branch of the decision tree. The SATE of each

Table 3
Available attributes in the Telco dataset.

Attribute names	
Churn (target)	Contract
Dependents	DeviceProtection
Gender	InternetService
MultipleLines	OnlineBackup
OnlineSecurity	PaperlessBilling
Partner	PaymentMethod
PhoneService	SeniorCitizen
StreamingMovies	StreamingTV
TechSupport	Tenure

Table 4
Attributes *Churn* and *Responder*.

Attribute ↓ Class value →	No	Yes
Churn	5174	1869
Responder	1869	5174

stratum differs from the CATE that is usually given by an ad hoc condition. The SATE corresponds to the effect of a clear rule that defines each segment using the most important variables, offering explainability to decision-makers.

The treated group receives the marketing campaign (e.g., special offer email), while the control group does not receive the campaign. The uplift is simply the difference between the probabilities of the treated and control groups. If the uplift of the stratum is positive, the campaign increases the probability of purchase. On the other hand, if it is negative, the campaign decreases the purchase probability.

Finally, profitable customers are found in the strata that show a positive uplift, i.e., where the treatment is worthwhile. As a result, customers who belong to a stratum with an uplift greater than zero are selected for the next campaign.

5. Dataset transformation

We intend to create a use case to illustrate the proposed model. In this case, we will first define the dataset and then apply the two-phase procedure of segmentation and uplift estimation. This section presents the dataset information and the transformations needed to perform the running example of the following sections.

Our choice fell on a public dataset, as public datasets offer accessible, standardized data that promotes transparency, reproducibility, and cost-effectiveness while reducing privacy concerns. This work utilizes the Telco public dataset [39], which comprises 7,043 customer records. In addition to the customer ID, this dataset includes 18 other covariates that may be related to the target variable, *Churn*. Table 3 presents the available attributes by name.

5.1. Potential outcome (Y) and treatment (T)

Some variable transformations are carried out to fit the variables in our problem adequately. The variables *Churn* and *Contract* are subject to transformation.

The binary attribute *Churn* presents the customer's status regarding churn or non-churn. To adequately address our problem, a new binary attribute, *Responder*, was created, which inverts the values of the *Churn* attribute. Table 4 presents the number of records by class for each attribute.

To simplify the problem, a new attribute, *ContractBin*, combines the contracts for one and two years in the same class, as shown in Table 5. Binary variables improve model interpretability and computation efficiency. Moreover, in this case, the attribute gets more balanced.

Table 5
ContractBin attribute.

ContractBin values	# observations
Month-to-month	3875
Long term (One year + Two year)	3168

Table 6
Attributes' correlation with the Responder and their actionability.

Attributes	Correlation	Actionable
Responder	1.000	...
ContractBin	0.405	yes
Tenure	0.352	no
InternetService	0.228	yes
TotalCharges	0.199	no
MonthlyCharges	0.193	yes
OnlineSecurity	0.171	yes

5.2. Direct Acyclic Graphs (DAG)

A causal model, represented by the Direct Acyclic Graphs (DAG), can be constructed either by experts with domain knowledge or through automated procedures known as causal discovery [40], such as the PC algorithm [41].

Expert domain knowledge implies the concept of actionable knowledge supported by actionable attributes. As emphasized by Cao [42], what organizations truly need is actionable knowledge, insights that not only describe what is happening but also guide specific actions to achieve desired outcomes. Actionable knowledge bridges this gap by producing insights that are directly tied to operational strategies. It involves actionable attributes, elements within the data that can be influenced or modified to produce favorable outcomes. These attributes, when linked to behavioral rules, empower decision-makers to implement precise and measurable actions, enhancing the overall utility of data mining efforts.

An *actionable attribute* is a data attribute that can be modified or influenced through specific interventions or decisions, with the potential to impact outcomes. Unlike non-actionable attributes, such as age or gender, which are fixed and cannot be changed, actionable attributes, like type of contract or pricing plans, can be adjusted by an organization to drive desired changes, such as improving customer retention or increasing satisfaction.

In this approach, we use human domain knowledge to define causal assumptions. First, the correlation of the attributes is ranked, and then the attributes are classified as actionable or non-actionable.

Table 6 presents the correlations between the attributes and the *Responder*, along with their actionability. *ContractBin* stands out as the actionable attribute with the highest correlation. In contrast, *Tenure* cannot be selected since it is a non-actionable attribute, analogous to age, and therefore not subject to change.

Direct Acyclic Graphs (DAGs) are valuable tools for representing causality. They delineate the causal assumptions inherent in each study [43]. The nodes within the graph correspond to variables (potential outcome Y, treatment T, and covariates X), and the arrows denote the potential associations between these nodes. The data description and DAG must precede the modeling process.

Regarding the choice of the potential outcome Y of the triple (Y, T, X), the variable "Responder" is selected. To choose the treatment T, we need to find an actionable variable with a high correlation with Y. The variable *ContractBin* shows a significant correlation with *Responder* and aligns well with marketing logic, as a long-term contract indicates a higher level of engagement with the service. The DAG obtained is illustrated in Fig. 7.

The final DAG incorporates not only correlation analysis but also domain-specific business knowledge. The causal assumptions rely on both factors. A weak correlation without domain knowledge or a strong correlation lacking domain knowledge should not be relied upon.

Table 7
Y* attribute.

Customer type	Rule	Y*	
		Class	# obs.
TR CN	ContractBin = Long term & Responder = Yes ContractBin = Month-to-month & Responder = No	1	4609
CR TN	ContractBin = Month-to-month & Responder = Yes ContractBin = Long term & Responder = No	0	2434

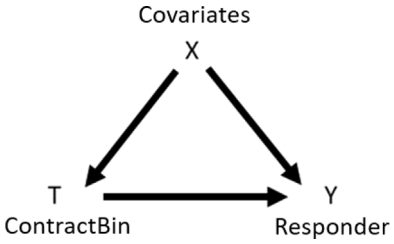

Fig. 7. DAG for Contract treatment.

5.3. Transformed outcome approach

Given the binary variables Y and T, the Transformed Outcome Approach (TOA) is applied, and the Y* attribute is created, following the equation Eq. (1) proposed by Jaskowski and Jaroszewicz [13].

Table 7 shows that Y* = 1 corresponds to customers who are considered TR (assigned to treatment and with a response) and CN (not assigned to treatment and with no response). On the other hand, Y* = 0 corresponds to customers considered CR (not assigned for treatment and with response) and TN (assigned for treatment and with no response).

6. Running example

This section presents a running example of the two-phase procedure for uplift modeling, which aims to identify high-value customers. The Telco dataset is transformed, and Y*, given by Y* = f0(Y, T), is calculated in the previous section. The R system, ver. x64 4.0.1, RStudio ver. 1.3.959, and related packages are used in the running example. The following subsections describe the application of the two phases of the procedure.

6.1. Phase 1: Stratification

In the first phase, machine learning algorithms are used to stratify the customers. The Y* attribute is the target attribute of the supervised algorithms. The remaining available attributes can be used as covariates, X. These covariates are input into the data mining algorithms, as previously defined, Y* = f1(X).

Models stratify data in different ways, falling into two categories. Scoring models assign scores to observations, such as logistic regression (LogReg) and linear discriminant analysis (LDA). In contrast, rule-based models, like decision trees, divide the input space into distinct strata.

For the first group of algorithms, the scoring algorithms, we used LDA from the R package 'MASS' [44] and LogReg from the R Package 'stats' [45]. LDA utilizes linear combinations of variables to predict the class of a given observation. LogReg uses a logistic function to transform the linear combination of variables into a probability value between 0 and 1, which identifies the class of an observation. Both algorithms score observations according to the likelihood of belonging to a target class.

For the second group of algorithms, the rule-based models, we consider using the C5.0 from the R package 'C50' [46] and CTree from the R package 'partykit' [47]. Both are rule-based algorithms based on decision trees. Each internal node in the tree signifies a decision based on the value of a specific input variable, while each leaf node represents a conclusive decision. Each leaf in the resulting tree represents a stratum that groups all observations that obey the rule that creates the leaf. The code area available in GitHub is at https://github.com/lcavique/uplift.

Table 8
Accuracy of applied models.

Model	Algorithm type	Accuracy
Lift Analysis	score-based	0.805
Linear Discriminant Analysis (LDA)	score-based	0.629
Logistic Regression (LogReg)	score-based	0.631
Decision tree C5.0	rule-based	0.700
Decision tree CTree	rule-based	0.661

Table 8 shows the results for each model's accuracy, including the lift model. The lift model uses LogReg in a configuration targeting Y to compare the proposed uplift model with a traditional predictive model.

Scoring algorithms create strata with the same number of elements or iso-cardinality segments. The scores obtained with the LDA and LogReg algorithms should be sorted and divided into several segments that the analyst defines. In our case, twelve groups were used to get the iso-cardinality solution.

Rule-based algorithms generate strata with a different number of elements or hetero-cardinality since each decision tree creates leaves with a diverse number of observations. In the hetero-cardinality segmenting solution, decision tree algorithms like C5.0 and CTree produce trees where we can observe the rules defining the strata formed by each tree leaf. Since the observations are already segmented into each tree leaf, it is only necessary to calculate the uplift for each leaf.

The charts in Fig. 8 show the uplift obtained by two scoring algorithms, LDA and LogReg, resulting in iso-cardinality charts. The user defines the 12 segments or strata. Each stratum has the same number of elements, where iso cardinal means the same number. With the same number of elements by stratum, the corresponding bar chart has the same bin size.

Using rule-based algorithms, decision tree algorithms like C5.0 and CTree produce trees where we can observe the rules defining the segments formed by each tree leaf. With these rule-based algorithms, since the observations were already segmented into each tree leaf, it was only necessary to calculate the Uplift for each leaf. Fig. 9 presents the application of the CTree algorithm.

Fig. 10 shows the uplift obtained by two rule-based models, C5.0 and CTree, resulting in hetero-cardinality charts. The decision trees define the 18th and 19th segments/strata, respectively. Each node of the decision tree includes a different number of elements. So, each stratum has a different number of elements, where hetero-cardinal means the different number. With the varying numbers of elements by stratum, the corresponding bar chart has a variable bin size.

Comparing Figs. 8 and 10 on segment cardinality reveals that segments from scoring models exhibit uniform cardinality (iso-cardinality), while those from rule-based models show variable cardinality (hetero-cardinality).

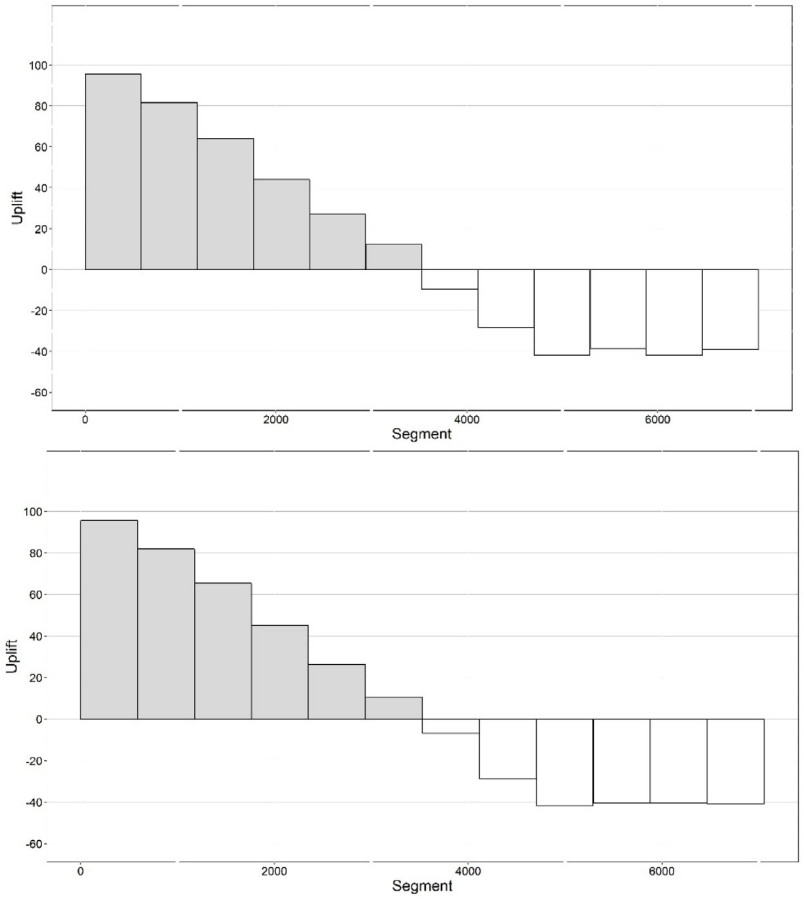

Fig. 8. Charts obtained by scoring algorithms LDA and LogReg.

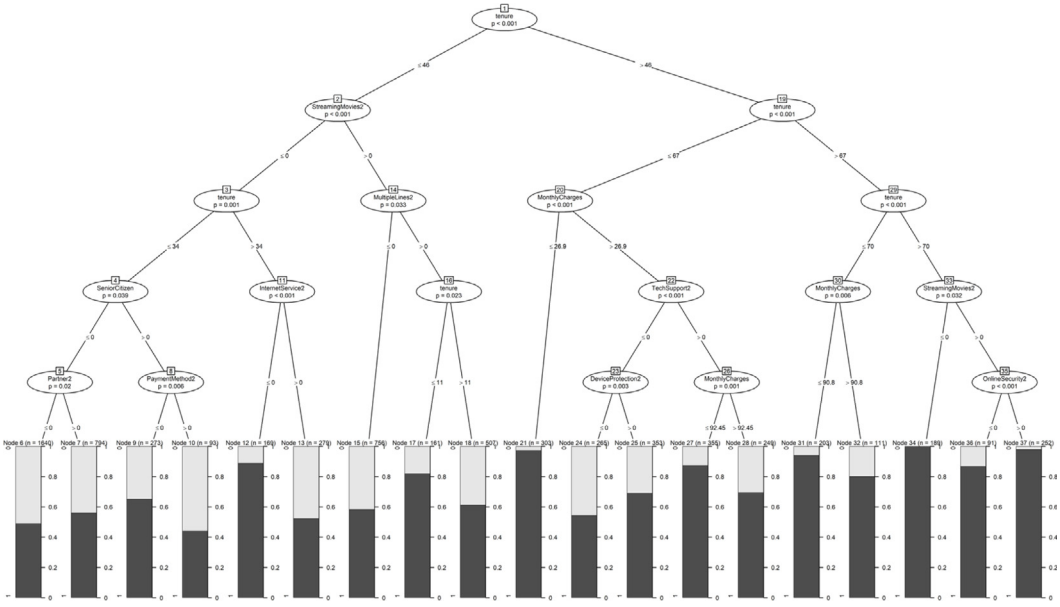

Fig. 9. Tree generated by the CTree algorithm.

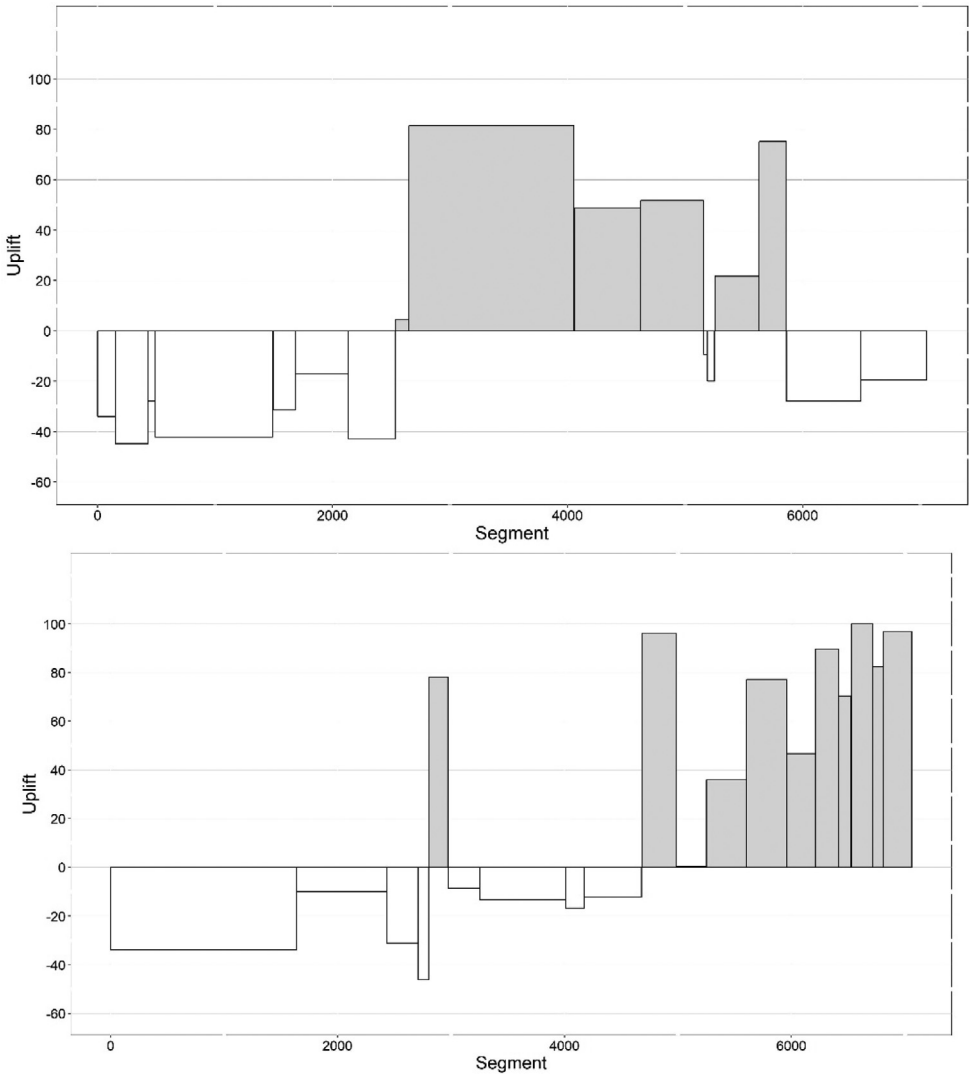

Fig. 10. Charts obtained by rule-based algorithms C5.0 and CTree.

6.2. Phase 2: Uplift estimation

After generating the set of strata S, we can calculate the uplift for each stratum. This calculation considers both the treatment T and the outcome Y, as expressed by the second procedure phase, Uplift(s) = f2(Y, T, S = s). Within the context of a marketing campaign (or different contracts), the uplift for a specific stratum s represents the difference between the response rates of the treatment group (Y^1) and the control group (Y^0). Stated differently, uplift measures the difference between conditional probabilities within a particular stratum S = s. Formally, the uplift can be expressed as follows, Eq. (10):

$$Uplift(s) = Pr(Y^1 = 1 \mid S = s) - Pr(Y^0 = 1 \mid S = s) \quad (10)$$

With the observations grouped into iso-cardinality or hetero-cardinality strata, the selection of profitable customers is determined by those with an uplift greater than zero.

In the case of the LDA and LogReg models with iso-cardinality segments, the profitable customers are those found in the segments from 1 to 6. In the case of the C5.0 model with hetero-cardinality segments, the profitable customers are those found in the leaves numbered 8 to 12, 14, and 15 of the decision tree. Finally, in the CTree model, the profitable customers are found in leaves 5 and 10 to 19 of the decision tree. Table 9 presents the number of profitable customers for each model. Scoring models display the same number of elements, while rule-based models present a diverse range.

Table 9
Number of observations in groups with uplift > 0.

Model	# Profitable customers
Lift Analysis	2935
Linear Discriminant Analysis (LDA)	3522
Logistic Regression (LogReg)	3522
Decision tree C5.0	3229
Decision tree CTree	2540

Finally, once the model's functioning is understood, it is essential to consider how it applies to new datasets. Given that the two types of models return diverse solutions, different approaches should be considered.

In the score models with the iso-cardinality solutions, each stratum with a positive uplift includes a set of observations that present a score given by the machine learning model. Since a previous order of this score was used for creating groups, we identified the observation with the lowest positive uplift. This score will serve as the minimum threshold for selecting new customers. As exemplified with the LDA model shown in Fig. 8, the limit should be obtained from the observation with the lowest score in group 6.

Table 10
Cohen's kappa results.

Model	Lift	Score-based models		Rule-based models	
		LDA	LogReg	C5.0	CTree
Lift	1.00	0.56	0.56	0.63	0.55
LDA		1.00	0.99	0.80	0.71
LogReg			1.00	0.80	0.71
C5.0				1.00	0.72
CTree					1.00

In rule-based models with heterogeneous cardinality solutions, the selection of new customers to be treated is determined by the rules defining the tree leaves with an uplift greater than zero. Exemplifying this with the CTree model and the generated tree shown in Fig. 9, the rule obtained for group 5 is (tenure = [34, 46] and streamingMovies = 0 and internetService = 0).

7. Models' performance

This section examines and compares the performance of the models, using Cohen's kappa statistics to assess whether different models identify the same profitable customers, the Qini index to evaluate the performance of each model, and purity to measure the homogeneity of the strata.

7.1. Cohen's kappa statistics

Cohen's kappa statistics check if the models find the same profitable customer. Cohen's kappa measures the agreement between two evaluators who classify N items into C mutually exclusive categories. In this work, $N = 5$ models and $C = 2$ since the threshold is given by the uplift greater than zero.

This indicator calculates an index with possible values ranging from −1.00 to 1.00, where values less than 0.00 indicate poor agreement, between 0.00 and 0.20 indicate slight agreement, between 0.21 and 0.40 indicate fair agreement, between 0.41 and 0.60 indicate moderate agreement, between 0.61 and 0.80 indicate substantial agreement, and between 0.81 and 1.00 indicate almost perfect agreement.

A significant aspect of the results obtained concerns the agreement between the proposed score and rule-based models and between the proposed models and the traditional model, the lift. The results obtained with Cohen's kappa indicator, presented in Table 10, show 'substantial agreement' between the four models, indicating that, despite using different machine learning models, the majority of profitable customers in all models are consistent.

7.2. Qini index

The Qini index (or coefficient) is essentially a variant of the Gini coefficient, explicitly tailored for uplift models. While the Gini coefficient is estimated from a conventional gain curve (where the Y-axis represents the number of responses), the Qini coefficient is derived from the uplift curve (where the Y-axis shows incremental gain). The Qini coefficient is simply a specialized form of Gini, with the primary distinction being that Qini measures the AUUC (area under the uplift curve). In contrast, Gini is a more general measure of AUC (area under the curve).

The incremental gain curve (also known as the Qini curve) is extremely valuable when comparing models. Unlike traditional gain charts, which focus on the response rate, the uplift curve is often represented on a graph of predicted incremental sales based on the number of individuals treated [31].

The observations are grouped into deciles ordered by the uplift value to compare the various models. The Gini values are calculated for treatment and control groups using the trapezoidal approximation to compute the area under a continuous function. The values presented in Table 11 show that all models achieve an index greater than 0.5, outperforming random targeting. LDA outperforms the other models. However, when it comes to retrieving the rules of the segments, the decision tree CTree performs better.

Table 11
Qini index.

Model	Qini
Linear Discriminant Analysis (LDA)	0.739
Logistic Regression (LogReg)	0.699
Decision tree C5.0	0.625
Decision tree CTree	0.644

Table 12
Cluster purity of the models.

Model	Purity-g	Purity-pc
Linear Discriminant Analysis (LDA)	0.663	0.692
Logistic Regression (LogReg)	0.667	0.698
Decision tree C5.0	0.691	0.719
Decision tree CTree	0.702	0.718

7.3. Strata similarity

To assess the homogeneity of strata, we use the clustering purity measure. Purity ranges from 0 to 1, with higher purity indicating more homogeneous clusters. The following expression gives the purity of a cluster with an attribute, shown in Eq. (11):

$$Purity = \frac{1}{N} \sum_{k} max \left| C_k \cap T_j \right| \tag{11}$$

Where N is the total number of data points, C_k is the set of data points in cluster k, T_j is the set of data points belonging to class j, and $|C_k \cap T_j|$ gives the number of data points in cluster k and class j.

The average cluster purity of each attribute gives the overall purity of a cluster with multiple attributes. Finally, the purity of a dataset is determined by the weighted average of the clusters, considering all attributes.

Table 12 displays the purity of the strata (or clusters), indicating the homogeneity of each stratum within the model. Purity-g refers to the general purity of the dataset, and Purity-pc refers to the purity only for profitable customers with an uplift greater than zero.

Rule-based models, such as decision trees, demonstrate slightly higher purity than score-based models, including LDA and logistic regression. Specifically, for profitable customers, the purity values exceed 0.71 in rule-based models, reflecting greater homogeneity and significant adaptability in the decision tree rules.

7.4. Summary

The overall performance of the models reveals that Cohen's kappa indicates substantial agreement among the four models. This suggests that, despite employing different machine learning approaches, the customers identified as most profitable remain consistent across models. Additionally, the Qini index for all models exceeds 0.5, indicating that each model outperforms random targeting.

In comparing score-based and rule-based models, rule-based models exhibit greater homogeneity and notable adaptability within the decision tree rules. Moreover, decision tree rules are generally more transparent, interpretable, and consistent than those generated by score-based models.

Unlike a numerical score, the clarity and interpretability of a rule can be illustrated with examples such as: (Tenure > 22 and MonthlyCharges < = 71). Moreover, some rules involve thousands of customers with a high uplift, representing a company's fundamental knowledge.

Extracting simple, interpretable rules with decision trees, which identify thousands of profitable customers, reveals essential organizational insights.

8. Conclusions

The study focuses on uplift modeling, identifying persuadable customers who respond exclusively when contacted. Uplift modeling alters the approach to churn modeling, shifting the focus from merely predicting whether a customer will stay or leave. Instead, the uplift modeling aims to identify high-value customers by focusing on customer groups that respond solely when contacted. Each customer can be classified into one of four quadrants: persuadable customers, always buyers, never buyers, and do-not-disturb. Persuadable customers are those who respond if and only if they are contacted.

Uplift modeling is clarified in [10] as addressing the knowledge gap, with a double perspective of 'what we know' and 'what we want to know'. This duality underscores the inherent complexity of uplift modeling, where the goal is not merely to predict customer behavior, as in traditional models, but to discern the causal impact of specific actions on persuadable individuals.

To overcome the knowledge gap, Jaskowski and Jaroszewicz [13] propose extracting persuadable customers by creating a new variable, Y*. The transformation of the variables Y and T into Y* is known as the Transformed Outcome Approach (TOA).

This work makes three main contributions to uplift modeling:

- Comprehensive and modular framework: proposes a procedure for uplift modeling that is straightforward to implement in organizations using machine learning techniques.
- Stratification using decision trees: explores decision trees as a tool for forming meaningful segments.
- Improved interpretability: associates each stratum to a specific decision rule, providing a stratum-specific average treatment effect (SATE).

In implementing the framework, our goal is to ensure transparency in the analytical process, enable full customization, and minimize dependency issues. To achieve this, we develop our code built on widely used standard libraries.

To clarify some causality concepts given the variables triple (Y, T, X), the following notation is used:

- $Y \in \{1,0\}$ to indicate the potential outcome of the treatment;
- $T \in \{1,0\}$ indicates whether an individual is treated or not;
- X corresponds to the covariates of the dataset;
- $Y^* \in \{1,0\}$ is the new potential outcome, where $Y^* = 1$ is a persuadable customer and $Y^* = 0$ is a defier or sleeping dog.

The two-phase procedure previously redefined the target, $Y^* = f0(Y, T)$, reinforcing the idea that customers respond to a campaign if and only if they are contacted. The first phase involves stratification, utilizing machine learning algorithms, where $Y^* = f1(X)$. Stratification mitigates the effects of bias by finding segments with uniform characteristics. Then, the uplift is estimated for each stratum, $Uplift(s) = f2(Y, T, S = s)$, and customers with an uplift greater than zero are identified as profitable.

Different machine learning techniques generate diverse types of strata. Scoring models approaches return iso-cardinality segments, while rule-based algorithms, such as decision trees, produce hetero-cardinality customer strata. The bar charts display distinct shapes; iso-cardinality charts present a set of ranked segments with the same bin size, while hetero-cardinality charts show non-ranked segments with variable bin sizes.

The core idea of the procedure is to segment the customer base into groups of homogeneous persuadable individuals. Each segment includes both treated and untreated customers, enabling accurate uplift estimation. This uplift-driven segmentation ultimately identifies the most profitable customers, transforming persuadable customers into profitable ones.

The models' performance indicates substantial agreement across four approaches, with Cohen's kappa consistently identifying profitable customers. The Qini index above 0.5 for all models confirms they outperform random targeting. Score-based models, especially decision trees, demonstrate greater homogeneity, adaptability, and transparency, making decision trees the best choice due to their interpretability and rule-based clarity.

In this work, we clarify uplift modeling techniques and analyze their computational results, providing a comprehensive and practical method for identifying profitable customers and bridging the gap between theory and implementation. Moreover, by applying decision trees for segmentation, organizations can derive interpretable rules affecting thousands of profitable customers, offering critical business insights.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The first author, Paulo Pinheiro, Ph.D., is the Chief of Software Development at CEDIS, Portugal.

The second and corresponding author, Luís Cavique, is a tenured Professor at the Universidade Aberta and a senior researcher at LASIGE, Portugal.

Acknowledgments

This work was supported by the LASIGE Research Unit, reference UID/00408/2025 – LASIGE.

Data availability

Data is public.

References

[1] D. Olaya, K. Coussement, W. Verbeke, A survey and benchmarking study of multitreatment uplift modeling, Data Min. Knowl. Discov. 34 (2020) 273–308, http://dx.doi.org/10.1007/s10618-019-00670-y.

[2] R. Gubela, A. Bequé, F. Gebert, S. Lessmann, Conversion uplift in e-commerce: A systematic benchmark of modeling strategies, Int. Res. Train. Group 1792 (2018) 2018–062.

[3] E. Ascarza, Retention futility: Targeting high-risk customers might be ineffective, J. Mark. Res. 55 (1) (2018) 80–98, http://dx.doi.org/10.1509/jmr.16.0163.

[4] R. Shrestha, B.K. Kadel, R. Mishra, A two-phase confirmatory factor analysis and structural equation modelling for customer-based brand equity framework in the smartphone industry, Decis. Anal. J. 8 (2023) 100306, http://dx.doi.org/10.1016/j.dajour.2023.100306.

[5] F. Akhavan, E. Hassannayebi, A hybrid machine learning with process analytics for predicting customer experience in the online insurance services industry, Decis. Anal. J. 11 (2024) 100452, http://dx.doi.org/10.1016/j.dajour.2024.100452.

[6] B. Rajendran, V. Babu, M. Anandhabalaji, A predictive modelling approach to decoding consumer intention for adopting energy-efficient technologies in food supply chains, Decis. Anal. J. 15 (2025) 100561, http://dx.doi.org/10.1016/j.dajour.2025.100561.

[7] R. Peto, M. Pike, P. Armitage, et al., Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design, Br. J. Cancer 34 (1976) 585–612, http://dx.doi.org/10.1038/bjc.1976.220.

[8] T.S. Jaffery, S. Liu, Measuring Campaign Performance By using Cumulative Gain and Lift Chart, in: paper 196, SAS Global Forum, 2009.

[9] W. Verbeke, C. Baesens, B. Bravo, Profit Driven Business Analytics, a Practitioners's Guide To Transforming Big Data Into Added Value, Wiley, 2018, ISBN-13: 978-1119286554.

[10] F. Devriendt, D. Moldovan, W. Verbeke, A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping-stone toward the development of prescriptive analytics, Big Data 6 (1) (2018) 13–41, http://dx.doi.org/10.1089/big.2017.0104.

[11] E. Siegel, Predictive Analytics: The Power To Predict Who Will Click, Buy, Lie, Or Die (Revised), Wiley, 2016, ISBN-13: 978-1118356852.
[12] P. Pinheiro, L. Cavique, Uplift modeling using the transformed outcome approach, in: G. Marreiros, B. Martins, A. Paiva, B. Ribeiro, A. Sardinha (Eds.), Progress in Artificial Intelligence, EPIA 2022, in: Lecture Notes in Computer Science, vol. 13566, Springer, Cham, 2022, http://dx.doi.org/10.1007/978-3-031-16474-3_51.
[13] M. Jaskowski, S. Jaroszewicz, Uplift modeling for clinical trial data, in: ICML 2012 Workshop on Clinical Data Analysis, 2012.
[14] S. Athey, G.W. Imbens, Machine Learning Methods for Estimating Heterogeneous Causal Effects, 2015.
[15] P. Gutierrez, J.-Y. Gérardy, Causal inference and uplift modeling: A review of the literature, Jmlr 67 (2016) 1–13.
[16] W. Zhang, J. Li, L. Liu, A unified survey of treatment effect heterogeneity modelling and uplift modelling, ACM Comput. Surv. 54 (8) (2022) 1–36, http://dx.doi.org/10.1145/3466818.
[17] G.W. Imbens, D.B. Rubin, Causal Inference in Statistics, Social, and Biomedical Sciences, Cambridge university press, 2015.
[18] D.B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, J. Educ. Psychol. 66 (5) (1974) 688–701, http://dx.doi.org/10.1037/h0037350.
[19] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55, http://dx.doi.org/10.2307/2335942.
[20] P.W. Holl, Statistics and causal inference, J. Amer. Statist. Assoc. 81 (396) (1986) 945–960.
[21] S. Cunningham, Causal Inference: The Mixtape, Yale University Press, 2021, ISBN-13: 978–0300251685.
[22] S.R. Künzel, J.S. Sekhon, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, Proc. Natl. Acad. Sci. USA 116 (10) (2019) 4156–4165, http://dx.doi.org/10.1073/pnas.1804597116.
[23] E. Bareinboim, Causal artificial intelligence: a roadmap for building casually intelligent systems, 2025, https://causalai-book.net.
[24] J.R. Zubizarreta, E.A. Stuart, D.S. Small, P.R. Rosenbaum (Eds.), Handbook of matching and weighting adjustments for causal inference, in: Handbooks of Modern Statistical Methods, Chapman & Hall/CRC, 2023, ISBN-13: 9780367609528.
[25] G.W. Imbens, J.D. Angrist, Identification and estimation of local average treatment effects, Econometrica 62 (2) (1994) 467–475, http://dx.doi.org/10.2307/2951620.
[26] J.D. Angrist, J.-S. Pischke, Mastering Metrics: The Path from Cause To Effect, Princeton University Press, 2014, ISBN-13: 978-0691152844.
[27] P.C. Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, Multivar. Behav. Res. 46 (3) (2011) 399–424, http://dx.doi.org/10.1080/00273171.2011.568786.
[28] P.R. Rosenbaum, D.B. Rubin, Reducing bias in observational studies using subclassification on the propensity score, J. Amer. Statist. Assoc. 79 (387) (1984) 516–524, http://dx.doi.org/10.2307/2288398.
[29] K. Kane, V.S.Y. Lo, Zheng J., True-lift modeling: Comparison of methods, J. Mark. Anal. 2 (2014) 218–238.
[30] S. Jaroszewicz, Uplift modeling, in: D. Phung, G.I. Webb, C. Sammut (Eds.), Encyclopedia of Machine Learning and Data Science, Springer, New York, NY, 2023, http://dx.doi.org/10.1007/978-1-4899-7502-7_911-2.
[31] N.J. Radcliffe, Using control groups to target on predicted lift: Building and assessing uplift models, Direct Mark. J. Direct Mark. Assoc. Anal. Counc. 1 (2007) 14–21.
[32] F. Devriendt, W. Van Belle, J. Verbeke, Learning to rank for uplift modeling, IEEE Trans. Knowl. Data Eng. 34 (10) (2022) 4888–4904, http://dx.doi.org/10.1109/TKDE.2020.3048510.
[33] V.S.Y. Lo, The true lift model: A novel data mining approach to response modeling in database marketing, in: ACM SIGKDD Explorations Newsletter, vol. 4, (2) 2002.
[34] Y. Zhao, D. Fang, X. Simchi-Levi, Uplift modeling with multiple treatments and general response types, in: Proceedings of the 2017 SIAM International Conference on Data Mining, 2017, pp. 588–596, http://dx.doi.org/10.1137/1.9781611974973.66.
[35] N.P. Suh, Axiomatic Design: Advances and Applications, Oxford University Press, ISBN: 019-513466-4, 2001.
[36] L. Cavique, A.B. M.S. Cavique, M. Mendes, Improving Information System Design: Using UML and Axiomatic Design, Computers in Industry, Cavique Elsevier, 2022, http://dx.doi.org/10.1016/j.compind.2021.103569.
[37] S.D. Eppinger, T.R. Browning, Design Structure Matrix Methods and Applications, The MIT Press, 2016, ISBN-13: 978-0262528887.
[38] A. Venkatasubramaniam, J. Wolfson, N. Mitchell, et al., Decision trees in epidemiological research, in: Emerging Themes in Epidemiology, vol. 14, 2017, p. 11, http://dx.doi.org/10.1186/s12982-017-0064-4.
[39] BlastChar, Telco Customer Churn, Kaggle Repository, 2018, www.kaggle.com/datasets/blastchar/telco-customer-churn.
[40] A. Zanga, F. Ozkirimli, E. Stella, A survey on causal discovery: Theory and practice, Internat. J. Approx. Reason. (ISSN: 0888-613X) 151 (2022) 101–129, http://dx.doi.org/10.1016/j.ijar.2022.09.004.
[41] P. Spirtes, C.N. Glymour, R. Scheines, Causation, Prediction, and Search, MIT Press, 2000, ISBN-13: 978–1461276500.
[42] L. Cao, Domain-driven data mining: challenges and prospects, IEEE Trans. Knowl. Data Eng. 22 (6) (2010) 755–769, http://dx.doi.org/10.1109/TKDE.2010.32.
[43] J. Pearl, N. Glymour, M. Jewell, Causal Inference in Statistics: A Primer, Wiley, 2016.
[44] B. Ripley, B. Venables, D.M. Bates, K. Homik, A. Gebhardt, D. Firth, Package 'MASS', 2023.
[45] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.
[46] M. Kuhn, S. Weston, M. Culp, N. Coulter, R. Quinlan, C5.0 Classification Models, 2020.
[47] T. Hothorn, H. Seibold, A. Zeileis, Package partykit, 2021.